Journal of Statistical Distributions and Applications
a SpringerOpen Journal

METHODOLOGY Open Access

A new multivariate two-sample test using regular minimum-weight spanning subgraphs

David M Ruth

Correspondence: druth@usna.edu
Department of Mathematics, United States Naval Academy, 572-C Holloway Rd, Annapolis, MD 21402, USA

Abstract

A new nonparametric test is proposed for the multivariate two-sample problem. Similar to Rosenbaum's cross-match test, each observation is considered to be a vertex of a complete undirected weighted graph; interpoint distances are edge weights. A minimum-weight, *r*-regular subgraph is constructed, and the mean cross-count test statistic is equal to the number of edges in the subgraph containing one observation from the first group and one from the second, divided by *r*. Unequal distributions will tend to result in fewer edges that connect vertices between different groups. The mean cross-count test is sensitive to a wide range of distribution differences and has impressive power characteristics. We derive the first and second moments of the mean cross-count test, and note that simulation studies suggest this test statistic is asymptotically normal regardless of underlying data distributions. A small simulation study compares the power of the mean cross-count test to Hotelling's T^2 test and to the cross-match test. This new test is a more powerful generalization of Rosenbaum's test (the cross-match test is the case $r = 1$) and constitutes a noteworthy addition to the class of multivariate, nonparametric two-sample tests.

Keywords: Distribution-free test; Graph-theoretic procedure; Change point

1 Background

1.1 Objective

Consider $N = m + n$ independent multivariate observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ and $\mathbf{Y}_{m+1}, \ldots, \mathbf{Y}_N$, where each $\mathbf{Y}_i$ is drawn from distribution F for $1 \le i \le m$ and from distribution G for $m + 1 \le i \le N$. The dimension of the observations does not depend on N. The covariates may be quantitative or categorical; there need only exist some function, d, that measures distance between observations. The null hypothesis is that $F = G$. The objective is a two-sample test that has little or no dependence on the underlying distribution of the data. Furthermore, this test should have sufficient power to be useful for applications.

1.2 Motivation

We follow in the vein of graph-theoretic tests for homogeneity: Consider each observation to be a vertex of a complete, undirected, weighted graph, $\mathcal{G}$, and assign interpoint distances as edge weights. The distribution of these distances is sensitive to departures from homogeneity; Maa et al. (1996) prove that two distributions are identical if and only if the distributions of inter-point distances within and between observations sampled from the two populations are the same. Friedman and Rafsky (1979, 1981) fit a minimum spanning tree

to $\mathcal{G}$ and count the number of edges in the tree that connect vertices from different groups to test whether the sampling distributions are the same. Schilling (1986), Henze (1988), and Hall and Tajvidi (2002) examine properties of nearest-neighbor subgraphs of $\mathcal{G}$ to test for homogeneity.

Rosenbaum (2005) provides a novel approach to this problem: Suppose N is even. Find a minimum-weight non-bipartite matching on $\mathcal{G}$, which is the lowest-weight spanning subgraph for which the degree of each vertex with respect to the subgraph is one and which consists of $N/2$ non-adjacent edges. Rosenbaum's cross-match statistic, A_1, counts the number of edges in the matching that include one vertex from each of the two groups. Under the null hypothesis of no group difference each vertex is equally likely to be paired with any other vertex. Rosenbaum (2005) shows that the exact null distribution of A_1 is found by combinatorial argument to be

$$P(A_1 = a_1) = \frac{2^{a_1}(N/2)!}{\binom{N}{m}\left(\frac{m-a_1}{2}\right)!a_1!\left(\frac{n-a_1}{2}\right)!} \tag{1}$$

for $a_1 \in \{0, 2, \ldots, \min(m, n)\}$ and m and n even, or $a_1 \in \{1, 3, \ldots, \min(m, n)\}$ and m and n odd; $P(A_1 = a_1) = 0$ otherwise. In the denominator of (1), $\frac{1}{2}(m-a_1)$ is the number of edges in the matching where both vertices are in the group of size m and $\frac{1}{2}(n-a_1)$ is the number of edges in the matching where both vertices are in the group of size n.

When the two groups are drawn from different distributions the number of within-group pairs tends to be higher than for the null case, so the null hypothesis of homogeneity is rejected if A_1 is sufficiently small. For odd N, this procedure may be simply modified by introducing a pseudo-observation, $\mathbf{Y}_0$ such that $d(\mathbf{Y}_0, \mathbf{Y}_i) = 0$ for all $i \in \{1, \ldots, N\}$, and randomly assigning it to one of the two groups. Then find a minimum-weight non-bipartite matching on this resulting graph with $N + 1$ vertices and compute A_1 with respect to observations $\mathbf{Y}_0, \ldots, \mathbf{Y}_N$.

That the exact null distribution of A_1 is known, regardless of the underlying data distribution, is a particularly attractive property for a multivariate two-sample test. Furthermore, the asymptotic normality of A_1 facilitates testing for large-sample problems. However, the cross-match test has relatively low power. Since only a single non-bipartite matching is considered in this test, information contained in the proximity of many pairs of points is ignored. Friedman and Rafsky (1979, 1981) observe that the power of their single-tree test is enhanced by evaluating successive disjoint low-weight spanning trees. Similarly, Ruth and Koyak (2011) show that ensembles of disjoint low-weight non-bipartite matchings carry significant information regarding whether a distributional change occurs over a sequence of independent observations. A drawback associated with examining collections of such subgraphs is that null distributions are extremely difficult to determine. Mindful of this caveat, we offer an extension of the cross-match test which exploits the information contained in the distances between many pairs of points.

2 Methods

2.1 Illustrating example

Consider the bivariate sample of size $N = 20$ listed in Table 1 and displayed in Figure 1. The sample consists of independent observations in groups 1 (○) and 2 (△); observations

Table 1 Bivariate data for illustrating example

Observation number	Group	Covariate 1	Covariate 2
1	1	-0.323	-1.389
2	1	1.020	-2.078
3	1	-0.269	-1.020
4	1	0.296	-0.144
5	1	0.602	1.021
6	1	0.814	-0.508
7	1	-0.475	-0.690
8	1	-0.079	1.360
9	1	-0.228	0.926
10	1	-0.481	1.958
11	2	1.269	1.275
12	2	0.954	2.133
13	2	-0.103	2.763
14	2	-0.581	-0.428
15	2	2.367	0.222
16	2	0.980	1.870
17	2	0.494	1.981
18	2	0.293	0.236
19	2	1.535	0.981
20	2	1.993	-0.120

within groups are identically distributed. For the purposes of this example, these data were simulated from distributions whose locations differ by one unit in each dimension. Figure 1 also shows the minimum-weight non-bipartite matching associated with this sample with respect to Euclidean distance. The present goal is to identify the distribution difference between these groups, making no assumptions about the underlying distributions.

The cross-match test is applicable here; for this example the value of the cross-match statistic is $A_1 = 4$ with a corresponding p-value = 0.433. So, the cross-match test is insufficiently powerful to identify a distribution difference in this case. In the next section, we introduce an extension of the cross-match test that enhances test power significantly.

2.2 The mean cross-count (MCC) test

As before, we assume an even number N of observations forming a complete, undirected, weighted graph, $\mathcal{G}$. Rather than find a minimum-weight non-bipartite matching, we find a *minimum-weight r-regular spanning subgraph of* $\mathcal{G}$, where $1 \le r \le N - 2$, denoted $\mathcal{G}_r^*$. That is, $\mathcal{G}_r^*$ is a subgraph of $\mathcal{G}$ with the following properties:

a) Every vertex in $\mathcal{G}$ is also in $\mathcal{G}_r^*$.
b) Every vertex in $\mathcal{G}_r^*$ has degree r.
c) The total weight of all edges in $\mathcal{G}_r^*$ is the lowest among all subgraphs of $\mathcal{G}$ which satisfy properties (a) and (b).

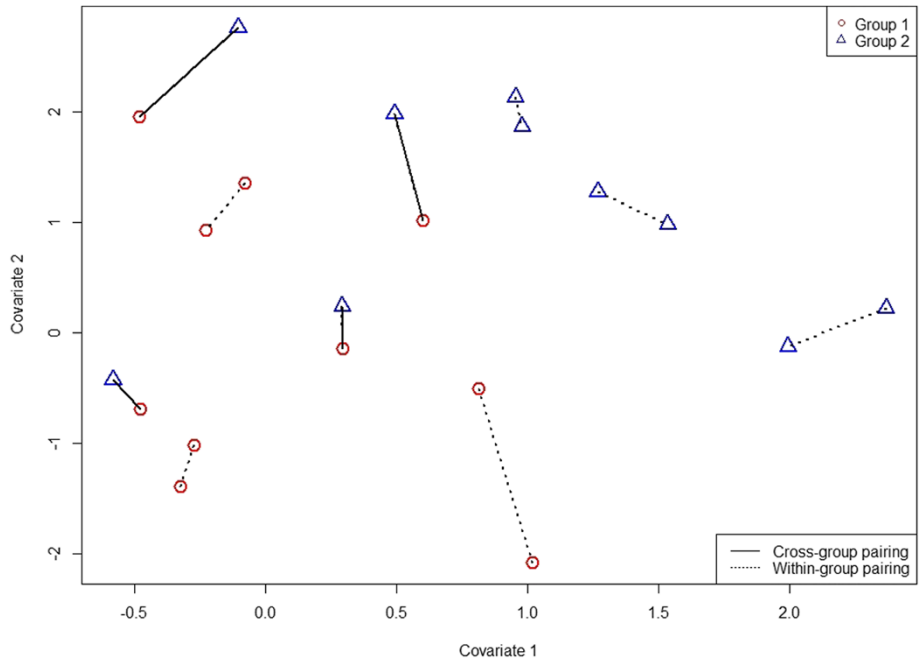

Figure 1 Bivariate data for illustrating example with optimal non-bipartite matching on groups 1 (○) and 2 (△) with $m = n = 10$. Cross-group pairs are connected by solid lines; within-group pairs by dotted lines.

In graph theory, an r-regular spanning subgraph of $\mathcal{G}$ is sometimes called an *r-factor of* $\mathcal{G}$. Note that $\mathcal{G}_1^*$ is the special case of a minimum-weight non-bipartite matching used by Rosenbaum (2005), and $\mathcal{G}_{N-1}^*$ is identical to $\mathcal{G}$. In practice, we are mainly interested in $2 \le r \le N/2$, although the theoretical details are not so constrained. Minimum-weight r-factors may be computed as follows: For any subgraph of $\mathcal{G}$, let x_{ij} be an indicator variable equal to 1 if the edge connecting vertices i and j is included in the subgraph and let d_{ij} be the distance between vertex i and vertex j. Then the edges of $\mathcal{G}_r^*$ solve following the combinatorial optimization problem:

$$\begin{aligned} &\min_{\mathbf{x}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} d_{ij} x_{ij} \\ &\text{subject to} \sum_{i=1}^{k-1} x_{ik} + \sum_{j=k+1}^{N} x_{kj} = r \quad \forall k \in \{1, \ldots, N\} \\ &x_{ij} \in \{0,1\} \ \forall j \in \{i+1, \ldots, N\}, \ \forall i \in \{1, \ldots, N-1\}. \end{aligned} \tag{2}$$

Anderson (1972) assures the existence of a solution for $r \le N/2$. Solutions for $r > N/2$ are guaranteed by the fact that the complement of an r-regular subgraph of $\mathcal{G}$ is an $(N-1-r)$ -regular graph. For this paper, solutions are found in R using the package "lpSolve" for $N \le 400$. For $N > 400$, solutions are found in R using the package "gurobi" due to the computational complexity of larger problems.

Similar to the cross-match test, we count the number of edges A_r in $\mathcal{G}_r^*$ that include a vertex from each group. We call $T_r = A_r/r$ the *mean cross-count* (*MCC*) *statistic*. The idea here is that the number of within-group edges in $\mathcal{G}_r^*$ will be higher for cases of a distribution difference than for the null case. So, small values of T_r are evidence against

the null hypothesis. Note that $T_1 = A_1$ is the cross-match statistic as before. One could use the total cross-count, A_r, as an equivalent test-statistic; however, we choose to scale this value to give some notion of "average cross-count per vertex degree" (hence the name "mean cross-count"). For odd N, randomly introduce a pseudo-observation in the same manner as the $r = 1$ case discussed in Section 1.2.

2.3 Illustrating example, continued

Figure 2 shows a minimum-weight 3-factor, $\mathcal{G}_3^*$, for the data in Table 1 with respect to Euclidean distance. Cross-group edges are shown with solid lines. For this case, $A_3 = 12 \Rightarrow T_3 = 4$, so the test statistic value here is the same as Rosenbaum's cross-match test statistic. A discussion of the distribution of T_r is in Section 3.1; for this example, we estimate the p-value for T_r by permutation test on the observation vertex labels. Using 10,000 permutations yields an estimated p-value = 0.146. While not enough evidence to conclude a group difference, this reduction in p-value relative to the $r = 1$ case (p-value = 0.433) suggests that considering minimum-weight r-factors for $r > 1$ may improve test power. In Section 3.2 we demonstrate significant power advantages that are realized for the MCC statistic.

3 Results and discussion

3.1 MCC moments and normal approximation

For the following discussion we assume N is even, adopting the convention that if the number of observations is odd then we will consider N to be the number of observations including a pseudo-observation as previously discussed. To find the mean and variance of T_r under the null hypothesis, we proceed as follows: Let $\mathcal{G}$ be the complete

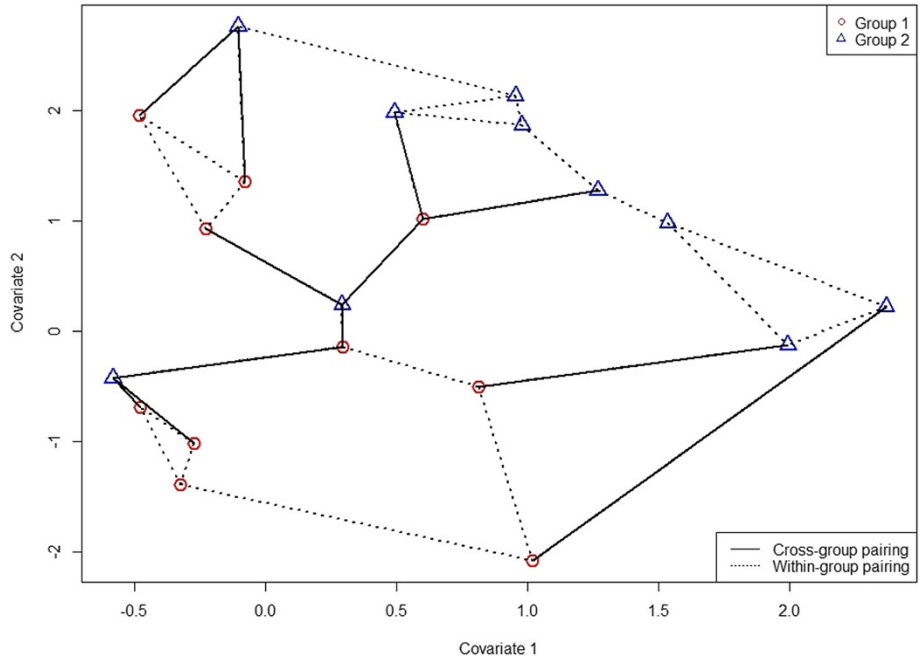

Figure 2 Bivariate data for illustrating example with optimal 3-factor on groups 1 (○) and 2 (△) with $m = n = 10$. Cross-group pairs are connected by solid lines; within-group pairs by dotted lines.

undirected graph $(\mathbb{Z}_N, E_N)$ where the vertex set $\mathbb{Z}_N$ consists of the indices $1, 2, \ldots, N$ and the edge set consists of all $N(N-1)/2$ pairs of vertices; by convention, write the pairs with smaller vertex first, so $E_N = \{(i,j) : 1 \le i < j \le N\}$. Partition $\mathbb{Z}_N$ into two sets S and T, with $|S| = m$ and $|T| = n$, so $m + n = N$. Denote $E_N^{(S,T)}$ as the set of all edges with one vertex in S and the other in T. Let X_{ij} be the random variable that indicates whether edge (i,j) is included in a minimum-weight r-regular subgraph, $\mathcal{G}_r^*$, with $1 \le r \le N-2$. By the r-regularity of $\mathcal{G}_r^*$, for each $i \in \mathbb{Z}_N$ we have $r = \sum_{j=1}^{i-1} X_{ji} + \sum_{j=i+1}^{N} X_{ij}$, and so

$$\begin{aligned} r = E[r] &= E\left[\sum_{j=1}^{i-1} X_{ji} + \sum_{j=i+1}^{N} X_{ij}\right] = \sum_{j=1}^{i-1} E\left[X_{ji}\right] + \sum_{j=i+1}^{N} E\left[X_{ij}\right] \\ &= \sum_{j=1}^{i-1} P\left(X_{ji} = 1\right) + \sum_{j=i+1}^{N} P\left(X_{ij} = 1\right). \end{aligned} \tag{3}$$

But under the null hypothesis, each edge is equally likely to be included in $\mathcal{G}_r^*$, so $r = (N-1)P(X_{12} = 1)$. Therefore, for all $(i,j) \in E_N$

$$E\left[X_{ij}\right] = P\left(X_{ij} = 1\right) = \frac{r}{N-1} \tag{4}$$

and

$$\mathrm{Var}\left[X_{ij}\right] = P\left(X_{ij} = 1\right)P\left(X_{ij} = 0\right) = \frac{r(N-1-r)}{(N-1)^2}\ . \tag{5}$$

The total cross-count, A_r, may be written

$$A_r = \sum_{(i,j) \in E_N^{(S,T)}} X_{ij} \tag{6}$$

resulting in

$$\begin{aligned} E[T_r] = \frac{1}{r}E[A_r] = \frac{1}{r}E\left[\sum_{(i,j) \in E_N^{(S,T)}} X_{ij}\right] &= \frac{mn}{r}E\left[X_{ij}\right] \\ &= \frac{mn}{N-1}\ . \end{aligned} \tag{7}$$

Finding the variance of T_r is slightly more involved. First take

$$\mathrm{Var}[A_r] = \mathrm{Var}\left[\sum_{(i,j) \in E_N^{(S,T)}} X_{ij}\right] = \sum_{(i,j) \in E_N^{(S,T)}} \mathrm{Var}\left[X_{ij}\right] + \sum_{\substack{(i,j),(k,l) \in E_N^{(S,T)} \\ (i,j) \ne (k,l)}} \mathrm{Cov}\left[X_{ij}, X_{kl}\right]\ . \tag{8}$$

The sum of variances is computed directly as

$$\sum_{(i,j) \in E_N^{(S,T)}} \mathrm{Var}\left[X_{ij}\right] = mn\mathrm{Var}\left[X_{ij}\right] = \frac{mnr(N-1-r)}{(N-1)^2}\ . \tag{9}$$

The sum of covariances may be partitioned into terms that include pairs of adjacent edges and terms that include disjoint (i.e., non-adjacent) edges:

$$\sum_{\substack{i,k\in S\\ j,l\in T\\ (i,j)\neq(k,l)}} \operatorname{Cov}\left[X_{ij},X_{kl}\right] = \sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \operatorname{Cov}\left[X_{ij},X_{kj}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{kl}\right] \quad (10)$$

For any two adjacent edges (k,l) and (i,j),

$$\begin{aligned} P\left(X_{ij}X_{kl}=1\right) &= P\left(X_{kl}=1|X_{ij}=1\right)P\left(X_{ij}=1\right) = \frac{(r-1)r}{(N-2)(N-1)} \\ &= E\left[X_{ij}X_{kl}\right], \end{aligned} \quad (11)$$

so

$$\begin{aligned} &\sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \operatorname{Cov}\left[X_{ij},X_{kj}\right] \\ &= (mn(n-1)+m(m-1)n)\left(\frac{(r-1)r}{(N-2)(N-1)} - \left(\frac{r}{N-1}\right)^2\right) \\ &= -\frac{mnr(N-1-r)}{(N-1)^2}. \end{aligned} \quad (12)$$

For any two disjoint edges (k,l) and (i,j),

$$\begin{aligned} P\left(X_{ij}X_{kl}=1\right) &= P\left(X_{kl}=1|X_{ij}=1\right)P\left(X_{ij}=1\right) = \frac{(r(N-4)+2)}{(N-3)(N-2)}\frac{r}{(N-1)} \\ &= E\left[X_{ij}X_{kl}\right], \end{aligned} \quad (13)$$

So

$$\begin{aligned} \sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{kl}\right] &= m(m-1)n(n-1)\left(\frac{(r(N-4)+2)}{(N-3)(N-2)}\frac{r}{(N-1)} - \left(\frac{r}{N-1}\right)^2\right) \\ &= \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2}. \end{aligned} \quad (14)$$

Combining terms yields

$$\begin{aligned} \operatorname{Var}[A_r] = \sum_{\substack{i\in S\\ j\in T}} \operatorname{Var}\left[X_{ij}\right] &+ \sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \operatorname{Cov}\left[X_{ij},X_{kj}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \operatorname{Cov}\left[X_{ij},X_{kl}\right] \\ &= \frac{mnr(N-1-r)}{(N-1)^2} - \frac{mnr(N-1-r)}{(N-1)^2} \\ &+ \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2} \\ &= \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2}. \end{aligned} \quad (15)$$

Therefore,

$$\mathrm{Var}[T_r] = \frac{1}{r^2}\mathrm{Var}[A_r] = \frac{2m(m-1)n(n-1)(N-1-r)}{r(N-3)(N-2)(N-1)^2} \ . \tag{16}$$

We note in particular that the first and second moment results in (7) and (16) match the results in Rosenbaum (2005) for the special case $r = 1$.

Simulation suggests that the null distribution of T_r is negatively skewed, but that for sufficiently large N and possibly certain conditions on r this distribution is asymptotically normal, independent of distribution function F. Rosenbaum (2005) proves that T_r is asymptotically normal for $r = 1$ for any distribution function; proof of this conjecture for $r > 1$ remains an open problem. This conjecture is supported by the normal QQ-plots shown in Figures 3 and 4 for 1,000 simulated values of $(T_r - E[T_r])/\sqrt{\mathrm{Var}(T_r)}$ at $r = 5$, 50 and $N = 150$, 600 with $m/n = 1/2$, under sampling from uniform distributions on $[-1, 1]^5$ and $[-1, 1]^{20}$. For the smaller sample size ($N = 150$), negative skewness is stronger for lower dimension and for higher r, with $r = 50$ and Dim = 5 being the most strongly skewed case shown. For the larger sample size ($N = 600$), skewness effects appear to vanish for all but the $r = 50$ and Dim = 5 case, and even in this case skewness is vastly diminished compared to the smaller sample size. Other distribution families and other values of m/n produce similar results.

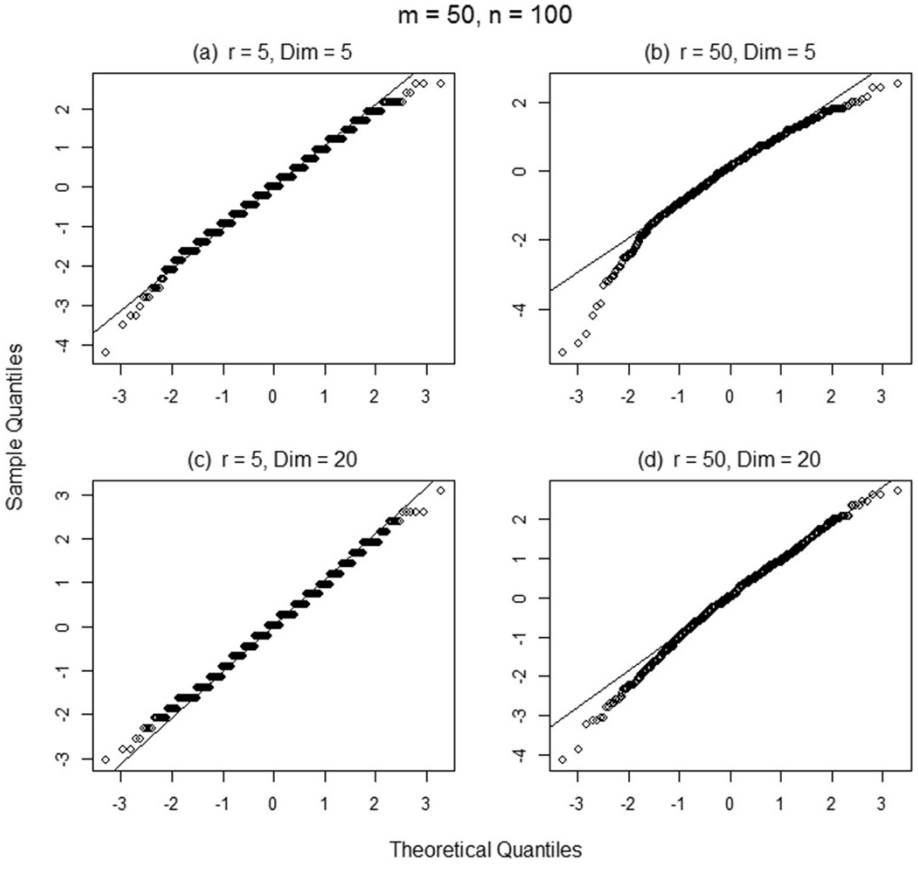

Figure 3 Normal QQ-plots of 1,000 simulated values of T_r for $m = 50$, $n = 100$ and $r = 5$, 50. Panels **(a)** and **(b)** are from independent samples of Unif $[-1, 1]^5$ variates. Panels **(c)** and **(d)** are from independent samples of Unif $[-1, 1]^{20}$ variates.

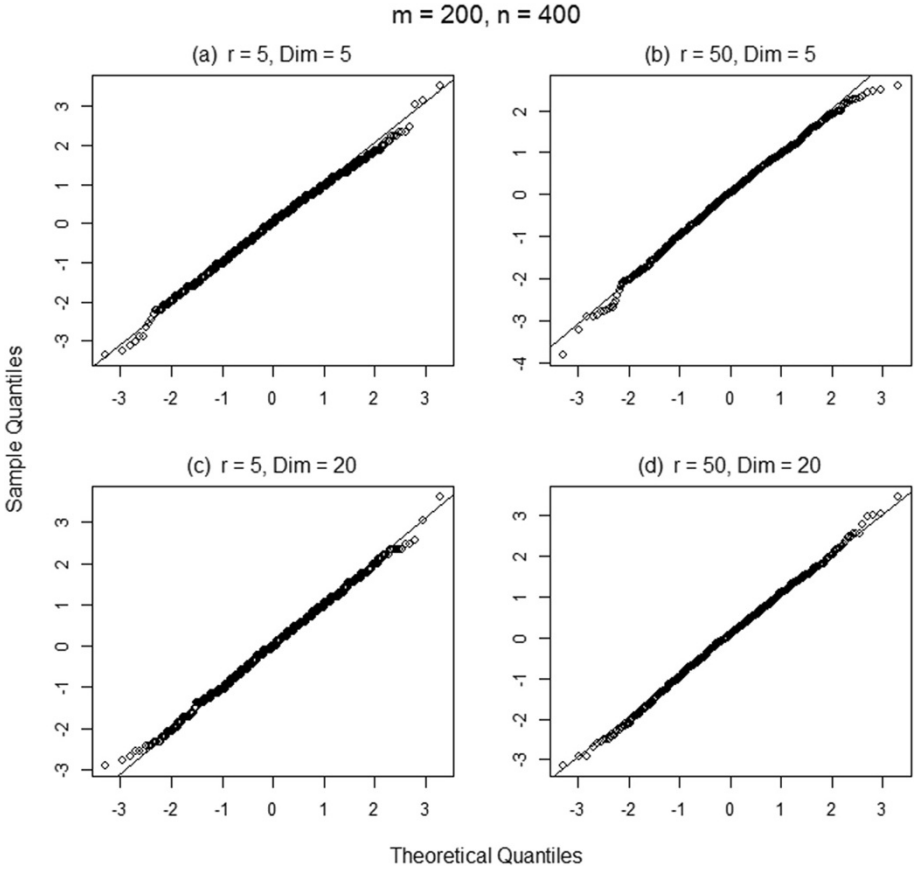

Figure 4 Normal QQ-plots of 1,000 simulated values of T_r for $m = 200$, $n = 400$ and $r = 5, 50$. Panels **(a)** and **(b)** are from independent samples of Unif $[-1, 1]^5$ variates. Panels **(c)** and **(d)** are from independent samples of Unif $[-1, 1]^{20}$ variates.

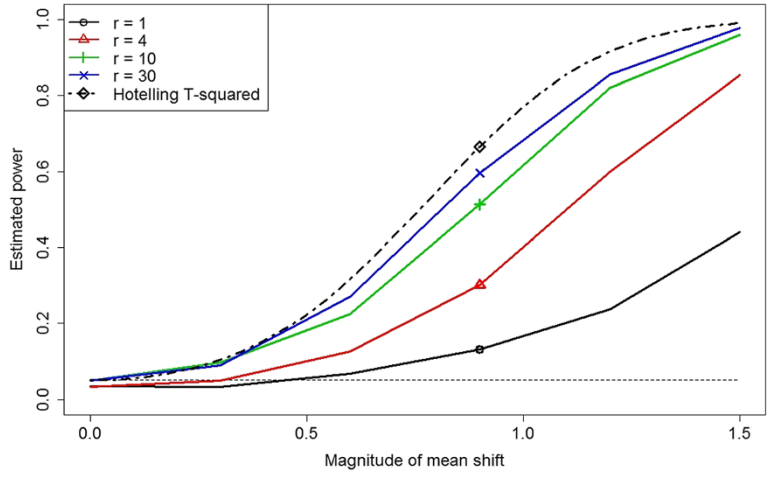

Figure 5 Power estimates for the MCC test at r = 1, 4, 10, and 30 and exact power for Hotelling's T^2 statistic for 5-variate normal mean alternatives with m = 20 and n = 40. The horizontal dashed line is test level α = .05.

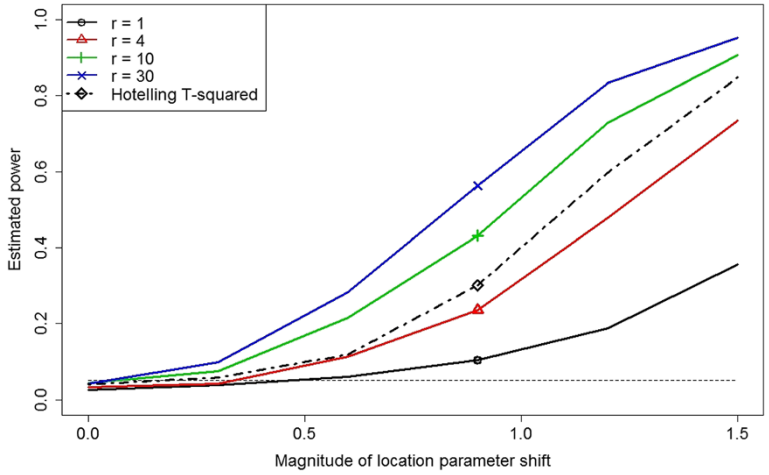

Figure 6 Power estimates for the MCC test at $r = 1, 4, 10$, and 30 and estimated power for Hotelling's T^2 statistic for 5-variate lognormal location parameter alternatives with $m = 20$ and $n = 40$. The horizontal dashed line is test level $\alpha = .05$.

Future work remains to bound rejection region probabilities in terms of sample size, dimension, and choice of r. In the absence of such theoretical bounds, for practical purposes a permutation test on observation indices serves as a suitable method to estimate p-values for the MCC test in cases where a normal approximation cannot be justified.

3.2 Small simulation study

We compare power characteristics of tests for two different location-shift scenarios. For each case, 1000 simulations are conducted for each shift in location parameter, group sizes are $m = 20$ and $n = 40$, and tests are conducted at significance level $\alpha = .05$. Distances are Euclidean. Estimated power is shown for MCC tests with $r = 1, 4, 10$, and 30 (where $r = 1$ is the cross-match test), and the performance of these tests is compared directly to that of Hotelling's T^2 test. Critical values for the MCC test were estimated through simulation for this study. All simulations were performed in R.

For the first example, the smaller group is drawn from a multivariate normal distribution with mean vector $\mathbf{0}$, identity covariance matrix, and dimension 5. The larger group is drawn from the same family, but the location vector of the second group is $\boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ ranges in magnitude from 0 to 1.5 by increments of 0.3. Hotelling's T^2 test is known to be the uniformly most powerful invariant test for location shift under these conditions (Bilodeau and Brenner 1999) and the exact power of the test is known for all location alternatives.

Figure 5 displays the estimated power results. We notice immediately that a modest increase of $r = 1$ to $r = 4$ substantially improves on the power of the cross-count test. As r continues to increase, MCC performance is even more impressive; the $r = 30 = N/2$ case performs nearly as well as Hotelling's T^2 test. Power estimates for cases $r > N/2$ are not shown. Not surprisingly, test power generally decreases as r increases beyond $N/2$ toward $N - 1$; in the extreme case T_{N-1} takes the fixed value $\frac{mn}{N-1}$ and hence the MCC test with $r = N - 1$ has power equal to zero against all alternatives.

For the second example, the first group is drawn from the multivariate log-normal distribution, where each of the 5 dimensions consists of independent, univariate log-normal draws with location parameter 0 and scale parameter 1. As before, the second group is drawn from the same family, but the location parameter vector for the second group is $\boldsymbol{\Delta}$, where the magnitude of $\boldsymbol{\Delta}$ ranges from 0 to 1.5 by increments of 0.3 and each dimension of $\boldsymbol{\Delta}$ takes equal value. The lognormal distribution is considered here to examine the effects of a skewed distribution on the tests in question. Since the underlying distributions are no longer multivariate normal, the power of Hotelling's T^2 test is estimated by simulation for this example.

Figure 6 displays the estimated power results. As before, we see that the power of the MCC test with $r = 4$ is much better than for $r = 1$. It is particularly noteworthy that for sufficiently large r the MCC test outperforms Hotelling's T^2 test.

4 Conclusions

The mean cross-count test is a powerful, non-parametric multivariate two-sample test that is applicable to any case where a notion of distance between observations exists. While this paper considers only location shifts, other simulations show that the MCC test has power in a variety of alternative cases as well. A shortcoming of the MCC test is that the null distribution for T_r is not simple (and perhaps not possible) to compute for all $r > 1$ and is not exactly distribution-free in these cases; in contrast, the test upon which it is based, the cross-match test with $r = 1$, has a known distribution that is independent of the distribution being tested.

It is known that T_1 is asymptotically normal, and while the mean and variance of T_r are derived herein and simulation suggests that the normal approximation for T_r is appropriate for sufficiently large N with $r > 1$, this property remains to be proven. This proof is part of ongoing work, as is sharpening the normal approximation via Edgeworth expansion based on higher moments of T_r. Likewise, finding useful criteria for choosing r is another area for future work. This choice is subject to competing factors: On the one hand, the power of T_r appears to improve as r increases to $N/2$ when group sizes are equal (i.e., $m = n = N/2$); therefore, $r = N/2$ seems a good choice for equal group sizes. On the other hand, the normal approximation appears to worsen as r increases; thus it may be desirable to restrict the size of r for this sake. Furthermore, an additional effect exists when group sizes are different. For example, assume $m < n$. If $r \geq m$, then at least one edge in $\mathcal{G}_r^*$ contains a vertex from each group and contributes to the cross-count, increasing the value of T_r. This is true even if the two groups are very different. Since a higher cross-count weakens the evidence against a group difference, this consideration suggests choosing $r < \min(m, n)$. A similar effect exists for multimodal distributions, suggesting that the size of r might be restricted as the number of modes grows. In any case, the best choice of r in practice clearly depends upon application specifics.

Competing interest
The author declares that he has no competing interests.

Received: 27 February 2014 Accepted: 26 August 2014
Published online: 04 November 2014

References

Anderson, I: Perfect matchings of a graph sufficient conditions for matchings. Proc Edinburgh Math Soc **18**, 129–136 (1972). Ser. 2
Bilodeau, M, Brenner, D: Theory of Multivariate Statistics. Springer, New York (1999)
Friedman, J, Rafsky, L: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Stat **7**, 697–717 (1979)
Friedman, J, Rafsky, L: Graphics for the multivariate two-sample problem. JASA **76**, 277–287 (1981)
Hall, P, Tajvidi, N: Permutation tests for equality of distributions in high-dimensional settings. Biometrika **89**, 359–374 (2002)
Henze, N: A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann Stat **16**, 772–783 (1988)
Maa, J, Pearl, D, Bartoszynski, R: Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. Ann Stat **24**, 1069–1074 (1996)
Rosenbaum, P: An exact distribution-free test comparing two multivariate distributions based on adjacency. JRSS B **67**, 515–530 (2005)
Ruth, D, Koyak, K: Nonparametric tests for homogeneity based on non-bipartite matching. JASA **106**, 1615–1625 (2011)
Schilling, M: Multivariate two-sample tests based on nearest neighbors. JASA **81**, 799–806 (1986)

doi:10.1186/s40488-014-0022-4
Cite this article as: Ruth: **A new multivariate two-sample test using regular minimum-weight spanning subgraphs.** *Journal of Statistical Distributions and Applications* 2014 **1**:22.