Journal of Multivariate Analysis 156 (2017) 57–69

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Moderate deviation principles for classical likelihood ratio tests of high-dimensional normal distributions

Hui Jiang [a], Shaochen Wang [b,*]

[a] Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing, PR China

[b] School of Mathematics, South China University of Technology, Guangzhou, PR China

ARTICLE INFO

Article history:
Received 4 December 2015
Available online 12 February 2017

AMS 2010 subject classifications:
60F10
62H15

Keywords:
High-dimensional normal distribution
Likelihood ratio tests
Moderate deviations

ABSTRACT

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a Gaussian random vector of dimension $p < n$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Based on this sample, we consider the moderate deviation principle of the modified likelihood ratio test (LRT) for the testing problem $\mathscr{H}_0 : \boldsymbol{\Sigma} = \lambda \mathbf{I}_p$ versus $\mathscr{H}_1 : \boldsymbol{\Sigma} \neq \lambda \mathbf{I}_p$ in the high-dimensional setting, where λ is some unknown constant (Jiang and Yang (2013)). We assume that both the dimension p and sample size n go to infinity in such a way that $p/n \to y \in (0, 1]$. Under $\mathscr{H}_0$, our results give the exponential convergence rate of the LRT statistic to the corresponding asymptotic distribution.

1. Introduction

Testing the mean and the covariance of a Normal distribution is an important topic in multivariate analysis. In the past few decades, due to technical limitations, most theoretical results about the asymptotic behavior of testing statistics have been obtained in the fixed dimensional case only. See, for instance, Anderson [1], Eaton [7] and Muirhead [21]. However, for many modern datasets such as financial, consumer, manufacturing and multimedia data, high-dimensional settings are now common. More examples can be found in Johnstone [16] and the references in [15].

We first review some known facts about the tests for the covariance matrix of a high-dimensional random vector. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent and identically distributed (iid) observations from a random vector of dimension $p < n$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We consider the hypothesis testing problem

$$\mathscr{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad \mathscr{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0. \tag{1}$$

If $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$, where p is fixed and $\mathbf{x}_1$ is normally distributed, the asymptotic behavior of the likelihood ratio test (LRT) for (1) has been well studied. See, for instance, Anderson [1], Eaton [7] and Muirhead [21]. See also below for more details. When the dimension p tends to infinity, Bai et al. [2] corrected the LRT under the large-dimensional limiting scheme $p/n \to y \in [0, 1)$. Their proofs depend on the Central Limit Theorem for the linear spectral statistics of the sample covariance matrix. Jiang et al. [13] extended it to the case $p/n \to y \in (0, 1]$ and $p < n$. However, their proofs were based heavily on an assumption on the population distribution.

More recently, Jiang [12] proposed Rao's score test for (1). This requires only the finiteness of the 4th moment of the population distribution and $p/n \to [0, \infty)$. Her proofs also depend on the Central Limit Theorem for the linear spectral

* Corresponding author.
E-mail addresses: huijiang@nuaa.edu.cn (H. Jiang), mascwang@scut.edu.cn (S. Wang).

http://dx.doi.org/10.1016/j.jmva.2017.02.004
0047-259X/

statistics of the sample covariance matrix. Finally, some other tests for covariance matrices can be found in Cai and Ma [4], Chen et al. [5], Wang and Yao [23], and others.

In this paper, we mainly study the moderate deviation principle (MDP) for the test statistic in the setting of Jiang and Yang [15]. More precisely, we consider the hypothesis testing problem

$$\mathcal{H}_0 : \mathbf{\Sigma} = \lambda \mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_1 : \mathbf{\Sigma} \neq \lambda \mathbf{I}_p \tag{2}$$

for the Normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with λ unspecified and $p/n \to y \in (0, 1]$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be iid $\mathbb{R}^p$-valued random variables with Normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Let

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \quad \text{and} \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \tag{3}$$

Mauchly [20] introduced the likelihood ratio test statistic of (2) as follows:

$$V_n = |\mathbf{S}| \times \left\{ \frac{\operatorname{tr}(\mathbf{S})}{p} \right\}^{-p}, \tag{4}$$

where $\operatorname{tr}(\cdot)$ is the trace operator and $|A|$ indicates the determinant of any square matrix A. When $p \geq n$, $\mathbf{S}$ is not of full rank with probability 1; see also Remark 3, Section 2. This indicates that the likelihood ratio test of (2) exists only when $p \leq n-1$. In what follows, we will not state this assumption again. For the case $p \geq n$, see Chen et al. [5], Ledoit and Wolf [19], and others.

We now consider the case when both n and p are large and $p \leq n-1$. For sake of clarity when taking limits, let $p = p_n$, i.e., p depends on n (if there is no confusion, we will frequently drop the subscript n). For all $n \geq 3$, we define

$$\boldsymbol{\mu}_n = -p - \left(n - p - \frac{3}{2}\right) \ln\left(1 - \frac{p}{n-1}\right) \tag{5}$$

and

$$\sigma_n^2 = -2\left\{ \frac{p}{n-1} + \ln\left(1 - \frac{p}{n-1}\right) \right\}. \tag{6}$$

We assume that $\lim_{n\to\infty} p/n = y \in (0, 1]$. Then, under $\mathcal{H}_0$ in (2), Jiang and Yang [15] proved that $(\ln V_n - \boldsymbol{\mu}_n)/\sigma_n$ converges in distribution to $\mathcal{N}(0, 1)$ as $n \to \infty$. Then, the performance of the test statistic V_n can be measured by a local measure [17], i.e., for any $x > 0$,

$$\lim_{a\to\infty} \lim_{n\to\infty} \frac{1}{a^2} \ln \Pr\left(\frac{|\ln V_n - \boldsymbol{\mu}_n|}{\sigma_n} \geq ax \right) = -\frac{x^2}{2}.$$

A natural question is whether one has

$$\lim_{n\to\infty} \frac{1}{a_n^2} \ln \Pr\left(\frac{|\ln V_n - \boldsymbol{\mu}_n|}{\sigma_n} \geq a_n x \right) = -\frac{x^2}{2}, \tag{7}$$

for any sequence (a_n) with $a_n \to \infty$. We call (7) the moderate deviation estimation or more generally moderate deviation principle (MDP). A standard reference for MDP theory is the book by Dembo and Zeitouni [6]. We note that (7) extends the conventional local asymptotic analysis for $\ln V_n$, focusing on σ_n-neighborhoods, i.e., $\Pr(|\ln V_n - \boldsymbol{\mu}_n| \geq \sigma_n x)$ (central limit theory), to the moderate deviation region, focusing on $a_n\sigma_n$-neighborhoods, i.e., $\Pr(|\ln V_n - \boldsymbol{\mu}_n| \geq a_n\sigma_n x)$. See Inglot and Kallenberg [10], Kallenberg [18], and Otsu [22].

In hypothesis testing problems, it is important to control the type I errors, i.e., the probability

$$\Pr\left(\frac{|\ln V_n - \boldsymbol{\mu}_n|}{a_n\sigma_n} \geq x \right), \quad x > 0.$$

As pointed by Gao and Zhao [9], if the MDP holds for the test statistic and if we use the test statistic to construct the rejection (or acceptance) region, then the probabilities related to both type I and type II errors tend to zero exponentially (see Remark 2). This decay rate can be used to give the minimum sample sizes for given type I and type II errors. Thus, from the viewpoint of the statistical cost of experiments, the study of moderate deviation estimation is quite meaningful.

In this paper, in addition to the sphericity hypothesis test model (2), we consider the MDP problem for several other classical likelihood ratio tests for mean and covariance matrices of high-dimensional Normal distributions. For example, the LRT statistics in testing that several components of a vector with distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ are independent; the hypothesis testing problem with $\mathcal{H}_0$: $\mathcal{N}_p(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1) = \cdots = \mathcal{N}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$; the test of the equality of the covariance matrices from several Normal distributions and others. Since the proofs are quite similar, we will state the results in the Appendix. We refer to Jiang and Yang [15] for a full study and a comparison with existing results, especially the powers of these new statistics.

The rest of this paper is organized as follows. In Section 2, we state our main results with some remarks. Section 3 is devoted to a simulation study of our MDP results. The proofs are postponed to Section 4. Moreover, we include the MDP for several other classical likelihood ratio tests for mean and covariance matrices of high-dimensional Normal distributions in the Appendix.