J. R. Statist. Soc. B (2005)
67, *Part* 4, *pp.* 515–530

An exact distribution-free test comparing two multivariate distributions based on adjacency

Paul R. Rosenbaum

University of Pennsylvania, Philadelphia, USA

[Received June 2004. Revised March 2005]

Summary. A new test is proposed comparing two multivariate distributions by using distances between observations. Unlike earlier tests using interpoint distances, the new test statistic has a known exact distribution and is exactly distribution free. The interpoint distances are used to construct an optimal non-bipartite matching, i.e. a matching of the observations into disjoint pairs to minimize the total distance within pairs. The cross-match statistic is the number of pairs containing one observation from the first distribution and one from the second. Distributions that are very different will exhibit few cross-matches. When comparing two discrete distributions with finite support, the test is consistent against all alternatives. The test is applied to a study of brain activation measured by functional magnetic resonance imaging during two linguistic tasks, comparing brains that are impaired by arteriovenous abnormalities with normal controls. A second exact distribution-free test is also discussed: it ranks the pairs and sums the ranks of the cross-matched pairs.

Keywords: Combinatorial optimization; Distribution-free test; Non-bipartite matching; Nonparametric test; Occupancy distribution; Rank test

1. Introduction: goal; notation; review

1.1. Goal and notation: an exact distribution-free permutation test for comparing multivariate responses in two groups

There are $N \geqslant 4$ subjects consisting of $n \geqslant 2$ subjects of one type, say treated subjects, and $m = N - n \geqslant 2$ subjects of a second type, say control subjects, and a multivariate response $\mathbf{Y}_i$ is recorded for each subject. The response $\mathbf{Y}_i$ may be, but need not be, a vector of continuous measurements, and, when it is a vector, the dimension of $\mathbf{Y}_i$ may be greater than N. Indeed, $\mathbf{Y}_i$ may be of infinite dimension, i.e. a $\mathbf{Y}_i$ may be a continuous curve, or a two- or three-dimensional image, e.g. the result of medical imaging. Alternatively, each $\mathbf{Y}_i$ may be a discrete sequence, such as the sequence of amino-acids in a protein, or the sequence of bases in strands of deoxyribonucleic acid; e.g. Durbin *et al.* (1999). There is some definition of a distance between two $\mathbf{Y}_i$s. The null hypothesis states that the distribution of $\mathbf{Y}_i$ is the same for all subjects, both treated and control.

Formally, there are N independent trials, $i = 1, \ldots, N$, and on each trial i a coin is flipped returning $Z_i = 1$ for heads with probability π and $Z_i = 0$ with probability $1 - \pi$; if $Z_i = 1$, then $\mathbf{Y}_i$ is drawn from a distribution $F(\cdot)$, but if $Z_i = 0$ then $\mathbf{Y}_i$ is drawn from a distribution $G(\cdot)$. The *null hypothesis* asserts that the treated distribution $F(\cdot)$ and the control distribution $G(\cdot)$ are the same: $F(\cdot) = G(\cdot)$. Write $\tilde{\mathbf{Y}} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ and $\mathbf{Z} = (Z_1, \ldots, Z_N)$. The number of treated subjects, $n = \Sigma_{i=1}^{N} Z_i$, is ancillary, so the test will condition on n, eliminating π. If $\mathbf{Y}_i$ were a

Address for correspondence: Paul R. Rosenbaum, Department of Statistics, Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA.
E-mail: rosenbaum@stat.wharton.upenn.edu

 1369–7412/05/67515

scalar, then this yields the conventional, unidimensional two-sample problem. This notation for a two-sample problem emphasizes that the subscript i carries no information, that the information is in $(\tilde{\mathbf{Y}}, \mathbf{Z})$, and the notation is convenient in Section 3.4 where $N \to \infty$ under alternative hypotheses, so that n increases. The same effect can be produced in empirical data simply by assigning subscripts i at random to observations $(\mathbf{Y}_i, Z_i)$, so the subscripts carry no information. Outside Section 3.4, all distributions condition on n, but this will not appear explicitly in the notation.

The goal is an exact distribution-free test, i.e., under the null hypothesis, the exact distribution of the test statistic should be a usable known distribution, and that known null distribution may depend on the sample sizes (n, m) but not on the unknown distribution $F(\cdot) = G(\cdot)$. For instance, in the case of a one-dimensional response with continuous distributions, the Wilcoxon rank sum test, Mood's median test, the Wald–Wolfowitz runs test, the Kolmogorov–Smirnov test, Wilks's (1962) 'empty block' test (page 446) and many other nonparametric tests are exact and distribution free. Certain multivariate, exact, distribution-free two-sample tests were discussed by Anderson (1966), section 4.

Given n, there are $\binom{N}{n}$ possible values of $\mathbf{Z}$, i.e. N-dimensional vectors $\mathbf{z}$ with binary coordinates such that $n = \sum_{i=1}^{N} z_i$. Place these $\binom{N}{n}$ possible $\mathbf{z}$ in a set Ω. Under the null hypothesis, $F(\cdot)$ and $G(\cdot)$ are identical, so $\Pr(\mathbf{Z} = \mathbf{z} | \tilde{\mathbf{Y}}) = \binom{N}{n}^{-1}$ for each $\mathbf{z} \in \Omega$.

The paper is organized as follows. Section 1.2 reviews the existing literature on multivariate nonparametric tests using interpoint distances, and Section 1.3 reviews algorithms for constructing an optimal, non-bipartite matching. Section 2 presents a motivating example involving brain activation during linguistic tasks in impaired and normal brains; in particular, a minimum distance non-bipartite matching is constructed and the cross-match statistic is defined. Properties of the cross-match statistic are developed in Section 3, including its exact distribution, its exact distribution-free property, its moments, a large sample approximation to the null distribution and some properties of the test under alternative hypotheses. The cross-match test is applied to the example in Section 3.2 and compared with Friedman and Rafsky's (1979) multivariate runs test that uses a minimum spanning tree instead of a non-bipartite matching. In the univariate case, a small simulation in Section 3.5 compares the power of the exact cross-match test with the power of the exact Kolmogorov–Smirnov test. Some matches are better than others, and perhaps better matches deserve more weight; therefore, in Section 4, the cross-match rank sum statistic is introduced, which ranks the pairs in some way and sums the ranks of cross-matched pairs.

1.2. Review of large sample tests based on interpoint distances

Several large sample tests have been proposed for the multidimensional problem by using $\binom{N}{2}$ 'distances' between the $\mathbf{Y}_i$s, say a distance δ_{ij} between the $\mathbf{Y}_i$ and $\mathbf{Y}_j$, $i < j$, where $\delta_{ij} \geqslant 0$ with equality if and only if $\mathbf{Y}_i = \mathbf{Y}_j$. Note carefully that δ_{ij} is calculated from the $\mathbf{Y}_i$s but does not use the Z_is.

In an early paper, Weiss (1959) drew the largest possible non-overlapping (open) spheres around the responses of each treated subject and counted the number of control responses within each sphere. In particular, he suggested using the number of spheres containing no controls as a test statistic, describing the test as a multivariate analogue of the Wald–Wolfowitz runs test, but he noted that the null distribution of this quantity is neither distribution free nor known, and he left aspects of practical implementation undeveloped.

In a clever, interesting, paper, Friedman and Rafsky (1979) used distances δ_{ij} to construct a minimum spanning tree, then removed the edges that connect a treated subject to a control

and used the resulting number of disjoint subtrees as a test statistic. Under the null hypothesis, they computed the conditional expectation and variance of their test statistic given $\tilde{\mathbf{Y}}$ and used these in conjunction with a permutational central limit theorem to obtain a normal approximation. The null conditional variance depends on $\tilde{\mathbf{Y}}$ through a structural feature of the tree, namely the number C of edge pairs that share a common node; as a result, the null distribution of their statistic depends on the unknown distribution of $\tilde{\mathbf{Y}}$, so it is not distribution free. Schilling (1986) and Henze (1988) used one or more nearest neighbours of each observation, counting the number of times that the nearest neighbours come from the same group. Maa *et al*. (1996) showed that, under mild conditions, two multivariate distributions differ if and only if the distributions of interpoint distances differ within and between the distributions. Henze and Penrose (1999) showed that the Friedman–Rafsky test is consistent against all alternatives.

The new procedure proposed in the current paper is similar in spirit to the several procedures just described but, unlike these procedures, the new test statistic A_1 in Section 3

(a) has a known, exact null distribution,
(b) this null distribution is exactly distribution free, i.e. the null distribution $\Pr(A_1 \leqslant a)$ does not depend on the common unknown distribution $F(\cdot) = G(\cdot)$, and
(c) the conditional null distribution of A_1 given $\tilde{\mathbf{Y}}$ equals its unconditional null distribution, $\Pr(A_1 \leqslant a|\tilde{\mathbf{Y}}) = \Pr(A_1 \leqslant a)$.

1.3. Review of optimal non-bipartite matching

The procedure of Friedman and Rafsky (1979) computed a minimum spanning tree from the distances δ_{ij}. In contrast, the procedure proposed here computes a minimum distance non-bipartite matching. Suppose first that N is even. Using the $\binom{N}{2}$ distances δ_{ij}, a minimum distance non-bipartite matching divides the N subjects into $I = N/2$ non-overlapping pairs of two subjects in such a way as to minimize the total of the $N/2$ distances within the $N/2$ pairs.

The minimum distance non-bipartite matching problem is a combinatorial optimization problem that can be solved quickly with a polynomial time algorithm. See Galil (1986) for a survey and see Papadimitriou and Steiglitz (1982), section 11.3, for one text-book discussion. In particular, the number of arithmetic operations that are required to find an optimal non-bipartite matching of N subjects is $O(N^3)$; see Papadimitriou and Steiglitz (1982), page 266. For comparison, if we multiply two $N \times N$ matrices in the conventional way, the calculation also requires $O(N^3)$ arithmetic operations.

An implementation of optimal non-bipartite matching in C may be downloaded free from `http://elib.zib.de/pub/Packages/mathprog/matching/weighted/index.html`. That implementation maximizes a total benefit β_{ij}, rather than minimizing a total distance δ_{ij}, so to use that implementation we must define $\beta_{ij} = \max_{b,c}(\delta_{bc}) - \delta_{ij}$. An implementation in Fortran was published by Derigs (1988). In statistics, optimal non-bipartite matching has been used to match with doses (Lu *et al*., 2001), to match with two control groups (Lu and Rosenbaum, 2004), and to match before random assignment in experiments (Greevy *et al*., 2004).

If N is odd, create a pseudosubject $N+1$, with $\delta_{i,N+1} = 0$ for $i = 1, \ldots, N$. Optimally match the $N+1$ subjects, and discard the one pair containing the pseudosubject. This results in a matching with $(N-1)/2$ pairs that minimizes the total distance between all matchings of the original N subjects into $(N-1)/2$ pairs which discard one subject. Rather than having separate notation for even and odd N, adopt the *convention* that, for odd N, all of the notation is adjusted to refer to the $N-1$ remaining subjects in $I = (N-1)/2$ pairs, i.e. the notation always refers to the case of even N, perhaps after discarding one subject in this way.

2. Motivating example: functional magnetic resonance imaging of brain activity

2.1. Brain activity during language tasks in subjects with arteriovenous abnormalities

Language tasks typically engage the brain's left hemisphere. What happens if the left hemisphere is impaired? Using functional magnetic resonance imaging (FMRI), Lehéricy *et al*. (2002) examined patients with arteriovenous malformations in the left hemisphere and normal controls. All patients and controls were right handed. Subjects performed various language tasks, including story listening and sentence repetition while undergoing FMRI. In the story listening task, subjects listened to a story. In the sentence repetition task, subjects listened to a sentence and then repeated it to themselves mentally. On the basis of FMRI, a continuous measure called the laterality index was computed, which measured the relative activation of the left and right hemispheres during the tasks. For instance, the laterality index was 1 if all the increased activation was on the left, −1 if all was on the right and 0 if the activation on both sides was about the same. Specifically, the laterality index was $(L - R)/(L + R)$ where L was the number of activated pixels in the left hemisphere's temporal lobe and R was the number activated in the right temporal lobe. See Lehéricy *et al*. (2002) for many details of this calculation.

Table 1 displays their data. For instance, during listening to a story, control 5 had increased activity only in the left temporal lobe but during sentence repetition had slightly more activation on the right than on the left. In contrast, patient 18 had increased activity only on the right. (Three patients had measurements both before and after embolization, a form of treatment, and the pre-embolization values appear in Table 1. The FMRI failed to produce a laterality index for three subjects, and they do not appear in Table 1.)

Table 1. Laterality index in the temporal lobe during story listening and sentence repetition tasks, for patients P and healthy controls C†

Subject identifier	*Group*	*Laterality index for the following tasks:*			
		Story listening	*Sentence repetition*	*Match identifier*	*Distance*
1	C	0.47	0.03	7	0.32
2	C	0.39	0.11	9	0.04
3	C	0.47	0.16	16	4.04
4	C	0.78	−0.10	5	0.23
5	C	1.00	−0.05	4	0.23
6	C	1.00	0.16	8	0.71
7	C	0.54	0.12	1	0.32
8	C	1.00	0.40	6	0.71
9	C	0.38	0.04	2	0.04
10	P	1.00	0.71	12	0.47
11	P	0.27	0.01	14	0.17
12	P	0.63	0.21	10	0.47
13	P	0.22	−0.18	18	0.58
14	P	0.00	−0.08	11	0.17
15	P	−1.00	−0.35	17	0.06
16	P	−0.42	0.26	3	4.04
17	P	−1.00	−0.60	15	0.06
18	P	−1.00	−1.00	13	0.58

†Source: Lehéricy *et al*. (2002).

Is there a systematic difference between patients and controls? If there is, it is either quite noisy or not of a simple form. For instance, patient 10 had more activation on the left than any control, patients 15, 17 and 18 had more activation on the right than any control, patient 14 was closer to balanced activation in the two hemispheres, $(0, 0)$, than any control, and patient 16 was unique in having more right activation with story listening (-0.42) and more left activation with sentence repetition (0.26). The null hypothesis asserts that the distribution of laterality indices is the same for patients and controls. Is the null hypothesis plausible?

2.2. Minimum distance pairing using $\mathbf{Y}_i$

A distance δ_{ij} between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ was defined as follows. The laterality indices for story listening were ranked from 1 to 18 with average ranks for ties. The same was done for sentence repetition. Each subject i now has two ranks, say $\mathbf{R}_i$ for the ranks of $\mathbf{Y}_i$. The two ranks exhibit a fairly strong relationship: the Spearman rank correlation is 0.63. Then δ_{ij} is defined to be the Mahalanobis distance between the two ranks for i and the two ranks for j, i.e.

$$\delta_{ij} = (\mathbf{R}_i - \mathbf{R}_j)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{R}_i - \mathbf{R}_j)$$

where $\mathbf{S}$ is the sample variance–covariance matrix of the ranks $\mathbf{R}_i$. There are $\binom{18}{2} = 153$ distinct pairwise distances δ_{ij}, $i < j$. The Mahalanobis distance takes appropriate account of the correlation and, because the correlation is fairly high, the Mahalanobis distance judges things quite differently from how the Euclidean distance would, i.e. the Mahalanobis distance that is attached to $\mathbf{R}_i - \mathbf{R}_j$ is constant on positively sloped ellipses centred at the origin, whereas the Euclidean distance is constant on circles.

The $N = 18$ subjects were optimally paired into nine pairs to minimize the total of the nine distances within pairs. The computations used the C algorithm that was mentioned in Section 1.3. The column 'Match identifier' in Table 1 gives the identity of the match for each subject. For instance, subject 1 is paired with subject 7, and of course subject 7 is paired with subject 1. The column 'Distance' in Table 1 gives the Mahalanobis distance.

Fig. 1 and Table 2 describe the $\binom{18}{2} = 153$ pairwise Mahalanbois distances before matching and the nine distances within the nine pairs. The median distance after matching is about a tenth of the median distance before matching. With one exception, all the distances after matching

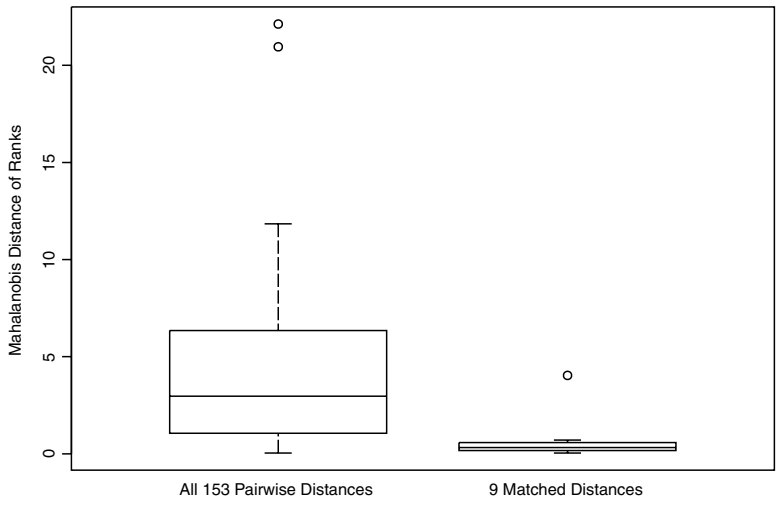

Fig. 1. Distances before and after optimal matching

Table 2. Five-number summaries of Mahalanobis distances using ranks, for all 153 pairs before matching and for nine matched pairs

	Mahalanobis distances for the following pairs:			
	All pairs		*Matched pairs*	
Median	2.97		0.32	
Quartiles	1.06	6.35	0.17	0.58
Extremes	0.04	22.12	0.04	4.04

are below the lower quartile of the distances before matching. The one exception is the pairing of control 3 with patient 16. As noted in Section 2.1, patient 16 exhibits a unique pattern, with more right activation for story listening and more left activation for sentence repetition. Judged using the Mahalanobis distance, patient 16 is not close to anyone else: the median distance subject between 16 and the other 17 subjects is 8.12 and the minimum distance is $\delta_{3,16} = 4.04$ with control 3. The two extremes in the box plot of all 153 distances are both with patient 16, specifically $\delta_{4,16} = 20.95$ and $\delta_{5,16} = 22.12$ and, of the 8/153 distances $\delta_{ij} \geqslant 10$, five are distances between patient 16 and another subject.

2.3. Graphical motivation for the test

Fig. 2 is a scatterplot of the laterality index for sentence repetition against the laterality index for story listening, with patients indicated by a triangle and controls indicated by a circle. Matched subjects are connected by a line. Keep in mind that Fig. 2 depicts Euclidean distance between the $\mathbf{Y}_i$s, which is quite different from Mahalanobis distance between the ranks $\mathbf{R}_i$s, owing mostly

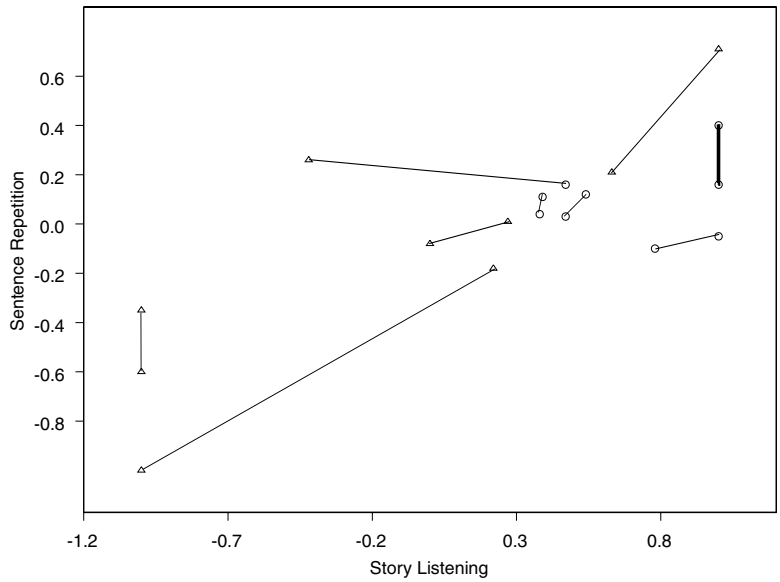

Fig. 2. Laterality index for sentence repetition and story listening for patients (△) and controls (○), with the optimal non-bipartite match

to the fairly strong correlation. For instance, the longest line in Fig. 2, the line that begins in the lower left-hand corner, connects patients 13 and 18 with $\delta_{13,18} = 0.58$, which is only the third largest distance among the nine pairs in Table 1 and is about half as large as the lower quartile of the 153 distances in Table 2 before matching. The short vertical bar on the right-hand side of Fig. 2 connects controls 6 and 8 with a distance between the ranks of $\delta_{6,8} = 0.71 > \delta_{13,18}$. The one segment with a negative slope, connecting a triangle in the upper left to a circle further right, connects patient 16 to control 3 with the largest distance, $\delta_{3,16} = 4.04$.

If the $\mathbf{Y}_i$s for patients are very different from the $\mathbf{Y}_i$s for controls, then we expect relatively few patients to be matched to controls, i.e. relatively few cross-matches. In Fig. 2, only one patient is matched to a control. Because $n = m = 9$, it was inevitable that at least one patient would be matched to a control, so Fig. 2 exhibits the smallest possible number of cross-matches. Counting cross-matches is the basis for the test.

3. The cross-match test

3.1. Exact null distribution

By the convention at the end of Section 1.3, there is an even number N of subjects. These subjects are optimally matched into $I = N/2$ non-overlapping pairs, using distances δ_{ij} that were computed from the responses $\tilde{\mathbf{Y}}$ alone, without using the group indicators $\mathbf{Z}$. Moreover, as discussed in Section 1.1, under the null hypothesis, $\Pr(\mathbf{Z} = \mathbf{z}|\tilde{\mathbf{Y}}) = \binom{N}{n}^{-1}$ for each $\mathbf{z} \in \Omega$.

Let A_k be the number of pairs with exactly k treated subjects, $k = 0, 1, 2$. In Table 1 and Fig. 2, $A_0 = 4$, $A_1 = 1$ and $A_2 = 4$. Because $A_0 + A_1 + A_2 = I$ and $A_1 + 2A_2 = n$, it follows that A_1 determines $A_2 = (n - A_1)/2$ and $A_0 = I - (n + A_1)/2$. Also, if n and m are both even, then A_1 can take even values from 0 to $\min(m, n)$, whereas, if n and m are both odd, then A_1 can take odd values from 1 to $\min(m, n)$. The null hypothesis of no effect is tested by using A_1. If $a_0 \geqslant 0$, $a_1 \geqslant 0$ and $a_2 \geqslant 0$ with $a_0 + a_1 + a_2 = I$ and $a_1 + 2a_2 = n$, then there are $2^{a_1} I!/a_0!\, a_1!\, a_2!$ treatment assignments $\mathbf{z} \in \Omega$ in which exactly a_1 pairs contain one treated subject and one control, and under the null hypothesis these each have probability $\binom{N}{n}^{-1}$, so the null distribution of A_1 is given by

$$\Pr(A_1 = a_1|\tilde{\mathbf{Y}}) = \frac{2^{a_1} I!}{\binom{N}{n} a_0!\, a_1!\, a_2!} = \pi_{a_1}, \tag{1}$$

say. Distribution (1) is a very special case of a restricted occupancy distribution; see Johnson *et al.* (1997), section 4.3, pages 186–187. Because distribution (1) depends on (m, n) but not on $\tilde{\mathbf{Y}}$, it follows that

$$\Pr(A_1 = a_1|\tilde{\mathbf{Y}}) = \Pr(A_1 = a_1). \tag{2}$$

Now equation (2) says that A_1 is distribution free under the null hypothesis, and distribution (1) gives the exact distribution. The null hypothesis is rejected if A_1 is small.

3.2. Example, continued: functional magnetic resonance imaging of brain activity

In the example, $n = m = 9$, $N = n + m = 18$ and there is one cross-match, $A_1 = 1$. Table 3 gives the distribution (1) of A_1 under the null hypothesis that the distribution of $\mathbf{Y}_i$ is the same for patients and controls. Each row is a possible occupancy distribution (A_0, A_1, A_2) for the types of pair, together with its probability. For Fig. 2, the significance level is $\Pr(A_1 \leqslant 1) = 0.0259$, so the null hypothesis is not plausible.

The cross-match test may be compared with the multivariate runs test that was proposed by Friedman and Rafsky (1979), and the formulae in this paragraph are from them. Fig. 3 shows the

Table 3. Exact null distribution of the number of cross-matches, $n = m = 9$†

A_0	A_1	A_2	$Pr(A_1 = k)$	$Pr(A_1 \leqslant k)$
4	1	4	0.0259	0.0259
3	3	3	0.2764	0.3023
2	5	2	0.4976	0.7999
1	7	1	0.1896	0.9895
0	9	0	0.0105	1.0000

†Rows are possible occupancy distributions and their probabilities.

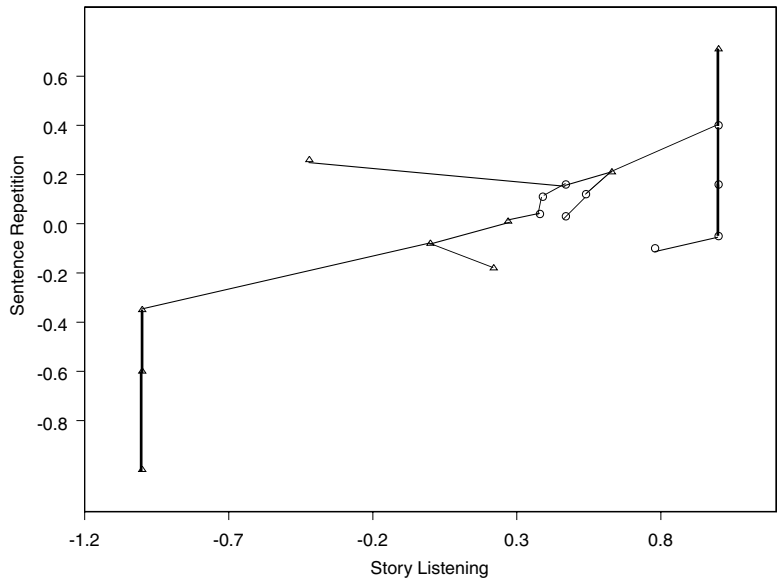

Fig. 3. Laterality index for sentence repetition and story listening for patients (△) and controls (○), with the minimum spanning tree

laterality indices together with a minimum spanning tree computed by using Kruskal's algorithm applied to the same set of distances δ_{ij}. If edges connecting a patient to a control are removed, there are $R = 7$ subtrees. Under the null hypothesis, $E(R) = 2mn/N + 1 = 2 \times 9 \times 9/18 + 1 = 10$. The variance of R given $\tilde{\mathbf{Y}}$ depends on $\tilde{\mathbf{Y}}$ through the number C of pairs of edges in the tree that share a common node. Hence, the null distribution of R depends on the unknown common distribution $F = G$. In Fig. 3, the number of edges that share a common node is $C = 8 + 4 \times 3 = 20$, as there are eight pairs of edges that share one node, and four triples of edges that share a node. Friedman and Rafsky refer $\{R - E(R)\}/\sqrt{\text{var}(R|C)} = (7 - 10)/\sqrt{4.094} = -1.483$ to the normal distribution, where $\sqrt{\text{var}(R|C)}$ is defined by their expression (14). By this measure, the null hypothesis would be judged to be marginally plausible.

3.3. *Moments; normal approximation*

The null expectation and variance of the test statistic A_1 have simple forms given by proposition 1 below. Renumber the N subjects after pairing them so that $\mathbf{Y}_{2i-1}$ is paired to $\mathbf{Y}_{2i}$ for $i = 1, \ldots, I$.

Write $B_i = Z_{2i-1} + Z_{2i} - 2Z_{2i-1}Z_{2i}$, so that $B_i = 1$ if pair i contains one treated subject and one control and $B_i = 0$ otherwise, and $A_1 = \Sigma_{i=1}^{I} B_i$. The moments of the B_i are given in lemma 1 below; they are used in the proof of proposition 2 and again in Section 4.

Lemma 1. Under the null hypothesis with $\Pr(\mathbf{Z}|\tilde{\mathbf{Y}}) = \binom{N}{n}^{-1}$,

$$E(B_i) = \frac{2nm}{N(N-1)} = \theta,$$

say,

$$\operatorname{var}(B_i) = \theta(1-\theta),$$

$$E(B_i B_j) = \frac{4n(n-1)m(m-1)}{N(N-1)\,(N-2)\,(N-3)} = \gamma,$$

say, for $i \neq j$, and

$$\operatorname{cov}(B_i, B_j) = \gamma - \theta^2$$

for $i \neq j$.

Proof. Write

$$\zeta_k = \frac{n(n-1)(n-2)\ldots(n-k+1)}{N(N-1)(N-2)\ldots(N-k+1)},$$

e.g. $\zeta_1 = n/N$ and $\zeta_2 = n(n-1)/N(N-1)$, and note that the product of k distinct Zs has expectation ζ_k. Hence, $E(B_i) = 2(\zeta_1 - \zeta_2) = \theta$, say, so $\operatorname{var}(B_i) = \theta(1-\theta)$. Also, for $i \neq j$,

$$\begin{aligned} E(B_i B_j) &= E\{(Z_{2i-1} + Z_{2i} - 2Z_{2i-1}Z_{2i})(Z_{2j-1} + Z_{2j} - 2Z_{2j-1}Z_{2j})\} \\ &= 4\zeta_2 - 8\zeta_3 + 4\zeta_4 = \gamma, \end{aligned}$$

so $\operatorname{cov}(B_i, B_j) = \gamma - \theta^2$.

Proposition 1. Under the null hypothesis with $\Pr(\mathbf{Z}|\tilde{\mathbf{Y}}) = \binom{N}{n}^{-1}$, the expectation and variance of A_1 are

$$E(A_1) = \frac{nm}{N-1}, \tag{3}$$

$$\operatorname{var}(A_1) = \frac{2n(n-1)m(m-1)}{(N-3)\,(N-1)^2}. \tag{4}$$

Proof. Because $A_1 = \Sigma_{i=1}^{I} B_i$, it follows that $E(A_1) = I\theta$ where $I = N/2$, and

$$\operatorname{var}(A_1) = I\theta(1-\theta) + I(I-1)(\gamma - \theta^2) = \frac{2n(n-1)m(m-1)}{(N-3)\,(N-1)^2}.$$

Proposition 2. Under the null hypothesis, the conditional distribution of A_1 given n in distribution (1) converges weakly to the normal distribution for $n/N \to \pi$:

$$\frac{A_1 - E(A_1)}{\sqrt{\operatorname{var}(A_1)}} \xrightarrow{\mathrm{D}} N(0,1). \tag{5}$$

Proof. From the structure that was assumed in Section 1.1, before conditioning on n, the Zs are independent and identically distributed binary random variables which equal 1 with probability π, and under the null hypothesis they are independent of $\tilde{\mathbf{Y}}$. Write $T_i = Z_{2i-1} + Z_{2i}$, so $n = \Sigma_{i=1}^{I} T_i$, and $A_1 = \Sigma_{i=1}^{I} B_i$, where the bivariate (T_i, B_i) are independent and identically

distributed before conditioning on n. Also, distribution (1) is the conditional distribution of A_1 given n. The multivariate central limit theorem (e.g. Rao (1973), section 2c.5(i) and section 2c.5(iv)) then implies that the unconditional joint distribution of (n, A_1) converges weakly to a bivariate normal distribution. Using this, theorem 2 of Holst (1979) implies that the conditional distribution of A_1 given n converges to the normal distribution.

3.4. Theoretical motivation for the test

This section provides some theoretical motivation for the cross-match test under alternative hypotheses in which $F(\cdot) \neq G(\cdot)$. In the current section, $N \to \infty$, so $n \to \infty$ also. Therefore, in the current section only, probability distributions are generated by the process in Section 1.1, but they are unconditional; unlike other sections, probabilities are not conditional given fixed n.

Suppose that $\mathbf{Y}$ takes only V values $\mathbf{y}_v$, $v = 1, \ldots, V$, with $\Pr(\mathbf{Y} = \mathbf{y}_v | Z = 1) = f_v$ and $\Pr(\mathbf{Y} = \mathbf{y}_v | Z = 0) = g_v$. Consider what happens as $N \to \infty$ with V fixed. If there are an even number, say N_v, of subjects with $\mathbf{Y} = \mathbf{y}_v$, then they can all be exactly paired to one another with a total distance of 0, but, if N_v is odd, then $N_v - 1$ subjects with $\mathbf{Y} = \mathbf{y}_v$ can be paired to one another, and the remaining one subject must be mismatched to someone with a different value of $\mathbf{Y}$. This means that at most V of the $N = \Sigma N_v$ subjects are mismatched for $\mathbf{Y}$, where $V/N \to 0$ as $N \to \infty$. The large sample behaviour of the statistic A_1 is not affected by these V mismatched pairs, so they are ignored throughout this section. Suppose that there are n_v treated subjects and m_v controls with $\mathbf{Y} = \mathbf{y}_v$. Then $E(n_v) = N\pi f_v$, and $E(m_v) = N(1-\pi)g_v$ and $E(N_v) = N\pi f_v + N(1-\pi)g_v$, and by the law of large numbers applied to the multinomial distribution the proportions converge in probability to their expectations, $n_v/N \to \pi f_v$, $m_v/N \to (1-\pi)\, g_v$ and $N_v/N \to \pi f_v + (1-\pi)\, g_v$. When a group of subjects has identical $\mathbf{Y}$s, among these identical $\mathbf{Y}$s, the algorithm returns a random pairing. (More precisely, the algorithm returns a pairing that is a deterministic function of the $\mathbf{Y}_i$s and their input order i, but the input order is random; see Section 1.1.) Hence, using equation (3) in proposition 1, as $N \to \infty$, we expect approximately

$$\begin{aligned} \frac{m_v n_v}{N_v - 1} &\doteq \frac{N^2 \pi\,(1-\pi)\, f_v g_v}{N\pi f_v + N(1-\pi) g_v} \\ &= N \frac{\pi\,(1-\pi)\, f_v g_v}{\pi f_v + (1-\pi)\, g_v} \end{aligned} \tag{6}$$

cross-matches among the expected $N\pi f_v + N(1-\pi)g_v$ subjects with $\mathbf{Y} = \mathbf{y}_v$. The expectation of A_1 under the alternative hypothesis is approximately the sum of approximation (6) over v, say $N\mu$, whereas the null expectation of A_1 is given by equation (3), or $mn/(N-1)$, which tends to $N\pi(1-\pi)$. These alternative and null expectations are related by Milne's inequality (specifically, inequality 67 of Hardy *et al.* (1952), section 2, page 61) which yields

$$\begin{aligned} \mu &= \sum_{v=1}^{V} \frac{\pi\,(1-\pi)\, f_v g_v}{\pi f_v + (1-\pi)\, g_v} \\ &\leqslant \frac{\sum \pi f_v \sum (1-\pi)\, g_v}{\sum \{\pi f_v + (1-\pi)\, g_v\}} = \pi\,(1-\pi)\,, \end{aligned}$$

with strict inequality unless $f_v = g_v$ for $v = 1, \ldots, V$. (Milne's inequality can be derived as a special case of a subtle refinement of the Cauchy inequality; see Daykin *et al.* (1969).) If $f_v \neq g_v$ for some v, then

$$A_1/N \xrightarrow{\mathrm{P}} \mu < \pi(1-\pi)$$

whereas

$$mn/N(N-1) \xrightarrow{\mathrm{P}} \pi(1-\pi),$$

so the standardized deviate on the left in expression (5) tends to $-\infty$, and the significance level based on A_1 tends to 0.

Essentially the quantity μ also plays a central role in the limiting behaviour of the Friedman–Rafsky runs statistic, and it is a member of the family of measures of distributional separation that was proposed by Györfi and Nemetz; see Henze and Penrose (1999), theorem 2.

In short, the cross-match test is a consistent test for comparing any two discrete distributions with finitely many mass points. Any two distributions may be approximated arbitrarily closely by two discrete distributions with finitely many mass points, and the cross-match test can consistently distinguish the two approximations. This is motivation, admittedly informal, for use of the cross-match test.

3.5. A small simulation

In both univariate and multivariate situations, we would not and should not use an omnibus test to detect expected changes in location or scale. The power of the exact cross-match test will be compared with that of the exact Kolmogorov–Smirnov two-sample test when comparing two distributions with the same expectation and variance, but with differing tendencies to clump at certain spots. Let $K \geqslant 2$ be an integer and σ be a real number, $0 \leqslant \sigma \leqslant 1$. The standard hot spot distribution $\mathcal{H}(K, \sigma)$ is an equal mixture of K equally spaced normal distributions, $N(\mu_k, \sigma^2)$ with

$$\mu_k = \theta\left(k - \frac{K+1}{2}\right),$$

and

$$\theta^2 = \frac{12(1-\sigma^2)}{K^2-1},$$

so the distribution has expectation 0 and variance 1 and is symmetric about zero. To see that this definition of θ yields variance 1, recall that

$$\frac{1}{K}\sum_{k=1}^{K}\left(k - \frac{K+1}{2}\right)^2 = \frac{K^2-1}{12},$$

e.g. Lehmann (1998), expression A.13, page 329. If $\sigma = 1$ then $\theta = 0$ and $\mathcal{H}(K, 1)$ is just the standard normal distribution for every K. If $\sigma = 0$ then $\mathcal{H}(K, 0)$ attaches equal probability $1/K$ to K equally spaced real numbers. Because $\mathcal{H}(K, \sigma)$ has mean 0 and variance 1 and is symmetric about zero for every K and σ, tests that are sensitive to location and scale differences are of little use in distinguishing between these distributions. Fig. 4 contrasts the cumulative distributions of $N(0, 1)$ and $\mathcal{H}(3, 1/5)$; they are quite different. In Tables 4 and 5, one sample will be drawn from the standard normal distribution, say $F(\cdot)$, and the other will be drawn from a standard hot spot distribution, say $G(\cdot)$, with parameters K and σ. When $\sigma = 1$, the simulation estimates the size of the test, which we know exactly from theory, whereas when $\sigma < 1$ the simulation estimates the power. Various other sampling situations are considered in Table 6.

With a scalar Y_i and $m+n$ even, the cross-match test takes a simple form, provided that distances between Ys are defined to be the absolute values of their differences. Then the optimal non-bipartite matching pairs adjacent order statistics from the combined sample, $Y_{(2l-1)}$ paired with $Y_{(2l)}$ for $l = 1, \ldots, (m+n)/2$. The cross-match statistic is the number of pairs which

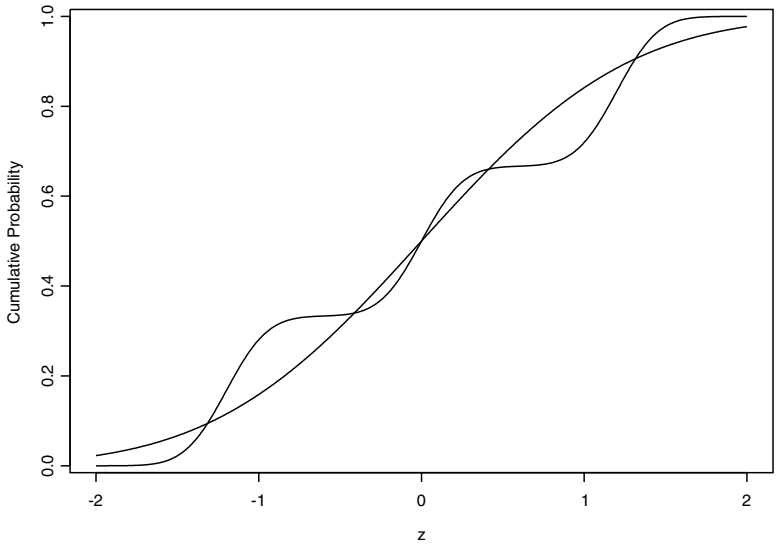

Fig. 4. Cumulative distributions for $N(0,1)$ and $\mathcal{H}(3,1/5)$

Table 4. Simulation results comparing the cross-match test with the Kolmogorov–Smirnov test, for sample size $n = m = 18$, comparing the standard normal with the standard hot spot distribution

K	σ	*Probabilities of rejection for the following tests:*	
		Kolmogorov–Smirnov	*Cross-match*
Level of test			
Exact	1.0	0.0207	0.0194
Simulated	1.0	0.0192	0.0196
Power of test			
5	0.05	0.04	0.22
3	0.1	0.06	0.36
3	0.05	0.08	0.64
2	0.05	0.29	0.91
2	0.2	0.16	0.30

contain one observation from $F(\cdot)$ and one from $G(\cdot)$, rejecting equality of $F(\cdot)$ and $G(\cdot)$ if there are few such pairs. Each situation was sampled independently 5000 times, so a proportion of successes has standard error less than or equal to $1/\sqrt{(4 \times 5000)} = 0.0071$.

A minor issue with any exact permutation test, including the Kolmogorov–Smirnov test and the cross-match test, is that discreteness typically creates a small gap between the nominal level of the test and the actual size of the test, so a 0.05-level test may reject true hypotheses slightly less than 5% of the time. To prevent this from materially favouring either statistic over the other, I selected two samples sizes, namely $m = n = 18$ and $m = n = 50$, for which the actual size of the two tests is nearly the same for nominal level 0.05.

Table 5. Simulation results comparing the cross-match test with the Kolmogorov–Smirnov test, for sample size $n = m = 50$, comparing the standard normal with the standard hot spot distribution

K	σ	*Probabilities of rejection for the following tests:*	
		Kolmogorov–Smirnov	*Cross-match*
Level of test			
Exact	1.0	0.0392	0.0372
Simulated	1.0	0.0332	0.0374
Power of test			
5	0.1	0.11	0.57
2	0.2	0.70	0.89
3	0.2	0.16	0.46
5	0.05	0.16	0.98
10	0.05	0.07	0.41

Table 6. Simulation results comparing the cross-match test with the Kolmogorov–Smirnov test, for sample size $n = m = 50$

F	G	*Probabilities of rejection for the following tests:*	
		Kolmogorov–Smirnov	*Cross-match*
$N(0, 1)$	t with 1 degree of freedom	0.11	0.19
$N(0, 1)$	t with 2 degrees of freedom	0.04	0.08
$N(0, 1)$	$N(\frac{1}{2}, 1)$	0.51	0.08
$N(0, 1)$	$N(1, 1)$	0.98	0.37
$N(0, 1)$	$N(0, 2)$	0.36	0.25
$N(0, 1)$	$\mathcal{C}_{0.25}(0, 100)$	0.09	0.19
$N(0, 1)$	$\mathcal{C}_{0.5}(2, 20)$	0.67	0.60
$N(0, 1)$	$\mathcal{C}_{0.25}(10, 1000)$	0.09	0.19
$N(0, 1)$	$\mathcal{C}_{0.5}(10, 1000)$	0.78	0.71
$E - 1$	$N(0, 1)$	0.22	0.24
$E - 1$	$-E + 1$	0.69	0.74
$\mathcal{H}(2, 0.5)$	$\mathcal{H}(5, 0.1)$	0.07	0.51
$\mathcal{H}(2, 0.1)$	$\mathcal{H}(3, 0.5)$	0.77	0.99
$\mathcal{H}(2, 0.8)$	$\mathcal{H}(2, 0.2)$	0.66	0.87

Table 4 displays results for $m = n = 18$, specifically the exact and simulated sizes of the two tests and the simulated power for several values of K and σ. For this sample size, the exact 0.05-level Kolmogorov–Smirnov two-sample test rejects for $J \geqslant 9$ and has size 0.0207; see Hollander and Wolfe (1999), section 5.4. Also, for this sample size, the 0.05-level exact cross-match test rejects for $A_1 \leqslant 4$ and has size 0.0194. The simulated sizes of the two tests are close to the exact sizes. In the cases that are considered in Table 4, the estimated power is higher for the cross-match test.

For $n = m = 50$, the exact 0.05-level Kolmogorov–Smirnov two-sample test rejects for $J \geqslant 14$ and has size 0.0392 (Drion (1952), Table 1), whereas the exact 0.05-level cross-match test rejects

for $A_1 \leqslant 18$ and has size 0.0372. In Table 5, the exact and simulated sizes of the two tests are close. Again, the power is higher for the cross-match test.

Finally, Table 6 compares the Kolmogorov–Smirnov test and the cross-match test for a variety of pairs of distributions. In Table 6, the distributions are as follows:

(a) t signifies a t-distribution,
(b) E signifies a standard exponential random variable so $E-1$ and $-E+1$ both have expectation 0 and variance 1,
(c) $\mathcal{C}_\iota(\varsigma, \eta)$ is a contaminated normal distribution formed as a mixture of the standard normal distribution with probability $1-\iota$ and the $N(\varsigma, \eta)$ distribution with probability ι and
(d) $\mathcal{H}(K, \sigma)$ is the standard hot spot distribution with K components having standard deviation σ.

Notable in Table 6 is the very poor relative performance of the cross-match test for a shift in location, and its good relative performance when the two distributions are similarly located and dispersed.

4. An extension: the cross-match rank sum

The cross-match test A_1 gives each of the I pairs equal weight. In contrast, the cross-match rank sum statistic Q, defined in the current section, will assign ranks, $1, \ldots, I$, to the I pairs by using $\tilde{\mathbf{Y}}$ or the δ_{ij} computed from $\tilde{\mathbf{Y}}$ without reference to $\mathbf{Z}$; then Q is the sum of the ranks of the A_1 cross-matched pairs. Unlike Wilcoxon's rank sum, the number A_1 of ranks that are added to form Q is a random variable. The null hypothesis $F(\cdot) = G(\cdot)$ is rejected for sufficiently small Q. In a related though different context, Schilling (1986), section 4, proposed weighted tests based on neighbours.

The null distribution of Q is not affected by the specific rule that is used to assign ranks to pairs, provided that the rule uses the information in $\tilde{\mathbf{Y}}$ or δ_{ij} without reference to $\mathbf{Z}$. Many ranking rules are possible, perhaps with a view to particular alternative hypotheses.

Let a be a fixed integer, $0 \leqslant a \leqslant I$, and let W_a be the sum of a numbers selected at random without replacement from $\{1, 2, \ldots, I\}$, so W_a has the null distribution of Wilcoxon's rank sum test comparing a subjects in one group with $I-a$ subjects in another. The probabilities $\lambda_{ak} = \Pr(W_a = k)$ for Wilcoxon's rank sum statistic may be computed by using the recursion formula that was given by Hájek *et al.* (1999), section 5.3.1. The null distribution of Q is the distribution of W_{A_1}, where A_1 has distribution (1), i.e.

$$\Pr(Q = k) = \sum_a \pi_a \lambda_{ak}.$$

For instance, in the example, with $n = m = 9$, the cross-match statistic A_1 takes values 1, 3, 5, 7 and 9 with the probabilities π_a that are given in Table 3. The 5% point of the distribution of Q is $\Pr(Q \leqslant 9) = 0.0489$. To see this, note that $\pi_1 = 0.0259$ from Table 3, and $\lambda_{1,k} = 1/9$ for $k = 1, 2, \ldots, 9$; also, $\pi_3 = 0.2764$ from Table 3, and $\lambda_{3,k} = \binom{9}{3}^{-1} = 0.0119$ for $k = 6, 7$, because $6 = 1+2+3$ and $7 = 1+2+4$, and $\lambda_{3,8} = 2\binom{9}{3}^{-1}$, because $8 = 1+2+5$ and $8 = 1+3+4$, and $\lambda_{3,9} = 3\binom{9}{3}^{-1}$, so

$$\Pr(Q \leqslant 9) = \pi_1 \times 9 \times \frac{1}{9} + \pi_3 \times 7 \times \binom{9}{3}^{-1} = 0.0259 + 0.2764 \times 0.0833 = 0.0489.$$

One might reason that a cross-match, $1 = B_i = Z_{2i-1} + Z_{2i} - 2Z_{2i-1}Z_{2i}$, in pair i is indicative of distributional overlap only if the distance in pair i, namely $\delta_{2i-1,\,2i}$, is small, so one might rank the $I = 9$ pair distances in Table 1 from largest to smallest, i.e. rank the largest distance 1, the

second largest distance 2, etc. In this case, $A_1 = 1$ and $Q = 1$, and the one-sided significance level is $\Pr(Q \leqslant 1) = \pi_1 \times \frac{1}{9} = 0.00288$, as opposed to $\Pr(A_1 \leqslant 1) = 0.0259$ in Section 3.2. In words, not only is there only $A_1 = 1$ cross-match in Fig. 2, but also that one cross-match reaches across the largest distance, $Q = 1$, so it provides even stronger evidence of limited overlap of the two distributions for patients and controls.

The null distribution of Q is the distribution of $\Sigma_{i=1}^{I} i B_i$. Using $\Sigma_{i=1}^{I} i = I(I+1)/2$ and $\Sigma_{i=1}^{I} i^2 = I(I+1)(2I+1)/6$ (e.g. Lehmann (1998), page 51), together with lemma 1, yields

$$E(Q) = \theta \frac{I(I+1)}{2}$$

and

$$\mathrm{var}(Q) = \theta(1-\theta)\frac{I(I+1)(2I+1)}{6} + (\gamma - \theta^2)\frac{I(I+1)(3I+2)(I-1)}{12}.$$

5. Discussion

The cross-match statistic is useful when it is natural to think of responses $\mathbf{Y}_i$ in terms of similarity, adjacency or distance, and when very general alternative hypotheses are of interest. In contrast, the cross-match statistic is not suited to problems in which the multivariate responses $\mathbf{Y}_i$ are partially ordered, and the alternatives of interest are also partially ordered, as would be true for detecting a positive shift in location for every co-ordinate of $\mathbf{Y}_i$. A multivariate rank test for partially ordered responses was discussed in Rosenbaum (1991).

Acknowledgements

This work was supported by grant SES-0345113 from the 'Methodology, measurement and statistics program' and the 'Statistics and probability program' of the US National Science Foundation.

References

Anderson, T. W. (1966) Some nonparametric multivariate procedures based on statistically equivalent blocks. In *Multivariate Analysis* (ed. P. R. Krishnaiah), pp. 5–27. New York: Academic Press.

Daykin, D. E., Eliezer, C. J. and Carlitz, L. (1969) Refinements of the Cauchy inequality. *Am. Math. Mnthly*, **76**, 98–100.

Derigs, U. (1988) Solving non-bipartite matching problems via shortest path techniques. *Ann. Oper. Res.*, **13**, 225–261.

Drion, E. F. (1952) Some distribution-free tests for the difference between two empirical cumulative distribution functions. *Ann. Math. Statist.*, **23**, 563–574.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

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

Galil, Z. (1986) Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, **18**, 23–38. (Available from `http://elib.zib.de/pub/Packages/mathprog/matching/weighted/index.html`.)

Greevy, R., Lu, B., Silber, J. H. and Rosenbaum, P. R. (2004) Optimal matching before randomization. *Biostatistics*, **5**, 263–275.

Hájek, J., Šidák, Z. and Sen, P. K. (1999) *Theory of Rank Tests*, 2nd edn. New York: Academic Press.

Hardy, G., Littlewood, J. E. and Polya, G. (1952) *Inequalities*. Cambridge: Cambridge University Press.

Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.*, **16**, 772–783.

Henze, N. and Penrose, M. D. (1999) On the multivariate runs test. *Ann. Statist.*, **27**, 290–298.

Hollander, M. and Wolfe, D. A. (1999) *Nonparametric Statistical Methods*, 2nd edn. New York: Wiley.
Holst, L. (1979) Two conditional limit theorems with applications. *Ann. Statist.*, **7**, 551–557.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*. New York: Wiley.
Lehéricy, S., Biondi, A., Sourour, N., Vlaicu, M., du Montcel, S. T., Cohen, L., Vivas, E., Capelle, L., Faillot, T., Casasco, A., Le Bihan, D. and Marsault, C. (2002) Arteriovenous brain malformations: is functional MR imaging reliable for studying language reorganization in patients? *Radiology*, **223,** 672–682.
Lehmann, E. L. (1998) *Nonparametrics*. Upper Saddle River: Prentice Hall.
Lu, B. and Rosenbaum, P. R. (2004) Optimal pair matching with two control groups. *J. Comput. Graph. Statist.*, **13**, 422–434.
Lu, B., Zanutto, E., Hornik, R. and Rosenbaum, P. R. (2001) Matching with doses in an observational study of a media campaign against drug abuse. *J. Am. Statist. Ass.*, **96**, 1245–1253.
Maa, J.-F., Pearl, D. K. and Bartoszynski, R. (1996) Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.*, **24**, 1069–1074.
Papadimitriou, C. H. and Steiglitz, K. (1982) *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs: Prentice Hall.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd edn. New York: Wiley.
Rosenbaum, P. R. (1991) Some poset statistics. *Ann. Statist.*, **19**, 1091–1097.
Schilling, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *J. Am. Statist. Ass.*, **81**, 799–806.
Weiss, L. (1959) Two-sample tests for multivariate distributions. *Ann. Math. Statist.*, **31**, 159–164.
Wilks, S. (1962) *Mathematical Statistics*. New York: Wiley.