Multivariate Two-Sample Tests Based on Nearest Neighbors

MARK F. SCHILLING*

A new class of simple tests is proposed for the general multivariate two-sample problem based on the (possibly weighted) proportion of all k nearest neighbor comparisons in which observations and their neighbors belong to the same sample. Large values of the test statistics give evidence against the hypothesis H of equality of the two underlying distributions. Asymptotic null distributions are explicitly determined and shown to involve certain nearest neighbor interaction probabilities. Simple infinite-dimensional approximations are supplied. The unweighted version yields a distribution-free test that is consistent against all alternatives; optimally weighted statistics are also obtained and asymptotic efficiencies are calculated. Each of the tests considered is easily adapted to a permutation procedure that conditions on the pooled sample. Power performance for finite sample sizes is assessed in simulations.

KEY WORDS: Distribution-free; Kth nearest neighbor; Infinite-dimensional approximation.

1. INTRODUCTION

A substantial number of nonparametric methods based on nearest neighbors have been developed in recent years for various multivariate situations. The popularity of these procedures has increased because of new theoretical developments, the expanding capabilities of modern high-speed computers, and efficient algorithms for nearest neighbor calculations, which mitigate to a great extent the computational obstacles involved. Classification, density estimation, and regression are areas that have received particular attention; more recently, distribution-free tests for multivariate goodness of fit based on nearest neighbors have been developed (see Bickel and Breiman 1983; Schilling 1983a,b) along with procedures for assessing multivariate association (Friedman and Rafsky 1983).

This article presents a new class of distribution-free tests for the general multivariate two-sample problem along with a related procedure for testing specific hypotheses. The tests are natural, require only mild assumptions, and are easy to implement. The basic setup is given in Section 2, followed by a brief history of related work. In Section 3 the elemental version of the test is introduced. The test statistic is simply the proportion of all k nearest neighbor comparisons in which a point and its neighbor belong to the same sample. The asymptotic null distribution is established and found to exhibit marked stability across both dimension and the number of neighbors used. Consistency against all alternatives is proven. Weighted versions are considered in Section 4. Optimal weights are found, and the question of asymptotic power is studied. The nondistribution-free weighted test appears nearly as efficient, when weighted properly for specific distributional models, as the (optimal) likelihood ratio test. Conditional tests are discussed in Section 5. Monte Carlo experiments support the various analytical results (Section 6).

2. PROBLEM AND HISTORY

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent random samples in $\mathbf{R}^d$ from unknown distributions $F(x)$ and $G(x)$, respectively, with corresponding densities $f(x)$ and $g(x)$ assumed to be continuous on their supports. The two-sample problem treated here is to test the hypothesis $H : F(x) = G(x)$ against the completely general alternative $K : F(x) \neq G(x)$. No knowledge of F or G is proclaimed by H—only their equivalence.

Take $n = n_1 + n_2$, $\Omega_1 = \{1, \ldots, n_1\}$, $\Omega_2 = \{n_1 + 1, \ldots, n\}$, and label the pooled sample as $Z_1, \ldots, Z_n$, where

$$Z_i = X_i, \qquad i \in \Omega_1$$
$$= Y_{i-n_1}, \qquad i \in \Omega_2.$$

Let $\|\cdot\|$ be a norm, and define the kth nearest neighbor to Z_i as that point Z_j satisfying $\|Z_{j'} - Z_i\| < \|Z_j - Z_i\|$ for exactly $k - 1$ values of j' $(1 \leq j' \leq n, j' \neq i, j)$. Ties are neglected, since they occur with probability zero. When ties occur in practice, however, because of rounding, limited resolution in measurement scales, and so forth, they can be easily handled by ranking neighbors in the following manner: Whenever exactly Q of the observations are equidistant from Z_i, with $k - 1$ other points strictly closer to Z_i, assign a random permutation of the appropriate ranks $k, k + 1, \ldots, k + Q - 1$ to these Q points in forming the nearest neighbor list for Z_i. This procedure has no effect on the validity of the results below.

Let $I[\cdot]$ represent the indicator function. Friedman and Steppel (1974) proposed two-sample tests for this framework based on the number of points C_i among the k nearest neighbors of each point Z_i that belong to (say) the first sample $\{X_1, \ldots, X_{n_1}\}$. Separate frequency distributions can be compiled from the counts for $i \in \Omega_1$ and $i \in \Omega_2$. When $F = G$ these counts $\{C_i; i = 1, \ldots, n\}$ are dependent hypergeometric variables with parameters $n - 1$, $n_1 - I[i \in \Omega_1]$, and k, and the frequency distributions have virtually identical expectations. For large n, Friedman and Steppel suggested contrasting $\{C_i, i \in \Omega_1\}$ with $\{C_i, i \in \Omega_2\}$ either by means of a t statistic or by comparing the

* Mark F. Schilling is Assistant Professor, Department of Mathematics, California State University, Northridge, CA 91330. Research for this article was supported in part by National Science Foundation Grants MCS 79-19141 and MCS 80-17103. The author wishes to thank the associate editor and referees for many detailed and helpful comments.

Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Theory and Methods

frequency distribution of all of the counts $C_1, \ldots, C_n$ (here each point is regarded as one of its own nearest neighbors) with the binomial distribution having parameters k and n_1/n. The latter idea can be implemented by a goodness-of-fit test of the χ^2 type. Since the C_i's are dependent because of the overlap of k nearest neighborhoods, the significance level α of such a test cannot be determined by ordinary binomial theory. Friedman and Steppel used a permutation procedure to estimate α.

Further results were obtained by Rogers (1976) under a different formulation. Let $S_{\alpha,j}$ represent the number of points Z_i ($i \in \Omega_\alpha$) for which exactly j of the k nearest neighbors have a common sample identity to Z_i for $\alpha = 1, 2$; $j = 1, \ldots, k$. These quantities are directly obtainable from the aforementioned frequency distributions. Rogers showed that the vector of $S_{\alpha,j}$ values, appropriately centered and scaled, is asymptotically multivariate normal under H with limiting covariance structure independent of $F = G$. Unfortunately, the asymptotic covariance matrix is analytically intractable and must be estimated via Monte Carlo methods. Rogers discussed tests based on linear combinations of the $S_{\alpha,j}$'s.

Friedman and Rafsky (1979) introduced procedures for the nonparametric two-sample problem that are based on the minimal spanning tree (MST) of the pooled sample—the graph of minimal length that provides a path between any two sample points. The graph connecting each point to its nearest neighbor is a subgraph of the MST. Friedman and Rafsky's tests are multivariate analogs of the Wald–Wolfowitz and Smirnov univariate two-sample tests. Conditional results are derived and power performance is estimated through Monte Carlo experiments.

3. THE UNWEIGHTED TEST

The tests that follow are modeled after those proposed by Friedman and Steppel (1974) and Rogers (1976). They are not only extremely simple conceptually but possess analytically tractable null distributions as well, which, it will be shown, are quite stable with regard to both the dimension of the observation space and number of neighbors used. Take $\|\cdot\|$ to be the Euclidean norm, and let $NN_i(r)$ represent the rth nearest neighbor to the sample point Z_i. Define

$$I_i(r) = 1 \quad \text{if } NN_i(r) \text{ belongs to the same sample as } Z_i$$
$$= 0 \quad \text{otherwise.}$$

The statistic considered initially for testing H is the quantity

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

which is simply the proportion of all k nearest neighbor comparisons in which a point and its neighbor are members of the same sample. One would expect $T_{k,n}$ to achieve a larger value under K than under H because of a lack of complete mixing of the two samples when the parent distributions are not identical; hence large values of $T_{k,n}$ are significant.

3.1 Asymptotic Null Distribution

Assume that n_1, n_2 tend to infinity in such a way that $\lambda_i = \lim_{n\to\infty} n_i/n$ exists for $i = 1, 2$. Consider the following events:

(i) $NN_1(r) = Z_2$, $NN_2(s) = Z_1$.
(ii) $NN_1(r) = NN_2(s)$.

We shall say that Z_1 and Z_2 are mutual neighbors if case (i) occurs for some r and s and that they share a neighbor if case (ii) occurs. Write $p_i(r, s)$, $i = 1, 2$, respectively, for the null probabilities of the preceding events.

The values of $p_1(r, s)$ and $p_2(r, s)$ in finite samples depend on the underlying density and are most difficult to compute. It is intuitively clear, however, that each is $O(n^{-1})$. It turns out that both $np_1(r, s)$ and $np_2(r, s)$ approach limits that are independent of $f = g$. Denote these limits by $p_i'(r, s)$, $i = 1, 2$, and write

$$\bar{p}_i' = k^{-2} \sum_{r=1}^{k} \sum_{s=1}^{k} p_i'(r, s), \qquad i = 1, 2.$$

The main result is that the asymptotic distribution of $T_{k,n}$ depends only on k, λ_1, λ_2, and $\bar{p}_1'$, $\bar{p}_2'$.

Theorem 3.1. If $n_1, n_2 \to \infty$ with n_i/n tending to λ_i for $i = 1, 2$, then $(nk)^{1/2}(T_{k,n} - \mu_k)/\sigma_k$ has a limiting standard normal distribution under H, where

$$\mu_k = \lim_{n\to\infty} E_H(T_{k,n}) = \lambda_1^2 + \lambda_2^2 \tag{3.1}$$

and

$$\sigma_k^2 = \lim_{n\to\infty} nk \operatorname{var}_H(T_{k,n})$$
$$= \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 k\bar{p}_1' - \lambda_1\lambda_2(\lambda_1 - \lambda_2)^2 k(1 - \bar{p}_2'). \tag{3.2}$$

The proof of Theorem 3.1 is given in the Appendix.

Expressions for the quantities $p_1'(r, s)$ for $p_2'(r, s)$ for general r and s are furnished in Schilling (1986) and are rather complex (though computable), particularly for the neighbor-sharing values $p_2'(r, s)$. The quantities $k\bar{p}_1'$ and $k(1 - \bar{p}_2')$ that appear in σ_k^2, however, are extremely stable in both k and d, which suggests the possibility of replacing them with limiting values. Support for this claim is provided by Schilling (1986, theorems 4.2 and 4.3), and numerical results for small k and d are also given there. The theorems are reproduced here for easy reference.

Theorem 3.2. For all d, $\lim_{k\to\infty} k\bar{p}_1'$ exists and equals 1.

Theorem 3.3. For all positive integers r and s, $\lim_{d\to\infty} p_2'(r, s)$ exists and equals 1.

The convergence indicated in Theorem 3.3 is very rapid.

The asymptotic variance takes particularly simple forms in the two cases given next. When $\lambda_1 = \lambda_2 = .5$, as is common in practice, the neighbor-sharing values $p_2(r, s)$ are not involved and (3.2) yields $\sigma_k^2 = (1 + k\bar{p}_1')/4$. For general λ_1 and λ_2 and $d = \infty$, p_2' is again absent (Theorem 3.3) and $k\bar{p}_1'$ takes a simple combinatorial form (Schilling

Table 1. σ_k^2 for $\lambda_1 = \lambda_2 = .5$ and $\lambda_1 = .25$, $\lambda_2 = .75$, for Selected k and d

	$\lambda_1 = \lambda_2 = .5$				$\lambda_1 = .25$, $\lambda_2 = .75$			
d	$k = 1$	$k = 2$	$k = 3$	$k = \infty$	$k = 1$	$k = 2$	$k = 3$	$k = \infty$
1	.417	.435	.445		.258	.268	.274	
2	.405	.428	.439		.257	.268	.275	
3	.398	.423	.435		.257	.269	.275	
4	.393	.419	.432		.257	.270	.276	
5	.389	.416	.430		.257	.271	.278	
10	.380	.410	.425		.257	.273	.282	
∞	.375	.406	.422	.500	.258	.275	.284	.328

1986); this yields

$$\sigma_k^2 = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2\left[1 - \binom{2k}{k}2^{-2k}\right] \doteq \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 \tag{3.3}$$

for k not too small. Note that the quantity $\lambda_1\lambda_2$ arises from the binomial character of $T_{k,n}$; the additional term is the result of the dependence between the nearest neighborhoods of the sample points and roughly doubles the asymptotic variance when the sample sizes are not too disparate.

The marked stability in both d and k of the asymptotic variance is demonstrated in Table 1 for $\lambda_1 = \lambda_2 = .5$ and $\lambda_1 = .25$, $\lambda_2 = .75$. It is evident that the infinite-dimensional variances in (3.3) are quite adequate replacements for the more complex finite-dimensional variances for most d.

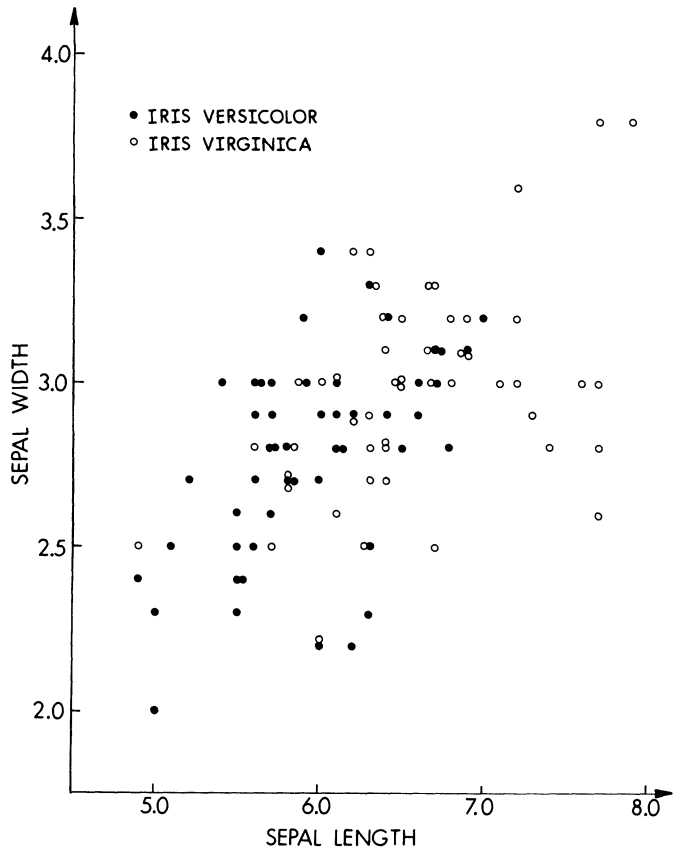

Figure 1. Sepal Measurements for Iris Versicolor and Iris Virginica ($n_1 = n_2 = 50$).

It is interesting to note the similarity of this distributional phenomenon (stability across dimension, simple infinite-dimensional limit) to that found for nearest neighbor goodness-of-fit tests (Schilling 1983b).

Limited simulation studies indicate that the asymptotic distribution in Theorem 3.1 serves well for small d (≤ 3) with moderate n (≥ 50–100) and adequately for higher values of d (≤ 10) using larger n (≥ 200). Difficulties arise when the dimension grows, because of fringe effects and the increasing emptiness of high-dimensional space; this leads to $\bar{p}_1'$ overestimating and $\bar{p}_2'$ underestimating the actual mutual neighbor and neighbor-sharing frequencies, respectively. Particular caution is in order for nonsmooth densities such as the d-dimensional uniform.

As an example of the utility of the preceding procedure for real data, $T_{k,n}$ tests with $k = 3$ and $k = 10$ were applied to a subset of the well-known iris data (Fisher 1936). The two similar species, *Versicolor* and *Virginica*, were compared using only the two variables (sepal length and sepal width) on which they are most alike (see Figure 1). The proportion of k nearest neighbors belonging to the same sample as their reference point was found to be $T_{k,n} = .600$ for $k = 3$ ($z = 2.746$) and $T_{k,n} = .615$ for $k = 10$ ($z = 5.541$), highly significant values in both cases. The infinite-dimensional approximation to the asymptotic distribution of $T_{k,n}$ based on (3.3) yields $z = 2.801$ and $z = 5.425$ for $k = 3$ and $k = 10$, respectively.

3.2 Consistency and Asymptotic Power

Theorem 3.4. The test based on $T_{k,n}$ is consistent against any alternative K.

To prove Theorem 3.4 it must be shown that $\lim\inf_{n\to\infty} E_K(T_{k,n}) > \lim_{n\to\infty} E_H(T_{k,n})$. Only the case $k = 1$ will be described; the situation for $k > 1$ follows similarly. We have

$$E_K(T_{k,n}) = (n_1/n)P_K(I_1(1) = 1) + (n_2/n)P_K(I_{n_1+1}(1) = 1). \tag{3.4}$$

It must be shown that $\lim\inf_{n\to\infty} P_K(I_1(1) = 1) > \lambda_1$ and $\lim\inf_{n\to\infty} P_K(I_{n_1+1}(1) = 1) > \lambda_2$. Now

$$P(I_1(1) = 1) = (n_1 - 1)P(NN_1(1) = Z_2)$$
$$= (n_1 - 1)\int_{\mathbf{R}^d} f(x_1)\int_{\mathbf{R}^d} f(x_2)\left\{1 - \int_S f(x)dx\right\}^{n_1-2} \times \left\{1 - \int_S g(x)dx\right\}^{n_2} dx_2dx_1,$$

where S is the sphere centered at x_1 having radius $\|x_2 - x_1\|$. Using first-order approximations to $\int_S f(x)dx$ and $\int_S g(x)dx$ and putting $\omega = n^{1/d}(x_2 - x_1)$ for the integral in x_2 produces

$$\lim_{n\to\infty} P(I_1(1) = 1)$$
$$= \int_{\mathbf{R}^d} f^2(x_1)\int_{\mathbf{R}^d} \exp\left\{-\left[f(x_1) + \frac{\lambda_2}{\lambda_1}g(x_1)\right]K_d\|\omega\|^d\right\}d\omega dx_1,$$

where K_d is the volume of a d-dimensional sphere of radius

1. Setting $\rho = \|\omega\|^d$ then yields

$$\lim_{n\to\infty} P(I_1(1) = 1) = \lambda_1 E_f[1/(\lambda_1 + \lambda_2 g(Z)/f(Z))]. \quad (3.5)$$

Similarly, one can obtain

$$\lim_{n\to\infty} P(I_{n_1+1}(1) = 1) = \lambda_2 E_g[1/(\lambda_2 + \lambda_1 f(Z)/g(Z))]. \quad (3.6)$$

The result follows from considering the random variable $g(Z)/f(Z)$ and applying Jensen's inequality.

After using (3.4)–(3.6), the asymptotic mean under K is found to be

$$\lim_{n\to\infty} E_K T_{k,n} = 1 - 2\lambda_1\lambda_2 \int \frac{f(x)g(x)dx}{\lambda_1 f(x) + \lambda_2 g(x)}.$$

This expression remains unchanged for general k.

We can get an indication of the large sample power performance of $T_{k,n}$ by means of its efficacy coefficient

$$\xi = (\lim_{n\to\infty} E_K T_{k,n} - \mu_k)/(\lim_{n\to\infty} n \operatorname{var}_H(T_{k,n})).$$

For the simplest case, $\lambda_1 = \lambda_2 = .5$,

$$\xi = (1/k + \bar{p}_1')^{-1/2} \int \frac{f^2(x) + g^2(x)}{f(x) + g(x)}\, dx. \quad (3.7)$$

Note from Theorem 3.2 that $\xi \sim O(\sqrt{k})$.

4. WEIGHTED VERSIONS

The statistic $T_{k,n}$ can be generalized in various ways in the hope of improving performance by weighting the contribution of each point by its value or by the ranks or values of those neighbors having the same sample identity. The search for asymptotically optimal weights will involve the following model: Let the null hypothesis H be that the common density of the observations is f_o, and consider a directional sequence of alternatives $\{K_n\}$ to H in which the densities $f = f_n$ and $g = g_n$ belong to a regular parametric family $\{q(\cdot, \theta), \theta \in \mathbf{R}\}$ with $f(x) = q(x, \theta_o + \Delta)$ and $g(x) = q(x, \theta_o - \Delta)$, where $\Delta = \Delta_n$ approaches 0 as $n \to \infty$ and $f_o(x) = q(x, \theta_o)$. Assume further that the first two derivatives of q with respect to θ (written as q', q'') exist at θ_0 and that derivatives can be passed under the integral sign when necessary.

The goal is to find weights that maximize power for the particular sequence of alternatives specified but still maintain adequate performance against other possible deviations from H.

4.1 Weighting by Rank of Neighbor

One possibility is to weight the contribution of each neighbor according to its rank in distance among neighbors and the sample membership of the reference point. To this end let $\mathbf{w}_1 = (w_1(1), \ldots, w_1(k))^T$ and $\mathbf{w}_2 = (w_2(1), \ldots, w_2(k))^T$ be vectors of weighting constants, write $\mathbf{w} = (\mathbf{w}_1^T, \mathbf{w}_2^T)^T$, and define

$$U_{k,n,\mathbf{w}} = \frac{1}{nk} \sum_{\alpha=1}^{2} \sum_{i\in\Omega_\alpha} \sum_{r=1}^{k} w_\alpha(r) I_i(r).$$

When $\mathbf{w}$ is a vector of ones, $U_{k,n,\mathbf{w}}$ reduces to the unweighted proportion statistic $T_{k,n}$. The asymptotic distribution of $U_{k,n,\mathbf{w}}$ can be obtained by a simple extension of the methods used for $T_{k,n}$.

A direct extension of calculations in Section 3.2 reveals that the optimal system of weights $\mathbf{w}^*$ does not depend on K and thus cannot be tuned for particular alternatives to H. Furthermore, regardless of the choice of weights, the asymptotic power of the $U_{k,n,\mathbf{w}}$ test is equal to the level of the test for alternative sequences $[K_n\}$ in which $\theta - \theta_o \sim o(n^{-1/4})$, which includes the $O(n^{-1/2})$ contiguous alternative models ordinarily used in calculating Pitman efficiencies. This observation parallels results for goodness-of-fit tests based on nearest neighbors (Bickel and Breiman 1983; Schilling 1983a). In view of the results of Le Cam (1973) and Birgé (1983) concerning optimal rates of convergence, one cannot hope to find a procedure achieving simultaneous power for densities converging at rate $O(n^{-1/2})$ for models as general as those used here; the metric entropy of the space of all continuous densities is too large.

The fourth power of the ratio of the efficacies of $U_{k,n,\mathbf{w}^*}$ to $T_{k,n}$, a natural analog to Pitman efficiency that measures the asymptotic ratio of sample sizes necessary for $T_{k,n}$ and $U_{k,n,\mathbf{w}^*}$ to achieve the same limiting power for sequences of density pairs converging at rate $n^{-1/4}$, turns out to be $\{\mathbf{e}^T(P_1' + I)^{-1}\mathbf{e}\ \mathbf{e}^T(P_1' + I)\mathbf{e}/k^2\}^2$, where $\mathbf{e}$ is a k vector of ones, I is the $k \times k$ identity matrix, and P_1' is the $k \times k$ matrix with elements $P_1'(r, s)$; $r, s = 1, \ldots, k$. This quantity was computed for the values $k = 2, 3, 5$, and 10 across dimensions $d = 1, 2, 3, 5, 10$, and ∞, and it fell into the extremely narrow range from 1.017 to 1.026 in all cases, leading to the conclusion that the weighted statistic $U_{k,n,\mathbf{w}^*}$ is not sufficiently superior to $T_{k,n}$ to be worth pursuing further.

4.2 Weighting by Reference Point Location

An alternative approach to weighting incorporates dependency on the actual position of each sample point with a statistic of the form

$$V_{k,n,\mathbf{w}} = \frac{1}{nk} \sum_{\alpha=1}^{2} \sum_{i\in\Omega_\alpha} w_\alpha(Z_i) \sum_{r=1}^{k} I_i(r) \quad (4.1)$$

for continuous weight functions $\mathbf{w} = (w_1, w_2)$. $V_{k,n,\mathbf{w}}$ can be designed to perform well against specific types of alternatives but unfortunately is not distribution-free. Asymptotic normality is supported by simulations.

The $V_{k,n,\mathbf{w}}$ test appears to achieve asymptotic power against any desired sequence of $O(n^{-1/2})$ alternatives to H when properly weighted. A procedure that has consistency against all fixed alternatives and asymptotic power in a chosen direction can be obtained consequently through a combination of the $T_{k,n}$ (or $U_{k,n,\mathbf{w}}$) and $V_{k,n,\mathbf{w}}$ tests (e.g., reject H if either $T_{k,n}$ or $V_{k,n,\mathbf{w}}$ rejects H at level $\alpha/2$). By the first Bonferroni bound, the level of the combined test does not exceed α.

Using the same model assumptions and expansions for alternatives as for $U_{k,n,\mathbf{w}}$, the efficacy coefficient of $V_{k,n,\mathbf{w}}$

Table 2. Efficiencies of $V_{k,n,\mathbf{w}}^*$ Relative to the Optimal Test for $\lambda_1 = \lambda_2 = .5$

d	$k = 1$	$k = 2$	$k = 3$
1	.89	.94	.96
2	.86	.93	.96
5	.83	.91	.94
10	.81	.90	.93
∞	.80	.89	.92

can be calculated and the Lagrange technique may be applied to find optimal weights for specific alternatives. This leads eventually to the optimal weight functions

$$w_i^*(x) = c_i(\lambda_1, \lambda_2, k) \frac{\partial}{\partial\theta} (\log q(x, \theta)) \mid_{\theta=\theta_o}, \qquad i = 1, 2,$$

where

$$c_i(\lambda_1, \lambda_2, k) = (-1)^{i+1}[(1 + 2\lambda_{3-i})(\lambda_i + \lambda_{3-i}k) - \lambda_{3-i}k(\lambda_i\bar{p}_1' + (\lambda_{3-i} - \lambda_i)\bar{p}_2')], \qquad i = 1, 2.$$

Note that these weight functions now depend on $\{K_n\}$ and are proportional to the first-order approximation to the likelihood ratio g/f; if each indicator function in the specification of $V_{k,n,\mathbf{w}}$ were set equal to 1 and the preceding weight functions were used, the $V_{k,n,\mathbf{w}}$ test would be virtually equivalent to the likelihood ratio test of H versus $\{K_n\}$.

The test statistic using these weight functions is denoted by $V_{k,n,\mathbf{w}^*}$. The limiting null moments of $V_{k,n,\mathbf{w}^*}$ are

$$\mu_{k,n,\mathbf{w}^*} = \lim E_H(V_{k,n,\mathbf{w}^*}) = 0$$

and

$$\sigma_{k,\mathbf{w}^*}^2 = \lim_{n\to\infty} nk \operatorname{var}_H(V_{k,n,\mathbf{w}^*}) = (\lambda_1 k + \lambda_2)\operatorname{var}(\lambda_1 w_1^*(Z)) + (\lambda_2 k + \lambda_1)\operatorname{var}(\lambda_2 w_2^*(Z)) + \lambda_1^2\lambda_2^2 k\bar{p}_1'\operatorname{var}(w_1^*(Z) + w_2^*(Z)) + \lambda_1\lambda_2 k(\bar{p}_2' + 2)\operatorname{var}(\lambda_1 w_1^*(Z) - \lambda_2 w_2^*(Z)),$$

where $Z \sim F = G$. The asymptotic variance is again well approximated by its infinite-dimensional limit. In the important special case $\lambda_1 = \lambda_2 = .5$, we have $w_2^*(x) = -w_1^*(x)$ and $\sigma_{k,\mathbf{w}^*}^2 = .25((3 + \bar{p}_2')k + 1)\operatorname{var} w_1^*$. The expression for the efficacy of $V_{k,n,\mathbf{w}^*}$ against $\{K_n\}$ is complex, but for $\lambda_1 = \lambda_2 = .5$ it reduces to

$$\xi^* = 2\Delta I^{1/2}(\theta_o)/\{3 + \bar{p}_2' + 1/k\}^{1/2}, \qquad (4.2)$$

where $I(\theta_o) = \int (q'(x, \theta_o))^2 q^{-1}(x, \theta_o)dx$ is the Fisher information number. The (optimal) likelihood ratio test for the case when q, θ_o, and Δ are known has an efficacy of $\Delta I^{1/2}(\theta_o)$ and thus the asymptotic efficiency of $V_{k,n,\mathbf{w}^*}$ relative to the optimal test of H versus $\{K_n\}$ when $\lambda_1 = \lambda_2$ is $4/(3 + \bar{p}_2' + 1/k)$. Since $\bar{p}_2'$ is generally near 1 (see Theorem 3.3 and Schilling 1986, table 3), efficiencies are quite high, as indicated in Table 2. It must be kept in mind, however, that for alternatives in other directions than that for which $\mathbf{w}^*$ was designed, power may be quite low.

5. CONDITIONAL TESTS

An alternative approach to nonparametric two-sample testing is to *condition* on the combined sample and use a permutation procedure. With conditioning, the distributions of the statistics previously considered now depend only on the graph-theoretical properties of the k nearest neighbor digraph that can be formed over the pooled sample by connecting each observation to its neighbors in the common sample, rather than requiring an intrinsic dimensionality for the data.

Consider, for example, (4.1), conditional on the values of $Z_1, \ldots, Z_n$. Defining scores

$$a_{ij} = (nk)^{-1} \sum_{\alpha=1}^{2} I(i, j \in \Omega_\alpha),$$

$$b_{ij} = \sum_{\alpha=1}^{2} I(i \in \Omega_\alpha) \sum_{r=1}^{k} w_\alpha(Z_i) I(NN_i(r) = Z_j),$$

(4.1) can be expressed as the generalized correlation coefficient (Daniels 1944) $\sum_{i=1}^n \sum_{j=1}^n a_{ij}b_{ij}$. As a result, conditional asymptotic normality holds under condition (18) of Friedman and Rafsky (1983); in particular, if k is either fixed or grows linearly with n, asymptotic normality holds in both the null and alternative cases.

The conditional null moments can be easily seen to be identical to those for the corresponding unconditional tests except that the neighbor configuration probabilities $p_1(r, s)$ and $p_2(r, s)$ are now replaced by the *proportions* of pairs (Z_i, Z_j), $i \neq j$, for which a mutual neighbor or shared neighbor relationship holds. Inasmuch as the quantities $np_1(r, s)$ and $np_2(r, s)$ have distribution-free limits, the asymptotic distributions obtained earlier are appropriate as approximations for conditional tests as well. Indeed, in those simulations in which both the conditional and unconditional normalized statistics were computed, the values were invariably quite close. Thus there appears to be little difference between the two approaches in practice. For those who object to permutation tests altogether on what may be termed philosophical grounds (see Basu 1980), this point may be reassuring.

6. MONTE CARLO RESULTS AND SUMMARY

6.1 Simulations

The performance of the various techniques introduced in the preceding sections was assessed for finite sample sizes by means of Monte Carlo experiments in $d = 1$, 2, 5, and 10 dimensions. The primary computational task is the identification of the k nearest neighbors of each sample point. This can be accomplished in $O(kn \log n)$ steps by means of an algorithm developed by Friedman, Bentley, and Finkel (1975); it should be noted, however, that computation time also grows rather significantly with d.

Tables 3 and 4 present Monte Carlo powers obtained for the $T_{k,n}$, $U_{k,n,\mathbf{w}^*}$, and $V_{k,n,\mathbf{w}^*}$ tests at level $\alpha = .05$ for multivariate normal samples differing in either location or scale,

Table 3. Monte Carlo Powers for Normal Location Alternatives [$F = N(\mathbf{0}, \mathbf{I})$, $G = N((\Delta, 0, \ldots, 0), \mathbf{I})$] at Level $\alpha = 5\%$, Based on 100 Trials Each ($n_1 = n_2 = 100$)

	$d = 1, \Delta = .3$			$d = 2, \Delta = .5$			$d = 5, \Delta = .75$			$d = 10, \Delta = 1.0$		
Statistic	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
$T_{k,n}$	5	9	12 (11)	7	21	26 (29)	38	50	66 (73)	43	61	67 (97)
$T_{k,n}$ (C)	5	9	12	7	22	27	41	57	74	47	67	76
$U_{k,n,\mathbf{w}}^*$	5	9	12	7	21	30	38	52	70	43	61	71
$V_{k,n,\mathbf{w}}^*$	71	68	72 (67)	100	100	100 (97)	99	99	97 (100)	100	100	100 (100)
$V_{k,n,\mathbf{w}}^*$ (C)	71	68	72	100	100	100	99	99	97	100	100	100
Combined	68	62	61 (55)	99	100	99 (93)	97	97	96 (100)	100	100	100 (100)

NOTE: Tests that condition on the pooled sample are indicated by (C). Power values in parentheses are theoretical values obtained from the asymptotics in Sections 3 and 4.

generated by means of the IMSL subroutine GGNOF. Also provided are the powers of the combined test (Section 4.2), which rejects if either $T_{k,n}$ or $V_{k,n,\mathbf{w}^*}$ rejects at level $\alpha/2$.

Both the unconditional and conditional (permutation test) versions were used, with the conditional tests indicated by (C) in Tables 3 and 4. Each number triplet gives the powers obtained by using $k = 1$, 2, and 3 neighbors, respectively. The parameters Δ and σ were selected to match those of Friedman and Rafsky (1979, table 1) in order to facilitate comparisons with their MST tests and standard parametric competitors. Values in parentheses indicate the corresponding theoretical powers for $k = 3$, obtained from the asymptotic developments outlined in Sections 3 and 4.

Several observations can be made based on the results shown in Tables 3 and 4. Weighting by rank of neighbor ($U_{k,n,\mathbf{w}^*}$) produced only slightly more detections overall than did the unweighted test $T_{k,n}$, which supports the efficiency remarks of Section 4.1. The performance of $T_{k,n}$ was generally comparable to that of Friedman and Rafsky's MST analog of the Wald–Wolfowitz runs test. Arguments similar to those in Section 4 suggest strongly that these tests, as with $T_{k,n}$ and $U_{k,n,\mathbf{w}}$, also have asymptotic power only against sequences of alternatives converging to H at rate $O(n^{-1/4})$ or faster. Letting $k = k(n) \to \infty$ as $n \to \infty$ would presumably yield power against somewhat faster converging sequences of alternatives for each of the tests under discussion; however, the problem of choosing the optimal rate for $k(n)$ is a difficult one that needs further study.

The $V_{k,n,\mathbf{w}^*}$ test achieved extremely high power in every case. Since $V_{k,n,\mathbf{w}}$ is designed for a much more restrictive model than the completely general situation appropriate to $T_{k,n}$ and $U_{k,n,\mathbf{w}^*}$, it is natural that $V_{k,n,\mathbf{w}^*}$ would perform substantially better within that model.

Powers computed by means of the infinite-dimensional approximating distributions (not shown) were never more than 3% different from the tabled values and in most cases agreed with the finite-dimensional powers exactly.

Asymptotic powers agree closely with the Monte Carlo results for low dimensions but are higher than the realized powers for $d = 5$ and 10. This can be understood by noting that k nearest neighborhoods are not sufficiently local in large-dimensional spaces, because of the sparsity of the sample points, to accurately reflect the relationship of f to g in the vicinity of the reference point unless k is quite small and the number of observations is quite large. Specifically, the diameter of k nearest neighborhoods is of order $O((k/n)^{1/d})$. This is also reflected in the fact that the gain in power for $T_{k,n}$ and $U_{k,n,\mathbf{w}^*}$ as k increases is less in higher dimensions than in the low-dimensional cases. It is probable, however, that choosing $k > 3$ would have produced higher powers than those shown in Tables 3 and 4 in all cases and particularly for $d = 1$ and 2.

The efficacy formulas (3.7) and (4.2) for $T_{k,n}$ and $V_{k,n,\mathbf{w}^*}$, respectively, along with Theorems 3.2 and 3.3, indicate that increasing k is much less important for $V_{k,n,\mathbf{w}^*}$ than for $T_{k,n}$. In fact, the simulations in Tables 3 and 4 show $V_{1,n,\mathbf{w}^*}$ generally outperforming $V_{3,n,\mathbf{w}^*}$. This can again be attributed to the phenomenon discussed before.

Results for the tests that condition on the pooled sample are similar to those for their unconditional counterparts. The agreement between the $p_1(r, s)$ and $p_2(r, s)$ counts and their limiting values was generally very good, although as d increases a tendency occurs for p_1 values to be smaller and p_2 values to be larger than their respective asymptotic limits, presumably because of the magnitude of fringe effects in large dimensional samples.

Table 4. Monte Carlo Powers for Normal Scale Alternatives [$F = N(\mathbf{0}, \mathbf{I})$, $G = N(\mathbf{0}, \sigma^2\mathbf{I})$] at Level $\alpha = 5\%$, Based on 100 Trials Each ($n_1 = n_2 = 100$)

	$d = 1, \sigma = 1.3$			$d = 2, \sigma = 1.2$			$d = 5, \sigma = 1.2$			$d = 10, \sigma = 1.1$		
Statistic	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
$T_{k,n}$	9	17	16 (14)	9	13	17 (15)	13	18	19 (41)	5	7	7 (20)
$T_{k,n}$ (C)	9	16	16	9	15	17	14	20	23	5	8	7
$U_{k,n,\mathbf{w}}^*$	9	16	19	9	15	16	13	19	21	5	8	6
$V_{k,n,\mathbf{w}}^*$	86	83	84 (82)	78	77	72 (81)	93	90	90 (99)	73	69	70 (90)
$V_{k,n,\mathbf{w}}^*$ (C)	85	83	85	78	77	72	93	90	90	64	65	62
Combined	79	74	74 (73)	66	66	65 (71)	87	87	87 (98)	50	52	52 (83)

NOTE: Tests that condition on the pooled sample are indicated by (C). Power values in parentheses are theoretical values obtained from the asymptotics in Sections 3 and 4.

The combined test, which "robustifies" the optimally weighted test V_{k,n,w^*}, appears to be not greatly inferior to V_{k,n,w^*} itself.

6.2 Summary

Intuitively simple nearest neighbor proportions tests are available for both the general model $H : F = G$ (unknown) versus $K : F \neq G$ and for models specifying a null distribution. In contrast to previous nearest neighbor and MST tests, explicit unconditional null distributions are available, along with extremely simple and accurate infinite-dimensional approximating distributions [see (3.3) and Table 1]. A corollary advantage is that power and efficiency can be assessed.

Asymptotic results provide a good fit to experiments with moderate sample sizes if d is not too large. The unweighted test is consistent against all alternatives and appears to achieve good power for the general model. Tests for specific distributions have power close to that of the likelihood ratio test.

Both classes of tests can be performed either as conditional or unconditional tests with little difference in the results. Computational requirements are moderate.

Extensions can be made in straightforward fashion to the p-sample problem for $p > 2$. It might also be possible to use $T_{k,n}$ as an estimator of the discrepancy between F and G—for example, by searching for a "reasonable" transformation of (say) $Y_1, \ldots, Y_{n_2}$ to $Y_1', \ldots, Y_{n_2}'$ that yields $T_{k,n}(X_1, \ldots, X_{n_1}, Y_1', \ldots, Y_{n_2}') = .5$, thus making the samples $X_1, \ldots, X_{n_1}$ and $Y_1', \ldots, Y_{n_2}'$ "well-mixed" in that sense.

APPENDIX: PROOF OF THEOREM 3.1

The statistic $T_{k,n}$ can be written in terms of the Rogers statistics as

$$T_{k,n} = \frac{1}{nk} \sum_{\alpha=1}^{2} \sum_{j=1}^{k} jS_{\alpha,j}.$$

Hence the limiting null distribution of $T_{k,n}$ is normal.

It is necessary to find the first two null moments. Only the variance requires any real effort. The mean is given by

$$E_H(T_{k,n}) = \frac{1}{nk} \sum_{i=1}^{n} \sum_{r=1}^{k} P_H(I_i(r) = 1).$$

Since $P_H(I_i(r) = 1) = (n_\alpha - 1)/(n - 1)$ for $i \in \Omega_\alpha$, $\alpha = 1, 2$, $r = 1, \ldots, k$, we easily obtain

$$E_H(T_{k,n}) = \frac{1}{n(n-1)} \sum_{\alpha=1}^{2} n_\alpha(n_\alpha - 1). \tag{A.1}$$

The variance is considerably more complex. We have

$$\mathrm{var}_H(nkT_{k,n}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{k} \sum_{s=1}^{k} P_H(I_i(r) = I_j(s) = 1) - \{nkE_H(T_{k,n})\}^2.$$

For terms in which $i = j \in \Omega_\alpha$ ($\alpha = 1, 2$), we readily obtain

$$P_H(I_i(r) = I_j(s) = 1) = \frac{n_\alpha - 1}{n - 1}\left(1 + \frac{n_\alpha - 2}{n - 1}\right) I(r \neq s). \tag{A.2}$$

When $i \neq j$ various nearest neighbor geometries come into play, with five mutually exclusive and exhaustive cases involved: (i) $NN_i(r) = Z_j$, $NN_j(s) = Z_i$; (ii) $NN_i(r) = NN_j(s)$; (iii) $NN_i(r) = Z_j$, $NN_j(s) \neq Z_i$; (iv) $NN_i(r) \neq Z_j$, $NN_j(s) = Z_i$; (v) $NN_i(r) \neq Z_j$, $NN_j(s) \neq Z_i$, $NN_i(r) \neq NN_j(s)$. These cases affect how many points are required to be from matching samples—either two or three, or two pairs. In particular, Z_i and Z_j are mutual neighbors if case (i) occurs for some r and s, and they share a neighbor if case (ii) occurs.

Let the null probabilities of these events be denoted by $p_1(r, s), \ldots, p_5(r, s)$, respectively. Note that these probabilities are independent of the sample identities of the points involved and symmetric to the arguments. It is easy to see that for $i \neq j$,

$$P_H(I_i(r) = I_j(s) = 1) = \sum_{a=1}^{5} c_a p_a(r, s), \tag{A.3}$$

where

$$\begin{aligned}
c_1 &= \sum_{\alpha=1}^{2} I[i, j \in \Omega_\alpha] \\
c_2 = c_3 = c_4 &= \sum_{\alpha=1}^{2} I[i, j \in \Omega_\alpha] \cdot \frac{n_\alpha - 2}{n - 2} \\
c_5 &= \sum_{\alpha=1}^{2} I[i, j \in \Omega_\alpha] \frac{(n_\alpha - 2)(n_\alpha - 3)}{(n - 2)(n - 3)} \\
&\quad + \sum_{\alpha=1}^{2} I[i \in \Omega_\alpha, j \in \Omega_{3-\alpha}] \frac{(n_1 - 1)(n_2 - 1)}{(n - 2)(n - 3)}.
\end{aligned} \tag{A.4}$$

Now using $p_1(r, s) = (n - 1)^{-1} P_H(NN_j(s) = Z_i \mid NN_i(r) = Z_j)$ we easily obtain

$$p_3(r, s) = p_4(r, s) = 1/(n - 1) - p_1(r, s) \tag{A.5}$$

and

$$p_5(r, s) = (n - 3)/(n - 1) + p_1(r, s) - p_2(r, s). \tag{A.6}$$

Thus $\mathrm{var}_H(T_{k,n})$ depends on $F = G$ only through $p_1(r, s)$ and $p_2(r, s)$, the probabilities associated with mutual neighbors and shared neighbors, respectively.

Using (A.3)–(A.6) and numerous algebraic manipulations, the following expression can be obtained for the variance:

$$\begin{aligned}
\mathrm{var}_H(nkT_{k,n}) &= [kn_1n_2/(n - 1)][1 - \{k/(n - 1)\} \\
&\quad \times \{(n_1 - n_2)^2/(n - 2) + 1\}] \\
&\quad + \frac{n_1n_2}{n - 3}\left[\frac{4(n_1 - 1)(n_2 - 1)}{n - 2} \sum_{r=1}^{k} \sum_{s=1}^{k} p_1(r, s)\right. \\
&\quad \left. + \left\{\frac{(n_1 - n_2)^2}{n - 2} - 1\right\} \sum_{r=1}^{k} \sum_{s=1}^{k} p_2(r, s)\right].
\end{aligned} \tag{A.7}$$

Theorem 3.1 follows directly upon taking limits in (A.1) and (A.7).

[*Received March 1983. Revised October 1985.*]

REFERENCES

Basu, D. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test" (with discussion), *Journal of the American Statistical Association*, 75, 575–595.

Bickel, P. J., and Breiman, L. (1983), "Sums of Functions of Nearest Neighbor Distances, Moment Bounds, Limit Theorems and a Goodness of Fit Test," *Annals of Probability*, 11, 185–214.

Birgé, L. (1983), "Approximation Dans les Espaces Métriques et Théorie de l'Estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65, 181–237.

Daniels, H. E. (1944), "The Relation Between Measures of Correlation in the Universe of Sample Permutations," *Biometrika*, 33, 120–135.
Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.
Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1975), "An Algorithm for Finding Best Matches in Logarithmic Time," Stanford Linear Accelerator Center Report SLAC-PUB-1549, Stanford University, Computer Science Dept.
Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate Generalizations of the Wald–Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics*, 7, 697–717.
——— (1983), "Graph-Theoretic Measures of Multivariate Association and Prediction," *The Annals of Statistics*, 11, 377–391.
Friedman, J. H., and Steppel, S. (1974), "A Nonparametric Procedure for Comparing Multivarite Point Sets," SLAC Computation Group (internal) Technical Memo 153 [U.S. Atomic Energy Contract AT(043)515], Stanford University.
Le Cam, L. (1973), "Convergence of Estimates Under Dimensionality Restrictions," *The Annals of Statistics*, 1, 38–53.
Rogers, W. H. (1976), "Some Convergence Properties of K-Nearest Neighbor Estimates," unpublished Ph.D. thesis, Stanford University, Dept. of Statistics.
Schilling, M. F. (1983a), "Goodness of Fit Testing Based on the Weighted Empirical Distribution of Certain Nearest Neighbor Statistics, *The Annals of Statistics*, 11, 1–12.
——— (1983b), "An Infinite-Dimensional Approximation for Nearest Neighbor Goodness of Fit Tests," *The Annals of Statistics*, 11, 13–24.
——— (1986), "Mutual and Shared Neighbor Probabilities: Finite and Infinite Dimensional Results," *Advances in Applied Probability*, 18.