Journal of Multivariate Analysis 141 (2015) 168–178

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On high dimensional two-sample tests based on nearest neighbors

Pronoy K. Mondal, Munmun Biswas, Anil K. Ghosh *

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

ARTICLE INFO

Article history:
Received 2 February 2015
Available online 21 July 2015

AMS 2010 subject classifications:
62G10
62H15

Keywords:
Central limit theorem
HDLSS data
Large sample test
Law of large numbers
Level and power of a test
Permutation test

ABSTRACT

In this article, we propose new multivariate two-sample tests based on nearest neighbor type coincidences. While several existing tests for the multivariate two-sample problem perform poorly for high dimensional data, and many of them are not applicable when the dimension exceeds the sample size, these proposed tests can be conveniently used in the high dimension low sample size (HDLSS) situations. Unlike Schilling (1986) [26] and Henze's (1988) test based on nearest neighbors, under fairly general conditions, these new tests are found to be consistent in HDLSS asymptotic regime, where the sample size remains fixed and the dimension grows to infinity. Several high dimensional simulated and real data sets are analyzed to compare their empirical performance with some popular two-sample tests available in the literature. We further investigate the behavior of these proposed tests in classical asymptotic regime, where the dimension of the data remains fixed and the sample size tends to infinity. In such cases, they turn out to be asymptotically distribution-free and consistent under general alternatives.

1. Introduction

In a two-sample problem, we test the equality of two d-dimensional distributions F and G based on two sets of independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1} \overset{i.i.d.}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2} \overset{i.i.d.}{\sim} G$. If F and G are assumed to be same except for their locations, it leads to a two-sample location problem, where we test the equality of the locations of F and G. For instance, if F and G are assumed to be Gaussian with a unknown common dispersion matrix, one uses the Hotelling's T^2 statistic to test the equality of their means. Nonparametric tests for the multivariate two-sample location problem include [4,22,23,18,9,17]. But, most of these tests perform poorly for high dimensional data, and none of them can be used when the dimension d exceeds the combined sample size $n = n_1 + n_2$. Two-sample location tests that can be used in high dimension, low sample size (HDLSS) situations include [1,8,21,28].

Several nonparametric tests have been proposed for the general two-sample problem as well, where we test the equality of two continuous distributions F and G without making any further assumptions on them. Friedman and Rafsky [11] used minimal spanning tree for multivariate generalizations of the Wald–Wolfowitz run test and the Kolmogorov–Smirnov maximum deviation test. Schilling [26] and Henze [16] developed two sample tests based on nearest neighbor type coincidences. Other non-parametric tests for the general two sample problem include [10,14,30,2,3,25,20,12]. Most of these tests are based on pairwise distances between the observations, and they can be used for HDLSS data.

Following [26,16], in this article, we develop multivariate two sample tests based on nearest neighbors. Like the nearest neighbor test of [26,16] (henceforth, we will refer to it as the NN test), these proposed tests have the large sample consistency

* Corresponding author.
E-mail addresses: pmpronoykanti96@gmail.com (P.K. Mondal), munmun.biswas08@gmail.com (M. Biswas), akghosh@isical.ac.in (A.K. Ghosh).

http://dx.doi.org/10.1016/j.jmva.2015.07.002
0047-259X/

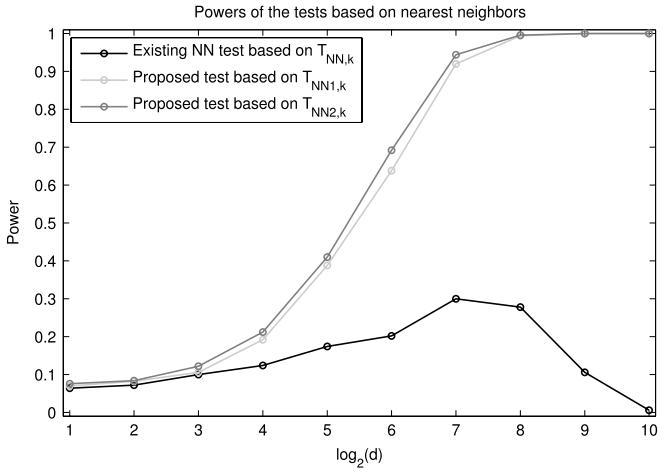

Fig. 1. Powers of nearest neighbor tests for varying choices of data dimension.

under general alternatives. However, this type of consistency in classical asymptotic regime is a rather trivial property of a test. The power of any reasonable test usually converges to unity when d remains fixed, and n increases. But, in HDLSS asymptotic regime, where n remains fixed and d tends to infinity, consistency of a test is no longer a trivial property. Many well known and popular tests fail to have the consistency in this set up (see e.g., [29,6]). We will see that the NN test also has a similar problem. To demonstrate this, let us consider a simple example, where the components of F are independent and identically distributed (i.i.d.) $N(0, 1)$, while those of G are i.i.d. $N(0.2, 1.2)$. We generated 20 observations from each distribution to test the null hypothesis $H_0 : F = G$ against the alternative hypothesis $H_1 : F \neq G$, and we carried out this experiment for different values of d ranging between 2 and 1024. For each value of d, we repeated the experiment 500 times, and the estimated power of the NN test (i.e., proportion of times it rejected H_0) is plotted in Fig. 1. Recall that the NN test rejects H_0 for large values of the statistic $T_{NN,k} = \frac{1}{kn}[\sum_{i=1}^{n_1}\sum_{r=1}^{k} I_{\mathbf{x}_i}(r) + \sum_{i=1}^{n_2}\sum_{r=1}^{k} I_{\mathbf{y}_i}(r)]$, where $I_{\mathbf{z}}(r)$ denotes the indicator variable that takes the value 1 if and only if $\mathbf{z}$ and its rth $(r \leq k)$ nearest neighbor come from the same distribution. For finding the neighbor of $\mathbf{z}$, here we use the leave-one-out method, where $\mathbf{z}$ itself is not considered as its neighbor. Throughout this article, for our numerical work, we use $k = 3$, which has been observed to perform well in the literature (see e.g., [26]).

Note that in this example, each and every component variable provides some evidence against H_0. So, one would expect the power of any reasonable test to increase to 1 as d increases. Surprisingly, that was not the case for the NN test. Initially its power increased with d, but then it dropped down to zero (see Fig. 1). Our proposed tests (described in Section 2) could overcome this limitation of the NN test. Their powers converged to unity as the dimension increased (see the power curves for tests based on $T_{NN1,k}$ and $T_{NN2,k}$ in Fig. 1). In the next section, we first investigate the reasons behind the failure of the NN test in the above example, and then we develop our proposed tests based on nearest neighbor type coincidences.

2. Proposed tests based on nearest neighbors

Let $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d}{\sim} F$, where the component variables are i.i.d. $N(\mu_1, \sigma_1^2)$, and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d}{\sim} G$, where the component variables are i.i.d. $N(\mu_2, \sigma_2^2)$. Clearly, $\|\mathbf{X}_1 - \mathbf{X}_2\|^2/2\sigma_1^2$ and $\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2/2\sigma_2^2$ both follow chi-square distribution with d degrees of freedom, while $\|\mathbf{X}_1 - \mathbf{Y}_1\|^2/(\sigma_1^2 + \sigma_2^2)$ follows non-central chi-square distribution with d degrees of freedom and the non-centrality parameter $(\mu_2 - \mu_1)^2/(\sigma_1^2 + \sigma_2^2)$. Here $\|\cdot\|$ denotes the usual Euclidean norm. It is easy to check that as $d \to \infty$, $d^{-1}\|\mathbf{X}_1 - \mathbf{X}_2\|^2 \xrightarrow{p} 2\sigma_1^2$, $d^{-1}\|\mathbf{Y}_1 - \mathbf{Y}_2\|^2 \xrightarrow{p} 2\sigma_2^2$ and $d^{-1}\|\mathbf{X}_1 - \mathbf{Y}_1\|^2 \xrightarrow{p} \sigma_1^2 + \sigma_2^2 + (\mu_2 - \mu_1)^2$. In fact, these above convergence results hold as long as the components of F and G are i.i.d. with finite second moments (follows from weak law of large numbers (WLLN)). In the example in Section 1, we had $\mu_1 = 0$, $\mu_2 = 0.2$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.2$ leading to $2\sigma_1^2 < \sigma_1^2 + \sigma_2^2 + (\mu_2 - \mu_1)^2 < 2\sigma_2^2$. Therefore, for large d, while each and every observation from F had its all k $(k = 3)$ neighbors from F, no observation from G had any of its neighbors from G. As a result, $T_{NN,k}$ attained the value 1/2, which was close to its expected value under H_0. Consequently, the NN test could not reject H_0 even on a single occasion. Now, let us define $T_{1,k} = \frac{1}{n_1 k}\sum_{i=1}^{n_1}\sum_{r=1}^{k} I_{\mathbf{x}_i}(r)$, the proportion of neighbors of $\mathbf{x}$-observations coming from F and $T_{2,k} = \frac{1}{n_2 k}\sum_{i=1}^{n_2}\sum_{r=1}^{k} I_{\mathbf{y}_i}(r)$, the proportion of neighbors of $\mathbf{y}$-observations coming from G. Under H_0, $T_{1,k}$ and $T_{2,k}$ are expected to be close to their expectations $E_{H_0}(T_{1,k}) = (n_1 - 1)/(n - 1)$ and $E_{H_0}(T_{2,k}) = (n_2 - 1)/(n - 1)$, respectively. But under H_1, the deviations $T_{1,k} - E_{H_0}(T_{1,k})$ and $T_{2,k} - E_{H_0}(T_{2,k})$ are supposed to be large. Note that the NN test statistic $T_{NN,k}$ is given by $T_{NN,k} = (n_1 T_{1,k} + n_2 T_{2,k})/n$, and hence $T_{NN,k} - E_{H_0}(T_{NN,k}) = \{n_1(T_{1,k} - E_{H_0}(T_{1,k})) + n_2(T_{2,k} - E_{H_0}(T_{2,k}))\}/n$. In our example, $T_{1,k}$ converges to 1 and $T_{2,k}$ converges to 0. So, in this type of examples, while $T_{1,k} - E_{H_0}(T_{1,k})$ turns out to be positive,