A Nonparametric Two-Sample Test Applicable to High Dimensional Data

Munmun Biswas and Anil K. Ghosh

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203, Barrackpore Trunk Road, Kolkata 700108, India.

E-mail : munmun_r@isical.ac.in, akghosh@isical.ac.in.

Abstract

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 set up. 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 when they are applied to high dimensional data. We also study the 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. Some simulated data sets are also analyzed to study its finite sample performance.

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

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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \overset{i.i.d.}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \overset{i.i.d.}{\sim} G$, 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 set up. 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 the multivariate two-sample location problem, the Hotelling T^2 test is often used. While it is 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 statistics that can be applied to high dimensional data (see e.g., Bai and Saranadasa, 1996; Srivastava, 2009; Chen and Qin, 2010), but these tests are 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 (1971), Randles and Peters (1990), Hettmansperger and Oja (1994), Möttönen and Oja (1995), Choi and Marden (1997) and Hettmansperger et. al. (1998). Liu and Singh (1993) and Rousson (2002) constructed nonparametric tests for multivariate two-sample location and scale problem. Some good reviews of most of these tests can be found in Oja and Randles (2004) and Oja (2010). 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.

Multivariate nonparametric tests for general two-sample problem have also been proposed in the literature. Friedman and Rafsky (1979) used the idea of minimal spanning tree (MST) to generalize the univariate run test in multi-dimension. Schilling (1986) and Henze (1988) proposed two-sample tests based on nearest neighbor type coincidences. Other nonparametric tests for the general two sample problem include Hall and Tajvidi (2002), Zech and Aslan (2003), Baringhaus and Franz (2004, 2010) and Liu and Modarres (2011). 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 (2005) 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 $\mathbf{X}, \mathbf{X}_* \overset{i.i.d.}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} G$, and $\|\cdot\|$ denotes the Euclidean norm (see Maa et. al., 1996). Such inter-point distances can be easily computed in any dimension. In this article, we use these inter-point distances to construct a new test for general two-sample problem.

In the next section, 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. Some simulated data sets are also analyzed to evaluate its finite sample performance. 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 (1979) multivariate generalization of the run test, the test based on nearest neighbor (NN) type coincidences (see, e.g., Schilling, 1986; Henze, 1988) and the test proposed by Baringhaus and Franz (2004). Henceforth, we will refer to them as the FR test, the NN test and the BF test, respectively. We computed powers of these tests for different values of d ranging from 2 to 500, and the results are presented in Figure 2.1.

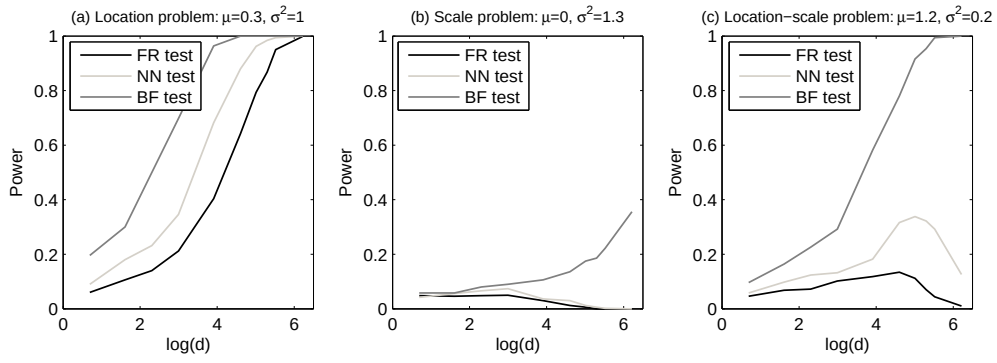

Figure 1: Powers of FR, NN and BF tests for various choices of d

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 location problem (see Figure 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 Figure 1(c)). In the case of scale problem, all of these three methods yielded poor performance (see Figure 1(b)). Note that if $\mathbf{X}, \mathbf{X}_* \overset{i.i.d.}{\sim} F$ and $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} G$, $\|\mathbf{X} - \mathbf{X}_*\|^2/2\sigma_1^2$ and $\|\mathbf{Y} - \mathbf{Y}_*\|^2/2\sigma_2^2$ both follow the chi-square distribution with d degrees of freedom (df). So, using Chebychev's inequality, one can show that $d^{-1}\|\mathbf{X} - \mathbf{X}_*\|^2 \xrightarrow{P} 2\sigma_1^2$ and $d^{-1}\|\mathbf{Y} - \mathbf{Y}_*\|^2 \xrightarrow{P} 2\sigma_2^2$ as $d \to \infty$. Again, $\|\mathbf{X} - \mathbf{Y}\|^2/(\sigma_1^2 + \sigma_2^2)$ follows a non-central chi-square distribution with d df and non-centrality parameter $d\mu^2/(\sigma_1^2+\sigma_2^2)$. In that, we have $d^{-1}\|\mathbf{X}-\mathbf{Y}\|^2 \xrightarrow{P} (\sigma_1^2+\sigma_2^2+\mu^2)$ as $d \to \infty$. Therefore, in the case of location problem, for large d, the distance between any two observations from the same distribution turned out to be smaller than the distance between any two observations belonging to two different distributions. In that situation, these three tests worked well. But this ordering of distances did not hold in other two problems, and that was the reason for the poor performance of FR and NN tests. In the case of scale problem, the overall average of intra-class distances (average of all $\mathbf{X}$-$\mathbf{X}_*$ and $\mathbf{Y}$-$\mathbf{Y}_*$ distances) was very close to the average inter-classes distance (average of all $\mathbf{X}$-$\mathbf{Y}$ distances), and that led to the poor performance by the BF test as well (see Section 3.2 for a detailed discussion on these issues). These limitations of the existing methods 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 $\mathbf{X}, \mathbf{X}_* \overset{i.i.d}{\sim} F$ and $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d}{\sim} G$. 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. (1996) 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. 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 equivalent null hypothesis $H_0'' : \boldsymbol{\mu}_{D_F} = \boldsymbol{\mu}_{D_G}$ against the alternative $H_1'' : \boldsymbol{\mu}_{D_F} \neq \boldsymbol{\mu}_{D_G}$. Our test statistic is very simple and easy to compute. From the data $\mathbf{x}_1, \ldots, \mathbf{x}_m \overset{i.i.d}{\sim} F$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{i.i.d}{\sim} G$, we calculate

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

and reject the null hypothesis for higher values of the test statistic $T_{m,n} = \|\widehat{\boldsymbol{\mu}}_{D_F} - \widehat{\boldsymbol{\mu}}_{D_G}\|^2$. When the sample size is small, we use the permutation principle to calculate out the cut-off. When it is large, this cut-off is chosen using the large sample distribution of $T_{m,n}$ (see Section 4).

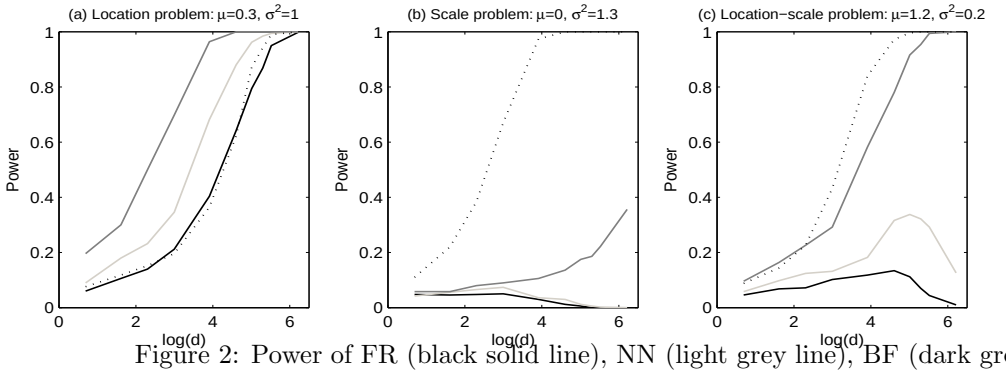

Figure 2: Power of FR (black solid line), NN (light grey line), BF (dark grey line) and proposed tests (black dotted line) for varying choices of d

Figure 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 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 $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ are the mean vector and the dispersion matrix of the distribution F, and $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ are the mean vector and the dispersion matrix of the distribution G. In order to carry out this investigation, following Hall et. al. (2005), we first make the following assumption

(A1) *There exist* $\sigma_1^2, \sigma_2^2 > 0$ *and* ν *such that* (*i*) $trace(\boldsymbol{\Sigma}_1)/d \to \sigma_1^2$ (*ii*) $trace(\boldsymbol{\Sigma}_2)/d \to \sigma_2^2$ *and* (*iii*) $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/d \to \nu^2$ *as* $d \to \infty$.

Note that in the examples with normal distributions discussed in Section 2, these conditions were

satisfied with $\sigma_1^2 = 1, \sigma_2^2 = \sigma^2$ and $\nu^2 = \mu^2$. 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. Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m \overset{i.i.d.}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \overset{i.i.d.}{\sim} G$. Now, if the components of $\mathbf{X}$ and $\mathbf{Y}$ vectors are i.i.d. with bounded second moments, as it was in the examples with normal distributions, we can show that for all $i \neq j$, $d^{-1}\|\mathbf{X}_i - \mathbf{X}_j\|^2 \xrightarrow{P} 2\sigma_1^2$, $d^{-1}\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 \xrightarrow{P} 2\sigma_2^2$ and for all i, j, $d^{-1}\|\mathbf{X}_i - \mathbf{Y}_j\|^2 \xrightarrow{P} \sigma_1^2 + \sigma_2^2 + \nu^2$ as $d \to \infty$, where σ_1^2, σ_2^2 and ν^2 have the same meaning as in (A1). This weak law of large numbers (WLLN) holds for a fairly general class of high dimensional probability models. Hall et. al. (2005) looked at the d-dimensional observations $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ and $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)})$ as infinite time series truncated at time d and studied the behavior of the inter-point distances as d increases. Here, we assume a similar set of assumptions, which 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.,* $sup_{1 \leq q < q' \leq \infty,\ |q-q'|>r} |corr\{(U^{(q)} - V^{(q)})^2,\ (U^{(q')} - V^{(q')})^2\}| \leq \rho(r)$ *where* $\rho(r) \to 0$ *as* $r \to \infty$.

Jung and Marron (2009) assumed similar conditions for the large dimensional consistency of estimated principal component directions. Andrews (1988) also assumed similar conditions to derive WLLN for mixingales. 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 $(U^{(q)}, V^{(q)})$ $\big[(U^{(q)}, V^{(q)}) = (X^{(q)}, X_*^{(q)}), (X^{(q)}, Y^{(q)})$ or $(Y^{(q)}, Y_*^{(q)})\big]$, 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 $\sigma_1^2 + \sigma_2^2 + \nu^2$. 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) $d^{-1/2}\|\mathbf{X}_i - \mathbf{Y}_j\| \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma^2 + \nu^2}$ 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 $d \to \infty$) 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 S_1 of side length $\sigma_1\sqrt{2}$. The other n vertices are the limits of n data points form G, and they form another regular simplex S_2 of edge length $\sigma_2\sqrt{2}$. The rest of the edges of the polyhedron are the edges connecting the vertices of S_1 to those of S_2, and they are of length $\sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$.

Under H_0, when we have $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$, the whole polyhedron turns out to a regular simplex of N points.

From the above discussion it is quite clear that for fixed m and n, under H_0, $T_{m,n}/d \xrightarrow{P} 0$ as $d \to \infty$. On the other hand, it converges to a positive value when $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. The following theorem shows that in such cases, the power of the proposed test converges to 1 as d tends to infinity.

Theorem 3.1: *Suppose that F and G satisfy (A1)-(A3) and also assume that either $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. Then, the power of the proposed test converges to 1 as d tends to infinity.*

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

The FR test statistic is given by $T_{m,n}^{FR} = 1 + \sum_{i=1}^{N-1} U_i$, where $N = m+n$, and U_i $(i = 1, 2, \ldots, N-1)$ 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 $T_{m,n}^{FR}$ 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 $T_{m,n,k}^{NN}$.

From results (a)-(c) in Section 3.1, it is clear that in the case of location problem, when we have $\sigma_1^2 = \sigma_2^2$ and $\nu^2 > 0$, 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.

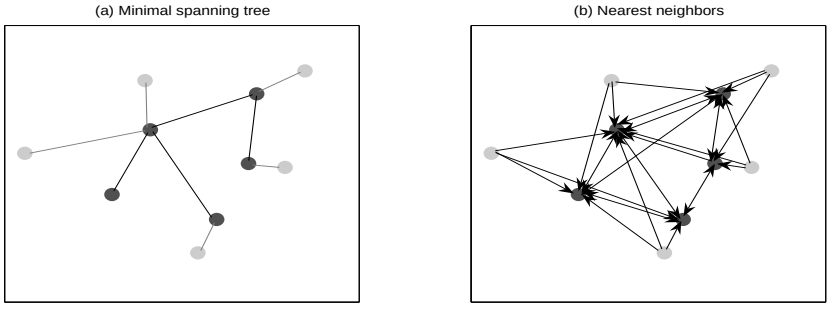

Figure 3: Minimal spanning tree and nearest neighbors of 10 ($m = n = 5$) observations.

However, this ordering of intra-class and inter-class distances does not hold in the case of high dimension small sample size data if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$. 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 Figure 2.2 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, $T^{FR}_{m,n}$ and $T^{NN}_{m,n,k}$ take the values, which are either equal or close to their expected values under H_0, and well below (above in the case of $T^{FR}_{m,n}$) 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 $\nu^2 < \sigma_2^2 - \sigma_1^2$ (interchange F and G, if required, and also interchange m and n accordingly).*
(a) If $n(n-m)/(n-1)(m+n) > \alpha$, the power of a level α test based on $T^{FR}_{m,n}$ converges to zero as $d \to \infty$.
(b) If $(n-1)/m > (1+\alpha)/(1-\alpha)$ and $k < \min\{m,n\}$, the power of a level α test based on $T^{NN}_{m,n,k}$ converges to zero as $d \to \infty$.

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

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 $\mathbf{X}, \mathbf{X}_* \overset{i.i.d.}{\sim} F$, $\mathbf{Y}, \mathbf{Y}_* \overset{i.i.d.}{\sim} G$, and equality holds iff $F = G$ (see Baringhaus and Franz, 2004). 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 $T^{BF}_{m,n}$ or a scaled version of it like $T^{BF}_{m,n}/\sqrt{d}$. Now, it is easy to see that when m and n are fixed and $d \to \infty$, under the assumptions (A1)-(A3), $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$, $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$, and hence $T^{BF}_{m,n}/\sqrt{d} \xrightarrow{P} 2\sqrt{\sigma_1^2+\sigma_2^2+\nu^2} - \sigma_1\sqrt{2} - \sigma_2\sqrt{2}$ $(= \gamma, \text{say})$, which is positive unless $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$. So, a consistency result similar to Theorem 3.1 can be proved for the BF test as well. But, in Section 2, we have seen 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 $\nu^2 < |\sigma_1^2 - \sigma_2^2|$. Now, $\nu^2 < |\sigma_1^2 - \sigma_2^2|$ implies that $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$ 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 sign. As a result, when they are added up, $T^{BF}_{m,n} = (\hat{\mu}_{FG}-\hat{\mu}_{FF}) + (\hat{\mu}_{FG}-\hat{\mu}_{GG})$ 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 cancelations 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 better performance in these two examples. One should also notice that $T_{m,n}$ can also be expressed as $T_{m,n} = \frac{1}{2}\left[(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG})^2 \ + \ (\hat{\mu}_{FF} - \hat{\mu}_{GG})^2\right]$, 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 location, 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 the case of pure location problem, 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.

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 is carried out 200 times as before. We estimated the power of the 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 the two other tests, one proposed by Hall and Tajvidi (2002) and the other proposed by Rosenbaum (2005). 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., 2011) and then used its null distribution to perform the test. Since the sizes of 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 $X^{(1)}, \ldots, X^{(d)}$ (and $Y^{(1)}, \ldots, Y^{(d)}$) 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 $(0, 0, \ldots, 0)^{'}$ and the dispersion matrix $\mathbf{\Sigma}_1 = \mathbf{S}$, those for G are taken to be $(\mu, \mu, \ldots, \mu)^{'}$

and $\mathbf{\Sigma}_2 = \sigma^2\mathbf{S}$, respectively, where $\mathbf{S} = ((s_{ij}))$ is of the form $s_{ij} = (0.5)^{|i-j|}$ for $i, j = 1, 2, \ldots, d$. Here also, we consider three different choices of μ and σ^2 $\left[(\mu, \sigma^2)=(0.25, 1), (0, 1.25) \text{ and } (0.1, 1.1)\right]$ to have three different types of problems. In cases of scale problem and location-scale problem, our proposed test yielded the highest power among all two-sample tests considered in this article. Only in the case of 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.

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

	Normal			Laplace			Normal vs.	AR(1)	AR(2)
	location	scale	loc-scale	location	scale	loc-scale	Laplace	process	process
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

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 location-scale problem, the proposed test performed better than the HT test, while in other two cases, they had nearly the same power.

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 AR(1) model and other with 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 $t = 1, 2, \ldots, 500$, where $X^{(0)}, U_1, U_2, \ldots, U_{500} \overset{i.i.d.}{\sim} N(0, 1)$. Observations in G were generated using another AR(1) model $Y^{(t)} = 0.25+0.4Y^{(t-1)}+V_t$, where $Y^{(0)}$,

$V_1, V_2, \ldots, V_{500} \overset{i.i.d.}{\sim} N(0,1)$. Note that in this example, F and G have difference both in locations and in scales but in the second example, F and G differ only in 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 $t = 1, 2, \ldots, 500$, and those in G were generated using the model $Y^{(t)} = 0.35Y^{(t-1)} + 0.25Y^{(t-2)} + V_t$ for $t = 1, 2, \ldots, 500$, where $X^{(0)}, X^{(-1)}, Y^{(0)}, Y^{(-1)}, U_1, U_2, \ldots, U_{500}, V_1, V_2, \ldots, V_{500}$ 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 is taken from the UCR Time Series Classification/Clustering Page (http://www.cs.
ucr.edu/∼eamonn/time_series _data/), and the other two are 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 power functions. Also, note that if we use the whole data set for testing, any test will either reject H_0 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 power. 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.

Figure 4: 'Normal' (on the left) and 'cyclic' (on the right) classes in control chart data

Next, we consider the control chart data. It is a synthetically generated time series data set, which contains 60-dimensional observations from 6 different 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 cyclic class contain some cyclic pattern (see Figure 4). There are 100 observations from each class, but for our analysis, 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.

Table 2: Powers of two-sample tests with 5% nominal level in real data sets.

Data sets	ECG			Control chart	Arcene		
(m,n)	(20,10)	(10,10)	(10,20)	(5,5)	(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

Arcene data set was obtained by merging three mass-spectometric 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, these 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 choices of m and n (see Table 2). In all these cases, FR and NN tests had the best performance. Our proposed test also had reasonably high power, and it outperformed BF and HT tests in 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* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ *and* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *from* F*, which has finite second moments. Also assume that as* $N = (m+n) \to \infty$*,* $m/N \to \lambda$ *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* $\sigma_0^2 = Var\{E(\|\mathbf{X}_1 - \mathbf{X}_2\||\mathbf{X}_1)\}$*, 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 $\sigma_0^2 = Cov(\|X_1 - X_2\|, \|X_1 - X_3\|) = E(\|X_1 - X_2\|\|X_1 - X_3\|) - E^2(\|X_1 - X_2\|)$. Now define

$$S_1 = \left[\binom{m}{3}^{-1} \sum_{1\le i<j<k\le m} \|X_i - X_j\|\|X_i - X_k\|\right] - \left[\binom{m}{2}^{-1} \sum_{1\le i<j\le m} \|X_i - X_j\|\right]^2$$

$$\text{and } S_2 = \left[\binom{n}{3}^{-1} \sum_{1\le i<j<k\le n} \|Y_i - Y_j\|\|Y_i - Y_k\|\right] - \left[\binom{n}{2}^{-1} \sum_{1\le i<j\le n} \|Y_i - Y_j\|\right]^2.$$

From the results on the probability convergence of U-Statistics (see e.g., Lee, 1990), 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 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 $T^*_{m,n}$ form the data, and for a test of nominal level α, we reject H_0 if $T^*_{m,n}$ exceeds $\chi^2_{1,\alpha}$, where $P(\chi_1^2 > \chi^2_{1,\alpha}) = \alpha$. 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 $m, n \to \infty$, $m/(m+n) \to \lambda$ 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.*

We carried out simulation studies to evaluate the performance of our proposed test based on the large sample distribution of $T^*_{m,n}$. Here also, we used 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, 1999; Schilling, 1986; Rosenbaum, 2005). 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 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. Throughout this section, for our simulation studies, we used $d = 5$ and $m = n = 100$. Each experiment was carried out 200 times, and the results are reported in Table 3.

Table 3 : Observed powers of the large sample tests with 5% nominal level.

	Normal		Laplace		Normal vs
	location	scale	location	scale	Laplace
BF	1.000	0.300	1.000	0.510	0.605
NN	0.955	0.190	0.995	0.140	0.380
FR	0.860	0.175	0.990	0.130	0.340
HT	0.995	0.670	0.985	0.365	0.685
Rosenbaum	0.630	0.135	0.780	0.085	0.265
Proposed	0.820	0.970	0.635	0.810	0.990

Here also we considered some examples with multivariate normal distributions. The location and the scatter parameters of F were taken as $(0, 0, \ldots, 0)'$ and $\mathbf{I}_d$, while those of G were of the form $(\mu, \mu, \ldots, \mu)'$ and $\sigma^2\mathbf{I_d}$. In the case of location problem ($\mu = 0.5, \sigma^2 = 1$), the BF test had the best performance, but in the case of scale problem ($\mu = 0, \sigma^2 = 1.5$), once again, our proposed test outperformed all of its competitors. We observed the same phenomenon when we carried out our experiment with Laplace distributions. Here also we assumed that the component variables of the Laplace distributions were i.i.d., and they had the same means and variances as in the

previous examples with of normal distributions. Clearly, these results are consistent with what we observed in Section 3. Like the high-dimensional case, we also considered another example, where the component variables in F are i.i.d. standard normal variates, and those in G are i.i.d. standard Laplace variates. In this example, the proposed method had an excellent performance. While it rejected the null hypothesis in 99% of the cases, the next best test had a power of 0.685 only.

5 Concluding remarks

In this article, we have proposed and investigated a two-sample test based on inter-point distances. This test is rotation invariant, 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 set up, 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 $H_0 : F_1 = F_2 = \ldots = F_M$ based on $n_1, n_2, \ldots, n_M$ independent observations from M distributions $F_1, F_2 \ldots, F_M$. 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 (2004) considered a similar generalization of the BF statistic. Another option is to use $\sum_{i<j} \|\hat{\boldsymbol{\mu}}_{D^0_{F_i}} - \hat{\boldsymbol{\mu}}_{D^0_{F_j}}\|^2$ as a test statistic, where $\hat{\boldsymbol{\mu}}_{D^0_{F_j}}$ is an estimate of $\boldsymbol{\mu}_{D^0_{F_j}}$, the mean of an M-dimensional distribution $D^0_{F_j}$. Here $D^0_{F_j}$ $(j = 1, 2, \ldots, M)$ denotes the distribution of $(\|\mathbf{Z}_0 - \mathbf{Z}_1\|, \|\mathbf{Z}_0 - \mathbf{Z}_2\|, \ldots, \|\mathbf{Z}_0 - \mathbf{Z}_M\|)$, where $\mathbf{Z}_i$'s are all independent, $\mathbf{Z}_0 \sim F_j$ and $\mathbf{Z}_i \sim F_i$ for all $i = 1, 2, \ldots, M$. Consistency results similar to Theorem 3.1 and Theorem 4.2 can proved for both of these generalizations. When the sample size is small, we can use the permutation test as before, but for the large sample test, we need to the 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 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.

Appendix: proofs and mathematical details

Lemma 1: Suppose that $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d}{\sim} G$. 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 (2004), 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 $F = G$. □

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

Proof of Lemma 2: Since the sequence of random variables $\{W^{(q)},\ q \geq 1\}$ has uniformly bounded second moments, we have $\sup_q V(W^{(q)}) < C$ for some $C > 0$. So, for any fixed d, we have

$$E\Big[\frac{1}{d}\sum_{q=1}^{d} W^{(q)} - \frac{1}{d}\sum_{q=1}^{d} E[W^{(q)}]\Big]^2 = V\Big[\frac{1}{d}\sum_{q=1}^{d} W^{(q)}\Big] \leq \frac{C}{d} + \frac{C}{d^2}\Big[\sum_{q \neq q'} Corr(W^{(q)}, W^{(q')})\Big].$$

Now, for every $\epsilon > 0$, one can choose an integer R_ϵ such that for every $r > R_\epsilon$, $|\rho(r)| < \epsilon/2C$. If we take $d > 6CR_\epsilon/\epsilon$, we have

$$\begin{aligned} V\Big[\tfrac{1}{d}\textstyle\sum_{q=1}^{d} W^{(q)}\Big] &\leq \tfrac{C}{d} + \tfrac{C}{d^2}\Big[\textstyle\sum_{(q,q'):1 \leq |q-q'| \leq R_\epsilon} Corr(W^{(q)}, W^{(q')})\Big] \\ &\qquad + \tfrac{C}{d^2}\Big[\textstyle\sum_{(q,q'):|q-q'|>R_\epsilon} Corr(W^{(q)}, W^{(q')})\Big] \\ &\leq \tfrac{C}{d} + \tfrac{2CR_\epsilon}{d} + \tfrac{\epsilon}{2} < \epsilon. \end{aligned}$$

This implies $\Big|\frac{1}{d}\sum_{q=1}^{d} W^{(q)} - \frac{1}{d}\sum_{q=1}^{d} E[W^{(q)}]\Big| \xrightarrow{P} 0$ as $d \to \infty$. □

Proof of Theorem 3.1: If F and G satisfy (A1)-(A3), for fixed m, n and $d \to \infty$, we have

$$\begin{aligned}
\|\mathbf{X}_i - \mathbf{X}_j\|/\sqrt{d} &\xrightarrow{P} \sigma_1\sqrt{2} \ \ \forall\ i \neq j, \ \ i,j \in \{1,2,\ldots,m\} \\
\|\mathbf{Y}_i - \mathbf{Y}_j\|/\sqrt{d} &\xrightarrow{P} \sigma_2\sqrt{2} \ \ \forall\ i \neq j, \ \ i,j \in \{1,2,\ldots,n\} \\
\|\mathbf{X}_i - \mathbf{Y}_j\|/\sqrt{d} &\xrightarrow{P} \sqrt{\sigma_1^2+\sigma_2^2+\nu^2} \ \ \forall\ i = 1,2,\ldots,m \text{ and } j = 1,2,\ldots,n.
\end{aligned}$$

So, one can see that $\hat{\mu}_{FF}/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$, $\hat{\mu}_{GG}/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$ and $\hat{\mu}_{FG}/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$. Now under H_0, we have $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$. Therefore, if $T^d_{m,n}$ (instead of $T_{m,n}$, we use $T^d_{m,n}$ for its dependence on d) denotes our test statistic, it is easy to show that $T^d_{m,n}/d = [(\hat{\mu}_{FF}/\sqrt{d} - \hat{\mu}_{FG}/\sqrt{d})^2 + (\hat{\mu}_{FG}/\sqrt{d} - \hat{\mu}_{GG}/\sqrt{d})^2] \xrightarrow{P} 0$ as $d \to \infty$. Let $c^d_{m,n}$ be the cutoff for our test, and we reject H_0 if $T^d_{m,n} > c^d_{m,n}$. Since $P_{H_0}\{T^d_{m,n}/d > c^d_{m,n}/d\} \leq \alpha$ for all $d \geq 1$, and $T^d_{m,n}/d \xrightarrow{P} 0$ under H_0, we have $c^d_{m,n}/d \to 0$ as $d \to \infty$. Now, under the alternative hypothesis, we have either $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. In that case $T^d_{mn}/d \xrightarrow{P} (\sqrt{2}\sigma_1 - \sqrt{\sigma_1^2+\sigma_2^2+\nu^2})^2 + (\sqrt{\sigma_1^2+\sigma_2^2+\nu^2} - \sqrt{2}\sigma_2)^2$, a positive quantity. So, the probability of rejecting H_0, $P_{H_1}(T^d_{mn}/d > c^d_{m,n}/d)$, tends to 1 as d tends to infinity. □

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

Proof of Lemma 3: Note that $0 \leq \nu^2 < \sigma_2^2 - \sigma_1^2$ implies $\sigma_1\sqrt{2} < \sqrt{\sigma_1^2+\sigma_2^2+\nu^2} < \sigma_2\sqrt{2}$. Therefore, if (A1)-(A3) hold, for every $i, i^{'} = 1,2,\ldots,m$ $(i \neq i^{'})$ and $j, j^{'} = 1,2,\ldots,n$ $(j \neq j^{'})$, $P(\|\mathbf{X}_i - \mathbf{X}_{i'}\| \leq \|\mathbf{X}_i - \mathbf{Y}_j\| \leq \|\mathbf{Y}_j - \mathbf{Y}_{j'}\|) \to 1$ as $d \to \infty$ (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 $T^{FR}_{m,n} - 1$ 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 Figure 2). So, as d tends to infinity, $T^{FR}_{m,n} \xrightarrow{P} n+1$. □

(b) Recall that the NN test statistic $T^{NN}_{m,n,k}$ 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. 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 $\sum_{i=1}^{m}\sum_{j=1}^{k} I_j(\mathbf{Y}_i) \xrightarrow{P} 0$ as $d \to \infty$. Consequently, $T^{NN}_{m,n,k} \xrightarrow{P} m/N$ as $d \to \infty$. □

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 $d \to \infty$. Let c_d be the cut-off for a level α 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. So, it is enough to show that as $\limsup_{d\to\infty} c_d < n+1$. First note that if $\frac{n(n-m)}{(n-1)(m+n)} > \alpha$, we can choose a $\delta > 0$ such that $\frac{n(n-m)-\delta}{(n-1)(m+n)-\delta}$ is also bigger than α. Now we will show that under the conditions of Theorem 3.2, for any given d, c_d is smaller than $n+1-\epsilon$, where $\epsilon = \delta/N$.

Note that $P_{H_0}(T_{m,n}^{FR} < c_d) \leq \alpha$ (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). Therefore, if $c_d \geq n+1-\epsilon$, we have $E_{H_0}(T_{m,n}^{FR}) \geq 2\alpha + (1-\alpha)(n+1-\epsilon) = 1+\alpha+(n-\epsilon)(1-\alpha)$. Now, from Friedman and Rafsky (1979), we have $E_{H_0}(T_{m,n}^{FR}) = \frac{2mn}{N}+1$. 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 the case.

So, we have $c_d < n+1-\epsilon$ for all d, and hence $\limsup_{d\to\infty} c_d < n+1$. Therefore, the power of the FR test converges to 0 as d tends to infinity. □

(b) The proof of 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 $d \to \infty$. Let C_d be the cut-off for a level α test based on $T_{m,n,k}^{NN}$ in dimension d, and here we reject H_0 if $T_{m,n,k}^{NN}$ exceeds C_d. So, it is enough to show that under the conditions of Theorem 3.2, C_d is larger than $m/N+\epsilon$ 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 $\epsilon = \delta/nN(N-1)$.

Note that $P_{H_0}(T_{m,n,k}^{NN} > C_d) \leq \alpha$, and $T_{m,n,k}^{NN}$ always lies between 0 and 1. Therefore, if C_d does not exceed $m/N+\epsilon$, we have $E_{H_0}(T_{m,n,k}^{NN}) \leq (1-\alpha)(m/N+\epsilon)+\alpha = (m+n\alpha)/N+\epsilon(1-\alpha)$. But from Schilling (1986), we know $E_{H_0}(T_{m,n,k}^{NN}) = \frac{m(m-1)+n(n-1)}{N(N-1)}$. Now, one can show that $\frac{m(m-1)+n(n-1)}{N(N-1)} \leq (m+n\alpha)/N+\epsilon(1-\alpha)$ implies $\frac{n-1-\delta}{m} \leq \frac{1+\alpha}{1-\alpha}$, which is not the case. □

Proof of Theorem 4.1: Note that

$$\begin{bmatrix} \frac{1}{\binom{m}{2}} \sum_{1\leq i_1<i_2\leq m} \|\mathbf{X}_{i_1}-\mathbf{X}_{i_2}\| - \frac{1}{mn}\sum_{i_1=1}^{m}\sum_{j_1=1}^{n}\|\mathbf{X}_{i_1}-\mathbf{Y}_{j_1}\| \\ \frac{1}{mn}\sum_{i_2=1}^{m}\sum_{j_2=1}^{n}\|\mathbf{X}_{i_2}-\mathbf{Y}_{j_2}\| - \frac{1}{\binom{n}{2}}\sum_{1\leq j_1<j_2\leq n}\|\mathbf{Y}_{j_1}-\mathbf{Y}_{j_2}\| \end{bmatrix}$$

can be expressed as a generalized bivariate U-statistic

$$U_{m,n} = \begin{pmatrix} U_{m,n}^{(1)} \\ U_{m,n}^{(2)} \end{pmatrix} = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{1\leq i_1<i_2\leq m} \sum_{1\leq j_1<j_2\leq n} \psi(\mathbf{X}_{i_1},\mathbf{X}_{i_2};\mathbf{Y}_{j_1},\mathbf{Y}_{j_2}),$$

where $\psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}; \mathbf{Y}j_1, \mathbf{Y}_{j_2}) = \begin{bmatrix} \|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\| - \|\|\mathbf{X}_{i_1} - \mathbf{Y}_{j_1}\| \\ \|\mathbf{X}_{i_2} - \mathbf{Y}_{j_2}\| - \|\mathbf{Y}_{j_1} - \mathbf{Y}_{j_2}\| \end{bmatrix}$. So, for any two-dimensional vector $\mathbf{t} = (t_1, t_2)'$, the linear projection $U^{\mathbf{t}}_{m,n} = \mathbf{t}'U_{m,n}$ can be viewed as a generalized U statistic with the univariate kernel function

$$\begin{aligned}\psi^{\mathbf{t}}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}; \mathbf{Y}_{j_1}, \mathbf{Y}_{j_2}) &= \mathbf{t}'\psi(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}; \mathbf{Y}_{j_1}, \mathbf{Y}_{j_2}) \\ &= \{t_1\|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\| - t_1\|\mathbf{X}_{i_1} - \mathbf{Y}_{j_1}\| + t_2\|\mathbf{X}_{i_2} - \mathbf{Y}_{j_2}\| - t_2\|\mathbf{Y}_{j_1} - \mathbf{Y}_{j_2}\|\}.\end{aligned}$$

Instead of $\psi^{\mathbf{t}}$, one can also use its symmetrized version $\psi^{\mathbf{t}}_0$ and express $U^{\mathbf{t}}_{m,n}$ as

$$U^{\mathbf{t}}_{m,n} = \frac{1}{\binom{m}{2}\binom{n}{2}} \sum_{1\le i_1<i_2\le m} \sum_{1\le j_1<j_2\le n} \psi^{\mathbf{t}}_0(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}; \mathbf{Y}_{j_1}, \mathbf{Y}_{j_2}), \quad \text{where}$$

$$\begin{aligned}\psi^{\mathbf{t}}_0(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2) &= (t_1\|\mathbf{x}_1 - \mathbf{x}_2\| - t_2\|\mathbf{y}_1 - \mathbf{y}_2\|) \\ &\quad + \tfrac{t_2 - t_1}{4}(\|\mathbf{x}_1 - \mathbf{y}_1\| + \|\mathbf{x}_1 - \mathbf{y}_2\| + \|\mathbf{x}_2 - \mathbf{y}_1\| + \|\mathbf{x}_2 - \mathbf{y}_2\|)\}.\end{aligned}$$

Define $\theta_1 = E(\|\mathbf{X}_1 - \mathbf{X}_2\|)$, $\theta_2 = E(\|\mathbf{Y}_1 - \mathbf{Y}_2\|)$, $\theta_3 = E(\|\mathbf{X}_1 - \mathbf{Y}_1\|)$ and $\boldsymbol{\theta} = (\theta_1 - \theta_3,\ \theta_3 - \theta_2)'$. Clearly, $E(U^{\mathbf{t}}_{m,n}) = \mathbf{t}'\boldsymbol{\theta}$. Let us express $U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta}$ as

$$\begin{aligned}U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta} &= \sum_{i=1}^{m} E(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta}|X_i) + \sum_{j=1}^{n} E(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta}|Y_j) + R^{\mathbf{t}}_{m,n} \\ &= \frac{2}{m}\sum_{i=1}^{m} \Phi^{\mathbf{t}}_{1,0}(X_i) + \frac{2}{n}\sum_{j=1}^{n} \Phi^{\mathbf{t}}_{0,1}(\mathbf{Y}_j) + R^{\mathbf{t}}_{m,n},\end{aligned}$$

where $\Phi^{\mathbf{t}}_{1,0}(\mathbf{X}_i) = E\psi^{\mathbf{t}}_0(\mathbf{X}_i, \mathbf{X}_{i'}; \mathbf{Y}_j, \mathbf{Y}_{j'} \mid \mathbf{X}_i) - \mathbf{t}'\boldsymbol{\theta}$, $\Phi^{\mathbf{t}}_{0,1}(\mathbf{Y}_j) = E\psi^{\mathbf{t}}_0(\mathbf{X}_i, \mathbf{X}_{i'};\ \mathbf{Y}_j, \mathbf{Y}_{j'} \mid \mathbf{Y}_j) - \mathbf{t}'\boldsymbol{\theta}$ and $R^{\mathbf{t}}_{m,n} = (U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta}) - \frac{2}{m}\sum_{i=1}^{m} \Phi^{\mathbf{t}}_{1,0}(X_i) - \frac{2}{n}\sum_{j=1}^{n} \Phi^{\mathbf{t}}_{0,1}(\mathbf{Y}_j)$. First we will show that $\sqrt{N}R^{\mathbf{t}}_{m,n} \xrightarrow{P} 0$ as $m, n \to \infty$. Let us define $L^{\mathbf{t}}_{m,n} = \frac{2}{m}\sum_{i=1}^{m} \Phi^{\mathbf{t}}_{1,0}(\mathbf{X}_i) + \frac{2}{n}\sum_{j=1}^{n} \Phi^{\mathbf{t}}_{0,1}(\mathbf{Y}_j)$. Note that

$$\begin{aligned}E(R^{\mathbf{t}}_{m,n})^2 &= E(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta} - L^{\mathbf{t}}_{m,n})^2 \\ &= E(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta})^2 + E(L^{\mathbf{t}}_{m,n})^2 - 2E((U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta})L^{\mathbf{t}}_{m,n}).\end{aligned}$$

Now, one can show that $E((U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta})L^{\mathbf{t}}_{m,n}) = E(L^{\mathbf{t}}_{m,n})^2$, and hence we have

$$E(R^{\mathbf{t}}_{m,n})^2 = E(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta})^2 - E(L^{\mathbf{t}}_{m,n})^2 = Var(U^{\mathbf{t}}_{m,n}) - Var(L^{\mathbf{t}}_{m,n})$$

Let $\Phi^{\mathbf{t}}_{a,b}$ be the conditional expectation of $\psi^{\mathbf{t}}_0(\mathbf{X}_i, \mathbf{X}_{i'}; \mathbf{Y}_j, \mathbf{Y}_{j'}) - \mathbf{t}'\boldsymbol{\theta}$ when a of the $\mathbf{X}$-variables of b of the $\mathbf{Y}$-variables are given $(a, b = 0, 1, 2)$, and $\sigma^{\mathbf{t}^2}_{ab} = Var(\Phi^{\mathbf{t}}_{a,b})$. Now, from the theory of generalized U-statistic (see e.g., Lee, 1990), we have

$$\begin{aligned}Var(U^{\mathbf{t}}_{m,n}) &= \textstyle\sum_{a=0}^{2}\sum_{b=0}^{2} \frac{\binom{2}{a}\binom{2}{b}\binom{m-2}{2-a}\binom{n-2}{2-b}}{\binom{m}{2}\binom{n}{2}}\sigma^{\mathbf{t}^2}_{ab} \\ &= \tfrac{4}{m}\sigma^{\mathbf{t}^2}_{10}(1 - \tfrac{1}{m-1})(1 - \tfrac{1}{n-1})(1 - \tfrac{2}{n-1}) \\ &\qquad + \tfrac{4}{n}\sigma^{\mathbf{t}^2}_{01}(1 - \tfrac{1}{n-1})(1 - \tfrac{1}{m-1})(1 - \tfrac{2}{m-1}) + O(\tfrac{1}{mn}).\end{aligned}$$

On the other hand, it is easy to check that $Var(L^{\mathbf{t}}_{m,n}) = \frac{4}{m}\sigma^{\mathbf{t}^2}_{10} + \frac{4}{n}\sigma^{\mathbf{t}^2}_{01}$. So, under the condition $m/N \to \lambda$ as $N \to \infty$, we have $E(\sqrt{N}R^{\mathbf{t}}_{m,n})^2 = N(Var(U^{\mathbf{t}}_{m,n}) - Var(L^{\mathbf{t}}_{m,n})) \to 0$, which implies $\sqrt{N}R^{\mathbf{t}}_{m,n} \xrightarrow{P} 0$ as $N \to \infty$. Hence, $\sqrt{N}(U^{\mathbf{t}}_{m,n} - \mathbf{t}'\boldsymbol{\theta})$ and $\sqrt{N}L^{\mathbf{t}}_{m,n}$ have the same asymptotic distribution. Now, using the Central Limit Theorem and the fact that $m/N \to \lambda$ as $N \to \infty$, we get

$$\sqrt{N}L^{\mathbf{t}}_{m,n} = 2\Big[\sqrt{\frac{N}{m}}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\Phi^{\mathbf{t}}_{1,0}(X_i) + \sqrt{\frac{N}{n}}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Phi^{\mathbf{t}}_{0,1}(Y_j)\Big] \xrightarrow{d} N\Big[0, 4\Big(\frac{1}{\lambda}\sigma^{\mathbf{t}^2}_{10} + \frac{1}{1-\lambda}\sigma^{\mathbf{t}^2}_{01}\Big)\Big].$$

It can be easily verified that $\Phi^{\mathbf{t}}_{1,0}(\mathbf{x}) = t_1\{E\|\mathbf{x}-\mathbf{X}\| - \theta_1\} + \frac{t_2-t_1}{2}\{E\|\mathbf{x}-\mathbf{Y}\| - \theta_3\}$ and $\Phi^{\mathbf{t}}_{0,1}(\mathbf{y}) = \frac{t_2-t_1}{2}\{E\|\mathbf{y}-\mathbf{X}\|-\theta_3\} - t_2\{E\|\mathbf{y}-\mathbf{Y}\|-\theta_2\}$. Now, under $H_0 : F = G$, we have $\theta_1 = \theta_2 = \theta_3$ (i.e. $\mathbf{t}'\boldsymbol{\theta} = 0$ for all $\mathbf{t}$) and $Var(E(\|X_1 - X_2\||X_1)) = Var(E(\|X_1 - Y_1\||X_1)) = Var(E(\|Y_1 - Y_2\||Y_1)) = \sigma^2_0$. We also have $E\|\mathbf{z}-\mathbf{X}\| = E\|\mathbf{z}-\mathbf{Y}\|$ for all $\mathbf{z} \in R^d$. Therefore, both $\sigma^{\mathbf{t}^2}_{10}$ and $\sigma^{\mathbf{t}^2}_{01}$ turn out to be $(\frac{t_1+t_2}{2})^2\sigma^2_0$. Now, taking $\mathbf{t} = (1,1)$, we get $\sigma^{\mathbf{t}^2}_{10} = \sigma^{\mathbf{t}^2}_{01} = \sigma^2_0$, and for $\mathbf{t} = (1,-1)$, we get $\sigma^{\mathbf{t}^2}_{10} = \sigma^{\mathbf{t}^2}_{01} = 0$. Therefore, we have

$$\sqrt{N}(U^{(1)}_{m,n} + U^{(2)}_{m,n}) \xrightarrow{d} N[0, 4\sigma^2_0/\lambda(1-\lambda)] \text{ and } \sqrt{N}(U^{(1)}_{m,n} - U^{(2)}_{m,n}) \xrightarrow{P} 0.$$

Now, $NT_{m,n} = N[(U^{(1)}_{m,n})^2 + (U^{(2)}_{m,n})^2] = \frac{1}{2}[\{\sqrt{N}(U^{(1)}_{m,n} + U^{(2)}_{m,n})\}^2 + \{\sqrt{N}(U^{(1)}_{m,n} - U^{(2)}_{m,n})\}^2]$. Therefore, $NT_{m,n} \xrightarrow{d} \frac{2\sigma^2_0}{\lambda(1-\lambda)}\chi^2_1$ as m and n both tend to infinity. □

Proof of Theorem 4.2 : Here, $T^*_{m,n} = N\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}^2_0 \xrightarrow{d} \chi^2_1$ as $m, n \to \infty$, and we reject H_0 at level α if $T^*_{m,n} > \chi^2_{1,\alpha}$. So, the power of the test is given by $P_{H_1}(T^*_{m,n} > \chi^2_{1,\alpha}) = P_{H_1}\Big[\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}^2_0 > \chi^2_{1,\alpha}/N\Big]$.

Now, from the results on probability convergence of 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 $m, n \to \infty$. This implies $\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}^2_0 \xrightarrow{P} \lambda(1-\lambda)\{(\mu_{FF} - \mu_{FG})^2 + (\mu_{FG} - \mu_{GG})^2\}/2\sigma^2_0$ 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 $m, n \to \infty$, $\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}^2_0$ converges (in probability) to a positive quantity, but $\chi^2_{1,\alpha}/N$ converges to 0. Therefore, the power of the test $P_{H_1}\Big[\hat{\lambda}(1-\hat{\lambda})T_{m,n}/2\hat{\sigma}^2_0 > \chi^2_{1,\alpha}/N\Big]$ converges to 1 as m and n both tend to infinity. □

References

[1] Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, **4**, 458–467.

[2] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two-sample problem. *Statist. Sinica*, **6**, 311-329.

[3] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190-206.

[4] Baringhaus, L. and Franz, C. (2010). Rigid motion invariant two-sample tests. *Statist. Sinica*, **20**, 1333-1361.

[5] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808-835.

[6] Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, **92**, 1581-1590.

[7] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.*, **7**, 697-717.

[8] Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Royal Statist. Soc. Ser. B*, **67**, 427-444.

[9] Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high dimensional settings. *Biometrika*, **89**, 359-374.

[10] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.*, **16**, 772-783.

[11] Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. Ann. Statist., **27**, 290-298.

[12] Hettmansperger, T. P., Möttönen, J., and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statist. Sinica*, **8**, 785-800.

[13] Hettmansperger, T. P. and Oja, H. (1994). Affine invariant multivariate multi-sample sign tests. *J. Royal Statist. Soc. Ser. B*, **56**, 235-249.

[14] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-4130.

[15] Lee, A. J. (1990). *U-Statistics: Theory and Practice*. Marcel Dekker, New York.

[16] Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, **88**, 252-260.

[17] Liu, Z. and Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *J. Nonparametr. Statist.*, **23**, 605-615.

[18] Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal non-bipartite matching and its statistical applications. *Amer. Statist.*, **65**, 21-30.

[19] Maa, J.-F., Pearl, D. K., and Bartoszyński, R. (1996). Reducing multidimensional two-sample data to one-dimensional inter-point comparisons. *Ann. Statist.*, **24**, 1069-1074.

[20] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametr. Statist.*, **5**, 201-213.

[21] Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer, New York.

[22] Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statist. Sci.*, **19**, 598-605.

[23] Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[24] Randles, R. H. and Peters, D. (1990). Multivariate rank tests for the two-sample location problem. *Comm. Statist. Theory Methods*, **19**, 42254238.

[25] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. Royal Statist. Soc. Ser. B*, **67**, 515-530.

[26] Rousson, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *J. Multivariate Anal.*, **80**, 43-57.

[27] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.*, **81**, 799-806.

[28] Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.*, **100**, 518-532.

[29] Székely, G. and Rizzo, M. (2004). Testing for equal distributions in high dimension. *Interstat.* **5**.

[30] Zech, G. and Aslan, B. (2003). A multivariate two-sample test based on the concept of minimum energy. *PHYSTAT*, 97-100.