Statistics and Probability Letters 106 (2015) 82–85

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A note on testing complete independence for high dimensional data

Guangyu Mao *

School of Economics and Management, Beijing Jiaotong University, China

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form 1 July 2015
Accepted 1 July 2015
Available online 14 July 2015

Keywords:
HDLSS
High dimension
Independence test
Universal asymptotics

ABSTRACT

The high dimensional independence test in Schott (2005) assumes $(m, n) \to \infty$ and $\frac{m}{n} \to \gamma \in (0, \infty)$, where m signifies the dimension and n denotes the sample size. This paper notes that without the restriction $\frac{m}{n} \to \gamma$, the test is still effective provided that $(m, n) \to \infty$, or n is fixed but $m \to \infty$.

1. Introduction

Given a Gaussian vector of dimension m with pairwise correlations $\rho_{ij} (1 \le i, j \le m)$, to test the complete independence, it is equivalent to test

$$H_0 : \rho_{ij} = 0 \quad \text{for } 1 \le i < j \le m \text{ vs. } H_1 : H_0 \text{ is not true.}$$

Denote the sample size as N and $Y_i = (Y_{i1}, \ldots, Y_{iN})'$ as the data vector of variable i for $1 \le i \le m$. When m is far smaller than N, testing for H_0 has been well understood for a very long time; see e.g. Wilks (1935). In this case, to approximate the distributions of the related test statistics, it is assumed that m is fixed but $N \to \infty$. When m is comparable to or even far larger than N, i.e., in the high dimensional case, it is generally more appropriate to postulate $m \to \infty$ when we make a statistical approximation. To test H_0 in high dimensions, Schott (2005) proposed the following statistic:

$$t_{nm} = \sum_{i=2}^{m} \sum_{j=1}^{i-1} r_{ij}^2 - \frac{m(m-1)}{2n},$$

where r_{ij} is the sample correlation between Y_i and Y_j, and $n = N - 1$. By assuming $(m, n) \to \infty$ but subject to $\frac{m}{n} \to \gamma \in (0, \infty)$, Schott proved that $\sigma_{t_{nm}}^2 = var(t_{nm}) = \frac{m(m-1)(n-1)}{n^2(n+2)} \to \gamma^2$ and $t_{nm} \xrightarrow{d} N(0, \gamma^2)$ under H_0, where $\xrightarrow{d}$ signifies convergence in distribution. As a result, $\sigma_{t_{nm}}^{-1} t_{nm}$ can be employed as the statistic to test H_0.

The restriction $\frac{m}{n} \to \gamma$ can ensure that t_{nm} converges to a non-degenerate distribution, nevertheless, which is not of paramount importance for the purpose of testing complete independence. In fact, to test H_0, it is sufficient to have $\sigma_{t_{nm}}^{-1} t_{nm} \xrightarrow{d} N(0, 1)$ under the null hypothesis. In this paper, we note that when $\frac{m}{n} \to \gamma$ is dropped, which implies that t_{nm} may

* Correspondence to: #926, Science and Technology Building, Beijing Jiaotong University, Shang Yuan Cun #3, Beijing, 100044, China.
E-mail address: gymao@bjtu.edu.cn.

http://dx.doi.org/10.1016/j.spl.2015.07.001
0167-7152/

converge to 0 or diverge, it still holds that $\sigma_{t_{nm}}^{-1} t_{nm} \xrightarrow{d} N(0, 1)$ under H_0 provided $(m, n) \to \infty$. In Paindaveine and Verdebout (forthcoming), this kind of asymptotics is called *universal* (n, m)-asymptotics in the sense that no restriction is imposed on the relative magnitude of the sample size and the dimension. In the literature, only a very limited number of papers consider high dimensional tests under the universal (n, m)-asymptotics; see Paindaveine and Verdebout (forthcoming) for a recent example. Besides, for a fixed n, we further note that $\sigma_{t_{nm}}^{-1} t_{nm} \xrightarrow{d} N(0, 1)$ as $m \to \infty$ under H_0. This amazing result suggests that Schott's test is also applicable to HDLSS data, where the abbreviation "HDLSS" introduced by Hall et al. (2005) signifies "High Dimension, Low Sample Size". In the HDLSS case, statistical theories are generally established based on the assumption that the dimension approaches infinity but the sample size is fixed.

Inspection of the simulation results in Table 1 in Schott (2005) shows that $\sigma_{t_{nm}}^{-1} t_{nm}$ performs well not only in the large-m and large-n case but also in the large-m and small-n case. In fact, the present paper can provide a theoretical explanation to this observation.

2. Theories and proofs

Our main findings are collected in the following theorem:

Theorem 1. *If the sample correlations r_{ij} $(1 \le j < i \le m)$ are computed using a random sample of size $n + 1$ from an m-dimensional Gaussian distribution, then $\sigma_{t_{nm}}^{-1} t_{nm} \xrightarrow{d} N(0, 1)$ as $(m, n) \to \infty$ under H_0, or $\sigma_{t_{nm}}^{-1} t_{nm} \xrightarrow{d} N(0, 1)$ as $m \to \infty$ under H_0 for fixed n.*

First, we give some results associated with moments of r_{ij}^2. It has been known that $r_{ij}^2 \sim Beta(\frac{1}{2}, \frac{n-1}{2})$ when $\rho_{ij} = 0$; see e.g. page 147 in Muirhead (1982). Thus, $E(r_{ij}^2) = \frac{1}{n}$. Denote $b_{ij} = r_{ij}^2 - \frac{1}{n}$. We further have

$$B_2 \equiv E(b_{ij}^2) = \frac{2(n-1)}{n^2(n+2)},$$
$$B_4 \equiv E(b_{ij}^4) = \frac{12(5n^3 - 18n^2 + 25n - 12)}{n^4(n^3 + 12n^2 + 44n + 48)}.$$

When $\rho_{ij} = 0$, r_{ij} can be represented by $\frac{U_i^* U_j^*}{\|U_i^*\| \, \|U_j^*\|}$ in terms of Theorem 5.1.1 in Muirhead (1982), where U_i^* and U_j^* depends only on Y_i and Y_j respectively, and are of mutually independent n-variate spherical distribution. As a result, according to 1.5.7 in Muirhead (1982), the distribution of r_{ij} is identical to the conditional distribution of r_{ij} given Y_i or Y_j. Therefore, for distinct u, v, j_1 and j_2, under H_0 we have

$$E(b_{uj_1} b_{uj_2}) = E[E(b_{uj_1} b_{uj_2} | Y_u)] = E[E(b_{uj_1} | Y_u) E(b_{uj_2} | Y_u)] = 0,$$
$$E(b_{uj_1} b_{vj_2}) = E(b_{uj_1}) E(b_{vj_2}) = 0,$$

which suggests $var(t_{nm}) = var(\sum_{i=2}^{m} \sum_{j=1}^{i-1} b_{ij}) = \sum_{i=2}^{m} \sum_{j=1}^{i-1} E(b_{ij}^2) = \frac{m(m-1)B_2}{2}$. Analogously, we can verify

$$E(b_{uj_1} b_{uj_2} b_{uj_3} b_{uj_4}) = \begin{cases} B_4 & j_1 = j_2 = j_3 = j_4 \\ B_2^2 & j_1 = j_2 \neq j_3 = j_4 \\ B_2^2 & \text{if } j_1 = j_3 \neq j_2 = j_4 \\ B_2^2 & j_1 = j_4 \neq j_2 = j_3 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and for $u \neq v$,

$$E(b_{ui_1} b_{ui_2} b_{vj_1} b_{vj_2}) = \begin{cases} B_2^2 & \text{if } i_1 = i_2,\ j_1 = j_2 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Now, we are in a position to prove Theorem 1.

Proof (Theorem 1)**.** We define $\mathcal{F}_{nu}$ as σ-field generated by $\{Y_1, \ldots, Y_u\}$, let $T_{nu} = \sigma_{t_{nm}}^{-1} t_{nu}$ for $u = 2, \ldots, m$ and $T_{n1} = 0$, and denote $X_{nu} = T_{nu} - T_{n,u-1} = \sigma_{t_{nm}}^{-1} \sum_{j=1}^{u-1} b_{uj}$. Using theorem 1.5.7 in Muirhead (1982), it can be easily verified that $E(T_{nu} | \mathcal{F}_{n,u-1}) = \sigma_{t_{nm}}^{-1} t_{n,u-1}$. Consequently, $\{X_{nu}, \mathcal{F}_{nu}, 2 \le u \le m\}$ is a MD (martingale difference) sequence when n is fixed or a MD array when both n and m tend to infinity. In the latter case, we may regard m as $m(n)$, an increasing function of n.

Since $\sigma_{t_{nm}}^{-1} t_{nm} = \sum_{u=2}^{m} X_{nu}$ is a sum of MD sequence or MD array, if we can prove the following two conditions:

$$(i): \sum_{u=2}^{m} E(X_{nu}^4) \to 0 \quad \text{and} \quad (ii): \sum_{u=2}^{m} X_{nu}^2 \xrightarrow{p} 1,$$

as $m \to \infty$, irrespective of whether n is fixed or diverging,