Sparse Estimation of High-dimensional Correlation Matrices

Ying Cui, Chenlei Leng and Defeng Sun [1]

Abstract

Estimating covariations of variables for high dimensional data is important for understanding their relations. Recent years have seen several attempts to estimate covariance matrices with sparsity constraints. A new convex optimization formulation for estimating correlation matrices, which are scale invariant, is proposed as opposed to covariance matrices. The constrained optimization problem is solved by the accelerated proximal gradient algorithm with fast convergence rate. An adaptive version of this approach is also discussed. Simulation results and an analysis of a cardiovascular microarray confirm its performance and usefulness.

Keywords: Accelerated proximal gradient; Correlation matrix; High-dimensionality; Positive definiteness; Sparsity.

1. Introduction

The covariance matrix plays a fundamental role and is a pivotal quantity in statistical analysis, for example in linear regression and multivariate analysis. Assume that we are given observations $x_i \in \mathbb{R}^p,\ i = 1, ..., n$ from the same distribution F. A simple way to estimate the population covariance matrix, which is assumed to be non-degenerate, is via the empirical covariance matrix

$$\Sigma_n = (\hat{\sigma}_{ij})_{1 \le i,j \le p} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T,$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ is the sample mean. When the dimensionality p is high compared to the sample size n, however, the sample covariance matrix becomes less useful or even degenerate if $p > n$.

To overcome this difficulty, a host of approaches have been proposed to estimate the covariance under the assumption that it is sparse or approximately so. Bickel and Levina (2008a,b) proposed to band or to threshold the entries of the sample covariance matrix. Rothman et al. (2009) studied more flexible thresholding rules. Cai and Liu (2011) advocated to adaptively threshold the entries according to their individual variability. Cai and Yuan (2012) applied blocked thresholding for adaptive estimation. A major drawback of these approaches is that the estimated covariance matrix is not guaranteed to be positive definite, a minimum requirement for a matrix to be a

[1] Cui is graduate student and Sun is Professor in the Department of Mathematics and Risk Management Institute, National University of Singapore. Leng is Professor in the Department of Statistics, University of Warwick. Leng is also affiliated with the Department of Statistics and Applied Probability, National University of Singapore. We are grateful to two reviewers for their constructive comments. Corresponding author: Chenlei Leng (Email: C.Leng@warwick.ac.uk).

covariance matrix. Lam and Fan (2009) outlined a unified analysis of various early approaches for estimating sparse matrices. Cai and Zhou (2012) discussed optimal rates of convergence for estimating sparse covariance matrices under various assumptions.

To simultaneously achieve sparsity and positive definiteness, Bien and Tibshirani (2011) applied the penalized likelihood method under Gaussianity, but their objective function is non-convex. Lin (2010) provided an algorithm for obtaining the local optimal solution of this formulation. Rothman (2012) suggested to minimize the squared Frobenius distance between the sample covariance matrix and the estimate by adding a sparsity penalty, and a log-determinant barrier that guarantees the positive definiteness. Xue et al. (2012) studied a constrained optimization formulation that enforces more explicitly the positive definite constraint. More specifically, they proposed to solve the following optimization problem

$$\tilde{\Sigma} = \arg\min_{\Sigma} \ \|\Sigma - \Sigma_n\|_F^2 + \rho|\Sigma|_1, \quad \text{such that } \Sigma \succeq \varepsilon I, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $|\cdot|_1$ is the element-wise ℓ_1-norm for sparsity (Tibshirani (1996)), and $\Sigma \succeq \varepsilon I$ means that $\Sigma - \varepsilon I$ is semipositive definite for a small positive constant ε. Thus, Σ itself is guaranteed positive definite.

There are potential problems with estimating the covariance matrix. The covariance matrix is not scale invariant. Should one scale the variables in x_i differently, $\tilde{\Sigma}$ would be different no matter how λ is chosen. A common practice is to normalize the variables to have zero mean and unit variances before the analysis, effectively making Σ_n a sample correlation matrix. However, in estimating Σ, this important prior information is ignored and Σ is treated as a usual covariance matrix as in Rothman (2012) and Xue et al. (2012). As we show in the theoretical study, this incurs p additional parameters in the diagonal of Σ that slows down the rate of convergence in terms of the spectral and the Frobenius norm.

To overcome the limitations elaborated above, we propose a new approach termed Sparse Estimation of the Correlation matrix (SEC). Instead of targeting a high-dimensional covariance matrix, we estimate a sparse correlation matrix by forcing the diagonal entries of the estimate to be unity. In addition, we formulate a general approach that adaptively penalizes the correlations according to the empirical ones.

Because estimating a correlation is notably much more challenging than estimating a covariance matrix, and in practice Σ_n may have large dimension so that it costs much to achieve a desirable solution, a new and efficient algorithm is highly needed. In this paper we take a dual approach to solve this constrained optimization by the accelerated proximal gradient algorithm (APG). As shown by Nesterov (1983), APG is a fast gradient method with the attractive $O(1/k^2)$ complexity of the function value, where k is the iteration number. The resulting estimate is guaranteed to be positive definite and a correlation matrix. Comparing to the estimation of a covariance matrix, the new estimate enjoys a faster rate of convergence. After this paper was completed, we became aware of Liu et al. (2014) where they used a similar ℓ_1 penalized formulation as ours and a similar algorithm as in Xue et al. (2012). As demonstrated in the simulation study, however, our algorithm is usually faster and the performance of the algorithm in Xue et al. (2012) and Liu et al. (2014) depends on a parameter usually difficult to tune.

The rest of the paper is organized as follows. In Section 2, we present the SEC method and discuss a weighted SEC scheme for adaptively estimating the correlations. In Section 3, we give some preliminaries on Moreau-Yosida regularization which will be used to design the algorithm later on. Then we introduce the framework of the APG algorithm to solve the dual problem of (3). Section 4 presents the statistical property of our SEC model. Section 5 reports the numerical performance and Section 6 draws the conclusion. All proofs are deferred to the appendix.

2. Sparse Estimation of Correlation

Let $R_n = D_n^{-1}\Sigma_n D_n^{-1}$ be the empirical correlation matrix, where D_n is the diagonal matrix with the square roots of the diagonal elements of Σ_n. We estimate the sparse correlation matrix by solving

$$\hat{R} = \arg\min_{R} \ \frac{1}{2}\|R - R_n\|_F^2 + \rho|R|_1, \quad \text{such that } R \succeq \varepsilon I,\ R_{jj} = 1,\ j = 1, ..., p. \tag{2}$$

The major difference between this approach and that of Xue et al. (2012) is that we add hard constraints $R_{jj} = 1,\ j = 1, ..., p$ to the formulation, making sure effectively that the correlation matrix is the main quantity of interest. In this work, we set $\varepsilon = 10^{-5}$. We note that the choice of ε makes little difference as long as it is small enough. In practice, we recommend to use an ε such that $\log_{10} \varepsilon \in [-8, -5]$.

Inspired by the adaptive lasso (Zou (2006)), we also consider a more general SEC problem with the weighted ℓ_1 penalty as $\rho|W \circ R|_1$. Here $\circ$ denotes the Hadamard product, i.e. $W \circ R = (W_{ij}R_{ij})_{p\times p}$. We aim to solve a general optimization problem as

$$\begin{aligned} \hat{R} = \quad \arg\min_{R} \quad & \tfrac{1}{2}\|R - R_n\|_F^2 + \rho|W \circ R|_1 \\ \text{s.t.} \quad & R_{ij} = b_{ij}, \quad (i,j) \in \Omega, \\ & R \succeq \varepsilon I. \end{aligned} \tag{3}$$

For the equality constraints in (2), $\Omega = \{(j,j) : j = 1, ..., p\}$, and $b_{ij} = 1$. To adaptively penalize the entries in R, one possible choice of the weight matrix W is $(\frac{1}{|(R_n)_{ij}|})_{p\times p}$, the componentwise inverse of the sample correlation matrix. The idea is to apply a larger amount of penalization to smaller empirical correlations.

Computationally, $(R_n)_{ij}$ may be close to zero sometimes if the true (ij)th correlation is close to zero, so that W_{ij} will be close to ∞. As a result, this will cause a great difficulty for computation since the constraint $R \succeq \varepsilon I$ strictly prohibits us from eliminating those infinity penalized entries from the objective function. Therefore, we naturally introduce an index set Ω to allow some fixed entries for $\hat{R}$. These entries include the diagonal part of a correlation matrix as well as the infinity ℓ_1 penalty coefficient, where the latter one will definitely induce zero components for the resulting estimator $\hat{R}$. That is, $\Omega = \{(i,j), (R_n)_{ij} = 0 \text{ or } i = j\}$, where we include (i,j) in Ω if $|(R_n)_{ij}| < \delta$. In theory, we need $\delta = o(\sqrt{\frac{\log p}{n}})$ and in practice we set $\delta = 10^{-8}$. Hence we take

$$W_{ij} = \begin{cases} \frac{1}{|(R_n)_{ij}|} & \text{for } (i,j) \notin \Omega, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad b_{ij} = \begin{cases} 0 & \text{for } |(R_n)_{ij}| < \delta, \\ 1 & \text{for } i = j. \end{cases}$$

In fact our algorithm also works when Ω includes other prior information. For example, some variables are correlated with fixed correlations, which frequently occur in finance. One can easily see that if W is a matrix with all entries equal to 1 in Ω^c, Ω is exactly the set of diagonal elements, (3) will reduce to (2), the sparse correlation estimation problem. Moreover, if Ω is an empty set, then (3) provides the sparse covariance matrix studied by Xue et al. (2012).

3. The Algorithm

Before stating our APG algorithm, first we show that the alternating direction method of multipliers (ADMM) used by Xue et al. (2012) can be applied to this model as well. Note that problem (3) can be formulated equivalently as follows:

$$\begin{aligned} \hat{R} = \quad & \arg\min_{R} \quad \tfrac{1}{2}\|R - R_n\|_F^2 + \rho|W \circ R|_1 \\ & \text{s.t.} \quad R_{ij} = b_{ij}, \quad (i,j) \in \Omega, \\ & \qquad\quad Y \succeq \varepsilon I, \\ & \qquad\quad R = Y. \end{aligned} \tag{4}$$

The augmented Lagrangian function of (4) is

$$L_\mu(R, Y, \Lambda) := \{\frac{1}{2}\|R - R_n\|_F^2 + \rho|W \circ R|_1 + \langle R - Y, \Lambda\rangle + \frac{1}{2\mu}\|R - Y\|_F^2,\ R_{ij} = b_{ij},\ (i,j) \in \Omega\ ;\ Y \succeq \varepsilon I\},$$

where Λ is the Lagrangian multiplier and μ is the penalty parameter. Therefore the k-th iteration for ADMM is

$$\begin{aligned} R^{k+1} &= \arg\min\{L_\mu(R, Y^k, \Lambda^k) | R_{ij} = b_{ij},\ (i,j) \in \Omega\}, \\ Y^{k+1} &= \arg\min\{L_\mu(R^{k+1}, Y, \Lambda^k) | Y \succeq \varepsilon I\}, \\ \Lambda^{k+1} &= \Lambda^k + \tfrac{1}{\mu}(R^{K+1} - Y^{k+1}). \end{aligned}$$

It is easy to see that the first two subproblems have closed form solution, which is respectively given by

$$R_{ij}^{k+1} = \begin{cases} b_{ij} & \text{for } (i,j) \in \Omega, \\ \text{sgn}((\tfrac{1}{\mu}Y^k + R_n - \Lambda^k)_{ij})(|(\tfrac{1}{\mu}Y^k + R_n - \Lambda^k)_{ij}| - \rho W_{ij})_+/(1 + \tfrac{1}{\mu}) & \text{for } (i,j) \in \Omega^c, \end{cases}$$

and

$$Y^{k+1} = \Pi_{S_+^n}(\mu\Lambda^k + R^{k+1}).$$

By the method stated above, one can solve the SEC problem. However, the numerical performance is not as good as the following dual approach. Detailed results are presented in Section 5.

3.1. The Moreau-Yosida regularization

In this subsection, we review some concepts and properties concerning Moreau-Yosida regularization. These results will be used to design the APG algorithm later on.

Let $f : \mathcal{X} \to (-\infty, +\infty]$ be a closed proper convex function, where $\mathcal{X}$ is a real finite dimensional Euclidean space with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. The Moreau (Moreau (1965))-Yosida (Yosida (1971)) regularization of f associated with a given parameter $\rho > 0$ is defined by

$$\psi_f^\rho(x) := \min_{y \in \mathcal{X}} \{ f(y) + \frac{1}{2\rho} \|y - x\|^2 \}, \quad x \in \mathcal{X}. \tag{5}$$

The unique minimizer of (5), written as $P_f^\rho(x)$, is called the proximal point mapping associated with f. There are some important properties for ψ_f^ρ and P_f^ρ which we summarize below without proof, see, e.g. Hiriart-Urruty and Lemaréchal (1993).

Proposition 1. *Let $f : \mathcal{X} \to (-\infty, +\infty]$ be a closed proper convex function, ψ_f^ρ be the Moreau-Yosida regularization of f, and P_f^ρ be the associated proximal point mapping. Then the following properties hold.*
(1) ψ_f^ρ is continuously differentiable with gradient given by

$$\nabla \psi_f^\rho(x) = \frac{1}{\rho}(x - P_f^\rho(x)).$$

Furthermore, $\nabla \psi_f^\rho$ is globally Lipschitz continuous with modulus $1/\rho$.
(2) For any $x, x' \in \mathcal{X}$, one has

$$\langle P_f^\rho(x) - P_f^\rho(x'), x - x' \rangle \geq \|P_f^\rho(x) - P_f^\rho(x')\|^2.$$

This implies that $P_f^\rho(\cdot)$ is globally Lipschitz continuous with modulus 1 by the Cauchy-Schwarz inequality.

The above proposition shows that for any closed proper convex function, not necessarily continuous, its Moreau-Yosida regularization is continuously differentiable with Lipschitz continuous gradient. This is of particular interest when the constrained optimization problem involves a nonsmooth term in the objective function like SEC, where directly applying the APG algorithm to the primal is not valid.

3.2. A dual accelerated proximal gradient framework

In this section we take a dual approach to solve (3) by the accelerated proximal gradient algorithm. Before that we briefly review the idea of this algorithm.

Let $\mathcal{X}$ be a real finite dimensional Euclidean space with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Let $g : \mathcal{X} \to (-\infty, +\infty]$ be a smooth convex function on an open set containing $\text{dom}\, g = \{x : g(x) < \infty\}$, and $h : \mathcal{X} \to (-\infty, +\infty]$ be a proper, convex, lower semicontinuous but possibly nonsmooth function. We further assume ∇g is Lipschitz continuous on dom g with modulus L, i.e.

$$\|\nabla g(x) - \nabla g(y)\|_* \leq L\|x - y\|, \ \forall x, y \in \text{dom}\, g, \tag{6}$$

where $\| \cdot \|_*$ stands for the dual norm of $\| \cdot \|$.

Consider a class of nonsmooth convex optimization problem

$$\min_x \ f(x) \equiv g(x) + h(x). \tag{7}$$

The classical proximal gradient algorithm solves (7) iteratively by minimizing its quadratic approximation at each iteration point as

$$\begin{aligned} x_{k+1} &= \arg\min_x \{g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 + h(x)\} \\ &= \arg\min_x \{\frac{L}{2}\|x - (x_k - \frac{1}{L}\nabla g(x_k))\|^2 + h(x)\}. \end{aligned} \tag{8}$$

It can be shown that for such algorithm, it satisfies $f(x_k) - \inf f \leq O(1/k)$ for any iteration number k, see, e.g. Nesterov and Nesterov (2004). Based on this idea, Nesterov (1983, 2005), Tseng (2008), and others proposed various kinds of accelerated proximal gradient (APG) methods, which have an attractive $O(1/k^2)$ complexity. Recently, Beck and Teboulle (2009) studied a fast shrinkage-thresholding algorithm (abbreviated FISTA), which is a special case of the classical accelerated proximal gradient method. They suggested to solve the following sequence iteratively:

$$\begin{cases} x_{k+1} = \arg\min_x \{g(y_k) + \langle \nabla g(y_k), x - y_k \rangle + \frac{L}{2}\|x - y_k\|^2 + h(x)\}, \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k). \end{cases} \tag{9}$$

The initial algorithm in Beck and Teboulle (2009) is designed for the vector case, where $\mathcal{X} = \mathcal{R}^n$. Later on many researchers extended the idea to matrix programming, see, e.g. Toh and Yun (2010). In this paper we shall adopt FISTA in matrix form to solve the dual problem of (3).

Let $\mathcal{S}^p$ and $\mathcal{S}^p_+$ be the space of $p \times p$ symmetric matrices and the cone of positive semidefinite matrices in $\mathcal{S}^p$ respectively. Denote $\| \cdot \|_F$ as the Frobenius norm induced by the standard trace inner product $\langle \cdot, \cdot \rangle$ in $\mathcal{S}^p$, i.e.

$\langle A, B\rangle = \mathrm{trace}(A^T B)$ for $A, B \in \mathcal{S}^p$. The Lagrangian dual problem of (3) is given by

$$\max_{Z\succeq 0} \inf_{\substack{R_{ij}=b_{ij},\\ (i,j)\in\Omega}} L(R,Z) := \frac{1}{2}\|R-R_n\|_F^2 + \rho|W\circ R|_1 - \langle Z, R-\varepsilon I\rangle.$$

Denote $X(Z) = Z + R_n$. Then we can solve the inner problem analytically as follows:

$$\begin{aligned} g(Z) \ : \ &= -\inf_{\substack{R_{ij}=b_{ij},\\ (i,j)\in\Omega}} \{\frac{1}{2}\|R-R_n\|_F^2 + \rho|W\circ R|_1 - \langle Z, R-\varepsilon I\rangle\} \\ &= -\inf_{\substack{R_{ij}=b_{ij},\\ (i,j)\in\Omega}} \{\frac{1}{2}\|R-X(Z)\|_F^2 + \rho|W\circ R|_1\} + \frac{1}{2}\|X(Z)-\varepsilon I\|_F^2 - \frac{1}{2}\|R_n-\varepsilon I\|_F^2 \\ &= -\tfrac{1}{2}\|S(Z)-X(Z)\|_F^2 - \rho|W\circ S(Z)|_1 + \tfrac{1}{2}\|X(Z)-\varepsilon I\|_F^2 - \tfrac{1}{2}\|R_n-\varepsilon I\|_F^2, \end{aligned}$$

where

$$S(Z)_{ij} = \begin{cases} b_{ij} & \text{for } (i,j)\in\Omega, \\ \mathrm{sgn}(X(Z)_{ij})\max\{|X(Z)_{ij}| - \rho W_{ij}, 0\} & \text{for } (i,j)\in\Omega^c \end{cases} \tag{10}$$

is the ℓ_1-norm soft thresholding in Ω^c.

Therefore the dual problem of (3) is

$$(D) \quad \min_Z \ f(Z) := g(Z) + \delta_{\mathcal{S}_+^p}(Z), \tag{11}$$

where $\delta_{\mathcal{S}_+^p}(Z)$ is the indicator function of $\mathcal{S}_+^p$.

From part (1) of Proposition 2 we know that the function $g(Z)$ is continuously differentiable, for which the gradient is given by

$$\nabla g(Z) = S(Z) - \varepsilon I. \tag{12}$$

Moreover, part (2) of this proposition shows that $\nabla g(Z)$ is globally Lipschitz continuous with modulus 1. Therefore we can apply the framework of FISTA to solve (3). Note that in the $(k+1)$th iteration of FISTA, one needs to obtain the solution of the subproblem

$$Y_{k+1} = \arg\min_{Z\succeq 0}\{g(Z_k) + \langle\nabla g(Z_k), Z-Z_k\rangle + \frac{1}{2}\|Z-Z_k\|^2\}. \tag{13}$$

Substituting (12) in, it is easy to derive that the unique solution of (13) can be expressed as

$$Y_{k+1} = \Pi_{\mathcal{S}_+^p}[Z_k - (S(Z_k) - \varepsilon I)],$$

where $\Pi_{\mathcal{S}_+^p}(\cdot)$ denotes the projection onto $\mathcal{S}_+^p$. It is well-known that if $X \in \mathcal{S}^p$ has the eigen-decomposition $X = \sum_{i=1}^{p} \lambda_i v_i v_i^T$, then $\Pi_{\mathcal{S}_+^p}(X) = \sum_{i=1}^{p} \max\{\lambda_i, 0\} v_i v_i^T$.

Hence we are ready to propose an accelerated proximal gradient algorithm to solve the SEC problem.

```
The Accelerated Proximal Gradient Algorithm.

Choose an initial point $Y_1,\ Z_1,\ t_1 = 1$. Set $k := 1$. Iterate until convergence:

Step 1. Compute $S(Z_k)$ given by (10). Then take $Y_{k+1} = \Pi_{\mathcal{S}^p_+}(Z_k - (S(Z_k) - \varepsilon I))$.

Step 2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

Step 3. compute $Z_{k+1} = Y_{k+1} + \frac{t_k-1}{t_{k+1}}(Y_{k+1} - Y_k)$.
```

There is a complete complexity theory of the APG algorithm to solve problems like (7). This work is first done by Nesterov (1983) with $h(x) = 0$, and extended by Beck and Teboulle (2009) with $p = 1$ and dom $h = \mathcal{R}^m$. More recently, Tseng (2008) presented a unified framework for general APG methods solving problems like (7) on a real linear space, and a similar complexity result was provided.

The following theorem states the $O(1/k^2)$ complexity of the APG algorithm which solves the dual problem of SEC. In fact this is the best complexity one can possibly have under the convex black box oracle, which means the only information we can get about function g at each iteration point x_k is the value $g(x_k)$ and $\nabla g(x_k)$. Since the proof is a straightforward extension to the matrix case of Beck and Teboulle (2009) [Theorem 4.4], we omit it here.

Theorem 1. *Assume f is defined in (11) and $\{Z_k\}$ are generated by the APG algorithm. Then for any optimal solution Z^* of $\min\limits_{Z} f(Z)$, we have*

$$f(Z_k) - f(Z^*) \leq \frac{2\|Z_0 - Z^*\|^2}{(k+1)^2}, \quad \forall\ k \geq 1.$$

3.3. A postprocessing: calibrating the estimator with equality and semidefinite constraints

Due to the large p, small n nature of the problem, the estimation error between our estimator $\hat{R}$ and the true correlation matrix R is quite large (one may find the details about the estimation performance in Section 5). Hence it is often of less use and inefficient for the algorithm to achieve high accuracy in practice. In fact our APG algorithm is able to compute the whole solution path of the adaptive SEC model with a modest scaled weight matrix W, say $\|W\|_\infty \leq 100$, in several minutes when n is up to 1000 with 10^{-8} relative accuracy, but we find that the performance of the resulting estimator is not much better than a less accurate one. One problem of generating merely a coarse solution is that the solution is not feasible, so it fails to be a correlation matrix with fixed entries in Ω. To bridge the gap, we need a postprocessing to calibrate the estimator into the feasible set.

We denote the APG solution of SEC model as $\hat{R}_{apg}$. If the Ω set only contains diagonal entries and those zero off-diagonal entries, we could apply a simple correction step as follows: Let λ be the smallest eigenvalue of $\hat{R}_{apg}$, write

$$\bar{R} = \hat{R}_{apg} - (\varepsilon - \max\{0, -\lambda\})I$$

and

$$D = \begin{pmatrix} \frac{\bar{R}_{11}}{b_{11}-\varepsilon} & & & \\ & \frac{\bar{R}_{22}}{b_{22}-\varepsilon} & & \\ & & \ddots & \\ & & & \frac{\bar{R}_{pp}}{b_{pp}-\varepsilon} \end{pmatrix},$$

then the corrected solution is given by

$$\hat{R}_c = D^{-\frac{1}{2}} \bar{R} D^{-\frac{1}{2}} + \varepsilon I.$$

In this way the resulting estimator $\hat{R}_c$ satisfies both the equality constraints and positive definiteness. Note that the above procedure cannot keep nonzero off-diagonal entries, it is not suitable if additional prior information is given in the Ω set. In that case, we can alternatively use the fast semismooth Newton-CG algorithm proposed by Qi and Sun (2006), which aims at computing the nearest correlation matrix with fixed diagonal and off diagonal elements.

4. Statistical Properties

Let $\Sigma^0 = (\sigma^0_{ij})_{1 \le i,j \le p}$ and $R^0 = (r^0_{ij})_{1 \le i,j \le p}$ be the population covariance and correlation matrix respectively. We denote the non-diagonal support of R^0 as $A_0 = \{(i,j) : \; i \neq j, \; r^0_{ij} \neq 0\}$ and its cardinality as s.

After an estimate $\hat{R}$ is available, the covariance matrix can be estimated as $\hat{\Sigma} = D_n \hat{R} D_n$. As in Cai and Liu (2011) and Xue et al. (2012), we discuss two general distributions that have interesting tail probabilities. The theorem is formally established for the estimator in (2).

Theorem 2. *Assume that the true correlation matrix R^0 is positive definite and that the marginal variances of the variables are bounded away from zero. Furthermore, assume that the variables follow either of the following distributions.*

a. (Exponential-type tails) Suppose that for some $\eta > 0$,

$$E\{\exp(tX_{ij}^2)\} \le K_1 \text{ for all } |t| < \eta,$$

for all i, j, and $\log(p) \le n$.

b. (Polynomial-type tails) Suppose that for some $\gamma, c > 0, p \le cn^\gamma$, and for some $\tau > 0$

$$E\{|X_{ij}|^{4\gamma+\tau+4}\} \le K_1$$

for all i, j.

If either a or b holds and $\rho = M\sqrt{\frac{\log p}{n}}$ for some constant M, we have for the estimator defined in (2),

$$\|\hat{R} - R^0\|_F^2 = O_p\Big(s\frac{\log p}{n}\Big).$$

Since R^0 is positive definite, we can specify a small constant ε to bound its smallest eigenvalue. In this paper, we use $\varepsilon = 10^{-5}$, which is realistically small enough. Note that for estimating the covariance matrix, the rate of convergence is $O_p((s+p)\log p/n)$ in the squared Frobenius norm (Cai and Liu (2011); Cai and Zhou (2012); Xue et al. (2012)). Thus, when the sample correlation is used for estimating a covariance matrix, the rate of convergence will be $O_p((s+p)\log p/n)$ if additional constraints on the diagonals are not enforced, as in Xue et al. (2012) and Rothman (2012). Our estimator of R^0 enjoys a fast rate of convergence by a difference of the order at least of p/n. This is due to the fact that the diagonals of R^0 need no estimation. In principal, we can allow $p \gg n$ in our approach as long as $n \gg s\log p$ to obtain the convergence result in terms of the Frobenius norm and spectral norm for estimating R^0. On the other hand, for estimating R^0 using the approach in Xue et al. (2012), in practice, p can at most be comparable to n to have a meaningful convergence result. More discussion of this difference can be found in Rothman et al. (2008) and Lam and Fan (2009). This rate can also be obtained via a penalized likelihood method in principle, if a multivariate Gaussian distribution is assumed (Lam and Fan (2009)). A major problem, however, is that the solution for this formulation is only a local optimum, because the likelihood-based loss function is nonconvex (Bien and Tibshirani (2011)).

Now, we discuss the adaptive estimator defined in (3).

Corollary 1. *Assume that either of the tail conditions in Theorem 2 holds. If $\delta = o(\sqrt{\frac{\log p}{n}})$, $\min_{(i,j)\in A_0} r_{ij}^0 \gg \sqrt{\frac{\log p}{n}}$, and we take $\rho = M\sqrt{\frac{\log p}{n}}\min_{(i,j)\in A_0}(R_n)_{ij}$ for some constant M, then*

$$\|\hat{R} - R^0\|_F^2 = O_p\Big(s\frac{\log p}{n}\Big).$$

Furthermore, if $v_p(R^0) \gg \sqrt{s\frac{\log p}{n}}$, where $v_1(R) \geq v_2(R) \geq \cdots \geq v_p(R)$ are eigenvalues of R, then we have: i) $\hat{r}_{ij} = 0$ for $(i,j) \in A_0^C$ and $i \neq j$, and ii) $\hat{r}_{ij} \neq 0$ for $(i,j) \in A_0$ with probability tending to one.

We note that the additional assumptions made for this corollary may not be tight. The assumption $\delta = o(\sqrt{\frac{\log p}{n}})$ is to make sure that we only set very small entries in $\hat{R}$ as zero. The assumption $\min_{(i,j)\in A_0} r_{ij}^0 \gg \sqrt{\frac{\log p}{n}}$ specifies the signal strength. In Lemma 2 in the appendix, we show that in R_n, the zero correlations will be estimated at a maximum magnitude $\sqrt{\frac{\log p}{n}}$, and nonzero correlations at a minimum magnitude much larger than $\sqrt{\frac{\log p}{n}}$ with probability tending to one. Thus, a small δ ensures that only some of the zero correlations would be fixed as zeros in their estimates. Although setting δ does not affect the theoretical results in Corollary 1 as long as it is small enough, as discussed before, having a small δ brings computational benefit. The eigenvalue assumption

$v_p(R^0) \gg \sqrt{s\frac{\log p}{n}}$ is to ensure that the adaptive estimator in (3) is estimated as a positive definite matrix without the positive definite assumption $R \succeq \varepsilon I$ with probability tending to one. The second part of the corollary states that the weighted estimator in (3) estimates the sparsity pattern of R^0 consistently, which is attractive from a model selection perspective.

5. Simulation and Data Analysis

5.1. Simulation

In this part we demonstrate the numerical performance of our SEC method in (3) by the accelerated proximal gradient algorithm. We use R_p, R_d and *gap* to denote the relative primal feasibility, dual feasibility and gap of primal-dual respectively, i.e.

$$R_p = \frac{\max\{\varepsilon - \lambda_1(S), 0\}}{1 + \|S\|_F}, \;\; R_d = \frac{\|\Pi_{\mathcal{S}^n_+}(Z - (S(Z) - \varepsilon I))\|_F}{1 + \|Z\|_F}, \;\; \mathit{gap} = \frac{|\mathit{pobj} - \mathit{dobj}|}{\mathit{1} + |\mathit{pobj}| + |\mathit{dobj}|},$$

where $\lambda_1(S)$ is the minimal eigenvalue of S, *pobj* and *dobj* are the respective values of the primal and dual objective function. We stop the APG algorithm when

$$\max\{R_P, R_D, \mathit{gap}\} \leq \mathit{5} \times \mathit{10}^{-\mathit{5}}.$$

Then we apply the postprocessing step as described in Section 3.2 to calibrate the solution as a correlation matrix with fixed entries in Ω.

We will consider the following three correlation models throughout our simulation study.

Example 1: (Banded matrix with ordering). $r^0_{ij} = (1 - \frac{|i-j|}{10})_+$.

Example 2: (Block diagonal matrix). Let $K = p/20$, and i_k denote the maximum index in I_k, then

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

Example 3: (Approximately sparse matrix). AR(1), where $r^0_{ij} = 0.3^{|i-j|}$.

The first two examples have been used previously by Bickel and Levina (2008b), Rothman (2012) and Xue et al. (2012). As opposed to the sparse models in Example 1 and 2, the third example is a non-sparse model. This example is meant to shed some light on the approximation power of various methods designed for estimating sparse matrices when the true matrix is dense.

We generate $n = 50$ independent p-variate samples for $p = 100$ or 500, and generate $n = 100$ samples for $p = 1000$ from the normal distribution $X \sim N(0_p, (r^0_{ij})_{p \times p})$. Then we standardize the variables with zero mean and unit variance such that what we obtain is the sample correlation matrix R_n.

Before showing the estimation performance of our adaptive SEC method, first we use these three examples to illustrate why we choose the APG algorithm to solve this model.

Note that the computational cost of the ADMM algorithm discussed at the beginning of Section 3 is essentially the same as our APG algorithm within each iteration, which is dominated by one eigenvalue decomposition. Generally speaking, there are two main reasons that we prefer the APG algorithm here. One is its attractive complexity stated in Theorem 1, which provides a theoretical guarantee for the iteration numbers before reaching the prescribed accuracy. And this property fails to hold for the ADMM algorithm. The other one is that the APG algorithm is free of tuning parameters, while there is an unknown parameter μ in the augmented Lagrangian function. And whether μ is suitable or not will largely influence the performance of ADMM in the sense of CPU time and iteration numbers. To illustrate more clearly about this part, we run the solution path for $\rho = \{0.01, 0.02, \cdots, 1\}$ with different choice of μ for all examples, and compare the results with the APG algorithm. The stopping criterion for ADMM is essentially of the same spirit with the APG algorithm, that is, we stop it when the primal feasibility, the dual feasibility and the duality gap are within the tolerance simultaneously. The number of iterations (iter) and CPU time (time, in seconds) are listed in Table 1 for different settings of $(p \mid n)$.

		Example 1		Example 1		Example 2		Example 2		Example 3		Example 3	
algo.	par.	$(100 \mid 50)$		$(1000 \mid 100)$		$(100 \mid 50)$		$(1000 \mid 100)$		$(100 \mid 50)$		$(1000 \mid 100)$	
	(μ)	iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
APG	-	271	1.1	1986	808	877	67	996	421	169	0.7	680	286
ADMM	0.1	3939	6.9	**6154**	**2657**	3693	390	3713	2440	2271	6.3	2647	1818
	0.5	981	2.9	7140	3083	**1359**	**104**	**1767**	**743**	598	1.9	**995**	**448**
	1	**667**	**2.0**	12197	5145	1882	150	2660	1160	408	1.3	1374	599
	2	715	2.2	18845	7964	3270	266	4808	2115	**380**	**1.3**	2355	1032

Table 1: Performance of the ADMM and APG algorithm.

Table 1 shows that for all the test instances, our APG algorithm substantially outperforms the ADMM in terms of computing time. For example, the ADMM takes at least three times more time to solve Example 1 with $p = 1000$ and $n = 100$ than our APG. Also, we can observe clearly that the performance of the ADMM is over sensitive to the parameter μ and no single value seems to fit for all scenarios. In model 1 with $(p, n) = (100, 50)$, $\mu = 1$ appears to be the best choice, but under this selection the ADMM performs much slower compared with $\mu = 0.1$ for the same model with $(p, n) = (1000, 100)$. This inconsistency may give rise to great difficulties in tuning the parameter μ, especially in many real applications. Therefore, the APG algorithm, which enjoys the optimal complexity in theory combined with simple implementation in practice, becomes a much more natural choice for the adaptive SEC method, and we employ this algorithm in the rest of our numerical experiment.

Next we concentrate on the estimation performance of our model. We choose two kinds of weight matrix W: one is the matrix with all entries equal to 1 in Ω_c, which represents the classical ℓ_1 penalty, and the other one is $W_{ij} = \frac{1}{|R_{ij}|}$ for $(i, j) \in \Omega_c$, the adaptive penalty. We compare the performance of our two estimators (SEC, Adaptive

SEC) with the covariance estimator (Hard thresholding, Xue's estimator) given by Bickel and Levina (2008a) and Xue et al. (2012) respectively. In order to make a fair comparison, the covariance estimators are normalized to have unit diagonal entries. For all estimators, we apply five-fold cross validation to select the optimal ρ by minimizing $\|\hat{R} - R_n\|_F^2$, where R_n is the sample correlation matrix of the fold of the data set not included in estimating $\hat{R}$. The effectiveness of V-fold cross validation is proved by Bickel and Levina (2008a).

We evaluate the performance of all estimators by the average relative error of Frobenius norm (relerr F-norm) and spectral norm (relerr S-norm) between the true correlation matrix $R^0 = (r_{ij}^0)_{p\times p}$ and the estimators. That is, the relative error is defined as $\|\hat{R} - R^0\|_*/\|R^0\|_*$ for some norm $\|\cdot\|_*$. Besides we test the ability of recovering the sparse correlation matrix by the true positive rate (TPR) together with the false positive rate (FPR), defined as respectively

$$\text{TPR} = \frac{\#\{(i,j), \hat{R}_{ij} \neq 0, R_{ij}^0 \neq 0\}}{\#\{(i,j), R_{ij}^0 \neq 0\}},$$

$$\text{FPR} = \frac{\#\{(i,j), \hat{R}_{ij} \neq 0, R_{ij}^0 = 0\}}{\#\{(i,j), R_{ij}^0 = 0\}},$$

as in Rothman et al. (2008, 2009).

Here we also report the estimator for which the sparsity structure is given (Prior information), that is, we first let $\hat{R}_{ij} = 0$ if the true correlation matrix $r_{ij}^0 = 0$ and then project them as a correlation matrix. Since this problem is much easier than SEC for which we have to find zero elements and estimate other entries simultaneously, we regard it as a benchmark to analyze our estimation quality.

$p \mid n$	Estimator	relerr (F-norm)	relerr (S-norm)	FPR	TPR
100|50	Sample	53.41% (0.07)	69.80% (0.09)	-	-
	Hard thresholding	29.28% (0.06)	25.38% (0.04)	0.0 (0.0)	68.7 (0.0)
	Xue's estimator	32.65% (0.06)	40.41% (0.04)	33.2 (0.1)	89.5 (0.0)
	Adaptive Xue's estimator	28.50% (0.07)	33.79% (0.04)	13.3 (0.0)	84.2 (0.0)
	SEC	32.52% (0.06)	40.21% (0.04)	32.5 (0.0)	89.7 (0.0)
	Adaptive SEC	27.86% (0.04)	31.70% (0.02)	10.5 (0.0)	82.9 (0.0)
	Prior information	25.18% (0.10)	28.55% (0.04)	0.0 (0.0)	100.0 (0.0)
500|50	Sample	122.6% (0.09)	240.4% (0.14)	-	-
	Hard thresholding	35.71% (0.06)	34.11% (0.04)	0.1 (0.0)	89.5 (0.0)
	Xue's estimator	42.74% (0.05)	53.16% (0.02)	10.4 (0.0)	59.7 (0.0)
	Adaptive Xue's estimator	36.17% (0.07)	45.52% (0.04)	6.2 (0.0)	79.1 (0.0)
	SEC	42.41% (0.05)	53.00% (0.02)	10.5 (0.0)	82.3 (0.0)
	Adaptive SEC	35.39% (0.06)	44.78% (0.03)	4.5 (0.0)	78.3 (0.0)
	Prior information	30.48% (0.15)	37.95% (0.03)	0.0 (0.0)	100.0 (0.0)
1000|100	Sample	122.4% (0.09)	252.2% (0.14)	-	-
	Hard thresholding	25.67% (0.05)	29.32% (0.04)	0.0 (0.0)	69.6 (0.0)
	Xue's estimator	33.96% (0.10)	43.41% (0.04)	7.5 (0.0)	87.0 (0.0)
	Adaptive Xue's estimator	26.75% (0.07)	34.10% (0.04)	4.7 (0.0)	82.8 (0.0)
	SEC	33.94% (0.10)	43.57% (0.04)	7.6 (0.0)	87.2 (0.0)
	Adaptive SEC	25.31% (0.07)	33.50% (0.04)	2.9 (0.0)	86.1 (0.0)
	Prior information	22.46% (0.10)	28.87% (0.02)	0.0 (0.0)	100.0 (0.0)

Table 2: Average (standard error) performance of the estimators for Example 1 over 100 replications.

$p \mid n$	Estimator	relerr(F-norm)	relerr(S-norm)	FPR	TPR
100|50	Sample	66.33% (0.05)	67.38% (0.08)	-	-
	Hard thresholding	51.10% (0.08)	40.22% (0.04)	2.19 (0.1)	84.62 (0.0)
	Xue's estimator	47.07% (0.07)	49.37% (0.05)	30.2 (0.0)	97.3 (0.0)
	Adaptive Xue's estimator	44.93% (0.09)	43.63% (0.04)	26.1 (0.0)	96.8 (0.0)
	SEC	47.05% (0.07)	49.29% (0.05)	29.2 (0.0)	96.8 (0.0)
	Adaptive SEC	44.75% (0.08)	44.08% (0.04)	20.8 (0.0)	95.8 (0.0)
	Prior information	26.89% (0.06)	28.63% (0.05)	0.0 (0.0)	100.0 (0.0)
500|100	Sample	151.8% (0.06)	225.0% (0.10)	-	-
	Hard thresholding	36.88% (0.07)	38.96% (0.04)	0.4 (0.1)	91.3 (0.0)
	Xue's estimator	62.57% (0.06)	67.95% (0.03)	16.3 (0.0)	94.2 (0.0)
	Adaptive Xue's estimator	55.70% (0.07)	62.76% (0.04)	13.3 (0.0)	93.5 (0.0)
	SEC	60.08% (0.06)	67.39% (0.03)	17.4 (0.0)	95.0 (0.0)
	Adaptive SEC	55.00% (0.07)	62.43% (0.06)	12.4 (0.0)	94.9 (0.0)
	Prior information	28.94% (0.07)	37.68% (0.04)	0.0 (0.0)	100.0 (0.0)
1000|100	Sample	151.7% (0.04)	239.0% (0.09)	-	-
	Hard thresholding	41.68% (0.09)	44.46% (0.06)	0.3 (0.1)	87.9 (0.0)
	Xue's estimator	47.95% (0.06)	55.42% (0.03)	10.4 (0.0)	99.7 (0.0)
	Adaptive Xue's estimator	41.27% (0.07)	53.72% (0.04)	8.6 (0.0)	98.8 (0.0)
	SEC	47.79% (0.06)	55.23% (0.03)	11.1 (0.0)	99.7 (0.0)
	Adaptive SEC	39.34% (0.07)	48.64% (0.04)	5.2 (0.0)	99.2 (0.0)
	Prior information	20.38% (0.07)	28.39% (0.03)	0.0 (0.0)	100.0 (0.0)

Table 3: Average (standard error) performance of the estimators for Example 2 over 100 replications.

$p \mid n$	estimator	relerr(F-norm)	relerr (S-norm)	FPR	TPR
100|50	Sample	129.7% (0.02)	252.9% (0.03)	-	-
	Hard thresholding	40.43% (0.03)	46.97% (0.04)	-	-
	Xue's estimator	37.63% (0.03)	46.36% (0.01)	-	-
	Adaptive Xue's estimator	37.52% (0.06)	46.85% (0.04)	-	-
	SEC	37.63% (0.03)	46.36% (0.01)	-	-
	Adaptive SEC	37.52% (0.04)	47.31% (0.02)	-	-
500|50	Sample	291.6% (0.02)	891.5% (0.05)	-	-
	Hard thresholding	40.63% (0.06)	47.06% (0.04)	-	-
	Xue's estimator	39.57% (0.01)	48.08% (0.04)	-	-
	Adaptive Xue's estimator	39.83% (0.06)	48.50% (0.03)	-	-
	SEC	39.57% (0.01)	48.08% (0.04)	-	-
	Adaptive SEC	39.71% (0.06)	48.43% (0.03)	-	-
1000|100	Sample	290.1% (0.01)	896.9% (0.03)	-	-
	Hard thresholding	39.27% (0.06)	51.68% (0.04)	-	-
	Xue's estimator	35.98% (0.04)	46.36% (0.01)	-	-
	Adaptive Xue's estimator	35.95% (0.06)	47.58% (0.04)	-	-
	SEC	35.97% (0.04)	46.36% (0.01)	-	-
	Adaptive SEC	35.80% (0.07)	47.52% (0.04)	-	-

Table 4: Average (standard error) performance of the estimators for Example 3 over 100 replications. Since the underlying correlation matrix in model 3 has no true zeros, the estimator with prior information is not available here.

Tables 2, 3 and 4 show the estimation results for the three models using different approaches to estimate the covariance matrix. Clearly, all the sparse estimators perform uniformly better than the sample covariance matrix in terms of the relative Frobenius norm loss and spectral norm loss, even for the nonsparse model in Example 3. This observation confirms the effectiveness of sparse methods for estimating large dimensional covariance matrices.

The correlation estimators, with or without adaptive weights, generally achieve a smaller estimation error than the covariance estimator in Xue et al. (2012), especially for Example 1 and 2. This is in agreement with the theoretical result in Theorem 2 that the convergence rate of SEC is faster than Xue et al's estimator. When the true covariance matrix is sparse as in Example 1 and 2, we see that the correlation estimator with adaptive weighting is more accurate, and performs better in terms of identifying the sparsity pattern. In addition, the correlation estimator without weighting and the covariance estimator perform similarly in recovering the sparsity pattern. For the nonsparse model in Example 3, we find that sparse estimation of the covariance matrix continues to outperform the sample estimator by a large margin. For this example, adaptive weighting for the correlation estimator does not improve the one without weighting, which is to be expected as this is not a sparse model. Note that the Frobenius norm loss and spectral norm loss generated by SEC and adaptive SEC are usually less than twice of that given prior information under Example 1 and 2. And especially in model 1, the adaptive SEC estimator is almost as good as the one given prior information. This indicates that the performance of the SEC method is close to the oracle. The hard-thresholding estimator also performs well by the above evaluation. But the potential trouble is that the resulting estimator cannot be guaranteed to be positive definite. Figure 1 shows the eigenvalues of the hard-thresholding for three examples respectively. All of them contain negative ones. Moreover, this drawback cannot be fixed by projecting the estimate to a positive definite cone which destroys the sparsity pattern.

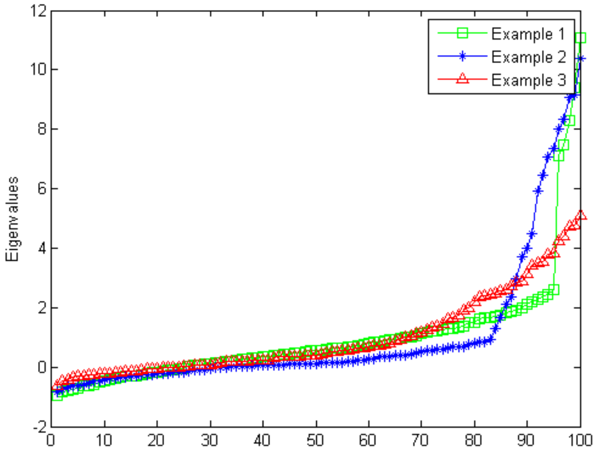


Figure 1: The eigenvalues of the hard thresholding estimator.

Observing that the difference between Xue's estimator and SEC is not large, we also present a pairwise comparison between the relative errors of the two norms in Table 5. One can see clearly that our SEC method is superior to the covariance estimation in terms of the spectral and the Frobenius norm for most instances. This again conforms the claim that the estimation of the correlation may be preferred over that of the covariance.

	Example 1			Example 2			Example 3		
$p\|n$	100 | 50	500 | 50	1000 | 100	100 | 50	500 | 50	1000 | 100	100 | 50	500 | 50	1000 | 100
F-norm	99/100	100/100	100/100	89/100	100/100	100/100	57/100	52/100	66/100
S-norm	97/100	71/100	62/100	98/100	100/100	71/100	81/100	63/100	54/100

Table 5: Pairwise comparison between Xue's estimator and SEC. The number in the table denotes the frequency that SEC is more accurate than Xue's estimator in terms of Frobenius norm and spectral norm respectively.

5.2. A microarray data

In this section we compare our SEC estimator with the regularized covariance estimator (Xue et al. (2012)) by analyzing a cardiovascular microarray experiment (Efron (2009, 2010)) for gene group-average agglomerative clustering. This data set has 63 training samples with 20436 gene expression values measured for each sample. Among them, 44 samples come from healthy controls and 19 come from cardiovascular patients.

As suggested by Rothman et al. (2009), we first calculate the F statistic

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

for all genes, where k is the number of classes, n, n_m are the total sample size and that of class m respectively, $\bar{x}$ and $\bar{x}_m$ are the overall mean and mean of class m, and $\hat{\sigma}_m^2$ is the sample variance of class m. This statistic represents how much discriminative information a gene can provide. In order to include both the informative and noninformative genes, we select the top 50 and bottom 150 genes from them. The resultant correlation matrix is expected to have a sparse structure. These 200 genes are standardized before the analysis is conducted. To choose the regularization parameter, we use five-fold cross validation.

First, we look at the covariance estimator by choosing the regularization term as $\rho\sum_{j\neq k}|\sigma_{jk}|$. The hope is to keep the unit diagonal entries by not penalizing the diagonal terms of the covariance matrix. However, as shown in Figure 2 the resulting estimator fails to be a correlation matrix, as the diagonal entries are usually larger than 1. Of course we may force the diagonal entries to be unity by taking $\hat{R} = D^{-1}\hat{\Sigma}D^{-1}$ to obtain an estimate of the correlation matrix, where D consists of the diagonal terms of $\hat{\Sigma}$. But doing so would slow down the convergence rate as discussed before. Thus, it is desirable, practically and theoretically, to enforce unity constraints of the diagonals in estimating the correlation matrix.

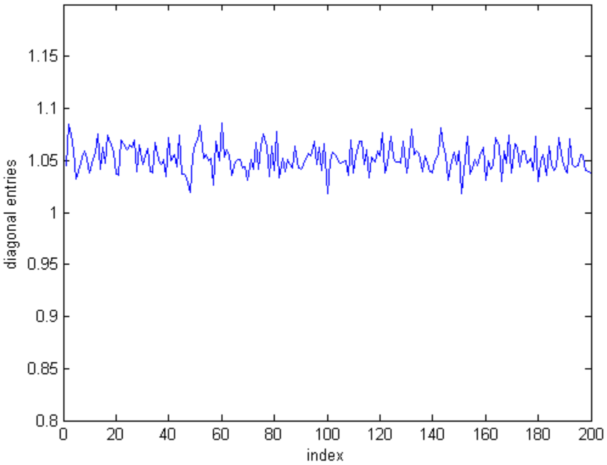
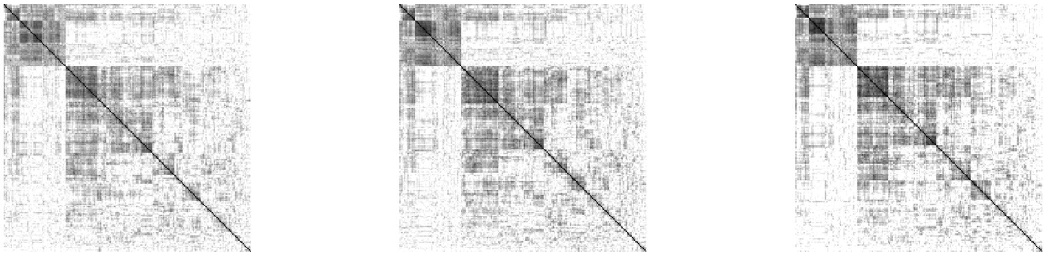


Figure 2: The diagonal entries of the covariance estimator.

Next, we plot in Figure 3 the heatmaps of the covariance estimator of Xue et al. (2012), the proposed SEC estimator without weighting, and the SEC estimator using the empirical correlation matrix in weighting. It is seen that the SEC estimator, especially the weighted one, gives sparser graph than that of the covariance estimator.



(a) Regularized covariance estimator (b) SEC with equal ℓ_1 penalty (c) SEC with adaptive penalty

Figure 3: Heat map of the absolute values of estimated correlation matrices. The genes are ordered by hierarchical clustering using the estimated correlations.

In order to compare these three estimates quantitatively, we implement cross validation again in the following manner. The selected 200 gene data set is randomly split into a training and a testing with sample size $2:1$. Then we choose the tuning parameter ρ by five-fold cross validation on the training data and come up with the estimated correlation matrices (and the covariance matrix) within this set. Following that we compare the resulting matrices with the empirical covariance matrix of the testing data via Frobenius norm and spectral norm together with their sparsity ($\#\{(i,j): \ i \neq j, \ \rho_{ij} \neq 0\}$). The results are listed in Table 6. We can see that our SEC estimators provide sparse structures and smaller estimation errors at the same time, which agrees with the simulation performance and reassuring our theoretical conclusions.

	$\hat{\Sigma}$	$\hat{R}$	$\hat{R}_{adapt.}$
F-norm error	52.9	52.8	52.4
S-norm error	35.5	32.5	34.3
zero numbers	9386	10530	14690

Table 6: Performance of the correlation (covariance) estimators for the cardiovascular microarray data.

6. Conclusion

We have proposed a new approach for estimating high dimensional correlation matrices, as opposed estimating covariance matrices in the literature. The proposed estimator can be efficiently computed, and enjoys attractive theoretical properties. We have also discussed an adaptive version of the estimator motivated by the adaptive lasso with the adaptive weights readily read off from the sample correlation matrix. Extensive simulation studies and an analysis of a cardiovascular microarray confirms that the proposed method performs better than its competitors. The Matlab code, implementing the SEC estimator, will appear on the second author's website.

Appendix

Only a sketch of the proof is provided here due to similarities to the proof in Cai and Liu (2011) or Xue et al. (2012).

Lemma 1. *Under the exponential-tail or the polynomial-tail conditions in Theorem 2, we have for some large constant C_1*

$$Pr\Big(max_{i,j}|\hat{\sigma}_{ij} - \sigma^0_{ij}| > C_1\sqrt{\frac{\log p}{n}}\Big) = o(1).$$

The proof of Lemma 1 follows straightforwardly from Cai and Liu (2011) and Xue et al. (2012). From Lemma 1 and the fact that $\hat{r}_{ij} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$, the following can be easily shown.

Lemma 2. *Under the exponential-tail or the polynomial-tail conditions in Theorem 2, if $\min_j \sigma^0_{jj}$ is bounded from below, we have for some large constant M*

$$Pr\Big(max_{i,j}|\hat{r}_{ij} - r^0_{ij}| > M\sqrt{\log p/n}\Big) = o(1).$$

Proof of Theorem 2. Let

$$\hat{\Delta} = \hat{R} - R^0 = \arg\min_{\Delta} F(\Delta), \quad \text{s.t. } R^0 + \Delta \succeq \varepsilon I \text{ and } \Delta_{jj} = 0,$$

where $F(\Delta) = \frac{1}{2}\|R^0 + \Delta - R_n\|^2_F + \rho|R^0 + \Delta|_1$.

Define the event $E = \{|\hat{r}_{ij} - r^0_{ij}| \leq \rho, \ \forall(i,j)\}$ where $\rho = M\sqrt{\frac{\log p}{n}}$ as in Lemma 2. We show that if event E holds, the global minimizer of our estimate defined in (2) satisfies $\|\hat{R} - R^0\|^2_F = O_p(s \log p/n)$.

Consider any $\Delta \in B$ for $B = \{\Delta : \Delta = \Delta^T, R^0 + \Delta \succeq \varepsilon I, \Delta_{jj} = 0, j = 1, ..., p, \|\Delta\|_F \geq 5s^{1/2}\rho\}$. Denote Δ_{A_0} as an matrix such that $[\Delta_{A_0}]_{ij} = [\Delta]_{ij}$ for $(i,j) \in A_0$ and $[\Delta_{A_0}]_{ij} = 0$ for $(i,j) \notin A_0$. We see that

$$\begin{aligned} F(\Delta) - F(0) &= \frac{1}{2}\|R^0 + \Delta - R_n\|_F^2 - \frac{1}{2}\|R^0 - R_n\|_F^2 + \rho(|R^0 + \Delta|_1 - |R^0|_1) \\ &= \frac{1}{2}\|\Delta\|_F^2 + \langle \Delta, R^0 - R_n \rangle + \rho|\Delta_{A_0^C}|_1 + \rho(|\Delta_{A_0} + R^0_{A_0}|_1 - |R^0_{A_0}|_1) \\ &\geq \frac{1}{2}\|\Delta\|_F^2 - \rho|\Delta|_1 + \rho|\Delta_{A_0^C}|_1 - \rho|\Delta_{A_0}|_1 = \frac{1}{2}\|\Delta\|_F^2 - 2\rho|\Delta_{A_0}|_1 \\ &\geq \frac{1}{2}\|\Delta\|_F^2 - 2\rho s^{1/2}\|\Delta\|_F \\ &> 0, \end{aligned}$$

where A_0 denotes the index set of the nonzero off-diagonal correlations in R^0 again. Because $G(\Delta) = F(\Delta) - F(0)$ is convex, $G(\Delta) = F(\Delta) - F(0) > 0$ for any $\Delta \in B$ and $G(0) = 0$, we see that the global optimal solution $\hat{\Delta}$ that minimizes $G(\Delta)$ must satisfy $\hat{\Delta} \ \notin B$ or $\|\hat{\Delta}\|_F \leq 5s^{1/2}\rho$, i.e.

$$\|\hat{R} - R^0\|_F \leq 5\sqrt{s}\rho.$$

Together with Lemma 2, we have proved

$$\|\hat{R} - R^0\|_F^2 = O_p\Big(s\frac{\log p}{n}\Big).$$

□

Proof of Corolloary 1. The proof to show $\|\hat{R} - R^0\|_F^2 = O_p\Big(s\frac{\log p}{n}\Big)$ is similar to that of Theorem 2, and is thus omitted. From it, we have

$$\|\hat{R} - R^0\|^2 \leq \|\hat{R} - R^0\|_F^2 = O_p\Big(s\frac{\log p}{n}\Big),$$

where $\|\cdot\|$ is the spectral norm. By the Weyl inequality $v_{i+j-1}(A_1 + A_2) \leq v_i(A_1) + v_j(A_2)$ for Hermitian matrices A_1 and A_2, we have $v_p(\hat{R}) \geq v_p(R^0) - v_1(R^0 - \hat{R}) > 0$ by assumption $v_p(R^0) \gg \sqrt{s\frac{\log p}{n}}$, meaning $\hat{R}$ that is positive definite with probability tending to one.

Note the above argument also holds true if the positive definite constraint $R \succeq \varepsilon I$ is removed. Therefore, with probability tending to one, $\hat{R}_{ij} = \text{sign}((R_n)_{ij})(|(R_n)_{ij}| - \frac{\rho}{W_{ij}})_+$, where $(a)_+ > 0$ for $a > 0$, and $(a)_+ = 0$ otherwise, which is the soft-thresholding estimator of $(R_n)_{ij}$ with an adaptive threshold $\frac{\rho}{W_{ij}}$. From the assumptions that $\min_{(i,j)\in A_0} r^0_{ij} \gg \sqrt{\frac{\log p}{n}}$, $\rho = M\sqrt{\frac{\log p}{n}}\min_{(i,j)\in A_0}(R_n)_{ij}$ and $W_{ij} = 1/|(R_n)_{ij}|$, and the conclusion of Lemma 2, it follows that: i) $\hat{r}_{ij} = 0$ for $(i,j) \in A_0^C$ and $i \neq j$, and ii) $\hat{r}_{ij} \neq 0$ for $(i,j) \in A_0$ with probability tending to one. □

References

Beck, A., Teboulle, M., Jan 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2 (1), 183202.

Bickel, P. J., Levina, E., Dec 2008a. Covariance regularization by thresholding. Ann. Statist. 36 (6), 25772604.

Bickel, P. J., Levina, E., Feb 2008b. Regularized estimation of large covariance matrices. Ann. Statist. 36 (1), 199227.

Bien, J., Tibshirani, R. J., Dec 2011. Sparse estimation of a covariance matrix. Biometrika 98 (4), 807820.

Cai, T., Liu, W., Jun 2011. Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association 106 (494), 672684.

Cai, T. T., Yuan, M., Aug 2012. Adaptive covariance matrix estimation through block thresholding. Ann. Statist. 40 (4), 20142042.

Cai, T. T., Zhou, H. H., Oct 2012. Optimal rates of convergence for sparse covariance matrix estimation. Ann. Statist. 40 (5), 23892420.

Efron, B., Sep 2009. Are a set of microarrays independent of each other? Ann. Appl. Stat. 3 (3), 922942.

Efron, B., 2010. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Vol. 1. Cambridge University Press.

Hiriart-Urruty, J.-B., Lemaréchal, C., 1993. Convex Analysis and Minimization Algorithms: Part 1: Fundamentals. Vol. 305. Springer.

Lam, C., Fan, J., Dec 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37 (6B), 42544278.

Lin, N., 2010. A penalized likelihood approach in covariance graphical model selection.

Liu, H., Wang, L., Zhao, T., Apr 2014. Sparse covariance matrix estimation with eigenvalue constraints. Journal of Computational and Graphical Statistics 23 (2), 439459.

Moreau, J.-J., 1965. Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France 93, 273–299.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Soviet Mathematics Doklady 27, 372–376.

Nesterov, Y., May 2005. Smooth minimization of non-smooth functions. Math. Program. 103 (1), 127152.

Nesterov, Y., Nesterov, I. E., 2004. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer.

Qi, H., Sun, D., Jan 2006. A quadratically convergent newton method for computing the nearest correlation matrix. SIAM Journal on Matrix Analysis and Applications 28 (2), 360385.

Rothman, A. J., Sep 2012. Positive definite estimators of large covariance matrices. Biometrika 99 (3), 733740.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics 2 (0), 494515.

Rothman, A. J., Levina, E., Zhu, J., Mar 2009. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association 104 (485), 177186.

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

Toh, K.-C., Yun, S., 2010. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pacific Journal of Optimization 6 (615-640), 15.

Tseng, P., 2008. On accelerated proximal gradient methods for convex-concave optimization. submitted to siam j. J. Optim.

Xue, L., Ma, S., Zou, H., 2012. Positive-definite ℓ_1-penalized estimation of large covariance matrices. Journal of the American Statistical Association 107 (500), 1480–1491.

Yosida, K., 1971. Functional analysis, 1980. Spring-Verlag, New York/Berlin.

Zou, H., Dec 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 14181429.