The Annals of Statistics
1988, Vol. 16, No. 2, 772–783

A MULTIVARIATE TWO-SAMPLE TEST BASED ON THE NUMBER OF NEAREST NEIGHBOR TYPE COINCIDENCES

BY NORBERT HENZE

University of Hannover

For independent d-variate random samples $X_1, \dots, X_{n_1}$ i.i.d. $f(x)$, $Y_1, \dots, Y_{n_2}$ i.i.d. $g(x)$, where the densities f and g are assumed to be continuous a.e., consider the number T of all k nearest neighbor comparisons in which observations and their neighbors belong to the same sample. We show that, if $f = g$ a.e., the limiting (normal) distribution of T, as $\min(n_1, n_2) \to \infty$, $n_1/(n_1 + n_2) \to \tau$, $0 < \tau < 1$, does not depend on f. An omnibus procedure for testing the hypothesis H_0: $f = g$ a.e. is obtained by rejecting H_0 for large values of T. The result applies to a general distance (generated by a norm on $\mathbb{R}^d$) for determining nearest neighbors, and it generalizes to the multisample situation.

1. Introduction. Let $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ be independent $\mathbb{R}^d$-valued random vectors ("observations, points"), $d \geq 1$. The distribution of X_i has unknown pdf $f(x)$, say, and the distribution of Y_j has unknown pdf $g(x)$, say. We assume that f and g are continuous a.e. with respect to Lebesgue measure. The two-sample problem (TSP), which represents one of the classical problems of the theory of nonparametric inference, is then to test the hypothesis

$$H_0\colon f = g \quad \text{a.e.} \tag{1.1}$$

versus the general alternative that f and g differ on a set of positive measure. Of course, any reasonable test of (1.1) should meet the minimum requirements:

(a) The probability of an error of the first kind does not depend on f (the testing procedure should be distribution free).

(b) As $\min(n_1, n_2) \to \infty$, the test statistic is asymptotically distribution free under H_0, and the limiting distribution is known.

(c) The test is consistent against general alternatives.

In the univariate case many tests for the TSP meeting the preceding requirements have been proposed, the most prominent of these being the tests of Smirnov (1939), Wald and Wolfowitz (1940), Cramér and von Mises [see Rosenblatt (1952)], Lehmann (1951) and the empty box test of Wilks (1962).

A common feature of these procedures is that they only use the information provided by the ranks of observations within the sorted list of the pooled sample. Consequently, the respective test statistics are distribution free under H_0, which in turn implies property (a).

The multivariate case seems to have been studied far less fully. An intrinsic difficulty for extending the tests of Smirnov and Cramér and von Mises to the

Received September 1986; revised August 1987.

AMS 1980 *subject classifications*. Primary 62H15; secondary 62G10.

Key words and phrases. Multivariate two-sample test, nearest neighbor-type coincidences.

case $d \geq 2$ is the fact that monotonic transformations of the respective coordinates do not necessarily carry an arbitrary distribution to the uniform distribution on the unit d-cube. This explains why the respective statistics are no longer distribution free under H_0 in the multivariate case.

Bickel (1969), by applying Fisher's permutation principle, shows that it is possible to construct consistent distribution free multivariate Smirnov tests by conditioning on the empirical cdf (ecdf) of the pooled sample. However, this test lacks property (b) and is thus not satisfactory for practical purposes.

Friedman and Rafsky (1979) propose a multivariate two-sample test (*multivariate run test*) based on the minimal spanning tree of the sample points as a multivariate generalization of the univariate sorted list. By conditioning on the ecdf of the pooled sample, their procedure is distribution free, and the limiting permutation distribution of the proposed statistic is shown to be normal. It is not known whether the multivariate run test satisfies postulates (b) and (c).

Further proposals [Anderson (1966) and Weiss (1960)] lack a proof of consistency, and a test of Lehmann (1951) involves postexperimental randomization as an intrinsic factor, which is an undesirable feature.

In this paper we present a multivariate two-sample test that possesses properties (a), (b) and (c). To state the procedure, let $|\cdot|$ denote a fixed but otherwise arbitrary norm on $\mathbb{R}^d$, and put

$$(1.2) \qquad \begin{aligned} Z_i &= X_i, & 1 \leq i \leq n_1, \\ &= Y_{i-n_1}, & n_1 + 1 \leq i \leq n, \end{aligned}$$

where $n = n_1 + n_2$ is the total sample size. Define the rth nearest neighbor to Z_i [denoted by $N_r(Z_i)$] as that point Z_j satisfying $|Z_\nu - Z_i| < |Z_j - Z_i|$ for exactly $r - 1$ values of ν, $1 \leq \nu \leq n$; $\nu \neq i, j$, and write

$$(1.3) \qquad \begin{aligned} I_i(r) &= 1, \quad \text{if } Z_i \text{ and } N_r(Z_i) \text{ belong to the same sample,} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The random variable to be studied is

$$(1.4) \qquad T_{n,k} = \sum_{i=1}^{n} \sum_{r=1}^{k} I_i(r),$$

which represents the number of all *k nearest neighbor type coincidences*. Rejection of H_0 is for large values of $T_{n,k}$. To make the procedure distribution free, we may condition on the pooled sample and conduct an exact permutation test.

The purpose of this paper is twofold: First, the restriction to the Euclidean metric imposed in previous work [Henze (1984) and Schilling (1986b)], which is undesirable in view of the well-known problem of commensurability of the different coordinates, is removed. Second, we give a proof of asymptotic normality of $T_{n,k}$ under H_0 (Section 3) via almost sure asymptotic normality of the conditional distribution of $T_{n,k}$ given the pooled sample together with stochastic convergence of the conditional variance of $T_{n,k}$ to a limit not depending on the underlying density f. It is interesting to note that almost sure conditional asymptotic normality follows as a special case from the work of Bloemena (1964), which seems to have been largely forgotten (see Section 2). Since the conditional variance consistently estimates the limiting variance (which does not seem to be computable except for the Euclidean case) the test, conducted as an approximate permutation test, is applicable in case of a general norm.

Consistency is proved in Section 4. The test may be easily adapted to the multisample situation [Henze (1985)], and weighted versions of $T_{n,k}$ are possible in order to achieve high power against specific sequences of alternatives [see Section 4 of Schilling (1986b)].

It is understood that the random variables $X_1, X_2, \dots; Y_1, Y_2, \dots$ (and the random variables $U_{n1}, \dots, U_{nn}$, $n \in \mathbb{N}$, introduced in Section 2) are defined on a common probability space whose formal description is left to the reader. Their joint distribution will be denoted by P_f (under H_0) or $P_{f,g}$ (in case of a general alternative). To avoid undefined expressions when taking limits, the sample sizes n_1, n_2 are tacitly assumed to be large enough whenever necessary (a lower bound may depend on f, g and the norm $|\cdot|$). $I(A)$ denotes the indicator function of an event A. For short, the dependence of events and random variables on n_1, n_2 will frequently be suppressed.

2. The permutation distribution of $T_{n,k}$. Let $Z_1, \dots, Z_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with common pdf $f(x)$, which represent the pooled sample without knowing sample identity. Independently of $Z_1, \dots, Z_n$, the distribution of $(U_{n1}, \dots, U_{nn})$ having $\{1, 2\}$-valued components U_{nj} is given by

$$P(U_{ni} = u_i; 1 \le i \le n) = \binom{n}{n_1}^{-1}, \quad \text{if } \sum_{i=1}^{n} I(u_i = 1) = n_1,$$
$$= 0, \qquad \text{otherwise.}$$

Z_j is defined to have *sample type* "X" ("Y"), if $U_{nj} = 1$ ($U_{nj} = 2$), $1 \le j \le n$. For $1 \le i \ne j \le n$; $r = 1, \dots, k$ we introduce the events

$$\begin{aligned} A_{ij}^{(r)} &= \{Z_j = N_r(Z_i)\} \\ &= \{\text{"}Z_j \text{ is the } r\text{th nearest neighbor of } Z_i\text{"}\}, \\ B_{ij} &= \{U_{ni} = U_{nj}\} \\ &= \{\text{"}Z_i \text{ and } Z_j \text{ are of the same sample type"}\}, \end{aligned}$$

and put $A_{ii}^{(r)} = B_{ii} = \varnothing$, $1 \le i \le n$. Clearly, under H_0, $T_{n,k}$ defined in (1.4) has the same distribution as

$$\tilde{T}_{n,k} = \sum_{i,j=1}^{n} \sum_{r=1}^{k} I(A_{ij}^{(r)}) I(B_{ij}),$$

and the permutation distribution to be studied is the conditional distribution of $\tilde{T}_{n,k}$ given (the pooled sample) $Z_i = z_i$, $1 \le i \le n$. We may assume (this event occurs with probability 1) that $z_1, \dots, z_n$ are distinct points in $\mathbb{R}^d$ having uniquely defined neighbors. Conditionally on $Z_i = z_i$, $1 \le i \le n$, $I(A_{ij}^{(r)}) = a_{ij}^{(r)}$, where

$$\begin{aligned} a_{ij}^{(r)} &= a_{ij}^{(r)}(z_1,\dots,z_n) \\ &= I(\text{“}z_j \text{ is the } r\text{th nearest neighbor of } z_i\text{”}), \qquad i \neq j. \end{aligned}$$

Letting $a_{ij}^{+} = \Sigma_{r=1}^{k} a_{ij}^{(r)}$, we have

$$a_{ij}^{+} \in \{0,1\}, \qquad a_{ii}^{+} = 0, \qquad 1 \leq i, j \leq n, \tag{2.1}$$

$$\sum_{j=1}^{n} a_{ij}^{+} = k, \qquad 1 \leq i \leq n. \tag{2.2}$$

In terms of graph theory, (a_{ij}^{+}) represents the *adjacency matrix* of the *directed k-nearest neighbor graph* (k-NNG) of $z_1,\dots,z_n$ and completely determines the distribution of the random variable

$$L_{n,k} = \sum_{i,j=1}^{n} a_{ij}^{+} I(B_{ij}). \tag{2.3}$$

For convenience, let $d_j^{(k)} = \Sigma_{i=1}^{n} a_{ij}^{+}$, $1 \leq j \leq n$,

$$c_n^{(k)} = \frac{1}{nk}\sum_{j=1}^{n}\left(d_j^{(k)} - k\right)^2 \quad \text{and} \quad v_n^{(k)} = \frac{1}{nk}\sum_{i,j=1}^{n} a_{ij}^{+} a_{ji}^{+}.$$

$d_j^{(k)}$ is the *indegree* of the vertex z_j in the k-NNG of $z_1,\dots,z_n$. Using (2.2), we have $\Sigma_j d_j^{(k)} = nk$, and thus $c_n^{(k)}$ may be regarded as the *variance of indegrees* of the k-NNG. By definition of a_{ij}^{+}, it follows that $v_n^{(k)} = (1/k)\Sigma_{r,s=1}^{k} v_n^{(r,s)}$, where $v_n^{(r,s)} = (1/n)\Sigma_{i,j=1}^{n} a_{ij}^{(r)} a_{ji}^{(s)}$ is the proportion of all observations that are the sth nearest neighbor to their own rth nearest neighbor.

PROPOSITION 2.1. *Let G_n be a directed graph having vertices $1,\dots,n$ and adjacency matrix $(a_{ij}^{+})_{1 \leq i,j \leq n}$ satisfying* (2.1) *and* (2.2), *and let*

$$m(n_1,n_2) = (n_1(n_1-1) + n_2(n_2-1))/(n-1),$$
$$q(n_1,n_2) = 4(n_1-1)(n_2-1)/((n-2)(n-3)).$$

Then $E[L_{n,k}] = km(n_1,n_2)$,

$$\operatorname{Var}(L_{n,k}) = k\frac{n_1 n_2}{n-1}\left(q(n_1,n_2)\left(1 + v_n^{(k)} - \frac{2k}{n-1}\right) + (1 - q(n_1,n_2))c_n^{(k)}\right). \tag{2.4}$$

PROOF. Letting $m_{ij} = a_{ij}^{+} + a_{ji}^{+}$, $Y = \Sigma_{i,j} m_{ij}(1 - I(B_{ij}))$, we have $L_{n,k} = kn - \frac{1}{2}Y$. The statistic Y has been studied in a more general context by Bloemena (1964) [see his definition (1.1.5)] so that the assertion follows easily from formulas (3.5.6) and (3.5.7) of Bloemena (1964), observing that, in his notation, $m_{i+} = k + d_i^{(k)}$, $m_{++} = 2kn$, $\Sigma_i(m_{i+} - (1/n)m_{++})^2 = knc_n^{(k)}$ and $\Sigma_{i,j} m_{ij}^2 = 2kn(1 + v_n^{(k)})$. □

PROPOSITION 2.2. *Let (G_n) be a sequence of directed graphs as in Proposition* 2.1. *Assume that there is a positive constant* $\mathfrak{C}$, $1 \le \mathfrak{C} < \infty$, *depending only on k such that*

$$\sup_{1 \le j \le n} d_j^{(k)} \le \mathfrak{C}, \qquad n \in \mathbb{N}. \tag{2.5}$$

If $n \to \infty$ with

$$0 < a \le n_1/n_2 \le b < \infty, \tag{2.6}$$

for positive constants a, b, then

$$\operatorname{Var}(L_{n,k})^{-1/2}(L_{n,k} - km(n_1, n_2)) \to_{\mathscr{D}} \mathscr{N}(0,1).$$

PROOF. The assertion is an immediate consequence of Theorem 4.1.2 of Bloemena (1964). □

From Corollary S1 of Bickel and Breiman (1983) (which may easily be generalized to kth nearest neighbors), it follows that condition (2.5) is satisfied almost surely for the indegrees of the k-NNG of $Z_1, \dots, Z_n$. If $n \to \infty$ and (2.6) holds, we therefore obtain P_f almost surely

$$\begin{aligned} \lim P\Big(\operatorname{Var}\big(\tilde{T}_{n,k} | Z_1, \dots, Z_n\big)^{-1/2}\big(\tilde{T}_{n,k} - km(n_1, n_2)\big) \le t | Z_1, \dots, Z_n\Big) \\ = \Phi(t), \end{aligned} \tag{2.7}$$

$t \in \mathbb{R}$, where $\Phi(t)$ is the cdf of the standardized normal distribution.

3. The asymptotic null distribution of $T_{n,k}$. In this section we derive the limiting null distribution of $\tilde{T}_{n,k}$. The main result (Theorem 3.4) and the equality in distribution of $\tilde{T}_{n,k}$ and $T_{n,k}$ under H_0 imply that $T_{n,k}$ is asymptotically distribution free under H_0.

In view of (2.7) it remains to investigate $\operatorname{Var}(\tilde{T}_{n,k} | Z_1, \dots, Z_n)$ as $n \to \infty$. By Proposition 2.2, this in turn requires a study of the random variables

$$C_n^{(k)} = \frac{1}{nk} \sum_{j=1}^{n} \left(D_j^{(k)} - k\right)^2 \tag{3.1}$$

and

$$V_n^{(k)} = \frac{1}{nk} \sum_{i,j=1}^{n} A_{ij}^{+} A_{ji}^{+}, \tag{3.2}$$

where $A_{ij}^{+} = \sum_{r=1}^{k} I(A_{ij}^{(r)})$, $D_j^{(k)} = \sum_{i=1}^{n} A_{ij}^{+}$.

To state the limiting behavior of $C_n^{(k)}$ and $V_n^{(k)}$, let λ be shorthand for Lebesgue measure and write $S(x, \delta) = \{y \in \mathbb{R}^d \colon |x - y| < \delta\}$ for the open $|\cdot|$-sphere with radius δ centered at x. $\mathbf{0} = (0, \dots, 0)$ is the origin in $\mathbb{R}^d$, and μ denotes $d - 1$-dimensional Hausdorff measure (surface area) normalized such that $\mu(\{x \colon |x| = 1\}) = 1$. Finally, let $A^1 = A$, $A^0 = A^c$.

Proposition 3.1. *As $n \to \infty$, we have $C_n^{(k)} \to_{P_f} c_\infty^{(k)}$, where*

$$c_\infty^{(k)} = c_\infty^{(k)}(d, |\cdot|) = 1 - k + \frac{1}{k} \sum_{r,s=1}^{k} c_\infty(r,s),$$

$$c_\infty(r,s) = \sum_{i,j=0}^{1} \sum_{l=0}^{\bar{l}} \frac{1}{l!\delta!\varepsilon!} \iint_{\Gamma_{i,j}} \lambda(S_1 \cap S_2)^l \lambda(S_1 \setminus S_2)^\delta$$

$$\times \lambda(S_2 \setminus S_1)^\varepsilon \exp[-\lambda(S_1 \cup S_2)]\, du_1\, du_2,$$

$$\bar{l} = \min(r + i - 2, s + j - 2),$$

$$\delta = r - l + i - 2, \qquad \varepsilon = s - l + j - 2,$$

$$\Gamma_{i,j} = \left\{(u_1, u_2) \in [\mathbb{R}^d]^2 \colon \mathbf{0} \in S(u_1, |u_1 - u_2|)^i \cap S(u_2, |u_1 - u_2|)^j\right\},$$

$$S_m = S(u_m, |u_m|), \qquad m = 1, 2.$$

Proof. Straightforward algebra and symmetry give

$$E\left[C_n^{(k)}\right] = 1 - k + \frac{1}{k} \sum_{r,s=1}^{k} (n-1)(n-2) P\left(A_{21}^{(r)} \cap A_{31}^{(s)}\right).$$

Following the reasoning of Schilling (1986a), page 392, and Section 3 of Henze (1987), we get $\lim (n-1)(n-2)P(A_{21}^{(r)} \cap A_{31}^{(s)}) = c_\infty(r,s)$ and thus $\lim E[C_n^{(k)}] = c_\infty^{(k)}$ as $n \to \infty$. The proof of $\lim \operatorname{Var}(C_n^{(k)}) = 0$ as $n \to \infty$ was given in Section 3 of Henze (1987) for the case $k = 1$. The general case $k > 1$ is handled similarly. □

Proposition 3.2. *As $n \to \infty$, we have $V_n^{(k)} \to_{P_f} v_\infty^{(k)}$, where*

$$v_\infty^{(k)} = v_\infty^{(k)}(d, |\cdot|) = \frac{1}{k} \sum_{r,s=1}^{k} v_\infty(r,s),$$

$$v_\infty(r,s) = \int_{|u|=1} \sum_{j=0}^{\kappa} \mathfrak{b}(r-1, j, p(u))\, \mathfrak{w}(r, s-1-j, q(u))\, \mu(du),$$

$$\kappa = \min(r-1, s-1),$$

$$p(u) = \frac{\lambda[S(\mathbf{0},1) \cap S(u,1)]}{\lambda[S(\mathbf{0},1)]}, \qquad q(u) = (2 - p(u))^{-1},$$

$$\mathfrak{b}(m,j,p) = \binom{m}{j} p^j (1-p)^{m-j}, \qquad \mathfrak{w}(m,j,p) = \binom{m-1+j}{m-1} p^m (1-p)^j.$$

The proof of Proposition 3.2 is given in Henze (1987). For the case of the Euclidean norm, numerical values of $c_\infty^{(k)}$ and $v_\infty^{(k)}$ are furnished by Schilling (1986a). Since $C_n^{(k)}$ and $V_n^{(k)}$ deal with problems of a local character not depending on the "local intensity" of observations, it is not surprising that $c_\infty^{(k)}$ and $v_\infty^{(k)}$ do not depend on f.

PROPOSITION 3.3. *If $n \to \infty$ with $n_1/n \to \tau$, $0 < \tau < 1$, then*

$$\operatorname{Var}\left(n^{-1/2}\tilde{T}_{n,k} | Z_1, \dots, Z_n\right) \to_{P_f} \sigma_k^2(\tau, d, |\cdot|),$$

where

$$\sigma_k^2(\tau, d, |\cdot|) = 4k\tau(1-\tau)\left(\tau(1-\tau)\left(1 + v_\infty^{(k)}\right) + \left(\tau - \tfrac{1}{2}\right)^2 c_\infty^{(k)}\right).$$

PROOF. The result follows immediately from (2.4), Proposition 3.1, Proposition 3.2 and the fact that, as $n \to \infty$ with $n_1/n \to \tau$, $\lim(1 - q(n_1, n_2)) = 4(\tau - \frac{1}{2})^2$. □

Using (2.7), Proposition 3.3 and a routine technique, we get the following main result.

THEOREM 3.4. *If $n \to \infty$ with $n_1/n \to \tau$, then*

$$n^{-1/2}\left(T_{n,k} - km(n_1, n_2)\right) \to_{\mathscr{D}_f} \mathscr{N}\left(0, \sigma_k^2(\tau, d, |\cdot|)\right),$$

$$\lim \operatorname{Var}_f\left(n^{-1/2}T_{n,k}\right) = \sigma_k^2(\tau, d, |\cdot|).$$

4. Consistency. In this section we consider the general setup of the beginning of Section 1. The first result is a weak limit theorem for $T_{n,k}$.

THEOREM 4.1. *If $n \to \infty$, $n_1/n \to \tau$, $0 < \tau < 1$, we have*

$$\frac{1}{nk} T_{n,k} \to_{P_{f,g}} D(f, g, \tau),$$

where

$$D(f, g, \tau) = \int\left(\tau^2 f^2(x) + (1-\tau)^2 g^2(x)\right) / \left(\tau f(x) + (1-\tau) g(x)\right) dx.$$

REMARK. Here and in what follows, we put $0/0 = 0$.

PROOF. We show that

$$\lim\left[\frac{1}{nk} T_{n,k}\right] = D(f, g, \tau), \tag{4.1}$$

$$\lim \operatorname{Var}\left(\frac{1}{nk} T_{n,k}\right) = 0. \tag{4.2}$$

Only the case $k = 1$ will be considered; the situation for $k > 1$ follows similarly. By symmetry,

$$E\left[n^{-1}T_{n,1}\right] = n_1 n^{-1} E\left[I_1(1)\right] + n_2 n^{-1} E\left[I_{n_1+1}(1)\right],$$

with $I_i(1)$ defined in (1.3), and so (4.1) is a consequence of the following lemma.

Lemma 4.2. *Let x be a point of continuity of both f and g. If $f(x) > 0$, we have*

$$\lim E\left[I_1(1) | X_1 = x\right] = \tau f(x)/(\tau f(x) + (1-\tau)g(x)).$$

If $g(x) > 0$, we have

$$\lim E\left[I_{n_1+1}(1) | Y_1 = x\right] = (1-\tau)g(x)/(\tau f(x) + (1-\tau)g(x)).$$

Proof. By symmetry, it suffices to prove the first assertion. Assume first that $g(x) > 0$. Let $\omega_d = \lambda(S(\mathbf{0}, 1))$, $R_x = \min\{|x - X_i|\colon\ 2 \le i \le n_1\}$, $R_y = \min\{|x - Y_j|\colon\ 1 \le j \le n_2\}$, $V_x = n_1 f(x)\omega_d R_x^d$, $V_y = n_2 g(x)\omega_d R_y^d$ and put $\rho = [\xi/(n_1\omega_d f(x))]^{1/d}$, where $\xi > 0$ is any fixed real number. From

$$P(V_x > \xi) = \left(1 - \int_{S(x,\rho)} f(y)\,dy\right)^{n_1 - 1}$$

and the continuity of f at x, which entails

$$\int_{S(x,\rho)} f(y)\,dy = \frac{\xi}{n_1} + o(n_1), \qquad n_1 \to \infty,$$

it follows that $\lim P(V_x > \xi) = \exp(-\xi)$. In the same way, $\lim P(V_y > \eta) = \exp(-\eta)$, $\eta > 0$. The joint independence of X_i, Y_j implies that, as $n_1, n_2 \to \infty$, V_x/V_y converges in distribution to a quotient Q, say, of independent unit exponential random variables yielding

$$\begin{aligned}\lim E\left[I_1(1)|X_1 = x\right] &= \lim P(R_x < R_y)\\ &= \lim P\big(V_x/V_y < n_1 f(x)/(n_2 g(x))\big)\\ &= P(Q < \tau f(x)/((1-\tau)g(x)))\\ &= \tau f(x)/(\tau f(x) + (1-\tau)g(x)),\end{aligned}$$

as asserted.

The case $g(x) = 0$ will be reduced to the preceding considerations. To this end, fix $\varepsilon > 0$ and take $\delta > 0$ such that

$$\lambda(S(x,\delta)) \le 1 \quad \text{and} \quad g(y) \le \varepsilon/2, \quad \text{whenever } |x - y| < \delta.$$

Independently of X_i, Y_j, let $J_1, \dots, J_{n_2}, W_1, \dots, W_{n_2}$ be independent random variables, J_ν being $\{0,1\}$-valued with

$$P(J_\nu = 1) = \big(\varepsilon\lambda(S(x,\delta)) - g_\delta\big)/(1 - g_\delta), \qquad 1 \le \nu \le n_2,$$

and W_ν having density

$$w(z) = (\varepsilon - g(z))/(\varepsilon\lambda(S(x,\delta)) - g_\delta)I(z \in S(x,\delta)), \qquad 1 \le \nu \le n_2,$$

where, for brevity, $g_\delta = \int_{S(x,\delta)} f(y)\,dy$. Putting

$$\begin{aligned} Y_\nu^* &= Y_\nu, \quad \text{if } |Y_\nu - x| < \delta \text{ or } |Y_\nu - x| \ge \delta \text{ and } J_\nu = 0,\\ &= W_\nu, \quad \text{if } |Y_\nu - x| \ge \delta \text{ and } J_\nu = 1,\end{aligned}$$

the density of Y_ν^* is given by

$$\begin{aligned} g^*(y) &= \varepsilon, & &\text{if } |y-x| < \delta, \\ &= g(y)(1 - \varepsilon\lambda(S(x,\delta)))/(1-g_\delta), & &\text{if } |y-x| \geq \delta, \end{aligned}$$

$1 \leq \nu \leq n_2$. $X_1, \dots, X_{n_1}, Y_1^*, \dots, Y_{n_2}^*$ are independent, and we have

$$P(R_x < R_y^*) \leq P(R_x < R_y), \tag{4.3}$$

where $R_y^* = \min\{|x - Y_\nu^*|: 1 \leq \nu \leq n_2\}$. From the results obtained for the case $g(x) > 0$ applied to X_i, Y_j^*, we get

$$\begin{aligned} \lim P(R_x < R_y^*) &= P(Q < \tau f(x)/((1-\tau)\varepsilon)) \\ &= \tau f(x)/(\tau f(x) + (1-\tau)\varepsilon), \end{aligned}$$

and thus Lemma 4.2 follows using (4.3) and letting ε approach 0. □

To complete the proof of Theorem 4.1, let x_1, x_2 be distinct points of continuity of both f and g with $f(x_j) > 0$, $1 \leq j \leq 2$. By symmetry and dominated convergence, to show (4.2) it suffices to demonstrate that

$$\lim E\left[I_1(1)I_2(1)|X_1 = x_1, X_2 = x_2\right] = \prod_{j=1}^{2}\left[\tau f(x_j)/(\tau f(x_j) + (1-\tau)g(x_j))\right].$$

This was proved in Henze (1984), page 270, for the case $g(x_j) > 0$, $1 \leq j \leq 2$. The modifications for the case $\min_{1 \leq j \leq 2} g(x_j) = 0$ follow the lines given previously. The details are omitted. □

The quantity $D(f, g, \tau)$ figuring in the statement of Theorem 4.1 is a member of a general class of separation measures of several probability distributions introduced and studied by Györfi and Nemetz (1975, 1977, 1978). From Theorem 1 and Corollary 1 of Györfi and Nemetz (1975), we have the following result.

PROPOSITION 4.3. *Let f_j be a pdf on $\mathbb{R}^d$, and let $\tau_j > 0$, $1 \leq j \leq s$, $\sum_{j=1}^{s}\tau_j = 1$, $s \geq 2$. Then*

$$\int \sum_{j=1}^{s} \tau_j^2 f_j^2(x) \Big/ \sum_{j=1}^{s} \tau_j f_j(x)\, dx \geq \sum_{j=1}^{s} \tau_j^2.$$

Equality holds if, and only if, the probability measures corresponding to $f_1, \dots, f_s$ coincide.

We now turn to the proof of consistency of a multivariate two-sample test based on $T_{n,k}$, carried out as an exact permutation test.

To this end, let $z_j = x_j$, $1 \leq j \leq n_1$, and $z_{n_1+l} = y_l$, $1 \leq l \leq n_2$, denote the observed values of $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$, and put $\mathbf{z}_n = (z_1, \dots, z_n)$. Given any level of significance α, $0 < \alpha < 1$, the critical value $c_{n,k}(\mathbf{z}_n; \alpha)$ and the probability of randomization $\gamma_{n,k}(\mathbf{z}_n; \alpha)$ for performing the test procedure are determined by

$$(4.4)\qquad 0 \le \gamma_{n,k}(z_n;\alpha) < 1,$$

$$(4.5)\quad P\big(L_n(z_n) > c_{n,k}(z_n;\alpha)\big) + \gamma_{n,k}(z_n;\alpha)P\big(L_n(z_n) = c_{n,k}(z_n;\alpha)\big) = \alpha,$$

where $L_n(z_n) = L_{n,k}$ is defined in (2.3).

THEOREM 4.4. *The test* φ_{n_1,n_2}, *defined by*

$$\begin{aligned}
&\varphi_{n_1,n_2}(x_1,\dots,x_{n_1},y_1,\dots,y_{n_2})\\
&\quad = 1, && \text{if } T_{n,k}(x_1,\dots,x_{n_1},y_1,\dots,y_{n_2}) > c_{n,k}(z_n;\alpha),\\
&\quad = \gamma_{n,k}(z_n;\alpha), && \text{if } T_{n,k}(x_1,\dots,x_{n_1},y_1,\dots,y_{n_2}) = c_{n,k}(z_n;\alpha),\\
&\quad = 0, && \text{if } T_{n,k}(x_1,\dots,x_{n_1},y_1,\dots,y_{n_2}) < c_{n,k}(z_n;\alpha),
\end{aligned}$$

and (4.4) *and* (4.5), *is consistent at level* α *for testing* H_0: $f = g$ *a.e.*; *i.e., if*

$$(4.6)\qquad f(x) \ne g(x), \quad \textit{on a set of positive measure},$$

we have, as $n \to \infty$, $n_1/n \to \tau$, $0 < \tau < 1$,

$$\lim E_{f,g}\big[\varphi_{n_1,n_2}(X_1,\dots,X_{n_1},Y_1,\dots,Y_{n_2})\big] = 1.$$

PROOF. Assume that (4.6) holds, and let Z_j be as in (1.2), $\mathscr{Z}_n = (Z_1,\dots,Z_n)$. Note that in contrast to Section 2, $Z_1,\dots,Z_n$ are independent but no longer identically distributed. From Corollary S1 of Bickel and Breiman (1983), generalized to k-nearest neighbors, it follows that condition (2.5) is satisfied $P_{f,g}$ a.s. for the sequence of k-NNGs with vertices $Z_1,\dots,Z_n$, and thus Propositions 2.1 and 2.2 yield

$$(4.7)\qquad U_{n_1,n_2}^{-1/2}\big(c_{n,k}(\mathscr{Z}_n;\alpha) - km(n_1,n_2)\big) \to \Phi^{-1}(1-\alpha), \quad P_{f,g}\text{ a.s.},$$

where

$$U_{n_1,n_2} = k\frac{n_1n_2}{n-1}\left(q(n_1,n_2)\left(1 + V_{n_1,n_2}^{(k)} - \frac{2k}{n-1}\right) + (1 - q(n_1,n_2))C_{n_1,n_2}^{(k)}\right)$$

and where $C_{n_1,n_2}^{(k)} = C_n^{(k)}$, $V_{n_1,n_2}^{(k)} = V_n^{(k)}$ are defined in (3.1) and (3.2), respectively (the notational change indicates that $\mathscr{Z}_n$ consists of two different samples). The inequalities $0 \le V_{n_1,n_2}^{(k)} \le k$, $P_{f,g}$ a.s., $0 \le C_{n_1,n_2}^{(k)} \le (\mathfrak{C} + k)^2$, $P_{f,g}$ a.s. imply that

$$U_{n_1,n_2} \le k\frac{n_1n_2}{n-1}\big((1+k)|q(n_1,n_2)| + |1 - q(n_1,n_2)|(\mathfrak{C}+k)^2\big), \quad P_{f,g}\text{ a.s.}$$

and on combining this with (4.7), we have

$$(4.8)\qquad \frac{1}{nk}c_{n,k}(\mathscr{Z}_n;\alpha) = n^{-1}m(n_1,n_2) + n^{-1/2}O_{P_{f,g}}(1),$$

where $O_{P_{f,g}}(1)$ denotes a random variable that is bounded in probability when taking limits. The assertion now follows from

$$E_{f,g}\left[\varphi_{n_1,n_2}(X_1,\dots,X_{n_1},Y_1,\dots,Y_{n_2})\right]$$
$$\geq P_{f,g}\left(\frac{1}{nk}T_{n,k}(X_1,\dots,X_{n_1},Y_1,\dots,Y_{n_2}) > \frac{1}{nk}c_{n,k}(\mathscr{Z}_n;\alpha)\right),$$

Theorem 4.1, (4.8), Proposition 4.3 and the fact that, as $n \to \infty$, $n_1/n \to \tau$,

$$\lim\left(n^{-1}m(n_1,n_2)\right) = \tau^2 + (1-\tau)^2. \qquad \square$$

5. Concluding remarks.

REMARK 5.1. For moderate or large sample sizes n_1, n_2, we may reject H_0 at (approximate) level α if

$$T_{n,k}(z_n) \geq c^*_{n,k}(z_n;\alpha),$$

where $z_n = (z_1,\dots,z_n) = (x_1,\dots,x_{n_1},y_1,\dots,y_{n_2})$,

$$c^*_{n,k}(z_n;\alpha) = km(n_1,n_2) + u_{n_1,n_2}(z_n)^{1/2}\phi^{-1}(1-\alpha),$$

with $u_{n_1,n_2}(z_n)$ given by the right-hand side of (2.4). The practical implementation of this *approximate permutation test* requires the determination of all k nearest neighbors [for efficient algorithms, cf. Friedman, Baskett and Shustek (1975) and Rohlf (1982)].

REMARK 5.2. The performance of the test based on $T_{n,k}$ for finite sample sizes (Euclidean metric) was assessed in Schilling (1986b) by means of Monte Carlo experiments for various values of k and d.

Acknowledgments. This work is based on a part of the author's Habilitationsschrift [Henze (1985)] written at the University of Hannover. The author wishes to thank Prof. Dr. D. Morgenstern for stimulating discussions, and Prof. Dr. W. R. van Zwet for drawing his attention to the work of Bloemena.

REFERENCES

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

BICKEL, P. J. (1969). A distribution free version of the Smirnov two-sample test in the multivariate case. *Ann. Math. Statist.* **40** 1–23.

BICKEL, P. J. and BREIMAN, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.* **11** 185–214.

BLOEMENA, A. R. (1964). *Sampling from a Graph. Mathematical Centre Tracts* **2.** Mathematisch Centrum, Amsterdam.

FRIEDMAN, J. H., BASKETT, F. and SHUSTEK, L. J. (1975). An algorithm for finding nearest neighbors. *IEEE Trans. Comput.* **24** 1000–1006.

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

GYÖRFI, L. and NEMETZ, T. (1975). *f*-dissimilarity: A general class of separation measures of several probability measures. *Topics in Information Theory. Colloq. Math. Soc. János Bolyai* **16** 309–321. Keszthely, Hungary.

GYÖRFI, L. and NEMETZ, T. (1977). On the dissimilarity of probability measures. *Problems Control Inform. Theory* **6** 263–267.

GYÖRFI, L. and NEMETZ, T. (1978). *f*-dissimilarity. A generalization of affinity of several distributions. *Ann. Inst. Statist. Math.* **30** 105–113.

HENZE, N. (1984). On the number of random points with nearest neighbor of the same type and a multivariate two-sample test. *Metrika* **31** 259–273. (In German.)

HENZE, N. (1985). A multivariate two- and multisample test based on the number of nearest neighbour type coincidences. Habilitationsschrift, Univ. Hannover. (In German.)

HENZE, N. (1987). On the fraction of random points with specified nearest neighbour interrelations and "degree of attraction." *Adv. in Appl. Probab.* **19** 873–895.

LEHMANN, E. L. (1951). Consistency and unbiasedness of certain non-parametric tests. *Ann. Math. Statist.* **22** 165–179.

ROHLF, F. J. (1982). Single link clustering algorithms. In *Handbook of Statistics* (P. R. Krishnaiah and L. N. Kanal, eds.) **2** 267–284. North-Holland, Amsterdam.

ROSENBLATT, M. (1952). Limit theorems associated with variants of the von-Mises statistic. *Ann. Math. Statist.* **23** 617–623.

SCHILLING, M. F. (1986a). Mutual and shared neighbor probabilities: Finite- and infinite-dimensional results. *Adv. in Appl. Probab.* **18** 388–405.

SCHILLING, M. F. (1986b). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806.

SMIRNOV, N. V. (1939). On the estimation of the discrepancy between empirical curves of distributions for two independent samples. *Bull. Moscow Univ.* **2** 3–6.

WALD, A. and WOLFOWITZ, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11** 147–162.

WEISS, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Statist.* **31** 159–164.

WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
UNIVERSITÄT HANNOVER
WELFENGARTEN 1
D-3000 HANNOVER 1
WEST GERMANY