Journal of Statistical Planning and Inference 180 (2017) 108–123

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

# Regularized LRT for large scale covariance matrices: One sample problem

Young-Geun Choi [a], Chi Tim Ng [b], Johan Lim [a,*]

[a] Department of Statistics, Seoul National University, Seoul, Republic of Korea

[b] Department of Statistics, Chonnam National University, Gwangju, Republic of Korea

ARTICLE INFO

*Article history:*
Received 7 July 2015
Received in revised form 15 June 2016
Accepted 21 June 2016
Available online 24 August 2016

*Keywords:*
Asymptotic normality
Covariance matrix estimator
Identity covariance matrix
High dimensional data
Linear shrinkage estimator
Linear spectral statistics
Random matrix theory
Regularized likelihood ratio test
Spiked covariance matrix

ABSTRACT

The main theme of this paper is a modification of the likelihood ratio test (LRT) for testing high dimensional covariance matrix. Recently, the correct asymptotic distribution of the LRT for a large-dimensional case (the case $p/n$ approaches to a constant $\gamma \in (0, 1]$) is specified by researchers. The correct procedure is named as corrected LRT. Despite of its correction, the corrected LRT is a function of sample eigenvalues that are suffered from redundant variability from high dimensionality and, subsequently, still does not have full power in differentiating hypotheses on the covariance matrix. In this paper, motivated by the successes of a linearly shrunken covariance matrix estimator (simply shrinkage estimator) in various applications, we propose a regularized LRT that uses, in defining the LRT, the shrinkage estimator instead of the sample covariance matrix. We compute the asymptotic distribution of the regularized LRT, when the true covariance matrix is the identity matrix and a spiked covariance matrix. The obtained asymptotic results have applications in testing various hypotheses on the covariance matrix. Here, we apply them to testing the identity of the true covariance matrix, which is a long standing problem in the literature, and show that the regularized LRT outperforms the corrected LRT, which is its non-regularized counterpart. In addition, we compare the power of the regularized LRT to those of recent non-likelihood based procedures.



## 1. Introduction

High dimensional data are now prevalent everywhere that include genomic data in biology, financial times series data in economics, and natural language processing data in machine learning and marketing. The traditional procedures that assume that sample size $n$ is large and dimension $p$ is fixed are not valid anymore for the analysis of high dimensional data. A significant amount of research are made to resolve the difficulty from the dimensionality of the data.

This paper considers the inference problem of large scale covariance matrix whose dimension $p$ is large compared to the sample size $n$. To be specific, we are interested in testing whether the covariance matrix equals to a given matrix; $\mathcal{H}_0 : \Sigma = \Sigma_0$, where $\Sigma_0$ can be set $\mathrm{I}_p$ without loss of generality. The likelihood ratio test (LRT) statistic for testing $\mathcal{H}_0 : \Sigma = \mathrm{I}_p$ is defined by

$$\mathrm{LRT} = \mathrm{tr}(\mathbf{S}_n) - \log|\mathbf{S}_n| - p = \sum_{i=1}^{p}(l_i - \log l_i - 1),$$

* Corresponding author.
*E-mail address:* johanlim@snu.ac.kr (J. Lim).

http://dx.doi.org/10.1016/j.jspi.2016.06.006
0378-3758/

where $\mathbf{S}_n$ is the unbiased and centered sample covariance matrix and $l_i$ is the $i$th largest eigenvalue of the sample covariance matrix. When $p$ is fixed, LRT follows the chi-square distribution with degrees of freedom $p(p+1)/2$ asymptotically. However, this does not hold when $p$ increases. Its correct asymptotic distribution is computed by Bai et al. (2009) for the case $p/n$ approaches $\gamma \in (0, 1)$ and both $n$ and $p$ increase. They further numerically show that their asymptotic normal distribution defines a valid procedure for testing $\mathcal{H}_0 : \Sigma = \mathrm{I}_p$. The results of Bai et al. (2009) are refined by Jiang et al. (2012), which include the asymptotic null distribution for the case $\gamma = 1$. Despite of the correction of the null distribution, the sample covariance is known to have redundant variability when $p$ is large, and it still remains a general question that the LRT is asymptotically optimal for testing problem in the $n$, $p$ large scheme.

In this paper, it is shown that the corrected LRT can be further improved by introducing a linear shrinkage component. In detail, we consider a modification of the LRT, denoted by regularized LRT (rLRT), defined by

$$\mathrm{rLRT} = \mathrm{tr}\big(\widehat{\Sigma}\big) - \log|\widehat{\Sigma}| - p = \sum_{i=1}^{p}\big(\psi_i - \log\psi_i - 1\big), \tag{1}$$

where $\widehat{\Sigma}$ is a regularized covariance matrix and $\psi_i$ is the $i$th largest eigenvalue of $\widehat{\Sigma}$. Here, we consider the regularization via linear shrinkage:

$$\widehat{\Sigma} \equiv \lambda\mathbf{S}_n + (1-\lambda)\mathrm{I}_p. \tag{2}$$

We also occasionally notate $\mathrm{rLRT}(\lambda)$ to emphasize the use of the value $\lambda$. The linearly shrunken sample covariance matrix (simply shrinkage estimator) is known to reduce expected estimation loss of the sample covariance matrix (Ledoit and Wolf, 2004). It is also successfully applied to many high-dimensional procedures to resolve the dimensionality problem. For example, Schäfer and Strimmer (2005) reconstruct a gene regulatory network from microarray gene expression data using the inverse of a regularized covariance matrix. Chen et al. (2011) propose a modified Hotelling's $T^2$-statistic for testing high dimensional mean vectors and apply it to finding differentially expressed gene sets. We are motivated by the success of above examples and inspect whether the power can be improved by the reduced variability via linear shrinkage. To the best of our knowledge, our work is the first time to apply the linear shrinkage to the covariance matrix testing problem itself.

We derive the asymptotic distribution of the proposed $\mathrm{rLRT}(\lambda)$ under two scenarios, (i) when $\Sigma = \mathrm{I}_p$ for the null distribution, and additionally (ii) when $\Sigma = \Sigma_{\mathrm{spike}}$ for power study. Here $\Sigma_{\mathrm{spike}}$ means a covariance matrix from the spiked population model (Johnstone, 2001), roughly it is defined as a covariance matrix whose eigenvalues are all 1's but some finite nonunit 'spike'. The spiked covariance matrix assumed here includes the well known compound symmetry matrix $\Sigma_{\mathrm{cs}}(\rho) = \mathrm{I}_p + \rho\mathrm{J}_p$, where $\mathrm{J}_p$ is the $p \times p$ matrix of ones. The main results show that $\mathrm{rLRT}(\lambda)$ has normal distribution in asymptotic under both (i) and (ii); their asymptotic means are different but the variances are same. The main results are useful in testing various one sample covariance matrices. To be specific, first, in testing $\mathcal{H}_0 : \Sigma = \mathrm{I}_p$, (i) provides the asymptotic null distribution of $\mathrm{rLRT}(\lambda)$. Second, combining (i) and (ii) provides the asymptotic power for an arbitrary spiked alternative covariance matrix including $\Sigma_{\mathrm{cs}}(\rho)$. Finally, the results with $\lambda = 1$ provide various asymptotic distributions of the corrected LRT. Among these many applications, in this paper, we particularly focus on the LRT for testing $\mathcal{H}_0 : \Sigma = \mathrm{I}_p$, which has long been studied by many researchers (Anderson, 2003; Ledoit and Wolf, 2002; Bai et al., 2009; Chen et al., 2010; Jiang et al., 2012).

The paper is organized as follows. In Section 2, we briefly review results of the random matrix theory that are essential to the asymptotic theory of the proposed rLRT. The results include the limit of empirical spectral distribution (ESD) of the sample covariance matrix and the central limit theorem (CLT) for linear spectral statistics (LSS). In Section 3, we formally define the rLRT, and prove the asymptotic normality of the rLRT when the true covariance matrix $\Sigma$ is $\mathrm{I}_p$ or $\Sigma_{\mathrm{spike}}$. In Section 4, the results developed in Section 3 are applied to testing $\mathcal{H}_0 : \Sigma = \mathrm{I}_p$. Numerical study is provided to compare the powers of the LRT and other existing methods including the corrected LRT and other non-LRT tests by Ledoit and Wolf (2002) and Chen et al. (2010). In Section 5, we conclude the paper with discussions of several technical details of the rLRT, for example, close spiked eigenvalues.

## 2. Random matrix theory

In this section, some useful properties of linear spectral statistics of the sample covariance matrix are introduced. The true covariance matrix $\Sigma$ is identity or that from a spiked population model.

The following notation is used throughout the paper. Let M be a real-valued symmetric matrix of size $p \times p$ and $\alpha_j(\mathrm{M})$ be the $j$th largest eigenvalue of the matrix M with natural labeling $\alpha_p(\mathrm{M}) \le \cdots \le \alpha_1(\mathrm{M})$. The *spectral distribution* (SD) for M is defined by

$$F^{\mathrm{M}}(t) := \frac{1}{p}\sum_{j=1}^{p}\delta_{\alpha_j(\mathrm{M})}(t), \quad t \in \mathbb{R},$$

where $\delta_\alpha(t)$ is a point mass function that can be also written, with notational abuse, as $\delta_\alpha(t) = I(\alpha \le t)$. Here, $I(A)$ denotes the indicator function of a set $A$.