Positive-Definite ℓ_1-Penalized Estimation of Large Covariance Matrices

Lingzhou XUE, Shiqian MA, and Hui ZOU

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 fix this drawback of thresholding estimation, we develop a positive-definite ℓ_1-penalized covariance estimator for estimating sparse large covariance matrices. We derive an efficient alternating direction method to solve the challenging optimization problem and establish its convergence properties. Under weak regularity conditions, nonasymptotic 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.

KEY WORDS: Alternating direction methods; Matrix norm; Positive-definite estimation; Soft-thresholding; Sparsity.

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 functional magnetic resonance imaging (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 (Marčenko and Pastur 1967; 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 decay 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 2012a). In this sense, thresholding is more robust than banding/tapering for real applications.

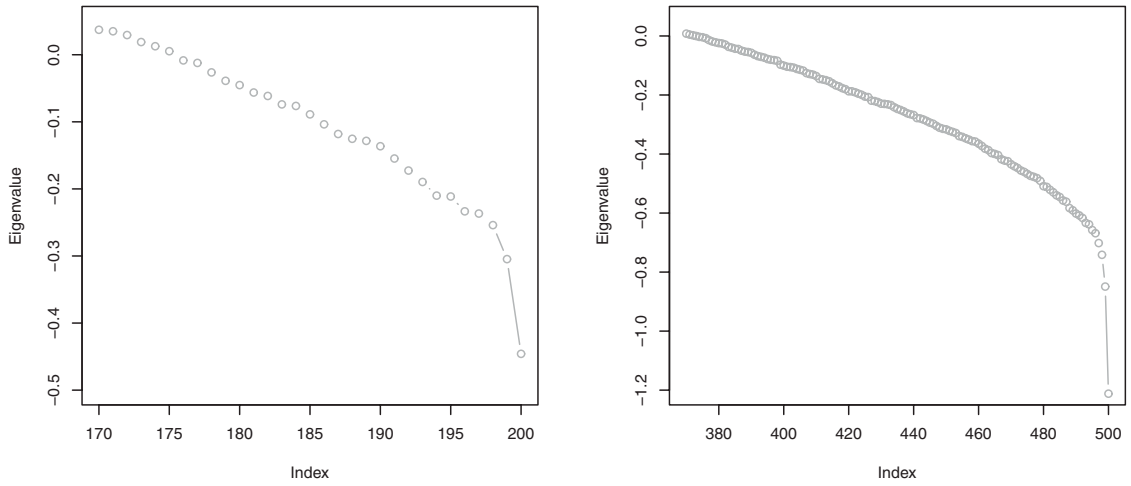
Let $\hat{\mathbf{\Sigma}}_n = (\hat{\sigma}_{ij})_{1\le i,j\le p}$ be the sample covariance matrix. Rothman, Levina, and Zhu (2009) defined the general thresholding covariance matrix estimator as $\hat{\mathbf{\Sigma}}_{\text{thr}} = \{s_\lambda(\hat{\sigma}_{ij})\}_{1\le i,j\le p}$, where $s_\lambda(z)$ is the generalized thresholding function. The generalized thresholding function covers a number of commonly used shrinkage procedures, for example, the hard thresholding $s_\lambda(z) = zI_{\{|z|>\lambda\}}$, the soft thresholding $s_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$, 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, for example, Bickel and Levina (2008a, b), El Karoui (2008), Rothman, Levina, and Zhu (2009), and Cai and Liu (2011). The recent articles by Cai and Zhou (2012a, b) have established the minimax optimality of the thresholding estimator for estimating a wide range of large sparse covariance matrices under commonly used matrix norms. The existing theoretical and empirical results show no clear favoritism to a particular thresholding rule. In this article, 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 off-diagonal elements. Then the soft-thresholding covariance estimator is equal to

$$\hat{\mathbf{\Sigma}} = \arg\min_{\mathbf{\Sigma}} \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Sigma}|_1. \quad (1)$$

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 chose p genes ($p = 200, 500$) and obtained the soft-thresholding sample correlation matrix for these genes. We repeated the process 10 times for $p = 200$ and 500, respectively, and each time the thresholding parameter λ was selected via the fivefold cross-validation. We found that none of these soft-thresholding estimators would become positive definite. On average, there are 22 and 124

Lingzhou Xue is Postdoctoral Research Associate, Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ 08544. Shiqian Ma is Assistant Professor, Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, Hong Kong. Hui Zou is Associate Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *zouxx019@umn.edu*). The article was completed when Lingzhou Xue was a Ph.D. student at the University of Minnesota and Shiqian Ma was a Postdoctoral Fellow in the Institute for Mathematics and Its Applications at the University of Minnesota. The authors thank Adam Rothman for sharing his code. We are grateful to the coeditor, the associate editor, and two referees for their helpful and constructive comments. Shiqian Ma was supported by the National Science Foundation postdoctoral fellowship through the Institute for Mathematics and Its Applications at the University of Minnesota. Lingzhou Xue and Hui Zou are supported in part by grants from the National Science Foundation and the Office of Naval Research.

Journal of the American Statistical Association
December 2012, Vol. 107, No. 500, Theory and Methods
DOI: 10.1080/01621459.2012.725386



(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.

negative eigenvalues for the soft-thresholding estimator for $p = 200$ and 500, respectively. Figure 1 displays the 30 smallest eigenvalues for $p = 200$ and the 130 smallest eigenvalues for $p = 500$.

From both methodological and practical perspectives, the positive-definite property is crucial for any covariance matrix estimator. First of all, any statistical procedure that uses the normal distribution requires a positive-definite covariance matrix, otherwise the density function is ill defined. Two well-known examples are the parametric bootstrap method and the quadratic discriminant analysis. Second, there are important statistical methods that do not use the normal distribution but still could not be carried out without a positive-definite covariance matrix estimator. The celebrated Markowitz portfolio optimization problem is one such example. We will provide more detailed discussion and examples in Section 3.2 to illustrate the importance of a positive-definite covariance matrix estimator in the Markowitz portfolio optimization problem. To deal with the indefiniteness issue in covariance matrix estimation, one possible solution is to use the eigen-decomposition of $\hat{\mathbf{\Sigma}}$ and project $\hat{\mathbf{\Sigma}}$ into the convex cone $\{\mathbf{\Sigma} \succeq 0\}$. Assume that $\hat{\mathbf{\Sigma}}$ has the eigen-decomposition $\hat{\mathbf{\Sigma}} = \sum_{i=1}^{p} \hat{\lambda}_i \boldsymbol{v}_i^{\tau} \boldsymbol{v}_i$, and then a positive semidefinite estimator $\tilde{\mathbf{\Sigma}}^{+}$ can be obtained by setting $\tilde{\mathbf{\Sigma}}^{+} = \sum_{i=1}^{p} \max(\hat{\lambda}_i, 0) \boldsymbol{v}_i^{\tau} \boldsymbol{v}_i$. However, this strategy does not work well for sparse covariance matrix estimation, because the projection destroys the sparsity pattern of $\hat{\mathbf{\Sigma}}$. Consider the Michigan data again, after semidefinite projection, the soft-thresholding 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

$$\hat{\mathbf{\Sigma}}^{+} = \arg\min_{\mathbf{\Sigma} \succeq 0} \frac{1}{2} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda |\mathbf{\Sigma}|_1. \tag{2}$$

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

$$\hat{\mathbf{\Sigma}}^{+} = \arg\min_{\mathbf{\Sigma} \succeq \epsilon \boldsymbol{I}} \frac{1}{2} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda |\mathbf{\Sigma}|_1. \tag{3}$$

It is important to note that ϵ is not a tuning parameter like λ. We simply include ϵ in the procedure to ensure that the smallest eigenvalue of the estimator is at least ϵ. If one knows that the smallest eigenvalue of the true covariance estimator is bounded below by a positive number δ', then ϵ can be δ'. To fix the idea, we use $\epsilon = 10^{-5}$ in all our numerical examples.

Despite its natural motivation, (3) is actually a very challenging optimization problem due to the positive-definite constraint. Rothman (2012) considered a slightly perturbed version of (3) by adding a log-determinant barrier function:

$$\check{\mathbf{\Sigma}}^{+} = \arg\min_{\mathbf{\Sigma} \succeq 0} \frac{1}{2} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 - \tau \log\det(\mathbf{\Sigma}) + \lambda |\mathbf{\Sigma}|_1, \tag{4}$$

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, Hastie, and Tibshirani 2008) that also has a log-determinant part and the elementwise ℓ_1-penalty. Rothman (2012) derived an iterative procedure to solve (4). Rothman's (2012) proposal is based on heuristic arguments and its convergence property is unknown. Another theoretical limitation of Rothman's (2012) proposal is that it requires the smallest eigenvalue of the true covariance matrix to be bounded away from zero, otherwise the influence of the perturbing term due to the log-determinant barrier could not be well controlled.

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

advantage, we also point out the methodological and theoretical advantages of our method over the log-barrier method.

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) and Glowinski and Le Tallec (1989) for more details on operator-splitting and alternating direction methods and a recent survey article by Boyd et al. (2011) for a more complete list of references.

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

$$(\hat{\boldsymbol{\Theta}}^+, \hat{\boldsymbol{\Sigma}}^+) = \arg\min_{\boldsymbol{\Theta},\boldsymbol{\Sigma}} \left\{ \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_n\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1 : \boldsymbol{\Sigma} = \boldsymbol{\Theta}, \boldsymbol{\Theta} \succeq \epsilon \boldsymbol{I} \right\}. \tag{5}$$

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, that is,

$$L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_n\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1 - \langle \boldsymbol{\Lambda}, \boldsymbol{\Theta} - \boldsymbol{\Sigma}\rangle + \frac{1}{2\mu}\|\boldsymbol{\Theta} - \boldsymbol{\Sigma}\|_F^2, \tag{6}$$

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

$$(\boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}^{i+1}) = \arg\min_{\boldsymbol{\Theta}\succeq\epsilon\boldsymbol{I},\boldsymbol{\Sigma}} L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}^i) \tag{7}$$

and then update the Lagrange multiplier $\boldsymbol{\Lambda}^{i+1}$ by

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

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

To sum up, the entire algorithm proceeds as follows:
For $i = 0, 1, 2, \ldots$, perform the following three steps sequentially till convergence:

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

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

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

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

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

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

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

Then the $\boldsymbol{\Sigma}$ step has a closed-form solution given as follows

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

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.

In our implementation, we use the soft-thresholding estimator as the initial value for both $\boldsymbol{\Theta}^0$ and $\boldsymbol{\Sigma}^0$, and we set $\boldsymbol{\Lambda}^0$ as a zero matrix. Note that our convergence result in Theorem 1 shows that Algorithm 1 globally converges for any $\mu > 0$. Unlike λ, μ does not change the final covariance estimator. In our numerical experiments, we fixed $\mu = 2$ just for simplicity. 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).

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

1. Input: μ, $\boldsymbol{\Sigma}^0$, and $\boldsymbol{\Lambda}^0$.
2. Iterative alternating direction augmented Lagrangian step: for the ith iteration
   - 2.1 Solve $\boldsymbol{\Theta}^{i+1} = (\boldsymbol{\Sigma}^i + \mu\boldsymbol{\Lambda}^i)_+$;
   - 2.2 Solve $\boldsymbol{\Sigma}^{i+1} = \frac{1}{1+\mu}\boldsymbol{S}(\mu(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Lambda}^i) + \boldsymbol{\Theta}^{i+1}, \lambda\mu)$;
   - 2.3 Update $\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\mu}(\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Sigma}^{i+1})$.
3. Repeat the above cycle till convergence.

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

	Model 1			Model 2		
p	100	200	500	100	200	500
Soft thresholding	0.2	1.3	5.7	0.1	1.3	5.6
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

3. NUMERICAL EXAMPLES

Before delving into theoretical analysis of our proposed algorithm and estimator, we first demonstrate the computational advantage of our estimator over the log-barrier method and then show an application of our estimator to the Markowitz portfolio selection problem using stock data.

3.1 Comparing Two Positive-Definite Thresholding Estimators

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 $\mathbf{\Sigma}_0 = (\sigma_{ij}^0)_{1 \le i,j \le p}$ for $p = 100$, 200, and 500. We considered two covariance models with different sparsity patterns:

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

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

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

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 that has been used in Rothman (2012).

First, we compare the run times of our estimator $\hat{\mathbf{\Sigma}}^+$ with the log-barrier estimator $\check{\mathbf{\Sigma}}^+$ by Rothman (2012). As shown in Table 1, our method is much faster than the log-barrier method.

In what follows, we compare the performance of $\hat{\mathbf{\Sigma}}^+$, $\check{\mathbf{\Sigma}}^+$, and the soft-thresholding estimator $\hat{\mathbf{\Sigma}}$. For all three regularized estimators, the thresholding parameter was chosen over 99 thresholding parameters $\lambda = \{0.01, 0.02, \ldots, 0.99\}$ by fivefold cross-validation (Bickel and Levina 2008b; Rothman, Levina, and Zhu 2009; Cai and Liu 2011). For $\check{\mathbf{\Sigma}}^+$, we set $\tau = 10^{-4}$ as in Rothman (2012). 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

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

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 over 100 replications and the percentage of positive definiteness to check the positive definiteness.

Tables 2 and 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 $\check{\mathbf{\Sigma}}^+$ 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 $\check{\mathbf{\Sigma}}^+$ are more accurate than $\hat{\mathbf{\Sigma}}$. As for the selection performance, $\hat{\mathbf{\Sigma}}^+$ and $\check{\mathbf{\Sigma}}^+$ achieve a slightly better true positive rate than $\hat{\mathbf{\Sigma}}$. Overall, $\hat{\mathbf{\Sigma}}^+$ is the best among all three regularized estimators.

To show the advantage of our method over Rothman's proposal, we further consider two gene expression datasets: one from a small round blue-cell tumors microarray experiment

Table 2. Comparison of the three regularized estimators for Model 1

	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	

NOTE: Each metric is averaged over 100 replications with the standard error shown in the bracket. NA means that the results for $\check{\mathbf{\Sigma}}^+$ (Rothman's method) are not available due to the extremely long run times.

Table 3. Comparison of the three regularized estimators for Model 2

	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	

NOTE: Each metric is averaged over 100 replications with the standard error shown in the bracket. NA means that the results for $\check{\mathbf{\Sigma}}^+$ (Rothman's method) are not available due to the extremely long run times.

(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 prefiltering step used in Khan et al. (2001) and then picked the top 40 and bottom 160 genes based on the F-statistic as done in Rothman, Levina, and Zhu (2009). The second dataset has 63 subjects with 44 healthy controls and 19 cardiovascular patients, and 20,426 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 2012), and our estimator. For all three estimators, the thresholding parameter was chosen by the fivefold cross-validation.

As evidenced in Figure 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 five cross-validations for both $\hat{\mathbf{\Sigma}}^+$ and $\check{\mathbf{\Sigma}}^+$, as shown in Table 4. It is obvious that our proposal is much more efficient.

3.2 Applications to Markowitz Portfolio Selection

To further support our proposal and illustrate the importance of positive definiteness, we consider the celebrated Markowitz

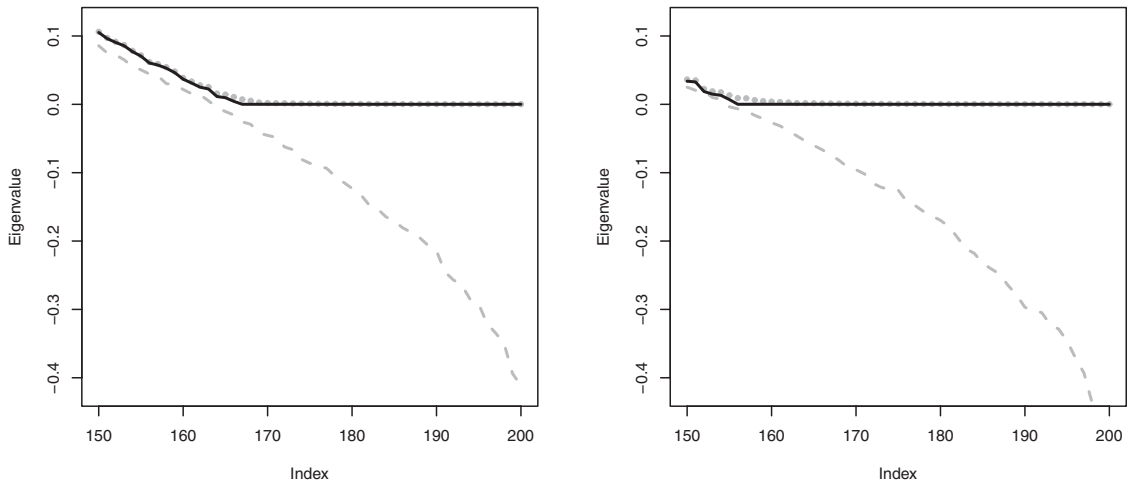


Figure 2. Plots of the bottom 50 eigenvalues of all regularized estimators for (A) the small round blue-cell data and (B) the cardiovascular data: $\hat{\mathbf{\Sigma}}$ (dashed line), $\hat{\mathbf{\Sigma}}^+$ (solid line), and $\check{\mathbf{\Sigma}}^+$ (dotted line).

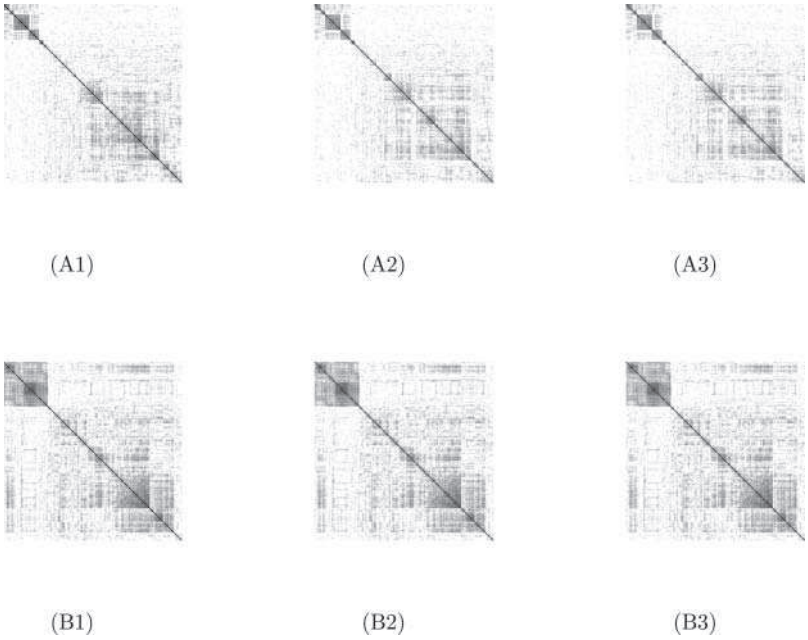


Figure 3. Heat maps of the absolute values of three regularized sample correlation matrix estimator for (A) the small round blue-cell data and (B) the cardiovascular data: $\hat{\mathbf{\Sigma}}$ (A1, B1), $\hat{\mathbf{\Sigma}}^+$ (A2, B2), and $\check{\mathbf{\Sigma}}^+$ (A3, B3). The genes are ordered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

portfolio selection problem (Markowitz 1952) in finance, which constructs the optimal mean-variance efficient portfolios by solving the following quadratic optimization problem:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \boldsymbol{w}'\mathbf{\Sigma}\boldsymbol{w} \quad \text{such that} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_P, \boldsymbol{w}'\boldsymbol{e} = 1, \quad (11)$$

where $\boldsymbol{e}$ is the p-dimensional vector whose entries are all equal to 1. The practical solution to the Markowitz problem is often obtained by solving the empirical version of (11) where the true covariance matrix $\mathbf{\Sigma}$ is replaced with the sample covariance matrix. In the recent literature, many researchers have revisited the classic Markowitz problem under the high-dimensional setting (Jagannathan and Ma 2003; Brodie et al. 2009; DeMiguel et al. 2009; El Karoui 2010; Fan, Zhang, and Yu 2012). In particular, El Karoui (2010) provided a detailed theoretical analysis to show the undesirable risk underestimation issue in the high-dimensional Markowitz problem (11) when $\mathbf{\Sigma}$ is simply estimated by the sample covariance matrix $\hat{\mathbf{\Sigma}}_n$. El Karoui (2010) further suggested using the thresholding covariance estimator to deal with this challenge.

Note that if any empirical estimator of $\mathbf{\Sigma}$ is used in the Markowitz problem, the estimator must be positive definite. Otherwise, the corresponding optimization problem is ill defined. To elaborate, we provide an empirical study to evaluate the performance of the Markowitz problem when $\mathbf{\Sigma}$ is estimated by the sample covariance estimator and the simple thresholding covariance estimator and our proposal, respectively. We considered the monthly stock return data of companies in the S&P 100 index from January 1990 to December 2007. We disregarded 16 companies that were listed after 1990 and only used the other $p = 84$ companies in the S&P 100 index. We followed Jagannathan and Ma (2003) and Brodie et al. (2009) to focus on the Markowitz problem with no-shortsale constraints, that is,

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \boldsymbol{w}'\mathbf{\Sigma}\boldsymbol{w} \quad \text{such that} \quad \boldsymbol{w} \geq 0, \boldsymbol{w}'\boldsymbol{e} = 1. \quad (12)$$

For ease of notation, we denote the Markowitz problem (12) with $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}_n$ as (P1) and the Markowitz problem (12) with $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}^+$ as (P2). The thresholding parameter in $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}^+$ is chosen by the threefold cross-validation. In the sequel, we consider the following strategy to construct two Markowitz portfolios using $n = 36$ historically monthly returns for each month from 1993 to 2007. At the beginning of each month, we use the past 36 months' stock return data to construct portfolios (P1) and (P2), and only the current month returns are recorded for (P1) and (P2). Then we repeat the above process for the next month. There are 180 monthly predictions in total.

We compared the performance of (P1) and (P2) by computing the average monthly return, the standard deviation of monthly returns, and the Sharpe ratio for the time period of 1993–2007. As shown in Table 5, (P2) significantly outperforms (P1) in terms of both returns and volatility. We also applied the log-barrier estimator to the Markowitz problem (12) with $\mathbf{\Sigma} = \check{\mathbf{\Sigma}}_n$, and we found that the outcomes are similar to that of (P2), but the log-barrier method required about 15–20 times more computing time than (P2). Moreover, we want to point out that the results of the simple soft-thresholding estimator are not reported because for the majority of the time period (137 out of 180 months), the soft-thresholding estimator is not positive definite thus cannot be used in the Markowitz problem (12) to produce a portfolio.

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

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

Table 5. Comparison of two Markowitz portfolio selection methods (P1) and (P2) for three different time periods in the monthly stock return data

Mean		Std. Dev.	Sharpe ratio
Time period		Jan. 1993–Dec. 2007	
P1	1.02	3.66	27.73
P2	1.13	3.42	33.06

NOTE: Monthly mean returns, standard deviations of monthly returns, and corresponding Sharpe ratios are all expressed in %.

4. THEORETICAL PROPERTIES

4.1 Convergence Analysis of the Algorithm

In this section, we prove that the sequence $(\boldsymbol{\Theta}^i, \boldsymbol{\Sigma}^i, \boldsymbol{\Lambda}^i)$ produced by the alternating direction method (Algorithm 1) converges to $(\hat{\boldsymbol{\Theta}}^+, \hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Lambda}}^+)$, where $(\hat{\boldsymbol{\Theta}}^+, \hat{\boldsymbol{\Sigma}}^+)$ is an optimal solution of (5) and $\hat{\boldsymbol{\Lambda}}^+$ 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 $2p \times 2p$ matrix defined as

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

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

$$\|U^i - U^*\|_G^2 - \|U^{i+1} - U^*\|_G^2 \geq \|U^i - U^{i+1}\|_G^2, \qquad (13)$$

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

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

Theorem 1. The sequence $\{(\boldsymbol{\Theta}^i, \boldsymbol{\Sigma}^i, \boldsymbol{\Lambda}^i)\}$ 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 $\boldsymbol{\Sigma}^0 = (\sigma^0_{jk})_{1\leq j,k\leq p}$ 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 $\boldsymbol{B}_{p\times p} \circ (I_{\{(j,k)\in A_0\}})_{1\leq j,k\leq p} = (b_{jk} \cdot I_{\{(j,k)\in A_0\}})_{1\leq j,k\leq p}$. 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.

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

$$E\left\{\exp\left(tX_{ij}^2\right)\right\} \leq K_1,$$

we also assume that $\log p \leq n$. 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

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

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

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

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

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

we also assume that $p \leq cn^\gamma$ 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

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

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 (2012b) that

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

However, to attain the same rate in the presence of the log-determinant barrier term, Rothman (2012) 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 $\hat{\boldsymbol{\Sigma}}_{st}$ will be positive definite with an overwhelming probability tending to 1 (Bickel and Levina 2008b; Cai and Zhou 2012a, b). Therefore, the theory requiring a lower bound on $\sigma_{\min}$ is not very appealing.

5. DISCUSSION

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 in practice, which prevents its use in many important applications such as quadratic discriminant analysis and Markowitz portfolio selection. In this article, 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.

The log-determinant barrier method is also a valid technique to achieve positive definiteness in sparse covariance matrix estimation. However, it is still unclear whether the iterative procedure proposed by Rothman (2012) actually finds the right solution to the log-determinant barrier perturbed criterion in (4), although the code seems to produce a reasonably good estimator. We have clearly demonstrated the computational advantage of our method. We have shown that unlike in Rothman (2012), our theory does not require a lower bound on the smallest eigenvalue of the true covariance matrix. Thus our theory is practically more relevant.

We would also like to argue that the main idea behind our method is much more flexible than the log-determinant barrier method in the sense that the former can be easily extended to compute the positive-definite ℓ_1-penalized covariance estimator under some additional constraints. For illustration, let us consider the scenario when we know some extra prior information that eigenvalues of the true covariance matrix are bounded by two positive constants α and β (d'Aspremont, Banerjee, and Ghaoui 2008; Lu 2010). Then we solve the following constrained optimization problem

$$\hat{\boldsymbol{\Sigma}}^{+}_{\alpha,\beta} = \arg\min_{\beta \boldsymbol{I} \succeq \boldsymbol{\Sigma} \succeq \alpha \boldsymbol{I}} \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_n\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1. \tag{14}$$

Define $(\boldsymbol{Z})_{\alpha,\beta}$ as the projection of a matrix $\boldsymbol{Z}$ onto the convex cone $\{\beta \boldsymbol{I} \succeq \boldsymbol{\Theta} \succeq \alpha \boldsymbol{I}\}$, that is, $(\boldsymbol{Z})_{\alpha,\beta} = \sum_{j=1}^{p} \min(\max(\lambda_j, \alpha), \beta)\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{v}_j$ when $\boldsymbol{Z}$ has the eigen-decomposition $\sum_{j=1}^{p} \lambda_j \boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{v}_j$. Hence, an efficient alternating direction algorithm can be obtained from Algorithm 1 by simply modifying its $\boldsymbol{\Theta}$ step as

$$\boldsymbol{\Theta}^{i+1} = \arg\min_{\beta \boldsymbol{I} \succeq \boldsymbol{\Sigma} \succeq \alpha \boldsymbol{I}} L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}^i; \boldsymbol{\Lambda}^i) = (\boldsymbol{\Sigma}^i + \mu \boldsymbol{\Lambda}^i)_{\alpha,\beta}.$$

However, if we apply the log-determinant barrier method to solve the new problem, the corresponding optimization criterion becomes

$$\hat{\boldsymbol{\Sigma}}^{+}_{\alpha,\beta} = \arg\min_{\beta \boldsymbol{I} \succeq \boldsymbol{\Sigma} \succeq \alpha \boldsymbol{I}} \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_n\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1 - \tau_1 \log\det(\boldsymbol{\Sigma} - \alpha \boldsymbol{I}) - \tau_2 \log\det(\beta \boldsymbol{I} - \boldsymbol{\Sigma}). \tag{15}$$

It is not clear how to extend the heuristic iterative procedure in Rothman (2012) to handle the above optimization problem which is considerably more complex than (4).

APPENDIX: TECHNICAL PROOFS

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

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

$$(\hat{\boldsymbol{\Sigma}}^{+} - \hat{\boldsymbol{\Sigma}}_n)_{jj} + \hat{\boldsymbol{\Lambda}}^{+}_{jj} = 0, \quad \forall j = 1, \ldots, p, \tag{A.2}$$

$$\hat{\boldsymbol{\Theta}}^{+} = \hat{\boldsymbol{\Sigma}}^{+}, \tag{A.3}$$

$$\hat{\boldsymbol{\Theta}}^{+} \succeq \epsilon \boldsymbol{I}, \tag{A.4}$$

and

$$\langle \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}^{+} \rangle \leq 0, \quad \forall \boldsymbol{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{A.5}$$

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

$$\left\langle \boldsymbol{\Lambda}^i - \frac{1}{\mu}(\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Sigma}^i), \boldsymbol{\Theta} - \boldsymbol{\Theta}^{i+1} \right\rangle \leq 0, \quad \forall \boldsymbol{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{A.6}$$

Using the updating formula for $\boldsymbol{\Lambda}^i$ in Algorithm 1, that is,

$$\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\mu}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Theta}^{i+1}), \tag{A.7}$$

(A.6) can be rewritten as

$$\left\langle \boldsymbol{\Lambda}^{i+1} - \frac{1}{\mu}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i), \boldsymbol{\Theta} - \boldsymbol{\Theta}^{i+1} \right\rangle \leq 0, \quad \forall \boldsymbol{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{A.8}$$

Now by letting $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{i+1}$ in (A.5) and $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{+}$ in (A.8), we can get that

$$\langle \hat{\boldsymbol{\Lambda}}^{+}, \boldsymbol{\Theta}^{i+1} - \hat{\boldsymbol{\Theta}}^{+} \rangle \leq 0, \tag{A.9}$$

and

$$\left\langle \boldsymbol{\Lambda}^{i+1} - \frac{1}{\mu}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i), \hat{\boldsymbol{\Theta}}^{+} - \boldsymbol{\Theta}^{i+1} \right\rangle \leq 0. \tag{A.10}$$

Summing (A.9) and (A.10) yields

$$\left\langle \boldsymbol{\Theta}^{i+1} - \hat{\boldsymbol{\Theta}}^{+}, (\boldsymbol{\Lambda}^{i+1} - \hat{\boldsymbol{\Lambda}}^{+}) + \frac{1}{\mu}(\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}) \right\rangle \geq 0. \tag{A.11}$$

The optimality conditions for the second subproblem in Algorithm 1, that is, the subproblem with respect to $\boldsymbol{\Sigma}$ in (8) are given by

$$0 \in (\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}_n)_{j\ell} + \lambda \partial \left|\boldsymbol{\Sigma}^{i+1}_{j\ell}\right| + \boldsymbol{\Lambda}^i_{j\ell} + \frac{1}{\mu}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Theta}^{i+1})_{j\ell}, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{A.12}$$

and

$$(\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}_n)_{jj} + \boldsymbol{\Lambda}^i_{jj} + \frac{1}{\mu}(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Theta}^{i+1})_{jj} = 0, \quad \forall j = 1, \ldots, p. \tag{A.13}$$

Note that by using (A.7), (A.12) and (A.13) can be, respectively, rewritten as

$$\frac{1}{\lambda}(-\boldsymbol{\Lambda}^{i+1} - \boldsymbol{\Sigma}^{i+1} + \hat{\boldsymbol{\Sigma}}_n)_{j\ell} \in \partial \left|\boldsymbol{\Sigma}^{i+1}_{j\ell}\right|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \text{ and } j \neq \ell, \tag{A.14}$$

and

$$(\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}_n)_{jj} + \boldsymbol{\Lambda}^{i+1}_{jj} = 0, \quad \forall j = 1, \ldots, p. \tag{A.15}$$

Using the fact that $\partial|\cdot|$ is a monotone function, (A.1), (A.2), (A.14), and (A.15) imply

$$\langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}, (\hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1}) + (\hat{\boldsymbol{\Sigma}}^{+} - \boldsymbol{\Sigma}^{i+1}) \rangle \geq 0. \tag{A.16}$$

The summation of (A.11) and (A.16) gives

$$\|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}\|_F^2 \leq \langle \boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^{+}, \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1} \rangle + \langle \hat{\boldsymbol{\Theta}}^{+} - \boldsymbol{\Theta}^{i+1}, \hat{\boldsymbol{\Lambda}}^{+} - \boldsymbol{\Lambda}^{i+1} \rangle + \frac{1}{\mu}\langle \hat{\boldsymbol{\Theta}}^{+} - \boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Sigma}^i \rangle. \tag{A.17}$$

Combining (A.17) with $\mathbf{\Theta}^{i+1} = \mu(\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}) + \mathbf{\Sigma}^{i+1}$ and $\hat{\mathbf{\Theta}}^+ = \hat{\mathbf{\Sigma}}^+$ leads to

$$\begin{aligned}\|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|_F^2 \le\ & \langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+, \hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1}\rangle \\ & + \langle \hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^{i+1} - \mu(\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}), \hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1}\rangle \\ & + \frac{1}{\mu}\langle \hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^{i+1} - \mu(\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}), \mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i\rangle.\end{aligned} \tag{A.18}$$

Simple algebraic derivation from (A.18) yields the following inequality:

$$\begin{aligned}&\|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|_F^2 - \langle \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle \\ &\le \mu\langle \mathbf{\Lambda}^{i+1} - \hat{\mathbf{\Lambda}}^+, \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}\rangle + \langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle\mu.\end{aligned} \tag{A.19}$$

Rearranging the right-hand side of (A.19) using $\hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^{i+1} = (\hat{\mathbf{\Lambda}}^+ - \mathbf{\Lambda}^i) + (\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1})$ and $\hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^{i+1} = (\hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^i) + (\mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1})$, then (A.17) can be reduced to

$$\begin{aligned}&\mu\langle \mathbf{\Lambda}^i - \hat{\mathbf{\Lambda}}^+, \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}\rangle + \frac{1}{\mu}\langle \mathbf{\Sigma}^i - \hat{\mathbf{\Sigma}}^+, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle \\ &\ge \mu\|\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}\|_F^2 + \frac{1}{\mu}\|\mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\|_F^2 + \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|_F^2 \\ &\quad - \langle \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle.\end{aligned} \tag{A.20}$$

Using the notation of U^i and U^*, (A.20) can be rewritten as

$$\begin{aligned}\langle U^i - U^*, U^i - U^{i+1}\rangle_G \ge\ & \|U^i - U^{i+1}\|_G^2 + \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|_F^2 \\ & - \langle \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle.\end{aligned} \tag{A.21}$$

Combining (A.21) with the following identity

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

we get

$$\begin{aligned}&\|U^i - U^*\|_G^2 - \|U^{i+1} - U^*\|_G^2 \\ &= 2\langle U^i - U^{i+1}, U^i - U^*\rangle - \|U^{i+1} - U^i\|_G^2 \\ &\ge 2\|U^i - U^{i+1}\|_G^2 + 2\|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|^2 - 2\langle \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle \\ &\quad - \|U^{i+1} - U^i\|_G^2 \\ &= \|U^i - U^{i+1}\|_G^2 + 2\|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^+\|^2 - 2\langle \mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}\rangle.\end{aligned} \tag{A.22}$$

Now, using (A.14) and (A.15) for i instead of $i+1$, we get,

$$\frac{1}{\lambda}(-\mathbf{\Lambda}^i - \mathbf{\Sigma}^i + \hat{\mathbf{\Sigma}}_n)_{j\ell} \in \partial|\mathbf{\Sigma}^i_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \quad \text{and} \quad j \ne \ell, \tag{A.23}$$

and

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

Combining (A.14), (A.15), (A.23), and (A.24) and using the fact that $\partial|\cdot|$ is a monotone function, we obtain

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

which immediately implies,

$$\langle \mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1}, \mathbf{\Lambda}^{i+1} - \mathbf{\Lambda}^i\rangle \ge \|\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^i\|_F^2 \ge 0. \tag{A.25}$$

By substituting (A.25) into (A.22), we get the desired result (13). □

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

(a) $\|U^i - U^{i+1}\|_G \to 0$;
(b) $\{U^i\}$ lies in a compact region;
(c) $\|U^i - U^*\|_G^2$ is monotonically non-increasing and thus converges.

It follows from (a) that $\mathbf{\Lambda}^i - \mathbf{\Lambda}^{i+1} \to 0$ and $\mathbf{\Sigma}^i - \mathbf{\Sigma}^{i+1} \to 0$. Then (A.7) implies that $\mathbf{\Theta}^i - \mathbf{\Theta}^{i+1} \to 0$ and $\mathbf{\Theta}^i - \mathbf{\Sigma}^i \to 0$. From (b), we obtain that U^i has a subsequence $\{U^{i_j}\}$ that converges to $\bar{U} = (\bar{\mathbf{\Lambda}}, \bar{\mathbf{\Sigma}})$, that is, $\mathbf{\Lambda}^{i_j} \to \bar{\mathbf{\Lambda}}$ 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 (A.14) and (A.13), respectively, imply that

$$\frac{1}{\lambda}(-\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}} + \hat{\mathbf{\Sigma}}_n)_{j\ell} \in \partial|\bar{\mathbf{\Sigma}}_{j\ell}|, \quad \forall j = 1, \ldots, p, \ell = 1, \ldots, p, \quad \text{and} \quad j \ne \ell, \tag{A.26}$$

and

$$(\bar{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_n)_{jj} + \bar{\mathbf{\Lambda}}_{jj} = 0, \quad \forall j = 1, \ldots, p, \tag{A.27}$$

and (A.8) implies that

$$\langle \bar{\mathbf{\Lambda}}, \mathbf{\Theta} - \bar{\mathbf{\Theta}}\rangle \le 0, \quad \forall \mathbf{\Theta} \succeq \epsilon \boldsymbol{I}. \tag{A.28}$$

(A.26), (A.27), and (A.28) 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 $1 \le i \le n$, $1 \le j \le p$. 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 $\{|\hat{\sigma}^n_{ij} - \sigma^0_{ij}| \le \lambda, \forall (i, j)\}$, we have

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

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, $\|\hat{\mathbf{\Delta}}\|_F \le 5\lambda(s+p)^{1/2}$ 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 $G(\hat{\mathbf{\Delta}}) \le G(\mathbf{0}) = 0$. Therefore, we can obtain the following probability bound

$$\Pr(\|\hat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^0\|_F \le 5\lambda(s+p)^{1/2}) \ge 1 - \Pr\left(\max_{i,j}|\hat{\sigma}^n_{ij} - \sigma^0_{ij}| > \lambda\right).$$

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

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

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

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

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

$$\begin{aligned}
&E\left[\left(X_{ij}X_{ik}-\sigma_{jk}^0\right)^2\cdot\exp\left(t_1\left|X_{ij}X_{ik}-\sigma_{jk}^0\right|\right)\right] \\
&\le 2E\left[X_{ij}^2X_{ik}^2\cdot\exp\left(\frac{1}{2}\eta\left|X_{ij}X_{ik}-\sigma_{jk}^0\right|\right)\right]+2\left(\sigma_{jk}^0\right)^2 \\
&\quad\cdot E\left[\exp\left(\frac{1}{2}\eta\left|X_{ij}X_{ik}-\sigma_{jk}^0\right|\right)\right] \\
&\le 8K_1\eta^{-2}\cdot\exp\left(\frac{1}{2}\eta\sigma_{jk}^0\right)+2\left(\sigma_{jk}^0\right)^2\cdot\exp\left(\frac{1}{2}\eta\sigma_{jk}^0\right) \\
&\quad\cdot E\left[\exp\left(\frac{1}{4}\eta\left(X_{ij}^2+X_{ik}^2\right)\right)\right] \\
&\le 8K_1\eta^{-2}\cdot\exp\left(\frac{1}{2}\eta\sigma_{\max}\right)+2\sigma_{\max}^2\cdot\exp\left(\frac{1}{2}\eta\sigma_{\max}\right) \\
&\quad\times\left(E\left[\exp\left(\frac{1}{2}\eta X_{ij}^2\right)\right]\right)^{1/2}\cdot\left(E\left[\exp\left(\frac{1}{2}\eta X_{ik}^2\right)\right]\right)^{1/2} \\
&\le 2K_1(4\eta^{-2}+\sigma_{\max}^2)\cdot\exp\left(\frac{1}{2}\eta\sigma_{\max}\right),
\end{aligned}$$

where we use the fact that $t_1 = \frac{1}{2}\eta(\frac{\log p}{n})^{1/2} \le \frac{1}{2}\eta < \eta$ in the first inequality and then use $|\sigma_{jk}^0| \le (\sigma_{jj}^0\sigma_{kk}^0)^{1/2} \le \sigma_{\max}$ in the third inequality.

Now, we can apply the Markov inequality to obtain the following probability bound:

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

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

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

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

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

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 $|E[Z_{ij}]| \le \delta_n^{-3}E[|X_{ij}|^4I_{\{|X_{ij}|>\delta_n\}}] \le K_2\delta_n^{-3} = o(\varepsilon_2)$. Now we can apply the Bernstein's inequality (Bernstein 1946; Bennett 1962) to obtain that

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

where the fact that $\mathrm{var}(Y_{ij}) \le E[X_{ij}^2] \le E[X_{ij}^2I_{\{|X_{ij}|\ge1\}}] + E[X_{ij}^2I_{\{|X_{ij}|\le1\}}] \le K_2+1$ is used in the second inequality. Besides, we can apply the Markov inequality to obtain that

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

Then, we can derive the following probability bound

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

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 $\sigma^0_{jk} = E[X_{ij}X_{ik}] = E[Y_{ij}Y_{ik}] + E[R_{ijk}]$. 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

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

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

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

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

$$\begin{aligned}&\Pr\left(\max_{j,k}\left|\sum_i \left(X_{ij}X_{ik} - \sigma^0_{jk}\right)\right| > n\varepsilon_3\right)\\ &\le \Pr\left(\max_{j,k}\left|\sum_i \{Y_{ij}Y_{ik} - E[Y_{ij}Y_{ik}]\}\right| > \frac{1}{2}n\varepsilon_3\right)\\ &\quad + \Pr\left(\max_{j,k}\left|\sum_i \{R_{ijk} - E[R_{ijk}]\}\right| > \frac{1}{2}n\varepsilon_3\right)\\ &\le 2\sum_{j,k}\Pr\left(\sum_i \{Y_{ij}Y_{ik} - E[Y_{ij}Y_{ik}]\} > \frac{1}{2}n\varepsilon_3\right)\\ &\quad + \Pr\left(\max_{j,k}\left|\sum_i \{R_{ijk} - o(\varepsilon_3)\}\right| > \frac{1}{2}n\varepsilon_3\right)\\ &\le O(p^{-M}) + \sum_{i,j}\Pr(|X_{ij}| > \delta_n)\\ &\le O(p^{-M}) + K_2p(\log n)^{2(1+\gamma+\varepsilon)}n^{-\gamma-\varepsilon}\end{aligned}$$

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

$$\begin{aligned}&\Pr\left(\max_{j,k}|\hat\sigma^n_{jk} - \sigma^0_{jk}| > \lambda\right)\\ &\le \Pr\left(\max_{j,k}\left|\sum_i \left\{X_{ij}X_{ik} - \sigma^0_{jk}\right\}\right| > n\varepsilon_3\right)\\ &\quad + \Pr\left(\max_j\left|\sum_i X_{ij}\right| > n\varepsilon_2\right)\\ &\le O(p^{-M}) + 3K_2p(\log n)^{2(1+\gamma+\varepsilon)}n^{-\gamma-\varepsilon}.\end{aligned}$$

[Received November 2011. Revised June 2012.]

REFERENCES

Anderson, T. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley. [1480]

Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002), "Gene-Expression Profiles Predict Survival of Patients With Lung Adenocarcinoma," *Nature Medicine*, 8, 816–824. [1480]

Bennett, G. (1962), "Probability Inequalities for the Sum of Independent Random Variables," *Journal of the American Statistical Association*, 57, 33–45. [1489]

Bernstein, S. (1946), *The Theory of Probabilities*, Moscow: Gostekhizdat. [1489]

Bickel, P., and Levina, E. (2008a), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [1480,1483]

——— (2008b), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [1480,1483,1486]

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [1482]

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009), "Sparse and Stable Markowitz Portfolios," *Proceedings of the National Academy of Sciences*, 106, 12267–12272. [1485]

Cai, T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684. [1480,1483]

Cai, T., Zhang, C., and Zhou, H. (2010), "Optimal Rates of Convergence for Covariance Matrix Estimation," *The Annals of Statistics*, 38, 2118–2144. [1480]

Cai, T., and Zhou, H. (2012a), "Minimax Estimation of Large Covariance Matrices Under ℓ_1-Norm" (with discussion), *Statistica Sinica*, 22, 1319–1348. [1480,1486]

——— (2012b), "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, to appear. [1480,1486]

d'Aspremont, A., Banerjee, O., and Ghaoui, L. (2008), "First-Order Methods for Sparse Covariance Selection," *SIAM Journal on Matrix Analysis and Applications*, 30, 56–66. [1487]

DeMiguel, V., Garlappi, L., Nogales, F., and Uppal, R. (2009), "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55, 798–812. [1485]

Douglas, J., and Rachford, H. H. (1956), "On the Numerical Solution of the Heat Conduction Problem in 2 and 3 Space Variables," *Transactions of the American Mathematical Society*, 82, 421–439. [1482]

Efron, B. (2009), "Are a Set of Microarrays Independent of Each Other?" *The Annals of Applied Statistics*, 3, 922–942. [1484]

——— (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge: Cambridge University Press. [1484]

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756. [1480]

——— (2010), "High-Dimensionality Effects in the Markowitz Problem and Other Quadratic Programs With Linear Constraints: Risk Underestimation," *The Annals of Statistics*, 38, 3487–3566. [1485]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1480]

Fan, J., Zhang, J., and Yu, K. (2012), "Vast Portfolio Selection with Gross-Exposure Constraints," *Journal of the American Statistical Association*, 107, 592–606. [1485]

Fortin, M., and Glowinski, R. (1983), *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, Amsterdam: North-Holland. [1482]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432. [1481]

Furrer, R., and Bengtsson, T. (2007), "Estimation of High-Dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants," *Journal of Multivariate Analysis*, 98, 227–255. [1480]

Glowinski, R., and Le Tallec, P. (1989), *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, Philadelphia, PA: SIAM. [1482]

Jagannathan, R., and Ma, T. (2003), "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *Journal of Finance*, 58, 1651–1684. [1485]

Johnstone, I. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of Statistics*, 29, 295–327. [1480]

Khan, J., Wei, J., Ringnér, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., and Meltzer, P. S. (2001), "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, 7, 673–679. [1484]

Lu, Z. (2010), "Adaptive First-Order Methods for General Sparse Inverse Covariance Selection," *SIAM Journal on Matrix Analysis and Applications*, 31, 2000–2016. [1487]

Marčenko, V. A., and Pastur, L. A. (1967), "Distribution of Eigenvalues for Some Sets of Random Matrices," *Sbornik: Mathematics*, 1, 457–483. [1480]

Markowitz, H. (1952), "Portfolio Selection," *Journal of Finance*, 7, 77–91. [1485]

Peaceman, D. H., and Rachford, H. H. (1955), "The Numerical Solution of Parabolic Elliptic Differential Equations," *SIAM Journal on Applied Mathematics*, 3, 28–41. [1482]

Rothman, A. (2012), "Positive Definite Estimators of Large Covariance Matrices," *Biometrika*, 99, 733–740. [1481,1483,1486,1487]

Rothman, A., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [1480,1483]

Scheinberg, K., Ma, S., and Goldfarb, D. (2010), "Sparse Inverse Covariance Selection via Alternating Linearization Methods," *Advances in Neural Information Processing Systems*. arXiv:1011.0097. [1482]

Subramaniana, A., Tamayoa, P., Moothaa, V., Mukherjeed, S., Eberta, B., Gillettea, M., Paulovichg, A., Pomeroyh, S., Goluba, T., Landera, E., Lander, E. S., and Mesirov, J. P. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550. [1480]

Wu, W., and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika*, 90, 831–844. [1480]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1480]