Generalized Thresholding of Large Covariance Matrices

Adam J. ROTHMAN, Elizaveta LEVINA, and Ji ZHU

We propose a new class of generalized thresholding operators that combine thresholding with shrinkage, and study generalized thresholding of the sample covariance matrix in high dimensions. Generalized thresholding of the covariance matrix has good theoretical properties and carries almost no computational burden. We obtain an explicit convergence rate in the operator norm that shows the tradeoff between the sparsity of the true model, dimension, and the sample size, and shows that generalized thresholding is consistent over a large class of models as long as the dimension p and the sample size n satisfy $\log p/n \to 0$. In addition, we show that generalized thresholding has the "sparsistency" property, meaning it estimates true zeros as zeros with probability tending to 1, and, under an additional mild condition, is sign consistent for nonzero elements. We show that generalized thresholding covers, as special cases, hard and soft thresholding, smoothly clipped absolute deviation, and adaptive lasso, and compare different types of generalized thresholding in a simulation study and in an example of gene clustering from a microarray experiment with tumor tissues.

KEY WORDS: Covariance; High-dimensional data; Regularization; Thresholding; Sparsity

1. INTRODUCTION

There is an abundance of problems in high-dimensional inference when an estimate of the covariance matrix is of interest: principal components analysis, classification by discriminant analysis, inferring a graphical model structure, and others. Examples of application areas in which these problems arise include gene arrays, functional MRI, text retrieval, image classification, spectroscopy, and climate studies. The properties of the traditional estimator, the sample covariance matrix, are by now fairly well understood (see, for example, Johnstone (2001) and references therein), and it is clear that alternative estimators are needed in high dimensions.

The existing literature on covariance estimation can be loosely divided into two categories. One large class of methods covers the situation in which variables have a natural ordering or there is a notion of distance between variables, as in longitudinal data, time series, spatial data, or spectroscopy. The implicit regularizing assumption here is that variables far apart are only weakly correlated, and estimators that take advantage of this have been proposed by Wu and Pourahmadi (2003); Bickel and Levina (2004); Huang et al. (2006); Furrer and Bengtsson (2007); Bickel and Levina (2008); Levina, Rothman, and Zhu (2008); and others.

There are, however, many applications in which an ordering of the variables is not available, such as genetics and social, financial, and economic data. Methods that are invariant to variable permutations (like the covariance matrix itself) are necessary in such applications. A common approach to permutation-invariant covariance regularization is encouraging sparsity. Adding a lasso penalty on the entries of the inverse covariance to the normal likelihood has been discussed by d'Aspremont, Banerjee, and El Ghaoui (2008); Yuan and Lin (2007); Rothman et al. (2008); and Friedman, Hastie, and Tibshirani (2008); and has extended to more general penalties by Lam and Fan (2007). Although relatively fast algorithms have been proposed by Friedman, Hastie, and Tibshirani (2008) and Rothman et al. (2008) for solving these penalized likelihood problems, they require computationally intensive methods and typically only provide a sparse estimate of the inverse, not of the covariance matrix itself.

A simple alternative to penalized likelihood is thresholding the sample covariance matrix, which has been analyzed by Bickel and Levina (2007) and El Karoui (2007). Thresholding carries essentially no computational burden, except for cross-validation for the tuning parameter (which is also necessary for penalized likelihood) and is thus an attractive option for problems in very high dimensions and real-time applications. However, in regression and wavelet shrinkage contexts (see, for example, Donoho et al. (1995) and Fan and Li (2001)), hard thresholding tends to do worse than more flexible estimators that combine thresholding with shrinkage—for example, soft thresholding or smoothly clipped absolute deviation (SCAD) (Fan and Li 2001). The estimates resulting from such shrinkage typically are continuous functions of the "naive" estimates, a desirable feature not shared by hard thresholding.

In this article, we generalize the thresholding approach to covariance estimation to a whole class of estimators based on elementwise shrinkage and thresholding. For any $\lambda \geq 0$, define a generalized thresholding operator to be a function $s_\lambda : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions for all $z \in \mathbb{R}$:

(i) $|s_\lambda(z)| \leq |z|$;
(ii) $s_\lambda(z) = 0$ for $|z| \leq \lambda$;
(iii) $|s_\lambda(z) - z| \leq \lambda$.

It is also natural, although not strictly necessary, to have $s_\lambda(z) = \text{sign}(z)\, s_\lambda(|z|)$. Condition (i) establishes shrinkage, condition (ii) enforces thresholding, and condition (iii) limits the amount of shrinkage to no more than λ. It is possible to have different parameters λ_1 and λ_2 in (ii) and (iii); for simplicity, we keep them the same. For a related discussion of penalties that have such properties, see also Antoniadis and Fan (2001).

The rest of this article is organized as follows. To make our definition of generalized thresholding concrete, we start by

Adam J. Rothman is a Ph.D. candidate, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 (E-mail: *ajrothma@umich.edu*). Elizaveta Levina is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 (E-mail: *elevina@umich.edu*). Ji Zhu is Associate Professor Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 (E-mail: *jizhu@umich.edu*). Elizaveta Levina's research is supported in part by grants from the National Science Foundation (NSF; DMS-0505424 and DMS-0805798). Ji Zhu's research is supported in part by grants from the NSF (DMS-0505432 and DMS-0705532). The authors thank an Associate Editor and two referees for helpful suggestions

Journal of the American Statistical Association
March 2009, Vol.104, No. 485, Theory and Methods
DOI 10.1198/jasa.2009.0101

giving examples in Section 2, and show that generalized thresholding covers many popular shrinkage/thresholding functions, including hard and soft thresholding, SCAD (Fan and Li 2001), and adaptive lasso (Zou 2006). In Section 3, we establish convergence rates for generalized thresholding of the sample covariance on a class of "approximately sparse" matrices, and show they are consistent as long as log p/n tends to 0. We also show that generalized thresholding is, in the terminology of Lam and Fan (2007), "sparsistent," meaning that in addition to being consistent it estimates true zeros as zeros with probability tending to 1, and, under an additional condition, estimates nonzero elements as nonzero, with the correct sign, with probability tending to 1. This property is sometimes referred to as "sign consistency." Simulation results are given in Section 4, where we show that although all the estimators in this class are guaranteed the same bounds on convergence rates and have similar performance in terms of overall loss, the more flexible penalties like SCAD are substantially better at getting the true sparsity structure. Last, Section 5 presents an application of the methods to gene expression data on small round blue-cell tumors. The Appendix contains all the proofs.

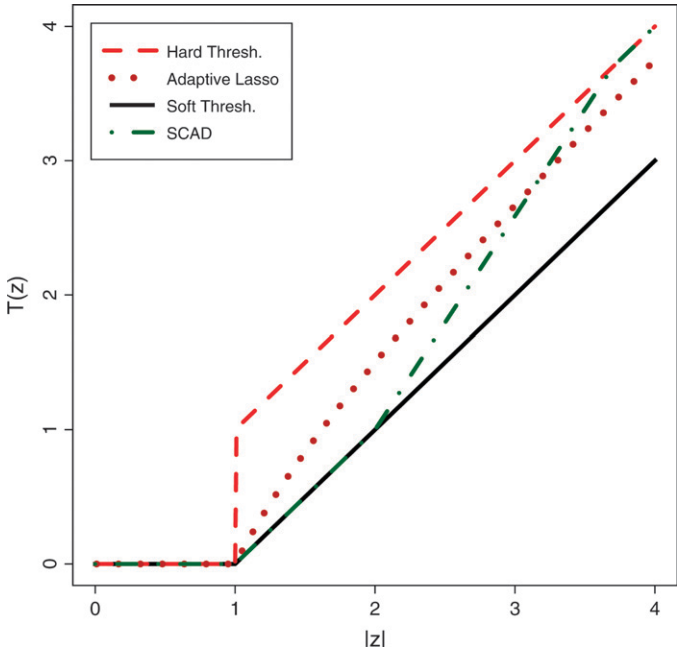

Figure 1. Generalized thresholding functions for $\lambda = 1$, $a = 3.7$, and $\eta = 1$.

2. EXAMPLES OF GENERALIZED THRESHOLDING

It turns out that conditions (i) through (iii), which define generalized thresholding, are satisfied by a number of commonly used shrinkage/thresholding procedures. These procedures are commonly introduced as solutions to penalized quadratic loss problems with various penalties. Because, in our case, the procedure is applied to each element separately, the optimization problems are univariate. Suppose $s_\lambda(z)$ is obtained as

$$s_\lambda(z) = \arg\min_\theta \left\{ \frac{1}{2}(\theta - z)^2 + p_\lambda(\theta) \right\}, \tag{1}$$

where p_λ is a penalty function. Next, we check that several popular penalties and thresholding rules satisfy our conditions for generalized thresholding. For more details on the relationship between penalty functions and resulting thresholding rules, see Antoniadis and Fan (2001).

The simplest example of generalized thresholding is the hard thresholding rule,

$$s_\lambda^H(z) = z1(|z| > \lambda), \tag{2}$$

where $1(\cdot)$ is the indicator function. Hard thresholding obviously satisfies conditions (i) through (iii).

Soft thresholding results from solving (1) with the lasso (λ_1) penalty function, $p_\lambda(\theta) = \lambda|\theta|$, and gives the rule

$$s_\lambda^s(z) = \text{sign}(z)(|z| - \lambda)_+. \tag{3}$$

Soft thresholding has been studied in the context of wavelet shrinkage by Donoho and Johnstone (1994) and Donoho et al. (1995), and in the context of regression by Tibshirani (1996). The soft-thresholding operator s_λ^S obviously satisfies conditions (i) and (ii). To check (iii), note that $|s_\lambda^S(z) - z| = |z|$ when $|z| \le \lambda$, and $|s_\lambda^S(z) - z| = \lambda$ when $|z| > \lambda$. Thus, soft thresholding corresponds to the maximum amount of shrinkage allowed by condition (iii), whereas hard thresholding corresponds to no shrinkage.

The SCAD penalty was proposed by Fan (1997) and Fan and Li (2001) as a compromise between hard and soft thresholding. Like soft thresholding, it is continuous in z, but the amount of shrinkage decreases as $|z|$ increases, and after a certain threshold there is no shrinkage, which results in less bias. The SCAD thresholding function is a linear interpolation between soft thresholding up to 2λ and hard thresholding after $a\lambda$ (see Fig. 1). The value $a = 3.7$ was recommended by Fan and Li (2001), and we use it throughout the article. See Fan and Li (2001) for the formulas of the SCAD thresholding function and the corresponding penalty function. The SCAD thresholding operator s_λ^{SC} satisfies conditions (i) through (iii): Condition (ii) is immediate, and (i) and (iii) follow from $|s^S(|z|)| \le |s^{SC}(|z|)| \le |s^H(|z|)|$.

Another idea proposed to mitigate the bias of lasso for large regression coefficients is adaptive lasso (Zou 2006). In regression context, the idea is to multiply each $|\beta_j|$ in the lasso penalty by a weight w_j, which is smaller for larger initial estimates $\hat{\beta}_j$. Thus, large coefficients get penalized less. One choice of weights proposed was $w_j = |\hat{\beta}_j|^{-\eta}$, where $\hat{\beta}_j$ is an ordinary least-squares estimate. Note that in the context of regression, the special case $\eta = 1$ is closely related to the nonnegative garrote (Breiman 1995). In our context, an analogous weight would be $|\hat{\sigma}_{ij}|^{-\eta}$. We can rewrite this as a penalty function $p_\lambda(\theta) = \lambda w(z)|\theta|$, where w is taken to be $C|z|^{-\eta}$, $\eta \ge 0$. Zou (2006) has $C = 1$ (it is absorbed in λ), but for us it is convenient to set $C = \lambda^\eta$, because then the resulting operator satisfies condition (ii)—in other words, thresholds everything below λ to 0. The resulting thresholding rule corresponding to $C = \lambda^\eta$, which we still call "adaptive lasso" for simplicity, is given by

$$s_\lambda^{AL}(z) = \text{sgn}(z)\left(|z| - \lambda^{n+1}|z|^{-\eta}\right)_+. \tag{4}$$

Conditions (i) and (ii) are obviously satisfied. To check (iii) for $|z| > \lambda$, note that $|s_\lambda^{AL}(z) - z| = \lambda^{\eta+1}|z|^{-\eta} \le \lambda$.

As illustrated in Figure 1, both SCAD and adaptive lasso fall in between hard and soft thresholding; any other function sandwiched between hard and soft thresholding will satisfy conditions (i) through (iii)—for example, the clipped L_1 penalty. For conditions on the penalty p_λ that imply the resulting operator is sandwiched between hard and soft thresholding, see Antoniadis and Fan (2001). In this article, we focus on the operators themselves rather than the penalties, because the penalties are never used directly.

3. CONSISTENCY AND SPARSITY OF GENERALIZED THRESHOLDING

In this section, we derive theoretical properties of the generalized thresholding estimator in the high-dimensional setting, meaning that both the dimension and the sample size are allowed to grow. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote iid p-dimensional random vectors sampled from a distribution F with $E\boldsymbol{X}_1 = 0$ (without loss of generality), and $E(\boldsymbol{X}_1\boldsymbol{X}_1^T) = \Sigma$. The convention in the literature is to assume that F is Gaussian. The key result underlying this theory, however, is the bound (A.4) we give in the Appendix. Bickel and Levina (2008) noted that for this result the normal assumption can be replaced with a tail condition on the marginal distributions—namely, that for all $1 \le j \le p$, if G_j in the cumulative distribution function of X_{1j}^2, then

$$\int_0^\infty \exp(\lambda t)\, dG_j(t) < \infty \quad \text{for} \quad 0 < |\lambda| < \lambda_0 \quad \text{for some } \lambda_0 > 0. \tag{5}$$

Let $\hat{\Sigma}$ denote the sample covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{X}_k - \bar{\boldsymbol{X}})(\boldsymbol{X}_k - \bar{\boldsymbol{X}})^T. \tag{6}$$

Let $s_\lambda(\boldsymbol{A}) = [s_\lambda(a_{ij})]$ denote the matrix resulting from applying a generalized thresholding operator s_λ to each of the elements of a matrix $\boldsymbol{A}$. Condition (ii) implies that $s_\lambda(\boldsymbol{A})$ is sparse for sufficiently large λ. As with hard thresholding and banding of the covariance matrix, the estimator $s_\lambda(\hat{\boldsymbol{\Sigma}})$ is not guaranteed to be positive definite, but instead we show that it converges to a positive definite limit with probability tending to 1.

We proceed to establish a bound on the convergence rate for $s_\lambda(\hat{\boldsymbol{\Sigma}})$. The result is uniform on a class of "approximately sparse" covariance matrices, which was introduced by Bickel and Levina (2007):

$$\mathcal{U}_\tau(q, c_0(p), M) = \left\{ \boldsymbol{\Sigma} : \sigma_{ii} \le M,\ \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p) \right\}, \tag{7}$$

for $0 \le q < 1$. When $q = 0$, this is a class of truly sparse matrices. For example, a d-diagonal matrix satisfies this condition with any $0 \le q < 1$ and $c_0(p) = M^q d$. Another example is the AR(1) covariance matrix, $\sigma_{ij} = \rho^{|i-j|}$, which satisfies the condition with $c_0(p) \equiv c_0$. Note that the condition of bounded variances, $\sigma_{ii} \le M$, is weaker than the often assumed bounded eigenvalues condition, $\lambda_{\max}(\boldsymbol{\Sigma}) \le M$. Also note that the constant $c_0(p)$ is allowed to depend on p and is thus not an explicit restriction on sparsity. The convergence will be established in the matrix operator norm (also known as "spectral" or "l_2 matrix norm"), $\|\boldsymbol{A}\|^2 = \lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T)$.

Theorem 1 (Consistency). Suppose s_λ satisfies conditions (i) through (iii) and F satisfies condition (5). Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large M', if $\lambda = M'\sqrt{(\log p/n)} = o(1)$,

$$\left\| s_\lambda(\hat{\Sigma}) - \Sigma \right\| = O_P\left(c_0(p) \left(\frac{\log p}{n} \right)^{\frac{1-q}{2}} \right).$$

Proof of Theorem 1 is given in the Appendix. For the case of hard thresholding, this theorem was established in Bickel and Levina (2007). Note that, through $c_0(p)$, the rate depends explicitly on how sparse the truth is. Also note that this rate is very similar to the rate of $\sqrt{(s \log p/n)}$ for a sparse estimator of the inverse covariance matrix established in Rothman et al. (2008), where s is the number of nonzero off-diagonal elements in the true inverse, even though the estimator is obtained by a completely different approach of adding a lasso penalty to the normal likelihood. The fundamental result underlying these different analyses, however, is the bound (A.4), which ultimately gives rise to similar rates.

Next, we state a sparsity result, which, together with Theorem 1, establishes the "sparsistency" property in the sense of Lam and Fan (2007).

Theorem 2 (Sparsity). Suppose s_λ satisfies conditions (i) through (iii), F satisfies (5), and $\sigma_{ii} \le M$ for all i. Then, for sufficiently large M', if $\lambda = M'\sqrt{(\log p/n)} = o(1)$,

$$s_\lambda(\hat{\sigma}_{ij}) = 0 \text{ for all } (i,j) \text{ such that } \sigma_{ij} = 0, \tag{8}$$

with probability tending to 1. If we additionally assume that all nonzero elements of $\boldsymbol{\Sigma}$ satisfy $|\sigma_{ij}| > \tau$, where $\sqrt{n}(\tau - \lambda) \to \infty$, we also have, with probability tending to 1,

$$\operatorname{sign}(s_\lambda(\hat{\sigma}_{ij}) \cdot \sigma_{ij}) = 1 \text{ for all } (i,j) \text{ such that } \sigma_{ij} \ne 0. \tag{9}$$

The proof is given in the Appendix. Note that Theorem 2 only requires that the true variances are bounded, and not the approximately sparse assumption. The additional condition on nonzero elements is analogous to the condition of El Karoui (2007) that nonzero elements are greater than $n^{-\alpha}$. If we assume the same (i.e., let $\tau = n^{-\alpha}$), the result holds under a slightly stronger condition $\log p/n^{1-2\alpha} \to 0$ instead of $\log p/n \to 0$. It may also be possible to develop further joint asymptotic normality results for nonzero elements along the lines of Fan and Peng (2004) or Lam and Fan (2007), but we do not pursue this further because of restrictive conditions required for the method of proof used there ($p^2/n \to 0$).

4. SIMULATION RESULTS

4.1 Simulation Settings

To compare the performance of various generalized thresholding estimators, both in terms of the overall covariance estimation and recovering the sparsity pattern, we conducted a simulation study with the following three covariance models:

Table 1. Average (standard error) operator norm loss for model 1

p	ρ	Sample	Hard	Soft	Adapt.lasso	SCAD
30	0.3	1.30 (0.02)	0.75 (0.01)	0.71 (0.01)	0.71 (0.01)	0.71 (0.01)
30	0.7	1.75 (0.04)	1.56 (0.04)	1.59 (0.05)	1.53 (0.04)	1.47 (0.04)
100	0.3	3.09 (0.03)	0.93 (0.01)	0.86 (0.01)	0.86 (0.01)	0.85 (0.01)
100	0.7	4.10 (0.07)	2.17 (0.04)	2.49 (0.03)	2.30 (0.04)	2.16 (0.04)
200	0.3	4.90 (0.03)	0.98 (0.01)	0.90 (0.00)	0.91 (0.01)	0.90 (0.00)
200	0.7	6.63 (0.08)	2.46 (0.03)	2.86 (0.02)	2.65 (0.03)	2.52 (0.03)
500	0.3	9.69 (0.04)	1.06 (0.01)	0.95 (0.00)	0.96 (0.00)	0.95 (0.00)
500	0.7	12.54 (0.08)	2.80 (0.02)	3.23 (0.02)	3.01 (0.02)	2.97 (0.02)

Model 1: AR(1), where $\sigma_{ij} = \rho^{|i-j|}$, for $\rho = 0.3$ and 0.7.
Model 2: MA(1), where $\sigma_{ij} = \rho 1(|i-j| = 1) + 1(i = j)$, for $\rho = 0.3$.
Model 3: "Triangular" covariance, $\sigma_{ij} = (1 - ((|i-j|)/k))_+$, for $k = \lfloor p/2 \rfloor$.

Models 1 and 2 are standard test cases in the literature. Note that even though these models come from time series, all estimators considered here are permutation invariant, and thus the order of the variables is irrelevant. Model 1 is "approximately sparse," because even though there are no true zeros, there are many very small entries away from the diagonal. Model 2 is a tridiagonal covariance matrix and is the most sparse of the three models. Model 3 has a linear decay in covariances as one moves away from the diagonal, and it provides a simple way to generate a positive definite matrix, with the level of sparsity controlled by the parameter k. With $k = p/2$, Model 3 is effectively the least sparse of the three models we consider. This covariance structure was considered by Wagaman and Levina (2007).

For each model, we generated $n = 100$ independent and identically distributed p-variate normal random vectors with mean 0 and covariance $\mathbf{\Sigma}$, for $p = 30$, 100, 200, and 500. The number of replications was fixed at 50. The tuning parameter λ for each method was selected by minimizing the Frobenius norm of the difference between $s_\lambda(\hat{\mathbf{\Sigma}})$ and the sample covariance matrix computed from 100 independently generated validation data observations. We note that the use of a validation set can be replaced with cross-validation without any significant change in results. We selected the Frobenius norm ($\|\boldsymbol{A}\|_F^2 = \sum_{i,j} a_{ij}^2$) for tuning because it had a slightly better performance than the operator norm or the matrix l_1 norm. Also, a theoretical justification for this choice for cross-validation has been provided by Bickel and Levina (2007).

4.2 Performance Evaluation

Keeping consistent with theory in Section 3, we defined the loss function for the estimators by the expected operator norm of the difference between the true covariance and the estimator:

$$L(s_\lambda(\hat{\Sigma}), \Sigma) = E\left\| s_\lambda(\hat{\Sigma}) - \Sigma \right\|.$$

The ability to recover sparsity was evaluated via the true-positive rate (TPR) in combination with the false-positive rate (FPR), defined as

$$\text{TPR} = \frac{\#\{(i,j) : s_\lambda(\hat{\sigma}_{ij}) \neq 0 \quad \text{and} \quad \sigma_{ij} \neq 0\}}{\#\{(i,j) : \sigma_{ij} \neq 0\}}, \tag{10}$$

$$\text{FPR} = \frac{\#\{(i,j) : s_\lambda(\hat{\sigma}_{ij}) \neq 0 \quad \text{and} \quad \sigma_{ij} = 0\}}{\#\{(i,j) : \sigma_{ij} = 0\}}. \tag{11}$$

Note that the sample covariance has TPR = 1, and a diagonal estimator has FPR = 0.

In addition, we compute a measure of agreement of principal eigenspaces between the estimator and the truth, which is relevant for principal components analysis. The measure we use to compare the eigenspaces spanned by the first q eigenvectors was defined by Krzanowski (1979) as

$$K(q) = \sum_{i=1}^{q} \sum_{j=1}^{q} \left(\hat{\boldsymbol{e}}_{(i)}^T \boldsymbol{e}_{(j)} \right)^2, \tag{12}$$

where $\hat{\boldsymbol{e}}_{(i)}$ denotes the estimated eigenvector corresponding to the i-th largest estimated eigenvalue, and $\boldsymbol{e}_{(i)}$ is the true eigenvector corresponding to the true i-th largest eigenvalue. Computing cosines of angles between all possible pairs of eigenvectors removes the problem of similar eigenvectors estimated in a different order. Note that $K(0) \equiv 0$ and $K(p) = p$. For any $0 < q < p$, perfect agreement between the two eigenspaces will result in $K(q) = q$. A convenient way to evaluate this measure is to plot $K(q)$ against q. Alternative measures of eigenvector agreement are available; for example, Fan, Wang, and Yao (2008) proposed using the measure

$$D(q) = 1 - \frac{1}{q} \sum_{i=1}^{q} \max_{1 \le j \le q} \left| \boldsymbol{e}_{(i)}^T \hat{\boldsymbol{e}}_j \right|,$$

which shares many of the properties of the Krzanowski's measure, such as invariance to permutations of the eigenvector order.

4.3 Summary of Results

Table 1 summarizes simulation results for the AR(1) model. Note that this model is not truly sparse, and thus TPRs and FPRs are not relevant. All generalized thresholding estimators improve over the sample covariance matrix under the operator norm loss. This improvement increases with dimension p. The thresholding rules are all quite similar for this model, with perhaps hard thresholding having a slight edge for $\rho = 0.7$ (more large entries) and being slightly worse than the others for $\rho = 0.3$.

Table 2. Average (standard error) operator norm loss, and TPRs and FPRs for model 2

p	Sample	Hard	Soft	Adapt.lasso	SCAD
			Operator norm loss		
30	1.34 (0.02)	0.69 (0.01)	0.61 (0.01)	0.62 (0.01)	0.63 (0.01)
100	2.99 (0.02)	0.88 (0.01)	0.70 (0.01)	0.73 (0.01)	0.72 (0.01)
200	4.94 (0.03)	0.94 (0.02)	0.75 (0.01)	0.78 (0.01)	0.76 (0.01)
500	9.65 (0.04)	1.01 (0.02)	0.81 (0.01)	0.85 (0.01)	0.81 (0.01)
			TPR/FPR		
30	NA	0.70/0.01	0.94/0.18	0.88/0.08	0.95/0.21
100	NA	0.49/0.00	0.87/0.07	0.78/0.03	0.92/0.12
200	NA	0.33/0.00	0.81/0.04	0.69/0.01	0.91/0.11
500	NA	0.20/0.00	0.70/0.02	0.57/0.01	0.89/0.08

NOTE: NA, not applicable.

Table 2 gives results for model 2, the tridiagonal sparse truth. We again see a drastic improvement in estimation performance of the "thresholded" estimates over the sample covariance matrix, which increases with dimension. This is expected because this is the sparsest model we consider. Under operator norm loss, the rules that combine thresholding with shrinkage all outperform hard thresholding, with soft thresholding performing slightly better than SCAD and adaptive lasso.

The 50 realizations of the values of TPR and FPR are also plotted in Figure 2, in addition to their average values given in Table 2. Here we see a big difference between the different thresholding rules. Hard thresholding tends to zero out too many elements, presumably because of its inability to shrink moderate values; thus, it has a very low FPR, but also a lower TPR than the other methods, particularly for large p. Overall, Figure 2 suggests that the SCAD thresholding has the best performance on sparsity for this model, particularly for large values of p.

Table 3 gives results for the "triangular" model with $k = p/2$, the least sparse of the three models we consider. Here we see only a small improvement of thresholded estimates over the sample covariance in the operator norm loss. All methods miss a substantial fraction of true zeros, most likely because a large number of small nonzero true entries leads to a choice of threshold that is too low. In this case, hard thresholding does somewhat better on false positives, which we conjecture may in general be the case for less sparse models. However, the plot of realizations of TPR and FPR in Figure 3 shows that the variance is very high and there is no clear best choice for estimating the sparsity structure in this case.

In Figure 4, we plot the average eigenspace agreement measure $K(q)$ defined in (12) versus q for $p = 200$ in all four models. For effectively sparser models AR(1) and MA(1), all

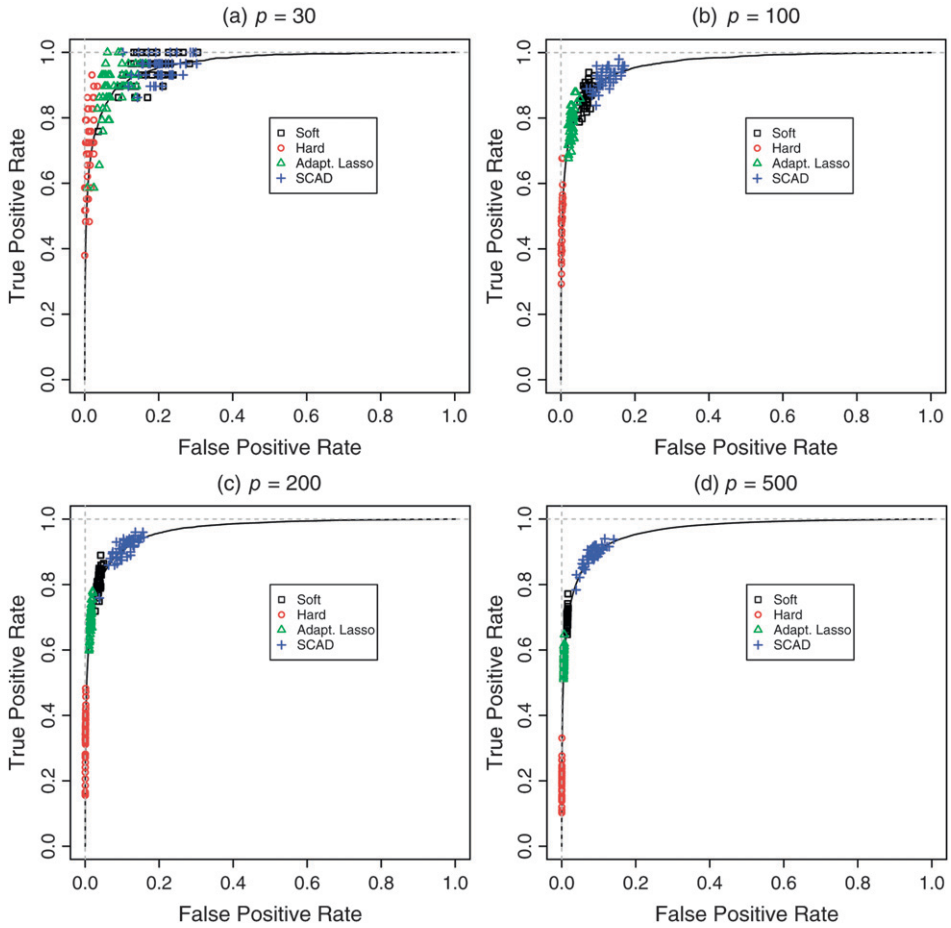

Figure 2. TPR versus FPR for model 2. The points correspond to 50 different realizations, with each method selecting its own threshold using validation data. The solid line is obtained by varying the threshold over the whole range (all methods have the same TPR and FPR for a fixed threshold).

Table 3. Average (standard error) operator norm loss, and TPRs and FPRs for model 3 (k = p/2)

p	Sample	Hard	Soft	Adapt.lasso	SCAD
			Operator norm loss		
30	2.55 (0.10)	2.40 (0.10)	2.33 (0.10)	2.34 (0.09)	2.39 (0.09)
100	8.67 (0.37)	8.10 (0.37)	8.05 (0.39)	7.99 (0.35)	8.11 (0.36)
200	17.66 (0.90)	16.81 (0.85)	16.42 (0.79)	16.21 (0.75)	16.69 (0.99)
500	43.71 (2.01)	40.49 (1.80)	42.75 (1.87)	41.08 (1.80)	40.60 (1.79)
			TPR/FPR		
30	NA	0.92/0.26	0.98/0.69	0.94/0.45	0.95/0.51
100	NA	0.91/0.28	0.98/0.72	0.94/0.54	0.94/0.46
200	NA	0.92/0.35	0.97/0.69	0.94/0.49	0.95/0.51
500	NA	0.90/0.39	0.98/0.79	0.94/0.54	0.95/0.59

NOTE: NA, not applicable.

thresholding methods improve on eigenspace estimation relative to the sample covariance, with SCAD and adaptive lasso showing the best performance. This effect is more pronounced for large p (plots not shown). For the less sparse triangular model, there is in fact no improvement relative to the covariance matrix, even though there is a slight improvement in overall operator norm loss. The eigenvalues corresponding to $q > 50$ here, however, are very small, and thus the differences in eigenspaces are inconsequential. The biggest improvement in eigenspace estimation across models is for AR(1) with $\rho = 0.7$, which is consistent with our expectations that these methods perform best for models with many small or zero entries and few large entries well separated from zero.

Overall, the simulations show that in truly sparse models, thresholding makes a big difference, and that penalties that combine the advantages of hard and soft thresholding tend to perform best at recovering the true zeros. When the true model is not sparse, the thresholded estimator does no worse than the sample covariance matrix, and thus in practice there does not seem to be any harm in applying thresholding even when there

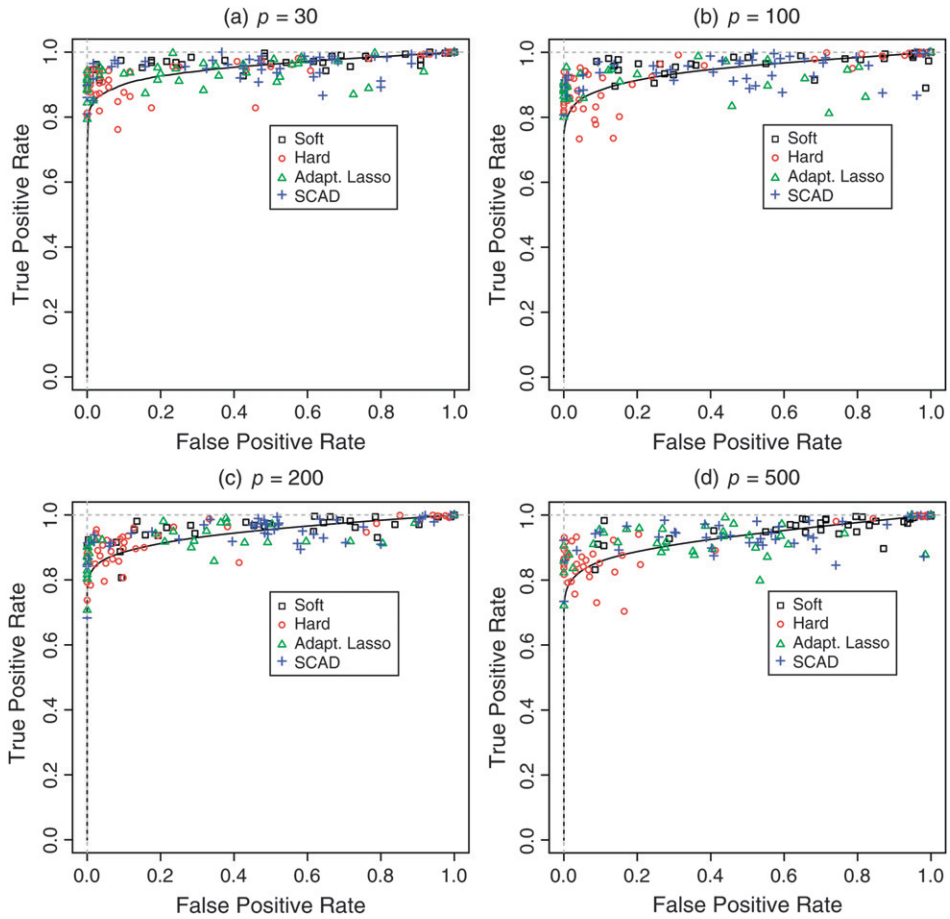

Figure 3. TPR versus FPR for model 3. The points correspond to 50 different realizations, with each method selecting its own threshold using validation data. The solid line is obtained by varying the threshold over the whole range (all methods have the same value of TPR and FPR for a fixed threshold).

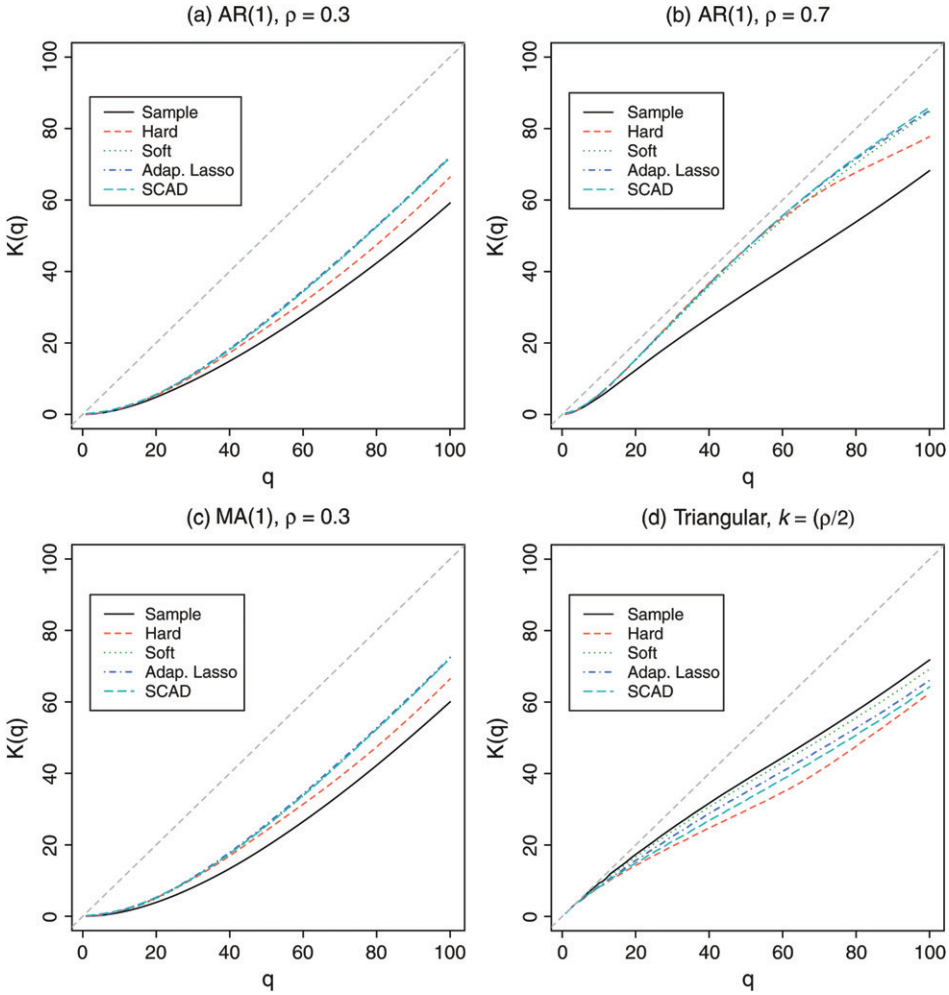

Figure 4. Average $K(q)$ versus q with $p = 200$.

is little or no prior information about the degree of sparsity of the true model.

5. EXAMPLE: GENE CLUSTERING VIA CORRELATIONS

Clustering genes using their correlations is a popular technique in gene expression data analysis (Eisen et al. 1998; Hastie et al. 2000). Here we investigate the effect of generalized thresholding on gene clustering using the data from a small round blue-cell tumors microarray experiment (Khan et al. 2001). The experiment had 64 training tissue samples, and 2,308 gene expression values recorded for each sample. The original dataset included 6,567 genes and was filtered down by requiring that each gene have a red intensity greater than 20 over all samples (for additional information, see Khan et al. (2001)). There are four types of tumors in the sample (EWS, BL-NHL, NB, and RMS).

First we ranked the genes by how much discriminative information they provide, using the F statistic:

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

where $k = 4$ is the number of classes, $n = 64$ is the number of tissue samples, n_m is the number of tissue samples of class m, $\bar{x}_m$ and $\hat{\sigma}_m^2$ are the sample mean and variance of class m, and $\bar{x}$ is the overall mean. Then we selected the top 40 and bottom 160 genes according to their F statistics, so that we have both informative and noninformative genes. This selection was done to allow visualizing the correlation matrices via heat maps.

We apply group average agglomerative clustering to genes using the estimated correlation in the dissimilarity measure,

$$d_{jj'} = 1 - |\hat{\rho}_{jj'}|, \tag{13}$$

where $\hat{\rho}_{jj'}$ is the estimated correlation between gene j and gene j'. We estimate the correlation matrix using hard, soft, adaptive lasso, and SCAD thresholding of the sample correlation matrix. The tuning parameter λ was selected via the resampling scheme described in Bickel and Levina (2008). The group-average agglomerative clustering is a bottom-up clustering method, which starts from treating all genes as singleton groups. Each step merges the two most similar groups, chosen to have the smallest average of pairwise dissimilarity between members of one group and the other. There are a total of $p - 1$ stages, and the last stage forms one group of size p. Figure 5 shows a heat map of the data, with rows (genes) sorted by hierarchical clustering based on the sample correlations, and

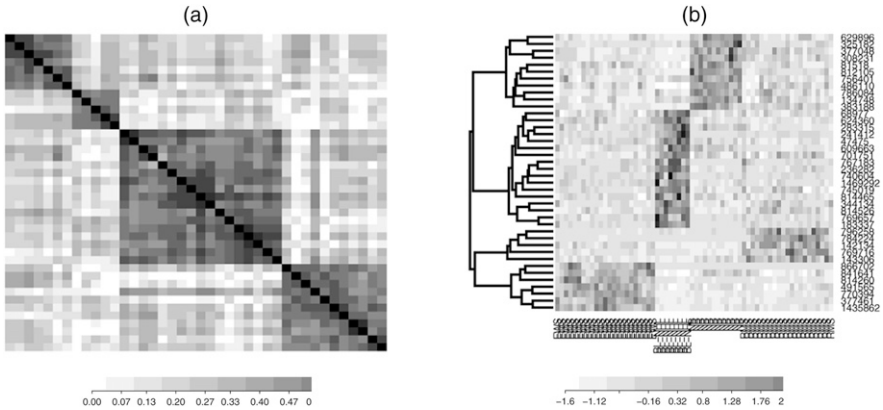

Figure 5. (a) Heat map of the absolute values of sample correlations of the top 40 genes. (b) Heat map of the gene expression data, with rows (genes) sorted by hierarchical clustering and columns sorted by tissue class.

columns (patients) sorted by tissue class for the 40 genes with the highest F statistics, along with a heat map of the sample correlations (absolute values) of the 40 genes ordered by hierarchical clustering. In all correlation heat maps, we plot absolute values rather than the correlations themselves, because here we are interested in the strength of pairwise association

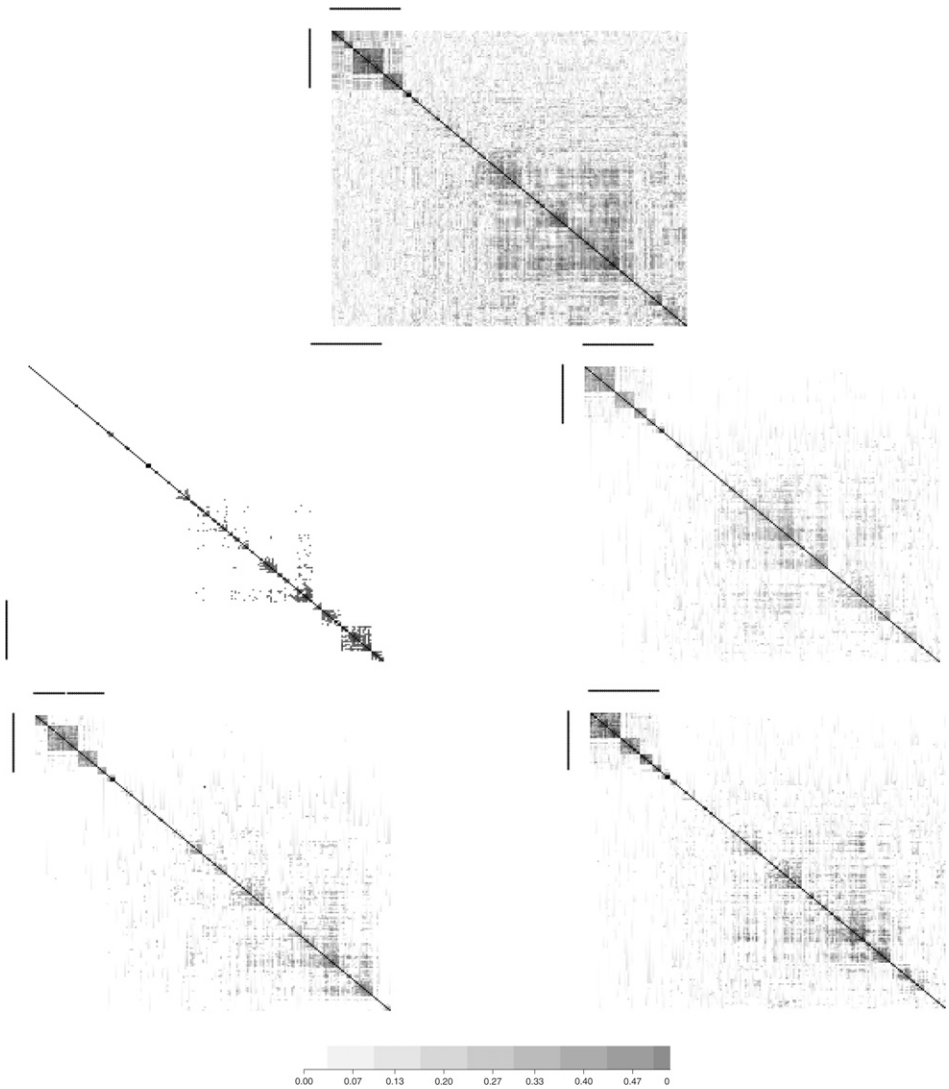

Figure 6. Heat maps of the absolute values of estimated correlations. The 40 genes with the largest F statistic are marked with lines. The genes are ordered by hierarchical clustering using estimated correlations. The percentage of off-diagonal elements estimated as zero is given in parentheses for each method.

between the genes regardless of its sign. It is clear that these 40 genes form strongly correlated blocks that correspond to different classes.

The resulting heat maps of the correlation matrix ordered by hierarchical clustering for each thresholding method are shown in Figure 6, along with the percentage of off-diagonal entries estimated as zero. Hard thresholding estimates many more zeros than other methods, resulting in a nearly diagonal estimator. This is consistent with hard-thresholding results in simulations, where hard thresholding tended to threshold too many entries, especially in higher dimensions. Also consistent with the simulation study is the performance of SCAD, which estimates the smallest number of zeros and appears to do a good job at cleaning up the signal without losing the block structure. As in simulations, the performance of adaptive lasso is fairly similar to SCAD. This example confirms that using a combination of thresholding and shrinkage, which is more flexible than hard thresholding, results in a cleaner and more informative estimate of the sparsity structure.

APPENDIX: PROOFS

We start from a lemma summarizing several earlier results we will use in the proof. The proofs and/or further references for these can be found in Bickel and Levina (2007).

Lemma 1. Under conditions of Theorem 1,

$$\max_i \sum_{j=1}^{p} |\hat{\sigma}_{ij}| 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| < \lambda) = O_P\left(c_0(p)\lambda^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)\lambda^{1-q}\right) \tag{A.1}$$

$$\max_i \sum_{j=1}^{p} |\sigma_{ij}| 1(|\hat{\sigma}_{ij}| < \lambda, |\sigma_{ij}| \geq \lambda) = O_P\left(c_0(p)\lambda^{-q}\sqrt{\frac{\log p}{n}} + c_0(p)\lambda^{1-q}\right) \tag{A.2}$$

$$\max_i \sum_{j=1}^{p} |\hat{\sigma}_{ij} - \sigma_{ij}| 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| \geq \lambda) = O_P\left(c_0(p)\lambda^{-q}\sqrt{\frac{\log p}{n}}\right) \tag{A.3}$$

$$P(\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| > t) \leq C_1 p^2 e^{-nC_2 t^2} + C_3 p e^{-nC_4 t} \tag{A.4}$$

where $t = o(1)$ and C_1, C_2, C_3, C_4 depend only on M.

Proof of Theorem 1. We start from the decomposition

$$\|s_\lambda(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\| \leq \|s_\lambda(\mathbf{\Sigma}) - \mathbf{\Sigma}\| + \|s_\lambda(\hat{\mathbf{\Sigma}}) - s_\lambda(\mathbf{\Sigma})\|. \tag{A.5}$$

For symmetric matrices, the operator norm satisfies (see, for example, Golub and Van Loan (1989)):

$$\| \boldsymbol{A} \| \leq \max_i \sum_j |a_{ij}|. \tag{A.6}$$

That is, the operator norm is bounded by the matrix l_1 or l_∞ norm, which coincide for symmetric matrices. From this point on, we bound all the operator norms by (A.6). For the first term in (A.5), note that by Assumptions (ii) and (iii) that define generalized thresholding,

$$\begin{aligned}\sum_{j=1}^{p} |s_\lambda(\sigma_{ij}) - \sigma_{ij}| &\leq \sum_{j=1}^{p} |\sigma_{ij}| 1(|\sigma_{ij}| \leq \lambda) + \sum_{j=1}^{p} \lambda 1(|\sigma_{ij}| > \lambda) \\ &= \sum_{j=1}^{p} |\sigma_{ij}|^q |\sigma_{ij}|^{1-q} 1(|\sigma_{ij}| \leq \lambda) + \sum_{j=1}^{p} \lambda^q \lambda^{1-q} 1(|\sigma_{ij}| > \lambda) \\ &\leq \lambda^{1-q} \sum_{j=1}^{p} |\sigma_{ij}|^q,\end{aligned}$$

and therefore by (A.6) and the definition (7), the first term in (A.5) is bounded by $\lambda^{1-q} c_0(p)$.

For the second term in (A.5), note that by (i) and (ii),

$$\begin{aligned}|s_\lambda(\hat{\sigma}_{ij}) - s_\lambda(\sigma_{ij})| \leq\ & |\hat{\sigma}_{ij}| 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| < \lambda) \\ &+ |\sigma_{ij}| 1(|\hat{\sigma}_{ij}| < \lambda, |\sigma_{ij}| \geq \lambda) \\ &+ \left(|\hat{\sigma}_{ij} - \sigma_{ij}| + |s_\lambda(\hat{\sigma}_{ij}) - \hat{\sigma}_{ij}| + |s_\lambda(\sigma_{ij}) - \sigma_{ij}|\right) \\ &\quad 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| \geq \lambda)\end{aligned} \tag{A.7}$$

The first three terms in (A.7) are controlled by (A.1), (A.2), and (A.3), respectively. For the fourth term, applying (iii), we have

$$\begin{aligned}&\max_i \sum_{j=1}^{p} |s_\lambda(\hat{\sigma}_{ij}) - \hat{\sigma}_{ij}| 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| \geq \lambda) \\ &\leq \max_i \sum_{j=1}^{p} \lambda^q \lambda^{1-q} 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| \geq \lambda) \\ &\leq \lambda^{1-q} \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q 1(|\sigma_{ij}| \geq \lambda) \leq \lambda^{1-q} c_0(p).\end{aligned}$$

Similarly, for the last term in (A.7) we have

$$\max_i \sum_{j=1}^{p} |s_\lambda(\sigma_{ij}) - \sigma_{ij}| 1(|\hat{\sigma}_{ij}| \geq \lambda, |\sigma_{ij}| \geq \lambda) \leq \lambda^{1-q} c_0(p).$$

Collecting all the terms, we obtain

$$\| s_{\lambda_n}(\hat{\Sigma}) - \Sigma \| = O_P\left(c_0(p)\left(\lambda^{1-q} + \lambda^{-q}\sqrt{\frac{\log p}{n}}\right)\right),$$

and the theorem follows by substituting $\lambda = M'\sqrt{(\log p/n)}$. ■

Proof of Theorem 2. To prove (8), apply (ii) to get

$$\begin{aligned}\{(i,j) : s_\lambda(\hat{\sigma}_{ij}) \neq 0, \sigma_{ij} = 0\} &= \{(i,j) : |\hat{\sigma}_{ij}| > \lambda, \sigma_{ij} = 0\} \\ &\subseteq \{(i,j) : |\hat{\sigma}_{ij} - \sigma_{ij}| > \lambda\}.\end{aligned}$$

Therefore,

$$P\left(\sum_{i,j} 1(s_\lambda(\hat{\sigma}_{ij}) \neq 0, \sigma_{ij} = 0) > 0\right) \leq P(\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| > \lambda). \tag{A.8}$$

Now we apply (A.4). With the choice $\lambda = M'\sqrt{(\log p/n)}$, the first term dominates the second one, so we only need to make

sure $C_1 p^2 e^{-nC_2\lambda^2} \to 0$. Because we can choose M' large enough so that $2 - C_2 M'^2 < 0$, the probability in (A.8) tends to 0.

Similarly, for (9) we have

$$\{(i,j) : s_\lambda(\hat{\sigma}_{ij}) \le 0, \sigma_{ij} > 0 \text{ or } s_\lambda(\hat{\sigma}_{ij}) \ge 0, \sigma_{ij} < 0\} \subseteq \{(i,j) : |\hat{\sigma}_{ij} - \sigma_{ij}| > \tau - \lambda\},$$

and applying the bound (A.4) and the additional condition $\sqrt{n}(\tau - \lambda) \to \infty$ gives

$$P\left(\sum_{i,j} 1(|\hat{\sigma}_{ij} - \sigma_{ij}| \ge \tau - \lambda) > 0\right) \le C_1 p^2 e^{-nC_2(\tau-\lambda)^2} \to 0.$$

■

[Received March 2008. Revised August 2008.]

REFERENCES

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelet Approximations," *Journal of the American Statistical Association,* 96, 939–955.

Bickel, P. J., 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.

——— (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics,* 36(6):2577–2604.

——— (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics,* 36(1):199–227.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics,* 37, 373–384.

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

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

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Pickard, D. (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 57, 301–369.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences of the United States of America,* 95, 14863–14868.

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large Dimensional Sparse Covariance Matrices," *The Annals of Statistics,* 36(6):2717–2756.

Fan, J. (1997), "Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis," *Journal of the Italian Statistical Association,* 6, 131–139.

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.

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics,* 32, 928–961.

Fan, J., Wang, M., and Yao, Q. (2008), "Modelling Multivariate Volatilities via Conditionally Uncorrelated Components," *Journal of the Royal Statistical Society,* Ser. B, 70, 679–702.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics (*Oxford, England), 9, 432–441.

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.

Golub, G. H., and Van Loan, C. F. (1989), *Matrix Computations* (2nd ed.), Baltimore: The John Hopkins University Press.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Botstein, D., and Brown, P. (2000), "Identifying Distinct Sets of Genes With Similar Expression Patterns via Gene Shaving," *Genome Biology,* 1, 1–21.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood," *Biometrika,* 93, 85–98.

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

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.

Krzanowski, W. (1979), "Between-Groups Comparison of Principal Components," *Journal of the American Statistical Association,* 74, 703–707.

Lam, C., and Fan, J. (2008), "Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation," Technical Report, Department of Operations Research and Financial Engineering, Princeton University.

Levina, E., Rothman, A. J., and Zhu, J. (2008), "Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty," *Annals of Applied Statistics,* 2, 245–263.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics,* 2, 494–515.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Ser. B, 58, 267–288.

Wagaman, A. S., and Levina, E. (2007), "Discovering Sparse Covariance Structures with the Isomap," Technical Report 472, University of Michigan.

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

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika,* 94, 19–35.

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