Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Adaptive Thresholding for Sparse Covariance Matrix Estimation

Tony CAI and Weidong LIU

In this article we consider estimation of sparse covariance matrices and propose a thresholding procedure that is adaptive to the variability of individual entries. The estimators are fully data-driven and demonstrate excellent performance both theoretically and numerically. It is shown that the estimators adaptively achieve the optimal rate of convergence over a large class of sparse covariance matrices under the spectral norm. In contrast, the commonly used universal thresholding estimators are shown to be suboptimal over the same parameter spaces. Support recovery is discussed as well. The adaptive thresholding estimators are easy to implement. The numerical performance of the estimators is studied using both simulated and real data. Simulation results demonstrate that the adaptive thresholding estimators uniformly outperform the universal thresholding estimators. The method is also illustrated in an analysis on a dataset from a small round blue-cell tumor microarray experiment. A supplement to this article presenting additional technical proofs is available online.

KEY WORDS: Frobenius norm; Optimal rate of convergence; Spectral norm; Support recovery; Universal thresholding.

1. INTRODUCTION

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a p-variate random vector with covariance matrix $\mathbf{\Sigma}_0$. Given an iid random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ from the distribution of $\mathbf{X}$, we wish to estimate the covariance matrix $\mathbf{\Sigma}_0$ under the spectral norm. This covariance matrix estimation problem is of fundamental importance in multivariate analysis with a wide range of applications. The high-dimensional setting, in which the dimension p can be much larger than the sample size n, is of particular current interest. In such a setting, conventional methods and results based on fixed p and large n are no longer applicable, and thus new methods and theories are needed. In particular, the sample covariance matrix

$$\mathbf{\Sigma}_n = (\hat{\sigma}_{ij})_{p\times p} := \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T, \qquad (1)$$

where $\bar{\mathbf{X}} = n^{-1}\sum_{k=1}^{n}\mathbf{X}_k$, performs poorly in this setting, and structural assumptions are required to estimate the covariance matrix consistently.

In this article we focus on estimating sparse covariance matrices. This problem has been considered in the literature. El Karoui (2008) and Bickel and Levina (2008) proposed thresholding of the sample covariance matrix $\mathbf{\Sigma}_n$ and obtained rates of convergence for the thresholding estimators. Rothman, Levina, and Zhu (2009) considered thresholding of the sample covariance matrix with more general thresholding functions. Cai and Zhou (2009, 2010) established the minimax rates of convergence under the matrix ℓ_1 norm and the spectral norm. Wang and Zou (2010) considered estimation of volatility matrices based on high-frequency financial data.

A common feature of the thresholding methods for sparse covariance matrix estimation proposed in the literature is that they all belong to the class of "universal thresholding rules"; that is, a single threshold level is used to threshold all the entries of the sample covariance matrix. Universal thresholding rules were originally introduced by Donoho and Johnstone (1994, 1998) for estimating sparse normal mean vectors in the context of wavelet function estimation (see also Antoniadis and Fan 2001). An important feature of the problems that those authors considered is that noise is homoscedastic. In such a setting, universal thresholding has demonstrated considerable success in nonparametric function estimation in terms of asymptotic optimality and computational simplicity.

In contrast to the standard homoscedastic nonparametric regression problems, sparse covariance matrix estimation is intrinsically a heteroscedastic problem, in the sense that the entries of the sample covariance matrix could have a wide range of variability. Although some universal thresholding rules have been shown to have desirable asymptotic properties, this is related mainly to the fact that the parameter space considered in the literature is relatively restrictive, which forces the covariance matrix estimation problem to be an essentially homoscedastic problem.

To illustrate this point, it is helpful to consider an idealized model in which

$$y_i = \mu_i + \gamma_i z_i, \qquad z_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1),\ 1 \le i \le p \qquad (2)$$

and one wishes to estimate the mean vector μ, which is assumed to be sparse. If the noise levels γ_i are bounded, say by B, then the universal thresholding rule $\hat{\mu}_i = y_i I(|y_i| \ge B\sqrt{2\log p})$ performs well asymptotically over a standard ℓ_q ball $\Theta_q(s_0)$, defined by

$$\Theta_q(s_0) = \left\{\boldsymbol{\mu} \in \mathbb{R}^p : \sum_{j=1}^{p}|\mu_j|^q \le s_0\right\}. \qquad (3)$$

In particular, $\Theta_0(s_0)$ is a set of sparse vectors with at most s_0 nonzero elements. Here the assumption that γ_i are bounded by B is crucial. The universal thresholding rule simply treats the heteroscedastic problem (2) as a homoscedastic one with all noise levels $\gamma_i = B$. It is intuitively clear that this method does not perform well when the range of γ_i is large, and that it

Tony Cai is Dorothy Silberberg Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: tcai@wharton.upenn.edu). Weidong Liu is Faculty Member, Department of Mathematics and Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China and Postdoctoral Fellow, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

Journal of the American Statistical Association
June 2011, Vol. 106, No. 494, Theory and Methods
DOI: 10.1198/jasa.2011.tm10560

fails completely without the uniform boundedness assumption on the γ_i's.

For sparse covariance matrix estimation, the following uniformity class of sparse matrices was considered by Bickel and Levina (2008) and Rothman, Levina, and Zhu (2009):

$$\mathcal{U}_q := \mathcal{U}_q(s_0(p)) = \left\{ \mathbf{\Sigma} : \mathbf{\Sigma} \succ 0,\ \max_i \sigma_{ii} \le K,\ \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \le s_0(p) \right\}$$

for some $0 \le q < 1$, where $\mathbf{\Sigma} \succ 0$ means that $\mathbf{\Sigma}$ is positive definite. Here each column of a covariance matrix in $\mathcal{U}_q(s_0(p))$ is assumed to be in the ℓ_q ball $\Theta_q(s_0(p))$. Define

$$\theta_{ij} := \mathsf{Var}((X_i - \mu_i)(X_j - \mu_j)), \tag{4}$$

where $\mu_i = \mathsf{E}X_i$. It is easy to see that in the Gaussian case, $\sigma_{ii}\sigma_{jj} \le \theta_{ij} \le 2\sigma_{ii}\sigma_{jj}$. The condition $\max_i \sigma_{ii} \le K$ for all i ensures that the variances of the entries of the sample covariance matrix is uniformly bounded. Bickel and Levina (2008) proposed a universal thresholding estimator $\hat{\mathbf{\Sigma}}_u = (\hat{\sigma}^u_{ij})$, where

$$\hat{\sigma}^u_{ij} = \hat{\sigma}_{ij} I\{\hat{\sigma}_{ij} \ge \lambda_n\}, \tag{5}$$

and showed that with a proper choice of the threshold λ_n, the estimator $\hat{\mathbf{\Sigma}}_u$ achieves a desirable rate of convergence under the spectral norm. Rothman, Levina, and Zhu (2009) considered a class of universal thresholding rules with more general thresholding functions than hard thresholding. Similar to the idealized model (2) discussed earlier, here a key assumption is that the variances σ_{ii} are uniformly bounded by K, which is crucial to make the universal thresholding rules well behaved. A universal thresholding rule in this case essentially treats the problem as if all $\sigma_{ii} = K$ when selecting the threshold λ.

For heteroscedastic problems, such as sparse covariance matrix estimation, it is arguably more desirable to use thresholds that capture the variability of individual variables instead of using a universal upper bound. This is particularly true when the variances vary over a wide range or when no obvious upper bound on the variances is known. A more natural and effective approach is to use thresholding rules with entry-dependent thresholds that automatically adapt to the variability of the individual entries of the sample covariance matrix. The main goal of the present work is to develop such an adaptive thresholding estimator and study its properties.

In this article we introduce an adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^\star = (\hat{\sigma}^\star_{ij})_{p\times p}$ with

$$\hat{\sigma}^\star_{ij} = s_{\lambda_{ij}}(\hat{\sigma}_{ij}), \tag{6}$$

where $s_\lambda(z)$ is a general thresholding function similar to those used by Rothman, Levina, and Zhu (2009), which we specify later. The individual thresholds λ_{ij} are fully data-driven and adapt to the variability of individual entries of the sample covariance matrix $\mathbf{\Sigma}_n$. We show that the adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^\star$ has excellent properties, both asymptotically and numerically. In particular, we consider the performance of the estimator $\hat{\mathbf{\Sigma}}^\star$ over a large class of sparse covariance matrices defined by

$$\mathcal{U}^\star_q := \mathcal{U}^\star_q(s_0(p)) = \left\{ \mathbf{\Sigma} : \mathbf{\Sigma} \succ 0,\ \max_i \sum_{j=1}^{p} (\sigma_{ii}\sigma_{jj})^{(1-q)/2} |\sigma_{ij}|^q \le s_0(p) \right\} \tag{7}$$

for $0 \le q < 1$. Compared with $\mathcal{U}_q(s_0(p))$, the columns of a covariance matrix in $\mathcal{U}^\star_q$ are required to be in a weighted ℓ_q ball instead of a standard ℓ_q ball, with the weight determined by the variance of the entries of the sample covariance matrix. A particular feature of $\mathcal{U}^\star_q$ is that it no longer requires the variances σ_{ii} to be uniformly bounded and allows $\max_i \sigma_{ii} \to \infty$. Note that $\mathcal{U}_q(s_0(p)) \subseteq \mathcal{U}^\star_q(K^{1-q}s_0(p))$, so the parameter space $\mathcal{U}^\star_q$ contains the uniformity class $\mathcal{U}_q$ as a subset. The parameter space $\mathcal{U}^\star_q$ also can be viewed as a weighted ℓ_q ball of correlation coefficients; see Section 3.1 for more discussion.

In Section 3 we show that $\hat{\mathbf{\Sigma}}^\star$ achieves the optimal rate of convergence,

$$s_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2},$$

over the parameter space $\mathcal{U}^\star_q(s_0(p))$. In comparison, we also show that the best universal thresholding estimator can only attain the rate $s_0^{2-q}(p)(\frac{\log p}{n})^{(1-q)/2}$ over $\mathcal{U}^\star_q(s_0(p))$, which is clearly suboptimal when $s_0(p) \to \infty$, because $q < 1$.

The choice of regularization parameters is important in any regularized estimation problem. The thresholds λ_{ij} used in (6) are based on an estimator of the variance of the entries $\hat{\sigma}_{ij}$ of the sample covariance matrix. More specifically, λ_{ij} are of the form

$$\lambda_{ij} = \delta\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}, \tag{8}$$

where $\hat{\theta}_{ij}$ are estimates of θ_{ij} defined in (4) and δ is a tuning parameter. The value of δ can be taken as fixed at $\delta = 2$ or can be chosen empirically through cross-validation. We investigate the theoretical properties of the resulting covariance matrix estimators using both methods, and show that the estimators attain the optimal rate of convergence under the spectral norm in both cases. We also consider support recovery of a sparse covariance matrix.

The adaptive thresholding estimators are easy to implement. We investigate the numerical performance of the estimators using both simulated and real data. Our simulation results indicate that the adaptive thresholding estimators perform favorably compared with existing methods. In particular, they uniformly outperform the universal thresholding estimators in the simulation studies. We also apply the procedure to analyze a dataset from a small round blue-cell tumor microarray experiment (Khan et al. 2001).

The article is organized as follows. Section 2 introduces the adaptive thresholding procedure for sparse covariance matrix estimation, and Section 3 considers asymptotic properties. It is shown that the adaptive thresholding estimator is rate-optimal over $\mathcal{U}^\star_q$, whereas the best universal thresholding estimator is proved to be suboptimal. Section 4 discusses data-driven selection of the thresholds using cross-validation (CV) and establishes asymptotic optimality of the resulting estimator. Section 5 investigates the numerical performance of the adaptive thresholding estimators by simulations and by an application to a dataset from a small round blue-cell tumor microarray experiment. Section 6 discusses methods based on the sample correlation matrix, and Section 7 provides proofs.

2. ADAPTIVE THRESHOLDING FOR SPARSE COVARIANCE MATRIX

In this section we introduce the adaptive thresholding method for estimating sparse covariance matrices. To motivate our estimator, consider again the sparse normal mean estimation problem (2). If the noise levels γ_i's are known or can be well estimated, then a good estimator of the mean vector is the hard thresholding estimator $\hat{\mu}_i = y_i I\{|y_i| \geq \gamma_i\sqrt{2\log p}\}$ or some generalized thresholding estimator with the same thresholds, $\gamma_i\sqrt{2\log p}$.

Similarly, for sparse covariance matrix estimation, a more effective thresholding rule than universal thresholding is one that adapts to the variability of the individual entries of the sample covariance matrix. Define θ_{ij} as in (4). Then, roughly speaking, sparse covariance matrix estimation is similar to the mean vector estimation problem based on the observations

$$\frac{1}{n}\sum_{k=1}^{n}(X_{ki}-\mu_i)(X_{kj}-\mu_j) = \sigma_{ij} + \sqrt{\frac{\theta_{ij}}{n}}z_{ij}, \qquad 1 \leq i,j \leq p, \tag{9}$$

with z_{ij} asymptotically standard normal. This analogy provides a good motivation for our adaptive thresholding procedure. If the θ_{ij} were known, then a natural thresholding estimator would be $(\hat{\sigma}^o_{ij})_{p\times p}$ with

$$\hat{\sigma}^o_{ij} = s_{\lambda^o_{ij}}(\hat{\sigma}_{ij}) \quad \text{with } \lambda^o_{ij} = 2\sqrt{\frac{\theta_{ij}\log p}{n}}, \tag{10}$$

where $s_\lambda(z)$ is a thresholding function. Compared with the universal thresholding rule of Bickel and Levina (2008), the variance factors θ_{ij} in the thresholds make the thresholding rule entry-dependent and lead to a more flexible estimator. In practice, θ_{ij} are typically unknown but can be well estimated. We propose the following estimator of θ_{ij}:

$$\hat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n}[(X_{ki}-\bar{X}^i)(X_{kj}-\bar{X}^j)-\hat{\sigma}_{ij}]^2, \qquad \bar{X}^i = n^{-1}\sum_{k=1}^{n}X_{ki}.$$

This leads to our adaptive thresholding estimator of the covariance matrix, $\mathbf{\Sigma}_0$,

$$\hat{\mathbf{\Sigma}}^\star(\delta) = (\hat{\sigma}^\star_{ij})_{p\times p} \quad \text{with } \hat{\sigma}^\star_{ij} = s_{\lambda_{ij}}(\hat{\sigma}_{ij}), \tag{11}$$

where

$$\lambda_{ij} := \lambda_{ij}(\delta) = \delta\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}. \tag{12}$$

Here $\delta > 0$ is a regularization parameter that can be fixed at $\delta = 2$ or chosen through CV. Good choices of δ will not affect the rate of convergence, but will affect the numerical performance of the resulting estimators. Selection of δ is thus of practical importance, and is addressed in more detail later in the article.

The analogy between the sparse covariance estimation problem and the idealized mean estimation problem (9) gives good motivation for the adaptive thresholding estimator defined in (11) and (12). Of course the matrix estimation problem is not exactly equivalent to the mean estimation problem (9) because noise is not exactly normal or iid and the loss is the spectral norm, not a vector norm or the Frobenius norm. We provide a more precise technical analysis in Sections 3 and 7.

In this article we consider simultaneously a class of thresholding functions $s_\lambda(z)$ that satisfy the following conditions:

(i) $|s_\lambda(z)| \leq c|y|$ for all z, y satisfy $|z-y| \leq \lambda$ and some $c > 0$
(ii) $s_\lambda(z) = 0$ for $|z| \leq \lambda$
(iii) $|s_\lambda(z) - z| \leq \lambda$, for all $z \in \mathbb{R}$.

These three conditions are satisfied by, for example, the soft thresholding rule $s_\lambda(z) = \mathrm{sgn}(z)(z-\lambda)_+$ and the adaptive lasso rule $s_\lambda(z) = z(1-|\lambda/z|^\eta)_+$ with $\eta \geq 1$ (Rothman, Levina, and Zhu 2009). We present a unified analysis of the adaptive thresholding estimators with the thresholding function $s_\lambda(z)$ satisfying the foregoing three conditions. It should be noted that condition (i) excludes the hard thresholding rule; however, all of the theoretical results in this article hold for the hard thresholding estimator under similar conditions. Here condition (i) is provided only to make the technical analysis in Section 7 work in a unified way for the class of thresholding rules. The results for the hard thresholding rule require slightly different proofs.

Rothman, Levina, and Zhu (2009) proposed generalized universal thresholding estimators

$$\hat{\mathbf{\Sigma}}_g = (\hat{\sigma}^g_{ij})_{p\times p}, \quad \text{where } \hat{\sigma}^g_{ij} = \bar{s}_{\lambda_n}(\hat{\sigma}_{ij})$$

and $\bar{s}_\lambda(z)$ satisfies (ii), (iii), and $|\bar{s}_\lambda(z)| \leq |z|$, which is slightly weaker than (i). Antoniadis and Fan (2001) introduced and studied similar general universal thresholding rules in the context of wavelet function estimation. We note that the generalized universal thresholding estimators $\hat{\mathbf{\Sigma}}_g$ suffer the same shortcomings as those of $\hat{\mathbf{\Sigma}}_u$, and like $\hat{\mathbf{\Sigma}}_u$ they are suboptimal over the class $\mathcal{U}^\star_q$.

3. THEORETICAL PROPERTIES OF ADAPTIVE THRESHOLDING

Here we consider the asymptotic properties of the adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^\star(\delta)$ defined in (11) and (12). We show that the estimator $\hat{\mathbf{\Sigma}}^\star(\delta)$ adaptively attains the optimal rate of convergence over the collection of parameter spaces $\mathcal{U}^\star_q(s_0(p))$.

We begin with some notation. Define the standardized variables

$$Y_i = (X_i - \mu_i)/(\mathsf{Var}(X_i))^{1/2},$$

where $\mu_i = \mathsf{E}X_i$, and let $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$. Throughout, write $|\mathbf{a}|_2 = \sqrt{\sum_{j=1}^p a_j^2}$ for the usual Euclidean norm of a vector $\mathbf{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, define the spectral norm $\|\mathbf{A}\|_2 = \sup_{|\mathbf{x}|_2\leq 1}|\mathbf{A}\mathbf{x}|_2$, the matrix ℓ_1 norm $\|\mathbf{A}\|_{L_1} = \max_{1\leq j\leq q}\sum_{i=1}^p |a_{i,j}|$, and the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ holds for all sufficiently large n, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

3.1 Rate of Convergence

In the covariance matrix estimation literature, it is conventional to divide the technical analysis into two cases according the the moment conditions on $\mathbf{X}$.

(C1) (*Exponential-type tails*) Suppose that $\log p = o(n^{1/3})$ and there exists some $\eta > 0$ such that

$$\mathsf{E}\exp(tY_i^2) \le K_1 < \infty \quad \text{for all } |t| \le \eta \text{ and } i. \tag{13}$$

Furthermore, assume that for some $\tau_0 > 0$,

$$\min_{ij} \mathsf{Var}(Y_iY_j) \ge \tau_0. \tag{14}$$

(C2) (*Polynomial-type tails*) Suppose that for some $\gamma, c_1 > 0$, $p \le c_1 n^\gamma$, and for some $\epsilon > 0$

$$\mathsf{E}|Y_i|^{4\gamma+4+\epsilon} \le K_1 \quad \text{for all } i. \tag{15}$$

Furthermore, assume that (14) holds.

Remark 1. Note that (C1) holds with $\tau_0 = 1$ in the Gaussian case where $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$. To this end, let ρ_{ij} be the correlation coefficient of Y_i and Y_j. We can then write $Y_i = \rho_{ij}Y_j + \sqrt{1-\rho_{ij}^2}Y$, where $Y \sim \mathrm{N}(0,1)$ is independent of Y_j. Thus we have $\mathsf{Var}(Y_iY_j) = 1 + \rho_{ij}^2 \ge 1$, and (14) holds with $\tau_0 = 1$.

The following theorem gives the rate of convergence over the parameter space $\mathcal{U}_q^\star$ under the spectral norm for the thresholding estimator $\hat{\boldsymbol{\Sigma}}^\star(\delta)$.

Theorem 1. Let $\delta \ge 2$ and $0 \le q < 1$.

(i) Under (C1), we have, for some constant $C_{K_1,\delta,c,q}$ depending only on δ, c, q and K_1,

$$\begin{aligned}\inf_{\boldsymbol{\Sigma}_0 \in \mathcal{U}_q^\star} \mathsf{P}\Bigg(&\|\hat{\boldsymbol{\Sigma}}^\star(\delta) - \boldsymbol{\Sigma}_0\|_2 \\ &\le C_{K_1,\delta,c,q} s_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\Bigg) \\ &\ge 1 - O\big((\log p)^{-1/2}p^{-\delta+2}\big).\end{aligned} \tag{16}$$

(ii) Under (C2), (16) holds with probability greater than $1 - O((\log p)^{-1/2}p^{-\delta+2} + n^{-\epsilon/8})$.

Although $\mathcal{U}_q^\star$ is larger than the uniformity class $\mathcal{U}_q$, the rates of convergence of $\hat{\boldsymbol{\Sigma}}^\star(\delta)$ over the two classes are of the same order, $s_0(p)(\log p/n)^{(1-q)/2}$.

Theorem 1 states the rate of convergence in terms of probability. The same rate of convergence holds in expectation with some additional mild assumptions. By (16) and some long but elementary calculations (see also the proof of Lemma 4), we have the following result on the mean squared spectral norm:

Proposition 1. Under (C1) and $p \ge n^\xi$ for some $\xi > 0$, we have for $\delta \ge 7 + \xi^{-1}$, $0 \le q < 1$, and some constant $C > 0$,

$$\sup_{\boldsymbol{\Sigma}_0 \in \mathcal{U}_q^\star} \mathsf{E}\|\hat{\boldsymbol{\Sigma}}^\star(\delta) - \boldsymbol{\Sigma}_0\|_2^2 \le C s_0^2(p)\left(\frac{\log p}{n}\right)^{1-q}. \tag{17}$$

Remark 2. Cai and Zhou (2010) established the minimax rates of convergence under the spectral norm for sparse covariance matrix estimation over $\mathcal{U}_q$. They showed that the optimal rate over $\mathcal{U}_q$ is $s_0(p)(\log p/n)^{(1-q)/2}$. Because $\mathcal{U}_q(s_0(p)) \subseteq \mathcal{U}_q^\star(K^{1-q}s_0(p))$, this immediately implies that the convergence rate attained by the adaptive thresholding estimator over $\mathcal{U}_q^\star$ in Theorem 1 and (17) is optimal.

Remark 3. The estimator $\hat{\boldsymbol{\Sigma}}^\star(\delta)$ immediately yields an estimate of the correlation matrix $\mathbf{R}_0 = (r_{ij})_{1\le i,j,\le p}$ which is the object of direct interest in some statistical applications. Denote the corresponding estimator of $\mathbf{R}_0$ by $\hat{\mathbf{R}}^\star(\delta) = (\hat{r}_{ij}^\star)_{1\le i,j,\le p}$, with $\hat{r}_{ij}^\star = \hat{\sigma}_{ij}^\star/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$. A parameter space for the correlation matrices is the following ℓ_q ball:

$$\begin{aligned}\mathcal{R}_q^\star &:= \mathcal{R}_q^\star(s_0(p)) \\ &= \left\{\mathbf{R} : \mathbf{R} \succ 0,\ \max_i \sum_{j=1}^p |r_{ij}|^q \le s_0(p)\right\}.\end{aligned} \tag{18}$$

Then Theorem 1 holds for estimating the correlation matrix $\mathbf{R}_0$ by replacing $\hat{\boldsymbol{\Sigma}}^\star(\delta)$, $\boldsymbol{\Sigma}_0$, and $\mathcal{U}_q^\star$ with $\hat{\mathbf{R}}^\star(\delta)$, $\mathbf{R}_0$, and $\mathcal{R}_q^\star$, respectively.

Note that the covariance matrix $\boldsymbol{\Sigma}_0$ can be written as $\boldsymbol{\Sigma}_0 = \mathbf{D}^{1/2}\mathbf{R}_0\mathbf{D}^{1/2}$, where $\mathbf{D} = \mathrm{diag}(\boldsymbol{\Sigma}_0)$. Thus the covariance matrix can be viewed as a weighted version of the correlation matrix with weights $\{(\sigma_{ii}\sigma_{jj})^{1/2}\}$. Correspondingly, the parameter space $\mathcal{U}_q^\star$ in (7) can be viewed as the weighted version of $\mathcal{R}_q^\star$ given in (18) with the same weights,

$$\mathcal{U}_q^\star := \left\{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} \succ 0,\ \max_i \sum_{j=1}^p (\sigma_{ii}\sigma_{jj})^{1/2}|r_{ij}|^q \le s_0(p)\right\}.$$

3.2 Support Recovery

A closely related problem to estimating a sparse covariance matrix under spectral norm is recovering the support of the covariance matrix. This problem has been considered by, for example, Rothman, Levina, and Zhu (2009). For support recovery, it is natural to consider the parameter space

$$\bar{\mathcal{U}}_0 := \bar{\mathcal{U}}_0(s_0(p)) = \left\{\boldsymbol{\Sigma} : \max_i \sum_{j=1}^p I\{\sigma_{ij} \ne 0\} \le s_0(p)\right\},$$

which assumes that the covariance matrix has at most $s_0(p)$ nonzero entries on each row.

Define the support of $\boldsymbol{\Sigma}_0 = (\sigma_{ij}^0)$ by $\Psi = \{(i,j) : \sigma_{ij}^0 \ne 0\}$. The following theorem shows that the adaptive thresholding estimator $\hat{\boldsymbol{\Sigma}}^\star(\delta)$ recovers the support Ψ exactly with high probability when the magnitudes of nonzero entries rises above a certain threshold.

Theorem 2. Suppose that $\boldsymbol{\Sigma}_0 \in \bar{\mathcal{U}}_0$. Let $\delta \ge 2$ and

$$|\sigma_{ij}^0| > (2 + \delta + \gamma)\sqrt{\frac{\theta_{ij}\log p}{n}} \quad \text{for all } (i,j) \in \Psi \text{ and some } \gamma > 0. \tag{19}$$

If either (C1) or (C2) holds, then we have

$$\inf_{\boldsymbol{\Sigma}_0 \in \bar{\mathcal{U}}_0} \mathsf{P}\big(\mathrm{supp}(\hat{\boldsymbol{\Sigma}}^\star(\delta)) = \mathrm{supp}(\boldsymbol{\Sigma}_0)\big) \to 1.$$

A similar support recovery result was established for the generalized universal thresholding estimator by Rothman, Levina, and Zhu (2009) under the condition $\max_i \sigma_{ii}^0 \le K$ and a lower bound condition similar to (19). Note that in Theorem 2, we do not require $\max_i \sigma_{ii}^0 \le K$.

Following Rothman, Levina, and Zhu (2009), we can evaluate the ability to recover the support via the true positive rate

(TPR) in combination with the false positive rate (FPR), defined respectively as

$$\text{TPR} = \frac{\#\{(i,j): \hat{\sigma}^{\star}_{ij} \neq 0 \text{ and } \sigma_{ij} \neq 0\}}{\#\{(i,j): \sigma_{ij} \neq 0\}} \quad \text{and}$$

$$\text{FPR} = \frac{\#\{(i,j): \hat{\sigma}^{\star}_{ij} \neq 0 \text{ and } \sigma_{ij} = 0\}}{\#\{(i,j): \sigma_{ij} = 0\}}.$$

It directly follows from Theorem 2 that $\mathsf{P}(\text{FPR} = 0) \to 1$ and $\mathsf{P}(\text{TPR} = 1) \to 1$ under the conditions of the theorem.

The next result shows that $\delta = 2$ is the optimal choice for support recovery in the sense that a thresholding estimator with any smaller choice of δ would fail to recover the support of $\mathbf{\Sigma}_0$ exactly with probability going to 1. We assume that $\mathbf{X}$ satisfies the following condition, which is weaker than the Gaussian assumption:

(C3) Suppose that

$$\mathsf{E}[(X_i - \mu_i)^2 (X_j - \mu_j)(X_k - \mu_k)] = 0,$$

$$\mathsf{E}[(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l)] = 0$$

if $\sigma^0_{j_1 j_2} = 0$ for all $j_1 \neq j_2 \in \{i, j, k, l\}$.

Theorem 3. Let $\lambda_{ij} = \tau\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}$ with $0 < \tau < 2$. Suppose that (C1) or (C2) holds. Under (C3) and $p = \exp(o(n^{1/5}))$, if $s_0(p) = O(p^{1-\tau_1})$ with some $\tau^2/4 < \tau_1 < 1$ and $p \to \infty$, then

$$\inf_{\mathbf{\Sigma}_0 \in \mathcal{U}_0} \mathsf{P}\big(\text{supp}(\hat{\mathbf{\Sigma}}^{\star}(\tau)) \neq \text{supp}(\mathbf{\Sigma}_0)\big) \to 1.$$

Remark 4. The condition $p = \exp(o(n^{1/5}))$ is used in the proof to ensure that the covariances of the samples $\{\mathbf{X}_n\}$ can be well approximated by normal vectors. It can be replaced by $p = \exp(o(n^{1/3}))$ if $\mathbf{X}$ is a multivariate normal population.

3.3 Comparison With Universal Thresholding

It is interesting to compare the asymptotic results for adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^{\star}(\delta)$ with the known results for universal thresholding estimators. We begin by comparing the rate of convergence of $\hat{\mathbf{\Sigma}}^{\star}(\delta)$ with that of the universal thresholding estimator $\hat{\mathbf{\Sigma}}_u$ introduced by Bickel and Levina (2008) in the case of polynomial-type tails. Suppose that (C2) holds. Bickel and Levina (2008) showed that

$$\|\hat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_0\|_2 = O_{\mathsf{P}}\left(s_0(p)\left(\frac{p^{1/(1+\gamma+\epsilon/2)}}{n^{1/2}}\right)^{1-q}\right) \tag{20}$$

for $\mathbf{\Sigma}_0 \in \mathcal{U}_q$. Clearly, the convergence rate given in Theorem 1 for the adaptive thresholding estimator is significantly faster than that in (20).

We next compare the rates over the class $\mathcal{U}^{\star}_q$, $0 \le q < 1$. For brevity, we focus on the Gaussian case $\mathbf{X} \sim \text{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_0)$. The following theorem gives the lower bound of the universal thresholding estimator.

Theorem 4. Assume that $n^{5q} \le p \le \exp(o(n^{1/3}))$ and $8 \le s_0(p) < \min\{p^{1/4}, 4(n/\log p)^{1/2}\}$. As $p \to \infty$, we have

$$\inf_{\lambda_n} \sup_{\mathbf{\Sigma}_0 \in \mathcal{U}^{\star}_q} \mathsf{P}\left(\|\hat{\mathbf{\Sigma}}_g - \mathbf{\Sigma}_0\|_2 > \frac{3}{64} s_0^{2-q}(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right) \to 1, \tag{21}$$

and thus, for large n,

$$\inf_{\lambda_n} \sup_{\mathbf{\Sigma}_0 \in \mathcal{U}^{\star}_q} \mathsf{E}\|\hat{\mathbf{\Sigma}}_g - \mathbf{\Sigma}_0\|_2^2 \ge \frac{1}{512} s_0^{4-2q}(p)\left(\frac{\log p}{n}\right)^{1-q}. \tag{22}$$

The rate in (21) is slower than the optimal rate $s_0(p) \times (\log p/n)^{(1-q)/2}$ given in (16) when $s_0(p) \to \infty$ as $p \to \infty$. Therefore, no universal thresholding estimators can be minimax-rate optimal under the spectral norm over $\mathcal{U}^{\star}_q$ if $s_0(p) \to \infty$.

If we assume the mean of $\mathbf{X}$ is 0 and ignore the term $\bar{\mathbf{X}}$ in $\mathbf{\Sigma}_n$, then the universal thresholding estimators given by Bickel and Levina (2008) and Rothman, Levina, and Zhu (2009) use the sample mean of the samples $\{X_{ki}X_{kj}; 1 \le k \le n\}$ to identify zero entries in the covariance matrix. The support of these estimators depends on the quantities $I\{|\hat{\sigma}_{ij}| \ge \lambda_n\}$. In the high-dimensional setting, the sample mean is usually unstable for non-Gaussian distributions with heavier tails. Non-Gaussian data can often arise from many practical applications such as in finance and genomics. For our estimator, instead of the sample mean, we use the Student t statistic $\hat{\sigma}_{ij}/\hat{\theta}_{ij}^{1/2}$ to distinguish zero and nonzero entries. Our support recovery depends on the quantities $I\{|\hat{\sigma}_{ij}|/\hat{\theta}_{ij}^{1/2} \ge 2\sqrt{\log p/n}\}$, which are more stable than $I\{|\hat{\sigma}_{ij}| \ge \lambda_n\}$, because the t statistic is much more stable than the sample mean (see Shao 1999 for the theoretical justification).

4. DATA–DRIVEN CHOICE OF δ

Section 3 analyzes the properties of the adaptive thresholding estimator with a fixed value of δ. Alternatively, δ can be selected empirically through CV. In the work of Bickel and Levina (2008), the value of the universal thresholding level λ_n was not fully specified, and the CV method was used to select λ_n empirically. The authors obtained the convergence rate under the Frobenius norm for an estimator based only on partial samples. Theoretical analysis of the rate of convergence under the spectral norm was lacking. In this section, we first briefly describe the CV method for choosing δ and then derive the theoretical properties of the resulting estimator under the spectral norm.

Divide the sample $\{\mathbf{X}_k; 1 \le k \le n\}$ into two subsamples at random. Let n_1 and $n_2 = n - n_1$ be the two sample sizes for the random split satisfying $n_1 \asymp n_2 \asymp n$, and let $\hat{\mathbf{\Sigma}}^v_1$ and $\hat{\mathbf{\Sigma}}^v_2$ be the two sample covariance matrices from the vth split, for $v = 1, \ldots, H$, where H is a fixed integer. Let $\hat{\mathbf{\Sigma}}^{\star v}_1(\delta)$ and $\hat{\mathbf{\Sigma}}^{\star v}_2(\delta)$ be defined as in (11) from the vth split and

$$\hat{R}(\delta) = \frac{1}{H}\sum_{v=1}^{H} \|\hat{\mathbf{\Sigma}}^{\star v}_1(\delta) - \hat{\mathbf{\Sigma}}^v_2\|_F^2.$$

Let $a_j = j/N$, $0 \le j \le 4N$ be $4N + 1$ points in $[0, 4]$ and take

$$\hat{\delta} = \hat{j}/N, \quad \text{where } \hat{j} = \arg\min_{0 \le j \le 4N} \hat{R}(j/N),$$

where $N > 0$ is a fixed integer. If several j's attain the minimum value, then $\hat{j}$ is chosen to be the smallest one. The final estimator of the covariance matrix $\mathbf{\Sigma}_0$ is given by $\hat{\mathbf{\Sigma}}^{\star}(\hat{\delta})$.

Theorem 5. Suppose that $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 \in \mathcal{U}_0$ and $\min_i \sigma_{ii}^0 \geq \tau_0$ for some $\tau_0 > 0$. Let $s_0(p) = O((\log p)^\gamma)$ for some $\gamma < 1$ and $n^\xi \leq p \leq \exp(o(n^{1/3}))$ for some $\xi > 0$. We then have

$$\inf_{\boldsymbol{\Sigma}_0 \in \mathcal{U}_0} \mathsf{P}\left(\|\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta}) - \boldsymbol{\Sigma}_0\|_2 \leq C s_0(p) \left(\frac{\log p}{n} \right)^{1/2} \right) \to 1.$$

Remark 5. The assumption that N is fixed is not a stringent condition, because we consider only δ belonging to the fixed interval $[0, 4]$. Moreover, we focus only on the matrices in $\mathcal{U}_0$, due to the complexity of the proof. Extending to the case $N \to \infty$ with certain rate and more general $\boldsymbol{\Sigma}_0$ is possible; however, this requires a far more complicated proof and is not considered further in this article.

Remark 6. The condition $s_0(p) = O((\log p)^\gamma)$ in the theorem is used purely for technical reasons, and we believe that it is not essentially needed and can be weakened. This condition is not stringent when $p = \exp(n^\alpha)$, and it becomes restrictive if $p = O(n^\alpha)$.

Similar to the fixed δ case, we also consider support recovery with the estimator $\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta})$.

Proposition 2. Suppose that the conditions in Theorem 5 hold. For $\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta})$, we have

$$\mathrm{FPR} = O_{\mathsf{P}}(s_0(p)/p) \to 0.$$

Moreover, because $\hat{\delta} \leq 4$, we have $\mathrm{TPR} = 1$ with probability tending to 1 if the lower bound in (19) holds, with $2+\delta$ replaced by 6.

5. NUMERICAL RESULTS

The adaptive thresholding procedure presented in Section 2 is easy to implement. In this section we study the numerical performance of the proposed adaptive thresholding estimator $\hat{\boldsymbol{\Sigma}}^\star(\delta)$ using Monte Carlo simulations. We consider both methods for choosing the regularization parameter δ and compare their performance with that of universal thresholding estimators. We illustrate the adaptive thresholding estimator in an analysis on a dataset from a small round blue-cell tumor microarray experiment.

5.1 Simulation

Two types of sparse covariance matrices are considered in the simulations to investigate the numerical properties of the adaptive thresholding estimator $\hat{\boldsymbol{\Sigma}}^\star(\delta)$:

- Model 1 (banded matrix with ordering). $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$, where $\mathbf{A}_1 = (\sigma_{ij})_{1 \leq i,j \leq p/2}$, $\sigma_{ij} = (1 - \frac{|i-j|}{10})_+$, $\mathbf{A}_2 = 4\mathbf{I}_{p/2 \times p/2}$. $\boldsymbol{\Sigma}_0$ is a two-block diagonal matrix, $\mathbf{A}_1$ is a banded and sparse covariance matrix, and $\mathbf{A}_2$ is a diagonal matrix with 4 along the diagonal.
- Model 2 (sparse matrix without ordering). $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2)$, where $\mathbf{A}_2 = 4\mathbf{I}_{p/2 \times p/2}$, $\mathbf{A}_1 = \mathbf{B} + \epsilon \mathbf{I}_{p/2 \times p/2}$, $\mathbf{B} = (b_{ij})_{p/2 \times p/2}$ with independent $b_{ij} = \mathrm{unif}(0.3, 0.8) \times \mathrm{Ber}(1, 0.2)$. Here $\mathrm{unif}(0.3, 0.8)$ is a random variable taking value uniformly in $[0.3, 0.8]$; $\mathrm{Ber}(1, 0.2)$ is a Bernoulli random variable that takes value 1 with probability 0.2 and value 0 with probability 0.8, and $\epsilon = \max(-\lambda_{\min}(\mathbf{B}), 0) + 0.01$ to ensure that $\mathbf{A}_1$ is positive definite.

Under each model, $n = 100$ iid p-variate random vectors are generated from the normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_0$, for $p = 30, 100, 200$. In each setting, 100 replications are used. We compare the numerical performance between the adaptive thresholding estimators $\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta})$ and $\hat{\boldsymbol{\Sigma}}_2^\star \equiv \hat{\boldsymbol{\Sigma}}^\star(2)$ and with the universal thresholding estimator $\hat{\boldsymbol{\Sigma}}_g$ of Rothman, Levina, and Zhu (2009). Here $\hat{\delta}$ is selected by fivefold CV in Section 4, and $\hat{\boldsymbol{\Sigma}}_2^\star$ is the adaptive thresholding estimator with fixed $\delta = 2$. The thresholding level λ_n in $\hat{\boldsymbol{\Sigma}}_g$ is selected by the fivefold CV method of Bickel and Levina (2008). For each procedure, we consider two types of thresholding functions, the hard thresholding and the adaptive lasso thresholding $s_\lambda(z) = x(1 - |\lambda/x|^\eta)$ with $\eta = 4$. The losses are measured by three matrix norms: the spectral norm, the matrix ℓ_1 norm, and the Frobenius norm. Tables 1 and 2 report the means and standard errors of these losses. We also carried out simulations with the SCAD thresholding function for both universal thresholding and adaptive thresholding. The phenomenon is very similar. The SCAD adaptive thresholding also outperforms the SCAD universal thresholding. For reasons of space, we do not report these results here.

Table 1. Comparison of average matrix losses for Model 1 over 100 replications. The standard errors are given in the parentheses

	Adaptive lasso			Hard		
p	$\hat{\boldsymbol{\Sigma}}_g$	$\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta})$	$\hat{\boldsymbol{\Sigma}}_2^\star$	$\hat{\boldsymbol{\Sigma}}_g$	$\hat{\boldsymbol{\Sigma}}^\star(\hat{\delta})$	$\hat{\boldsymbol{\Sigma}}_2^\star$
			Operator norm			
30	3.53 (0.13)	1.72 (0.05)	2.39 (0.07)	3.50 (0.14)	1.77 (0.05)	1.77 (0.04)
100	7.94 (0.11)	2.72 (0.05)	4.68 (0.06)	8.64 (0.07)	2.57 (0.05)	3.04 (0.05)
200	8.95 (0.004)	3.23 (0.05)	5.70 (0.05)	8.95 (0.004)	3.02 (0.05)	3.77 (0.05)
			Matrix ℓ_1 norm			
30	5.29 (0.15)	2.57 (0.08)	3.34 (0.09)	5.71 (0.15)	2.60 (0.09)	2.70 (0.06)
100	9.03 (0.05)	4.15 (0.07)	6.39 (0.09)	9.24 (0.03)	4.17 (0.07)	4.87 (0.09)
200	9.35 (0.01)	4.90 (0.07)	7.64 (0.07)	9.35 (0.01)	4.89 (0.07)	5.97 (0.09)
			Frobenius norm			
30	5.97 (0.10)	3.15 (0.05)	3.68 (0.05)	6.58 (0.09)	3.29 (0.05)	3.29 (0.04)
100	15.93 (0.12)	6.57 (0.05)	8.92 (0.06)	16.88 (0.03)	6.79 (0.06)	7.53 (0.05)
200	24.23 (0.01)	9.62 (0.05)	14.20 (0.07)	24.24 (0.01)	9.97 (0.06)	11.68 (0.05)

Table 2. Comparison of average matrix losses for Model 2 over 100 replications. The standard errors are given in the parentheses

	Adaptive lasso			Hard		
p	$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$	$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$
			Operator norm			
30	1.48 (0.02)	1.24 (0.03)	1.19 (0.03)	1.50 (0.02)	1.25 (0.03)	1.21 (0.03)
100	5.31 (0.01)	2.82 (0.05)	4.71 (0.03)	5.31 (0.01)	2.69 (0.05)	3.97 (0.04)
200	10.74 (0.01)	6.78 (0.08)	10.52 (0.02)	10.74 (0.01)	6.58 (0.10)	10.04 (0.03)
			Matrix ℓ_1 norm			
30	1.70 (0.03)	1.33 (0.04)	1.22 (0.03)	1.70 (0.02)	1.32 (0.04)	1.24 (0.03)
100	6.16 (0.01)	4.10 (0.05)	5.52 (0.03)	6.16 (0.01)	4.20 (0.06)	5.22 (0.03)
200	12.70 (0.01)	9.81 (0.08)	12.31 (0.04)	12.70 (0.01)	10.06 (0.08)	12.06 (0.04)
			Frobenius norm			
30	4.08 (0.03)	2.52 (0.04)	2.57 (0.04)	4.10 (0.03)	2.50 (0.04)	2.45 (0.04)
100	12.77 (0.01)	7.57 (0.05)	10.96 (0.04)	12.78 (0.02)	8.07 (0.06)	10.00 (0.05)
200	25.51 (0.01)	16.94 (0.07)	24.67 (0.03)	25.52 (0.01)	18.69 (0.07)	24.05 (0.03)

Under models 1 and 2, both adaptive thresholding estimators $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ and $\hat{\mathbf{\Sigma}}_2^\star$ uniformly outperform the universal thresholding rule $\hat{\mathbf{\Sigma}}_g$ significantly, regardless of the thresholding function or loss function used. Between $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ and $\hat{\mathbf{\Sigma}}_2^\star$, $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ performs better than $\hat{\mathbf{\Sigma}}_2^\star$ in general. Between the two thresholding functions, the hard thresholding rule outperforms the adaptive lasso thresholding rule for $\hat{\mathbf{\Sigma}}_2^\star$, whereas the difference is not significant for $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$. For both models, the hard and adaptive lasso universal thresholding rules behave very similarly. They both tend to "overthreshold" and remove many nonzero off-diagonal entries of the covariance matrices.

For support recovery, again both $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ and $\hat{\mathbf{\Sigma}}_2^\star$ outperform $\hat{\mathbf{\Sigma}}_g$. The values of TPR and FPR based on the off-diagonal entries are reported in Tables 3 and 4. For model 1, $\hat{\mathbf{\Sigma}}_g$ tends to estimate many nonzero off-diagonal entries by 0 when p is large. To better illustrate the recovery performance elementwise of the two models, heat maps of the nonzeros identified out of 100 replications when $p = 60$ are shown in Figures 1 and 2. These heat maps suggest that the sparsity patterns recovered by $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ and $\hat{\mathbf{\Sigma}}_2^\star$ have significantly closer resemblance to the true model than $\hat{\mathbf{\Sigma}}_g$.

5.2 Correlation Analysis on Real Data

We now apply the adaptive thresholding estimator $\hat{\mathbf{\Sigma}}^\star(\delta)$ to a dataset from a small round blue-cell tumor (SRBC) microarray experiment (Khan et al. 2001) and compare the ability of support recovery with that of the universal thresholding estimator $\hat{\mathbf{\Sigma}}_g$. We do not consider the estimator $\hat{\mathbf{\Sigma}}_2^\star$ here, because the simulation results in Section 5.1 show that $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ outperforms $\hat{\mathbf{\Sigma}}_2^\star$ when the sample size is not large. The SRBC dataset was analyzed by Rothman, Levina and Zhu (2009), who considered the universal thresholding rules. To make the results comparable, we follow the same steps as done by Rothman, Levina and Zhu (2009).

The SRBC dataset contains 63 training tissue samples, with 2308 gene expression values recorded for each sample. The original dataset had 6567 genes and was reduced to 2308 genes after an initial filtering (see Khan et al. 2001). The 63 tissue samples contain four types of tumors (23 EWS, 8 BL-NHL, 12 NB, and 20 RMS). As done by Rothman, Levina, and Zhu (2009), we first ranked the genes by the amount of discriminative information based on the F-statistic,

$$F = \frac{1}{k-1}\sum_{m=1}^{k} n_m(\bar{x}_m - \bar{x})^2 \Big/ \left(\frac{1}{n-k}\sum_{m=1}^{k}(n_m-1)\hat{\sigma}_m^2\right),$$

where $n = 63$ is the sample size, $k = 4$ is the number of classes, n_m, $1 \le m \le 4$ are the sample sizes of the four types of tumors, $\bar{x}_m$ and $\hat{\sigma}_m$ are the sample mean and sample variance of the class m, and $\bar{x}$ is the overall sample mean. Based on the F values, we chose the top 40 and bottom 160 genes. We also ordered

Table 3. Comparison of support recovery for Model 1 over 100 replications

		Adaptive lasso			Hard		
p		$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$	$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$
30	TPR	0.57	0.84	0.72	0.46	0.79	0.72
	FPR	0.07	0.01	0.00	0.05	0.003	0.00
100	TPR	0.15	0.76	0.57	0.01	0.69	0.57
	FPR	0.01	0.01	0.00	0.00	0.00	0.00
200	TPR	0.00	0.73	0.51	0.00	0.65	0.51
	FPR	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. Comparison of support recovery for Model 2 over 100 replications

		Adaptive lasso			Hard		
p		$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$	$\hat{\mathbf{\Sigma}}_g$	$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$	$\hat{\mathbf{\Sigma}}_2^\star$
30	TPR	0.02	0.95	0.88	0.00	0.91	0.88
	FPR	0.00	0.01	0.00	0.00	0.00	0.00
100	TPR	0.00	0.80	0.33	0.00	0.66	0.33
	FPR	0.00	0.01	0.00	0.00	0.00	0.00
200	TPR	0.00	0.68	0.09	0.00	0.49	0.09
	FPR	0.00	0.01	0.00	0.00	0.00	0.00

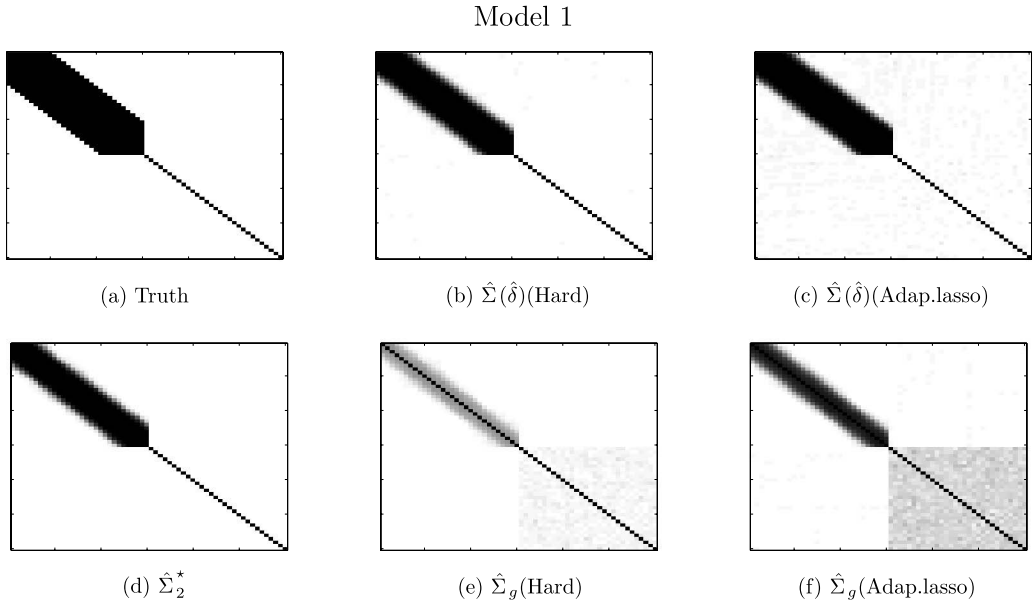


Figure 1. Heat maps of the frequency of the 0s identified for each entry of the covariance matrix (when $p = 60$) out of 100 replications. White indicates 100 0s identified out of 100 runs; black, 0/100.

the first 40 genes according to the ordering of Rothman, Levina, and Zhu (2009). Based on the 200 genes, we considered the performance of the two estimators $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ and $\hat{\mathbf{\Sigma}}_g$. We selected the tuning parameters $\hat{\delta}$ and λ_n by fivefold CV. To this end, we needed to divide the 63 samples into five groups of nearly equal size. Because there are four types of tumors in the samples, we let the proportions of the four types of tumors in each group be nearly equal, so that each fold was a good representative of the whole. We also used threefold CV in this way and obtained similar results.

Figure 3 plots the heat maps of $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ with hard thresholding [$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ Hard], $\hat{\mathbf{\Sigma}}_g$ with hard thresholding ($\hat{\mathbf{\Sigma}}_g$ Hard), $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ with adaptive lasso thresholding [$\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ AL], $\hat{\mathbf{\Sigma}}_g$ with adaptive lasso thresholding ($\hat{\mathbf{\Sigma}}_g$ AL). $\hat{\mathbf{\Sigma}}_g$ AL and $\hat{\mathbf{\Sigma}}_g$ Hard result in very sparse estimators, with 97.88% zero elements in off diagonal positions. The estimator $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ AL is the least sparse with 69.78% 0s, whereas $\hat{\mathbf{\Sigma}}^\star(\hat{\delta})$ hard has 83.11% 0s. The over-thresholding phenomenon in the real data analysis is consistent with that observed in the simulations. The universal thresholding rule removes many nonzero off diagonal entries and results in an oversparse estimate, whereas adaptive thresholding with different individual levels results in a clean but more informative estimate of the sparsity structure.

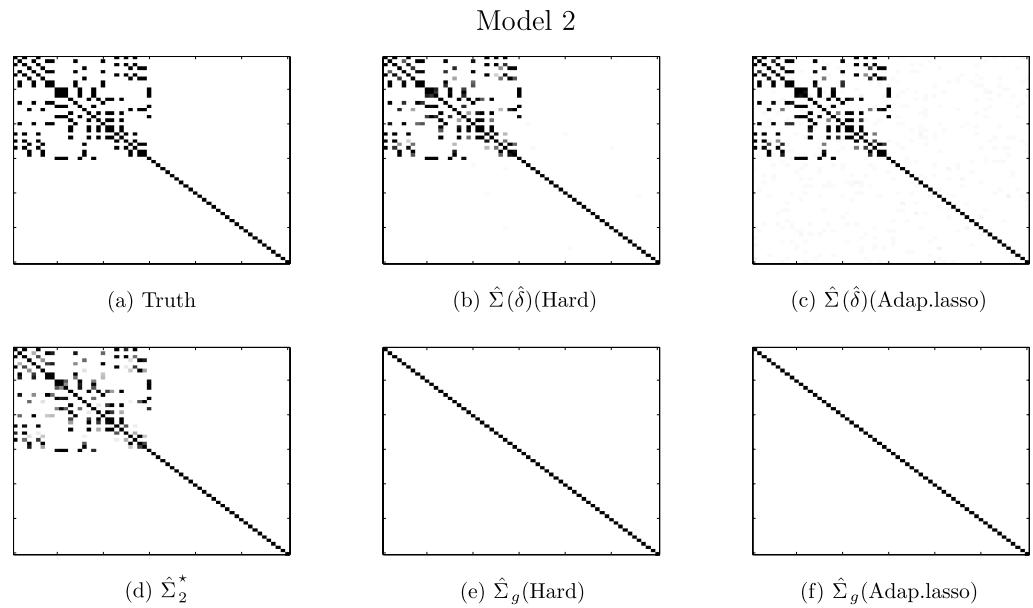
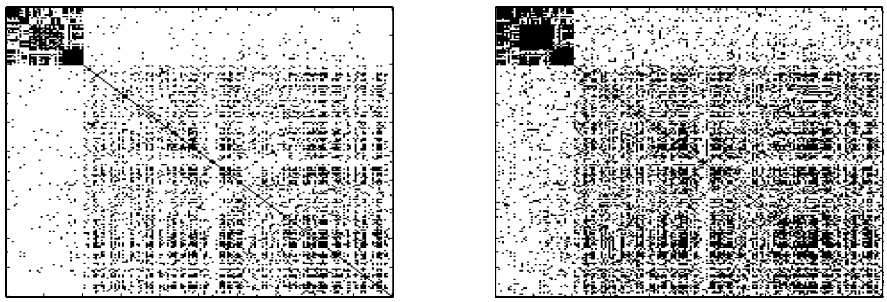
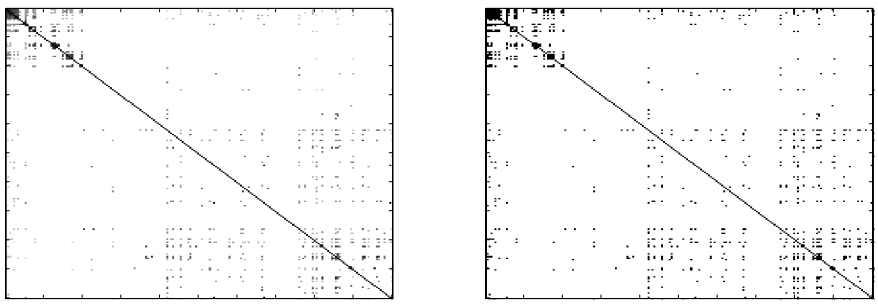


Figure 2. Heat maps of the frequency of the 0s identified for each entry of the covariance matrix (when $p = 60$) out of 100 replications. White indicates 100 0s identified out of 100 runs; black, 0/100.



(a) $\hat{\Sigma}(\hat{\delta})$ Hard (83.11% zeros) (b) $\hat{\Sigma}(\hat{\delta})$ AL (69.78% zeros)



(c) $\hat{\Sigma}_g$ Hard (97.88% zeros) (d) $\hat{\Sigma}_g$ AL (97.88% zeros)

Figure 3. Heatmaps of the estimated supports.

6. DISCUSSION

This article introduces an adaptive entry-dependent thresholding procedure for estimating sparse covariance matrices. Our proposed estimator $\hat{\mathbf{\Sigma}}^{\star}(\delta) = (\hat{\sigma}^{\star}_{ij})$ demonstrates excellent performance both theoretically and numerically. In particular, $\hat{\mathbf{\Sigma}}^{\star}(\delta)$ attains the optimal rate of convergence over $\mathcal{U}_q^{\star}$ given in (7), whereas universal thresholding estimators are suboptimal. The main reason that universal thresholding does not perform well is that the sample covariances can have a wide range of variability. A simple and natural way to deal with this heteroscedasticity is to first estimate the correlation matrix $\mathbf{R}_0$ and then renormalize by the sample variances to obtain an estimate of the covariance matrix. Here we discuss two approaches based on this idea.

Denote the sample correlation matrix by $\hat{\mathbf{R}} = (\hat{r}_{ij})_{1\le i,j\le p}$ with $\hat{r}_{i,j} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$. An estimate of the correlation matrix $\mathbf{R}_0$ can be obtained by thresholding $\hat{r}_{ij}$. Define the universal thresholding estimator of the correlation matrix by $\hat{\mathbf{R}}(\lambda_n) = (\hat{r}^{thr}_{ij})_{p\times p}$ with

$$\hat{r}^{thr}_{ij} = \hat{r}_{ij} I\{|\hat{r}_{ij}| \ge \lambda_n\}$$

and the corresponding estimator of the covariance matrix by $\hat{\mathbf{\Sigma}}_R = \mathbf{D}_n^{1/2}\hat{\mathbf{R}}(\lambda_n)\mathbf{D}_n^{1/2}$, where $\mathbf{D}_n = \operatorname{diag}(\mathbf{\Sigma}_n)$. It is easily seen that a good choice of the threshold λ_n is $\lambda_n = C\sqrt{(\log p)/n}$ for some constant $C > 0$. Choosing C is difficult, however, because the choice depends on the unknown underlying distribution. Assuming that the constant C is chosen sufficiently large, it can be shown that the resulting estimator $\hat{\mathbf{\Sigma}}_R$ attains the same minimax rate of convergence. However, the estimator $\hat{\mathbf{\Sigma}}_R$ is less efficient than $\hat{\mathbf{\Sigma}}^{\star}(\delta)$ for support recovery. In fact, $\hat{\mathbf{\Sigma}}_R$ is unable to recover the support of $\mathbf{\Sigma}_0$ exactly for a class of non-Gaussian distributions of $\mathbf{X}$. Denote by $\mathcal{V}(\gamma, \delta, K_1)$ the class of distributions F of $\mathbf{X}$ satisfying the conditions of Theorem 2. Then it can be shown that for any $\gamma > 0$, $\delta \ge 2$, and some $K_1 = K_1(\gamma) > 0$,

$$\inf_{\lambda_n} \sup_{F\in\mathcal{V}(\gamma,\delta,K_1)} \mathsf{P}(\operatorname{supp}(\hat{\mathbf{\Sigma}}_R) \ne \operatorname{supp}(\mathbf{\Sigma}_0)) \to 1. \tag{23}$$

The sample correlation coefficients $\hat{r}_{ij}$ are not homoscedastic, although its range of variability is smaller than that of sample covariances. This in fact is the main reason for the negative result on support recovery given in Equation (23). A natural approach to dealing with the heteroscedasticity of the sample correlation coefficients is to first stabilize the variance by using Fisher's z-transformation, then threshold, and finally obtain the estimator by inverse transformation. Applying Fisher's z-transformation to each correlation coefficient yields

$$\hat{Z}_{ij} = \frac{1}{2}\ln\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}.$$

When $\mathbf{X}$ is multivariate normal, it is well known that $\hat{Z}_{ij}$ is asymptotically normal with mean $(1/2)\ln((1 + r_{ij})/(1 - r_{ij}))$ and variance $1/(n-3)$. The behavior of $\hat{Z}_{ij}$ in the non-Gaussian case is more complicated. In general, the asymptotic variance of $\hat{Z}_{ij}$ depends on $\mathsf{E}X_i^2X_j^2$ even when $r_{ij} = 0$ (see Hawkins 1989). Similar to the method of thresholding the sample correlation coefficients discussed earlier, universally thresholding $(\hat{Z}_{ij})_{p\times p}$ is unable to recover the support of $\mathbf{\Sigma}_0$ exactly for a class of non-Gaussian distributions of $\mathbf{X}$ satisfying the conditions in Theorem 2.

In conclusion, the two natural approaches based on the sample correlation matrix discussed above are not as efficient as the entry-dependent thresholding method that we proposed in Section 2. For reasons of space, we omit the proofs of the results stated in this section. We will explore these issues in detail elsewhere.

7. PROOFS

We begin by stating a few technical lemmas that are essential for the proofs of the main results. The first lemma is an exponential inequality on the partial sums of independent random variables.

Lemma 1. Let $\xi_1, \ldots, \xi_n$ be independent random variables with mean 0. Suppose that there exists some $\eta > 0$ and $\bar{B}_n$ such that $\sum_{k=1}^n \mathsf{E}\xi_k^2 e^{\eta|\xi_k|} \le \bar{B}_n^2$. Then for $0 < x \le \bar{B}_n$,

$$\mathsf{P}\left(\sum_{k=1}^n \xi_k \ge C_\eta \bar{B}_n x\right) \le \exp(-x^2), \tag{24}$$

where $C_\eta = \eta + \eta^{-1}$.

Proof. By the inequality $|e^s - 1 - s| \le s^2 e^{s\max(s,0)}$, we have, for any $t \ge 0$,

$$\begin{aligned}\mathsf{P}\left(\sum_{k=1}^n \xi_k \ge C_\eta \bar{B}_n x\right) &\le \exp(-tC_\eta \bar{B}_n x) \prod_{k=1}^n \mathsf{E}\exp(t\xi_k) \\ &\le \exp(-tC_\eta \bar{B}_n x) \prod_{k=1}^n \bigl(1 + t^2 \mathsf{E}\xi_k^2 e^{t|\xi_k|}\bigr) \\ &\le \exp\left(-tC_\eta \bar{B}_n x + \sum_{k=1}^n t^2 \mathsf{E}\xi_k^2 e^{t|\xi_k|}\right).\end{aligned}$$

Take $t = \eta(x/\bar{B}_n)$. It follows that

$$\mathsf{P}\left(\sum_{k=1}^n \xi_k \ge C_\eta \bar{B}_n x\right) \le \exp(-\eta C_\eta x^2 + \eta^2 x^2) = \exp(-x^2),$$

which completes the proof.

The second and third lemmas are on the asymptotic behaviors of the largest entry of the sample covariance matrix and $\hat{\theta}_{ij}$. The proof of Lemma 2 is given in the supplementary material.

Lemma 2. (i) Under (C1), we have, for any $\delta \ge 2$, $\varepsilon > 0$, and $M > 0$,

$$\begin{aligned}&\mathsf{P}\Bigl(\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}^0| / \hat{\theta}_{ij}^{1/2} \ge \delta\sqrt{\log p/n}\Bigr) \\ &\quad = O\bigl((\log p)^{-1/2} p^{-\delta+2}\bigr),\end{aligned} \tag{25}$$

$$\mathsf{P}\Bigl(\max_{ij} |\hat{\theta}_{ij} - \theta_{ij}| / \sigma_{ii}^0 \sigma_{jj}^0 \ge \varepsilon\Bigr) = O(p^{-M}), \tag{26}$$

and

$$\mathsf{P}\Bigl(\max_i |\bar{X}^i| / (\sigma_{ii}^0)^{1/2} \ge C\sqrt{\log p/n}\Bigr) = O(p^{-M}) \tag{27}$$

for some $C > 0$.

(ii) Under (C2), (25)–(27) still hold if we replace $O((\log p)^{-1/2} p^{-\delta+2})$ and $O(p^{-M})$ with $O((\log p)^{-1/2} p^{-\delta+2} + n^{-\epsilon/8})$ and $O(n^{-\epsilon/8})$ respectively.

Lemma 3. Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a mean-0 random vector. Suppose that $\mathsf{Cov}(\mathbf{X}) = I_{p\times p}$, (C3) holds and $p \to \infty$. Then, under (C1) or (C2), we have, for any $\delta > 0$,

$$\mathsf{P}\left(\max_{1\le i<j\le p} (n\theta_{ij})^{-1} \left|\sum_{k=1}^n X_{ki}X_{kj}\right|^2 \ge (4-\delta)\log p\right) \to 1.$$

Proof. We arrange the two dimensional indices $\{(i,j) : 1 \le i < j \le p\}$ in any order and set them as $\{(i_m, j_m) : 1 \le m \le p(p-1)/2 =: L\}$. Let

$$Y_{km} = \theta_{ij}^{-1/2} X_{ki_m} X_{kj_m}, \qquad S_m = n^{-1/2} \sum_{k=1}^n Y_{km},$$

$$A_m = \{|S_m| \ge \sqrt{(4-\delta)\log p}\}, \qquad 1 \le m \le L.$$

Define $\bar{Y}_{km} = Y_{km} I\{|Y_{km}| \le \delta_n \sqrt{n/(\log p)^3}\}$ and $\hat{Y}_{km} = \bar{Y}_{km} - \mathsf{E}\bar{Y}_{km}$, where $\delta_n \to 0$ sufficiently slow. Then, by (C1) or (C2) when n is large, we have

$$\begin{aligned}&\mathsf{P}\left(\max_{1\le i<j\le p} (n\theta_{ij})^{-1} \left|\sum_{k=1}^n X_{ki}X_{kj}\right|^2 \ge (4-\delta)\log p\right) \\ &\quad \ge \mathsf{P}\left(\max_{1\le m\le L} n^{-1} \left|\sum_{k=1}^n \hat{Y}_{km}\right|^2 \ge (4-2\delta)\log p\right) \\ &\qquad - O\bigl(p^{-M} + n^{-\epsilon/8}\bigr) \\ &\quad \ge \mathsf{P}\left(\max_{1\le m\le L} n^{-1} \left|\sum_{k=1}^n \hat{Y}_{km}\right|^2 \ge 4\log p - \log\log p + x\right) \\ &\qquad - O\bigl(p^{-M} + n^{-\epsilon/8}\bigr)\end{aligned} \tag{28}$$

for any $M > 0$ and $x < 0$. Set $y_n = \sqrt{4\log p - \log\log p + x}$ and

$$\hat{A}_m = \left\{n^{-1/2}\left|\sum_{k=1}^n \hat{Y}_{km}\right| \ge y_n\right\}.$$

Then, by Bonferroni's inequality, for any fixed l, we have

$$\begin{aligned}&\mathsf{P}\left(\max_{1\le m\le L} n^{-1}\left|\sum_{k=1}^n \hat{Y}_{km}\right|^2 \ge y_n^2\right) \\ &\qquad \ge \sum_{d=1}^{2l} (-1)^{d-1} \sum_{1\le i_1<\cdots<i_d\le L} \mathsf{P}\left(\bigcap_{j=1}^d \hat{A}_{i_j}\right).\end{aligned} \tag{29}$$

Write

$$\hat{\mathbf{Y}}_k = \bigl(\hat{Y}_{ki_1}, \ldots, \hat{Y}_{ki_d}\bigr)^T, \qquad 1 \le k \le n.$$

By theorem 1 of Zaitsev (1987), we have

$$\begin{aligned}&\mathsf{P}\bigl(|\hat{\mathbf{N}}|_{d,\infty} \ge y_n - \delta_n^{1/2}(\log p)^{-1/2}\bigr) + c_1 \exp\bigl(-c_2\delta_n^{-1/2}\log p\bigr) \\ &\quad \ge \mathsf{P}\left(\left|n^{-1/2}\sum_{k=1}^n \hat{\mathbf{Y}}_k\right|_{d,\infty} \ge y_n\right) \\ &\quad \ge \mathsf{P}\bigl(|\hat{\mathbf{N}}|_{d,\infty} \ge y_n + \delta_n^{1/2}(\log p)^{-1/2}\bigr) \\ &\qquad - c_1 \exp\bigl(-c_2\delta_n^{-1/2}\log p\bigr),\end{aligned} \tag{30}$$

where c_1 and c_2 are positive constant depending only on d, $|\cdot|_{d,\infty}$ means $|\mathbf{a}|_{d,\infty} = \min_{1\le i\le d} |a_i|$ for $\mathbf{a} = (a_1, \ldots, a_d)^T$, and $\hat{\mathbf{N}}$ is a d-dimensional normal random vector with mean 0 and covariance matrix $\mathsf{Cov}(\hat{\mathbf{Y}}_k)$. Set

$$\hat{B}_{i_1,\ldots,i_d}^{\pm} = \bigl\{|\hat{\mathbf{N}}|_{d,\infty} \ge y_n \mp \delta_n^{1/2}(\log p)^{-1/2}\bigr\}.$$

We can check that $\|\operatorname{Cov}(\hat{\mathbf{N}}_k) - I_{d\times d}\|_2 = O(1/(\log p)^8)$. Let $\mathbf{Z}$ be a standard d-dimensional normal vector. Then we have

$$\begin{aligned}\mathsf{P}\big(\hat{B}^+_{i_1,\ldots,i_d}\big) &\le \mathsf{P}\big(|\mathbf{Z}|_{d,\infty} \ge y_n - 2\delta_n^{1/2}(\log p)^{-1/2}\big)\\ &\quad + \mathsf{P}\big(\|\operatorname{Cov}(\hat{\mathbf{N}}_k) - I_{d\times d}\|_2 |\mathbf{Z}|_2 \ge \delta_n^{1/2}(\log p)^{-1/2}\big)\\ &= (1+o(1))\bigg(\frac{1}{\sqrt{2\pi}}p^{-2}\exp(-x/2)\bigg)^d\\ &\quad + O\big(\exp(-C(\log p)^2)\big). \end{aligned} \tag{31}$$

Similarly, we can get

$$\begin{aligned}\mathsf{P}\big(\hat{B}^-_{i_1,\ldots,i_d}\big) &\ge (1-o(1))\bigg(\frac{1}{\sqrt{2\pi}}p^{-2}\exp(-x/2)\bigg)^d\\ &\quad - O\big(\exp(-C(\log p)^2)\big). \end{aligned} \tag{32}$$

Submitting (30)–(32) into (29), we can get

$$\begin{aligned}&\liminf_{n\to\infty}\mathsf{P}\bigg(\max_{1\le m\le L} n^{-1}\bigg|\sum_{k=1}^n \hat{Y}_{km}\bigg|^2 \ge y_n^2\bigg)\\ &\quad\ge \sum_{d=1}^{2l}(-1)^{d-1}\bigg(\frac{1}{\sqrt{8\pi}}\exp(-x/2)\bigg)^d\Big/ d!\\ &\quad\to 1-\exp\bigg(-\frac{1}{\sqrt{8\pi}}\exp(-x/2)\bigg)\end{aligned} \tag{33}$$

as $l\to\infty$. Letting $x\to-\infty$, we prove the lemma by (28) and (33).

Proof of Theorem 1. By (C1) or (C2), we have $\theta_{ij} \le C_{K_1}\sigma^0_{ii}\sigma^0_{jj}$. In the event that $\{|\hat{\sigma}_{ij} - \sigma^0_{ij}| \le \lambda_{ij} \text{ for all } i,j\} \cap \{\hat{\theta}_{ij} \le 2\theta_{ij} \text{ for all } i,j\}$, we have, by conditions (i)–(iii) on $s_\lambda(z)$, that

$$\begin{aligned}&\sum_{j=1}^p \big|s_{\lambda_{ij}}(\hat{\sigma}_{ij}) - \sigma^0_{ij}\big|\\ &\quad= \sum_{j=1}^p \big|s_{\lambda_{ij}}(\hat{\sigma}_{ij}) - \sigma^0_{ij}\big| I\{|\hat{\sigma}_{ij}| \ge \lambda_{ij}\} + \sum_{j=1}^p |\sigma^0_{ij}| I\{|\hat{\sigma}_{ij}| < \lambda_{ij}\}\\ &\quad\le 2\sum_{j=1}^p \lambda_{ij} I\{|\sigma^0_{ij}| \ge \lambda_{ij}\}\\ &\qquad+ \sum_{j=1}^p |s_{\lambda_{ij}}(\hat{\sigma}_{ij}) - \sigma^0_{ij}| I\{|\hat{\sigma}_{ij}| \ge \lambda_{ij}, |\sigma^0_{ij}| < \lambda_{ij}\}\\ &\qquad+ \sum_{j=1}^p |\sigma^0_{ij}| I\{|\sigma^0_{ij}| < 2\lambda_{ij}\}\\ &\quad\le 2\sum_{j=1}^p \lambda_{ij}^{1-q}|\sigma^0_{ij}|^q + (1+c)\sum_{j=1}^p |\sigma^0_{ij}| I\{|\sigma^0_{ij}| < \lambda_{ij}\}\\ &\qquad+ \sum_{j=1}^p |\sigma^0_{ij}| I\{|\sigma^0_{ij}| < 2\lambda_{ij}\}\\ &\quad\le C_{q,c}\sum_{j=1}^p \lambda_{ij}^{1-q}|\sigma^0_{ij}|^q\\ &\quad\le C_{K_1,\delta,c,q} s_0(p)\bigg(\frac{\log p}{n}\bigg)^{(1-q)/2}.\end{aligned}$$

The proof follows from Lemma 2 and the fact $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_{L_1}$ for any symmetric matrix.

Proof of Theorems 2 and 3. Theorem 2 follows immediately from Lemma 2. We now prove Theorem 3. For each $1 \le i \le p$, let A_1 be the largest subset of $\{1,\ldots,p\}$ such that X_i is uncorrelated with $\{X_k, k \in A_1\}$. Let $i_1 = \arg\min\{|j-i| : j \in A_1\}$. We then have $|i_1 - i| \le s$. Also, $\operatorname{Card}(A_1) \ge p - s$. Similarly, let A_l be the largest subset of A_{l-1} such that $X_{i_{l-1}}$ is uncorrelated with $\{X_k, k \in A_l\}$ and $i_l = \arg\min\{|j - i_{l-1}| : j \in A_l\}$. We can see that $|i_l - i| \le ls$ and $\operatorname{Card}(A_l) \ge \operatorname{Card}(A_{l-1}) - s \ge p - sl$. Take $l = [p^{\tau_2}]$ with $\tau^2/4 < \tau_2 < \min(\tau^2/3, \tau_1)$. Then $X_{i_0},\ldots,X_{i_l}$ are pairwise uncorrelated random variables, where we set $i_0 = i$. Clearly, $i_1,\ldots,i_l \in B_i = \{j : \sigma^0_{ij} = 0; j \ne i\}$. Without loss of generality, we assume that $X_1,\ldots,X_l$ are pairwise uncorrelated. Note that $|s_\lambda(z)| \ge |z| - \lambda$. It suffices to show that for some $\varepsilon_0 > 0$,

$$\mathsf{P}\bigg(\max_{1\le i<j\le l}\{\lambda_{nij}^{-1}|\hat{\sigma}_{ij}|\} > 1+\varepsilon_0\bigg) \to 1. \tag{34}$$

Clearly, we can assume $\mathsf{E}\mathbf{X} = 0$ and $\operatorname{Var}(X_i) = 1$ for $1 \le i \le l$. By Lemma 2 and (14), we have $\min_{ij}\lambda_{nij} > 0$ with probability tending to 1. By Lemma 2 it suffices to show that for any $0 < \tau < 2$,

$$\begin{aligned}A_n &:= \mathsf{P}\bigg(\max_{1\le i<j\le l}\bigg\{(n\theta_{ij})^{-1/2}\bigg|\sum_{k=1}^n X_{ki}X_{kj}\bigg|\bigg\} \ge \tau\sqrt{\log p}\bigg)\\ &\quad\to 1.\end{aligned} \tag{35}$$

Because $\tau^2\log p \le (4-\delta)\log l$ for $0 < \delta < 4 - \tau^2/\tau_2$ and large n, (35) follows from Lemma 3.

Lemmas 4 and 5, proved in the supplementary material, are needed to prove Theorems 4 and 5.

Lemma 4. Suppose that $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 \in \bar{\mathcal{U}}_0$. Let $s_0(p) = O((\log p)^\gamma)$ for some $\gamma < 1$ and $n^\xi \le p \le \exp(o(n^{1/3}))$ for some $\xi > 0$. Let $\delta > \sqrt{2}$. Then there are at most $O(s_0(p))$ nonzero elements in each row of $\hat{\boldsymbol{\Sigma}}^\star(\delta)$. Furthermore,

$$\begin{aligned}&\inf_{\boldsymbol{\Sigma}_0\in\bar{\mathcal{U}}_0}\mathsf{P}\bigg(\|\hat{\boldsymbol{\Sigma}}^\star(\delta) - \boldsymbol{\Sigma}_0\|_2\\ &\qquad\le C_{\gamma,\delta,M}\max_i \sigma^0_{ii} s_0(p)\bigg(\frac{\log p}{n}\bigg)^{1/2}\bigg) \ge 1 - O(p^{-M})\end{aligned} \tag{36}$$

for any $M > 0$, where $C_{\gamma,\delta,M}$ is a constant depending only on γ, δ, M, and

$$\sup_{\boldsymbol{\Sigma}_0\in\bar{\mathcal{U}}_0}\mathsf{E}\|\hat{\boldsymbol{\Sigma}}^\star(\delta) - \boldsymbol{\Sigma}_0\|_2^2 \le C s_0^2(p)\frac{\log p}{n} \tag{37}$$

for some constant $C > 0$.

Lemma 5. Let $\lambda_{ij} = \tau\sqrt{\frac{\hat{\theta}_{ij}\log p}{n}}$ with $0 < \tau < \sqrt{2}$. Under the conditions of Lemma 4,

$$\mathsf{P}\bigg(\min_i \sum_{j\in B_i} I\{|\hat{\sigma}_{ij}| \ge \lambda_{nij}(\tau)\} \ge p^{2\epsilon_0}\bigg) \to 1 \tag{38}$$

with any $\epsilon_0 < (1-\tau^2/2)/2$, where $B_i = \{j : \sigma_{ij}^0 = 0; j \neq i\}$. Thus, for some constant $C > 0$,

$$\inf_{\mathbf{\Sigma}_0 \in \mathcal{U}_0} \mathsf{P}\left(\|\hat{\mathbf{\Sigma}}^\star(\tau) - \mathbf{\Sigma}_0\|_2 \geq C \min_i \sigma_{ii}^0 p^{\epsilon_0/2} s_0(p) \left(\frac{\log p}{n}\right)^{1/2} \right) \to 1.$$

Proof of Theorem 4. To simplify notation, we write s_0 for $s_0(p)$. We construct a matrix $\mathbf{\Sigma}_0 \in \mathcal{U}_q^\star$. Let $s_1 = [(s_0 - 1)^{1-q}(\log p/n)^{-q/2}] + 1$ and $(X_1, \ldots, X_{s_1}), X_{s_1+1}, \ldots, X_p$ be independent. Let $\sigma_{ii}^0 = s_0$ for all $i > s_1$, $\sigma_{ii}^0 = 1$ for $1 \leq i \leq s_1$, and $\sigma_{ij}^0 = 4^{-1} s_0 \sqrt{\log p/n}$ for $1 \leq i \neq j \leq s_1$. Note that $\sigma_{ij}^0 = 0$ for $i \neq j > s_1$. Because $s_0 < 4\sqrt{n/\log p}$, $\mathbf{\Sigma}_0$ is a positive definite covariance matrix belonging to $\mathcal{U}_q^\star$. Set $\mathbf{M}_n = (\sigma_{ij}^0)_{1 \leq i,j \leq s_1}$. We first suppose that $\lambda_n \leq 3^{-1}\sigma_{pp}^0 \sqrt{2\log p/n}$. Lemma 5 yields

$$\mathsf{P}\left(\sum_{j=s_1+1}^{p} I\left\{ |\hat{\sigma}_{pj}| \geq \frac{\sqrt{2}}{2} \sigma_{pp}^0 \sqrt{\frac{\log p}{n}} \right\} \geq p^{2\epsilon_0} \right) \to 1,$$

with any $\epsilon_0 < 3/8$. Take $\epsilon_0 = 7/20$, and note that $p^{1/4} \geq s_0$, $p^{1/10} \geq n^{q/2}$. By the inequality $|s_\lambda(z)| \geq z - \lambda$,

$$\inf_{\lambda_n \leq 3^{-1}\sigma_{pp}^0 \sqrt{2\log p/n}} \sup_{\mathcal{U}_q^\star} \mathsf{P}\Bigg(\|\hat{\mathbf{\Sigma}}_g - \mathbf{\Sigma}_0\|_2 > \frac{\sqrt{2}}{6} s_0^2(p) \left(\frac{\log p}{n}\right)^{(1-q)/2} \Bigg) \to 1. \quad (39)$$

We next consider the case $\lambda_n > 3^{-1}\sigma_{pp}^0 \sqrt{2\log p/n}$. We have

$$\|\hat{\mathbf{\Sigma}}_g - \mathbf{\Sigma}_0\|_2 \geq \|\hat{\mathbf{M}}_n - \mathbf{M}_n\|_2,$$

where $\hat{\mathbf{M}}_n = (\hat{\sigma}_{ij}^g)_{1 \leq i,j \leq s_1}$. As in Lemma 2, for any $\gamma > 0$, we can get

$$\mathsf{P}\Big(\max_{1 \leq i,j \leq s_1} |\hat{\sigma}_{ij} - \sigma_{ij}^0| \geq \sqrt{2\gamma \log p/n} \Big) \leq C s_1^2 (\log p)^{-1/2} p^{-\gamma}.$$

Taking $\gamma = 1$, we have, with probability tending to 1, $\max_{1<i<j\leq s_1} |\hat{\sigma}_{ij}| \leq (4^{-1}s_0 + \sqrt{2})\sqrt{\log p/n}$, which implies that $\hat{\sigma}_{ij}^g = 0$ for $1 \leq i \neq j \leq s_1$. Thus, with probability tending to 1,

$$\|\hat{\mathbf{M}}_n - \mathbf{M}_n\|_2 \geq (4^{-1} - \sqrt{2}s_0^{-1}) s_1 s_0 \sqrt{\frac{\log p}{n}} \geq \frac{3}{64} s_0^{2-q} \left(\frac{\log p}{n}\right)^{(1-q)/2}.$$

This and (39) together imply (21).

Proof of Theorem 5 and Proposition 2. For brevity, we consider only the case where $H = 1$. The proof for general H is similar. We first show that for any $\varepsilon > 0$,

$$\mathsf{P}(\hat{\delta} \geq \sqrt{2} - \varepsilon) \to 1. \quad (40)$$

Because the random split is independent with the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, we can assume that the two samples are $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\{\mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_n\}$. Let $\hat{\mathbf{\Sigma}}_2$ be the sample covariance matrix from $\{\mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_n\}$, and let $\hat{\mathbf{\Sigma}}_1^\star(\delta)$ be defined as in (11) from $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$. Define

$$\hat{\delta}_o = \hat{j}_o/N, \quad \text{where } \hat{j}_o = \arg\min_{0 \leq j \leq 4N} \|\hat{\mathbf{\Sigma}}_1^\star(j/N) - \mathbf{\Sigma}_0\|_F^2.$$

Set $a_n = p^{-1}\|\hat{\mathbf{\Sigma}}_1^\star(\hat{\delta}) - \mathbf{\Sigma}_0\|_F^2$ and $r_n = p^{-1}\|\hat{\mathbf{\Sigma}}_1^\star(\hat{\delta}_o) - \mathbf{\Sigma}_0\|_F^2$. By the proof of Theorem 1, we have $\mathsf{P}(\|\hat{\mathbf{\Sigma}}_1^\star(2) - \mathbf{\Sigma}_0\|_{L_1} \leq C_1 s_0(p)(\log p/n)^{1/2}) \to 1$ for some $C_1 > 0$. Using the inequality $p^{-1}\|\mathbf{A}\|_F^2 \leq |\mathbf{A}|_\infty \|\mathbf{A}\|_{L_1}$ for any $p \times p$ symmetric matrix $\mathbf{A}$ and the definition of $\hat{\delta}_o$, we have $\mathsf{P}(r_n \leq C_2 s_0(p) \log p/n) \to 1$ for some $C_2 > 0$. Note that

$$\mathsf{E}|(\mathbf{V}, \hat{\mathbf{\Sigma}}_2 - \mathbf{\Sigma}_0)|^2 \leq C n^{-1}$$

for any $p \times 1$ vector $\mathbf{V}$ with $\|\mathbf{V}\|_F = 1$. By the proof of theorem 3 of Bickel and Levina (2008) and the assumption that N is fixed, we can see that

$$a_n \leq O_{\mathsf{P}}\left(\frac{1}{n^{1/2}}\right) a_n^{1/2} + O_{\mathsf{P}}\left(\frac{1}{n^{1/2}}\right) r_n^{1/2} + r_n. \quad (41)$$

Thus, for some $C_3 > 0$,

$$\mathsf{P}(a_n \leq C_3 s_0(p) \log p/n) \to 1. \quad (42)$$

Note that by applying Lemma 5 to the samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$,

$$\mathsf{P}(a_n \leq C_3 s_0(p) \log p/n, \hat{\delta} < \sqrt{2} - \varepsilon) = o(1).$$

This, together with (42), shows that

$$\mathsf{P}(\hat{\delta} < \sqrt{2} - \varepsilon) \leq \mathsf{P}(\hat{\delta} < \sqrt{2} - \varepsilon, a_n \leq C_3 s_0(p) \log p/n) + o(1) = o(1),$$

and thus (40) holds. Because N is fixed, we have $|\hat{\sigma} - \sqrt{2}| \geq \varepsilon_0$ for some fixed $\varepsilon_0 > 0$ that depends on N. This, together with (40), implies that

$$\mathsf{P}(\hat{\delta} \geq \sqrt{2} + \epsilon) \to 1 \quad (43)$$

for some $\epsilon > 0$. By Lemma 4, we see that with probability tending to 1, for each i, there are at most $O(s_0(p))$ nonzero numbers of $\{|s_{\lambda_{ij}}(\hat{\sigma}_{ij})|; j \in B_i\}$, and by Lemma 2, these are of order $O(\max_i \sigma_{ii}^0 \sqrt{\log p/n})$. Let $\Psi_i = \{j : \sigma_{ij}^0 \neq 0\}$ and $\hat{\Psi}_i = \{j : \hat{\sigma}_{ij}^\star \neq 0\}$. Then, by the conditions on $s_\lambda(z)$, we have

$$\|\hat{\mathbf{\Sigma}}^\star(\hat{\delta}) - \mathbf{\Sigma}_0\|_{L_1} \leq \max_i \sum_{j \in \Psi_i \cup \hat{\Psi}_i} |s_{\lambda_{ij}}(\hat{\sigma}_{ij}) - \sigma_{ij}^0| \leq C \max_i \sigma_{ii}^0 s_0(p) \left(\frac{\log p}{n}\right)^{1/2} \quad (44)$$

with probability tending to 1. The proof of Theorem 5 is completed. Finally, Proposition 2 is proved by (43) and Lemmas 2 and 4.

SUPPLEMENTARY MATERIALS

Additional proofs: A supplement to the main article contains additional technical arguments including the proofs of Lemmas 2, 4, and 5. (Supplement.pdf)

[Received September 2010. Revised January 2011.]

REFERENCES

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelet Approximations," *Journal of the American Statistical Association*, 96, 939–967. [672,674]

Cai, T. T., and Zhou, H. H. (2009), "Minimax Estimation of Large Covariance Matrices Under l_1 Norm," technical report, Department of Statistics, University of Pennsylvania. [672]

——— (2010), "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," technical report, Department of Statistics, University of Pennsylvania. [672,675]

Bickel, P., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [672-674,676,677,683]

Donoho, D. L., and Johnstone, J. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [672]

——— (1998), "Minimax Estimation via Wavelet Shrinkage," *The Annals of Statistics*, 26, 879–921. [672]

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large-Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756. [672]

Hawkins, D. L. (1989), "Using U Statistics to Derive the Asymptotic Distribution of Fisher's Z Statistic," *Journal of the American Statistical Association*, 43, 235–237. [680]

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. (2001), "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, 7, 673–679. [673,678]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [672-679]

Shao, Q. M. (1999), "A Cramér Type Large Deviation Result for Student's t-Statistic," *Journal of Theoretical Probability*, 12, 385–398. [676]

Wang, Y., and Zou, J. (2010), "Vast Volatility Matrix Estimation for High-Frequency Financial Data," *The Annals of Statistics*, 38, 943–978. [672]

Zaitsev, A. Yu. (1987), "Estimates of the Lévy–Prokhorov Distance in the Multivariate Central Limit Theorem for Random Variables With Finite Exponential Moments," *Theory of Probability and Its Applications*, 31, 203–220. [681]