An Approach to the Multivariate Two-Sample Problem Using Classification and Regression Trees and Minimum-Weight Spanning Subgraphs

David M. Ruth[1], Samuel E. Buttrey[2], Lyn R. Whitaker[2]

[1]United Sates Naval Academy, 572-C Holloway Rd, Annapolis, MD 21402
[2]Naval Postgraduate School, 1411 Cunningham Road, Monterey, CA 93943-5219

Abstract

The multivariate two-sample problem is one of continued interest in statistics. Approaches to this problem normally require a dissimilarity measure on the observation sample space; such measures are typically restricted to numeric variables. In order to accommodate both categorical and numeric variables, we use a new dissimilarity measure based on a set of classification and regression trees. We briefly discuss this new measure and then incorporate it into a recently developed graph-based multivariate test. The test statistic counts the number of intergroup edges in a minimum-weight regular spanning subgraph; unequal distributions will tend to result in fewer edges in this count. Test performance is examined via simulation study, and test efficacy investigated using real-world data.

Key Words: Multivariate two-sample problem, graph-based test, classification and regression trees, nonparametric statistics

1. Introduction

Testing whether two samples can be considered as a random sample from a common distribution is a central problem in statistics. We consider the multivariate problem of testing for a difference between two groups when each group member has many measured attributes other than its group label. Ruth (2014) proposed the Mean Cross-Count (MCC) test as a graph-theoretic approach to the multivariate two-sample test: N observations are considered as vertices on a complete graph, interpoint differences are assigned as edge weights, and a test statistic is computed by counting the number of cross-group edges included in a minimum-weight regular subgraph of the complete graph. This method demonstrates impressive power when the vertex degrees in the minimum-weight subgraph are near $N/2$. The MCC test carries with it very few assumptions, but it does require the specification of a dissimilarity measure on the observation space.

While the context of a problem in some cases suggests a reasonable choice of dissimilarity measure, it may be unclear how best to specify what it means for two observations to be "close." Distance measures such as Euclidean, Mahalanobis, Manahattan, and others are widely used for data that include quantitative variables only. Several dissimilarity measures have been proposed for purely categorical variables; see Boriah, Chandola, and Kumar (2008) for a summary of many of these. Options are more limited for mixed data. This paper highlights a new approach to measuring dissimilarity on mixed data using

classification and regression tree (CART) analysis, introduced by Buttrey and Whitaker (2015). Using that measure, we apply the MCC test to detect a difference between two groups involving mixed data. Additionally, we present a useful known combinatorial optimization result which greatly simplifies some of the computational difficulties that can affect graph-theoretic tests such as MCC.

2. Background

2.1 Dissimilarity measures for mixed data

2.1.1 Gower dissimilarity

A typical standard of measure for dissimilarity in mixed data is Gower distance (Gower, 1971); this is the standard to which we will compare alternative methods. Consider N p-variate observations $X_1, X_2, \dots, X_N$ with possibly quantitative and categorical covariates. The dissimilarity $d_{ij,k}$ between observations X_i and X_j on covariate k is given by

$$d_{ij,k} = \begin{cases} 0 & \text{if covariate } k \text{ is categorical and } x_{ik} = x_{jk}\,, \\ 1 & \text{if covariate } k \text{ is categorical and } x_{ik} \neq x_{jk}\,, \\ \dfrac{|x_{ik} - x_{jk}|}{R_k} & \text{if covariate } k \text{ is quantitative,} \end{cases}$$

where x_{ik} and x_{jk} are the i and j entries, respectively, in the column associated with covariate k, and R_k is the range of covariate k. Gower's dissimilarity measure is a weighted average of these covariate-wise dissimilarities, given by

$$d_{\text{Gower}}(X_i, X_j) = \frac{\sum_{k=1}^{p} \partial_{ij,k} d_{ij,k}}{\sum_{k=1}^{p} \partial_{ij,k}},$$

where weights $\partial_{ij,k} = 1$ in except in special cases such as that of missing values. In some implementations, other weighting values may be chosen.

2.1.2 treeClust dissimilarity

Buttrey and Whitaker (2015) present a novel and robust approach to measuring dissimilarity in mixed data. For each covariate $k \in \{1, \dots, p\}$, construct a classification tree (if covariate k is categorical) or a regression tree (if covariate k is quantitative), modeling covariate k as the response variable and including all other covariates as predictor variables in the tree. Trees may be pruned to avoid overfitting, and those that are pruned back to the root are discarded. Call the remaining trees T_k, $k \in \{1, \dots, K;\ K \leq p\}$. Each observation gets assigned to one leaf in each tree. The key idea then is that *observations are considered dissimilar with respect to T_k when they fall in different leaves of T_k.*

Over the collection of all K resulting trees, a variety of options exist to measure the dissimilarity between observations. For example, let $I_k(i,j)$ be the indicator function that observations i and j fall in different leaves of T_k. Then a natural dissimilarity measure is

$$d_{\text{treeClust}}(X_i, X_j) = \frac{1}{K} \sum_{k=1}^{K} I_k(i,j)\ ;$$

that is, $d_{\text{treeClust}}(X_i, X_j)$ is the proportion of trees in which X_i and X_j fall in different leaves. Variations on this measure include weighting trees based on goodness of fit, pruning trees back to maximal trees in which X_i and X_j fall in the same leaf and then computing appropriate ratios of differences in deviances, and other approaches; see Buttrey and Whitaker (2015) for details. In particular, the treeClust approach demonstrates an appealing resistance to noise, a feature we will exploit later in this paper.

2.2 The Mean Cross-Count (MCC) test

2.2.1 Description

Graph-theoretic approaches offer robust and sometimes powerful tests for sample heterogeneity in the multivariate setting; see, for example, Friedman and Rafsky (1979), Rosenbaum (2005), Ruth and Koyak (2011), and Chen and Freidman (2016). Given a dissimilarity measure, such approaches consider observations as vertices in a graph, with an edge between each pair of points weighted by interpoint dissimilarity. An optimal subgraph is computed, where optimality is determined by including only certain low-weight edges. The edges in the optimal subgraph contain information regarding whether two samples are drawn from different populations.

The MCC test is one of this type. Consider $N = n + m$ independent observations $X_1, X_2, \dots, X_m$ and $X_{m+1}, X_{m+2}, \dots, X_{m+n}$ where each X_i is assumed to be drawn from distribution F for $1 \le i \le m$ and from distribution G for $m + 1 \le i \le N$. The goal is to test the null hypothesis $F = G$. Given some interpoint dissimilarity measure d, let $\mathcal{G}$ be the complete graph with each observation X_i constituting a vertex and each pair of observations (X_i, X_j) constituting an undirected edge with weight $d(X_i, X_j)$. We assume N is even for this discussion, but odd N can be easily accommodated. Pick an integer $r \in \{1, \dots, N/2\}$ and find a minimum-weight r-regular spanning subgraph, $\mathcal{G}_r^*$. Note that $\mathcal{G}_r^*$ does not depend on group labels. Count the number of edges in $\mathcal{G}_r^*$ that have one vertex in the first group and the other in the second group; call this count A_r. The Mean Cross Count statistic is defined as $T_r = \frac{A_r}{r}$. Since unequal distributions will tend to result in fewer edges that connect vertices between different groups, T_r will tend to be small when $F \neq G$. Approximate p-values for T_r may be computed easily using a permutation test on the observation group labels. The MCC test has been shown to have impressive power over a broad range of alternatives; see Ruth (2014) for details.

2.2.2 Illustrating example

Figure 1a shows 20 bivariate iid observations plotted along with the minimum-weight 3-regular spanning subgraph with respect to Euclidean distance. The number of edges in $\mathcal{G}_3^*$ connecting observations in Group 1 to those in Group 2 is 20, so the MCC statistic is $T_3 = \frac{20}{3} \approx 6.67$. Figure 1b shows the same situation, except Group 1 is shifted by one unit in both covariates. A new $\mathcal{G}_3^*$ results, and in this case the cross count is reduced to 10 and $T_3 = \frac{10}{3} \approx 3.33$. This demonstrates the intuitive notion that the mean cross count goes down as group locations move apart. We note here that the MCC test is not confined to location alternatives but it is less effective against scale alternatives, particularly in high dimension. See Chen and Freidman (2016) for an attractive graph-theoretic approach to accommodate scale alternatives.

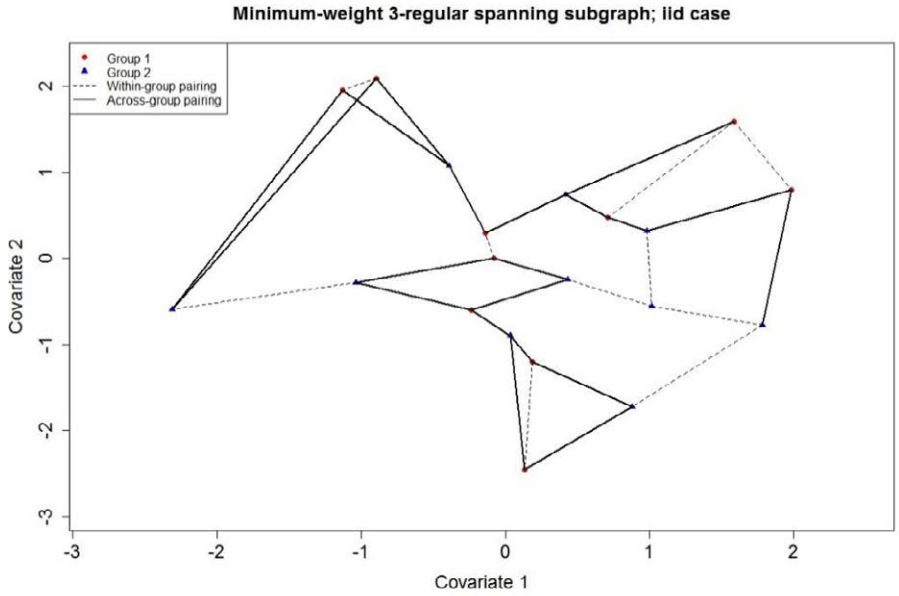

Figure 1a: Twenty bivariate iid observations consisting of two groups of equal size. The minimum-weight 3-regular spanning subgraph with respect to Euclidean distance shows across-group pairings as solid lines and within-group pairings as dashed lines. The cross count is $A_3 = 20$, so the MCC test statistic is $T_r \approx 6.67$.

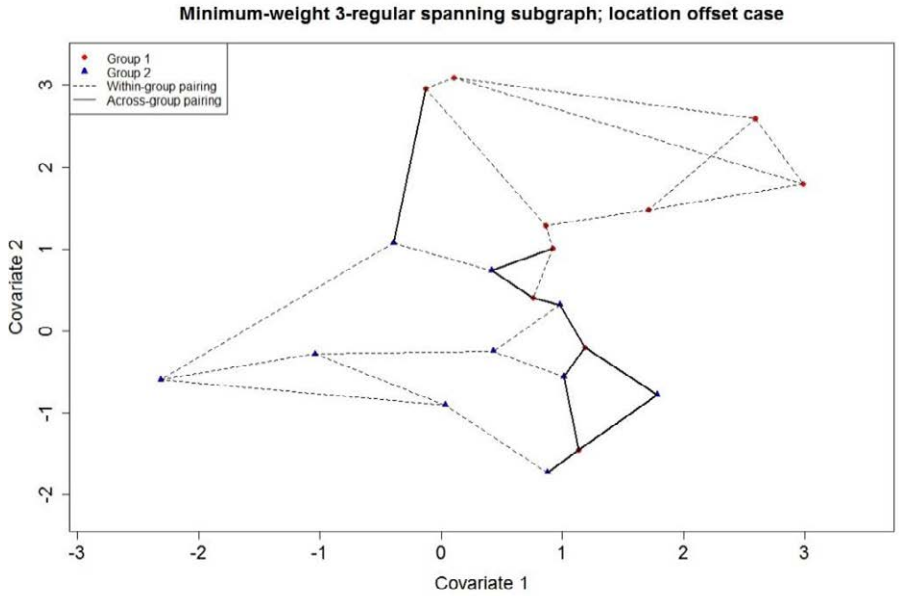

Figure 1b: Twenty bivariate observations with a location offset between two groups of equal size. The minimum-weight 3-regular spanning subgraph with respect to Euclidean distance shows across-group pairings as solid lines and within-group pairings as dashed lines. The cross count is $A_3 = 10$, so the MCC test statistic is $T_r \approx 3.33$.

3. Analysis and Performance

This section examines the performance of the MCC test in a mixed data setting employing the Gower dissimilarity measure compared to that with the treeClust measure.

3.1 Heart data

3.1.1 Description

The data used for this study are the Cleveland Heart Disease Data, consisting of test results of 303 patients undergoing angiography at the Cleveland Clinic in Ohio. The data include 76 attributes, but we use a subset of 14 of them which have been analyzed many times in published experiments using these data (Blake and Merz, 1998). We group observations by angiographic disease status (binary); the other five quantitative and eight categorical explanatory variables are used to compute interpoint dissimilarities. Table 1 shows the values of the response variables for six of the 303 observations.

Table 1: Leading rows of response variables for Cleveland Heart Disease Data. *Age, trestbps, chol, thalach,* and *oldpeak* are quantitative; all others are categorical.

age	*sex*	*cp*	*trestbps*	*chol*	*fbs*	*restecg*	*thalach*	*exang*	*oldpeak*	*slope*	*ca*	*thal*
63	1	1	145	233	1	2	150	0	2.3	3	0	6
67	1	4	160	286	0	2	108	1	1.5	2	3	3
67	1	4	120	229	0	2	129	1	2.6	2	2	7
37	1	3	130	250	0	0	187	0	3.5	3	0	3
41	0	2	130	204	0	2	172	0	1.4	1	0	3
56	1	2	120	236	0	0	178	0	0.8	1	0	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A first question of interest is: *Are the two diagnosis groups statistically different with respect to the explanatory variables?* For these data in raw form, the two groups are not very difficult to differentiate; in fact, a univariate two-sample t-test on *trestbps* (resting blood pressure) is sufficient to detect a difference between groups at significance level < 0.01. In the interest of detecting a difference when heterogeneity is more subtle, we add noise to the data as described below and seek to answer a second question: *Can a statistical difference be found between the two diagnosis groups with respect to the explanatory variables in the presence of noise?*

3.1.2 Modifications

In order to make group difference detection more difficult, we modify the data in the following ways:

1) For the first analysis, we permute some fraction of the diagnosis labels and then examine estimated power of the MCC test for different permutation fractions.
2) For the second analysis, we make the original explanatory variables "noisy" by randomly permuting each column 20 different times and then appending the 20 permuted columns to the original data set. This adds 260 columns of noise, where each added column has the same marginal distribution as one of the original data columns. Then we permute some fraction of the diagnosis labels and examine estimated power of the MCC test for different permutation fractions, as in (1) above.

3.2 Performance comparison

3.2.1 No noise added to explanatory variables

In this case, we ran 1000 simulations under the following conditions: Shuffle some fraction, λ, of the diagnosis labels, then apply the MCC test at significance level $\alpha = 0.05$ for dissimilarity measure d_{Gower} and then $d_{\text{treeClust}}$.[1] Perform these simulations for values of λ decreasing from 1 down to 0.5. Figure 2a shows the estimated power for these simulations plotted against $1 - \lambda$, that is, against the fraction of unshuffled labels. When all labels are shuffled estimated test power is approximately equal to α, which confirms that both tests respect significance level. The MCC test performs similarly under each dissimilarity measure, with d_{Gower} outperforming $d_{\text{treeClust}}$ by about 7% in the mid-power range (where approximately 75% of diagnosis labels are shuffled).

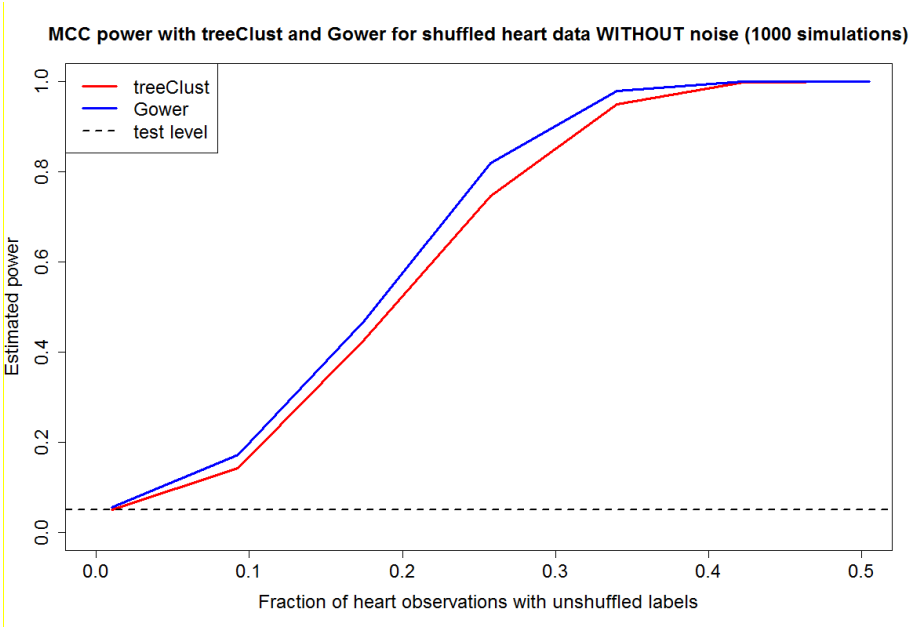

Figure 2a: Estimated power for the MCC test at significance level $\alpha = 0.05$ as a function of the fraction of diagnosis labels unshuffled, without noise columns added. Gower dissimilarity is the top curve (blue); treeClust dissimilarity is the bottom curve (red). Test significance level 0.05 is the horizontal line (dashed). Zero unshuffled labels means all diagnosis labels are assigned randomly.

3.2.2 Noise added to explanatory variables

In this case, before finding optimal subgraphs, we add noise to the data as described in Section 3.1.2, and then estimate power for varying fractions of shuffled labels as in the no-noise case. Figure 2b shows both tests respect test level as before; however, in this case $d_{\text{treeClust}}$ solidly outperforms d_{Gower}. In the mid-power range (with approximately 70% of diagnosis labels shuffled), estimated MCC test power under $d_{\text{treeClust}}$ exceeds power under d_{Gower} by about 35%. Additionally, comparing Figure 2b to Figure 2a we note that MCC test power under $d_{\text{treeClust}}$ suffers little degradation in the noise case relative to the no-noise case while under d_{Gower} the power degradation is fairly severe.

[1] Dissimilarities were computed using R packages "cluster" (Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2015) for d_{Gower} and "treeClust" (Buttrey, 2016) for $d_{\text{treeClust}}$. MCC tests were performed using R package "AcrossTic" (Ruth and Buttrey, 2016).

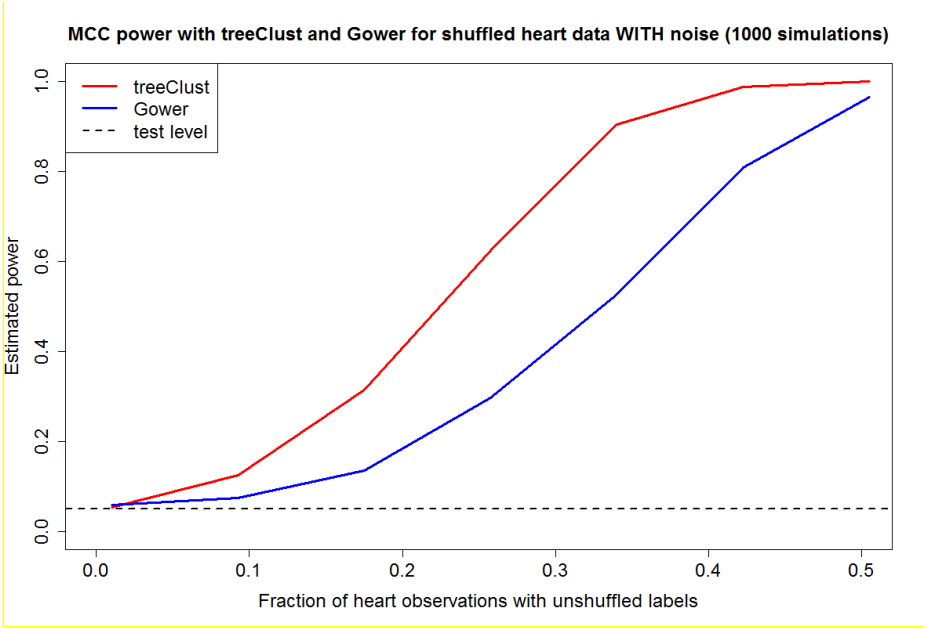

Figure 2b: Estimated power for the MCC test at significance level $\alpha = 0.05$ as a function of the fraction of diagnosis labels unshuffled, with noise columns added. Gower dissimilarity is the bottom curve (blue); treeClust dissimilarity is the top curve (red). Test significance level 0.05 is the horizontal line (dashed). Zero unshuffled labels means all diagnosis labels are assigned randomly.

3.3 Computational efficiencies

Minimum-weight regular spanning subgraphs of a complete graph can found directly using binary integer linear programming. In general, solutions may take a long time to obtain using readily-available solvers in cases of large data. However, this particular problem falls into a special class of combinatorial optimization problems called "b-matchings." A formulation of this problem is

$$\min_{\mathbf{y}} \sum_{i<j} d(X_i, X_j) y_{ij}$$

subject to

$$\sum_{j=1}^{i-1} y_{ji} + \sum_{j=i+1}^{N} y_{ij} = b_i \quad \forall i \in \{1, \dots, N\},$$

$$y_{ij} \in \{0,1\} \ \forall i, j \in \{1, \dots, N\}, i < j,$$

where $y_{ij} = 1$ if the edge connecting X_i and X_j is in the optimal subgraph and $y_{ij} = 0$ if it is not, and b_i is the degree of vertex i in the optimal subgraph. In our case, $b_i = r \ \forall i$. An appealing known result is that when the condition $y_{ij} \in \{0,1\}$ is relaxed to $y_{ij} \in [0,1]$, the optimal values of y_{ij} must always be 0, $\frac{1}{2}$, or 1 when r is an integer and in fact 0 or 1 when r is even (for example, see Hirai, 2013). So, *we can remove the integrality constraint in this linear program and compute the MCC test statistic for bigger data sets using readily available solvers.*

4. Conclusions

Graph-theoretic methods can be effective at detecting differences between two groups in a multivariate setting. Such methods generally require a dissimilarity measure, and tree clustering provides this in a novel way that is useful for mixed data, particularly when noise is present. For the MCC test, which employs minimum-weight regular spanning subgraphs, combinatorial optimization integrality properties improve the computational speed of pairing algorithms. Packages "treeClust" and "AcrossTic" enable R users to put these tools into practice.

Acknowledgements

Thanks to Professor David Phillips at the US Naval Academy for his insight regarding fractional *b*-matchings and to Dr. Lawrence Rafsky at Acquire Media for his pioneering work in this field and helpful feedback.

References

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases, University of California, Irvine, Department of Information and Computer Sciences.

Boriah, S., Chandola, V., and Kumar, V. (2008). "Similarity Measures for Categorical Data: A Comparative Evaluation." Presented at SIAM Conference on Data Mining, Atlanta, GA, 2008.

Buttrey, S. and Whitaker, L. (2015). "treeClust: An R Package for Tree-Based Clustering Dissimilarities," *R Journal*, Vol. 7, No. 2, pp. 227–236.

Buttrey, S. (2016). treeClust: Cluster Distances Through Trees, R package version 1.1-6.

Chen, H. and Friedman, J. (2016). "A New Graph-Based Two-Sample Test for Multivariate and Object Data," *Journal of the American Statistical Association*, accepted author version posted online: 11 Feb 2016, retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.2016.1147356.

Friedman, J. and Rafsky, L. (1979). "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics,* Vol. 7, No. 4, pp. 697–717.

Gower, J. (1971). "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, Vol. 27, No. 4, pp. 857–871.

Hirai, H. (2013). "Half-Integrality of Node-Capacitated Multiflows and Tree-Shaped Facility Locations on Trees," *Mathematical Programming Series A,* Vol 137, pp. 503–530.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2015). cluster: Cluster Analysis Basics and Extensions, R package version 2.0.3.

R Core Team (2015). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rosenbaum, P. (2005). "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society Series B*, Vol. 67, No. 4, pp. 515–530.

Ruth, D. and Koyak, R. (2011). "Nonparametric Tests for Homogeneity Based on Non-Bipartite Matching," *Journal of the American Statistical Association*, Vol. 106, No. 496, pp. 1615–1625.

Ruth, D. (2014). "A New Multivariate Two-Sample Test Using Regular Minimum-Weight Spanning Subgraphs," *Journal of Statistical Distributions and Applications*, 1:22.

Ruth, D. and Buttrey, S. (2016). AcrossTic: A Cost-Minimal Regular Spanning Subgraph with TreeClust. R package version 1.0-3.