Statistics and Probability Letters 140 (2018) 132–141

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

On Schott's and Mao's test statistics for independence of normal random vectors

Shuhua Chang [a], Yongcheng Qi [b,*]

[a] Coordinated Innovation Center for Computable Modeling in Management Science, Tianjin University of Finance and Economics, Tianjin 300222, PR China

[b] Department of Mathematics and Statistics, University of Minnesota Duluth, 1117 University Drive, Duluth, MN 55812, USA

ARTICLE INFO

Article history:
Received 1 November 2017
Received in revised form 8 February 2018
Accepted 3 May 2018
Available online 15 May 2018

Keywords:
High dimension
Complete independence
Normal distribution
Limiting distribution

ABSTRACT

Consider a random sample of n independently and identically distributed p-dimensional normal random vectors. A test statistic for complete independence of high-dimensional normal distributions, proposed by Schott (2005), is defined as the sum of squared Pearson's correlation coefficients. A modified test statistic has been proposed by Mao (2014). Under the assumption of complete independence, both test statistics are asymptotically normal if the limit $\lim_{n\to\infty} p/n$ exists and is finite. In this paper, we investigate the limiting distributions for both Schott's and Mao's test statistics. We show that both test statistics, after suitably normalized, converge in distribution to the standard normal as long as both n and p tend to infinity. Furthermore, we show that the distribution functions of the test statistics can be approximated very well by a chi-square distribution function with $p(p-1)/2$ degrees of freedom as n tends to infinity regardless of how p changes with n.

1. Introduction

In classical multivariate analysis, statistical methods have been developed mainly for data from designed experiments and dimensions of the data are fixed or very small compared with the sample size. Nowadays, new technology has generated various types of high-dimensional datasets such as financial data, consumer data, modern manufacturing data, multimedia data, hyperspectral image data, internet data, microarray and DNA data. A common feature for all these datasets is that their dimensions can be very large compared with their sample sizes. See, e.g., Schott (2001, 2005, 2007), Ledoit and Wolf (2002), Fan et al. (2005), Bai et al. (2009), Chen et al. (2010), Chen and Qin (2010), Fujikoshi et al. (2010), Bühlmann and van de Geer (2011), Jiang et al. (2012), Srivastava and Reid (2012).

Throughout the paper, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. We assume that $\boldsymbol{\Sigma}$ is positive definite. Write $\boldsymbol{\Sigma} = (\sigma(i,j))_{1\le i,j\le p}$. Then, $\boldsymbol{\Gamma} = (\rho_{ij})_{1\le i,j\le p}$ is the correlation matrix of $\boldsymbol{\Sigma}$ given by $\rho_{ij} = \sigma(i,j)/\sqrt{\sigma(i,i)\sigma(j,j)}$.

Assume that a p-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_p)'$ has a distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We are interested in testing whether the p components $x_1, x_2, \ldots, x_p$ are independent or equivalently testing whether the covariance matrix $\boldsymbol{\Sigma}$ is diagonal. Then, the test can be written as

$$H_0 : \boldsymbol{\Gamma} = \mathbf{I}_p \ \text{ vs } \ H_a : \boldsymbol{\Gamma} \ne \mathbf{I}_p. \tag{1.1}$$

In literature, (1.1) is known as the test of complete independence.

* Corresponding author.
E-mail addresses: szhang@tjufe.edu.cn (S. Chang), yqi@d.umn.edu (Y. Qi).

https://doi.org/10.1016/j.spl.2018.05.009
0167-7152/

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. from $N_p(\mu, \mathbf{\Sigma})$. Write

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})', \quad k = 1, \ldots, n.$$

Define

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \cdot \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}, \tag{1.2}$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ and $\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. Then, $\mathbf{R}_n := (r_{ij})_{p\times p}$ is the sample correlation matrix based on the p-dimensional random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

In classic multivariate analysis when p is a fixed integer, the likelihood method is a nice approach to test (1.1). From Bartlett (1954) or Morrison (2005), the likelihood ratio test statistic is a function of the determinant of $\mathbf{R}_n$. When $p = p_n$ depends on n and $p_n \to \infty$, the limiting distribution of the likelihood ratio test statistic has been obtained in Jiang and Yang (2013), Jiang et al. (2013) and Jiang and Qi (2015), and the likelihood ratio method can still be used to test (1.1). However, the likelihood ratio method fails when $p \geq n$, since the sample correlation matrix $\mathbf{R}_n$ is singular and the corresponding test statistic is degenerate. A natural requirement for non-singularity of $\mathbf{R}_n$ is $p < n$.

Schott (2005) considers the following test statistic

$$t_{np} = \sum_{1\leq j<i\leq p} r_{ij}^2.$$

Assume that the null hypothesis of (1.1) holds and $\lim_{n\to\infty} p/n = \gamma \in (0, \infty)$. Schott (2005) proves that $t_{np} - \frac{p(p-1)}{2(n-1)}$ converges in distribution to a normal distribution with mean 0 and variance γ^2, that is,

$$t_{np}^* := \frac{t_{np} - \frac{p(p-1)}{2(n-1)}}{\tau_{np}} \xrightarrow{d} N(0, 1), \tag{1.3}$$

where $\tau_{np}^2 = \frac{p(p-1)(n-2)}{(n-1)^2(n+1)}$.

It is worth noting that the same test statistic t_{np} is also proposed by Srivastava (2005). Srivastava (2005, 2006) also considers a test statistic which is based on Fisher's z-transformation and originally proposed by Chen and Mudholkar (1990):

$$Q_{np} = \frac{(n-3)\sum_{1\leq j<i\leq p} z_{ij}^2 - \frac{1}{2}p(p-1)}{\sqrt{p(p-1)}},$$

where $z_{ij} = \frac{1}{2}\log\frac{1+r_{ij}}{1-r_{ij}}$. From Srivastava (2005), such a test has not been designed for large p. Instead, Srivastava (2005) proposes a test statistic T_3 which is related to the sample covariances only. See Srivastava (2005, 2006) for details. Under certain conditions, Srivastava (2005) shows that T_3 converges in distribution to the standard normal under the null hypothesis in (1.1). A simulation study in Srivastava (2006) indicates that Q_{np} statistic is inferior as the test does not give a consistent nominal level when n and p are close.

Very recently, Mao (2014) proposes a new test for complete independence. The new test statistic is closely related to Schott's test and is defined by

$$T_{np} = \sum_{1\leq j<i\leq p} \frac{r_{ij}^2}{1 - r_{ij}^2}.$$

It has been proved in Mao (2014) that T_{np} is asymptotically normal under the null hypothesis of (1.1) and assumption that $\lim_{n\to\infty} p/n = \gamma \in (0, \infty)$.

In this paper, we will remove the condition imposed on p and assume only that $p = p_n \to \infty$ as $n \to \infty$. We will show that both T_{np} and t_{np} are asymptotically normal. We also establish a unified chi-square approximation for the distributions of T_{np} and t_{np} regardless of how p changes with n.

The rest of the paper is organized as follows. The main results of the paper are given in Section 2 and their proofs are postponed until Section 4. A simulation study to compare the performance of several different approaches is reported in Section 3.

2. Main results

Our main results include three theorems. We first obtain the limiting distribution of the test statistic T_{np} in a larger range for p, and then establish a unified chi-square approximation for all $p \geq 2$. The corresponding limiting distributions of t_{np} are given in the third theorem.

The first theorem states that Mao's (2014) test statistic T_{np} is asymptotically normal as long as $p = p_n \to \infty$ as $n \to \infty$.