Journal of Statistical Planning and Inference 142 (2012) 3097–3106

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A class of multivariate distribution-free tests of independence based on graphs

R. Heller [a,*], M. Gorfine [b], Y. Heller

[a] Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel
[b] Faculty of Industrial Engineering and Management, Technion, Haifa, Israel

ARTICLE INFO

Article history:
Received 13 December 2011
Received in revised form
3 June 2012
Accepted 6 June 2012
Available online 18 June 2012

Keywords:
Independence test
Random vectors
High-dimensional response
Multivariate association

ABSTRACT

A class of distribution-free tests is proposed for the independence of two subsets of response coordinates. The tests are based on the pairwise distances across subjects within each subset of the response. A complete graph is induced by each subset of response coordinates, with the sample points as nodes and the pairwise distances as the edge weights. The proposed test statistic depends only on the rank order of edges in these complete graphs. The response vector may be of any dimensions. In particular, the number of samples may be smaller than the dimensions of the response. The test statistic is shown to have a normal limiting distribution with known expectation and variance under the null hypothesis of independence. The exact distribution free null distribution of the test statistic is given for a sample of size 14, and its Monte-Carlo approximation is considered for larger sample sizes. We demonstrate in simulations that this new class of tests has good power properties for very general alternatives.

1. Introduction

A high dimensional response vector is measured on a group of subjects. Important applications examine whether there is a relationship between two subsets of response coordinates. In genomics, for example, it is of interest to test for associations between the measured signal on genes or on sets of genes. Moreover, it is often of interest to combine two platforms, and test for associations between one signal coming from one platform (say Chip-seq) and another signal coming from another platform (say gene expression) on the same gene or set of genes.

Classical tests for independence for bivariate populations are the Pearson and Spearman test, among others see Hollander and Wolfe (1999). For multivariate data, classical tests in Puri and Sen (1971) are not applicable if the dimension exceeds the sample size. Related tests for higher dimensions may be found in Taskinen et al. (2005). These methods base the tests on the componentwise ranking, and are ineffective for testing non-monotone types of dependence (Szekely et al., 2007).

A recent approach by Szekely and Rizzo (2009) suggests a test based on distance correlation. The latter test stands apart from other tests in two major ways. First, it is a consistent test against all alternatives. Specifically, it has power against non-monotone relationship, as opposed to the classical univariate tests and their multivariate extensions. Second, it is applicable in any dimensions. In particular, the number of samples may be smaller than the dimensions of the response vectors being tested for independence. The asymptotic null distribution of the test statistic suggested by Szekely and Rizzo

* Corresponding author.
E-mail addresses: ruheller@post.tau.ac.il (R. Heller), gorfinm@ie.technion.ac.il (M. Gorfine), heller.yair@gmail.com (Y. Heller).

0378-3758
http://dx.doi.org/10.1016/j.jspi.2012.06.003

(2009) is that of a nonnegative quadratic form of centered Gaussian random variables, with coefficients that depend on the distribution of the two subsets of response coordinates. The asymptotic null distribution for the test is not distribution free. An upper bound on the null distribution is distribution free but typically too conservative, and therefore the authors recommend using a permutation test instead.

We propose a new class of multivariate distribution-free test statistics for independence using the graph structure of the sample points on the two subsets of the multivariate response vector. We show how one can define test statistics based on the ranks of the distances on the two graphs. As in Szekely and Rizzo (2009), these tests are powerful against very general alternatives and can be applied in arbitrary dimensions. Moreover, our test statistics under the null hypothesis have a known and easily calculable asymptotic distribution and the exact distribution-free null distribution can be very well approximated by Monte-Carlo sampling. The implications are that a look-up table of the quantiles of the null distribution can be created before the study is analyzed, and repeating the study several times will not require recomputing the null distribution as long as the sample size is fixed. In contrast, the permutation test of Szekely and Rizzo (2009) will require recomputing the null distribution in every repetition of the study, since the null distribution depends on the observed data. The computational advantages of our proposed test are further addressed in the Discussion section.

Our proposed test will make use of a tree created from the two graphs of sample points. Trees, especially minimal spanning trees have been used in the literature for the purpose of comparing two groups, see Friedman and Rafsky (1979). Relatedly, nearest neighbor tests have been used for comparing two groups (Henze, 1988) or for testing goodness of fit (Bickel and Breiman, 1983). We will use the proximity among the sample points in constructed trees to weigh evidence against independence.

As an example, we consider the study of Sakaue-Sawano et al. (2008) that followed two proteins from birth to division in HeLa cells. We focused on the question of independence between the two proteins. There were 62 measurements over time for each protein, for 20 independent cells. The number of variables was three times larger than the sample size. There was no reason to believe that the relationship is monotone between the protein measurements, and therefore the concern was that a test targeted towards finding monotone relationships may not reject the null hypothesis due to the more complex nature of the relationship. We return to this example in Section 6.

In Section 2 we present the problem. Section 3 presents the nonparametric test and its null distribution. Section 4 discusses variations of the proposed test. Section 5 shows the results of a simulation study comparing the power of these tests and other tests. In particular, the simulated examples in Section 5 show the power advantage of our approach over the test in Szekely and Rizzo (2009) for small sample sizes. In Section 7 we give final remarks and further extensions.

2. The problem

We have a random vector $\mathbf{Y}$ of dimension M. Let $\mathbf{s_j}=(s_{j1},\ldots,s_{jm},\ldots,s_{jM}), j\in\{0,1\}$, be an M-dimensional vector of 0's and 1's with at least one 1, and let $\mathbf{Y}(\mathbf{s}_j)$ be the sub-vector of $\mathbf{Y}$ of dimension $\sum_{m=1}^{M} s_{jm}$ containing the coordinates for which $s_{jm}=1$. We are interested in testing whether there is a relationship between the outcomes represented by the two (disjoint) sub-vectors $\mathbf{Y}(\mathbf{s_0})$ and $\mathbf{Y}(\mathbf{s_1})$. The null hypothesis states that the two subsets of response coordinates are independent

$$H_0 : \mathcal{L}(\mathbf{Y}(\mathbf{s_0}),\mathbf{Y}(\mathbf{s_1}))=\mathcal{L}(\mathbf{Y}(\mathbf{s_0}))\mathcal{L}(\mathbf{Y}(\mathbf{s_1})),$$

where $\mathcal{L}$ refers to the "law" or "distribution". We are interested in the general alternative that the two sub-vectors are dependent

$$H_1 : \mathcal{L}(\mathbf{Y}(\mathbf{s_0}),\mathbf{Y}(\mathbf{s_1}))\neq\mathcal{L}(\mathbf{Y}(\mathbf{s_0}))\mathcal{L}(\mathbf{Y}(\mathbf{s_1})).$$

There are N independent subjects and a multivariate response $\mathbf{Y}_i$ is recorded for each subject i. The dimension of $\mathbf{Y}_i$, M, may be much higher than N. (Note that an equivalent formulation is the following: we have two random vectors $W_1\in\Re^q$ and $W_2\in\Re^p$ and N independent copies from the joint distribution of W_1 and W_2 for testing whether these random vectors are independent. In our notation $W_1=\mathbf{Y}(\mathbf{s_0})$ and $W_2=\mathbf{Y}(\mathbf{s_1})$.)

The distance covariance test in Szekely and Rizzo (2009) may be computed as follows. First, all pairwise Euclidean distances between sample values of one sub-vector and separately for the other sub-vector are computed: $a_{kl}=\left|\mathbf{Y_k}(\mathbf{s_0})-\mathbf{Y_l}(\mathbf{s_0})\right|_{\sum_{i=1}^{M} s_{0i}}$, $b_{kl}=\left|\mathbf{Y_k}(\mathbf{s_1})-\mathbf{Y_l}(\mathbf{s_1})\right|_{\sum_{i=1}^{M} s_{1i}}$, $k,l=1,\ldots,N$. Then the resulting two distance matrices are centered

$$A_{kl}=a_{kl}-\frac{1}{N}\sum_{l=1}^{N}a_{kl}-\frac{1}{N}\sum_{k=1}^{N}a_{kl}+\frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}a_{kl}$$

and

$$B_{kl}=b_{kl}-\frac{1}{N}\sum_{l=1}^{N}b_{kl}-\frac{1}{N}\sum_{k=1}^{N}b_{kl}+\frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}b_{kl}.$$

Next, the componentwise product matrix of the two centered distance matrices is averaged: $1/N^2\sum_{l=1}^{N}\sum_{k=1}^{N}A_{kl}B_{kl}$. This is the squared distance covariance between the two sub-vectors, called dCov, and it is their test statistic for testing the

independent null hypothesis. The dCov test is implemented in the R package *energy* (R Development Core Team, 2011) as a permutation test.

We suggest a new approach for testing for independence against very general alternatives. Similar to Szekely and Rizzo (2009), this approach is based on the distances between the outcomes of the N subjects on the two sub-vectors. In our approach, the distances used may be any similarity measure $d(\cdot)$ between two vectors of outcomes, so $a_{kl}=d(\mathbf{Y_k}(\mathbf{s_0}),\mathbf{Y_l}(\mathbf{s_0}))$, $b_{kl}=d(\mathbf{Y_k}(\mathbf{s_1}),\mathbf{Y_l}(\mathbf{s_1}))$, $k=1,\ldots,N$, $l=1\ldots,N$. Moreover, the resulting test statistics have a very simple form with a known null distribution.

3. The graph approach to the test of independence

An edge weighted graph is a graph with a real number assigned to each edge. A complete graph of the N sample data points on a sub-vector $\mathbf{Y}(\mathbf{s})$ is an edge weighted graph with $\binom{N}{2}$ edges linking all pairs of points. The weight associated with each edge is the distance between the nodes (points) defining it. We have a complete graph induced by $\{\mathbf{Y}_i(\mathbf{s_0}),\ i=1,\ldots,N\}$ and a graph induced by $\{\mathbf{Y}_i(\mathbf{s_1}),\ i=1,\ldots,N\}$. These graphs are fixed. Let $U_1,\ldots,U_N$ be the node labels of the graph induced by the sub-vector $\mathbf{s_0}$ and $W_1,\ldots,W_N$ be the node labels of the graph induced by the sub-vector $\mathbf{s_1}$. These node labels are a permutation of $\{1,\ldots,N\}$. When the independence null hypothesis is true, knowing the permutation $U_1,\ldots,U_N$ gives no information about the permutation $W_1,\ldots,W_N$ so all $N!$ permutations are equally likely.

3.1. The minimum spanning tree

A *path* between two prescribed nodes is an alternating sequence of nodes and edges with the prescribed nodes as first and last elements, all other nodes distinct, and each edge linking the two nodes adjacent to it in the sequence. A *connected graph* has a path between any two distinct nodes. A *tree* is a connected graph with no cycles. A *minimal spanning tree* (MST) of an edge weighted graph is a spanning tree for which the sum of edge weights is minimum.

MSTs have two important properties that make them appropriate for our application. First, they connect all the nodes with $N-1$ edges. Second, the node pairs defining the edges represent points that tend to be close together (i.e. with small distance or dissimilarity). We will use the MST based on one of the sub-vectors, say the graph induced by $\{\mathbf{Y}_i(\mathbf{s_0}),\ i=1,\ldots,N\}$, to select the edges in each step. In step 1 of the construction, we will select a node at random and select an edge in the MST that starts from that node. In step 2, we will select another edge that is in the MST and connected to one of the nodes already visited in step 1, and so forth.

3.2. Construction of the test statistic

To explain and illustrate the construction method for our proposed test, we first consider a toy example with $N=5$ sample data points. Fig. 1 (top) are the distance weighted complete graphs G_0 and G_1 induced by $\mathbf{Y}(\mathbf{s_0})$ and $\mathbf{Y}(\mathbf{s_1})$, respectively. If the two sub-vectors are independent, there is no reason to expect that sample points connected by edges with low weight in the left graph also have low weight edges in the right graph. Therefore, under the null assumption of independence, we expect that if we choose some edges based on information from G_0 only, and then look at their ranks in G_1, these ranks will be randomly distributed. Under the alternative we expect that given the MST of G_0, displayed in Fig. 1 bottom, the weight of these edges in G_1 will be small.

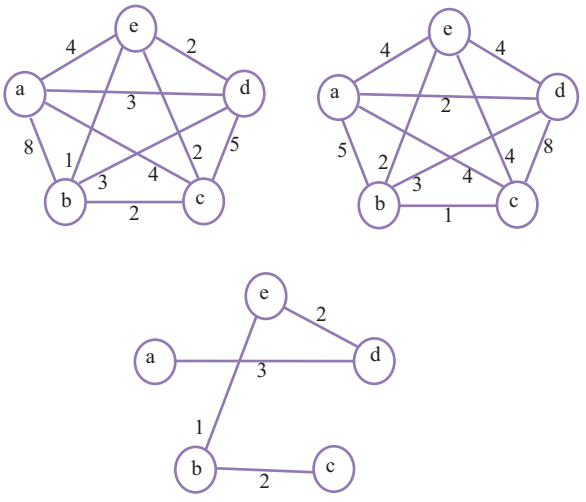

Fig. 1. A toy example: G_0 (top left), G_1 (top right), the MST based on G_0 (bottom).

We now traverse the edges matching the MST in G_1 in the following way. We start, for example, from vertex **a** and look at edge (**a**,**d**)—note that it was chosen only based on information from G_0 (Fig. 1 bottom). The rank of edge (**a**,**d**) among all the edges coming out of **a** is 1 since it is shortest of all the four edges. We now go to **d** (we could also have stayed in **a** if **a** had more edges coming out of it). The edge (**d**, **e**) (also chosen only based on information from G_0) is the second shortest among the three edges coming out of it and going to new edges (i.e. we ignore the edge (**d**,**a**)) so its rank is 2. We now continue to the vertex **e**. Yet again it is the shortest between itself and (**e**,**c**) so its rank is 1 (out of 2). In summary, in this example we got ranks lower than expected so it seems possible that there is some dependency between $\mathbf{Y}(\mathbf{s_0})$ and $\mathbf{Y}(\mathbf{s_1})$.

More generally, we consider the MST based on G_0. We select a random traversal of the tree, where at each step we start at a node already visited, and move forward to a new node. Therefore, the tree is traversed in $N-1$ steps. The traversal may be represented by $\{v_1^j, v_2^j : j=1,\ldots,N-1\}$, where v_1^j and v_2^j denote the index of the first and second node selected at step j, and $v_1^j \in \{v_1^1, v_2^1, v_2^2, \ldots, v_2^{j-1}\}$ and $v_2^j \notin \{v_1^1, v_2^1, v_2^2, \ldots, v_2^{j-1}\}$. We compute the following:

Step 1 The rank of the weight of edge $e_1 = (v_1^1, v_2^1)$ in the subgraph of $\mathbf{Y}(\mathbf{s_1})$ among the $N-1$ weights of the edges connecting v_1^1 with the $N-1$ other nodes, call this rank R_1 ($R_1 \in \{1,\ldots,N-1\}$).

Step 2 The rank of the weight of edge $e_2 = (v_1^2, v_2^2)$ in the subgraph of $\mathbf{Y}(\mathbf{s_1})$ among the edges connecting v_1^2 with $\{v_2^2, \ldots, v_2^{N-1}\}$, call this rank R_2 ($R_2 \in \{1,\ldots,N-2\}$).

$\vdots$

Step j The rank of the weight of edge $e_j = (v_1^j, v_2^j)$ in the subgraph of $\mathbf{Y}(\mathbf{s_1})$ among the edges connecting v_1^j with $\{v_2^j, \ldots, v_2^{N-1}\}$, call this rank R_j ($R_j \in \{1,\ldots,N-j\}$).

$\vdots$

Step $N-2$ The rank of the weight of edge $e_{N-2} = (v_1^{N-2}, v_2^{N-2})$ in the subgraph of $\mathbf{Y}(\mathbf{s_1})$ among the edges connecting v_1^{N-2} with $\{v_2^{N-2}, v_2^{N-1}\}$, call this rank R_{N-2} ($R_{N-2} \in \{1,2\}$).

The null distribution of the $N-2$ ranks is given in Lemma 3.1 below. Moreover, the proposition states that these ranks are independent. The independence of the $N-2$ ranks will be exploited in the construction of a powerful test statistic.

Lemma 3.1. *Under the null hypothesis of no association, R_i is uniformly distributed on $\{1,2,\ldots,N-i\}$, $i=1,\ldots,N-2$. Moreover, $R_1,\ldots,R_{N-2}$ are mutually independent.*

The proof of Lemma 3.1 is provided in Appendix A for the more general construction of a tree that includes the MST as a special case, see Lemma A.1.

The construction method results in $N-2$ ranks, $R_1,\ldots,R_{N-2}$. How can we combine the $N-2$ ranks $(R_1,\ldots,R_{N-2})$ into a test statistic? Many methods have been suggested to combine p-values, and it was shown that the combining method materially affects the power but that the optimal combining method depends on the distributions of the p-values under the alternative (Loughin, 2004). Fisher's method takes the product of the p-values as the combined evidence against the null. This combining method was investigated to have good power properties for a broad family of alternative distributions, e.g. Wallis (1942), Loughin (2004), and Benjamini and Heller (2008).

We can view the $N-2$ steps in the construction as $N-2$ tests against the null hypothesis of independence. The p-value at step j is therefore $P_j = R_j/(N-j)$. Fisher's combining method results in the test statistic $F_N = -2\sum_{j=1}^{N-2} \log P_j$. This test statistic has the desired property that when N is large, it is enough that only one of the ranks is very small for the test statistic to be large and highly significant.

3.3. The exact, asymptotic and Monte-Carlo approximate tail probabilities

The null expectation and variance of $F_{Nj} = -2\log R_j/(N-j)$ are

$$E_0(F_{Nj}) = 2\log\left[\frac{N-j}{((N-j)!)^{1/(N-j)}}\right], \quad \mathrm{Var}_0(F_{Nj}) = \frac{4}{N-j}\sum_{k=1}^{N-j}\left[\log\frac{k}{((N-j)!)^{1/(N-j)}}\right]^2.$$

As $N\to\infty$ for $j=o(N)$, F_{Nj} goes in distribution to a chi-squared random variable with 2 degrees of freedom, so the expectation and variance of F_{Nj} go to 2 and 4, respectively. From Lemma 3.1 it follows that the test statistic $F_N = \sum_{j=1}^{N-2} F_{Nj}$ is the sum of $N-2$ independent (non-identically distributed) random variables, with null expectation and variance $E_0 F_N = \sum_{j=1}^{N-2} E_0(F_{Nj}) = 2\sum_{j=2}^{N-1} \log j/(j!)^{1/j}$ and $\mathrm{Var}_0 F_N = \sum_{j=1}^{N-2} \mathrm{Var}_0(F_{Nj})$. Moreover, the asymptotic null distribution is normal.

Theorem 3.1. *When the null hypothesis of independence is true, $\mathcal{L}((F_N - E_0F_N)/\sqrt{\mathrm{Var}_0\, F_N}) \to N(0,1)$ as $N\to\infty$.*

Table 1
The exact (column 2), normal approximated (column 3) and Monte-Carlo (column 4) p-values for a sample size of $N=14$.

Test statistic, F	Exact p-value, 1-$CDF_0(F)$	Normal approximation, $1-\Phi\left(\frac{F-E_0(F)}{\sqrt{var_0(F)}}\right)$	Monte-Carlo approximation, $\frac{\sum_{b=1}^{10^6} I[F(b)\geq F]}{10^6}$
31.710259	0.000308	0.000098	0.000304
29.038958	0.002122	0.000986	0.002044
25.777678	0.014430	0.009985	0.014331
25.147138	0.019896	0.014684	0.019741
23.886213	0.036750	0.029935	0.036622
23.330950	0.046785	0.039968	0.046721
22.892610	0.056676	0.049687	0.056455
22.499919	0.067333	0.059916	0.067054

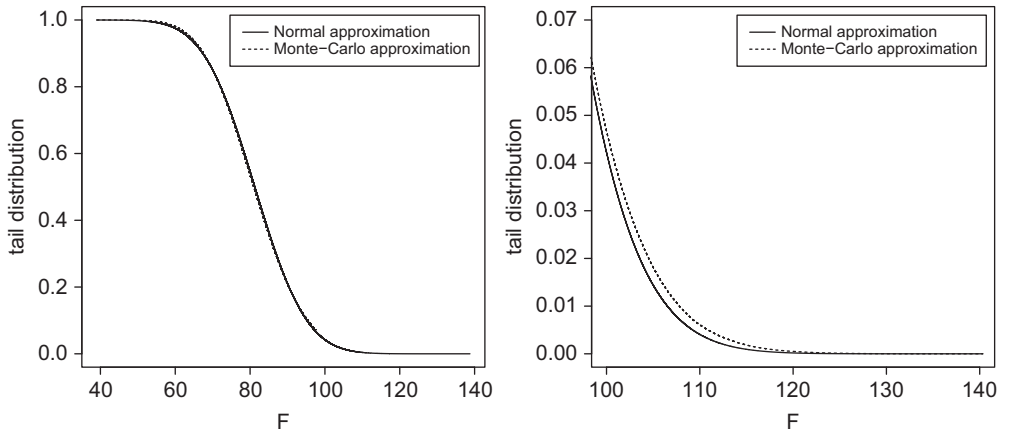

Fig. 2. For $N=50$, the tail distribution of the p-value based on the normal approximation (solid line) and Monte-Carlo approximation (dashed line), respectively. Left and right panels show the entire distribution and a zoom on the relevant range of large F values, respectively.

The proof is given in Appendix B.

How good is the tail normal approximation? Specifically, when the approximate p-value is at most 0.0001, 0.001, 0.01, 0.02, 0.03, 0.04, or 0.05, Table 1 shows the exact computation along with the asymptotic approximation for $N=14$. The normally approximated p-values (column 3) are smaller than the exact p-values (column 2). The greatest relative errors of the normal approximation are for small p-values in the extreme tail, where the approximated p-value may be more than three times smaller than the actual p-value.

It is not computationally feasible to compute the exact distribution for $N>14$ on a personal computer. However, it is possible to produce with a modern computer a supposedly endless flow of random variables from the distribution of the test statistic $F=F_N$ by sampling from the relevant discrete uniform distributions. Using standard Monte-Carlo method in R (R Development Core Team, 2011) producing 10^6 random variables from the exact distribution for $N=14$ is extremely fast, see code in Appendix C. For a test statistic F and a Monte-Carlo sample $F(1),\ldots,F(B)$, the Monte-Carlo p-value is $\sum_{b=1}^{B} I[F(b)\geq F]/B$. Table 1 shows that the Monte-Carlo p-values (column 4) based on 10^6 samples are very close to the exact p-values even in the extreme tail.

Fig. 2 shows the tail distribution of the p-value for $N=50$, based on the normal approximation (solid line) and Monte-Carlo approximation (dashed line), respectively. The agreement between these two approximations is very good, except possibly at the far tails. The 0.010 and 0.050 tail area of the test statistic based on the normal approximation are 106.5254 and 99.1266, respectively. These values correspond to a tail area of 0.013 and 0.0544, respectively using the Monte-Carlo approximation. For $N=100$, for a test statistic $F=204.63$, the p-value based on the normal approximation is 0.0500 and the p-value based on the Monte-Carlo approximation is 0.0540.

Since the Monte-Carlo approximation can be made arbitrary close to the exact distribution for finite N, and since computing the Monte-Carlo approximation is very fast, in all our computations henceforth we use the Monte-Carlo approximation to the exact null distribution of the test statistic. Note that contrary to permutation methods, where the null distribution is obtained by conditioning on the observed sample, the Monte-Carlo approximation is based on random draws from the null distribution of the ranks and thus is the same for all observed samples of fixed size N.

Other large sample approximations that give closer or more conservative estimates than the normal approximation may be derived. For example, since the null distribution of our test statistic is skewed, more accurate results may be obtained by using a chi-squared approximation, as noted by Hall (1983). However, since for most practical applications

(in particular, the ones considered in this paper) the Monte-Carlo approximation can be easily computed, we chose to use it and omit any further developments on the large sample approximation of our test statistic.

4. Another construction of the tree

The MST is just one of the possible ways to select short edges. More generally, we could select some of the edges based on G_0 and look at their rank in G_1 and some of the edges based on G_1 and look at their rank in G_0. The most general form is given in Appendix A.

We consider below a construction that is based on optimal pairing. Suppose that there are $N=2I$ nodes and a distance d_{ij} between node i and node j. A minimum distance nonbipartite matching pairs the nodes into I non-overlapping pairs to minimize the total distance within pairs. For notational convenience, the nodes are renumbered after pairing so that in the new order subject $2i-1$ and subject $2i$ are paired for $i=1,\ldots,I$. The nonbipartite matching has minimal distance if $\sum_{i=1}^{I} d_{(2i-1,2i)}$ is smallest over all possible pairings. The minimum distance nonbipartite matching problem is a standard combinatorial problem that can be solved in $O(I^3)$ operations, implemented in the R package nbpMatching, see Lu et al. (2011). The distance matrix $\{d_{ij} : i=1,\ldots,2I, j=1,\ldots,2I\}$ has to be symmetric and positive, but need not satisfy the triangle inequality. For an odd number of subjects, a pseudo subject is added with distance 0 from all other subjects and the one actual subject who is paired with the pseudo subject is discarded.

The basic idea behind the construction of the tree based on the optimal pairings is that we alternate between taking edges that are the optimal pairs based on $\mathbf{Y}(\mathbf{s_0})$ and based on $\mathbf{Y}(\mathbf{s_1})$. We do however have to be careful not to create a cycle (i.e. an alternating sequence of nodes and edges where a node is repeated) therefore we do the following. In step 1 of the construction, we will select the shortest edge among the I edges that form the optimal pairing based on $\mathbf{Y}(\mathbf{s_0})$. In step 2, if the edge selected in step 1 is also optimally paired based on $\mathbf{Y}(\mathbf{s_1})$ (i.e. a cycle is formed if we choose this edge again), then in step 2 we select a vertex from step 1 and choose the shortest edge based on the graph of $\mathbf{Y}(\mathbf{s_1})$ among the $N-2$ edges starting from it that have not yet been selected, i.e. that go to a new vertex. Otherwise, we select an edge among the I edges that form the optimal pairing based on $\mathbf{Y}(\mathbf{s_1})$ that starts in a vertex used in step 1. In step j, for j odd (even), if we can choose an edge from the optimal pairing based on $\mathbf{Y}(\mathbf{s_0})$ ($\mathbf{Y}(\mathbf{s_1})$) that starts in an edge we already visited and goes to an edge that we have not yet visited, we do so. Otherwise in order to avoid a cycle, we select a vertex we already visited and pick the shortest edge from it to a new vertex based on the graph of $\mathbf{Y}(\mathbf{s_0})$ ($\mathbf{Y}(\mathbf{s_1})$).

As an illustration, consider the toy example in Fig. 1. An optimal pairing based on $\mathbf{Y}(\mathbf{s_0})$ is edges (**e**,**b**) and (**a**,**d**). An optimal pairing based on $\mathbf{Y}(\mathbf{s_1})$ is (**b**,**c**) and (**a**,**d**). We construct the tree as follows: we start for example from vertex **e** and select edge (**e**,**b**) since it is the shortest optimally paired edge based on G_0. The rank of edge (**e**,**b**) among all four edges coming out of **e** in G_1 is one since it is the shortest of all the four edges. We now go to **b**. The edge (**b**,**c**) is selected based on the optimal pairing of G_1. In G_0 this edge is the shortest among the three edges coming out of **b**, excluding edge (**e**,**b**). Therefore its rank is 1 out of 3. Finally, we need to select based on G_0 an edge that starts from **e**,**b**,**c** excluding the three edges between these nodes. Since there is no optimal pair not yet visited to be selected, we can instead select the shortest edge out of all possible ones in G_0. The shortest such edge is edge (**e**,**d**). It is the same distance as (**e**,**a**) in G_1 so its rank is 1.5.

5. Simulations

In all simulations, the dCov test was applied by calling the function *dcov.test* implemented in the R package *energy* (Szekely and Rizzo, 2009) with 10 000 permutation samples.

5.1. Bivariate distributions

We consider first the six simulated examples of unusual bivariate distributions in Newton (2009). These examples mimic those at the wikipedia.org page on Pearson correlation. For $N=100$ sample points the relations are already manifest by eye, as can be seen from the example data in Fig. 3. The example of four independent clouds is an example of a null distribution. Table 2 shows the power comparison between the following tests: Pearson's correlation test, Spearman's correlation test, dCov, the proposed test that uses optimal pairing in the construction, and the proposed test that uses MST in the construction. All tests maintain the correct size for the example of four independent clouds. In the other examples, while the power of the tests based on Pearson and Spearman correlations remain small, the power of the tests based on dCov, optimal pairing and MST increases towards 1 as the sample size increases. Large differences can be observed for smaller sample sizes. The most pronounced difference is observed for the circle relation, where at $N=100$ the power of the tests based on optimal pairing and MST are 0.81 and 0.54, respectively, whereas dCov had no power to detect this relation.

5.2. Multivariate distributions

Szekely et al. (2007) considered the following example of a non-linear relation, where none of the likelihood ratio type of tests they considered performs well. Using our notation, the distribution of $\mathbf{Y}(\mathbf{s_0})$ is standard multivariate normal with five dimensions, and $\mathbf{Y}(\mathbf{s_1})$ is $\log(\mathbf{Y^2}(\mathbf{s_0}))$. Table 3 shows the power of a test at level 0.05 for each of the following tests: dCov, the test using optimal pairing in the construction, and the test using MST in the construction. The proposed test based on

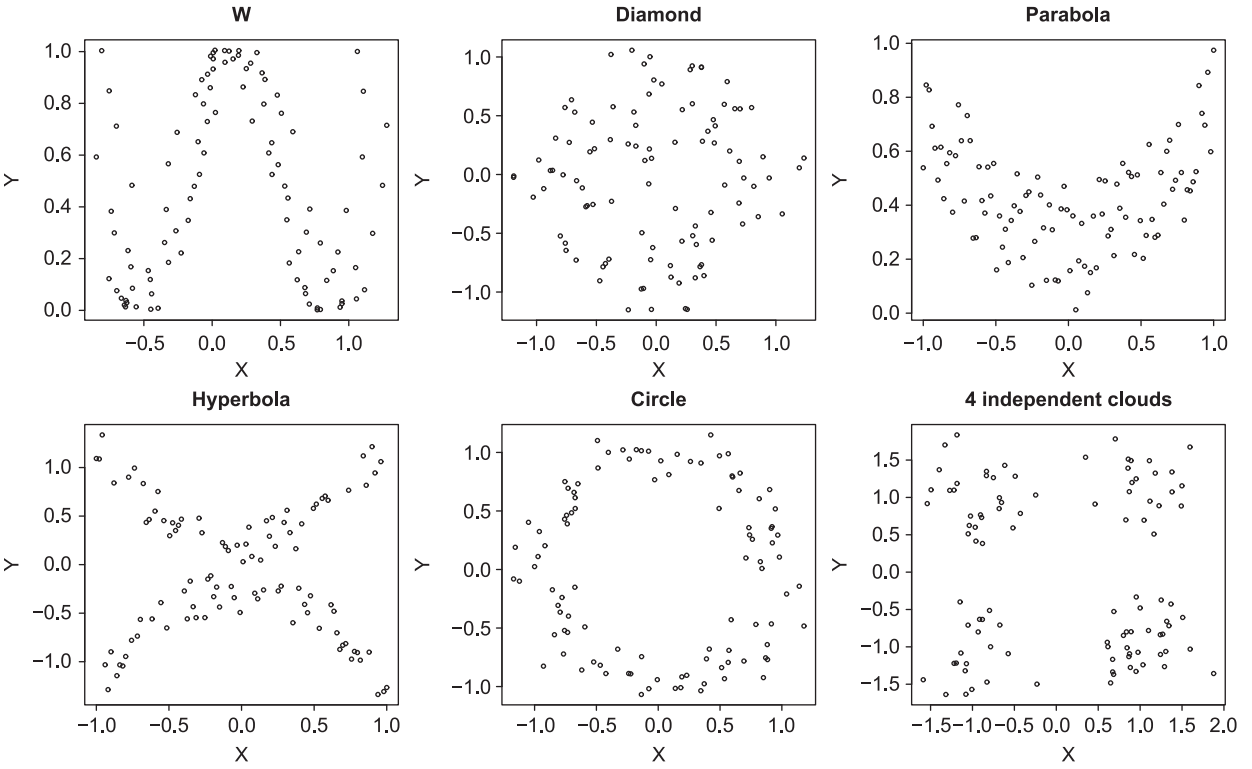

Fig. 3. Six simulated examples of unusual bivariate distributions; a sample of size $N=100$ from each distribution.

Table 2
The power (SE), based on 10 000 repetitions, for a test at level 0.05 per sample size from the joint distributions that generate the unusual bivariate relation in wikipedia.org page on Pearson correlation for various sample sizes. The tests compared are Pearson's correlation test (column 3), Spearman's correlation test (column 4), the distance correlation test dCov (column 5), the proposed test that uses optimal pairing in the construction (column 6), and the proposed test that uses MST in the construction (column 7).

N	Distribution	Pearson	Spearman	dCov	Optimal pairing	MST
50	W	0 (0)	0 (0)	0.8480 (0.0036)	0.1165 (0.0032)	**0.9969 (0.0006)**
	Diamond	0.0029 (0.0005)	0.0139 (0.0012)	0.0446 (0.0020)	**0.1852 (0.0039)**	0.0782 (0.0027)
	Parabola	0.0045 (0.0007)	0.0019 (0.0004)	**0.9718 (0.0017)**	0.5694 (0.0050)	0.7645 (0.0042)
	Hyperbola	0.1537 (0.0036)	0.1397 (0.0035)	0.2975 (0.0046)	0.6310 (0.0048)	**0.9446 (0.0023)**
	Circle	0 (0)	0 (0)	0 (0)	**0.4060 (0.0049)**	0.2281 (0.0042)
	Four clouds	0.0492 (0.0022)	0.0491 (0.0022)	0.0487 (0.0022)	0.0495 (0.0022)	0.0518 (0.0022)
100	W	0 (0)	0 (0)	**1 (0)**	0.4599 (0.0050)	**1 (0)**
	Diamond	0.0019 (0.0004)	0.0134 (0.0011)	0.1395 (0.0035)	**0.3112 (0.0046)**	0.1012 (0.0030)
	Parabola	0.0044 (0.0005)	0.0023 (0.0005)	**1 (0)**	0.8447 (0.0036)	0.9464 (0.0023)
	Hyperbola	0.1380 (0.0034)	0.1294 (0.0034)	0.9535 (0.0021)	0.9100 (0.0029)	**0.9973 (0.0005)**
	Circle	0 (0)	0 (0)	0.0001 (0.0001)	**0.8120 (0.0039)**	0.5367 (0.0050)
	Four clouds	0.0476 (0.0021)	0.0486 (0.0022)	0.0473 (0.0021)	0.0516 (0.0022)	0.0473 (0.0021)

MST has excellent power even for samples of size 30, and the proposed test based on optimal pairing has very good power when the sample size is at least 60. Both tests are clearly superior to the dCov test in this example.

6. Example

Sakaue-Sawano et al. (2008) followed the fluorescence level of two fluorescent proteins, one that labels G_1 phase nuclei in red and the other that labels $S/G_2/M$ phase nuclei in green, from birth to division in HeLa cells. We analyzed a subset of the data, kindly provided to us by Sivan Pearl from the research group of Professor Nathalie Questembert-Balaban at the Hebrew university. Our subset consisted of the time series of the two proteins in a sample of 20 independent cells. Examination within cell of the fluorescence of these two proteins will give a strong association because the expression of both proteins depends on cell cycle progression. However, by the examination of the time curves across cells, we can ask whether changes in the expression curve (over cell progression) in one protein is predictive of changes in the expression

Table 3
The power (SE) of a test at level 0.05 per sample size from the joint distribution that generates Example 3 in Szekely et al. (2007): $\mathbf{Y}(\mathbf{s_0})$ is standard multivariate normal with five dimensions and $\mathbf{Y}(\mathbf{s_1}) = \log(\mathbf{Y^2}(\mathbf{s_0}))$.

N	dCov	Optimal pairing	MST
20	0.159 (0.012)	0.279 (0.014)	**0.488** (0.016)
30	0.296 (0.014)	0.493 (0.016)	**0.854** (0.011)
40	0.443 (0.016)	0.720 (0.014)	**0.980** (0.004)
50	0.636 (0.015)	0.871 (0.011)	**0.999** (0.001)
60	0.750 (0.014)	0.939 (0.008)	**1.000** (0.000)
70	0.910 (0.009)	**0.994** (0.002)	**1.000** (0.000)
80	0.955 (0.007)	**0.997** (0.002)	**1.000** (0.000)

curve of the other protein. In other words, by the examination of the time curves across cells, we can test for independence between the expression curves (from birth to division of the cell) of the two proteins.

For simplicity, the data included only the first 62 time points, since this was the length of the life cycle of the shortest of the 20 cells. The proposed test that uses optimal pairing and MST in the construction resulted in p-values of 0.12 and 0.01, respectively. The dCov test of independence resulted in a p-value of 0.08. This analysis suggests that there is some evidence of dependence, as reflected in the fairly small p-value of the test that uses MST in the construction.

We further considered the test of independence after standardizing each time series to have mean zero and a standard deviation of one. The standardization had a remarkable effect. After standardization of the data, the three tests of independence suggest strongly that the two random vectors are dependent, so that relative changes in the expression curve of one protein are reflected in relative changes in the expression curve of the other protein. The proposed test that uses optimal pairing and MST in the construction resulted in p-values of 0.0002 and 0.0015, respectively. The dCov test of independence resulted in a p-value of 0.00001. Since the choice of role of the two random vectors in the construction was arbitrary, we reversed their roles and received corresponding p-values of 0.0002 and 0.0016. Clearly, the choice of role in the construction was not relevant in this particular example.

7. Discussion

We proposed two distribution-free tests of independence based on graphs. In our tests, the distance or similarity measure can be very general and the dimensions of the response vector may be larger than the sample size. Moreover, the null distribution of our test statistics is known and easily approximated for large sample size. We showed that our tests have good power properties even when the sample size is small for non-linear relationships between the two subsets of response coordinates tested for independence. We recommend using these tests when the subsets of the response vector of interest are suspected to have complex, non-monotone relationships. We observe that the choice of method of construction of the tree matters, but the better choice performance-wise depends on the alternative. The test using optimal pairing in the construction treats both sub-vectors similarly, so the test results may differ only slightly if the roles of the two sub-vectors are reversed.

Having an easily calculable distribution-free test of the null hypothesis is very important computationally in multiple testing settings. Consider the following example from Genomics research. In microarray studies, the signal in many genes is simultaneously measured, and it may be of interest to test the independence between various groups of genes, called gene sets. Suppose we have M (in the order of hundreds or thousands) gene sets and $\binom{M}{2}$ hypotheses of independence between the gene sets. In order to adjust for multiplicity, the significance at the far tail of the null distribution needs to be calculated. A permutation test, such as the one described in Szekely and Rizzo (2009), will require $O(M^2)$ permutations for each test to be in the $O(1/M^2)$ tail of the null distribution. Note that it is necessary to be this far in the tail for most popular multiplicity correction methods, not only the conservative Bonferroni correction but also for the corrections advocated by Benjamini and Hochberg (1995), when the fraction of true findings is small. Thus $O(M^4)$ permutations are necessary for all tests. For a sample of size N, the cost of computing the test statistic is $O(N^2)$ (this is the cost of computing the $N \times N$ distance matrix, or of multiplying the entries of two $N \times N$ distance matrices), and therefore the total computational complexity is $O(M^4 \times N^2)$. To apply our test, a look-up table can be created for the null distribution of a sample of size N with a computational cost of order $O(M^2 \times N)$ in advance for N small, or the asymptotic approximation may be used for N large. The computational complexity is therefore only $O(M^2 \times N^2)$, where $O(N^2)$ is the cost of computing the distance matrix and the minimal spanning tree (Seth and Vijaya, 2002), and this is a substantial reduction in computational cost when testing several hundred or several thousands gene sets for pairwise independence.

Acknowledgments

We thank the authors of Sakaue-Sawano et al. (2008) for providing the raw data of the fluorescence images, and Sivan Pearl for the processed data that was used for the example. We also thank Sivan Pearl and Professor Nathalie Questembert-Balaban for helpful discussions of the example.

Appendix A. The general construction of the test statistic

In the construction method below, let v_1^j and v_2^j denote the index of the first and second node selected at step j, from the possible values $\{1,\ldots,N\}$.

Step 1 Begin at any node v_1^1. From v_1^1 using only information from $\mathbf{Y}(\mathbf{s}_{i_1})$ $(i_1 \in \{0,1\})$ choose a new node v_2^1. Calculate the rank of the weight of edge $e_1=(v_1^1,v_2^1)$ in the subgraph of $\mathbf{Y}(\mathbf{s_{1-i_1}})$ that connects v_1^1 with the $N-1$ other nodes, call this rank R_1 $(R_1 \in \{1,\ldots,N-1\})$.

Step 2 Go to node $v_1^2 \in \{v_1^1,v_2^1\}$ and use only information from $\mathbf{Y}(\mathbf{s_{i_2}})$ $(i_2 \in \{0,1\})$ to choose a new node v_2^2 that we have not visited yet, i.e. $v_2^2 \notin \{v_1^1,v_2^1\}$. Calculate the rank of the weight of edge $e_2=(v_1^2,v_2^2)$ in the subgraph of $\mathbf{Y}(\mathbf{s_{1-i_2}})$ that connects v_1^2 with the $N-2$ other nodes we have not visited yet, call this rank R_2 $(R_2 \in \{1,\ldots,N-2\})$.

$\vdots$

Step j Go to node $v_1^j \in \{v_1^1,v_2^1,v_2^2,\ldots,v_2^{j-1}\}$ and use only information from $\mathbf{Y}(\mathbf{s}_{i_j})$ $(i_j \in \{0,1\})$ to choose a new node v_2^j that we have not visited yet, i.e. $v_2^j \notin \{v_1^1,v_2^1,v_2^2,\ldots,v_2^{j-1}\}$. Calculate the rank of the weight of edge $e_j=(v_1^j,v_2^j)$ in the subgraph of $\mathbf{Y}(\mathbf{s_{1-i_j}})$ that connects v_1^j with the $N-j$ other nodes we have not visited yet, call this rank R_j $(R_j \in \{1,\ldots,N-j\})$.

$\vdots$

Step $N-2$ Go to node $v_1^{N-2} \in \{v_1^1,v_1^2,v_2^2,\ldots,v_2^{N-3}\}$ and use only information from $\mathbf{Y}(\mathbf{s_{i_{N-2}}})$ $(i_{N-2} \in \{0,1\})$ to choose a new node v_2^{N-2} that we have not visited yet, i.e. $v_2^{N-2} \notin \{v_1^1,v_1^2,v_2^2,\ldots,v_2^{N-3}\}$. Calculate the rank of the weight of edge $e_{N-2}=(v_1^{N-2},v_2^{N-2})$ in the subgraph of $\mathbf{Y}(\mathbf{s_{1-i_{N-2}}})$ that connects v_1^{N-2} with the two other nodes we have not visited yet, call this rank R_{N-2} $(R_{N-2} \in \{1,2\})$.

The null distribution of the $N-2$ ranks is given in Lemma 3.1 below. Moreover, the proposition states that these ranks are independent. The independence of the $N-2$ ranks will be exploited in the construction of a powerful test statistic.

Lemma A.1. *Under the null hypothesis of no association, R_i is uniformly distributed on $\{1,2,\ldots,N-i\}$, $i=1,\ldots,N-2$. Moreover, $R_1,\ldots,R_{N-2}$ are mutually independent.*

Proof. The proof of the lemma is by induction. For $k=1$, since all permutations of the nodes 1,2,3,…,N are equally likely on the graph $Y(s_{1-i_1})$, then once we fix v_1^1 and v_2^1 using information on the subgraph $Y(s_{i_1})$, then the weight of $e_1=(v_1^1,v_2^1)$ in the subgraph of $Y(s_{1-i_1})$ may be any of the possible weights with equal probability under the null hypothesis of independence. In particular, fixing v_1^1 in the subgraph $Y(s_{1-i_1})$, then the weight of $e_1=(v_1^1,v_2^1)$ in the subgraph of $Y(s_{1-i_1})$ may be any of the $N-1$ possible weights with equal probability. So R_1 is uniformly distributed on $\{1,\ldots,N-1\}$ (assuming no ties).

For $k=2$, note that v_1^2 is already fixed in both subgraphs. We choose $v_2^2 \notin \{v_1^1,v_2^1\}$ using information on the subgraph $Y(s_{i_1})$. Therefore, under the null hypothesis of independence the weight of $e_2=(v_1^2,v_2^2)$ in the subgraph of $Y(s_{1-i_1})$ may be any of the $N-2$ possible weights (excluding the weight of e_1) with equal probability, regardless of the value of R_1. Therefore, R_2 is uniformly distributed on $\{1,2,\ldots,N-2\}$ and independent of R_1.

Assuming that the lemma is true of $i<j$, then for $k=j$, note that v_1^j is already fixed in both subgraphs. We choose $v_2^j \notin \{v_1^1,v_2^1,v_2^2,\ldots,v_2^{j-1}\}$ using information from $Y(s_{i_j})$. The weight of $e_j=(v_1^j,v_2^j)$ in the subgraph of $Y(s_{1-i_j})$ can be any of the $N-1-(j-1)$ possible weights (excluding the weights on the edges connecting v_1^j with the nodes already visited, i.e. with $\{v_1^1,v_2^1,v_2^2,\ldots,v_2^{j-1}\}/v_1^j$) with equal probability, regardless of the values of $R_1,\ldots,R_{j-1}$. Therefore, R_j is uniformly distributed on $\{1,\ldots,N-j\}$ and independent of $R_1,\ldots,R_{j-1}$. □

Appendix B. Proof of Theorem 3.2

Proof. From Lemma A.1 it follows that F_{Nj}, $j=1,\ldots,N-2$ are independent non-identically distributed random variables. It is straightforward to show that for a fixed j, $E_0(F_{Nj}) \le 2$ and $\mathrm{Var}_0(F_{Nj}) \le 4$. It therefore follows that $\mathrm{Var}_0 F \le 4N$, but we now show that in addition, $\mathrm{Var}_0 F = O(N)$:

$$\mathrm{Var}_0 F_N = 4\sum_{n=2}^{N-1}\left[\left(\frac{1}{n}\sum_{x=1}^{n}(\log x)^2\right)-\left(\frac{1}{n^2}(\log n!)^2\right)\right] \ge 4\sum_{n=\sqrt{N}}^{N-1}\left[\left(\frac{1}{n}\sum_{x=1}^{n}(\log x)^2\right)-\left(\frac{1}{n^2}(\log n!)^2\right)\right]$$

$$= 4\sum_{n=\sqrt{N}}^{N-1}\left[\left(2-\frac{2}{n}-2\log n+(\log n)^2\right)-(\log n-1)^2+o(1)\right] = 4\sum_{n=\sqrt{N}}^{N-1}[1+o(1)] = O(N).$$

The Lindeberg–Feller central limit theorem states that if the following condition is satisfied:

$$\lim_{N\to\infty}\frac{1}{\mathrm{Var}_0 F_N}\sum_{j=1}^{N-2}E_0[(F_{Nj}-E_0F_{Nj})^2 I((F_{Nj}-E_0F_{Nj})^2 > \epsilon^2\,\mathrm{Var}_0 F_N)] = 0 \quad \forall\epsilon > 0 \tag{1}$$

then $(F_N-E_0F_N)/\sqrt{\mathrm{Var}_0 F_N}$ converges in distribution to a standard normal random variable as $N\to\infty$. Therefore it remains to prove (1) above. Since

$$(F_{Nj}-E_0(F_{Nj}))^2 = (2\log(N-j)-2\log R_j-E_0(F_{Nj}))^2 \le (2\log N)^2,$$

it follows that

$$\frac{1}{\mathrm{Var}_0 F_N}\sum_{j=1}^{N-2}E_0[(F_{Nj}-E_0F_{Nj})^2 I((F_{Nj}-E_0F_{Nj})^2 > \epsilon^2\,\mathrm{Var}_0 F_N)] \le \frac{1}{\mathrm{Var}_0 F_N}\sum_{j=1}^{N-2}E_0[(2\log N)^2 I((F_{Nj}-E_0F_{Nj})^2 > \epsilon^2\,\mathrm{Var}_0 F_N)]$$
$$\le \frac{(2\log N)^2}{\mathrm{Var}_0 F_N}\sum_{j=1}^{N-2}\frac{\mathrm{Var}_0(F_{Nj})}{\epsilon^2\,\mathrm{Var}_0 F_N} = \frac{(2\log N)^2}{\mathrm{Var}_0 F_N}\frac{1}{\epsilon^2},$$

where the last inequality is a direct application of Markov's inequality. Since $\mathrm{Var}_0 F_N = O(N)$, it follows that $\lim_{N\to\infty}((2\log N)^2/\mathrm{Var}_0 F_N)(1/\epsilon^2) = 0$ and thus condition (1) is satisfied. □

Appendix C. Monte-Carlo approximation of the null distribution

The R code for computing the Monte-Carlo p-value in Table 1.

```
c=3; n=14; m=n-c; B=1 000 000;

bmat=matrix(NA,nrow=B,ncol=n-c)

for (i in1:(n-c)){
  Ri=rwilcox(B,1, (n-i-1))+1
  bmat[,i]=-2*log(Ri/(n-i))
}

T=apply(bmat,1,sum)

c(sum(T>=31.710259)/B,sum(T>=29.038958)/B,sum(T>=25.777678)/B,...
sum(T>=25.147138)/B,sum(T>=23.886213)/B,sum(T>=23.330950)/B,...
sum(T>=22.892610)/B,sum(T>=22.499919)/B)
```

References

Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64, 1215–1222.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Methodological 57 (1), 289–300.

Bickel, P., Breiman, L., 1983. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. The Annals of Probability 11 (1), 185–214.

Friedman, J., Rafsky, L., 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics 7 (4), 697–717.

Hall, P., 1983. Chi squared approximations to the distribution of a sum of independent random variables. The Annals of Probability 11 (4), 1028–1036.

Henze, N., 1988. A multivariate two-sample test based on the number of nearest neighbor coincidences. Annals of Statistics 16, 772–783.

Hollander, M., Wolfe, D., 1999. Nonparametric Statistical Methods, 2nd ed. John Wiley & Sons Inc., New York.

Loughin, T., 2004. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics and Data Analysis 47, 467–485.

Lu, B., Robert, G., Xinyi, X., Beck, C., 2011. Optimal nonbipartite matching and its statistical applications. The American Statistician 65 (1), 21–30.

Newton, M., 2009. Introducing the discussion paper by Szekely and Rizzo. The Annals of Applied Statistics 3 (4), 1233–1235.

Puri, M., Sen, P., 1971. Nonparametric Methods in Multivariate Analysis. John Wiley & Sons Inc., New York.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Sakaue-Sawano, A., Kurokawa, H., Morimura, T., Hanyu, A., Hama, H., Osawa, H., Kashiwagi, S., Fukami, K., Miyata, T., Miyoshi, H., Imamura, T., Ogawa, M., Msai, H., Miyawaki, A., 2008. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell 132 (3), 487–498.

Seth, P., Vijaya, R., 2002. An optimal minimum spanning tree algorithm. Journal of the Association for Computing Machinery 49 (1), 16–34.

Szekely, G., Rizzo, M., 2009. Brownian distance covariance. The Annals of Applied Statistics 3 (4), 1236–1265.

Szekely, G., Rizzo, M., Bakirov, N., 2007. Measuring and testing independence by correlation of distances. The Annals of Statistics 35, 2769–2794.

Taskinen, S., Oja, H., Randles, R., 2005. Multivariate nonparametric tests of independence. American Statistical Association 100 (471), 916–925.

Wallis, W., 1942. Compounding probabilities from independent significance tests. Econometrica 10 (3/4), 229–248.