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 $\ell_1$ 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 $\ell_\infty$ norm and spectral norm [19]. Other penalties were also studied before. For example, the $\ell_1$ 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 $\ell_1$ 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 $\ell_\infty$ 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 $l_\infty$ 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 $\ell_1$ norm $\|\boldsymbol{A}\|_{L_1} = \max_{1\le j\le q} \sum_{i=1}^p |a_{ij}|$, the matrix $\infty$ 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 $\ell_1$ 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 $i$th row and $j$th 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.