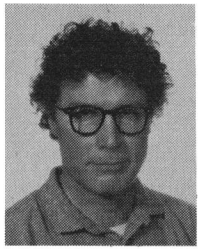
[8] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Ass.*, vol. 66, pp. 846-850, 1971.
[9] F. J. Rohlf, "Methods of comparing classifications," *Annu. Rev. Ecol. Syst.*, vol. 5, pp. 101-113, 1974.
[10] J. Rubin, "Optimal classification into groups: An approach to solving the taxonomy problem," *J. Theoretical Biol.*, vol. 15, pp. 103-144, 1967.



William H. E. Day was born in Boston, MA. He received the A.B. degree in astronomy from Harvard University, Cambridge, MA, in 1958 and the D.Sc. degree in applied mathematics and computer science from Washington University, St. Louis, MO, in 1975.

He was a systems programmer at IBM Corporation for ten years before attending graduate school. He has held faculty appointments at Southern Methodist University and Memorial University of Newfoundland. He is interested in applying graph theoretic and algorithm design techniques to solve computational problems in numerical taxonomy. He is a SIAM Visiting Lecturer.

Robert S. Wells received the B.Sc. degree in mathematics in 1979 and the B.Sc. (Honors) degree in computer science in 1981 from Memorial University of Newfoundland, St. John's.

He is employed by Statistics Canada, Ottawa, Ont., and is involved with the design and implementation of large scale application packages.

Mr. Wells is a member of the Association for Computing Machinery.

Testing for Uniformity in Multidimensional Data

STEPHEN P. SMITH, MEMBER, IEEE, AND ANIL K. JAIN, MEMBER, IEEE

Abstract—**Testing for uniformity in multidimensional data is important in exploratory pattern analysis, statistical pattern recognition, and image processing. The goal of this paper is to determine whether the data follow the uniform distribution over some compact convex set in K-dimensional space, called the sampling window. We first provide a simple, computationally efficient method for generating a uniformly distributed sample over a set which approximates the convex hull of the data. We then test for uniformity by comparing this generated sample to the data by using Friedman-Rafsky's minimal spanning tree (MST) based test. Experiments with both simulated and real data indicate that this MST-based test is useful in deciding if data are uniform.**

Index Terms—**Clustering tendency, convex hull, exploratory pattern analysis, minimal spanning tree.**

I. Introduction

THIS PAPER addresses the problem of testing the uniformity of multidimensional data. Data are represented as patterns (points) in a K-dimensional ($K > 2$) space. The need for a test of uniformity versus clustering arose in our research in the area of clustering tendency [5]. A major problem with clustering algorithms is that they impose a clustering structure on the data even if such structure is not inherent in the data. To avoid elaborate interpretation of uniform data, a test is needed to screen such data before clustering is applied. Thus for our purposes it will be enough to know that the data in Fig. 1(a) are "uniform" and the data in Fig. 1(b) and (c) are "clustered." In some other applications, one may be interested in knowing that the data in Fig. 1(e) form an "S."

A test for uniformity is useful in a number of application areas. For example, before finding a region of high point density in a Hough transform space one should determine if a "significant" cluster of such points exists. In image segmentation by clustering each pixel is represented as a point in a multidimensional space; these data should be tested for uniformity. A measure of uniformity of data is also useful in any pattern classification problem where the training samples are unlabeled.

A. Null Hypothesis

Our stochastic model for uniform data will be the continuous uniform distribution over some compact convex set in K-dimenisonal space, called the *sampling window*. Using this definition, the only data set in Fig. 1 which is uniform is Fig. 1(a).

Another possible null hypothesis of no structure is the Poisson point process [11], which is popular in forestry and ecology applications. A Poisson process, restricted to a bounded sampling window, induces a uniform density over that sampling window. A Poisson process null hypothesis allows the distribution of certain statistics for uniformity assessment to be derived. However, the Poisson process extends over the entire Euclidean space, so edge correction factors need

Manuscript received August 30, 1982; revised February 10, 1983. This work was supported by the National Science Foundation under Grant ECS 8007106.

S. P. Smith was with the Department of Computer Science, Michigan State University, East Lansing, MI 48824. He is now with the Northrop Research Center, Palos Verdes Peninsula, CA 90274.

A. K. Jain is with the Department of Computer Science, Michigan State University, East Lansing, MI 48824.

0162-8828/84/0100-0073$01.00

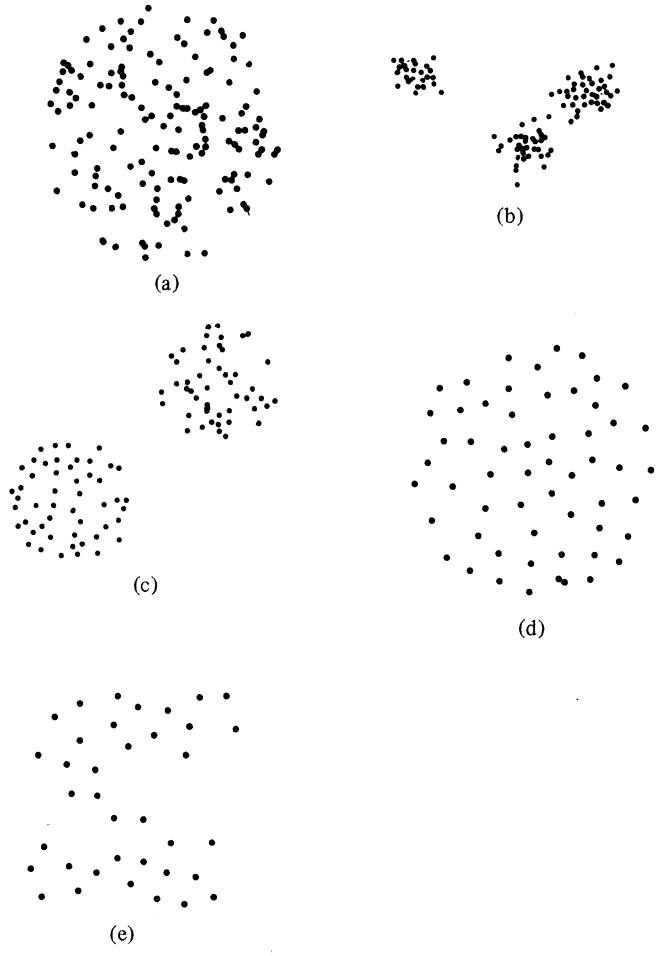


Fig. 1. Data sets exhibiting different structures. (a) Uniform data. (b) Clustered data. (c) Two-cluster data. (d) Regular data. (e) "S" shaped data.

to be introduced to analyze data over a bounded sampling window. A popular edge correction method, which can only be used for hyper-rectangular sampling windows, is the *wrap-around* method of edge correction [11]. A hyper-rectangular sampling window can be considered as a torus, so that opposite faces are considered to be identical. Thus, interpoint distances can wrap around the boundaries of the hyper-rectangle.

The difficulty of testing uniformity of a set of patterns is twofold. First, the sampling window is often unknown and must be estimated from the data. Second, the test for uniformity must be performed in the K-dimensional space. Popular one-dimensional tests for uniformity against a general alternative, such as the Kolmogorov-Smirnov test, are difficult to extend to high dimensions with an unknown sampling window. The distribution of uniformly distributed points which are projected into a lower dimensional space by any of the popular projection algorithms [2] is unknown in the projected space. Further, checking only for marginal uniformity may not be sufficient.

B. Alternative Hypothesis

The alternative to uniformity which is of most interest here is one of clustering. A well-known stochastic model for clustering is the Neyman-Scott cluster process [11]. It uses a set of points uniformly distributed over the sampling window as cluster centers and generates patterns around each cluster center. The Neyman-Scott process has three parameters:

1) N, the number of points in the realization,
2) μ, the average number of points in a cluster, and
3) σ, the spread of each cluster.

Cluster centers are uniformly distributed over the sampling window and the number of patterns around each cluster center has a Poisson distribution with parameter μ. Each pattern is a random vector following the normal density whose mean is the cluster center and whose covariance matrix is $\sigma^2 I$, where I is the identity matrix. Cluster centers and patterns are generated sequentially until N patterns have been obtained. A Neyman-Scott process over a hyper-rectangular sampling window can be generated with or without wrap-around. With wrap-around, points which fall outside the sampling window are moved back into the sampling window by referring to the torus topology. Without wrap-around, any point falling outside the window is rejected and a new point is generated.

The antithesis of a clustering is one of regularity. A well-known model for a regular alternative is the hardcore model [11]. This process places N patterns in sampling window S consecutively according to the rule that the ith pattern is distributed uniformly over the set of all points in S at Euclidean distance at least d from all previously located patterns. The parameter of the process is its packing density,

$$\rho = N A_k (d/2)^k / \vartheta(S)$$

where A_k is the volume of the unit hypersphere in K dimensions and $\vartheta(S)$ is the volume of the sampling window. The value of ρ, barring edge effects, is the proportion of S covered by N nonintersecting spheres of diameter d.

C. The Sampling Window

A *sampling window* can be defined as the compact convex support set for the underlying distribution. The crucial role of the sampling window in assessing the uniformity of a set of patterns can be seen from Fig. 2. Fig. 2(a) shows a small square inside the unit square over which 100 points have been generated uniformly. If the sampling window is taken to be the small square, then the data should be viewed as uniform. However, if for some *a priori* reason the unit square is taken as the sampling window, then the data might be viewed as a single cluster in the middle of the unit square. The need for a convex sampling window is shown in Fig. 2(b). This data set should intuitively be considered as consisting of two clusters. However, the 100 data points are uniformly distributed over two small circles. Hence, the data could be considered uniform over a region which is the union of these two circles. To exclude such a situation, we make the restriction that sampling windows be compact convex sets.

In practical situations, one is given N points in K dimensions and asked to determine if they are uniformly distributed. The sampling window is generally not known and must be estimated. Ripley and Rasson [10] have shown theoretically that in two dimensions the convex hull of the data is a good sampling window estimator; this has also been verified experimentally [12]. Unfortunately, determining the convex hull

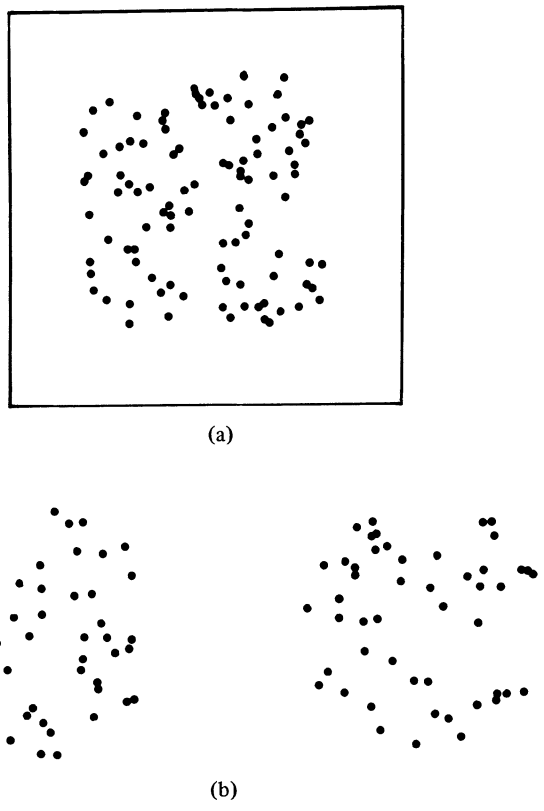


Fig. 2. Importance of the sampling window. (a) Data uniform over small subsquare. (b) Data uniform over two disjoint circles.

of high-dimensional data is computationally burdensome. The only algorithm for computing the convex hull in high dimensions known to us [3] requires $0(N^{f(K)})$ time. In Section II-A, we provide a simple heuristic for generating points uniformly distributed over a set which approximates the convex hull of the data.

D. Background

Testing the uniformity of a univariate sample is a well-studied problem [8], [13]. Unfortunately, few extensions to higher dimensional samples have appeared [6]. Using a Poisson process null hypothesis, there have been extensions to the two-dimensional case [11], but with the sampling window assumed known. Two of these tests, the Hopkins and the Cox-Lewis test, have recently been extended to higher dimensions [4], [9]. Both have shown higher power against clustered alternatives. Section II introduces a test which first generates a uniformly distributed sample over a set which approximates the convex hull of the data which is to be tested for uniformity. The test then determines whether this generated sample and the given data belong to the same population using the minimal spanning tree (MST) of the pooled samples. Section III contains an experimental study of the MST-based test on both simulated and real data, while Section IV presents our conclusions and suggestions for future research.

II. A Minimal Spanning Tree Based Test

We define a test which does not explicitly require any knowledge of the true sampling window. It is assumed that the convex hull of the data is a reasonable estimate of the sampling window, but there is no need to compute the convex hull. The idea of this test comes from a multivariate extension of the Wald-Wolfowitz runs test proposed by Friedman and Rafsky [6]. The Friedman-Rafsky test determines if two sets of high-dimensional sample points belong to the same distribution. The test statistic is determined from the minimal spanning tree (MST) of the pooled sample points. We adopt this test for our purposes by bootstrapping. The given data which are to be tested for uniformity constitute one sample. The other sample needed for the Friedman-Rafsky test is obtained by generating points uniformly distributed over the convex hull of the given data. If the null hypothesis that the two samples belong to the same population is accepted, then we say that the given data are uniformly distributed over the convex hull. One of the problems, of course, is generating uniform data over the convex hull, since it is not computationally feasible to form the convex hull of high-dimensional data.

A. Generating Uniform Points over the Convex Hull

We describe a heuristic which produces uniform points over a set which is approximately equivalent to the convex hull of the data. It may be possible that results from computational geometry could provide an exact, computationally feasible algorithm for generating points uniformly in the convex hull, but this is unsubstantiated by published results. Note that elegant results on expected time behavior of certain algorithms, such as Bentley *et al.*'s work on maximal vector formation [1], cannot be applied in this situation since the "independent and distinct" property of the data cannot be guaranteed to hold. We will see that the heuristic which follows leads to a fast algorithm and, more importantly, gives an adequate bootstrap sample for the uniformity test to be applied.

The overall procedure will be as follows. A (relatively simple) convex set containing the data is determined. Uniformly distributed points are generated over this set and those that fall in the convex hull of the data are retained. This rejection technique would then approximate a set of uniform points over the convex hull. The rejection procedure will use the following property of the convex hull $H(X)$ of the given data $X = \{Xi\}$. A point Y is not in $H(X)$ if and only if there exists a hyperplane, with normal vector n, passing through Y, such that $((Xi - Y) \cdot n) > 0$ for all $i = 1, 2, \cdots, N$, where $(v \cdot w)$ is the inner product of vectors v and w. This follows from the definition of $H(X)$ as the intersection of all convex subsets containing $\{X_i\}$. It should be clear that, for a point Y not in $H(X)$, one normal vector that will always satisfy the above positivity constraint is the vector $n^* = Z - Y$, where Z is the unique point in $H(X)$ closest to Y.

We want to estimate n^* from the given data. If the data are uniform over $H(X)$, one expects to see points in the data set which are near the point Z. These points could estimate n^*. We choose the following estimator which takes a weighted average over all points in the data set:

$$\hat{n}^* = N^{-1} \sum_{i=1}^{N} (Xi - Y)/(||Xi - Y||_2)^{K+1}.$$

This estimator is the sum over all i of the unit vectors from Y to Xi weighted by an amount inversely proportional to the

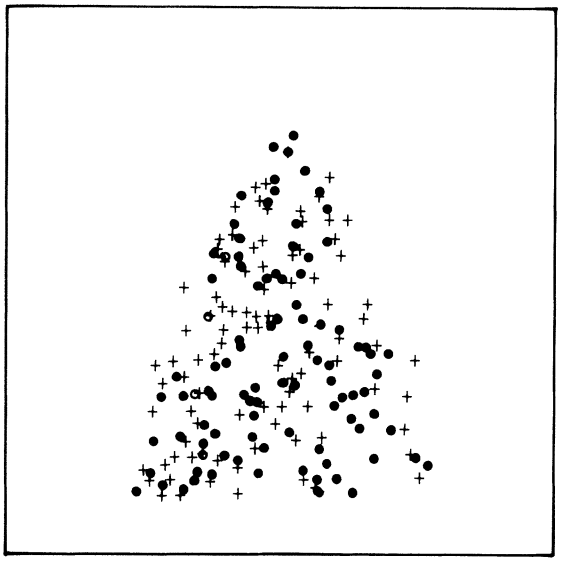


Fig. 3. Points generated by the rejection technique. 100 points "○" are generated uniformly in the triangle. Another 100 points are shown as "+" after passing the rejection procedure. A total of 406 points were generated randomly over the unit square to obtain this second sample.

Kth power of the distance from Y to Xi. The Kth power of the distance penalizes points far from Y. This penalty is proportional to the volume of a hypersphere centered at Y and passing through Xi.

The procedure to compute a sample of points uniformly distributed over (approximately) the convex hull of the data is as follows. A simple compact convex set is placed around the data. For example, this set could either be the smallest hypersphere enclosing the data [12] or a hyper-rectangle $[ai, bi]_{i=1}^{K}$ aligned with the coordinate axes of the space, where ai and bi are, respectively, the minimum and maximum values in the ith coordinate among the data points. Points are then generated which are uniformly distributed over this set. A point Y is rejected if all the data lie in one half-space of the hyperplane passing through Y with normal vector $\hat{n}^*$, (i.e., $\hat{n}^*$ satisfies the positivity constraint mentioned above). The procedure continues until the desired number of points has been generated. This algorithm is less costly than computing the convex hull explicitly if the initial convex set used to enclose the data is not too large. For instance, if 100 points are uniformly distributed in the unit hypersphere in 10 dimensions, the time taken to generate a sample of 100 points by the rejection technique over the smallest aligned hyper-rectangle is approximately 10 CPU seconds on a CDC CYBER 750/175.

This procedure is demonstrated in Fig. 3. One hundred uniform points are given inside a traingle contained in the unit square. An additional 100 uniform points were generated using the rejection procedure by first generating points over the unit square. We note that the points generated by the rejection procedure (denoted +) appear to be uniformly scattered like the original data in the triangle (denoted ○).

B. Definition of the Test

The Friedman-Rafsky test is based on the minimal spanning tree (MST) of the pooled sample points. The MST has been used extensively in unsupervised pattern recognition, chiefly as a basis for clustering data. Establishing the MST for points in an Euclidean space involves computation of a complete weighted graph whose nodes represent the points. The edges in the graph are weighted by the Euclidean distance between the points. The MST is that subgraph which is a spanning tree (a spanning tree is a connected graph with no cycles) and which has minimal sum of edge weights. If the set of distances between points has no ties then its MST is unique.

In the Friedman-Rafsky test, the MST of the pooled samples is first computed. Let the N data points in one sample be labeled X and the M points in the second sample be labeled Y. The number of edges in the MST linking a point labeled X to a point labeled Y is found. Denote this X-Y join count as T. Under the null hypothesis that the two samples are from the same population, Friedman and Rafsky show that

$$E[T] = 2MN/L$$

and

$$\operatorname{var}[T|C] = \frac{2MN}{L(L-1)}\left\{\frac{2MN-L}{L} + \frac{C-L+2}{(L-2)(L-3)}\left[L(L-1)-4MN+2\right]\right\}$$

where C is the number of edge pairs in the MST sharing a common node and $L = M + N$. Further, the permutation distribution of T, conditioned on C is asymptotically normal.

In the context of our situtation, the points labeled X are the given data points and the points labeled Y are uniformly generated over a set which approximates $H(X)$. If the given data are uniform, one expects the null hypothesis of the Friedman-Rafsky test to be true. In the case of clustered data, many of the points labeled Y are expected to be generated between clusters. This would produce an unusually high number of X-X and Y-Y joins, thus reducing the value of the statistic T. Fig. 4 shows an example of this.

Thus the test for uniformity against a clustered alternative is as follows. Reject the data as uniform when

$$\frac{T - E[T]}{\sqrt{\operatorname{var}[T|C]}} < Z(\mathfrak{L})$$

where $Z(\mathfrak{L})$ is the $\mathfrak{L}$ quantile of the standard normal distribution. One could, of course, perform the analogous upper tail test for uniformity against the alternative hypothesis of regularity.

The MST-based test to analyze a data set containing N points over unknown sampling window can be summarized as follows. The number of points to include in the uniformly distributed sample is open. For simplicity, we choose to have the two samples of equal size.

1) Determine a convex set containing the data.

2) Using the rejection technique, generate N uniformly distributed points over a set which approximates the convex hull of the data.

3) Pool the N data points and the N uniform points generated in step 2) and compute their MST.

4) Determine the test statistic T. Reject the data as uniform

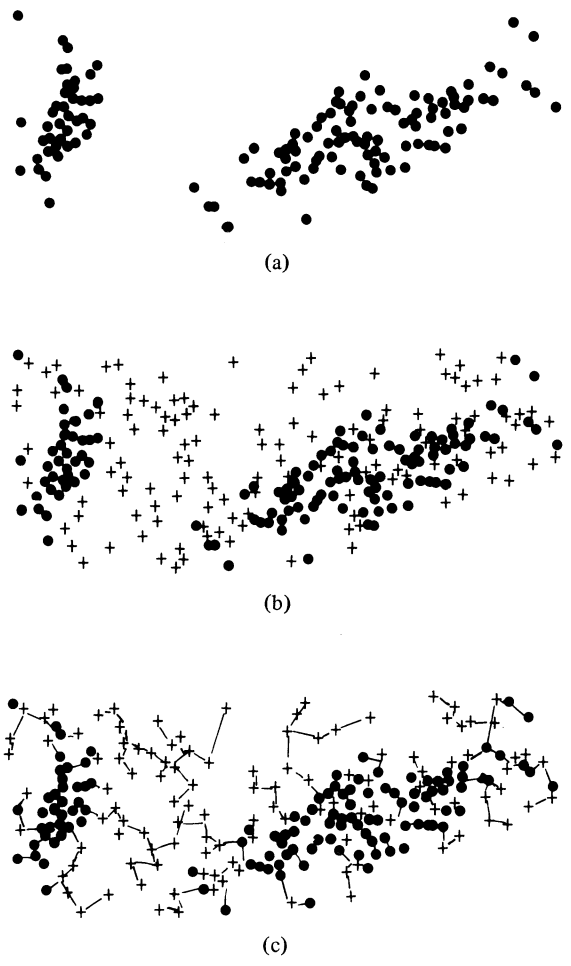


Fig. 4. MST-based test on clustered data. The value of the MST statistic is −6.13. (a) Original data. (b) Original data and points generated by the rejection technique. (c) MST of the data and the points generated by the rejection technique.

in favor of a clustered alternative if T is too small. Reject the data as uniform in favor of a regular alternative if T is too large.

If the sampling window is known, replace steps 1) and 2) by generating N points uniformly over this window.

III. Performance of the MST-Based Test

In this section we analyze the performance of the MST-based test by simulation. Only the rejection rates for a one-sided test against a clustered alternative at both the 0.05 and 0.02 levels are reported. The entries in the tables are $(R(0.05), R(0.02))$, where $R(\pounds)$ is the percent rejections of the null hypothesis of the $\pounds$ level. At the 0.05 level, we expect 5 rejections out of 100 under the null hypothesis. The parameters K and N are the number of dimensions and the number of patterns, respectively. The number of Monte-Carlo runs is fixed to be 100.

A. Sampling Window Known

In the following experiments we assume that the sampling window over which the data is generated is known. Thus the set of uniform points is generated over this window.

1) Uniform Data: Table I reports the results when a sample of uniform data in the unit hypercube is subjected to the MST-based test. Since all the entries in Table I are within their expected values (at the 0.05 level of significance), it is concluded that one can set the size of the MST-based test at a given level. These simulations have also shown that the size of the one-sided test against a regular alternative may also be set using the asymptotic distribution of T.

TABLE I
Size of the MST-Based Test for Uniform Data in Unit Hypercube

		K		
		2	5	10
N	50	(1, 1)	(4, 0)	(3, 2)
	100	(5, 3)	(6, 4)	(7, 3)
	200	(5, 2)	(5, 4)	(5, 2)

TABLE II
Power of the MST-Based Test Against a Neyman-Scott Process (Wrapped) in Unit Hypercube. $N = 200$.

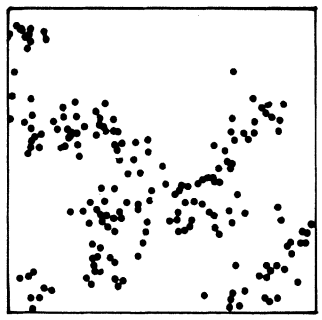
		σ			
		0.05	0.1	0.2	
μ	16	(100, 100)	(86, 74)	(12, 7)	$K = 2$
	8	(100, 100)	(56, 37)	(4, 1)	
	1	(46, 28)	(11, 2)	(5, 1)	
μ	16	(100, 100)	(100, 100)	(46, 32)	$K = 5$
	8	(100, 100)	(100, 100)	(29, 18)	
	1	(100, 100)	(99, 94)	(15, 6)	

2) Neyman-Scott Process: Table II gives estimates of the power of the MST-based test for Neyman-Scott clustering alternatives. To compare our results with previous studies, the wrap-around paradigm is used, both for generating the data inside the unit hypercube and computing the interpoint distances. The MST defined with wrap-around is then a tree on a torus.

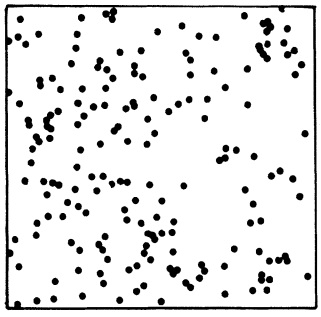
The MST-based test shows the expected increase in power with increasing μ and decreasing σ. It also shows an increase in power with dimensionality. Fig. 5 shows two realizations of the Neyman-Scott process. With large σ, the Neyman-Scott process approaches a uniform density. The performance of the MST-based test can be compared with two well-known distance-based tests: the Hopkins test [4] and the Cox-Lewis test [9]. Table III gives these comparisons for $\mu = 16$. The MST-based test gives significantly higher power (at the 0.001 level) against all other tests for $K = 2$ with $\sigma = 0.05$ and 0.1 and for $K = 5$ with $\sigma = 0.2$.

3) Other Data Types: To estimate the power of the MST-based test against a hardcore alternative, one must change the test's critical region to the 1-$\pounds$ upper tail of the normal distribution since very few X-Y joins are expected under a regular alternative. Rejection rates of (64, 46), (100, 100), and (100, 100) are obtained in 2, 4, and 5 dimensions, respectively, for a hardcore process over the unit hypercube with $\rho = 0.1$, and with $N = 200$. Fig. 6 shows a realization of this hardcore process.

Finally, the power of the MST-based test for detecting a normal swarm of points is investigated. The normally distributed points (with zero mean and identity covariance matrix) are forced into the unit radius hypersphere by dividing all the points by the maximum norm of the data. The second sample required for the MST-based test is then generated uniformly



(a)



(b)

Fig. 5. Two realizations of 200 points from the Neyman-Scott process over the unit square with wrap-around. (a) Realization with $\mu = 8$, $\sigma = 0.05$. (b) Realization with $\mu = 8$, $\sigma = 0.2$.

TABLE III
COMPARISON OF THE POWERS OF THE HOPKINS, COX-LEWIS, AND MST-BASED TESTS. ENTRIES ARE PERCENT REJECTIONS FOR HOPKINS, COX-LEWIS, MST (H, C, M) $N = 200$, $\mu = 16$.

	H, C, M	H, C, M	H, C, M
$K = 2$	80, 53, 100	32, 7, 74	*, 2, 7
$K = 5$	91, 100, 100	94, 90, 100	*, 8, 32
	0.05	0.1	0.2
		σ	

*Entries not provided by Cross and Jain [4].

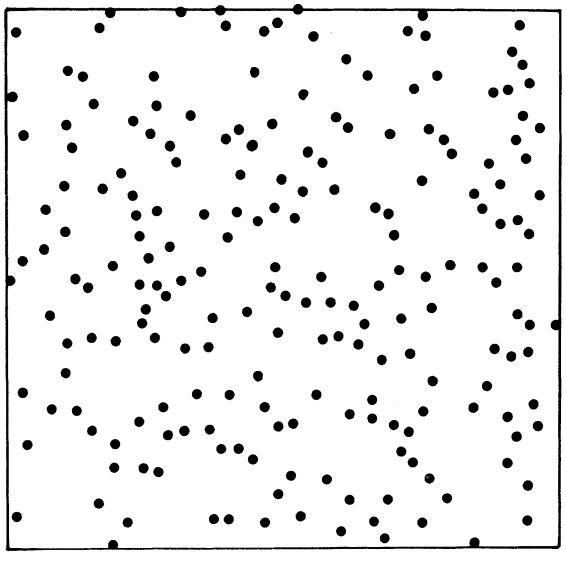


Fig. 6. Realization of the hardcore process with 200 points over a unit square with $\rho = 0.1$.

TABLE IV
SIZE OFF THE MST-BASED TEST FOR UNIFORM DATA IN AN UNKNOWN UNIT HYPERCUBE

		K			
		2	3	5	10
N	50	(4, 1)	(2, 0)	(0, 0)	(0, 0)
	100	(4, 2)	(5, 2)	(6, 0)	(0, 0)
	200	(6, 1)	(3, 2)	(0, 0)	(0, 0)

TABLE V
POWER OF THE MST-BASED TEST AGAINST THE NEYMAN-SCOTT PROCESS (WRAPPED) IN AN UNKNOWN UNIT HYPERCUBE. $N = 200$.

μ	16	(100, 100)	(78, 72)	(14, 6)	$K = 2$
	8	(100, 99)	(40, 30)	(11, 4)	
	1	(47, 37)	(12, 8)	(3, 0)	
μ	16	(100, 100)	(100, 100)	(21, 14)	$K = 5$
	8	(100, 97)	(68, 51)	(6, 5)	
	1	0.05	0.1	(3, 1)	
			σ	0.2	

over this hypersphere. When $N = 200$, rejection rates of 100 percent, for $K = 2$, 5, and 10 are obtained; however, when N is decreased to 50, the rejection rates (at the 0.05 level) become 50, 84, and 76 for 2, 5, and 10 dimensions, respectively. With 50 points, the Gaussian alternative is barely distinguishable from uniform data.

B. Sampling Window Unknown

Here we determine the size and power of the MST-based test for unknown sampling windows. The second sample of uniform points is generated by the rejection technique as mentioned in Section II-A.

1) Uniform Data: Table IV reports the size estimates for uniform data in a unit hypercube. Similar results have been obtained with data generated randomly over a hypersphere, a hyperellipse and simplexes. Note that all entries are within or below their expected values. As dimensionality increases, the test becomes more conservative, i.e., the observed number of rejections of the null hypothesis is less than expected. This can be explained by noting from the simulations that the mean of the test statistic T increases as dimensionality increases. This is due to the fact that the volume of the convex hull of the data underestimates the volume of the true sampling window. Thus the uniformly distributed points inside the convex hull decrease the pattern-to-pattern (X-X) joins more than would be expected under the null hypothesis of the Friedman-Rafsky test. This may result in less power against clustered alternatives. In addition, this precludes using the MST-based test as a test for uniformity versus a regular alternative, since the proper size of the test cannot be set with an unknown sampling window, except by Monte-Carlo simulation.

2) Neyman-Scott Process: In Table V, we use the Neyman-Scott cluster alternative over a unit hypercube with wrap-around. Of course, since the sampling window is unknown, distances cannot be computed using wrap-around. The power of the MST-based test with unknown sampling

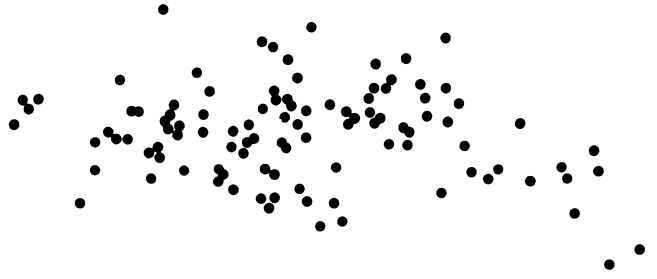


Fig. 7. IRIS23 data projected by the principal component method.

window (Table V) is similar to the case when the sampling window in known (Table II), although a slight loss in power can be seen. Similar results were obtained with a Neyman-Scott process generated inside a hypersphere.

C. Experiments with Some Real Data

To demonstrate the applicability of the MST-based test in practical situations, we used it to test for uniformity in some data from actual studies in pattern recognition. We assume that no information about the sampling window is available. In addition, we do not utilize any category information (pattern labels). The data sets used in this study are the following.

1) IRIS—This is a well-known data set containing measurements on three species of iris. It consists of 50 patterns from each species on each of 4 features. See Fig. 4(a) for a projection of the IRIS data to two dimensions by the principal component method [2].

2) IRIS23—This is a subset of the IRIS data containing measurements for only two of the species (versicolor and virginica). These 100 patterns are known to be well separated from the patterns corresponding to the setosa species. Fig. 7 shows these data projected to two dimensions by the principal component method.

3) 80X—The 80X data set is derived from the Munson handprinted Fortran character set. Included are 15 patterns from each of the characters "8," "0," and "X." Each pattern consists of 8 feature measurements [5]. Fig. 8 shows the 80X data projected to two dimensions by principal component analysis.

The data sets are tested in the following two configurations.

1) The original feature space.

2) The patterns are transformed so that the data have zero mean and identity covariance matrix. This is done by whitening the data, i.e., applying the principal component transformation followed by the whitening transformation [7]. Using this configuration shows the effect of scaling the data.

Each configuration is tested by the MST-based test in two ways. First, the rejection technique is used to generate the second sample of uniform points. Second, the smallest aligned hyper-rectangle and the smallest hypersphere enclosing the data are found and the estimate with the smaller volume is used as the true sampling window. The second sample of uniform points is then generated inside this sampling window.

The results of these experiments are shown in Table VI. Except in a few instances, all the entries in the table are significant for rejecting the null hypothesis of uniformity at the 0.05 level (the 0.05 quantile of the normal distribution is -1.65). The exception is the transformed 80X data tested by using the rejection technique. Even in this case, the value of the test statistic indicates the presence of a slight clustering in the data. There is a general trend in Table VI which shows that the test statistic using the rejection technique is larger than when the best fitting sampling window is used. This is expected since the high-dimeniosnal data do not usually fit very well inside the smallest aligned hyper-rectangle or smallest hypersphere sampling windows, and the data may look like a single cluster in the center of the window. This effect also arises due to the conservative nature of the MST-based test when using the rejection technique.

One problem with the MST-based test is that repeating the test with a different uniform sample can yield a different value of the test statistic. For instance, if the transformed 80X data are retested by both the rejection and the best fitting window methods, the test statistic values are 0.21 and -1.90, respectively. These new values suggest that this data set is less clustered than indicated previously. If the original IRIS data are

Fig. 8. 80X data projected by the principal component method.

TABLE VI

THE PERFORMANCE OF THE MST-BASED TEST ON SOME REAL DATA SETS. ENTRIES ARE THE VALUE OF THE NORMALIZED FRIEDMAN-RAFSKY STATISTIC. THE TOP NUMBER IS THE VALUE USING THE REJECTION TECHNIQUE. THE BOTTOM NUMBER IS THE VALUE USING EITHER THE SMALLEST HYPER-RECTANGULAR OR SMALLEST HYPERSPHERE SAMPLING WINDOW.

Configuration	Data Sets		
	IRIS	IRIS23	80X
Original	-11.08	-3.59	-1.90
	-12.91	-8.77	-4.64
Transformed	-5.42	-2.83	-.90
	-6.46	-5.45	-3.79

retested we get about the same value of the test statistic for the rejection method, but the value using the smallest aligned hyper-rectangle decreases to -17.20. This suggests that the test will view well-clustered data as well-clustered no matter what uniform sample is used. However, if the value of the test statistic is close to the critical value (for the 0.05 level) then the interpretation requires caution. One solution is to perform the test with various samples and average the resulting statistic values. Under the null hypothesis, this average should again have an approximate normal distribution.

We conclude that the MST-based test is able to provide reliable information about the uniformity of real data.

IV. Conclusions and Discussion

The focus of this research has been how to determine whether a given set of high-dimensional patterns is distributed uniformly over some region (the sampling window). We developed a test, based on the minimal spanning tree, which assumes only that the convex hull is a reasonable estimate of the sampling window. This MST-based test uses the Freidman-Rafsky test to determine if two samples come from the same population. In our application, one of the samples is the given data; the other sample is generated uniformly over the sample window. The generation of this second sample is straightforward when the sampling window is known. For unknown sampling windows, we present a heuristic that generates uniformly distributed points over a set that approximates the convex hull of the given data.

We found that if the sampling window is known, then the size of the MST-based test could be determined, even if the number of patterns is small. The power against the clustered alternative is significantly higher (at the 0.001 level) than other tests available in the literature. For unknown sampling windows, the MST-based test is conservative against a clustered alternative but still showed good power. However, the MST-based test could not be used as a test against a regular alternative in this environment. The MST-based test gave acceptable performance on some real data. Of course, in high dimensions, one must have enough patterns to reliably assess the uniformity of the data.

It is unlikely that a single test will provide all the information needed to determine the structure of real data. We envision a number of tests, including a test for uniformity, being applied to the data and the results of these tests combined in some manner. There is a definite need to explore the strategies for doing this. Even if a data set is rejected as uniform it may still not be very interesting from the point of view of clustering; for example, the data may be unimodal Gaussian. Therefore, it is necessary to look more closely at this "nonuniform" data. For example, it would be extremely useful to know the number of clusters present in the data.

References

[1] J. L. Bentley *et al.*, "On the average number of maxima in a set of vectors and applications," *J. Ass. Comput. Mach.*, vol. 25, pp. 536-543, 1978.

[2] G. Biswas *et al.*, "Evaluation of projection algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-3, pp. 701-708, 1981.

[3] D. R. Chand and S. S. Kapur, "An algorithm for convex polytopes," *J. Ass. Comput. Mach.*, vol. 17, pp. 78-86, 1970.

[4] G. R. Cross and A. K. Jain, "Measurement of clustering tendency," in *Proc. IFAC Symp. Digital Contr.*, New Delhi, India, 1982, pp. 24-29.

[5] R. C. Dubes and A. K. Jain, "Clustering methodologies in exploratory data analysis," in *Advances in Computers*, vol. 19, M. Yovits, Ed. New York: Academic, 1980, pp. 113-228.

[6] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, pp. 697-717, 1979.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.

[8] D. E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, vol. 2, 2nd ed. Reading, MA: Addison-Wesley, 1981.

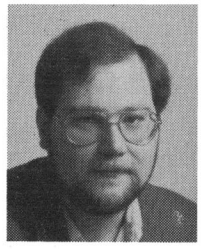
[9] E. Panayirci and R. C. Dubes, "A new statistic for assessing gross structure of multidimensional patterns," Dep. Comput. Sci., Michigan State Univ., East Lansing, Tech. Rep. TR81-04, 1981.

[10] B. D. Ripley and J. P. Rasson, "Finding the edge of a Poisson forest," *J. Appl. Probability*, vol. 14, pp. 483-491, 1977.

[11] B. D. Ripley, *Spatial Statistics.* New York: Wiley, 1981.

[12] S. P. Smith, "Structure of multidimensional patterns," Ph.D. dissertation, Dep. Comput. Sci., Michigan State Univ., East Lansing, 1982.

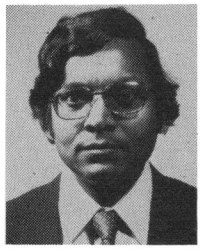
[13] D. L. Young, "The linear nearest neighbor statistic," *Biometrika*, vol. 69, pp. 477-480, 1982.



Stephen P. Smith (S'79-M'82) was born in Cincinnati, OH, on September 4, 1955. He received the B.S., M.S., and Ph.D. degrees in computer science from Michigan State University, East Lansing, in 1977, 1979, and 1982, respectively.

Currently at Northrop's Research and Technology Center, he is responsible for the application of artificial intelligence and pattern recognition technologies to problems in target recognition, sensor fusion, and industrial automation. He is principal investigator on research into knowledge-based systems for industrial inspection. From 1977 to 1982, he held the position of Graduate Research Assistant in the Pattern Recognition and Image Processing Laboratory at Michigan State University where he performed novel research in the areas of scene analysis, shape matching, and exploratory data analysis. During this time, he was also consulting at the Babcock and Wilcok Co. in the application of signal processing and pattern recognition techniques to nondestructive examination of nuclear reactor components. He has published a number of articles in technical journals and has presented talks at various technical conferences and research establishments.

Dr. Smith is a member of the Association for Computing Machinery and Phi Kappa Phi.



Anil K. Jain (S'70-M'72) was born in Basti, India, on August 5, 1948. He received the B.Tech degree with distinction from the Indian Institute of Technology, Kanpur, India, in 1969, and the M.S. and Ph.D. degrees in electrical engineering from Ohio State University, Columbus, in 1970 and 1973, respectively.

From 1971 to 1972, he was a Research Associate in the Communications and Control Systems Laboratory, Ohio State University. Then from 1972 to 1974, he was an Assistant Profes-

sor in the Department of Computer Science, Wayne State University, Detroit, MI. In 1974, he joined the Department of Computer Science, Michigan State University, where he is currently a Professor. He served as the Program Director of the Intelligent Systems Program at the National Science Foundation from September 1980 to August 1981. His research interests are in the areas of pattern recognition and image processing.

Dr. Jain is a member of the Association for Computing Machinery, the Pattern Recognition Society, and Sigma Xi. He is also an Advisory Editor of *Pattern Recognition Letters*.

K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality

SHOKRI Z. SELIM AND M. A. ISMAIL, MEMBER, IEEE

Abstract—**The *K*-means algorithm is a commonly used technique in cluster analysis. In this paper, several questions about the algorithm are addressed. The clustering problem is first cast as a nonconvex mathematical program. Then, a rigorous proof of the finite convergence of the *K*-means-type algorithm is given for any metric. It is shown that under certain conditions the algorithm may fail to converge to a local minimum, and that it converges under differentiability conditions to a Kuhn–Tucker point. Finally, a method for obtaining a local-minimum solution is given.**

Index Terms—**Basic ISODATA, cluster analysis, *K*-means algorithm, *K*-means convergence, numerical taxonomy.**

I. INTRODUCTION

K-MEANS-type algorithms for exploratory data clustering and analysis are very popular and well known [1]-[8]. The main idea behind these techniques is the minimization of a certain criterion function usually taken up as a function of the deviations between all patterns from their respective cluster centers. Usually, the minimization of such a criterion function is sought utilizing an iterative scheme which starts with an arbitrary chosen initial cluster configuration of the data, then alters the cluster membership in an iterative manner to obtain a better configuration. The sum of squared Euclidean distances criterion has been adopted in most of the studies related to these algorithms, due to its computational simplicity, since the cluster at each iteration can be calculated in a straightforward manner. A *K*-means algorithm alternates between two major steps until a stopping criterion is satisfied. These steps are mainly the distribution of patterns among clusters utilizing a specific classifier (usually the minimum Euclidean distance classifier: MEDC), and the updating of cluster centers [9]-[11].

Incorporation of some heuristic procedures into the above-mentioned iterative scheme results in the well-known ISODATA algorithm of Ball and Hall [12], which may be considered as another sophisticated form of the original *K*-means. The most important heuristic procedures in ISODATA are those allowing cluster lumping and cluster splitting.

Several extensive studies dealing with comparative analysis of different clustering methods have been conducted recently, utilizing both simulated data sets, e.g., [2], [13], and practical data, e.g., [6], [14]-[16]. These studies recommend the *K*-means algorithm as one of the best clustering methods available. Several evaluation criteria were adopted in such studies and a variety of techniques were compared simultaneously.

Moreover, fuzzy versions of the *K*-means algorithm have been reported in a series of papers by Ruspini [17], [18] Dunn [19], and Bezdek [19]-[22], where each pattern is allowed to have membership functions to all clusters rather than having a distinct membership to exactly one cluster.

The usefulness of the *K*-means-type algorithms is not questionable, and the extensive experimentation with these algorithms using practical data suggests and establishes the applicability and practicality of such techniques. Although it is found that such algorithms converge when applied to different data sets from a wide range of applications, no rigorous theorem for the convergence of the *K*-means-type algorithms exists to date, and the question of convergence of such methods remain open [1], [3], and [22].

In this paper, a rigorous proof of convergence of the *K*-means-type algorithm is given in a generalized form. Moreover, local optimality of solutions obtained has been investigated, where it is shown that under certain conditions, the *K*-means algo-

Manuscript received July 12, 1982; revised January 24, 1983. This work was supported by the University of Petroleum and Minerals, Dhahran, Saudi Arabia.

S. Z. Selim is with the Department of Systems Engineering, University of Petroleum and Minerals, Dhahran, Saudi Arabia.

M. A. Ismail is with the Department of Computer Science, University of Windsor, Windsor, Canada.

0162-8828/84/0100-0081$01.00