Journal of Multivariate Analysis 149 (2016) 124–143

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

More powerful tests for sparse high-dimensional covariances matrices

Liuhua Peng[a,*], Song Xi Chen[a,b], Wen Zhou[c]

[a] Department of Statistics, Iowa State University, Ames, IA 50011, USA

[b] Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, PR China

[c] Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 2 September 2015
Available online 6 April 2016

AMS 2010 subject classification:
62H15

Keywords:
Banding estimation
High-dimensional inference
Hypothesis testing
Sparse covariance matrix

ABSTRACT

This paper considers improving the power of tests for the identity and sphericity hypotheses regarding high dimensional covariance matrices. The power improvement is achieved by employing the banding estimator for the covariance matrices, which leads to significant reduction in the variance of the test statistics in high dimension. Theoretical justification and simulation experiments are provided to ensure the validity of the proposed tests. The tests are used to analyze a dataset from an acute lymphoblastic leukemia gene expression study for an illustration.

1. Introduction

This paper is interested in testing hypothesis for high-dimensional covariance matrices, $\mathbf{\Sigma}$, of a p-dimensional random vector $\mathbf{X}$. In practice, it is often of scientific interest to test whether or not a prescribed dependence structure is supported by data, for instance

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0 \quad \text{vs.} \quad H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0 \tag{1.1}$$

and

$$H_0 : \mathbf{\Sigma} = \sigma^2\mathbf{\Sigma}_0 \quad \text{vs.} \quad H_1 : \mathbf{\Sigma} \neq \sigma^2\mathbf{\Sigma}_0 \quad \text{for some unknown } \sigma^2 > 0, \tag{1.2}$$

for a known non-degenerate covariance matrix $\mathbf{\Sigma}_0$. Among many practical applications, genomic studies usually motivate (1.1) or (1.2): it is not uncommon to postulate a grouping structure among genes of interest such that genes are not correlated across groups (Katsani et al. [25]), i.e. $\mathbf{\Sigma}$ is presumed in a diagonal block shape upon permutations. Additionally, in the fields of image segmentation, epidemiology and ecology, large numbers of pixels or population abundances are collected across the spatial domain. Certain spatial autocorrelations are usually prescribed for fitting data to a parametric or semiparametric model for predictions (Bolker [9], Cressie [16]). It is important to verify whether or not these hypothetical dependence structures are supported by data.

[*] Corresponding author.
E-mail addresses: penglh@iastate.edu (L. Peng), songchen@iastate.edu, csx@gsm.pku.edu.cn (S.X. Chen), riczw@stat.colostate.edu (W. Zhou).

http://dx.doi.org/10.1016/j.jmva.2016.03.008

For identically and independently distributed data $\mathbf{X}_1, \dots, \mathbf{X}_n$ with unknown common mean μ and covariance $\mathbf{\Sigma}_0$, linear transform $\mathbf{\Sigma}_0^{-1/2}\mathbf{X}_i$ reduces (1.1) and (1.2) to

$$H_0 : \mathbf{\Sigma} = \mathbf{I}_p \quad \text{vs.} \quad H_1 : \mathbf{\Sigma} \neq \mathbf{I}_p, \tag{1.3}$$

and

$$H_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p \quad \text{vs.} \quad H_1 : \mathbf{\Sigma} \neq \sigma^2\mathbf{I}_p, \tag{1.4}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix. Hypotheses (1.3) and (1.4) are called the identity and sphericity hypothesis, respectively. For fixed p, likelihood ratio test has been developed and widely applied. We refer to Anderson [1] for more details. Let $\widehat{\mathbf{\Sigma}}$ be the sample covariance matrix. John [23,24] and Nagao [27] showed that for a fixed p, test statistics

$$\mathrm{V}_n = p^{-1}\mathrm{tr}\{(\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)^2\} \quad \text{and} \quad \mathrm{U}_n = p^{-1}\mathrm{tr}[\{p\widehat{\mathbf{\Sigma}}/\mathrm{tr}(\widehat{\mathbf{\Sigma}}) - \mathbf{I}_p\}^2] \tag{1.5}$$

provide the most powerful invariant tests for both the identity and sphericity hypotheses against the local alternatives. Traditional tests, however, are not applicable to the large p, small n paradigm since the sample covariance matrix is singular with probability one whenever $p > n$ and is no longer a consistent estimator if p is not a smaller order of n (Bai and Yin [5], Bai et al. [4]).

Tests for covariance matrices suited for the high dimensionality have been developed over the recent years. Ledoit and Wolf [26] established the asymptotic properties of statistics in (1.5) for $p/n \to c \in (0, +\infty)$ and proposed tests for identity (1.4) and sphericity (1.4) under the Gaussian assumption. Jiang [22] developed a sphericity test based on the max-type statistic $L_n = \max_{1\le i<j\le p} |\hat{\rho}_{ij}|$, where $\hat{\rho}_{ij}$ denotes the sample correlation coefficient between the ith and jth components of $\mathbf{X}$. With the aid of the random matrix theory, Bai et al. [2] proposed a modified likelihood ratio statistic for testing (1.3) for $p/n \to y \in (0, 1)$. To avoid the issue of inconsistency of $\widehat{\mathbf{\Sigma}}$ when $p > n$, Chen et al. [14] proposed U-statistic based testing procedures for both the identity and sphericity hypotheses. Their tests require much relaxed assumptions on the data distribution, and allow p diverges in n in any rates. See, for example, Cai and Jiang [10], Hallin and Paindaveine [21], Srivastava and Yanagihara [35], Srivastava and Reid [34], Srivastava et al. [36], Zou et al., [42] for alternative test formulations, and Bai et al. [2], Schott [32], Zheng et al., [41], Qiu and Chen [29] for related works. One limitation of these high dimensional tests is a loss of power under sparse high dimension situations, largely due to a rapid increase in the variance of the test statistic as the p gets larger. For instance, in the formulation of the identity test, estimation of the discrepancy measure $p^{-1}\mathrm{tr}\{(\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)^2\}$ involves all the entries of the sample covariance. As a result, the test statistics incurs larger variation as the dimension gets larger. The increased variance dilutes the signal $p^{-1}\mathrm{tr}\{(\widehat{\mathbf{\Sigma}} - \mathbf{I}_p)^2\}$ of the test and hence brings down its power.

While we are gathering more dimensions in the data as more features are recorded, the information content of the data is not necessarily increasing at the same rate as the dimension. Indeed, it is commonly acknowledged that parameters associated with high dimensional data can be sparse in the sense of that many of the parameters are either zero or taking small values. This was the rationale behind the proposal of LASSO in Tibshirani [37] as well as other regularization-based estimations in regression and covariance matrices; see Bickel and Levina [7], Cai et al. [11], Fan and Li [18], Rothman et al. [31]. We consider in this paper tests for covariance matrices by utilizing the regularization-based estimation constructed for a specific class of sparse covariance matrices, the so-called bandable covariances, introduced by Bickel and Levina [8]. The bandable class is naturally suited as alternative hypotheses to the null identity and the sphericity hypotheses. Specifically, we formulate the test statistics by employing the banded covariance estimator proposed in Bickel and Levina [8]. This allows us to take advantage of the knowledge of sparsity in the $\mathbf{\Sigma}$. We demonstrate in this paper that the new test formulations have a remarkable power enhancement over the existing high dimensional tests for the covariance which do not utilize the sparsity information.

The rest of the paper is organized as follows. We introduce our motivations in Section 2 and present the testing procedures in Section 3. The theoretical properties of the proposed tests are also investigated in Section 3. Section 4 is devoted to a discussion on the selection of k for the proposed tests. Numerical results are displayed in Section 5 to investigate the performance of the tests in practice. Both simulation studies and applications of the proposed tests to an acute lymphoblastic leukemia gene expression dataset are reported. The last section concludes the article with a brief discussion, and technical proofs are given in the Appendix A. Supplementary material contains more details on the numerical studies (see Appendix B).

2. Motivations and preliminaries

Our investigation is motivated by the notion of the bandable covariance class introduced by Bickel and Levina [8], which is defined as

$$\begin{aligned}\mathcal{U}(\varepsilon_0, \mathrm{C}, \alpha) = \Bigg\{ \mathbf{\Sigma} = (\sigma_{ij})_{1\le i,j\le p} : \max_j \sum_{|i-j|>k_0} |\sigma_{ij}| \le \mathrm{C}k_0^{-\alpha} \quad \text{for all } k_0 \ge 0, \\ 0 < \varepsilon_0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le 1/\varepsilon_0 \Bigg\},\end{aligned} \tag{2.1}$$