Computational Statistics and Data Analysis 104 (2016) 10–23

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Power computation for hypothesis testing with high-dimensional covariance matrices

Ruitao Lin [a], Zhongying Liu [b], Shurong Zheng [b,*], Guosheng Yin [a]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, China

[b] School of Mathematics & Statistics and KLAS, Northeast Normal University, China

ARTICLE INFO

Article history:
Received 4 June 2015
Received in revised form 23 February 2016
Accepted 11 May 2016
Available online 6 June 2016

Keywords:
Central limit theorem
Confidence interval
High-dimensional covariance matrix
Hypothesis testing
Power calculation
Stieltjes transform

ABSTRACT

Based on the random matrix theory, a unified numerical approach is developed for power calculation in the general framework of hypothesis testing with high-dimensional covariance matrices. In the central limit theorem of linear spectral statistics for sample covariance matrices, the theoretical mean and covariance are computed numerically. Based on these numerical values, the power of the hypothesis test can be evaluated, and furthermore the confidence interval for the unknown parameters in the high-dimensional covariance matrix can be constructed. The validity of the proposed algorithms is well supported by a convergence theorem. Our numerical method is assessed by extensive simulation studies, and a real data example of the S&P 100 index data is analyzed to illustrate the proposed algorithms.

1. Introduction

High-dimensional data become commonplace in the modern big data era, particularly in the fields of finance, genomics, informatics, image analysis, and so on. Estimation and inference based on the sample covariance matrix are fundamentally important in multivariate statistical analysis. However, when the dimensionality grows with the sample size, the conventional estimate of the covariance matrix is inconsistent (Bai and Yin, 1993), which thus leads to poor inference procedures in the high-dimensional setting. To address the inconsistency issue, extensive research has been conducted on estimation of high-dimensional covariance matrices. In particular, the random matrix theory is able to attenuate the randomness of the sample covariance matrix, and thus is widely used as a popular tool to achieve efficient statistical inference on the high-dimensional covariance matrices. Bai and Silverstein (2004) and Zheng et al. (2015) established the central limit theorem (CLT) of linear spectral statistics (LSS) for large dimensional sample covariance matrices when the dimension p and the sample size n increase proportionally; that is, $p/n \to c$ where c is some constant in $(0, \infty)$.

Based on the CLT of LSS of sample covariance matrices, extensive research has been carried out on testing the high-dimensional covariance matrix structures. For example, Bai et al. (2009) proposed to test whether the high-dimensional covariance matrix is equal to a specified matrix. Nevertheless, there is no explicit power function, because the asymptotic mean and covariance in the CLT of LSS of large dimensional covariance matrices are expressed by the contour integrals that cannot be derived in closed forms. In hypothesis testing, power functions are important for evaluating the effectiveness of the testing methods. For example, Chen et al. (2012), Onatski et al. (2013) and Wang (2014) studied the power functions for

* Corresponding author.
E-mail address: zhengsr@nenu.edu.cn (S. Zheng).

http://dx.doi.org/10.1016/j.csda.2016.05.008

different covariance testing problems. However, all these power functions are derived for some specific high-dimensional covariance structure, and may not be generally applicable.

Our method is motivated by a real example in financial studies. As a preliminary step, the equicorrelation structure is commonly adopted for analyzing the correlation structures among financial assets (Ledoit and Wolf, 2004; Kyj et al., 2009; Engle and Kelly, 2012). Given the return data on a set of stocks, it is of great importance to examine whether the stock returns are equicorrelated. Therefore, the hypothesis testing procedure can be used to reduce the risk of inappropriate modeling for the covariance matrix. In addition to testing the covariance structure, the confidence intervals of the equicorrelation coefficients can provide useful information on the movement of stock prices. Unfortunately, it is often challenging to compute the power and confidence intervals as the number of assets is typically large relative to the number of observations.

Our goal is to provide a general numerical algorithm for computing the asymptotic mean and covariance in the CLT of LSS of large-dimensional sample covariance matrices in Zheng et al. (2015). The numerical algorithm can be used to obtain the power functions of high-dimensional covariance tests. Moreover, when the covariance matrix $\boldsymbol{\Sigma}$ is characterized by some low dimensional parameters $\boldsymbol{\rho}$, say $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\rho})$, the confidence region of $\boldsymbol{\rho}$ can be numerically constructed as well.

The rest of the paper is arranged as follows. Section 2 establishes a general lemma based on Zheng et al. (2015) and provides a numerical algorithm for computation of the theoretical mean and variance in the CLT of LSS of sample covariance matrices. The power function for hypothesis testing and confidence intervals for the parameters in the covariance matrix are also derived. Section 3 exhibits some simulation studies, and Section 4 conducts a real data analysis on the S&P 100 index data for illustration of our proposal. Section 5 concludes with a discussion and technical details are delineated to Appendix.

2. Power and confidence interval

2.1. Motivating example: Hypothesis testing of a covariance matrix

Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a sample from the population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\mathbf{y}_i$ is a p-dimensional random vector, $i = 1, \ldots, n$. It is of interest in multivariate statistical analysis to test

$$H_0 : \boldsymbol{\Sigma} = \mathbf{I}_p,$$

where $\mathbf{I}_p$ is the identity matrix of dimension p (Anderson, 2003). For a more general framework, the identity matrix $\mathbf{I}_p$ in the null hypothesis H_0 can be replaced by any positive-definite covariance matrix $\boldsymbol{\Sigma}_0$. Many well-known test statistics are constructed as functionals of the eigenvalues of the sample covariance matrix, which is defined as

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})^T ,$$

where $\bar{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} \mathbf{y}_i$ is the sample mean and the superindex "^{T}" denotes the transpose. Hereafter, such statistics are referred as linear spectral statistics (LSS). For example, the likelihood ratio test (LRT) statistic for the Gaussian case is given by

$$T_n = \sum_{j=1}^{p} \mathrm{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p,$$

where "tr" denotes the trace of a matrix. Bai et al. (2009) established the CLT for T_n under the null hypothesis H_0 for high-dimensional cases when $c_n = p/n \to c \in (0, \infty)$, while the theoretical power function was not derived despite its importance in hypothesis testing. Moreover, when $\boldsymbol{\Sigma}$ is characterized by some parameters $\boldsymbol{\rho}$, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\rho})$, it is also of interest to construct the confidence region for $\boldsymbol{\rho}$.

2.2. CLT of LSS for sample covariance matrices

Let $\{\lambda_1 \le \cdots \le \lambda_p\}$ be the eigenvalues of the population covariance matrix $\boldsymbol{\Sigma}$. The empirical spectral distribution (ESD) of $\boldsymbol{\Sigma}$ is given by

$$H_p(x) = \frac{1}{p} \sum_{j=1}^{p} \delta\{\lambda_j \le x\}, \quad \text{for any } x \in R$$

where $\delta\{\cdot\}$ is the indicator function and R is the real line. Let $\{\hat{\lambda}_1 \le \cdots \le \hat{\lambda}_p\}$ be the eigenvalues of the sample covariance matrix $\mathbf{S}_n$, and correspondingly the ESD of $\mathbf{S}_n$ is

$$F_n(x) = \frac{1}{p} \sum_{j=1}^{p} \delta\{\hat{\lambda}_j \le x\}, \quad \text{for any } x \in R.$$

The Marčenko–Pastur scheme is adopted; that is,

$$c_n = p/n \to c \in (0, \infty), \quad \text{as } p, n \to \infty.$$