Journal of Multivariate Analysis 135 (2015) 153–162

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Fast and adaptive sparse precision matrix estimation in high dimensions

Weidong Liu [a], Xi Luo [b,c,d,*]

[a] Department of Mathematics and Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China
[b] Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, RI, USA
[c] Brown Institute for Brain Science, Brown University, Providence, RI, USA
[d] Initiative for Computation in Brain and Mind, Brown University, Providence, RI, USA

HIGHLIGHTS

- We propose a new procedure for sparse precision matrix estimation.
- We are among the first to establish the theory of cross validation for this problem.
- The conditions are slightly weaker than an important penalized likelihood method.
- Improved numerical performance is observed in several examples.

ARTICLE INFO

Article history:
Received 7 December 2013
Available online 18 December 2014

AMS subject classifications:
62H12
62F12

Keywords:
Adaptivity
Coordinate descent
Cross validation
Gaussian graphical models
Lasso
Convergence rates

ABSTRACT

This paper proposes a new method for estimating sparse precision matrices in the high dimensional setting. It has been popular to study fast computation and adaptive procedures for this problem. We propose a novel approach, called Sparse Column-wise Inverse Operator, to address these two issues. We analyze an adaptive procedure based on cross validation, and establish its convergence rate under the Frobenius norm. The convergence rates under other matrix norms are also established. This method also enjoys the advantage of fast computation for large-scale problems, via a coordinate descent algorithm. Numerical merits are illustrated using both simulated and real datasets. In particular, it performs favorably on an HIV brain tissue dataset and an ADHD resting-state fMRI dataset.

1. Introduction

Estimating covariance matrices is fundamental in multivariate analysis. It has been popular to estimate the inverse covariance (or precision) matrix in the high dimensional setting, where the number of variables p goes to infinity with the sample size n (more precisely, in this paper, $p \gg n$ and $(\log p)/n = o(1)$). Inverting the sample covariance matrix has been known to be unstable for estimating the precision matrix. Recent proposals usually formulate this objective as regularized/penalized optimization problems, where regularization is employed to control the sparsity of the precision matrix. Besides the challenge of solving such large optimization problems, there is an important issue on how to choose an

* Corresponding author at: Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, RI, USA.
E-mail address: xi.rossi.luo@gmail.com (X. Luo).

http://dx.doi.org/10.1016/j.jmva.2014.11.005
0047-259X/

appropriate regularization level that is adaptive to the data. To address these two challenges, we propose a fast and adaptive method, and establish the theoretical properties when the regularization level is chosen by cross validation.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-variate random vector with a covariance matrix $\boldsymbol{\Sigma}$ or its corresponding precision matrix $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$. Suppose we observe independent and identically distributed random samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ from the distribution of $\boldsymbol{X}$. To encourage a sparse and stable estimate for $\boldsymbol{\Omega}$, regularized/penalized likelihood approaches have been proposed. Here, sparsity means that most of the entries in $\boldsymbol{\Omega}$ are exactly zero. Popular penalties include the ℓ_1 penalty [22] and its extensions, for example, [24,12,8,20]. In particular, [12] developed an efficient algorithm, glasso, to compute the penalized likelihood estimator, and its convergence rates were obtained under the Frobenius norm [20] and the elementwise ℓ_∞ norm and spectral norm [19]. Other penalties were also studied before. For example, the ℓ_1 penalty was replaced by the nonconvex SCAD penalty [11,14,10]. Due to the complexity of the penalized likelihood objective, theoretical analysis and computation are rather involved. Moreover, the theory usually relies on some theoretical assumptions of the penalty, and thus it provides limited guidance for applications.

Recently, column-wise or neighborhood based procedures have caught much attention, due to the advantages in both computation and analysis. [18] proposed to recover the support of $\boldsymbol{\Omega}$ using ℓ_1 penalized regression, aka LASSO [22], in a row by row fashion. This can be computed efficiently via path-following coordinate descent [13] for example. A Dantzig selector proposal, replacing the LASSO approach, was proposed recently by [23], and the computation is based on standard solvers for linear programming. [5] proposed a procedure, CLIME, which seeks a sparse precision matrix under a matrix inversion constraint. Their procedure is also solved column by column via linear programming. Compared with the regularized likelihood approaches, their convergence rates were obtained under several matrix norms mentioned before, without imposing the mutual incoherence condition [19], and were improved when $\boldsymbol{X}$ follows polynomial tail distributions. However, all these procedures can be computational expensive for very large p, and again these estimators were analyzed based on theoretical choices of the penalty.

Cross validation on the other hand has gained popularity for choosing the penalty levels or tuning parameters, because it is adaptive and usually yields superior performance in practice. Unfortunately, the theoretical understanding of cross validation is sparse. For a related problem on estimating sparse covariance matrices, [1] analyzed the performance of covariance thresholding where the threshold is based on cross validation. [4] provided a different approach using self-adaptive thresholding. However, these covariance estimation results cannot be extended to the inverse covariance setting, partly due to the problem complexity. This paper will provide theoretical justification for cross validation when estimating the precision matrix. This result is made possible because we propose a new column-wise procedure that is easy to compute and analyze. To the best of our knowledge, this paper is among the first to provide theoretical justification of cross validation for sparse precision matrix estimation.

The contributions of this paper are several folds. First, we propose a novel and penalized column-wise procedure, called Sparse Columnwise Inverse Operator (SCIO), for estimating the precision matrix $\boldsymbol{\Omega}$. Second, we establish the theoretical justification under mild conditions when its penalty is chosen by cross validation. The theory for cross validation is summarized as follows. A matrix is called s_p-sparse if there are at most s_p non-zero elements on each row. It is shown that the error between our cross validated estimator $\hat{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}$ satisfies $\|\hat{\boldsymbol{\Omega}}^{-1} - \boldsymbol{\Omega}\|_F^2/p = O_P(s_p(\log p)/n)$, where $\|\cdot\|_F$ is the Frobenius norm. Third, theoretical guarantees for the SCIO estimator are also obtained under other matrix norms, for example the element-wise ℓ_∞ norm which achieves graphical model selection [15]. Fourth, we provide a fast and simple algorithm for computing the estimator. Because our algorithm exploits the advantages of conjugate gradient and coordinate descent, and thus it provides superior performance in computational speed and cost. In particular, we reduce two nested loops in glasso [12] to only one. An R package of our method, *scio*, has been developed, and is publicly available on CRAN.

The rest of the paper is organized as follows. In Section 2, after basic notations and definitions are introduced, we present the SCIO estimator. Finite sample convergence rates are established with the penalty level chosen both by theory in Section 3 and by cross validation in Section 4. The algorithm for solving SCIO is introduced in Section 5. Its numerical merits are illustrated using simulated and real datasets. Further discussions on the connections and differences of our results with other related work are given in Section 6. The supplementary material includes additional results for the numerical examples in Section 5 and the proof of the main results (see Appendix A).

The notations in this paper are collected here. Throughout, for a vector $\mathbf{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, define $|\mathbf{a}|_1 = \sum_{j=1}^p |a_j|$ and $|\mathbf{a}|_2 = \sqrt{\sum_{j=1}^p a_j^2}$. All vectors are column vectors. For a matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we define the elementwise l_∞ norm $|\boldsymbol{A}|_\infty = \max_{1\le i\le p, 1\le j\le q} |a_{ij}|$, the spectral norm $\|\boldsymbol{A}\|_2 = \sup_{|\mathbf{x}|_2\le 1} |\boldsymbol{A}\mathbf{x}|_2$, the matrix ℓ_1 norm $\|\boldsymbol{A}\|_{L_1} = \max_{1\le j\le q} \sum_{i=1}^p |a_{ij}|$, the matrix ∞ norm $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le q} \sum_{j=1}^p |a_{ij}|$, the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, and the elementwise ℓ_1 norm $\|\boldsymbol{A}\|_1 = \sum_{i=1}^p \sum_{j=1}^q |a_{ij}|$. $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{\cdot j}$ denote the ith row and jth column respectively. $\boldsymbol{I}$ denotes an identity matrix. $1\{\cdot\}$ is the indicator function. The transpose of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$. For any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of proper sizes, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \sum_i \left(\boldsymbol{A}^T\boldsymbol{B}\right)_{ii}$. For any two index sets T and T' and a matrix $\boldsymbol{A}$, we use $\boldsymbol{A}_{TT'}$ to denote the $|T| \times |T'|$ matrix with rows and columns of $\boldsymbol{A}$ indexed by T and T' respectively. The notation $\boldsymbol{A} \succ 0$ means that $\boldsymbol{A}$ is positive definite. For two real sequences $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for large n, $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Write $a_n = O_P(b_n)$ if $a_n = O(b_n)$ holds with the probability going to 1. The constants $C, C_0, C_1, \ldots$ may represent different values at each appearance.

2. Methodology

Our estimator is motivated by adding the ℓ_1 penalty [22] to a column loss function, which is related to conjugate descent and a constrained minimization approach CLIME [5]. The technical derivations that lead to the estimator is provided in the supplementary material (see Appendix A). Denote the sample covariance matrix by $\hat{\mathbf{\Sigma}}$. Let a vector $\hat{\boldsymbol{\beta}}_i$ be the solution to the following equation:

$$\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\frac{1}{2}\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}\boldsymbol{\beta} - \boldsymbol{e}_i^T\boldsymbol{\beta} + \lambda_{ni}|\boldsymbol{\beta}|_1\right\}, \tag{1}$$

where $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip})^T$, $\boldsymbol{e}_i$ is the ith column of a $p \times p$ identity matrix, and $\lambda_{ni} > 0$ is a tuning parameter. The tuning parameter could be different from column to column, adapting to different magnitude and sparsity of each column.

One can formulate a precision matrix estimate where each column is the corresponding $\hat{\boldsymbol{\beta}}_i$. However, the resulting matrix may not be symmetric. Similar to a symmetrization step employed in CLIME, we define the SCIO estimator $\hat{\mathbf{\Omega}} = (\hat{\omega}_{ij})_{p\times p}$, using the following symmetrization step,

$$\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\beta}_{ij}1\{|\hat{\beta}_{ij}| < |\hat{\beta}_{ji}|\} + \hat{\beta}_{ji}1\{|\hat{\beta}_{ij}| \geq |\hat{\beta}_{ji}|\}. \tag{2}$$

As we will establish in Section 3, similar to the results of CLIME, the convergence rates shall not change if the diagonal of the sample covariance $\hat{\mathbf{\Sigma}}$ is added by a small positive amount, as long as in the order of $n^{-1/2}\log^{1/2} p$. With this modification, (1) is then strictly convex and has a unique solution. In Section 5, we will present an efficient coordinate descent algorithm to solve it.

The SCIO estimator, like other penalized estimators, depends on the choice of λ_{ni}. We allow λ_{ni} to be different from column to column, so that it is possible to adapt to each column's magnitude and sparsity, as we will illustrate in Section 4. More importantly, due to the simplified column loss function (1), we are able to establish, in Section 4, the theoretical guarantees when λ_{ni} is chosen by cross validation. In comparison, the theory of cross validation for glasso [12] and CLIME [5] has not been established before, to the best of our knowledge.

3. Theoretical guarantees

3.1. Conditions

Let $\mathcal{S}_i$ be the support of $\mathbf{\Omega}_{\cdot,i}$, the ith column of $\mathbf{\Omega} = (\omega_{ij})_{p\times p}$. Define the s_p-sparse matrices class

$$\mathcal{U} = \left\{\mathbf{\Omega} \succ 0 : \max_{1\leq j\leq p}\sum_{i=1}^{p} 1\{\omega_{ij} \neq 0\} \leq s_p,\ \|\mathbf{\Omega}\|_{L_1} \leq M_p,\ c_0^{-1} \leq \Lambda_{\min}(\mathbf{\Omega}) \leq \Lambda_{\max}(\mathbf{\Omega}) \leq c_0\right\},$$

where c_0 is a positive constant, $\Lambda_{\min}(\mathbf{\Omega})$ and $\Lambda_{\max}(\mathbf{\Omega})$ are the minimum and maximum eigenvalues of $\mathbf{\Omega}$ respectively. The sparsity s_p is allowed to grow with p, as long as it satisfies the following condition.

(C1). Suppose that $\mathbf{\Omega} \in \mathcal{U}$ with

$$s_p = o\left(\sqrt{\frac{n}{\log p}}\right) \tag{3}$$

and

$$\max_{1\leq i\leq p}\left\|\mathbf{\Sigma}_{\mathcal{S}_i^c\mathcal{S}_i}\left(\mathbf{\Sigma}_{\mathcal{S}_i\mathcal{S}_i}\right)^{-1}\right\|_\infty \leq 1-\alpha \tag{4}$$

for some $0 < \alpha < 1$.

As we will see from Theorem 1, condition (3) is required for proving the consistency. Condition (4) is in the same spirit as the mutual incoherence or irrepresentable condition for glasso [19], but it is slightly relaxed, see Remark 2. In general, this type of conditions is believed to be almost necessary for penalization methods to recover support.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T = \mathbf{\Omega}\boldsymbol{X} - \mathbf{\Omega}\boldsymbol{\mu}$ where $\boldsymbol{\mu} = \mathsf{E}\boldsymbol{X}$. The covariance matrix of $\boldsymbol{Y}$ is thus $\mathbf{\Omega}$. The second condition is on the moments of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

(C2). (Exponential-type tails) Suppose that $\log p = o(n)$. There exist positive numbers $\eta > 0$ and $K > 0$ such that

$$\mathsf{E}\exp\left(\eta(X_i-\mu_i)^2\right) \leq K, \qquad \mathsf{E}\exp\left(\eta Y_i^2\right) \leq K \quad \text{for all } 1 \leq i \leq p.$$

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

$$\mathsf{E}|X_i-\mu_i|^{4\gamma+4+\delta} \leq K, \qquad \mathsf{E}|Y_i|^{4\gamma+4+\delta} \leq K \quad \text{for all } i.$$

We will assume either one of these two types of tails in our main analysis. These two conditions are standard for analyzing sparse precision matrix estimation, see [5] and references within.

3.2. Convergence rates of $\hat{\mathbf{\Omega}} - \mathbf{\Omega}$

The first theorem is on the convergence rate under the spectral norm. It implies the convergence rates of eigenvalues and eigenvectors, which are essential in principle component analysis for example. The convergence rate under the spectral norm may also be important for classification, for example linear/quadratic discriminant analysis as we illustrate in Section 5.

Theorem 1. *Let $\lambda_{ni} = C_0\sqrt{\log p/n}$ with C_0 being a sufficiently large number. Under* (C1)*, and* (C2) *(or* (C2*)*), we have*

$$\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_2 \leq CM_p s_p \sqrt{\frac{\log p}{n}}$$

with probability greater than $1 - O\left(p^{-1}\right)$ (or $1 - O(p^{-1} + n^{-\delta/8})$ under (C2*)*), where $C > 0$ depends only on c_0, η, C_0 and K (or c_0, c_1, γ, δ, C_0 and K under* (C2*)*).*

Remark 1. If $M_p s_p n^{-1/2} \log^{1/2} p = o(1)$, then $\hat{\mathbf{\Omega}}$ is positive definite with probability tending to one. We can also revise $\hat{\mathbf{\Omega}}$ to $\hat{\mathbf{\Omega}}_\tau$ with

$$\hat{\mathbf{\Omega}}_\tau = \hat{\mathbf{\Omega}} + \tau \boldsymbol{I},$$

where $\tau = (|\Lambda_{\min}(\hat{\mathbf{\Omega}})| + n^{-1/2})1\{\Lambda_{\min}(\hat{\mathbf{\Omega}}) \leq 0\}$. By Theorem 1, assuming $\tau \leq CM_p s_p n^{-1/2} \log^{1/2} p$, we have with probability greater than $1 - O(p^{-1})$ (or $1 - O(p^{-1} + n^{-\delta/8})$) that

$$\left\|\hat{\mathbf{\Omega}}_\tau - \mathbf{\Omega}\right\|_2 \leq CM_p s_p \sqrt{\frac{\log p}{n}}.$$

Such a simple perturbation will make the revised estimator $\hat{\mathbf{\Omega}}_\tau$ to have a larger minimal eigenvalue, for stability concerns. The later results on support recovery and other norms will also hold under such a small perturbation.

Remark 2 ([19])**.** imposed the following irrepresentable condition on glasso: for some $0 < \alpha < 1$,

$$\left\|\mathbf{\Gamma}_{\Psi^c\Psi}\left(\mathbf{\Gamma}_{\Psi\Psi}\right)^{-1}\right\|_\infty \leq 1 - \alpha, \tag{5}$$

where Ψ is the support of $\mathbf{\Omega}$, $\mathbf{\Gamma} = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$, and $\otimes$ denotes the Kronecker matrix product. To make things concrete, we now compare our conditions using the examples given in [19]:

1. In the diamond graph, let $p = 4$, $\sigma_{ii} = 1$, $\sigma_{23} = 0$, $\sigma_{14} = 2\rho^2$ and $\sigma_{ij} = \rho$ for all $i \neq j$, $(i, j) \neq (2, 3)$ and $(2, 4)$. For this matrix, (5) is reduced to $4|\rho|(|\rho| + 1) < 1$ and so it requires $\rho \in (-0.208, 0.208)$. Our relaxed condition (4) only needs $\rho \in (-0.5, 0.5)$.
2. In the star graph, let $p = 4$, $\sigma_{ii} = 1$, $\sigma_{1,j} = \rho$ for $j = 2, 3, 4$, $\sigma_{ij} = \rho^2$ for $1 < i < j \leq 4$. For this model, (5) requires $|\rho|(|\rho| + 2) < 1$ (i.e. $\rho \in (-0.4142, 0.4142)$), while our condition (4) holds for all $\rho \in (-1, 1)$.

We have the following result on the convergence rates under the element-wise l_∞ norm and the Frobenius norm.

Theorem 2. *Under the conditions of Theorem 1, we have with probability greater than $1 - O(p^{-1})$ under* (C2) *(or $1 - O(p^{-1} + n^{-\delta/8})$ under* (C2*))

$$\left|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right|_\infty \leq CM_p \sqrt{\frac{\log p}{n}} \tag{6}$$

and

$$\frac{1}{p}\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_F^2 \leq Cs_p \frac{\log p}{n}. \tag{7}$$

Remark 3. The convergence rate under the Frobenius norm does not depend on M_p. In comparison, [6] obtained the minimax lower bound, when $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$,

$$\frac{1}{p}\min_{\hat{\mathbf{\Omega}}}\max_{\mathbf{\Omega}\in\mathcal{U}} \mathrm{E}\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_F^2 \geq CM_p^2 s_p \frac{\log p}{n}. \tag{8}$$

They also showed that this rate is achieved by sequentially running two CLIME estimators, where the second CLIME estimator uses the first CLIME estimate as input. Though CLIME allows a weaker sparsity condition where our ℓ_0 ball bound s_p in $\mathcal{U}$ is replaced by an ℓ_q ball bound ($0 \leq q < 1$), our rate in (7) is faster than CLIME, because M_p^2 in (8) could grow with p. The faster rate is due to the fact that we consider the condition (4). Under a slightly stronger condition (5) (see Remark 2), [19]

proved that the glasso estimator $\hat{\mathbf{\Omega}}_{glasso}$ has the following convergence rate

$$\frac{1}{p}\left\|\hat{\mathbf{\Omega}}_{glasso}-\mathbf{\Omega}\right\|_F^2=O_P\left(\kappa_\Gamma^2 s_p\frac{\log p}{n}\right), \tag{9}$$

where $\kappa_\Gamma=\left\|(\mathbf{\Gamma}_{\Psi\Psi})^{-1}\right\|_{L_1}$. Our convergence rate is also faster than theirs in (9) if $\kappa_\Gamma\to\infty$.

3.3. Support recovery

As discussed in the introduction, support recovery is related to Gaussian graphical models. The support of $\mathbf{\Omega}$ is recovered by SCIO, with high probability by the following theorem. Recall $\Psi=\{(i,j):\omega_{ij}\neq 0\}$ be the support of $\mathbf{\Omega}$, and similarly

$$\hat{\Psi}=\{(i,j):\hat{\omega}_{ij}\neq 0\}.$$

The next theorem gives the result on support recovery.

Theorem 3. (i). *Under the conditions of Theorem 1, we have $\hat{\Psi}\subseteq\Psi$ with probability greater than $1-O(p^{-1})$ under* (C2) *(or $1-O(p^{-1}+n^{-\delta/8})$ under* (C2*)). (ii). *In addition, suppose that for a sufficiently large number $C>0$,*

$$\min_{(i,j)\in\Psi}\left|\omega_{ij}\right|\geq CM_p\sqrt{\frac{\log p}{n}}. \tag{10}$$

Then under the conditions of Theorem 1, we have $\hat{\Psi}=\Psi$ with probability greater than $1-O(p^{-1})$ under (C2) *(or $1-O(p^{-1}+n^{-\delta/8})$ under* (C2*)).

The condition (10) on the signal strength is standard for support recovery, see [19,5] for example. We also note that the CLIME method [5] requires an additional thresholding step for support recovery, while SCIO does not need this step.

4. Theory for data-driven penalty

This section analyzes a cross validation scheme for choosing the tuning parameter λ_{ni}, and we establish the theoretical justification of this data-driven procedure.

We consider the following cross validation method for simplicity, similar to the one analyzed in [1]. Divide the sample $\{\boldsymbol{X}_k;\,1\leq k\leq n\}$ into two subsamples at random. Let n_1 and $n_2=n-n_1$ be the two sample sizes of the random splits satisfying $n_1\asymp n_2\asymp n$, and let $\hat{\mathbf{\Sigma}}_1^l$, $\hat{\mathbf{\Sigma}}_2^l$ be the sample covariance matrices from the two samples n_1 and n_2 respectively in the lth split, for $l=1,\ldots,H$, where H is a fixed integer. For each i, let $\hat{\boldsymbol{\beta}}_i^l(\lambda)$ be the estimator minimizing the average out-of-sample SCIO loss, over λ,

$$\hat{R}_i(\lambda)=\frac{1}{H}\sum_{v=1}^{H}\left[\frac{1}{2}(\hat{\boldsymbol{\beta}}_i^l(\lambda))^T\hat{\mathbf{\Sigma}}_2^l\hat{\boldsymbol{\beta}}_i^l(\lambda)-\boldsymbol{e}_i^T\hat{\boldsymbol{\beta}}_i^l(\lambda)\right] \tag{11}$$

where $\hat{\boldsymbol{\beta}}_i^l(\lambda)$ is calculated from the n_1 samples with a tuning parameter λ to be determined. For implementation purposes, instead of searching for continuous λ, we will divide the interval $[0,4]$ by a grid $\lambda_0<\lambda_1<\cdots<\lambda_N$, where $\lambda_i=\frac{4i}{N}$. The number 4 comes from the CLIME constraint, see the supplementary material (see Appendix A). The tuning parameter on the grid is chosen by, for each i,

$$\hat{\lambda}_i=\operatorname*{argmin}_{0\leq j\leq N}\hat{R}_i(\lambda_j). \tag{12}$$

It is important to note that the size N should be sufficiently large but not too large, see the first two conditions on N in Theorem 4, and the convergence rate will then hold even if we only perform cross validation on a grid. The choice of $\hat{\lambda}_i$ could be different for estimating each column of the precision matrix using the column loss function (11). This allows the procedure to adapt to the magnitude and sparsity of each column, compared with the standard glasso estimator with a single choice of λ for the whole matrix. Though it is possible to specify different λ for each column (even each entry) in glasso, searching over all possible combinations of λ's over high dimensional grids, using a non-column-wise loss (e.g. the likelihood), is computationally untrackable. Our column loss thus provides a simple and computationally trackable alternative for choosing adaptive λ.

As described before, the complexity of the likelihood function may make it difficult to analyze the glasso estimator using cross validation. Though CLIME uses a constrained approach for estimation, its constrained objective function cannot be directly used for cross validation. [5] proposed to use the likelihood function as the cross validation loss, which makes it difficult to establish the theory of cross validated CLIME. For a different setting of estimating the covariance matrix, [1] obtained the convergence rate under the Frobenius norm, using covariance thresholding. The threshold is also based on

sample splitting like ours. However, to the best of our knowledge, it has been an open problem on establishing the theoretical justification of cross validation when estimating the precision matrix. Theorem 4 below fills the gap, showing that the estimator based on $\hat{\lambda}_i$ from (12) attains the optimal rate under the Frobenius norm. For simplicity, we set $H = 1$ as in [1].

Our theory adopts the following condition on the sub-Gaussian distribution, which was used in [7] for example.

(C3). There exist positive numbers $\eta' > 0$ and $K' > 0$ such that

$$\max_{|\boldsymbol{v}|_2=1} \mathsf{E} \exp\left(\eta'(\boldsymbol{v}^T(\boldsymbol{X}-\boldsymbol{\mu}))^2\right) \le K'.$$

This condition is slightly stronger than (C2), because our next theorem adapts to unknown $\boldsymbol{\Omega}$ using cross validation, instead of the theoretical choice λ_{ni}. It is easy to see that (C3) holds for the multivariate normal distribution as a special case.

Denote the unsymmetrized $\hat{\boldsymbol{\Omega}}_1^1 := (\hat{\omega}_{ij1}^1) = (\hat{\boldsymbol{\beta}}_1^1(\hat{\lambda}_1), \ldots, \hat{\boldsymbol{\beta}}_p^1(\hat{\lambda}_p))$ and recall the symmetrized matrix $\hat{\boldsymbol{\Omega}}^1$ as

$$\hat{\omega}_{ij}^1 = \hat{\omega}_{ji}^1 = \hat{\omega}_{ij1}^1 1\{|\hat{\omega}_{ij1}^1| < |\hat{\omega}_{ji1}^1|\} + \hat{\omega}_{ji1}^1 1\{|\hat{\omega}_{ij1}^1| \ge |\hat{\omega}_{ji1}^1|\}.$$

The following theorem shows that the estimator $\hat{\boldsymbol{\Omega}}^1 = (\hat{\omega}_{ij}^1)$ attains the minimax optimal rate under the Frobenius norm.

Theorem 4. *Under the conditions* $\log N = O(\log p)$, $\sqrt{n/\log p} = o(N)$, *and* (C3), *we have as* $n, p \to \infty$,

$$\frac{1}{p}\left\|\hat{\boldsymbol{\Omega}}^1 - \boldsymbol{\Omega}\right\|_F^2 = O_P\left(s_p \frac{\log p}{n}\right).$$

The convergence rate using cross validation is the same as (7) in Theorem 2 with the theoretical choice of λ. Using similar arguments in Theorem 4 of [1], this result can be extended to multiple folds $H > 1$. To the best of our knowledge, Theorem 4 is the first result on the theoretical justification of cross validation when estimating the sparse precision matrix.

5. Numerical examples

5.1. Algorithm

Recall that the SCIO estimator is obtained by applying symmetrization (2) to the solution from (1), where each column $\hat{\boldsymbol{\beta}}_i$ is given by the following

$$\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}_i \in \mathbb{R}^p}{\operatorname{argmin}} \left\{\frac{1}{2}\boldsymbol{\beta}_i^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_i - \boldsymbol{e}_i^T \boldsymbol{\beta}_i + \lambda|\boldsymbol{\beta}_i|_1\right\} \tag{13}$$

for any $\lambda > 0$. We propose to employ an iterative coordinate descent algorithm to solve (13) for each i. In contrast, the R package glasso employs an outside loop over the columns of the precision matrix, while having another inside loop over the coordinates of each column. Our algorithm does not need an outside loop because our loss function is column-wise.

The iterative coordinate descent algorithm for each i goes as follows. In each iteration, we fix all but one coordinate in $\boldsymbol{\beta}$, and optimize over that fixed coordinate. Without loss of generality, we consider optimizing over the pth coordinate β_p while all other coordinates of $\boldsymbol{\beta}$ (denoted by $\boldsymbol{\beta}_{-p}$) are fixed. The solution is in an explicit form by the following simple proposition. The solution when fixing other coordinates is similar, simply by permuting the matrix. We then loop through the coordinates until the updates are smaller than a user-specified threshold, say 10^{-4}.

Proposition 1. *Let the subvector partition* $\boldsymbol{\beta} = (\boldsymbol{\beta}_{-p}, \beta_p)$ *and partition* $\hat{\boldsymbol{\Sigma}}$ *accordingly as follows*

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12}^T \\ \hat{\boldsymbol{\Sigma}}_{12} & \hat{\Sigma}_{22} \end{pmatrix}.$$

Fixing $\boldsymbol{\beta}_{-p}$, *the minimizer of* (13) *is*

$$\beta_p = \mathcal{T}\left(1\{p=i\} - \boldsymbol{\beta}_{-p}^T \hat{\boldsymbol{\Sigma}}_{12}, \lambda\right) / \hat{\Sigma}_{22}$$

where the soft thresholding rule $\mathcal{T}(x, \lambda) = \operatorname{sign}(x)\max(|x| - \lambda, 0)$.

We implement this algorithm in an R package, *scio*, available on CRAN. All the following computation is performed using R on an AMD Opteron processor (2.6 GHz) with 32 Gb memory. The glasso and CLIME estimators are computed using its R packages *glasso* (version 1.7) and *clime* (version 0.4.1) respectively. The path-following strategy with warm-starts [13] is enabled in all methods.

5.2. Simulations

In this section, we compare the performance with glasso and CLIME on several measures using simulated data. In order to compare the adaptivity of the procedures, the covariance matrices that generate the data all contain two block diagonals

Table 1
Comparison of average losses of SCIO, SCIOcv, CLIME, and glasso over 100 simulation runs. The best performance is highlighted in bold. All standard errors of the results are smaller than 0.1.

p	Decay				Sparse				Block			
	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso	SCIO	SCIOcv	CLIME	glasso
Spectral norm												
50	**10.00**	11.24	11.62	12.10	**2.73**	4.03	5.70	3.86	**7.24**	9.55	8.03	9.61
100	**11.89**	12.68	12.29	13.11	**4.51**	5.57	6.54	5.70	**9.63**	9.78	9.13	9.77
200	**12.88**	13.46	12.91	13.84	**7.93**	8.31	8.43	8.48	9.88	9.85	10.05	**9.83**
400	**13.63**	13.87	14.09	14.07	**10.88**	11.60	11.63	11.11	9.92	9.91	10.31	**9.87**
800	14.13	**14.05**	14.10	14.71	15.58	**15.48**	15.60	16.08	9.96	**9.95**	10.01	10.63
1600	14.15	**14.12**	**14.12**	14.83	20.94	**20.90**	20.94	21.61	9.97	**9.96**	10.15	10.68
Frobenius norm												
50	**16.22**	18.54	19.25	20.18	**6.71**	7.95	12.66	8.14	**16.10**	20.98	17.58	21.68
100	**27.48**	29.58	28.40	30.92	**12.93**	14.84	18.48	14.91	**30.83**	31.02	28.72	31.15
200	42.93	45.12	**42.80**	47.00	**24.34**	24.67	26.60	26.11	44.49	44.23	44.92	**44.19**
400	**65.61**	66.60	68.65	68.10	**36.65**	38.99	40.67	37.76	62.91	62.73	65.38	**62.54**
800	97.52	**96.09**	97.25	102.67	59.08	**57.55**	59.97	66.30	88.98	**88.78**	88.63	96.42
1600	138.09	**136.90**	137.74	147.11	83.85	**82.87**	84.50	96.90	125.85	125.64	**125.41**	137.27

of different magnitude, where the second block is 4 times the first one. Similar examples were used in [4] in comparing adaptive covariance estimation. The first block is generated from the following models respectively.

1. **decay**: $\omega_{ij} = 0.6^{|i-j|}$.
2. **sparse**: Let the prototype $\boldsymbol{\Omega}_0 = \boldsymbol{O} + \delta \boldsymbol{I}$, where each off-diagonal entry in $\boldsymbol{O}$ is generated independently, and equals to 0.5 with probability 0.1 and 0 with probability 0.9. δ is chosen such that the conditional number (the ratio of maximal and minimal singular values of a matrix) equals to p. Finally, the block matrix is standardized to have unit diagonals.
3. **block**: A block diagonal matrix with block size 5 where each block has off-diagonal entries equal to 0.5 and diagonal 1. The resulting matrix is then randomly permuted.

100 independent and identically distributed observations constituting a training dataset are generated from each multivariate Gaussian covariance model with mean zero, and 100 additional observations are generated from the same model as a validating dataset. Using the training data alone, a series of penalized estimators with 50 different tuning parameters λ is computed. For a fair comparison, we first pick the tuning parameters in glasso, CLIME, and SCIO to produce the smallest Bregman loss on the validation sample. The Bregman loss is defined by

$$L(\boldsymbol{\Sigma}, \boldsymbol{\Omega}) = \langle \boldsymbol{\Omega}, \boldsymbol{\Sigma} \rangle - \log \det(\boldsymbol{\Omega}).$$

We also compare with our cross validation scheme in Section 4, where the cross validation loss is the column-wise adaptive loss (11). The resulting estimator is denoted by SCIOcv. We consider different values of $p = 50, 100, 200, 400, 800, 1600$, and replicate 100 times.

Table 1 compares the estimation performance of SCIO, SCIOcv, CLIME, and glasso under the spectral norm and the Frobenius norm. It shows that SCIO and SCIOcv almost uniformly outperform all other methods under both norms. SCIO has better performance when $p \leq 400$, while SCIOcv has better performance when $p \geq 800$. The glasso estimator has the worst performance overall, but it has slightly improved performance than other methods in the block model for $p = 200$ and 400. The CLIME estimator has slightly worse performance than our estimators overall, except for a few cases.

As discussed before, support recovery carries important consequences for graphical model estimation. The frequencies of correct zero/nonzero identification are summarized in Table 1 of the supplementary material (see Appendix A). In there, the SCIO and SCIOcv estimates are sparser than the CLIME and glasso estimates in general. To further illustrate this, we plot the heatmaps of support recovery in Fig. 1 using $p = 100$ as a representing example. These heatmaps confirm that the SCIO estimates usually contain less zeros than glasso and CLIME. By visual inspection, these SCIO estimates also tend to be closer to the truth, especially under the sparse model. In particular, they adapt to different magnitude. In contrast, glasso yields some interference patterns and artificial stripes, especially under the sparse model.

5.3. A genetic dataset on HIV-1 associated neurocognitive disorders

Antiretroviral therapy (ART) has greatly reduced mortality and morbidity of HIV patients; however, HIV-1 associated neurocognitive disorders (HAND) are becoming common, which cause greatly degradation of life quality. We here apply our graphical models to a gene expression dataset [2] to study how their genetic interactions/pathways are altered between treated and untreated HAND patients, and compare with other methods using classification. The supplementary material includes the full description of the dataset, the modeling approach, and additional results (see Appendix A).

Fig. 2(a) compares classification accuracy between treated and untreated HAND. The results comparing HAND and controls are not shown because all methods have a constant area-under-the-curve value 1. Because the number of nonzero off-diagonal elements may depend on the different scales of the penalization parameters in each method, we plot the classification accuracy against the average percentages of nonzero off-diagonals of these two classes (treated and untreated),

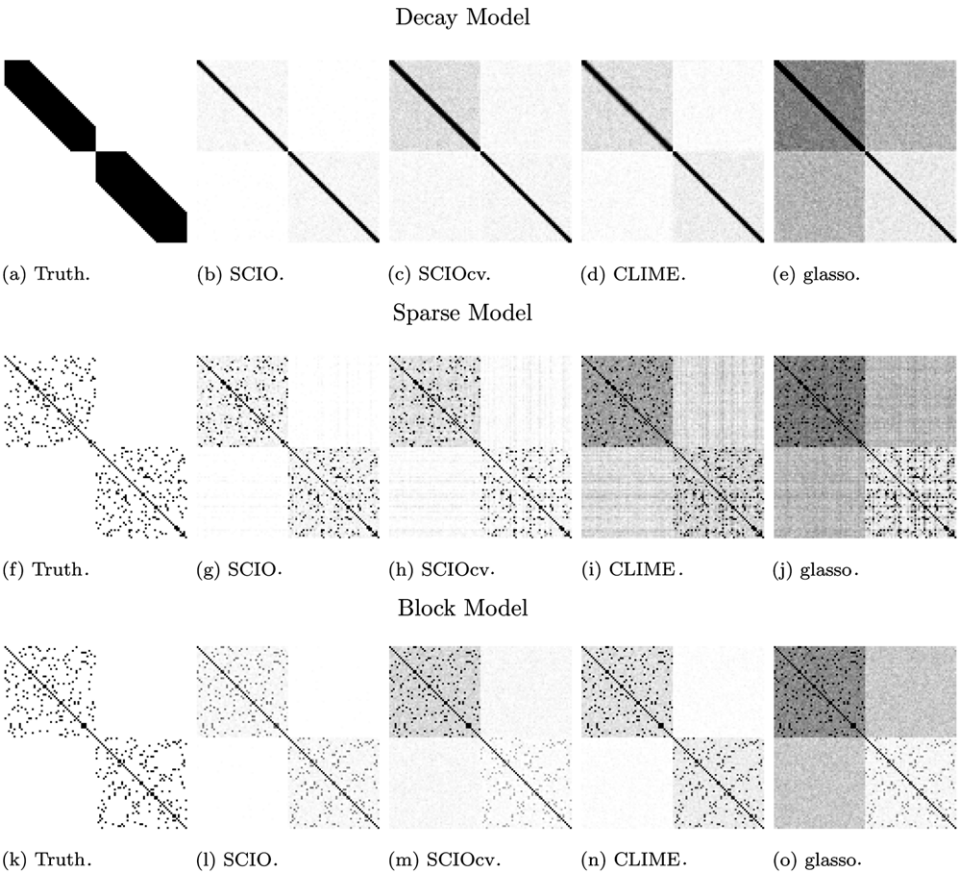


Fig. 1. Heatmaps of support recovery over 100 simulation runs (black is 100/100, white is 0/100).

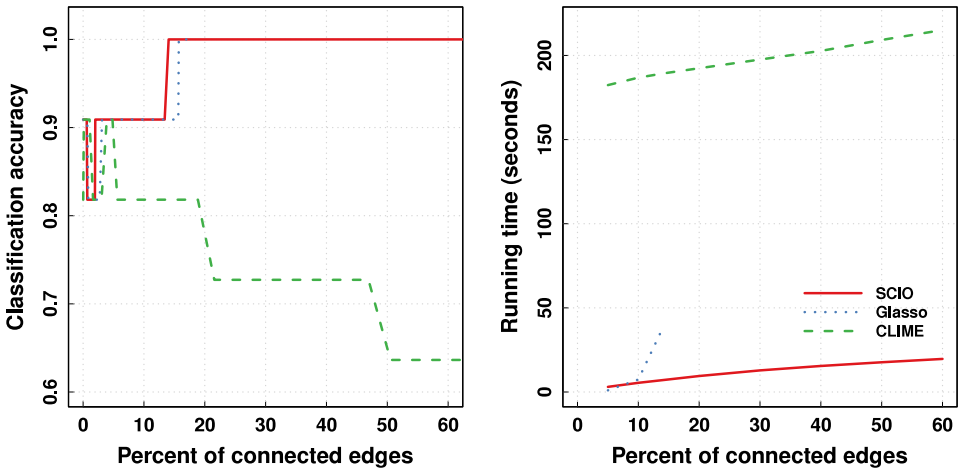
i.e. the average percentages of connected edges in two recovered graphical models for the treated and untreated respectively. The SCIOcv estimators (not shown) only differs from SCIO on how to pick λ in a data-driven way, and thus it has the identical performance as SCIO under the same λ. This figure shows that in most cases SCIO outperforms glasso and CLIME when both methods use the same number of connected edges. The SCIO estimators are stable in classification performance even if the number of connected edges increases. We are not able to plot the performance of glasso with more than 14% connected edges (corresponding to small penalization parameters), because the glasso package does not converge within 120 h. CLIME shows decreased performance when the number of connected edges increases. As a comparison with other classification algorithms, we use the same data to compare with a few other classification methods, including random forest [3], AIC penalized logistic regression, and ℓ_1 penalized logistic regression with 5-fold cross validation. Their classification accuracies are 78.6%, 90.9% and 45.6% respectively. Our classification rule compares favorably with these competing methods on this dataset.

Fig. 2(b) compares the running times against the percentages of connected edges. Because it is known that path-following algorithms may compute a sequence of solutions much faster than for a single one, we use 50 log-spaced penalization parameters from the largest (0% edges) to the designated percentages of edges, including 5%, 10%, 14%, 20%, 30%, 40%, 50% and 60%. As reported before, we are unable to plot the running times for glasso beyond 14% due to nonconvergence. SCIO takes about 2 s more than glasso when computing for 5% edges, but is much faster than glasso for 10% and more. For example, it compares favorably in the 14% case where SCIO takes only a quarter of the time of glasso. In general, the running time of SCIO grows linearly with the number of connected edges, while glasso shows exponential growth in computation time. CLIME is the slowest among all methods.

Fig. 1 of the supplementary material compares the performance of support recovery, and it shows similar advantages of SCIO as in the simulations (see Appendix A).

5.4. An fMRI dataset on attention deficit hyperactivity disorders

Attention Deficit Hyperactivity Disorder (ADHD) causes substantial impairment among about 10% of school-age children in United States. A neuroimaging study showed that the correlations between brain regions are different between typically developed children and children with such disorders [9]. The description of the data and additional results are provided in the supplementary material (see Appendix A). In there, we compare the performance of support recovery using the data



(a) Classification accuracy. (b) Running times.

Fig. 2. Comparison of classification accuracy and running times using SCIO, CLIME and glasso for the HIV dataset. Red solid lines are SCIO, green dash lines are CLIME, and blue dotted lines are glasso.

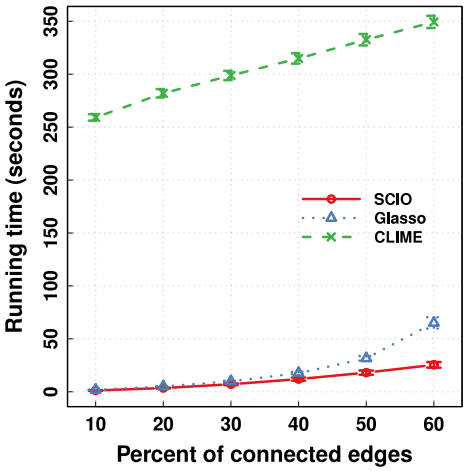


Fig. 3. Comparison of average (± 1 SE) running times for the ADHD dataset. The red solid line with circle marks is SCIO, the green dashed line with crosses is CLIME, and the blue dotted line with triangles is glasso.

from each subject, and the results suggest that SCIO has competitive performance with CLIME and glasso in recovering brain connectivities for both healthy and ADHD children.

Fig. 3 compares the running times of SCIO, CLIME, and glasso. Similar to the procedure described before, for each subject, we use path following algorithms in all methods up to the designated edge percentages, including 10%, 20%, 30%, 40%, 50% and 60%. This plot shows that the running times of SCIO grows almost linearly, and it is about 2 times faster than glasso with 60% connected edges. CLIME again is the slowest among all methods.

6. Discussion

It is possible to achieve adaptive estimation via other approaches. During the preparation of this paper, it comes to our attention that recently [21] applied a new adaptive penalized regression procedure, Scale Lasso, to the inverse covariance matrix estimation. [6] proposed an improved CLIME estimator, which runs the CLIME estimation sequentially twice. We instead analyzed cross validation as an alternative approach for this goal because cross validation remains to be popular among practitioners. It would be interesting to study the theory of cross validation for these other estimators, and to study if these adaptive approaches can also be applied to our loss.

Choosing the tuning parameters is an important problem in the practice of penalization procedures, though most of the prior theoretical results are based on some theoretical assumptions of the tuning parameters. This paper is among the first to demonstrate that a cross validated estimator for the problem of precision matrix estimation achieves the $n^{-1/2}\log^{1/2} p$ rate under the Frobenius norm. This rate may not be improved in general, because it should be minimax optimal [6], though

a rigorous justification is needed. We also note that the distribution condition (C3) in Theorem 4 is slightly stronger than (C2) and (C2*). It is an interesting problem to study if the result in Theorem 4 can be extended to more general distributions. Moreover, it would be interesting to study whether minimax rates can also be achieved under other matrix norms, such as the operator norm, using cross validation.

The rate for support recovery in Theorem 3 also coincides with the minimax optimal rate in [6]. However, $\mathcal{U}$ together with (4) is actually a smaller class than theirs. It would be interesting to explore if their minimax rate can be improved in this important sub-class. It would also be interesting to study if our results can be extended to their general matrix class.

We employ the ℓ_1 norm to enforce sparsity due to computational concerns. It has been pointed out before that the ℓ_1 penalty inheritably introduces biases, and thus it would be interesting to replace the ℓ_1 norm by other penalty forms, such as Adaptive Lasso [25] or SCAD [10]. Such extensions should be easy to implement because our loss is column-wise, similar to penalized regression. We are currently implementing these variants for future releases of our R package.

There are several other interesting directions. It would be interesting to study the precision matrix estimation under the setting that the data are generated from statistical models, while the covariance estimation problem under this setting was studied by [17]. It is also of interest to consider extending SCIO to the nonparanormal family distributions [16].

Finally, this paper only considers the setting that all the data are observed. It is an interesting problem to study the inverse covariance matrix estimation when some observations are missing. It turns out that the SCIO procedure can also be applied to the missing data setting, with additional modifications. Due to the space limitation, we will report these results elsewhere.

Acknowledgments

We would like to thank the Associate Editor and two anonymous referees for their very helpful comments, which have led to improved presentation of this paper. Weidong Liu was supported by the National Natural Science Foundation of China grants 11201298, 11322107 and 11431006, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Pujiang Program, Foundation for the Author of National Excellent Doctoral Dissertation of China and a grant from Australian Research Council. Xi Luo was partially supported by the National Institutes of Health grants P01AA019072, P20GM103645, P30AI042853, R01NS052470, and S10OD016366, a Brown University Research Seed award, a Brown Institute for Brain Science Pilot award, a Brown University faculty start-up fund, and a developmental research award from Lifespan/Brown/Tufts Center for AIDS Research.

Appendix A. Supplementary data

Supplementary material online includes the motivation of our estimator, additional descriptions of the numerical examples, and proof of the main results.

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2014.11.005.

References

[1] P.J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (6) (2008) 2577–2604. http://dx.doi.org/10.1214/08-AOS600.
[2] A. Borjabad, S. Morgello, W. Chao, S.-Y. Kim, A.I. Brooks, J. Murray, M.J. Potash, D.J. Volsky, Significant effects of antiretroviral therapy on global gene expression in brain tissues of patients with hiv-1-associated neurocognitive disorders, PLoS Pathogens 7 (9) (2011) e1002213.
[3] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[4] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 672–684.
[5] T. Cai, W. Liu, X. Luo, A constrained ℓ_1 minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 594–607.
[6] T.T. Cai, W. Liu, H.H. Zhou, Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation, Ann. Statist. (2015) in press.
[7] T. Cai, C. Zhang, H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Statist 38 (2010) 2118–2144.
[8] A. d'Aspremont, O. Banerjee, L. El Ghaoui, First-order methods for sparse covariance selection, SIAM J. Matrix Anal. Appl. 30 (1) (2008) 56–66.
[9] D.P. Dickstein, C. Gorrostieta, H. Ombao, L.D. Goldberg, A.C. Brazel, C.J. Gable, C. Kelly, D.G. Gee, X.-N. Zuo, F.X. Castellanos, et al., Fronto-temporal spontaneous resting state functional connectivity in pediatric bipolar disorder, Biol. Psychiatry 68 (9) (2010) 839–846.
[10] J. Fan, Y. Feng, Y. Wu, Network exploration via the adaptive lasso and scad penalties, Ann. Appl. Stat. 3 (2009) 521–541.
[11] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[12] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.
[13] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22.
[14] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Ann. Statist. 37 (6B) (2009) 4254–4278.
[15] S.L. Lauritzen, Graphical Models, Oxford University Press, 1996.
[16] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: semiparametric estimation of high dimensional undirected graphs, J. Mach. Learn. Res. 10 (2009) 2295–2328.
[17] X. Luo, Recovering model structures from large low rank and sparse covariance matrix estimation, arXiv preprint arXiv:1111.1133.
[18] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[19] P. Ravikumar, M.J. Wainwright, G. Raskutti, B. Yu, et al., High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, Electron. J. Stat. 5 (2011) 935–980.
[20] A.J. Rothman, P.J. Bickel, E. Levina, J. Zhu, Sparse permutation invariant covariance estimation, Electron. J. Stat. 2 (2008) 494–515. http://dx.doi.org/10.1214/08-EJS176.
[21] T. Sun, C.-H. Zhang, Sparse matrix inversion with scaled lasso, J. Mach. Learn. Res. 14 (1) (2013) 3385–3418.
[22] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.
[23] M. Yuan, High dimensional inverse covariance matrix estimation via linear programming, J. Mach. Learn. Res. 11 (2010) 2261–2286.
[24] M. Yuan, Y. Lin, Model selection and estimation in the gaussian graphical model, Boimetrika 94 (1) (2007) 19–35.
[25] S. Zhou, S. van de Geer, P. Bühlmann, Adaptive lasso for high dimensional regression and gaussian graphical modeling, arXiv preprint arXiv:0903.2515.