2011-12

# Nonparametric tests for homogeneity based on non-bipartite matching

Ruth, David M.; Koyak, Robert A.

Journal of the American Statistical Association

December 2011, Vol. 106, No. 496, Theory and Methods
http://hdl.handle.net/10945/47656

# Nonparametric Tests for Homogeneity Based on Non-Bipartite Matching

David M. RUTH and Robert A. KOYAK

Given a sequence of observations, has a change occurred in the underlying probability distribution with respect to observation order? This problem of detecting change points arises in a variety of applications including health prognostics for mechanical systems, syndromic disease surveillance in geographically dispersed populations, anomaly detection in information networks, and multivariate process control in general. Detecting change points in high-dimensional settings is challenging, and most change-point methods for multidimensional problems rely upon distributional assumptions or the use of observation history to model probability distributions. We present three new nonparametric statistical tests for heterogeneity based on the combinatorial properties of minimum non-bipartite matching (MNBM). The key idea underlying each of these tests is that if a sequence of independent random observations undergoes a change in distribution—either an abrupt "shift" or a gradual "drift"—a MNBM based on inter-point distances tends to produce pairings that are closer in the sequence labeling than would be the case if the observations were drawn from the same distribution. Our tests follow on the work of Rosenbaum (2005) who used MNBM to derive a simple cross-match test statistic for the two-sample problem based on this idea. Similar ideas are present in the minimum spanning tree (MST) test derived by Friedman and Rafsky (1979, 1981). We extend these approaches by utilizing ensembles of orthogonal MNBMs which greatly increase information extraction from the data, leading to tests that compare favorably to parametric procedures while maintaining level and good power properties across distributions.

KEY WORDS: Change point; Distribution-free test; Graph-theoretic procedure.



## 1. INTRODUCTION

Testing if data can be treated as a random sample from a single population is a fundamental problem in statistics. We consider the problem of testing for homogeneity in multivariate data that are sequenced in some manner, such as with respect to time: the goal is to determine whether the sampling distribution undergoes a change, such as a jump or gradual drift in mean, beginning at an unspecified point in the sequence. Change need not be limited to measures of centrality, but may also be signaled in dispersion or other aspects of the probability distribution. "Change detection" is fundamental to a growing number of applications, including health prognostics for mechanical systems (Hess, Duke, and Kogut 2001), syndromic disease surveillance in geographically dispersed populations (Rogerson and Yamada 2004), anomaly detection in information networks (Tartakovsky et al. 2006a, 2006b), and multivariate process control generally. Such applications often employ multivariate cumulative sum (MCUSUM) methods to detect change.

We propose new tests for detecting change in a multivariate data sequence that extend the classical multivariate two-sample testing paradigm. In a two-sample test, let $\mathbf{X}_1, \ldots, \mathbf{X}_m, \mathbf{X}_{m+1}, \ldots, \mathbf{X}_N$ denote a sequence of $p$-dimensional observations that are arranged so that the first $m$ observations are sampled independently from distribution $F_1$, and the last $n$ observations ($N = m + n$) are sampled independently from distribution $F_2$. The null hypothesis is $H_0 : F_1 = F_2$; the alternative $H_1 : F_1 \neq F_2$ implies that $m + 1$ is a change point of the sequence. Our tests do not require that the change point be specified in advance, and they demonstrate power against general directional-change ("drift") alternatives. Like in the works of Rosenbaum (2005) and Heller et al. (2010), our tests use minimum non-bipartite matching (MNBM) to determine a system of observation pairings that minimize a sum of distances. If the distribution changes at some point, either a jump or a gradual drift away from the initial distribution, MNBM tends to pair observations that are closer in sequence than under homogeneity. This fact is exploited in several ways to produce test statistics that are sensitive to distributional change. But, a single MNBM contains limited information about the structure of a dataset, and tests based on them are not very powerful. Additional information is extracted by fitting an ensemble of *orthogonal* MNBMs (i.e., MNBMs that share no common edges) to the data. Each MNBM in the ensemble minimizes the sum of distances of edge-connected points subject to the exclusion of edges that appear in prior matchings. Friedman and Rafsky (1979) originally proposed the use of orthogonal minimum spanning tree (MST) ensembles in the two-sample problem. We develop a cumulative sums test based on MNBM ensembles that exhibits remarkable power against a broad range of alternatives.

Optimal matching has statistical applications other than homogeneity testing. For example, several authors have investigated both bipartite and non-bipartite matching to create pairings of subjects in experiments and observational studies (Lu et al. 2001; Lu and Rosenbaum 2004; Heller et al. 2010).

## 2. GRAPH-THEORETIC APPROACHES TO HOMOGENEITY TESTING

Suppose that data consist of $N$ observations ordered with respect to time (or any other sequencing variable) and we want to test whether the observations are sampled from the same distribution, or from distributions that change with respect to this ordering. For example, we might want to test for a jump or a drift

David Ruth is Permanent Military Professor, Mathematics Department, United States Naval Academy, 572-C Holloway Rd, Annapolis, MD 21402 (E-mail: *druth@usna.edu*). Robert Koyak is Associate Professor, Operations Research Department, Naval Postgraduate School, 1411 Cunningham Road, Monterey, CA 93943-5219 (E-mail: *rakoyak@nps.edu*). The authors thank Bo Lu for providing non-bipartite matching software that was used in their research; Sam Buttrey for computing assistance; David Aldous and Don Gaver for helpful suggestions; and the editor, associate editor, and referees for comments that greatly improved the quality of the article.

**In the Public Domain**
**Journal of the American Statistical Association**
**December 2011, Vol. 106, No. 496, Theory and Methods**
**DOI: 10.1198/jasa.2011.tm10576**

beginning at some unknown point in the sequence. The observations, which may be multivariate, are assumed to be independent and sampled from continuous distributions. If $F_i$ denotes the distribution of $\mathbf{X}_i$, the null hypothesis of homogeneity asserts that $F_1 = F_2 = \cdots = F_N$ without specifying the common distribution. The alternative hypothesis asserts the existence of a point up to which the sampling distribution is the same, and after which the distribution changes. One way to state this alternative hypothesis is that there exists an integer $k$, between $k_0 \geq 1$ and $k_1 \leq N-1$, such that $F_1 = F_2 = \cdots = F_k, F_k \neq F_{k+1}$, and $\delta(F_1, F_j) - \max_{k+1 \leq \ell \leq j-1} \delta(F_\ell, F_j)$ is strictly positive and nondecreasing over $j = k+2, \ldots, N$, where $\delta(F, G)$ is a measure of distance between two probability distributions. Usually we take $k_0 = 1$ and $k_1 = N-1$ but in some cases it may be desired to restrict the change point to a narrower interval. The alternative hypothesis includes simple jump and directional drift alternatives as special cases.

Consider the complete, 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 $E_N$ consists of all $N(N-1)/2$ segments joining each pair of vertices. Associated with edge $(i, j)$ is a weight $d_{ij} = \text{Distance}(\mathbf{X}_i, \mathbf{X}_j)$, where $\text{Distance}(\cdot, \cdot)$ is an inter-point distance function on $\mathbb{R}^p$ (e.g., Euclidean). We do not specify the distance function other than to note that it should have sensible properties: $d_{ij} \geq 0$, $d_{jj} = 0$, $d_{ij} = d_{ji}$, and $d_{ij} \leq d_{ih} + d_{hj}$ for all $i, j$, and $h$. Of interest are families of subgraphs of $(\mathbb{Z}_N, E_N)$ that satisfy certain properties—let $\mathcal{G}_N$ denote such a family—and minimum-weight instances that we denote $\hat{G} = (\hat{V}, \hat{E}) \in \mathcal{G}_N$. That is, $\hat{G} = \text{argmin}_{G=(V,E)\in\mathcal{G}_N} \sum_{(i,j)\in E} d_{ij}$. A test statistic $\hat{\psi} = \psi(\hat{G})$ is defined, based on the minimum subgraph, to detect departure from homogeneity. Its null distribution is obtained using simple permutation principles.

This paradigm underlies several procedures that have appeared in statistical literature. Friedman and Rafsky (1979, 1981) proposed fitting a MST to data combined from two different samples to test whether the sampling distributions are the same. A spanning tree is an acyclic subgraph of $(\mathbb{Z}_N, E_N)$ that connects all vertices. The test statistic, $\hat{\psi}_{\text{MST}}$, is the number of subtrees obtained by removing edges that join vertices belonging to different samples. The null hypothesis is rejected for small values of $\hat{\psi}_{\text{MST}}$, giving a multivariate extension of the Wald–Wolfowitz runs test. The use of MSTs in statistical applications also was explored by Filliben, Kafadar, and Shier (1983) who derived a MST-based test statistic to detect spatial homogeneity in a two-dimensional surface of chemical concentrations. Edges connect adjacent vertices on a gridded surface; edge weights are the (positive) concentration difference between adjacent grid regions. Their test statistic is the sum of all edge weights in a MST with small values giving evidence against homogeneity.

The cross-match test proposed by Rosenbaum (2005), which uses MNBM, also is based on counting features of a minimum subgraph that relate observations from two different samples. A MNBM is a subgraph in which observations are paired so that the sum of inter-point distances of paired vertices is minimized. If $N$ is even, each vertex is incident to exactly one edge and the matching is called *perfect*. If $N$ is odd, exactly one vertex remains unmatched. The cross-match statistic, $\hat{\psi}_{\text{CM}}$, is the number of edges that join vertices belonging to different samples. Two samples that have different distributions produce larger numbers of within-group matches and fewer cross-matches than in the case that the distributions are the same. We discuss Rosenbaum's test in more detail in Section 3.

Also noteworthy are tests based on nearest-neighbor (directed) graphs that have been proposed by a number of authors including Schilling (1986), Henze (1988), and Hall and Tajvidi (2002). Graph-theoretic statistical procedures derive their power from information in the inter-point distances $d_{ij}$. That these distances are sensitive to departures from homogeneity follows from the work of Maa, Pearl, and Bartoszynski (1996) who proved that two (multivariate) distributions are identical if and only if the distributions of inter-point distances, taken within and between observations sampled from the two populations, are the same.

Following Rosenbaum (2005) we focus on statistical procedures derived from MNBMs fit to a sequence of observations so that if $(V, E) \in \mathcal{G}_N$, then $V$ has cardinality of at least $N-1$ and each vertex has degree 1 (i.e., no two distinct edges in $E$ share a common vertex). It is easy to see that $\mathcal{G}_N$ contains $2^{-n}N!/n!$ elements for all $N \geq 2$, where $n = \lfloor N/2 \rfloor$ is the number of edges in the matching. An important idea in hypothesis testing is that *a MNBM $\hat{G}$ is uniformly distributed over $\mathcal{G}_N$ if the observations are independently sampled from a continuous distribution*. The same is not true for minimum spanning trees or nearest-neighbor graphs. And unlike MSTs, *it is always possible to fit an ensemble of n orthogonal MNBMs to the data in a recursive manner.* We show in Section 7 that tests based on ensembles of MNBMs are vastly superior to those based on fitting a single subgraph to the data.

## 3. AN EXTENSION OF ROSENBAUM'S CROSS-MATCH TEST

Consider the case where the first $k$ observations are classified as Group 1 and the remaining $N-k$ observations are classified as Group 2. This is the setting for testing equality of distributions in the two-sample problem. Let $M$ denote the number of pairs that are matched within Group 1, and $M^C$ the number of pairs cross-matched between the two groups. The two statistics are equivalent due to the relationship $2M + M^C = k$. Rosenbaum (2005) called $\hat{\psi}_{\text{CM}} = M^C$ the *cross-match statistic* and derived its exact null distribution, with small values constituting evidence against homogeneity. For convenience we focus on $M$. Following Rosenbaum (2005) the null distribution of $M$ is given by

$$g(r; k, N) = P(M = r) = 2^{k-2r} \binom{n}{k-r} \binom{k-r}{r} \binom{N}{k}^{-1}, \quad r = 0 \vee (k-n), \ldots, \lfloor k/2 \rfloor, \tag{1}$$

with $g(r; k, N) = 0$ for $r < 0 \vee (k-n)$ and $x \vee y = \max(x, y)$. The null hypothesis is rejected if $M > q_k(\alpha)$, where

$$q_k(\alpha) = \min\left\{ q : \sum_{r=0}^{q} g(r; k, N) \geq 1 - \alpha \right\}. \tag{2}$$

We return to the situation where the observations are sequentially ordered to test against the alternative that at some

unknown $k \in \{k_0, k_0+1, \ldots, k_1\}$ the distribution undergoes a change as described in Section 2. Let $M_k$ denote the number of pairs that are matched among the first $k$ observations $2 \le k \le N-1$. The $\alpha$-level *simultaneous accumulated match* (SAM) *test* rejects the null hypothesis if $M_N^* = \max_{k_0 \le k \le k_1}\{M_k - q_k(\alpha_k)\} > 0$, where the individual test levels $\alpha_k$ are chosen so that $P(M_N^* \le 0)$ is minimized subject to $P(M_N^* \le 0) \ge 1-\alpha$. Suppose that $q_k \equiv q_k(\alpha_k)$ are fixed. To evaluate $P(M_N^* \le 0)$ we start with the identity

$$P(M_N^* \le 0) = \sum_{r=0}^{q_{k_1}} \pi(r; k_1, N) \cdot g(r; k_1, N), \tag{3}$$

where $\pi(r; k, N) = P(M_j - q_j \le 0, k_0 \le j \le k-1 | M_k = r)$. The fact that $M_k = r$ implies either $M_{k-1} = r$ or $M_{k-1} = r-1$ suggests the recursion

$$\begin{aligned}\pi(r; k, N) = {} & \frac{2r}{k}\pi(r-1; k-1, N) I(r-1 \le q_{k-1}) \\ & + \frac{k-2r}{k}\pi(r; k-1, N) I(r \le q_{k-1}), \\ & k = k_0+1, \ldots, k_1;\ r = 0 \vee (k-n), \ldots, q_k,\end{aligned} \tag{4}$$

where $I(A)$ is the indicator of an event $A$. The recursion is initialized by taking $\pi(r; k_0, N) \equiv 1$. To find a critical region start with all $\alpha_k$ equal to a common value and adjust the individual test levels to bring the simultaneous test level as close to $\alpha$ as desired.

While the framework of the SAM test seems to invite the use of established multiple comparison procedures, we note that in the present context we are concerned with a *single* null hypothesis, namely, that the underlying distribution for each observation is identical. In general, multiple comparison procedures simultaneously test a family of distinct hypotheses, some of which may be true while others may not be. In such cases, an often-useful alternative to testing whether all hypotheses are jointly true (i.e., controlling family-wise error rate) is to use a procedure which controls the false discovery rate (FDR), that is, the expected proportion of falsely rejected hypotheses relative to all rejected hypotheses. Benjamini and Hochberg (1995) presented a FDR-controlling procedure for independent test statistics; Benjamini and Yekutieli (2001) proved that this procedure also controls the FDR under positive dependence of the individual tests. A simulation exercise suggests that the Benjamini and Hochberg procedure is quite conservative in the present context: To achieve a 0.05 level test with $N = 100$, the SAM test has exact level 0.048 using $\alpha_k = 0.0046$ for all $k$. For 100,000 simulations of $N = 100$ i.i.d. observations from a standard multivariate ($p = 5$) normal distribution, the simulated test level for Benjamini–Hochberg was 0.008. Benjamini–Hochberg improves marginally on the Bonferroni test, with exact level 0.006 using $\alpha_k = 0.0005$, which we believe is due to strong positive dependence across the individual tests.

## 4. THE SUM OF PAIR MAXIMA (SPM) TEST

For indexing purposes assume that the edge set of a MNBM is ordered by edge weight. If $(u, v)$ is the $j$th edge in this ordering, let $R_{j1} = \min\{u, v\}$ and $R_{j2} = \max\{u, v\}$. It is easy to see that $M_k$ is equal to the number of $R_{j2}$ less than or equal to $k$. If $R_{[j]2}$ denotes the $j$th order statistic of the vertex maxima, an efficient implementation of the SAM test is to reject the null hypothesis if $R_{[q_k+1]2} \le k$ holds for at least one $k \in \{k_0, k_0+1, \ldots, k_1\}$ (taking $R_{[n+1]2} = \infty$, where $n = \lfloor N/2 \rfloor$ as before). Because small values of $R_{j2}$ are evidence against homogeneity, we propose the *sum of pair maxima* (SPM) $T_N = \sum_{j=1}^{n} R_{j2}$ as an alternate test statistic. Its equivalence to the sum of vertex differences follows from the relationship

$$T_N = N(N+1)/4 + \sum_{j=1}^{n}(R_{j2} - R_{j1})/2. \tag{5}$$

A small value of $T_N$ indicates that observations close together in sequence order tend to be matched.

Under the null hypothesis $R_{12}, \ldots, R_{n2}$ are exchangeable random variables and their moments are obtained from straightforward calculations; see the work of Ruth (2009). For $N$ even we have

$$\begin{aligned}E(R_{12}) &= 2(N+1)/3, \\ \mathrm{Var}(R_{12}) &= (N+1)(N-2)/180, \\ \mathrm{Cov}(R_{12}, R_{22}) &= -4(N+1)/45\end{aligned} \tag{6}$$

and

$$\begin{aligned}\mu_N &= E(T_N) = N(N+1)/3, \\ \sigma_N^2 &= \mathrm{Var}(T_N) = N(N-2)(N+1)/180.\end{aligned} \tag{7}$$

For odd sample sizes one observation is not paired in the MNBM, and the mean and variance are given by

$$\begin{aligned}\mu_N &= E(T_N) = (N-1)(N+1)/3, \\ \sigma_N^2 &= \mathrm{Var}(T_N) = (N-1)(N+2)(N+1)/180.\end{aligned} \tag{8}$$

Unlike the cross-match statistic, the exact distribution of $T_N$ is difficult to obtain. The following establishes a useful asymptotic result for $T_N$ which we prove in the Appendix:

*Theorem 1.* Let $W_N = \sigma_N^{-1}(T_N - \mu_N)$. Then for all $t \in (-\infty, \infty)$

$$|P(W_N \le t) - \Phi(t)| \to 0 \quad \text{as } N \to \infty, \tag{9}$$

where $\Phi(t)$ is the standard normal cumulative distribution function.

Under a normal approximation rejection of the null hypothesis is indicated for $T_N$ less than $\mu_N - \Phi^{-1}(1-\alpha)\sigma_N - 0.5$. A sharper approximation using an Edgeworth expansion is given by

$$\begin{aligned}&P(W_N \le t) \\ &\quad \approx \Phi(t) + c_0 \frac{N+3}{N\sqrt{(N-2)(N+1)}}(t^2-1)\exp(-0.5t^2),\end{aligned} \tag{10}$$

where $c_0 = \sqrt{5/441\pi} \approx 0.06$. Details are provided by Ruth (2009).

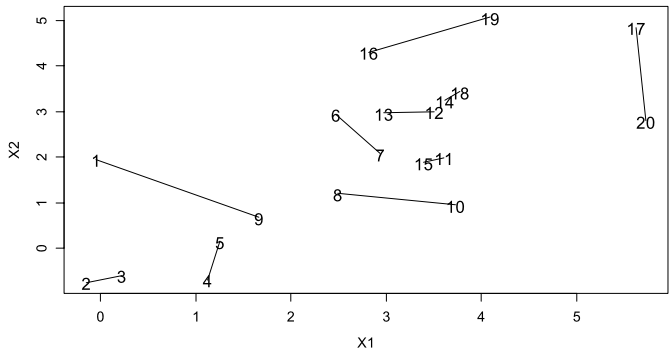


Figure 1. A minimum-cost non-bipartite graph fit to $N = 20$ bivariate observations.

## 5. CONSISTENCY PROPERTIES

Although graph-theoretic nonparametric tests are intuitively appealing, establishing their theoretical properties under alternative hypotheses is challenging. Universal consistency of the generalized runs test of Friedman and Rafsky (1979) was proved by Henze and Penrose (1999) twenty years later. Rosenbaum (2005) showed that the cross-match test is consistent under less general conditions. In both cases consistency is established for the "two-sample" problem in which group labeling is used explicitly. Ruth (2009) showed that the SPM test is consistent against general jump alternatives (without specifying the change point) under the same conditions that the cross-match test is consistent. Extensive simulations, which we discuss below, lead us to believe that both the SPM and SAM tests are consistent under a broadly stated alternative.

## 6. ILLUSTRATIVE EXAMPLE

We introduce a simple example for illustration. A sample of $N = 20$ bivariate observations was randomly generated, which are plotted by their sequence numbers in Figure 1 and summarized in Table 1. These observations exhibit strong mean drift, and the MNBM (based on Euclidean distance) shows a tendency to pair observations that occur close together in sequence. The solution was obtained using Derigs' (1988) algorithm, which was provided to us by Bo Lu at Ohio State University in the form of a dynamically loaded library (DLL) module and executed in R (R Development Core Team 2009).

Table 1. Homogeneity test calculations for the data shown in Figure 1

| Pair | Maximum | $M_k$ | $q_k$ |
|---|---|---|---|
| (2, 3) | 3 | $M_3 = 1$ | $q_3 = 1$ |
| (4, 5) | 5 | $M_5 = 2$ | $q_5 = 2$ |
| (6, 7) | 7 | $M_7 = 3$ | $q_7 = 3$ |
| (1, 9) | 9 | $M_9 = 4$ | $q_9 = 3^*$ |
| (8, 10) | 10 | $M_{10} = 5$ | $q_{10} = 4^*$ |
| (12, 13) | 13 | $M_{13} = 6$ | $q_{13} = 6$ |
| (11, 15) | 15 | $M_{15} = 7$ | $q_{15} = 7$ |
| (14, 18) | 18 | $M_{18} = 8$ | $q_{18} = 9$ |
| (16, 19) | 19 | $M_{19} = 9$ | $q_{19} = 9$ |
| (17, 20) | 20 | $M_{20} = 10$ | |
| | $T_N = 119$ | | |

NOTE: Asterisks indicate violations of the acceptance region in a nominal $\alpha = 0.05$ simultaneous accumulated match (SAM) test.

For the SAM test we first determine the non-simultaneous critical values $q_k, k = 2, \ldots, 19$ that achieve a simultaneous test level close to the nominal $\alpha = 0.05$. Using (4) produces an achieved simultaneous test level of 0.046 with critical values listed in Table 1. Rejection of the null hypothesis is signaled if any $M_k - q_k$ is greater than zero, or equivalently if $R_{[q_k+1]2} \le k$ for any $k$. The latter reduces to checking the following critical conditions: $R_{[2]2} \le 3$, $R_{[3]2} \le 5$, $R_{[4]2} \le 9$, $R_{[5]2} \le 10$, $R_{[7]2} \le 13$, $R_{[8]2} \le 15$, and $R_{[10]2} \le 19$. Because $R_{[4]2} = 9$, rejection of the null hypothesis is signaled at $k = 9$ using the SAM test. ($R_{[5]2} = 10$ also signals rejection at $k = 10$.)

For the SPM test we obtain $T_N = 119$. Under the null hypothesis the expected value and standard deviation of $T_N$ are given by $\mu_N = 20(21)/3 = 140, \sigma_N = \sqrt{20(18)(21)/180} = 6.48$. From (10) the (asymptotic) critical value for a $\alpha = 0.05$ level test is 129 which agrees exactly with the critical value obtained using 10,000 simulations. The normal approximation $p$-value is 0.0008, which suggests that the pattern shown in Figure 1 is not indicative of homogeneity.

## 7. ENSEMBLE TESTS

The power of a graph-theoretic test is enhanced by considering collections of orthogonal subgraphs that we call *ensembles*. Recall that two graphs are orthogonal if they share no common edges. A *k-ensemble* consists of $k$ subgraphs that are pairwise orthogonal. If $k = N - 1$ we say that an ensemble of MNBMs is *complete*. Complete MNBM ensembles are solutions to round-robin scheduling problems, and are isomorphic to the set of symmetric Latin squares of order $N - 1$ with permutations of the integers $1, \ldots, N - 1$ on the main diagonal. Friedman and Rafsky (1979, 1983) suggested that ensembles of MSTs with $k \ll N/2$ be used to enhance the sensitivity of their generalized runs test. Similarly, we define a *recursively optimal k-ensemble* of MNBMs to be a sequence $G_1, \ldots, G_k$ where $G_1$ is a solution to the MNBM problem and $G_{r+1}$ is a solution to MNBM problem subject to $G_{r+1} \in \mathcal{G}_N^{\perp}(G_1, \ldots, G_r)$ where $\mathcal{G}_N^{\perp}(G_1, \ldots, G_r)$ denotes the subclass of MNBMs in $\mathcal{G}_N$ that contain no common edges with any of $G_1, \ldots, G_r$. It is easy to find examples where recursive construction fails to yield a complete MNBM ensemble, but it can be shown always to produce a "half" ensemble ($k = \lfloor N/2 \rfloor$) (Anderson 1972) which is sufficient for our purposes.

The $n$-ensemble version of the SPM test is based on the quantities

$$S_{N,\nu} = \sum_{j=1}^{\nu} T_{N,j}, \qquad \nu = 1, \ldots, n = \lfloor N/2 \rfloor, \tag{11}$$

where $T_{N,j}$ is the SPM statistic for the $j$th MNBM of the ensemble. The first two moments of $S_{N,\nu}$ are obtained from the following result, which we prove in the Appendix:

*Theorem 2.* Let $G_1, \ldots, G_n$ be a recursively optimal $n$-ensemble of MNBMs based on distances between $N = 2n$ points constituting a random sample from a continuous distribution in $\mathbb{R}^p$, and let $T_{N,j}$ denote the SPM statistic based on $G_j, j = 1, \ldots, n$. Then $(T_{N,1}, \ldots, T_{N,n})$ has identical univariate marginal distributions and moments up to second order,

with

$$\begin{aligned} E(T_{N,j}) &= \mu_N = N(N+1)/3, \\ \mathrm{Var}(T_{N,j}) &= \sigma_N^2 = N(N-2)(N+1)/180, \\ \mathrm{Cov}(T_{N,i}, T_{N,j}) &= -\sigma_N^2/(N-2) \\ &= -N(N+1)/180 \qquad (i \neq j). \end{aligned} \tag{12}$$

The following are immediate from Theorem 2 under the null hypothesis of homogeneity:

$$\xi_{N,\nu} = E(S_{N,\nu}) = \nu\mu_N = \nu N(N+1)/3 \tag{13}$$

and

$$\mathrm{Cov}(S_{N,\nu}, S_{N,\omega}) = \left(\frac{\nu}{N-1}\right)\left(1 - \frac{\omega}{N-1}\right)c_N^2, \quad 1 \le \nu \le \omega \le N-1, \tag{14}$$

where $c_N^2 = N(N+1)(N-1)^2/180$.

Theorem 1 and (14) lead us to expect that $S_{N,\nu}$ is asymptotically normal and that the process

$$B_N\left(\frac{\nu}{N-1}\right) = \frac{\xi_{N,\nu} - S_{N,\nu}}{c_N}, \qquad \nu \in \{1, 2, \ldots, n\}; B_N(0) \equiv 0, \tag{15}$$

can be approximated by a Brownian bridge $B(t)$, $0 \le t \le 0.5$, for $N$ sufficiently large. This leads us to consider a test based on the quantity

$$B_N^* = \max_{\nu \in \{0,1,\ldots,n\}} B_N\left(\frac{\nu}{N-1}\right). \tag{16}$$

We call $B_N^*$ the *ensemble sum of pair maxima* (ESPM) statistic. From Brownian bridge theory we have

$$P\Big(\sup_{0 \le t \le \lambda} B(t) > b\Big) = 1 - \Phi(b/\sqrt{\lambda(1-\lambda)}) + \exp(-2b^2) \times \Phi(b(2\lambda - 1)/\sqrt{\lambda(1-\lambda)}) \tag{17}$$

(Shorack and Wellner 1986, p. 38). Setting $\lambda = 0.5$ gives an approximate $\alpha$-level test that rejects the null hypothesis of homogeneity if $B_N^*$ exceeds $b_\alpha$, where $b_\alpha$ is the solution to the equation

$$1 - \alpha - \Phi(2b) + 0.5\exp(-2b^2) = 0. \tag{18}$$

Simulations, however, point to the surprising result that (15) is not well approximated by a Gaussian process. This is apparent in the normal QQ-plots shown in Figure 2 for 10,000 simulated values of $B_N(t)$ at $t = \nu/(N-1)$, $N = 200$, $\nu = 1, 50, 100$ under sampling from uniform distributions on $[0,1]^2$ and $[0,1]^{100}$, respectively. Although Figure 2(a) is anticipated by Theorem 1 for a single SPM statistic, $B_N(t)$ has an increasingly skewed distribution as $t$ increases, especially in the lower-dimensional example. Why this happens is not well understood by us, but it may be due to long-range dependence resulting from the restriction of distance matrices to manifolds of approximately the same dimension as the number of variables. To examine how dimensionality affects the higher-order random behavior of orthogonal matchings we conducted a simulation in which 1000 replicates of $N = 100$ and $N = 500$ samples were generated from a uniform distribution on $[0,1]^p$ for $p = 2, 5, 10, 20$. Of interest is the proportion of "triangle edges," $\tau_N(p)$, appearing in the third orthogonal matching.

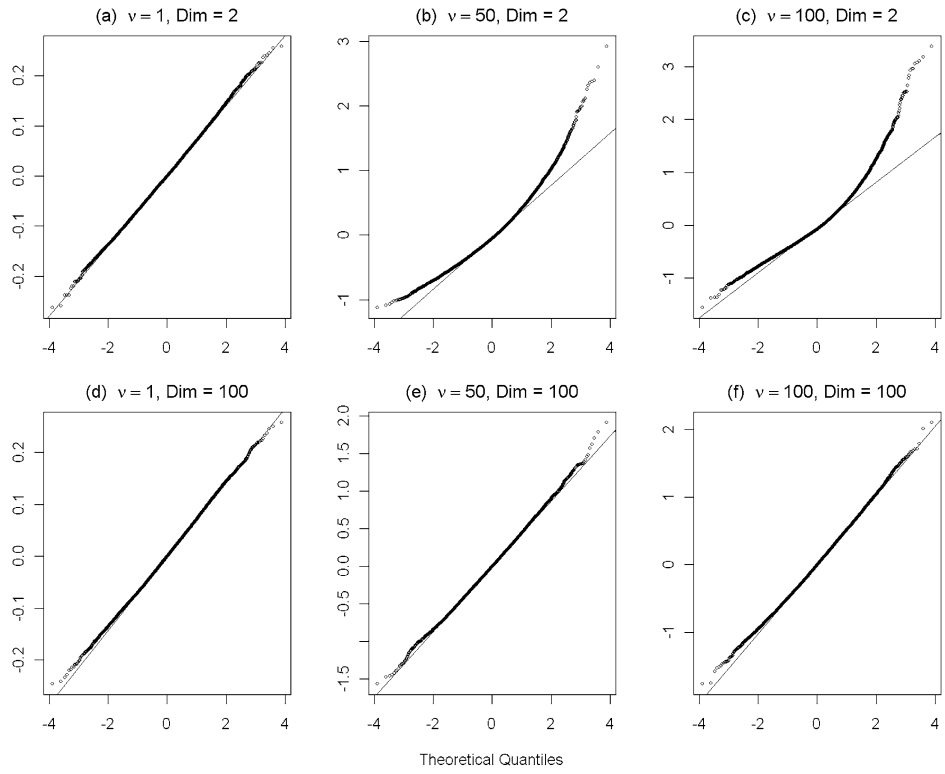


Figure 2. Normal quantile–quantile plots of 10,000 simulated values of $B_N(\nu/(N-1))$ for $N = 200$, $\nu = 1, 50, 100$. Panels (a)–(c) are from independent samples of Unif$[0,1]^2$ variates. Panels (d)–(f) are from independent samples of Unif$[0,1]^{100}$ variates.

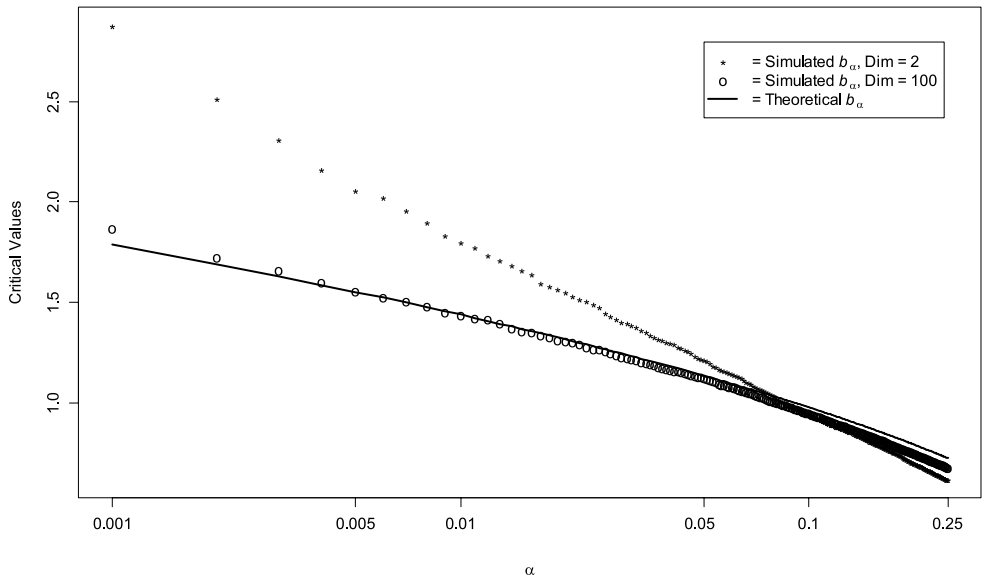


Figure 3. Comparison of critical values of $B_N^*$, $N = 200$, approximated using 10,000 simulations of dimensions 2 and 100, and Brownian bridge theory (theoretical $b_\alpha$).

A triangle edge $(u, v)$ is one where there is an integer $w$ such that the edges $(u, w)$ and $(v, w)$ appear in the first two matchings. We find that the proportion of triangle edges decreases sharply as $p$ increases for both sample sizes. The estimated values are $\hat{\tau}_{100}(2) = 0.72$, $\hat{\tau}_{100}(5) = 0.42$, $\hat{\tau}_{100}(10) = 0.24$, $\hat{\tau}_{100}(20) = 0.13$ for $N = 100$; $\hat{\tau}_{500}(2) = 0.68$, $\hat{\tau}_{500}(5) = 0.37$, $\hat{\tau}_{500}(10) = 0.20$, $\hat{\tau}_{500}(20) = 0.09$ for $N = 500$. This exercise suggests that ensembles exhibit stochastic effects of higher than second order that depend on $p$ but do not dissipate as $N$ increases.

Figure 3 confirms that use of $b_\alpha$ as a critical value may violate the nominal level. At $\alpha = 0.05$ the simulation critical value for $p = 2$ is 1.218, which is greater than $b_{0.05} = 1.133$. In high-dimensional settings skewness is less evident and Figure 3 suggests that $b_\alpha$ can provide a reasonable critical value.

Table 2 provides critical values for test levels $\alpha = 0.01, 0.05$ under different values of $N$ and number of variables $p$ obtained from 100,000 simulated values of $B_N^*$. Samples were taken from uniform distributions on unit hypercubes, and Euclidean distances were used. Simulations were performed at the Naval Postgraduate School's High Performance Computing Center. We have found that these critical values are stable with respect to the choice of sampling distribution and distance function. A more extensive table of critical values can be found in the work of Ruth (2009).

Table 2. Critical values of the ensemble test statistic $B_N^*$

| | $p$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $N$ | 1 | 2 | 3 | 4 | 5 | 10 | 20 | 50 |
| | | | | $\alpha = 0.01, b_{0.01} = 1.438$ | | | | |
| 20 | *1.72* | 1.66 | 1.60 | 1.56 | *1.53* | 1.46 | 1.43 | 1.38 |
| 40 | 1.83 | 1.74 | 1.68 | 1.63 | 1.59 | 1.50 | 1.47 | 1.43 |
| 60 | 1.85 | 1.76 | 1.70 | 1.65 | 1.62 | 1.53 | 1.50 | 1.44 |
| ≥80 | 1.86 | 1.78 | 1.72 | 1.67 | 1.63 | 1.54 | 1.50 | 1.45 |
| | | | | $\alpha = 0.05, b_{0.05} = 1.133$ | | | | |
| 20 | 1.13 | *1.12* | 1.10 | 1.10 | 1.09 | 1.07 | 1.07 | *1.03* |
| 40 | 1.20 | 1.17 | 1.15 | 1.14 | 1.13 | 1.10 | 1.09 | 1.08 |
| 60 | 1.20 | 1.18 | 1.16 | 1.15 | 1.14 | 1.11 | 1.10 | 1.09 |
| ≥80 | 1.21 | 1.19 | 1.18 | 1.16 | 1.15 | 1.13 | 1.11 | 1.10 |

NOTE: Values obtained from 100,000 simulations per instance of sample size ($N$) and number of variables ($p$) taken from Ruth (2009). Samples are generated as Unif$[0, 1]^2$; distances are Euclidean. Critical values $b_\alpha$, $\alpha = 0.01, 0.05$, for a Brownian bridge approximation also are shown. Estimated standard errors are less than 0.015 with the exception of italicized values that have estimated standard errors of approximately 0.02.

## 8. ENSEMBLE TESTS: AN ILLUSTRATION

Table 3 shows age-adjusted breast cancer mortality rates relative to a 1968 base for the counties of Philadelphia, PA and Schuylkill, PA for the twenty-year period 1969 to 1988. Data are from the Compressed Mortality File (1968–1989) provided by the National Center for Health Statistics (2000) which bears no responsibility for our use of the data. The reported value of 1.017 for Philadelphia in 1969 implies that the breast cancer mortality rate was 1.7% greater in 1969 than in 1968. Of interest is whether these rates have remained stable over time, which is the subject of the wider study of Rogerson and Yamada (2004). We present our example for illustrative purposes only. Table 4 shows the results of fitting a recursively optimal $n = 10$ ensemble of MNBMs to the data using four distance functions: Euclidean, Mahalanobis with estimated covariance matrix centered on variable means, Mahalanobis with estimated covariance matrix centered on a scatterplot smoother applied to the variables in time sequence, and Manhattan. It also shows results for the SPM and SAM tests, which are based on fitting a single MNBM to the data. For comparison we also consider the test (JJS) of James, James, and Siegmund (1992) which is designed to detect a mean-change point in multivariate normal data with the modified likelihood ratio test statistic

$$T_{\mathrm{JJS}} = \max_{k_0 \le k \le k_1} \frac{N}{k(N-k)} \left( \mathbf{S}_k - \frac{k}{N} \mathbf{S}_N \right)' \times \left( \mathbf{U}_N - \frac{1}{N} \mathbf{S}_N \mathbf{S}_N' \right)^{-1} \left( \mathbf{S}_k - \frac{k}{N} \mathbf{S}_N \right), \quad (19)$$

Table 3. Age-adjusted breast cancer mortality rates relative to a 1968 base for the Pennsylvania counties of Philadelphia and Schuylkill, 1969–1988

| Year | Philadelphia | Schuylkill |
|---|---|---|
| 1969 | 1.017 | 1.034 |
| 1970 | 1.069 | 1.044 |
| 1971 | 0.943 | 1.260 |
| 1972 | 1.002 | 1.320 |
| 1973 | 0.955 | 1.239 |
| 1974 | 1.037 | 1.274 |
| 1975 | 1.008 | 0.974 |
| 1976 | 0.946 | 0.936 |
| 1977 | 1.134 | 1.329 |
| 1978 | 1.077 | 1.664 |
| 1979 | 0.989 | 1.095 |
| 1980 | 1.040 | 1.274 |
| 1981 | 1.140 | 1.299 |
| 1982 | 1.049 | 1.313 |
| 1983 | 1.209 | 1.319 |
| 1984 | 1.133 | 1.342 |
| 1985 | 1.274 | 1.528 |
| 1986 | 1.073 | 1.543 |
| 1987 | 1.171 | 1.060 |
| 1988 | 1.228 | 1.463 |

NOTE: The age distribution of the nine states of the U.S. Northeast obtained from the 1980 Census was used for adjustment. Values greater than 1.0 indicate an increased mortality rate over the 1968 rate for the same county.

Table 4. Derivation of the ESPM statistic $B_N^*$ from values of $B_N(\nu/N-1)$ for the data in Table 3 ($N=20$) with respect to four distance functions

| | Distance function | | | |
|---|---|---|---|---|
| | $L^2(I)$ | $L^2(\hat{\Sigma})$ | $L^2(\hat{\Sigma}_C)$ | $L^1$ |
| $\nu=1$ | 0.069 | 0.103 | 0.103 | 0.069 |
| $\nu=2$ | 0.620 | 0.551 | 0.620 | 0.655 |
| $\nu=3$ | 0.896 | 0.724 | 1.103 | 0.896 |
| $\nu=4$ | 0.896 | 0.999 | 1.309 | 0.896 |
| $\nu=5$ | 1.034 | 0.827 | 1.309 | 1.034 |
| $\nu=6$ | 1.171 | 1.034 | 1.240 | 1.309 |
| $\nu=7$ | 1.413 | 0.827 | 1.378 | 1.654 |
| $\nu=8$ | 1.723 | 1.068 | 1.619 | 1.895 |
| $\nu=9$ | 2.102 | 1.344 | 1.826 | 2.274 |
| $\nu=10$ | 2.240 | 1.344 | 1.723 | 2.515 |
| $B_{20}^*$ | 2.240 | 1.344 | 1.826 | 2.515 |
| $p$-value | $p<0.01$ | $0.01<p<0.05$ | $p<0.01$ | $p<0.01$ |

NOTE: $L^2(I)$ = Euclidean; $L^2(\hat{\Sigma})$ = Mahalanobis; $L^2(\hat{\Sigma}_C)$ = Mahalanobis with estimated covariance matrix centered on a scatterplot smoother applied to the variables in time sequence; $L^1$ = Manhattan. Value of $B_{20}^*$ and associated $p$-value are listed for each case. Critical values for $B_{20}^*$ are 1.66 ($\alpha=0.01$) and 1.22 ($\alpha=0.05$).

where $\mathbf{S}_k=\sum_{i=1}^k \mathbf{X}_i$, $\mathbf{U}_k=\sum_{i=1}^k \mathbf{X}_i\mathbf{X}_i'$, and $k_0$ and $k_1$ are the lower and upper limits, respectively, of an interval containing a putative change point. Although the JJS test bears resemblance to a MCUSUM procedure, the former is a nonsequential test designed to detect change after data have been collected.

Figure 4 shows the first two MNBMs of the ensemble (using Euclidean distance). Neither suggests temporal change, and indeed neither of the single-MNBM tests (SAM and SPM) finds evidence to that effect (SAM $p$-values exceed 0.2 in all cases and SPM $p$-values exceed 0.3 in all cases). The ESPM test does, however, detect change with each distance function considered, achieving $p$-values less than 0.01 in all but one case (mean-centered Mahalanobis). The JJS test also strongly signals change in this example with a $p$-value of 0.0024. While all three MNBM-based tests maintain test level, this example shows that the ensemble test does so with superior power. We examine the performance of these tests in more detail in the next section.

It is interesting to note that both of the unscaled distances (Euclidean and Manhattan) outperformed either of the Mahalanobis-based distances in this example. If pronounced directional drift is present, the use of an estimated covariance matrix in distance calculations can dampen the effect that the test was designed to detect. Smoothing offers only partial improvement. Individually standardizing the variables is sensible if they are measured on widely different scales, but we have found little to recommend scaling using an estimated covariance matrix.

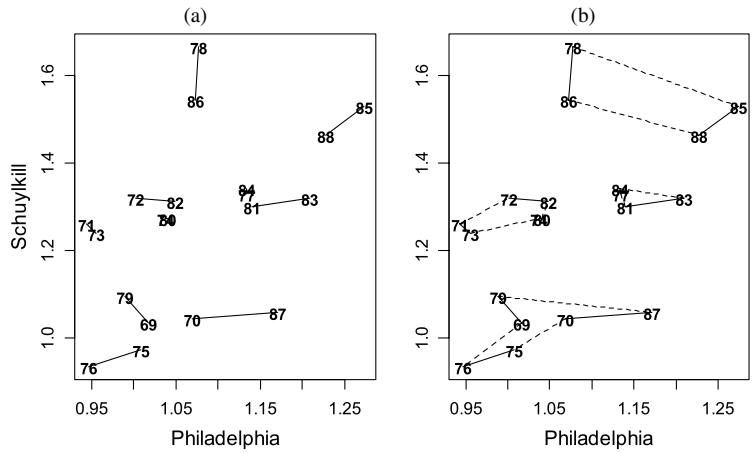


Figure 4. Age-adjusted breast cancer mortality rates relative to a 1968 base for the Pennsylvania counties of Philadelphia and Schuylkill, 1969–1988. Source: National Center for Health Statistics (1968–1988). Numbers plotted are years, where 69 = 1969, 70 = 1970, ..., 88 = 1988. Solid line segments indicate the minimum-cost non-bipartite matching based on Euclidean distances. Dashed line segments in (b) indicate the minimum-cost orthogonal matching.

## 9. SIMULATION STUDY

Simulations by Ruth (2009) suggest that ensemble tests are powerful across a broad range of alternatives. Table 5 displays a subset of his findings. A total of 1000 samples were generated for each of 30 vignettes, and four different tests are compared: the three MNBM-based tests developed by Ruth (2009), and the JJS test. The vignettes vary by dimensionality ($p = 5, 20$); type of change (jump or drift); multivariate distribution type (normal, normal mixture, or Weibull); the parameter $\theta$ affected (mean vector, covariance matrix, or scale parameter); and magnitude of change ($\Delta$). In all vignettes the true change point is $k = 101$. For mean-change vignettes, all samples start with a mean vector of zero and end with the mean vector having norm $\Delta$, either as a jump or as a linear drift starting at the change point. For vignettes in which the covariance matrix or scale parameter changes, all samples start with unit parameters and end with the parameter multiplied by $1 + \Delta$, but only for the first variate. Multivariate normal mixtures have zero mean and identity covariance with probability 0.9, zero mean, and 16 times the identity covariance with probability 0.1. Multivariate Weibull vectors consist of independent, univariate Weibulls with shape parameters 1.5 and scale parameters 1. The nominal test level is $\alpha = 0.05$; power at $\Delta = 0$ should be near the test level up to simulation error.

Simulated powers for the SAM, SPM, ESPM, and JJS tests in Table 5 clearly demonstrate the value of using ensembles in graph-theoretic procedures: the ESPM test consistently dominates both the SAM and SPM tests, and it is surprisingly competitive with the JJS test on multivariate normal data. On non-normal data the ESPM test dominates the JJS test with the latter unable to maintain its level in outlier-prone situations or in the presence of skewness, conditions for which the JJS test was not designed.

## 10. FUTURE WORK

The information-extraction capability of graph-theoretic statistical procedures is impressive when ensembles are used. Procedures based on minimum non-bipartite matching are conceptually appealing and they lend themselves well to ensemble extensions. We highlight additional work that is needed to bring these procedures into wide usage, particularly in modern-scale applications:

(a) *Theoretical development*. Whether the change point is specified or not, conditions have not yet been defined under which ensemble procedures such as ESPM are consistent. Formal articulation of their properties under alternative hypotheses outside of the two-sample setting also remains an open problem.

(b) *The need for speed*. MNBMs can be found in polynomial time, but not as efficiently as some other well-known minimum-weight subgraphs such as minimum spanning trees and bipartite matchings. The "Blossom algorithm" and its successors (Edmonds 1965a, 1965b; Gabow 1973; Derigs 1988; Cook and Rohe 1999; Mehlhorn and Schäfer 2002; Kolmogorov 2009) define the state of the art for solving this problem. But the Blossom algorithm's $O(N^3)$ worst-case and empirical complexity (Kolmogorov 2009) make it prohibitively expensive to implement on large samples. This contrasts with the $O(N \log N)$ times that can be achieved for near-minimum spanning trees

Table 5. Simulated powers for simultaneous change-point detection tests under different distributions and change scenarios based on 1000 simulations, $\alpha = 0.05$, $N = 200$, and change point $k = 101$

| | Jump | | | | Drift | | | |
|---|---|---|---|---|---|---|---|---|
| $\Delta$ | SAM | SPM | ESPM | JJS | SAM | SPM | ESPM | JJS |
| | | | | (a) Multivariate normal, $\theta$ = mean, $p = 5$ | | | | |
| 0 | 0.05 | 0.06 | 0.04 | 0.05 | 0.05 | 0.04 | 0.06 | 0.07 |
| 0.5 | 0.09 | 0.10 | 0.60 | 0.52 | 0.05 | 0.07 | 0.27 | 0.22 |
| 1.0 | 0.33 | 0.41 | 1.00 | 1.00 | 0.16 | 0.20 | 0.84 | 0.85 |
| | | | | (b) Multivariate normal, $\theta$ = mean, $p = 20$ | | | | |
| 0 | 0.05 | 0.05 | 0.05 | 0.03 | 0.05 | 0.05 | 0.05 | 0.04 |
| 0.5 | 0.07 | 0.09 | 0.33 | 0.20 | 0.05 | 0.07 | 0.13 | 0.09 |
| 1.0 | 0.16 | 0.22 | 0.95 | 0.95 | 0.09 | 0.11 | 0.56 | 0.49 |
| | | | | (c) Multivariate normal, $\theta$ = covariance matrix, $p = 5$ | | | | |
| 0 | 0.05 | 0.06 | 0.05 | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 |
| 0.5 | 0.42 | 0.51 | 0.97 | 0.15 | 0.20 | 0.27 | 0.52 | 0.27 |
| 1.0 | 0.99 | 0.99 | 1.00 | 0.24 | 0.77 | 0.79 | 1.00 | 0.54 |
| | | | | (d) Multivariate normal mixture, $\theta$ = mean, $p = 5$ | | | | |
| 0 | 0.05 | 0.05 | 0.04 | 0.27 | 0.04 | 0.04 | 0.06 | 0.28 |
| 0.5 | 0.08 | 0.09 | 0.56 | 0.38 | 0.07 | 0.07 | 0.21 | 0.33 |
| 1.0 | 0.25 | 0.36 | 0.99 | 0.85 | 0.12 | 0.15 | 0.76 | 0.55 |
| | | | | (e) Multivariate Weibull, $\theta$ = scale parameter, $p = 5$ | | | | |
| 0 | 0.05 | 0.05 | 0.06 | 0.09 | 0.05 | 0.06 | 0.05 | 0.09 |
| 0.5 | 0.10 | 0.13 | 0.70 | 0.70 | 0.07 | 0.08 | 0.35 | 0.46 |
| 1.0 | 0.27 | 0.34 | 0.99 | 1.00 | 0.20 | 0.23 | 0.86 | 0.94 |

NOTE: SAM = simultaneous accumulated match test, SPM = sum of pair maxima test, ESPM = ensemble sum of pair maxima test, JJS = approximate normal likelihood ratio test.

in "coordinate spaces" (Bentley and Friedman 1978) or nearest neighbors (Friedman, Bentley, and Finkel 1977; Arya et al. 1998). For an ensemble test that requires $\lfloor N/2 \rfloor$ MNBM calculations, the problem of speed is acute. Reported run times by Ruth (2009) for MNBM ensembles using Derigs' (1988) algorithm are on the order of $N^4$. These sobering observations point to the need for (i) more efficient, perhaps suboptimal algorithms for MNBM for near-real time applications, and (ii) a "cost-benefit analysis" for MNBM compared to more computationally efficient suboptimal approaches or alternate graph-theoretic methods, for example, MST or nearest neighbors.

(c) *Performance in high-dimensional problems*. Not surprisingly, the power of graph-theoretic methods such as ESPM decreases as the number of variables increases. Modern applications of anomaly detection routinely encounter situations in which the number of variables exceeds the number of observations. The options are to conduct many low-dimensional monitoring procedures in parallel (with loss of information about the relationships between variables) or to reduce dimensionality. Principal component analysis or its sparse extensions (e.g., Zou, Hastie, and Tibshirani 2006) are natural choices, but the directions in high-dimensional space that align with distributional change are not necessarily the same as those that maximize variability.

(d) *Extensions beyond the iid case*. Data used in applications are often correlated, serially or spatially, and any procedure that minimizes a criterion based on inter-point distances is sensitive to this. Also, control variables or covariates that are not of interest to anomaly detection may be present. The latter applies, for example, to detection of a change in residual distribution after model fitting. Currently, graph-theoretic procedures do not accommodate these issues; however, we believe that extensions might be possible in some cases. For example, to deal with autocorrelation matching can be prohibited between observations taken near to each other in time.

(e) *One-sided detection*. Often the goal in applications such as biosurveillance is to detect "signed" change, such as increases in reported incidents of an infectious disease. Although several approaches to developing directionally sensitive multivariate procedures have been considered (Fricker 2007; Fricker, Knitt, and Hu 2008; Joner et al. 2008), existing graph-theoretic procedures are not designed for this setting.

## APPENDIX: PROOFS

### Proof of Theorem 1

Initially, we assume $N = 2n$ is an even integer, and then extend our result to the case of odd sample sizes. Our strategy is to exploit the combinatorial structure of $T_N$ by constructing a closely coupled random variable $\tilde{T}_N$ where $(T_N, \tilde{T}_N)$ is an exchangeable pair. Stein's method (Stein 1986) is then used to prove that $W_N$ is asymptotically standard normal under the null hypothesis of homogeneity. We assume that $G$ consists of $n$ pairs ordered in some manner, such as by smaller pair element. Select a pair of distinct integers $(u, v)$ at random from $1, 2, \ldots, n$ and then swap one element chosen at random from each of the $u$th and $v$th pairs in $G$. This results in a new MNBM $\tilde{G}$ and we take $\tilde{T}_N$ to be the resulting SPM statistic. Then $(T_N, \tilde{T}_N)$ and $(W_N, \tilde{W}_N)$ both are exchangeable pairs where $\tilde{W}_N = \sigma_N^{-1}(\tilde{T}_N - \mu_N)$, due to the exchangeability of $(G, \tilde{G})$ under the random switching procedure. Write $\tilde{T}_N = T_N + \Delta_N$, where $\Delta_N = \tilde{R}_{u2} + \tilde{R}_{v2} - R_{u2} - R_{v2}$. Some algebra shows that

$$E(\tilde{R}_{j2}|W_N) = 2(N+1)(N-1)/3(N-2) - 2T_N/N(N-2) \quad \text{(A.1)}$$

and $E(\tilde{W}_N|W_N) = (1-\lambda_N)W_N$, where $\lambda_N = 4(N-1)/N(N-2)$. Applying Stein's method through theorem 2.5 from the article of Rinott and Rotar (2000) gives

$$|P(W_N \le t) - \Phi(t)| \le 6\lambda_N^{-1}\sqrt{\mathrm{Var}\{E[(\tilde{W}_N - W_N)^2|W_N]\}} + 6\lambda_N^{-1/2}\sqrt{E\{|\tilde{W}_N - W_N|^3\}} \quad \text{(A.2)}$$

which reduces to

$$|P(W_N \le t) - \Phi(t)| \le 360N^{-2}\sqrt{\mathrm{Var}\{E[\Delta_N^2|W_N]\}} + 360N^{-7/4}\sqrt{E\{|\Delta_N|^3\}}. \quad \text{(A.3)}$$

Because $|\Delta_N| \le N - 3$ it follows that $N^{-7/4}[E\{|\Delta_N|^3\}]^{1/2} \to 0$ so we focus on the first term on the right side of (A.3). It is sufficient to show that $N^{-4}\mathrm{Var}\{E[\Delta_N^2|G]\} \to 0$ due to $\mathrm{Var}\{E[\Delta_N^2|G]\} \ge \mathrm{Var}\{E[\Delta_N^2|W_N]\}$. Observe that

$$E[\Delta_N^2|G] = 2\sum_{r<s}^{n}\sum \Delta_{r,s}^2 \big/ n(n-1), \quad \text{(A.4)}$$

where $\Delta_{r,s}^2$ is the expected value of $(\tilde{R}_{r2} + \tilde{R}_{s2} - R_{r2} - R_{s2})^2$ with respect to the switching of one randomly selected element from each of the pairs $r$ and $s$ of $G$. Taking the variance gives

$$\begin{aligned}\mathrm{Var}\{E[\Delta_N^2|G]\} &= 8\,\mathrm{Var}(\Delta_{1,2}^2)/N(N-2) \\ &\quad + 8(N-4)\,\mathrm{Cov}(\Delta_{1,2}^2, \Delta_{1,3}^2)/N(N-2) \\ &\quad + (N-4)(N-6)\,\mathrm{Cov}(\Delta_{1,2}^2, \Delta_{3,4}^2)/N(N-2). \quad \text{(A.5)}\end{aligned}$$

The fact that $|\Delta_{r,s}| \le N - 3$ is sufficient to ensure that the first two terms on the right side in (A.5) go to zero when multiplied by $N^{-4}$. For the third term we note that $\mathrm{Cov}(\Delta_{1,2}^2, \Delta_{3,4}^2) = \mathrm{Cov}[g(\pi_1), g(\pi_2)]$ for some function $g(\cdot)$, where $(\pi_1, \pi_2)$ is a pair of permutations of size $m = 4$ randomly sampled from $\{1, 2, \ldots, N\}$ with no common element. We claim that

$$\mathrm{Cov}[g(\pi_1), g(\pi_2)] = -\mathrm{Cov}_1[g(\pi_1), g(\pi_2)](p_{N,m}^{-1} - 1), \quad \text{(A.6)}$$

where the covariance on the right side of (A.6) is taken over randomly sampled pairs of permutations that have at least one common element, and $p_{N,m} = (N-m)!(N-m)!/N!(N-2m)!$ is the probability that a pair of independently sampled permutations has no common element. This identity follows from the fact that

$$\begin{aligned}0 &= \mathrm{Cov}_0[g(\pi_1), g(\pi_2)] \\ &= p_{N,m}\,\mathrm{Cov}[g(\pi_1), g(\pi_2)] \\ &\quad + (1 - p_{N,m})\,\mathrm{Cov}_1[g(\pi_1), g(\pi_2)], \quad \text{(A.7)}\end{aligned}$$

where $\mathrm{Cov}_0$ refers to sampling the two permutations independently. Thus, $|\mathrm{Cov}(\Delta_{1,2}^2, \Delta_{3,4}^2)| \le (N-3)^4(p_{N,4}^{-1} - 1)$ where $p_{N,4}^{-1} - 1 = \mathrm{O}(N^{-1})$. This proves Theorem 1 for the case that $N$ is even.

Next let $N = 2n + 1$ and construct $\dot{T}_{N-1}$ by replacing the sequence value $N$ by the (smaller) unmatched value if $N$ is matched in $G$; otherwise take $\dot{T}_{N-1} = T_N$. Now $T_{N-1}$ and $\dot{T}_{N-1}$ have the same distribution, and it is easy to verify that $W_N - \dot{W}_{N-1}$ goes to zero in probability. Because $\sigma_N^{-1}(\tilde{T}_N - \mu_N)\mathrm{N}(0, 1)$ over even integers, this proves Theorem 1 for the case that $N$ is odd.

### Proof of Theorem 2

We begin by observing that every pair of orthogonal MNBMs on $N$ vertices $(G, \tilde{G})$ can be decomposed into $c$ disjoint components (cycles), each of which contains an even number of at least four vertices, where $1 \le c \le \lfloor N/4 \rfloor$. This can be seen by considering $(G, \tilde{G})$ as a regular graph of degree 2. To construct the decomposition explicitly take $v_1^{(1)} = 1$ and transition to its matched vertex $v_2^{(1)}$ in $G$. Next, transition to the matched vertex of $v_2^{(1)}$ in $\tilde{G}$ (call it $v_3^{(1)}$) and keep alternating between $G$ and $\tilde{G}$ in this manner until the cycle is completed when it returns to $v_1^{(1)}$. It is possible that the entire set $\{1, 2, \ldots, N\}$ will be exhausted and the number of cycles is $c = 1$. If not, define a new cycle by taking $v_1^{(2)}$ to be the smallest number excluded from the first cycle. Repeat this process until all vertices are exhausted. We will say that the *cycle order* of $(G, \tilde{G})$ is an integer set $z = \{\varsigma_1, \ldots, \varsigma_c\}$ where $\varsigma_1 + \cdots + \varsigma_c = n, 2 \le \varsigma_1 \le \cdots \le \varsigma_c \le n$ with cycle $j$ having $2\varsigma_j$ vertices.

Let $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)$ denote the class of orthogonal MNBM pairs which we express as $(\mathcal{G}_N, \tilde{\mathcal{G}}_N) = \bigcup_z (\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ where $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ consists of all orthogonal MNBM $N$-graph pairs with cycle order $z$, the union taken over all cycle orders. Let $\mathcal{P}_N$ denote the class of permutations of $\{1, 2, \ldots, N\}$.

*Lemma A.1.* For every cycle order $z$ there exists a mapping from $\mathcal{P}_N$ onto $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ such that the pre-image contains a constant number of elements (depending on $z$).

*Proof.* Let $z = \{\varsigma_1, \ldots, \varsigma_c\}$ and take $\mathbf{b} = (b_1, b_2, \ldots, b_N)'$ to be any element of $\mathcal{P}_N$. For $j = 1, 2, \ldots, n$ let $s_r = \sum_{i=0}^{r} \varsigma_i$ where $\varsigma_0 \equiv 0$ and $r(j) = \min\{r : j \le s_r\}$. If $j = s_{r(j)}$ take $\eta(j) = 2s_{r(j)-1} + 1$; otherwise, $\eta(j) = 2j+1$. This uniquely determines a new permutation vector $\tilde{\mathbf{b}} = (\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_N)'$ where $\tilde{b}_{2j-1} = b_{2j}$ and $\tilde{b}_{2j} = b_{\eta(j)}$. Now let $G$ and $\tilde{G}$ contain edges that connect vertices $\{(b_{2j-1}, b_{2j}), j = 1, \ldots, n\}$ and $\{(\tilde{b}_{2j-1}, \tilde{b}_{2j}), j = 1, \ldots, n\}$, respectively. Note that every element of $(G, \tilde{G}) \in (\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ can be constructed in this manner for some choice of $\mathbf{b} \in \mathcal{P}_N$ and that the number of such vectors is a constant equal to $\omega_N(z) = 2^c \prod_{j=1}^{c} \varsigma_j \prod_{k=2}^{n} \mu_k!$ where $\mu_k = \#\{r : \varsigma_r = k\}$.

We will call $\mathbf{b}$ a *characteristic vector* of $(G, \tilde{G})$ and the set $\mathcal{B}$ of characteristic vectors its *characteristic set*. Let $\mathbf{X}$ be an $N \times d$ matrix where each row is an independent observation from a common probability distribution in $\mathbb{R}^p$ and consider a random orthogonal MNBM pair $(G, \tilde{G})(\mathbf{X}) \in (\mathcal{G}_N, \tilde{\mathcal{G}}_N)$ determined from $\mathbf{X}$. Let $z(\mathbf{X})$ denote the cycle order of $(G, \tilde{G})(\mathbf{X})$ and let $\mathcal{B}(\mathbf{X})$ denote the characteristic set. We will say that $(G, \tilde{G})(\mathbf{X})$ is *permutation equivariant* if, for every $N \times N$ permutation matrix $\mathbf{V}$, $z(\mathbf{VX}) = z(\mathbf{X})$ and $\mathcal{B}(\mathbf{VX}) = \{\mathbf{Vb} : \mathbf{b} \in \mathcal{B}(\mathbf{X})\}$. Note that for any characteristic set $\mathcal{B}$ the operation $\{\mathbf{Vb} : \mathbf{b} \in \mathcal{B}\}$ maps onto all characteristic sets of elements in $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ as $\mathbf{V}$ varies over all permutation matrices.

*Lemma A.2.* If $(G, \tilde{G})(\mathbf{X})$ is permutation equivariant, then $(G, \tilde{G})(\mathbf{X})$ is uniformly distributed over $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ conditional on $z(\mathbf{X}) = z$.

*Proof.* Let $\mathbf{X}_0$ denote an ordering of the rows of $\mathbf{X}$ that is invariant under row permutations (e.g., sorting on the $L^2$ norms of the rows). Then $\mathbf{X} = \mathbf{VX}_0$ where $\mathbf{V}$ is uniformly distributed over all $N \times N$ permutation matrices. Lemma A.1 implies, conditional on $z(\mathbf{X}) = z(\mathbf{X}_0) = z$, that $\mathcal{B}(\mathbf{X})$ is equally likely to be any of the possible characteristic sets that uniquely determine elements of $(\mathcal{G}_N, \tilde{\mathcal{G}}_N)_z$ and the claim now follows.

Because the distribution of $z(\mathbf{X})$ depends on the sampling distribution, $(G, \tilde{G})(\mathbf{X})$ is not distribution-free under the null hypothesis. It is true, however, that $G(\mathbf{X})$ and $\tilde{G}(\mathbf{X})$ each are marginally distributed as uniform over $\mathcal{G}_N$ regardless of the sampling distribution. It is also true, although less obviously, that the first two moments of SPM statistics obtained from $(G, \tilde{G})(\mathbf{X})$ do not depend on the sampling distribution:

*Lemma A.3.* Let $(T, \tilde{T})$ denote the SPM statistic pair corresponding to $(G, \tilde{G})(\mathbf{X})$. Under the conditions of Lemma A.2, $\mathrm{Cov}(T, \tilde{T}) = -N(N+1)/180$.

*Proof.* Notation from the proof of Lemma A.1 yields the following expressions:

$$T = \sum_{j=1}^{n} \max(b_{2j-1}, b_{2j}), \qquad \tilde{T} = \sum_{j=1}^{n} \max\bigl(b_{2j}, b_{\eta(j)}\bigr). \tag{A.8}$$

Condition on the cycle order and use the fact that $\eta(j)$ always is odd to obtain $E[T\tilde{T}] = Nm_{12,23} + N(N-4)m_{12,34}/4$, where $m_{12,23} = E[\max(b_1, b_2) \cdot \max(b_2, b_3)]$ and $m_{12,34} = E[\max(b_1, b_2) \cdot \max(b_3, b_4)]$. This shows that $E[T\tilde{T}]$ does not depend on the cycle order. Straightforward calculations yield $m_{12,23} = 7(N+1)(4N+3)/60$ and $m_{12,34} = 4(N+1)(5N+4)/45$ from which the result follows.

To conclude the proof of Theorem 2 we note that any pair of SPM statistics obtained from a recursively optimal $n$-ensemble of MNBMs based on an i.i.d. sample from a continuous distribution meets the permutation equivariance condition of Lemma A.2, and Lemma A.3 therefore applies.

*[Received September 2010. Revised July 2011.]*

## REFERENCES


Anderson, I. (1972), "Sufficient Conditions for Matchings," *Proceedings of the Edinburgh Mathematical Society* (*Series 2*), 18, 129–136. [1618]

Arya, S., Mount, D., Netanyahu, N., Silverman, R., and Wu, A. (1998), "An Optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions," *Journal of the ACM*, 45, 891–923. [1623]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57 (1), 289–300. [1617]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, 29 (4), 1165–1188. [1617]

Bentley, J., and Friedman, J. (1978), "Fast Algorithms for Constructing Minimal Spanning Trees in Coordinate Spaces," *IEEE Transactions on Computers*, 27 (2), 97–105. [1623]

Cook, W., and Rohe, A. (1999), "Computing Minimum-Weight Perfect Matchings," *INFORMS Journal on Computing*, 11 (2), 138–148. [1622]

Derigs, U. (1988), "Solving Non-Bipartite Matching Problems via Shortest Path Techniques," *Annals of Operations Research*, 13, 225–261. [1618,1622, 1623]

Edmonds, J. (1965a), "Maximum Matching and a Polyhedron with 0–1 Vertices," *Journal of Research at the National Bureau of Standards*, 69B, 125–130. [1622]

——— (1965b), "Path, Trees, and Flowers," *Canadian Journal of Mathematics*, 17, 449–467. [1622]

Filliben, J., Kafadar, K., and Shier, D. (1983), "Testing for Homogeneity of Two-Dimensional Surfaces," *Mathematical Modeling*, 4, 167–189. [1616]

Fricker, R., Jr. (2007), "Directionally Sensitive Multivariate Statistical Process Control Procedures With Applications to Syndromic Surveillance," *Advances in Disease Surveillance*, 3, 1–17. [1623]

Fricker, R., Jr., Knitt, M., and Hu, C. (2008), "Comparing Directionally Sensitive MCUSUM and MEWMA Procedures With Application to Biosurveillance," *Quality Engineering*, 20, 478–494. [1623]

Friedman, J., and Rafsky, L. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics*, 7 (4), 697–717. [1615,1616,1618]

——— (1981), "Graphics for the Multivariate-Two-Sample Problem," *Journal of the American Statistical Association*, 76, 374, 277–287. [1615,1616]

——— (1983), "Graph-Theoretic Measures of Multivariate Association and Prediction," *The Annals of Statistics*, 11 (2), 377–391. [1618]

Friedman, J., Bentley, J., and Finkel, R. (1977), "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Transactions on Mathematical Software*, 3, 209–226. [1623]

Gabow, H. (1973), "Implementation of Algorithms for Maximum Matching on Non-Bipartite Graphs," unpublished Ph.D. thesis, Stanford University, Computer Science Dept. [1622]

Hall, P., and Tajvidi, N. (2002), "Permutation Tests for Equality of Distributions in High-Dimensional Settings," *Biometrika*, 89 (2), 359–374. [1616]

Heller, R., Jensen, S., Rosenbaum, P., and Small, D. (2010), "Sensitivity Analysis for the Cross-Match Test, With Applications in Genomics," *Journal of the American Statistical Association*, 105 (1), 1005–1013. [1615]

Henze, N. (1988), "A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences," *The Annals of Statistics*, 16, 772–783. [1616]

Henze, N., and Penrose, M. (1999), "On the Multivariate Runs Test," *The Annals of Statistics*, 27 (1), 290–298. [1618]

Hess, R., Duke, A., and Kogut, D. (2001), "The IMD HUMS as a Tool for Rotorcraft Health Management and Diagnostics," *2001 IEEE Aerospace Conference Proceedings*, 6, 3039–3058. [1615]

James, B., James, K., and Siegmund, D. (1992), "Asymptotic Approximations for Likelihood Ratio Tests and Confidence Regions for a Change-Point in the Mean of a Multivariate Normal Distribution," *Statistica Sinica*, 2 (1), 69–90. [1620]

Joner, M., Jr., Woodall, W., Reynolds, M., Jr., and Fricker, R., Jr. (2008), "A One-Sided MEWMA Chart for Health Surveillance," *Quality and Reliability Engineering International*, 24, 503–519. [1623]

Kolmogorov, V. (2009), "Blossom V: A New Implementation of a Minimum Cost Perfect Matching Algorithm," *Mathematical Programming Computation*, 1 (1), 43–67. [1622]

Lu, B. and Rosenbaum, P. R. (2004), "Optimal Pair Matching With Two Control Groups," *Journal of Computational and Graphical Statistics*, 13 (2), 422–434. [1615]

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association*, 96, 1245–1253. [1615]

Maa, J., Pearl, D., and Bartoszynski, R. (1996), "Reducing Multidimensional Two-Sample Data to One-Dimensional Interpoint Comparisons," *The Annals of Statistics*, 24 (3), 1069–1074. [1616]

Mehlhorn, K., and Schäfer, G. (2002), "Implementation of $O(nm \log n)$ Weighted Matchings in General Graphs: The Power of Data Structures," *Journal of Experimental Algorithmics*, 7 (4), 1–19. [1622]

National Center for Health Statistics (2000), "Data File Documentation," Compressed Mortality File, 1968–1988 (CD-ROM Series 20, No. 2A), Hyattsville, MD. [1620]

R Development Core Team (2009). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, available at *http://www.R-project.org*. [1618]

Rinott, Y., and Rotar, V. (2000), "Normal Approximations by Stein's Method," *Decisions in Economics and Finance*, 23, 15–29. [1623]

Rogerson, P., and Yamada, Y. (2004), "Monitoring Change in Spatial Patterns of Disease: Comparing Univariate and Multivariate Cumulative Sum Approaches," *Statistics in Medicine*, 23, 2195–2214. [1615,1620]

Rosenbaum, P. (2005), "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society, Ser. B*, 67 (4), 515–530. [1615,1616,1618]

Ruth (2009), "Applications of Assignment Algorithms to Nonparametric Tests for Homogeneity," unpublished Ph.D. thesis, Naval Postgraduate School, Dept. of Operations Research. [1617,1618,1620,1622,1623]

Schilling, M. (1986), "Multivariate Two-Sample Tests Based on Nearest Neighbors," *Journal of the American Statistical Association*, 81 (395), 799–806. [1616]

Shorack, G., and Wellner, J. (1986), *Empirical Processes With Applications to Statistics*, New York: Wiley. [1619]

Stein, C. (1986), *Approximate Computation of Expectations*. *Lecture Notes—Monograph Series 7*, ed. S. Gupta, Hayward, CA: Institute of Mathematical Statistics. [1623]

Tartakovsky, A., Rozovskii, B., Blažek, R., and Kim, H. (2006a), "A Novel Approach to Detection of Intrusions in Computer Networks via Adaptive Sequential and Batch-Sequential Change-Point Detection Methods," *IEEE Transactions on Signal Processing*, 54 (9), 3372–3382. [1615]

——— (2006b), "Detection of Intrusions in Information Systems by Sequential Change-Point Methods," *Statistical Methodology*, 3 (3), 252–293. [1615]

Zou, H., Hastie, T., and Tibshirani, R. (2006), "Sparse Principal Components Analysis," *Journal of Computational and Graphical Statistics*, 15, 265–286. [1623]