Computational Statistics and Data Analysis 117 (2018) 128–137

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Testing independence in high dimensions using Kendall's tau

Guangyu Mao [*]

School of Economics and Management, Beijing Jiaotong University, China

ARTICLE INFO

Article history:
Received 14 March 2017
Received in revised form 16 July 2017
Accepted 28 July 2017
Available online 23 August 2017

Keywords:
High dimensions
Independence test
Kendall's tau
Rank correlation

ABSTRACT

To check the total independence of a random vector without Gaussian assumption in high dimensions, Leung and Drton (forthcoming) recently developed a test by virtue of pairwise Kendall's taus. However, as their simulation shows, the test suffers from noticeable size distortion when the sample size is small. The present paper provides a theoretical explanation about this phenomenon, and accordingly proposes a new test. The new test can be justified when both the dimension and the sample size go to infinity simultaneously, and moreover, it can be even justified when the dimension tends to infinity but the sample size is fixed, which implies that the test can perform well in the cases of small sample size. Simulation studies confirm the theoretical findings, and show that the newly proposed test can bring remarkable improvement. To illustrate the use of the new test, a real data set is also analyzed.

1. Introduction

Due to the increasing availability of high-dimensional data sets, the past decade has witnessed a surge of interest in the theoretical analysis about these kinds of data sets in the literature. Typically, a high-dimensional data set consisting of p variables has the characteristic that p is comparable to, or far larger than n, where n is the sample size. Therefore, unlike the traditional cases, it is more reasonable to assume that p diverges as $n \to \infty$ when investigating statistical issues such as estimation, inference, or model selection in high dimensions. Under this assumption, denoted by $(p, n) \to \infty$ below, various statistical tools for different high-dimensional settings have been developed in the literature. For a recent summary and bibliography, interested reader may refer to Koch (2013), Pourahmadi (2013), Giraud (2014), and Yao et al. (2015).

In this paper, we are interested in testing total independence among a p-dimensional random vector $(X_1, \ldots, X_p)$ using n independent observations in high dimensions. Our aim is to develop new tests under the following assumption:

Assumption 1. $(X_1, \ldots, X_p)$ is both a jointly and marginally continuous random vector whose joint density is $f(x_1, \ldots, x_p)$ and marginal densities are $f_k(x_k)$ $(k = 1, \ldots, p)$.

This assumption excludes the cases of non-continuous variables, but it can apply to lots of real-world data sets. Based on it, the null hypothesis of interest and the corresponding alternative can be stated by

$$H_0 : f(x_1, \ldots, x_p) = \prod_{k=1}^{p} f_k(x_k) \text{ against } H_1 : H_0 \text{ does not hold.}$$

When $(X_1, \ldots, X_p)$ are jointly Gaussian, the targeted null hypothesis is equivalent to

$$\tilde{H}_0 : r_{kl} = 0 \text{ for } 2 \leq k \leq p \text{ and } 1 \leq l < k - 1,$$

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

http://dx.doi.org/10.1016/j.csda.2017.07.012
0167-9473/© 2017 Elsevier B.V. All rights reserved.

where r_{kl} is Pearson's correlation coefficient of X_k and X_l. As a result, in the Gaussian case any test that is applicable to $\tilde{H}_0$ under $(p, n) \to \infty$ can be directly implemented to test H_0. In the literature, available tests include, but are not limited to, those in Schott (2005), Srivastava (2005), Cai and Jiang (2011), Srivastava et al. (2011), Qiu and Chen (2012), and Mao (2014).

In practice, the Gaussian assumption does not always hold. When the population underlying the data is not Gaussian, $\tilde{H}_0$ is not equivalent to H_0. Under this circumstance, it may be inappropriate to employ the tests in the above listed papers to test H_0, primarily due to the fact that Pearson's correlation coefficient can be defined only if the variances of the variables are finite. In practice, some data sets may be of infinite variance. For instance, returns of risky assets in financial markets are believed not to have finite variances in some settings. If this is the case, testing H_0 by the existing tests for $\tilde{H}_0$ can by no means be justified. Therefore, it is better to drop the Gaussian assumption and abandon the use of Pearson's correlation coefficient (or covariance) if H_0 is the main concern.

To circumvent the drawbacks of the existing tests to formulate effective tests for H_0, several papers recently have shifted attention to rank-based methods. In the traditional cases, rank-based tests usually have the merit that they only require weak distributional assumptions as documented, e.g. by Hájek et al. (1999). For example, to test independence between two variables by the commonly used rank correlations: Spearman's rho (Spearman, 1904) or Kendall's tau (Kendall, 1938), it suffices to require that the two variables have continuous densities. This is a weak assumption since it can accommodate lots of distributions, and allow infinite moments. Based on different rank statistics, Han and Liu (2014), Leung and Drton (Forthcoming) and Mao (Forthcoming) recently have developed several tests for H_0 under $(p, n) \to \infty$ without strong distributional assumptions, which shows that the merit of rank-based tests can continue to hold in high dimensions.

In this paper, we will develop a new test based on the sum of squares of pairwise Kendall's taus. The sum has been employed by Leung and Drton (Forthcoming) to construct a test for H_0. However, their simulation shows that test tends to suffer from noticeable size distortion when n is small. We will provide an explanation about this phenomenon, and then propose a new statistic, which can bring remarkable improvement according to our simulation studies. Besides, it is worth noting that Leung and Drton (Forthcoming) justified their test under $(p, n) \to \infty$. In contrast, we justify the new test not only under $(p, n) \to \infty$ but also under $p \to \infty$ with fixed n. Thus, the new test applies to HDLSS data, where "HDLSS" is an abbreviation of "High Dimension, Low Sample Size" introduced by Hall et al. (2005). As we will show, this property is not shared with the test of Leung and Drton (Forthcoming).

The remainder of this paper is organized as follows. In the next section, we theoretically explain why the test of Leung and Drton (Forthcoming) tends to be oversized for small n, propose our test statistic, and discuss related statistical properties. Section 3 is devoted to simulation studies about the new test. Section 4 illustrates the use of the new test by a real example. Section 5 is a short conclusion. All technical proofs are postponed to Appendix.

2. Tests based on Kendall's tau

Suppose $\{X_{ki}\}_{i=1}^n$ and $\{X_{li}\}_{i=1}^n$ are two random samples of X_k and X_l, and accordingly $\{R_{ki}\}_{i=1}^n$ and $\{R_{li}\}_{i=1}^n$ are the corresponding ranks, respectively. The so called Kendall's tau (Kendall, 1938) of the two variables can be defined by

$$\tau_{kl} = \frac{2S_{kl}}{n(n-1)}, \tag{1}$$

where $S_{kl} = \sum_{i=2}^n \sum_{j=1}^{i-1} sgn(R_{ki} - R_{kj}) sgn(R_{li} - R_{lj})$. As a measure of ordinal association, Kendall's tau can be employed to test independence between two variables; see e.g. Hájek et al. (1999). When τ_{kl} is close to zero in the statistical sense, there is evidence of independence.

To test $\tilde{H}_0$, Schott (2005) employed the statistic $\sum_{k=2}^p \sum_{l=1}^{k-1} \hat{r}_{kl}^2$, where $\hat{r}_{kl}$ is the sample Pearson's correlation coefficient of X_k and X_l. $\tilde{H}_0$ will be rejected for large values of the statistic. Motivated by Schott (2005), it is natural to use $T_{np} \triangleq \sum_{k=2}^p \sum_{l=1}^{k-1} \tau_{kl}^2$ to construct effective statistics for testing H_0, where $\triangleq$ signifies "is defined to be equal to". Leung and Drton (Forthcoming) centered T_{np} by its mean under H_0, and then proved by virtue of theories about U-statistics in their Theorem 4.1 that under H_0,

$$S_\tau \triangleq \frac{9n}{4p}\left[T_{np} - \frac{p(p-1)(2n+5)}{9n(n-1)}\right] \xrightarrow{d} N(0, 1) \text{ as } (p, n) \to \infty,$$

where $\xrightarrow{d}$ denotes convergence in distribution. Given a nominal significance level α, the S_τ-based test rejects H_0 as long as the actual value of S_τ is larger than the $1 - \alpha$ quantile of the standard Gaussian distribution.

Even though the S_τ-based test can be justified in the asymptotic sense, as the simulation of Leung and Drton (Forthcoming) shows, the test generally suffers from severe size distortion when n is small. To understand this phenomenon, we note that the centered T_{np} in S_τ is scaled by $(9n)^{-1}4p$, rather than $\sigma_{np} \triangleq \sqrt{Var(T_{np})}$. If the distribution of the centered T_{np} that is scaled by its exact standard deviation can be well approximated by the standard Gaussian distribution (which will be empirically justified by the simulation studies below), it is reasonable to explore how much $(9n)^{-1}4p$ deviate from σ_{np} in order to understand the size distortion. To do so, we need to know the exact forms of the variances and related covariances of the squared Kendall's taus. In terms of Property $*$ 2 in Brown and Eagleson (1984), under H_0,

$$\tau_{kl} \text{ for } 2 \le k \le p \text{ and } 1 \le l < k-1 \text{ are of pairwise independence.} \tag{2}$$