Positive Definite ℓ_1 Penalized Estimation of Large Covariance Matrices

Lingzhou Xue, Shiqian Ma and Hui Zou

University of Minnesota

December 13, 2011

Accepted by JASA, August 2012

Abstract

The thresholding covariance estimator has nice asymptotic properties for estimating sparse large covariance matrices, but it often has negative eigenvalues when used in real data analysis. To simultaneously achieve sparsity and positive definiteness, we develop a positive definite ℓ_1-penalized covariance estimator for estimating sparse large covariance matrices. An efficient alternating direction method is derived to solve the challenging optimization problem and its convergence properties are established. Under weak regularity conditions, non-asymptotic statistical theory is also established for the proposed estimator. The competitive finite-sample performance of our proposal is demonstrated by both simulation and real applications.

Keywords: Alternating direction methods; Large covariance matrices; Matrix norm; Positive-definite estimation; Sparsity; Soft-thresholding.

1 Introduction

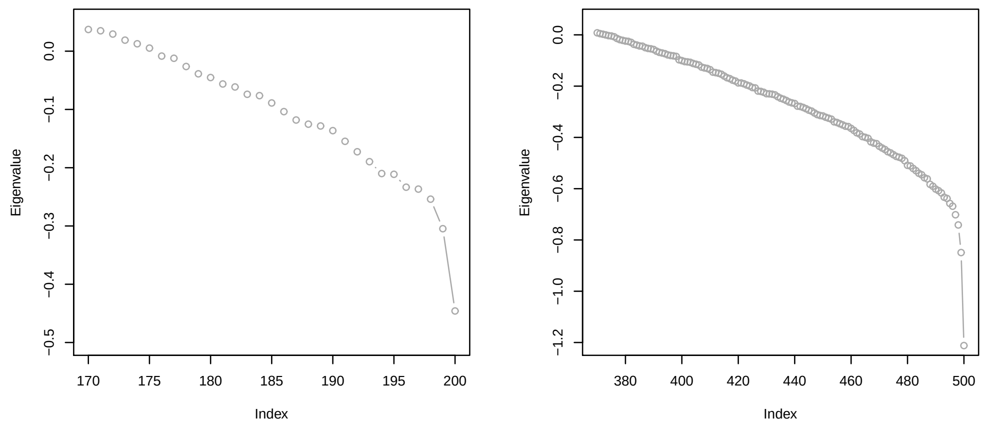
Estimating covariance matrices is of fundamental importance for an abundance of statistical methodologies. Nowadays, the advance of new technologies has brought massive high-dimensional data into various research fields, such as fMRI imaging, web mining, bioinformatics, climate studies and risk management, and so on. The usual sample covariance matrix is optimal in the classical setting with large samples and fixed low dimensions (Anderson, 1984), but it performs very poorly in the high-dimensional setting (Johnstone, 2001). In the recent literature, regularization techniques have been used to improve the sample covariance matrix estimator, including banding (Wu and Pourahmadi, 2003; Bickel and Levina, 2008a), tapering (Furrer and Bengtsson, 2007; Cai, Zhang, and Zhou, 2010) and thresholding (Bickel and Levina, 2008b; El Karoui, 2008; Rothman, Levina, and Zhu, 2009). Banding or tapering is very useful when the variables have a natural ordering and off-diagonal entries of the target covariance matrix decays to zero as they move away from the diagonal. On the other hand, thresholding is proposed for estimating permutation-invariant covariance matrices. Thresholding can be used to produce consistent covariance matrix estimators when the true covariance matrix is bandable (Bickel and Levina, 2008b; Cai and Zhou, 2011a). In this sense, thresholding is more robust than banding/tapering for real applications.

Let $\hat{\mathbf{\Sigma}}_n = (\hat{\sigma}_{ij})_{1\le i,j\le p}$ be the sample covariance matrix. Rothman, Levina, and Zhu (2009) defined the general thresholding covariance matrix estimator as $\hat{\mathbf{\Sigma}}_{thr} = \{s_\lambda(\hat{\sigma}_{ij})\}_{1\le i,j\le p}$, where $s_\lambda(z)$ is the generalized thresholding function. The generalized thresholding function covers a number of commonly used shrinkage procedures, e.g. the hard thresholding $s_\lambda(z) = zI_{\{|z|>\lambda\}}$, the soft thresholding $s_\lambda(z) = \mathrm{sign}(z)(|z|-\lambda)_+$, the smoothly clipped absolute deviation thresholding (Fan and Li, 2001) and the adaptive lasso thresholding (Zou, 2006). Consistency results and explicit rates of convergence have been obtained for these regularized estimators in the literature, e.g. Bickel and Levina (2008a,b), El Karoui (2008), Rothman et al. (2009), Cai and Liu (2011). The recent work by Cai and Zhou (2011a) has established the minimax rate of convergence under the ℓ_1 matrix norm over a fairly wide range of classes of large covariance matrices, where the thresholding estimator is shown to be minimax rate optimal. The existing theoretical and empirical results show no clear favoritism to a particular thresholding rule. In this paper we focus on the soft-thresholding because it can be formulated as the solution of a convex optimization problem. Let $\|\cdot\|_F$ be the Frobenius norm and $|\cdot|_1$ be the element-wise ℓ_1-norm of all non-diagonal elements. Then the soft-thresholding covariance estimator is equal to

$$\hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma}} \ \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_n\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1. \tag{1}$$

However, there is no guarantee that the thresholding estimator is always positive definite. Although the positive definite property is guaranteed in the asymptotic setting with high probability, the actual estimator can be an indefinite matrix, especially in real data analysis. To illustrate this issue, we consider the Michigan lung cancer gene-expression data (Beer et al., 2002) which have 86 tumor samples from patients with lung adenocarcinomas and 5217 gene expression values for each sample. More details about this dataset are referred to Beer et al. (2002) and Subramaniana et al. (2005). We randomly choose p genes ($p = 200, 500$), and obtain the soft-thresholding sample correlation matrix for these genes. We repeat the process ten times for $p = 200$ and 500 respectively, and each time the thresholding parameter λ is selected via the 5-fold cross validation. We found that none of the soft-thresholding estimators would become positive definite for both $p = 200$ and 500. On average, there exist 22 and 124 negative eigenvalues for the soft-thresholding estimator for $p = 200$ and $p = 500$, respectively. Figure 1 displays the 30 smallest eigenvalues for $p = 200$ and the 130 smallest eigenvalues for $p = 500$.

To deal with the indefiniteness, one possible solution is to utilize the eigen-decomposition of $\hat{\boldsymbol{\Sigma}}$, and project $\hat{\boldsymbol{\Sigma}}$ into the convex cone $\{\boldsymbol{\Sigma} \succeq 0\}$. Assume that $\hat{\boldsymbol{\Sigma}}$ has the eigen-decomposition $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{p} \hat{\lambda}_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_i$, and then a positive semidefinite estimator $\tilde{\boldsymbol{\Sigma}}^{+}$ can be obtained by setting $\tilde{\boldsymbol{\Sigma}}^{+} = \sum_{i=1}^{p} \max(\hat{\lambda}_i, 0) \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_i$. However, this strategy does not work well for sparse covariance matrix estimation, because the projection destroys the sparsity pattern of $\hat{\boldsymbol{\Sigma}}$. Consider the Michigan data again. After semidefinite projection, the soft-thresholding



(A) $p = 200$: the minimal 30 eigenvalues

(B) $p = 500$: the minimal 130 eigenvalues

Figure 1: Illustration of the indefinite soft-thresholding estimator in the Michigan lung cancer data

estimator has no zero entry.

In order to simultaneously achieve sparsity and positive semidefiniteness, a natural solution is to add the positive semidefinite constraint to (1). Consider the following constrained ℓ_1 penalization problem

$$\hat{\mathbf{\Sigma}}^+ = \arg\min_{\mathbf{\Sigma} \succeq 0} \; \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 + \lambda|\mathbf{\Sigma}|_1. \tag{2}$$

Note that the solution to (2) could be positive semidefinite. To obtain a positive definite covariance estimator, we can consider the positive definite constraint $\{\mathbf{\Sigma} \succeq \epsilon \boldsymbol{I}\}$ for some arbitrarily small $\epsilon > 0$. Then the modified $\hat{\mathbf{\Sigma}}^+$ is always positive definite. In this work, we focus on solving the positive definite $\hat{\mathbf{\Sigma}}^+$ as follows

$$\hat{\mathbf{\Sigma}}^+ = \arg\min_{\mathbf{\Sigma} \succeq \epsilon \boldsymbol{I}} \; \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 + \lambda|\mathbf{\Sigma}|_1. \tag{3}$$

Despite its natural motivation, (3) is actually a very challenging optimization problem due to the positive semidefinite constraint. To our best knowledge, the first attempt for solving (3) was recently proposed by Rothman (2011) who added the log-determinant barrier function to (3):

$$\breve{\mathbf{\Sigma}}^{+} = \arg\min_{\mathbf{\Sigma}\succ 0} \ \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 - \tau \log\det(\mathbf{\Sigma}) + \lambda|\mathbf{\Sigma}|_1, \tag{4}$$

where the barrier parameter τ is a small positive constant, say 10^{-4}. From the optimization viewpoint, (4) is similar to the graphical lasso criterion (Friedman et al., 2008) which also has a log-determinant part and the element-wise ℓ_1-penalty. Rothman (2011) derived an iterative procedure to solve (4) . Rothman (2011)'s proposal is based on heuristic arguments and its convergence property is unknown.

In this paper we present an alternating direction algorithm for solving (3) directly. Numerical examples show that our algorithm is much faster than the log-barrier method. We further prove the convergence properties of our algorithm and discuss the statistical properties of the positive-definite constrained ℓ_1 penalized covariance estimator.

2 Alternating Direction Algorithm

We use an alternating direction method to solve (3) directly. The alternating direction method is closely related to the operator-splitting method that has a long history back to 1950s for solving numerical partial differential equations, see e.g., Douglas and Rachford (1956); Peaceman and Rachford (1955). Recently, the alternating direction method has been revisited and successfully applied to solving large scale problems arising from different applications. For example, Scheinberg, Ma, and Goldfarb (2010) introduced the alternating linearization methods to efficiently solve the graphical lasso optimization problem. We refer to Fortin and Glowinski (1983); Glowinski and Le Tallec (1989) for more details on operator-splitting and alternating direction methods.

In the sequel, we propose an alternating direction method to solve the ℓ_1 penalized covariance matrix estimation problem (3) under the positive-semidefinite constraint. We first introduce a new variable $\mathbf{\Theta}$ and an equality constraint as follows

$$(\hat{\mathbf{\Theta}}^{+}, \hat{\mathbf{\Sigma}}^{+}) = \arg\min_{\mathbf{\Theta},\mathbf{\Sigma}} \ \{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 + \lambda|\mathbf{\Sigma}|_1 : \ \mathbf{\Sigma} = \mathbf{\Theta}, \ \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}\}. \tag{5}$$

The solution to (5) gives the solution to (3). To deal with the equality constraint in (5), we shall minimize its augmented Lagrangian function for some given penalty parameter μ, i.e.

$$L(\mathbf{\Theta}, \mathbf{\Sigma}; \mathbf{\Lambda}) = \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 + \lambda|\mathbf{\Sigma}|_1 - \langle \mathbf{\Lambda}, \mathbf{\Theta} - \mathbf{\Sigma} \rangle + \|\mathbf{\Theta} - \mathbf{\Sigma}\|_F^2/(2\mu), \tag{6}$$

where $\mathbf{\Lambda}$ is the Lagrange multiplier. We iteratively solve

$$(\mathbf{\Theta}^{i+1}, \mathbf{\Sigma}^{i+1}) = \arg\min L(\mathbf{\Theta}, \mathbf{\Sigma}; \mathbf{\Lambda}^i) \tag{7}$$

and then update the Lagrangian multiplier $\mathbf{\Lambda}^{i+1}$ by

$$\mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i - (\mathbf{\Theta}^{i+1} - \mathbf{\Sigma}^{i+1})/\mu.$$

For (7) we do it by alternatingly minimizing $L(\mathbf{\Theta}, \mathbf{\Sigma}; \mathbf{\Lambda}^i)$ with respect to $\mathbf{\Theta}$ and $\mathbf{\Sigma}$.

To sum up, the entire algorithm proceeds as follows:

For $i = 0, 1, 2, \ldots$, solve the following three sub-problems sequentially till convergence

$$\mathbf{\Theta} \text{ step}: \quad \mathbf{\Theta}^{i+1} = \arg\min_{\mathbf{\Theta} \succeq \epsilon \boldsymbol{I}} L(\mathbf{\Theta}, \mathbf{\Sigma}^i; \mathbf{\Lambda}^i) \tag{8}$$

$$\mathbf{\Sigma} \text{ step}: \quad \mathbf{\Sigma}^{i+1} = \arg\min_{\mathbf{\Sigma}} L(\mathbf{\Theta}^{i+1}, \mathbf{\Sigma}; \mathbf{\Lambda}^i) \tag{9}$$

$$\mathbf{\Lambda} \text{ step}: \quad \mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i - (\mathbf{\Theta}^{i+1} - \mathbf{\Sigma}^{i+1})/\mu. \tag{10}$$

To further simplify the alternating direction algorithm, we derive the closed-form solutions for (8)–(9). Consider the $\mathbf{\Theta}$ step. Define $(\boldsymbol{Z})_+$ as the projection of a matrix $\boldsymbol{Z}$ onto the convex cone $\{\mathbf{\Theta} \succeq \epsilon \boldsymbol{I}\}$. Assume that $\boldsymbol{Z}$ has the eigen-decomposition $\sum_{i=1}^p \lambda_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_i$, and then $(\boldsymbol{Z})_+$ can be obtained as $\sum_{i=1}^p \max(\lambda_i, \epsilon) \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_i$. Then the $\mathbf{\Theta}$ step can be analytically solved as follows

$$\begin{aligned}
\mathbf{\Theta}^{i+1} &= \arg\min_{\mathbf{\Theta} \succeq \epsilon \boldsymbol{I}} L(\mathbf{\Theta}, \mathbf{\Sigma}^i; \mathbf{\Lambda}^i) \\
&= \arg\min_{\mathbf{\Theta} \succeq \epsilon \boldsymbol{I}} -\langle \mathbf{\Lambda}^i, \mathbf{\Theta} \rangle + \|\mathbf{\Theta} - \mathbf{\Sigma}^i\|_F^2/(2\mu) \\
&= (\mathbf{\Sigma}^i + \mu \mathbf{\Lambda}^i)_+.
\end{aligned}$$

Next, define an entry-wise soft-thresholding rule for all the non-diagonal elements of a matrix $\boldsymbol{Z}$ as $\boldsymbol{S}(\boldsymbol{Z}, \tau) = \{s(z_{ij}, \tau)\}_{1 \le i,j \le p}$ with

$$s(z_{ij}, \tau) = \text{sign}(z_{ij}) \max(|z_{ij}| - \tau, 0) I_{\{i \neq j\}} + z_{ij} I_{\{i=j\}}.$$

Then the $\mathbf{\Sigma}$ step has a closed-form solution given below

$$\begin{aligned}
\mathbf{\Sigma}^{i+1} &= \arg\min_{\mathbf{\Sigma}} L(\mathbf{\Theta}^{i+1}, \mathbf{\Sigma}; \mathbf{\Lambda}^i) \\
&= \arg\min_{\mathbf{\Sigma}} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2/2 + \lambda|\mathbf{\Sigma}|_1 + \langle \mathbf{\Lambda}^i, \mathbf{\Sigma} \rangle + \|\mathbf{\Sigma} - \mathbf{\Theta}^{i+1}\|_F^2/(2\mu) \\
&= \{\boldsymbol{S}(\mu(\hat{\mathbf{\Sigma}}_n - \mathbf{\Lambda}^i) + \mathbf{\Theta}^{i+1}, \lambda\mu)\}/(1+\mu).
\end{aligned}$$

Algorithm 1 shows the complete details of our alternating direction method for (3). In Section 4 we provide the convergence analysis of Algorithm 1 and prove that Algorithm 1 always converges to the optimal solution of (5) from any starting point.

Algorithm 1 Our alternating direction method for the ℓ_1 penalized covariance estimator

1. Input: μ, $\mathbf{\Sigma}^0$ and $\mathbf{\Lambda}^0$.

2. Iterative alternating direction augmented Lagrangian step: for the i-th iteration

   2.1 Solve $\mathbf{\Theta}^{i+1} = (\mathbf{\Sigma}^i + \mu\mathbf{\Lambda}^i)_+$;

   2.2 Solve $\mathbf{\Sigma}^{i+1} = \{\boldsymbol{S}(\mu(\hat{\mathbf{\Sigma}}_n - \mathbf{\Lambda}^i) + \mathbf{\Theta}^{i+1}, \lambda\mu)\}/(1+\mu)$;

   2.3 update $\mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i - (\mathbf{\Theta}^{i+1} - \mathbf{\Sigma}^{i+1})/\mu$.

3. Repeat the above cycle till convergence.

In our implementation we use the soft-thresholding estimator as the initial value for both $\mathbf{\Theta}^0$ and $\mathbf{\Sigma}^0$, and we set $\mathbf{\Lambda}^0$ as a zero matrix. The value for μ is 2. Before invoking Algorithm 1, we always check whether the soft-thresholding estimator is positive definite. If yes, then the soft-threhsolding estimator is the final solution to (3).

3 Numerical Examples

3.1 Simulation

Before delving into theoretical analysis of the algorithm and the resulting estimator, we first use simulation to show the competitive performance of our proposal. In all examples we standardize the variables to have zero mean and unit variance. In each simulation model, we generated 100 independent datasets, each with $n = 50$ independent p-variate random vectors from the multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}_0 = (\sigma^0_{ij})_{1\le i,j\le p}$ for $p = 100, 200$ & 500. We considered two covariance models with different sparsity patterns:

Model 1: $\sigma^0_{ij} = (1 - |i-j|/10)_+$.

Model 2: partition the indices $\{1, 2, \ldots, p\}$ into $K = p/20$ non-overlapping subsets of equal size, and let i_k denote the maximum index in I_k.

$$\sigma^0_{ij} = 0.6 I_{\{i=j\}} + 0.4 \sum_{k=1}^{K} I_{\{i\in I_k, j\in I_k\}} + 0.4 \sum_{k=1}^{K-1} (I_{\{i=i_k, j\in I_{k+1}\}} + I_{\{i\in I_{k+1}, j=i_k\}})$$

Model 1 has been used in Bickel and Levina (2008a) and Cai and Liu (2011), and Model 2 is similar to the overlapping block diagonal design used in Rothman (2011).

First, we compare the run times of our estimator $\hat{\mathbf{\Sigma}}^+$ with the log-barrier estimator $\breve{\mathbf{\Sigma}}^+$ by Rothman (2011). As shown in Table 1, our method is much faster than the log-barrier method.

In what follows, we compare the performance of $\hat{\mathbf{\Sigma}}^+$, $\breve{\mathbf{\Sigma}}^+$ and the soft-thresholding estimator $\hat{\mathbf{\Sigma}}$. For all three regularized estimators, the thresholding parameter was chosen by 5-fold cross-validation (Bickel and Levina, 2008b; Rothman et al., 2009; Cai and Liu, 2011). The estimation performance is measured by the average losses under both the Frobenius norm and the spectral norm. The selection performance is examined by the false positive rate

$$\frac{\#\{(i,j) :\ \hat{\sigma}_{ij} \neq 0 \ \&\ \sigma_{ij} = 0\}}{\#\{(i,j) :\ \sigma_{ij} = 0\}}$$

Table 1: Total time (in seconds) for computing a solution path with 99 thresholding parameters $\lambda = \{0.01, 0.02, \cdots, 0.99\}$. Timing was carried out on an `AMD` 2.8GHz processor.

	Model 1			Model 2		
p	100	200	500	100	200	500
Our method	9.2	65.2	1156.0	7.5	51.1	986.6
Rothman's method	84.1	822.1	35911.8	51.3	611.1	32803.0

and the true positive rate

$$\frac{\#\{(i,j):\ \hat{\sigma}_{ij} \neq 0 \ \&\ \sigma_{ij} \neq 0\}}{\#\{(i,j):\ \sigma_{ij} \neq 0\}}.$$

Moreover, we compare the average number of negative eigenvalues and the percentage of positive-definiteness to check the positive-definiteness.

Table 2 and Table 3 show the average metrics over 100 replications. The soft-thresholding estimator $\hat{\mathbf{\Sigma}}$ is positive definite in 19 or fewer out of 100 simulation runs, while $\hat{\mathbf{\Sigma}}^{+}$ and $\breve{\mathbf{\Sigma}}^{+}$ can always guarantee a positive-definite estimator. The larger the dimension, the less likely for the soft-thresholding estimator to be positive definite. In terms of estimation, both $\hat{\mathbf{\Sigma}}^{+}$ and $\breve{\mathbf{\Sigma}}^{+}$ are more accurate than $\hat{\mathbf{\Sigma}}$. As for the selection performance, $\hat{\mathbf{\Sigma}}^{+}$ and $\breve{\mathbf{\Sigma}}^{+}$ achieve a slightly better true positive rate than $\hat{\mathbf{\Sigma}}$. Overall, $\hat{\mathbf{\Sigma}}^{+}$ is the best among all three regularized estimators.

3.2 Real data

To demonstrate our proposal we further consider two gene expression datasets: one from a small round blue-cell tumors microarray experiment (Khan et al., 2001) and the other one from a cardiovascular microarray study (Efron, 2009, 2010). The first dataset has 64 training tissue samples with four types of tumors (23 EWS, 8 BL-NHL, 12 NB, and 21 RMS), and 6567 gene expression values for each sample. We applied the pre-filtering step used in Khan et al. (2001) and then picked the top 40 and bottom 160 genes based on the F-statistic

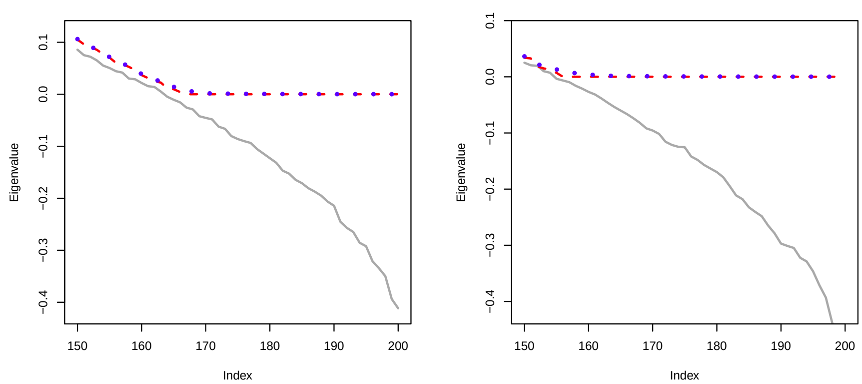
Table 2: Comparison of the three regularized estimators for Model 1. Each metric is averaged over 100 replications with the standard error shown in the bracket. NA means that the results for $\breve{\boldsymbol{\Sigma}}^{+}$ (Rothman's method) are not available due to the extremely long run times.

	Frobenius norm	Spectral norm	False positive	True positive	Negative eigenvalues	Positive definiteness
			$p = 100$			
soft thresholding	8.41 (0.06)	4.02 (0.04)	24.5 (0.1)	87.6 (0.0)	2.24 (0.14)	53/100
our method	8.40 (0.06)	4.02 (0.04)	24.8 (0.1)	87.8 (0.0)	0.00 (0.00)	100/100
Rothman's method	8.40 (0.06)	4.02 (0.04)	24.5 (0.1)	87.7 (0.0)	0.00 (0.00)	100/100
			$p = 200$			
soft thresholding	13.82 (0.06)	4.70 (0.03)	14.3 (0.4)	83.2 (0.3)	3.74 (0.22)	23/100
our method	13.80 (0.06)	4.69 (0.03)	14.6 (0.4)	83.5 (0.3)	0.00 (0.00)	100/100
Rothman's method	13.81 (0.05)	4.69 (0.03)	14.6 (0.4)	83.5 (0.3)	0.00 (0.00)	100/100
			$p = 500$			
soft thresholding	25.15 (0.11)	5.28 (0.04)	6.3 (0.2)	78.1 (0.3)	4.64 (0.60)	7/100
our method	25.10 (0.11)	5.28 (0.04)	6.5 (0.2)	78.3 (0.3)	0.00 (0.00)	100/100
Rothman's method	NA NA	NA NA	NA NA	NA NA	NA NA	NA

Table 3: Comparison of the three regularized estimators for Model 2. Each metric is averaged over 100 replications with the standard error shown in the bracket. NA means that the results for $\breve{\boldsymbol{\Sigma}}^{+}$ (Rothman's method) are not available due to the extremely long run times.

	Frobenius norm	Spectral norm	False positive	True positive	Negative eigenvalues	Positive definiteness
	$p = 100$					
soft thresholding	9.81 (0.07)	4.87 (0.05)	29.5 (0.0)	97.2 (0.0)	1.54 (0.14)	19/100
our method	9.78 (0.07)	4.85 (0.05)	30.2 (0.0)	97.3 (0.0)	0.00 (0.00)	100/100
Rothman's method	9.78 (0.07)	4.85 (0.05)	30.0 (0.0)	97.3 (0.0)	0.00 (0.00)	100/100
	$p = 200$					
soft thresholding	15.95 (0.12)	5.90 (0.06)	17.1 (0.4)	94.1 (0.3)	3.93 (0.27)	7/100
our method	15.81 (0.12)	5.84 (0.06)	18.8 (0.3)	95.0 (0.3)	0.00 (0.00)	100/100
Rothman's method	15.83 (0.12)	5.85 (0.06)	18.3 (0.4)	94.6 (0.3)	0.00 (0.00)	100/100
	$p = 500$					
soft thresholding	29.46 (0.18)	6.92 (0.07)	7.6 (0.1)	87.7 (0.5)	3.84 (0.78)	4/100
our method	29.17 (0.20)	6.84 (0.06)	8.7 (0.2)	88.8 (0.6)	0.00 (0.00)	100/100
Rothman's method	NA NA	NA NA	NA NA	NA NA	NA NA	NA

as done in Rothman et al. (2009). The second dataset has 63 subjects with 44 healthy controls and 19 cardiovascular patients, and 20426 genes measured for each subject. We used the F-statistic to pick the top 50 and bottom 150 genes. By doing so, it is expected that there is weak dependence between the top and the bottom genes. We considered the soft-thresholding estimator (Bickel and Levina, 2008b), the log-barrier estimator (Rothman, 2011) and our estimator. For all three estimators, the thresholding parameter was chosen by 5-fold cross validation.

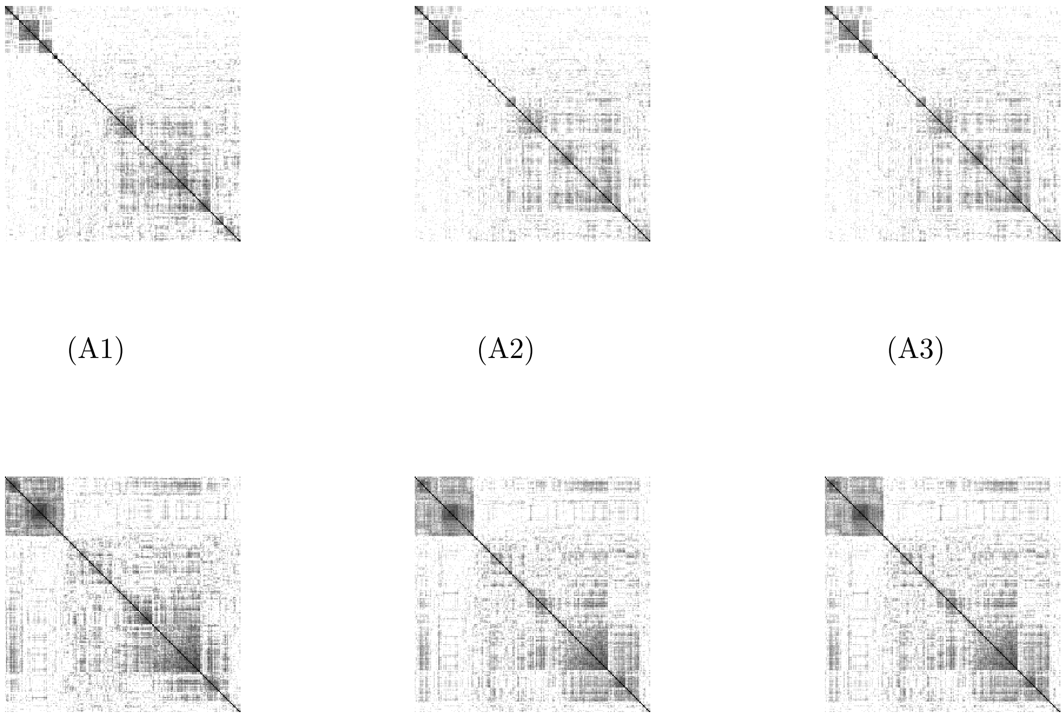


(A) (B)

Figure 2: Plots of the bottom 50 eigenvalues of all three regularized estimators for the small round blue-cell data (A) and the cardiovascular data (B): $\hat{\mathbf{\Sigma}}$ (solid), $\hat{\mathbf{\Sigma}}^+$ (dashed) and $\breve{\mathbf{\Sigma}}^+$ (dotted).

As evidenced in Plot 2, the soft-thresholding estimator yields an indefinite matrix for both real examples whereas the other two regularized estimators guarantee the positive-definiteness. The soft-thresholding estimator contains 37 negative eigenvalues in the small round blue-cell data, and 46 negative eigenvalues in the cardiovascular data. Regularized correlation matrix estimation has a natural application in clustering when the dissimilarity measure is constructed using the correlation among features. For both datasets we did hierarchical clustering using the three regularized estimators. The heat maps are shown in Figure 3 in which the estimated sparsity pattern well matches the expected sparsity pattern.

Finally, we compared the average run times over 5 cross validations for both $\hat{\mathbf{\Sigma}}^+$ and $\breve{\mathbf{\Sigma}}^+$,



(A1) (A2) (A3)

(B1) (B2) (B3)

Figure 3: Heat maps of the absolute values of three regularized sample correlation matrix estimator for the small round blue-cell data (A) and the cardiovascular data (B): $\hat{\mathbf{\Sigma}}$ (A1, B1), $\hat{\mathbf{\Sigma}}^{+}$ (A2, B2) and $\breve{\mathbf{\Sigma}}^{+}$ (A3, B3). The genes are ordered by hierarchical clustering using the estimated correlations.

Table 4: Total time (in seconds) for computing a solution path with 99 thesholding parameters. Timing was carried out on an `AMD` 2.8GHz processor.

	Blue cell data	Cardiovascular data
Our method	74.7	66.3
Rothman's method	1302.7	1575.3

as shown in Table 4. It is obvious that our proposal is much more efficient.

4 Theoretical properties

4.1 Convergence analysis of the algorithm

In this section, we prove that the sequence $(\mathbf{\Theta}^i, \mathbf{\Sigma}^i, \mathbf{\Lambda}^i)$ produced by the alternating direction method (Algorithm 1) converges to $(\hat{\mathbf{\Theta}}^+, \hat{\mathbf{\Sigma}}^+, \hat{\mathbf{\Lambda}}^+)$, where $(\hat{\mathbf{\Theta}}^+, \hat{\mathbf{\Sigma}}^+)$ is an optimal solution of (5) and $\hat{\mathbf{\Lambda}}^+$ is the optimal dual variable. This automatically implies that Algorithm 1 gives an optimal solution of (3).

We define some necessary notation for ease of presentation. Let G be a $2p$ by $2p$ matrix defined as

$$G = \begin{pmatrix} \mu \boldsymbol{I}_{p\times p} & 0 \\ 0 & (1/\mu)\boldsymbol{I}_{p\times p} \end{pmatrix}.$$

Define the norm $\|\cdot\|_G^2$ as $\|U\|_G^2 = \langle U, GU\rangle$ and the corresponding inner product $\langle \cdot, \cdot\rangle_G$ as $\langle U, V\rangle_G = \langle U, GV\rangle$. Before we give the main theorem about the global convergence of Algorithm 1, we need the following lemma.

Lemma 1. *Assume that $(\hat{\mathbf{\Theta}}^+, \hat{\mathbf{\Sigma}}^+)$ is an optimal solution of* (5) *and $\hat{\mathbf{\Lambda}}^+$ is the corresponding optimal dual variable associated with the equality constraint $\mathbf{\Sigma} = \mathbf{\Theta}$. Then the sequence $\{(\mathbf{\Theta}^i, \mathbf{\Sigma}^i, \mathbf{\Lambda}^i)\}$ produced by Algorithm 1 satisfies*

$$\|U^i - U^*\|_G^2 - \|U^{i+1} - U^*\|_G^2 \geq \|U^i - U^{i+1}\|_G^2, \tag{11}$$

where $U^* = (\hat{\mathbf{\Lambda}}^+,\ \hat{\mathbf{\Sigma}}^+)^{\mathrm{T}}$ and $U^i = (\mathbf{\Lambda}^i,\ \mathbf{\Sigma}^i)^{\mathrm{T}}$.

Now we are ready to give the main convergence result of Algorithm 1.

Theorem 1. *The sequence $\{(\mathbf{\Theta}^i, \mathbf{\Sigma}^i, \mathbf{\Lambda}^i)\}$ produced by Algorithm 1 from any starting point converges to an optimal solution of* (5).

4.2 Statistical analysis of the estimator

Define $\boldsymbol{\Sigma}^0$ as the true covariance matrix for the observations $\boldsymbol{X} = (X_{ij})_{n\times p}$, and define the active set of $\boldsymbol{\Sigma}^0 = (\sigma^0_{jk})_{1\le j,k\le p}$ as $A_0 = \{(j,k) : \sigma^0_{jk} \neq 0, j \neq k\}$ with the cardinality $s = |A_0|$. Denote by $\boldsymbol{B}_{A_0}$ the Hadamard product $\boldsymbol{B}_{p\times p} \circ (I_{\{(j,k)\in A_0\}})_{1\le j,k\le p} = (b_{jk} \cdot I_{\{(j,k)\in A_0\}})_{1\le j,k\le p}$. Define $\sigma_{\max} = \max_j \sigma^o_{jj}$ as the maximal true variance in $\boldsymbol{\Sigma}^0$.

Theorem 2. *Assume that the true covariance matrix $\boldsymbol{\Sigma}^0$ is positive definite.*

(i) Under the exponential-tail condition that for all $|t| \le \eta$ and $1 \le i \le n, 1 \le j \le p$

$$E\{\exp(tX^2_{ij})\} \le K_1,$$

we also assume that $\log p \le n$. For any $M > 0$, we pick the thresholding parameter as

$$\lambda = c_0^2 \frac{\log p}{n} + c_1 \left(\frac{\log p}{n}\right)^{1/2},$$

where

$$c_0 = \frac{1}{2} e K_1 \eta^{1/2} + \eta^{-1/2}(M+1)$$

and

$$c_1 = 2K_1(\eta^{-1} + \frac{1}{4}\eta\sigma^2_{\max})\exp(\frac{1}{2}\eta\sigma_{\max}) + 2\eta^{-1}(M+2).$$

With probability at least $1 - 3p^{-M}$, we have

$$\|\hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^0\|_F \le 5\lambda(s+p)^{1/2}.$$

(ii) Under the polynomial-tail condition that for all $\gamma > 0$, $\varepsilon > 0$and $1 \le i \le n, 1 \le j \le p$

$$E\{|X_{ij}|^{4(1+\gamma+\varepsilon)}\} \le K_2,$$

we also assume that $p \le cn^\gamma$ for some $c > 0$. For any $M > 0$, we pick the thresholding parameter as

$$\lambda = 8(K_2+1)(M+1)\frac{\log p}{n} + 8(K_2+1)(M+2)\left(\frac{\log p}{n}\right)^{1/2},$$

With probability at least $1 - O(p^{-M}) - 3K_2 p(\log n)^{2(1+\gamma+\varepsilon)} n^{-\gamma-\varepsilon}$, we have

$$\|\hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^0\|_F \le 5\lambda(s+p)^{1/2}.$$

Define $d = \max_j \sum_k I_{\{\sigma_{jk} \neq 0\}}$ and assume that $\sigma_{\max}$ is bounded by a fixed constant, then we can pick $\lambda = O((\log p/n)^{1/2})$ to achieve the minimax optimal rate of convergence under the Frobenius norm as in Theorem 4 of Cai and Zhou (2011b) that

$$\frac{1}{p}\|\hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^0\|_F^2 = O_p\left((1 + \frac{s}{p})\frac{\log p}{n}\right) = O_p\left(d\frac{\log p}{n}\right).$$

However, to attain the same rate in the presence of the log-determinant barrier term, Rothman (2011) instead would require that $\sigma_{\min}$, the minimal eigenvalue of the true covariance matrix, should be bounded away from zero by some positive constant, and also that the barrier parameter should be bounded by some positive quantity. We would like to point out that if $\sigma_{\min}$ is bounded away from zero, then the soft-thresholding estimator $\hat{\mathbf{\Sigma}}_{st}$ will be positive-definite with an overwhelming probability tending to 1, (Bickel and Levina, 2008b; Cai and Zhou, 2011a,b). Therefore the theory requiring a lower bound on $\sigma_{\min}$ is not very appealing.

5 Conclusions

The soft-thresholding estimator has been shown to enjoy good asymptotic properties for estimating large sparse covariance matrices. But its positive definiteness property can be easily violated, which means the soft-thresholding estimator could be in principle an inadmissible estimator for covariance matrices. In this paper we have put the soft-thresholding estimator in a convex optimization framework and considered a natural modification by imposing the positive definiteness constraint. We have developed a fast alternating direction method to solve the constrained optimization problem and the resulting estimator retains the sparsity and positive definiteness properties simultaneously. The algorithm and the new estimator are supported by numerical and theoretical results.

Acknowledgement

We thank Adam Rothman for sharing his code. Shiqian Ma's research is supported by the National Science Foundation postdoctoral fellowship through Institute for Mathematics and Its Applications at University of Minnesota. Hui Zou's research is supported in part by grants from the National Science Foundation and the Office of Naval Research.

Appendix: Technical Proofs

Proof of Lemma 1. Since $(\hat{\mathbf{\Theta}}^+, \hat{\mathbf{\Sigma}}^+, \hat{\mathbf{\Lambda}}^+)$ is optimal to (5), it follows from the KKT conditions that the followings hold.

$$(-\hat{\mathbf{\Lambda}}^+ - \hat{\mathbf{\Sigma}}^+ + \hat{\mathbf{\Sigma}}_n)_{j\ell}/\lambda \in \partial|\hat{\mathbf{\Sigma}}^+_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p \text{ and } j \neq \ell, \tag{12}$$

$$(\hat{\mathbf{\Sigma}}^+ - \hat{\mathbf{\Sigma}}_n)_{jj} + \hat{\mathbf{\Lambda}}^+_{jj} = 0, \quad \forall j = 1, \ldots, p, \tag{13}$$

$$\hat{\mathbf{\Theta}}^+ = \hat{\mathbf{\Sigma}}^+, \tag{14}$$

$$\hat{\mathbf{\Theta}}^+ \succeq \epsilon \boldsymbol{I}, \tag{15}$$

and

$$\langle \hat{\mathbf{\Lambda}}^+, \mathbf{\Theta} - \hat{\mathbf{\Theta}}^+ \rangle \leq 0, \quad \forall \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{16}$$

Note that the optimality conditions for the first subproblem in Algorithm 1, i.e. the subproblem with respect to $\mathbf{\Theta}$ in (8), are given by

$$\langle \mathbf{\Lambda}^i - (\mathbf{\Theta}^{i+1} - \mathbf{\Sigma}^i)/\mu, \mathbf{\Theta} - \mathbf{\Theta}^{i+1} \rangle \leq 0, \quad \forall \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{17}$$

Using the updating formula for $\mathbf{\Lambda}^i$ in Algorithm 1, i.e.,

$$\mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i - (\mathbf{\Sigma}^{i+1} - \mathbf{\Theta}^{i+1})/\mu, \tag{18}$$

(17) can be rewritten as

$$\langle \mathbf{\Lambda}^{i+1} - (\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i)/\mu, \mathbf{\Theta} - \mathbf{\Theta}^{i+1} \rangle \leq 0, \quad \forall \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{19}$$

Now by letting $\mathbf{\Theta} = \mathbf{\Theta}^{i+1}$ in (16) and $\mathbf{\Theta} = \hat{\mathbf{\Theta}}^+$ in (19), we can get that

$$\langle \hat{\mathbf{\Lambda}}^+, \mathbf{\Theta}^{i+1} - \hat{\mathbf{\Theta}}^+ \rangle \leq 0, \tag{20}$$

and

$$\langle \mathbf{\Lambda}^{i+1} - (\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i)/\mu, \hat{\mathbf{\Theta}}^+ - \mathbf{\Theta}^{i+1} \rangle \leq 0. \tag{21}$$

Summing (20) and (21) yields

$$\langle \mathbf{\Theta}^{i+1} - \hat{\mathbf{\Theta}}^+, (\mathbf{\Lambda}^{i+1} - \hat{\mathbf{\Lambda}}^+) + (\mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1})/\mu \rangle \geq 0. \tag{22}$$

The optimality conditions for the second subproblem in Algorithm 1, i.e., the subproblem with respect to $\mathbf{\Sigma}$ in (8) are given by

$$0 \in (\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}_n)_{j\ell} + \lambda \partial |\mathbf{\Sigma}^{i+1}_{j\ell}| + \mathbf{\Lambda}^i_{j\ell} + (\mathbf{\Sigma}^{i+1} - \mathbf{\Theta}^{i+1})_{j\ell}/\mu, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{23}$$

and

$$(\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}_n)_{jj} + \mathbf{\Lambda}^i_{jj} + (\mathbf{\Sigma}^{i+1} - \mathbf{\Theta}^{i+1})_{jj}/\mu = 0, \quad \forall j = 1, \ldots, p. \tag{24}$$

Note that by using (18), (23) and (24) can be respectively rewritten as:

$$(-\mathbf{\Lambda}^{i+1} - \mathbf{\Sigma}^{i+1} + \hat{\mathbf{\Sigma}}_n)_{j\ell}/\lambda \in \partial |\mathbf{\Sigma}^{i+1}_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{25}$$

and

$$(\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}_n)_{jj} + \mathbf{\Lambda}^{i+1}_{jj} = 0, \quad \forall j = 1, \ldots, p. \tag{26}$$

Using the fact that $\partial | \cdot |$ is a monotone function, (12), (13), (25) and (26) imply

$$\langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+, (\hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1}) + (\hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^{i+1}) \rangle \geq 0. \tag{27}$$

The summation of (22) and (27) gives

$$\begin{aligned} &\langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+, \hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1} \rangle + \langle \hat{\mathbf{\Theta}}^+ - \mathbf{\Theta}^{i+1}, \hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1} \rangle \\ &+ \langle \hat{\mathbf{\Theta}}^+ - \mathbf{\Theta}^{i+1}, \mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i \rangle/\mu \geq \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|_F^2. \end{aligned} \tag{28}$$

Combining (28) with $\boldsymbol{\Theta}^{i+1} = \mu(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}) + \boldsymbol{\Sigma}^{i+1}$ and $\hat{\boldsymbol{\Theta}}^+ = \hat{\boldsymbol{\Sigma}}^+$ leads to

$$\begin{aligned}&\langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Lambda}}^+ - \boldsymbol{\Lambda}^{i+1}\rangle + \langle \hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^{i+1} - \mu(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}), \hat{\boldsymbol{\Lambda}}^+ - \boldsymbol{\Lambda}^{i+1}\rangle \\ &+\langle \hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^{i+1} - \mu(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}), \boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i\rangle/\mu \geq \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2.\end{aligned} \tag{29}$$

Simple algebraic derivation from (29) yields the following inequality:

$$\begin{aligned}&\mu\langle \boldsymbol{\Lambda}^{i+1} - \hat{\boldsymbol{\Lambda}}^+, \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}\rangle + \langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle\mu \\ &\geq \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - \langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle.\end{aligned} \tag{30}$$

Rearranging the terms on the left hand side of (30) using $\hat{\boldsymbol{\Theta}}^+ - \boldsymbol{\Theta}^{i+1} = (\hat{\boldsymbol{\Theta}}^+ - \boldsymbol{\Theta}^i) + (\boldsymbol{\Theta}^i - \boldsymbol{\Theta}^{i+1})$ and $\hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^{i+1} = (\hat{\boldsymbol{\Sigma}}^+ - \boldsymbol{\Sigma}^i) + (\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1})$, then (28) can be reduced to

$$\begin{aligned}&\mu\langle \boldsymbol{\Lambda}^i - \hat{\boldsymbol{\Lambda}}^+, \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}\rangle + \langle \boldsymbol{\Sigma}^i - \hat{\boldsymbol{\Sigma}}^+, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle/\mu \\ \geq\ &\mu\|\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}\|_F^2 + \|\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\|_F^2/\mu + \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - \langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle.\end{aligned} \tag{31}$$

Using the notation of U^i and U^*, (31) can be rewritten as

$$\langle U^i - U^*, U^i - U^{i+1}\rangle_G \geq \|U^i - U^{i+1}\|_G^2 + \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - \langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle. \tag{32}$$

Combining (32) with the following identity

$$\|U^{i+1} - U^*\|_G^2 = \|U^{i+1} - U^i\|_G^2 - 2\langle U^k - U^{i+1}, U^i - U^*\rangle_G + \|U^i - U^*\|_G^2,$$

we get

$$\begin{aligned}&\|U^i - U^*\|_G^2 - \|U^{i+1} - U^*\|_G^2 \\ =\ &2\langle U^i - U^{i+1}, U^i - U^*\rangle - \|U^{i+1} - U^i\|_G^2 \\ \geq\ &2\|U^i - U^{i+1}\|_G^2 + 2\|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|^2 - 2\langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle - \|U^{i+1} - U^i\|_G^2 \\ =\ &\|U^i - U^{i+1}\|_G^2 + 2\|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|^2 - 2\langle \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle.\end{aligned} \tag{33}$$

Now, using (25) and (26) for i instead of $i+1$, we get,

$$(-\boldsymbol{\Lambda}^i - \boldsymbol{\Sigma}^i + \hat{\boldsymbol{\Sigma}}_n)_{j\ell}/\lambda \in \partial|\boldsymbol{\Sigma}^i_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{34}$$

and

$$(\boldsymbol{\Sigma}^i - \hat{\boldsymbol{\Sigma}}_n)_{jj} + \boldsymbol{\Lambda}^i_{jj} = 0, \quad \forall j = 1, \ldots, p. \tag{35}$$

Combining (25), (26), (34), (35) and using the fact that $\partial|\cdot|$ is a monotone function, we obtain,

$$\langle \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}, \mathbf{\Lambda}^{i+1} - \mathbf{\Lambda}^i + \mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i \rangle \geq 0,$$

which immediately implies,

$$\langle \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}, \mathbf{\Lambda}^{i+1} - \mathbf{\Lambda}^i \rangle \geq \|\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i\|_F^2 \geq 0. \tag{36}$$

By substituting (36) into (33), we get the desired result (11). □

Proof of Theorem 1. From Lemma 1 we can easily get that

(i) $\|U^i - U^{i+1}\|_G \to 0$;

(ii) $\{U^i\}$ lies in a compact region;

(iii) $\|U^i - U^*\|_G^2$ is monotonically non-increasing and thus converges.

It follows from (i) that $\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1} \to 0$ and $\mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1} \to 0$. Then (18) implies that $\mathbf{\Theta}^i - \mathbf{\Theta}^{i+1} \to 0$ and $\mathbf{\Theta}^i - \mathbf{\Sigma}^i \to 0$. From (ii) we obtain that, U^i has a subsequence $\{U^{i_j}\}$ that converges to $\bar{U} = (\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Sigma}})$, i.e., $\mathbf{\Lambda}^{i_j} \to \bar{\mathbf{\Lambda}}$ and $\mathbf{\Sigma}^{i_j} \to \bar{\mathbf{\Sigma}}$. From $\mathbf{\Theta}^i - \mathbf{\Sigma}^i \to 0$ we also get that $\mathbf{\Theta}^{i_j} \to \bar{\mathbf{\Theta}} := \bar{\mathbf{\Sigma}}$. Therefore, $(\bar{\mathbf{\Theta}}, \bar{\mathbf{\Sigma}}, \bar{\mathbf{\Lambda}})$ is a limit point of $\{(\mathbf{\Theta}^i, \mathbf{\Sigma}^i, \mathbf{\Lambda}^i)\}$.

Note that (25) and (24) respectively imply that

$$(-\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}} + \hat{\mathbf{\Sigma}}_n)_{j\ell}/\lambda \in \partial|\bar{\mathbf{\Sigma}}_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{37}$$

and

$$(\bar{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_n)_{jj} + \bar{\mathbf{\Lambda}}_{jj} = 0, \quad \forall j = 1, \ldots, p, \tag{38}$$

and (19) implies that

$$\langle \bar{\mathbf{\Lambda}}, \mathbf{\Theta} - \bar{\mathbf{\Theta}} \rangle \leq 0, \quad \forall \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{39}$$

(37), (38) and (39) together with $\bar{\mathbf{\Theta}} = \bar{\mathbf{\Sigma}}$ mean that $(\bar{\mathbf{\Theta}}, \bar{\mathbf{\Sigma}}, \bar{\mathbf{\Lambda}})$ is an optimal solution to (5). Therefore, we showed that any limit point of $\{(\mathbf{\Theta}^i, \mathbf{\Sigma}^i, \mathbf{\Lambda}^i)\}$ is an optimal solution to (5). □

Proof of Theorem 2. Without loss of generality, we may always assume that $E(X_{ij}) = 0$ for all $1 \leq i \leq n, 1 \leq j \leq p$. By the condition that $\mathbf{\Sigma}^0$ is positive definite, we can always choose some very small $\epsilon > 0$ such that ϵ is smaller than the minimal eigenvalue of $\mathbf{\Sigma}^0$. We introduce $\mathbf{\Delta} = \mathbf{\Sigma} - \mathbf{\Sigma}^0$, and then we can write (3) in terms of $\mathbf{\Delta}$ as follows,

$$\hat{\mathbf{\Delta}} = \arg \min_{\mathbf{\Delta}:\mathbf{\Delta}=\mathbf{\Delta}^{\mathrm{T}}, \mathbf{\Delta}+\mathbf{\Sigma}^0 \succeq \epsilon \boldsymbol{I}} \frac{1}{2}\|\mathbf{\Delta} + \mathbf{\Sigma}^0 - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Delta} + \mathbf{\Sigma}^0|_1 \quad (\equiv F(\mathbf{\Delta})).$$

Note that it is easy to see that $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^0$.

Now we consider $\mathbf{\Delta} \in \{\mathbf{\Delta} : \mathbf{\Delta} = \mathbf{\Delta}^{\mathrm{T}}, \mathbf{\Delta} + \mathbf{\Sigma}^0 \succeq \epsilon \boldsymbol{I}, \|\mathbf{\Delta}\|_F = 5\lambda s^{1/2}\}$. Under the probability event $\{|\hat{\sigma}_{ij}^n - \sigma_{ij}^0| \leq \lambda, \ \forall (i,j)\}$, we have

$$\begin{aligned}
F(\mathbf{\Delta}) - F(\mathbf{0}) &= \frac{1}{2}\|\mathbf{\Delta} + \mathbf{\Sigma}^0 - \hat{\mathbf{\Sigma}}_n\|_F^2 - \frac{1}{2}\|\mathbf{\Sigma}^0 - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Delta} + \mathbf{\Sigma}^0|_1 - \lambda|\mathbf{\Sigma}^0|_1 \\
&= \frac{1}{2}\|\mathbf{\Delta}\|_F^2 + < \mathbf{\Delta}, \mathbf{\Sigma}^0 - \hat{\mathbf{\Sigma}}_n > + \lambda|\mathbf{\Delta}_{A_0^c}|_1 + \lambda(|\mathbf{\Delta}_{A_0} + \mathbf{\Sigma}_{A_0}^0|_1 - |\mathbf{\Sigma}_{A_0}^0|_1) \\
&\geq \frac{1}{2}\|\mathbf{\Delta}\|_F^2 - \lambda(|\mathbf{\Delta}|_1 + \sum_i |\Delta_{ii}|) + \lambda|\mathbf{\Delta}_{A_0^c}|_1 - \lambda|\mathbf{\Delta}_{A_0}|_1 \\
&\geq \frac{1}{2}\|\mathbf{\Delta}\|_F^2 - 2\lambda(|\mathbf{\Delta}_{A_0}|_1 + \sum_i |\Delta_{ii}|) \\
&\geq \frac{1}{2}\|\mathbf{\Delta}\|_F^2 - 2\lambda(s+p)^{1/2}\|\mathbf{\Delta}\|_F \\
&\geq \frac{5}{2}\lambda^2(s+p) \\
&> 0
\end{aligned}$$

Note that $\hat{\mathbf{\Delta}}$ is also the optimal solution to the following convex optimization problem

$$\hat{\mathbf{\Delta}} = \arg \min_{\mathbf{\Delta}:\mathbf{\Delta}=\mathbf{\Delta}^{\mathrm{T}}, \mathbf{\Delta}+\mathbf{\Sigma}^0 \succeq \epsilon \boldsymbol{I}} F(\mathbf{\Delta}) - F(\mathbf{0}) \quad (\equiv G(\mathbf{\Delta})).$$

Under the same probability event, $\|\hat{\mathbf{\Delta}}\|_F \leq 5\lambda(s+p)^{1/2}$ would always hold. Otherwise, the fact that $G(\mathbf{\Delta}) > 0$ for $\|\mathbf{\Delta}\|_F = 5\lambda(s+p)^{1/2}$ should contradict with the convexity of $G(\cdot)$ and $G(\hat{\mathbf{\Delta}}) \leq G(\mathbf{0}) = 0$. Therefore, we can obtain the following probability bound

$$\Pr(\|\hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^0\|_F \leq 5\lambda(s+p)^{1/2}) \geq 1 - \Pr(\max_{i,j} |\hat{\sigma}_{ij}^n - \sigma_{ij}^0| > \lambda).$$

Now we shall prove the probability bound under the exponential-tail condition. First it is easy to verify two simple inequalities that $1 + u \leq \exp(u) \leq 1 + u + \frac{1}{2}u^2 \exp(|u|)$ and

$v^2\exp(|v|) \le \exp(v^2+1)$. The first inequality can be proved by using the Taylor expansion, and the second one can be easily derived using the obvious facts that $\exp(v^2+1) \ge \exp(2|v|)$ and $\exp(|v|) \ge v^2$.

Let $t_0 = (\eta\frac{\log p}{n})^{1/2}$, $c_0 = \frac{1}{2}eK_1\eta^{1/2} + \eta^{-1/2}(M+1)$ and $\varepsilon_0 = c_0(\frac{\log p}{n})^{1/2}$. For any $M > 0$, we can apply the Markov inequality to obtain that

$$
\begin{aligned}
\Pr(\sum_i X_{ij} > n\varepsilon_0) &\le \exp(-t_0 n\varepsilon_0)\cdot\prod_{i=1}^n E[\exp(t_0X_{ij})] \\
&\le \exp(-t_0 n\varepsilon_0)\cdot\prod_{i=1}^n \left\{1 + \frac{t_0^2}{2}E[X_{ij}^2\exp(t_0|X_{ij}|)]\right\} \\
&\le p^{-c_0\eta^{1/2}}\cdot\exp(\frac{t_0^2}{2}\sum_{i=1}^n E[X_{ij}^2\exp(t_0|X_{ij}|)]) \\
&\le p^{-c_0\eta^{1/2}}\cdot\exp(\frac{t_0^2}{2}\sum_{i=1}^n E[\exp(t_0^2X_{ij}^2+1)]) \\
&\le p^{-c_0\eta^{1/2}}\cdot\exp(\frac{1}{2}eK_1\eta\log p) \quad (= p^{-M-1}),
\end{aligned}
$$

where we apply $\exp(u) \le 1 + u + \frac{1}{2}u^2\exp(|u|)$ in the second inequality and $1+u \le \exp(u)$ in the third inequality, and then use $v^2\exp(|v|) \le \exp(v^2+1)$ in the fourth inequality. Moreover, the simple facts that $E[X_{ij}] = 0$ $(1 \le i \le n)$ and $t_0^2 = \eta\frac{\log p}{n} \le \eta$ are also used.

Let $t_1 = \frac{1}{2}\eta(\frac{\log p}{n})^{1/2}$ and $c_1 = 2K_1(\eta^{-1} + \frac{1}{4}\eta\sigma_{\max}^2)\exp(\frac{1}{2}\eta\sigma_{\max}) + 2\eta^{-1}(M+2)$. Define $\varepsilon_1 = c_1(\frac{\log p}{n})^{1/2}$. For any $M > 0$, we first apply the Cauchy inequality to obtain that

$$
\begin{aligned}
E[X_{ij}^2X_{ik}^2\cdot\exp(\frac{1}{2}\eta|X_{ij}X_{ik}|)] &\le E[X_{ij}^2X_{ik}^2\cdot\exp(\frac{1}{4}\eta(X_{ij}^2+X_{ik}^2))] \\
&\le (E[X_{ij}^4\exp(\eta X_{ij}^2/2)])^{1/2}\cdot(E[X_{ik}^4\exp(\eta X_{ik}^2/2)])^{1/2} \\
&\le 4\eta^{-2}\cdot(E[\exp(\eta X_{ij}^2)])^{1/2}\cdot(E[\exp(\eta X_{ik}^2)])^{1/2} \\
&\le 4K_1\eta^{-2},
\end{aligned}
$$

where we use the simple inequality $\exp(|v|) \ge v^2$ in the third inequality. Then, combining this result with the Cauchy inequality again yields that

$$
\begin{aligned}
&E[(X_{ij}X_{ik} - \sigma_{jk}^0)^2\cdot\exp(t_1|X_{ij}X_{ik} - \sigma_{jk}^0|)] \\
\le\ & 2E[X_{ij}^2X_{ik}^2\cdot\exp(\frac{1}{2}\eta|X_{ij}X_{ik} - \sigma_{jk}^0|)] + 2(\sigma_{jk}^0)^2\cdot E[\exp(\frac{1}{2}\eta|X_{ij}X_{ik} - \sigma_{jk}^0|)]
\end{aligned}
$$

$$
\begin{aligned}
&\leq 8K_1\eta^{-2}\cdot\exp(\frac{1}{2}\eta\sigma_{jk}^0)+2(\sigma_{jk}^0)^2\cdot\exp(\frac{1}{2}\eta\sigma_{jk}^0)\cdot E[\exp(\frac{1}{4}\eta(X_{ij}^2+X_{ik}^2))]\\
&\leq 8K_1\eta^{-2}\cdot\exp(\frac{1}{2}\eta\sigma_{\max})+2\sigma_{\max}^2\cdot\exp(\frac{1}{2}\eta\sigma_{\max})\big(E[\exp(\frac{1}{2}\eta X_{ij}^2)]\big)^{1/2}\cdot\big(E[\exp(\frac{1}{2}\eta X_{ik}^2)]\big)^{1/2}\\
&\leq 2K_1(4\eta^{-2}+\sigma_{\max}^2)\cdot\exp(\frac{1}{2}\eta\sigma_{\max})
\end{aligned}
$$

where we use the fact that $t_1=\frac{1}{2}\eta(\frac{\log p}{n})^{1/2}\leq\frac{1}{2}\eta<\eta$ in the first inequality, and then use $|\sigma_{jk}^0|\leq(\sigma_{jj}^0\sigma_{kk}^0)^{1/2}\leq\sigma_{\max}$ in the third inequality. Now, we can apply the Markov inequality to obtain the following probability bound

$$
\begin{aligned}
&\Pr(\sum_i\{X_{ij}X_{ik}-\sigma_{jk}^0\}>n\varepsilon_1)\\
\leq\ &\exp(-t_1n\varepsilon_1)\cdot\prod_{i=1}^n E\left[\exp(t_1(X_{ij}X_{ik}-\sigma_{jk}^0))\right]\\
\leq\ &p^{-\frac{1}{2}c_1\eta}\cdot\prod_{i=1}^n\left\{1+\frac{1}{2}t_1^2\cdot E\left[(X_{ij}X_{ik}-\sigma_{jk}^0)^2\cdot\exp(t_1|X_{ij}X_{ik}-\sigma_{jk}^0|)\right]\right\}\\
\leq\ &p^{-\frac{1}{2}c_1\eta}\cdot\exp\left(\frac{1}{2}t_1^2\cdot\sum_{i=1}^n E\left[(X_{ij}X_{ik}-\sigma_{jk}^0)^2\cdot\exp(t_1|X_{ij}X_{ik}-\sigma_{jk}^0|)\right]\right)\\
\leq\ &p^{-\frac{1}{2}c_1\eta}\cdot\exp\left(K_1(1+\frac{1}{4}\eta^2\sigma_{\max}^2)\cdot\exp(\frac{1}{2}\eta\sigma_{\max})\cdot\log p\right)\quad(=p^{-M-2}),
\end{aligned}
$$

where we apply $\exp(u)\leq 1+u+\frac{1}{2}u^2\exp(|u|)$ and $E[X_{ij}X_{ik}]=\sigma_{jk}^0$ for $i=1,2,\cdots,n$ in the second inequality, and we use $1+u\leq\exp(u)$ in the third inequality.

Recall that $\lambda=c_0\frac{\log p}{n}+c_1(\frac{\log p}{n})^{1/2}=\varepsilon_0^2+\varepsilon_1$ and

$$
\hat{\sigma}_{jk}^n-\sigma_{jk}^0=(\frac{1}{n}\sum_i X_{ij}X_{ik}-\sigma_{jk}^0)-(\frac{1}{n}\sum_i X_{jk})\cdot(\frac{1}{n}\sum_i X_{ik}).
$$

Therefore, we can complete the probability bound under the exponential-tail condition as follows

$$
\begin{aligned}
\Pr(\max_{j,k}|\hat{\sigma}_{jk}^n-\sigma_{jk}^0|>\lambda)\ &\leq\ p^2\Pr(\sum_i X_{ij}X_{ik}>n(\sigma_{jk}^0+\varepsilon_1))+2p\Pr(\sum_i X_{ij}>n\varepsilon_0)\\
&\leq\ 3p^{-M}.
\end{aligned}
$$

In the sequel we shall prove the probability bound under the polymonial-tail condition. First, we define $c_2=8(K_2+1)(M+1)$ and $\varepsilon_2=c_2(\frac{\log p}{n})^{1/2}$. Define $\delta_n=n^{1/4}(\log n)^{-1/2}$,

$Y_{ij} = X_{ij} I_{\{|X_{ij}| \le \delta_n\}}$ and $Z_{ij} = X_{ij} I_{\{|X_{ij}| > \delta_n\}}$. Then we have $X_{ij} = Y_{ij} + Z_{ij}$ and $E[X_{ij}] = E[Y_{ij}] + E[Z_{ij}]$. By construction, $|Y_{ij}| \le \delta_n$ are bounded random variables, and $E[Z_{ij}]$ are bounded by $o(\varepsilon_2)$ due to the fact that $|E[Z_{ij}]| \le \delta_n^{-3} E[|X_{ij}|^4 I_{\{|X_{ij}|>\delta_n\}}] \le K_2 \delta_n^{-3} = o(\varepsilon_2)$. Now we can apply the Bernstein's inequality (Bernstein, 1946; Bennett, 1962) to obtain that

$$\begin{aligned} \Pr(\sum_i \{Y_{ij} - E[Y_{ij}]\}\} > \frac{1}{2} n\varepsilon_2) &\le \exp\left(\frac{-n\varepsilon_2^2}{8var(Y_{ij}) + \frac{4}{3}\delta_n \varepsilon_2}\right) \\ &\le \exp\left(\frac{-c_2 \log p}{8K_2 + 8 + O(n^{-1/4})}\right) \\ &= O(p^{-M-1}), \end{aligned}$$

where the fact that $var(Y_{ij}) \le E[X_{ij}^2] \le E[X_{ij}^2 I_{\{|X_{ij}| \ge 1\}}] + E[X_{ij}^2 I_{\{|X_{ij}| \le 1\}}] \le K_2 + 1$ is used in the second inequality. Besides, we can apply the Markov inequality to obtain that

$$\Pr(|X_{ij}| > \delta_n) \le \delta_n^{-4(1+\gamma+\varepsilon)} E\left[|X_{ij}|^{4(1+\gamma+\varepsilon)}\right] \le K_2 (\log n)^{2(1+\gamma+\varepsilon)} n^{-1-\gamma-\varepsilon}.$$

Then, we can derive the following probability bound

$$\begin{aligned} \Pr(\sum_i X_{ij} > n\varepsilon_2) &= \Pr(\sum_i \{Y_{ij} + Z_{ij} - E[Y_{ij} + Z_{ij}]\} > n\varepsilon_2) \\ &\le \Pr(\sum_i \{Y_{ij} - E[Y_{ij}]\} > \frac{1}{2} n\varepsilon_2) + \Pr(\sum_i \{Z_{ij} - E[Z_{ij}]\} > \frac{1}{2} n\varepsilon_2) \\ &\le O(p^{-M-1}) + \Pr(\sum_i \{Z_{ij} - o(\varepsilon_2)\} > \frac{1}{2} n\varepsilon_2) \\ &\le O(p^{-M-1}) + \sum_i \Pr(|X_{ij}| > \delta_n) \\ &\le O(p^{-M-1}) + K_2 (\log n)^{2(1+\gamma+\varepsilon)} n^{-\gamma-\varepsilon}. \end{aligned}$$

Let $c_3 = 8(K_2+1)(M+2)$ and $\varepsilon_3 = c_3(\frac{\log p}{n})^{1/2}$. Recall that $\delta_n = (\frac{n}{\log(n)})^{1/4}$, and define $R_{ijk} = X_{ij} X_{ik} I_{\{|X_{ij}| > \delta_n \text{ or } |X_{ik}| > \delta_n\}}$. Then we have $X_{ij} X_{ik} = Y_{ij} Y_{ik} + R_{ijk}$ and $\sigma_{jk}^0 = E[X_{ij} X_{ik}] = E[Y_{ij} Y_{ik}] + E[R_{ijk}]$. By construction, $|Y_{ij} Y_{ik}| \le \delta_n^2$ are bounded random variables, and $E[R_{ijk}]$ is bounded by $o(\varepsilon_3)$ due to the fact that

$$\begin{aligned} |E[R_{ijk}]| &\le |E[X_{ij} X_{ik} I_{\{|X_{ij}| > \delta_n\}}]| + |E[X_{ij} X_{ik} I_{\{|X_{ik}| > \delta_n\}}]| \\ &\le \delta_n^{-2-4\gamma} E[X_{ij}^{4(1+\gamma)} I_{\{|X_{ij}| > \delta_n\}}] \cdot E[X_{ik}^2] + \delta_n^{-2-4\gamma} E[X_{ik}^{4(1+\gamma)} I_{\{|X_{ik}| > \delta_n\}}] \cdot E[X_{ij}^2] \\ &\le 2K_2 \delta_n^{-2-4\gamma} \quad (= o(\varepsilon_3)). \end{aligned}$$

Again, we can apply the Bernstein's inequality to obtain that

$$
\begin{aligned}
\Pr(\sum_i \{Y_{ij}Y_{ik} - E[Y_{ij}Y_{ik}]\} > \frac{1}{2}n\varepsilon_3) &\leq \exp\left(\frac{-n\varepsilon_3^2}{8K_2 + 8 + \frac{4}{3}\delta_n^2\varepsilon_3}\right) \\
&\leq \exp\left(\frac{-c_3 \log p}{8K_2 + 8 + O((\log n)^{-1/2})}\right) \\
&= O(p^{-M-2}),
\end{aligned}
$$

where the fact that $var(Y_{ij}Y_{ik}) \leq E[X_{ij}^2 X_{ik}^2] \leq (E[X_{ij}^4]E[X_{ik}^4])^{1/2} \leq K_2 + 1$ is used.

$$
\begin{aligned}
& \Pr(\max_{j,k} |\sum_i (X_{ij}X_{ik} - \sigma_{jk}^0)| > n\varepsilon_3) \\
\leq\ & \Pr(\max_{j,k} |\sum_i \{Y_{ij}Y_{ik} - E[Y_{ij}Y_{ik}]\}| > \frac{1}{2}n\varepsilon_3) + \Pr(\max_{j,k} |\sum_i \{R_{ijk} - E[R_{ijk}]\}| > \frac{1}{2}n\varepsilon_3) \\
\leq\ & 2\sum_{j,k} \Pr(\sum_i \{Y_{ij}Y_{ik} - E[Y_{ij}Y_{ik}]\} > \frac{1}{2}n\varepsilon_3) + \Pr(\max_{j,k} |\sum_i \{R_{ijk} - o(\varepsilon_3)\}| > \frac{1}{2}n\varepsilon_3) \\
\leq\ & O(p^{-M}) + \sum_{i,j} \Pr(|X_{ij}| > \delta_n) \\
\leq\ & O(p^{-M}) + K_2 p (\log n)^{2(1+\gamma+\varepsilon)} n^{-\gamma-\varepsilon}
\end{aligned}
$$

Recall that $\lambda = c_2 \frac{\log p}{n} + c_3 (\frac{\log p}{n})^{1/2} = \varepsilon_2^2 + \varepsilon_3$. Therefore, we can prove the desired probability bound under the polynomial-tail condition as follows

$$
\begin{aligned}
& \Pr(\max_{j,k} |\hat{\sigma}_{jk}^n - \sigma_{jk}^0| > \lambda) \\
\leq\ & \Pr(\max_{j,k} |\sum_i \{X_{ij}X_{ik} - \sigma_{jk}^0\}| > n\varepsilon_3) + \Pr(\max_j |\sum_i X_{ij}| > n\varepsilon_2) \\
\leq\ & O(p^{-M}) + 3K_2 p (\log n)^{2(1+\gamma+\varepsilon)} n^{-\gamma-\varepsilon}.
\end{aligned}
$$

□

References

Anderson, T. (1984), "An introduction to multivariate statistical analysis", John Wiley & Sons New York.

Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D., et al. (2002), "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Med.*, 8, 816–824.

Bennett, G. (1962), "Probability inequalities for the sum of independent random variables," *J. Amer. Statist. Assoc.*, 33–45.

Bernstein, S. (1946), "The Theory of Probabilities", Gostekhizdat, Moscow.

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

— (2008b), "Covariance regularization by thresholding," *Ann. Statist.*, 36, 2577–2604.

Cai, T. and Liu, W. (2011), "Adaptive thresholding for sparse covariance matrix estimation," *J. Amer. Statist. Assoc.*, 1–13.

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

Cai, T. and Zhou, H. (2011a), "Minimax estimation of large covariance matrices under ℓ_1-norm," *Statist. Sinica, to appear.*

— (2011b), "Optimal rates of convergence for sparse covariance matrix estimation," *Manuscript.*

Douglas, J. and Rachford, H. H. (1956), "On the numerical solution of the heat conduction problem in 2 and 3 space variables," *Trans. Amer. Math. Soc.*, 82, 421–439.

Efron, B. (2009), "Are a set of microarrays independent of each other?" *Ann. Appl. Stat.*, 3, 922–942.

— (2010), "Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction", Cambridge University Press.

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

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Fortin, M. and Glowinski, R. (1983), “Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems(Book)”, North-Holland. Co.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical lasso,” *Biostat.*, 9, 432.

Furrer, R. and Bengtsson, T. (2007), “Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants,” *J. Multivariate Anal.*, 98, 227–255.

Glowinski, R. and Le Tallec, P. (1989), “Augmented Lagrangian and operator-splitting methods in nonlinear mechanics”, SIAM, Philadelphia, Pennsylvania.

Johnstone, I. (2001), “On the distribution of the largest eigenvalue in principal components analysis,” *Ann. Statist.*, 295–327.

Khan, J., Wei, J., Ringnér, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., et al. (2001), “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Med.*, 7, 673–679.

Peaceman, D. H. and Rachford, H. H. (1955), “The numerical solution of parabolic elliptic differential equations,” *SIAM J. Appl. Math.*, 3, 28–41.

Rothman, A. (2011), “Positive definite estimators of large covariance matrices,” *Manuscript.*

Rothman, A., Levina, E., and Zhu, J. (2009), “Generalized thresholding of large covariance matrices,” *J. Amer. Statist. Assoc.*, 104, 177–186.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010), “Sparse inverse covariance selection via alternating linearization methods,” *Adv. Neural Inf. Process. Syst.*

Subramaniana, A., Tamayoa, P., Moothaa, V., Mukherjeed, S., Eberta, B., Gillettea, M., Paulovichg, A., Pomeroyh, S., Goluba, T., Landera, E., et al. (2005), “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci.*, 102, 15545–15550.

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

Zou, H. (2006), “The adaptive lasso and its oracle properties,” *J. Amer. Statist. Assoc.*, 101, 1418–1429.