Journal of Multivariate Analysis 123 (2014) 160–171

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A nonparametric two-sample test applicable to high dimensional data

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 29 October 2012
Available online 25 September 2013

AMS subject classification:
62G10
62H15

Keywords:
High dimensional asymptotics
Inter-point distances
Large sample distribution
Permutation test
U-statistic
Weak law of large numbers

ABSTRACT

The multivariate two-sample testing problem has been well investigated in the literature, and several parametric and nonparametric methods are available for it. However, most of these two-sample tests perform poorly for high dimensional data, and many of them are not applicable when the dimension of the data exceeds the sample size. In this article, we propose a multivariate two-sample test that can be conveniently used in the high dimension low sample size setup. Asymptotic results on the power properties of our proposed test are derived when the sample size remains fixed, and the dimension of the data grows to infinity. We investigate the performance of this test on several high-dimensional simulated and real data sets, and demonstrate its superiority over several other existing two-sample tests. We also study some theoretical properties of the proposed test for situations when the dimension of the data remains fixed and the sample size tends to infinity. In such cases, it turns out to be asymptotically distribution-free and consistent under general alternatives.

1. Introduction

In a two-sample testing problem, we test the null hypothesis $H_0 : F = G$, which suggests the equality of two distributions F and G, against the alternative hypothesis $H_1 : F \neq G$. Usually, we have two sets of independent d-dimensional observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \overset{\text{i.i.d.}}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} G$, and using these observations, we compute a test statistic to perform the test. Instead of considering a general two-sample problem, sometimes we make some assumptions on F and G and test $H_0 : F = G$ in that restricted setup. For instance, if F and G are assumed to be same except for their locations (and/or scales), one can test for the equality of their locations (and/or scales). For a multivariate two-sample location problem, the Hotelling T^2 test is often used. While it is the most powerful invariant test for normally distributed data, other nonparametric tests outperform the Hotelling T^2 test for a wide variety of non-Gaussian distributions. Moreover, it cannot be used when the dimension of the data exceeds the sample size. Several attempts have been made in the literature to construct Hotelling T^2 type test statistics that can be applied to high dimensional data (see e.g., Bai and Saranadasa [2], Srivastava [28], Chen and Qin [5]), but these tests are also based on several model assumptions, and they are suitable only for location problems. Popular nonparametric tests for two-sample location problem include Puri and Sen [23], Randles and Peters [24], Hettmansperger and Oja [13], Möttönen and Oja [20], Choi and Marden [6] and Hettmansperger et al. [12]. Liu and Singh [17] and Rousson [26] constructed nonparametric tests for multivariate two-sample location and scale problems. Some good reviews of most of these tests can be found in Oja and Randles [22] and Oja [21]. However, all these above mentioned nonparametric tests perform poorly when applied to high dimensional data, and in practice, none of them can be used when the dimension of the data is larger than the sample size.

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

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.09.004

Multivariate nonparametric tests for a general two-sample problem have also been proposed in the literature. Friedman and Rafsky [7] used the idea of minimal spanning tree (MST) to generalize the univariate run test in multi-dimension. Schilling [27] and Henze [10] proposed two-sample tests based on nearest neighbor type coincidences. Other nonparametric tests for the general two sample problem include Hall and Tajvidi [9], Zech and Aslan [30], Baringhaus and Franz [3,4] and Liu and Modarres [16]. All these tests are rotation invariant, and they can be used even when the dimension of the data is larger than the sample size. Rosenbaum's [25] test can also be used in high dimension low sample size situations if the test statistic is computed using the Euclidean distance. Another interesting feature of these tests is that all of them are based on inter-point distances. These inter-point distances contain useful information about the separability between two distributions F and G. Under mild conditions, F and G differ if and only if $\|\mathbf{X}-\mathbf{X}_*\|$, $\|\mathbf{X}-\mathbf{Y}\|$ and $\|\mathbf{Y}-\mathbf{Y}_*\|$ differ in their distributions, where $\mathbf{X}, \mathbf{X}_* \overset{\text{i.i.d.}}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \overset{\text{i.i.d.}}{\sim} G$, and $\|\cdot\|$ denotes the Euclidean norm (see Maa et al. [19]). Such inter-point distances can be easily computed in any dimension. In this article, we use them to construct a new test for a general two-sample problem.

In Section 2, we begin with some simple examples that show the limitations of some of the popular two-sample tests in high dimension low sample size situations. In Section 3, we propose a new test to overcome these limitations and study the power properties of the proposed test when the sample size remains fixed, and the dimension of the data grows to infinity. Some high dimensional simulated and real data sets are also analyzed to compare its empirical performance with some existing two-sample tests. In Section 4, we study the asymptotic behavior of the power function of the proposed test in situations where the dimension of the data remains fixed and the sample size tends to infinity. We prove that the proposed test is asymptotically distribution-free and consistent under general alternatives. Finally, Section 5 contains a brief summary of the work and ends with a discussion on possible directions for further research. All proofs and mathematical details are given in the Appendix.

2. Some illustrative examples

Let us consider a two-sample problem, where the observations in F and G are distributed as $N_d((0,\ldots,0)', \mathbf{I}_d)$ and $N_d((\mu,\ldots,\mu)', \sigma^2\mathbf{I}_d)$, respectively. Here, N_d stands for a d-variate normal distribution, and $\mathbf{I}_d$ denotes the $d\times d$ identity matrix. We considered three different choices of μ and σ^2, namely, $(\mu = 0.3, \sigma^2 = 1)$, $(\mu = 0, \sigma^2 = 1.3)$ and $(\mu = 0.2, \sigma^2 = 1.2)$, and in each case, we generated 20 observations from each distribution to test $H_0 : F = G$. Note that these three choices of μ and σ^2 lead to a location problem, a scale problem and a location–scale problem, respectively. In each case, the experiment was repeated 200 times, and the proportion of times a test rejected H_0 was considered as an estimate of its power. These estimated powers were computed for three popular two-sample tests, namely, Friedman and Rafsky's [7] multivariate generalization of the run test, the test based on nearest neighbor (NN) type coincidences (see, e.g., Schilling [27], Henze [10]) and the test proposed by Baringhaus and Franz [3]. Henceforth, we will refer to them as the FR test, the NN test and the BF test, respectively. The FR test constructs an MST using $m+n$ sample observations, and the test statistic is given by $T^{FR}_{m,n} = 1+\sum_{i=1}^{N-1} U_i$, where $N = m+n$, and U_i $(i = 1, 2, \ldots, N-1)$ is an indicator variable that takes the value 1 if the i-th edge of the MST joins two observations from different populations, and 0 otherwise. Naturally, H_0 is rejected if $T^{FR}_{m,n}$ is small. The NN test statistic $T^{NN}_{m,n,k}$ (instead of $T^{NN}_{m,n}$, we use $T^{NN}_{m,n,k}$ for its dependence on the number of neighbors k) can be expressed as $T^{NN}_{m,n,k} = \frac{1}{Nk}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} I_j(\mathbf{x}_i) + \sum_{i=1}^{n}\sum_{j=1}^{k} I_j(\mathbf{y}_i)\right]$, where $I_j(\mathbf{z})$ is an indicator function takes the value 1 if $\mathbf{z}$ and its j-th neighbor belong to the same population, and 0 otherwise. This test rejects H_0 for large values of $T^{NN}_{m,n,k}$. The BF test is motivated by the result that $2E\|\mathbf{X}-\mathbf{Y}\| - E\|\mathbf{X}-\mathbf{X}_*\| - E\|\mathbf{Y}-\mathbf{Y}_*\| \geq 0$, where $\mathbf{X}, \mathbf{X}_* \overset{\text{i.i.d.}}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \overset{\text{i.i.d.}}{\sim} G$, and the equality holds iff $F = G$ (see Baringhaus and Franz [3]). The BF test statistic $T^{BF}_{m,n}$ is constructed by replacing the expectations with their empirical analogs, and the test rejects H_0 for large values of $T^{BF}_{m,n}$. We computed powers of these tests for different values of d ranging from 2 to 500, and the results are presented in Fig. 1.

Note that in each of these examples, as d increases, the separability between F and G also increases. So, one should expect the powers of these tests to tend to unity as d increases. We observed that in the case of the location problem (see Fig. 1(a)), but not in other two cases. In the location–scale problem, although the power of the BF test increased with d, those of the other two tests dropped down to zero as d increased (see Fig. 1(c)). In the case of the scale problem, all of these three methods yielded poor performance (see Fig. 1(b)). The reasons for such limitations of these existing methods will be discussed later (see Section 3.2). These limitations clearly show the necessity to develop a new test for high dimensional data. We construct one such test in the next section.

3. A new test based on inter-point distances

Consider four independent random vectors $\mathbf{X}, \mathbf{X}_* \overset{\text{i.i.d}}{\sim} F$ and $\mathbf{Y}, \mathbf{Y}_* \overset{\text{i.i.d}}{\sim} G$. Let D_{FF}, D_{GG} and D_{FG} denote the distributions of $\|\mathbf{X}-\mathbf{X}_*\|$, $\|\mathbf{Y}-\mathbf{Y}_*\|$ and $\|\mathbf{X}-\mathbf{Y}\|$, respectively, and μ_{FF}, μ_{GG} and μ_{FG} be their respective means. Under mild conditions, Maa et al. [19] proved that D_{FF}, D_{GG} and D_{FG} are identical if and only if $F = G$. Now, $(\|\mathbf{X}-\mathbf{X}_*\|, \|\mathbf{X}-\mathbf{Y}\|)$ follows a bivariate distribution, say D_F, with marginals D_{FF} and D_{FG}, respectively. Again, $(\|\mathbf{Y}-\mathbf{X}\|, \|\mathbf{Y}-\mathbf{Y}_*\|)$ follows another bivariate distribution, say D_G, with marginals D_{FG} and D_{GG}, respectively. So, when F and G differ, D_F and D_G differ as well, and vice versa. If $\boldsymbol{\mu}_{D_F}$ and $\boldsymbol{\mu}_{D_G}$ denote the mean vectors of D_F and D_G, respectively, we have $\boldsymbol{\mu}_{D_F} = \boldsymbol{\mu}_{D_G} \Leftrightarrow \mu_{FF} = \mu_{FG} = \mu_{GG}$, and that happens if and only if $F = G$ (see Lemma 1 in the Appendix). Therefore, instead of testing $H_0 : F = G$, we can test an