The Annals of Statistics
1999, Vol. 27, No. 1, 290–298

ON THE MULTIVARIATE RUNS TEST

BY NORBERT HENZE AND MATHEW D. PENROSE

Universität Karlsruhe and University of Durham

For independent d-variate random variables $X_1, \dots, X_m$ with common density f and $Y_1, \dots, Y_n$ with common density g, let $R_{m,n}$ be the number of edges in the minimal spanning tree with vertices $X_1, \dots, X_m$, $Y_1, \dots, Y_n$ 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 $X_1, X_2, X_3, \dots$ are independent d-dimensional variables with common probability density function f, and independently, $Y_1, Y_2, \dots$ 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 $\mathscr{X}_m := \{X_1, \dots, X_m\}$ and $\mathscr{Y}_n := \{Y_1, \dots, Y_n\}$, 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 $l(\mathscr{T}) \leq l(\mathscr{T}')$ for all spanning trees $\mathscr{T}'$. 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

$$R_{m,n} = R(\mathscr{X}_m, \mathscr{Y}_n).$$

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.

In fact, Friedman and Rafsky consider $1 + R_{m,n}$, which is the number of disjoint subtrees that result from removing all edges of $\mathcal{T}(\mathcal{X}_m \cup \mathcal{Y}_n)$ 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 $m \to \infty$ and $n \to \infty$ in a linked manner so that $m/(m+n) \to p \in (0,1)$, which we shall call the *usual limiting regime*. Set $q = 1 - p$ and $r = 2pq$, and write $\to_{\mathcal{D}}$ for convergence in distribution. Let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with expectation μ and variance σ^2. For $\lambda > 0$, let $\mathcal{P}_\lambda$ denote a homogeneous Poisson process on $\mathbb{R}^d$ of rate λ, with a point added at the origin.

THEOREM 1. *In the usual limiting regime, under* H_0,

$$(m+n)^{-1/2}\left(R_{m,n} - \frac{2mn}{m+n}\right) \to_{\mathcal{D}} \mathcal{N}(0, \sigma_d^2),$$

where

$$\sigma_d^2 = r\big(r + \tfrac{1}{2}\operatorname{Var}(D_d)\,(1-2r)\big).$$

Here D_d is the degree of the vertex at 0 in the MSF $\mathcal{T}(\mathcal{P}_1)$.

THEOREM 2. *In the usual limiting regime,*

$$\frac{R_{m,n}}{m+n} \to 2pq \int \frac{f(x)g(x)}{pf(x)+qg(x)}\,dx \quad \textit{almost surely}. \tag{1}$$

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

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

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-α 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 $\operatorname{Var}(D_d)$ for low dimensions are given in Section 2, along with a proof of Theorem 1. Interestingly, the dependence of σ_d^2 on the dimension d via $\operatorname{Var}(D_d)$ vanishes if $p = 1/2$ since then $\sigma_d^2 = 1/4$. It is also of interest to compare σ_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 $\alpha_{k,d}$ $(= P(D_d = k))$ *and* $\mathrm{Var}(D_d)$

	k								
d	**1**	**2**	**3**	**4**	**5**	**6**	**7**	$\widehat{\mathbf{Var}}(\boldsymbol{D_d})$	
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	
∞	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

$$\tilde{\sigma}_d^2 = r(1 + v_d) + \tfrac{1}{2}\,\mathrm{Var}(\tilde{D}_d)\;(1 - 2r)$$

(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 $\tilde{\sigma}_d^2 = (1 + v_d)/2$ 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 $\mathrm{Var}(D_d)$ and thus on σ_d^2 may be obtained from Table 1 which presents estimates $\hat{\alpha}_{k,d}$ of the probabilities $\alpha_{k,d} = P(D_d = k)$ and hence also an estimate $\widehat{\mathrm{Var}}(D_d)$ of $\mathrm{Var}(D_d)$ 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 dth 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, ..., 1,000th nearest neighbor of 0 in $\mathscr{T}(\mathscr{P}_1)$ 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, second-nearest, ..., 1,000th 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 $\alpha_{k,d} \to \alpha_k$ as $d \to \infty$, where

$$\alpha_k = \int_0^1 \exp(-\varphi(u))\frac{\varphi(u)^{k+1}}{(k+1)!}\,du$$

and

$$\varphi(u) = \int_0^u \frac{\log(1/x)}{1-x}\,dx, \qquad u < 1$$

(see [1], page 385). If D_∞ denotes a variable with $P[D_\infty = k] = \alpha_k$ ($k = 1, 2, 3, \ldots$), then $E[D_\infty] = 2$ (see [1]) and $\mathrm{Var}(D_d) \to \mathrm{Var}(D_\infty)$ as $d \to \infty$. 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 "∞" in Table 1 contains numerical values for α_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 $E(D_\infty) = 1.994$ 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

$$\begin{aligned}&\mathrm{Var}(R_{m,n} | \mathscr{X}_m \cup \mathscr{Y}_n)\\ &\quad= \frac{2mn}{N(N-1)}\\ &\qquad\times \left(\frac{2mn-N}{N} + \frac{C_N - N + 2}{(N-2)(N-3)}[N(N-1) - 4mn + 2]\right),\end{aligned} \tag{2}$$

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

$$\tilde{R}_{m,n} = \frac{R_{m,n} - 2mn/(m+n)}{\mathrm{Var}(R_{m,n} | \mathscr{X}_m \cup \mathscr{Y}_n)^{1/2}},$$

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

$$\frac{\mathrm{Var}(R_{m,n} | \mathscr{X}_m \cup \mathscr{Y}_n)}{m+n} = r\left(r + \left(\frac{C_N}{N} - 1\right)(1-2r)\right) + o_P(1),$$

it remains to prove

$$\frac{C_N}{N} - 1 \to \frac{1}{2}\mathrm{Var}(D_d) \quad \text{in probability.}$$

To this end, note first that $E[D_d] = 2$ by Lemma 7 of [2], so $\frac{1}{2}\mathrm{Var}(D_d) = \frac{1}{2}E[D_d^2] - 2$. Note also that $C_N = 1/2\sum_{i=1}^N G_i^2 - (N-1)$, where G_i is the degree of the ith vertex in $\mathscr{T}(\mathscr{X}_m \cup \mathscr{Y}_n)$, 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 $\mathscr{T}(\mathscr{X}_m \cup Y_n)$ 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 $P(D_d = k)$ ([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*,

$$|R(S \cup \{x\}, T) - R(S, T)| \leq K_d, \tag{3}$$

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 ϕ and ϕ_k, $k \geq 1$, are probability density functions on $\mathbb{R}^d$ with identical support, and with $\phi_k(x)/\phi(x) \to 1$ as $k \to \infty$, 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 ϕ 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 ϕ; see, for example, [20], Theorem 7.7.

PROPOSITION 1. *Let $h: \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ 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 $V_1^k, V_2^k, \dots, V_k^k$ be independent d-dimensional variables with common density function ϕ_k, and set $\mathscr{V}_k = \{V_1^k, \dots, V_k^k\}$. Then*

$$\lim_{k\to\infty} k^{-1} E \sum\sum_{1 \leq i < j \leq k} h(V_i^k, V_j^k)\mathbf{1}\{(V_i^k, V_j^k) \in \mathscr{T}(\mathscr{V}_k)\} = \int_{\mathbb{R}^d} h(x, x)\phi(x)\, dx. \tag{4}$$

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 $aS = \{aX: X \in S\}$ and $S - x = \{X - x : X \in S\}$. Let $\to_{\mathscr{D}}$ denote weak convergence of point processes as $k \to \infty$, where the topology on point measures on $\mathbb{R}^d$ is as described in [2].

Let x be a Lebesgue point of ϕ with $\phi(x) > 0$. Let $\mathscr{V}_k^x$ be the point process $\{x, V_2^k, V_3^k, \dots, V_k^k\}$, and let $\mathscr{W}_k^x = k^{1/d}(\mathscr{V}_k^x - x)$. By Proposition 3.21 of [19] and Theorem 7.10 of [20], $\mathscr{V}_k^x \to_{\mathscr{D}} \phi(x)^{-1/d}\mathscr{P}_{\phi(x)}$, 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 $\tilde{\mathscr{W}}_k^x$ and $\tilde{\mathscr{P}}_{\phi(x)}$ with the same distribution as $\mathscr{W}_k^x$ and $\mathscr{P}_{\phi(x)}$, respectively, satisfying $\tilde{\mathscr{W}}_k^x \to \tilde{\mathscr{P}}_{\phi(x)}$ as $k \to \infty$, almost surely. By Lemma 6(a) of [2],

$$\liminf_{k\to\infty} \Delta(0; \tilde{\mathscr{W}}_k^x) \ge \Delta(0; \tilde{\mathscr{P}}_{\phi(x)}) \quad \text{a.s.}$$

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

$$2 \le E \liminf_{k\to\infty} \Delta(0; \tilde{\mathscr{W}}_k^x) \le \liminf_{k\to\infty} E\Delta(0; \mathscr{W}_k^x). \tag{5}$$

Similarly, for any $K > 0$,

$$E\Delta_K(0; \mathscr{P}_{\phi(x)}) \le \liminf_{k\to\infty} E\Delta_K(0; \mathscr{W}_k^x). \tag{6}$$

By (5) and Fatou's lemma again,

$$\begin{aligned} 2 = \int 2\phi(x)\,dx &\le \int \liminf_{k\to\infty} E\Delta(0; \mathscr{W}_k^x)\phi_k(x)\,dx \\ &\le \int \limsup_{k\to\infty} E\Delta(0; \mathscr{W}_k^x)\phi_k(x)\,dx \le \limsup_{k\to\infty} \int E\Delta(0; \mathscr{W}_k^x)\phi_k(x)\,dx. \end{aligned} \tag{7}$$

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

$$\lim_{k\to\infty} E\Delta(0; \mathscr{W}_k^x) = 2, \tag{8}$$

and by (6),

$$\limsup_{k\to\infty} E[\Delta^K(0; \mathscr{W}_k^x)] \le 2 - E\Delta_K(0; \mathscr{P}_{\phi(x)}). \tag{9}$$

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

$$\begin{aligned} &E \sum_{j=2}^{k} |h(x, V_j^k) - h(x, x)| \mathbf{1}\{V_j^k \in B(x; Kk^{-1/d})\} \\ &\quad = (k-1)\int_{B(x;\,Kk^{-1/d})} \big|(h(x, y)\phi_k(y) - h(x, x)\phi_k(x)) \\ &\qquad\qquad + h(x, x)(\phi_k(x) - \phi_k(y))\big|\,dy, \end{aligned}$$

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

$$\limsup_{k\to\infty} E \sum_{j=2}^{k} |h(x, V_j^k) - h(x, x)| \mathbf{1}\{(x, V_j^k) \in \mathscr{T}(\mathscr{V}_k^x)\} \tag{10}$$
$$\leq \limsup_{k\to\infty} E\Delta^K(0; \mathscr{W}_k^x),$$

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 $\phi(x) > 0$,

$$E \sum_{j=2}^{k} h(x, V_j^k)\mathbf{1}\{(x, V_j^k) \in \mathscr{T}(\mathscr{V}_k^x)\} = h(x, x)E\Delta(x; \mathscr{V}_k^x) + o(1). \tag{11}$$

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 $2h(x, x)$ by (8). Hence, by the dominated convergence theorem,

$$\begin{aligned} k^{-1} E &\sum\sum_{1\le i<j\le k} h(V_i^k, V_j^k)\mathbf{1}\{(V_i^k, V_j^k) \in \mathscr{T}(\mathscr{V}_k)\} \\ &= \tfrac{1}{2} E \sum_{j=2}^{k} h(V_1^k, V_j^k)\mathbf{1}\{(V_1^k, V_j^k) \in \mathscr{T}(\mathscr{V}_k)\} \\ &= \tfrac{1}{2} \int \phi_k(x)\, dx E \sum_{j=2}^{k} h(x, V_j^k)\mathbf{1}\{(x, V_j^k) \in \mathscr{T}(\mathscr{V}_k^x)\} \\ &\to \int \phi(x) h(x, x)\, dx. \end{aligned}$$

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 $\{X_i\}$ and $\{Y_j\}$. Let $\mathscr{X}'_m$ and $\mathscr{Y}'_n$ be the Poisson processes $\{X_1, \dots, X_{M_m}\}$ and $\{Y_1, \dots, Y_{N_n}\}$, respectively. Set $R'_{m,n} = R(\mathscr{X}'_m, \mathscr{Y}'_n)$. By Lemma 1,

$$|R'_{m,n} - R_{m,n}| \leq K_d(|M_m - m| + |N_n - n|). \tag{12}$$

We shall prove below that in the usual limiting regime,

$$\frac{E[R'_{m,n}]}{m+n} \to 2pq \int \frac{f(x)g(x)}{pf(x) + qg(x)}\, dx. \tag{13}$$

This will suffice, since $(m+n)^{-1}E|R'_{m,n} - R_{m,n}| \to 0$ by (12), so that $ER_{m,n}/(m+n)$ 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 $\mathscr{X}'_m \cup \mathscr{Y}'_n$ are conditionally independent, given their positions. To make this precise, for each m, n let $Z_1^{m,n}, Z_2^{m,n}, Z_3^{m,n}, \dots$ be independent variables with common density $\phi_{m,n}(x) := (mf(x) + ng(x))/$

$(m+n)$, $x \in \mathbb{R}^d$. Let $L_{m,n}$ be an independent Poisson variable with mean $m+n$. Let $\mathscr{Z}'_{m,n} = \{Z_1^{m,n}, \dots, Z_{L_{m,n}}^{m,n}\}$, a nonhomogeneous Poisson process of rate $mf + ng$.

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

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

$$h_{m,n}(x, y) := \frac{mf(x)ng(y) + ng(x)mf(y)}{(mf(x)+ng(x))(mf(y)+ng(y))}.$$

Then

$$E[\tilde{R}'_{m,n}|\mathscr{Z}'_{m,n}] = \sum\sum_{i<j\le L_{m,n}} h_{m,n}(Z_i^{m,n}, Z_j^{m,n}) \mathbf{1}\{(Z_i^{m,n}, Z_j^{m,n}) \in \mathscr{T}(\mathscr{Z}'_{m,n})\}. \tag{14}$$

Set

$$h(x, y) = \frac{pq(f(x)g(y) + g(x)f(y))}{(pf(x)+qg(x))(pf(y)+qg(y))}.$$

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

$$\begin{aligned} &E[\tilde{R}'_{m,n}] \\ &\quad = E \sum\sum_{i<j\le L_{m,n}} h(Z_i^{m,n}, Z_j^{m,n}) \mathbf{1}\{(Z_i^{m,n}, Z_j^{m,n}) \in \mathscr{T}(\mathscr{Z}'_{m,n})\} + o(m+n). \end{aligned} \tag{15}$$

Let $\mathscr{Z}_{m,n}$ be the non-Poisson point process $\{Z_1^{m,n}, Z_2^{m,n}, \dots, Z_{m+n}^{m,n}\}$. By the proof of Lemma 1 and the fact that $E[|M_m + N_n - m - n|] = o(m+n)$,

$$E[\tilde{R}'_{m,n}] = E \sum\sum_{i<j\le m+n} h(Z_i^{m,n}, Z_j^{m,n}) \mathbf{1}\{(Z_i^{m,n}, Z_j^{m,n}) \in \mathscr{T}(\mathscr{Z}_{m,n})\} + o(m+n).$$

Set $\phi(x) = pf(x) + qg(x)$. Then $\phi_{m,n}(x)/\phi(x) \to 1$, uniformly on $\{x: \phi(x) > 0\}$. By Proposition 1,

$$\frac{E\tilde{R}'_{m,n}}{m+n} \to \int h(x, x)\phi(x)\,dx = \int \frac{2pqf(x)g(x)}{pf(x)+qg(x)}\,dx.$$

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 jth nearest neighbour to its own kth 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