Biometrika (2014), **101**, 4, *pp.* 913–926
Printed in Great Britain

doi: 10.1093/biomet/asu045
Advance Access publication 23 October 2014

A distribution-free two-sample run test applicable to high-dimensional data

BY MUNMUN BISWAS

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
munmun.biswas08@gmail.com

MINERVA MUKHOPADHYAY

Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
minervamukherjee@gmail.com

AND ANIL K. GHOSH

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
akghosh@isical.ac.in

SUMMARY

We propose a multivariate generalization of the univariate two-sample run test based on the shortest Hamiltonian path. The proposed test is distribution-free in finite samples. While most existing two-sample tests perform poorly or are even inapplicable to high-dimensional data, our test can be conveniently used in high-dimension, low-sample-size situations. We investigate its power when the sample size remains fixed and the dimension of the data grows to infinity. Simulated and real datasets demonstrate our method's superiority over existing nonparametric two-sample tests.

Some key words: Distribution-free property; High-dimension, low-sample-size data; Shortest Hamiltonian path; Two-sample run test.

1. INTRODUCTION

In a two-sample testing problem, we test the equality of two distributions F and G based on independent observations $x_1, \ldots, x_m$ from F and $y_1, \ldots, y_n$ from G. This is a well-studied problem, especially in the univariate setting, and several nonparametric methods have been proposed for it. The Wilcoxon–Mann–Whitney rank test, the Kolmogorov–Smirnov maximum deviation test and the Wald–Wolfowitz run test (Gibbons & Chakraborti, 2003) are univariate rank-based distribution-free tests. The first is mainly used when the alternative hypothesis suggests a stochastic ordering between F and G, and the others are used for more general alternatives.

Multivariate rank-based tests for the two-sample location problem include those of Puri & Sen (1971), Randles & Peters (1990), Hettmansperger & Oja (1994), Möttönen & Oja (1995), Choi & Marden (1997) and Hettmansperger et al. (1998). However, none of these tests is distribution-free. Liu & Singh (1993) used ranks based on data depth to develop distribution-free tests for location and scale problems, and depth-based distribution-free tests were also considered by Rousson (2002), but these tests cannot be used when the dimension of the data exceeds the sample size.

Multivariate nonparametric methods have been developed for general two-sample problems as well. Many of these can be used in high-dimension, low-sample-size situations, but they are not distribution-free (Friedman & Rafsky, 1979; Schilling, 1986; Henze, 1988; Hall & Tajvidi, 2002; Baringhaus & Franz, 2004; Liu & Modarres, 2011). Even the most natural generalization of the Kolmogorov–Smirnov statistic is not distribution-free in two or higher dimensions (Bickel, 1969). In such cases, one either uses the conditional test based on the permutation principle or the large-sample test based on the asymptotic null distribution of the test statistic. Ferger (2000) proposed a distribution-free two-sample test from the perspective of change point detection, but for its proper implementation one needs to find a suitable weight function and an appropriate asymmetric kernel function. Rosenbaum (2005) proposed a simpler distribution-free test based on the idea of optimal non-bipartite matching; it can also be used for high-dimension, low-sample-size data if the Euclidean metric is used for distance computation.

In this article, we propose a multivariate generalization of the Wald–Wolfowitz run test using the shortest Hamiltonian path. This test is based on pairwise distances between sample observations, and hence the test statistic is invariant under location change, rotation and homogeneous scale transformation. It is distribution-free in finite-sample situations and can be used for high-dimensional data or even functional data taking values in a Banach space. Compared to the run test of Friedman & Rafsky (1979), which is based on the notion of the minimal spanning tree, it enjoys better power in high-dimension, low-sample-size settings.

2. MULTIVARIATE RUN TEST BASED ON THE SHORTEST HAMILTONIAN PATH

2·1. *Shortest Hamiltonian path and the test statistic*

Consider a graph $\mathcal{G}$ on N vertices. A Hamiltonian path $\mathcal{H}$ in $\mathcal{G}$ is defined to be a connected, acyclic subgraph of $\mathcal{G}$ with $N-1$ edges, where no vertex has a degree bigger than two. In other words, $\mathcal{H}$ is a path in $\mathcal{G}$ that visits each vertex of $\mathcal{G}$ exactly once. For any given $\mathcal{G}$, a Hamiltonian path may or may not exist, but if $\mathcal{G}$ is a complete graph on N vertices, there are $N!$ Hamiltonian paths. However, for every path, there is another in the reverse order; so if we consider the two as the same path, then there are $N!/2$ distinct Hamiltonian paths. Now, take $\mathcal{G}$ to be a complete graph on N vertices, where each of the $N(N-1)/2$ edges has a cost associated with it. For instance, the distance between the two vertices of an edge can be used as its cost. For each $\mathcal{H}$ in $\mathcal{G}$, one can compute the sum of the costs corresponding to its $N-1$ edges, and define this to be the cost of $\mathcal{H}$. The Hamiltonian path having the minimum cost is said to be the shortest Hamiltonian path, denoted by $\mathcal{H}^*$. For a graph $\mathcal{G}$, $\mathcal{H}^*$ may not be unique, but if the costs corresponding to different edges come from continuous distributions, $\mathcal{H}^*$ becomes unique with probability one. Figure 1 shows a complete graph on four vertices $\{z_1, z_2, z_3, z_4\}$, along with the costs corresponding to different edges. There are 12 distinct Hamiltonian paths in this graph, and the path $z_2 \to z_1 \to z_3 \to z_4$, or equivalently $z_4 \to z_3 \to z_1 \to z_2$, is the shortest Hamiltonian path.

In a two-sample problem, define $z_i = x_i$ for $i = 1, \ldots, m$ and $z_{m+i} = y_i$ for $i = 1, \ldots, n$. Consider a complete graph on $N = m + n$ vertices $z_1, \ldots, z_N$, where the edge between z_i and z_j $(1 \leqslant i < j \leqslant N)$ has the cost $\|z_i - z_j\|$, the Euclidean distance between z_i and z_j. We find the shortest Hamiltonian path $\mathcal{H}^*$ in this graph and count the number of runs along $\mathcal{H}^*$, which is $T_{m,n} = 1 + \sum_{i=1}^{N-1} U_i$ where each U_i is an indicator variable that equals 1 if and only if the ith edge of $\mathcal{H}^*$ connects two observations from two different distributions. In Fig. 1, if we assume that z_1 and z_2 are from F while z_3 and z_4 are from G, the number of runs turns out to be 2, corresponding to $z_2 \to z_1$ and $z_3 \to z_4$. If F and G are widely separated, one would expect $T_{m,n}$ to be small, while under $H_0 : F = G$ it is expected to be large; therefore, we reject H_0 for small

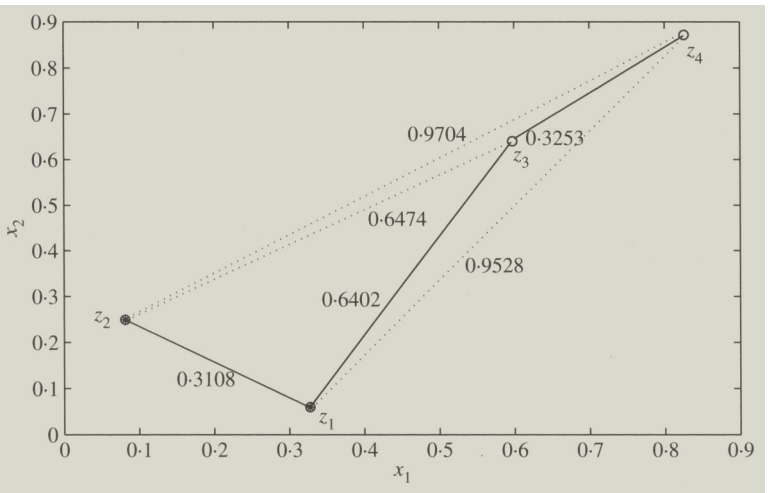

Fig. 1. Example of a shortest Hamiltonian path in a complete graph on four vertices.

values of $T_{m,n}$. Friedman and Rafsky's run test uses the test statistic $T^{\mathrm{FR}}_{m,n} = 1 + \sum_{i=1}^{N-1} W_i$, where W_i denotes the indicator variable that takes the value 1 if and only if the ith edge of the minimal spanning tree on $z_1, \ldots, z_N$ connects two observations from two different distributions. This test rejects H_0 for small values of $T^{\mathrm{FR}}_{m,n}$. If F and G are one-dimensional distributions, both the shortest Hamiltonian path and the minimal spanning tree are obtained by joining the observations $z_1, \ldots, z_N$ in either increasing or decreasing order, and in that case $T_{m,n}$ and $T^{\mathrm{FR}}_{m,n}$ match the univariate run statistic. Therefore, Friedman and Rafsky's test and our test can both be viewed as multivariate generalizations of the univariate run test.

2·2. *Distribution of $T_{m,n}$*

From the above discussion, it is clear that the proposed test has the distribution-free property in one dimension, where it matches the univariate run test. Unlike Friedman and Rafsky's test, our generalization successfully retains this distribution-free property in higher dimensions. Note that $T_{m,n}$ and the univariate run statistic are the same function of ranks of the observations from the two distributions, with $T_{m,n}$ using the ranks computed along $\mathcal{H}^*$. Now, under H_0, because of the exchangeability of $z_1, \ldots, z_N$, irrespective of the underlying distribution and the data dimension, this rank vector has the same distribution as in the univariate case. So $T_{m,n}$ has the distribution-free property, and its null distribution exactly matches that of the univariate run statistic, which is given by (Wald & Wolfowitz, 1940; Gibbons & Chakraborti, 2003)

$$\mathrm{pr}_0(T_{m,n} = 2k) = 2\binom{m-1}{k-1}\binom{n-1}{k-1}\Big/\binom{N}{m}, \quad k = 1, \ldots, \min\{m, n\},$$

$$\mathrm{pr}_0(T_{m,n} = 2k-1) = \left\{\binom{m-1}{k-1}\binom{n-1}{k-2} + \binom{m-1}{k-2}\binom{n-1}{k-1}\right\}\Big/\binom{N}{m},$$
$$k = 2, \ldots, \min\{m, n\} + 1.$$

In that sense, $T_{m,n}$ can be viewed as a natural multivariate generalization of the univariate run statistic. If m and n are small, then to carry out our test we can use statistical tables for the

univariate run test. Because of the discreteness of $T_{m,n}$, one may need to use randomization at the cut-off point to match the size of the test with the significance level. However, if m and n are large, one can also use the test based on the asymptotic null distribution of $T_{m,n}$. Under H_0, the expectation and variance of $T_{m,n}$ are $E_0(T_{m,n}) = 2mn/N + 1$ and $\mathrm{var}_0(T_{m,n}) = 2mn(2mn - N)/\{N^2(N-1)\}$, respectively. Let us assume that as $N \to \infty$, $m/N \to \lambda$ for some $\lambda \in (0, 1)$. Under this condition, $E_0(T_{m,n}/N) \to 2\lambda(1-\lambda)$ and $\mathrm{var}_0(T_{m,n}/\sqrt{N}) \to 4\lambda^2(1-\lambda)^2$ as $N \to \infty$. In this case, one can show (Wald & Wolfowitz, 1940) that under H_0, $T^*_{m,n} = \sqrt{N}\{T_{m,n}/N - 2\lambda(1-\lambda)\}$ converges in distribution to $N\{0, 4\lambda^2(1-\lambda)^2\}$.

2·3. *Computation of $T_{m,n}$ for multivariate data*

Unless m and n are very small, finding $\mathcal{H}^*$ in a complete graph $\mathcal{G}$ is equivalent to the travelling salesman problem, which is NP-complete (Garey & Johnson, 1979). However, some good heuristic search algorithms are available (Lawler et al., 1985). In this article, we adopt a popular method based on Kruskal's algorithm (Kruskal, 1956). This algorithm first sorts the edges of $\mathcal{G}$ in order of increasing cost, and then starts from the edge with minimum cost and selects edges one by one according to their costs. However, if an edge together with previously chosen edges makes a cycle, or if it makes the degree of a vertex more than 2, the method does not select that edge. The algorithm terminates when $N - 1$ edges have been chosen. The Hamiltonian path formed by these $N - 1$ edges is considered to be shortest. This algorithm works well for our test, and the reasons for its success are discussed in § 3.

2·4. *An illustrative example*

We have already mentioned that our run test and Friedman and Rafsky's run test can both be used even when the dimension of the data exceeds the sample size. Now we consider a simple example to investigate how these two tests perform in high-dimension, low-sample-size situations. Let F and G be d-variate normal distributions $N_d\{(0, \ldots, 0)^{\mathrm{T}}, I_d\}$ and $N_d\{(\mu, \ldots, \mu)^{\mathrm{T}}, \sigma^2 I_d\}$, respectively, where I_d denotes the $d \times d$ identity matrix. We consider the two parameter combinations ($\mu = 0{\cdot}3$, $\sigma^2 = 1$) and ($\mu = 0$, $\sigma^2 = 1{\cdot}3$), which lead to a location problem and a scale problem, respectively. In each case, we generated 20 observations from each distribution to test $H_0 : F = G$ against $H_1 : F \neq G$. Each experiment was repeated 500 times, and the proportion of times a test rejected H_0 was used to estimate its power. In the case of Friedman and Rafsky's test, which is not distribution-free, we used the conditional test with 500 permutations. We used values of d ranging from 3 to 3000; see Fig. 2. Like our test, that of Rosenbaum (2005) based on non-bipartite matching is applicable to general two-sample problems and is distribution-free. To make it applicable to high-dimension, low-sample-size data, we used the Euclidean metric for distance computation. For Rosenbaum's test, we used both the distances between the observations and the distances between the coordinate-wise rank vectors. Since the former yielded better results, we report its estimated power in Fig. 2. Throughout this article, all tests have 5% nominal level, although our findings also hold for other nominal levels.

In both location and scale problems, the separation between F and G increases with d, so one should expect the powers of the tests to tend to unity. We observed this in the location problem but not in the scale problem. Figure 2(a) shows that in the location problem, all three tests were comparable, though our test had an edge. But the result was more interesting in the case of the scale problem: in Fig. 2(b), the powers of the proposed test and Rosenbaum's test increased with d, but the latter increased very slowly. While the power of the proposed run test rapidly increased to unity, that of Friedman and Rafsky's run test dropped to zero as d increased. In the next section, we investigate the reasons for this behaviour.

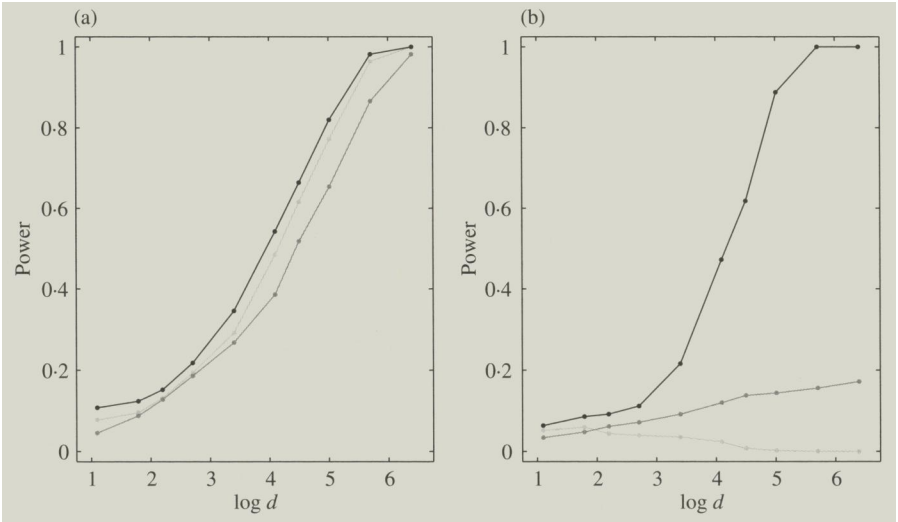

Fig. 2. Powers of Friedman and Rafsky's test (light grey), Rosenbaum's test (dark grey) and the proposed test (black) when: (a) $\mu = 0{\cdot}3$ and $\sigma^2 = 1$; (b) $\mu = 0$ and $\sigma^2 = 1{\cdot}3$.

3. Behaviour of the multivariate run tests for high-dimensional data

We assume m and n to be fixed and study the power functions of the run tests as $d \to \infty$. In usual large-sample asymptotics, we assume d to be fixed and expect to gain more information about the separation of F and G as the sample sizes increase. Here, however, we consider the sample sizes to be fixed and expect to get more information as the dimension increases. We have m independent observations on $X = (X^1, \ldots, X^d)^{\mathrm{T}}$ from F and n independent observations on $Y = (Y^1, \ldots, Y^d)^{\mathrm{T}}$ from G, and the observations on X and Y are independent. Following Hall et al. (2005), we consider the assumptions given below.

Assumption 1. The fourth moments of X^q and Y^q are uniformly bounded for $q \geqslant 1$.

Assumption 2. Let X_1 and X_2 be two independent copies of X, and let Y_1 and Y_2 be two independent copies of Y. Under some permutation of the X^q and the same permutation of the Y^q, for $(U^q, V^q) = (X_1^q, X_2^q)$, (X_1^q, Y_1^q) or (Y_1^q, Y_2^q) we have

$$\sup_{|q_1 - q_2| > r} \left|\mathrm{corr}\{(U^{q_1} - V^{q_1})^2,\ (U^{q_2} - V^{q_2})^2\}\right| \leqslant \rho(r)$$

where $\rho(r) \to 0$ as $r \to \infty$.

Assumption 3. There exist $\sigma_1^2, \sigma_2^2 > 0$ and ν^2 such that:

(i) $d^{-1} \sum_{q=1}^{d} \{E(X^q) - E(Y^q)\}^2 \to \nu^2$;
(ii) $d^{-1} \sum_{q=1}^{d} \mathrm{var}(X^q) \to \sigma_1^2$; and
(iii) $d^{-1} \sum_{q=1}^{d} \mathrm{var}(Y^q) \to \sigma_2^2$ as $d \to \infty$.

Hall et al. (2005) treated d-dimensional observations as infinite time series truncated at time d and studied the behaviour of the interpoint distances as d increases. Here we look at these observations from a multivariate data perspective, so we make some minor modifications to the assumptions of Hall et al. (2005), in particular Assumption 2. Ahn et al. (2007) studied the geometry of high-dimension, low-sample-size data under similar assumptions. Jung & Marron (2009)

assumed almost the same set of conditions for the high-dimensional consistency of estimated principal component directions.

Under Assumptions 1 and 2, the weak law holds for the sequence $\{(U^q - V^q)^2\}$. The proof is straightforward, so it is omitted. Again, depending on the choice of $(U^q, V^q) = (X_1^q, X_2^q)$, (X_1^q, Y_1^q) or (Y_1^q, Y_2^q), under Assumption 3 $d^{-1}\sum_{q=1}^{d} E(U^q - V^q)^2$ converges to $2\sigma_1^2$, $\sigma_1^2 + \sigma_2^2 + \nu^2$ or $2\sigma_2^2$ as d tends to infinity. So, under Assumptions 1–3, the pairwise distance between any two independent observations, when divided by $d^{1/2}$, converges in probability to a positive constant. If both observations are from the same distribution, then depending on whether they are from F or G this scaled distance converges to $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$. If one of the observations is from F and the other is from G, the distance converges to $(\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$. However, if the components of X and Y are independent and identically distributed, as for the normal examples in § 2·4, we only require the existence of second-order moments of the component variables for these convergence results. In this case, we do not need Assumption 1, while Assumptions 2 and 3 hold automatically under the existence of second-order moments. Under Assumptions 1–3, if we have $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$, the power of the proposed test converges to unity as the dimension increases.

THEOREM 1. *Let F and G satisfy Assumptions* 1–3. *Let m and n be such that $m!\,n!/(m+n-1)! \leqslant \alpha$, where α is a predefined positive number. If $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$, the power of the proposed run test of level α converges to* 1 *as $d \to \infty$.*

Note that $m!\,n!/(m+n-1)! < 0{\cdot}05$ for all $m, n \geqslant 5$. So, for the high-dimensional consistency of the proposed test with 5% nominal level, it is enough to have five observations from each distribution. Boxplots in Fig. 3(b) show the distributions of $T_{m,n}$ for different choices of d in the location problem discussed in § 2·4: clearly $T_{m,n} \to 2$ in probability as d increases when $\nu^2 > |\sigma_1^2 - \sigma_2^2|$; but if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$, $T_{m,n}$ converges in probability to 3 as $d \to \infty$; see Fig. 3(d), which displays results for the scale problem. If $\sigma_1^2 > \sigma_2^2$, $\mathcal{H}^*$ starts and ends with observations from F, with all observations from G in the middle. Otherwise, $\mathcal{H}^*$ starts with some observations from G, followed by all observations from F, and then ends with observations from G again. This phenomenon can be explained using the proof of Theorem 1 given in the Appendix.

Theorem 1 holds even for the implemented version of the test, where $T_{m,n}$ is computed along the path obtained by Kruskal's algorithm. If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$, this algorithm first selects $m-1$ edges of XX type and $n-1$ edges of YY type to form two disjoint paths, and then joins these paths with an XY-type edge, giving $T_{m,n} = 2$. When $\nu^2 \leqslant |\sigma_1^2 - \sigma_2^2|$, under the conditions of Theorem 1 we have $|\sigma_1^2 - \sigma_2^2| > 0$. Without loss of generality, let us assume $\sigma_1^2 < \sigma_2^2$, which implies that $2\sigma_1^2 < \sigma_1^2 + \sigma_2^2 + \nu^2 \leqslant 2\sigma_2^2$. So Kruskal's algorithm first selects $m-1$ edges of XX type to form a path on m nodes corresponding to observations from F. Only two of these m nodes will have degree 1 and the rest will have degree 2. Since all nodes in $\mathcal{H}$ have degrees less than or equal to 2, $\mathcal{H}$ cannot have more than two XY-type edges, and so $T_{m,n} \leqslant 3$.

Instead of leading to the actual $\mathcal{H}^*$, Kruskal's algorithm sometimes yields a suboptimal path; but the test is unaffected if the number of runs along that path remains the same. In order to study the behaviour of Kruskal's algorithm, we carried out an experiment with the location problem considered in § 2·4. We chose $d = 3000$ and $m = n = 5$ so that we could compute $\mathcal{H}^*$ by complete enumeration. In most cases, $\mathcal{H}^*$ yielded two runs, where all observations from one distribution were followed by all observations from the other distribution. Clearly, any rearrangement of the observations from a distribution changes the cost of the path but not the value of the run statistic. In many cases, Kruskal's algorithm led to such a rearrangement. Figure 4(a) shows

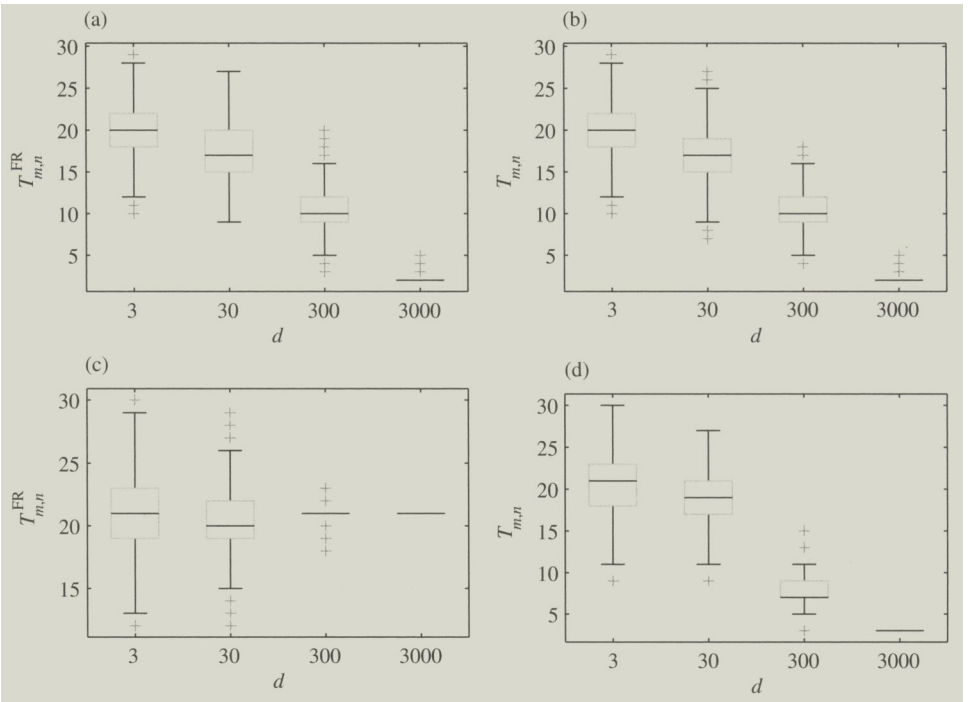

Fig. 3. Boxplots showing distributions of: (a) Friedman and Rafsky's run statistic when $\mu = 0{\cdot}3$ and $\sigma^2 = 1$; (b) the proposed test statistic when $\mu = 0{\cdot}3$ and $\sigma^2 = 1$; (c) Friedman and Rafsky's run statistic when $\mu = 0$ and $\sigma^2 = 1{\cdot}3$; (d) the proposed test statistic when $\mu = 0$ and $\sigma^2 = 1{\cdot}3$.

boxplots of the efficiency scores of Kruskal's algorithm for different dimensions, where each efficiency score is computed as the ratio of the cost of the actual $\mathcal{H}^*$ to that of the path obtained by Kruskal's algorithm. Figure 4(b) shows the distribution of the difference between the test statistics computed along the two paths. These plots demonstrate that Kruskal's algorithm works well, and its performance improves as the dimension increases. For $d = 3000$, the test statistics computed along the two paths were the same in more than 95% of cases. We observed a similar phenomenon for the scale problem as well, and this can be explained using a similar argument.

Under Assumptions 1–3, the performance of Friedman and Rafsky's test depends on the ordering of XX-type, XY-type and YY-type distances. If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$, i.e., if both $\sigma_1\sqrt{2}$ and $\sigma_2\sqrt{2}$ are smaller than $(\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$, then as $d \to \infty$, all XY-type distances become larger than all XX-type and YY-type distances. In that case, each observation from F tends to have its first $m - 1$ neighbours from F, and each observation from G tends to have its first $n - 1$ neighbours from G. As a result, $T^{FR}_{m,n}$ attains its minimum value 2 with probability tending to 1. We observed this phenomenon in Fig. 3(a) for the location problem discussed in § 2·4, where we had $\sigma_1^2 = \sigma_2^2 = 1$ and $\nu^2 = 0{\cdot}09$. So, in this case, unless m and n are very small, the power of Friedman and Rafsky's test converges to 1 as $d \to \infty$. However, the situation is different if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$, i.e., if either $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ exceeds $(\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$. Without loss of generality, let us assume $\sigma_2^2 - \sigma_1^2 > \nu^2$, as in the case of the scale problem in § 2·4. In this case, each observation from F has its first $m - 1$ neighbours from F as before, but now each observation from G has all of its first m neighbours from F also. Thus, $T^{FR}_{m,n}$ converges in probability to $n + 1$; see Fig. 3(c). This limiting value of $n + 1$ is equal to or bigger than the expected value of $T^{FR}_{m,n}$ under H_0 if $m = n$ or $m < n$, respectively. Since this limiting value is much higher than the cut-off, the test performed

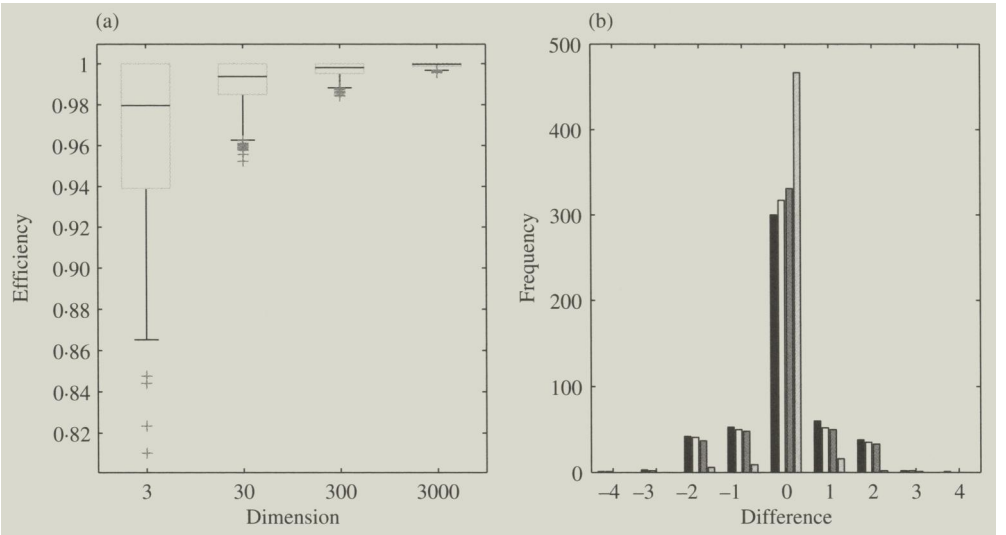

Fig. 4. (a) Boxplots of efficiency scores for Kruskal's algorithm. (b) Distributions of the difference in the values of the run statistic for $d = 3$ (black), 30 (white), 300 (dark grey) and 3000 (light grey).

poorly in the scale problem. In fact, in such cases, depending on m and n, the power of this test may even tend to zero as $d \to \infty$.

THEOREM 2. *Suppose that F and G satisfy Assumptions* 1–3. *Consider a predefined positive number α.*

(i) *If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ and $\max(\lfloor N/n \rfloor, \lfloor N/m \rfloor)/\binom{m+n}{m} \leqslant \alpha$, the power of Friedman and Rafsky's test of level α converges to* 1 *as $d \to \infty$, where $\lfloor r \rfloor$ denotes the greatest integer that does not exceed r.*

(ii) *If $\nu^2 < \sigma_1^2 - \sigma_2^2$ and $m/n > (1+\alpha)/(1-\alpha)$ or if $\nu^2 < \sigma_2^2 - \sigma_1^2$ and $n/m > (1+\alpha)/(1-\alpha)$, the power of Friedman and Rafsky's test of level α converges to* 0 *as $d \to \infty$.*

Part (ii) of Theorem 2 gives only a sufficient condition for the failure of Friedman and Rafsky's test, but this test may fail in other cases. For instance, the scale problem in § 2·4 had $m/n = 1$, but the power of the test dropped to 0 as d increased.

4. ANALYSIS OF SIMULATED DATASETS

We compare the performance of the proposed test with some popular nonparametric two-sample tests. Along with Friedman and Rafsky's test and Rosenbaum's test, here we consider two other tests: a test based on nearest-neighbour-type coincidences (Schilling, 1986; Henze, 1988) and the Cramér test (Baringhaus & Franz, 2004). We took $m = n = 20$ and $m = n = 50$. Unlike the proposed test and Rosenbaum's test, the other three tests are not distribution-free, so for them we used the conditional tests based on 500 permutations. Each experiment was repeated 500 times, and the estimated powers for $d = 30$ and 90 are reported in Table 1.

In Examples 1 and 2, we consider the location and scale problems discussed in § 2·4. In the location problem of Example 1, the Cramér test showed the best performance, followed by the nearest-neighbour test. Our test had the third best performance in this example. On the other hand, in the scale problem of Example 2, our test outperformed all competitors. In view of Theorems 1 and 2, good performance of the proposed test and poor performance of Friedman and Rafsky's

Table 1. *Empirical power (%) of two-sample tests (with 5% nominal level) applied to simulated datasets*

		Example 1		Example 2		Example 3		Example 4		Example 5		Example 6	
N	d	30	90	30	90	30	90	30	90	30	90	30	90
40	FR	29	62	04	01	39	44	07	03	00	00	09	09
	R	27	52	09	14	32	39	12	15	31	43	14	39
	C	83	100	10	15	08	05	06	07	49	67	04	02
	NN	49	87	07	04	48	59	10	09	01	00	18	25
	Prop	35	66	22	62	44	50	18	55	85	99	28	56
100	FR	62	96	06	02	94	97	13	10	00	00	18	19
	R	50	87	11	16	87	93	18	27	71	85	42	88
	C	99	100	14	42	15	13	09	09	92	99	06	06
	NN	84	99	08	04	99	100	23	20	01	00	35	54
	Prop	69	98	39	95	98	99	42	94	99	100	67	98

FR, Friedman and Rafsky's test; R, Rosenbaum's test; C, the Cramér test; NN, nearest-neighbour test; Prop, our proposed test.

test are to be expected in Example 2. Like Friedman and Rafsky's test, the power of the nearest-neighbour test also dropped to zero as the dimension increased. The reason for such behaviour of the nearest-neighbour test can be explained by a result similar to Theorem 2. The reason for the poor performance of the Cramér test in high-dimensional scale problems was discussed in Biswas & Ghosh (2014).

In Examples 3–6, we take $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, where ν^2, σ_1^2 and σ_2^2 have the same meaning as in Assumption 3. We use these examples to investigate how our test performs when the assumptions of Theorem 1 do not hold. Examples 3 and 4 deal with two multivariate normal distributions, where F and G differ only in their correlation structures. In Example 3, F has the scatter matrix Σ_F, whose (i, j)th entry is $0{\cdot}35^{|i-j|}$, and G has the scatter matrix Σ_G, whose (i, j)th entry is $(-0{\cdot}35)^{|i-j|}$. In Example 4, all off-diagonal elements of Σ_F are $0{\cdot}1$, and those of Σ_G are $0{\cdot}3$. Assumptions 1–3 hold in Example 3, but Assumption 2 is violated in Example 4. In Example 3, the nearest-neighbour test had the best performance, followed by our run test. In this example, the Cramér test failed to compete with the other methods. In Example 4, the proposed test outperformed all competitors. Rosenbaum's test had the next best performance, but its power was not comparable to that of our test. In Example 5 two distributions, the multivariate normal distribution $N_d\{(0, \ldots, 0)^{\mathrm{T}}, 3I_d)\}$ and the standard multivariate t-distribution with three degrees of freedom, have the same mean vector and the same dispersion matrix, but differ in their shapes. The proposed test had excellent performance in this example as well. While Friedman and Rafsky's test and the nearest-neighbour test both failed to reject H_0 on even a single occasion, the proposed test could reject it in almost all cases. In Example 6, F is an equal mixture of two normal distributions $N_d(0{\cdot}3\,1_d, I_d)$ and $N_d(-0{\cdot}3\,1_d, 4I_d)$, and G is an equal mixture of two normal distributions $N_d(0{\cdot}3\beta_d, I_d)$ and $N_d(-0{\cdot}3\beta_d, 4I_d)$. Here $1_d = (1, \ldots, 1)^{\mathrm{T}}$ denotes the d-dimensional vector whose elements are all 1, and $\beta_d = (1, \ldots, 1, -1, \ldots, -1)^{\mathrm{T}}$ has its first $d/2$ elements equal to 1 and the remaining $d/2$ elements equal to -1. Again, in this example, the proposed test outperformed the competitors.

Finally, we consider two examples with autoregressive processes of order 1 and order 2. In one example, the observations $X = (X^1, \ldots, X^{500})^{\mathrm{T}}$ in F were generated using the AR(1) model $X^t = 0{\cdot}25 + 0{\cdot}3X^{t-1} + U_t$ for $t = 1, \ldots, 500$, and the observations $Y = (Y^1, \ldots, Y^{500})^{\mathrm{T}}$ in G were generated using another AR(1) model, $Y^t = 0{\cdot}25 + 0{\cdot}5Y^{t-1} + V_t$, where X^0, Y^0, $U_1, \ldots, U_{500}$ and $V_1, \ldots, V_{500}$ are independent $N(0, 1)$ variates. In the other example, the observations in F were generated using the AR(2) model $X^t = 0{\cdot}3X^{t-1} + 0{\cdot}2X^{t-2} + U_t$ for $t =$

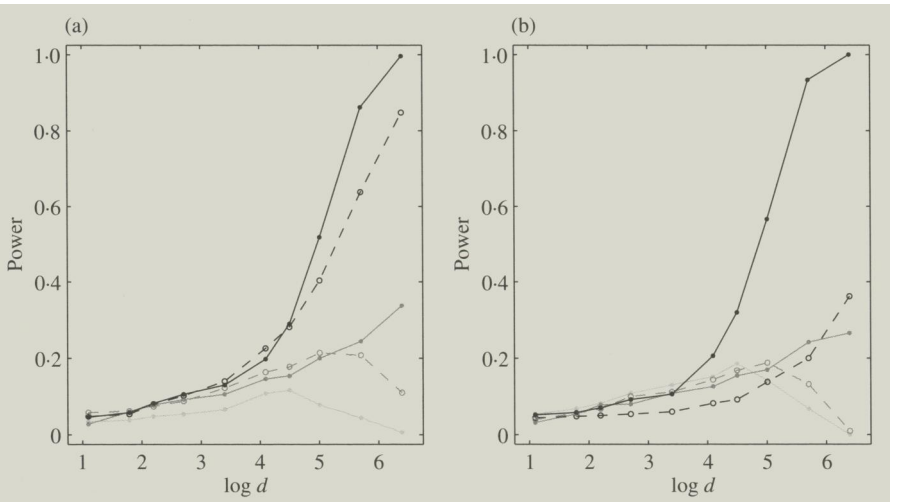

Fig. 5. Powers of Friedman and Rafsky's test (light grey), Rosenbaum's test (dark grey), the Cramér test (black dashed), the nearest-neighbor test (dark grey dashed) and the proposed test (black) in examples involving: (a) two AR(1) models; (b) two AR(2) models.

$1, \ldots, 500$, and those in G were generated using the AR(2) model $Y^t = 0{\cdot}4Y^{t-1} + 0{\cdot}3Y^{t-2} + V_t$ for $t = 1, \ldots, 500$, where X^0, X^{-1}, Y^0, Y^{-1}, $U_1, \ldots, U_{500}$ and $V_1, \ldots, V_{500}$ are all independent standard normal variates. In both examples, we generated 20 observations from each class to form the sample, and the experiment was repeated 500 times. We performed this experiment for d ranging from 3 to 3000; see Fig. 5. The superiority of the proposed test in high dimensions is apparent from this figure, especially in the second example.

5. ANALYSIS OF BENCHMARK DATASETS

We analyse five benchmark datasets. The Trace dataset is obtained from the University of California, Riverside's time series classification/clustering page, at www.cs.ucr.edu/~eamonn/time_series_data/. The Colon dataset is available within the R package dprep. The rest of the datasets are taken from the University of California, Irvine's machine learning repository, at http://archive.ics. uci.edu/ml/datasets/. Detailed descriptions of these datasets are available at their sources; they have been extensively used in the literature on supervised classification. In all cases, we have reasonable separation between two competing classes, so we can assume $H_1 : F \neq G$ to be true, and different tests can be compared on the basis of their power functions. However, if we use the whole dataset for testing, any test will either reject H_0 or accept it. Based on that single experiment, it would be difficult to compare different test procedures. Therefore, in each of the cases, we repeated the experiment 500 times based on 500 random subsets of the same size chosen from the whole dataset. We formed these subsets by taking an equal number of observations from the two classes, and the results for different subset sizes are shown in Fig. 6.

The ionosphere dataset contains 34-dimensional observations on 126 good and 225 bad radar returns. Figure 6(a) shows that for this dataset, the proposed test and the Cramér test performed better than their competitors. For sample sizes below 20, the latter had a slight edge over the proposed test, but the proposed test had higher power thereafter. These two tests had power 1 for samples of size 40 or greater. The performances of the other three tests were also comparable. Among them, the nearest-neighbour test yielded the best performance.

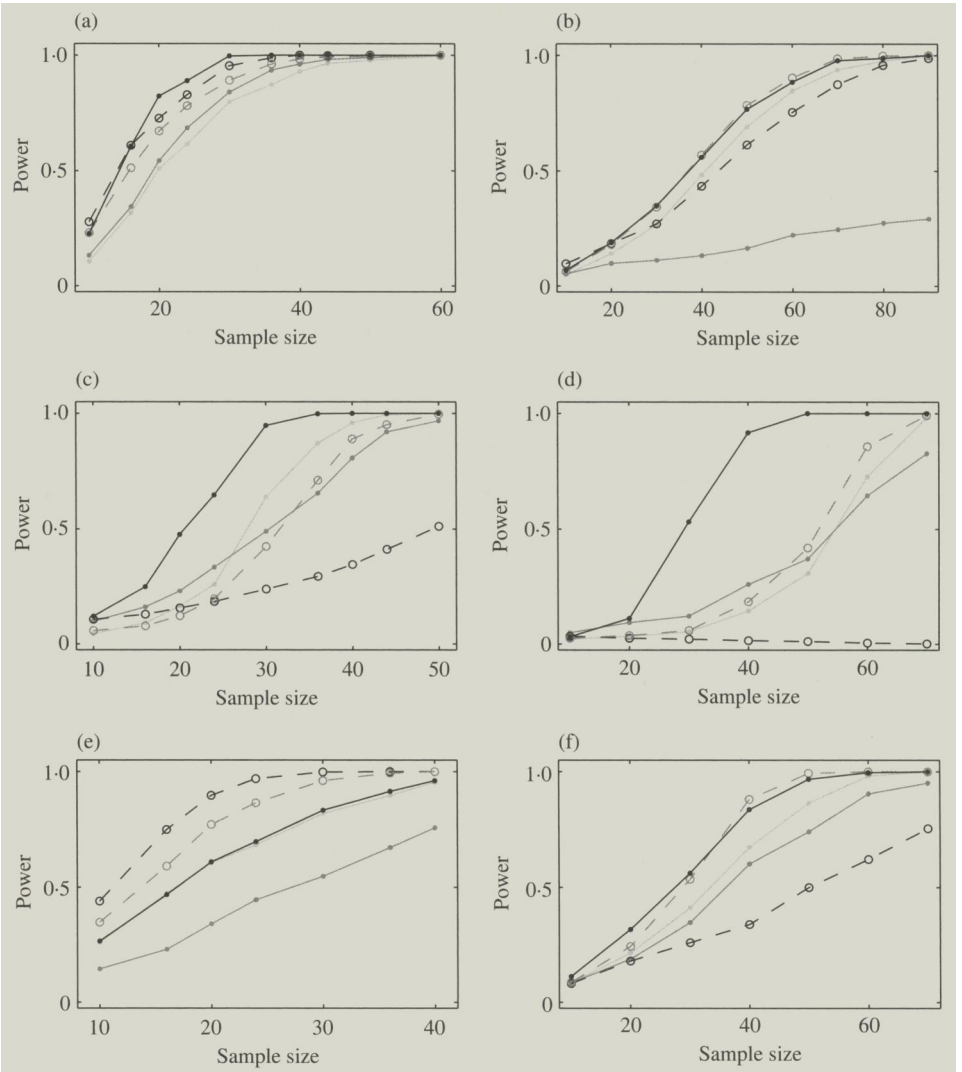

Fig. 6. Powers of Friedman and Rafsky's test (light grey), Rosenbaum's test (dark grey), the Cramér test (black dashed), the nearest-neighbour test (dark grey dashed) and the proposed test (black) in the analysis of: (a) the ionosphere dataset; (b) the sonar dataset; (c) Trace data-1; (d) Trace data-2; (e) the Colon dataset; (f) the Arcene dataset.

The sonar dataset contains 111 patterns obtained by bouncing sonar signals off a metal cylinder, together with 97 patterns obtained from rocks. Each number in a 60-dimensional pattern represents the energy within a particular frequency band integrated over a certain period of time. For this dataset, the nearest-neighbour test had the best overall performance, closely followed by the proposed test; see Fig. 6(b). In all cases, the difference between the powers of these two tests was less than 0·02. Friedman and Rafsky's test also had comparable performance. The Cramér test had the highest power for sample size 10, but it was outperformed by the nearest-neighbour test and the proposed test for all larger sample sizes. Rosenbaum's test did not have satisfactory performance on this dataset.

The Trace dataset consists of four classes, each containing 50 instances of length 275. We consider two two-sample problems relating to this dataset, one between the first and second classes and the other between the third and fourth classes; we refer to these as Trace data-1 and

Trace data-2, respectively. From Fig. 6(c) and (d), it is clear that in both these cases, our test had substantially higher power than all the other tests considered. The Cramér test had very poor performance on these datasets, especially in the case of Trace data-2.

Next, we consider two datasets with data dimensions larger than 1000. The Colon dataset is a microarray gene expression dataset that contains expression levels of 2000 genes for each of 62 samples, 40 from colon cancer tissue and 22 from normal tissue. Figure 6(e) shows that for this dataset, the Cramér test yielded the best performance, while the nearest-neighbour test came in second place. Friedman and Rafsky's run test and the proposed test had very similar performances, and they both performed better than Rosenbaum's test.

Finally, we analyse the Arcene dataset, for which the samples include ovarian or prostate cancer patients and healthy patients. The data contain 7000 features indicating the abundance of proteins in human sera having a given mass value. In addition, 3000 probes were used to increase the number of features to 10 000. The data repository includes separate training, test and validation sets. For our analysis, we used random subsets from the training set consisting of 44 cancer patients and 56 healthy patients. For this dataset, the proposed test and the nearest-neighbour test outperformed all other tests considered here. Figure 6(f) shows that the proposed test had an edge over the nearest-neighbour test for small sample sizes, but for samples of size 40 and 50, the latter test had the higher power. Both tests had power 1 for samples of size 60 or greater.

In terms of overall performance on these benchmark datasets, the proposed test is comparable to, if not better than, the other nonparametric two-sample tests considered here. The multisample extension of the proposed test is straightforward, and the null distribution of the test statistic can be found in Mood (1940). The idea underlying the proposed test, based on shortest Hamiltonian path, can also be used as a general recipe for distribution-free multivariate generalizations of many other univariate rank-based tests. For instance, by using this idea, one can develop distribution-free multivariate versions of the Wilcoxon–Mann–Whitney statistic or the two-sample Kolmogorov–Smirnov statistic, which can be used for testing even when the dimension of the data exceeds the sample size.

ACKNOWLEDGEMENT

We thank an associate editor and two referees for their careful reading of an earlier version of the manuscript and for helpful comments. We are also grateful to Sasthi C. Ghosh for his suggestions and help in writing the computer programs.

APPENDIX

Proof of Theorem 1. Recall that $T_{m,n}$ has the same null distribution as the univariate run statistic, and hence $\mathrm{pr}_0(T_{m,n} \leqslant 3) = m!\, n!/(m+n-1)!$. Since $\mathrm{pr}_0(T_{m,n} \leqslant 3) \leqslant \alpha$, both $T_{m,n} = 2$ and $T_{m,n} = 3$ lead to the rejection of H_0; so it is enough to prove that $\mathrm{pr}_1(T_{m,n} > 3) \to 0$ as $d \to \infty$, where pr_1 denotes the probability under H_1.

Define $a = \sigma_1\sqrt{2}$, $b = \sigma_2\sqrt{2}$ and $c = (\sigma_1^2 + \sigma_2^2 + \nu^2)^{1/2}$. As d tends to infinity, $\|x_i - x_j\|/\sqrt{d}$ converges in probability to a for $1 \leqslant i < j \leqslant m$, $\|y_i - y_j\|/\sqrt{d}$ converges in probability to b for $1 \leqslant i < j \leqslant n$, and $\|x_i - y_j\|/\sqrt{d}$ converges in probability to c for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Clearly $2c \geqslant a + b$, where equality holds if and only if $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. Let $\mathcal{H}$ be a Hamiltonian path in the graph on $m+n$ vertices. Now, $\mathcal{H}$ can either: (i) start and end with observations from same distribution; or (ii) start with an observation from one distribution and end with an observation from the other distribution.

In case (i), $T_{m,n}$ can take only odd values, i.e., $T_{m,n} = 2k+1$ for some integer $k > 0$. Now, if $\mathcal{H}$ starts and ends with observations from F, then $\mathcal{H}$ contains $m-k-1$ edges of XX type, $n-k$ edges of YY type and $2k$ edges of XY type; so the total cost of $\mathcal{H}$ converges in probability to $(m-k-1)a + (n-k)b + 2kc =$

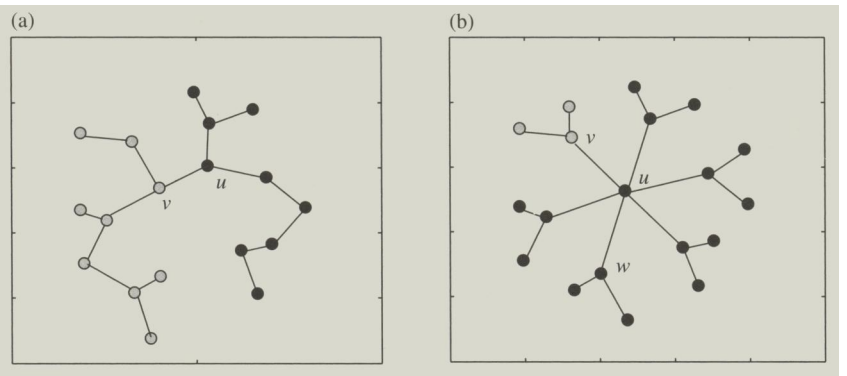

Fig. A1. Minimal spanning trees with $T^{\mathrm{FR}}_{m,n} = 2$ when: (a) $m = n = 9$; (b) $m = 16$ and $n = 3$.

$(m-1)a + nb + k(2c - a - b)$. Similarly, if $\mathcal{H}$ starts and ends with observations from G, the total cost of $\mathcal{H}$ converges to $(m-k)a + (n-k-1)b + 2kc = ma + (n-1)b + k(2c - a - b)$. Under the condition that $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$, we have $2c > a + b$. So, irrespective of whether $\mathcal{H}$ starts and ends with F or G, the cost of $\mathcal{H}$ is minimum when $k = 1$. Therefore $\mathcal{H}^*$, the shortest Hamiltonian path, cannot have more than three runs or, in other words, $\mathrm{pr}_1(T_{m,n} > 3 \mid T_{m,n}$ is odd$) \to 0$ as $d \to \infty$.

In case (ii), we have $T_{m,n} = 2k$ for some integer $k > 0$. In this case, there are $m - k$ edges of XX type, $n - k$ edges of YY type and $2k - 1$ edges of XY type in $\mathcal{H}$; so the total cost of $\mathcal{H}$ converges to $(m-k)a + (n-k)b + (2k-1)c = (m-1)a + (n-1)b + c + (k-1)(2c - a - b)$, which is minimum when $k = 1$. Therefore, $\mathrm{pr}_1(T_{m,n} > 2 \mid T_{m,n}$ is even$) \to 0$ as $d \to \infty$. □

Proof of Theorem 2. (i) Under the condition that $\nu^2 > |\sigma_1^2 - \sigma_2^2|$, $T^{\mathrm{FR}}_{m,n}$ converges in probability to 2 as $d \to \infty$; see Fig. 3(a) and the discussion in § 3. So there exist a subtree $\mathcal{T}_1$ on m vertices corresponding to m observations from F and another subtree $\mathcal{T}_2$ on n vertices corresponding to n observations from G. These two subtrees are connected by an edge $e = \{uv\}$, where u and v correspond to vertices of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Now let us compute $\mathrm{pr}(T^{\mathrm{FR}}_{m,n} = 2)$ under the permutation distribution. If $\mathcal{T}_1$ and $\mathcal{T}_2$ both contain some vertices labelled as F and some labelled as G, then $T^{\mathrm{FR}}_{m,n}$ cannot be 2. Thus, if $m = n$, there are only two possibilities: either all vertices of $\mathcal{T}_1$ or all vertices of $\mathcal{T}_2$ should be labelled F, as in Fig. A1(a). In that case, $\mathrm{pr}(T^{\mathrm{FR}}_{m,n} = 2)$ turns out to be $2m!\,n!/(m+n)!$. Now, without loss of generality, let us assume $m > n$. In this case, all vertices of $\mathcal{T}_2$ should have the same label. If all of them are labelled G, all vertices of $\mathcal{T}_1$ will get the label F as in Fig. A1(b). If all vertices of $\mathcal{T}_2$ are labelled F, to count the number of favourable cases, first note that u must have label F; also, at most one of its neighbours, i.e., the vertices that share an edge with u, can have label G. Suppose that $w \neq v$ is the neighbour having label G. Consider the collection C_w of all vertices in $\mathcal{T}_1$ that connect to u through w. All vertices in this collection, including w, should have label G, and no other vertices in $\mathcal{T}_1$ can have label G. So the cardinality of C_w must be n. Similarly, the other neighbours of u can have label G only if the corresponding collection has cardinality n. Therefore, if the collection corresponding to each of the k neighbours of u including v has cardinality n, the vertex w can be chosen in $k - 1$ different ways, and the total number of favourable cases turns out to be k, including the one where all vertices of $\mathcal{T}_2$ have label G. If u does not have any neighbour labelled G, then, instead of on u, the same argument can be used on each of the neighbours of u barring v. In order to have these k favourable cases, we need $kn + 1 \leqslant N$ or $(N-1)/n > k$, where $N = m + n$. So we cannot have more than $\lfloor (N-1)/n \rfloor$ favourable cases. Similarly, if $n > m$, the number of favourable cases cannot exceed $\lfloor (N-1)/m \rfloor$. Recall that if $N/n = N/m = 2$, i.e., $m = n$, the number of favourable cases is 2. Thus, combining all these results, under the permutation distribution we get $\mathrm{pr}(T^{\mathrm{FR}}_{m,n} = 2) \leqslant k\,m!\,n!/(m+n)!$, where $k = \max\{\lfloor N/n \rfloor, \lfloor N/m \rfloor\}$. If this upper bound is smaller than α, the power of Friedman and Rafsky's run test of level α converges to 1 as d tends to infinity.

(ii) Under the given condition, $T = T^{FR}_{m,n} - 1$ converges to m in probability; see Fig. 3(c) and the discussion in § 3. Note that T is a nonnegative random variable, and $E(T \mid \mathcal{Z})$, the conditional expectation of the permutation distribution of T given the data $\mathcal{Z} = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$, does not depend on $\mathcal{Z}$ (Friedman & Rafsky, 1979) and is given by $E(T \mid \mathcal{Z}) = 2mn/N$ for all $\mathcal{Z}$, where $N = m + n$. Therefore, using the Markov inequality, we have $\text{pr}(T \geqslant m \mid \mathcal{Z}) \leqslant 2n/N \Rightarrow \text{pr}(T < m \mid \mathcal{Z}) \geqslant (m - n)/N$. Now, $m/n > (1 + \alpha)/(1 - \alpha)$ implies $(m - n)/N > \alpha$, and this completes the proof. □

REFERENCES

AHN, J., MARRON, J. S., MULLER, K. M. & CHI, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–6.

BARINGHAUS, L. & FRANZ, C. (2004). On a new multivariate two-sample test. *J. Mult. Anal.* **88**, 190–206.

BICKEL, P. J. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Statist.* **40**, 1–23.

BISWAS, M. & GHOSH, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *J. Mult. Anal.* **123**, 160–71.

CHOI, K. & MARDEN, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *J. Am. Statist. Assoc.* **92**, 1581–90.

FERGER, D. (2000). Optimal tests for the general two-sample problem. *J. Mult. Anal.* **74** 1–35.

FRIEDMAN, J. H. & RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**, 697–717.

GAREY, M. & JOHNSON, D. (1979). *Computers and Intractability: A Guide to the Theory of NP Completeness*. San Francisco: W.H. Freeman & Co.

GIBBONS, J. D. & CHAKRABORTI, S. (2003). *Nonparametric Statistical Inference*. New York: Marcel Dekker.

HALL, P., MARRON, J. S. & NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

HALL, P. & TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89**, 359–74.

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16**, 772–83.

HETTMANSPERGER, T. P., MÖTTÖNEN, J. & OJA, H. (1998). Affine invariant multivariate rank tests for several samples. *Statist. Sinica* **8**, 785–800.

HETTMANSPERGER, T. P. & OJA, H. (1994). Affine invariant multivariate multi-sample sign tests. *J. R. Statist. Soc.* B **56**, 235–49.

JUNG, S. & MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37**, 4104–30.

KRUSKAL, J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. Am. Math. Soc.* **7**, 48–50.

LAWLER, E., LENSTRA, J., KAN, A. & SHMOYS, D. (1985). *The Travelling Salesman Problem*. New York: Wiley.

LIU, R. Y. & SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Am. Statist. Assoc.* **88**, 252–60.

LIU, Z. & MODARRES, R. (2011). A triangle test for equality of distribution functions in high dimensions. *J. Nonparam. Statist.* **23**, 605–15.

MOOD, A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.* **11**, 367–92.

MÖTTÖNEN, J. & OJA, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparam. Statist.* **5**, 201–13.

PURI, M. L. & SEN, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. New York: Wiley.

RANDLES, R. H. & PETERS, D. (1990). Multivariate rank tests for the two-sample location problem. *Commun. Statist.* A **19**, 4225–38.

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Statist. Soc.* B **67**, 515–30.

ROUSSON, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *J. Mult. Anal.* **80**, 43–57.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Am. Statist. Assoc.* **81**, 799–806.

WALD, A. & WOLFOWITZ, J. (1940). On a test whether two samples are from the same distribution. *Ann. Math. Statist.* **11**, 147–62.

[*Received August* 2013. *Revised July* 2014]