

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.

© 2013 Elsevier Inc. All rights reserved.

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 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \stackrel{\text{i.i.d.}}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \stackrel{\text{i.i.d.}}{\sim} G$, 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).

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 $\mathbf{X}, \mathbf{X}_* \stackrel{\text{i.i.d.}}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \stackrel{\text{i.i.d.}}{\sim} G$, 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 $N_d((0, \dots, 0)', \mathbf{I}_d)$ and $N_d((\mu, \dots, \mu)', \sigma^2 \mathbf{I}_d)$, 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 $T_{m,n}^{FR} = 1 + \sum_{i=1}^{N-1} U_i$, where $N = m+n$, and U_i ($i = 1, 2, \dots, N-1$) 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 $T_{m,n}^{FR}$ is small. The NN test statistic $T_{m,n,k}^{NN}$ (instead of $T_{m,n}^{NN}$, we use $T_{m,n,k}^{NN}$ for its dependence on the number of neighbors k) can be expressed as $T_{m,n,k}^{NN} = \frac{1}{Nk} \left[\sum_{i=1}^m \sum_{j=1}^k I_j(\mathbf{x}_i) + \sum_{i=1}^n \sum_{j=1}^k I_j(\mathbf{y}_i) \right]$, 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 $T_{m,n,k}^{NN}$. 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 $\mathbf{X}, \mathbf{X}_* \stackrel{\text{i.i.d.}}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \stackrel{\text{i.i.d.}}{\sim} G$, and the equality holds iff $F = G$ (see Baringhaus and Franz [3]). The BF test statistic $T_{m,n}^{BF}$ is constructed by replacing the expectations with their empirical analogs, and the test rejects H_0 for large values of $T_{m,n}^{BF}$. 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 $\mathbf{X}, \mathbf{X}_* \stackrel{\text{i.i.d.}}{\sim} F$ and $\mathbf{Y}, \mathbf{Y}_* \stackrel{\text{i.i.d.}}{\sim} G$. 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 μ_{D_F} and μ_{D_G} denote the mean vectors of D_F and D_G , respectively, we have $\mu_{D_F} = \mu_{D_G} \Leftrightarrow \mu_{FF} = \mu_{FG} = \mu_{GG}$, 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

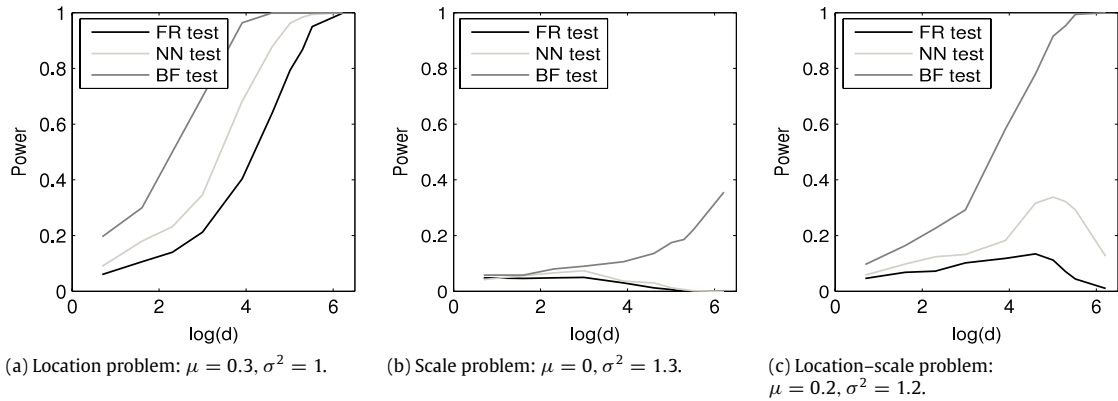


Fig. 1. Powers of FR, NN and BF tests for various choices of d .

equivalent null hypothesis $H_0'' : \mu_{D_F} = \mu_{D_G}$ against the alternative $H_1'' : \mu_{D_F} \neq \mu_{D_G}$. Our test statistic is very simple and easy to compute. From the data $\mathbf{x}_1, \dots, \mathbf{x}_m \stackrel{\text{i.i.d.}}{\sim} F$ and $\mathbf{y}_1, \dots, \mathbf{y}_n \stackrel{\text{i.i.d.}}{\sim} G$, we calculate

$$\hat{\mu}_{D_F} = \left[\hat{\mu}_{FF} = \binom{m}{2}^{-1} \sum_{i=1}^m \sum_{j=i+1}^m \|\mathbf{x}_i - \mathbf{x}_j\|, \hat{\mu}_{FG} = (mn)^{-1} \sum_{i=1}^m \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{y}_j\| \right],$$

$$\hat{\mu}_{D_G} = \left[\hat{\mu}_{FG} = (mn)^{-1} \sum_{i=1}^m \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{y}_j\|, \hat{\mu}_{GG} = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=i+1}^n \|\mathbf{y}_i - \mathbf{y}_j\| \right],$$

and reject H_0 for higher values of the test statistic $T_{m,n} = \|\hat{\mu}_{D_F} - \hat{\mu}_{D_G}\|^2$. When the sample size is small, we use the permutation principle to calculate the cut-off. When it is large, this cut-off is chosen using the large sample distribution of $T_{m,n}$ (see Section 4). Note that like FR, NN and BF tests, this test is invariant under location change, rotation and homogeneous scale transformation.

Fig. 2 shows the performance of this proposed test in the three examples with normal distributions discussed in Section 2. Unlike what we observed earlier, in all these three cases, the power of the proposed test converged to 1 as the dimension increased. In the scale problem and the location-scale problem, it outperformed all the three tests discussed earlier. Only in the case of the location problem, the BF test had the best performance. In the next two sections, we will investigate the reasons for such behavior of the power functions of these tests.

3.1. Behavior of the proposed test in high dimensions

We have already seen that our proposed test can be used for high dimensional data even when the dimension is much larger than the sample size. In this section, we investigate the limiting behavior of this test when m and n are fixed, while d diverges to infinity. Suppose that μ_1 and Σ_1 are the mean vector and the dispersion matrix of the distribution F , and μ_2 and Σ_2 are the mean vector and the dispersion matrix of the distribution G . In order to carry out our investigation, following Hall et al. [8], we first make the following assumption

(A1) There exist $\sigma_1^2, \sigma_2^2 > 0$ and ν such that (i) $\text{trace}(\Sigma_1)/d \rightarrow \sigma_1^2$ (ii) $\text{trace}(\Sigma_2)/d \rightarrow \sigma_2^2$ and (iii) $\|\mu_1 - \mu_2\|^2/d \rightarrow \nu^2$ as $d \rightarrow \infty$.

In conventional asymptotics, we get more information about the separability between F and G as the sample sizes increase, but here we consider the sample sizes to be fixed, and unless $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$, under (A1), we expect to get more information as the dimension increases.

Hall et al. [8] looked at the d -dimensional observations $\mathbf{X} = (X^{(1)}, X^{(2)}, \dots, X^{(d)})$ and $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \dots, Y^{(d)})$ as infinite time series truncated at time d and studied the behavior of the inter-point distances as d increases. Here, we look at these observations from a multivariate data perspective. So, here we make some minor modifications to the assumptions of Hall et al. [8]. Our assumptions are given below.

(A2) Fourth moments of the components of \mathbf{X} and \mathbf{Y} are uniformly bounded.

(A3) Under some permutation of the $X^{(q)}$ s (and the same permutation of the $Y^{(q)}$ s), for $(U^{(q)}, V^{(q)}) = (X^{(q)}, X_*^{(q)}), (X^{(q)}, Y^{(q)})$ and $(Y^{(q)}, Y_*^{(q)})$, the sequence $\{(U^{(q)} - V^{(q)})^2, q \geq 1\}$ is ρ mixing, i.e., $\sup_{1 \leq q < q' \leq \infty, |q - q'| > r} |\text{corr}\{(U^{(q)} - V^{(q)})^2, (U^{(q')} - V^{(q')})^2\}| \leq \rho(r)$ where $\rho(r) \rightarrow 0$ as $r \rightarrow \infty$.

Jung and Marron [14] assumed similar conditions for the large dimensional consistency of estimated principal component directions. Andrews [1] also assumed similar conditions to derive the weak law of large numbers (WLLN) for mixingales.

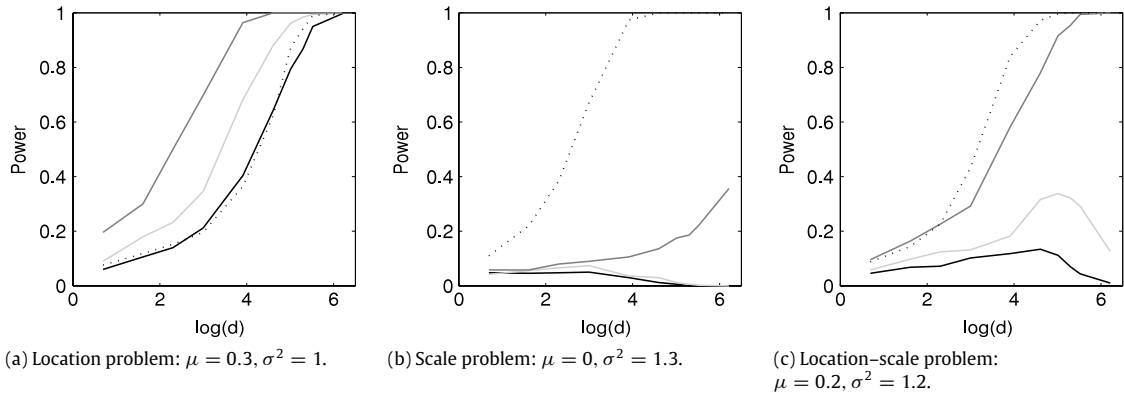


Fig. 2. Power of FR (black solid line), NN (light gray line), BF (dark gray line) and proposed tests (black dotted line) for varying choices of d .

Under the assumptions on uniformly bounded moments (A2) and weak dependence among component variables (A3), we have the WLLN for the sequence $\{(U^{(q)} - V^{(q)})^2, q \geq 1\}$ (see Lemma 2 in the Appendix). Again, depending on the choice of $(U^{(q)}, V^{(q)})$ $[(U^{(q)}, V^{(q)}) = (X^{(q)}, X_*^{(q)}), (X^{(q)}, Y^{(q)})$ or $(Y^{(q)}, Y_*^{(q)})]$, under the assumption (A1), $d^{-1} \sum_{q \geq 1} E(U^{(q)} - V^{(q)})^2$ converges to $2\sigma_1^2$, $2\sigma_2^2$ or $\sigma_1^2 + \sigma_2^2 + v^2$. So, under (A1)–(A3), as d tends to infinity, we have

- (a) $d^{-1/2} \|\mathbf{X}_i - \mathbf{X}_j\| \xrightarrow{P} \sigma_1 \sqrt{2}$ for $1 \leq i < j \leq m$.
- (b) $d^{-1/2} \|\mathbf{Y}_i - \mathbf{Y}_j\| \xrightarrow{P} \sigma_2 \sqrt{2}$ for $1 \leq i < j \leq n$.
- (c) $d^{-1/2} \|\mathbf{X}_i - \mathbf{Y}_j\| \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Therefore, after re-scaling by the factor $d^{-1/2}$, the $N = m + n$ points of the combined sample are asymptotically (as $d \rightarrow \infty$) located at the vertices of a N -polyhedron in $(N - 1)$ -dimensional space. Note that m of these vertices are the limits of m data points from F , and they form a regular simplex δ_1 of side length $\sigma_1 \sqrt{2}$. The other n vertices are the limits of n data points from G , and they form another regular simplex δ_2 of edge length $\sigma_2 \sqrt{2}$. The rest of the edges of the polyhedron are the edges connecting the vertices of δ_1 to those of δ_2 , and they are of length $\sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$. So, after re-scaling by the factor $d^{-1/2}$, the whole data cloud looks like a deterministic polyhedron in the limit, and the randomness appears only as a random rotation of this polyhedron (see also Hall et al. [8]). Under H_0 , when we have $\sigma_1^2 = \sigma_2^2$ and $v^2 = 0$, the whole polyhedron turns out to a regular simplex of N points. Note that if the components of \mathbf{X} and \mathbf{Y} vectors are independent and identically distributed with finite second moments, as it was the case in the examples with normal distributions in Section 2, (A1) and (A3) hold automatically. Further, instead of (A2), the finiteness of second order moments of the component variables is sufficient for the results (a)–(c) and hence the above high dimensional geometry to hold. We will keep this geometry in mind while investigating the behavior of the proposed and other two-sample tests in the case of high dimension low sample size data.

Theorem 3.1. Suppose that we have m independent observations from each of F and G , which satisfy (A1)–(A3). Also assume that either $v^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. Then, unless m is very small (i.e., $\binom{2m}{m} \leq 2/\alpha$), the power of the proposed test of level α converges to 1 as d tends to infinity.

Note that for all choices of m and n , as $d \rightarrow \infty$, under the assumptions (A1)–(A3), the results (a)–(c) hold, and hence we have $T_{m,n}/d \xrightarrow{P} \gamma_0$, where $\gamma_0 = (\sigma_1 \sqrt{2} - \sqrt{\sigma_1^2 + \sigma_2^2 + v^2})^2 + (\sigma_2 \sqrt{2} - \sqrt{\sigma_1^2 + \sigma_2^2 + v^2})^2$. Now, from the proof of Theorem 3.1 (see Appendix), it is clear that for $m = n$, the limiting p -value of the permutation test, i.e., the limiting value (as $d \rightarrow \infty$) of $P(T_{m,n}/d \geq \gamma_0)$ under the permutation distribution is $2/\binom{2m}{m}$. So, for a test of level 0.05 (or 0.01), it is enough to have four (or five) observations from each class for the convergence of the power to unity. The case $m \neq n$ calls for more complicated calculations, but for $m \geq 4$ and $n \geq 4$ (or $m \geq 5$ and $n \geq 5$), it can be viewed as the case $m = n = 4$ (or $m = n = 5$) with some additional information on at least one of the distributions. So, the resulting test is expected to have more power, and one can expect it to have the large dimensional consistency for all such choices of m and n . Fig. 3, which shows the limiting p -values for different choices of $m \geq 4$ and $n \geq 4$, justifies this claim. In this figure, one can notice that for each of the three examples discussed in Section 2, the limiting p -values were almost the same, and for any fixed choice of m or n , the limiting p -value was non-increasing in n or m , respectively. We carried out our investigation for various other choices of the parameters v , σ_1^2 and σ_2^2 , but this basic pattern remained the same.

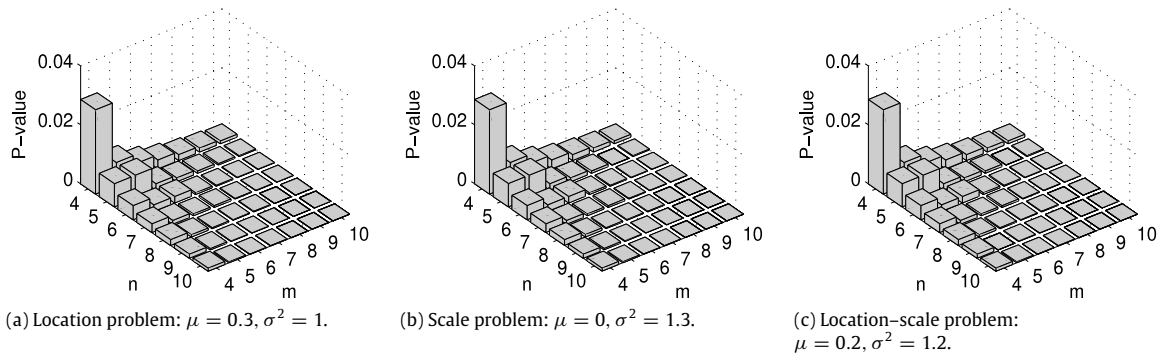


Fig. 3. Limiting p -values for different choices of m and n .

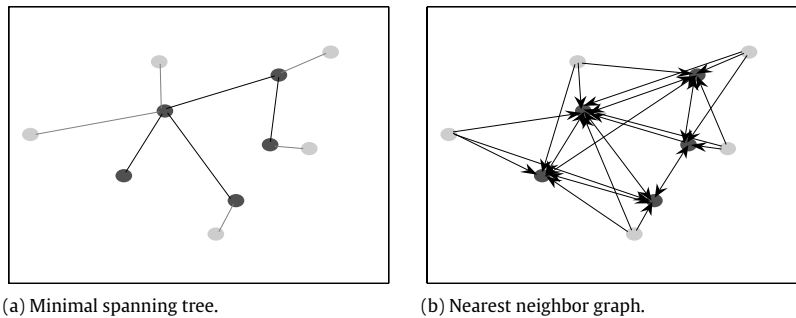


Fig. 4. Minimal spanning tree and nearest neighbor graph for 10 ($m = n = 5$) 500-dimensional observations.

3.2. Behavior of FR, NN and BF tests in high dimensions

Recall the definitions of the FR test statistic $T_{m,n}^{FR}$ and the NN test statistic $T_{m,n,k}^{NN}$ given in Section 2. From results (a)–(c) in Section 3.1, it is clear that in the case of the location problem, when we have $\sigma_1^2 = \sigma_2^2$ and $v^2 > 0$, for large d , each and every observation from F (or G , respectively) has all of its first $m - 1$ (or $n - 1$, respectively) nearest neighbors from F (or G , respectively) itself. As a result, $T_{m,n}^{FR}$ attains its minimum value, and $T_{m,n,k}^{NN}$ (throughout this article, we use $k = 3$, which has been reported to perform well in the literature) attains its maximum value with probability tending to one. So, these two tests are expected to perform well.

However, this ordering of intra-class and inter-class distances does not hold in the case of high dimension small sample size data if $v^2 < |\sigma_1^2 - \sigma_2^2|$. Without loss of generality, let us assume $\sigma_2^2 > v^2 + \sigma_1^2$ as it was the case in the scale problem and the location-scale problem discussed in Section 2. In these cases, each observation from F has its first $m - 1$ neighbors from F as before, but each observation from G has all of its first m nearest neighbors from F (see Fig. 4 that shows the MST and NN graph for $m = n = 5, k = 3$ and $d = 500$, where an arrow from ‘A’ to ‘B’ indicates that ‘B’ is one of the first k nearest neighbors of ‘A’). As a result, for higher values of d , $T_{m,n}^{FR}$ and $T_{m,n,k}^{NN}$ take the values, which are either equal or close to their expected values under H_0 , and well below (above in the case of $T_{m,n}^{FR}$) the cut-off. This is precisely the reason why these two tests failed in cases of scale and location-scale problems discussed in Section 2. In fact, in such cases, depending on m and n , the power of these tests may even tend to zero as d tends to infinity. This result is given by the following theorem.

Theorem 3.2. Suppose that the distributions F and G satisfy (A1)–(A3) and $v^2 < \sigma_2^2 - \sigma_1^2$ (interchange F and G , if required, and also interchange m and n accordingly).

- If $n(n - m)/(n - 1)(m + n) > \alpha$, the power of a level α test based on the FR test statistic $T_{m,n}^{FR}$ converges to zero as $d \rightarrow \infty$.
- If $(n - 1)/m > (1 + \alpha)/(1 - \alpha)$ and $k < \min\{m, n\}$, the power of a level α test based on the NN test statistic $T_{m,n,k}^{NN}$ converges to zero as $d \rightarrow \infty$.

Note that this theorem gives only two sufficient conditions under which the NN test and the FR test fail. These tests may fail in many other cases like the examples with normal distributions discussed in Section 2, where we had $m = n = 20$.

Now consider the BF test statistic given by $T_{m,n}^{BF} = 2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG}$. It is easy to see that when m and n are fixed and $d \rightarrow \infty$, under the assumptions (A1)–(A3), $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$, $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$, and hence the scaled version of the BF test statistic, $T_{m,n}^{BF}/\sqrt{d}$, converges to $2\sqrt{\sigma_1^2 + \sigma_2^2 + v^2} - \sigma_1\sqrt{2} - \sigma_2\sqrt{2}$ ($= \gamma$, say) in probability.

Now, γ is positive unless $\sigma_1^2 = \sigma_2^2$ and $v^2 = 0$, and a consistency result similar to [Theorem 3.1](#) can be proved for the BF test as well. But, in [Section 2](#), we observed that in the location–scale problem and the scale problem, especially in the latter case, it did not perform well. Note that in such cases, we had $v^2 < |\sigma_1^2 - \sigma_2^2|$. Now, $v^2 < |\sigma_1^2 - \sigma_2^2|$ implies that $\sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$ lies between $\sigma_1\sqrt{2}$ and $\sigma_2\sqrt{2}$. So, even when both $(\hat{\mu}_{FG} - \hat{\mu}_{FF})$ and $(\hat{\mu}_{FG} - \hat{\mu}_{GG})$ are significantly different from zero, they are likely to be of different signs. As a result, when they are added up, $T_{m,n}^{BF} = (\hat{\mu}_{FG} - \hat{\mu}_{FF}) + (\hat{\mu}_{FG} - \hat{\mu}_{GG})$ may take a value close to zero, and consequently, H_0 may get accepted. We observed that several times in the location–scale problem and the scale problem in [Section 2](#). In the case of scale problem, γ was also close to zero. So, even for $d = 500$, the BF test did not have satisfactory power. But, if we take the sum of $(\hat{\mu}_{FG} - \hat{\mu}_{FF})^2$ and $(\hat{\mu}_{FG} - \hat{\mu}_{GG})^2$, such cancellations are not possible, and H_0 is more likely to be rejected. That is why our test based on $T_{m,n} = (\hat{\mu}_{FG} - \hat{\mu}_{FF})^2 + (\hat{\mu}_{FG} - \hat{\mu}_{GG})^2$ had a better performance in these two examples. One should also notice that $T_{m,n}$ can also be expressed as $T_{m,n} = \frac{1}{2} [(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG})^2 + (\hat{\mu}_{FF} - \hat{\mu}_{GG})^2]$, where the first part $(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ is the square of the BF test statistic. We have seen that the BF test works well when F and G differ in their locations, but it is not very sensitive against small changes in scale. The second part $(\hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ compensates for that and makes the test sensitive against scale alternatives. However, in cases of pure location problems, the term $(\hat{\mu}_{FF} - \hat{\mu}_{GG})^2$ serves as noise. Therefore, in such cases, our proposed test is unlikely to outperform the BF test, and that is what we observed in our experiment.

Baringhaus and Franz [\[4\]](#) proposed a class of rigid motion invariant two-sample tests that includes the BF test discussed earlier. They considered a continuous function $\phi : [0, \infty) \rightarrow [0, \infty)$ and defined $\mu_{FF}^\phi = E\phi(\|\mathbf{X} - \mathbf{X}_*\|^2)$, $\mu_{GG}^\phi = E\phi(\|\mathbf{Y} - \mathbf{Y}_*\|^2)$ and $\mu_{FG}^\phi = E\phi(\|\mathbf{X} - \mathbf{Y}\|^2)$. They also proved that if ϕ is non decreasing and it satisfies some appropriate regularity conditions, the inequality $2\mu_{FG}^\phi - \mu_{FF}^\phi - \mu_{GG}^\phi \geq 0$ is satisfied, where the equality holds iff $F = G$. So, replacing μ_{FF}^ϕ , μ_{GG}^ϕ and μ_{FG}^ϕ by their empirical analogs, a class of test statistics $2\hat{\mu}_{FG}^\phi - \hat{\mu}_{FF}^\phi - \hat{\mu}_{GG}^\phi$ was constructed. Note that when H_0 fails to hold, due to monotonicity of ϕ , depending on the ordering of the three types of distances, here also $\hat{\mu}_{FG}^\phi$ can lie between $\hat{\mu}_{FF}^\phi$ and $\hat{\mu}_{GG}^\phi$. In such cases, due to the cancellation of positive and negative terms, the test statistic may take small values leading to the acceptance of H_0 . But such cancellations are not possible if we use $(\hat{\mu}_{FG}^\phi - \hat{\mu}_{FF}^\phi)^2 + (\hat{\mu}_{FG}^\phi - \hat{\mu}_{GG}^\phi)^2$ as the test statistic. As a consequence, the resulting test can have better power properties in such situations.

3.3. Results from the analysis of simulated data sets

We carried out simulation studies to evaluate the performance of our proposed method in high dimensional data. For this study, we used some examples involving 500 dimensional normal and Laplace distributions as well as some examples involving auto-regressive processes. In all these cases, we generated 20 observations from each of two distributions (say, F and G) to constitute the sample and used it to test $H_0 : F = G$ against $H_1 : F \neq G$. Each experiment was carried out 200 times as before. We estimated the power of the proposed test by the proportion of times it rejected H_0 , and the results are reported in [Table 1](#). To facilitate comparison, results are also reported for some other popular two-sample tests that can be used for high dimensional data. Along with FR, NN and BF tests, we used two other tests, one proposed by Hall and Tajvidi [\[9\]](#) and the other proposed by Rosenbaum [\[25\]](#). In future, we will refer to them as the HT test and the Rosenbaum test, respectively. The Rosenbaum test statistic has the distribution free property. We used the R package ‘nbpMatching’ to compute the value of this statistic (using the Euclidean distance) based on non-bipartite matching (see e.g., Lu et al. [\[18\]](#)) and then used its null distribution to perform the test. Since the sizes of the samples were small, for all other methods, we used the conditional tests based on the permutation principle.

Let us begin with some examples involving normal distributions. In [Section 2](#), we used some examples with multivariate normal distributions, where the component variables $X^{(1)}, \dots, X^{(d)}$ (and $Y^{(1)}, \dots, Y^{(d)}$) were independent and identically distributed. So, here we consider some examples, where both in F and G , the component variables are positively correlated. While F has the mean vector $(0, 0, \dots, 0)'$ and the dispersion matrix $\Sigma_1 = \mathbf{S}$, those for G are taken to be $(\mu, \mu, \dots, \mu)'$ and $\Sigma_2 = \sigma^2 \mathbf{S}$, respectively, where $\mathbf{S} = ((s_{ij}))$ is of the form $s_{ij} = (0.5)^{|i-j|}$ for $i, j = 1, 2, \dots, d$. Here also, we consider three different choices of μ and σ^2 $[(\mu, \sigma^2) = (0.25, 1), (0, 1.25) \text{ and } (0.1, 1.1)]$ to have three different types of problems. In cases of the scale problem and the location–scale problem, our proposed test yielded the highest power among all two-sample tests considered in this article. Only in the case of the location problem, the BF test and the NN test performed better than the proposed test. However, the proposed test and the HT test had comparable performance in this example as well, and they yielded much higher powers than those of the Rosenbaum test and the FR test.

We obtained similar results when we carried out our experiment with Laplace distributions, where the component variables in F and G were assumed to be independent and identically distributed. We considered three different types of problems (location, scale and location–scale) as before, and in each case, the component variables in F and G had the same means and variances as in the corresponding previous example with normal distributions. Again, in the location problem, the BF test and the NN test had the best performance, but in other two cases, the proposed test and the HT outperformed their competitors. In the case of the location–scale problem, the proposed test performed better than the HT test, while in other two cases, they had nearly the same power.

Table 1Observed powers of two-sample tests with 5% nominal level ($d = 500$, $m = n = 20$).

	Normal			Laplace			Normal vs. Laplace	AR(1) process	AR(2) process
	Location	Scale	Loc-scale	Location	Scale	Loc-scale			
BF	1.000	0.200	0.370	1.000	0.280	0.485	1.000	0.125	0.085
NN	0.945	0.000	0.115	1.000	0.005	0.150	0.000	0.060	0.060
FR	0.770	0.000	0.075	0.910	0.000	0.060	0.000	0.030	0.040
HT	0.835	1.000	0.925	0.700	1.000	0.790	1.000	0.885	0.965
Rosenbaum	0.675	0.090	0.110	0.865	0.100	0.115	0.885	0.105	0.060
Proposed	0.820	1.000	0.940	0.695	1.000	0.855	1.000	0.910	0.990

Next, we consider an example, where the component variables in F are i.i.d. standard normal variates, while those in G are i.i.d. standard Laplace variates. In this example, while the proposed test, the HT test and the BF test rejected H_0 in all of the 200 cases, the NN test and the FR test could not reject it even in a single occasion. The Rosenbaum test had power 0.885.

Finally, we consider two examples with auto-regressive (AR) processes, one with the AR(1) model and other with the AR(2) model. In the first example, we generated the observations in F using the AR(1) model $X^{(t)} = 0.25 + 0.3X^{(t-1)} + U_t$ for $t = 1, \dots, 500$, where $X^{(0)}, U_1, U_2, \dots, U_{500} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$. Observations in G were generated using another AR(1) model $Y^{(t)} = 0.25 + 0.4Y^{(t-1)} + V_t$, where $Y^{(0)}, V_1, V_2, \dots, V_{500} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$. Note that in this example, F and G have differences both in locations and scales. In the second example, F and G differ only in their scales. In this example, the observations in F were generated using the AR(2) model $X^{(t)} = 0.3X^{(t-1)} + 0.2X^{(t-2)} + U_t$ for $t = 1, 2, \dots, 500$, and those in G were generated using the model $Y^{(t)} = 0.35Y^{(t-1)} + 0.25Y^{(t-2)} + V_t$ for $t = 1, 2, \dots, 500$, where $X^{(0)}, X^{(-1)}, Y^{(0)}, Y^{(-1)}, U_1, U_2, \dots, U_{500}, V_1, V_2, \dots, V_{500}$ are all i.i.d. standard normal variates. In these two examples, the proposed method had excellent performance, and it outperformed all its competitors. While NN, FR, BF and Rosenbaum tests failed to yield satisfactory results (see Table 1), in these two examples, it had powers 0.91 and 0.99, respectively.

3.4. Results from the analysis of benchmark data sets

We analyzed three benchmark data sets, namely, the ECG data, the Synthetic Control Chart data and the Arcene data, for further evaluation of the proposed method. The ECG data set was taken from the UCR Time Series Classification/Clustering Page (http://www.cs.ucr.edu/~eamonn/time_series_data/), and the other two were taken from the UCI machine learning repository (<http://www.ics.uci.edu/ml/datasets>). Detailed descriptions of these data sets are available at these repositories. In the case of control chart data, though there are observations from six classes, for our analysis, we considered the two classes labeled as 'cyclic' and 'normal'. Several researchers have extensively investigated these three data sets, mainly in context of supervised classification. It is also well known that in all these data sets, we have reasonable separability between two competing classes. So in each of these cases, we can assume the alternative hypothesis to be the true, and different tests can be compared on the basis of their powers. Also, note that if we use the whole data set for testing, any test will either reject the null hypothesis or accept it. Based on that single experiment, it is difficult to compare among different test procedures. So in each of these cases, we repeated the experiment 500 times based on 500 different subsets chosen from the data at random, and the results are reported in Table 2. In all these cases, we chose m and n to be small compared to the dimension of the data.

ECG data consist of 200 observations, each of which is a time series recorded at 96 different time points. There are distinct training and test sets containing 100 observations each. For our analysis, we chose random subsets from the pooled data set consisting of 200 observations, 133 of them from one class (labeled as 'normal') and the rest from the other class (labeled as 'abnormal'). We considered subsets of three different sizes, and in each case, the experiment was repeated 500 times. In the case of $m = 20$, $n = 10$, i.e., when the subset sizes were proportional to the number of observations from that class in the pooled sample, the BF test, the NN test and our proposed test performed better than other three tests. In the case of equal subset size $m = n = 10$, the NN test had the best performance, but the performance of the proposed test and that of BF and HT tests were also comparable. The FR test and the Rosenbaum test had relatively low powers. In the case of $m = 10$ and $n = 20$, all methods except the Rosenbaum test rejected H_0 in more than 92% of the cases, while the NN test had the best performance.

Next, we consider the control chart data. It is a synthetically generated time series data set, which contains 60-dimensional observations from 6 classes. However, we considered only two classes ('normal' and 'cyclic') for our analysis. The time series in the 'normal' class are purely white noise, while those in the 'cyclic' class contain some cyclic pattern (see Fig. 5). There are 100 observations from each class, but we used subsets of size 5 (i.e., $m = n = 5$). In this data set, the HT test and our proposed test rejected H_0 in all the 500 cases, while the BF test failed only once. The NN test had power 0.916, but the FR test and the Rosenbaum test yielded poor performance.

The Arcene data set was obtained by merging three mass-spectrometric data sets. It contains 10,000-dimensional observations from two classes of 'cancer patients' (ovarian or prostate cancer) and 'healthy patients'. In the UCI repository, there is a training set and a validation set containing 100 observations each. We chose random subsets from the pooled data set of size 200 (88 cancer patients and 112 healthy patients) to carry out our experiment. Here also, we considered three

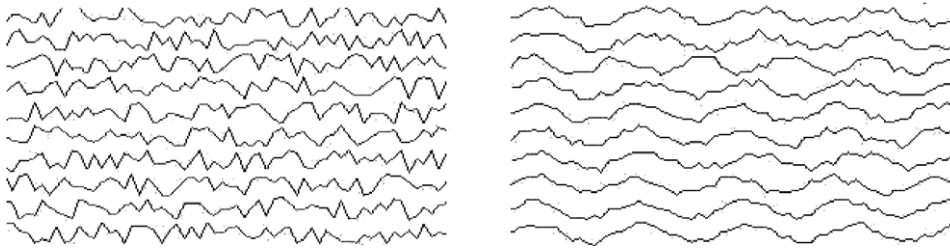


Fig. 5. 'Normal' (on the left) and 'cyclic' (on the right) classes in control chart data.

Table 2

Observed powers of two-sample tests with 5% nominal level in real data sets.

Data sets (m, n)	ECG			Control chart (5, 5)	Arcene		
	(20, 10)	(10, 10)	(10, 20)		(25, 30)	(25, 25)	(30, 25)
BF	0.978	0.862	0.958	0.998	0.758	0.696	0.736
NN	0.980	0.902	0.988	0.916	0.994	0.988	0.992
FR	0.880	0.752	0.922	0.298	0.972	0.944	0.966
HT	0.898	0.856	0.946	1.000	0.708	0.662	0.674
Rosenbaum	0.612	0.548	0.730	0.184	0.872	0.810	0.852
Proposed	0.960	0.868	0.928	1.000	0.808	0.752	0.794

choices of m and n (see Table 2). In all these cases, the NN test had the best performance closely followed by the FR test. Our proposed test also had reasonably high power, and it outperformed BF and HT tests on all these three occasions.

4. Large sample behavior of the proposed test

So far, we have investigated the behavior of our proposed test in high dimension low sample size situations. In this section, we study its large sample properties when the dimension of the data remains fixed. Here also, we use the test statistic $T_{m,n}$ to test $H_0 : F = G$ against $H_1 : F \neq G$ and reject the null hypothesis for higher values of $T_{m,n}$. However, it is computationally expensive to use the permutation method when m and n are too large. So, here we construct the test based on the large sample distribution of $T_{m,n}$. This asymptotic distribution is given by the following theorem.

Theorem 4.1. Suppose that we have two sets of independent observations $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$ and $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ from F , which has finite second moments. Also assume that as $N = (m + n) \rightarrow \infty$, $m/N \rightarrow \lambda$ for some $\lambda \in (0, 1)$. Then, $NT_{m,n}$ is asymptotically distributed as $\frac{2\sigma_0^2}{\lambda(1-\lambda)}\chi_1^2$, where $\sigma_0^2 = \text{Var}\{E(\|\mathbf{X}_1 - \mathbf{X}_2\| \|\mathbf{X}_1\|)\}$, and χ_1^2 denotes the chi-square distribution with 1 degree of freedom.

To construct a test based on this asymptotic distribution, one needs to find consistent estimates for λ and σ_0^2 . From the condition stated in the theorem, it is clear that $\hat{\lambda} = m/(m + n)$ is consistent for λ . To find a consistent estimate for σ_0^2 , first note that it can also be expressed as $\sigma_0^2 = \text{Cov}(\|\mathbf{X}_1 - \mathbf{X}_2\|, \|\mathbf{X}_1 - \mathbf{X}_3\|) = E(\|\mathbf{X}_1 - \mathbf{X}_2\| \|\mathbf{X}_1 - \mathbf{X}_3\|) - E^2(\|\mathbf{X}_1 - \mathbf{X}_2\|)$. Now define

$$S_1 = \left[\binom{m}{3}^{-1} \sum_{1 \leq i < j < k \leq m} \|\mathbf{X}_i - \mathbf{X}_j\| \|\mathbf{X}_i - \mathbf{X}_k\| \right] - \left[\binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} \|\mathbf{X}_i - \mathbf{X}_j\| \right]^2$$

$$\text{and } S_2 = \left[\binom{n}{3}^{-1} \sum_{1 \leq i < j < k \leq n} \|\mathbf{Y}_i - \mathbf{Y}_j\| \|\mathbf{Y}_i - \mathbf{Y}_k\| \right] - \left[\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \|\mathbf{Y}_i - \mathbf{Y}_j\| \right]^2.$$

From the results on the probability convergence of the U-Statistic (see e.g., Lee [15]), one can check that S_1 and S_2 both are consistent for σ_0^2 . Consequently, one can use $\hat{\sigma}_0^2 = (mS_1 + nS_2)/(m + n)$ as a consistent estimator of σ_0^2 and show that under H_0 , $T_{m,n}^* = (m + n)\hat{\lambda}(1 - \hat{\lambda})T_{mn}/2\hat{\sigma}_0^2 \sim \chi_1^2$. So, any test based on $T_{m,n}^*$ turns out to be asymptotically distribution-free. We compute $T_{m,n}^*$ from the data, and for a test of nominal level α , we reject H_0 if $T_{m,n}^*$ exceeds $\chi_{1,\alpha}^2$, where $P(\chi_1^2 > \chi_{1,\alpha}^2) = \alpha$. The following theorem shows that when d remains fixed, and m and n both tend to infinity, the power of this proposed test converges to one. So, it is consistent under general alternatives. Note that unlike our proposed test, the BF test and the HT test do not have the asymptotic distribution-free property. So, one has to use the bootstrap or the permutation method to find the cut-off, and this increases the computing cost substantially.

Theorem 4.2. Suppose that F and G both have finite second moments, and as $m, n \rightarrow \infty$, $m/(m + n) \rightarrow \lambda$ for some $\lambda \in (0, 1)$. Then, the power of our proposed test based on $T_{m,n}^*$ converges to 1 as m and n both tend to infinity.

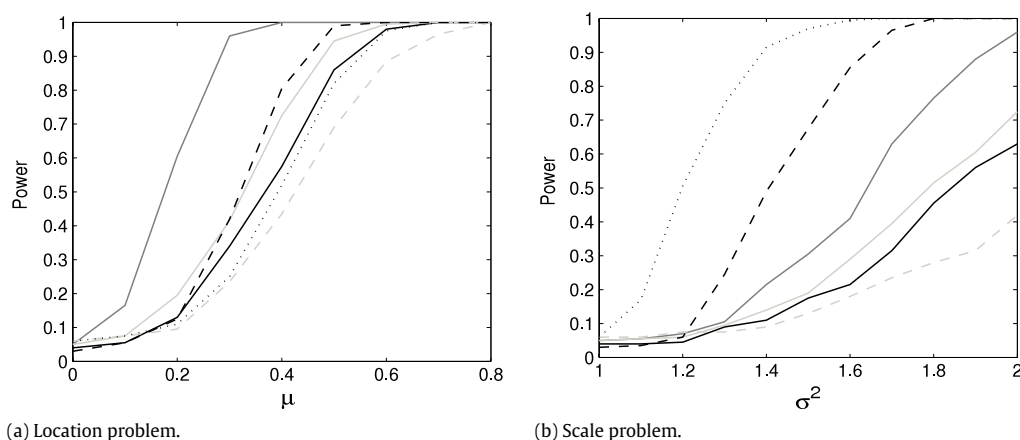


Fig. 6. Power curves of BF (dark gray solid line), FR (black solid line), NN (light gray solid line), Rosenbaum (light gray dashed line), HT (black dashed line) and proposed (black dotted line) tests for normal location and scale problems.

Fig. 6 shows the power curves of our proposed test based on the large sample distribution of $T_{m,n}^*$ in normal location and scale problems. It also shows the power curves of BF, NN, FR, HT and Rosenbaum tests for comparison. For FR, NN and Rosenbaum tests, we used the tests based on large sample distributions of the corresponding test statistics (see e.g., Henze and Penrose [11], Schilling [27], Rosenbaum [25]). For the NN test, we used the codes available at the R package 'MTSKNN', and for the FR test, we used our own codes. In the case of the BF test, we used the codes for the large sample test based on bootstrap approximation available at the R package 'cramer'. Since the large sample distribution of the HT test statistic is not known, we used its conditional version based on the permutation principle. Here F was considered to be a multivariate normal distribution with the location parameter $(0, 0, \dots, 0)'$ and the scatter matrix \mathbf{I} , while G differed from F either in location (μ, μ, \dots, μ) or in scatter $\sigma^2 \mathbf{I}$. Each experiment was carried out 200 times to estimate the powers of different tests, when we used $d = 5$ and $m = n = 100$. In the location problem, the BF test had the best performance, but in the case of the scale problem, once again, our proposed test outperformed all of its competitors. We observed the same phenomenon when we carried out our experiment with Laplace distributions. Therefore, to save space, we do not report it here. Clearly, these results are consistent with what we observed in Section 3. We also considered another example, where all the five component variables in F were i.i.d. standard normal variates, and those in G were i.i.d. standard Laplace variates. In this example, the proposed method had an excellent performance. While BF, NN, FR, HT and Rosenbaum tests had powers 0.605, 0.380, 0.340, 0.635 and 0.265, respectively, it rejected the null hypothesis in 99% of the cases.

5. Concluding remarks

In this article, we have proposed and investigated a two-sample test based on inter-point distances. This test is conceptually very simple and computationally very efficient. It can be conveniently used for high dimensional data or even for functional data if a distance function is defined in the functional space. In the high dimension low sample size setup, while many popular two-sample tests cannot be used at all, and many others lead to poor performance, this proposed test performs well. When the sample size remains fixed, and the dimension of the data increases, the power of this test converges to unity for a fairly general class of alternatives. Moreover, the large sample test based on this proposed test statistic has the distribution-free property, and it is consistent under general alternatives. In almost all simulated and real data sets that we have analyzed in this article, if not better, its performance was comparable to other two-sample tests used in this article.

Our proposed test can be generalized for multi-sample problems as well. Suppose that we want to test $H_0 : F_1 = F_2 = \dots = F_M$ based on n_1, n_2, \dots, n_M independent observations from M ($M > 2$) distributions F_1, F_2, \dots, F_M . In that case, we can compute T_{n_i, n_j} for each pair of classes and use $\sum_{i < j} T_{n_i, n_j}$ as the test statistic. Székely and Rizzo [29] considered a similar generalization of the BF statistic. Another option is to use $\sum_{i < j} \|\hat{\mu}_{D_{F_i}^0} - \hat{\mu}_{D_{F_j}^0}\|^2$ as a test statistic, where $\hat{\mu}_{D_{F_j}^0}$ is an estimate of $\mu_{D_{F_j}^0}$, the mean of an M -dimensional distribution $D_{F_j}^0$. Here $D_{F_j}^0$ ($j = 1, 2, \dots, M$) denotes the distribution of $(\|\mathbf{Z}_0 - \mathbf{Z}_1\|, \|\mathbf{Z}_0 - \mathbf{Z}_2\|, \dots, \|\mathbf{Z}_0 - \mathbf{Z}_M\|)$, where \mathbf{Z}_i 's are all independent, $\mathbf{Z}_0 \sim F_j$ and $\mathbf{Z}_i \sim F_i$ for all $i = 1, 2, \dots, M$. When the sample sizes are small, we can use the permutation test as before, but for the large sample test, we need to derive the asymptotic null distribution of the test statistic. One also needs to investigate the empirical performance of these tests for high dimensional data.

Though our proposed test performed well in almost all simulated and real data sets we analyzed in this article, in the case of the location problem, the BF test had better performance. So, it could be helpful if we use the union-intersection type method to combine these two tests or if we can come up with a method that uses the available data to automatically decide which of these two tests to be used in a particular problem.

Acknowledgments

We would like to thank two anonymous reviewers for their careful reading of an earlier version of the article and for providing us with several helpful comments.

Appendix. Proofs and mathematical details

Lemma 1. Suppose that $\mathbf{X}_1, \mathbf{X}_2 \stackrel{\text{i.i.d.}}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \stackrel{\text{i.i.d.}}{\sim} G$. Also assume that $\mu_{FF} = E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$, $\mu_{GG} = E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|)$ and $\mu_{FG} = E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ exist. Then, μ_{FF} , μ_{GG} and μ_{FG} are equal if and only if $F = G$.

Proof of Lemma 1. If $F = G$, there is nothing to prove. So, let us prove the ‘only if’ part. If $E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$, $E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|)$ and $E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ are equal, we have $2E(\|\mathbf{X}_1 - \mathbf{Y}_1\|) - E(\|\mathbf{X}_1 - \mathbf{X}_2\|) - E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|) = 0$. Now, from Baringhaus and Franz [3], we know that for F and G with finite expected norm, $2E(\|\mathbf{X}_1 - \mathbf{Y}_1\|) - E(\|\mathbf{X}_1 - \mathbf{X}_2\|) - E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|) \geq 0$, where equality holds if and only if F and G are identical. So, $E(\|\mathbf{X}_1 - \mathbf{X}_2\|) = E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|) = E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ implies $F = G$. \square

Lemma 2. If a sequence of random variables $\{W^{(q)}, q \geq 1\}$ has uniformly bounded second moments, and $\sup_{1 \leq q, q' < \infty, |q - q'| > r} |\text{corr}(W^{(q)}, W^{(q')})| < \rho(r)$, where $\rho(r) \rightarrow 0$ as $r \rightarrow \infty$, WLLN holds for the sequence $\{W^{(q)}, q \geq 1\}$.

Proof of Lemma 2. The proof is straightforward, and hence it is omitted.

Proof of Theorem 3.1. If F and G satisfy (A1)–(A3), using the results (a)–(c) stated in Section 3.1, for fixed m, n and $d \rightarrow \infty$, we have $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$ and $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$. Let these three limiting values be denoted by β_1, β_2 and β_3 , respectively. So, a re-scaled version of our test statistic, T_{mn}^d/d (instead of $T_{m,n}$, here we use $T_{m,n}^d$ to show its dependence on d) converges to $\gamma_0 = (\beta_1 - \beta_2)^2 + (\beta_2 - \beta_3)^2$ in probability.

Now, let us consider the permutation distribution of $T_{m,n}^d$. Note that if $m - r$ observations ($r = 0, 1, \dots, m$) from F and r observations from G are assumed to come from one distribution and the rest from the other distribution, as $d \rightarrow \infty$, the value of the test statistic converges to $\gamma_r = (\beta_{1,r} - \beta_{2,r})^2 + (\beta_{2,r} - \beta_{3,r})^2$ in probability, where $\beta_{1,r} = \left[\binom{m-r}{2} \beta_1 + (m-r)r \beta_2 + \binom{r}{2} \beta_3\right] / \binom{m}{2}$, $\beta_{2,r} = [(m-r)r \beta_1 + \{(m-r)^2 + r^2\} \beta_2 + r(m-r) \beta_3] / m^2$ and $\beta_{3,r} = \left[\binom{m-r}{2} \beta_3 + (m-r)r \beta_2 + \binom{r}{2} \beta_1\right] / \binom{m}{2}$. So, as $d \rightarrow \infty$, the permutation distribution tend to have $(m+1)$ mass points $\gamma_0, \gamma_1, \dots, \gamma_m$ with probabilities $\binom{m}{m} \binom{m}{0} / \binom{2m}{m}, \binom{m}{m-1} \binom{m}{1} / \binom{2m}{m}, \dots, \binom{m}{0} \binom{m}{m} / \binom{2m}{m}$, respectively.

Now, we will show that $\gamma_r \leq \gamma_0$ for all choices of r , where the equality holds for $r = 0$ and $r = m$. First note that under the given condition ($\sigma_1^2 \neq \sigma_2^2$ or $v^2 > 0$), we have $2\beta_2 - \beta_1 - \beta_3 > 0$. Also, note that $\beta_{1,r}, \beta_{2,r}$ and $\beta_{3,r}$ can be expressed as $\beta_{1,r} = \beta_1 + (m-r)r(\beta_2 - \beta_1) / \binom{m}{2} + \binom{r}{2}(\beta_3 - \beta_1) / \binom{m}{2}$, $\beta_{2,r} = \beta_2 + (m-r)r(\beta_1 + \beta_3 - 2\beta_2) / m^2$ and $\beta_{3,r} = \beta_3 + (m-r)r(\beta_2 - \beta_3) / \binom{m}{2} + \binom{r}{2}(\beta_1 - \beta_3) / \binom{m}{2}$. So, we have $\beta_{3,r} - \beta_{1,r} = (\beta_3 - \beta_1) \left[\binom{m}{2} - r(m-r) - 2\binom{r}{2}\right] / \binom{m}{2}$. Now, it is easy to check that $-1 \leq \left[\binom{m}{2} - r(m-r) - 2\binom{r}{2}\right] / \binom{m}{2} \leq 1$, which implies $|\beta_{3,r} - \beta_{1,r}| \leq |\beta_3 - \beta_1|$ or $(\beta_{3,r} - \beta_{1,r})^2 \leq (\beta_3 - \beta_1)^2$. Unless $\beta_1 = \beta_3$, here the equality holds only when $r = 0$ or $r = m$. Again, we have $2\beta_{2,r} - \beta_{1,r} - \beta_{3,r} = (2\beta_2 - \beta_1 - \beta_3) \left[1 - r(m-r)/m^2 - r(m-r) / \binom{m}{2}\right]$. Since $-1 \leq \left[1 - r(m-r)/m^2 - r(m-r) / \binom{m}{2}\right] \leq 1$, we have $(2\beta_{2,r} - \beta_{1,r} - \beta_{3,r})^2 \leq (2\beta_2 - \beta_1 - \beta_3)^2$, where the equality holds only when $r = 0$ or $r = m$. So, we have $\gamma_r = \frac{1}{2}[(2\beta_{2,r} - \beta_{1,r} - \beta_{3,r})^2 + (\beta_{3,r} - \beta_{1,r})^2] \leq \frac{1}{2}[(2\beta_2 - \beta_1 - \beta_3)^2 + (\beta_3 - \beta_1)^2] = \gamma_0$, where equality holds only for $r = 0$ and $r = m$.

Therefore, as $d \rightarrow \infty$, under the permutation distribution, the test statistic takes the value γ_0 or higher with probability tending to $2 / \binom{m}{2}$. So, for all m with $2 / \binom{m}{2} < \alpha$, the proposed test rejects H_0 with probability tending to 1 as d tends to infinity. \square

Lemma 3. Suppose that F and G satisfy (A1)–(A3) and $v^2 < \sigma_2^2 - \sigma_1^2$ (interchange F and G , if required, and also interchange m and n accordingly). Then, (a) $T_{m,n}^{FR} \xrightarrow{P} n + 1$ as $d \rightarrow \infty$ (b) for any $k < \min\{m, n\}$, $T_{m,n,k}^{NN} \xrightarrow{P} \frac{m}{N}$ as $d \rightarrow \infty$.

Proof of Lemma 3. Note that $0 \leq v^2 < \sigma_2^2 - \sigma_1^2$ implies $\sigma_1\sqrt{2} < \sqrt{\sigma_1^2 + \sigma_2^2 + v^2} < \sigma_2\sqrt{2}$. Therefore, if (A1)–(A3) hold, for every $i, i' = 1, 2, \dots, m$ ($i \neq i'$) and $j, j' = 1, 2, \dots, n$ ($j \neq j'$), $P(\|\mathbf{X}_i - \mathbf{X}_{i'}\| \leq \|\mathbf{X}_i - \mathbf{Y}_j\| \leq \|\mathbf{Y}_j - \mathbf{Y}_{j'}\|) \rightarrow 1$ as $d \rightarrow \infty$ (follows from the results (a)–(c) stated in Section 3.1). So, for each \mathbf{X}_i as well as for each \mathbf{Y}_j , all of its first k nearest neighbors come from F with probability tending to one.

(a) A minimal spanning tree (MST) on N vertices contains $N - 1$ edges, and $T_{m,n}^{FR} - 1$ gives the total number of edges (out of those $N - 1$ edges) that connect two nodes from two different populations. Now, from the above discussion, it is quite transparent that under the given conditions, the MST contains a sub-tree on m vertices corresponding to m observations from F , and each \mathbf{Y}_j is connected to one of the \mathbf{X}_i 's (see Fig. 4). So, as d tends to infinity, $T_{m,n}^{FR} \xrightarrow{P} n + 1$. \square

(b) Recall that the NN test statistic $T_{m,n,k}^{NN}$ can be expressed as

$$T_{m,n,k}^{NN} = \frac{1}{Nk} \left[\sum_{i=1}^m \sum_{j=1}^k I_j(\mathbf{X}_i) + \sum_{i=1}^n \sum_{j=1}^k I_j(\mathbf{Y}_i) \right],$$

where $I_j(\mathbf{Z})$ is an indicator function that takes the value 1 if \mathbf{Z} and its j -th neighbor belong to the same population and 0 otherwise. Now, from the above discussion, it is quite clear that $\sum_{i=1}^m \sum_{j=1}^k I_j(\mathbf{X}_i) \xrightarrow{P} km$ and $\sum_{i=1}^n \sum_{j=1}^k I_j(\mathbf{Y}_i) \xrightarrow{P} 0$ as $d \rightarrow \infty$. Consequently, $T_{m,n,k}^{NN} \xrightarrow{P} m/N$ as $d \rightarrow \infty$. \square

Proof of Theorem 3.2. (a) In Lemma 3, we have seen that under the given alternative, $T_{m,n}^{FR} \xrightarrow{P} n+1$ as $d \rightarrow \infty$. Now, given a sample, let c_d be the cut-off for a level α permutation test based on $T_{m,n}^{FR}$ in dimension d , and we reject H_0 if the observed value of $T_{m,n}^{FR}$ is smaller than c_d (for ease of notation, here we do not show the dependence of c_d on the given sample). So, it is enough to show that for sufficiently large d , $c_d < n+1$ almost everywhere. First note that if $\frac{n(n-m)}{(n-1)(m+n)} > \alpha$, we can choose a $\delta > 0$ such that $\frac{n(n-m)-\delta}{(n-1)(m+n)-\delta} > \alpha$. Now we will show that under the conditions of Theorem 3.2, for any given d , c_d is smaller than $n+1-\epsilon$ almost everywhere, where $\epsilon = \delta/N$.

Given an observed sample, let \mathbb{P} and \mathbb{E} denote the permutation probability distribution of the FR statistic and the expectation under it. Then, note that $\mathbb{P}(T_{m,n}^{FR} < c_d) \leq \alpha$ (equality holds in non-randomized cases), and $T_{m,n}^{FR}$ is bounded below by 2 (when there is only one edge between two observations belonging to two different populations). Now, if possible, suppose that $c_d \geq n+1-\epsilon$. Then we have $\mathbb{E}(T_{m,n}^{FR}) \geq 2\alpha + (1-\alpha)(n+1-\epsilon) = 1+\alpha + (n-\epsilon)(1-\alpha)$. Now, following Friedman and Rafsky [7], one can check that the expectation of the permutation distribution of the FR statistic is independent of the sample, and it is $\frac{2mn}{N} + 1$. Now, $\frac{2mn}{N} + 1 \geq 1+\alpha + (n-\epsilon)(1-\alpha) \Rightarrow \frac{n(n-m)-\delta}{(n-1)(m+n)-\delta} \leq \alpha$, but this is not possible. Hence $c_d < n+1-\epsilon$ almost everywhere.

(b) The proof is quite similar to that of part (a). From Lemma 3, we know that under the given condition, $T_{m,n,k}^{NN} \xrightarrow{P} m/N$ as $d \rightarrow \infty$. Now, given a sample, let C_d be the cut-off for a level α permutation test based on $T_{m,n,k}^{NN}$ in dimension d , and we reject H_0 if $T_{m,n,k}^{NN}$ exceeds C_d (here also, for the ease of notation, we do not show the dependence of C_d on the given sample). So, it is enough to show that under the conditions of Theorem 3.2, C_d is larger than $m/N + \epsilon$ almost everywhere for some $\epsilon > 0$ and all $d \geq 1$. Now, note that if $(n-1)/m > (1+\alpha)/(1-\alpha)$, we can always choose a $\delta > 0$ such that $(n-1-\delta)/m > (1+\alpha)/(1-\alpha)$. Let us choose $\epsilon = \delta/nN(N-1)$.

Now, given a sample, let \mathbb{P} and \mathbb{E} denote the permutation probability distribution of the NN statistic and expectation under it. Then note that $\mathbb{P}(T_{m,n,k}^{NN} > C_d) \leq \alpha$, and $T_{m,n,k}^{NN}$ always lies between 0 and 1. Therefore, if C_d does not exceed $m/N + \epsilon$, for that given sample, we have $\mathbb{E}(T_{m,n,k}^{NN}) \leq (1-\alpha)(m/N + \epsilon) + \alpha = (m+n\alpha)/N + \epsilon(1-\alpha)$. But following Schilling [27], one can check that the expectation of the permutation distribution of the NN statistic is independent of sample, and we have $\mathbb{E}(T_{m,n,k}^{NN}) = \frac{m(m-1)+n(n-1)}{N(N-1)}$. Now, $\frac{m(m-1)+n(n-1)}{N(N-1)} \leq (m+n\alpha)/N + \epsilon(1-\alpha)$ implies $\frac{n-1-\delta}{m} \leq \frac{1+\alpha}{1-\alpha}$, which is not possible. \square

Proof of Theorem 4.1. Note that $NT_{m,n}$ can be expressed as $NT_{m,n} = \frac{1}{2} \left[\left\{ \sqrt{N}(\hat{\mu}_{FF} - \hat{\mu}_{GG}) \right\}^2 + \left\{ \sqrt{N}T_{m,n}^{BF} \right\}^2 \right]$, where $T_{m,n}^{BF}$

is the BF statistic. From Baringhaus and Franz [3], under H_0 , we have $NT_{m,n}^{BF} = O_p(1)$, and hence $\sqrt{N}T_{m,n}^{BF} \xrightarrow{P} 0$ as $N = (m+n) \rightarrow \infty$. Again, under H_0 , we have $\mu_{FF} = \mu_{GG}$, and hence $\sqrt{N}(\hat{\mu}_{FF} - \hat{\mu}_{GG}) = \sqrt{N}\{(\hat{\mu}_{FF} - \mu_{FF}) - (\hat{\mu}_{GG} - \mu_{GG})\}$. Now, $\hat{\mu}_{FF} - \mu_{FF} = \binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} (\|\mathbf{X}_i - \mathbf{X}_j\| - \mu_{FF})$ is a U -statistic with a symmetric kernel function. Therefore, from standard results on the U -statistic (see e.g., Lee [15]), we have $\sqrt{m}(\hat{\mu}_{FF} - \mu_{FF}) \xrightarrow{d} N(0, 4\sigma_0^2)$, where $\sigma_0^2 = \text{Var}(E(\|\mathbf{X}_1 - \mathbf{X}_2\| | \mathbf{X}_1))$. Similarly, $\sqrt{n}(\hat{\mu}_{GG} - \mu_{GG}) \xrightarrow{d} N(0, 4\sigma_0^2)$, and they are independent. So, using the fact that $m/N \rightarrow \lambda$ as $N \rightarrow \infty$, we have $\sqrt{N}(\hat{\mu}_{FF} - \hat{\mu}_{GG}) = \sqrt{N/m} [\sqrt{m}(\hat{\mu}_{FF} - \mu_{FF})] - \sqrt{N/n} [\sqrt{n}(\hat{\mu}_{GG} - \mu_{GG})] \xrightarrow{d} N(0, (\frac{1}{\lambda} + \frac{1}{1-\lambda})4\sigma_0^2)$, and hence $NT_{m,n} \xrightarrow{d} \frac{2\sigma_0^2}{\lambda(1-\lambda)} \chi_1^2$ as m and n both tend to infinity. \square

Proof of Theorem 4.2. Here, $T_{m,n}^* = N\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}_0^2 \xrightarrow{d} \chi_1^2$ as $m, n \rightarrow \infty$, and we reject H_0 at level α if $T_{m,n}^* > \chi_{1,\alpha}^2$. So, the power of the test is given by $P_{H_1}(T_{m,n}^* > \chi_{1,\alpha}^2) = P_{H_1}[\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}_0^2 > \chi_{1,\alpha}^2/N]$.

Now, from the results on probability convergence of the U -statistic, we have $\hat{\mu}_{FF} \xrightarrow{P} \mu_{FF}$, $\hat{\mu}_{GG} \xrightarrow{P} \mu_{GG}$ and $\hat{\mu}_{FG} \xrightarrow{P} \mu_{FG}$ as $m, n \rightarrow \infty$. This implies $\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}_0^2 \xrightarrow{P} \lambda(1-\lambda)\{(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2\}/2\sigma_0^2$ as m and n tend to infinity. Since $(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2 = 0 \Leftrightarrow \mu_{FF} = \mu_{FG} = \mu_{GG}$, from Lemma 1, we have $(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2 = 0$ iff $F = G$. So, under H_1 , as $m, n \rightarrow \infty$, $\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}_0^2$ converges (in probability) to a positive quantity, but $\chi_{1,\alpha}^2/N$ converges to 0. Therefore, the power of the test $P_{H_1}[\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}_0^2 > \chi_{1,\alpha}^2/N]$ converges to 1 as m and n both tend to infinity. \square

References

- [1] D.W.K. Andrews, Laws of large numbers for dependent non-identically distributed random variables, *Econom. Theory* 4 (1988) 458–467.
- [2] Z. Bai, H. Saranadasa, Effect of high dimension: by an example of a two-sample problem, *Statist. Sinica* 6 (1996) 311–329.
- [3] L. Baringhaus, C. Franz, On a new multivariate two-sample test, *J. Multivariate Anal.* 88 (2004) 190–206.
- [4] L. Baringhaus, C. Franz, Rigid motion invariant two-sample tests, *Statist. Sinica* 20 (2010) 1333–1361.
- [5] S.X. Chen, Y.-L. Qin, A two-sample test for high-dimensional data with applications to gene-set testing, *Ann. Statist.* 38 (2010) 808–835.
- [6] K. Choi, J. Marden, An approach to multivariate rank tests in multivariate analysis of variance, *J. Amer. Statist. Assoc.* 92 (1997) 1581–1590.
- [7] J.H. Friedman, L.C. Rafsky, Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, *Ann. Statist.* 7 (1979) 697–717.
- [8] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005) 427–444.
- [9] P. Hall, N. Tajvidi, Permutation tests for equality of distributions in high dimensional settings, *Biometrika* 89 (2002) 359–374.
- [10] N. Henze, A multivariate two-sample test based on the number of nearest neighbor type coincidences, *Ann. Statist.* 16 (1988) 772–783.
- [11] N. Henze, M.D. Penrose, On the multivariate runs test, *Ann. Statist.* 27 (1999) 290–298.
- [12] T.P. Hettmansperger, J. Möttönen, H. Oja, Affine invariant multivariate rank tests for several samples, *Statist. Sinica* 8 (1998) 785–800.
- [13] T.P. Hettmansperger, H. Oja, Affine invariant multivariate multi-sample sign tests, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 56 (1994) 235–249.
- [14] S. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, *Ann. Statist.* 37 (2009) 4104–4130.
- [15] A.J. Lee, *U-Statistics: Theory and Practice*, Marcel Dekker, New York, 1990.
- [16] Z. Liu, R. Modarres, A triangle test for equality of distribution functions in high dimensions, *J. Nonparametr. Stat.* 23 (2011) 605–615.
- [17] R.Y. Liu, K. Singh, A quality index based on data depth and multivariate rank tests, *J. Amer. Statist. Assoc.* 88 (1993) 252–260.
- [18] B. Lu, R. Greevy, X. Xu, C. Beck, Optimal non-bipartite matching and its statistical applications, *Amer. Statist.* 65 (2011) 21–30.
- [19] J.-F. Maa, D.K. Pearl, R. Bartoszyński, Reducing multidimensional two-sample data to one-dimensional inter-point comparisons, *Ann. Statist.* 24 (1996) 1069–1074.
- [20] J. Möttönen, H. Oja, Multivariate spatial sign and rank methods, *J. Nonparametr. Stat.* 5 (1995) 201–213.
- [21] H. Oja, *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*, Springer, New York, 2010.
- [22] H. Oja, R.H. Randles, Multivariate nonparametric tests, *Statist. Sci.* 19 (2004) 598–605.
- [23] M.L. Puri, P.K. Sen, *Nonparametric Methods in Multivariate Analysis*, Wiley, New York, 1971.
- [24] R.H. Randles, D. Peters, Multivariate rank tests for the two-sample location problem, *Comm. Statist. Theory Methods* 19 (1990) 4225–4238.
- [25] P.R. Rosenbaum, An exact distribution-free test comparing two multivariate distributions based on adjacency, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005) 515–530.
- [26] V. Rousson, On distribution-free tests for the multivariate two-sample location-scale model, *J. Multivariate Anal.* 80 (2002) 43–57.
- [27] M.F. Schilling, Multivariate two-sample tests based on nearest neighbors, *J. Amer. Statist. Assoc.* 81 (1986) 799–806.
- [28] M.S. Srivastava, A test for the mean vector with fewer observations than the dimension under non-normality, *J. Multivariate Anal.* 100 (2009) 518–532.
- [29] G. Székely, M. Rizzo, Testing for equal distributions in high dimension, *Interstat*, 2004.
- [30] G. Zech, B. Aslan, A multivariate two-sample test based on the concept of minimum energy, in: *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT 2003)*, SALC, Menlo Park, Stanford, California, USA, September 8–11, 2003, pp. 97–100.