Journal of Statistical Planning and Inference 175 (2016) 78–86

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

On some exact distribution-free tests of independence between two random vectors of arbitrary dimensions

Munmun Biswas, Soham Sarkar, Anil K. Ghosh *

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form 22 February 2016
Accepted 28 February 2016
Available online 16 March 2016

Keywords:
Distance correlation
Edge weighted graph
Level and power of a test
Minimal spanning tree
Prim's algorithm

ABSTRACT

Several nonparametric methods are available in the literature to test the independence between two random vectors. But, many of them perform poorly for high dimensional data and are not applicable when the dimension of one of these vectors exceeds the sample size. Moreover, most of these tests are not distribution-free in the general multivariate set up. Recently, Heller et al. (2012) proposed a test of independence, which is distribution-free and can be conveniently used even when the dimensions are larger than the sample size. In this article, we point out some limitations of this test and propose some modifications to overcome them retaining its distribution-free property. Some simulated and real data sets are analyzed to demonstrate the utility of our proposed modifications.

1. Introduction

Let $\begin{pmatrix}\mathbf{x}_1\\ \mathbf{y}_1\end{pmatrix}, \begin{pmatrix}\mathbf{x}_2\\ \mathbf{y}_2\end{pmatrix}, \ldots, \begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix}$ be n independent realizations of a $(p+q)$-dimensional continuous random vector $\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$, where $\mathbf{x}$ and $\mathbf{y}$ are of dimensions p and q, respectively. Here, we want to test whether $\mathbf{x}$ and $\mathbf{y}$ are independent. Under the assumption of normality of $\mathbf{z}$, this is equivalent to test the uncorrelatedness between $\mathbf{x}$ and $\mathbf{y}$. Based on this normality assumption, one can construct the likelihood ratio test based on the Wilks' Λ statistic (Wilks, 1935). Tests somewhat similar to Wilks Λ test are those based on Roy's largest root, Hotelling–Lawley trace and Pillai–Bartlett trace. A power comparison among these tests can be found in Pillai and Jayachandran (1967).

Several nonparametric tests have also been proposed in the literature, and they are often preferred over parametric tests because of their flexibility and robustness. For the univariate case (i.e., $p = q = 1$), Blomqvist (1950) proposed a test based on quadrant statistic. Distribution-free tests of independence based on empirical distribution function were developed in Hoeffding (1948) and Blum et al. (1961). One can construct distribution-free tests based on Spearman's ρ and Kendall's τ statistics (see e.g., Gibbons and Chakraborti (2011)) as well. In the multivariate case (i.e., $p > 1$ or $q > 1$), perhaps the simplest among the nonparametric tests of independence are those based on co-ordinate wise signs and ranks (see e.g., Puri and Sen, 1971). Gieser and Randles (1997) proposed a multivariate extension of the quadrant test based on interdirections. Using spatial signs and ranks, Taskinen et al. (2003) and Taskinen et al. (2005) proposed multivariate extensions of the tests based on Spearman's ρ, Kendall's τ and Blomqvist's quadrant statistics. A summary of most of these multivariate nonparametric tests can be found in Oja and Randles (2004) and Oja (2010). But, none of these above mentioned multivariate tests can be used when the dimension of either $\mathbf{x}$ or $\mathbf{y}$ exceeds the sample size. Moreover, none of them are distribution-free in finite sample situations. In such cases, one either uses the test based on the large sample distribution of the test statistic or the permutation test.

* Corresponding author.
E-mail addresses: munmun.biswas08@gmail.com (M. Biswas), sohamsarkar1991@gmail.com (S. Sarkar), akghosh@isical.ac.in (A.K. Ghosh).

http://dx.doi.org/10.1016/j.jspi.2016.02.007
0378-3758/

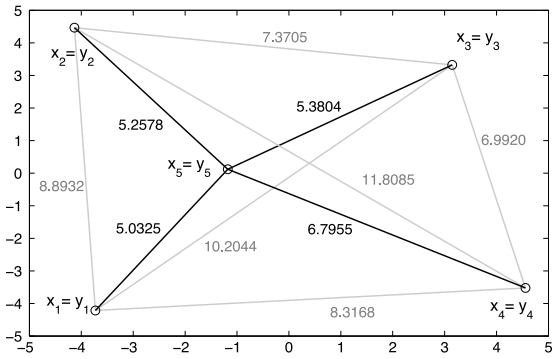

Fig. 1. The complete graph $\mathcal{G}_1$ (or $\mathcal{G}_2$) and the minimal spanning tree $\mathcal{T}_1$ (or $\mathcal{T}_2$).

Nonparametric tests of independence that can be used for high dimension, low sample size data include Bakirov et al. (2006), Székely et al. (2007) and Heller et al. (2013). These tests are consistent against general alternatives; but they are not distribution-free in finite sample situations. In all these cases, the authors suggested to use conditional tests based on the permutation principle.

Friedman and Rafsky (1983) was the first to construct some graph based tests of independence between two random vectors of arbitrary dimensions. Following similar ideas, Heller et al. (2012) constructed a multivariate test of independence (henceforth referred to as the HGH test) based on random traversals of minimal spanning trees (MST). Like Friedman and Rafsky's tests, the HGH test can be conveniently used even when the dimensions of the random vectors exceed the sample size. Moreover, this test has the distribution-free property in finite sample situations. However, it is not above all limitations. In the next section, we point out some shortcomings of the HGH test and propose some modifications to overcome them. Our modified tests retain the distribution-free property, and they can also be used in high dimension, low sample size situations. Further, they usually yield better performance than the HGH test, which we will see in Section 2. In Sections 3 and 4, we analyze some simulated and real data sets, respectively, to further demonstrate the utility of our proposed modifications. We also compare the performance of these proposed tests with some popular tests of independence. Finally, some concluding remarks are given in Section 5.

2. HGH test and its modifications

Consider two edge weighted complete graphs $\mathcal{G}_1$ on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and $\mathcal{G}_2$ on $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, where the observations are taken as nodes and the Euclidean distance between two observations is taken as the weight associated with the edge connecting them. Heller et al. (2012) first considered two minimal spanning trees $\mathcal{T}_1$ in $\mathcal{G}_1$ and $\mathcal{T}_2$ in $\mathcal{G}_2$. They chose one of these trees at random ($\mathcal{T}_1$, say), performed a random traversal of its $n-1$ edges and computed the ranks of the corresponding edges in the other graph ($\mathcal{G}_2$, say). A random traversal of $\mathcal{T}_1$ starts with a randomly chosen node $v_{11}^{(1)}$ of $\mathcal{G}_1$, and at the first step, it randomly selects one of the edges of $\mathcal{T}_1$ adjacent to $v_{11}^{(1)}$ to visit a new node $v_{12}^{(1)}$. In each of the subsequent steps, it chooses an edge of $\mathcal{T}_1$ adjacent to one of the visited nodes to visit a new node. Suppose that at the ith step ($i = 1, 2, \ldots, n-1$), it chooses an edge $e_i = (v_{i1}^{(1)}, v_{i2}^{(1)})$ that connects an already visited node $v_{i1}^{(1)}$ with an unvisited node $v_{i2}^{(1)}$. Let $e_i' = (v_{i1}^{(2)}, v_{i2}^{(2)})$ be the corresponding edge in $\mathcal{G}_2$. Then R_i is defined as the rank of e_i' (rank of the weight of e_i') among the $n-i$ edges (weights of the edges) of the form $(v_{i1}^{(2)}, v_j^{(2)})$, where $v_j^{(1)}$ is not visited before the ith step. Note that, $\mathcal{T}_1$ is traversed in $n-1$ steps, but we get non-degenerate ranks $R_1, R_2, \ldots, R_{n-2}$ in the first $n-2$ steps only. Heller et al. (2012) argued that under the null hypothesis of independence, $R_1, R_2, \ldots, R_{n-2}$ are mutually independent, and R_i ($i = 1, 2, \ldots, n-2$) follows a discrete uniform distribution with mass points $1, 2, \ldots, n-i$. So, their proposed test statistic $T^{HGH} = -2\sum_{i=1}^{n-2} \log(\frac{R_i}{n-i})$ has the exact distribution-free property in finite sample situations. The null hypothesis is rejected for smaller values of the R_is and hence for higher values of T^{HGH} (see Heller et al., 2012 for details).

Now, consider a simple example with $p = q = 2$ and $n = 5$, where $\mathbf{x}_i = \mathbf{y}_i$ for $i = 1, 2, \ldots, 5$. Clearly, in this case, there is an extreme dependence between $\mathbf{x}$ and $\mathbf{y}$. Fig. 1 shows the scatter plot of the observations on $\mathbf{x}$ (and $\mathbf{y}$) and the corresponding complete graph $\mathcal{G}_1$ (and $\mathcal{G}_2$) along with all its edge weights. It also shows the MST $\mathcal{T}_1$ (and $\mathcal{T}_2$) and the weights of its edges in black. Now consider the following random traversal of $\mathcal{T}_1$: $(v_{11}^{(1)} = \mathbf{x}_5, v_{12}^{(1)} = \mathbf{x}_4)$, $(v_{21}^{(1)} = \mathbf{x}_5, v_{22}^{(1)} = \mathbf{x}_3)$, $(v_{31}^{(1)} = \mathbf{x}_5, v_{32}^{(1)} = \mathbf{x}_2)$, $(v_{41}^{(1)} = \mathbf{x}_5, v_{42}^{(1)} = \mathbf{x}_1)$. Therefore, R_1, the rank of the weight of $(\mathbf{y}_5, \mathbf{y}_4)$ among the weights of $(\mathbf{y}_5, \mathbf{y}_1)$, $(\mathbf{y}_5, \mathbf{y}_2)$, $(\mathbf{y}_5, \mathbf{y}_3)$ and $(\mathbf{y}_5, \mathbf{y}_4)$, turns out to 4. Similarly, we have $R_2 = 3$ and $R_3 = 2$. So, in all these cases, R_i ($i = 1, 2, 3$) takes its highest possible value. As a result, the test based on T^{HGH} fails to reject H_0, the null hypothesis of independence. This example shows that even in the case of extreme dependence, random traversal can yield misleading results. This problem with random traversal becomes more evident in high dimensions, where some of the nodes often have much higher degrees in the MST (like the node $\mathbf{x}_5$ in the above example) compared to the rest. In the computer science literature, these nodes are called hubs. Note that in one dimension, a node v_0 can be the nearest neighbor of at most two other nodes, one located on the left and