# ON THE MULTIVARIATE RUNS TEST 

By Norbert Henze and Mathew D. Penrose<br>Universität Karlsruhe and University of Durham


#### Abstract

For independent $d$-variate random variables $X_{1}, \ldots, X_{m}$ with common density $f$ and $Y_{1}, \ldots, Y_{n}$ with common density $g$, let $R_{m, n}$ be the number of edges in the minimal spanning tree with vertices $X_{1}, \ldots, X_{m}$, $Y_{1}, \ldots, Y_{n}$ that connect points from different samples. Friedman and Rafsky conjectured that a test of $H_{0}: f=g$ that rejects $H_{0}$ for small values of $R_{m, n}$ should have power against general alternatives. We prove that $R_{m, n}$ is asymptotically distribution-free under $H_{0}$, and that the multivariate two-sample test based on $R_{m, n}$ is universally consistent.


1. Introduction and results. Suppose $X_{1}, X_{2}, X_{3}, \ldots$ are independent $d$-dimensional variables with common probability density function $f$, and independently, $Y_{1}, Y_{2}, \ldots$ are independent $d$-dimensional variables with common density function $g$. An important and challenging problem in multivariate statistics is the two-sample problem: given observations of $\mathscr{X}_{m}:=$ $\left\{X_{1}, \ldots, X_{m}\right\}$ and $\mathscr{Y}_{n}:=\left\{Y_{1}, \ldots, Y_{n}\right\}$, find a good test for the null hypothesis $H_{0}: f=g$, against a general alternative. A number of well-understood tests are known in the case $d=1$; these are based on the ranks of observations within the sorted list of the pooled sample and hence are distribution-free under $H_{0}$. For samples in $\mathbb{R}^{d}, d \geq 2$, the problem has been studied far less fully (see [3], [4], [6], [7], [13], [21]).

The subject of this paper is the multivariate runs test proposed by Friedman and Rafsky [8], which is defined as follows. Given a finite set $S \subset \mathbb{R}^{d}$, a spanning tree on $S$ is a connected graph $\mathscr{T}$ with vertex-set $S$ and no cycles; its length $l(\mathscr{T})$ is the total of its Euclidean edge lengths. A minimal spanning tree (MST) is a spanning tree with $l(\mathscr{T}) \leq l\left(\mathscr{T}^{\prime}\right)$ for all spanning trees $\mathscr{T}^{\prime}$. Denote $S \subset \mathbb{R}^{d}$ nice if it is locally finite and all interpoint distances among elements of $S$ are distinct. If $S$ is nice and finite, it has a unique MST (see, e.g., [2] or [16]). If $S$ is nice and infinite, an analogous notion of minimal spanning forest (MSF) was developed by Aldous and Steele in [2] and denoted $g(S)$ there. In this paper, for nice $S \subset \mathbb{R}^{d}$ we denote the MST (if $S$ is finite) or MSF (if infinite) by $\mathscr{T}(S)$.

Given finite sets $S$ and $T$ in $\mathbb{R}^{d}$ such that $S \cup T$ is nice, let $R(S, T)$ denote the number of edges of $\mathscr{T}(S \cup T)$ which connect a point of $S$ to a point of $T$. Friedman and Rafsky's test statistic $R_{m, n}$ is given by

$$
R_{m, n}=R\left(\mathscr{X}_{m}, \mathscr{Y}_{n}\right) .
$$

[^0]In fact, Friedman and Rafsky consider $1+R_{m, n}$, which is the number of disjoint subtrees that result from removing all edges of $\mathscr{T}\left(\mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)$ that join vertices of different samples. They conjecture that rejection of $H_{0}$ for small values of $R_{m, n}$ "can be expected to have power against general alternatives" ([8], page 708). We verify this by proving the consistency of the multivariate runs test against general alternatives. Furthermore, we show that the test statistic is asymptotically distribution-free under $H_{0}$.

For asymptotics, we take $m \rightarrow \infty$ and $n \rightarrow \infty$ in a linked manner so that $m /(m+n) \rightarrow p \in(0,1)$, which we shall call the usual limiting regime. Set $q=1-p$ and $r=2 p q$, and write $\rightarrow_{g}$ for convergence in distribution. Let $\mathscr{N}\left(\mu, \sigma^{2}\right)$ denote the normal distribution with expectation $\mu$ and variance $\sigma^{2}$. For $\lambda>0$, let $\mathscr{P}_{\lambda}$ denote a homogeneous Poisson process on $\mathbb{R}^{d}$ of rate $\lambda$, with a point added at the origin.

THEOREM 1. In the usual limiting regime, under $H_{0}$,

$$
(m+n)^{-1 / 2}\left(R_{m, n}-\frac{2 m n}{m+n}\right) \rightarrow_{\mathscr{O}} \mathscr{N}\left(0, \sigma_{d}^{2}\right)
$$

where

$$
\sigma_{d}^{2}=r\left(r+\frac{1}{2} \operatorname{Var}\left(D_{d}\right)(1-2 r)\right)
$$

Here $D_{d}$ is the degree of the vertex at 0 in the $\operatorname{MSF} \mathscr{T}\left(\mathscr{P}_{1}\right)$.
Theorem 2. In the usual limiting regime,

$$
\begin{equation*}
\frac{R_{m, n}}{m+n} \rightarrow 2 p q \int \frac{f(x) g(x)}{p f(x)+q g(x)} d x \quad \text { almost surely. } \tag{1}
\end{equation*}
$$

REMARK 1. The right-hand side of (1) equals $1-\delta(f, g, p)$, where

$$
\delta(f, g, p)=\int \frac{p^{2} f^{2}(x)+q^{2} g^{2}(x)}{p f(x)+q g(x)} d x
$$

is a member of a general class of separation measures of several probability distributions (see [9], [10] and [11]). From Theorem 1, Theorem 2 and the fact that the inequality $\delta(f, g, p) \geq \delta(f, f, p)=p^{2}+q^{2}$ is strict for densities $f$ and $g$ differing on a set of positive measure (see [9], Theorem 1 and Corollary 1), it follows that a level- $\alpha$ test which rejects $H_{0}$ for small values of $R_{m, n}$ is consistent against general alternatives. Such a test may be carried out as an exact permutation test.

REmark 2. Numerical estimates of $\operatorname{Var}\left(D_{d}\right)$ for low dimensions are given in Section 2, along with a proof of Theorem 1. Interestingly, the dependence of $\sigma_{d}^{2}$ on the dimension $d$ via $\operatorname{Var}\left(D_{d}\right)$ vanishes if $p=1 / 2$ since then $\sigma_{d}^{2}=1 / 4$. It is also of interest to compare $\sigma_{d}^{2}$ with the asymptotic variance of a closely related two-sample statistic considered in [21] and [13], namely the number

TABLE 1 Estimates of $\alpha_{k, d}\left(=P\left(D_{d}=k\right)\right)$ and $\operatorname{Var}\left(D_{d}\right)$

|  | $\boldsymbol{k}$ |  |  |  |  |  |  |  |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{d}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\widehat{\operatorname{Var}}\left(\boldsymbol{D}_{\boldsymbol{d}}\right)$ |  |
| 2 | 0.221 | 0.566 | 0.206 | 0.007 | 0.000 | - | - | 0.455 | cf. [22] |
| 2 | 0.2108 | 0.5694 | 0.2121 | 0.0077 | 0.0000 | - | - | 0.453 |  |
| 3 | 0.2858 | 0.4595 | 0.2216 | 0.0314 | 0.0017 | 0.0000 | 0.0000 | 0.648 |  |
| 4 | 0.3021 | 0.4238 | 0.2209 | 0.0478 | 0.0052 | 0.0002 | 0.0000 | 0.763 |  |
| $\infty$ | 0.40658 | 0.32429 | 0.17112 | 0.06835 | 0.02201 | 0.00593 | 0.00138 | 1.192 |  |

$\mathbf{N}_{m, n}$ of elements of the pooled sample $\mathscr{X}_{m} \cup \mathscr{Y}_{n}$ that have a nearest neighbor from the same sample. The asymptotic variance of $\mathbf{N}_{m, n}$ under $H_{0}$ is

$$
\tilde{\sigma}_{d}^{2}=r\left(1+v_{d}\right)+\frac{1}{2} \operatorname{Var}\left(\tilde{D}_{d}\right)(1-2 r)
$$

(see [13], Proposition 3.3). Here $v_{d}$ is the probability that 0 is the nearest neighbor of its own nearest neighbor in $\mathscr{P}_{1}$, and $\tilde{D}_{d}$ stands for the number of points of $\mathscr{P}_{1}$ which have the origin as their nearest neighbor. If $p=1 / 2$, then $\tilde{\sigma}_{d}^{2}=\left(1+v_{d}\right) / 2$ so that, in contrast to the Friedman-Rafsky statistic, there is still a dependence of $\tilde{\sigma}_{d}^{2}$ on $d$ via the probability $v_{d}$ for the "reciprocity" of the nearest neighbor relation. A closed-form expression for $v_{d}$ is given in [18] (see also [12]).
2. The limiting null distribution. Some limited information on $\operatorname{Var}\left(D_{d}\right)$ and thus on $\sigma_{d}^{2}$ may be obtained from Table 1 which presents estimates $\hat{\alpha}_{k, d}$ of the probabilities $\alpha_{k, d}=P\left(D_{d}=k\right)$ and hence also an estimate $\widehat{\operatorname{Var}}\left(D_{d}\right)$ of $\operatorname{Var}\left(D_{d}\right)$ for the cases $d=2,3,4$.

The first row reproduces the estimates $\hat{\alpha}_{k, 2}$ obtained in [22] as the average fraction of observed vertices of degree $k$ from 20 independently generated minimal spanning trees, each tree formed by 65,536 vertices taken independently at random from the unit square. The entries in the $d$ th row, where $d=2,3,4$, are the average fractions out of 10,000 independent replications of the MST formed by 0 and the nearest, second-nearest, $\ldots, 1,000$ th nearest neighbor of 0 in $\mathscr{T}\left(\mathscr{P}_{1}\right)$ on $\mathbb{R}^{d}$, in which the degree of the vertex at 0 is $k$. Since, for low dimensions such as 2,3 or 4 , the union of the nearest, secondnearest, $\ldots, 1,000$ th nearest neighbor of 0 should with high probability be a "blocking set around the origin" in the language of [16], this simulation design should produce a variable with a distribution very close to that of $D_{d}$. Computations were carried out at the Rechenzentrum of the University of Karlsruhe using an IBM RS/6000 SP parallel computer. The CPU computing time for the case $d=4$ was about 15 hours.

It is known [17] that $\alpha_{k, d} \rightarrow \alpha_{k}$ as $d \rightarrow \infty$, where

$$
\alpha_{k}=\int_{0}^{1} \exp (-\varphi(u)) \frac{\varphi(u)^{k+1}}{(k+1)!} d u
$$

and

$$
\varphi(u)=\int_{0}^{u} \frac{\log (1 / x)}{1-x} d x, \quad u<1
$$

(see [1], page 385). If $D_{\infty}$ denotes a variable with $P\left[D_{\infty}=k\right]=\alpha_{k}(k=1,2$, $3, \ldots$ ), then $E\left[D_{\infty}\right]=2$ (see [1]) and $\operatorname{Var}\left(D_{d}\right) \rightarrow \operatorname{Var}\left(D_{\infty}\right)$ as $d \rightarrow \infty$. This can be proved using the methods of [17], in particular Lemma 3 and the proof of Lemma 4 from that paper.

The row denoted " $\infty$ " in Table 1 contains numerical values for $\alpha_{k}$. These were obtained using an IMSL routine (Gauss-Kronrod numerical integration) and, complemented by $\alpha_{8}=0.00028$ and $\alpha_{9}=0.00005$, should be accurate up to five digits, in contrast with the values given in [1], page 396, which gives $E\left(D_{\infty}\right)=1.994$ when it should be 2 (the values in [1] were reported incorrectly in [17]).

Proof of Theorem 1. The conditional variance of $R_{m, n}$ given the pooled sample $\mathscr{X}_{m} \cup \mathscr{Y}_{n}$, is

$$
\begin{align*}
& \operatorname{Var}\left(R_{m, n} \mid \mathscr{X}_{m} \cup \mathscr{Y}_{n}\right) \\
& =  \tag{2}\\
& \quad \frac{2 m n}{N(N-1)} \\
& \quad \times\left(\frac{2 m n-N}{N}+\frac{C_{N}-N+2}{(N-2)(N-3)}[N(N-1)-4 m n+2]\right),
\end{align*}
$$

where $N=m+n$ is the total sample size, and $C_{N}$ is the number of edge pairs in $\mathscr{T}\left(\mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)$ that share a common vertex (see [8], page 701). Putting

$$
\tilde{R}_{m, n}=\frac{R_{m, n}-2 m n /(m+n)}{\operatorname{Var}\left(R_{m, n} \mid \mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)^{1 / 2}},
$$

Theorem 4.1.2 of [5] yields almost sure asymptotic normality of $\tilde{R}_{m, n}$ under the usual limiting regime, that is, $\lim P\left(\tilde{R}_{m, n} \leq t \mid \mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)=\Phi(t)$ almost surely for each $t \in \mathbb{R}$, where $\Phi$ is the standard normal distribution function. Since, in the usual limiting regime,

$$
\frac{\operatorname{Var}\left(R_{m, n} \mid \mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)}{m+n}=r\left(r+\left(\frac{C_{N}}{N}-1\right)(1-2 r)\right)+o_{P}(1),
$$

it remains to prove

$$
\frac{C_{N}}{N}-1 \rightarrow \frac{1}{2} \operatorname{Var}\left(D_{d}\right) \quad \text { in probability. }
$$

To this end, note first that $E\left[D_{d}\right]=2$ by Lemma 7 of [2], so $\frac{1}{2} \operatorname{Var}\left(D_{d}\right)=$ $\frac{1}{2} E\left[D_{d}^{2}\right]-2$. Note also that $C_{N}=1 / 2 \sum_{i=1}^{N} G_{i}^{2}-(N-1)$, where $G_{i}$ is the degree of the $i$ th vertex in $\mathscr{T}\left(\mathscr{X}_{m} \cup \mathscr{Y}_{n}\right)$, and the vertices are numbered completely at
random. Furthermore,

$$
\frac{1}{N} \sum_{i=1}^{N} G_{i}^{2}=\sum_{k=1}^{K_{d}} k^{2} \frac{V_{k}(N)}{N}
$$

where $V_{k}(N)$ is the number of vertices in $\mathscr{T}\left(\mathscr{X}_{m} \cup Y_{n}\right)$ with degree $k$, and $K_{d}$ is the largest possible degree of any vertex of any MST in $\mathbb{R}^{d}$ (see [2], Lemma 4). Since $V_{k}(N) / N$ converges almost surely to $P\left(D_{d}=k\right)$ ([17], page 1905), the proof is complete.

## 3. Proof of Theorem 2.

Lemma 1. If $S, T$ and $\{x\}$ are disjoint sets in $\mathbb{R}^{d}$ such that $S \cup T \cup\{x\}$ is nice,

$$
\begin{equation*}
|R(S \cup\{x\}, T)-R(S, T)| \leq K_{d}, \tag{3}
\end{equation*}
$$

where $K_{d}$ is given in the proof of Theorem 1.
Proof. By the revised add and delete algorithm of Lee [16], page 1000, the graph $\mathscr{T}(S \cup T)$ can be modified to get $\mathscr{T}(S \cup\{x\} \cup T)$ by adding at most $K_{d}$ edges [those edges of $\mathscr{T}(S \cup\{x\} \cup T)$ which have an endpoint at $\{x\}$ ] and deleting at most $K_{d}-1$ other edges of $\mathscr{T}(S \cup T)$. Then (3) follows.

In the next result, suppose $\phi$ and $\phi_{k}, k \geq 1$, are probability density functions on $\mathbb{R}^{d}$ with identical support, and with $\phi_{k}(x) / \phi(x) \rightarrow 1$ as $k \rightarrow \infty$, uniformly on $\{x: \phi(x)>0\}$. The most interesting special case has $\phi_{k} \equiv \phi$, but the more general case is needed later on. Recall that $x \in \mathbb{R}^{d}$ is a Lebesgue point of $\phi$ if the average of $|\phi(\cdot)-\phi(x)|$ over small balls centered at $x$ tends to zero. Almost every $x \in \mathbb{R}^{d}$ is a Lebesgue point of $\phi$; see, for example, [20], Theorem 7.7.

Proposition 1. Let $h: \mathbb{R}^{d} \times \mathbb{R}^{d} \rightarrow[0,1]$ be a symmetric, jointly measurable function, such that for almost every $x \in \mathbb{R}^{d}, h(x, \cdot)$ is measurable with $x$ a Lebesgue point of the function $\phi(\cdot) h(x, \cdot)$. For each $k$, let $V_{1}^{k}, V_{2}^{k}, \ldots, V_{k}^{k}$ be independent d-dimensional variables with common density function $\phi_{k}$, and set $\mathscr{V}_{k}=\left\{V_{1}^{k}, \ldots, V_{k}^{k}\right\}$. Then
(4) $\lim _{k \rightarrow \infty} k^{-1} E \sum_{1 \leq i<j \leq k} h\left(V_{i}^{k}, V_{j}^{k}\right) \mathbf{1}\left\{\left(V_{i}^{k}, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}\right)\right\}=\int_{\mathbb{R}^{d}} h(x, x) \phi(x) d x$.

Proof. Given any nice $S \subset \mathbb{R}^{d}$, and given $x \in S$, let $\Delta(x ; S)$ denote the degree of vertex $x$ in the MST or MSF $\mathscr{T}(S)$. Let $\Delta_{K}(x ; S)$ be the total number of edges of $\mathscr{T}(S)$, of length at most $K$, with one end at $x$. Let $\Delta^{K}(x ; S)=$ $\Delta(x ; S)-\Delta_{K}(x ; S)$. For $a \in \mathbb{R}$, and $x \in \mathbb{R}^{d}$, set $a S=\{a X: X \in S\}$ and $S-x=\{X-x: X \in S\}$. Let $\rightarrow_{\mathscr{O}}$ denote weak convergence of point processes as $k \rightarrow \infty$, where the topology on point measures on $\mathbb{R}^{d}$ is as described in [2].

Let $x$ be a Lebesgue point of $\phi$ with $\phi(x)>0$. Let $\mathscr{V}_{k}^{x}$ be the point process $\left\{x, V_{2}^{k}, V_{3}^{k}, \ldots, V_{k}^{k}\right\}$, and let $\mathscr{W}_{k}^{x}=k^{1 / d}\left(\mathscr{V}_{k}^{x}-x\right)$. By Proposition 3.21 of [19] and Theorem 7.10 of [20], $\mathscr{V}_{k}^{x} \rightarrow_{\mathscr{D}} \phi(x)^{-1 / d} \mathscr{P}_{\phi(x)}$, with $\mathscr{P}_{\lambda}$ as defined in Section 1.

We follow pages $253-254$ of [2]. By the Skorohod representation theorem, we can take coupled point processes $\tilde{\mathscr{H}}_{k}^{x}$ and $\tilde{\mathscr{P}}_{\phi(x)}$ with the same distribution as $\mathscr{W}_{k}^{x}$ and $\mathscr{P}_{\phi(x)}$, respectively, satisfying $\tilde{\mathscr{V}}_{k}^{x} \rightarrow \tilde{\mathscr{P}}_{\phi(x)}$ as $k \rightarrow \infty$, almost surely. By Lemma 6(a) of [2],

$$
\liminf _{k \rightarrow \infty} \Delta\left(0 ; \tilde{\mathscr{W}}_{k}^{x}\right) \geq \Delta\left(0 ; \tilde{\mathscr{P}}_{\phi(x)}\right) \quad \text { a.s. }
$$

By Lemma 7 of $[2], E\left[\Delta\left(0 ; \mathscr{P}_{\phi(x)}\right)\right]=2$. So by Fatou's lemma,

$$
\begin{equation*}
2 \leq E \liminf _{k \rightarrow \infty} \Delta\left(0 ; \tilde{\mathscr{W}}_{k}^{x}\right) \leq \liminf _{k \rightarrow \infty} E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right) \tag{5}
\end{equation*}
$$

Similarly, for any $K>0$,

$$
\begin{equation*}
E \Delta_{K}\left(0 ; \mathscr{P}_{\phi(x)}\right) \leq \liminf _{k \rightarrow \infty} E \Delta_{K}\left(0 ; \mathscr{W}_{k}^{x}\right) \tag{6}
\end{equation*}
$$

By (5) and Fatou's lemma again,

$$
\begin{align*}
2 & =\int 2 \phi(x) d x \leq \int \liminf _{k \rightarrow \infty} E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right) \phi_{k}(x) d x \\
& \leq \int \limsup _{k \rightarrow \infty} E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right) \phi_{k}(x) d x \leq \limsup _{k \rightarrow \infty} \int E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right) \phi_{k}(x) d x \tag{7}
\end{align*}
$$

Since the total number of edges of $\mathscr{T}\left(\mathscr{V}_{k}\right)$ is $k-1$, it follows that $E \Delta\left(V_{i}^{k} ; \mathscr{V}_{k}\right)=$ $2-2 / k$ for each $i$, and hence $\int E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right) \phi_{k}(x) d x=2-(2 / k)$, so the inequalities in (7) are all equalities. In particular, for almost all $x$ with $\phi(x)>0$,

$$
\begin{equation*}
\lim _{k \rightarrow \infty} E \Delta\left(0 ; \mathscr{W}_{k}^{x}\right)=2 \tag{8}
\end{equation*}
$$

and by (6),

$$
\begin{equation*}
\limsup _{k \rightarrow \infty} E\left[\Delta^{K}\left(0 ; \mathscr{W}_{k}^{x}\right)\right] \leq 2-E \Delta_{K}\left(0 ; \mathscr{P}_{\phi(x)}\right) \tag{9}
\end{equation*}
$$

Let $B(x, r)=\{y:|y-x| \leq r\}$. For any positive $K$,

$$
\begin{aligned}
& E \sum_{j=2}^{k}\left|h\left(x, V_{j}^{k}\right)-h(x, x)\right| \mathbf{1}\left\{V_{j}^{k}\right.\left.\in B\left(x ; K k^{-1 / d}\right)\right\} \\
&=(k-1) \int_{B\left(x ; K k^{-1 / d}\right)} \mid\left(h(x, y) \phi_{k}(y)-h(x, x) \phi_{k}(x)\right) \\
&+h(x, x)\left(\phi_{k}(x)-\phi_{k}(y)\right) \mid d y
\end{aligned}
$$

which tends to zero provided $x$ is a Lebesgue point of both $\phi$ and $h(x, \cdot) \phi(\cdot)$. Therefore, since $h$ has range [0,1],

$$
\begin{align*}
& \limsup _{k \rightarrow \infty} E \sum_{j=2}^{k}\left|h\left(x, V_{j}^{k}\right)-h(x, x)\right| \mathbf{1}\left\{\left(x, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}^{x}\right)\right\}  \tag{10}\\
& \quad \leq \limsup _{k \rightarrow \infty} E \Delta^{K}\left(0 ; \mathscr{W}_{k}^{x}\right)
\end{align*}
$$

and by (9), this can be made arbitrarily small by choice of $K$. Hence the left side of (10) is zero, so for almost all $x$ with $\phi(x)>0$,

$$
\begin{equation*}
E \sum_{j=2}^{k} h\left(x, V_{j}^{k}\right) \mathbf{1}\left\{\left(x, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}^{x}\right)\right\}=h(x, x) E \Delta\left(x ; \mathscr{V}_{k}^{x}\right)+o(1) . \tag{11}
\end{equation*}
$$

Since $h$ has range [ 0,1 ], the left-hand side of (11) is bounded by $K_{d}$ (defined in the proof of Theorem 1), while the right-hand side which tends to $2 h(x, x)$ by (8). Hence, by the dominated convergence theorem,

$$
\begin{aligned}
k^{-1} E & \sum_{1 \leq i<j \leq k} h\left(V_{i}^{k}, V_{j}^{k}\right) \mathbf{1}\left\{\left(V_{i}^{k}, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}\right)\right\} \\
\quad= & \frac{1}{2} E \sum_{j=2}^{k} h\left(V_{1}^{k}, V_{j}^{k}\right) \mathbf{1}\left\{\left(V_{1}^{k}, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}\right)\right\} \\
& =\frac{1}{2} \int \phi_{k}(x) d x E \sum_{j=2}^{k} h\left(x, V_{j}^{k}\right) \mathbf{1}\left\{\left(x, V_{j}^{k}\right) \in \mathscr{T}\left(\mathscr{V}_{k}^{x}\right)\right\} \\
& \rightarrow \int \phi(x) h(x, x) d x .
\end{aligned}
$$

Proof of Theorem 2. Let $M_{m}$ and $N_{n}$ be Poisson variables with mean $m$ and $n$, respectively, independent of one another and of $\left\{X_{i}\right\}$ and $\left\{Y_{j}\right\}$. Let $\mathscr{X}_{m}^{\prime}$ and $\mathscr{Y}_{n}^{\prime}$ be the Poisson processes $\left\{X_{1}, \ldots, X_{M_{m}}\right\}$ and $\left\{Y_{1}, \ldots, Y_{N_{n}}\right\}$, respectively. Set $R_{m, n}^{\prime}=R\left(\mathscr{X}_{m}^{\prime}, \mathscr{Y}_{n}^{\prime}\right)$. By Lemma 1,

$$
\begin{equation*}
\left|R_{m, n}^{\prime}-R_{m, n}\right| \leq K_{d}\left(\left|M_{m}-m\right|+\left|N_{n}-n\right|\right) . \tag{12}
\end{equation*}
$$

We shall prove below that in the usual limiting regime,

$$
\begin{equation*}
\frac{E\left[R_{m, n}^{\prime}\right]}{m+n} \rightarrow 2 p q \int \frac{f(x) g(x)}{p f(x)+q g(x)} d x . \tag{13}
\end{equation*}
$$

This will suffice, since $(m+n)^{-1} E\left|R_{m, n}^{\prime}-R_{m, n}\right| \rightarrow 0$ by (12), so that $E R_{m, n} /(m+n)$ also converges to the right side of (13). By Lemma 1, we can then apply Theorem 2.3 of [14] (with $d_{m, n}$ of that paper equal to a constant), to obtain (1).

It remains to prove (13). The point of the Poissonization is that the sample identities of the points of $\mathscr{X}_{m}^{\prime} \cup \mathscr{Y}_{n}^{\prime}$ are conditionally independent, given their positions. To make this precise, for each $m, n$ let $Z_{1}^{m, n}, Z_{2}^{m, n}, Z_{3}^{m, n}, \ldots$ be independent variables with common density $\phi_{m, n}(x):=(m f(x)+n g(x)) /$
$(m+n), x \in \mathbb{R}^{d}$. Let $L_{m, n}$ be an independent Poisson variable with mean $m+n$. Let $\mathscr{\mathscr { O }}_{m, n}^{\prime}=\left\{Z_{1}^{m, n}, \ldots, Z_{L_{m, n}}^{m, n}\right\}$, a nonhomogeneous Poisson process of rate $m f+n g$.

Assign a mark from the set $\{1,2\}$ to each point of $\mathscr{F}_{m, n}^{\prime}$, a point at $x$ being assigned the mark 1 with probability $m f(x) /(\underset{\sim}{m} f(x)+n g(x))$ and a mark 2 otherwise, independently of other points. Let $\tilde{\mathscr{X}}_{m}^{\prime}$ be the set of points of $\mathscr{g}^{\prime}{ }_{m, n}$ marked 1, and let $\tilde{\mathscr{Y}}_{n}^{\prime}$ be the set of points of $\mathscr{\mathscr { P }}_{m, n}^{\prime}$ marked 2. By the marking theorem [15], $\tilde{\mathscr{X}}_{m}^{\prime}$ and $\tilde{\mathscr{Y}}_{n}^{\prime}$ are independent Poisson processes with the same distribution as $\mathscr{X}_{m}^{\prime}$ and $\mathscr{Y}_{n}^{\prime}$, respectively. Hence $\tilde{R}_{m, n}^{\prime}:=R\left(\tilde{\mathscr{X}}_{m}^{\prime}, \tilde{\mathscr{Y}}_{n}^{\prime}\right)$ has the same distribution as $R_{m, n}^{\prime}$, and it suffices to prove (13) with $R_{m, n}^{\prime}$ replaced by $\tilde{R}_{m, n}^{\prime}$.

Given points of $\mathscr{F}_{m, n}^{\prime}$ at $x$ and $y$, the probability that they have different marks is given by

$$
h_{m, n}(x, y):=\frac{m f(x) n g(y)+n g(x) m f(y)}{(m f(x)+n g(x))(m f(y)+n g(y))}
$$

Then

$$
\begin{equation*}
E\left[\tilde{R}_{m, n}^{\prime} \mid \mathscr{\mathscr { O }}_{m, n}^{\prime}\right]=\sum_{i<j \leq L_{m, n}} h_{m, n}\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \mathbf{1}\left\{\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \in \mathscr{T}\left(\mathscr{\mathscr { P }}_{m, n}^{\prime}\right)\right\} \tag{14}
\end{equation*}
$$

Set

$$
h(x, y)=\frac{p q(f(x) g(y)+g(x) f(y))}{(p f(x)+q g(x))(p f(y)+q g(y))}
$$

Observe that both $h_{m, n}$ and $h$ have range [0, 1]. In the usual limiting regime, $h_{m, n} \rightarrow h$ uniformly. Taking expectations in (14), we have

$$
\begin{align*}
& E\left[\tilde{R}_{m, n}^{\prime}\right] \\
& \quad=E \sum_{i<j \leq L_{m, n}} h\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \mathbf{1}\left\{\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \in \mathscr{T}\left(\mathscr{\mathscr { O }}_{m, n}^{\prime}\right)\right\}+o(m+n) \tag{15}
\end{align*}
$$

Let $\mathscr{\mathscr { F }}_{m, n}$ be the non-Poisson point process $\left\{Z_{1}^{m, n}, Z_{2}^{m, n}, \ldots, Z_{m+n}^{m, n}\right\}$. By the proof of Lemma 1 and the fact that $E\left[\left|M_{m}+N_{n}-m-n\right|\right]=o(m+n)$,

$$
E\left[\tilde{R}_{m, n}^{\prime}\right]=E \sum_{i<j \leq m+n} \sum_{i} h\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \mathbf{1}\left\{\left(Z_{i}^{m, n}, Z_{j}^{m, n}\right) \in \mathscr{T}(\mathscr{g} m, n)\right\}+o(m+n)
$$

Set $\phi(x)=p f(x)+q g(x)$. Then $\phi_{m, n}(x) / \phi(x) \rightarrow 1$, uniformly on $\{x: \phi(x)>0\}$. By Proposition 1,

$$
\frac{E \tilde{R}_{m, n}^{\prime}}{m+n} \rightarrow \int h(x, x) \phi(x) d x=\int \frac{2 p q f(x) g(x)}{p f(x)+q g(x)} d x
$$

Acknowledgment. The authors thank Nora Gürtler for assistance in computation.

## REFERENCES

[1] Aldous, D. (1990). A Random tree model associated with random graphs. Random Structures Algorithms 1 383-401.
[2] Aldous, D. and Steele, J. M. (1992). Asymptotics for Euclidean minimal spanning trees on random points. Probab. Theory Related Fields 92 247-258.
[3] Anderson, N. H., Hall, P. and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. J. Multivariate Anal. 50 41-54.
[4] BAHR, R. (1996). A new test for the multivariate two-sample problem with general alternatives (in German). Doctoral thesis, Univ. Hannover.
[5] Bloemena, A. R. (1964). Sampling from a graph. Mathematical Centre Tracts 2. Math. Centrum, Amsterdam.
[6] Einmahl, J. H. J. and Khmaladze, E. V. (1998). The two-sample problem in $\mathbb{R}^{m}$ and measure-valued martingales. Report S98-2, Dept. Statistics, Univ. New South Wales, Sidney.
[7] Ferger, D. (1997). Optimal tests for the general two-sample problem. Dresdener Schriften zur Mathematischen Stochastik Technische Univ. Dresden.
[8] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests. Ann. Statist. 7 697-717.
[9] Györfi, L. and Nemetz, T. (1975). $f$-dissimilarity: A general class of separation measures of several probability measures. In Topics in Information Theory. Colloq. Math. Soc. János Bolyai 16 309-321.
[10] Györfi, L. and Nemetz, T. (1977). On the dissimilarity of probability measures. Problems Control Inform. Theory 6 263-267.
[11] Györfi, L. and Nemetz, T. (1978). $f$-dissimilarity. A generalization of affinity of several distributions. Ann. Inst. Statist. Math. 30 105-113.
[12] Henze, N. (1986). On the probability that a random point is the $j$ th nearest neighbour to its own $k$ th nearest neighbour. J. Appl. Probab. 23 221-226.
[13] HEnze, N. (1988). A multivariate two-sample test based on the number of nearest-neighbor type coincidences. Ann. Statist. 16 772-783.
[14] Henze, N. and Voigt, B. (1992). Almost sure convergence of certain slowly changing symmetric one- and multi-sample statistics. Ann. Probab. 20 1086-1098.
[15] Kingman, J. F. C. (1993). Poisson Processes. Oxford Univ. Press.
[16] Lee, S. (1997). The central limit theorem for Euclidean minimal spanning trees I. Ann. Appl. Probab. 7 996-1020.
[17] Penrose, M. D. (1996). The random minimal spanning tree in high dimensions. Ann. Probab. 24 1903-1925.
[18] Pickard, D. K. (1982). Isolated nearest neighbours. J. Appl. Probab. 19 444-449.
[19] Resnick, S. I. (1987). Extreme Values, Regular Variation, and Point Processes. Springer, New York.
[20] Rudin, W. (1987). Real and Complex Analysis, 3rd ed. McGraw-Hill, New York.
[21] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. J. Amer. Statist. Assoc. 81 799-806.
[22] Steele, J. M., Shepp, L. A. and Eddy, W. F. (1987). On the number of leaves of a Euclidean minimal spanning tree. J. Appl. Prob. 24 809-826.

Institut für Mathematische Stochastik
Universität Karlsruhe
Englerstr. 2
D-76128 KarLSRUHE
Germany
E-MAIL: norbert.henze@math.uni-karlsruhe.de

Department of Mathematical Sciences
University of Durham
South Road
Durham DH1 3LE
United Kingdom
E-MAIL: mathew.penrose@durham.ac.uk


[^0]:    Received July 1998; revised December 1998.
    AMS 1991 subject classifications. Primary 62H15; secondary 62G10, 60F05, 60F15.
    Key words and phrases. Multivariate two-sample problem, minimal spanning tree, multivariate runs test, homogeneous Poisson process.

