A Distribution Free Adjacency Based Test for the Similarity of Spatial Locations of Ribosomal Proteins in Two Groups

REU Site: Interdisciplinary Program in High Performance Computing

Nil Mistry[1], Jordan Ramsey[2], Benjamin Wiley[3], and Jackie Yanchuck[4],
Graduate RAs: Xuan Huang[5] and Andrew Raim[5],
Faculty Mentors: Matthias K. Gobbert[5] and Nagaraj K. Neerchal[5],
Client: Philip J. Farabaugh[6]

[1]Department of Mathematics and Statistics, University of Connecticut
[2]Department of Computer Science and Electrical Engineering, UMBC
[3]Department of Mathematics and Statistics, University of New Mexico
[4]Department of Mathematics, Seton Hill University
[5]Department of Mathematics and Statistics, UMBC
[6]Department of Biological Sciences, UMBC

Abstract

It is of interest to test if spatial locations of ribosomal proteins are related to the occurrence of phenotypes. We apply an adjacency based test statistic to investigate this question in a specific data set. The Mahalanobis distance is computed between each pair of protein locations, and the optimal pairing is determined by minimizing the sum of the within-pair distances. We created a code that allows a user to compute Mahalanobis distances, to determine the optimal pairing, and to test whether the two groups are statistically different. The user can also compute an exact p-value for this distribution rather than rely on an approximation. Our Codes also produce useful graphics to help understand and explain the data and results.

1 Introduction

Motivation for the methodology presented in this paper came from a problem presented to us by Dr. Philip Farabaugh, Professor of Biological Sciences at UMBC. His experimental data is a set of three dimensional coordinates for RNA proteins contained within a ribosome. Each protein is either phenotype-related or not. About 80 proteins exist within the structure of the ribosome, each assuming its own unique position identified by a three dimensional vector of coordinates. Certain proteins are carriers of a particular phenotype, while others are not and are identified in this study as non-phenotype proteins. Phenotypes are physical manifestations of a particular characteristic that results from a particular genotype and its relationship with the surrounding environment. The two sets of ribosomal proteins are depicted in Figure 1.1: one containing phenotype traits (p) and another containing non-phenotype (n) traits. The objective is to compare the distributions of the spatial locations of these two sets of points. More specifically, are the proteins corresponding to the phenotype group sufficiently separated away from the other group. Dissimilar distributions between categories could indicate dissimilar traits between phenotype and non-phenotype categories.

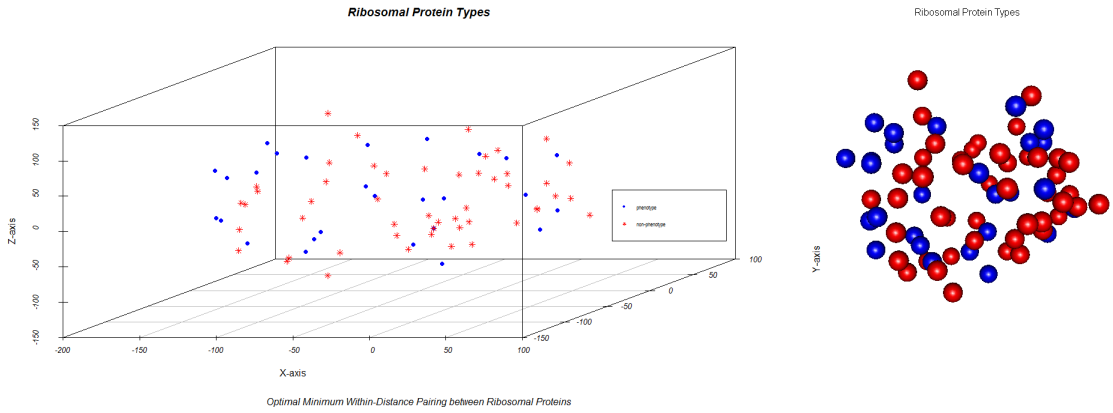

Figure 1.1: Ribosomal protein types.

Figure 1.2: Two-dimensional comparison.

To further examine the scatter cloud of Figure 1.1, we consider the data in all possible pairs of two dimensions at a time as shown in Figure 1.2. This figure shows the a scatter plot for each view of two variables at a time, a histogram and a box plot for each comparison between any two axes. Note that visual comparison of the graphics (histogram or boxplot) for the p group vs the corresponding graphics for the n group seem to suggest that the two groups are not significantly different. The objective is to extend this comparison and obtain a formal conclusion that takes into account of all three dimensions simultaneously. [In other words, the client wants one p-value – not three p-values!]

We will be focusing on two key aspects of this problem. First, we want to make minimal parametric assumptions (such as Normality) about the distribution of the spatial locations of the proteins. Second, we are looking for a generalization of the univariate nonparametric test statistic which are usually based on adjacency of the data points. While several methods

exist under this approach, [6] has introduced an exact, distribution-free test which compares multivariate distributions based on adjacency.

The test statistic proposed by [6] is based on the idea that if the two sets of points are identically distributed then the closest neighbor of each point is equally likely to belong to either phenotype or non-phenotype group. Thus, we would choose a distance measure to determine closeness and determine an optimal algorithm to divide the combined (p's and n's) dataset into pairs. Then the test statistic is the number of pairs containing one point from each group. Clearly, if the two groups are well-separated this statistic will be small and if the two groups are well-mixed then it will be large. [6] recommends that we use Mahalanobis distance (to be introduced later) for measuring distance, and derives the distribution of this test statistic under the null hypothesis that the protein locations of the two groups are random samples drawn from the same distribution. He also provides a normal approximation that is valid when the number of points is large.

While the test statistic is easily explained (and hopefully understood), it turns out that its computation involves a rather challenging combinatorial optimization. [6] uses a non-bipartite matching algorithm (using an already existing C code) to accomplish this. The main contribution of our project is to create a single R function which takes a typical data set as input, computes the Mahalanobis distance matrix, passes the data into the C code in appropriate formatting, and outputs the test statistic and associated p-value. The function also produces insightful graphics that the researchers can use in explaining the results and include in their scientific publications.

2 Background

Wilcoxon Rank Sum (WRS) tests and Wilcoxon Signed Rank Test are two of the most commonly used test statistics for comparing univariate distributions. As suggested by their names, these are based on ranks of data points from the two groups under comparison. For example, WRS test statistic is the sum of the ranks of members of the first group in the combined ranking of members of both groups. Thus, the univariate test statistic rely on our ability to totally order all the data, and therefore do not generalize directly to multivariate observations. Furthermore, the probability distribution of WRS under the null hypothesis that the two groups are random samples from the same population, does not depend on specific parametric distributional assumption on the population. It turns out that this "distribution-free" property also does not convey for some straightforward generalizations.

First, we note that total ordering is not possible in multivariate data. For example, while a single variable from two different sets, such as length, may be relatively simple to order on a scale, one may have more difficulty comparing multiple variables, such as age and height, from two different sets. That is, while 6 cm is longer than 5 cm, how may one compare 6 cm north and 4 cm west with respect to a center point, in relation to another point which is 5 cm south and 3 cm east? Comparing entire data sets with respect to each other might provide a better approach. These adjacency based "neighbor count" are not necessarily distribution free. Friedman and Rafsky [1] implemented pairwise distances to construct a

minimum spanning tree, and removed edges in the tree that connected the two different groups. Accordingly, the resulting number of remaining disjoint sub-trees may be utilized as a test statistic. Another test created by Schilling [8] and Henze [3] paired up data using the nearest neighbor to each subject, and then counted the number of times that subjects from the same group were paired together. However, these tests are not distribution-free [6]. To be distribution free, the distribution of the test statistic should be a known distribution that depends on the sample size, but not on the distributions of the input data.

Rosenbaum [6] developed a test statistic based on the the Mahalanobis distance to analyze multi-dimensional data and showed that it is distribution-free. The N data points are broken down into $\frac{N}{2}$ pairs, optimally, to minimize the sum of the within-pair distances. Each pair must contain either zero, one, or two points from the first group, and a test statistic referring to the equality of distributions is produced from the number of pairs with exactly one data point from the first group (denoted by A_1).

The ideal pairing is produced from optimal, non-bipartite matching. That is, the sum of the within-pair distances must be minimized, and to accomplish this, each distance must be considered in relation to the others. The number of arithmetic operations required to derive the Rosenbaum test statistic is $O(N^3)$ [6]. For example, Papadimitriou and Steiglitz [5] provides several various algorithms, such as the Hungarian Method and the Weighted Matching Algorithm. In this paper, implementation of optimal, non-bipartite matching is obtained from E. Rothberg's C algorithm [7].

3 Numerical Methods

3.1 Mahalanobis Distance

The Mahalanobis distance is obtained by normalizing the Euclidean distance with respect to the an estimated variance-covariance matrix. In our case, we consider two groups of RNA proteins: one containing phenotype traits and the other containing non-phenotype traits. We shall therefore use the pooled estimate of the variance-covariance matrix. Below, we provide the formula for this estimate.

Let the number of observations in the two groups be denoted by n_1 and n_2. Let $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijm})^T$ denote the j^{th} vector of measurements from the i^{th} group, $j = 1, 2, \cdots, n_i; i = 1, 2$. Let $\bar{X}_i = (\bar{X}_{i1}, \bar{X}_{i2}, \cdots, \bar{X}_{im})^T$ and S_i denote the vector of mean responses and the m-by-m variance-covariance matrix for the i^{th} group, respectively. That is, for $i = 1, 2$,

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \text{ and } S_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T.$$

The pooled variance-covariance matrix is defined as

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.1}$$

Then the Mahalanobis distance between any two observations is defined as

$$D^2 = (X_{ij} - X_{i'j'})^T S^{-1} (X_{ij} - X_{i'j'}). \tag{3.2}$$

If the covariance-variance matrix is replaced with the identity matrix, the above equation reduces to the Euclidean norm. Consequently, one may imagine the Mahalanobis distance as similar to the Euclidean distance only it also takes into account for the differences in variances among measurements and their mutual dependencies.

3.2 Distance Computation

R code was developed that takes the length of the data and ranks each vector of data being analyzed. Ranks are computed for the x-, y-, and z-coordinates individually. In the case of ties, we take the minimum of all ranks included in the tie. For example, if there was a tie for first, second, and third place, then all elements included in the tie would receive a rank equal to one. The covariance matrix, S, is computed for these ranks. Since we know the ranks and covariance matrix, we use the Mahalanobis distance for every combination between pairs within the matrix. The computed data is output to a text file in order to read it into C code.

3.3 Non-Bipartite Matching

The C code is an optimal non-bipartite combinatorial optimization matching sorting algorithm. The optimal pairing is determined by minimizing the sum of the within-pair distances between ribosomal proteins. In this case, there are $\binom{N}{2}$ total possible pairings, where $N(= 76)$ is the total number of proteins. The minimum distance that is calculated divides 76 proteins into 38 non overlapping pairs, and minimizes the sum of the 38 within-pair distances. We develop R code to match protein pairs, based on phenotype and non-phenotype matches. Based on the proportion of phenotype and non-phenotype proteins within the data set using the number of non-matching pairs we derive p-value (statistical significance) for the comparison of distribution with respect to the distances.

The optimal non-bipartite matching algorithm is written in C by Ed Rothberg [7]. The code implements H. Gabow's N-cubed weighting matching algorithm [2]. The algorithm maximizes the sum of benefits, benefit defined as β_{ij},

$$\beta_{ij} = \max_{b,c}(\delta_{bc}) - \delta_{ij}. \tag{3.3}$$

where δ_{ij} are the distances, which is clearly equivalent to minimizing sum of the distances as required by [6].

3.4 Significant Figures

The optimal, non-bipartite algorithm implemented in C accepts only integers as input values for distances between two points. If the Mahalanobis distance vector entered into the code

contains decimal places, the C function will remove the additional figures in the conversion from double precision floating-point numbers to integers. As a result, the optimal number of non-matching pairs may be inaccurate due to the potential of failing to account for significant figures contained within the distance sums. One solution to avoiding this error is multiplying the Mahalanobis distance vector by a large power of 10, thereby ensuring that the conversion to integer format encompasses these significant figures that were not previously considered in the former number conversion process. The R function created for this project allows the user to specify the appropriate power of 10.

3.5 Null Distribution

The following results are given in [6]. The null distribution for A_1, the total number of cross-matched pairs, is given by

$$Pr(A_1 = a_1) = \frac{2^{a_1}\, I!}{\binom{N}{N/2}\, a_0!\, a_1!\, a_2!}, \tag{3.4}$$

where a_k is defined as the number of pairs with k subjects from group 1 (phyontypes), and $a_0 + a_1 + a_2 = N/2$ is the number of total pairs for the entire set of N data points, and n is the number of cofactor phenotype ribosomal proteins [6].

As proven in [6], the null distribution for A_1 converges to the Normal distribution:

$$z = \frac{A_1 - E(A_1)}{\sqrt{Var(A_1)}}. \tag{3.5}$$

where the $E(A_1)$ and $Var(A_1)$ are given in [6].

While it has been shown that the distribution of the cross-match statistic converges to the normal distribution, the exact probability is quite easily computed. Note, however, that the computation of the value of the test statistic from the data itself is quite difficult, because an integer programming over all possible $(\frac{N}{2})!$ non-overlapping pairs is involved. We are using an already published code wrapped in an R program. Also, included in the R program are several steps of data handling as well as preliminary computations such as the distance matrix. Therefore as additional check, we simulated the null distribution using our program and compared to the exact theoretical result given above. This is done by randomly assigning each each data point to a particular group. We then ran our cross-match tests in order to see what the distribution of the cross-matches would be like if the distribution was random. We repeated this process 5,000 times in order to see how many cross-matches would occur.

4 Results

After computing the Mahalanobis distance between all possible pairs of ribosomal proteins, the optimal, non-bipartite sorting algorithm determined that there are 17 non-matching

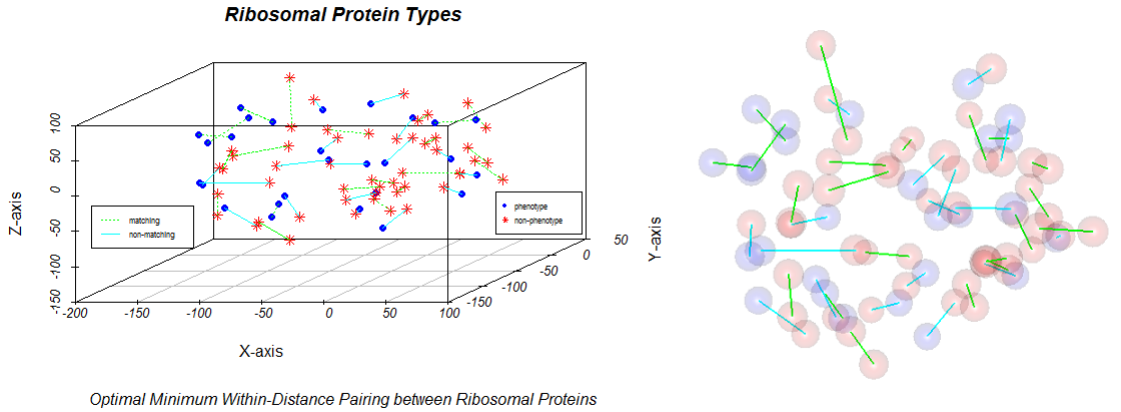

Figure 4.1: Optimal within-distance pairing between ribosomal protein types, all matching.

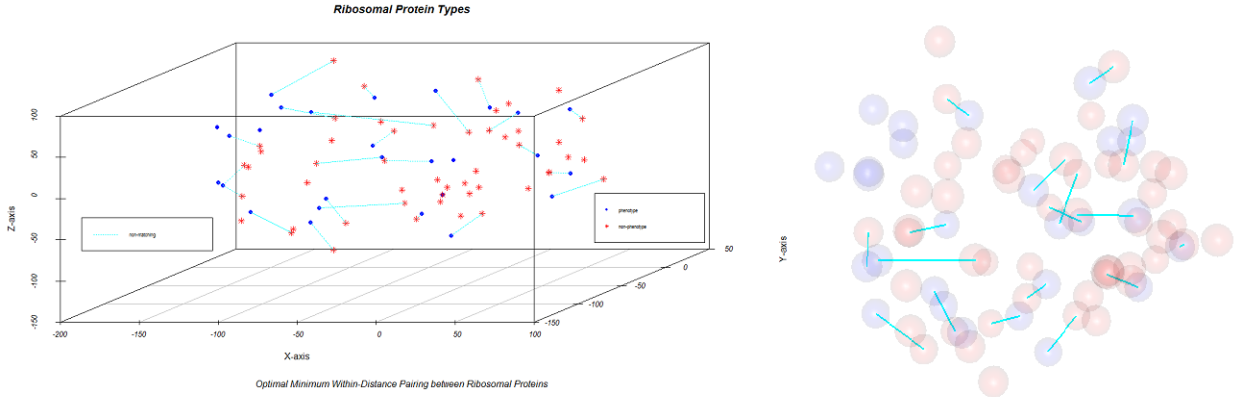

Figure 4.2: Optimal within-distance pairing between ribosomal protein types, non-matching.

pairs in the closest within-pair distance pairing. Figure 4.1 shows the optimal connections between all pairs, where the cyan lines represent the connections between phenotype and non-phenotype proteins. The maximum number of non-matching pairs is 38(= 76/2). In effect, roughly 45% of optimal pairings are non-matching.

Alternately, to emphasize the non-matched pairs between the two categories, Figure 4.2 displays the optimal within-distance pairing while only showing the connections between phenotypes to non-phenotypes. In this computation, the number of decimal digits in the distance computations were specified as 3, which is equivalent to multiplying the Mahalanobis distance vector by 10^3. The same conclusion was also obtained for 10^5, and therefore 10^3 is large enough to offset the rounding that occurs due to truncation by the C code.

The null hypothesis is that the distribution of three-dimensional positioning is the same for phenotype and non-phenotype proteins. Computing a p-value using the Normal approximation for the null distribution of A_1 given in equation (3.6) of Section 3.5, we have that $p - value = 0.821447$. Consequently, we fail to reject the null hypothesis and therefore we

Table 4.1: Summary statistics for non-matching pairs.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	15.00	17.00	17.61	19.00	25.00

may conclude that the distributions of three-dimensional positions are statistically similar for the two groups.

Determining whether substantial variance is present within the two groups, the Euclidean distance is computed in place of the Mahalanobis distance and the deviation from the above result is noted. Using the 3×3 identity matrix rather than the variance-covariance matrix, the 2-norm is determined for all possible within-pair distances and optimal, non-bipartite matching is utilized to obtain the minimum within-distance sum. The number of non-matching pairs is equal to 21, which is slightly greater than the above Mahalanobis result. Computing a p-value from these pairings, it follows that $p = 0.236081$. Again, the null hypothesis has failed to be rejected at an alpha value of 0.05, and therefore the distributions of three-dimensional positions are still statistically similar even not taking into account inter-group variance. In this case, the number of significant digits is expressed to four significant figures in the integer conversion.

Finally, performing a permutation test in which randomly assigned data in the set are identified as either phenotype or non-phenotype, consistent with the proportion present in the original set, the number of optimal non-matching pairs was computed at each iteration. This process was repeated $5,000$ times to note the distribution of non-matching pairs from each instance. Table 4.1 shows the summary statistics of these non-matching pairs in this experiment, and one should confidently notice that the above number, 17, is well within the middle range.

Figure 4.3 is the histogram for our permutation test, which allowed us to calculate a more exact p-value, with a violin and box plot to reinforce the above numerical conclusion.

5 Conclusions

After computing the Mahalanobis distance between each protein pair within the data set, it was determined by the optimal, non-bipartite sorting algorithm that there are 17 non-matching pairs of ribosomal proteins matching phenotype to non-phenotype traits. Since the data set contains 76 elements, the maximum possible non-matching pairs is $76/2 = 38$ pairs, and therefore one should notice the high proportion of non-matching pairs $A_1 = 17$. This value of the test statistic corresponds to a p-value equal to 0.821447. That is, we fail to reject the null hypothesis (at the conventional $\alpha = 0.05$), and conclude that the distributions comparing adjacency between phenotype and non-phenotype groups are statistically equivalent.

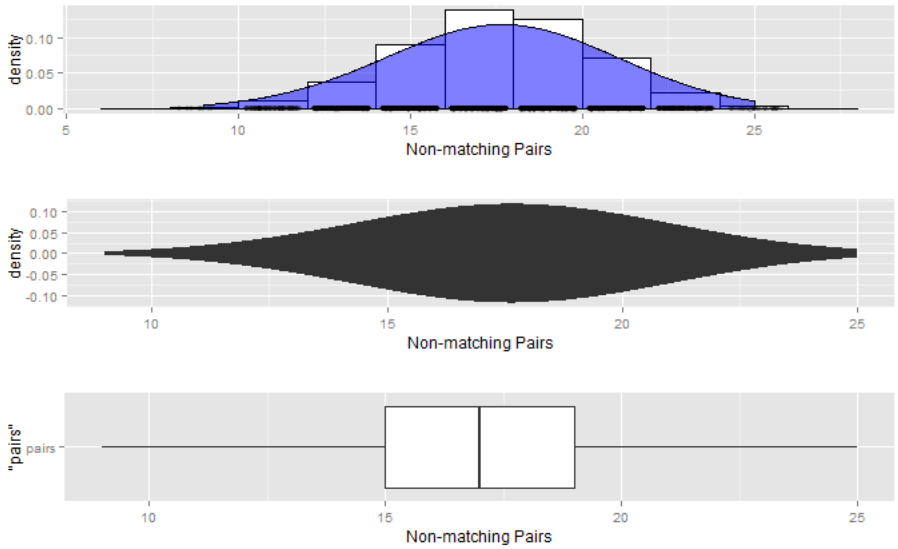

Figure 4.3: Histogram for non-matching pairs overlaid with kernel density curve, with violin and box plots.

Acknowledgments

These results were obtained as part of the REU Site: Interdisciplinary Program in High Performance Computing (`www.umbc.edu/hpcreu`) in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County (UMBC) in Summer 2013, where they were originally reported in the tech. rep. [4]. This program is funded jointly by the National Science Foundation and the National Security Agency (NSF grant no. DMS–1156976), with additional support from UMBC, the Department of Mathematics and Statistics, the Center for Interdisciplinary Research and Consulting (CIRC), and the UMBC High Performance Computing Facility (HPCF). HPCF (`www.umbc.edu/hpcf`) is supported by the National Science Foundation through the MRI program (grant nos. CNS–0821258 and CNS–1228778) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from UMBC. Co-author Jordan Ramsey was supported, in part, by the UMBC National Security Agency (NSA) Scholars Program though a contract with the NSA. Graduate RAs Xuan Huang and Andrew Raim were supported by UMBC as HPCF RAs.

References

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

[2] H. Gabow, *Implementation of algorithms for maximum matching on nonbipartite graphs*. Ph.D. thesis, Stanford University, 1973.

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

[4] N. Mistry, J. Ramsey, B. Wiley, J. Yanchuck, X. Huang, A. Raim, M. K. Gobbert, N. K. Neerchal, and P. J. Farabaugh, *Clustering of multidimensional data sets with applications to spatial distributions of ribosomal proteins*, Tech. Rep. HPCF–2013–10, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2013.

[5] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice Hall, Englewood Cliffs, 1982.

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

[7] E. Rothberg, *MATHPROG: Solver for the maximum weight matching problem*, 1999. `http://elib.zib.de/pub/Packages/mathprog/matching/weighted/index.html`, accessed December 11, 2013.

[8] M. F. Schilling, *Multivariate two-sample tests based on nearest neighbors*, J. Am. Statist. Ass., 81 (1986), pp. 799–806.