Computational Statistics and Data Analysis 95 (2016) 75–82

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Testing the order of a population spectral distribution for high-dimensional data

Yingli Qin [a], Weiming Li [b,*]

[a] Department of Statistics and Actuarial Science, University of Waterloo, Canada N2L 3G1

[b] Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Article history:
Received 27 October 2014
Received in revised form 5 September 2015
Accepted 16 September 2015
Available online 1 October 2015

Keywords:
Covariance matrix
High-dimension
Hypothesis testing
Population spectral distribution

ABSTRACT

Large covariance matrices play a fundamental role in various high-dimensional statistics. Investigating the limiting behavior of the eigenvalues can reveal informative structures of large covariance matrices, which is particularly important in high-dimensional principal component analysis and covariance matrix estimation. In this paper, we propose a framework to test the number of distinct population eigenvalues for large covariance matrices, i.e. the order of a Population Spectral Distribution. The limiting distribution of our test statistic for a Population Spectral Distribution of order 2 is developed along with its (N, p) consistency, which is clearly demonstrated in our simulation study. We also apply our test to two classical microarray datasets.

1. Introduction

Large-scale statistical inference involving covariance matrices is developing rapidly. This type of inference often relies on some specific structural assumption about covariance matrices; e.g., in terms of population eigenvalues and eigenvectors. Johnstone (2001) outlines a spiked covariance model which assumes that there only exist a fixed number r of population eigenvalues separated from the bulk. Cai et al. (2013) establish the optimal rates of convergence for estimating the principal subspace under sparse Principal Component Analysis (PCA) assumptions. Sparse PCA assumes that r leading eigenvectors are sparse and the number of leading eigenvalues can grow with the sample size N and the dimension p. Berthet and Rigollet (2013) consider an extreme case in which $r = 1$ and the covariance matrix can be modeled as $I_p + \theta \nu \nu^T$, where I_p is the $p \times p$ identity matrix, ν is a unit length sparse vector and $\theta \in \mathbb{R}^+$. Birnbaum et al. (2013) establish minimax rates of convergence and adaptive estimation of $r \geq 1$ individual leading eigenvectors when the ordered entries of each eigenvector have rapid decay. Clearly, it is of great interest to investigate the behavior of eigenvalues of large covariance matrices, or equivalently their spectral distributions.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a sequence of i.i.d. random vectors in $\mathbb{R}^p$ with a common population covariance matrix Σ_p. The sample covariance matrix is

$$S_n = \frac{1}{n} \sum_{k=1}^{N} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})',$$

* Corresponding author.
E-mail addresses: yingli.qin@uwaterloo.ca (Y. Qin), liwm601@gmail.com (W. Li).

http://dx.doi.org/10.1016/j.csda.2015.09.009
0167-9473/

where $\bar{\mathbf{x}} = (1/N)\sum_{k=1}^{N}\mathbf{x}_k$ and $n = N - 1$ represents the degree of freedom. Let $(\sigma_i)_{1\le i\le p}$ be the p eigenvalues of the population covariance matrix Σ_p. We are particularly interested in the following spectral distribution:

$$H_p := \frac{1}{p}\sum_{i=1}^{p}\delta_{\sigma_i},$$

where δ_b denotes the Dirac point measure at b. Following the random matrix theory, both the dimension p and the sample size N grow to infinity. It is then natural to assume that H_p weakly converges to a limiting distribution H, as $p \to \infty$. We refer both H_p and its limit H as Population Spectral Distribution (PSD).

In "large p, large N" framework, El Karoui (2008) proposes a nonparametric estimator of the PSD. For $N = 500$ and $p = 100$, the proposed estimator performs relatively well for three models considered in their simulation: $\Sigma_p = I_p$, $H = 0.5\delta_1 + 0.5\delta_2$, and a Toeplitz covariance matrix. The second model suggests that the population eigenvalues can be grouped into two clusters of equal probability, which is a special case of order 2 discrete PSD, while $\Sigma_p = I_p$ is of order 1 and the third model is of order infinite (continuous PSD). In this paper, we consider testing for the order of PSDs of large covariance matrices. Hypotheses of interest can be viewed as plausible cases of spiked models in Cai et al. (2013), and extensions of the spiked model in Johnstone (2001) and Berthet and Rigollet (2013). Such extensions allow r, the number of leading eigenvalues, to grow with N and p. Furthermore, our hypotheses of interest certainly lay the ground for various estimation procedures of discrete PSD, e.g., Rao et al. (2008) and Bai et al. (2010).

Specifically, assume that there are only two distinct population eigenvalues a_1 and a_2, and their multiplicities are $p - r$ and r, respectively. Assume also that

$$\frac{p-r}{p} \to w_1 \quad \text{and} \quad \frac{r}{p} \to w_2,$$

as $p \to \infty$, then the PSD of the covariance matrix can be modeled as

$$H = w_1\delta_{a_1} + w_2\delta_{a_2}, \qquad w_1 + w_2 = 1.$$

This model has been considered in El Karoui (2008), Rao et al. (2008), Bai et al. (2010), Li et al. (2013) and Li and Yao (2014), where only the estimation of H is concerned. The question that remains unanswered is whether we can find statistical evidence to support the assumption that only two distinct mass points appear in H.

To exclude the possibility of $a_1 = a_2$, one can use sphericity tests in Fisher et al. (2010), Fisher (2012) and some references therein. To claim that distinct mass points in H are no more than two, we consider the following hypotheses:

$$H_0 : k \le 2 \quad \text{vs.} \quad H_1 : k > 2, \tag{1}$$

where k represents the order of H. The main goal of this paper is to propose an original test procedure for hypotheses in (1). It should be noticed that the order of H may differ from H_p. For instance, the single spike model (Johnstone, 2001) is such a case, where the order of H_p is always 2 but the order of H is 1.

The rest of the paper is organized as follows. In the next section, we propose a framework to test the order of a PSD. In Section 3, we formulate a specific test for a PSD of order $k \le 2$. Section 4 reports simulation results and Section 5 presents real data analysis. Conclusions and remarks are presented in the last section.

2. A general strategy for testing the order of a PSD

In this section, we propose a new framework to test the following hypotheses:

$$H_0 : k \le k_0 \quad \text{vs.} \quad H_1 : k > k_0, \tag{2}$$

where k stands for the order of a PSD H and $k_0 \in \mathbb{N}$.

For $l = 0, 1, 2, \ldots,$ let

$$\gamma_l = \int t^l dH_p(t) \quad \text{and} \quad \tilde{\gamma}_l = \int t^l dH(t)$$

be the lth integer moments of H_p and H, respectively. Their common estimators are denoted by $\hat{\gamma}_l$. Note that $\hat{\gamma}_0 = \gamma_0 = \tilde{\gamma}_0 = 1$.

Define the qth Hankel matrix related to H as

$$\Gamma(H, q) = \begin{pmatrix} \tilde{\gamma}_0 & \tilde{\gamma}_1 & \cdots & \tilde{\gamma}_{q-1} \\ \tilde{\gamma}_1 & \tilde{\gamma}_2 & \cdots & \tilde{\gamma}_q \\ \vdots & \vdots & & \vdots \\ \tilde{\gamma}_{q-1} & \tilde{\gamma}_q & \cdots & \tilde{\gamma}_{2q-2} \end{pmatrix}.$$

Similarly, we may give the Hankel matrix related to H_p, denoted by $\Gamma(H_p, q)$, by replacing $\tilde{\gamma}_l$'s with γ_l's in the matrix.