Journal of Multivariate Analysis 143 (2016) 107–126

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Inference for high-dimensional differential correlation matrices[☆]

T. Tony Cai, Anru Zhang[*]

Department of Statistics, The Wharton School, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 21 November 2014
Available online 11 September 2015

AMS 2010 subject classification:
primary 62H12
secondary 62F12

Keywords:
Adaptive thresholding
Covariance matrix
Differential co-expression analysis
Differential correlation matrix
Optimal rate of convergence
Sparse correlation matrix
Thresholding

ABSTRACT

Motivated by differential co-expression analysis in genomics, we consider in this paper estimation and testing of high-dimensional differential correlation matrices. An adaptive thresholding procedure is introduced and theoretical guarantees are given. Minimax rate of convergence is established and the proposed estimator is shown to be adaptively rate-optimal over collections of paired correlation matrices with approximately sparse differences. Simulation results show that the procedure significantly outperforms two other natural methods that are based on separate estimation of the individual correlation matrices. The procedure is also illustrated through an analysis of a breast cancer dataset, which provides evidence at the gene co-expression level that several genes, of which a subset has been previously verified, are associated with the breast cancer. Hypothesis testing on the differential correlation matrices is also considered. A test, which is particularly well suited for testing against sparse alternatives, is introduced. In addition, other related problems, including estimation of a single sparse correlation matrix, estimation of the differential covariance matrices, and estimation of the differential cross-correlation matrices, are also discussed.

1. Introduction

Statistical inference on the correlation structure has a wide array of applications, ranging from gene co-expression network analysis [10,21,33,12,14] to brain intelligence analysis [26]. For example, understanding the correlations between the genes is critical for the construction of the gene co-expression network. See [19,20], and [14]. Driven by these and other applications in genomics, signal processing, empirical finance, and many other fields, making sound inference on the high-dimensional correlation structure is becoming a crucial problem.

In addition to the correlation structure of a single population, the difference between the correlation matrices of two populations is of significant interest. Differential gene expression analysis is widely used in genomics to identify disease-associated genes for complex diseases. Conventional methods mainly focus on the comparisons of the mean expression levels between the disease and control groups. In some cases, clinical disease characteristics such as survival or tumor stage do not have significant associations with gene expression, but there may be significant effects on gene co-expression related to the clinical outcome [27,16,1]. Recent studies have shown that changes in the correlation networks from different stages

[☆] The research was supported in part by National Science Foundation Grants DMS-1208982 and DMS-1403708, and National Institutes of Health Grant R01 CA127334.

[*] Corresponding author.
E-mail address: anrzhang@wharton.upenn.edu (A. Zhang).

http://dx.doi.org/10.1016/j.jmva.2015.08.019
0047-259X/Published by Elsevier Inc.

of disease or from case and control groups are also of importance in identifying dysfunctional gene expressions in disease. See, for example, [11]. This differential co-expression network analysis has become an important complement to the original differential expression analysis as differential correlations among the genes may reflect the rewiring of genetic networks between two different conditions (see [27,1,11,17,13]).

Motivated by these applications, we consider in this paper optimal estimation of the differential correlation matrix. Specifically, suppose we observe two independent sets of p-dimensional i.i.d. random samples $\mathbf{X}^{(t)} = \{\mathbf{X}_1^{(t)}, \dots, \mathbf{X}_{n_t}^{(t)}\}$ with mean $\boldsymbol{\mu}_t$, covariance matrix $\boldsymbol{\Sigma}_t$, and correlation matrix $\mathbf{R}_t$, where $t = 1$ and 2. The goal is to estimate the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. A particular focus of the paper is on estimating an approximately sparse differential correlation matrix in the high dimensional setting where the dimension is much larger than the sample sizes, i.e., $p \gg \max(n_1, n_2)$. The estimation accuracy is evaluated under both the spectral norm loss and the Frobenius norm loss.

A naive approach to estimating the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ is to first estimate the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ separately and then normalize to obtain estimators $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$ of the individual correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, and finally take the difference $\hat{\mathbf{D}} = \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ as the estimator of the differential correlation matrix $\mathbf{D}$. A simple estimate of a correlation matrix is the sample correlation matrix. However, in the high-dimensional setting, the sample correlation matrix is a poor estimate. Significant advances have been made in the last few years on optimal estimation of a high-dimensional covariance matrix. Regularization methods such as banding, tapering, and thresholding have been proposed. In particular, Cai et al. [8] established the optimal rate of convergence and Cai and Yuan [7] developed an adaptive estimator of bandable covariance matrices. For sparse covariance matrices where each row and each column has relatively few nonzero entries, Bickel and Levina [4] introduced a thresholding estimator and obtained rates of convergence; Cai and Liu [5] proposed an adaptive thresholding procedure and Cai and Zhou [9] established the minimax rates of convergence for estimating sparse covariance matrices.

Structural assumptions on the individual correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ are crucial for the good performance of the difference estimator. These assumptions, however, may not hold in practice. For example, gene transcriptional networks often contain the so-called hub nodes where the corresponding gene expressions are correlated with many other gene expressions. See, for example, [3,2]. In such settings, some of the rows and columns of $\mathbf{R}_1$ and $\mathbf{R}_2$ have many nonzero entries which mean that $\mathbf{R}_1$ and $\mathbf{R}_2$ are not sparse. In genomic applications, the correlation matrices are rarely bandable as the genes are not ordered in any particular way.

In this paper, we propose a direct estimation method for the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ without first estimating $\mathbf{R}_1$ and $\mathbf{R}_2$ individually. This direct estimation method assumes that $\mathbf{D}$ is approximately sparse, but otherwise does not impose any structural assumptions on the individual correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. An adaptive thresholding procedure is introduced and analyzed. The estimator can still perform well even when the individual correlation matrices cannot be estimated consistently. For example, direct estimation can recover the differential correlation network accurately even in the presence of hub nodes in $\mathbf{R}_1$ and $\mathbf{R}_2$ as long as the differential correlation network is approximately sparse. The key is that sparsity is assumed for $\mathbf{D}$ and not for $\mathbf{R}_1$ or $\mathbf{R}_2$.

Theoretical performance guarantees are provided for direct estimator of the differential correlation matrix. Minimax rates of convergence are established for the collections of paired correlation matrices with approximately sparse differences. The proposed estimator is shown to be adaptively rate-optimal. In comparison to adaptive estimation of a single sparse covariance matrix considered in Cai and Liu [5], both the procedure and the technical analysis of our method are different and more involved. Technically speaking, correlation matrix estimators are harder to analyze than those of covariance matrices and the two-sample setting in our problem further increases the difficulty.

Numerical performance of the proposed estimator is investigated through simulations. The results indicate significant advantage of estimating the differential correlation matrix directly. The estimator outperforms two other natural alternatives that are based on separate estimation of $\mathbf{R}_1$ and $\mathbf{R}_2$. To further illustrate the merit of the method, we apply the procedure to the analysis of a breast cancer dataset from the study by van de Vijver et al. [30] and investigate the differential co-expressions among genes in different tumor stages of breast cancer. The adaptive thresholding procedure is applied to analyze the difference in the correlation alternation in different grades of tumor. The study provides evidence at the gene co-expression level that several genes, of which a subset has been previously verified, are associated with the breast cancer.

In addition to optimal estimation of the differential correlation matrix, we also consider hypothesis testing of the differential correlation matrices, $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$ vs. $H_1 : \mathbf{R}_1 - \mathbf{R}_2 \neq 0$. We propose a test which is particularly well suited for testing again sparse alternatives. The same ideas and techniques can also be used to treat other related problems. We also consider estimation of a single sparse correlation matrix from one random sample, estimation of the differential covariance matrices as well as estimation of the differential cross-correlation matrices.

The rest of the paper is organized as follows. Section 2 presents in detail the adaptive thresholding procedure for estimating the differential correlation matrix. The theoretical properties of the proposed estimator are analyzed in Section 3. In Section 4, simulation studies are carried out to investigate the numerical performance of the thresholding estimator and Section 5 illustrates the procedure through an analysis of a breast cancer dataset. Hypothesis testing on the differential correlation matrices is discussed in Section 6.1, and other related problems are considered in the rest of Section 6. All the proofs are given in the Appendix.