Biometrika (2012), **99**, 3, *pp.* 733–740
© 2012 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/ass025
Advance Access publication 18 June 2012

Miscellanea

Positive definite estimators of large covariance matrices

BY ADAM J. ROTHMAN

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
arothman@umn.edu

SUMMARY

Using convex optimization, we construct a sparse estimator of the covariance matrix that is positive definite and performs well in high-dimensional settings. A lasso-type penalty is used to encourage sparsity and a logarithmic barrier function is used to enforce positive definiteness. Consistency and convergence rate bounds are established as both the number of variables and sample size diverge. An efficient computational algorithm is developed and the merits of the approach are illustrated with simulations and a speech signal classification example.

Some key words: Barrier function; Classification; Convex optimization; High-dimensional data; Sparsity.

1. INTRODUCTION

The problem of estimating a covariance or precision matrix when the number of variables p is greater than the sample size n has attracted significant attention in the last decade; see Pourahmadi (2011) for a detailed review. We focus our attention on estimators that are invariant to permutations of variable labels and are sparse, meaning they introduce zeros as a form of regularization. In addition to being applicable in several multivariate analysis problems, these sparse estimators provide an estimate of the edges in a Gaussian graphical model.

For a matrix M, let $|M|_q = \|\mathrm{vec}(M)\|_q$ denote the q-norm of the vector formed by stacking the columns of M. Let $\|M\|_F \equiv |M|_2$ and $\|M\|$ denote the Frobenius and spectral norms of M respectively. When M is a square matrix, let $|M|$ denote its determinant, $\mathrm{tr}(M)$ denote its trace, and $\varphi_j(M)$ denote its jth largest eigenvalue. Let $M \succ 0$ indicate that M is symmetric and positive definite. Let M^+ denote a diagonal matrix with the same diagonal as M and define $M^- = M - M^+$.

Many sparse estimators of the covariance and precision matrix can be expressed as $\hat{\Psi} = \mathrm{argmin}_{\Psi}\{L(\Psi, S) + P(\Psi)\}$, where S is the sample covariance, L is a loss function, and P is a nondifferentiable penalty function used to encourage sparse solutions. Since inversion tends to destroy sparsity and may introduce noise when p is large, the problems of estimating a sparse covariance matrix and sparse precision matrix are generally considered separately.

To estimate a sparse precision matrix, Banerjee et al. (2008) and Yuan & Lin (2007) proposed the penalized normal likelihood precision matrix estimator

$$\hat{\Sigma}^{-1} = \underset{\Sigma^{-1} \succ 0}{\mathrm{argmin}}\{\mathrm{tr}(\Sigma^{-1}S) - \log|\Sigma^{-1}| + \lambda|\Sigma^{-1-}|_1\}. \tag{1}$$

With these choices of the loss and penalty functions, the optimization in (1) is convex. If $\lambda > 0$, the log-determinant term ensures the existence of a unique global positive definite minimizer, even when $p \geqslant n$. High-dimensional asymptotic analyses (Rothman et al., 2008; Lam & Fan, 2009; Ravikumar et al., 2011) demonstrated the merits of this estimator when $p \geqslant n$ and a fast computational algorithm (Friedman et al., 2008) enabled its application to problems with thousands of variables.

With the goal of estimating a sparse or approximately sparse covariance matrix, several types of elementwise thresholding of the sample covariance matrix have been proposed and analysed. These include hard thresholding (Bickel & Levina, 2008a; El Karoui, 2008), soft thresholding with generalizations (Rothman et al., 2009), and adaptive thresholding (Cai & Liu, 2011) which is able to achieve an optimal rate of convergence in the spectral norm, as shown in a 2010 unpublished technical report by T. Cai and H. Zhou. In general, elementwise thresholding estimators have a very low computational cost and good rates of convergence under high-dimensional asymptotics, but may have negative eigenvalues in finite samples. These negative eigenvalues can be avoided by lasso-type penalized normal likelihood (Lam & Fan, 2009), but the resulting estimator $\hat{\Sigma} = \text{argmin}_{\Sigma \succ 0}\{\text{tr}(\Sigma^{-1}S) - \log|\Sigma^{-1}| + \lambda|\Sigma^{-}|_1\}$ requires nonconvex optimization and is not unique. Recently, Bien & Tibshirani (2011) developed a majorize-minimize algorithm that computes this estimator as a special case.

Our interest is to construct a sparse covariance estimator via convex optimization that performs well in high-dimensional settings and is positive definite in finite samples. Positive definiteness is desirable when the covariance estimator is applied to methods for supervised learning. Many of these methods either require a positive definite covariance estimator, or use optimization that is convex only if the covariance estimator is nonnegative definite, e.g., quadratic discriminant analysis and covariance regularized regression (Witten & Tibshirani, 2009).

2. A positive definite sparse covariance estimator

2·1. *Methodology*

We assume that the sample covariance matrix S is computed from n independent copies of $X = (X_{(1)}, \ldots, X_{(p)})^{\mathrm{T}}$, where $E(X) = 0$ and $E(XX^{\mathrm{T}}) = \Sigma_0$. Let Θ_0 denote the population correlation matrix and let R denote the sample correlation matrix. We propose the correlation matrix estimator

$$\hat{\Theta}_\lambda = \underset{\Theta \succ 0}{\text{argmin}}(\|\Theta - R\|_F^2/2 - \tau \log|\Theta| + \lambda|\Theta^{-}|_1), \qquad (2)$$

where $\lambda \geqslant 0$ is a tuning parameter and $\tau > 0$ is fixed at a small value. The logarithmic barrier term ensures the existence of a positive definite solution, since $-\tau \log|\Theta| = -\tau \sum_{j=1}^{p} \log \varphi_j(\Theta)$ when $\Theta \succ 0$. The lasso-type penalty is used to encourage sparse solutions, analogous to its use in (1).

We propose the covariance matrix estimator $\hat{\Sigma}_\lambda = (S^+)^{1/2}\hat{\Theta}_\lambda(S^+)^{1/2}$. Motivated by Rothman et al. (2008) and Lam & Fan (2009), regularizing on the correlation scale enables us to prove a faster convergence rate bound and produces a covariance estimator $\hat{\Sigma}_\lambda$ that is invariant to scaling of the variables.

To see the effect of the logarithmic barrier term in (2), suppose that $\lambda = 0$. This implies that $\hat{\Theta}_0$ and R have the same eigenvectors and the eigenvalues of $\hat{\Theta}_0$ are an inflation of the eigenvalues of R. Specifically, $\varphi_j(\hat{\Theta}_0) = \varphi_j(R)/2 + \{\varphi_j^2(R) + 4\tau\}^{1/2}/2$ $(j = 1, \ldots, p)$. If $p \geqslant n$ and $j \geqslant n$, then $\varphi_j(R) = 0$, implying that $\varphi_j(\hat{\Theta}_0) = \tau^{1/2}$. Also, one can show that

$$\hat{\Theta}_0^{-1} = \underset{\Theta^{-1} \succ 0}{\text{argmin}}\{\text{tr}(\Theta^{-1}R) - \log|\Theta^{-1}| + \tau\|\Theta^{-1}\|_F^2/2\}. \qquad (3)$$

The objective function in (3), with R replaced by S, is equivalent to the negative normal loglikelihood with a ridge penalty, for which Witten & Tibshirani (2009) derived the minimizer in closed form. Thus, $\hat{\Theta}_0^{-1}$ can be viewed as a ridge penalized likelihood inverse correlation matrix estimator.

If $\tau = 0$ and the feasible set in (2) is expanded to $\{\Theta : \Theta = \Theta^{\mathrm{T}}\}$, then $\hat{\Theta}_\lambda$ is obtained by elementwise soft thresholding (Donoho & Johnstone, 1994; Rothman et al., 2009) of R^{-} at λ. If $\tau > 0$ and $\lambda > 0$, a closed-form solution is unavailable and we use an efficient iterative algorithm described in § 3·1.

2·2. *Asymptotic analysis*

Let s denote the number of nonzero off-diagonal entries in Σ_0. We present convergence rate bounds for the proposed covariance estimator, where n, p, and s are allowed to diverge. Proofs are given in the Appendix.

Assumption 1. For all p, $0 < \kappa_1 \leqslant \varphi_{\min}(\Sigma_0) \leqslant \varphi_{\max}(\Sigma_0) \leqslant \kappa_2 < \infty$, where κ_1 and κ_2 are constants.

Assumption 2. For all $j = 1, \ldots, p$, we have $E\{\exp(tX_{(j)}^2)\} \leqslant C_1 < \infty$ for $0 < |t| < t_0$.

Unlike entry-wise thresholding estimators of the covariance matrix, the lower bound on $\varphi_{\min}(\Sigma_0)$ in Assumption 1 is needed because of the log-determinant term in the objective function, which ensures that the proposed estimator is positive definite in finite samples. Assumption 2 holds, for example, if $X_{(1)}, \ldots, X_{(p)}$ are Gaussian.

THEOREM 1. *Let K_1 be a sufficiently large constant and suppose that Assumptions 1 and 2 hold. If $\lambda = K_1(n^{-1}\log p)^{1/2}$, $\tau = O\{(n^{-1}s\log p)^{1/2}\|\Theta_0^{-1}\|_F^{-1}\}$, and $(s+1)\log p = o(n)$, then* (i) $\|\hat{\Theta}_\lambda - \Theta_0\|_F = O_P\{(n^{-1}s\log p)^{1/2}\}$ *and* (ii) $\|\hat{\Sigma}_\lambda - \Sigma_0\| = O_P[\{n^{-1}(s+1)\log p\}^{1/2}]$.

The bounds in Theorem 1 coincide with those established for a particular local minimizer of the lasso penalized normal likelihood (Lam & Fan, 2009). We obtain slower convergence rate bounds by relaxing Assumption 2.

Assumption 3. For some $\alpha \geqslant 2$, it is the case that $E(|X_{(j)}|^{2\alpha}) \leqslant C_2 < \infty$, for all $j = 1, \ldots, p$.

Assumption 3 was used by Bickel & Levina (2008a) in the context of covariance thresholding.

THEOREM 2. *Let K_2 be a sufficiently large constant and suppose that Assumptions 1 and 3 hold. If $\lambda = K_2(n^{-1}p^{4/\alpha})^{1/2}$, $\tau = O\{(n^{-1}sp^{4/\alpha})^{1/2}\|\Theta_0^{-1}\|_F^{-1}\}$, and $(s+1)p^{4/\alpha} = o(n)$, then* (i) $\|\hat{\Theta}_\lambda - \Theta_0\|_F = O_P\{(n^{-1}sp^{4/\alpha})^{1/2}\}$ *and* (ii) $\|\hat{\Sigma}_\lambda - \Sigma_0\| = O_P[\{n^{-1}(s+1)p^{4/\alpha}\}^{1/2}]$.

3. COMPUTATION

3·1. *Algorithm*

We develop an iterative algorithm to compute the solution to (2) and more generally to compute the positive definite minimizer of the objective function h, defined as

$$h(\Sigma) = \frac{1}{2}\|\Sigma - Q\|_F^2 - \tau\log|\Sigma| + \lambda|\Sigma^-|_1,$$

where Q is a symmetric $p \times p$ matrix with positive diagonal entries. The algorithm is similar to the graphical lasso algorithm (Friedman et al., 2008) and the graphical elastic net algorithm, developed in a 2010 Stanford University PhD thesis by G. Allen, both of which compute a sparse precision matrix estimate.

Since h is strictly convex, there exists a unique zero subgradient of h at the global minimizer $\hat{\Sigma} \succ 0$, which follows from Theorem 3.4.3 of Bazaraa et al. (2006). Express this zero subgradient as a matrix equation $\hat{\Sigma} - Q - \tau\hat{\Sigma}^{-1} + \lambda\hat{\Gamma} = 0$, where the $p \times p$ matrix $\hat{\Gamma}$ has entries $\hat{\gamma}_{ij} = 1(i \neq j, \hat{\sigma}_{ij} \neq 0)\text{sign}(\hat{\sigma}_{ij}) + u_{ij}1(i \neq j, \hat{\sigma}_{ij} = 0)$, $\hat{\sigma}_{ij}$ is the (i, j)th entry of $\hat{\Sigma}$, and $-1 \leqslant u_{ij} \leqslant 1$. The algorithm solves the matrix equation by iteratively solving

$$0 = \hat{\Sigma} - Q - \tau\hat{\Omega} + \lambda\hat{\Gamma}, \quad 0 = \hat{\Sigma}\hat{\Omega} - I. \tag{4}$$

Let (Σ, Ω) denote our current iterate. Following the approach introduced by Banerjee et al. (2008) and employed by Friedman et al. (2008), we partition (4) as

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{\mathrm{T}} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{12}^{\mathrm{T}} & q_{22} \end{pmatrix} - \tau\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^{\mathrm{T}} & \omega_{22} \end{pmatrix} + \lambda\begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}^{\mathrm{T}} & 0 \end{pmatrix} = 0, \tag{5}$$

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{\mathrm{T}} & \sigma_{22} \end{pmatrix}\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^{\mathrm{T}} & \omega_{22} \end{pmatrix} - \begin{pmatrix} I & 0 \\ 0^{\mathrm{T}} & 1 \end{pmatrix} = 0. \tag{6}$$

We continue by describing how to update the final row/column of our iterate (Σ, Ω). From the final row/column of (5) and (6), we obtain

$$\Sigma_{12} - Q_{12} - \tau\Omega_{12} + \lambda\Gamma_{12} = 0, \tag{7}$$

$$\sigma_{22} - q_{22} - \tau\omega_{22} = 0, \tag{8}$$

$$\Omega_{11}\Sigma_{12} + \sigma_{22}\Omega_{12} = 0, \tag{9}$$

$$\Sigma_{12}^{\mathrm{T}}\Omega_{12} + \sigma_{22}\omega_{22} = 1. \tag{10}$$

From (8), we compute the update $\tilde{\sigma}_{22} = q_{22} + \tau\omega_{22}$. From (9), we substitute $\Omega_{12} = -\Omega_{11}\Sigma_{12}/\sigma_{22}$ and $\tilde{\sigma}_{22}$ into (7), yielding $\Sigma_{12} - Q_{12} + \tau\Omega_{11}\Sigma_{12}/\tilde{\sigma}_{22} + \lambda\Gamma_{12} = (I + \tau\Omega_{11}/\tilde{\sigma}_{22})\,\Sigma_{12} - Q_{12} + \lambda\Gamma_{12} = 0$. This zero subgradient is equivalent to the zero subgradient of the lasso penalized regression

$$\tilde{\Sigma}_{12} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2}\beta^{\mathrm{T}} \left(I + \frac{\tau}{\tilde{\sigma}_{22}}\Omega_{11} \right) \beta - \beta^{\mathrm{T}} Q_{12} + \lambda\|\beta\|_1 \right\}, \tag{11}$$

which we solve with cyclical coordinate descent (Fu, 1998; Friedman et al., 2007, 2008). If $\tau = 0$, the solution to (11) is obtained by soft thresholding the final row of Q at λ, excluding its diagonal element. In general, if $\Omega \succ 0$, then $\tilde{\sigma}_{22} > 0$, implying that (11) is strictly convex, admitting a unique global minimizer $\tilde{\Sigma}_{12}$.

We substitute $\tilde{\Sigma}_{12}, \tilde{\sigma}_{22}$ into (9) and (10) and obtain

$$\tilde{\Omega}_{12} = -\Omega_{11}\tilde{\Sigma}_{12}/\tilde{\sigma}_{22}, \quad \tilde{\omega}_{22} = (1 - \tilde{\Sigma}_{12}^{\mathrm{T}}\tilde{\Omega}_{12})/\tilde{\sigma}_{22}. \tag{12}$$

We have described how to update the final row of our iterate (Σ, Ω) with $\tilde{\Sigma}_{12}, \tilde{\sigma}_{22}, \tilde{\Omega}_{12}$, and $\tilde{\omega}_{22}$. The algorithm continues by permuting the rows and columns so the jth row/column is updated, for $j = 1, \ldots, p$. Let $M_{-j,j}$ denote the vector formed from the jth column of M with its jth element removed, and let $M_{-j,-j}$ denote a matrix formed by removing the jth row and column from M. In the description of the steps in Algorithm 1, we update symmetric matrices and it is understood that $M_{-j,j} = M_{j,-j}$.

Algorithm 1. Given input Q, λ, and τ, initialize $(\Sigma^{(0)}, \Omega^{(0)})$. Perform Steps 1–3 for $j = 1, \ldots, p$ and repeat until convergence.

Step 1. Compute $\sigma_{jj}^{(k+1)} = q_{jj} + \tau\omega_{jj}^{(k)}$ and solve the lasso penalized regression

$$\Sigma_{j,-j}^{(k+1)} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2}\beta^{\mathrm{T}} \left(I + \frac{\tau}{\sigma_{jj}^{(k+1)}}\Omega_{-j,-j}^{(k)} \right) \beta - \beta^{\mathrm{T}} Q_{-j,j} + \lambda\|\beta\|_1 \right\}.$$

Step 2. Compute $\Omega_{j,-j}^{(k+1)} = -\Omega_{-j,-j}^{(k)}\Sigma_{j,-j}^{(k+1)}/\sigma_{jj}^{(k+1)}$.

Step 3. Compute $\omega_{jj}^{(k+1)} = (1 - \Sigma_{j,-j}^{(k+1)T}\Omega_{j,-j}^{(k+1)})/\sigma_{jj}^{(k+1)}$.

Remark 1. If Algorithm 1 is initialized with $\Omega^{(0)} \succ 0$ and Q is symmetric with positive diagonal entries, then all future iterates $\Omega^{(k)}$ are positive definite.

We prove Remark 1 in the Appendix. An initial positive definite iterate of $\Sigma^{(0)} = Q^{+}$ and $\Omega^{(0)} = (\Sigma^{(0)})^{-1}$ is sensible regardless of the rank of Q. Although all iterates $\Omega^{(k)}$ are positive definite, small values of k may produce an iterate $\Sigma^{(k)}$ which is indefinite; in practice $\Sigma^{(k)}$ becomes positive definite after only a few iterations. Following Friedman et al. (2008), we use the criterion $|\Sigma^{(k+1)} - \Sigma^{(k)}|_1 < \epsilon|Q^{-}|_1$ to determine convergence. The convergence tolerance ϵ should be chosen relative to τ. For example, if $\tau = 10^{-4}$, we recommend setting $\epsilon = 10^{-7}$. In every numerical example we have tried, Algorithm 1 converged to a solution point that solved the zero subgradient equations; however, we are unable to prove that Algorithm 1 converges to the global minimizer.

Algorithm 1 has computational complexity $O(p^3)$, which is the same order as the graphical lasso algorithm of Friedman et al. (2008) for sparse precision matrix estimation.

Table 1. *Averages and standard errors, in parentheses, of the spectral norm and Frobenius norm losses, computed from* 500 *realizations*

		Spectral norm loss			Frobenius norm loss		
Model	p	Samp. Cov.	Soft Est.	Prop. Est.	Samp. Cov.	Soft Est.	Prop. Est.
Σ_1	30	2·05 (0·01)	0·85 (0·00)	0·85 (0·00)	4·29 (0·01)	2·39 (0·01)	2·39 (0·01)
	100	4·86 (0·01)	0·96 (0·00)	0·96 (0·00)	14·09 (0·02)	4·93 (0·01)	4·93 (0·01)
	200	8·14 (0·02)	1·00 (0·00)	1·00 (0·00)	28·07 (0·02)	7·34 (0·01)	7·34 (0·01)
Σ_2	30	2·59 (0·03)	2·40 (0·02)	2·40 (0·02)	4·37 (0·02)	3·83 (0·02)	3·83 (0·02)
	100	6·97 (0·06)	5·51 (0·03)	5·48 (0·03)	14·29 (0·04)	10·77 (0·05)	10·73 (0·04)
	200	11·52 (0·06)	6·42 (0·02)	6·38 (0·02)	28·23 (0·05)	16·90 (0·05)	16·76 (0·04)

Samp. Cov., sample covariance; Soft Est., soft thresholding estimator; Prop. Est., proposed estimator.

3·2. *Tuning parameter selection*

We select the tuning parameter λ using a method introduced by Bickel & Levina (2008b) and analysed by Bickel & Levina (2008a). The data are randomly partitioned N times into a training set of size n_1 and a validation set of size n_2, where $n_2 = \lfloor n/\log n \rfloor$ and $n_1 = n - n_2$. The selected tuning parameter is $\hat{\lambda} = \mathrm{argmin}_{\lambda} \sum_{m=1}^{N} \|\hat{\Sigma}_{\lambda}^{(m,n_1)} - S^{(m,n_2)}\|_F^2$, where $\hat{\Sigma}_{\lambda}^{(m,n_1)}$ is the covariance estimator, with penalty parameter λ, computed with the training set of the mth split and $S^{(m,n_2)}$ is the sample covariance computed with the validation set of the mth split.

When computing the proposed correlation matrix estimator $\hat{\Theta}_{\lambda}$, we found that selecting $\tau = 10^{-4}$ leads to a stable solution of Algorithm 1 and performs well in simulations.

4. Numerical studies

4·1. *Simulation*

We compared the performance of the proposed covariance estimator with a soft thresholding covariance estimator (Rothman et al., 2009) formed by soft thresholding the off-diagonal elements of the sample correlation matrix and then rescaling by sample standard deviations. The sample covariance matrix was also included as a benchmark.

We used two models for the population covariance matrix. The first is a tridiagonal model where the population covariance matrix Σ_1 has entries $\sigma_{1ij} = 0{\cdot}4\ 1(|i-j|=1) + 1(i=j)$. The second is an overlapping block-diagonal model, where the population covariance matrix Σ_2 has unit diagonal entries and off-diagonal entries defined in the following way. The indices $1,\ldots,p$ are partitioned into K ordered blocks of equal size. We set $\sigma_{2ij} = 0{\cdot}4$ if i and j are in the same block, or if i and j are in adjacent blocks and $\min(i,j)$ is the maximum index of a block. All other off-diagonal entries of Σ_2 are zero. We used $K=3$, 5 and 10 for $p=30$, 100 and 200 respectively.

Using these covariance models, we generated $n=50$ realizations of independent p-variate normal random vectors with mean zero. We selected the tuning parameters for both methods with the random splitting scheme described in § 3·2 using ten random splits. The resulting covariance estimators were compared to the population covariance matrix using the spectral and Frobenius norms of their difference. This procedure was repeated 500 times.

For both covariance models, the averages and standard errors of the spectral and Frobenius norm losses from these 500 replications are reported in Table 1. The proposed estimator and soft thresholding have very similar performance in all settings for both covariance models. For the tridiagonal model, the soft thresholding estimator was always positive definite; however, for the overlapping block-diagonal model, it was positive definite only in 210 for $p=100$ and 91 for $p=200$ of the 500 realizations.

We have performed simulations from other covariance models and found that sparser models yield a higher percentage of positive definite realizations for soft thresholding, and in terms of matrix norm losses, the proposed estimator, which is always positive definite, performs as well as soft thresholding.

4·2. *Speech signal classification example*

We evaluated the performance of the proposed covariance estimator when it is used in quadratic discriminant analysis to discriminate between healthy individuals and individuals with Parkinson's disease, using features extracted from speech signals. The data, introduced by Little et al. (2009), were obtained from the UCI machine learning data repository. There is a total of 195 speech signals, of which 147 are from individuals with Parkinson's disease. Each signal had 22 numerical features extracted. Although there are multiple signals recorded from the same individual, we treated the 195 numerical feature vectors as realizations of independent random vectors.

To assess high-dimensional classification performance, we randomly partitioned the data 500 times into 65 training cases and 130 testing cases, where 49 of the training cases and 98 of the testing cases were from individuals with Parkinson's disease. Letting x_i denote the feature vector for the ith testing case, we classify x_i as healthy, $\hat{k}_i = 1$, or as having Parkinson's disease, $\hat{k}_i = 2$, with the quadratic discriminant rule $\hat{k}_i = \text{argmax}_k\{\log|\hat{\Sigma}_k^{-1}|/2 - (x_i - \hat{\mu}_k)^{\mathrm{T}}\hat{\Sigma}_k^{-1}(x_i - \hat{\mu}_k)/2 + \log\hat{\pi}_k\}$, where $\hat{\mu}_k$, $\hat{\pi}_k$, and $\hat{\Sigma}_k$ are the class k training sample mean, class k proportion of training cases, and the class k training covariance estimate, respectively. The proposed covariance estimator, the soft thresholding covariance estimator, and a diagonal estimator S^+ were used in place of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. Tuning parameters were selected using the method described in § 3·2 with 10 random splits.

The average percentage of misclassified testing cases, based on the 500 random partitions, was 21·8 for the proposed covariance estimator and 29·1 for the diagonal covariance estimator. The soft thresholding estimator was inapplicable in 120 of the 500 random partitions since it did not give positive definite matrices for both $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. When it was applicable, its average percentage of misclassified testing cases was 22·6.

Acknowledgement

The author thanks the editor, associate editor, and referees for helpful suggestions, as well as Tiefeng Jiang for a helpful discussion. The author's research is supported in part by a grant from the U.S. National Science Foundation.

Appendix

Proofs of the main results

Let $C_3, \ldots, C_9$ denote positive constants. To prove Theorems 1 and 2 we use the following Lemma.

Lemma A1. *Suppose that Assumption 1 holds, $\lambda \leqslant \varepsilon s^{-1/2}/16$, and $\tau \leqslant \varepsilon\|\Theta_0^{-1}\|_F^{-1}/8$. Then, for ε positive and sufficiently small,* $\max_{i\neq j}|r_{ij} - \theta_{0ij}| \leqslant \lambda$ *implies* $\|\hat{\Theta}_\lambda - \Theta_0\|_F \leqslant \varepsilon$.

Proof. Let f denote the objective function in (2), $f(\Theta) = \text{tr}(\Theta^2)/2 - \text{tr}(\Theta R) - \tau\log|\Theta| + \lambda|\Theta^-|_1$. Analogous to Rothman et al. (2008), define the set $\mathcal{B}_\varepsilon = \{\Delta : \Delta = \Delta^{\mathrm{T}}, \|\Delta\|_F = \varepsilon\}$. By convexity of f and the fact that $\hat{\Theta}_\lambda$ is its global minimizer, $\inf\{f(\Theta_0 + \Delta) : \Delta \in \mathcal{B}_\varepsilon\} > f(\Theta_0)$ implies $\|\hat{\Theta}_\lambda - \Theta_0\|_F \leqslant \varepsilon$. Equivalently, $\inf\{G(\Delta) : \Delta \in \mathcal{B}_\varepsilon\} > 0$ implies $\|\hat{\Theta}_\lambda - \Theta_0\|_F \leqslant \varepsilon$, where $G(\Delta) = f(\Theta_0 + \Delta) - f(\Theta_0)$. A lower bound for $G(\Delta)$ for $\Delta \in \mathcal{B}_\varepsilon$ is obtained by writing

$$\begin{aligned}
G(\Delta) = {} & \text{tr}\{(\Theta_0 + \Delta)^2\}/2 - \text{tr}\{(\Theta_0 + \Delta)R\} - \tau\log|\Theta_0 + \Delta| + \lambda|\Theta_0^- + \Delta^-|_1 \\
& - \text{tr}(\Theta_0^2)/2 + \text{tr}(\Theta_0 R) + \tau\log|\Theta_0| - \lambda|\Theta_0^-|_1 \\
= {} & \text{tr}(\Delta\Theta_0 - \Delta R + \Delta^2/2) - \tau(\log|\Theta_0 + \Delta| - \log|\Theta_0|) + \lambda(|\Theta_0^- + \Delta^-|_1 - |\Theta_0^-|_1) \\
= {} & \|\Delta\|_F^2/2 - \tau(\log|\Theta_0 + \Delta| - \log|\Theta_0|) + \text{tr}\{\Delta(\Theta_0 - R)\} + \lambda(|\Theta_0^- + \Delta^-|_1 - |\Theta_0^-|_1) \\
= {} & \|\Delta\|_F^2/2 + T_1 + T_2 + T_3.
\end{aligned}$$

Assumption 1 implies that $\varphi_{\min}(\Theta_0) \geqslant C_3 > 0$. From this, a lower bound for $T_1 = -\tau(\log|\Theta_0 + \Delta| - \log|\Theta_0|)$ is obtained by applying Taylor's theorem to $g(t) = \log|\Theta_0 + t\Delta|$, similarly to

Rothman et al. (2008). We have

$$T_1 = -\tau \operatorname{tr}(\Delta\Theta_0^{-1}) + \tau \operatorname{vec}(\Delta)^{\mathrm{T}} \left\{ \int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} \, \mathrm{d}v \right\} \operatorname{vec}(\Delta) = T_{11} + T_{12}.$$

Following a similar argument used in Rothman et al. (2008),

$$\varphi_{\min} \left\{ \int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} \, \mathrm{d}v \right\} \geqslant \frac{1}{2} \min_{0 \leqslant u \leqslant 1} \{\varphi_{\min}^2 (\Theta_0 + u\Delta)^{-1}\} \geqslant 0,$$

which implies that $T_{12} \geqslant 0$. This and the Cauchy–Schwartz inequality imply that

$$T_1 \geqslant T_{11} \geqslant -\tau |\operatorname{tr}(\Delta\Theta_0^{-1})| \geqslant -\tau \|\Delta\|_F \|\Theta_0^{-1}\|_F. \tag{A1}$$

We have

$$T_2 \geqslant -|\operatorname{tr}\{\Delta(\Theta_0 - R)\}| = -\left| \sum_{i \neq j} \delta_{ij}(\theta_{0ij} - r_{ij}) \right| \geqslant -\max_{i \neq j} |r_{ij} - \theta_{0ij}| |\Delta^-|_1. \tag{A2}$$

A lower bound for T_3 is obtained with an argument identical to that in Rothman et al. (2008). Define the index set $\mathcal{S} = \{(i, j) : \theta_{0ij} \neq 0,\ i \neq j\}$ and for a matrix M, let $M_{\mathcal{S}}$ denote a matrix with (i, j)th entry $m_{ij} 1\{(i, j) \in \mathcal{S}\}$. We have $T_3 = \lambda(|\Theta_0^- + \Delta^-|_1 - |\Theta_0^-|_1) = \lambda(|\Theta_{0\mathcal{S}}^- + \Delta_{\mathcal{S}}^-|_1 + |\Delta_{\mathcal{S}^c}^-|_1 - |\Theta_{0\mathcal{S}}^-|_1) \geqslant \lambda\{|\Theta_{0\mathcal{S}}^- + \Delta_{\mathcal{S}}^-|_1 + |\Delta_{\mathcal{S}^c}^-|_1 - (|\Theta_{0\mathcal{S}}^- + \Delta_{\mathcal{S}}^-|_1 + |\Delta_{\mathcal{S}}^-|_1)\} = \lambda(|\Delta_{\mathcal{S}^c}^-|_1 - |\Delta_{\mathcal{S}}^-|_1)$. This lower bound, (A1), and (A2) imply that $G(\Delta) \geqslant \|\Delta\|_F^2/2 - \tau \|\Delta\|_F \|\Theta_0^{-1}\|_F - \max_{i \neq j} |r_{ij} - \theta_{0ij}| |\Delta^-|_1 + \lambda(|\Delta_{\mathcal{S}^c}^-|_1 - |\Delta_{\mathcal{S}}^-|_1)$. Since $\max_{i \neq j} |r_{ij} - \theta_{0ij}| \leqslant \lambda$,

$$\begin{aligned} G(\Delta) &\geqslant \|\Delta\|_F^2/2 - \tau \|\Delta\|_F \|\Theta_0^{-1}\|_F - \lambda |\Delta^-|_1 + \lambda(|\Delta_{\mathcal{S}^c}^-|_1 - |\Delta_{\mathcal{S}}^-|_1) \\ &= \|\Delta\|_F^2/2 - \tau \|\Delta\|_F \|\Theta_0^{-1}\|_F - \lambda(|\Delta_{\mathcal{S}^c}^-|_1 + |\Delta_{\mathcal{S}}^-|_1) + \lambda(|\Delta_{\mathcal{S}^c}^-|_1 - |\Delta_{\mathcal{S}}^-|_1) \\ &= \|\Delta\|_F^2/2 - \tau \|\Delta\|_F \|\Theta_0^{-1}\|_F - 2\lambda |\Delta_{\mathcal{S}}^-|_1. \end{aligned} \tag{A3}$$

By a standard vector norm inequality, $|\Delta_{\mathcal{S}}^-|_1 \leqslant s^{1/2} |\Delta^-|_2 \leqslant s^{1/2} \|\Delta\|_F$. Since $\lambda \leqslant \varepsilon s^{-1/2}/16$, $\tau \leqslant \varepsilon \|\Theta_0^{-1}\|_F^{-1}/8$, and $\|\Delta\|_F = \varepsilon$ when $\Delta \in \mathcal{B}_\varepsilon$, from (A3), $G(\Delta) \geqslant \varepsilon^2/2 - \varepsilon^2/8 - \varepsilon^2/8 = \varepsilon^2/4 > 0$. □

Proof of Theorem 1. Since Assumptions 1 and 2 hold, a fact employed by Rothman et al. (2008) is that for ν sufficiently small, $\operatorname{pr}(\max_{i \neq j} |r_{ij} - \theta_{0ij}| > \nu) \leqslant p^2 C_4 \exp(-C_5 n \nu^2)$. Since $(s + 1) \log p = o(n)$, we apply this bound and Lemma A1 with $\varepsilon = K_3 (n^{-1} s \log p)^{1/2}$ and $\lambda = K_3 (n^{-1} \log p)^{1/2}/16$, to obtain $\operatorname{pr}\{\|\hat{\Theta}_\lambda - \Theta_0\|_F \leqslant K_3 (n^{-1} s \log p)^{1/2}\} \geqslant \operatorname{pr}\{\max_{i \neq j} |r_{ij} - \theta_{0ij}| \leqslant K_3 (n^{-1} \log p)^{1/2}/16\} \geqslant 1 - C_6 p^{2 - C_7 K_3^2}$, which is arbitrarily close to one by making K_3 sufficiently large. The second claim follows from the same argument used in the proof of Theorem 2 of Rothman et al. (2008). □

Proof of Theorem 2. Under Assumptions 1 and 3, one can modify a result of Bickel & Levina (2008a) and show that for ν sufficiently small, $\operatorname{pr}(\max_{i \neq j} |r_{ij} - \theta_{0ij}| > \nu) \leqslant p^2 C_8 n^{-\alpha/2} \nu^{-\alpha}$. Since $(s + 1) p^{4/\alpha} = o(n)$, we apply this bound and Lemma A1 with $\varepsilon = K_4 (n^{-1} s p^{4/\alpha})^{1/2}$ and $\lambda = K_4 (n^{-1} p^{4/\alpha})^{1/2}/16$, to obtain $\operatorname{pr}\{\|\hat{\Theta}_\lambda - \Theta_0\|_F \leqslant K_4 (n^{-1} s p^{4/\alpha})^{1/2}\} \geqslant \operatorname{pr}\{\max_{i \neq j} |r_{ij} - \theta_{0ij}| \leqslant K_4 (n^{-1} p^{4/\alpha})^{1/2}/16\} \geqslant 1 - C_9 K_4^{-\alpha}$, which is arbitrarily close to one by making K_4 sufficiently large. The second claim follows from the same argument used in the proof of Theorem 2 of Rothman et al. (2008). □

Proof of Remark 1. Assume that our current iterate $\Omega \succ 0$. This implies that $\Omega_{11} \succ 0$ and hence $|\Omega_{11}| > 0$. A single iteration of the algorithm updates the final row and column of Ω. We will show the new iterate $\tilde{\Omega} \succ 0$. Since $|\tilde{\Omega}| = |\Omega_{11}|(\tilde{\omega}_{22} - \tilde{\Omega}_{12}^{\mathrm{T}} \Omega_{11}^{-1} \tilde{\Omega}_{12})$ and $\Omega_{11} \succ 0$, it suffices to show that $\tilde{\omega}_{22} - \tilde{\Omega}_{12}^{\mathrm{T}} \Omega_{11}^{-1} \tilde{\Omega}_{12} > 0$. From (12), $\tilde{\omega}_{22} = \tilde{\sigma}_{22}^{-1} (1 + \tilde{\Sigma}_{12}^{\mathrm{T}} \Omega_{11} \tilde{\Sigma}_{12} / \tilde{\sigma}_{22})$ and

$$\tilde{\omega}_{22} - \tilde{\Omega}_{12}^{\mathrm{T}} \Omega_{11}^{-1} \tilde{\Omega}_{12} = \frac{1 + \tilde{\Sigma}_{12}^{\mathrm{T}} \Omega_{11} \tilde{\Sigma}_{12} / \tilde{\sigma}_{22}}{\tilde{\sigma}_{22}} - \frac{\tilde{\Sigma}_{12}^{\mathrm{T}} \Omega_{11} \Omega_{11}^{-1} \Omega_{11} \tilde{\Sigma}_{12}}{\tilde{\sigma}_{22}^2} = \frac{1}{\tilde{\sigma}_{22}} = \frac{1}{q_{22} + \tau \omega_{22}} > 0. \qquad \square$$

References

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.* **9**, 485–516.

Bazaraa, M. S., Sherali, H. D. & Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*. New Jersey: Wiley. 3rd edition.

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

Bickel, P. J. & Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.

Bien, J. & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–20.

Cai, T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 672–84.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717–56.

Friedman, J., Hastie, T. J. & Tibshirani, R. J. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–32.

Friedman, J., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Fu, W. (1998). Penalized regressions: The bridge versus the lasso. *J. Comp. Graph. Statist.* **7**, 397–416.

Lam, C. & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Ann. Statist.* **37**, 4254–78.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J. & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **56**, 1015–22.

Pourahmadi, M. (2011). Modeling covariance matrices: The GLM and regularization perspectives. *Statist. Sci.* **26**, 369–87.

Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–80.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

Rothman, A. J., Levina, E. & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Am. Statist. Assoc.* **104**, 177–86.

Witten, D. M. & Tibshirani, R. J. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. R. Statist. Soc.* B **71**, 615–36.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

[*Received June* 2011. *Revised February* 2012]