Sparse Covariance Matrix Estimation with Eigenvalue Constraints

Han Liu[1] **and Lie Wang**[2] **and Tuo Zhao**[3]

[1] *Department of Operations Research and Financial Engineering, Princeton University*
[2] *Department of Mathematics, Massachusetts Institute of Technology*
[3] *Department of Computer Science, Johns Hopkins University*

Abstract: We propose a new approach for estimating high dimensional positive-definite covariance matrices. Our method extends the generalized thresholding operator by adding an explicit eigenvalue constraint. The estimated covariance matrix simultaneously achieves sparsity and positive definiteness. The estimator is rate optimal in the minimax sense and we develop an efficient iterative soft-thresholding and projection algorithm based on the alternating direction method of multipliers. Empirically, we conduct thorough numerical experiments on simulated data sets as well as real data examples to illustrate the usefulness of our method.

Keywords and phrases: high dimensional data, covariance matrix estimation, positive-definiteness guarantee, explicit eigenvalue constraint.

1. Introduction

We study the sparse covariance matrix estimation problem. Let $\boldsymbol{X} := (X_1,\ldots,X_d)^T$ be a d-dimensional random vector with covariance matrix $\Sigma^*_{\text{cov}} := \left[(\Sigma^*_{\text{cov}})_{jk}\right]_{1\leq j,k\leq d}$, where $(\Sigma^*_{\text{cov}})_{jk} := \mathbb{E}X_jX_k - \mathbb{E}X_j\mathbb{E}X_k$. Our goal is to estimate Σ^*_{cov} from n observational data points. The covariance matrix estimation problem plays an essential role in multivariate methods such as time series analysis (Box et al., 2011), spatial data analysis (Cressie, 1992), and longitudinal data analysis (Searle et al., 2009). In this paper we are mainly interested in estimating covariance matrix in high dimensional settings where $d \gg n$.

The problem of estimating high dimensional covariance matrices has been extensively studied. By assuming the true covariance matrix is sparse, Wu and Pourahmadi (2003) analyzed longitudinal data where variables have a natural order and are weakly correlated when they are far apart. By utilizing different forms of (weak) sparsity, researchers proposed many regularized estimators (Bickel and Levina, 2004, 2008b; Cai and Liu,

1. Email: hanliu@princeton.edu
2. Email: liewang@math.princeton.edu
3. Email: tour@cs.jhu.edu

2011; Furrer and Bengtsson, 2007; Huang et al., 2006; Levina et al., 2008; Rothman et al., 2010; Cai et al., 2010; Cai and Liu, 2011). These estimators usually only require elementwise thresholding procedures (or combined with the Cholesky decomposition) and are computationally efficient. For other applications in which the variables do not have natural orders (e.g., genomic data, social science data and economic data), Bickel and Levina (2008a); Karoui (2008) showed that the thresholded sample covariance matrix also works well when the true covariance matrix is (weakly) sparse. Recently, Rothman et al. (2009) established more comprehensive theoretical results for a generalized thresholding operator family including the soft-thresholding (Tibshirani, 1996; Chen et al., 1998), hard-thresholding, SCAD-thresholding (Fan and Li, 2001), and adaptive thresholding operators (Zou, 2006). These thresholding estimators, though simple, do not guarantee the positive definiteness of the estimated covariance matrix. Such indefiniteness may cause trouble in downstream analysis such as evaluating the predictive likelihood or linear discriminant analysis. The only exceptions are Lam and Fan (2009) and Rothman (2012), which both use a Log-Determinant function to explicitly enforce the positive definiteness of the estimated covariance matrix. The difference between them is that Rothman (2012) adopt a convex least square formulation, while Lam and Fan (2009) adopt the penalized Gaussian likelihood approach, which results in a non-convex formulation and there is no guarantee to find the global optimal solution in polynomial time.

In this paper, we propose an alternative approach for estimating high dimensional positive-definite covariance matrices. Our method can be viewed as an extension of the generalized thresholding operator (Rothman et al., 2009). More specifically, we provide a convex formulation which secures the positive-definiteness of the estimator by enforcing an eigenvalue constraint. Compared with the logarithmic barrier method of Rothman (2012), our approach explicitly constrains the smallest eigenvalue of the estimated covariance matrix and can be naturally extended to more flexible settings. Explicitly constraining the eigenvalues has its practical implications. For example, for certain downstream analysis of covariance matrices estimation, such as discriminant analysis and graphical estimation, we require the estimated covariance matrix to have a good condition number. Sometimes, we have prior knowledge about the smallest and largest eigenvalues and may want to incorporate such knowledge into the inference. Computationally, we adopt the alternating direction method of multipliers (ADMM) to develop an efficient ISP (Iterative Soft-thresholding and Projection) algorithm (Gabay and Mercier, 1976; He and Yuan, 2012). The ISP algorithm iteratively conducts the soft-thresholding operation and eigenvalue decomposition. We prove that, with high probability, the obtained estimator is the same as the generalized thresholding estimator (when it is positive definite). This result implies that our procedure preserves the nice asymptotic properties of the generalized thresholding estimator. More precisely, let $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$ be our proposed covariance matrix estimator and $\|\widehat{\Sigma}\|_2$ the spectral norm of $\widehat{\Sigma}$, we have

$$\sup_{\Sigma^*_{\rm cov}\in\mathcal{U}(q)} \mathbb{E}\left\|\widehat{\Sigma}-\Sigma^*_{\rm cov}\right\|_2 \asymp M_d\left(\sqrt{\frac{\log d}{n}}\right)^{1-q}, \tag{1.1}$$

where $\mathcal{U}$ denotes some parameter class (will be explained in Section 4) and M_d is some quantity that may scale with d. This rate of convergence is optimal in the minimax sense. Besides these nice asymptotic properties, our procedure has good finite sample performance. Empirically, we report thorough numerical experimental results on both simulated and real data sets to illustrate the usefulness of the proposed method.

While this paper was under review, we learnt that a similar positive definite covariance estimator was independently proposed by Xue et al. (2012). There are several major differences between our estimator and theirs as listed in the following:

(1) Our estimator is different from that of Xue et al. (2012). Instead of estimating covariance matrix directly, we first estimate the correlation matrix and then combine the estimated correlation matrix with the marginal sample standard deviations to get the final covariance estimator. This same idea is also used in Lam and Fan (2009) and Rothman et al. (2008). The main advantage is that our estimator is adaptive to the marginal variability and more amendable to theoretical analysis (In particular, we provide explicit rate of convergence for $\mathbb{E}\left\|\widehat{\Sigma}-\Sigma^*_{\rm cov}\right\|_2$, while Xue et al. (2012) only provide large probability bound for $\left\|\widehat{\Sigma}-\Sigma^*_{\rm cov}\right\|_{\sf F}$.)

(2) Besides the ℓ_1 penalty function, we also study other non-convex penalty functions. An interesting fact we observed is that our proposed estimator can be combined with the non-convex minimum concavity (MC+) penalty function without losing a global convex formulation. Therefore we can simultaneously reduce the estimation bias and preserves the global convexity of the problem. In Xue et al. (2012), only the ℓ_1 penalty function is studied.

(3) Xue et al. (2012) mainly focus on the optimization convergence property of the propose ADMM-based computational algorithm. In contrast, we focus more on analyzing the statistical properties of the proposed method. We also illustrate the relationship between our proposed estimator and the generalized thresholding estimator. Such a relationship allows us to establish the minimax optimal rates of convergence of the resulting sparse covariance estimates under Forbenius and spectral norms.

Moreover, we should point out that the ISP algorithm was first developed in Bien and Tibshirani (2011) to solve a subproblem obtained by their iterative majorize/minimize procedure. Their subproblem is identical to our proposed convex program in the numerical form. More details about this point will be discussed in Section 3.1.

The rest of this paper is organized as follows. In Section 2, we briefly review the covariance matrix estimation problem. In Section 3, we introduce our new method and present the ISP algorithm. In Section 4, we analyze the theoretical properties of the proposed estimator. Numerical simulations and real data experiments are reported in Section 5. We give brief conclusions and discussions in the last section. All the proofs are in the appendix.

2. Notation and Background

Let $\mathbf{A} = \left[\mathbf{A}_{jk}\right] \in \mathbb{R}^{d\times d}$, $\mathbf{B} = \left[\mathbf{B}_{jk}\right] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1,\ldots,v_d)^T \in \mathbb{R}^d$. We denote by $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of $\mathbf{A}$. The inner product of $\mathbf{A}$ and $\mathbf{B}$ is defined as $\left\langle \mathbf{A},\mathbf{B}\right\rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$. We define the vector norms: $\|\boldsymbol{v}\|_1 = \sum_j |v_j|$, $\|\boldsymbol{v}\|_2^2 = \sum_j v_j^2$, $\|\boldsymbol{v}\|_\infty = \max_j |v_i|$. We also define matrix operator and elementwise norms: $\|\mathbf{A}\|_{1,\mathrm{off}} = \sum_{j\neq k} |\mathbf{A}_{jk}|$, $\|\mathbf{A}\|_{\infty,\mathrm{off}} = \max_{j\neq k} |\mathbf{A}_{jk}|$, $\|\mathbf{A}\|_2^2 = \Lambda_{\max}(\mathbf{A}^T\mathbf{A}), \|\mathbf{A}\|_F^2 = \sum_{j,k} \mathbf{A}_{jk}^2$. We use $\mathbf{S}_{\mathrm{cov}}$ and $\mathbf{S}$ to denote the sample covariance and correlation matrices. The true covariance and correlation matrices are denoted by Σ^*_{cov} and Σ^*, respectively. We write $a_n \asymp b_n$ if there are positive constants c_1 and c_2 independent of n such that $c_1 b_n \leq a_n \leq c_2 b_n$.

2.1. *Connection between the Covariance and Correlation Estimation*

A closely related problem is the estimation of correlation matrix, in which we are interested in estimating the correlation matrix $\Sigma^* := (\boldsymbol{\Theta}^*)^{-1}\Sigma^*_{\mathrm{cov}}(\boldsymbol{\Theta}^*)^{-1}$, where $\boldsymbol{\Theta}^* := \mathrm{diag}(\theta_1^*,...,\theta_d^*)$ is a diagonal matrix with θ_j^* denoting the standard deviation of the j-th variable. Using this decomposition, we can construct a covariance matrix estimator $\widehat{\Sigma}_{\mathrm{cov}}$ from $\widehat{\Sigma}$:

$$\widehat{\Sigma}_{\mathrm{cov}} = \widehat{\boldsymbol{\Theta}}\widehat{\Sigma}\widehat{\boldsymbol{\Theta}} \text{ with } \widehat{\boldsymbol{\Theta}} = \mathrm{diag}(\widehat{\theta}_1,\widehat{\theta}_2,...,\widehat{\theta}_d). \tag{2.1}$$

where $\widehat{\theta}_j$ is the sample standard deviation for the j-th variable. Such a procedure is more amenable to theoretical analysis (more details can be found in Remark 4.4 in Section 4) and also makes the estimator adaptive to the marginal variability (more details can be found in Rothman et al. (2008); Lam and Fan (2009)).

2.2. *Soft-thresholding Operator*

Recall that $\mathbf{S}$ is the sample correlation matrix, the generalized thresholding operator (Rothman et al., 2009) with ℓ_1 penalty solves the following optimization problem:

$$\widehat{\Sigma}^{\mathrm{STO}} = \underset{\Sigma}{\mathrm{argmin}}\, \frac{1}{2}\|\mathbf{S}-\Sigma\|_{\mathsf{F}}^2 + \lambda\|\Sigma\|_{1,\mathrm{off}}. \tag{2.2}$$

This simple formulation has a closed-form solution:

$$\widehat{\Sigma}^{\mathrm{STO}}_{jk} = \begin{cases} \mathrm{sign}(\mathbf{S}_{jk})\cdot\max\{|\mathbf{S}_{jk}|-\lambda,0\} & \text{if } j\neq k \\ \mathbf{S}_{jk} & \text{otherwise} \end{cases}, \tag{2.3}$$

which is well known as the univariate soft-thresholding operator. Once the STO estimate $\widehat{\Sigma}^{\mathrm{STO}}$ is obtained, we plug it into (2.1) for estimating the covariance matrix $\widehat{\Sigma}_{\mathrm{cov}}$. Since (2.2) has no control of the positive-definiteness, $\widehat{\Sigma}^{\mathrm{STO}}$ may be indefinite.

3. Our Proposed Method

Our method, named EC2 (Estimation of Covariance with Eigenvalue Constraints), extends the STO method by explicitly adding eigenvalue constraints. Recall that $\mathbf{S}$ is the sample correlation matrix, the EC2 estimator is defined as

$$\widehat{\Sigma}^{\mathrm{EC2}} := \underset{\Sigma_{jj}=1}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{S}-\Sigma\|_{\mathsf{F}}^2 + \lambda\|\Sigma\|_{1,\mathrm{off}} \quad \text{s.t. } \tau \leq \Lambda_{\min}(\Sigma) \tag{3.1}$$

where $\lambda > 0$ is a regularization parameter, and $\tau > 0$ is a desired minimum eigenvalue lower bound of the estimator which also be chosen in a data-dependent way. The EC2 method simultaneously conducts sparse estimation and guarantees the positive-definiteness of the solution. The equality constraint $\Sigma_{jj} = 1$ ensures all diagonal entries of $\widehat{\Sigma}$ to be 1, since we are estimating the correlation matrix. Once $\widehat{\Sigma}^{\mathrm{EC2}}$ is obtained, we convert it to our covariance matrix estimator as (2.1). The next proposition shows that the formulation in (3.1) is convex:

Proposition 3.1. *The optimization problem in* (3.1) *is a convex program.*

Proof. The proof of this proposition is presented in Appendix A. □

We proceed to develop an efficient algorithm to solve (3.1) using the ADMM. The resulting algorithm is iterative and solves a closed-form soft-thresholding step and a spectral projection step within each iteration.

3.1. Iterative Soft-thresholding and Projection Algorithm

We first reparametrize (3.1) by introducing an auxiliary variable $\mathbf{\Gamma}$:

$$(\widehat{\Sigma}, \widehat{\mathbf{\Gamma}}) = \underset{\Sigma_{jj}=1,\ \tau\leq\Lambda_{\min}(\mathbf{\Gamma})}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{S}-\mathbf{\Gamma}\|_{\mathsf{F}}^2 + \lambda\|\Sigma\|_{1,\mathrm{off}} \quad \text{s. t. } \mathbf{\Gamma} = \Sigma. \tag{3.2}$$

This reparametrization decouples the computational dependency in the optimization problem. Therefore a complicated problem decouples into multiple simpler sub-problems, which can be solved in closed-forms. More specifically, (3.2) can be rewritten in the following equivalent saddle point problem

$$(\widehat{\Sigma}, \widehat{\mathbf{\Gamma}}, \widehat{\mathbf{U}}) = \underset{\Sigma_{jj}=1,\ \tau\leq\Lambda_{\min}(\mathbf{\Gamma})}{\operatorname{argmin}} \max_{\mathbf{U}} \frac{1}{2}\|\mathbf{S}-\mathbf{\Gamma}\|_{\mathsf{F}}^2 + \lambda\|\Sigma\|_{1,\mathrm{off}} + \left\langle \mathbf{U}, \mathbf{\Gamma}-\Sigma \right\rangle + \frac{\rho}{2}\|\mathbf{\Gamma}-\Sigma\|_{\mathsf{F}}^2, \tag{3.3}$$

where $\rho > 0$ is the penalty multiplier, and $\mathbf{U} \in \mathbb{R}^{d\times d}$ is the Lagrange multiplier matrix. The ADMM works in an iterative fashion. Suppose we obtain the solution $\boldsymbol{\Sigma}^{(t)}, \boldsymbol{\Gamma}^{(t)}, \mathbf{U}^{(t)}$ at the t-th iteration, the algorithm proceeds as follows:

Step 1. Update $\boldsymbol{\Sigma}$ by

$$\boldsymbol{\Sigma}^{(t+1)} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}}\, \lambda\|\boldsymbol{\Sigma}\|_{1,\mathrm{off}} + \frac{\rho}{2}\Big\|\frac{1}{\rho}\mathbf{U}^{(t)} + \boldsymbol{\Gamma}^{(t)} - \boldsymbol{\Sigma}\Big\|_{\mathsf{F}}^2. \tag{3.4}$$

Let $\widetilde{\boldsymbol{\Sigma}} := \boldsymbol{\Gamma}^{(t)} + \frac{1}{\rho}\mathbf{U}^{(t)}$, (3.4) has the closed-form solution by soft-thresholding,

$$\Sigma_{jk}^{(t+1)} = \begin{cases} \operatorname{sign}(\widetilde{\Sigma}_{jk})(|\widetilde{\Sigma}_{jk}| - \lambda/\rho) & \text{if } j \neq k \\ 1 & \text{otherwise} \end{cases}. \tag{3.5}$$

Step 2. Given $\boldsymbol{\Sigma}^{(t+1)}$, we then update $\boldsymbol{\Gamma}$ by

$$\boldsymbol{\Gamma}^{(t+1)} = \underset{\tau \leq \Lambda_{\min}(\boldsymbol{\Gamma})}{\operatorname{argmin}} \Big\|\boldsymbol{\Gamma} - \frac{\mathbf{S} + \rho\boldsymbol{\Sigma}^{(t+1)} - \mathbf{U}^{(t)}}{(1+\rho)}\Big\|_{\mathsf{F}}^2. \tag{3.6}$$

Equation (3.6) has a closed-form solution $\boldsymbol{\Gamma}^{(t+1)} = \mathcal{P}_+\left(\frac{\mathbf{S}+\rho\boldsymbol{\Sigma}^{(t+1)}-\mathbf{U}^{(t)}}{(1+\rho)}, \tau\right)$, where $\mathcal{P}_+(\cdot,\cdot)$ is a spectral projection operator and is characterized by the following lemma.

Lemma 3.2. *Suppose* $\mathbf{A}$ *has the spectral decomposition:* $\mathbf{A} = \sum_{j=1}^d \widehat{\delta}_j \boldsymbol{v}_j \boldsymbol{v}_j^T$*, where* $\widehat{\delta}_j$*'s are the eigenvalues and* $\boldsymbol{v}_j$*'s are the corresponding eigenvectors. Let* $\widetilde{\delta}_j = \max(\widehat{\delta}_j, \tau)$*, we have* $\mathcal{P}_+(\mathbf{A}, \tau) = \sum_{j=1}^d \widetilde{\delta}_j \boldsymbol{v}_j \boldsymbol{v}_j^T$.

Proof. The proof of this lemma is presented in Appendix B. □

Updating $\boldsymbol{\Gamma}$ requires an eigenvalue decomposition. The eigenvalue decomposition has been intensively studied in the literatures of numerical analysis and parallel computing. Many sophisticated parallel algorithms such as the Lanczos algorithm have been developed and the eigenvalue decomposition now can scale to large problems using GPGPU (general-purpose graphics processing unit) or multicore techniques.

Step 3. Update $\mathbf{U}$ by

$$\mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} + \rho(\boldsymbol{\Gamma}^{(t+1)} - \boldsymbol{\Sigma}^{(t+1)}). \tag{3.7}$$

We stop the iteration when the following convergence criterion is satisfied

$$\max\left\{\frac{\|\boldsymbol{\Sigma}^{(t+1)} - \boldsymbol{\Sigma}^{(t)}\|_F^2}{\|\boldsymbol{\Sigma}^{(t)}\|_{\mathsf{F}}^2}, \frac{\|\boldsymbol{\Sigma}^{(t+1)} - \boldsymbol{\Gamma}^{(t+1)}\|_{\mathsf{F}}^2}{\|\boldsymbol{\Sigma}^{(t)}\|_{\mathsf{F}}^2}\right\} \leq \epsilon, \tag{3.8}$$

where $\epsilon > 0$ is a precision tolerance parameter. In applications, we set $\epsilon = 10^{-6}$. Xue et al. (2012) also developed a similar computational algorithm, and our ISP algorithm has a different updating formula for Σ due to the equality constraint $\Sigma_{jj} = 1$ for the correlation matrix estimation. Moreover, as a reviewer pointed out, such an ADMM type algorithm was first derived in Bien and Tibshirani (2011) for sparse covariance estimation problem. More specifically, they adopted the majorize/minimize procedure to relax the non-conconvex formation. The resulting subproblem is identical to (3.1), and solved by the same computational algorithm. For more details, please refer to Appendix 3 in Bien and Tibshirani (2011).

Remark 3.3. *The convergence rate of our ISP algorithm has been established in He and Yuan (2012). It achieves an $O(1/t)$ rate of convergence in the sense of variational inequality, where t is the number of iterations. Please refer to He and Yuan (2012) for more technical details. A similar result is also established in Xue et al. (2012). Compared to existing analysis, a more interesting problem would be establishing a possible linear convergence rate since the least square loss function in* (3.1) *is strongly convex. However, this is beyond the scope of this paper and will be left for future investigation.*

3.2. Relationship between EC2 and Soft-thresholding Operator

By comparing the STO estimator $\widehat{\Sigma}^{\mathrm{STO}}$ and EC2 estimator $\widehat{\Sigma}^{\mathrm{EC2}}$ in (2.2) and (3.1), we can see that $\widehat{\Sigma}^{\mathrm{STO}} = \widehat{\Sigma}^{\mathrm{EC2}}$ holds, if $\widehat{\Sigma}^{\mathrm{STO}}$ falls in the feasible region of (3.1), i.e., $\tau \le \Lambda_{\min}(\widehat{\Sigma}^{\mathrm{STO}})$. In the next section, we will show that this relationship holds with high probability under appropriate scaling. This result has important implications for both computation and theoretical analysis. Empirically, we use $\widehat{\Sigma}^{\mathrm{STO}}$ as the initial estimator and if $\widehat{\Sigma}^{\mathrm{STO}}$ is already feasible, then the algorithm stops immediately.

3.3. Covariance Estimation with Adaptive Penalty

As mentioned earlier, the generalized thresholding operator family contains many thresholding functions induced by other non-convex penalty functions such as SCAD. Similarly we can also combined our EC2 with other non-convex regularization to reduce the estimation bias. Usually a non-convex regularization makes the penalized least square formulation non-convex and there is no guarantee to obtain a global solution in polynomial time. To retain the global convexity of the formulation, we usually choose the adaptive penalty in (Zou, 2006), which results in the following optimization problem

$$\widehat{\Sigma}^{\mathrm{EC2}} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{S} - \Sigma\|_{\mathsf{F}}^2 + \lambda\|\mathbf{W} \circ \Sigma\|_{1,\mathrm{off}} \ \text{ s. t. } \tau \le \Lambda_{\min}(\Sigma) \tag{3.9}$$

where $\mathbf{W}$ represents the weight matrix and $\circ$ denotes the Hadamard product, i.e., $\mathbf{W} \circ \Sigma = [\mathbf{W}_{jk} \cdot \Sigma_{jk}]_{j,k=1}^{d}$ There are multiple choices for $\mathbf{W}$, and Zou (2006) suggest to use the reciprocal of the ordinary least square estimate as the weight, i.e., $\mathbf{W}_{jk} = (\mathbf{S}_{jk})^{-1}$ in our case. Some recent result in Xue and Zou (2012) also implies that ℓ_1 penalized estimator is potentially a better choice, i.e., $\mathbf{W}_{jk} = \left(|\widehat{\Sigma}_{jk}^{\mathrm{STO}}| + 1/n\right)^{-1}$ in our case, where $1/n$ is to avoid the reciprocal of zero elements in the ℓ_1 penalized estimator. We will compare these two weight matrices in numerical experiments in Section 5.

Adding the weights does not change the problem formulation. Therefore we can solve the problem in (3.9) using the same ISP algorithm with Step 1 replaced by the following Step 1a,

Step 1a. Update Σ by

$$\Sigma^{(t+1)} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}}\, \lambda \|\mathbf{W} \circ \Sigma\|_{1,\mathrm{off}} + \frac{\rho}{2}\left\|\frac{1}{\rho}\mathbf{U}^{(t)} + \mathbf{\Gamma}^{(t)} - \Sigma\right\|_{\mathsf{F}}^2. \tag{3.10}$$

Let $\widetilde{\Sigma} := \mathbf{\Gamma}^{(t)} + \frac{1}{\rho}\mathbf{U}^{(t)}$, (3.10) has the closed-form solution by soft-thresholding,

$$\mathbf{\Sigma}_{jk}^{(t+1)} = \begin{cases} \operatorname{sign}(\widetilde{\Sigma}_{jk})(|\widetilde{\Sigma}_{jk}| - \lambda \mathbf{W}_{jk}/\rho) & \text{if } j \neq k \\ 1 & \text{otherwise} \end{cases}, \tag{3.11}$$

and the same convergence guarantee holds for the EC2 with adaptive penalty.

3.4. EC2 with MC+ penalty

Most non-convex penalty functions result in non-convex formulations of EC2, there is usually no guarantee to get a global solution in polynomial time. However, one exception is the minimax concave (MC+) penalty, it is possible to retain the global convexity with a suitable choice of tuning parameter. Recall the minimax concave penalty in Zhang (2010):

$$\begin{aligned} \lambda P(\Sigma, \lambda, \gamma) &= \lambda \sum_{k \neq j} \int_0^{|\Sigma_{jk}|} \left(1 - \frac{t}{\gamma\lambda}\right)_+ dt \\ &= \lambda \sum_{k \neq j} \left\{\left(|\Sigma_{jk}| - \frac{|\Sigma_{jk}|^2}{2\lambda\gamma}\right) I(|\Sigma_{jk}| < \lambda\gamma) + \frac{\lambda\gamma}{2} I(|\Sigma_{jk}| \geq \lambda\gamma)\right\}, \end{aligned} \tag{3.12}$$

where $\gamma > 1$ is a pre-fixed parameter. By varying from $\gamma \to \infty$ to $\gamma \to 0^+$, we can obtain a continuum of penalties and threshold operators for every $\lambda > 0$. The EC2 with MC+ penalty is formulated as

$$\widehat{\Sigma}^{\mathrm{EC2}} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}}\, \frac{1}{2}\|\mathbf{S} - \Sigma\|_{\mathsf{F}}^2 + \lambda P(\Sigma, \lambda, \gamma) \;\text{ s. t. } \tau \leq \Lambda_{\min}(\Sigma) \tag{3.13}$$

The next proposition shows that the problem in (3.13) in convex whenever the MC+ parameter $\gamma > 1$. Therefore, by applying the MC+ penalty function to EC2, we enjoy the bias reduction of the non-convex penalty without losing global convexity.

Proposition 3.4. *The optimization problem in* (3.13) *is convex when* $\gamma > 1$.

Proof. The proof of this proposition is presented in Appendix C. □

Remark 3.5. *A heuristic approach to solve* (3.13) *is to modify Step 1 of the ISP algorithm described in Subsection 3.1. More specifically, we replace Step 1 with Step 1b described as follows.*
Step 1b. Update current values of $\mathbf{\Gamma}^{(t)}$ *and* $\mathbf{U}^{(t)}$*, we update* Σ *by*

$$\Sigma^{(t+1)} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}} \lambda P(\Sigma, \lambda, \gamma) + \frac{\rho}{2}\|\mathbf{U}^{(t)} + \mathbf{\Gamma}^{(t)} - \Sigma\|_{\mathsf{F}}^2. \tag{3.14}$$

Let $\widetilde{\Sigma} := \mathbf{\Gamma}^{(t)} + \mathbf{U}^{(t)}$*, for* $\gamma > 1$*,* (3.10) *has a closed-form solution which is a piecewise-linear thresholding function,*

$$\Sigma_{jk}^{(t+1)} = \begin{cases} 0 & \textit{if } j \neq k \textit{ and } |\widetilde{\Sigma}_{jk}| \leq \dfrac{\lambda}{\rho} \\ \dfrac{\operatorname{sign}(\widetilde{\Sigma}_{jk})(|\widetilde{\Sigma}_{jk}| - \lambda/\rho)}{1 - 1/\gamma} & \textit{if } j \neq k \textit{ and } \dfrac{\lambda}{\rho} \leq |\widetilde{\Sigma}_{jk}| \leq \dfrac{\lambda\gamma}{\rho} \\ \widetilde{\Sigma}_{jk} & \textit{if } j \neq k \textit{ and } |\widetilde{\Sigma}_{jk}| > \dfrac{\lambda\gamma}{\rho} \\ 1 & \textit{otherwise} \end{cases}. \tag{3.15}$$

However, the ADMM cannot guarantee the convergence to the global solution of (3.13) *since existing analysis relies on the convexity of both the loss and penalty functions. Our empirical studies show that such an ADMM often fails to converge. Therefore, instead of using this algorithm, we propose an alternative algorithm using the local linear approximation (LLA).*

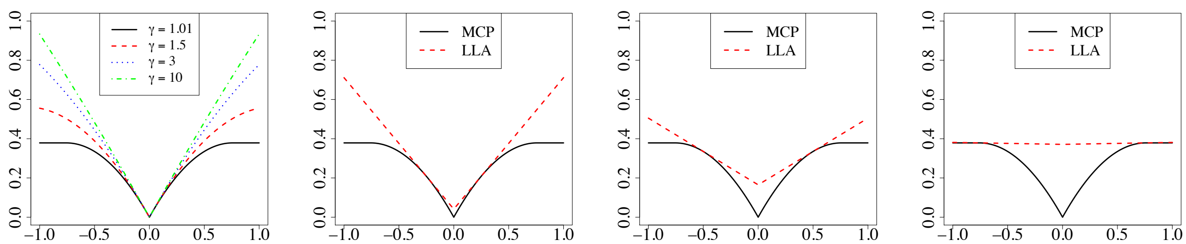
The LLA is also an iterative algorithm, and suppose at the t-th iteration, we already have the current solution $\Sigma^{(t)}$. We consider the linearization of MC+ at $\Sigma = \Sigma^{(t)}$, which is defined as

$$\lambda L(\Sigma, \Sigma^{(t)}) = \lambda \sum_{k \neq j} \left\{ \left(1 - \frac{|\Sigma_{jk}^{(t)}|}{\gamma\lambda} \right)_+ \left(|\Sigma_{jk}| - |\Sigma_{jk}^{(t)}| \right) \right\}, \tag{3.16}$$

then instead of directly minimizing (3.13), we solve the following convex relaxation of:

$$\widehat{\Sigma}^{(t+1)} = \underset{\Sigma_{jj}=1}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{S} - \Sigma\|_{\mathsf{F}}^2 + \lambda L(\Sigma, \Sigma^{(t)}) \ \text{ s. t. } \tau \leq \Lambda_{\min}(\Sigma) \tag{3.17}$$

The LLA is usually used to solve non-convex minimization problem and has a provable convergence to the local optimum. Since the formulation of (3.13) is convex, a global solution can be obtained. Several illustrative examples of the MC+ and LLA are shown in Figure 1.



(a) MC+ (b) LLA at $\Sigma_{jk} = 0.25$ (c) LLA at $\Sigma_{jk} = 0.5$ (d) LLA at $\Sigma_{jk} = 0.75$

Fig 1. *Illustration of the minimax concave penalty function and its local linear approximations. In (a), we illustrate the MC+ penalty with different γ's. In (b), (c), (d), the LLA approximations of different MC+ penalties are plotted.*

3.5. EC2 with Largest Eigenvalue Constraint

In this subsection, we extend the EC2 by also constraining the largest eigenvalue. In particular, we consider the following optimization problem:

$$\widehat{\Sigma}^{\mathrm{EC2}} := \operatorname*{argmin}_{\Sigma_{jj}=1} \frac{1}{2}\|\mathbf{S}-\Sigma\|_{\mathsf{F}}^2 + \lambda\|\Sigma\|_{1,\mathrm{off}} \ \text{ s. t. } \tau_1 \le \Lambda_{\min}(\Sigma) \le \Lambda_{\max}(\Sigma) \le \tau_2 \tag{3.18}$$

The formulation in (3.18) is also convex:

Proposition 3.6. *The optimization problem in* (3.18) *is a convex program.*

Proof. The proof of this proposition is presented in Appendix A. □

To derive the algorithm, we only need to modify Step 2 of the ISP algorithm described in Subsection 3.1. More specifically, we replace Step 2 with Step 2a described as follows. Step 2a. Given $\Sigma^{(t+1)}$, we update $\mathbf{\Gamma}$ by

$$\mathbf{\Gamma}^{(t+1)} = \operatorname*{argmin}_{\tau_1 \le \Lambda_{\min}(\mathbf{\Gamma}) \le \Lambda_{\max}(\mathbf{\Gamma}) \le \tau_2} \left\| \mathbf{\Gamma} - \frac{\mathbf{S} + \rho\Sigma^{(t+1)} - \rho\mathbf{U}^{(t)}}{(1+\rho)} \right\|_{\mathsf{F}}^2. \tag{3.19}$$

Equation (3.19) also has a closed-form solution

$$\mathbf{\Gamma}^{(t+1)} = \mathcal{P}_+'\left(\frac{\mathbf{S} + \rho\Sigma^{(t+1)} - \rho\mathbf{U}^{(t)}}{(1+\rho)}, \tau_1, \tau_2 \right), \tag{3.20}$$

where $\mathcal{P}_+^{'}(\cdot,\cdot,\cdot)$ is a spectral projection operator and is characterized by the following lemma.

Lemma 3.7. *Suppose* $\mathbf{A}$ *has the spectral decomposition:* $\mathbf{A} = \sum_{j=1}^{d} \widehat{\delta}_j \boldsymbol{v}_j \boldsymbol{v}_j^T$, *where* $\widehat{\delta}_j$*'s are the eigenvalues and* $\boldsymbol{v}_j$*'s are the corresponding eigenvectors. Let* $\widetilde{\delta}_j = \tau_1 \cdot I\Big(\widehat{\delta}_j < \tau_1\Big) + \widehat{\delta}_j \cdot I\Big(\tau_1 \le \widehat{\delta}_j \le \tau_2\Big) + \tau_2 \cdot I\Big(\widehat{\delta}_j > \tau_2\Big)$, *we have* $\mathcal{P}_+^{'}(\mathbf{A}, \tau_1, \tau_2) = \sum_{j=1}^{d} \widetilde{\delta}_j \boldsymbol{v}_j \boldsymbol{v}_j^T$.

Proof. The proof of this lemma is presented in Appendix B. □

Therefore, using an iterative soft-thresholding and spectral thresholding procedure, we simultaneously control the smallest and largest eigenvalues of the estimated covariance matrix. By replacing Step 2 with Step 2a, the rate of convergence of the ISP algorithm preserves. The same convergence guarantee also holds for the EC2 with the largest eigenvalue constraint.

4. Statistical Properties

In this section, we analyze the theoretical properties of the EC2 estimator. In Section 3, we introduced several EC2 estimators: EC2 with ℓ_1-penalty, EC2 with MC+ penalty, EC2 with adaptive penalty, and EC2 with largest eigenvalue constraint. All these estimators have similar statistical properties. For conciseness, we focus on analyzing the EC2 estimator with ℓ_1-penalty as defined in (3.1) and (2.1) in this section. The same results hold for the other variants of the EC2 estimators by following the same argument.

Since EC2 first estimates the correlation matrix, we proceed to derive the asymptotic error bound for $\widehat{\boldsymbol{\Sigma}}^{\mathrm{EC2}}$. Let $0 \le q < 1$, we consider the following class of sparse correlation matrices:

$$\mathcal{M}(q, M_d, \delta) := \bigg\{\boldsymbol{\Sigma} : \max_{1\le j\le d} \sum_{k\ne j} \big|\Sigma_{jk}\big|^q \le M_d \text{ and } \Sigma_{jj} = 1 \text{ for all } j, \Lambda_{\min}\big(\boldsymbol{\Sigma}\big) \ge \delta\bigg\}.$$

We also define a class of covariance matrices:

$$\mathcal{U}(\kappa, q, M_d, \delta) := \bigg\{\boldsymbol{\Sigma} : \max_{j} \Sigma_{jj} \le \kappa \text{ and } \boldsymbol{\Theta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{-1} \in \mathcal{M}(q, M_d, \delta)\bigg\}, \tag{4.1}$$

where $\boldsymbol{\Theta}^{-1} = \mathrm{diag}(\sqrt{\Sigma_{11}}, \ldots, \sqrt{\Sigma_{dd}})$. The definition of this class is similar to the "universal thresholding class" defined by Bickel and Levina (2008b); Rothman et al. (2009). The main difference is that instead of assuming each column of the covariance matrix belongs to a ℓ_q-ball, we impose such conditions on the correlation matrix. Such a class is more suitable for our setting since the EC2 first estimates the correlation matrix.

Here we list down all the required assumptions:

(A1) Each marginal distribution of $\boldsymbol{X} := (X_1, \dots, X_d)$ is sub-Gaussian, i.e., there exists a finite constant K, such that

$$\max_{1\le j\le d} \mathbb{E}\exp\Big(t(X_j - \mathbb{E}X_j)\Big) \le \exp\left(\frac{K^2 t^2}{2}\right) \text{ for all } t. \tag{4.2}$$

(A2) The true covariance matrix $\boldsymbol{\Sigma}^*_{\rm cov} \in \mathcal{U}(\kappa, q, M_d, \delta_{\min})$ for some $\delta_{\min} > 0$.

(A3) The dimensionality d and sample size n satisfy $\limsup_{n\to\infty} (M_d)^{\frac{2}{1-q}} \dfrac{\log d}{n} = 0$.

These conditions are mild. Similar assumptions have been used in Rothman et al. (2009) to analyze the asymptotic properties of the STO estimator. Though this paper focuses on the sub-Gaussian family (as shown in Condition A1), the methodology and theory can be further extended to the semparametric Gaussian copula family (Liu et al., 2009, 2012), which further benefits Gaussian copula-based statistical methods such as Gaussian copula discriminant analysis and Gaussian copula component analysis in Han et al. (2012); Han and Liu (2012).

In the rest of this section, we provide both asymptotic and non-asymptotic analysis. For the asymptotic result, we show that, with high probability, the EC2 estimator in (3.1) is the same as the STO estimator in (2.2). This result suggests that the EC2 estimator preserves all the asymptotic properties of the STO estimator and is thus rate optimal in the minimax sense. However, such an asymptotic result holds only for large enough sample size n, which may not be satisfied in the finite-sample settings. To gain more insights, we also provide a non-asymptotic result of the EC2 estimator. Our non-asymptotic analysis achieves the same rate of convergence as that in Rothman (2012), which is optimal in terms of Frobenius norm error. To our best knowledge, EC2 is the first estimator that simultaneously achieves sparse and positive definite estimates and owes the optimal rate of convergence in terms of the spectral norm error. The optimal rate of convergence under the spectral norm has not been established in Rothman (2012); Lam and Fan (2009); Bien and Tibshirani (2011). It is still an open problem in the literature on whether their estimators can achieve the optimal rate of convergence under the spectral norm.

4.1. Asymptotic Results

We first present the asymptotic analysis, which suggests that, for large enough sample size n, the EC2 estimator is the same as the STO estimator. In this section, we always consider the high dimensional settings where $d > n$. The analysis of the low-dimensional case when $d < n$ is trivial.

Theorem 4.1. *Let $\widehat{\boldsymbol{\Sigma}}^{\rm EC2}$ be defined as in* (3.1). *Under assumptions (A1) to (A3), there exist universal constants c_0 and c_1, such that, by taking $\lambda = c_0\sqrt{\log d/n}$ and $\tau \le \delta_{\min}/2$, whenever*

$n \geq \left(\dfrac{2c_1 M_d}{\delta_{\min}}\right)^{\frac{2}{1-q}} \cdot \log d$, *we have* $\mathbb{P}\left(\widehat{\Sigma}^{\text{EC2}} = \widehat{\Sigma}^{\text{STO}}\right) \geq 1 - \frac{1}{d^5}$.

Proof. The proof of this theorem is presented in Appendix D. □

Theorem 4.1 shows that the solution to STO is also the solution to EC2 with high probability. Therefore our ISP algorithm always first checks the minimum eigenvalue of the STO estimator. The algorithm proceeds only when the feasibility condition is not satisfied. Empirically, in many situations, the soft-thresholding operator meets the requirement in Theorem 4.1. In these cases, the computation of the EC2 estimator is as efficient as the simple STO estimator. Also, in applications, we do not require exact knowledge of $\delta_{\min}$, a lower bound is enough. In fact, in later numerical experiment section, we propose a fully data-dependent approach to select τ.

The next theorem bounds the error term between the EC2 covariance estimate $\widehat{\Sigma}_{\text{cov}}^{\text{EC2}}$ and the true covariance Σ_{cov}^*.

Theorem 4.2. *Let* $\widehat{\Sigma}_{\text{cov}}^{\text{EC2}}$ *be defined as in* (2.1). *Under assumptions (A1) to (A3) and* $d > n$, *there exist universal constants* c_0 *and* c_1, *such that, by taking* $\lambda = c_0\sqrt{\log d/n}$ *and* $\tau \leq \delta_{\min}/2$, *whenever* $n \geq \left(\dfrac{2c_1 M_d}{\delta_{\min}}\right)^{\frac{2}{1-q}} \cdot \log d$, *we have*

$$\sup_{\Sigma_{\text{cov}}^* \in \mathcal{U}(\kappa, q, M_d, \delta_{\min})} \mathbb{E}\left\|\widehat{\Sigma}_{\text{cov}}^{\text{EC2}} - \Sigma_{\text{cov}}^*\right\|_2 \leq c_1 \cdot M_d \left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}. \tag{4.3}$$

Proof. The proof of this theorem is presented in Appendix E. □

As has been proved in Cai and Zhou (2012), the rate of convergence in (4.3) is minimax optimal over the class $\mathcal{U}(\kappa, q, M_d, \delta_{\min})$. Therefore, we have

$$\sup_{\Sigma_{\text{cov}}^* \in \mathcal{U}(\kappa, q, M_d, \delta_{\min})} \mathbb{E}\left\|\widehat{\Sigma}_{\text{cov}}^{\text{EC2}} - \Sigma_{\text{cov}}^*\right\|_2 \asymp M_d \left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}. \tag{4.4}$$

Theorem 4.2 shows that the EC2 estimator is asymptotically rate optimal. The main catch of this result is that it requires the condition $n \geq \left(\dfrac{2c_1 M_d}{\delta_{\min}}\right)^{\frac{2}{1-q}} \cdot \log d$. Under Assumption (A3), this would not be a problem in the asymptotic sense. However, in finite sample settings, this may no longer be true. In the next section, we provide a non-asymptotic result of the EC2 estimator. This result is comparable to the theoretical result as in Rothman (2012).

4.2. *Non-asymptotic Results*

In this section we derive the non-asymptotic error bound to further analyze the performance of EC2. To make our results comparable to that in Rothman (2012), we analyze the error rate in Frobenius norm and only consider the case that $\Sigma^*_{\text{cov}} \in \mathcal{U}(\kappa, 0, M_d, \delta_{\min})$. Our results show that EC2 achieves the optimal rate in Forbunius norm.

The EC2 solves a convex optimization problem, which results in an M-estimator. Therefore we can apply similar technique as in Negahban et al. (2012) to derive the following non-asymptotic error bound.

Theorem 4.3. *Let* $\widehat{\Sigma}^{\text{EC2}}$ *be defined as in* (3.1). *We assume (A1) and (A3) hold. For (A2), we assume* $\Sigma^*_{\text{cov}} \in \mathcal{U}(\kappa, 0, M_d, \delta_{\min})$. *We denote* N_d *to be the number of nonzero off-diagonal elements in* Σ^*_{cov}. *Then there exist universal constants* c_2 *and* c_3 *such that , by taking* $\lambda = c_2\sqrt{\log d/n}$ *and* $\tau \le \delta_{\min}$, *we have*

$$\mathbb{P}\left(\|\widehat{\Sigma}^{\text{EC2}} - \Sigma^*\|_{\mathsf{F}} \le c_3\sqrt{\frac{N_d \log d}{n}}\right) \ge 1 - \frac{1}{d^5}. \tag{4.5}$$

Proof. The proof of this theorem is presented in Appendix F. □

Remark 4.4. *If we directly exploit EC2 to estimate* Σ^*_{cov}, *we do not have* $(\Sigma^*_{\text{cov}})_{jj} = (\mathbf{S}_{\text{cov}})_{jj}$ *in (F.5). The proof will end up with a rate* $O_P(\sqrt{(d+N_d)\log d/n})$ *under the Frobenius norm. The extra term* $d \log d$ *on the numerator reflects the effort for estimating the diagonal elements.*

The rate of convergence under the Frobenius norm in Theorem 4.3 matches the minimax lower bound for correlation matrix estimation and is the same as the one obtained by Rothman (2012). Such a result leads to an error bound under the spectral norm as in the following corollary.

Corollary 4.5. *Under the same conditions as in Theorem 4.3 and* $d > n$, *we have*

$$\sup_{\Sigma^*_{\text{cov}} \in \mathcal{U}(\kappa,0,M_d,\delta_{\min})} \mathbb{E}\left\|\widehat{\Sigma}^{\text{EC2}}_{\text{cov}} - \Sigma^*_{\text{cov}}\right\|_2 \asymp \sqrt{\frac{N_d \log d}{n}}. \tag{4.6}$$

Proof. The proof follows from the same argument as Theorem 4.2, combined with the trivial bound

$$\|\widehat{\Sigma}^{\text{EC2}} - \Sigma^*\|_2 \le \|\widehat{\Sigma}^{\text{EC2}} - \Sigma^*\|_{\mathsf{F}}. \tag{4.7}$$

The detailed proof is omitted for conciseness. □

5. Numerical Experiments

We compare different variants of EC2 with the estimators proposed by Rothman et al. (2009) and Rothman (2012). More specifically, we consider nine methods:

- HTO: generalized thresholding with L_0-penalty (or the hard-thresholding estimator),
- STO: generalized thresholding with ℓ_1-penalty (or the soft-thresholding estimator),
- ATO-S: adaptive STO using the sample correlation matrix as the weight matrix,
- ATO-G: adaptive STO using the STO estimator as the weight matrix,
- MCTO: generalized thresholding estimator using the MC+ penalty,
- EC2-L1: EC2 with ℓ_1-penalty,
- AEC2-S: adaptive EC2 using the sample correlation matrix as the weight matrix,
- AEC2-G: adaptive EC2 using the STO estimator as the weight matrix.
- EC2-M: EC2 with MC+ penalty
- Log-Det: the estimator proposed in Rothman (2012) [1].

We select the tuning parameters of these methods in a data-dependent way as described in the next subsection.

5.1. *Tuning Parameter Selection*

We choose the tuning parameter by subsampling (Bickel and Levina, 2008a). More specifically, we randomly split the dataset K times. For splitting, the data is divided into a validation set of size n_2 and a training set of size $n_1 = n - n_2$. Theoretically, we can choose $n_2 \asymp n/\log n$. Practically, we choose $n_1 = \frac{3}{4}n$. Then we select the tuning parameter $\gamma = 1.01$ for EC2-M, which makes the performance of EC2-M close to HTO. We choose λ for STO and MCTO as

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \sum_{k=1}^{K} \|\widehat{\Sigma}_{\lambda}^{k,n_1} - \mathbf{S}^{k,n_2}\|_{\mathsf{F}}^2, \tag{5.1}$$

where $\widehat{\Sigma}_{\lambda}^{k,n_1}$ is the STO or MCTO correlation estimator, with λ computed with the training set of the k-th split and $\mathbf{S}^{k,n_2}$ is the sample correlation of the validation set of the k-th split. For EC2-L1 and EC2-M, we choose λ as the same as STO and MCTO respectively, and then tune τ as

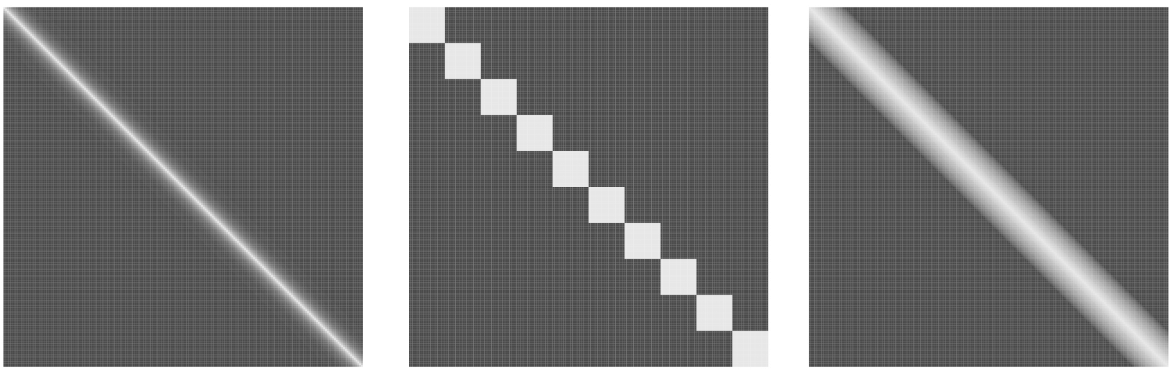
$$\widehat{\tau} = \underset{\tau}{\operatorname{argmin}} \sum_{k=1}^{K} \|\widehat{\Sigma}_{\tau}^{k,n_1} - \mathbf{S}^{k,n_2}\|_{\mathsf{F}}^2, \tag{5.2}$$

[1] The R package is available on `http://cran.r-project.org/web/packages/PDSCE/index.html`

where $\widehat{\Sigma}_{\tau}^{k,n_1}$ is the EC2-L1 or EC-M estimator using a similar notion in (5.1). For ATO, we tune λ over the gird using the same criterion as (5.1). For AEC2, we choose the same τ as EC2 and the same λ as ATO. For Log-Det, we take $\tau = 10^{-4}$ as suggested in Rothman (2012) and tune λ over the grid using the same criterion as (5.1).

5.2. Simulated Datasets

We use three models for the population covariance matrix. The heatmaps of these three covariance models with $d = 200$ are shown in Figure 2:



(a) Toeplitz (b) Block (c) Banded

Fig 2. *Heatmaps of three simulated covariance matrices for $d = 200$.*

(1) Toeplitz matrix: $\Sigma^*_{jk} = 0.75^{|j-k|}$. We know that for $d = 100, 200$ or 400, $\Lambda_{\min}(\Sigma^*) \approx 0.143$;

(2) Block matrix: The indices $1, \ldots, d$ are evenly divided into 10 groups with $\Sigma^*_{jk} = 0.8$ if k and j ($k \neq j$) belong to the same group and 0 otherwise. We know that for $d = 100, 200$ or 400, $\Lambda_{\min}(\Sigma^*) = 0.2$;

(3) Banded matrix: $\Sigma^*_{jk} = 1 - \frac{|j-k|}{20}$ if $|j-k| \leq 20$ and 0 otherwise. We know that $\Lambda_{\min}(\Sigma^*) \approx 0.034, 0.010, 0.002$ for $d = 100, 200, 400$ respectively.

Using these covariance models, we generate $n = 200$ independent data points from a d-dimensional Gaussian random variable with mean $\mathbf{0}$. The resulting covariance estimates are compared to the population covariance matrix using the spectral and Frobenius norms as evaluation metrics. We repeat this procedure 200 times and summarize the averaged estimation errors in Tables 1-3 with the corresponding standard errors in the parentheses. The histograms of the minimum eigenvalues of the HTO, STO, ATO-G, ATO-S and MCTO estimators are provided in Figures 3-5.

From these results, we see that STO, EC2-L1 and Log-Det achieve similar estimation performance in both Frobenius norm and spectral norm errors. In particularly, EC2-L1

Table 1
Quantitive comparison among 9 different methods for the Toeplitz setting

	d	HTO	STO	ATO-G	ATO-S	MCTO	EC2-L1	AEC2-G	AEC2-S	EC2-M	Log-Det
$\|\cdot\|_F$	100	4.9884 (0.2365)	4.8897 (0.3145)	4.5851 (0.2588)	4.5836 (0.2632)	4.9873 (0.2523)	4.8871 (0.3154)	4.4231 (0.2592)	4.4660 (0.2657)	4.6597 (0.2753)	4.8884 (0.3077)
	200	7.5706 (0.2147)	7.7281 (0.2982)	7.0908 (0.2515)	7.5720 (0.2389)	7.5826 (0.2486)	7.7254 (0.2992)	6.6514 (0.2346)	7.1443 (0.2386)	7.0286 (0.2486)	7.7528 (0.3025)
	400	11.292 (0.2305)	11.946 (0.3293)	10.779 (0.2462)	13.170 (0.2631)	11.262 (0.2453)	11.945 (0.3299)	9.8182 (0.2386)	11.748 (0.2487)	10.387 (0.2723)	11.901 (0.3301)
$\|\cdot\|_2$	100	1.8999 (0.2452)	2.3238 (0.2321)	1.8246 (0.2301)	1.8762 (0.2310)	1.8957 (0.2543)	2.3248 (0.2318)	1.8487 (0.2370)	1.8946 (0.2329)	1.9094 (0.2552)	2.3808 (0.2333)
	200	2.1781 (0.1920)	2.7299 (0.1686)	2.0674 (0.1995)	2.1844 (0.1931)	2.1775 (0.1569)	2.7306 (0.1683)	2.1221 (0.1939)	2.2052 (0.1891)	2.2203 (0.1562)	2.7352 (0.1678)
	400	2.4055 (0.1886)	3.0652 (0.1488)	2.2739 (0.1749)	2.6387 (0.1521)	2.3408 (0.1924)	3.0654 (0.1488)	2.3680 (0.1799)	2.5857 (0.1511)	2.3789 (0.2032)	3.0593 (0.1501)
P. D.	100	0/200	200/200	12/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200
	200	0/200	200/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200
	400	0/200	200/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200

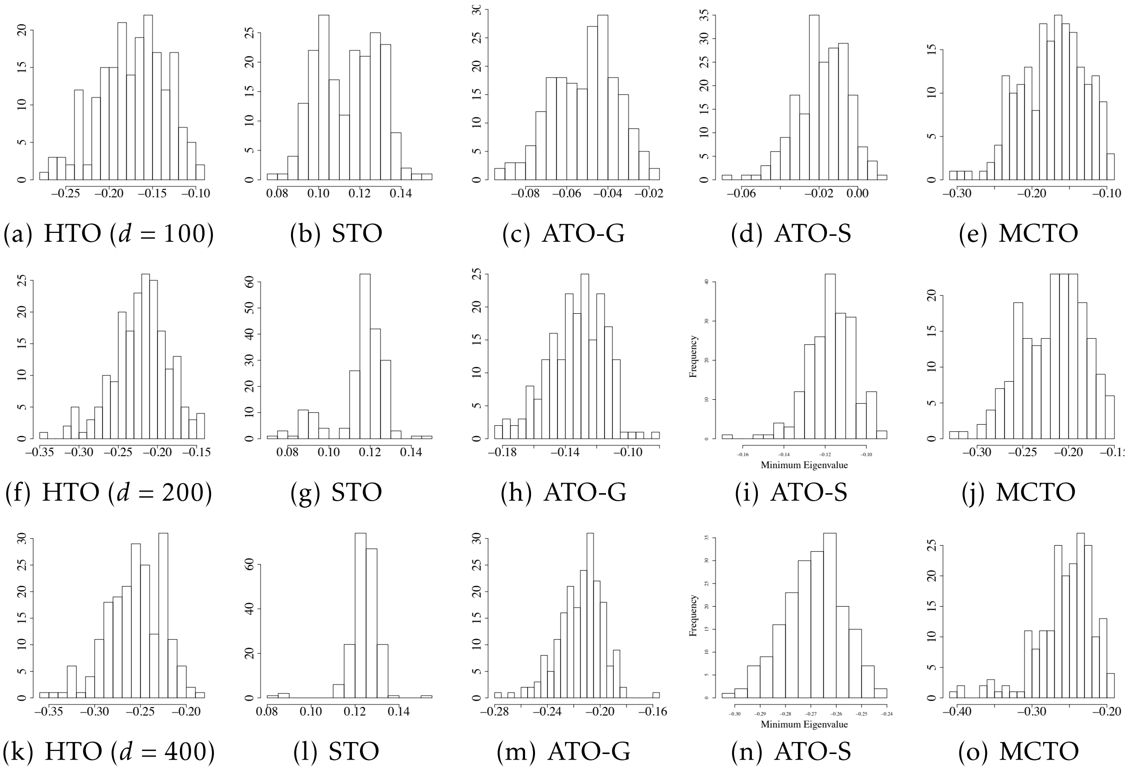
Table 2
Quantitive comparison among 9 different methods for the block setting

	d	HTO	STO	ATO-G	ATO-S	MCTO	EC2-L1	AEC2-G	AEC2-S	EC2-M	Log-Det
$\|\cdot\|_F$	100	2.8596 (0.5048)	4.2816 (0.5283)	3.5540 (0.5561)	3.5612 (0.6116)	2.8512 (0.5733)	4.2713 (0.5318)	3.1565 (0.5425)	3.3175 (0.5623)	2.6381 (0.5439)	4.2735 (0.5251)
	200	5.7245 (1.0861)	8.4954 (1.0666)	7.0914 (1.0696)	7.1843 (1.1559)	5.9215 (1.1145)	8.4488 (1.0813)	6.1891 (1.0291)	6.5814 (1.0736)	5.4628 (1.0063)	8.4821 (1.0774)
	400	11.5622 (1.8796)	17.3455 (2.1898)	14.6285 (2.1387)	14.8318 (2.4461)	11.085 (1.8598)	17.2004 (2.2313)	12.8227 (2.0607)	13.6331 (2.2910)	10.254 (1.8910)	17.331) (2.1517)
$\|\cdot\|_2$	100	1.6136 (0.4352)	2.1871 (0.3730)	1.8590 (0.4381)	1.8723 (0.4235)	1.5957 (0.4191)	2.1904 (0.3734)	1.7173 (0.4151)	1.7960 (0.3964)	1.4968 (0.4065)	2.1911 (0.3901)
	200	3.1632 (0.8901)	4.4013 (0.6786)	3.5856 (0.8145)	3.6534 (0.8067)	3.2964 (0.8873)	4.4140 (0.6811)	3.3320 (0.7236)	3.5143 (0.7233)	3.0491 (0.7064)	4.5012 (0.6978)
	400	6.4316 (1.5812)	8.8578 (1.3325)	7.4399 (1.6952)	7.5318 (1.7206)	6.0839 (1.6128)	8.8864 (1.3349)	6.8903 (1.5598)	7.2267 (1.5976)	5.7409 (1.5215)	8.8488 (1.3035)
P. D.	100	198/200	200/200	7/200	98/200	200/200	200/200	200/200	200/200	200/200	200/200
	200	200/200	179/200	0/200	2/200	200/200	200/200	200/200	200/200	200/200	200/200
	400	200/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200	200/200

Table 3
Quantitive comparison among 9 different methods for the banded setting

	d	HTO	STO	ATO-G	ATO-S	MCTO	EC2-L1	AEC2-G	AEC2-S	EC2-M	Log-Det
$\|\cdot\|_F$	100	5.8906 (1.0158)	6.2322 (1.0459)	5.6524 (1.0382)	5.7007 (1.0497)	5.9849 (1.0566)	6.2309 (1.0463)	5.5816 (1.0401)	5.6515 (1.0559)	5.4667 (1.1640)	6.2311 (1.0515)
	200	9.0050 (0.7644)	10.2387 (1.0703)	8.9650 (0.9514)	8.9833 (0.9171)	8.8706 (0.8509)	10.2301 (1.0734)	8.7203 (0.9565)	8.8152 (0.9314)	7.9104 (1.0280)	10.2201 (1.0561)
	400	13.5656 (0.8763)	16.1438 (1.0779)	13.9411 (1.1702)	14.8573 (1.1959)	13.620 (0.8347)	16.1224 (1.0814)	13.2626 (1.1416)	14.3028 (1.1758)	12.014 (0.9258)	(16.1389) 1.0991
$\|\cdot\|_2$	100	3.3106 (1.0020)	4.0640 (0.9660)	3.4677 (0.9929)	3.5170 (0.9852)	3.2976 (1.0265)	4.0643 (0.9661)	3.4701 (0.9830)	3.5184 (0.9832)	3.2817 (1.0044)	4.0553 (0.9881)
	200	3.8574 (0.7660)	5.2307 (0.8045)	4.2794 (0.8243)	4.3174 (0.7712)	3.8045 (0.7759)	5.2328 (0.8047)	4.2751 (0.8023)	4.3199 (0.7752)	3.7769 (0.8422)	5.2225 (0.8182)
	400	4.3551 (0.7719)	6.1805 (0.6044)	4.9620 (0.8302)	5.3328 (0.8063)	4.3890 (0.8208)	6.1846 (0.6043)	4.9356 (0.7235)	5.2862 (0.7599)	4.2781 (0.7020)	6.1708 (0.6001)
P. D.	100	0/200	0/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200
	200	0/200	0/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200
	400	0/200	0/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200

and Log-Det slightly outperforms STO in Frobenius norm errors, while STO has a slightly better spectral norm performance. For both block covariance and banded covariance models, EC2-L1 always guarantees the positive definiteness of the estimated covariance



(a) HTO ($d = 100$) (b) STO (c) ATO-G (d) ATO-S (e) MCTO

(f) HTO ($d = 200$) (g) STO (h) ATO-G (i) ATO-S (j) MCTO

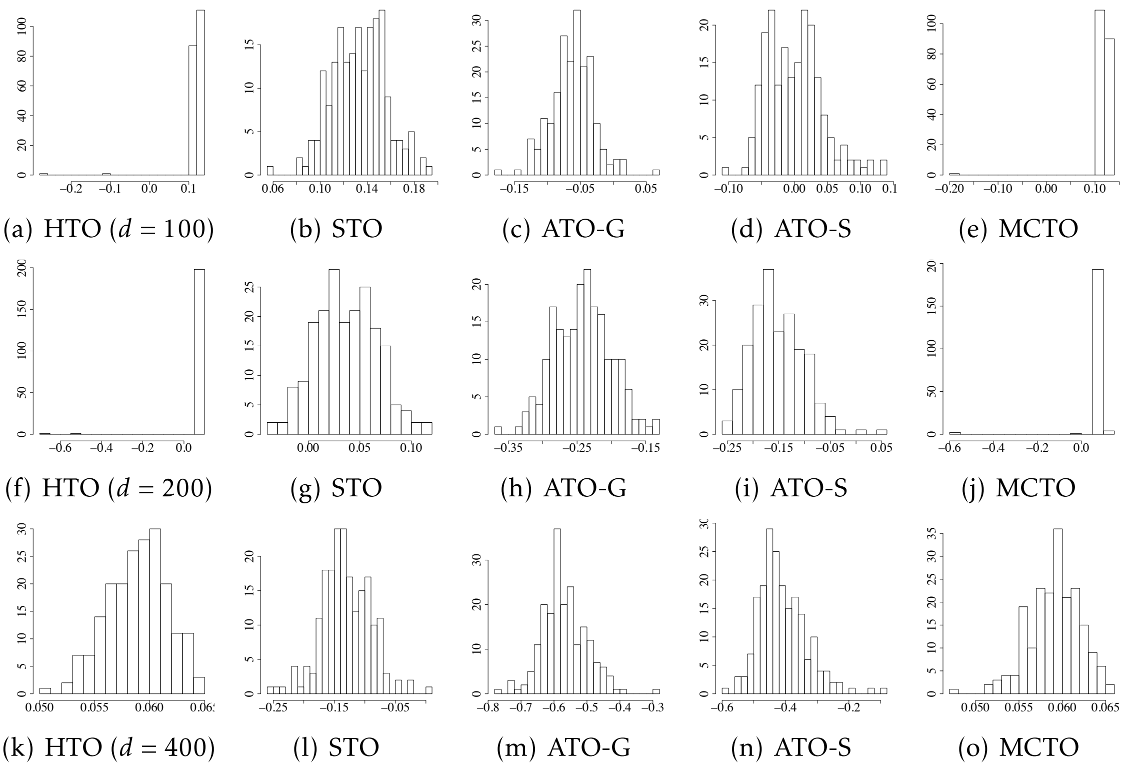
(k) HTO ($d = 400$) (l) STO (m) ATO-G (n) ATO-S (o) MCTO

Fig 3. *The histograms of the minimum eigenvalues for the Toeplitz setting.*

matrices, in contrast, STO almost never delivers positive definite estimates (especially for $d = 400$). This is similar to the results obtained by Rothman (2012).

Since the adaptive penalty reduces the estimation bias, ATO-S and ATO-G perform better than STO. ATO-G performs better than ATO-S. Similarly, both AEC2-S and AEC2-G outperform EC2-L1 with AEC2-G performs better than AEC2-S. Comparing AEC2-G with ATO-G, we see that AEC2-G outperforms ATO-G in most experimental settings. This results suggests that by utilizing the eigenvalue constraint, AEC2-G further improves the estimation performance.

HTO has a comparable performance to AEC2-G. For the Toeplitz covariance model, AEC2-G clearly outperforms HTO. For the block covariance model, HTO outperforms AEC2-G. For the banded covariance model, these two methods behave similarly: HTO slightly outperforms AEC2-G in spectral norm error, while AEC2-G is slightly superior to HTO in Forbenius norm error. The main reason that HTO achieves good estimation performance for the block covariance model is due to the fact that the hard-thresholded covariance matrix is always positive definite in this setting. The performance of HTO decreases on both Toeplitz and banded covariance models since the hard-thresholded covariance matrices are never positive definite in these two settings. This again confirms the fact that utilizing the eigenvalue constraints improves the estimation performance of



(a) HTO ($d = 100$) (b) STO (c) ATO-G (d) ATO-S (e) MCTO
(f) HTO ($d = 200$) (g) STO (h) ATO-G (i) ATO-S (j) MCTO
(k) HTO ($d = 400$) (l) STO (m) ATO-G (n) ATO-S (o) MCTO

Fig 4. *The histograms of the minimum eigenvalues for the block setting.*

an estimator.

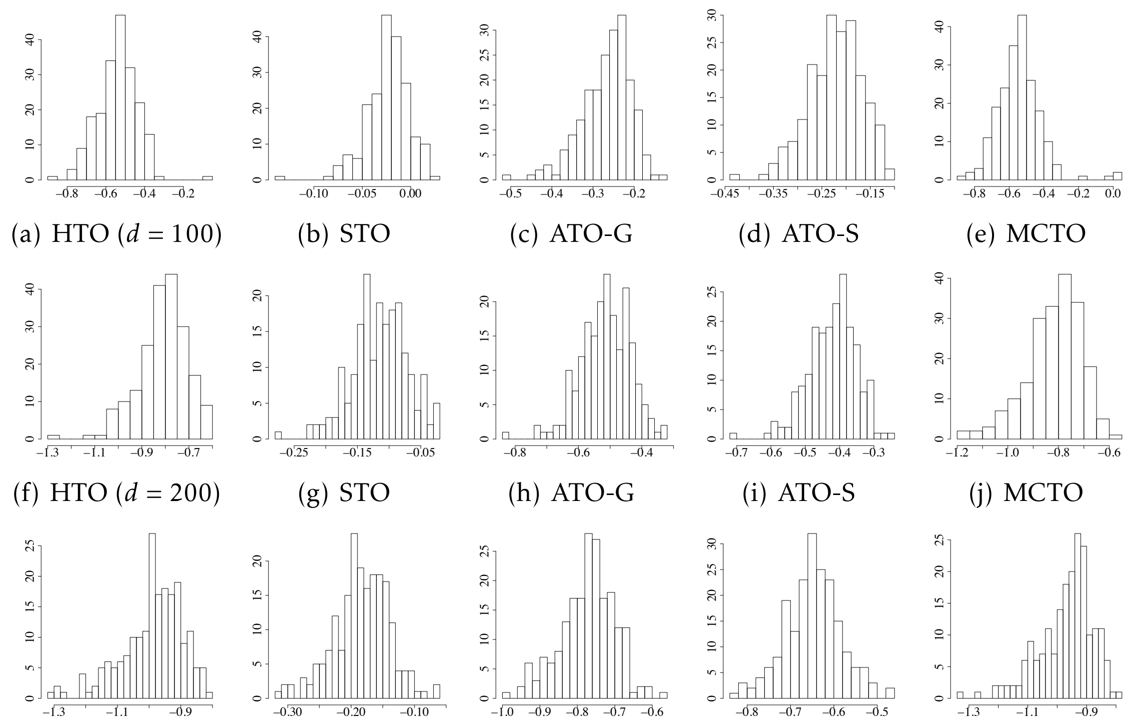
EC2-M and MCTO achieves good estimation performance in many settings. Especially for both block and banded covariance models, EC2-M achieves the best estimation performance in both Frobenius and spectral norm errors, and MCTO falls slightly behind the EC2-M due to the indefiniteness.

In summary, by adding the eigenvalue constraints we can obtain equally or better estimation than the naive thresholding. Also, by examining the histograms for minimum eigenvalues of the HTO, STO, ATO-G, ATO-S and MCTO estimates in Figures 3, 4, and 5, we see that all these thresholded estimators suffer different extents of indefiniteness, especially for the banded setting, no positive definite output is found over all replications.

5.3. Real Dataset

To illustrate the efficacy of the proposed methods, we adopt the Parkinsons disease dataset from the UCI machine learning data repository[2], which is the same as Rothman (2012). It contains 195 speech signals, of which 147 are from cases and 48 are from controls. Each signal is converted into 22 numerical variables. We plug the estimated covariance

[2]http://archive.ics.uci.edu/ml/



(a) HTO ($d = 100$) (b) STO (c) ATO-G (d) ATO-S (e) MCTO

(f) HTO ($d = 200$) (g) STO (h) ATO-G (i) ATO-S (j) MCTO

(k) HTO ($d = 400$) (l) STO (m) ATO-G (n) ATO-S (o) MCTO

Fig 5. *The histograms of the minimum eigenvalues for the banded setting.*

matrix into quadratic discriminant analysis (QDA), and the performance of the proposed covariance estimators are evaluated based on the classification error rates. Similar to the simulated datasets, we randomly partition the data 10 times into 65 training samples (49 from cases and 16 from controls) and 130 testing cases (98 from cases and 32 from controls).

Table 4
Classification error rates among 9 different estimators when plugged into QDA.

	HTO	STO	ATO-G	ATO-S	MCTO	EC2-L1	AEC2-G	AEC2-S	EC2-M	Log-Det
Test. Err.	N.A. (N.A.)	0.2291 (0.0266)	N.A. (N.A.)	N.A. (N.A.)	N.A. (N.A.)	0.2252 (0.0179)	0.2182 (0.0199)	0.2198 (0.0202)	0.2145 (0.0225)	22.56 (0.200)
P. D.	0/200	136/200	0/200	0/200	0/200	200/200	200/200	200/200	200/200	200/200

As is shown in Table 4, for all 200 partitions, HTO, ATO-G, ATO-S, MCTO are not applicable to QDA due to the indefiniteness. The STO estimator is positive definite and applicable to QDA for 136 out of 200 random partitions. The average classification error rate of all the 136 applicable STO estimates is 22.91%. While all EC2 estimators are guaranteed to be positive definite and work well for QDA. In terms of classification error, EC2-M achieves the best performance with an error rate 21.45%. The performance of

AEC2 is close to EC2-M. This result again confirms that the EC2 estimators achieve better estimation performance than the simple thresholded estimators.

6. Conclusions

We propose a novel approach named EC2 (Estimation of Covariance with Eigenvalue Constraints) for estimating high dimensional sparse covariance matrices with explicit eigenvalue constraints. EC2 is computationally tractable and theoretically justifiable. In particularly, EC2 achieves the optimal rates of convergence in terms of both spectral norm and Frobenius norm errors. Practically, we adopt a data-dependent approach for tuning parameter selection and numerical experiments on both simulated and real datasets are used to illustrate the usefulness of our method.

Acknowledgment

We thank the editor, associate editor and two referees for their helpful comments. Han Liu and Tuo Zhao are supported by NSF Grant III-1116730, and Lie Wang is supported by NSF Grant DMS-1005539.

References

Bickel, P. and Levina, E. (2004). Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010.

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.

Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Annals of Statistics* **36** 199–227.

Bien, J. and Tibshirani, R. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820.

Box, G., Jenkins, G. and Reinsel, G. (2011). *Time series analysis: forecasting and control*. Wiley.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.

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

Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics* **40** 2389–2420.

Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20** 33–61.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova* **4** 613–617.

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.

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.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* **2** 17–40.

Han, F. and Liu, H. (2012). Semiparametric principal component analysis. *Advances in Neural Information Processing Systems* **25** 171–179.

Han, F., Zhao, T. and Liu, H. (2012). Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research* **14** 629–671.

He, B. and Yuan, X. (2012). On non-ergodic convergence rate of douglas-rachford alternating direction method of multipliers. Tech. rep., Nanjing University.

Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.

Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics* **36** 2717–2756.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 42–54.

Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics* **2** 245–263.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics* **40** 2293–2326.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10** 2295–2328.

Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**.

Rothman, A. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99** 733–740.

Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104** 177–186.

Rothman, A., Levina, E. and Zhu, J. (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550.

Searle, S., Casella, G. and McCulloch, C. (2009). *Variance components*, vol. 419. Wiley-Interscience.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288.

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

Xue, L., Ma, S. and Zou, H. (2012). Positive definite l_1 penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107**.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics* **40**.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

Appendix for Sparse Covariance Matrix Estimation with Eigenvalue Constraints

Han Liu[1] **and Lie Wang**[2] **and Tuo Zhao**[3]

[1] *Department of Operations Research and Financial Engineering, Princeton University*
[2] *Department of Mathematics, Massachusetts Institute of Technology*
[3] *Department of Computer Science, Johns Hopkins University*

Appendix A: Proof of Propositions 3.1 and 3.6

Proof. It suffices to show that (3.6) is a convex program. We define two sets

$$\mathcal{A}_1 := \Big\{\Sigma : \text{ such that } \Sigma = \Sigma^T \text{ and } \tau_1 \leq \Lambda_{\min}(\Sigma)\Big\}, \tag{A.1}$$

$$\mathcal{A}_2 := \Big\{\Sigma : \text{ such that } \Sigma = \Sigma^T \text{ and } \tau_2 \geq \Lambda_{\max}(\Sigma)\Big\}. \tag{A.2}$$

We only need to prove that both $\mathcal{A}_1$ and $\mathcal{A}_2$ are convex sets. To show that $\mathcal{A}_1$ is a convex set, let $\Sigma_1, \Sigma_2 \in \mathcal{A}_1$ and $\alpha \in [0,1]$. Let

$$\widetilde{\Sigma} = \alpha\Sigma_1 + (1-\alpha)\Sigma_2. \tag{A.3}$$

Let $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, then for any $0 \leq \alpha \leq 1$ we have

$$\Lambda_{\min}\big(\widetilde{\Sigma}\big) = \min_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)\boldsymbol{v} \tag{A.4}$$

$$\geq \alpha \min_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T\Sigma_1\boldsymbol{v} + (1-\alpha)\min_{\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^T\Sigma_2\boldsymbol{w} \tag{A.5}$$

$$= \alpha\Lambda_{\min}(\Sigma_1) + (1-\alpha)\Lambda_{\min}(\Sigma_2) \tag{A.6}$$

$$= \tau_1. \tag{A.7}$$

1. Email: hanliu@princeton.edu
2. Email: liewang@math.princeton.edu
3. Email: tour@cs.jhu.edu

Therefore, $\widetilde{\Sigma} \in \mathcal{A}_1$. We show that $\mathcal{A}_1$ is a convex set. To show that $\mathcal{A}_2$ is a convex set, we follow a similar argument:

$$\begin{aligned}
\Lambda_{\max}\left(\widetilde{\Sigma}\right) &= \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)\boldsymbol{v} && \text{(A.8)}\\
&\le \alpha \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T\Sigma_1\boldsymbol{v} + (1-\alpha)\max_{\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^T\Sigma_2\boldsymbol{w} && \text{(A.9)}\\
&= \alpha\Lambda_{\max}(\Sigma_1) + (1-\alpha)\Lambda_{\max}(\Sigma_2) && \text{(A.10)}\\
&= \tau_2. && \text{(A.11)}
\end{aligned}$$

We thus complete the proof. □

Appendix B: Proof of Lemmas 3.2 and 3.7

Proof. Here we only prove Lemma 3.2, Lemma 3.7 follows the same argument and is omitted.

We rewrite the spectral decomposition of $\mathbf{A}$ as $\mathbf{A} = \mathbf{V}\mathbf{Z}\mathbf{V}^T$, where

$$\mathbf{Z} = \mathrm{diag}(\widehat{\delta}_1, \widehat{\delta}_2, \ldots, \widehat{\delta}_d) \text{ and } \mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d). \tag{B.1}$$

Since the Frobenius norm is invariant to the unitary matrix $\mathbf{V}$, we have

$$\min_{\mathbf{B}\succeq 0} \|\mathbf{B}-\mathbf{A}\|_{\mathsf{F}}^2 = \min_{\mathbf{B}\succeq 0} \|\mathbf{V}^T(\mathbf{B}-\mathbf{A})\mathbf{V}\|_{\mathsf{F}}^2 = \min_{\mathbf{B}\succeq 0} \|\mathbf{V}^T\mathbf{B}\mathbf{V}-\mathbf{Z}\|_{\mathsf{F}}^2. \tag{B.2}$$

Let $\widetilde{\delta}_j = \max\{\widehat{\delta}_j, \tau\}$, it is easy to check that when $\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{R} := \mathrm{diag}(\widetilde{\delta}_1, \widetilde{\delta}_2, \ldots, \widetilde{\delta}_d)$, we minimize (B.2). Therefore $\mathbf{B} = \mathbf{V}\mathbf{R}\mathbf{V}^T$ is the solution to the projection. □

Appendix C: Proof of Proposition 3.4

Proof. From the proof of Proposition 3.1, we know that the constraint sets are convex. Therefore, it suffices so show that the objective function in (3.13) is convex. Since (3.13) is elementwise decomposable, we only need to show that the following function is convex

$$F(\Sigma_{jk}) = \frac{1}{2}(\mathbf{S}_{jk} - \Sigma_{jk})^2 + \lambda \int_0^{|\Sigma_{jk}|} \left(1 - \frac{t}{\gamma\lambda}\right)_+ dt. \tag{C.1}$$

For $|\Sigma_{jk}| \ge \lambda\gamma$, the result follows from the fact that the penalty becomes a constant. For $|\Sigma_{jk}| < \lambda\gamma$, we have

$$F(\Sigma_{jk}) = \frac{1}{2}\mathbf{S}_{jk}^2 - \mathbf{S}_{jk}\Sigma_{jk} + \lambda|\Sigma_{jk}| + \left(1 - \frac{1}{\gamma}\right)\frac{|\Sigma_{jk}|^2}{2}. \tag{C.2}$$

Thus $F(\Sigma_{jk})$ is convex in $(-\infty, -\lambda\gamma)$, $(-\lambda\gamma, 0)$, $(0, \lambda\gamma)$, and $(\lambda\gamma, +\infty)$ respectively. Moreover, let $F'_{-}(\Sigma_{jk})$ and $F'_{+}(\Sigma_{jk})$ be the left and right derivates of $F(\Sigma_{jk})$ respectively, then we further check three critical points $-\gamma\lambda$, 0 and $\gamma\lambda$ and obtain

$$F'_{-}(-\gamma\lambda) = -\mathbf{S}_{jk} - \gamma\lambda = F'_{+}(-\gamma\lambda), \tag{C.3}$$

$$F'_{-}(0) = -\mathbf{S}_{jk} - \lambda \leq -\mathbf{S}_{jk} + \lambda = F'_{+}(0), \tag{C.4}$$

$$F'_{-}(\gamma\lambda) = -\mathbf{S}_{jk} + \gamma\lambda = F'_{+}(\gamma\lambda), \tag{C.5}$$

which further implies that the EC2 is globally convex when $\gamma > 1$. □

Appendix D: Proof of Theorem 4.1

Proof. By a simple modification of Theorem 1 in Rothman et al. (2009), we know that there exist universal constants c_0 and c_1, such that, by taking $\lambda = c_0\sqrt{\log d/n}$, we have

$$\mathbb{P}\left(\left\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\right\|_2 \leq c_1 M_d \left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}\right) \geq 1 - \frac{1}{d^5}. \tag{D.1}$$

We denote $\boldsymbol{v}$ to be the eigenvector that corresponds to the smallest eigenvalue of Σ^*. Then we have

$$\begin{aligned}
\left\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\right\|_2 &= \max_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{w}\|_2=1} \left\langle \boldsymbol{u}, (\widehat{\Sigma}^{\text{STO}} - \Sigma^*)\boldsymbol{w} \right\rangle && \text{(D.2)}\\
&\geq \left\langle \boldsymbol{v}, \left(\widehat{\Sigma}^{\text{STO}} - \Sigma^*\right)\boldsymbol{v} \right\rangle && \text{(D.3)}\\
&= \left\langle \boldsymbol{v}, \widehat{\Sigma}^{\text{STO}}\boldsymbol{v} \right\rangle - \left\langle \boldsymbol{v}, \Sigma^*\boldsymbol{v} \right\rangle && \text{(D.4)}\\
&\geq \Lambda_{\min}(\widehat{\Sigma}^{\text{STO}}) - \Lambda_{\min}(\Sigma^*). && \text{(D.5)}
\end{aligned}$$

By symmetry, we can show that

$$\left\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\right\|_2 \geq \Lambda_{\min}(\Sigma^*) - \Lambda_{\min}(\widehat{\Sigma}^{\text{STO}}). \tag{D.6}$$

Therefore, $\left|\Lambda_{\min}(\widehat{\Sigma}^{\text{STO}}) - \Lambda_{\min}(\Sigma^*)\right| \leq \|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\|_2 \leq c_1 M_d \left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}$, whenever

$$n > \left(\frac{2c_1 M_d}{\delta_{\min}}\right)^{\frac{2}{1-q}} \cdot \log d, \tag{D.7}$$

we have $\left|\Lambda_{\min}(\widehat{\Sigma}^{\text{STO}}) - \Lambda_{\min}(\Sigma^*)\right| \leq \frac{\delta_{\min}}{2}$, which implies that

$$\mathbb{P}\left(\Lambda_{\min}(\widehat{\Sigma}^{\text{STO}}) \geq \tau\right) \geq 1 - \frac{1}{d^5}, \tag{D.8}$$

which completes the proof. □

Appendix E: Proof of Theorem 4.2

Proof. In this analysis, we adopt a generic constant c , i.e., its value may varies from line to line. First, we have the following decomposition.

$$
\begin{aligned}
&\mathbb{E}\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 && \text{(E.1)}\\
=&\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} = \widehat{\Sigma}^{\mathrm{STO}}\right)\right] + \mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] && \text{(E.2)}\\
\leq&\mathbb{E}\left\|\widehat{\Sigma}^{\mathrm{STO}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 + \mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right], && \text{(E.3)}
\end{aligned}
$$

We first control the term $\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right]$, for this

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] && \text{(E.4)}\\
\leq\;& \mathbb{E}\left[\left\|\Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] + \mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] && \text{(E.5)}\\
\leq\;& \left\|\Sigma^*_{\mathrm{cov}}\right\|_2 \mathbb{E}\left[I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] + \sqrt{\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}}\right\|_2^2\right]}\sqrt{\mathbb{E} I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.6)}\\
\leq\;& \kappa \cdot d \cdot \mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right) + \sqrt{\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}\right\|_2^2 \|\widehat{\boldsymbol{\Theta}}\|_2^2\right]}\sqrt{\mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.7)}\\
\leq\;& \kappa \cdot d \cdot \mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right) + d\sqrt{\mathbb{E}\left[\|\widehat{\boldsymbol{\Theta}}\|_2^2\right]}\sqrt{\mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.8)}\\
\leq\;& \kappa \cdot d \cdot \mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right) + d\sqrt{\mathbb{E}\left[\left(\sum_{j=1}^{d} \widehat{\boldsymbol{\Theta}}_{jj}\right)^2\right]}\sqrt{\mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.9)}\\
\leq\;& \kappa \cdot d \cdot \mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right) + d\sqrt{d\left[\left(\sum_{j=1}^{d} \mathbb{E}\widehat{\boldsymbol{\Theta}}_{jj}^2\right)\right]}\sqrt{\mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.10)}
\end{aligned}
$$

From (4.2), it is easy to show that

$$
\max_{1\leq j\leq d} \mathbb{E}\widehat{\boldsymbol{\Theta}}_{jj}^2 \leq C_K, \tag{E.11}
$$

where C_K is a constant only depends on K as in (4.2). Therefore

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\Sigma}^{\mathrm{EC2}}_{\mathrm{cov}} - \Sigma^*_{\mathrm{cov}}\right\|_2 I\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)\right] && \text{(E.12)}\\
\leq\;& \kappa \cdot d \cdot \mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right) + d^2\sqrt{C_K}\sqrt{\mathbb{P}\left(\widehat{\Sigma}^{\mathrm{EC2}} \neq \widehat{\Sigma}^{\mathrm{STO}}\right)} && \text{(E.13)}\\
\leq\;& \frac{\kappa}{d^4} + \sqrt{\frac{C_K}{d}}. && \text{(E.14)}
\end{aligned}
$$

We then control the term $\mathbb{E}\left\|\widehat{\Sigma}^{\text{STO}}_{\text{cov}} - \Sigma^*_{\text{cov}}\right\|_2$. Let $\mathbf{\Theta}^* := \text{diag}(\theta^*_1, \ldots, \theta^*_d)$ be the diagonal matrix with θ^*_j denotes the standard deviation of the j-th variable X_j. The proof proceeds with the following decompositions:

$$
\begin{aligned}
&\widehat{\Sigma}^{\text{STO}}_{\text{cov}} - \Sigma^*_{\text{cov}} && \text{(E.15)}\\
=&\widehat{\mathbf{\Theta}}\widehat{\Sigma}^{\text{STO}}\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\Sigma^*\mathbf{\Theta}^* && \text{(E.16)}\\
=&(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)\Sigma^*\mathbf{\Theta}^* + (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)(\widehat{\Sigma}^{\text{STO}}\widehat{\mathbf{\Theta}} - \Sigma^*\mathbf{\Theta}^*) + \mathbf{\Theta}^*(\widehat{\Sigma}^{\text{STO}}\widehat{\mathbf{\Theta}} - \Sigma^*\mathbf{\Theta}^*).
\end{aligned}
$$

We further have

$$
\widehat{\Sigma}^{\text{STO}}\widehat{\mathbf{\Theta}} - \Sigma^*\mathbf{\Theta}^* = (\widehat{\Sigma}^{\text{STO}} - \Sigma^*)\mathbf{\Theta}^* + (\widehat{\Sigma}^{\text{STO}} - \Sigma^*)(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*) + \Sigma^*(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*). \tag{E.17}
$$

Since $\|\mathbf{\Theta}^*\|_2 = \max_j \mathbf{\Theta}^*_{jj} \le \sqrt{\kappa}$, then let $\eta_1 := \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2$ and $\eta_2 := \|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\|_2$, we have

$$
\begin{aligned}
&\|\widehat{\Sigma}^{\text{STO}}_{\text{cov}} - \Sigma^*_{\text{cov}}\|_2\\
\le&\underbrace{2\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2\|\Sigma^*\|_2\|\mathbf{\Theta}^*\|_2}_{T_1} + \underbrace{2\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\|_2\|\mathbf{\Theta}^*\|_2}_{T_2}\\
&+\underbrace{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2^2\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\|_2}_{T_3} + \underbrace{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2^2\|\Sigma^*\|_2}_{T_4} + \underbrace{\|\widehat{\Sigma}^{\text{STO}} - \Sigma^*\|_2\|\mathbf{\Theta}^*\|_2^2}_{T_5}\\
\le&\underbrace{2\sqrt{\kappa}\Lambda_{\max}\left(\Sigma^*\right)\eta_1}_{T_1} + \underbrace{\sqrt{\kappa}\eta_1\eta_2}_{T_2} + \underbrace{\eta_1^2\eta_2}_{T_3} + \underbrace{\Lambda_{\max}\left(\Sigma^*\right)\eta_1^2}_{T_4} + \underbrace{\kappa\eta_2}_{T_5}.
\end{aligned} \tag{E.18}
$$

To bound the term $\Lambda_{\max}\left(\Sigma^*\right)$, we have, for $0 \le q < 1$,

$$
\Lambda_{\max}\left(\Sigma^*\right) \le \|\Sigma^*\|_1 \le 1 + M_d^q \le 1 + M_d. \tag{E.19}
$$

We now bound the term η_1. By the following decomposition

$$
\begin{aligned}
\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\mathbf{\Theta}^*\|_2 &= \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)(\widehat{\mathbf{\Theta}} + \mathbf{\Theta}^*)\|_2\\
&\le \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2 + 2\|\mathbf{\Theta}^*\|_2) \le \eta_1^2 + 2\eta_1\sqrt{\kappa},
\end{aligned} \tag{E.20}
$$

we have $\mathbb{P}(\eta_1^2 + 2\eta_1\sqrt{\kappa} \ge \epsilon) \le \mathbb{P}(\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\mathbf{\Theta}^*\|_2 \ge \epsilon) \le 2d\exp(-c_1 n\epsilon^2)$, where the second inequality follows from the chi-square tail bound. Assume $\epsilon \le 3\kappa$, then by solving the inequality $\eta_1^2 + 2\eta_1\sqrt{\kappa} \ge \epsilon$, we have

$$
\mathbb{P}\left(\eta_1 \ge \frac{\epsilon}{\sqrt{\kappa} + \sqrt{\kappa + \epsilon}} \ge \frac{\epsilon}{3\sqrt{\kappa}}\right) \le 2d\exp(-c_1 n\epsilon^2), \tag{E.21}
$$

Therefore, there exists a constant c_1, such that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_2 \le c_1\sqrt{\frac{\kappa\log d}{n}}\right) \ge 1-\frac{1}{d^5}. \tag{E.22}$$

By a simple modification of Theorem 1 in Rothman et al. (2009), we know that there exist universal constants c_0 and c_1, such that, by taking $\lambda = c_0\sqrt{\log d/n}$, we have

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{STO}}-\boldsymbol{\Sigma}^*\right\|_2 \le c_2 M_d\left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}\right) \ge 1-\frac{1}{d^5}. \tag{E.23}$$

By plugging (E.22) and (E.23) into (E.18), we further obtain

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{STO}}_{\mathrm{cov}}-\boldsymbol{\Sigma}^*_{\mathrm{cov}}\|_2 \\
&\le \underbrace{2\sqrt{\kappa}\Lambda_{\max}\left(\boldsymbol{\Sigma}^*\right)\eta_1}_{T_1}+\underbrace{2\sqrt{\kappa}\eta_1\eta_2}_{T_2}+\underbrace{\eta_1^2\eta_2}_{T_3}+\underbrace{\Lambda_{\max}\left(\boldsymbol{\Sigma}^*\right)\eta_1^2}_{T_4}+\underbrace{\kappa\eta_2}_{T_5} \\
&\le \underbrace{2c_1\kappa(1+M_d)\sqrt{\frac{\log d}{n}}}_{T_1}+\underbrace{2c_1c_2\kappa M_d\left(\frac{\log d}{n}\right)^{1-\frac{q}{2}}}_{T_2}+\underbrace{c_1^2c_2\kappa M_d\left(\frac{\log d}{n}\right)^{\frac{3}{2}-\frac{q}{2}}}_{T_3} \\
&\quad+\underbrace{\kappa(1+M_d)\frac{\log d}{n}}_{T_4}+\underbrace{c_2\kappa M_d\left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}}_{T_5}
\end{aligned} \tag{E.24}$$

with probability at least $1-2/d^5$. Under Assumption A3, T_1, T_2, T_3 and T_4 can be asymptotically ignored because they converge to 0 with a faster rate than T_5. Consequently, using the same argument as from (E.4) to (E.14), we have that, there exist generic constants c_3 and c_4, such that

$$\mathbb{E}\left\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{STO}}_{\mathrm{cov}}-\boldsymbol{\Sigma}^*_{\mathrm{cov}}\right\|_2 \le c_3 M_d\left(\frac{\log d}{n}\right)^{\frac{1-q}{2}}+c_4\sqrt{\frac{C_K}{d}}, \tag{E.25}$$

where C_K is a constant only depends on K as in (4.2). The desired result then follows by piecing all these terms together. □

Appendix F: Proof of Theorem 4.3

Proof. Since $\tau \le \delta_{\min}$, $\boldsymbol{\Sigma}^*$ is a feasible solution to (3.1) and we have

$$\frac{1}{2}\|\mathbf{S}-\widehat{\boldsymbol{\Sigma}}^{\mathrm{EC2}}\|_{\mathsf{F}}^2+\lambda\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{EC2}}\|_{1,\mathrm{off}} \le \frac{1}{2}\|\mathbf{S}-\boldsymbol{\Sigma}^*\|_{\mathsf{F}}^2+\lambda\|\boldsymbol{\Sigma}^*\|_{1,\mathrm{off}}. \tag{F.1}$$

Let $\widehat{\boldsymbol{\Delta}} = \widehat{\Sigma}^{\text{EC2}} - \Sigma^*$, then after simple manipulation we have

$$\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 - \left\langle \mathbf{S} - \Sigma^*, \widehat{\boldsymbol{\Delta}} \right\rangle + \lambda\|\Sigma^* + \widehat{\boldsymbol{\Delta}}\|_{1,\text{off}} - \lambda\|\Sigma^*\|_{1,\text{off}} \le 0. \tag{F.2}$$

Let $\text{supp}(\Sigma^*) := \left\{(j,k) \in \mathbb{R}^{d\times d} \,|\, \Sigma^*_{jk} \neq 0\right\}$, we define $\mathcal{E}$ and $\mathcal{E}^{\perp}$ as two subets of $\mathbb{R}^{d\times d}$,

$$\mathcal{E} = \{\mathbf{A} \in \mathbb{R}^{d\times d} | \mathbf{A}_{jk} = 0 \text{ for all } (j,k) \notin \text{supp}(\Sigma^*)\}, \text{and } \mathcal{E}^{\perp} = \mathbb{R}^{d\times d} \setminus \mathcal{E}.$$

We use $\mathbf{A}_{\mathcal{E}}$ to denote the projection of the matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ to the subspace $\mathcal{E}$. More specifically, the element on the j-th row and k-th column of $\mathbf{A}_{\mathcal{E}}$ is the same as $\mathbf{A}$ if $(j,k) \in \text{supp}(\Sigma^*)$. Otherwise, the value is 0. Moreover, let $\mathbf{A}_{\mathcal{E}^{\perp}} = \mathbf{A} - \mathbf{A}_{\mathcal{E}}$. Since $\|\cdot\|_{1,\text{off}}$ is decomposable, we have

$$\begin{aligned}
\|\Sigma^* + \widehat{\boldsymbol{\Delta}}\|_{1,\text{off}} &= \|\Sigma^*_{\mathcal{E}} + \Sigma^*_{\mathcal{E}^{\perp}} + \widehat{\boldsymbol{\Delta}}_{\mathcal{E}} + \widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} \\
&\ge \|\Sigma^*_{\mathcal{E}}\|_{1,\text{off}} + \|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} - \|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} - \|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}}.
\end{aligned} \tag{F.3}$$

Combining (F.3) with the decomposition $\|\Sigma^*\|_{1,\text{off}} = \|\Sigma^*_{\mathcal{E}}\|_{1,\text{off}} + \|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}}$, we have

$$\|\Sigma^* + \widehat{\boldsymbol{\Delta}}\|_{1,\text{off}} - \|\Sigma^*\|_{1,\text{off}} \ge \|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} - \|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} - 2\|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}}. \tag{F.4}$$

By the Cauchy-Schwartz inequality, we also have

$$\begin{aligned}
-\left\langle \mathbf{S} - \Sigma^*, \widehat{\boldsymbol{\Delta}} \right\rangle &\ge -\Big|\sum_{j\neq k}(\mathbf{S}_{jk} - \Sigma^*_{jk})\widehat{\boldsymbol{\Delta}}_{jk}\Big| - \Big|\sum_{j=1}^{d}(\mathbf{S}_{jj} - \Sigma^*_{jj})\widehat{\boldsymbol{\Delta}}_{j,j}\Big| \\
&= -\Big|\sum_{j\neq k}(S_{jk} - \Sigma^*_{jk})\widehat{\boldsymbol{\Delta}}_{jk}\Big| \\
&\ge -\|\mathbf{S} - \Sigma^*\|_{\max,\text{off}}\|\widehat{\boldsymbol{\Delta}}\|_{1,\text{off}},
\end{aligned} \tag{F.5}$$

where the equality comes from $\boldsymbol{\Sigma}^*_{jj} = \mathbf{S}_{jj} = 1$ and $\|\mathbf{A}\|_{\max,\text{off}} := \max_{j\neq k}|\mathbf{A}_{jk}|$.

Consider the event $\|\mathbf{S} - \Sigma^*\|_{\max,\text{off}} \le \lambda/2$, by combining (F.2), (F.4) and (F.5), we have

$$\begin{aligned}
0 &\ge \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 - \frac{\lambda}{2}(\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} + \|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}}) + \lambda\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} - \lambda\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} - 2\lambda\|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}} \\
&= \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 - \frac{3\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} + \frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} - 2\lambda\|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}}.
\end{aligned} \tag{F.6}$$

Since $\|\boldsymbol{\Delta}\|_{\mathsf{F}} \ge 0$, we have $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}^{\perp}}\|_{1,\text{off}} \le 3\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} + 4\|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}}$. Therefore by combining with (F.6), we have

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 \le 3\lambda\|\widehat{\boldsymbol{\Delta}}_{\mathcal{E}}\|_{1,\text{off}} + 4\lambda\|\Sigma^*_{\mathcal{E}^{\perp}}\|_{1,\text{off}} \le 3\lambda\sqrt{N_d}\|\boldsymbol{\Delta}_{\mathcal{E}}\|_{\mathsf{F}}. \tag{F.7}$$

The second equality comes from the fact that $\mathcal{E}$ has any most N_d non-zero elements and $\Sigma^*_{\mathcal{E}^\perp} = 0$.

Since the marginal distributions of $\mathbf{X}$ are sub-Gaussian, there exists a constant C such that

$$\mathbb{P}\Big(|\mathbf{S}_{jk} - \Sigma^*_{jk}| > \epsilon\Big) \le 2\exp(-Cn\epsilon^2). \tag{F.8}$$

By taking the union bound we have

$$\mathbb{P}\left(\|\mathbf{S} - \Sigma^*\|_{\max,\mathrm{off}} > \epsilon\right) \le 2d^2 \exp(-Cn\epsilon^2) = 2\exp(-Cn\epsilon^2 + 2\log d). \tag{F.9}$$

Let c_2 be a large enough constant, we have, by taking $\lambda = c_2\sqrt{\log d/n}$, there exists a constant c_3, such that $\mathbb{P}\left(\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}} \ge c_3\sqrt{\frac{N_d \log d}{n}}\right) \le \frac{1}{d^5}$. This finishes the proof. □

References

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104** 177–186.