Statistics and Probability Letters 88 (2014) 184–189

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Asymptotic power of likelihood ratio tests for high dimensional data

Cheng Wang *

Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

ARTICLE INFO

Article history:
Received 4 October 2013
Received in revised form 11 February 2014
Accepted 14 February 2014
Available online 22 February 2014

Keywords:
Covariance matrix
High dimensional data
Identity test
Likelihood ratio test
Power
Stein's loss

ABSTRACT

This paper studies the asymptotic power of the likelihood ratio test (LRT) for the identity test when the dimension p is large compared to the sample size n. The asymptotic distribution under local alternatives is derived and a simulation study is carried out to compare LRT with other tests. All these studies show that LRT is a powerful test to detect small eigenvalues.

1. Introduction

In multivariate analysis for high dimensional data, testing the structure of population covariance matrices is an important problem. See, for example, Johnstone (2001), Ledoit and Wolf (2002), Srivastava (2005), Schott (2006), Chen et al. (2010), Cai and Jiang (2011) and Li and Chen (2012), among others. To specify the problem considered here, let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) from a multivariate normal distribution $N_p(\mu_p, \Sigma_p)$ where μ_p is the mean vector and Σ_p is the population covariance matrix. In many studies, a hypothesis test of significant interest is to test

$$H_0 : \Sigma_p = I_p \quad \text{vs.}\ H_1 : \Sigma_p \neq I_p, \tag{1.1}$$

where I_p is the p-dimensional identity matrix. Note that the identity matrix in (1.1) can be replaced by any other positive definite matrix Σ_0 through multiplying the data by $\Sigma_0^{-1/2}$.

To test (1.1), we usually need the sample covariance matrix which is defined as

$$S_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})',$$

where $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$. The likelihood ratio test (LRT) can be defined as

$$T_n = \operatorname{tr}(S_n) - \log|S_n| - p, \tag{1.2}$$

* Correspondence to: Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China.
E-mail addresses: cescwang@gmail.com, wwcc@mail.ustc.edu.cn.

http://dx.doi.org/10.1016/j.spl.2014.02.010
0167-7152/

and when p is fixed and n tends to infinity, nT_n converges to a chi-squared distribution with $p(p+1)/2$ degrees of freedom under H_0 (Anderson, 2003). For high dimensional data (p is large), the failure of classical LRT was first observed by Dempster (1958) and later in a pioneer work by Bai et al. (2009), authors proposed corrections to LRT when $p/n \to c \in (0, 1)$ and $\mu_p = 0$. Successive works included Jiang et al. (2012) which extended the results of Bai et al. (2009) to Gaussian data with general μ_p and our work (Wang et al., 2013) where we studied the LRT for general μ_p and non-Gaussian data. About the LRT for other related problems, see also two recent works by Jiang and Yang (2013) and Wang and Yao (2013).

As we know, the existing results about LRT (Bai et al., 2009; Jiang et al., 2012; Wang et al., 2013) in high dimensional data have only derived asymptotic null distribution and we know little about the asymptotic point-wise power of LRT under the alternative hypothesis. In this work, we will consider the asymptotic distribution of LRT when $\Sigma_p \neq I_p$ but $\text{tr}(\Sigma_p - I_p)^2 = o(p)$. From these results, we find that LRT is powerful to detect eigenvalues around zero. Simulations will also be conducted to compare LRT with two other tests proposed by Chen et al. (2010) and Cai and Ma (2013).

The rest of the paper is organized as follows. Section 2 introduces the basic data structure and establishes the asymptotic power of LRT while Section 3 reports simulation studies. All the proofs are included in Appendix.

2. Main results

To relax the Gaussian assumptions, we assume that the observations $X_1, \ldots, X_n$ satisfy a multivariate model (Chen et al., 2010)

$$X_i = \Sigma_p^{1/2} Y_i + \mu_p, \quad \text{for } i = 1, \ldots, n, \tag{2.3}$$

where μ_p is a p-dimensional constant vector and the entries of $\mathcal{Y}_n = (Y_{ij})_{p\times n} = (Y_1, \ldots, Y_n)$ are i.i.d. with $EY_{ij} = 0$, $EY_{ij}^2 = 1$ and $EY_{ij}^4 = 3 + \Delta$.

When $y_n = p/n < 1$, Bai et al. (2009) proposed a correction to the classic LRT and redefined LRT as

$$L_n = \frac{1}{p}\text{tr}(S_n) - \frac{1}{p}\log|S_n| - 1 - d(y_n), \tag{2.4}$$

where $d(x) = 1 + (1/x - 1)\log(1 - x),\ 0 < x < 1$. Under the null hypothesis, Bai et al. (2009) derived the asymptotic distribution of L_n for Gaussian data with known means. Our previous work (Wang et al., 2013) extended this result to the multivariate model (2.3) which can accommodate unknown means and non-Gaussian data and the following is the details of the main results in Wang et al. (2013).

Theorem 2.1 (*Theorem 2.1 of Wang et al., 2013*). *When $\Sigma_p = I_p$ and $y_n = p/n \to y \in (0, 1)$,*

$$\frac{pL_n - \mu_n}{\sigma_n} \xrightarrow{D} N(0, 1),$$

where $\mu_n = y_n(\Delta/2 - 1) - 3/2\log(1 - y_n)$, $\sigma_n^2 = -2y_n - 2\log(1 - y_n)$ and $\xrightarrow{D}$ denotes convergence in distribution.

When $X_1, \ldots, X_n$ are i.i.d. distributed from $N_p(\mu_p, \Sigma_p)$, Jiang et al. (2012) derived a similar result as Theorem 2.1 by using the Selberg integral and they also considered the special situation where $p/n \to 1$. Based on the asymptotic normality under the respective null hypothesis, an asymptotic level α test based on L_n is given by

$$\phi = I\left(\frac{pL_n - \mu_n}{\sigma_n} > z_{1-\alpha}\right), \tag{2.5}$$

where $I(\cdot)$ is the indicator function, and $z_{1-\alpha}$ denotes the $100 \times (1 - \alpha)$th percentile of the standard normal distribution.

In the classical LRT test, if S_n be seen as the estimator of Σ_p, LRT actually is an estimator for

$$R(\Sigma_p) = \text{tr}(\Sigma_p) - \log|\Sigma_p| - p, \tag{2.6}$$

which can be regarded as a special case of Stein's loss function (James and Stein, 1961). Denoting the eigenvalues of Σ_p as $u_1 \geq \cdots \geq u_p > 0$, we have

$$R(\Sigma_p) = \sum_{k=1}^{p}(u_k - \log u_k - 1), \tag{2.7}$$

which is 0 when $\Sigma_p = I_p$ and positive when $\Sigma_p \neq I_p$. In the following theorem, we establish the convergence of L_n under the local alternatives where $\text{tr}(\Sigma_p - I_p)^2 = o(p)$.

Theorem 2.2. *When $\text{tr}(\Sigma_p - I_p)^2/p \to 0$ and $y_n = p/n \to y \in (0, 1)$,*

$$\frac{pL_n - R(\Sigma_p) - \mu_n}{\sigma_n} \xrightarrow{D} N(0, 1).$$