Computational Statistics and Data Analysis 93 (2016) 390–403

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Sparse estimation of high-dimensional correlation matrices

Ying Cui [a], Chenlei Leng [b,c,*], Defeng Sun [a]

[a] Department of Mathematics, National University of Singapore, Singapore
[b] Department of Statistics, University of Warwick, United Kingdom
[c] Department of Statistics and Applied Statistics, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 12 December 2013
Received in revised form 30 September 2014
Accepted 1 October 2014
Available online 13 October 2014

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

ABSTRACT

Several attempts to estimate covariance matrices with sparsity constraints have been made. A convex optimization formulation for estimating correlation matrices as opposed to covariance matrices is proposed. An efficient accelerated proximal gradient algorithm is developed, and it is shown that this method gives a faster rate of convergence. An adaptive version of this approach is also discussed. Simulation results and an analysis of a cardiovascular microarray confirm its performance and usefulness.

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. Given observations $x_i \in \mathbb{R}^p,\ i = 1, \ldots, n$ 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 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. Yi and Zou (2013) studied a tapering procedure and Maurya (2014) developed a doubly convex method for estimating the inverse of a covariance matrix.

* Correspondence to: Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom. Tel.: +44 0 2476150779; fax: +44 0 2476524532.
E-mail address: C.Leng@warwick.ac.uk (C. Leng).

http://dx.doi.org/10.1016/j.csda.2014.10.001
0167-9473/

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

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

The major difference between this approach and that of Xue et al. (2012) is that we add hard constraints $R_{jj} = 1,\ j = 1, \ldots, p$ 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

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

For the equality constraints in (2), $\Omega = \{(j,j) : j = 1, \ldots, p\}$, 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