The Annals of Statistics
1979, Vol. 7, No. 4, 697–717

MULTIVARIATE GENERALIZATIONS OF THE WALD-WOLFOWITZ AND SMIRNOV TWO-SAMPLE TESTS[1]

BY JEROME H. FRIEDMAN AND LAWRENCE C. RAFSKY

Stanford Linear Accelerator Center and ADP Network Services

Multivariate generalizations of the Wald-Wolfowitz runs statistic and the Smirnov maximum deviation statistic for the two-sample problem are presented. They are based on the minimal spanning tree of the pooled sample points. Some null distribution results are derived and a simulation study of power is reported.

1. Introduction. The nonparametric two-sample problem has been well studied (see Gibbons, 1971, Ch. 7). Classical univariate tests designed for general alternatives are the Smirnov (1939) maximum deviation test and the Wald-Wolfowitz (1940) runs test. Straightforward extensions of the Smirnov test (and related tests, Darling (1957)) using the multivariate empirical cdf lead to procedures that are not distribution free. Bickel (1969) shows that it is possible to construct consistent distribution free multivariate Smirnov tests by conditioning on the empirical cdf of the pooled sample. Distribution free multivariate generalizations of the Wald-Wolfowitz test have not previously been proposed.

The basic operational procedure employed in both the Wald-Wolfowitz and Smirnov tests is the sorting of pooled univariate observations in ascending order without regard to sample identity. Statistics are then computed based on the ranks of sample members in this sorted list. The difficulty in extending this procedure to multivariate observations is that the notion of a sorted list cannot be immediately generalized.

We propose using the minimal spanning tree (MST) of the sample points as a multivariate generalization of the univariate sorted list. We show how one can define two-sample test statistics based on the MST in analogy with those based on the sorted list. (In fact, in one dimension, the MST is defined precisely by the sorted list.)

Section 2 presents a formal statement of the problem and reviews the Wald-Wolfowitz and Smirnov univariate tests. Section 3 introduces the MST and reviews some standard theorems that indicate its appropriateness for the two-sample problem. Section 4 presents the multivariate generalization of the Wald-Wolfowitz runs test and Section 5 presents our multivariate generalization of the Smirnov maximum deviation test. Section 6 discusses multivariate two-sample tests based on

Received March 1978; revised June 1978.

[1]Supported by Department of Energy and, in part, by Chase Manhattan Bank, New York City.

AMS 1970 *subject classifications.* Primary 62G10, 62H15; secondary 05C05, 62G30, 62H20.

Key words and phrases. Nonparametric two-sample tests, multivariate observations, runs, Wald-Wolfowitz test, Smirnov test, minimal spanning trees.

2×2 contingency tables. Section 7 presents further extensions of the runs test using orthogonal spanning trees. Section 8 shows the results of a simulation study comparing the power of these tests and other multivariate tests.

2. The problem. Consider samples of size m and n respectively from distributions F_X and F_Y, both defined on R^p. The hypothesis H_0 to be tested specifies that $F_X = F_Y$. We are interested in general alternative hypotheses $F_X \neq F_Y$.

The Wald-Wolfowitz test (for $p = 1$) begins by sorting the $N = m + n$ univariate observations in ascending order. Each observation is then replaced by a label "X" or "Y" depending upon the sample to which it originally belonged. The test statistic is the total number of *runs*, R. A run is a consecutive sequence of identical labels. Rejection of H_0 is for small values of R. The null distribution of the test statistic can be derived by a straightforward combinatorial argument. Asymptotically, the quantity

$$W = \frac{R - \frac{2mn}{N} - 1}{\left(\frac{2mn(2mn - N)}{N^2(N-1)}\right)^{\frac{1}{2}}} \tag{1}$$

has a standard normal distribution. The test is consistent if the ratio m/n is bounded away from 0 and ∞ as $m, n \to \infty$.

The Smirnov test also begins by sorting the univariate observations in ascending order irrespective of sample identity. For each i, $1 \leqslant i \leqslant N$, the quantity

$$d_i = \frac{r_i}{m} - \frac{s_i}{n} \tag{2}$$

is calculated where $r_i(s_i)$ is the number of $X(Y)$ observations for which the rank (in the sorted list) is less than or equal to i. The test statistic is

$$D = \max|d_i| \qquad 1 \leqslant i \leqslant N. \tag{3}$$

Rejection of H_0 is for large values of D. The distribution of D has been extensively tabulated and asymptotic approximations are known. This test is consistent under the same conditions as the runs test.

3. Minimal spanning trees. We begin by reviewing some terms from graph theory. A *graph* consists of a set of *nodes* and a set of node pairs called *edges*. We say that an edge *links* the two nodes defining it and that it is *incident* on both of them. The *degree* of a node is the number of edges incident on it. A *path* between two prescribed nodes is an alternating sequence of nodes and edges with the prescribed nodes as first and last elements, all other nodes distinct, and each edge linking the two nodes adjacent to it in the sequence. The *length* of a path is the number of edges it contains. A *connected graph* has a path between any two distinct nodes. A *cycle* is a path beginning and ending with the same node. A *tree* is a connected graph with no cycles. A *subgraph* of a given graph is a graph with all of its nodes and edges in the given graph. Connected subgraphs of trees are trees and

are called *subtrees*. Subgraphs having no nodes in common are called *disjoint*; subgraphs having no edges in common are called *orthogonal*. A *spanning subgraph* of a given graph is a subgraph with node set identical to the node set of the given graph. A *spanning tree* of a graph is a spanning subgraph that is a tree. Note that there is a (unique) path between every two nodes in a tree, and thus a spanning tree of a (connected) graph provides a path between every two nodes of the graph.

An *edge weighted graph* is a graph with a real number assigned to each edge. A *minimal spanning tree* (MST) of an edge weighted graph is a spanning tree for which the sum of edge weights is a minimum. (The terminology of graph theory has not been standardized; for a general discussion, see Harary (1969)).

In the two-sample problem, consider the edge weighted graph consisting of the N pooled sample data points in R^p as nodes, and edges linking all pairs. This "complete" graph has $N(N-1)/2$ edges. Take the weight associated with each edge to be Euclidean distance or a generalized dissimilarity between the nodes (points) defining it. The MST of this graph is thus the subgraph of minimum total distance (dissimilarity) that provides a path between every two nodes. Note that it is unique if there are no ties among the $N(N-1)/2$ interpoint distances (dissimilarities).

Minimal spanning trees have two important properties that make them appropriate for application to the two-sample problem: (1) they connect all of the nodes with $N-1$ edges and (2) the node pairs defining the edges represent points that tend to be close together (small distance or dissimilarity). The first property follows from the fact that the MST is a spanning tree, and the second from the requirement that the sum of the edge weights be a minimum. This is amplified by the following two theorems from Prim (1957) (see also Kruskal (1956)):

THEOREM (1). *An* MST *contains as a subgraph the "nearest neighbor graph". That is, there is an edge linking each node and the node closest to it (or one of them if there are ties).*

THEOREM (2). *If any edge of an* MST *is deleted, thereby dividing the graph into two disjoint connected subgraphs, and thus dividing the points into two disjoint subsets, the deleted edge weight is the smallest interpoint distance between the two subsets.*

MSTs are well known in pattern recognition (Zahn, 1971) and cluster analysis (Hartigan, 1975) for providing excellent descriptions of point sets. Figures 1b, 2b and 3b display MSTs for some two-dimensional point sets. Computational considerations for constructing MSTs (and the test statistics defined below) are discussed in the Appendix.

4. Multivariate number of runs test. For a univariate sample, the edges of the MST are defined by adjacent points in the sorted list. The Wald-Wolfowitz runs test described above can be alternately described as follows: (1) construct the MST of the pooled sample (univariate) data points, (2) remove all edges for which the defining nodes originate from different samples, and (3) define the test statistic R

as the number of disjoint subtrees that result. This will be one more than the number of edges deleted. Rejection of H_0 is for a small number of subtrees (runs).

When described in this manner the multivariate generalization becomes straightforward. In step 1 above, the MST of the multivariate pooled sample is constructed and used in step 2. This generalization preserves the spirit of the Wald-Wolfowitz test since the sorted list is used primarily to link points that are close together in R^1. The multivariate MST links points that are close together in R^p.

These concepts are illustrated in Figures 1, 2 and 3. The data shown in Figure 1 are two samples of 25 points each drawn from a standard bivariate normal distribution. In Figure 2, the samples are drawn from bivariate normal distributions where the location of one is translated by two standard deviations. In Figure 3, the underlying bivariate normal distributions have identical location, but the covariance matrix of one is three times that of the other. Figures 1a, 2a and 3a show the data points in the plane with their sample identities. Figures 1b, 2b and 3b superimpose the MST of the pooled sample, while Figures 1c, 2c and 3c delete those edges linking nodes from different samples.

Under H_0, the mean and variance of R can be calculated using a minor extension of the indicator variable approach for the univariate case (see e.g., Gibbons, 1971, Ch. 3). Number the $N-1$ edges of the MST arbitrarily and define Z_i, $1 \leqslant i \leqslant N-1$, as follows:

$Z_i = 1$ if the ith edge links nodes from different samples.

$= 0$ otherwise.

Then

$$R = \sum_{i=1}^{N-1} Z_i + 1 \quad \text{and} \quad E[R] = \sum_{i=1}^{N-1} E[Z_i] + 1. \tag{4}$$

Note

$$E[Z_i] = \Pr\{Z_i = 1\}. \tag{5}$$

Now $\Pr\{Z_i = 1\}$ is the probability that the two nodes defining this edge are labeled X and Y or Y and X. These probabilities are

$$\Pr\{XY\} = \Pr\{YX\} = \frac{mn}{N(N-1)},$$

so that

$$\Pr\{Z_i = 1\} = \frac{2mn}{N(N-1)}, \tag{6}$$

and from (4)

$$E[R] = \frac{2mn}{N} + 1. \tag{7}$$

This is the same result as in the univariate case (Wald-Wolfowitz, 1940).

The variance of the runs distribution (under H_0) can be calculated similarly:

$$\text{(8)}\qquad \begin{aligned}\operatorname{Var}[R] &= \operatorname{Var}\left[\Sigma_{i=1}^{N-1} Z_i\right] \\ &= \Sigma_{i=1}^{N-1} \operatorname{Var}[Z_i] + 2\Sigma_{i<j} \operatorname{Cov}[Z_i, Z_j].\end{aligned}$$

Clearly,

$$\text{(9)}\qquad \Sigma_{i=1}^{N-1} \operatorname{Var}[Z_i] = \frac{2mn}{N} - \frac{4m^2n^2}{N^2(N-1)}.$$

Now consider

$$\text{(10)}\qquad \operatorname{Cov}(Z_i, Z_j) = E[Z_iZ_j] - (E[Z_i])^2.$$

Observe that

$$\text{(11)}\qquad E[Z_iZ_j] = \Pr\{Z_iZ_j = 1\}.$$

This probability depends on whether or not the ith and jth edges share a common node. If they do, the two edges are defined by three nodes and there are two possible label sequences for which $Z_iZ_j = 1$: XYX or YXY with probabilities respectively

$$\Pr\{XYX\} = \frac{mn(m-1)}{N(N-1)(N-2)}$$

and

$$\Pr\{YXY\} = \frac{mn(n-1)}{N(N-1)(N-2)},$$

so that

$$\text{(12)}\qquad E[Z_iZ_j|\text{common node}] = \frac{mn}{N(N-1)}.$$

If Z_i and Z_j do not share a common node, there are four nodes defining the two edges and four possible labelings that lead to $Z_iZ_j = 1$: $(XY)(XY)$, $(XY)(YX)$, $(YX)(XY)$, $(YX)(YX)$, each with the same probability, so that

$$\text{(13)}\qquad E[Z_iZ_j|\text{no common node}] = 4\frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)}.$$

Let C be the number of edge pairs that share a common node. The total number of edge pairs is $\binom{N-1}{2}$. Combining this with (5–13), one has (after some algebraic simplification):

$$\text{(14)}\qquad \begin{aligned}\operatorname{Var}[R|C] = {} & \frac{2mn}{N(N-1)} \\ & \left\{\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}[N(N-1)-4mn+2]\right\}.\end{aligned}$$

The value of C depends upon the configuration (topology) of the MST: it is determined by the node degrees. In the univariate case, these degrees are fixed independent of the sample: there are two nodes of degree one and $N - 2$ nodes of degree two. In this case, $C = N - 2$ and (14) reduces to the Wald-Wolfowitz result.

In the general case ($p > 1$), MSTs with a variety of node degree values are possible and the variance of R under H_0 depends upon the common F. To make a distribution free test, we condition on the observed MST of the pooled sample points. It is sufficient (but not necessary) to condition on the pooled sample itself. Then C is fixed and (14) is the variance of R under the distribution induced by the $N!$ permutations of sample identities among the observations. From (7), one sees that the conditional and unconditional expectations are the same.

The observed MST topology completely determines the permutation distribution of the test statistic. One can compute higher moments in the same manner as the variance calculation above. For small enough sample sizes, it may be feasible to calculate directly the distribution of R over all permutations of the sample identities. For large sample sizes ($m, n \to \infty$ with m/n bounded away from 0 and ∞), the permutation distribution of

$$W = \frac{R - E[R]}{(\mathrm{Var}[R])^{\frac{1}{2}}}$$

approaches the standard normal distribution given any realized sequence of trees for which C is $O(N)$. For MSTs based on Euclidean, as well as more general (e.g., qth power) distances, sphere packing properties of p-space (Leech and Sloane, 1971) imply that C cannot exceed κN, κ a constant depending only on p.

The asymptotic normality of R is most easily seen by casting R in the form of a generalized correlation coefficient (Kendall, 1962) between the interpoint distances and the sample identities. Let a_{ij} and b_{ij} be scores for every pair of points:

(15) $a_{ij} = \frac{1}{2}$ if point i and point j define an edge of the MST
$= 0$ otherwise;

(16) $b_{ij} = 1$ if point i and point j are from the same sample
$= 0$ otherwise.

With this scoring define

(17) $$\gamma = \Sigma_{i=1}^{N} \Sigma_{j=1}^{N} a_{ij} b_{ij}.$$

Clearly, $\gamma = N - R$. Under the stated conditions, the asymptotic normality of γ, and hence R, induced by the $N!$ permutations of the sample identities follows directly from arguments in Daniels (1944). It is apparent from (15–17) that the Wald-Wolfowitz runs test can be viewed as rejecting the null hypothesis when "closeness" is too highly correlated with sample identity.

5. Multivariate maximum deviation (Smirnov) test. Under H_0, the distribution of the maximum absolute deviation (3) follows the Smirnov distribution for any assignment of integer ranks $1 \leqslant i \leqslant N$ to the N observations, provided that the ranking does not involve the sample identities. Specific ranking schemes are dictated by considerations of power. For a test to have reasonable power against general alternatives, it is desirable that there be a strong relationship between the absolute difference in rank between pairs of points and their distance in the observation space. For one-dimensional observations, this can be achieved by ranking the observations in order of their values. For higher dimensions, it is not generally possible to achieve as strong a relationship. Nonetheless, reasonable power can be achieved if this goal can be approximated reasonably well.

This section presents a convenient procedure for ranking multivariate observations based on their MST. As observed above, the MST tends to connect points that are close and is, therefore, a natural vehicle for such a procedure.

In order to describe this ranking procedure, it is necessary to introduce additional concepts from graph theory. The *eccentricity* of a node in a tree is the number of edges in a path with greatest length beginning with that node. The node at the other end of such a longest path is called an *antipode* of the node. The path between a node with largest eccentricity and its antipode is called a *diameter*. Define a *center* node of a tree as a node for which the eccentricity is minimum. For an MST in one dimension (equivalent to the sorted list), the eccentricity of a node (point) with rank i is $\max[i, N - i]$. The antipode of each node is one of the two end points of the list, the diameter is the path through the entire graph, and a center is the median (if the sample size is odd).

A *rooted tree* has one of its nodes designated as the *root*. (It is possible to make an MST a rooted tree by assigning one of its nodes to be the root; the MST of Figure 2b is represented as a rooted tree in Figure 4.) We associate with each node of a rooted tree its *depth*: the depth of the root is zero; the depth of any other node is the length of the (unique) path between it and the root. The *height* of a rooted tree is the maximum depth of any node in the tree. The *parent* of a given node is the penultimate node encountered on the path from the root to the given node; all nodes but the root have a parent. The *daughters* of a node are those nodes that are not its parent but are linked to it. The *ancestors* of a given node are those nodes on the path between the root and it, excluding the given node. The *descendants* of a node are all nodes for which it is an ancestor. The subgraph connecting a node and its descendants is a rooted subtree with that node as its root.

Our multivariate ranking procedure begins by rooting the MST at a node with largest eccentricity. The nodes (points) are then ranked in the order in which they are visited in a "height directed preorder" (HDP) traversal of the tree, which we define recursively:

(1) visit the root;

(2) HDP traverse in ascending order of height the subtrees rooted at the daughters of the root. (Resolve ties by visiting first subtrees with roots closer in Euclidean distance to the node visited in Step 1).

Figure 5 illustrates this ranking procedure for the MST of Figure 2b. For univariate observations, the MST is equivalent to the sorted list and (as in our generalization of the Wald-Wolfowitz test) our test reduces to the usual Smirnov test.

The univariate Smirnov test is known to have relatively low power against alternatives for which the two populations differ mainly in scale (Capon, 1965). Its power against scale alternatives can be substantially increased (at the expense of rendering the test ineffective against location alternatives) by ranking on absolute distance from the median of the pooled sample (as suggested by Siegel and Tukey (1960) for the Wilcoxon test). A multivariate generalization of this procedure is to root the MST at a center node and assign ranks such that nodes with larger depth receive higher rank than those with smaller depth. Nodes with the same depth can be ordered on their interpoint distance from the center node. As with its univariate counterpart, the resulting "radial Smirnov" test will be most sensitive to alternatives having similar location and differing primarily in scale.

Note that the particular ranking obtained is affected by the choice of a largest eccentricity node (there are at least two) or center nodes (there may be one or two). Nonetheless, our experience with these procedures has indicated that the value of the test statistic is generally little changed.

These two generalizations of the Smirnov test involve establishing a ranking of the points in R^p and applying the standard univariate Smirnov test to the resulting sequence. This ranking can clearly be used to generalize other nonparametric tests based on ranks. The effectiveness of this approach will depend on the extent to which the ranking reflects the interpoint distance relationships in R^p. In particular, one could apply the standard univariate Wald-Wolfowitz runs test to this sequence, yielding a different multivariate generalization. However, this would result in a test that is generally inferior to the one described in the previous section since that test directly deals with the interpoint distances through the MST and does not involve the approximation associated with a ranking.

6. 2×2 tests based on the MST. Multivariate two-sample tests can be based on 2×2 contingency tables. If the pooled observations are divided into two mutually exclusive categories, based on any criterion not involving the "X" or "Y" labels, one can test for the number of "X" observations in the first category. Under H_0, this quantity follows the usual hypergeometric distribution.

Since the MST of the pooled data points does not involve the sample identities, it can be used to partition the observations into two categories. A particularly useful partitioning is based on the degree of each node. Nodes of degree one form one category, while those with degree greater than one form the other. The test statistic is the number of degree one nodes from the "X" sample. Nodes of degree one tend to be found at the edges of the point scatter so that one might expect this test to be sensitive to scale alternatives.

7. Orthogonal spanning trees. Results were derived in Section 4 for the mean and variance of a runs statistic based on the MST. However, the derivations do not require that the set of edges considered form an MST or even a tree. The results are valid for any graph with exactly $N - 1$ edges, and moreover, for any graph containing the N points: simply substitute the number of edges in the graph as the upper summation limit in (4) and (8). Consider the test statistic S, the number of edges deleted. Let E be the number of edges in the graph and (as before) let C be the number of edge pairs that share a common node. Then, with $P = 2mn/N(N - 1)$, one has

$$E[S|E] = PE \tag{18}$$

and

$$\text{Var}[S|E, C] = P\left\{E + C + \frac{2(m-1)(n-1)}{(N-2)(N-3)}[E(E-1) - 2C] - PE^2\right\}. \tag{19}$$

The permutation distribution of S, conditioned on the realized graph, is asymptotically normal when C is $O(N)$. This follows directly from (15–17) by changing the scores a_{ij}: use $a_{ij} = 1/2$ if points i and j define an edge in the graph, and $a_{ij} = 0$ otherwise. Choices among different possible graphs are dictated by considerations of power. To have reasonable power against general alternatives, it is necessary that the edges generally link points that are close in the observation space. As pointed out in Section 3, this motivated our choice of the MST. The graph that links every point to its nearest neighbor(s) is another possibility. (For other multivariate tests based on near neighbors, see Weiss (1960) and Rogers (1976).)

For increasing sample size N, the MST edges link a decreasing fraction of the $N(N - 1)/2$ point pairs, and there are many close pairs that are not MST edges. Including these pairs could increase the power of the test. The concept behind the MST can be extended to produce larger graphs that maintain its desirable properties for two-sample testing. These extensions are based on the notion of orthogonal spanning trees.

An MST connects all of the points with minimum total distance. A *second* MST connects all of the points with minimum total distance subject to the constraint that it be orthogonal to the first MST. A *third* MST connects the points with minimum total distance subject to the constraint that it be orthogonal to both the first and second MSTs. Generally, the kth MST is a minimal spanning tree orthogonal to the $(k - 1)$th through the first MST. If $k \ll N/2$, the graph defined by all the edges of the first k MSTs should mainly connect close points and be appropriate for a two-sample runs test. The number of edges in this graph is $E = k(N - 1)$. The number of edge pairs, C, that share a common node is computed in the same manner (see Appendix) as for an ordinary MST. If k is held constant, then C is $O(N)$. But even if k is $O(N)$, implying C is $O(N^3)$, Daniels (1944) can still be used to demonstrate asymptotic normality of the test statistic.

8. Power comparisons. The utility of the tests presented in the previous sections lies in their power to discriminate against a wide variety of alternative hypotheses. In this section, we present the results of computer simulations that compare the power of these tests for several alternate hypotheses in various dimensions.

Table 1 shows results for normal populations where the alternatives are either differences in location (Table 1a) or scale (Table 1b). The tests compared are the multivariate runs test ["runs"] using the first, second, and third Euclidean distance MSTs, the multivariate maximum absolute deviation test ["Smirnov"], the number of degree one nodes from the first sample ["X (deg 1)"], and the multivariate radial Smirnov test ["radial Smirnov"]. Two additional tests are included in this comparison: the likelihood ratio criterion for normal populations with all parameters unknown ["normal theory"] which is asymptotically most powerful (Anderson, 1958), and the corresponding inverse normal scores test ["normal scores"] which has asymptotic relative efficiency one for normal populations (Puri and Sen, 1971). Comparisons are made in one, two, five, ten and twenty dimensions. In each case, the specific alternate hypothesis was chosen so that the tests have moderate power.

For Table 1a, each of the two populations is a standard normal distribution (unit covariance matrix) with mean vectors separated by a distance Δ. For Table 1b, the locations of the two populations are identical but the covariance matrix of one is scaled by σ.

Table 2 shows results for products of independent log normal distributions with alternatives differing in log location Δ. Changing Δ changes both the location and scale of a log normal population.

Table 1
Number of trials (out of 100 trials)
with significance less than 5%.
Normal data, $m = n = 100$.

	Location Alternatives				
	$p = 1$ $\Delta = 0.3$	$p = 2$ $\Delta = 0.5$	$p = 5$ $\Delta = 0.75$	$p = 10$ $\Delta = 1.0$	$p = 20$ $\Delta = 1.2$
Normal theory	45	77	91	92	84
Normal scores	21	41	40	45	30
Smirnov	46	55	51	44	24
Radial Smirnov	5	17	28	31	14
Runs—1st 2nd 3rd MST	14 17 18	17 26 35	44 56 64	53 70 78	71 79 86
X (deg 1)	5	6	5	5	5
	Scale alternatives				
	$p = 1$ $\sigma = 1.3$	$p = 2$ $\sigma = 1.2$	$p = 5$ $\sigma = 1.2$	$p = 10$ $\sigma = 1.1$	$p = 20$ $\sigma = 1.075$
Normal theory	62	54	61	19	10
Normal scores	5	5	5	5	5
Smirnov	12	8	5	5	7
Radial Smirnov	46	26	33	16	22
Runs—1st 2nd 3rd MST	16 18 24	15 11 14	9 13 21	6 3 9	10 12 13
X (deg 1)	5	5	36	39	52

Table 2
Number of trials (out of 100 trials)
with significance less than 5%.
Product log normal data, $m = n = 100$

	Log location alternatives				
	$p = 1$ $\Delta = 0.4$	$p = 2$ $\Delta = 0.4$	$p = 5$ $\Delta = 0.3$	$p = 10$ $\Delta = 0.3$	$p = 20$ $\Delta = 0.3$
Normal Scores	37	46	24	42	35
Smirnov	66	47	19	13	9
Radial Smirnov	5	58	72	87	93
Runs—1st 2nd 3rd MST	12 15 20	23 32 41	33 41 47	46 66 68	62 84 85
X (Deg 1)	5	5	5	24	61

The univariate Wald-Wolfowitz runs test is well known to be generally one of the less powerful nonparametric tests. This is verified in the results presented in Table 1 and Table 2. For the univariate case ($p = 1$), the Smirnov test is seen to dominate the other nonparametric tests (having comparable power to the likelihood ratio criterion) for location alternatives. This relationship tends to hold for the respective multivariate generalizations in low dimensions ($p < 5$). However, the relative power of the multivariate maximum deviation test is seen to decrease with increasing dimension. This reflects the increasing discrepancy between the relative ranks in the sequence and the interpoint distances in R^p, for increasing p. For high dimensions, the runs tests, which do not invoke a ranking, are seen to dominate approaching the power of the normal theory test at $p = 20$.

The performance of the runs test is seen to be generally improved by including the second and third MSTs. As might be expected, the incremental improvement is less for adding the third MST.

For low dimensions, none of the nonparametric tests are seen to have high power against scale alternatives, the radial maximum deviation test doing the best. For higher dimensions ($p > 5$), the X (deg 1) test achieves high power against scale alternatives.

For the product log normal data simulations, the radial maximum deviation test is seen to dominate for all $p > 2$ with the runs tests catching up at high dimensions ($p > 10$).

These simulations (and others not shown) indicate that the maximum deviation tests tend to have high power for low dimensions ($p < 5$) while the runs tests dominate for higher dimensions. In any particular situation, of course, the best choice will depend upon the specific underlying distributions.

9 Discussion. The MST is determined by the order of the sorted $N(N-1)/2$ distances or dissimilarities between the observations. Changing the dissimilarities between data points can change the MST. If the observations can be represented as points in a coordinate space, there are a variety of ways of defining interpoint distance. A common choice is Euclidean distance since it is invariant to rotations of the coordinate system. Within the choice of a distance measure is the choice of the relative scaling of individual coordinates (or their linear combinations). The

MST is known to be reasonably robust under moderate changes of this type (Zahn, 1971) but it is not invariant to such changes. Results for the null hypothesis are valid for any choice of dissimilarity measure or relative scaling, but the power against specific alternatives is affected by such choices. The considerations that lead to specific choices are the same as for any other procedure involving interpoint distances (Fukunaga and Hostetler, 1973).

There exist multisample extensions to both the univariate runs test (Mood, 1940) and maximum deviation test (Kiefer, 1959). Although we have only discussed multivariate generalizations for the two-sample case, our procedures can be used in a straightforward manner to generalize the usual multisample extensions. Moreover, our procedures clearly apply to related tests such as the Cramér-von Mises (Darling, 1957) which employ alternative statistics based upon absolute deviations.

10. Summary. Four multivariate two-sample tests have been presented. The Smirnov and radial Smirnov generalizations involve a sequencing of the sample points in R^p, while the runs test (with one or several MSTs) and the 2×2 test are direct multivariate procedures. The runs test and the Smirnov test can be expected to have power against general alternatives, while the radial Smirnov and 2×2 tests sacrifice generality in order to have increased power against scale alternatives. The simulation results presented in Table 1 and Table 2 indicate that the Smirnov generalizations have higher power in low dimensions ($p \leqslant 5$), while the runs and 2×2 tests are more sensitive in higher dimensions.

APPENDIX

Computational considerations. The computational problems involved in constructing MSTs have been extensively studied. The classical algorithms intended for complete graphs are due to Kruskal (1956) and Prim (1957). Their efficient implementation on a computer is described in Dijkstra (1959).

Prim gave several definitions and adduced two principles for constructing minimal spanning trees. Define an *isolated node* as a node to which, at a given stage of construction, no links have yet been made. A *fragment* is a spanning tree of a subgraph. An *isolated fragment* is a fragment that, at a given stage of construction, is not connected to the rest of the graph. The *distance* of a node to a fragment of which it is not a member is the minimum of its interpoint distances from the individual nodes comprising the fragment. A *nearest neighbor of a node* is one whose distance from the specified node is at least as small as that of any other. A *nearest neighbor of a fragment* is a node whose distance from the specified fragment is at least as small as that of any other.

With these definitions, Prim's construction principles for minimal spanning trees are:

Principle 1—any isolated node can be linked to a nearest neighbor.

Principle 2—any isolated fragment can be linked to a nearest neighbor by a shortest available edge.

Prim shows that an MST can be constructed by making $N - 1$ connections in accordance with these principles.

With Prim's principles, constructing MSTs becomes a straightforward procedure. For bivariate data represented as points in the plane, an MST can easily be built "based on visual judgments of relative distance, perhaps augmented by a pair of dividers in a few close instances" (Prim, 1957). If the matrix of dissimilarities is given, an MST can be constructed without too much effort for small data sets.

For large data sets in coordinate spaces or for large dissimilarity matrices, a computer is necessary. Prim's principles can be embodied into a fast computer algorithm (Whitney, 1972) that requires computation time $O(N^2)$ where N is the number of data points. For less than a few hundred points, these algorithms are the fastest known. For larger data sets in the plane, Shamos and Hoey (1975) have developed an algorithm with computation time never greater than $O(N \log N)$. For higher dimensions, Bentley and Friedman (1975), and Rohlf (1977), have presented algorithms for which the computation time has been measured to be on the average $O(N \log N)$.

After construction of the MST, the various test statistics can be evaluated in time $O(N)$. For the runs test, one simply counts the number of edges connecting nodes with different labels. The number of edge pairs sharing a common node, C, must also be counted to evaluate the variance (14, 19). If d_i is the degree of the ith node, then

$$C = \tfrac{1}{2}\Sigma_{i=1}^{N} d_i(d_i - 1). \tag{20}$$

The degree of each node can be found by counting the number of times the node appears as a member of a pair defining an MST edge.

For the maximum deviation test, rooting and traversing the MST can be done in time $O(N)$. (For general information on tree traversal, see Aho, Hopcroft, and Ullman, 1974.) For the radial maximum deviation test, it is necessary to find a center node of the MST. This is facilitated by the observation that a center node of the entire MST is also a center node of the subgraph defined by one of its diameters. Moreover, the antipode of any MST node must lie on one of the diameters. A center node can then be found by the following sequence of operations, each involving time $O(N)$. Choose an arbitrary node as root and find an MST node of greatest depth (this is an antipode). Choose this antipode as the root and find its antipode. These two nodes form the end points of a diameter of the MST. With one of these nodes as a root, find a node on this diameter for which the depth is as close as possible to one-half of the depth of its antipode. This is a center node of the MST. (See also Hakimi, 1964).

If sufficient storage is available for the distance matrix of the point set, then one can construct multiple MSTs, each in turn, using Prim's construction principles. As

each MST is completed, the entries in the matrix corresponding to its edges are set to infinity before constructing the next MST. If the distance matrix is too large for complete storage, the distances must be recomputed for each MST and the edges of the previous MSTs must be stored in a table that permits rapid searching for the existence of a particular edge. This is accomplished by storing each MST in a space efficient representation. Consider a list L of length $N - 1$. As each node i is linked to its nearest fragment at node j (using Prim's principles), one sets $L(i) = j$. Upon completion, the integer pairs $[i, L(i)]$ label the node pairs defining the MST. If an arbitrary node pair $[r, s]$ is an edge in this MST, it must be true that $L(r) = s$ or $L(s) = r$. K MSTs can be stored in an array A, dimensioned K by $N - 1$, such that the integer pair $[i, A(m, i)]$ labels two nodes defining an edge of the mth MST. If during the construction of the $(m + 1)$th MST the distance between i and j is required, the array is first checked to see if (i, j) is an edge in a previous MST. For this to be true, one of the $A(n, i)$ must be equal to j, or one of the $A(n, j)$ must be equal to i, for $1 \leqslant n \leqslant m$. If the pair (i, j) is found to be an edge of a previous MST, then their distance is set to infinity, otherwise their actual distance is computed.

A FORTRAN program implementing the tests described is available from either author.

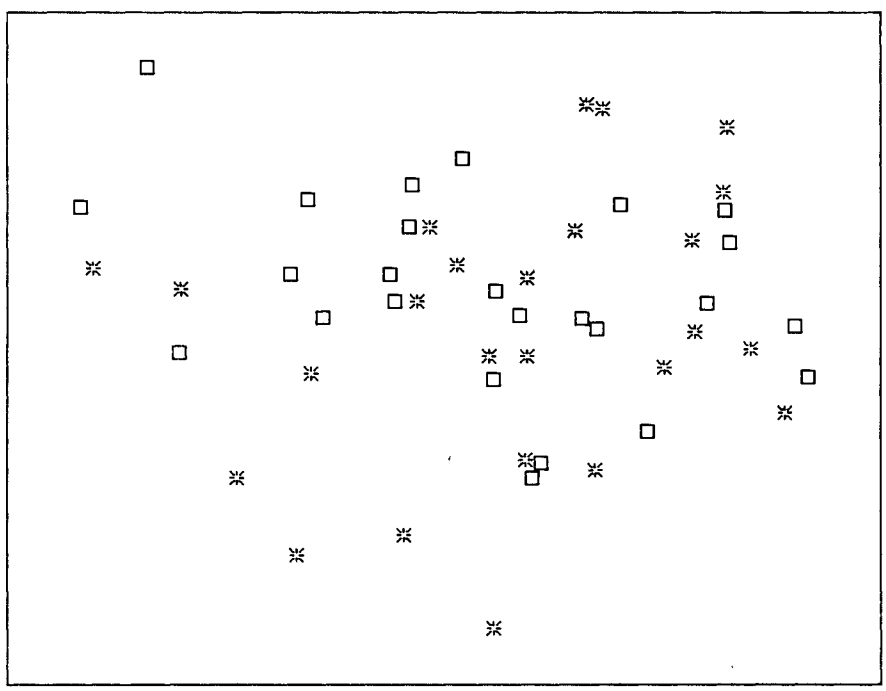

FIG. 1a.

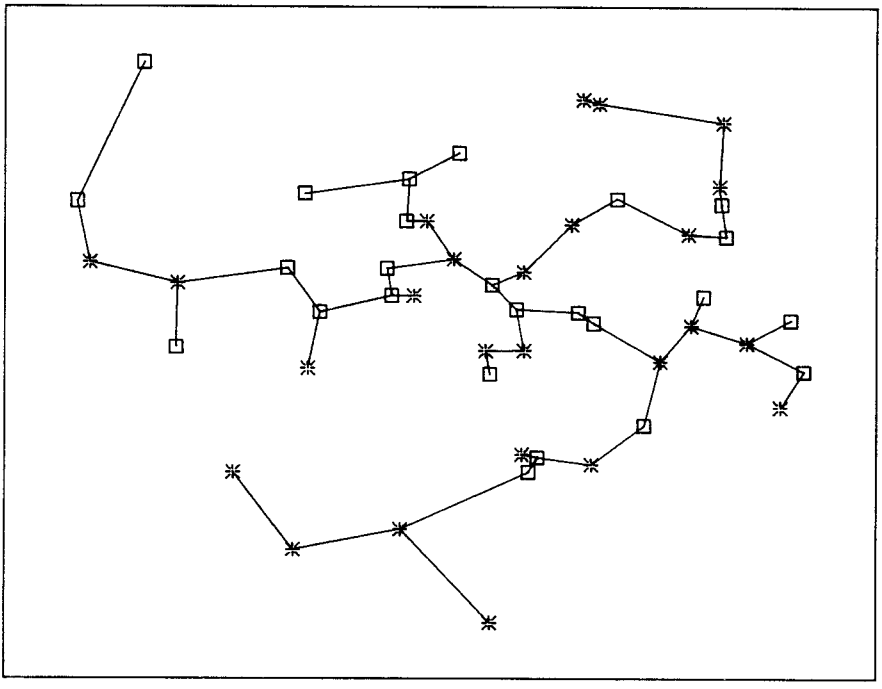

FIG. 1b.

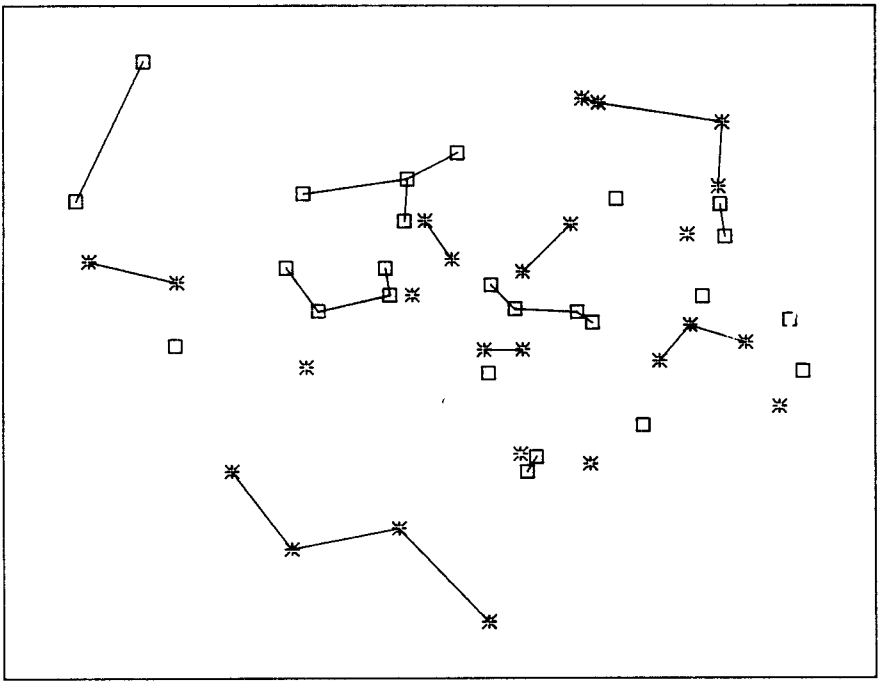

FIG. 1c.

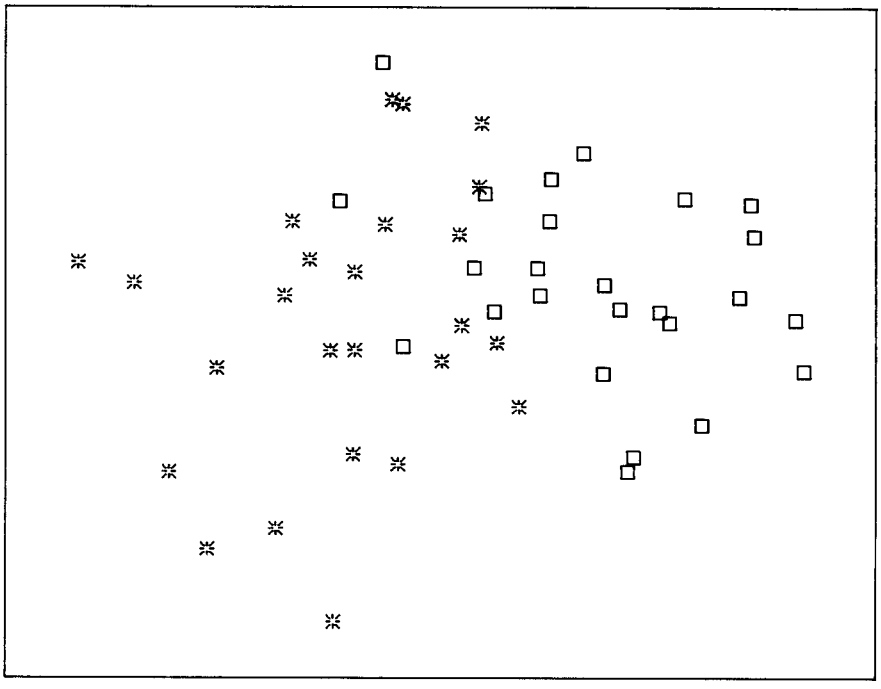

FIG. 2a.

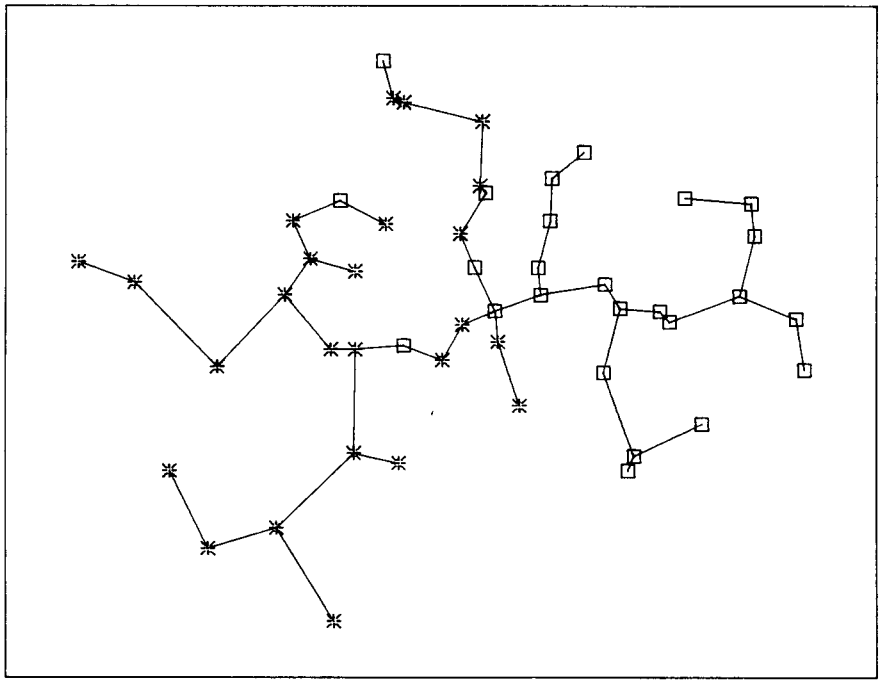

FIG. 2b.

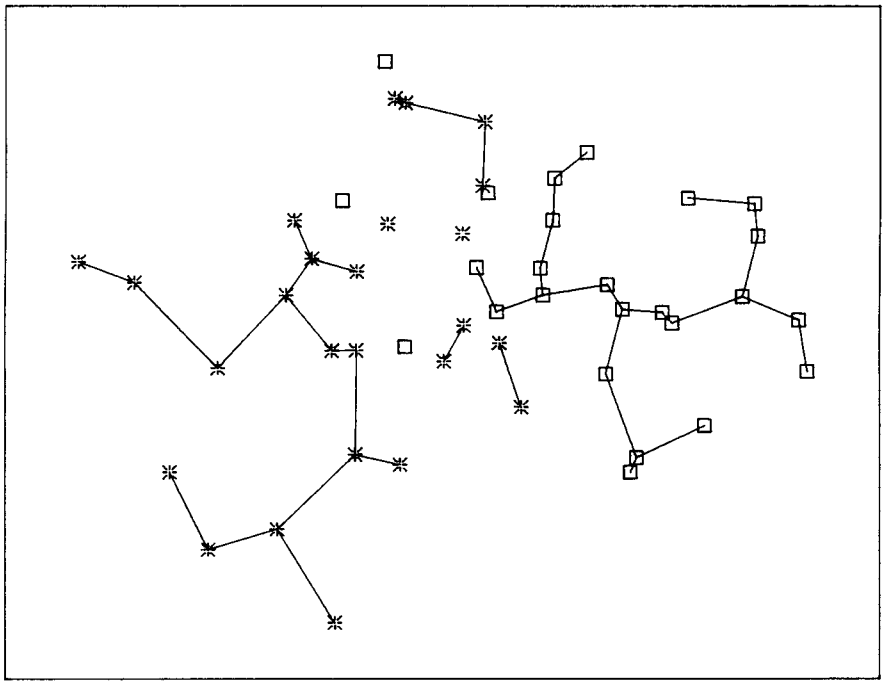

FIG. 2c.

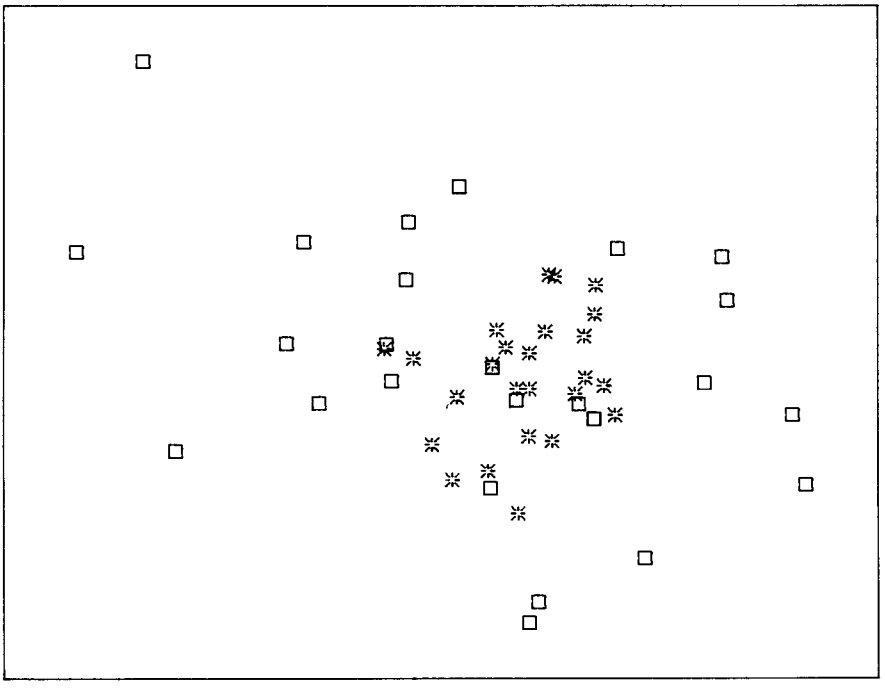

FIG. 3a.

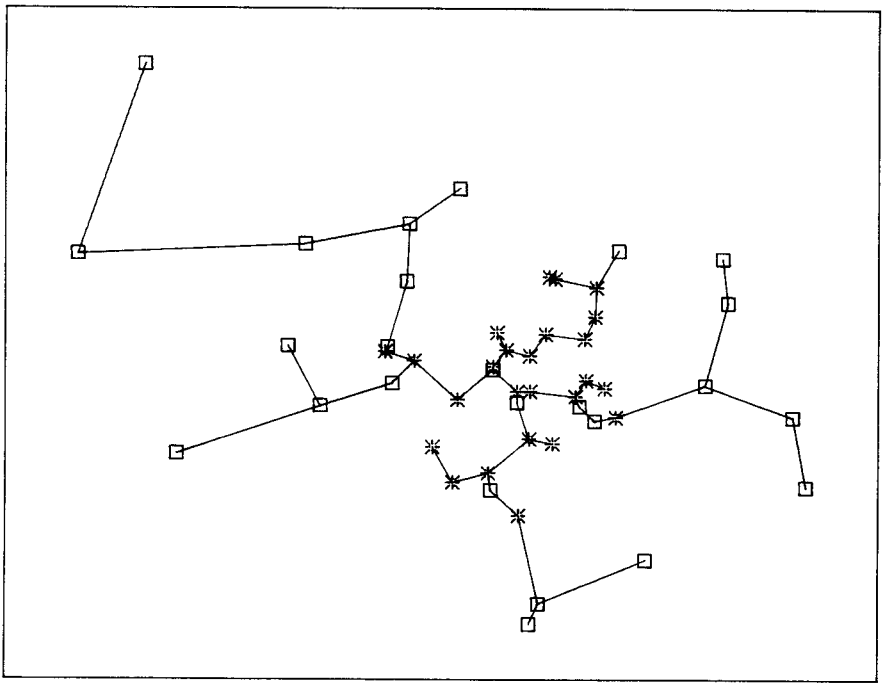

FIG. 3b.

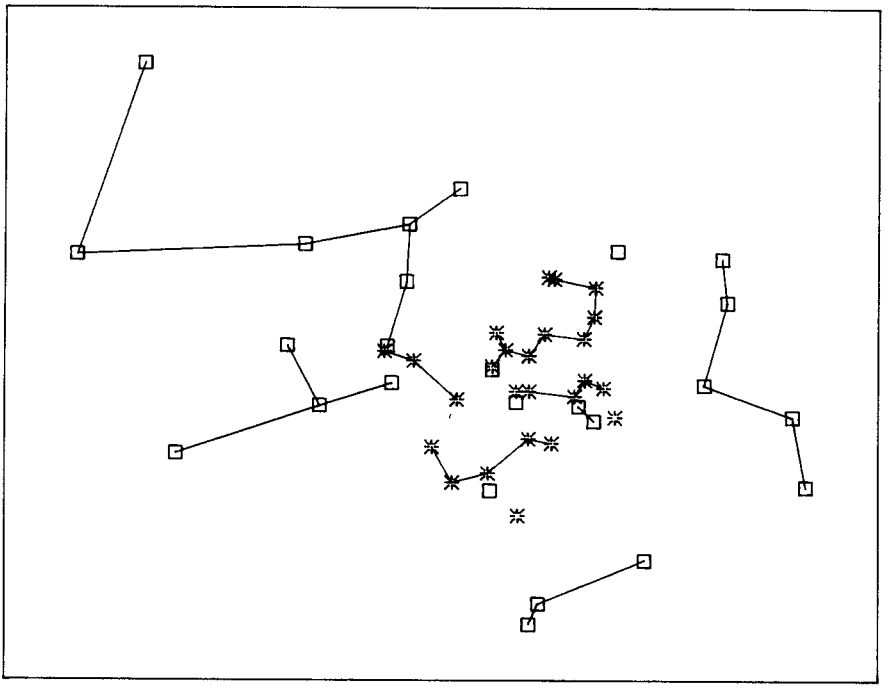

FIG. 3c.

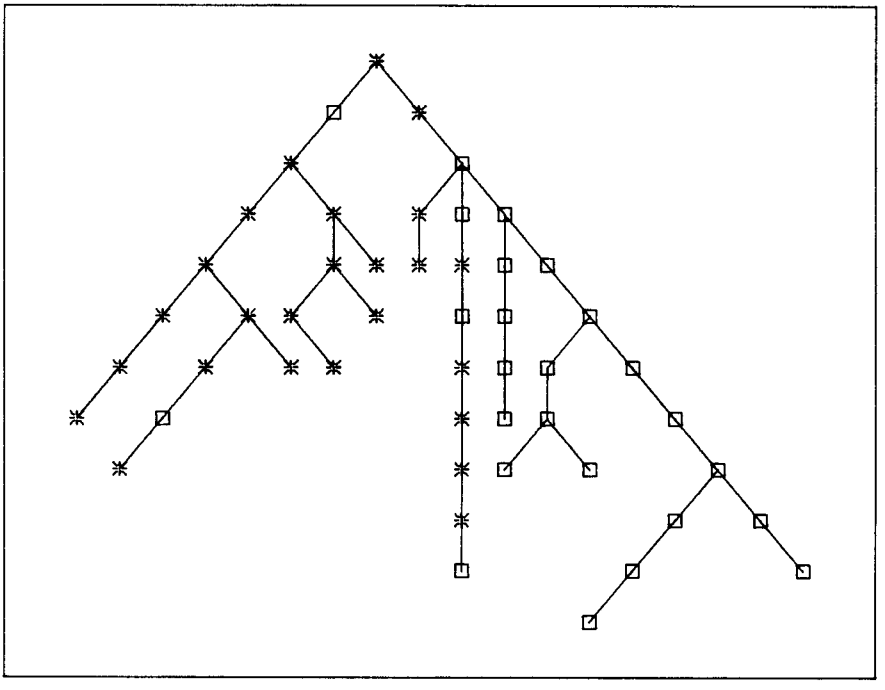

FIG. 4.

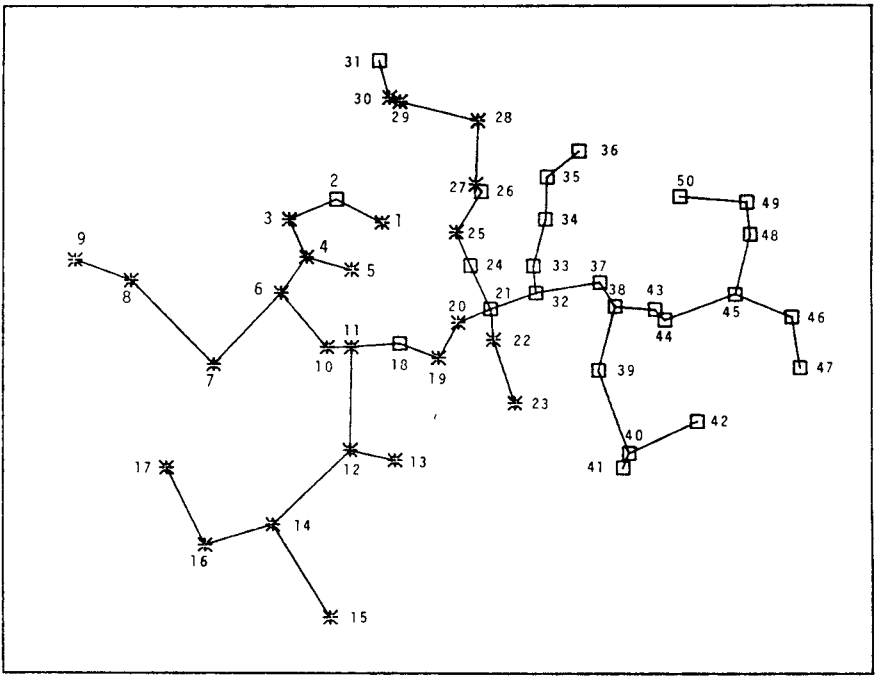
31
30
29
28
36
35
27
26
2
3
1
34
25
9
4
5
24
33
37
6
8
21
32
38
43
50
49
48
20
11
22
44
45
46
10
18
19
39
47
7
23
42
40
41
17
12
13
14
16
15

FIG. 5.

Acknowledgment. We thank John W. Tukey for bringing to our attention the notion of the second MST. We also thank Jon L. Bentley, Frank Boesch, Persi Diaconis and Andrew C. Yao for helpful discussions.

REFERENCES

AHO, A., HOPCROFT, J. and ULLMAN, J. (1974). *The Design and Analysis of Computer Algorithms*. 52–55. Addison-Wesley.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. 251–256. John Wiley & Sons.

BENTLEY, J. L. and FRIEDMAN, J. H. (1975). Fast algorithms for constructing minimal spanning trees in coordinate spaces. Stanford Linear Accelerator Report (SLAC) PUB-1665.

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

CAPON, J. (1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test. *J. Amer. Statist. Assoc.* **60** 843–853.

DANIELS, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33** 120–135.

DARLING, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Statist.* **28** 823–838.

DIJKSTRA, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math.* **1** 269–271.

FUKUNAGA, K. and HOSTETLER, L. D. (1973). Optimization of K-nearest neighbor density estimates. *IEEE Trans. Information Theory* **IT-19** 320–326.

GIBBONS, J. D. (1971). *Nonparametric Statistical Interference*. McGraw-Hill.

HAKIMI, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Res.* **12** 450–459.

HARARY, F. (1969). *Graph Theory*. Ch. 2. Addison-Wesley.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. 201–205. John Wiley & Sons.

KENDALL, M. G. (1962). *Rank Correlation Methods*. Ch. 2. Charles Griffin & Co.

KIEFER, J. (1959). K-sample analogs of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Ann. Math. Statist.* **30** 420–447.

KRUSKAL, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* **7** 48–50.

LEECH, J. and SLOANE, N. J. A. (1971). Sphere packings and error correcting codes. *Canad. J. Math.* **23** 718–745.

MOOD, A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.* **11** 367–392.

PRIM, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Tech. J.* **36** 1389–1401.

PURI, M. L. and SEN, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. 398–399. John Wiley & Sons.

ROGERS, W. H. (1976). Some convergence properties of k-nearest neighbor estimates. Ph.D. dissertation, Statist. Dept., Stanford Univ.

ROHLF, F. J. (1977). A probabilistic minimum spanning tree algorithm. IBM Research Report C6502.

SHAMOS, M. I. and HOEY, D. (1975). Closest point problems. 16th Ann. Symp. Foundations Computer Science, IEEE, 151–162.

SIEGEL, S. and TUKEY, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *J. Amer. Statist. Assoc.* **55** 429–445.

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

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

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

WHITNEY, V. K. M. (1972). Algorithm 422, minimal spanning tree. *Comm. ACM* **15** 273–274.
ZAHN, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Computers* **C20** 68–86.

STANFORD LINEAR ACCELERATOR CENTER
P. O. BOX 4349
STANFORD, CALIFORNIA 94305

ADP NETWORK SERVICES
ANN ARBOR, MICHIGAN