TESTING FOR EQUAL DISTRIBUTIONS IN HIGH DIMENSION

Gábor J. Székely [*†]
Bowling Green State University

Maria L. Rizzo [‡]
Ohio University

October 30, 2004

Abstract

We propose a new nonparametric test for equality of two or more multivariate distributions based on Euclidean distance between sample elements. Several consistent tests for comparing multivariate distributions can be developed from the underlying theoretical results. The test procedure for the multisample problem is developed and applied for testing the composite hypothesis of equal distributions, when distributions are unspecified. The proposed test is universally consistent against all fixed alternatives (not necessarily continuous) with finite second moments. The test is implemented by conditioning on the pooled sample to obtain an approximate permutation test, which is distribution free. Our Monte Carlo power study suggests that the new test may be much more sensitive than tests based on nearest neighbors against several classes of alternatives, and performs particularly well in high dimension. Computational complexity of our test procedure is independent of dimension and number of populations sampled. The test is applied in a high dimensional problem, testing microarray data from cancer samples.

Keywords: homogeneity, two-sample problem, multisample problem, permutation test, e-distance, E-statistics, energy statistics.

1. INTRODUCTION

[*]This work was partially supported by NSA Grant MDA904-02-1-0091.

[†]Gábor J. Székely, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, email: gabors@bgnet.bgsu.edu.

[‡]Maria L. Rizzo, Department of Mathematics, Ohio University, Athens, OH 45701, email: rizzo@math.ohiou.edu.

We propose a new multivariate test for equal distributions, which is practical and powerful for high dimensional data. The univariate two-sample problem has been studied extensively, and several classical tests are available. Classical approaches to the two-sample problem in the univariate case based on comparing empirical distribution functions, such as the Kolmogorov–Smirnov and Cramér-von Mises tests, do not have a natural distribution-free extension to the multivariate case. Since tests for the multivariate problem based on maximum likelihood depend on distributional assumptions about the underlying populations, these tests do not apply to the general two-sample or k-sample problem, and may not be robust to departures from these assumptions. Bickel (1969) constructed a consistent distribution free multivariate extension of the univariate Smirnov test by conditioning on the pooled sample. Friedman and Rafsky (1979) proposed distribution free multivariate generalizations of the Wald-Wolfowitz runs test and Smirnov test for the two-sample problem, based on the minimal spanning tree of the pooled sample. A class of consistent, asymptotically distribution free tests for the multivariate problem is based on nearest neighbors in the Euclidean distance metric (see Bickel and Breiman 1983, Henze 1988, and Schilling 1986). Henze (1988) generalized the nearest neighbor tests to distances based on norms other than the Euclidean norm. The nearest neighbor tests apply to testing the k-sample hypothesis when all distributions are continuous.

A formal statement of the two-sample and k-sample problem is as follows. Suppose that $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent random samples of R^d-valued random vectors, $d \geq 1$, with respective distributions F_1 and F_2. The two-sample problem is to test

$$H_0 : F_1 = F_2 \tag{1}$$

versus the composite alternative $F_1 \neq F_2$. For k samples from distributions $F_1, \ldots, F_k$, the corresponding k-sample hypothesis is

$$H_0 : F_1 = \cdots = F_k \tag{2}$$

versus the composite alternative $F_i \neq F_j$ for some $1 \leq i < j \leq k$. The notation $X \stackrel{\mathcal{D}}{=} Y$ is defined to mean that $F_X = F_Y$, where F_X and F_Y are the distributions of X and Y respectively.

This paper presents a new approach to testing the k-sample hypothesis (2) using the $\mathcal{E}$ statistic (defined below) based on Euclidean distances between sample elements. The proposed test statistic is a function of e-distances, a multivariate measure of the distance between distributions that is introduced in Section 2. Székely and Móri (2001) proposed the statistic for the special case of testing diagonal symmetry (see also Székely 2000b). These test statistics belong to a class

of multivariate statistics ($\mathcal{E}$-statistics or energy statistics) proposed by Székely (2000a).

We show (see Appendix) that *the limiting distribution of the test statistic exists under the hypothesis of equal distributions, and the test is consistent against fixed general alternatives.* The $\mathcal{E}$ test is more general than tests based on ranks of neighbors in the sense that no assumptions about continuity of the underlying distributions are necessary. Computational complexity of the $\mathcal{E}$ test does not depend on dimension or number of samples. Calculations do not involve sorting, and the resulting procedure has certain computational advantages over the distribution free procedures above. These properties suggest that the $\mathcal{E}$ test is a practical new approach to the multivariate k-sample problem that is applicable in arbitrarily high dimension, and our empirical results show that the $\mathcal{E}$ test is powerful competitor to nearest neighbor tests. In our empirical study, the $\mathcal{E}$ test was notably superior to the nearest neighbor test in higher dimensions, suggesting potential application in problems with high dimensional data such as microarray data.

In Section 2 multivariate e-distance and its properties are introduced. The two-sample and k-sample test statistics are presented in Section 3, and details of implementation follow in Section 4. The Monte Carlo power comparison is presented in Section 5. An application to testing gene expression data of cancer samples is discussed in Section 6, followed by a Summary in Section 7.

2. MULTIVARIATE e-DISTANCE

Several new tests for comparing multivariate distributions, including the test for equal distributions that is the subject of this paper, are based on the inequality (4) below, relating expected values of the distances between random vectors.

First, we introduce the e-distance between finite sets. Let $\mathcal{A} = \{a_1, \ldots, a_{n_1}\}$ and $\mathcal{B} = \{b_1, \ldots, b_{n_2}\}$ be disjoint nonempty subsets of R^d. Euclidean norm is denoted by $\|\cdot\|$. Define the e-distance $e(\mathcal{A}, B)$ between $\mathcal{A}$ and $\mathcal{B}$ as

$$e(\mathcal{A}, B) = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|a_i - b_j\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|a_i - a_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|b_i - b_j\|\right). \quad (3)$$

The non-negativity of $e(\mathcal{A}, B)$ is a special case of the following inequality proved in Székely and Rizzo (2003a):

If X, X', Y, Y' are independent random vectors in R^d with finite

expectations, $X \stackrel{\mathcal{D}}{=} X'$ and $Y \stackrel{\mathcal{D}}{=} Y'$, then

$$2E\|X-Y\| - E\|X-X'\| - E\|Y-Y'\| \geq 0, \tag{4}$$

and equality holds if and only if X and Y are identically distributed.

The above inequality has been applied to develop consistent rotation-invariant goodness-of-fit tests, a consistent test of multivariate normality (Rizzo 2002, Székely and Rizzo 2005), test for Poisson distribution (Székely and Rizzo 2004) and clustering procedure (Székely and Rizzo 2003b). In this paper we develop a consistent test for the k-sample hypothesis (2) based on the e-distance defined in (3).

We begin with the two-sample problem. Consider independent random samples $\mathcal{A} = \{X_1, \ldots, X_{n_1}\}$ and $\mathcal{B} = \{Y_1, \ldots, Y_{n_2}\}$ of R^d valued random vectors X and Y respectively, where $E\|X\| < \infty$ and $E\|Y\| < \infty$. Let $\mu_{AB} := E\|X-Y\|$, $\mu_A := E\|X_1 - X_2\|$, and $\mu_B := E\|Y_1 - Y_2\|$. Then

$$\begin{aligned} E[e(\mathcal{A},\mathcal{B})] =& \frac{n_1 n_2}{n_1+n_2}\left(2\mu_{AB} - \frac{n_1-1}{n_1}\,\mu_A - \frac{n_2-1}{n_2}\,\mu_B\right) \\ =& \frac{n_1 n_2}{n_1+n_2}\,(2\mu_{AB} - \mu_A - \mu_B) + \frac{n_2\mu_A}{n_1+n_2} + \frac{n_1\mu_B}{n_1+n_2}. \end{aligned}$$

If X and Y are identically distributed, then $\mu_{AB} = \mu_A = \mu_B$, implying that $2\mu_{AB} - \mu_A - \mu_B = 0$ and

$$E[e(\mathcal{A},\mathcal{B})] = \frac{n_2\mu_A + n_1\mu_B}{n_1+n_2} = \mu_{AB} = E\|X-Y\|,$$

for all positive integers n_1 and n_2. If X and Y are not identically distributed, it follows from (4) that $2\mu_{AB} - \mu_A - \mu_B = c > 0$. Hence if X and Y are not identically distributed, and $n = n_1 + n_2$, $E[e(\mathcal{A},\mathcal{B})]$ is asymptotically a positive constant times n, provided n_1/n converges to a constant in $(0,1)$. As the sample size n tends to infinity, under the null hypothesis $E[e(\mathcal{A},\mathcal{B})]$ tends to a positive constant, while under the alternative hypothesis $E[e(\mathcal{A},\mathcal{B})]$ tends to infinity. Not only the expected value of e, but e itself, converges (in distribution) under the null hypothesis, and tends to infinity (stochastically) otherwise.

Note that conditional on the observed samples, we have

$$\begin{aligned} 2\mu_{AB} - \mu_A - \mu_B =& \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|x_i - y_j\| \\ & - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|x_i - x_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|y_i - y_j\|\Bigg). \end{aligned}$$

Hence, conditional on the observed samples, $e(\mathcal{A},\mathcal{B})$ is exactly $n/(n_1 n_2)$ times the nonnegative constant $2\mu_{AB} - \mu_A - \mu_B$, where $2\mu_{AB} - \mu_A -$

$\mu_B = 0$ if and only if the distributions of X and Y are equal. Clearly large e-distance corresponds to different distributions, and measures the distance between distributions in a similar sense as the univariate empirical distribution function (edf) statistics. In contrast to edf statistics, however, e-distance does not depend on the notion of a sorted list, and e-distance is by definition a multivariate measure of distance between distributions. These observations show that e-distance determines a multivariate test for equal distributions, which is more formally developed in the next section.

3. THE MULTIVARIATE $\mathcal{E}$ TEST FOR EQUAL DISTRIBUTIONS

Suppose $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are independent random samples of random vectors in R^d, $d \geq 1$. If $\mathcal{A}$ denotes the first sample, and $\mathcal{B}$ denotes the second sample, we propose the statistic $\mathcal{E}_{n_1,n_2} = e(\mathcal{A}, \mathcal{B})$ for testing the two-sample hypothesis (1). The two-sample test statistic is

$$\mathcal{E}_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2} \Bigg(\frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} \|X_i - Y_m\| - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|X_i - X_j\| - \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{m=1}^{n_2} \|Y_\ell - Y_m\| \Bigg). \quad (5)$$

Let n denote the total sample size of the pooled sample. Under the null or alternative hypothesis, a random permutation $W_1^{(\pi)}, \dots, W_n^{(\pi)}$ of the pooled sample $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ is equal in distribution to a random sample size n from the mixture W, where W is the random variable obtained by sampling from the distribution of X_1 with probability n_1/n and sampling from the distribution of Y_1 with probability n_2/n.

Let $\alpha \in (0,1)$ be fixed, and let c_α be the constant satisfying $\lim_{n\to\infty} P(\mathcal{E}_n > c_\alpha) = \alpha$, where P denotes the limiting probability for two independent random samples size n_1 and n_2 from the mixture W. Here we assume that n_1/n converges to a constant in $(0,1)$. Refer to the Appendix for a proof of the existence of c_α. The size α test of $H_0 : F_X = F_Y$ based on $\mathcal{E}_n$ that rejects the null hypothesis if $\mathcal{E}_n > c_\alpha$ is universally consistent against a very general class of alternatives (all distributions F_X, F_Y with finite second moments) provided n_1/n converges to a constant $(0,1)$. The asymptotic properties of $\mathcal{E}_n$ are intuitively obvious from the properties of e-distance given in the preceding section. Proofs of the existence of the limiting distribution of $\mathcal{E}_n$ and consistency are outlined in the Appendix.

The two-sample $\mathcal{E}$ statistic is easily generalized to the k-sample problem. Suppose $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$ are $k \geq 2$ independent random sam-

ples of random vectors in R^d, sizes $n_1, n_2, \ldots, n_k$ respectively, and let $n = n_1 + \cdots + n_k$. For each pair of samples $(\mathcal{A}_i, \mathcal{A}_j)$, $i \neq j$, let $\mathcal{E}_{n_i,n_j}(\mathcal{A}_i, \mathcal{A}_j) = e(\mathcal{A}_i, \mathcal{A}_j)$ denote the corresponding two-sample $\mathcal{E}$ statistic. The k-sample test statistic is obtained by summing the e-distances between samples over all $k(k-1)/2$ pairs of samples:

$$\mathcal{E}_n = \sum_{1 \le i < j \le k} \mathcal{E}_{n_i,n_j}(\mathcal{A}_i, \mathcal{A}_j) = \sum_{1 \le i < j \le k} e(\mathcal{A}_i, \mathcal{A}_j), \tag{6}$$

where $\mathcal{E}_{n_i,n_j}(\mathcal{A}_i, \mathcal{A}_j)$ is given by (5). Large values of $\mathcal{E}_n$ are significant, and the asymptotic properties of the k-sample test follow immediately from the results for the two-sample test by induction, provided all second moments of the sampled populations are finite and n_i/n converges to a constant in $(0, 1)$, $i = 1, \ldots, k$.

4. IMPLEMENTATION

The null distribution of the test statistic $\mathcal{E}_n$ depends on the unknown common distribution F of the sampled populations. Although one can estimate the null distribution in the case of a completely specified F, it is of much greater practical interest to develop a test procedure for the case where the only information available about the sampled populations is contained in the observed samples. We condition on the pooled observed samples to obtain a distribution free test procedure. This procedure is essentially the permutation test approach outlined e.g. by Efron (1993, Chapter 15).

Suppose $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ are $k \geq 2$ independent random samples in R^d from distributions $F_1, \ldots, F_k$ respectively. Consider the pooled sample $\{W_1, \ldots, W_n\} = \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_k$. Under the null hypothesis, $W_1, \ldots, W_n$ are independent and identically distributed (iid), with common distribution function F. If the desired significance level is α, resample (without replacement) from $W_1, \ldots, W_n$ a suitable number B of random samples size n so that $(B+1)\alpha$ is an integer. Let $m_j = \sum_{i=1}^{j} n_i$ and $m_0 = 0$. For each bootstrap sample $W_1^{(b)}, \ldots, W_n^{(b)}$, $b = 1, \ldots, B$, compute $\mathcal{E}_n^{(b)}$ determined by the k samples $\mathcal{A}_i^{(b)} = \{W_{m_{i-1}+1}^{(b)}, \ldots, W_{m_i}^{(b)}\}$, $i = 1, \ldots, k$. The bootstrap estimate of $P_n(\mathcal{E}_n \leq t)$ is $\frac{1}{B} \sum_{b=1}^{B} I(\mathcal{E}_n^{(b)} \leq t)$, where $I(\cdot)$ is the indicator function. Reject $H_0 : F_1 = \cdots = F_k$ if the observed $\mathcal{E}_n$ exceeds $100(1-\alpha)\%$ of the replicates $\mathcal{E}_n^{(b)}$.

If all possible distinct permutations of the pooled sample into k samples of size $n_1, \ldots, n_k$ are drawn, the resulting test is exact. In this case we have consistency for an even larger class of alternatives, all fixed alternatives with finite first moments. For realistic sample sizes in

multivariate problems, however, the computational resources required for the exact test are excessive. Our implementation approximates the exact permutation test procedure using B random permutations.

The resulting bootstrap implementation of the test, which is an approximate permutation test, may be performed by sampling with or without replacement. Computationally, sampling without replacement is simpler because it is not necessary to update the distance matrix for each permutation. Therefore, in our test implementation the resampling method is without replacement.

Software implementing the multivariate energy test for equal distributions is available in the `energy` package, distributed for R through the Comprehensive R Archive Network, `http://cran.r-project.org/`.

5. EMPIRICAL RESULTS

In this section, the empirical power performance of our proposed test for equal distributions is compared with nearest neighbor (NN) tests. Nearest neighbor tests are among only a few universally consistent tests available for testing equality of multivariate distributions. The test applies to continuous distributions in arbitrary dimension.

The nearest neighbor test statistic and procedure is as follows (see Schilling 1986 or Henze 1988). In the notation of the preceding section, we define the r^{th} nearest neighbor of W_i to be the sample element W_j satisfying $\|W_i - W_\ell\| < \|W_i - W_j\|$ for exactly $r-1$ indices $1 \le \ell \le n$, $\ell \ne i$. Denote the r^{th} nearest neighbor of a sample element W_i by $NN_r(W_i)$. For $i = 1, \ldots, n$ define indicator function $I_i(r)$ as $I_i(r) = 1$ if W_i and $NN_r(W_i)$ belong to the same sample, and otherwise $I_i(r) = 0$. The J^{th} nearest neighbor statistic measures the proportion of first through J^{th} nearest neighbor coincidences:

$$T_{n,J} = \frac{1}{nJ} \sum_{i=1}^{n} \sum_{r=1}^{J} I_i(r). \tag{7}$$

Under the hypothesis of equal distributions, the pooled sample has on average less nearest neighbor coincidences than under the alternative hypothesis, so the test rejects the null hypothesis for large values of $T_{n,J}$. Henze (1988) proved that the limiting distribution of a class of nearest neighbor statistics is normal for any distance generated by a norm on R^d. Schilling (1986) derived the mean and variance of the distribution of $T_{n,2}$ for selected values of n_1/n and d in the case of Euclidean norm. In general, the parameters of the normal distribution may be difficult to obtain analytically. If we condition on the pooled sample to implement an exact permutation test, the procedure is distribution free. Our implementation approximates this procedure using the same bootstrap implementation (sampling without replacement) described in the previous section.

The empirical power of the third nearest neighbor test $T_{n,3}$ was compared in a Monte Carlo simulation versus our proposed test statistic $\mathcal{E}_n$. In each case, the empirical power was estimated from simulation of $10,000$ pairs of samples. Sample data was generated using standard routines available in Splus. Empirically we found that the number of bootstrap replicates required for NN tests to adequately approximate the exact test was much higher than that required for the $\mathcal{E}$ test. In particular, the first two NN statistics $T_{n,1}$ and $T_{n,2}$ did not achieve close to nominal significance levels in our simulations with a reasonably large number of replicates, $B = 499$. For this reason, we have reported comparisons for the $T_{n,3}$ statistic only. Both the $\mathcal{E}_n$ and $T_{n,3}$ statistics achieved approximately correct empirical significance in our simulations using $B = 499$ replicates for all sample sizes. For example, see case $\delta = 0$ in Table 1.

For each pair of simulated samples, calculation of $\mathcal{E}_n$ and $T_{n,3}$ test statistics were implemented in our external libraries coded in C, and the test decision was based on $B = 499$ bootstrap replicates. Empirical results are given below for selected alternatives, sample sizes, and dimension, at significance level $\alpha = 0.1$.

Table 1 lists empirical results for comparison of simulated bivariate normal location alternatives from $F_1 = N_2((0,0), I)$ and $F_2 = N_2((0,\delta), I)$ for equal sample sizes. The empirical significance levels ($\delta = 0$) for both tests are approximately equal to the nominal significance level, although $T_{n,3}$ may be slightly higher. These alternatives differ in location only, and the empirical evidence suggests that $\mathcal{E}_n$ is more powerful than $T_{n,3}$ against this class of alternatives.

The notation $F^{(d)}$ denotes d-variate distributions with iid marginal distributions F. We have compared $\mathcal{E}_n$ and $T_{n,3}$ across sample sizes $n = 50, 75, 100$ and $d = 1, 2, 5, 10, 20, 50$ in Tables 2 and 3. Against the alternatives $F_1 = N_d(0, I)$; $F_2 = t(5)^{(d)}$, empirical evidence exhibited in Table 2 suggests that the tests perform about the same in the univariate case, but $\mathcal{E}_n$ is the superior test in higher dimensions. The improvement in power of $\mathcal{E}_n$ relative to $T_{n,3}$ is quite dramatic as dimension increases. Against the alternatives $\text{Uniform}(a,b)^{(d)}$ compared in Table 3, $\mathcal{E}_n$ is superior to $T_{n,3}$. In $d = 5, 10$ and 20 the empirical power of $\mathcal{E}_n$ is about twice that of $T_{n,3}$ against this alternative.

Since nearest neighbor tests are based on ranks, the tests are not appropriate for comparing discrete distributions. Thus, our comparisons did not include discrete alternatives, but in fact, $\mathcal{E}_n$ is applicable and very effective for comparing multivariate discrete distributions.

If the combined sample size n is quite large, we can alternately define an incomplete U-statistic and associated V-statistic (see Appendix) with kernel (10), averaging over a random sample of the n^2 distances. A great reduction of computational effort is achieved while maintaining whatever level of significance is desired. From Janson

(1984) the asymptotic distributions of the incomplete version and the complete version of the statistic $\mathcal{E}_n = (n_1 n_2/n)V_n$ are equal. For this reason and other considerations such as bootstrap size, $\mathcal{E}_n$ has certain computational advantages over the nearest neighbor tests.

Both the nearest neighbor (NN) tests and the $\mathcal{E}$ test require initially computing all the pairwise distances. Thus, both types of tests have at least $O(n^2)$ time complexity for the observed test statistic. Ideally, there are sufficient resources to store the distance matrix, requiring $O(n^2)$ space complexity. An efficient approach to computing the $\mathcal{E}$ statistic is to permute only the indices rather than the actual sample vectors, and use the permutation vector to retrieve the distances from the distance matrix. This approach also applies to the NN tests. When the number of bootstrap replicates is large relative to total sample size n, the bootstrap operations dominate the calculations, so that the computing time increases proportional to the time per replicate for each statistic given the stored distance matrix, which is $O(n \log n)$ for $T_{n,3}$ and $O(n^2)$ for $\mathcal{E}_n$. The difference on a Pentium 4 2.0 GHz computer is 0.071-0.047 = 0.024 seconds per test decision when $n = 200$ and $B = 499$. Note that the time complexity of the approximate permutation test for $\mathcal{E}$ or NN tests is independent of dimension and number of sampled populations.

6. APPLICATION: CANCER MICROARRAY DATA

In this section we discuss an application testing for differences between gene expression data in cancer samples, an example involving very high dimension and small sample sizes. Microarrays are the focus of much scientific research to study the variation among tumors. Székely and Rizzo (2003b) compared three hierarchical clustering methods applied to the NCI60 microarray data discussed in Chapter 14 of Hastie, Tibshirani, and Friedman (2001). An extensive analysis of the NCI60 data appears in Ross, et al. (2000). Classifications produced by hierarchical clustering methods did not reliably distinguish between lung cancer and ovarian cancer samples. We applied the $\mathcal{E}_n$ and $T_{n,3}$ tests to determine whether gene expression data of lung cancer cells and ovarian cancer cells are significantly different.

The raw data are expression levels from cDNA microarrays, in 60 cell cancer cell lines used in the screen for anti-cancer drugs by the National Cancer Institute. The data [http://genome-www.stanford.edu/nci60/nci.data] is an array of 6830 gene expression measurements for 64 human cancer samples. The samples include nine types of cancers: breast (7), central nervous system (CNS) (5), colon (7), leukemia (6), melanoma (8), non-small-cell-lung-carcinoma (NSCLC) (9), ovarian (6), prostate (2), renal (9), and unknown cancer (1). The data has been centered to row median zero and column median zero, and

for this analysis we have removed variables with missing values. The resulting data has dimension 1962.

Table 4 gives the two-sample test results comparing lung cancer with other cancers. All sample sizes are comparable and small (total size less than 20 in each test). Both the $\mathcal{E}_n$ and $T_{n,3}$ tests are highly significant for differences between lung cancer and breast, CNS, colon, leukemia, melanoma and renal cancer, but neither test is significant at $\alpha = 0.1$ when lung cancer is compared with ovarian cancer. This helps to explain why cluster analysis methods applied to this data did not reliably discriminate between lung cancer and ovarian cancer.

7. SUMMARY

We have presented and implemented a multivariate test for equal distributions that is consistent against a very general class of alternatives. The statistic $\mathcal{E}_n$ can be applied to the multivariate k-sample problem for discrete or continuous data, in arbitrary dimension $d \geq 1$. Theoretical properties of the two-sample $\mathcal{E}_n$ test and the bootstrap implementation are generalized to the k-sample problem. The test can also be extended to an arbitrarily large sample implementation using incomplete V-statistics, with the same asymptotic properties. The bootstrap test procedure is demonstrated to be a practical, distribution free approach to testing for equal multivariate distributions, and our empirical results show that the empirical performance of $\mathcal{E}_n$ compares favorably with the nearest neighbor tests. Our results suggest that $\mathcal{E}_n$ may be one of the most powerful tests available for high dimensional data.

Table 1: Significant tests (nearest whole percent) at $\alpha = 0.1$ of bivariate normal location alternatives $F_1 = N_2((0,0), I)$, $F_2 = N_2((0,\delta), I)$.

		$\delta = 0$		$\delta = .5$		$\delta = .75$		$\delta = 1$	
n_1	n_2	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$
10	10	10	12	23	19	40	29	58	42
15	15	9	11	30	21	53	34	75	52
20	20	10	12	37	23	64	38	86	58
25	25	10	11	43	25	73	42	93	65
30	30	10	11	48	25	81	47	96	70
40	40	11	10	59	28	90	52	99	78
50	50	10	11	69	29	95	58	100	82
75	75	10	11	85	37	99	69	100	93
100	100	10	10	92	40	100	79	100	100

Table 2: Significant tests (nearest whole percent) of $H_0 : F_1 = F_2$ at $\alpha = .1$, where $F_1 = N_d(0, I)$; $F_2 = t(5)^{(d)}$.

n_1, n_2	50		75		100	
d	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$
1	12	13	14	14	15	15
2	15	15	18	18	23	20
5	24	18	34	23	46	26
10	40	21	61	28	78	34
20	68	27	90	35	98	44
50	97	27	100	36	100	46

Table 3: Significant tests (nearest whole percent) of $H_0 : F_1 = F_2$ at $\alpha = .1$, where $F_1 = \text{Uniform}(0, 1)^{(d)}$; $F_2 = \text{Uniform}(0, .9)^{(d)}$.

n_1, n_2	50		75		100	
d	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$	$\mathcal{E}_n$	$T_{n,3}$
1	22	25	29	30	36	35
2	29	24	39	33	48	41
5	44	24	61	32	74	37
10	64	31	83	39	93	46
20	84	43	97	54	100	64
50	99	62	100	77	100	87

Table 4: Tests comparing samples of NSCLC gene expression data ($d = 1962$) with other types of cancer samples.

Type	n_1	Type	n_2	$\mathcal{E}_n$	p-value	$T_{n,3}$	p-value
NSCLC	9	Breast	7	61.711	0.027	0.646	0.004
NSCLC	9	CNS	5	55.860	0.005	0.762	0.000
NSCLC	9	Colon	7	78.075	0.000	0.729	0.000
NSCLC	9	Leukemia	6	91.413	0.000	0.821	0.000
NSCLC	9	Melanoma	8	85.996	0.000	0.882	0.001
NSCLC	9	Ovarian	6	44.880	0.131	0.533	0.197
NSCLC	9	Renal	9	54.769	0.001	0.685	0.001

APPENDIX

We restate the following fundamental inequality proved by Székely and Rizzo (2005) (see also Székely 2000a for a more general result and Székely 1996 for an interesting special case).

Theorem 1. *If X, X', Y, Y' are independent random vectors in R^d with finite expectations, $X \overset{\mathcal{D}}{=} X'$, and $Y \overset{\mathcal{D}}{=} Y'$, then*

$$2E\|X - Y\| - E\|X - X'\| - E\|Y - Y'\| \geq 0, \tag{8}$$

and equality holds if and only if X and Y are identically distributed.

Proofs of the existence of the limiting distribution under H_0 of the two-sample statistic $\mathcal{E}_n$ and consistency results follow.

Suppose $X_1, \ldots, X_{n_1}$ are iid R^d valued random vectors, and $Y_1, \ldots, Y_{n_2}$ are iid R^d valued random vectors independent of $X_1, \ldots, X_{n_1}$, $d \geq 1$. Let $h(x_i, x_j; y_\ell, y_m)$ be a real valued function that is symmetric within each argument (x_i, x_j) and (y_ℓ, y_m). The generalized U-statistic corresponding to the kernel h is obtained by averaging over all pairs of distinct combinations (x_i, x_j), (y_ℓ, y_m), $1 \leq i < j \leq n_1$, $1 \leq \ell < m \leq n_2$. The associated V-statistic with kernel h is

$$V_{n_1,n_2} = n_1^{-2} n_2^{-2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \sum_{\ell=1}^{n_2} \sum_{m=1}^{n_2} h(x_i, x_j; y_\ell, y_m). \tag{9}$$

Our proposed test statistic for the two-sample problem is based on a V-statistic with kernel

$$h(x_i, x_j; y_\ell, y_m) = \|x_i - y_\ell\| + \|x_j - y_m\| - \|x_i - x_j\| - \|y_\ell - y_m\|. \tag{10}$$

If X and Y are identically distributed, $Eh(X_1, X_2; Y_1, Y_2) = 0$ follows from inequality (8). Then since

$$g(x, y) = Eh(x, X_1; y, Y_1) = 0$$

for almost all (x, y), V_{n_1,n_2} is a degenerate kernel V-statistic. Asymptotic properties of a k-sample V-statistic with a given kernel h parallel those of the U-statistic with the same kernel. The asymptotic results for V-statistics can be inferred from the corresponding U-statistics results of Hoeffding (1948) for the case $k = 1$. If U_n is a degenerate U-statistic then nU_n has a non-degenerate limiting distribution

$$\sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1), \tag{11}$$

provided $Eh^2 < \infty$, where the constants λ_i are the eigenvalues associated with the kernel h, and Z_i^2 are independent $\chi^2(1)$ random variables.

With the necessary moment conditions, essentially the same proof applies to show that the corresponding V-statistic V_n with degenerate kernel h satisfies

$$nV_n \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i Z_i^2. \tag{12}$$

Following similar reasoning for the two-sample degenerate case

$$\mathcal{E}_{n_1,n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} V_{n_1,n_2} \tag{13}$$

has a proper (non-degenerate) limiting distribution, provided $Eh^2 < \infty$. For an explicit proof of the weak convergence of (13) under $X \overset{\mathcal{D}}{=} Y$, we apply Theorem 1.1 of Neuhaus (1977). Also see Theorems 4.5.3 and 5.6.1 in Koroljuk and Borovskich (1994).

Theorem 2 (Consistency)**.** *Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent random samples of R^d valued random vectors, $d \geq 1$, with respective distribution functions F_1 and F_2 and sample size $n_1+n_2 = n$. Assume that $\lim_{n\to\infty} n_i/n = \lambda_i \in (0,1)$, $i = 1, 2$. Then $\mathcal{E}_n$ determines a test of the hypothesis of equal distributions (1) that is consistent against all fixed alternatives satisfying $E\|X^2\| < \infty$ and $E\|Y^2\| < \infty$.*

Proof. Suppose that $F_1 \neq F_2$, $E\|X^2\| < \infty$ and $E\|Y^2\| < \infty$. Then we have strict inequality in inequality (8), so $\lim_{n\to\infty} V_{n_1,n_2} = c > 0$ with probability one. By the theory of degenerate two-sample V-statistics, under the null hypothesis there exists a constant c_α satisfying $\lim_{n\to\infty} P[\mathcal{E}_n > c_\alpha] = \alpha$, and the size α test rejects the null hypothesis for $\mathcal{E}_n > c_\alpha$. Under the alternative hypothesis

$$\lim_{n\to\infty} P(\mathcal{E}_n > c_\alpha) = \lim_{n\to\infty} P((n_1 n_2/n)\, V_{n_1,n_2} > c_\alpha) = 1.$$

□

By the corresponding theory of generalized V-statistics, distributional results above can be generalized to the k-sample problem.

For the $\mathcal{E}$ permutation test of the k-sample hypothesis, we have consistency under an even more general class of alternatives, those with finite first moments.

Theorem 3 (Consistency of permutation $\mathcal{E}_n$ test)**.** *Suppose $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ are $k \geq 2$ independent random samples in R^d, sizes $n_1, \ldots, n_k$ respectively, where $n = n_1 + \cdots + n_k$ and $\lim_{n\to\infty} n_i/n = \lambda_i \in (0,1)$, $i = 1, \ldots, k$. Then the exact permutation test based on the k-sample test statistic*

$$\mathcal{E}_n = \sum_{1\leq i<j\leq k} \mathcal{E}_{n_i,n_j}(\mathcal{A}_i, \mathcal{A}_j) = \sum_{1\leq i<j\leq k} e(\mathcal{A}_i, \mathcal{A}_j)$$

is consistent against all fixed alternatives with finite first moments.

Proof. Let $\lambda_{in} = n_i/n,\ i = 1, \ldots, k$, and let $\mu_{ij} := E\|X_i - X_j\|$, where X_i denotes the random variable corresponding to the i^{th} sample. Under the hypothesis of equal distributions we have $E[\mathcal{E}_n] = (k(k-1)/2)\mu_{11}$. Under any alternative $F_1, \ldots, F_k$ with finite first moments,

$$E[\mathcal{E}_n] = \sum_{1 \le i < j \le k} \left[\frac{n_i n_j}{n_i + n_j}(2\mu_{ij} - \mu_{ii} - \mu_{jj}) + \frac{n_j \mu_{ii} + n_i \mu_{jj}}{n_i + n_j} \right],$$

and conditional on the observed sample $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2 \ldots \mathcal{A}_k\}$,

$$E[\mathcal{E}_n|\mathcal{A}] = \sum_{1 \le i < j \le k} \frac{n_i n_j}{n_i + n_j}(2\mu_{ij} - \mu_{ii} - \mu_{jj}) = \mathcal{E}_n.$$

(The extra term in $E[\mathcal{E}_n]$ arises because of the bias in using the V-statistic to estimate $2E\|X - Y\| - E\|X - X'\| - E\|Y - Y'\|$.) Then by inequality (8) and the fact that $\lim_{n\to\infty}(\lambda_{in}\lambda_{jn})/(\lambda_{in} + \lambda_{jn}) = \lambda_i\lambda_j/(\lambda_i + \lambda_j) \in (0, \infty)$,

$$\lim_{n\to\infty} E[\mathcal{E}_n|\mathcal{A}] = \lim_{n\to\infty} \sum_{1 \le i < j \le k} \frac{n\ \lambda_{in}\lambda_{jn}}{\lambda_{in} + \lambda_{jn}}(2\mu_{ij} - \mu_{ii} - \mu_{jj})$$

exists if and only if $F_i = F_j$ for all $1 \le i < j \le k$. Otherwise $\lim_{n\to\infty} P[\mathcal{E}_n > c_\alpha] = \lim_{n\to\infty} P[\mathcal{E}_n > c_\alpha|\mathcal{A}] = 1$. □

REFERENCES

Bickel, P. J. (1969). A distribution free version of the Smirnov two-sample test in the multivariate case. *Annals of Mathematical Statistics* **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. *Annals of Probability* **11** 185–214.

Efron, B. (1993). An introduction to the bootstrap. *Chapman and Hall*, New York.

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

Hastie, T. Tibshirani, R., And Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

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

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **49** 293–325.

Janson, S. (1984). The asymptotic distributions of incomplete U-statistics. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **66** 495–505.

Koroljuk, V. S. and Borovskich, Y. V. (1994). Theory of U-statistics. Kluwer, Dordrecht.

Neuhaus, G. (1977). Functional limit theorems for U-statistics in the degenerate case. *Journal of Multivariate Analysis* **7** (3), 424–439.

Rizzo, M. L. (2002). A new rotation invariant goodness–of–fit test. Ph.D. dissertation, Bowling Green State University.

Ross, D.T., Sherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van De Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., And Brown, P.O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics, 24* 227–235.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81** 799–806.

Székely, G. J. (1996). *Contests in Higher Mathematics*. Springer, New York.

Székely, G. J. (2000a). Technical Report 03-05: $\mathcal{E}$-statistics: energy of statistical samples, *Department of Mathematics and Statistics, Bowling Green State University*.

Székely, G. J. (2000b). Pre-limit and post-limit theorems for statistics. In *Statistics for the 21st Century* (C. R. Rao and G. J. Székely. Eds.), pp. 411–422. Marcel Dekker, New York.

Székely, G. J. and Móri, T. F. (2001). A characteristic measure of asymmetry and its application for testing diagonal symmetry. *Communications in Statistics – Theory and Methods* **30** (8 & 9) 1633–1639.

Székely, G. J. and Rizzo, M. L. (2005). A new test of multivariate normality. *Journal of Multivariate Analysis* **93** 58–80.

Székely, G. J. and Rizzo, M. L. (2003b). Hierarchical clustering via joint between-within distances, submitted.

Székely, G. J. and Rizzo, M. L. (2004). Mean distance test of Poisson distribution. *Statistics and Probability Letters*, **67**/3 241–247.