SLAC Computation Group
Stanford, California

CGTM No. 153
November 1974

A NONPARAMETRIC PROCEDURE FOR COMPARING MULTIVARIATE POINT SETS

Jerome H. Friedman and Sam Steppel

ABSTRACT

An algorithm that tests the hypothesis that two multivariate point samples were drawn from the same unknown probability density, is described. The algorithm provides a measure of the discrepancy between the two point distributions, as well as providing information concerning the regions of the multivariate space where the discrepancies occur. A procedure for calculating the significance level of the test in any given situation is presented. On multivariate normal data, this nonparametric test is found to be considerably more powerful than the normal theory likelihood ration test when the two samples differ primarily in scale.

Introduction

Consider two samples of N_1 and N_2 observations taken on vector random variables $\vec{x}$ and $\vec{y}$ with unknown probability density functions $f(\vec{x})$ and $g(\vec{y})$. We wish to test the null hypothesis, H_o, that $f(\vec{x}) = g(\vec{y})$ for all $\vec{x}$ and $\vec{y}$.

It is often useful in data analysis problems to compare two multivariate point sets to determine to what extent they are similar or different. At the most straightforward level, two experiments can be compared for their compatibility. Since the test is able to make the comparison in the full dimensionality of the data measurement space, all of the information contained in the experiments is used.

More generally, one usually wishes to determine if changing a property, y, of the data has any effect or consequences on the resulting joint probability density function, $p(\vec{x})$, of the experimental measurements, $\vec{x}$. That is, whether these experimental measurables can be used in any way as predictors for y. Although it may be impossible to formulate a specific model for the dependence (as one does with regression), it may still be important to know if there exists any information in the aggregate of measured variables, $\vec{x}$, having a bearing on the response variable, y.

This can be accomplished by partitioning the data set on the value of the dependent variable, y, and treating the resulting data sets as independent samples. These samples can then be compared for compatiblilty in the full dimensionality of the data space. If these multivariate point sets are shown to be compatible, then there exists no relationship between the measurement variables, $\vec{x}$, and the response or dependent variable, y.

If, on the contrary, a relationship is shown to exist, then it is often useful to be able to determine those simultaneous values of the measurement variables where the response is strongest. That is, those regions of the

multivariate data space where data samples most disagree in their relative densities. This information can often give considerable insight as to the nature of the dependence of y on the measurement variables, $\vec{x}$, as well as allowing the researcher to maximize the predictive power of his observations.

As an example, consider discrimination analysis in pattern classification. Here the dependent variable, y, is an indicator of the correct classification of a pattern, and the measurement variables, $\vec{x}$, are a chosen set of pattern features. Different patterns can be compared in the feature space to determine the discrimination ability of the chosen set of features. Those regions of the feature space that yield the lowest error rate for each pattern can be identified. The relative merits of various different feature spaces can also be evaluated.

For another example, consider medical diagnosis. Here the data is partitioned into two samples on the basis of incidence of a disease. In this case, the dependent variable, y, is binary and simply indicates the presence or absence of the disease. These two samples are then compared in the data space, $\vec{x}$, of all clinical measurements performed on the subjects. If the two samples are shown to be compatible, then that particular set of clinical observations cannot be used to diagnose the disease. If the samples are different, then the set of simultaneous values of the measurements that are most likely to indicate the presence (or absence) of the disease can be identified. This information can be used to diagnose future patients.

When $f(\vec{x})$ and $g(\vec{y})$ are both multivariate normal, the point distributions can differ only in their location vectors and dispersion matrices. For non-normal distributions the data sets can differ in a variety of ways. The identity of location vectors and of dispersion matrices, while necessary, is not sufficient to guarantee the identity of the two distributions. For many

applications, the underlying density functions may not be known, requiring more general techniques for comparing distributions. For these cases, nonparametric procedures are usually required. Nonparametric procedures do not work in terms of a set of parameters and require few assumptions about the underlying multivariate distributions. A desirable property of these procedures is that they be at least roughly distribution-free in the sense that they perform similarly (at least as far as the probability of falsely finding a significant distinction) for differing underlying multivariate distributions.[1] In the multivariate case, nonparametric procedures often have been difficult to implement and have proved to be far from distribution-free.

This report describes a nonparametric procedure for comparing two multidimensional point distributions that is straightforward to implement and whose test statistic distribution is reasonably independent of the underlying density function. In addition, a permutation technique is described for estimating the significance level of the test in any given application. This procedure not only gives a measure of the compatibility of the two-point distributions, but also gives information as to those regions of the multidimensional data space where the correspondence between the point sets is good (in terms of their relative densities), and where they are most discrepant.

The finite sample statistical properties of this test are studied by applying it to a wide variety of simulated data. These simulations show that when applied to multivariate normal data this nonparametric test has comparable power to the normal theory likelihood ratio test for location differences and is considerably more powerful for differences in scale.

Comparison Procedure

The algorithm for testing the null hypothesis, H_o, that two multivariate point samples (classes) were drawn from the same unknown probability density function (p.d.f.), proceeds as follows. The two samples of size N_1 and N_2 respectively, are combined into a single sample of size $N = N_1 + N_2$ with each point tagged as to the class from which it originates. The closest k points to each point are examined and the number, k_1, originating from class one (or the corresponding number, $k_2 = k-k_1$, originating from class two), is determined. Thus, associated with each point in this combined sample is a measure of the composition of the points closest to it. The observed frequency distribution of k_1, $n(k_1)$, for all the sample points, is recorded. This frequency distribution is then compared to that expected under the null hypothesis.

There are a variety of ways of testing whether the observed distribution for k_1 conforms to that expected under H_o. One technique involves comparing the frequency distribution of k_1, $n_1(k_1)$, evaluated in the neighborhoods centered at class one points to the frequency distribution of k_1, $n_2(k_1)$, evaluated in the neighborhoods centered at class two points. Under H_o these two distributions are expected to be the same. Asymptotically, differing multivariate p.d.f.'s for the two classes results in higher concentrations of class one points in near neighborhoods of class one centers and, similarly, higher concentrations of class two points in near neighborhoods of class two centers. This causes $n_1(k_1)$ to be shifted towards higher values of k_1 while $n_2(k_1)$ will be shifted toward lower values, from their expected distributions, $n_o(k_1)$, under H_o. This leads naturally to a t-statistic of the form

$$t = \left[\bar{k}_1(1) - \bar{k}_1(2) \right] \Big/ \sqrt{\frac{V_1(1)}{N_1} + \frac{V_1(2)}{N_2}} \quad , \tag{1}$$

where

$$\bar{k}_1(1) = \frac{1}{N_1} \sum_{k_1=0}^{k} k_1 n_1(k_1)$$

$$\bar{k}_1(2) = \frac{1}{N_2} \sum_{k_1=0}^{k} k_1 n_2(k_1)$$

$$V_1(1) = \frac{1}{N_1} \sum_{k_1=0}^{k} [k_1 - \bar{k}_1(1)]^2 \, n_1(k_1)$$

$$V_1(2) = \frac{1}{N_2} \sum_{k_1=0}^{k} [k_1 - \bar{k}_1(2)]^2 \, n_2(k_1) \quad ,$$

as a candidate for a test statistic comparing the two distributions. Alternatively, any one of the many well known statistical techniques for directly testing the compatibility of two univariate distributions can be used.

Although, asymptotically, this procedure may be totally optimal, for finite sample sizes it works well only when the two classes differ primarily in their locations. The test is relatively insensitive to differences in scale (extent) between the two multivariate point samples. This is caused by the tendency of near neighboring to choose points preferentially in regions of higher density. Even points in the relatively sparse regions of the space will tend to choose nearest neighbors preferentially in dense regions. Only those sparse points that lie very far from a dense region will have mostly sparse points as nearest neighbors. For the case where the two point samples differ mainly in scale, the class with smaller scale (for example, class one) will populate the space mainly in regions of high density, while the points

from the other class (class two) will lie principally in regions of lower density. The class one points will clearly have an excess of class one neighbors. However, due to the tendency of near neighboring to choose points from regions of high density, the class two points will also have an excess of class one neighbors. In fact, for small to moderate differences in scale between the multivariate distributions, there is very little difference between $n_1(k_1)$ and $n_2(k_1)$ even though they both deviate from the expected null distribution, $n_o(k_1)$. This deficiency can be overcome by directly comparing the detailed distributions of $n_1(k_1)$ and $n_2(k_1)$ to their expected distribution, $n_0(k_1)$, under H_o. The problem of comparing two multivariate point distributions is, in this way, reduced to a univariate goodness-of-fit test.

If each of the N k-neighborhoods were mutually exclusive, then the relative frequency of the possible values of k_1 would (under the null hypothesis) conform to a binomial distribution over 0, 1, 2, ..., k with probability $p = N_1/N$; that is, $n_o(k_1)$ would be a binomial distribution with k-degrees of freedom. These neighborhoods cannot be mutually exclusive,

however, since there are N neighborhoods - each containing k points - with only N total sample points. Thus, there is no reason to expect the distribution of k_1 values to be compatible with such a binomial distribution. The precise distribution in the general case is difficult to derive, but Monte Carlo calculations for a wide variety of cases indicate very little discrepancy between the true distribution and a binomial. Thus, a difference between the two multivariate samples can be measured by comparing the distribution observed for the k_1 values with the corresponding binomial distribution. Any one of the many well known univariate goodness-of-fit tests may be used for this purpose.

For the experiments described in this report, the two k_1-distributions, $n_1(k_1)$ and $n_2(k_1)$ were summed to a single distribution, $n(k_1) = n_1(k_1) + n_2(k_1)$, and the resulting sum compared to that predicted by the null hypothesis, $n_o(k_1)$. Counting each sample point itself as one of its k-nearest neighbors and applying binomial statistics, one obtains

$$n_o(k_1) = N_1 \binom{k-1}{k_1-1} \frac{(N_1-1)^{k_1-1} N_2^{k-k_1}}{(N-1)^{k-1}} + N_2 \binom{k-1}{k_1} \frac{N_1^{k_1} (N_2-1)^{k-k_1-1}}{(N-1)^{k-1}} \quad (0 < k_1 \leq k) \tag{2}$$

with the conventions $\binom{k-1}{-1} \equiv 0$ and $\binom{k-1}{k} \equiv 0$. The mean value of k_1 for this distribution is

$$\bar{k}_1 = \frac{N_1}{N} k$$

as expected for a binomial process.

The observed number of counts, $n(k_1)$, for each value of k_1 is then compared to $n_o(k_1)$. This comparison can be made using any goodness of fit test statistic. For the experiments described in this report, a statistic analogous to Pearson's χ^2 test statistic

$$T = \sum_{k_1=0}^{k} [n(k_1) - n_o(k_1)]^2 / n_o(k_1) \qquad (3)$$

was used.[2]

In order for this algorithm to be useful, the significance level (P-value), $\alpha(T)$, for the experimentally obtained value of the test statistic, T, must be determined. This can be accomplished by employing a permutation procedure to estimate the significance level of the test for each application, directly from the data. This permutation test proceeds as follows: the two samples (classes) are combined into a single sample of size $N = N_1+N_2$, as described above. But instead of assigning each sample point to the class from which it originated, it is randomly assigned to one of the two classes. These random assignments are made subject to the constraint that the assignments preserve the original proportion N_1/N_2. The comparison procedure is then applied to these two newly defined samples and a test statistic value (eqn 3) is obtained. The points are then given another such random assignment and the test re-applied, obtaining another permuted test statistic value. Repeated application of this random permutation procedure yields a series of test statistic values that closely approximate the null test statistic distribution for the given problem. In particular, the fraction of these permuted test statistic values that are larger than the value, T, obtained for the unpermutated case, is an estimate of the significance level, $\alpha(T)$, for the test.

This comparison procedure can also be used to identify those regions of the multidimensional space where the two point samples most disagree (or agree) in their relative densities. This is because the algorithm assigns such an estimate to each point in the combined sample. Those points for which k_1 is near its expected mean value, $\bar{k}_1 = (N_1/N)k$ are located in regions where the agreement is good, while those points for which k_1 is far from this expected mean value are located in regions of the multidimensional space where the agreement is bad. One can use the experimentally determined values of k_1 for each point to select and isolate those points that give rise to the most disagreement (or agreement) between the two point sets. The coordinates of these selected points can then be examined to determine where these points lie in the multidimensional space.

Statistical Properties

In order to investigate the variation of the null test statistic cumulative distribution with sample size, dimensionality, and underlying p.d.f., a large number of simulations were performed.[3] Tables 1 - 3 show the cumulative distribution of the test statistic, T, (eqn 3) for a wide variety of simulated data samples. Each of these samples were randomly divided into two subsamples of equal size, and these halves were then compared. (For all comparisons, except in Table 7, the number of near neighbors used was k=20.) The T-distributions were obtained by calculating the value of T (eqn 3) for 100 independent samples for each situation. The corresponding distributions in parentheses are for 100 repeated random permutations of a single sample. In all cases, the distributions obtained from the random permutation procedure are not significantly different (to within the simulation accuracy) from those obtained from the independently drawn samples.

Table 1 shows that the T-distribution seems to have very little dependence on sample size. On the other hand, Table 2 shows that the T-distribution exhibits a marked dependence on the dimensionality, d, of the data for very low dimensionality. This dependence rapidly diminishes with increasing dimensionality. Table 3 shows the T-distribution for several different underlying multivariate p.d.f's, which are simply the d-fold products of the corresponding univariate p.d.f's. These results indicate that there is some dependence on the underlying p.d.f., but at least for those cases considered, this dependence is rather small.

These tables are not meant to provide an exhaustive summary of the statistical properties of the test. They do, however, give an indication that the null T-distribution remains remarkably similar in a wide variety of situations. The algorithm can be applied to any situation, of course, since the permutation procedure can be employed to determine the significance level of the test directly from the data for any particular case.

The discussion, so far, has centered on determining the significance level under the null hypothesis. For a goodness-of-fit test this is all that can be determined. For the test to be useful, however, it must be able to discriminate against hypotheses alternate to H_o. The class of all alternate hypotheses is sufficiently general for some alternate hypothesis to give an arbitrarily good comparison. Thus this test (like any goodness-of-fit test) cannot rule out all alternate hypotheses. It is, however, instructive to consider various classes of alternate hypotheses that continuously approach H_o and to study the ability of the test to discriminate between each of them and H_o.

For this purpose the test statistic distributions for various alternate hypotheses were compared with the corresponding null distribution. This is summarized by estimating the significance level from the null distribution

of the estimated 50% power points (medians) of the alternate test statistic distributions. (See Appendix for details of these calculations.) Table 4a shows the results when a sample drawn from a 10-dimensional normal, located at the origin with unit dispersion matrix, is compared to another similarly drawn sample whose location is systematically displaced a distance Δ along one axis. ($\Delta = 0$ represents the null hypothesis.) Shown in Table 4a is the median value estimated for the alternate T-distribution and its corresponding significance level from the null distribution. Also shown for the same value of Δ is the corresponding quantity for the likelihood ratio test assuming normally distributed data.[4] Asymptotically, this likelihood ratio test is the most powerful parametric test. Table 4b similarly compares two multivariate normal point samples, but where their locations are the same, and the scale of one is systematically varied along each axis by the same factor, σ. ($\sigma = 1$ represents the null hypothesis.)

As one might expect, the median T-value for the alternate hypothesis increases with its deviation from the null hypothesis, resulting in a decrease in its corresponding significance level (under the null hypothesis). Also, as expected, this effect is enhanced for larger sample size. When the alternate hypothesis represents a difference in location, this nonparametric test is seen to have somewhat less power than the normal theory test on the normally distributed data (as is usual for nonparametric procedures). For differences in scale between the two multivariate distributions, however, this nonparametric test is seen to be considerably more powerful than the normal theory test on the normal data. The nonparametric test has 50% power with 9% acceptance at $\sigma = 1.07$, while the corresponding point for the normal theory test is around $\sigma = 1.15$. This high relative efficiency for differences in scale of the nonparametric test is a property of the finite size of samples

(since asymptotically the likelihood ratio test is uniformly most powerful) and is related to the tendency (described above) for the near-neighboring algorithm to choose nearest neighbors, preferentially, in regions of relatively high density. This causes the difference between the observed k_1 frequency distributions, $n_1(k_1)$ and $n_2(k_1)$, and the null distribution, $n_o(k_1)$, to be increased, resulting in higher values of the test statistic, T.

Table 5 shows the results of comparing a point sample drawn from this same multivariate normal p.d.f. to another sample drawn from a p.d.f. whose location and scale are the same as the normal, but where its shape systematically approaches the multivariate normal. This second p.d.f. is a "multivariate" Student's t-distribution where the number of degrees of freedom, n, is systematically increased. This "multivariate" Student's distribution is formed by simply taking the product of ten one-dimensional distributions with zero location and unit scale. As n approaches infinity, this distribution approaches a multivariate normal. For finite n it has larger kurtosis along each axis and departs from spherical symmetry. For the extreme case (n = 1), it becomes a product Cauchy distribution. Shown in Table 5 for various values of n, are the kurtosis along each axis, a measure of the spherical symmetry of the multivariate Student's t-distribution, and the value of the test statistic obtained in its comparison with the multivariate unit normal. The spherical symmetry measure is the ratio of the radial variance along a line at 45° to all axes, to the variance along the axes. Since the radial variance is minimal along such a diagonal and maximal along the axes, this variance ratio is most sensitive to departure from spherical symmetry. Values of zero for kurtosis and one for the variance ratio correspond to the multivariate normal p.d.f.

The results in Table 5 show that the test is reasonably sensitive to small departures in shape between the two multivariate distributions under comparison.

In the above simulations, the samples under consideration were drawn from radially symmetric p.d.f.'s. Table 6 compares samples that are drawn from the radially skew "multivariate" log-normal,

$$p(\vec{x}) = \frac{1}{(2\pi)^5 \sigma^{10}} \prod_{i=1}^{10} \frac{1}{x_i} \exp\left[-(\log x_i - \mu)^2 / 2\sigma^2\right], \qquad x_i > 0.$$

For the null hypothesis, $\mu=0$ and $\sigma=1$ for both samples. In Table 6a, the log-location along the first axis of one of the samples is systematically displaced by a distance Δ, while in Table 6b the log-scale, σ, is varied along all axes.

Table 7 compares the performance of the test for differing number, k, of near neighbors. Here the results of Table 4 (k=20) are compared to the results of the test for k=5 and k=10. It is seen that the performance of the test tends to be reduced with smaller values of k but the effect is not dramatic.

These experiments, while not exhaustive, do indicate that the comparison procedure performs reasonably well in determining whether the two multivariate point distributions were drawn from the same underlying p.d.f. The test appears to be sensitive to differences of location, scale, and shape between the two distributions. The performance of the test seems to have only a moderate dependence on the choice of the number of near neighbors. To the accuracy of the simulations, the test appears to be unbiased for the cases tested. The test is also very robust against extreme outliers. Since the procedure involves only ordering and counting and does not use the distances directly, it is easy to see that a few extreme outliers will have little effect on the results.

Measurement Variables Scaling and Metric

This procedure leaves to the researcher's discretion the choice of the coordinate variables and metric, as well as the number of nearest neighbors, k. As shown in Table 7, this procedure is reasonably insensitive to the choice of k, provided that it is not too small. In order for the test to be consistent, k should be a function of the total sample size such that[5]

$$\lim_{N\to\infty} k(N) = \infty \ , \quad \text{and} \quad \lim_{N\to\infty} \frac{k(N)}{N} = 0.$$

Experimentation has shown that the choice of k is not important so long as k > 10-20. Clearly, k should be small compared to the total sample size, N.

This algorithm is somewhat more sensitive to choice of measurement variables and metric. Unfortunately, there are no good guidelines for their choice. For very large sample size, the algorithm is clearly invariant to changes in coordinate variables and metric since these changes simply alter the shape of the volume element containing the evaluation point. Since these volumes are infinitesimally small, their shape doesn't matter.

For finite sample sizes, however, the shape does matter. Changes in the volume shape that result in changes of the identities of the nearest neighbors can have an effect on the performance of the algorithm. Fukunaga and Hostetler[6] show that for those data distributions that can be made spherically symmetric by a linear transformation, the optimum metric is the inverse covariance matrix of the underlying distribution, $p(\vec{x})$. If this covariance matrix is estimated by the data sample covariance matrix, then this is equivalent to scaling each of the coordinates so that they have equal variance along the principal axes of the data (sphericalizing the data).

If there is no a priori information concerning the data, then this is probably the best procedure. Another reasonable procedure is to simply scale

the data to have equal variance along the original measurement coordinates. On the other hand, different experimental measurement accuracy or different characteristic length of density variation can dictate unequal scales among the various coordinates. Changing the scale of a coordinate changes its relative importance in determining the goodness-of-fit. Thus, if the researcher has information as to which coordinates are most important, they should be given larger scales.

The number and specific choice of coordinate variables also affect the performance of this test. Increasing the number of coordinates only improves the performance when those variables contain information concerning the hypothesis under test. In fact, coordinates that do not contain such information (noise coordinates) dilute the power of the test. This is because these dimensions add statistical uncertainty to the k_1 estimates without providing information helpful to the comparison. Even a coordinate that does contain some additional information may not help because the increase in statistical variance that it introduces hurts more than the information increase helps. The precision of this test can be increased greatly if the researcher's knowledge and intuition lead him to a judicious choice of coordinate variables.

After the specific measurement variables and their scales are chosen, there is still the choice of the metric or multidimensional dissimilarity measure. That is, given a one-dimensional dissimilarity measure for each coordinate, how are these combined to define a distance in the multidimensional space. One-dimensional distance is nearly always defined as the absolute value of the difference in coordinate values

$$d_{mn}^{(i)} = \left| X_m^{(i)} - X_n^{(i)} \right| ,$$

where the subscripts label the points and the superscripts label the coordinate. The most commonly used multidimensional distance measures are the Minkowski p-metrics,

$$d_p(\vec{x}_m,\vec{x}_n) = \left[\sum_{i=1}^{d} |x_m^{(i)} - x_n^{(i)}|^p\right]^{\frac{1}{p}} .$$

Of these, the three most popular are the

$p = 1$ (city block or taxi cab distance)

$p = 2$ (Euclidean distance measure)

$p = \infty$ (maximum axis distance measure)

i.e., $$d_\infty(\vec{x}_m,\vec{x}_n) = \max_{1\le i\le d}\{|x_m^{(i)} - x_n^{(i)}|\} .$$

The advantage of the $p = 1$ measure is that it can be calculated rapidly with no multiplications. It is also relatively robust against changes in relative axis scales. The advantage of the $p = 2$ distance measure is that it is the natural extension of the familiar Euclidean distance measure used in plane and solid geometry. It is somewhat less robust to changes in axis scales than the $p = 1$ metric. Like the $p = 1$ metric, the $p = \infty$ metric can be calculated rapidly with no multiplications, and it has the additional advantage that the nearest neighbors can be found much more efficiently for this distance measure (especially for high dimensionality) than with the other p-metrics.[7,8] This metric has the greatest sensitivity to the relative coordinate scales.

The best choice for a distance measure depends upon the problem at hand, and is related to the underlying density distribution of the multidimensional data. As for the choice of variables and scales, there are no good guidelines. It can be shown that the $p = 2$ metric is optimal when the data points are multivariate normal. However, there are no general results for other types of distributions.

Although the choice of measurement variables, scales and metric is arbitrary, they can have an effect on the performance of the test. There is usually an optimum choice for each particular problem. Fortunately, one can use the algorithm itself to determine the effect of differing choices and even search for the best one. The test can be applied with various differing choices to observe the extent to which the significance level, $\alpha(T)$, changes. If there is little change, then the specific choice is probably not important while if there is a great deal of change, then the particular set that minimizes the significance level is a good candidate for the best choice. In particular, by choosing various sets of measurement variables, one can search for the best feature space for pattern classification.

Computational Requirements

Almost the entire computational cost of employing this procedure is in the calculations required for finding the k-nearest neighbors to each sample point. The most straightforward procedure is to simply calculate the distance from each point to all of the other points and identify the k-smallest. This is known as the brute force (BF) method and requires computation proportional to the dimensionality, d, and proportional to the square of the sample size,

$$C_{BF}(d,k,N) = \alpha_{BF}(k)dN^2 .$$

Recently, several new algorithms for finding nearest neighbors have been reported[7,8] that are faster than the brute force method for sufficiently large N and small d. The method of Friedman, Baskett, and Shustek (FBS)[7] finds the k-nearest neighbors with computation

$$C_{FBS}(d,k,N) = \alpha_{FBS}{}^d \left[\frac{kd\Gamma\left(\frac{d}{2}\right)}{2}\right]^{\frac{1}{d}} N^{2-\frac{1}{d}}$$

while the method of Bentley (B)[8] finds them with computation

$$C_B(d,k,N) \simeq \alpha_B(k)2^d \, N\log_2 N \quad .$$

The advantage of the BF-method is it's low overhead requirements in both computation per distance calculation and additional memory. The FBS method introduces a small amount of computational overhead and a substantial increase in memory requirements (depending upon the optimization level used with the algorithm). The B-algorithm introduces a large computational overhead per distance calculation and requires an amount of additional memory comparable to the FBS method.

The best algorithm to use depends upon the combined sample size, $N = N_1+N_2$, the dimensionality, d, to a smaller extent the number of near neighbors, k, and on the amount of memory available. Simulations have shown (see reference 9) that for dimensionalities less than ten and combined sample sizes less than 3000 to 5000, the FBS algorithm performs the best. For the same range of dimensionalities but larger sample sizes, the B algorithm is fastest. For high dimensionalities and small sample sizes, a brute force method that is highly tailored to a particular computer is most effective.

For very large sample sizes, one might consider an "egg crating" technique. That is, the multivariate space is partitioned into several cells and the comparison performed separately on the sample points in each cell. The results of the several comparisons can then be statistically combined to yield a significance level for the test over the entire multivariate space.

Acknowledgment

Helpful discussions with William H. Rogers are gratefully acknowledged.

FOOTNOTES AND REFERENCES

1. The word "nonparametric" is often used to mean "distribution free." It will not be so used here.

2. Although this procedure may appear somewhat naive, our simulation experiments indicate that this test has similar power and robustness to other more sophisticated approaches. In particular, we could measure no significant difference between this approach and one involving separate comparisons of $n_o(k_1)$ to $n_1(k_1)$ and $n_2(k_1)$ individually, or employing more sophisticated test statistics for the univariate comparison. Also, this test performed as well as the t-statistic (eqn 1) when the multivariate distributions differed only in location. (For the reasons described above, it, of course, performed much better for differences in scale between the two multivariate distributions.)

3. All of these simulation results were obtained using the specific comparison procedure described in the preceding section, namely, comparing $n_1(k_1) + n_2(k_1)$ to $n_o(k_1)$ (eqn 2) using the test statistic of eqn 3. As discussed in the preceding section, this choice is somewhat arbitrary. However, additional simulations with a variety of other procedures indicate very little difference in characteristics or performance, so that the simulation results presented here are characteristic of the general properties of the test and not of the details of its implementation.

4. Anderson, T.W., An Introduction to Multivariate Statistical Analysis, New York: John Wiley and Sons, Inc., 1958. pp 250-251.

5. Loftsgaarden, D O., and Quesenberry, C.P., "A Nonparametric Estimate of a Multivariate Density Function," Ann. Math. Statist., Vol. 36, pp. 2049-1051, 1965.

6. Fukanaga, K. and Hostetler, L.D., "Optimization of k-Nearest Neighbor Density Estimates," IEEE Trans. Info. Theory, Vol. IT-19, pp. 320-326, May 1973.

7. Friedman, J.H., Baskett, F., and Shustek, L.J., "A Relatively Efficient Algorithm for Finding Nearest Neighbors," Stanford Linear Accelerator Center, Report No. SLAC-PUB-1488, June 1974. Also, Computer Repository R74-234, pp. 67, November 1974.

8. Bentley, J.L., "Multidimensional Binary Search Trees used for Associative Searching," Stanford University, Dept. of Computer Science preprint (1974).

9. Friedman, J H., "COMPAR - A Program for Comparing Multidimensional Point Sets," Stanford Linear Accelerator Center, Computation Research Group Technical Memo No. 162, October 1974.

APPENDIX

This section details the calculational procedure used to obtain the results presented in Tables 4 - 7.

For Table 4, three hundred normally distributed samples of size 200 were generated and each sample equally divided into two 100-point samples. The comparison procedure was applied to each pair yielding 300 values for the test statistic (eqn 3). These values were ordered and the median estimate was taken to be the average of the 150th and 151st values. For the alternate hypotheses, the first 100 points of each sample were either translated along the first axis a distance, Δ, (Table 4a) or were scaled in all axes by a factor, σ, (Table 4b). The second 100 points of each sample remained untranslated and unscaled. The median T-value, med(T), was estimated, as above, from these 300 two-sample comparisons. Next, the second 100 points of each sample were similarly translated or scaled while the first 100 points remained untranslated and unscaled. Another med(T) was estimated from these 300 comparisons and averaged with the previous estimate. The identical set of three hundred 200-point samples was used for all hypotheses. This minimizes the statistical uncertainty in <u>relative</u> values of med(T) due to sampling fluctuations. The significance levels were estimated by the fraction of T-values in the 300 null comparisons that were larger than med(T) for each alternate hypotheses. The notation <0.3% indicates no null T-values from the 300 were larger than med(T) for that alternate hypothesis.

The calculations for the 1000 point estimates were obtained in a similar manner by dividing twenty-four 1000-point samples each into two 500-point samples and estimating med(T) as the average of the 12th and 13th ordered value in each case. The log normal calculations (Table 6) were identical to the normal ones except that the coordinates of each point were exponentiated be-

fore performing the comparison. The calculations in Table 7 were identical to those in Table 4 except that differing numbers of near neighbors were used.

The Student's-t calculations of Table 5 were performed in a similar manner, except that n/2+1 uniform random numbers are required to obtain one random number distributed according to a Student's distribution with n-degrees of freedom. For a given value of n, the uniform random numbers generated for all smaller values of n were reused to minimize relative sampling fluctuations.

For the purpose of the relative comparison with the nonparametric procedure, the calculations for the parametric normal theory estimates (Table 4) were performed in the identical manner on the identical data as the nonparametric estimates to which they are compared.

As a consistency check on this procedure, the values for the parametric normal theory test were recalculated using 1000 (instead of 300) 200-point samples for the null as well as all alternate hypotheses. The values so obtained differed very little from those presented in Table 4, obtained using the more approximate procedure.

TABLE 1

Dependence of the percent points of the test statistic null distribution on total sample size for 10-dimensional spherical normal distributions. The corresponding values for the permutation tests are in parentheses.

Total Area	100 Points per Sample	250 Points per Sample	500 Points per Sample
75%	16 (17)	16 (16)	15 (16)
50%	22 (25)	23 (25)	21 (24)
25%	33 (35)	33 (38)	34 (36)
10%	49 (53)	51 (58)	57 (58)

TABLE 2

Dependence of the percent points of the test statistic null distribution on dimensionality of data space for spherical normal distributions with 100 points per sample. The corresponding values for the permutation tests are shown in parentheses.

Total Area	One Dimension	Five Dimensions	Ten Dimensions
75%	27 (34)	17 (15)	16 (17)
50%	42 (44)	23 (21)	22 (25)
25%	64 (64)	33 (31)	33 (35)
10%	108 (107)	43 (41)	49 (53)

TABLE 3

Dependence of the percent points of the test statistic null distribution on underlying multivariate distribution in 10-dimensions with 250 points per data sample. The corresponding values for the permutations tests are shown in parentheses.

Total Area	Normal	Uniform	Cauchy	log-normal
75%	16 (16)	15 (16)	21 (22)	17 (17)
50%	23 (25)	22 (21)	29 (30)	27 (25)
25%	33 (38)	31 (29)	42 (43)	36 (35)
10%	51 (58)	44 (41)	68 (82)	55 (52)

TABLE 4

Variation of the median (50% power point) of the test statistic distribution and corresponding null significance level, obtained in comparing a 10-dimensional spherical normal distribution with zero location and unit scale to various other 10-dimensional spherical normals. Also shown are the corresponding significance levels for the normal theory likelihood ratio (L.R.) test. (See Ref. 4)

a) Variation in Location

	100 data points per sample			500 data points per sample		
location difference	med(T)	Significance Level This Test	Significance Level L.R. Test	med(T)	Significance Level This Test	Significance Level L.R. Test
0.0	22	50%	50%	21	50%	50%
0.4	24	46%	29%	71	6%	0.6%
0.6	40	17%	7%	262	< 0.3%	< 0.3%
0.8	89	3%	1%			
1.0	234	< 0.3%	< 0.3%			

b) Variation in Scale

Scale Factor						
1.0	22	50%	50%	21	50%	50%
1.05	37	21%	44%	111	1%	16%
1.07	55	9%	35%	200	< 0.3%	4%
1.10	96	1%	25%			
1.15	231	< 0.3%	7%			

TABLE 5

Variation of test statistic value obtained in comparing a 10-dimensional spherical normal distribution to a 10-dimensional "Student's-t" distribution with the same location and scale, but varying degrees of freedom.

100 Points per Sample

Degrees of Freedom	Kurtosis Along Axes	Diagonal to Axis Variance Ratio (Sphericality)	Test Statistic Value
1 (Cauchy)	∞	0.000	4.9×10^6
2	∞	0.000	2.2×10^5
4	∞	0.426	1200
6	3.00	0.597	479
8	1.50	0.690	178
12	0.75	0.787	75
16	0.50	0.838	49
24	0.30	0.891	39
32	0.21	0.917	31

TABLE 6

Variation of the median (50% power point) of the test statistic distribution and corresponding null significance level, obtained in comparing a 10-dimensional "product log normal" distribution with zero log-location and unit log-dispersion to various other 10-dimensional "product log normal" distributions.

100 Data Points per Sample

a) Variation in log location

Log Location Difference	med(T)	Significance Level
-1.0	128	1%
-0.8	62	7%
-0.6	41	17%
-0.4	31	36%
0.0	25	50%
0.4	29	39%
0.6	48	12%
0.8	212	0.3%
1.0	1177	< 0.3%

b) Variation in log scale

Log Scale Factor	med(T)	Significance Level
1/1.15	126	1%
1/1.10	66	7%
1/1.07	43	15%
1/1.05	36	28%
1.00	25	50%
1.05	28	39%
1.07	35	29%
1.10	47	13%
1.15	91	4%

TABLE 7

Variation of the null significance level of the 50% power point obtained in comparing a 10-dimensional spherical normal distribution with zero location and unit scale to various other 10-dimensional spherical normals, for several different numbers of near neighbors k.

100 Data Points per Sample

a) Variation in Location

Location Difference	Significance Level k = 5	k = 10	k = 20
0.4	45%	41%	46%
0.6	31%	24%	17%
0.8	8%	5%	3%
1.0	1%	1%	< 0.3%

b) Variation in Scale

Scale Factor			
1.05	32%	25%	21%
1.07	18%	13%	9%
1.10	6%	4%	1%
1.15	0.7%	0.3%	< 0.3%