ADAPTIVE COVARIANCE MATRIX ESTIMATION THROUGH BLOCK THRESHOLDING

BY T. TONY CAI[1] AND MING YUAN[2]

University of Pennsylvania and Georgia Institute of Technology

Estimation of large covariance matrices has drawn considerable recent attention, and the theoretical focus so far has mainly been on developing a minimax theory over a fixed parameter space. In this paper, we consider adaptive covariance matrix estimation where the goal is to construct a single procedure which is minimax rate optimal simultaneously over each parameter space in a large collection. A fully data-driven block thresholding estimator is proposed. The estimator is constructed by carefully dividing the sample covariance matrix into blocks and then simultaneously estimating the entries in a block by thresholding. The estimator is shown to be optimally rate adaptive over a wide range of bandable covariance matrices. A simulation study is carried out and shows that the block thresholding estimator performs well numerically. Some of the technical tools developed in this paper can also be of independent interest.

1. Introduction. Covariance matrix estimation is of fundamental importance in multivariate analysis. Driven by a wide range of applications in science and engineering, the high-dimensional setting, where the dimension p can be much larger than the sample size n, is of particular current interest. In such a setting, conventional methods and results based on fixed p and large n are no longer applicable, and in particular, the commonly used sample covariance matrix and normal maximum likelihood estimate perform poorly.

A number of regularization methods, including banding, tapering, thresholding and ℓ_1 minimization, have been developed in recent years for estimating a large covariance matrix or its inverse. See, for example, Ledoit and Wolf (2004), Huang et al. (2006), Yuan and Lin (2007), Banerjee, El Ghaoui and d'Aspremont (2008), Bickel and Levina (2008a, 2008b), El Karoui (2008), Fan, Fan and Lv (2008), Friedman, Hastie and Tibshirani (2008), Rocha, Zhao and Yu (2008), Rothman et al. (2008), Lam and Fan (2009), Rothman, Levina and Zhu (2009), Cai, Zhang and Zhou (2010), Yuan (2010), Cai and Liu (2011) and Cai, Liu and Luo (2011), among many others.

Let $X^{(1)}, \ldots, X^{(n)}$ be n independent copies of a p dimensional Gaussian random vector $X = (X_1, \ldots, X_p)^\mathsf{T} \sim N(\mu, \Sigma)$. The goal is to estimate the covariance matrix Σ and its inverse Σ^{-1} based on the sample $\{X^{(i)} : i = 1, \ldots, n\}$. It is now well known that the usual sample covariance matrix

$$\bar{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^\mathsf{T},$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$, is not a consistent estimator of the covariance matrix Σ when $p \gg n$, and structural assumptions are required in order to estimate Σ consistently.

One of the most commonly considered classes of covariance matrices is the "bandable" matrices, where the entries of the matrix decay as they move away from the diagonal. More specifically, consider the following class of covariance matrices introduced in Bickel and Levina (2008a):

$$\mathcal{C}_\alpha = \mathcal{C}_\alpha(M_0, M) := \left\{ \Sigma : \max_j \sum_i \{ |\sigma_{ij}| : |i - j| \geq k \} \leq M k^{-\alpha} \; \forall k, \right.$$

(1.1)

$$\left. \text{and } 0 < M_0^{-1} \leq \lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma) \leq M_0 \right\}.$$

Such a family of covariance matrices naturally arises in a number of settings, including temporal or spatial data analysis. See Bickel and Levina (2008a) for further discussions. Several regularization methods have been introduced for estimating a bandable covariance matrix $\Sigma \in \mathcal{C}_\alpha$. Bickel and Levina (2008a) suggested banding the sample covariance matrix $\bar{\Sigma}$ and estimating Σ by $\bar{\Sigma} \circ B_k$ where B_k is a banding matrix

$$B_k = \big(\mathbb{I}(|i - j| \leq k) \big)_{1 \leq i, j \leq p}$$

and $\circ$ represents the Schur product, that is, $(A \circ B)_{ij} = A_{ij} B_{ij}$ for two matrices of the same dimensions. See Figure 1(a) for an illustration. Bickel and Levina (2008a) proposed to choose $k \asymp (n/\log p)^{1/(2(\alpha+1))}$ and showed that the resulting banding estimator attains the rate of convergence

(1.2)
$$\|\bar{\Sigma} \circ B_k - \Sigma\| = O_p\left(\left(\frac{\log p}{n} \right)^{\alpha/(2\alpha+2)} \right)$$

uniformly over $\mathcal{C}_\alpha$, where $\| \cdot \|$ stands for the spectral norm. This result indicates that even when $p \gg n$, it is still possible to consistently estimate $\Sigma \in \mathcal{C}_\alpha$, so long as $\log p = o(n)$.

Cai, Zhang and Zhou (2010) established the minimax rate of convergence for estimation over $\mathcal{C}_\alpha$ and introduced a tapering estimator $\bar{\Sigma} \circ T_k$ where the tapering matrix T_k is given by

$$T_k = \left(\frac{2}{k} \{ (k - |i - j|)_+ - (k/2 - |i - j|)_+ \} \right)_{1 \leq i, j \leq p},$$
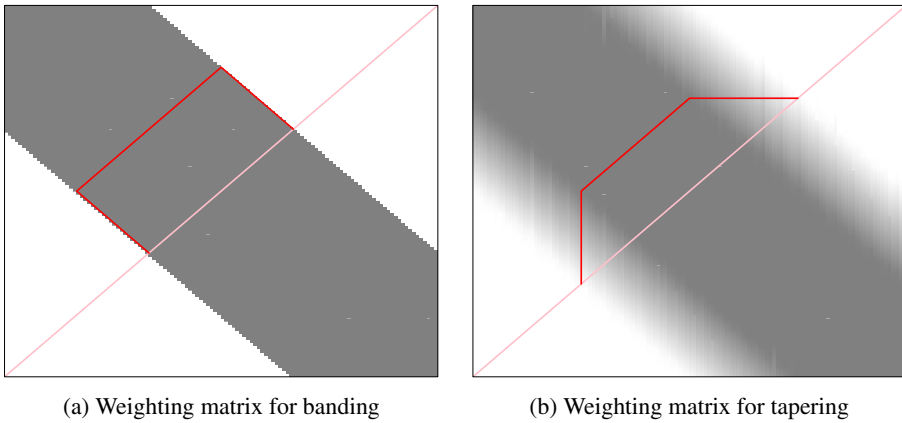
(a) Weighting matrix for banding   (b) Weighting matrix for tapering

FIG. 1.   *Both banding and tapering estimators can be expressed as the Schur product of the sample covariance matrix and a weighting matrix. Subfigures of* (a) *and* (b) *illustrate the weighting matrix for both estimators.*

with $(x)_+ = \max(x, 0)$. See Figure 1(b) for an illustration. It was shown that the tapering estimator $\bar{\Sigma} \circ T_k$ with $k \asymp n^{1/(2\alpha+1)}$ achieves the rate of convergence

$$(1.3) \qquad \|\bar{\Sigma} \circ T_k - \Sigma\| = O_p\left(n^{-\alpha/(2\alpha+1)} + \left(\frac{\log p}{n}\right)^{1/2}\right)$$

uniformly over $\mathcal{C}_\alpha$, which is always faster than the rate in (1.2). This implies that the rate of convergence given in (1.2) for the banding estimator with $k \asymp (n/\log p)^{1/(2(\alpha+1))}$ is in fact sub-optimal. Furthermore, a lower bound argument was given in Cai, Zhang and Zhou (2010) which showed that the rate of convergence given in (1.3) is indeed optimal for estimating the covariance matrices over $\mathcal{C}_\alpha$.

The minimax rate of convergence in (1.3) provides an important benchmark for the evaluation of the performance of covariance matrix estimators. It is, however, evident from its construction that the rate optimal tapering estimator constructed in Cai, Zhang and Zhou (2010) requires explicit knowledge of the decay rate α which is typically unknown in practice. It is also clear that a tapering estimator designed for a parameter space with a given decay rate α performs poorly over another parameter space with a different decay rate. The tapering estimator mentioned above is thus not very practical.

This naturally leads to the important question of adaptive estimation: Is it possible to construct a single estimator, not depending on the decay rate α, that achieves the optimal rate of convergence simultaneously over a wide range of the parameter spaces $\mathcal{C}_\alpha$? We shall show in this paper that the answer is affirmative. A fully data-driven adaptive estimator $\hat{\Sigma}$ is constructed and is shown to be simultaneously

rate optimal over the collection of the parameter spaces $\mathcal{C}_\alpha$ for all $\alpha > 0$. That is,

$$\sup_{\Sigma \in \mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \asymp \min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\} \qquad \text{for all } \alpha > 0.$$

In many applications, the inverse covariance matrix is of significant interest. We introduce a slightly modified version of $\hat{\Sigma}^{-1}$ and show that it adaptively attains the optimal rate of convergence for estimating Σ^{-1}.

The adaptive covariance matrix estimator achieves its adaptivity through block thresholding of the sample covariance matrix $\bar{\Sigma}$. The idea of adaptive estimation via block thresholding can be traced back to nonparametric function estimation using Fourier or wavelet series. See, for example, Efromovich (1985) and Cai (1999). However, the application of block thresholding to covariance matrix estimation poses new challenges. One of the main difficulties in dealing with covariance matrix estimation, as opposed to function estimation or sequence estimation problems, is the fact that the spectral norm is not separable in its entries. Another practical challenge is due to the fact that the covariance matrix is "two-directional" where one direction is along the rows and another along the columns. The blocks of different sizes need to be carefully constructed so that they fit well in the sample covariance matrix and the risk can be assessed based on their joint effects rather than their individual contributions. There are two main steps in the construction of the adaptive covariance matrix estimator. The first step is the construction of the blocks. Once the blocks are constructed, the second step is to estimate the entries of the covariance matrix Σ in groups and make simultaneous decisions on all the entries within a block. This is done by thresholding the sample covariance matrix block by block. The threshold level is determined by the location, block size and corresponding spectral norms. The detailed construction is given in Section 2.

We shall show that the proposed block thresholding estimator $\hat{\Sigma}$ is simultaneously rate-optimal over every $\mathcal{C}_\alpha$ for all $\alpha > 0$. The theoretical analysis of the estimator $\hat{\Sigma}$ requires some new technical tools that can be of independent interest. One is a concentration inequality which shows that although the sample covariance matrix $\bar{\Sigma}$ is not a reliable estimator of Σ, its submatrices could still be a good estimate of the corresponding submatrices of Σ. Another useful tool is a so-called norm compression inequality which reduces the analysis on the whole matrix to a matrix of much smaller dimensions, whose entries are the spectral norms of the blocks.

In addition to the analysis of the theoretical properties of the proposed adaptive block thresholding estimator, a simulation study is carried out to investigate the finite sample performance of the estimator. The simulations show that the proposed estimator enjoys good numerical performance when compared with nonadaptive estimators such as the banding and tapering estimators.

Besides bandable matrices considered in the present paper, estimating sparse covariance matrices and sparse precision matrices has also been actively studied in

the recent literature. Bickel and Levina (2008b) proposed a thresholding estimator for sparse covariance matrices and obtained the rate of convergence. Cai and Zhou (2011) developed a new general minimax lower bound technique and established the minimax rate of convergence for estimating sparse covariance matrices under the spectral norm and other matrix operator norms. Cai and Liu (2011) introduced an adaptive thresholding procedure for estimating sparse covariance matrices that automatically adjusts to the variability of individual entries. Estimation of sparse precision matrices has also drawn considerable attention due to its close connections to Gaussian graphical model selection. See Yuan and Lin (2007), Yuan (2010), Ravikumar et al. (2011) and Cai, Liu and Luo (2011). The optimal rate of convergence for estimating sparse inverse covariance matrices was established in Cai, Liu and Zhou (2011).

The rest of the paper is organized as follows. Section 2 presents a detailed construction of the data-driven block thresholding estimator $\check{\Sigma}$. The theoretical properties of the estimator are investigated in Section 3. It is shown that the estimator $\hat{\Sigma}$ achieves the optimal rate of convergence simultaneously over each $\mathcal{C}_\alpha(M_0, M)$ for all $\alpha, M_0, M > 0$. In addition, it is also shown that a slightly modified version of $\hat{\Sigma}^{-1}$ is adaptively rate-optimal for estimating Σ^{-1} over the collection $\mathcal{C}_\alpha(M_0, M)$. Simulation studies are carried out to illustrate the merits of the proposed method, and the numerical results are presented in Section 4. Section 5 discusses extension to subguassian noise, adaptive estimation under the Frobenius norm and other related issues. The proofs of the main results are given in Section 6.

2. Block thresholding. In this section we present in detail the construction of the adaptive covariance matrix estimator. The main strategy in the construction is to divide the sample covariance matrix into blocks and then apply thresholding to each block according to their sizes and dimensions. We shall explain these two steps separately in Sections 2.1 and 2.2.

2.1. *Construction of blocks.* As mentioned in the Introduction, the application of block thresholding to covariance matrix estimation requires more care than in the conventional sequence estimation problems such as those from nonparametric function estimation. We begin with the blocking scheme for a general $p \times p$ symmetric matrix. A key in our construction is to make blocks larger for entries that are farther away from the diagonal and take advantage of the approximately banding structure of the covariance matrices in $\mathcal{C}_\alpha$. Before we give a precise description of the construction of the blocks, it is helpful to graphically illustrate the construction in the following plot.

Due to the symmetry, we shall focus only on the upper half for brevity. We start by constructing blocks of size $k_0 \times k_0$ along the diagonal as indicated by the darkest squares in Figure 2. Note that the last block may be of a smaller size if k_0 is not a divisor of p. Next, new blocks are created successively toward the top right corner. We would like to increase the block sizes along the way. To this end,
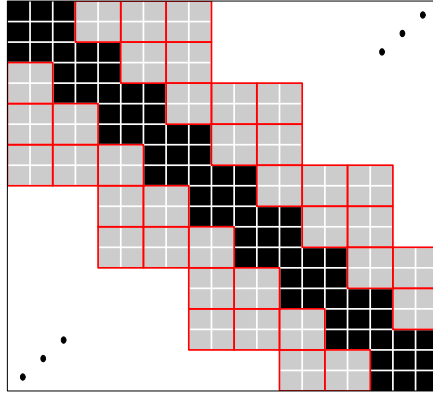
FIG. 2. *Construction of blocks with increasing dimensions away from the diagonal. The solid black blocks are of size $k_0 \times k_0$. The gray ones are of size $2k_0 \times 2k_0$.*

we extend to the right from the diagonal blocks by either two or one block of the same dimensions ($k_0 \times k_0$) in an alternating fashion. After this step, as exhibited in Figure 2, the odd rows of blocks will have three $k_0 \times k_0$ blocks, and the even rows will have two $k_0 \times k_0$ in the upper half. Next, the size of new blocks is doubled to $2k_0 \times 2k_0$. Similarly to before, the last block may be of smaller size if $2k_0$ is not a divisor of p, and for the most part, we shall neglect such a caveat hereafter for brevity. The same procedure is then followed. We extend to the right again by three or two blocks of the size $2k_0 \times 2k_0$. Afterwards, the block size is again enlarged to $2^2k_0 \times 2^2k_0$ and we extend to the right by three or two blocks of size $2^2k_0 \times 2^2k_0$. This procedure will continue until the whole upper half of the $p \times p$ matrix is covered. For the lower half, the same construction is followed to yield a symmetric blocking of the whole matrix.

The initial block size k_0 can take any value as long as $k_0 \asymp \log p$. In particular, we can take $k_0 = \lfloor \log p \rfloor$. The specific choice of k_0 does not impact the rate of convergence, but in practice it may be beneficial sometimes to use a value different from $\lfloor \log p \rfloor$. In what follows, we shall keep using k_0 for the sake of generality.

For notational purposes, hereafter we shall refer to the collection of index sets for the blocks created in this fashion as $\mathcal{B} = \{B_1, \ldots, B_N\}$ where $B_k = I_k \times J_k$ for some subintervals $I_k, J_k \subset \{1, \ldots, p\}$. It is clear that $\mathcal{B}$ forms a partition of $\{1, 2, \ldots, p\}^2$, that is,

$$B_{k_1} \cap B_{k_2} = \varnothing \quad \text{if } k_1 \neq k_2 \quad \text{and} \quad B_1 \cup B_2 \cup \cdots \cup B_N = \{1, 2, \ldots, p\}^2.$$

For a $p \times p$ matrix $A = (a_{ij})_{1 \leq i, j \leq p}$ and an index set $B = I \times J \in \mathcal{B}$, we shall also write $A_B = (a_{ij})_{i \in I, j \in J}$, a $|I| \times |J|$ submatrix of A. Hence A is uniquely determined by $\{A_B : B \in \mathcal{B}\}$ and the partition $\mathcal{B}$. With slight abuse of notation, we shall also refer to an index set B as a block when no confusion occurs, for the sake of brevity.

Denote by $d(B)$ the dimension of B, that is,

$$d(B) = \max\{\operatorname{card}(I), \operatorname{card}(J)\}.$$

Clearly by construction, most of the blocks in $\mathcal{B}$ are necessarily square in that $\operatorname{card}(I) = \operatorname{card}(J) = d(B)$. The exceptions occur when the block sizes are not divisors of p, which leaves the blocks along the last row and column in rectangles rather than squares. We opt for the more general definition of $d(B)$ to account for these rectangle blocks.

2.2. *Block thresholding.* Once the blocks are constructed, the next step is to estimate the entries of the covariance matrix Σ, block by block, through thresholding the corresponding blocks of the sample covariance matrix based on the location, block size and corresponding spectral norms.

We now describe the procedure in detail. Denote by $\hat{\Sigma}$ the block thresholding estimator, and let $B = I \times J \in \mathcal{B}$. The estimate of the block Σ_B is defined as follows:

(a) *keep the diagonal blocks*: $\hat{\Sigma}_B = \bar{\Sigma}_B$ if B is on the diagonal, that is, $I = J$;
(b) *"kill" the large blocks*: $\hat{\Sigma}_B = \mathbf{0}$ if $d(B) > n/\log n$;
(c) *threshold the intermediate blocks*: For all other blocks B, set

$$(2.1) \quad \hat{\Sigma}_B = T_{\lambda_0}(\bar{\Sigma}_B) = \bar{\Sigma}_B \cdot \mathbb{I}\left(\|\bar{\Sigma}_B\| > \lambda_0 \sqrt{\|\bar{\Sigma}_{I \times I}\| \|\bar{\Sigma}_{J \times J}\|} \sqrt{\frac{d(B) + \log p}{n}}\right),$$

where $\lambda_0 > 0$ is a turning parameter. Our theoretical development indicates that the resulting block thresholding estimator is optimally rate adaptive whenever λ_0 is a sufficiently large constant. In particular, it can be taken as fixed at $\lambda_0 = 6$. In practice, a data-driven choice of λ_0 could potentially lead to further improved finite sample performance.

It is clear from the construction that the block thresholding estimate $\hat{\Sigma}$ is fully data-driven and does not require the knowledge of α. The choice of the thresholding constant λ_0 comes from our theoretical and numerical studies. See Section 5 for more discussions on the choice of λ_0.

We should also note that, instead of the hard thresholding operator T_{λ_0}, more general thresholding rules can also be applied in a similar blockwise fashion. In particular, one can use block thresholding rules $T_{\lambda_0}(\bar{\Sigma}_B) = \bar{\Sigma}_B \cdot t_{\lambda_B}(\|\bar{\Sigma}_B\|)$ where

$$\lambda_B = \lambda_0 \sqrt{\|\bar{\Sigma}_{I \times I}\| \|\bar{\Sigma}_{J \times J}\|} \sqrt{\frac{d(B) + \log p}{n}},$$

and t_{λ_B} is a univariate thresholding rule. Typical examples include the soft thresholding rule $t_{\lambda_B}(z) = (|z| - \lambda_B)_+ \operatorname{sgn}(z)$ and the so-called adaptive lasso rule $t_{\lambda_B}(z) = z(1 - |\lambda_B/z|^\eta)_+$ for some $\eta \geq 1$, among others. Rothman, Levina and Zhu (2009) considered entrywise universal thresholding for estimating sparse covariance matrix. In particular, they investigate the class of univariate thresholding rules

t_{λ_B} such that (a) $|t_{\lambda_0}(z)| \leq |z|$; (b) $t_{\lambda_B}(z) = 0$ if $|z| \leq \lambda_B$; and (c) $|t_{\lambda_B}(z) - z| \leq \lambda_B$. Although we will focus on the hard thresholding rule in the present paper for brevity, all the theoretical results developed here apply to the more general class of block thresholding rules as well.

3. Adaptivity. We now study the properties of the proposed block thresholding estimator $\hat{\Sigma}$ and show that the estimator simultaneously achieves the minimax optimal rate of convergence over the full range of $\mathcal{C}_\alpha$ for all $\alpha > 0$. More specifically, we have the following result.

THEOREM 3.1. *Let $\hat{\Sigma}$ be the block thresholding estimator of Σ as defined in the Section 2. Then*

$$(3.1) \qquad \sup_{\Sigma \in \mathcal{C}_\alpha(M_0, M)} \mathbb{E} \| \hat{\Sigma} - \Sigma \|^2 \leq C \min \left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n} \right\}$$

for all $\alpha > 0$, where C is a positive constant not depending on n and p.

Comparing with the minimax rate of convergence given in Cai, Zhang and Zhou (2010), this shows that the block thresholding estimator $\hat{\Sigma}$ is optimally rate adaptive over $\mathcal{C}_\alpha$ for all $\alpha > 0$.

REMARK 1. The block thresholding estimator $\hat{\Sigma}$ is positive definite with high probability, but it is not guaranteed to be positive definite. A simple additional step, as was done in Cai and Zhou (2011), can make the final estimator positive semi-definite and still achieve the optimal rate of convergence. Write the eigen-decomposition of $\hat{\Sigma}$ as $\hat{\Sigma} = \sum_{i=1}^{p} \hat{\lambda}_i v_i v_i^T$, where $\hat{\lambda}_i$'s and v_i's are, respectively, the eigenvalues and eigenvectors of $\hat{\Sigma}$. Let $\hat{\lambda}_i^+ = \max(\hat{\lambda}_i, 0)$ be the positive part of $\hat{\lambda}_i$, and define

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

Then $\hat{\Sigma}^+$ is positive semi-definite, and it can be shown easily that $\hat{\Sigma}^+$ attains the same rate as $\hat{\Sigma}$. See Cai and Zhou (2011) for further details. If a strictly positive definite estimator is desired, one can also set $\hat{\lambda}_i^+ = \max(\hat{\lambda}_i, \varepsilon_n)$ for some small positive value ε_n, say $\varepsilon_n = O(\log p / n)$, and the resulting estimator $\hat{\Sigma}^+$ is then positive definite and attains the optimal rate of convergence.

The inverse of the covariance matrix, $\Omega := \Sigma^{-1}$, is of significant interest in many applications. An adaptive estimator of Ω can also be constructed based on our proposed block thresholding estimator. To this end, let $\hat{\Sigma} = \hat{U} \hat{D} \hat{U}^T$ be its

eigen-decomposition, that is, $\hat{U}$ is an orthogonal matrix, and $\hat{D}$ is a diagonal matrix. We propose to estimate Ω by

$$\hat{\Omega} = \hat{U} \operatorname{diag}(\min\{\hat{d}_{ii}^{-1}, n\}) \hat{U}^{\mathsf{T}},$$

where $\hat{d}_{ii}$ is the ith diagonal element of $\hat{D}$. The truncation of $\hat{d}_{ii}^{-1}$ is needed to deal with the case where $\hat{\Sigma}$ is near singular. The result presented above regarding $\hat{\Sigma}$ can be used to show that $\hat{\Omega}$ adaptively achieves the optimal rate of convergence for estimating Ω.

THEOREM 3.2. *Let $\hat{\Omega}$ be defined as above. Then*

$$\sup_{\Sigma \in \mathcal{C}_\alpha} \mathbb{E}\|\hat{\Omega} - \Omega\|^2 \le C \min\left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n} \right\}$$

for all $\alpha > 0$, where $C > 0$ is a constant not depending on n and p.

The proof of the adaptivity results is somewhat involved and requires some new technical tools. The main ideas in the theoretical analysis can be summarized as follows:

- The different $\hat{\Sigma} - \Sigma$ can be decomposed into a sum of matrices such that each matrix in the sum only consists of blocks in $\mathcal{B}$ that are of the same size. The individual components in the sum are then bounded separately according to their block sizes.
- Although the sample covariance matrix $\bar{\Sigma}$ is not a reliable estimator of Σ, its submatrix, $\bar{\Sigma}_B$, could still be a good estimate of Σ_B. This is made precise through a concentration inequality.
- The analysis on the whole matrix is reduced to the analysis of a matrix of much smaller dimensions, whose entries are the spectral norms of the blocks, through the application of a so-called norm compression inequality.
- With high probability, large blocks in $\{\bar{\Sigma}_B : B \in \mathcal{B}\}$, which correspond to negligible parts of the true covariance matrix Σ, are all shrunk to zero because by construction they are necessarily far away from the diagonal.

We shall elaborate below these main ideas in our analysis and introduce some useful technical tools. The detailed proof is relegated to Section 6.

3.1. *Main strategy.* Recall that $\mathcal{B}$ is the collection of blocks created using the procedure in Section 2.1, and it forms a partition of $\{1, 2, \ldots, p\}^2$. We analyze the error $\hat{\Sigma} - \Sigma$ by first decomposing it into a sum of matrices such that each matrix in the sum only consists of blocks in $\mathcal{B}$ that are of the same size. More precisely, for a $p \times p$ matrix A, define $S(A; l)$ to be a $p \times p$ matrix whose (i, j) entry equals
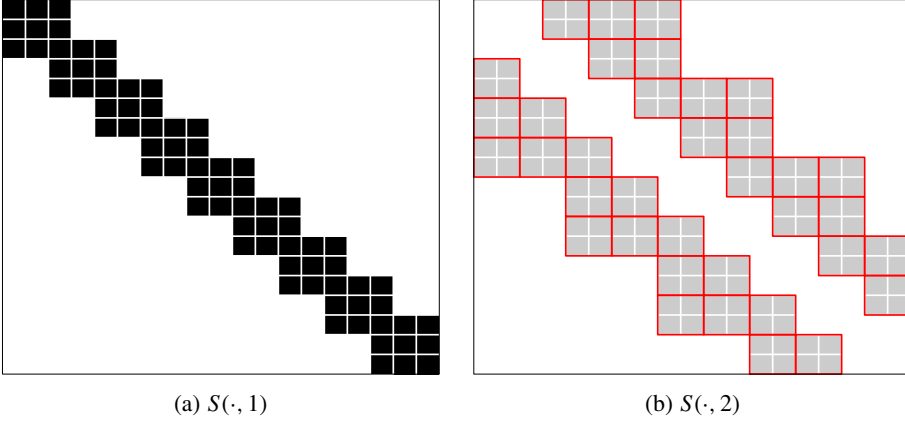
(a) $S(\cdot, 1)$ (b) $S(\cdot, 2)$

FIG. 3. *Decompose a matrix into the sum of matrices of different block sizes*: $S(\cdot, 1)$ *on the left and* $S(\cdot, 2)$ *on the right. All entries in the unshaded area are zero.*

that of A if (i, j) belongs to a block of dimension $2^{l-1}k_0$, and zero otherwise. In other words,

$$S(A, l) = \sum_{B \in \mathcal{B}:d(B)=2^{l-1}k_0} A \circ \mathbb{I}\big((i, j) \in B\big)_{1 \le i, j \le p}.$$

With this notation, $\hat{\Sigma} - \Sigma$ is decomposed as

$$\hat{\Sigma} - \Sigma = S(\hat{\Sigma} - \Sigma, 1) + S(\hat{\Sigma} - \Sigma, 2) + \cdots.$$

This decomposition into the sum of blocks of different sizes is illustrated in Figure 3 below.

We shall first separate the blocks into two groups, one for big blocks and another for small blocks. See Figure 4 for an illustration. By the triangle inequality, for any $L \ge 1$,

$$(3.2) \qquad \|\hat{\Sigma} - \Sigma\| \le \sum_{l \le L} \|S(\hat{\Sigma} - \Sigma, l)\| + \left\| \sum_{l > L} S(\hat{\Sigma} - \Sigma, l) \right\|.$$

The errors on the big blocks will be bounded as a whole, and the errors on the small blocks will be bounded separately according to block sizes. With a careful choice of the cutoff value L, it can be shown that there exists a constant $c > 0$ not depending on n and p such that for any $\alpha > 0$ and $\Sigma \in \mathcal{C}_\alpha$,

$$(3.3) \qquad \mathbb{E}\left(\sum_{l \le L} \|S(\hat{\Sigma} - \Sigma, l)\| \right)^2 = c \min\left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n} \right\},$$

and

$$(3.4) \qquad \mathbb{E}\left\| \sum_{l > L} S(\hat{\Sigma} - \Sigma, l) \right\|^2 = c \min\left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n} \right\},$$
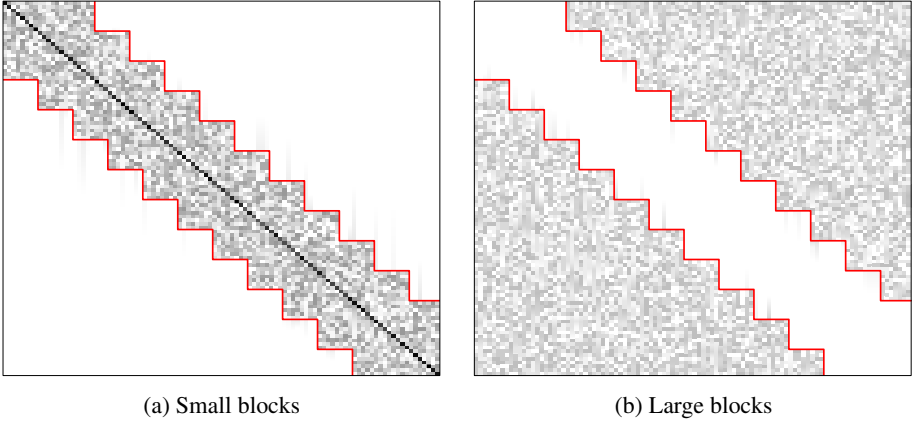
(a) Small blocks　　　　　　　　　　　　(b) Large blocks

FIG. 4.　*Small blocks and large blocks are treated separately. Small blocks are necessarily close to the diagonal and large blocks are away from the diagonal.*

which then implies Theorem 3.1 because

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \le 2\mathbb{E}\left(\sum_{l \le L} \|S(\hat{\Sigma} - \Sigma, l)\|\right)^2 + 2\mathbb{E}\left\|\sum_{l > L} S(\hat{\Sigma} - \Sigma, l)\right\|^2.$$

The choice of the cutoff value L depends on p and n and different approaches are taken to establish (3.3) and (3.4). In both cases, a key technical tool we shall use is a concentration inequality on the deviation of a block of the sample covariance matrix from its counterpart of the true covariance matrix, which we now describe.

3.2. *Concentration inequality.* The rationale behind our block thresholding approach is that although the sample covariance matrix $\bar{\Sigma}$ is not a reliable estimator of Σ, its submatrix, $\bar{\Sigma}_B$, could still be a good estimate of Σ_B. This observation is formalized in the following theorem.

THEOREM 3.3.　*There exists an absolute constant $c_0 > 0$ such that for all $t > 1$,*

$$\mathbb{P}\left(\bigcap_{B = I \times J \in \mathcal{B}} \left\{\|\bar{\Sigma}_B - \Sigma_B\| < c_0 t \sqrt{\|\Sigma_{I \times I}\| \|\Sigma_{J \times J}\|} \sqrt{\frac{d(B) + \log p}{n}}\right\}\right)$$

$$\ge 1 - p^{-(6t^2 - 2)}.$$

In particular, we can take $c_0 = 5.44$.

Theorem 3.3 enables one to bound the estimation error $\hat{\Sigma} - \Sigma$ block by block. Note that larger blocks are necessarily far away from the diagonal by construction.

For bandable matrices, this means that larger blocks are necessarily small in the spectral norm. From Theorem 3.3, if $\lambda_0 > c_0$, with overwhelming probability,

$$\|\bar{\Sigma}_B\| \le \|\Sigma_B\| + c_0\sqrt{\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\sqrt{\frac{d(B)+\log p}{n}}$$

$$< \lambda_0\sqrt{\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\sqrt{\frac{d(B)+\log p}{n}}$$

for blocks with sufficiently large sizes. As we shall show in Section 6, $\|\Sigma_{I\times I}\|$ and $\|\Sigma_{J\times J}\|$ in the above inequality can be replaced by their respective sample counterparts. This observation suggests that larger blocks are shrunken to zero with our proposed block thresholding procedure, which is essential in establishing (3.4).

The treatment of smaller blocks is more complicated. In light of Theorem 3.3, blocks of smaller sizes can be estimated well, that is, $\bar{\Sigma}_B$ is close to Σ_B for B of smaller sizes. To translate the closeness in such a blockwise fashion into the closeness in terms of the whole covariance matrix, we need a simple yet useful result based on a matrix norm compression transform.

3.3. *Norm compression inequality.* We shall now present a so-called norm compression inequality which is particularly useful for analyzing the properties of the block thresholding estimators. We begin by introducing a matrix norm compression transform.

Let A be a $p \times p$ symmetric matrix, and let $p_1, \ldots, p_G$ be positive integers such that $p_1 + \cdots + p_G = p$. The matrix A can then be partitioned in a block form as

$$A = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1G} \\ A_{21} & A_{22} & \ldots & A_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ A_{G1} & A_{G2} & \ldots & A_{GG} \end{pmatrix},$$

where A_{ij} is a $p_i \times p_j$ submatrix. We shall call such a partition of the matrix A a regular partition and the blocks A_{ij} regular blocks. Denote by $\mathcal{N}: \mathbb{R}^{p\times p} \mapsto \mathbb{R}^{G\times G}$ a norm compression transform

$$A \mapsto \mathcal{N}(A; p_1, \ldots, p_G) = \begin{pmatrix} \|A_{11}\| & \|A_{12}\| & \ldots & \|A_{1G}\| \\ \|A_{21}\| & \|A_{22}\| & \ldots & \|A_{2G}\| \\ \vdots & \vdots & \ddots & \vdots \\ \|A_{G1}\| & \|A_{G2}\| & \ldots & \|A_{GG}\| \end{pmatrix}.$$

The following theorem shows that such a norm compression transform does not decrease the matrix norm.

THEOREM 3.4 (Norm compression inequality). *For any $p \times p$ matrix A and block sizes $p_1, p_2, \ldots, p_G$ such that $p_1 + \cdots + p_G = p$,*

$$\|A\| \le \|\mathcal{N}(A; p_1, \ldots, p_G)\|.$$

Together with Theorems 3.3 and 3.4 provides a very useful tool for bounding $S(\hat{\Sigma} - \Sigma, l)$. Note first that Theorem 3.4 only applies to a regular partition, that is, the divisions of the rows and columns are the same. It is clear that $S(\cdot, 1)$ corresponds to regular blocks of size $k_0 \times k_0$ with the possible exception of the last row and column which can be of a different size, that is, $p_1 = p_2 = \cdots = k_0$. Hence, Theorem 3.4 can be directly applied. However, this is no longer the case when $l > 1$.

To take advantage of Theorem 3.4, a new blocking scheme is needed for $S(\cdot, l)$. Consider the case when $l = 2$. It is clear that $S(l, 2)$ does not form a regular blocking. But we can form new blocks with $p_1 = p_2 = \cdots = k_0$, that is, half the size of the original blocks in $S(\cdot, 2)$. Denote by the collection of the new blocks $\mathcal{B}'$. It is clear that under this new blocking scheme, each block B of size $2k_0$ consists of four elements from $\mathcal{B}'$. Thus

$$S(A, 2) = \sum_{\substack{B \in \mathcal{B} \\ d(B)=2k_0}} A \circ \mathbb{I}\big((i, j) \in B\big) = \sum_{\substack{B' \in \mathcal{B}' \\ \exists B \in \mathcal{B} \text{ such that } d(B)=2k_0 \\ \text{and } B' \subset B}} A \circ \mathbb{I}\big((i, j) \in B'\big).$$

Applying Theorem 3.4 to the regular blocks $\mathcal{B}'$ yields

$$\big\| S(A, 2) \big\| \leq \big\| \mathcal{N}\big(S(A, 2); k_0, \ldots, k_0\big) \big\|,$$

which can be further bounded by

$$\big\| \mathcal{N}\big(S(A, 2); k_0, \ldots, k_0\big) \big\|_{\ell_1},$$

where $\| \cdot \|_{\ell_1}$ stands for the matrix ℓ_1 norm. Observe that each row or column of $\mathcal{N}(S(A, 2); k_0, \ldots, k_0)$ has at most 12 nonzero entries, and each entry is bounded by

$$\max_{\substack{B' \in \mathcal{B}' \\ \exists B \in \mathcal{B} \text{ such that } d(B)=2k_0 \\ \text{and } B' \subset B}} \| A_{B'} \| \leq \max_{\substack{B \in \mathcal{B} \\ d(B)=2k_0}} \| A_B \|$$

because $B' \subset B$ implies $\| A_{B'} \| \leq \| A_B \|$. This property suggests that $\| S(\hat{\Sigma} - \Sigma, l) \|$ can be controlled in a block-by-block fashion. This can be done using the concentration inequalities given in Section 3.2.

The case when $l > 2$ can be treated similarly. Let $p_{2j-1} = (2^{l-1} - 3)k_0$ and $p_{2j} = 3k_0$ for $j = 1, 2, \ldots$. It is not hard to see that each block B in $\mathcal{B}$ of size $2^{l-1}k_0$ occupies up to four blocks in this regular blocking. And following the same argument as before, we can derive bounds for $S(A, l)$.

The detailed proofs of Theorems 3.1 and 3.2 are given in Section 6.

4. Numerical results. The block thresholding estimator $\hat{\Sigma}$ proposed in Section 2 is easy to implement. In this section we turn to the numerical performance of the estimator. The simulation study further illustrates the merits of the proposed block thresholding estimator. The performance is relatively insensitive to the choice of k_0, and we shall focus on $k_0 = \lfloor \log p \rfloor$ throughout this section for brevity.

We consider two different sets of covariance matrices. The setting of our first set of numerical experiments is similar to those from Cai, Zhang and Zhou (2010). Specifically, the true covariance matrix Σ is of the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p, \\ \rho|i - j|^{-2}u_{ij}, & 1 \le i \neq j \le p, \end{cases}$$

where the value of ρ is set to be 0.6 to ensure positive definiteness of all covariance matrices, and $u_{ij} = u_{ji}$ are independently sampled from a uniform distribution between 0 and 1.

The second settings are slightly more complicated, and the covariance matrix Σ is randomly generated as follows. We first simulate a symmetric matrix $A = (a_{ij})$ whose diagonal entries are zero and off-diagonal entries a_{ij} ($i < j$) are independently generated as $a_{ij} \sim N(0, |i - j|^{-4})$. Let $\lambda_{\min}(A)$ be its smallest eigenvalue. The covariance matrix Σ is then set to be $\Sigma = \max(0, -1.1\lambda_{\min}(A))I + A$ to ensure its positive definiteness.

For each setting, four different combinations of p and n are considered, $(n, p) = (50, 50)$, $(100, 100)$, $(200, 200)$ and $(400, 400)$, and for each combination, 200 simulated datasets are generated. On each simulated dataset, we apply the proposed block thresholding procedure with $\lambda_0 = 6$. For comparison purposes, we also use the banding estimator of Bickel and Levina (2008a) and tapering estimator of Cai, Zhang and Zhou (2010) on the simulated datasets. For both estimators, a tuning parameter k needs to be chosen. The two estimators perform similarly for the similar values of k. For brevity, we report only the results for the tapering estimator because it is known to be rate optimal if k is appropriately selected based on the true parameter space. It is clear that for both our settings, $\Sigma \in \mathcal{C}_\alpha$ with $\alpha = 1$. But such knowledge would be absent in practice. To demonstrate the importance of knowing the true parameter space for these estimators and consequently the necessity of an adaptive estimator such as the one proposed here, we apply the estimators with five different values of α, $0.2, 0.4, 0.6, 0.8$ and 1. We chose $k = \lfloor n^{1/(2\alpha+1)} \rfloor$ for the tapering estimator following Cai, Zhang and Zhou (2010).The performance of these estimators is summarized in Figures 5 and 6 for the two settings, respectively.

It can be seen in both settings that the numerical performance of the tapering estimators critically depends on the specification of the decay rate α. Misspecifying α could lead to rather poor performance by the tapering estimators. It is perhaps not surprising to observe that the tapering estimator with $\alpha = 1$ performed the best among all estimators since it correctly specifies the true decay rate and
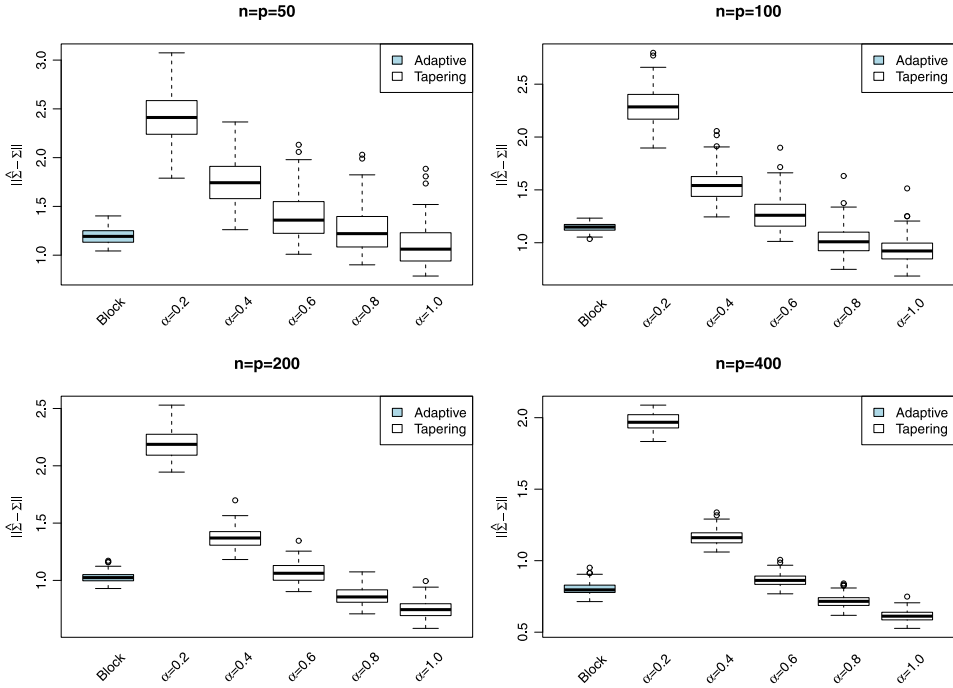
FIG. 5.   *Comparison between the tapering and adaptive block thresholding estimators—simulation setting* 1: *each panel corresponds to a particular combination of sample size n and dimension p. In each panel, boxplots of the estimation errors, measured in terms of the spectral norm, are given for the block thresholding estimator with* $\lambda_0 = 6$ *and the tapering estimator with* $\alpha = 0.2, 0.4, 0.6, 0.8$ *and* 1.

therefore, in a certain sense, made use of the information that may not be known a priori in practice. In contrast, the proposed block thresholding estimator yields competitive performance while not using such information.

5. Discussion.   In this paper we introduced a fully data-driven covariance matrix estimator by blockwise thresholding of the sample covariance matrix. The estimator simultaneously attains the optimal rate of convergence for estimating bandable covariance matrices over the full range of the parameter spaces $\mathcal{C}_\alpha$ for all $\alpha > 0$. The estimator also performs well numerically.

As noted in Section 2.2, the choice of the thresholding constant $\lambda_0 = 6$ is based on our theoretical and numerical studies. Similar to wavelet thresholding in nonparametric function estimation, in principle other choices of λ_0 can also be used. For example, the adaptivity results on the block thresholding estimator holds as long as $\lambda_0 \geq 5.44$ ($= \sqrt{24}/(1 - 2e^{-3})$) where the value 5.44 comes from the concentration inequality given in Theorem 3.3. Our experience suggests the performance of the block thresholding estimator is relatively insensitive to a small
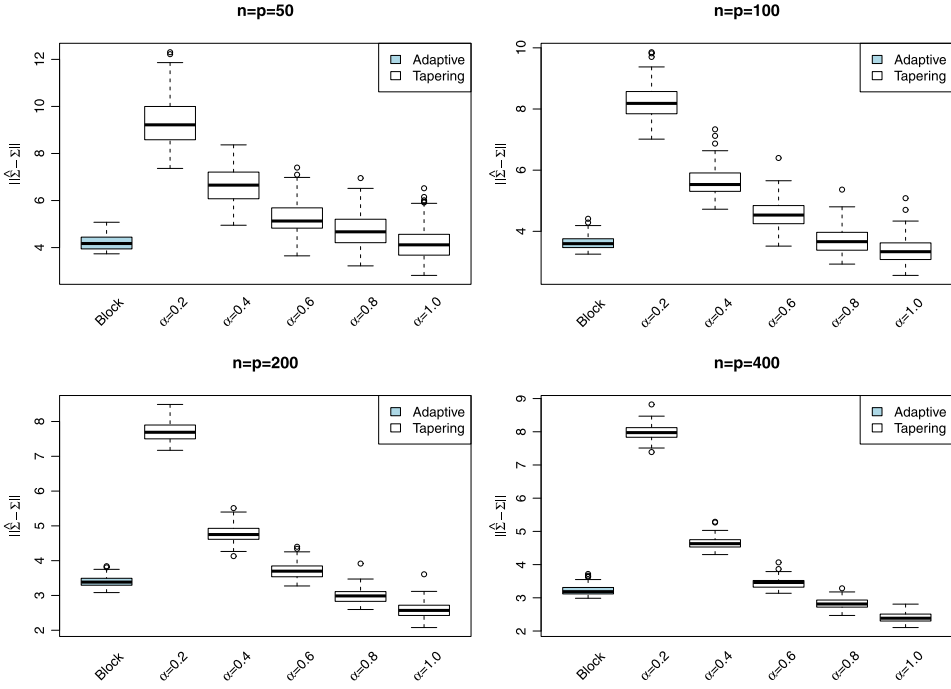
FIG. 6. *Comparison between the tapering and adaptive block thresholding estimators—simulation setting* 2: *each panel corresponds to a particular combination of sample size n and dimension p. In each panel, boxplots of the estimation errors, measured in terms of the spectral norm, are given for the block thresholding estimator with* $\lambda_0 = 6$ *and the tapering estimator with* $\alpha = 0.2, 0.4, 0.6, 0.8$ *and* 1.

change of λ_0. However, numerically the estimator can sometimes be further improved by using data-dependent choices of λ_0.

Throughout the paper, we have focused on the Gaussian case for ease of exposition and to allow for the most clear description of the block thresholding estimator. The method and the results can also be extended to more general subgaussian distributions. Suppose that the distribution of the $X^{(i)}$'s is subgaussian in the sense that there exists a constant $\sigma > 0$ such that

$$(5.1) \qquad \mathbb{P}\{|\mathbf{v}^\mathsf{T}(X - \mathbb{E}X)| > t\} \le e^{-t^2/2\sigma^2} \qquad \text{for all } t > 0 \text{ and } \|\mathbf{v}\| = 1.$$

Let $\mathcal{F}_\alpha(\sigma, M_0, M)$ denote the collection of distributions satisfying both (1.1) and (5.1). Then for any given $\sigma_0 > 0$, the block thresholding estimator $\hat{\Sigma}$ adaptively attains the optimal rate of convergence over $\mathcal{F}_\alpha(\sigma, M_0, M)$ for all α, $M_0, M > 0$ and $0 < \sigma \le \sigma_0$ whenever λ_0 is chosen sufficiently large.

In this paper we have focused on estimation under the spectral norm. The block thresholding procedure, however, can be naturally extended to achieve adaption under other matrix norms. Consider, for example, the Frobenius norm. In this case, it is natural and also necessary to threshold the blocks based on their respective

Frobenius norms instead of the spectral norms. Then following a similar argument as before, it can be shown that this Frobenius norm based block thresholding estimator can adaptively achieve the minimax rate of convergence over every $\mathcal{C}_\alpha$ for all $\alpha > 0$. It should also be noted that adaptive estimation under the Frobenius norm is a much easier problem because the squared Frobenius norm is entrywise decomposable, and the matrix can then be estimated well row by row or column by column. For example, applying a suitable block thresholding procedure for sequence estimation to the sample covariance matrix, row-by-row would also lead to an adaptive covariance matrix estimator.

The block thresholding approach can also be used for estimating sparse covariance matrices. A major difference in this case from that of estimating bandable covariance matrices is that the block sizes cannot be too large. With suitable choices of the block size and thresholding level, a fully data-driven block thresholding estimator can be shown to be rate-optimal for estimating sparse covariance matrices. We shall report the details of these results elsewhere.

6. Proofs. In this section we shall first prove Theorems 3.3 and 3.4 and then prove the main results, Theorems 3.1 and 3.2. The proofs of some additional technical lemmas are given at the end of the section.

6.1. *Proof of Theorem* 3.3. The proof relies the following lemmas.

LEMMA 1. *Let A be a 2×2 random matrix following the Wishart distribution $W(n, A_0)$ where*

$$A_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Then

$$\mathbb{P}(|A_{12} - \rho| \geq x) \leq 2\mathbb{P}(|W_n - n| \geq nx),$$

where $W_n \sim \chi_n^2$.

PROOF. Let $Z = (Z_1, Z_2)^\mathsf{T} \sim N(0, A_0)$ and $Z^{(1)}, \ldots, Z^{(n)}$ be n independent copies of Z. Let

$$S = \frac{1}{n} \sum_{i=1}^{n} Z^{(i)}(Z^{(i)})^\mathsf{T}$$

be its sample covariance matrix. It is clear that $S =_d A$. Hence

$$\mathbb{P}(|A_{12} - \rho| \geq x) = \mathbb{P}(|S_{12} - \rho| \geq x).$$

Note that

$$S_{12} - \rho = \frac{1}{4}\left(\frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} + Z_2^{(i)})^2 - 2(1+\rho)) - \frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} - Z_2^{(i)})^2 - 2(1-\rho))\right).$$

Therefore,

$$\mathbb{P}(|S_{12} - \rho| \geq x)$$

$$\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} + Z_2^{(i)})^2 - 2(1 + \rho))\right| \geq 2(1 + \rho)x\right)$$

$$+ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} - Z_2^{(i)})^2 - 2(1 - \rho))\right| \geq 2(1 - \rho)x\right).$$

Observe that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} + Z_2^{(i)})^2 - 2(1 + \rho))\right| \geq 2(1 + \rho)x\right)$$

$$= \mathbb{P}\left(\left|\sum_{i=1}^{n}\frac{(Z_1^{(i)} + Z_2^{(i)})^2}{2(1 + \rho)} - n\right| \geq x\right)$$

$$= \mathbb{P}(|W_n - n| \geq x).$$

Similarly,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}((Z_1^{(i)} - Z_2^{(i)})^2 - 2(1 - \rho))\right| \geq 2(1 - \rho)x\right) = \mathbb{P}(|W_n - n| \geq x).$$

The proof is now complete. $\square$

LEMMA 2. *Let $B = I \times J \subset [1, p]^2$. There exists an absolute constant $c_0 > 0$ such that for any $t > 1$,*

$$\mathbb{P}\left\{\|\bar{\Sigma}_B - \Sigma_B\| < c_0 t \sqrt{\|\Sigma_{I \times I}\|\|\Sigma_{J \times J}\|}\sqrt{\frac{d(B) + \log p}{n}}\right\} \geq 1 - p^{-6t^2}.$$

In particular, we can take $c_0 = 5.44$.

PROOF. Without loss of generality, assume that $\mathrm{card}(I) = \mathrm{card}(J) = d(B) = d$. Let A be a $d \times d$ matrix, $\mathbf{u}_1, \mathbf{u}_2$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{S}^{d-1}$ where $\mathcal{S}^{d-1}$ is the unit sphere in the d dimensional Euclidean space. Observe that

$$|\mathbf{u}_1^\mathsf{T} A \mathbf{v}_1| - |\mathbf{u}_2^\mathsf{T} A \mathbf{v}_2| \leq |\mathbf{u}_1^\mathsf{T} A \mathbf{v}_1 - \mathbf{u}_2^\mathsf{T} A \mathbf{v}_2|$$

$$= |\mathbf{u}_1^\mathsf{T} A(\mathbf{v}_1 - \mathbf{v}_2) + (\mathbf{u}_1 - \mathbf{u}_2)^\mathsf{T} A \mathbf{v}_2|$$

$$\leq |\mathbf{u}_1^\mathsf{T} A(\mathbf{v}_1 - \mathbf{v}_2)| + |(\mathbf{u}_1 - \mathbf{u}_2)^\mathsf{T} A \mathbf{v}_2|$$

$$\leq \|\mathbf{u}_1\|\|A\|\|\mathbf{v}_1 - \mathbf{v}_2\| + \|\mathbf{u}_1 - \mathbf{u}_2\|\|A\|\|\mathbf{v}_2\|$$

$$= \|A\|(\|\mathbf{v}_1 - \mathbf{v}_2\| + \|\mathbf{u}_1 - \mathbf{u}_2\|),$$

where as before, we use $\|\cdot\|$ to represent the spectral norm for a matrix and ℓ_2 norm for a vector. As shown by Böröczky and Wintsche [(2005), e.g., Corollary 1.2], there exists an δ-cover set $Q_d \subset \mathcal{S}^{d-1}$ of $\mathcal{S}^{d-1}$ such that

$$\mathrm{card}(Q_d) \leq \frac{c\cos\delta}{\sin^d\delta}d^{3/2}\log(1+d\cos^2\delta) \approx c\delta^{-d}d^{3/2}\log(1+d)$$

for some absolute constant $c > 0$. Note that

$$(6.1) \qquad \|A\| = \sup_{\mathbf{u},\mathbf{v}\in\mathcal{S}^{d-1}}\mathbf{u}^\mathsf{T}A\mathbf{v} \leq \sup_{\mathbf{u},\mathbf{v}\in Q_d}\mathbf{u}^\mathsf{T}A\mathbf{v} + 2\delta\|A\|.$$

In other words,

$$(6.2) \qquad \|A\| \leq (1-2\delta)^{-1}\sup_{\mathbf{u},\mathbf{v}\in Q_d}\mathbf{u}^\mathsf{T}A\mathbf{v}.$$

Now consider $A = \bar{\Sigma}_B - \Sigma_B$. Let $X_I = (X_i : i \in I)^\mathsf{T}$ and $X_J = (X_i : i \in J)^\mathsf{T}$. Then

$$\bar{\Sigma}_B = \frac{1}{n}\sum_{i=1}^n (X_I^{(i)} - \bar{X}_I)(X_J^{(i)} - \bar{X}_J)^\mathsf{T},$$

where

$$\bar{X}_I = (\bar{X}_i : i \in I)^\mathsf{T} \quad\text{and}\quad \bar{X}_J = (\bar{X}_i : i \in J)^\mathsf{T}.$$

Similarly, $\Sigma_B = \mathbb{E}(X_I - \mathbb{E}X_I)(X_J - \mathbb{E}X_J)^\mathsf{T}$. Therefore,

$$A = \frac{1}{n}\sum_{i=1}^n (X_I^{(i)}(X_J^{(i)})^\mathsf{T} - \mathbb{E}X_I X_J^\mathsf{T}) - (\bar{X}_I\bar{X}_J^\mathsf{T} - \mathbb{E}X_I\mathbb{E}X_J^\mathsf{T}).$$

Clearly the distributional properties of A are invariant to the mean of X. We shall therefore assume without loss of generality that $\mathbb{E}X = 0$ in the rest of the proof.

For any fixed $\mathbf{u}, \mathbf{v} \in \mathcal{S}^{d-1}$, we have

$$\mathbf{u}^\mathsf{T}A\mathbf{v} = \frac{1}{n}\sum_{i=1}^n (Y_1^{(i)}Y_2^{(i)} - \mathbb{E}Y_1Y_2) - \bar{Y}_1\bar{Y}_2,$$

where $Y_1 = \mathbf{u}^\mathsf{T}X_I$, $Y_2 = \mathbf{v}^\mathsf{T}X_J$, and similarly, $Y_1^{(i)} = \mathbf{u}^\mathsf{T}X_I^{(i)}$, $Y_2^{(i)} = \mathbf{v}^\mathsf{T}X_J^{(i)}$. It is not hard to see that

$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix}\mathbf{u}^\mathsf{T}\Sigma_{I\times I}\mathbf{u} & \mathbf{u}^\mathsf{T}\Sigma_{I\times J}\mathbf{v}\\ \mathbf{v}^\mathsf{T}\Sigma_{J\times I}\mathbf{u} & \mathbf{v}^\mathsf{T}\Sigma_{J\times J}\mathbf{v}\end{pmatrix}\right),$$

and $\mathbf{u}^\mathsf{T}A\mathbf{v}$ is simply the difference between the sample and population covariance of Y_1 and Y_2. We now appeal to the following lemma:

Applying Lemma 1, we obtain

$$\mathbb{P}\{|\mathbf{u}^\mathsf{T}A\mathbf{v}| \geq x\} \leq 2\mathbb{P}\left(|W_n - n| \geq \frac{nx}{((\mathbf{u}^\mathsf{T}\Sigma_{I\times I}\mathbf{u})(\mathbf{v}^\mathsf{T}\Sigma_{J\times J}\mathbf{v}))^{1/2}}\right),$$

where $W_n \sim \chi_n^2$. By the tail bound for χ^2 random variables, we have

$$\mathbb{P}\left(|W_n - n| \geq \frac{nx}{((\mathbf{u}^\mathsf{T}\Sigma_{I\times I}\mathbf{u})(\mathbf{v}^\mathsf{T}\Sigma_{J\times J}\mathbf{v}))^{1/2}}\right) \leq \exp\left(-\frac{nx^2}{4\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\right).$$

See, for example, Lemma 1 of Laurent and Massart (2000). In summary,

$$\mathbb{P}\{|\mathbf{u}^\mathsf{T}A\mathbf{v}| \geq x\} \leq 2\exp\left(-\frac{nx^2}{4\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\right).$$

Now an application of union bound then yields

$$\mathbb{P}\big(\|\bar{\Sigma}_B - \Sigma_B\| \geq x\big) \leq \mathbb{P}\Big\{\sup_{\mathbf{u},\mathbf{v}\in Q_d} \mathbf{u}^\mathsf{T}A\mathbf{v} \geq (1-2\delta)x\Big\}$$

$$\leq 2\operatorname{card}(Q_d)^2 \exp\left(-\frac{n(1-2\delta)^2x^2}{4\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\right)$$

$$\leq c\delta^{-2d}d^3\log^2(1+d)\exp\left(-\frac{n(1-2\delta)^2x^2}{4\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\right)$$

for some constant $c > 0$. In particular, taking

$$x = c_0 t\sqrt{\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|}\sqrt{\frac{d+\log p}{n}}$$

yields

$$\mathbb{P}\big(\|\bar{\Sigma}_B - \Sigma_B\| \geq x\big) \leq c\delta^{-2d}d^3\log^2(1+d)\exp\left(-\frac{c_0^2 t^2}{4}(1-2\delta)^2(d+\log p)\right).$$

Let $\delta = e^{-3}$ and

$$c_0 > \frac{\sqrt{24}}{1-2\delta} = 5.44.$$

Then

$$\mathbb{P}\big(\|\bar{\Sigma}_B - \Sigma_B\| \geq x\big) \leq p^{-6t^2}. \qquad \square$$

We are now in position to prove Theorem 3.3. It is clear that the total number of blocks can be upper bounded by $\operatorname{card}(\mathcal{B}) \leq (p/k_0)^2 < p^2$. It follows from the union bound and Lemma 2 that

$$\mathbb{P}\Big\{\bigcup_{B\in\mathcal{B}} \|\bar{\Sigma}_B - \Sigma_B\| \geq c_0 t\big(\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|\big)^{1/2}\big(n^{-1}(d(B)+\log p)\big)^{1/2}\Big\}$$

$$\leq \sum_{B\in\mathcal{B}} \mathbb{P}\big\{\|\bar{\Sigma}_B - \Sigma_B\| \geq c_0 t\big(\|\Sigma_{I\times I}\|\|\Sigma_{J\times J}\|\big)^{1/2}\big(n^{-1}(d(B)+\log p)\big)^{1/2}\big\}$$

$$\leq p^{-6t^2+2}.$$

6.2. *Proof of Theorem* 3.4. Denote by $\mathbf{u}, \mathbf{v}$ the left and right singular vectors corresponding to the leading singular value of A, that is, $\mathbf{u}^\mathsf{T} A \mathbf{v} = \|A\|$. Let $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_G)^\mathsf{T}$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_G)^\mathsf{T}$ be partitioned in the same fashion as X, for example, $\mathbf{u}_g, \mathbf{v}_g \in \mathbb{R}^{p_g}$. Denote by $\mathbf{u}_* = (\|\mathbf{u}_1\|, \ldots, \|\mathbf{u}_G\|)^\mathsf{T}$ and $\mathbf{v}_* = (\|\mathbf{v}_1\|, \ldots, \|\mathbf{v}_G\|)^\mathsf{T}$. It is clear that $\|\mathbf{u}_*\| = \|\mathbf{v}_*\| = 1$. Therefore,

$$\|\mathcal{N}(A)\| \geq \mathbf{u}_*^\mathsf{T} \mathcal{N}(A) \mathbf{v}_* = \sum_{j,k=1}^{G} \|\mathbf{u}_j\| \|\mathbf{v}_k\| \|\Sigma_{jk}\|$$

$$\geq \sum_{j,k=1}^{G} \mathbf{u}_j^\mathsf{T} \Sigma_{jk} \mathbf{v}_k = \mathbf{u}^\mathsf{T} \Sigma \mathbf{v} = \|A\|.$$

6.3. *Proof of Theorem* 3.1. With the technical tools provided by Theorems 3.3 and 3.4, we now show that $\hat{\Sigma}$ is an adaptive estimator of Σ as claimed by Theorem 3.1. We begin by establishing formal error bounds on the blocks using the technical tools introduced earlier.

6.3.1. *Large blocks.* First treat the larger blocks. When $\Sigma \in \mathcal{C}_\alpha$, large blocks can all be shrunk to zero because they necessarily occur far away from the diagonal and therefore are small in spectral norm. More precisely, we have:

LEMMA 3. *For any $B \in \mathcal{B}$ with $d(B) \geq 2k_0$,*
$$\|\Sigma_B\| \leq M d(B)^{-\alpha}.$$

Together with Theorem 3.3, this suggests that

$$\|\bar{\Sigma}_B\| \leq \|\bar{\Sigma}_B - \Sigma_B\| + \|\Sigma_B\|$$
$$\leq c_0 \big(\|\Sigma_{I \times I}\| \|\Sigma_{J \times J}\|\big)^{1/2} \big(n^{-1}(d(B) + \log p)\big)^{1/2} + M d(B)^{-\alpha},$$

with probability at least $1 - p^{-4}$. Therefore, when

(6.3) $$d(B) \geq c \min\left\{ n^{1/(2\alpha+1)}, \left(\frac{n}{\log p}\right)^{1/(2\alpha)} \right\}$$

for a large enough constant $c > 0$,

(6.4) $$\|\bar{\Sigma}_B\| < \tfrac{1}{2}(c_0 + \lambda_0)\big(\|\Sigma_{I \times I}\| \|\Sigma_{J \times J}\|\big)^{1/2} \big(n^{-1}(d(B) + \log p)\big)^{1/2}.$$

The following lemma indicates that we can further replace $\|\Sigma_{I \times I}\|$ and $\|\Sigma_{J \times J}\|$ by their respective sample counterparts.

LEMMA 4. *Denote by $\mathcal{I} = \{I : I \times J \in \mathcal{B}\}$. Then for all $I \in \mathcal{I}$,*
$$1 - \frac{\sqrt{\mathrm{card}(I)} + t}{\sqrt{n}} \leq \frac{\|\bar{\Sigma}_{I \times I}\|}{\|\Sigma_{I \times I}\|} \leq 1 + \frac{\sqrt{\mathrm{card}(I)} + t}{\sqrt{n}},$$
with probability at least $1 - 4p^2 \exp(-t^2/2)$.

In the light of Lemma 4, (6.4) implies that, with probability at least $1 - 2p^{-4}$, for any $B \in \mathcal{B}$ such that $d(B) \le n/\log n$ and (6.3) holds,

$$\|\bar{\Sigma}_B\| < \lambda_0 \big(\|\bar{\Sigma}_{I \times I}\| \|\bar{\Sigma}_{J \times J}\|\big)^{1/2} \big(n^{-1}(d(B) + \log p)\big)^{1/2},$$

whenever $n/\log p$ is sufficiently large. In other words, with probability at least $1 - 2p^{-4}$, for any $B \in \mathcal{B}$ such that (6.3) holds, $\hat{\Sigma}_B = \mathbf{0}$.

6.3.2. *Small blocks.* Now consider the smaller blocks. From the discussions in Section 3.3, we have

$$(6.5) \qquad \big\|S(\hat{\Sigma} - \Sigma, l)\big\| \le 12 \max_{B \in \mathcal{B} : d(B) = 2^{l-1} k_0} \|\hat{\Sigma}_B - \Sigma_B\|.$$

Observe that by the definition of $\hat{\Sigma}$,

$$\|\hat{\Sigma}_B - \Sigma_B\| \le \|\hat{\Sigma}_B - \bar{\Sigma}_B\| + \|\bar{\Sigma}_B - \Sigma_B\|$$
$$\le \lambda_0 \big(\|\bar{\Sigma}_{I \times I}\| \|\bar{\Sigma}_{J \times J}\|\big)^{1/2} \big(n^{-1}(d(B) + \log p)\big)^{1/2} + \|\bar{\Sigma}_B - \Sigma_B\|.$$

By Lemma 4, the spectral norm of $\bar{\Sigma}_{I \times I}$ and $\bar{\Sigma}_{J \times J}$ appeared in the first term on the rightmost-hand side can be replaced by their corresponding population counterparts, leading to

$$\|\hat{\Sigma}_B - \Sigma_B\| \le \lambda_0 \big(\|\Sigma_{I \times I}\| \|\Sigma_{J \times J}\|\big)^{1/2} \big(n^{-1}(d(B) + \log p)\big)^{1/2} + \|\bar{\Sigma}_B - \Sigma_B\|$$
$$\le \lambda_0 M_0 \big(n^{-1}(d(B) + \log p)\big)^{1/2} + \|\bar{\Sigma}_B - \Sigma_B\|,$$

where we used the fact that $\|\Sigma_{I \times I}\|, \|\Sigma_{J \times J}\| \le M_0$. This can then be readily bounded, thanks to Theorem 3.3:

$$\|\hat{\Sigma}_B - \Sigma_B\| \le (\lambda_0 M_0 + c_0) \big(n^{-1}(d(B) + \log p)\big)^{1/2}.$$

Together with (6.5), we get

$$(6.6) \qquad \big\|S(\hat{\Sigma} - \Sigma, l)\big\| \le C \big(n^{-1}(k_0 2^{l-1} + \log p)\big)^{1/2}.$$

6.3.3. *Bounding the estimation error.* To put the bounds on both small and big blocks together, we need only to choose an appropriate cutoff L in (3.2). In particular, we take

$$(6.7) \quad L = \begin{cases} \lceil \log_2(p/k_0) \rceil, & \text{if } p \le n^{1/(2\alpha+1)}, \\ \lceil \log_2(n^{1/2\alpha+1}/k_0) \rceil, & \text{if } \log p < n^{1/(2\alpha+1)} \text{ and } n^{1/(2\alpha+1)} \le p, \\ \lceil \log_2(\log p/k_0) \rceil, & \text{if } n^{1/(2\alpha+1)} \le \log p, \end{cases}$$

where $\lceil x \rceil$ stands for the smallest integer that is no less than x.

Small p. If $p \leq n^{1/(2\alpha+1)}$, all blocks are small. From the bound derived for small blocks, for example, equation (6.6), we have

$$\|\hat{\Sigma} - \Sigma\| \leq \sum_l \|S(\hat{\Sigma} - \Sigma, l)\| \leq C \sum_l (n^{-1}(2^{l-1}k_0 + \log p))^{1/2} \leq C(p/n)^{1/2},$$

with probability at least $1 - 2p^{-4}$. Hereafter we use $C > 0$ as a generic constant that does not depend on p, n or α, and its value may change at each appearance. Thus

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 = \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \mathbb{I}\{\|\hat{\Sigma} - \Sigma\| \leq C(p/n)^{1/2}\}$$
$$+ \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \mathbb{I}\{\|\hat{\Sigma} - \Sigma\| > C(p/n)^{1/2}\}.$$

It now suffices to show that the second term on the right-hand side is $O(p/n)$. By the Cauchy–Schwarz inequality,

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \mathbb{I}(\|\hat{\Sigma} - \Sigma\| > C(p/n)^{1/2})$$
$$\leq (\mathbb{E}\|\hat{\Sigma} - \Sigma\|^4 \mathbb{P}\{\|\hat{\Sigma} - \Sigma\| > C(p/n)^{1/2}\})^{1/2}$$
$$\leq (2p^{-4}\mathbb{E}\|\hat{\Sigma} - \Sigma\|^4)^{1/2}.$$

Observe that

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^4 \leq \mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}^4 \leq Cp^4/n^2,$$

where $\|\cdot\|_{\mathrm{F}}$ stands for the Frobenius norm of a matrix. Thus,

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|^2 \mathbb{I}\{\|\hat{\Sigma} - \Sigma\| > C(p/n)^{1/2}\} \leq Cp/n.$$

Medium p. When $\log p < n^{1/(2\alpha+1)}$ and $n^{1/(2\alpha+1)} \leq p$, by the analysis from Section 6.3.1, all large blocks will be shrunk to zero with overwhelming probability, that is,

$$\mathbb{P}\left\{\sum_{l>L} S(\hat{\Sigma}, l) = 0\right\} \geq 1 - 2p^{-4}.$$

When this happens,

$$\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\| = \left\|\sum_{l>L} S(\Sigma, l)\right\| \leq \left\|\sum_{l>L} S(\Sigma, l)\right\|_{\ell_1}.$$

Recall that $\|\cdot\|_{\ell_1}$ stands for the matrix ℓ_1 norm, that is, the maximum row sum of the absolute values of the entries of a matrix. Hence,

$$\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\| \leq ML^{-\alpha} \leq Cn^{-\alpha/(2\alpha+1)}.$$

As a result,

$$\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^2 = \mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^2 \mathbb{I}\left\{\sum_{l>L} S(\hat{\Sigma}, l) = 0\right\}$$

$$+ \mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^2 \mathbb{I}\left\{\sum_{l>L} S(\hat{\Sigma}, l) \neq 0\right\}.$$

It remains to show that

$$\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^2 \mathbb{I}\left\{\sum_{l>L} S(\hat{\Sigma}, l) \neq 0\right\} = O(n^{-2\alpha/(2\alpha+1)}).$$

By the Cauchy–Schwarz inequality,

$$\mathbb{E}\left\{\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^2 \mathbb{I}\left(\sum_{l>L} S(\hat{\Sigma}, l) \neq 0\right)\right\}$$

$$\leq \left(\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^4 \mathbb{P}\left\{\sum_{l>L} S(\hat{\Sigma}, l) \neq 0\right\}\right)^{1/2}.$$

Observe that

$$\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^4 \leq \left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|_F^4 = \left(\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|_F^2\right)^2$$

$$\leq \left(\left\|\sum_{l>L} S(\bar{\Sigma} - \Sigma, l)\right\|_F^2 + \left\|\sum_{l>L} S(\Sigma, l)\right\|_F^2\right)^2$$

$$\leq 2\left(\left\|\sum_{l>L} S(\bar{\Sigma} - \Sigma, l)\right\|_F^4 + \left\|\sum_{l>L} S(\Sigma, l)\right\|_F^4\right),$$

where the second inequality follows from the fact that $\hat{\Sigma} = \bar{\Sigma}$ or $\mathbf{0}$. It is not hard to see that

$$\mathbb{E}\left\|\sum_{l>L} S(\bar{\Sigma} - \Sigma, l)\right\|_F^4 \leq \mathbb{E}\|\bar{\Sigma} - \Sigma\|_F^4 \leq Cp^4/n^2.$$

On the other hand,

$$\left\|\sum_{l>L} S(\Sigma, l)\right\|_F^4 \leq \left(\sum_{i,j:|i-j|>k_0 2^{L-1}} \sigma_{ij}^2\right)^2 \leq \left(\sum_{i,j:|i-j|>k_0 2^{L-1}} |\sigma_{ij}|\right)^4$$

$$\leq Cn^{-4\alpha/(2\alpha+1)}.$$

Therefore,

$$\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma} - \Sigma, l)\right\|^4 \leq C(p^4/n^2 + n^{-4\alpha/(2\alpha+1)}).$$

Together with Theorem 3.3, we conclude that

$$\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma}-\Sigma, l)\right\|^2 \mathbb{I}\left\{\sum_{l>L} S(\hat{\Sigma}, l) \neq 0\right\} \leq Cn^{-1}.$$

Large p. Finally, when p is very large in that $\log p > n^{1/(2\alpha+1)}$, we can proceed in the same fashion. Following the same argument as before, it can be shown that

$$\mathbb{E}\left\|\sum_{l>L} S(\hat{\Sigma}, l)\right\|^2 \leq C(n^{-1}\log p).$$

The smaller blocks can also be treated in a similar fashion as before. From equation (6.6),

$$\sum_{l\leq L} \|S(\hat{\Sigma}-\Sigma, l)\| \leq C(n^{-1}\log p),$$

with probability at least $1 - 2p^{-4}$. Thus, it can be calculated that

$$\mathbb{E}\left(\sum_{l\leq L} \|S(\hat{\Sigma}-\Sigma, l)\|\right)^2 \leq C(n^{-1}\log p).$$

Combining these bounds, we conclude that $\mathbb{E}\|\hat{\Sigma}-\Sigma\|^2 \leq C(n^{-1}\log p)$. In summary,

$$\sup_{\Sigma\in\mathcal{C}_\alpha} \mathbb{E}\|\hat{\Sigma}-\Sigma\|^2 \leq C\min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\},$$

for all $\alpha > 0$. In other words, the block thresholding estimator $\hat{\Sigma}$ achieves the optimal rate of convergence simultaneously over every $\mathcal{C}_\alpha$ for all $\alpha > 0$.

6.4. *Proof of Theorem 3.2.* Observe that

$$\mathbb{E}\|\hat{\Omega}-\Omega\|^2 = \mathbb{E}\big(\|\hat{\Omega}-\Omega\|^2\mathbb{I}\{\lambda_{\min}(\hat{\Sigma}) \geq \tfrac{1}{2}\lambda_{\min}(\Sigma)\}\big)$$
$$+ \mathbb{E}\big(\|\hat{\Omega}-\Omega\|^2\mathbb{I}\{\lambda_{\min}(\hat{\Sigma}) < \tfrac{1}{2}\lambda_{\min}(\Sigma)\}\big),$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a symmetric matrix. Under the event that

$$\lambda_{\min}(\hat{\Sigma}) \geq \tfrac{1}{2}\lambda_{\min}(\Sigma),$$

$\hat{\Sigma}$ is positive definite and $\hat{\Omega} = \hat{\Sigma}^{-1}$. Note also that

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

Therefore,

$$\mathbb{E}\left(\|\hat{\Omega} - \Omega\|^2 \mathbb{I}\left\{\lambda_{\min}(\hat{\Sigma}) \geq \frac{1}{2}\lambda_{\min}(\Sigma)\right\}\right) \leq 4\|\Omega\|^2 \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2$$

$$\leq C\min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\}$$

by Theorem 3.1. On the other hand,

$$\mathbb{E}(\|\hat{\Omega} - \Omega\|^2 \mathbb{I}\{\lambda_{\min}(\hat{\Sigma}) < \tfrac{1}{2}\lambda_{\min}(\Sigma)\})$$

$$\leq \mathbb{E}((\|\hat{\Omega}\| + \|\Omega\|)^2 \mathbb{I}\{\lambda_{\min}(\hat{\Sigma}) < \tfrac{1}{2}\lambda_{\min}(\Sigma)\})$$

$$\leq (n + \|\Omega\|)^2 \mathbb{P}\{\lambda_{\min}(\hat{\Sigma}) < \tfrac{1}{2}\lambda_{\min}(\Sigma)\}.$$

Note that

$$\mathbb{P}\{\lambda_{\min}(\hat{\Sigma}) < \tfrac{1}{2}\lambda_{\min}(\Sigma)\} \leq \mathbb{P}\{\|\hat{\Sigma} - \Sigma\| > \tfrac{1}{2}\lambda_{\min}(\Sigma)\}.$$

It suffices to show that

$$n^2 \mathbb{P}\left\{\|\hat{\Sigma} - \Sigma\| > \frac{1}{2}\lambda_{\min}(\Sigma)\right\} \leq C\min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\}.$$

Consider first the case when p is large. More specifically, let

$$p > n(48\lambda_0 M^2)^{-2}.$$

As shown in the proof of Theorem 3.1,

$$\mathbb{P}\{\|\hat{\Sigma} - \Sigma\| > \tfrac{1}{2}\lambda_{\min}(\Sigma)\} \leq 4p^{-4}.$$

It is not hard to see that this implies the desired claim.

Now consider the case when

$$p \leq n(48\lambda_0 M^2)^{-2}.$$

Observe that for each $B = I \times J \in \mathcal{B}$,

$$\|\hat{\Sigma}_B - \bar{\Sigma}_B\| \leq \lambda_0(\|\bar{\Sigma}_{I\times I}\|\|\bar{\Sigma}_{J\times J}\|)^{1/2}(n^{-1}(d(B) + \log p))^{1/2}$$

$$\leq \lambda_0\|\bar{\Sigma}\|(n^{-1}(d(B) + \log p))^{1/2}.$$

It can then be deduced from the norm compression inequality, in a similar spirit as before, that

$$\|\hat{\Sigma} - \bar{\Sigma}\| \leq \sum_l \|S(\hat{\Sigma} - \bar{\Sigma}, l)\|$$

$$\leq 12\lambda_0\|\bar{\Sigma}\| \sum_l (n^{-1}(2^{l-1}k_0 + \log p))^{1/2}$$

$$\leq 12\lambda_0\|\bar{\Sigma}\|(p/n)^{1/2}.$$

By the triangle inequality,

$$\|\hat{\Sigma} - \Sigma\| \le \|\bar{\Sigma} - \Sigma\| + \|\hat{\Sigma} - \bar{\Sigma}\|,$$

and

$$\|\bar{\Sigma}\| \le \|\bar{\Sigma} - \Sigma\| + \|\Sigma\|.$$

Under the event that

$$\|\bar{\Sigma} - \Sigma\| > \frac{(1/2)\lambda_{\min}(\Sigma) - 12\lambda_0(p/n)^{1/2}\lambda_{\max}(\Sigma)}{1 + 12\lambda_0(p/n)^{1/2}} \ge \frac{1}{5}\lambda_{\min}(\Sigma),$$

we have

$$\|\hat{\Sigma} - \Sigma\| > \tfrac{1}{2}\lambda_{\min}(\Sigma).$$

Now by Lemma 2,

$$\mathbb{P}\left\{\|\hat{\Sigma} - \Sigma\| > \frac{1}{2}\lambda_{\min}(\Sigma)\right\} \le \mathbb{P}\left\{\|\bar{\Sigma} - \Sigma\| > \frac{1}{5}\lambda_{\min}(\Sigma)\right\} \le \exp\left(-\frac{cn\lambda_{\min}^2(\Sigma)}{\lambda_{\max}^2(\Sigma)}\right),$$

for some constant $c > 0$, which concludes the proof.

6.5. *Proof of Lemma* 3. The proof relies on the following simple observation.

LEMMA 5. *For any $B \in \mathcal{B}$ with dimension $d(B) \ge 4k_0$,*

$$\min_{(i,j)\in B} |i - j| \ge d(B).$$

PROOF. Note that for any $B \in \mathcal{B}$, there exists an integer $r > 0$ such that $d(B) = 2^{r-1}k_0$. We proceed by induction on r. When $r = 3$, it is clear by construction, blocks of size $4k_0 \times 4k_0$ are at least one $2k_0 \times 2k_0$ block away from the diagonal. See Figure 2 also. This implies that the statement is true for $r = 3$. From $r + 1$ to $r + 2$, one simply observes that all blocks of size $2^{r+1}k_0 \times 2^{r+1}k_0$ is at least one $2^r k_0 \times 2^r k_0$ block away from blocks of size $2^{r-1}k_0 \times 2^{r-1}k_0$. Therefore,

$$\min_{(i,j)\in B} |i - j| \ge 2^r k_0 + 2^r k_0 = 2^{r+1}k_0,$$

which implies the desired statement.   □

We are now in position to prove Lemma 3 which states that big blocks of the covariance matrix are small in spectral norm. Recall that the matrix ℓ_1 norm is defined as

$$\|A\|_{\ell_1} = \sup_{\mathbf{x}\in\mathbb{R}^p : \|\mathbf{x}\|_{\ell_1}=1} \|A\mathbf{x}\|_{\ell_1} = \max_{1\le j\le n} \sum_{i=m}^{p} |a_{ij}|,$$

for an $m \times n$ matrix $A = (a_{ij}))_{1 \leq i \leq m, 1 \leq j \leq n}$. Similarly the matrix ℓ_∞ norm is defined as

$$\|A\|_{\ell_\infty} = \sup_{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_{\ell_\infty} = 1} \|A\mathbf{x}\|_{\ell_\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|.$$

It is well known [see, e.g., Golub and Van Loan (1996)] that

$$\|A\|^2 \leq \|A\|_{\ell_1} \|A\|_{\ell_\infty}.$$

Immediately from Lemma 5, we have

$$\|\Sigma_B\|_{\ell_1}, \|\Sigma_B\|_{\ell_\infty} \leq \max_{1 \leq i \leq p} \sum_{j : |j-i| \geq 2^r k_0} |\sigma_{ij}| \leq M d(B)^{-\alpha},$$

which implies $\|\Sigma_B\| \leq M d(B)^{-\alpha}$.

6.6. *Proof of Lemma* 4. For any $I \in \mathcal{I}$, write $Z_I = \Sigma_{I \times I}^{-1/2} Y$. Then the entries of Z_I are independent standard normal random variables. From the concentration bounds on the random matrices [see, e.g., Davidson and Szarek (2001)], we have

$$1 - \frac{\sqrt{\text{card}(I)} + t}{\sqrt{n}} \leq \lambda_{\min}^{1/2}(\bar{\Sigma}_{Z_I}) \leq \lambda_{\max}^{1/2}(\bar{\Sigma}_{Z_I}) \leq 1 + \frac{\sqrt{\text{card}(I)} + t}{\sqrt{n}}$$

with probability at least $1 - 2\exp(-t^2/2)$ where $\bar{\Sigma}_{Z_I}$ is the sample covariance matrix of Z_I. Applying the union bound to all $I \in \mathcal{I}$ yields that with probability at least $1 - 2p^2 \exp(-t^2/2)$, for all I

$$1 - \frac{\sqrt{\text{card}(I)} + t}{\sqrt{n}} \leq \lambda_{\min}^{1/2}(\bar{\Sigma}_{Z_I}) \leq \lambda_{\max}^{1/2}(\bar{\Sigma}_{Z_I}) \leq 1 + \frac{\sqrt{\text{card}(I)} + t}{\sqrt{n}}.$$

Observe that $\bar{\Sigma}_{I \times I} = \Sigma_{I \times I}^{1/2} \bar{\Sigma}_{Z_I} \Sigma_{I \times I}^{1/2}$. Thus

$$\lambda_{\min}(\bar{\Sigma}_{Z_I}) \lambda_{\max}(\Sigma_{I \times I}) \leq \lambda_{\max}(\bar{\Sigma}_{I \times I}) \leq \lambda_{\max}(\bar{\Sigma}_{Z_I}) \lambda_{\max}(\Sigma_{I \times I}),$$

which implies the desired statement.

REFERENCES

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BÖRÖCZKY, K. and WINTSCHE, G. (2005). Covering the sphere by equal spherical balls. Available at http://www.renyi.hu/~carlos/spherecover.ps.

CAI, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924. MR1724035

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., LIU, W. and ZHOU, H. H. (2011). Optimal estimation of large sparse precision matrices. Unpublished 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. MR2676885

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

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, *Vol. I* 317–366. North-Holland, Amsterdam. MR1863696

EFROMOVICH, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–661.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, T. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

ROCHA, G., ZHAO, P. and YU, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation. Technical report, Dept. Statistics, Univ. California, Berkeley.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu

SCHOOL OF INDUSTRIAL
   AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA
E-MAIL: myuan@isye.gatech.edu