This is to certify that the

dissertation entitled

STRUCTURE OF MULTIDIMENSIONAL PATTERNS

presented by

Stephen Phillip Smith

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Computer Science

Major professor

Date August 18, 1982

STRUCTURE OF MULTIDIMENSIONAL PATTERNS

By

Stephen Phillip Smith

A DISSERTATION

Submitted to

Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

1982

ABSTRACT

STRUCTURE OF MULTIDIMENSIONAL PATTERNS

By

Stephen Phillip Smith

The problem of describing the structure of multidimensional data is important in exploratory data analysis, statistical pattern recognition, and image processing. We view a data set as a collection of points embedded in a high dimensional space. The primary goal of this research is to determine if the data have any clustering structure; such a structure implies the presence of class information (categories) in the data.

We wish to use a statistical hypothesis test in our decision making. To this end, we define data with no structure as data following the uniform distribution over some compact convex set in K-dimensional space, called the sampling window.

This thesis defines two new tests for uniformity along with various sampling window estimators. The first test is a volume-based test which captures density changes in the data. The second test compares a uniformly distributed sample to the data by using the minimal spanning tree (MST) of the pooled samples. We provide sampling window estimators for simple sampling windows and use the convex hull of the data as a general sampling window estimator.

For both of the tests for uniformity, we provide theoretical results on their size, and study their size and power by Monte-Carlo simulations. Both tests show good power against clustered alternatives. We also use simulation to study the efficacy of the sampling window estimators. These estimates perform well, but the convex hull estimator is too computationally burdensome to apply in high dimensions. Since the MST-based test can be performed without explicitly computing the convex hull of the data, we conclude that it is more reasonable to apply to real data. Experiments with some real data sets also demonstrate the power of the MST-based test.

ACKNOWLEDGEMENTS

I wish to thank my major professor, Anil K. Jain, for his help and guidance, both in thesis preparation and in his encouragement of my research efforts. Without him, my Ph.D. would never have been completed. Special thanks should also go to Professor Richard C. Dubes, who provides an excellent role model for all those associated with him. It was his interest in the evaluation of clustering methods that lead me to a study of the clustering tendency problem.

I thank Professors Carl Page and Joseph Gardiner for agreeing to serve on my Ph.D. committee. In addition, acknowledgement is due to all those graduate students and visiting professors who made working in the Pattern Recognition and Image Processing Laboratory such a stimulating experience: Dr. Tom Bailey, Dr. Eric Backer, soon to be Doctors Gautam Biswas and James Coggins, Dr. George Cross, Dr. Qichao He, Wei-Chung Lin, Dr. Erdal Panayirci, and Neal Wyse.

Acknowledgement should also be made of the financial support I recieved during my long stay at MSU from NSF grant ECS-8007106, DER, and John Flora at the Babcock and Wilcox Company.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Problem Statement

This thesis addresses the problem of describing the structure of multidimensional data. We are interested in data that are represented as points in a K-dimensional (K>2) space. We assume that little prior information about the data is available and we wish to make as few assumptions about the data as possible. This restricts us to a preliminary assessment of the structure and interrelationship among the points.

Figure 1 shows a number of data sets in two dimensions. Obviously, a complete description of some of these data sets would take considerable effort. However, the descriptors used would depend on the end goal. For some applications, it might be enough to know that the data in Figure 1(a) are 'uniform' and the data in Figures 1(b) and 1(c) are 'clustered', while other applications may be interested in knowing that the data in Figure 1(e) form an "S".

We are interested in a gross description of the data. We will try to decide if a data set has some 'unusual internal structure'. To do so, we will define data with no 'structure'. There would probably be

little interest in further analyzing data sets classified as unstructured.

There are three main questions which every Pattern Recognition researcher should be interested in asking about data. These are

(1) Do the data suggest, by their own internal structure, any 'classes' in the data?,

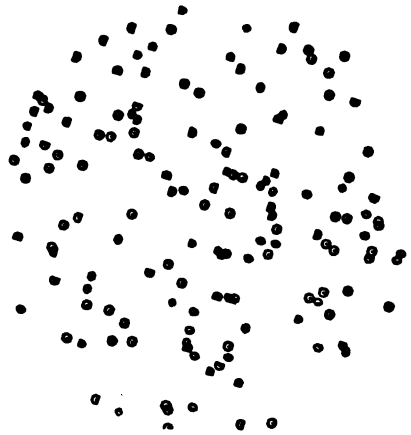
(2) Do these classes correspond to a priori pattern classes or to other extraneous factors in the application environment?,

(3) What measurements best extract the a priori pattern classes? What type of classifier best embodies these class distinctions and how can one best learn about the parameters of these classes?

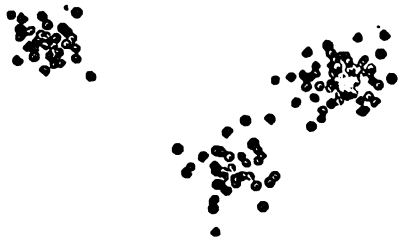
Classical Pattern Recognition theory deals mostly with the questions posed in (3). Some techniques are available in Exploratory Data Analysis and Pattern Recognition to gain answers to questions (1) and (2). This thesis deals with a way of providing information to answer question (1). We ask if there is any 'structure' in the data.

To make a decision of 'structure' versus 'no structure' for a particular data set, we would like to phrase this problem as a standard statistical hypothesis test. This compels us to define a stochastic model for unstructured data and one for structured data. Our stochastic model for unstructured data will be the continuous uniform distribution over some compact convex set in K-dimensional space, called the sampling window. Using this definition, the only data set in Figure 1 which is

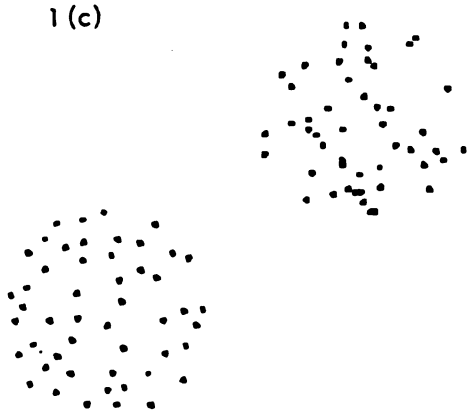
1 (a)



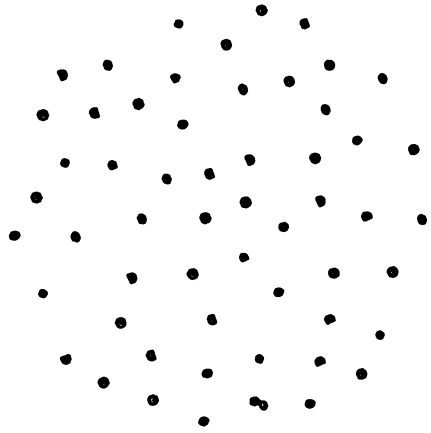
1 (b)



1 (c)



1 (d)



1 (e)

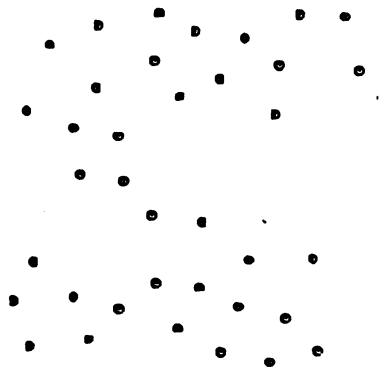


FIGURE 1: Data Sets Exhibiting Different Structures
(a) Uniform data
(b) Clustered data
(c) Two-Cluster data
(d) Regular data
(e) 'S' shaped data

unstructured is Figure 1(a). This is reasonable since it is the only data set in Figure 1 which has no meaningful higher level description than 'randomly dispersed data inside a circle'.

There are many possible alternatives to unstructured data. Since a primary motivation for this work is in assessing the 'clustering tendency' [Dub80, Cro80, Cro82] of a data set, an important stochastic model for structured data is one of clustering or aggregation. Clusters in the data would represent the 'classes' of interest to a Pattern Recognition researcher. The antithesis of clustered data is lattice regularity, shown in Figure 1(d). Under our definition, this regular data should also be categorized as structured, although this structure is not of significant interest in Pattern Recognition.

To decide if a data set is structured or unstructured using a statistical hypothesis testing paradigm, we need to define some test statistic which will capture this difference. The primary goal of this thesis is to find a test statistic whose distribution is known under the null hypothesis of uniformity and all possible alternative hypotheses for all dimensions and for all sampling windows. We will see that this goal is overly ambitious. We at least demand that the null distribution of the statistic be available with known sampling window. This allows one to set the size of the test based on the statistic. We also require that the statistic be applicable to high dimensional data. Little study has been done to evaluate test statistics when the sampling window is unknown, and we merely begin such a study here. Our basic method of study will be to perform a Monte-Carlo simulation of a statistic to

check its size and determine its power. We will use data over both known and unknown sampling windows.

In the remainder of this chapter, we will briefly present background on Pattern Recognition, the field from which this thesis originates. Since clustering techniques in Exploratory Data Analysis are aimed at providing information on the 'class' structure of data, we review some of these techniques. Our notion of structure is driven by, and closely related to, the concept of clustering tendency, and we define this concept. Finally, we give the organization of this thesis.

1.2 Pattern Recognition

Pattern Recognition techniques form the backbone of important methods used in the fields of machine intelligence and machine perception. Pattern Recognition can be defined as "the categorization of input data into indentifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail" [Gon78]. The categorization of input data is treated extensively in the book by Duda and Hart [Dud73]. We are primarily interested in testing if the data have any 'significant features' to recommend it for further study. In terms of the applications, research methods, and research techniques, this thesis is under the broad umbrella of Pattern Recognition. Thus we emphasize multi-dimensional data sets and are sensitive to computational considerations. We do not explicitly treat any application areas. We are concerned with, but hopefully not dominated by, theoretical issues in statistical analysis

[Bar75] and probability theory [Har74]. To facilitate further discussion, we now define some standard terms used in Pattern Recognition studies.

Pattern Recognition can be broken into two broad subfields: the geometric approach and the structural approach. The structural approach essentially views 'patterns' as complex parts formed from idealized simpler parts in the presence of distortion [Gre76, Gre78]. It can be further subdivided into grammatical techniques [Fu74] and heuristic techniques [Pav77], depending on how the parts and their relationships are described. The geometric approach, with which this thesis deals, views objects as being represented as points between which proximities are given or can be computed. Geometric Pattern Recognition can be further subdivided into statistical versus non-statistical approaches. We work in statistical Pattern Recognition. There are two forms of data presentation in statistical Pattern Recognition algorithms: the pattern matrix or the proximity matrix. In a proximity matrix, N patterns are represented by an N by N matrix, whose (i,j)th entry specifies the proximity (similarity or dissimilarity) between pattern i and pattern j. This type of data occurs most frequently in applications from the social and behavioral sciences. We deal with the pattern matrix, which is an N by K matrix, where each row is a pattern and each column denotes a feature. The K features are viewed as a set of orthogonal axes and each pattern is then seen to be a point or vector in a K-dimensional space called the pattern space.

Another dichotomy in Pattern Recognition is that of labeled patterns (supervised learning) versus unlabeled patterns (unsupervised learning). One may assign a priori labels to each pattern representing the 'class' to which that pattern belongs. This set of labeled patterns constitutes the training samples which can be used to learn the structure of each pattern class or determine the decision boundaries between the classes. If it is assumed that the patterns from a class follow some parametric statistical distribution, then we have a parametric statistical decision problem. Otherwise, we must either estimate the density function or use some non-parametric decision rule. We assume little information is available about the patterns in our data sets and, therefore, we work in the unsupervised learning mode. Further, we assume that we have no knowledge about the number of possible classes present in the data. Work in this mode can be categorized under the general heading of Exploratory Data Analysis, which is detailed in the next section.

1.3 Exploratory Data Analysis

Exploratory Data Analysis [Gna77, Tuk77] is a "generic term for a body of mathematical, statistical and heuristic operations whose goals are to help an investigator get acquainted with data taken at a preliminary stage of scientific inquiry" [Pan81]. As the word 'exploratory' implies, we are interested in a preliminary assessment of the gross structure of a data set, rather than confirming some

application-derived model of the data. The difficulty in an intuitive interpretation of data embedded in high dimensional space is obvious. Even in two and three dimensions the use of these techniques may result in a better and more systematic categorization of the data set than can be done by the naked eye. Also, the large volumes of such data that occur in numerous scientific fields necessitate computer processing.

The technique of Exploratory Data Analysis in which we are most interested is called clustering. Clustering attempts to find natural groupings of patterns in a data set such that patterns within groups are more 'similar' than patterns across groups. There are many clustering algorithms [And73, Eve74, Har75] and each essentially represents its own definition of what is meant by a 'natural' grouping. Techniques range from graph-theoretic clustering methods [Zah71] to minimum square-error clustering methods [And73]. One major problem with clustering algorithms is that they impose a clustering structure on the data set even if such structure is not inherent in the data. For instance, clustering algorithms will almost always find clusters in uniformly distributed data. Thus, quite often, clusters found in data are artifacts of the clustering method. We wish to avoid elaborate interpertation of uniform data, and so we will refuse to apply clustering algorithms to any 'unstructured' data. This is essentially the methodological paradigm involving assessing the 'clustering tendency' of the data, set forth by Dubes and Jain [Dub80]. For other problems in Cluster Analysis see Everitt [Eve79] and Dubes and Jain [Dub76, Dub79, Dub80].

1.4 Clustering Tendency

The term 'clustering tendency' refers to the problem of deciding whether the data exhibit a predisposition to cluster, in other words to form natural groups. We are interested in assessing if the structural arrangement of the points is unusual, either on the side of aggregation of the data, or on the other extreme when the data is aligned in a near lattice arrangement. Basically, clustering tendency assessment implies categorizing a given data set into one of the following three broad descriptions:

(1) data are arranged randomly

(2) data are aggregated

(3) data are regularly spaced.

1.4.1 Proximity Matrix

Most of the work in clustering tendency assessment reported in the literature deals with proximity matrix data. The entries in the proximity matrix are rank ordered, that is only the ranks of the similarities are meaningful. The null hypothesis of randomness is stated to mean that all proximity matrices are equally likely. This is called the Random Graph null hypothesis by Dubes and Jain [Dub80], since there is a one to one correspondence between an N by N rank order matrix

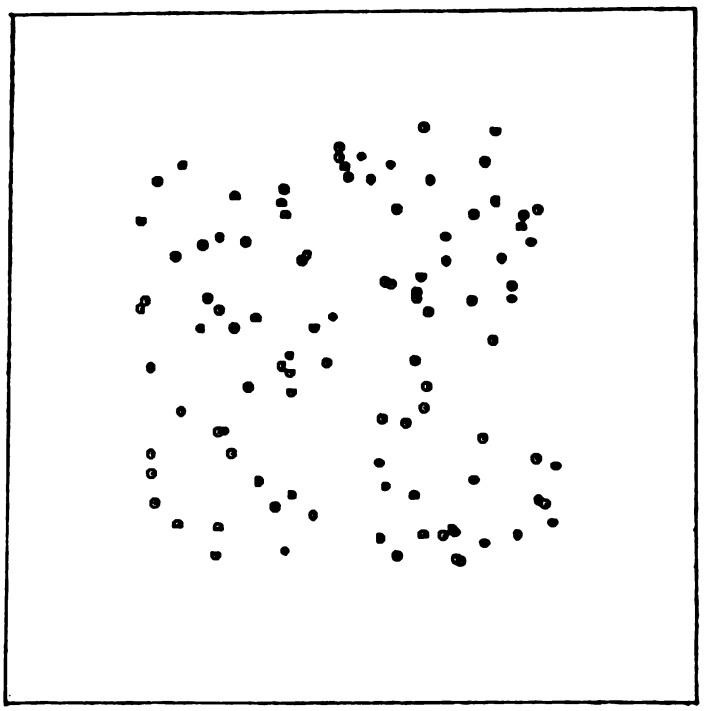
and an undirected, weighted, labeled graph on N nodes.

Statistics used to assess clustering tendency under the Random Graph null hypothesis include the number of edges needed to connect the graph [Fil71, Lin75], the distribution of node degrees in a threshold graph [Fil71], the number of cycles in such a graph [Fil71], and the number of nodes with incident edges in a threshold graph [Lin73]. However, Bailey [Bai78] points out that the Random Graph null hypothesis is inappropriate for points distributed randomly is space. This is because the metric space in which the points lie impose some additional constraints on data configurations.

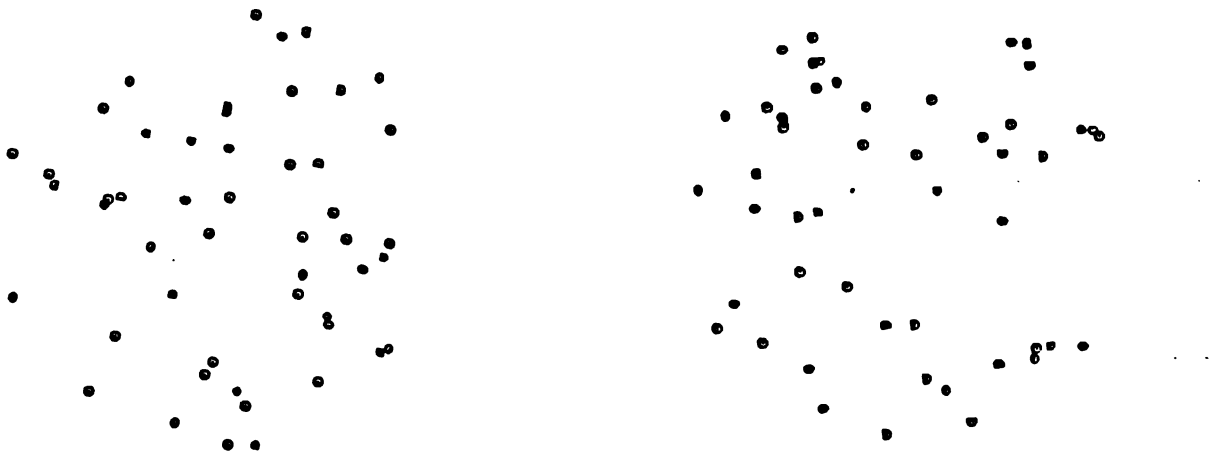
1.4.2 Pattern Matrix

The null hypothesis of no structure here is the continuous uniform distribution over some compact convex set $S \subset R^K$. This null hypothesis can also be viewed as a spatial Poisson process over R^K restricted to set S. The set S is called the sampling window. Thus a <u>sampling window</u> can be defined as the compact convex support set for the underlying distribution.

The crucial role of the sampling window in assessing the structure of a set of patterns can be seen from Figure 2. Figure 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, and hence the data has no structure. However, if for some a priori reason the unit square



2 (a)



2 (b)

FIGURE 2: Importance of the Sampling Window
(a) Data uniform over small subsquare
(b) Data uniform over two disjoint circles

is taken as the sampling window, then the data would have to be considered as structured in this sampling window. Perhaps one would wish to call it a single cluster in the middle of the unit square. The need for a convex sampling window is shown in Figure 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 convex sets.

The statistical test of hypothesis can thus be stated as:

Ho: The data are uniform over the sampling window
versus
H1: The data are not uniform over the sampling window.

The difficulty of testing uniformity of a pattern matrix is twofold. First, the sampling window is unknown and must be estimated from the data. Second, the test for uniformity must be performed in the K-dimensional space. The distribution of uniformly distributed points which are projected into a lower dimensional space by any of the popular projection algorithms [Bis81] is unknown in the projected space. Further, checking only for marginal uniformity may not be sufficient. As an example, consider the following non-uniform density function over the unit square whose marginal densties are uniform.

$$f(x,y) = 4(x^2 - x - y^2 + y)\ (2y - 1)\ (2x - 1) + 1 \qquad \text{for } 0 \le x,y \le 1$$
$$= 0 \qquad \text{otherwise.}$$

1.4.3 Spatial Point Processes

We now provide a mathematical framework to introduce spatial point processes and, in particular, the Poisson process [Rip77, Cox80, Ish81]. We imagine a probabilistic mechanism scattering points throughout K-dimensional Euclidean space. Each realization of the process is a countable number of points over the space. The important random variables are N(B), where B is a Borel subset of R^K, and N, which is a measurable mapping from the Borel sets into the natural numbers and counts the number of points in B. A model of a point process determines the distribution of N(B) for all Borel subsets of R^K. The intensity, L, of a homogeneous process is the expected number of points per unit volume and summarizes the first moment structure of the family {N(B)}.

Let u(.) denote K-dimensional Lebesgue measure. For a Poisson process, we demand (i) N(B) has a Poisson distribution with parameter L•u(B) for all bounded Borel sets B and (ii) {N(Bi)} is a set of independent random variables whenever {Bi} is a class of disjoint sets. In this case, the model is completely determined by the parameter L. A Poisson process restricted to a bounded set, such as the sampling window S, generates the continuous uniform distribution over S, conditioned on N(S).

For an alternative hypothesis we have a number of choices. An example of a clustered process is the classic Neyman-Scott process [Ney72]. Details of this process are given in Appendix A. For a regular alternative, we use a hardcore or inhibitory model [Mat60, Rip77] which is also described in detail in Appendix A. Strauss [Str75, Kel76] shows a theoretical relationship between many of these point process models.

When a Poisson process is used as a null hypothesis and the distribution of a statistic is derived under this assumption, one must decide how to approach data in some sampling window $S \subset R^K$. This is because 'edge effects' which arise from having a bounded sampling window can invalidate the distribution of the statistic [Rip81]. These edge effects become increasingly dominant as dimensionality increases. Some statistics (such as Ripley's D(t), mentioned later) contain their own edge correction factors. There are two general approaches to this problem:

(1) Analyze points only inside $W \subset S$ but allow measurements between the points in W and those that remain in S-W. In general, one does not know the relative size of W as compared to S needed to eliminate edge effects. This has been called the border, or guard area, method of edge correction for obvious reason.

(2) A hyper-rectangular sampling window can be regarded as a torus, so that opposite faces are considered to be close. Thus interpoint distances can 'wrap around' the boundaries of the

hyper-rectangle. This is the so called wrap around method of edge effect correction.

Most of the studies which deal with the null hypothesis of a Poisson process (described in Chapter 2) have used the wrap around method of edge correction.

1.5 Organization of the Thesis

Chapter 2 contains a literature review of tests for assessing structure in a data set. Chapter 3 introduces a new test, called the volume-based test. The theory underlying the test is also given. The volume-based test requires precise knowledge of the sampling window so Chapter 4 studies estimators for various types of windows. Chapter 5 presents experimental results when using the volume-based test over both known and unknown sampling windows. Since the conclusion of Chapter 5 is that the volume-based test is not computationally feasible in high dimensions with unknown sampling window, Chapter 6 presents a new test, called the MST-based test, which handles this case. Finally, Chapter 7 presents the contributions of this thesis, our conclusions, and suggestions for future research.

CHAPTER 2

TESTS FOR STRUCTURE IN DATA

2.1 Introduction

The problem of deciding if data have structure has been addressed, in a slightly different format, in both the ecological literature [Pie77] and the geographical literature [Rog74]. The recent book by Ripley [Rip81] provides a good overview of the statistical methods used. Basically, both fields are interested in testing if there is some non-random mechanism at work in the spatial distribution of the populations under study. Unfortunately, both fields deal with points in two dimensions and normally assume that the sampling window is known. These two assumptions are rarely valid in Pattern Recognition studies. One problem with many of the tests for spatial arrangement is that the distribution of the test statistic even under the null hypothesis of uniformity is not known [Rip77]. In some instances, when the distribution is known, it is applicable only under the assumption of an infinite Poisson process. In applying these tests to a finite sampling window, edge effects dominate, especially in high dimensions. Here we give a brief overview of tests used in clustering tendency, keeping in mind the need to extend the tests to higher dimensions.

2.2 The Scan Test

Tests for structure based on the number of points in the most populous region of the sampling window are intuitively appealing. An abnormally large count would indicate the presence of clustering. The size of the region must, for statistical reasons, be fixed a priori and either a continuous scan (overlapping windows) or a disjoint partition of regions is used. The choice of region size is not obvious. The model of randomness is a uniform distribution over the sampling window. Attempts have been made, mostly for the one-dimensional case, to derive the null distribution of such statistics [Nau66, Wal74]. Unfortunately, even in one dimension, determining the size of the scan test is computationally infeasible. Conover et. al. [Con79] and Naus [Nau65] have attempted to apply this test to two dimensions. It does not appear possible to extend this test to high dimensions.

2.3 Quadrat Analysis

The basic idea of the quadrat method [Rog74] is simple. We divide the sampling window into squares of equal size (hypercubes in K dimensions), called quadrats, and record the number of points which fall in each quadrat. A data set containing a regular arrangement of points would be expected to generate relatively equal quadrat counts, an aggregated data set would generate a few quadrats with most of the points and a uniform data set would lead to a situation somewhere

between these two extremes.

As quadrats are disjoint and of equal volume, the set of counts should follow a Poisson distribution under the null hypothesis of no structure. Typically a Chi-squared test is performed to determine if this hypothesis holds. A significant drawback of the quadrat test is its inability to detect and test spatial arrangement at more than one scale, set by the quadrat mesh. The Grieg-Smith approach [Gri64] and Mead's approach [Mea74] are attempts to correct this deficiency. Another problem with quadrat tests is that the number of quadrats becomes enormous in high dimensions, most of them being empty.

One possible solution to this problem is the use of transect sampling. Transects are narrow tubes inserted at random through the data. Counts are taken only on data that fall within these tubes, thus providing a linear strip of counts. Cross [Cro80] discussed this possibility. Unfortunately, because of the sparseness of data in practical situations, transect sampling rarely provides adequate information for assessing the structure of the data.

2.4 Second Moment Estimators

Another class of tests for structure rests on computing an estimate of the variance of a point process. As shown by Ripley [Rip77], the second moment structure of a process may be reduced to a function $D(t)$ defined on $(0,\infty)$ such that, for a process with intensity L, the following properties hold.

(1) $L^2 \cdot D(t)$ is the expected number of ordered pairs of distinct points less than distance t apart when the first point is in a given set of unit area, and

(2) $L \cdot D(t)$ is the expected number of additional points within a distance t of an arbitrary point in the process.

Ripley provides an unbiased estimator of D(t) for a sample containing N points, given by

$$\hat{D}(t) = (N)^{-2} \sum k(x,y),$$

where the sum is over ordered pairs of points (x,y) closer than a distance t. Here k(x,y) is an edge correction factor such that 1/k(x,y) is the proportion of the boundary of the hypersphere centered at x and passing through y which is within the sampling window S. Unfortunately, very little is known about the sampling fluctuations of $\hat{D}(t)$ even in two dimensions. Ripley [Rip77] resorts to Monte-Carlo simulations of $\hat{D}(t)$ for fitting models to data, while in [Rip79] he uses the maximum deviation of a normalized version of $\hat{D}(t)$ from its expected value.

In two dimensions, Liebetrau and Rothman [Lie77, Lie77b, Lie78] use estimates of Var[N(C)]/E[N(C)] for a rectangle C aligned within a rectangular sampling window S, where N(C) is the number of data points in rectangle C. Their test statistics are

$$\sum Q(x,y) \quad \text{and} \quad \sum Q^2(x,y),$$

where the sums are over all pairs of points (x,y) and

$$Q((x1,x2),(y1,y2)) = [\ c1 - |x1-y1|\ ] * [\ c2 - |x2-y2|\ ].$$

The function [c] denotes the maximum of c and zero. The values of c1 and c2 determine the size of the rectangle. Again the choice of c1 and c2 is critical, though the authors are able to show asymptotic joint normality of the test statistic using various values of c1 and c2 simultaneously. It is unclear how to extend this test to other types of sampling windows.

2.5 Distance-Based Tests

The existence of a 'structure' in a given set of points could be defined based on some interrelationship among the points that has unexpected characteristics [Moo74, Ala81]. One gross measure of such a structural relationship is simply the interpoint distances. While the use of interpoint distances, without additional information such as which point pair generated which distance, may not capture important details, the use of interpoint distances has much appeal. First, distances are invariant under the group of Euclidean motions which is consistent with our intuitive notion of a 'structure' as being invariant under rotation and translation. Second, interpoint distances are easy to compute in K dimensions.

2.5.1 Using All Interpoint Distances

The naive way to use these distances is to compute all N(N-1)/2 interpoint distances, find the resulting histogram or the empirical distribution function, and compare this function with the distribution

function under the null hypothesis of uniformity (or any other hypothesis). This procedure runs into two main problems. First, the theoretical distribution, which depends on the size and shape of the sampling window, is unknown even for simple shaped sampling windows like a hypersphere. Another problem is that we have no means of testing the equality of the empirical distribution and the theoretical distribution. This is because the known non-parametric tests (K-S or Chi-squared) assume that the sample points are independent. This is certainly not the case for all the interpoint distances. Therefore, we do not know the level of a test based on, say, the K-S test statistic. One solution to these problems is suggested by Bartlett [Bar64] who adjusts the critical values of the Chi-squared test statistic based on the correlation among the distances. Another solution to these problems would be to use Monte-Carlo techniques to compute the exact significance level of the test [Dig79].

2.5.2 Using Subsets of Distances

The joint distribution of all the interpoint distances for N points is unknown in an arbitrary sampling window. However, we know the distribution of the distance between two points placed at random in a hypersphere [Ham50, Lor54, Ala76]. This distribution can also be worked out when two points are placed randomly in a hypercube, though the derivation is tedious. For a general sampling window the results appear out of reach. An obvious simplification over using the actual distribution of interpoint distances among N points is to use the theoretical distribution of distance between two random points and

ignore the dependencies. Cross [Cro80] has shown that this method leads to spurious rejections of the null hypothesis of uniformity. Another possible procedure [Cro80] is to select independent distances from all N(N-1)/2 interpoint distances. Cross shows that the key factor in using such a test is the sampling window. Since the known theoretical distribution is between two points in a hypersphere, the data set in question must somehow be scaled to fit into a hypersphere. How such a scaling should be done is an open question.

Another approach when using distance-based methods is to observe only the small interpoint distances. This has intuitive appeal since the interpoint distance distribution should be flat near zero when the points are regularly spaced, should have a mode near zero when the points are clustered and should have a shape between these two when the points are uniform. Also, as Ripley [Rip78] mentions, the minimum interpoint distance can be shown to be the Uniformly Most Powerful test of uniformity against a hard-core alternative. Using small distances allows us to derive asymptotic distributions for some test statistics [Sau77, Sil78]. The only extension of these tests to K dimensions (K>2) is by Smith and Dubes [Smi81]. Unfortunately, the reliance on asymptotic theory makes the applicability of these tests to real data doubtful.

Other subsets of the total N(N-1)/2 interpoint distances can be used in summarizing the structure of the data and defining tests for uniformity. Generally speaking, such subsets are chosen on the basis of their being part of some structural relationship between the points.

For instance, the most popular subsets are some form of near neighbor distances. Given points from a Poisson field, we know the joint distribution of the distances from an arbitrary point in the space to its first M nearest neighbors among these points [Cro80]. For a finite number of points over a bounded region, one needs to modify the distribution to take into account the edge effects and the effects of near neighbors common between points. In the literature, generally, only the nearest neighbor information has been used. Clark and Evans [Cla54] suggest a statistic based on the average nearest neighbor distance among the sample points. With corrections to reduce interdependence and edge effects given by Ripley [Rip79] for two-dimensions, this statistic approximately follows a standard normal distribution.

Brown [Bro75] and Brown and Rothery [Bro78] suggest the coefficient of variation of the squared nearest neighbor distances and the ratio of the geometric mean to the arithmetic mean of these distances as possible test statistics. This could be extended to K dimensions by taking the Kth power of these distances. No adequate approximation to the sampling distribution of these statistics is known, though the asymptotic results of Silverman and Brown [Sil79] could be used on the small near neighbor distances. Also, simulation studies [Rip79] indicate that these statistics are not powerful against clustered alternatives.

Near neighbor distances have the unfortunate characteristic of capturing only the local structure. This occasionally makes deriving theoretical results possible but has the disadvantage that much of the

structural information about the data set is ignored. Perhaps distances from the minimal spanning tree [Har72] of the data set or from other structural graphs such as the Delaunay tesselation [Ahu81] or the Relative Neighborhood graph [Tou80] may be of interest in capturing more global information. However, even in the two-dimensional case, no firm results are available on the use of such structural graphs.

2.5.3 Sampling Origins

One technique that overcomes some of the inadequacies of using the nearest neighbors between the sample points is the use of sampling origins. Sampling origins are distinguished points fixed by the researcher in the sampling window, usually at random. The need to know the sampling window to make this technique meaningful is obvious. Several statistics using nearest neighbor distances between sampling origins and data points and nearest neighbor distances between data points are available. Diggle et. al. [Dig76] and Hines and Hines [Hin79] give extensive simulation studies on the performance of these statistics in the two-dimensional case. Cross [Cro80] extended this to higher dimensions.

In these studies, one statistic that showed high power against clustered alternatives is the Hopkins statistic [Hop54]. Cross and Jain [Cro82] study the performance of the Hopkins statistic in high dimensions. Let {Yi} be M sampling origins placed at random in the sampling window and let {Xi} be the N data points. Let Uj be the minimum distance from Yj to points in {Xi}, j=1,2,...M. Let Wj,

j=1,2,..,M be a random sample of size M from the N nearest neighbor distances among the data points. Under the null hypothesis of a Poisson process {Uj} and {Wj} have identical distributions. The Hopkins statistic is given by

$$\sum_{j=1}^{M} \frac{(Uj)^K}{(Uj)^K + (Wj)^K}$$

which has a Beta distribution with parameters (M,M) under the null hypothesis.

Panayirci and Dubes [Pan81] present a detailed study of the extension of another statistic, called the Cox-Lewis statistic [Cox76], to K dimensions. The Cox-Lewis statistic measures second-order information from the data in the following manner. First, it computes the distance between a sampling origin and the origin's nearest neighbor, say Xi, among the data points. It then computes the distance from Xi to its nearest neighbor among the remaining data points. By properly normalizing the ratio of these two distances Panayirci and Dubes obtain a statistic that follows the uniform distribution on the interval [0,1] for the null hypothesis of a Poisson process. To obtain information from more than one local area, several sampling origins are used. The Cox-Lewis statistic is then the average of the normalized distance ratios for a number of sampling origins. It appears to be as powerful in detecting clustering as the Hopkins statistic [Pan81].

The null distribution of both Hopkins and the Cox-Lewis statistics relies on a Poisson process null hypothesis. This effects their usage in the following ways. First, all the distances measured for a sampling origin are assumed to be independent from those for other sampling origins. In finite data sets this implies that the number of sampling origins be small. Cross and Jain [Cro82] suggest choosing M to be equal to 5% of N for the Hopkins statistic. Panayirci and Dubes [Pan81] use this choice in their simulation study of the Cox-Lewis statistic. Second, to reduce edge effects, both studies have used data over a hyper-rectangular sampling window with wrap around. The near neighbor distances were also computed using the wrap around method of edge correction.

2.6 Summary

We have reviewed tests for spatial randomness and clustering tendency. We wish to use such tests to determine the structure of high dimensional data. Many of the proposed tests are inadequate for our application. First, they may not be extendable to high dimensions or to situations when the sampling window is unknown. In fact, no study has been made on the effect of unknown sampling window for any test except the brief study by Cross [Cro80] for tests based on interpoint distance distributions. Many statistics have an unknown null distribution. Few deal with the null hypothesis of uniformity and rather choose to use a Poisson process null hypothesis which invariably leads to problems

involving edge effects and sample size. The next chapter introduces the volume-based test which is able to deal with both edge effects and sample size.

CHAPTER 3

THE VOLUME PARADIGM AND TEST

3.1 Introduction

We have seen in the previous chapter that most of the tests for spatial randomness have some limitations and restrictions. Much of the distributional theory available for the test statistics is either asymptotic or it deals with data from a planar Poisson point process. Reliance on asymptotic distributional theory forces one to include heuristic 'edge effect' correction factors when computing a statistic. This may limit one to rectangular sampling windows where torus wrapping can be accomplished. In Pattern Recognition applications, this wrap around is not appropriate. Also, extension of these tests to high dimensions is not straightforward.

In this chapter we propose a test that

(a) is applicable in all dimensions and to all sampling windows,

(b) has an exact null distribution known for all sample sizes, and

(c) eliminates the need for edge effect correction.

The above properties require that our null hypotheis of randomness be the continuous uniform distribution over the sampling window. This is

of course equivalent to having a Poisson process restricted to the sampling window, conditioned on N, the number of data points.

Testing for uniformity rather than a Poisson process insures two things. First, it allows us to deal directly with N and second, the null hypothesis of uniformity involves the sampling window (the set over which the uniform density is nonzero). This eliminates the need to apply an edge correction factor.

This chapter first presents the theorem from which various tests for uniformity can be defined. It then gives some examples of tests, and describes a test, called the volume-based test, that will be used in following experiments. The chapter concludes with a discussion of the factors that must be considered when using the volume-based test.

3.2 Volume Paradigm

The defining property of a uniform sample of points is the equidensity of points throughout the sampling window. To test uniformity, we wish to measure the change in this density over the sampling window. The main question one must answer in trying to use density to test the uniformity of a given sample of points is how the density is expected to change under uniformity.

Since density changes are volume related, we choose to use a certain sequence of volumes in our test. The volume-based test is derived from the following theorem which creates a paradigm for various

tests of K-dimensional uniformity. The theorem tells us the distribution of the volumes of certain sets for random data.

Theorem 1: Let $\{Xi\}_{i=1}^{N}$ be i.i.d. random vectors with the uniform distribution over sampling window $S \subset R^K$. Let u be K-dimensional Lebesgue measure. Let $< Wz \mid z\in(0,\infty) >$ be an ordered class of subsets of R^K such that,

(1) for all $z\in(0,\infty)$, $Wz \subseteq S$,

(2) for z_1, z_2, $z_1 \leq z_2$ if and only if $Wz_1 \subseteq Wz_2$ and

(3) there exists a function F: (0,u(S)) --> $(0,\infty)$ such that if F(y) = z then u(Wz) = y. Thus, given a particular value for volume, we can pull out the set in the sequence with that volume.

Let V(Xi) = Wz where $z = \inf\{ z \mid Xi\in Wz \}$, i=1,2,..,N. In other words, V(Xi) is the first subset in the sequence <Wz> which contains Xi.

Then { u(V(Xi)) } is a set of random variables which are i.i.d. uniform over (0,u(S)).

Proof:

Since the {Xi} are i.i.d., we need only prove that $u(V(X_1))$ has the uniform distribution over (0,u(S)). That is we need to prove

$$P[\, u(V(X_1)) \leq y \,] = y / u(S) \quad \text{for } y\in(0,u(S)).$$

For all $y\in(0,u(S))$, the existence of F guarantees the existence of

a Wz in $\langle Wz \rangle$ such that $u(Wz)=y$. In fact, this index z is $F(y)$. Also, from the definition of $V(X_1)$ and property (2), for $z \in (0,\infty)$, $X_1 \in Wz$ if and only if $V(X_1) \subseteq Wz$. Thus $X_1 \in Wz$ if and only if $u(V(X_1)) \leq u(Wz)$. Finally, since X_1 is uniformly distributed over S and for all z, $Wz \subseteq S$ then,

$$P[\, X_1 \in Wz \,] = u(Wz) \,/\, u(S).$$

So, for all $y \in (0,u(S))$,

$$\begin{aligned} P[\, u(V(X_1)) \leq y \,] &= P[\, u(V(X_1)) \leq u(W_{F(y)}) \,] \\ &= P[\, X_1 \in W_{F(y)}] \\ &= u(W_{F(y)}) \,/\, u(S) = y \,/\, u(S). \end{aligned}$$

QED.

In other words, the theorem states the following. We have a sequence of monotone increasing subsets in the sampling window S. The sequence is further constrained by the fact that for each volume from zero to u(S) we can choose the element in the sequence with that volume. If we then associate with each data point the first subset in the sequence that contains the point and measure the volume of this subset, this volume is then a uniform random variable on the interval (0,u(S)). Further, the volumes associated with the data points are independent. Thus we have taken uniform random vectors in K dimensions and transformed them into uniform random variables in one dimension. Of course, the theorem does not tell us the key point in making this transformation: how to define the sequence of subsets <Wz>. We deal with that next.

3.3 Examples of the Volume Paradigm

We now give some applications of Theorem 1, showing how the sequence of subsets may be defined. As before, let {Xi} be i.i.d. uniform random vectors over sampling window S.

3.3.1 Marginal Uniformity in a Hypercube

Let $S = [0,1]^K$, the unit hypercube in K dimensions. Define

$$Wz = \{ X=(x_1,x_2,\ldots,x_K) \in S \mid x_1 \leq z \} \text{ for each } z \in (0,1).$$

Note that the conditions of the theorem hold since we can define F to be the identity function. With <Wz> defined in this way we get $V(Xi) = Wx_{i1}$, where x_{i1} is the first coordinate of Xi. Thus u(V(Xi)) is just the value of the first coordinate of Xi. Then the theorem states that given random vectors i.i.d. uniform over the unit hypercube, the first coordinate of each of the vectors is a uniform random variable between zero and one. This is a trivial and unexciting result, but it shows the generality of the theorem.

3.3.2 Uniform Volumes about a Point

Let S be an arbitrary sampling window and let P be a point in S. Define

$$Wz = \{ X \in S \mid ||X-P||_2 \leq z \} \text{ for each } z \text{ in } (0,\infty),$$

where $||.||_2$ is the Euclidean distance metric. Then F can be taken as the inverse of the function that relates a radius about P to the portion of the volume of the hypersphere with that radius that is inside the sampling window.

Note that for small radii the hypersphere about P may be wholely contained in S and this function is analytically derivable. However, as the radii increase, the relation between distance and volume may not be amenable to analysis for arbitrary S.

When <Wz> is defined in this manner, V(Xi) is then $\{ X \in S \mid ||X-P||_2 \leq ||Xi-P||_2 \}$, i=1,..,N. See Figure 3 for an example. Then by Theorem 1, { u(V(Xi)) } is a set of i.i.d. uniform random variables on (0,u(S)). Note that any distance metric could have been used in defining <Wz> without altering the result.

This approach is similar to the use of the joint distribution of the first M near neighbors distances of a given point P in a Poisson process [Cro80]. The advantages of using volumes rather than distances is twofold. First, the results are exact for N points following a uniform distribution over any sampling window, rather than for an infinite Poisson process. Second, unlike distances, the sequence of volumes are independent random variables, which simplifies their joint distribution.

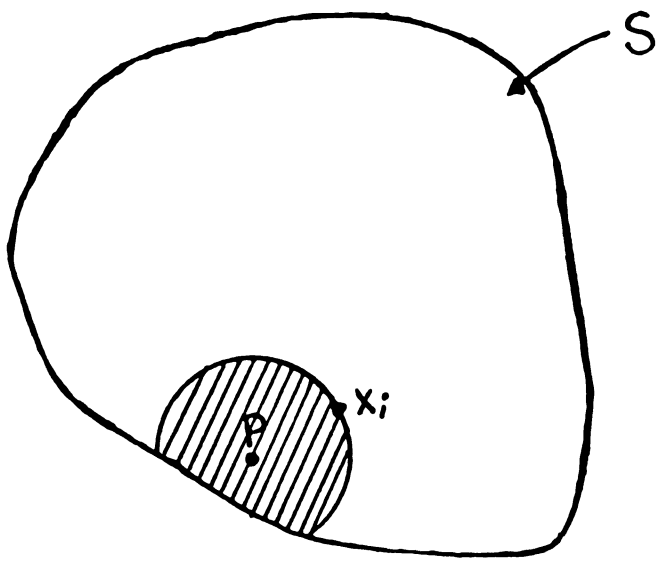


FIGURE 3: Definition of V(Xi).
The shaded area is V(Xi) for point Xi in sampling window S using the Euclidean metric.

3.3.3 Uniformity from the Border of the Sampling Window

Let S be a arbitrary sampling window. Define

$$Wz = \{ X \in S \mid \inf ||X-Y|| \leq z \} \text{ for all } z \in (0,\infty)$$

where the infimum is taken over all Y in the complement of S. Thus Wz is the set of all points in S within a distance z of the boundary of S. The conditions of the theorem are satisfied if F is taken to be the inverse of the function which relates this distance z to the volume of Wz. Then

$$V(Xi) = \{ X \in S \mid \inf ||X-Y|| \leq \inf ||Xi-Y|| \} \text{ for all } Xi,$$

where the infimums are over the complement of S. So, by Theorem 1, $\{u(V(Xi))\}$ is a set of i.i.d. uniform random variables over $(0,u(S))$.

3.4 The Volume-Based Test

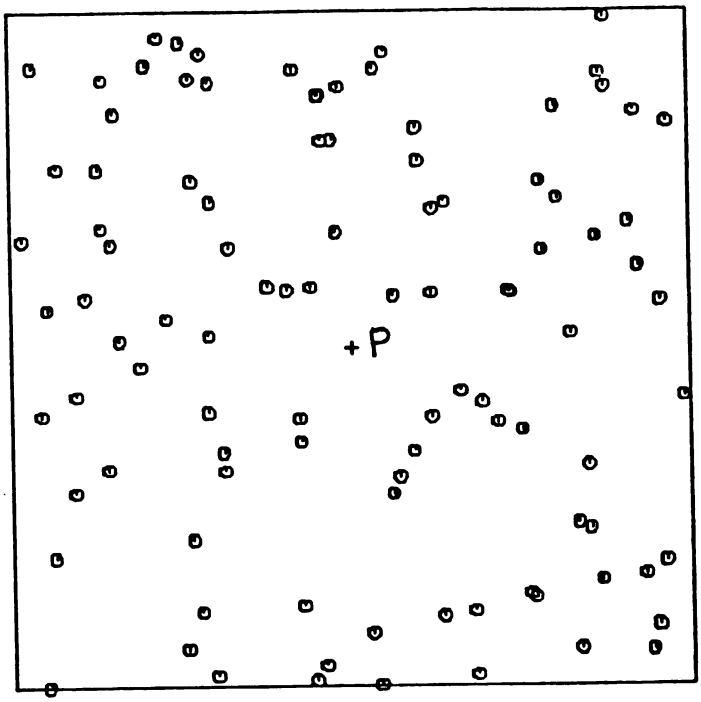
In the preceding examples we used a distance as the index parameter to the sequence of subsets. This provides a meaningful interpretation of the sets V(Xi). In Section 3.3.1, the distances were from a line to Xi, in the next example (Section 3.3.2), from a point to Xi and finally in Section 3.3.3 from a (possibly) complicated K-1 manifold to Xi.

For practical reasons, we limit ourselves to a test for uniformity against a general alternative based on the example of Section 3.3.2. That is, given some point P in the sampling window, we take a ball of

radius ||Xi-P|| centered at P and measure the volume of the intersection of this ball with the sampling window. According to Theorem 1 the set of volumes for all the Xi is a set of i.i.d. random variables uniformly distributed between zero and the volume of the sampling window. The Kolmogorov-Smirnov test is used to determine if this set of volumes is uniformly distributed.

This application of the theorem yields a test that is simpler than, say, that in Section 3.3.3, where computing the infimums is quite complicated. Further, it can be generalized to various sampling windows, unlike the application in Section 3.3.1. This test is still computationally expensive because of the need to compute the volumes of the intersections of sets.

The proposed volume-based test is intuitively appealing since it effectively measures the density of the data points near the point P and the density of the points far from P. Figure 4 shows 200 points generated uniformly inside the unit square. Also shown is the graph of ordered volumes about P (where distance from point P is measured by the supremum metric) versus the total number of points captured in this volume. Note the linear nature of this graph. For clustered data there are two possibilites. If P is placed at the center of a cluster then the points in the cluster will be abnormally close to P, thus generating smaller volumes, as shown in Figure 5. If P is placed outside a cluster, in a region of low density, we expect to see a few small volumes followed by an aggregation of volumes generated by points in the cluster. Figure 6 shows 200 points with one cluster in the center of



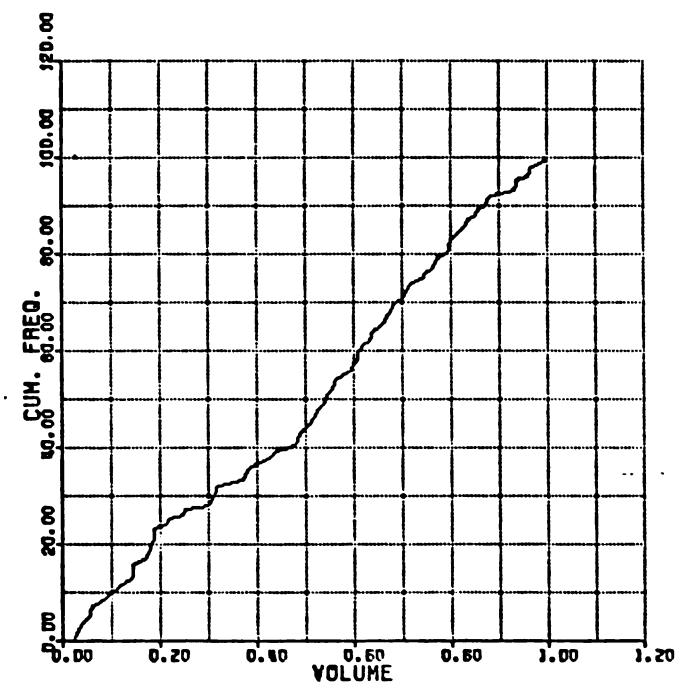
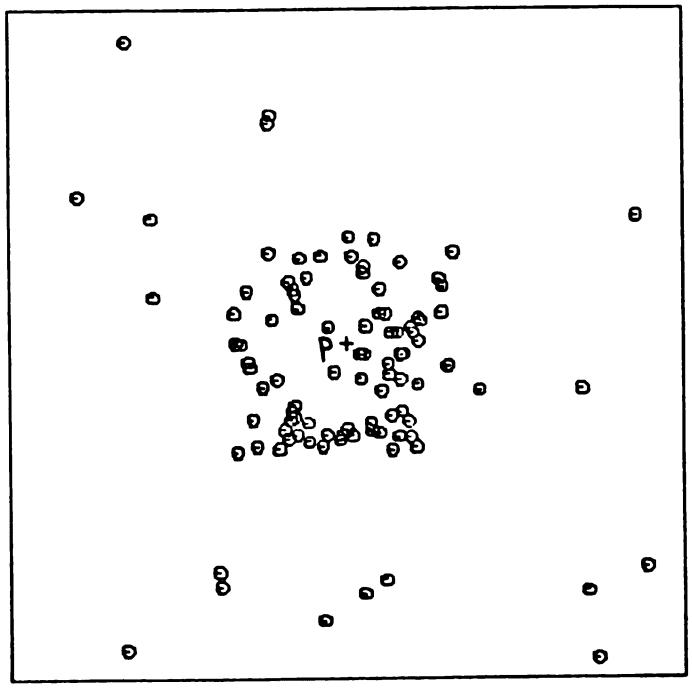


FIGURE 4: Uniform Data and its Volume Graph.
Point P is shown as "+"



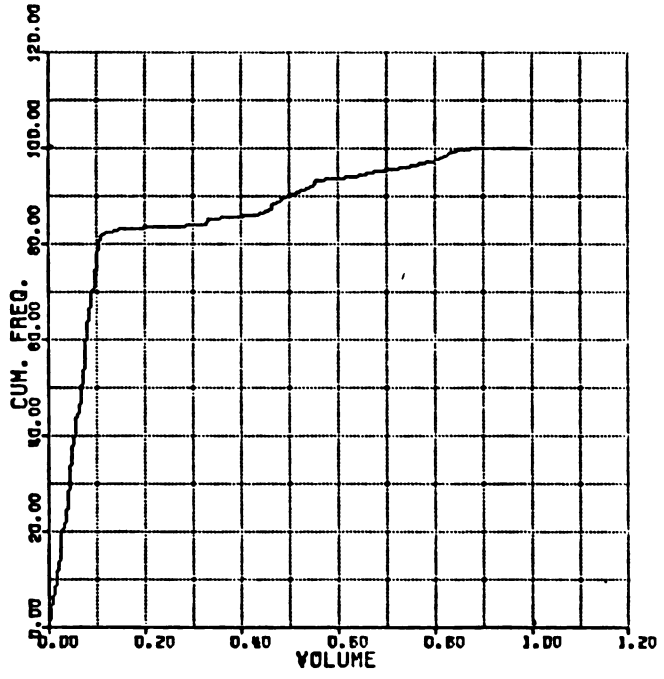
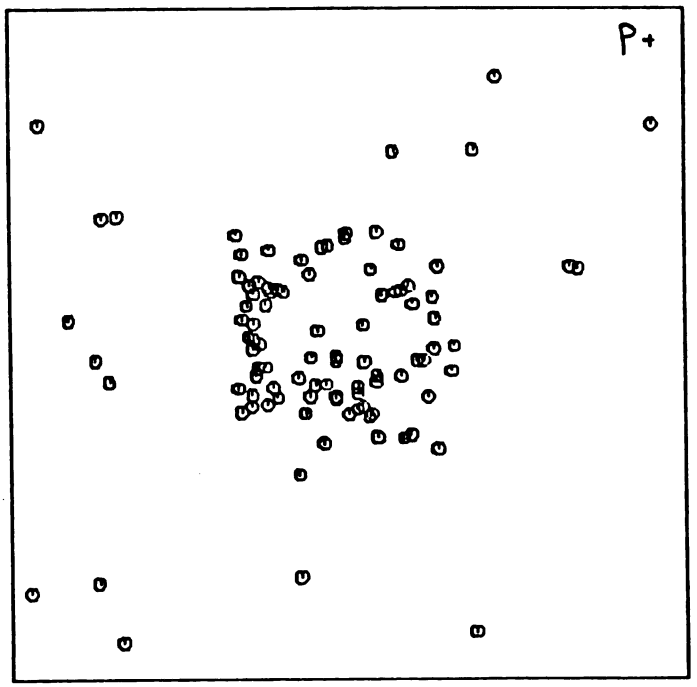


FIGURE 5: Clustered Data and Volume Graph with P inside the Cluster, Point P is shown as "+"



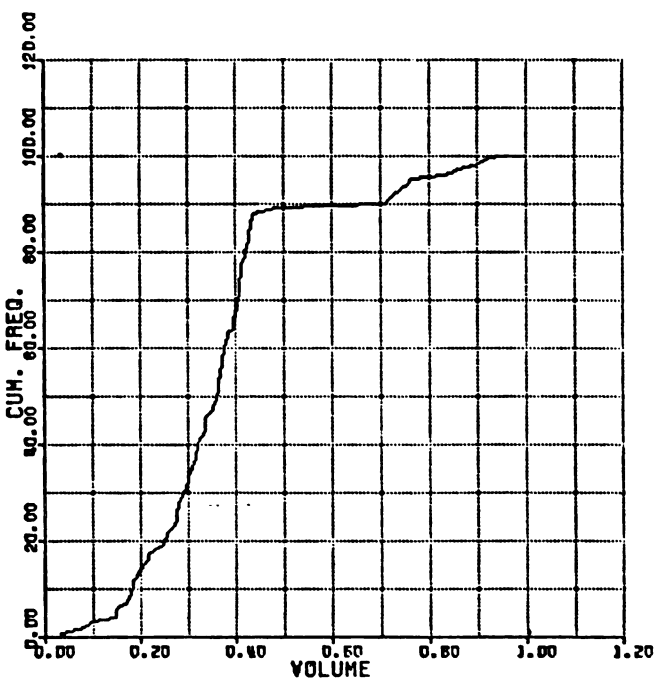


FIGURE 6: Clustered Data and Volume Graph with P outside the Cluster, Point P is shown as "+"

the square. Also shown is the graph of the volume versus total number of points contained in this volume when P is placed slightly outside the cluster.

One problem with using a single point P in computing the sequence of volumes is obvious. Different placements of P will yield a different view of the data under consideration. However, uniform data has the property that all placements of P should yield uniform volumes. Using a single point P also has the following undesirable property. Data points with approximately the same distance from P (and hence approximately the same volume) need not be spatially adjacent, especially if the distance is large. Thus, while data points which generate small volumes about P measure the density of the points near P, those that generate large volumes contribute to an estimation of the density over a much more spatially varied portion of the sampling window. The meaning of this comment is clear when one considers that the volume of a hypersphere about a point P in K-dimensional space is proportional to the Kth power of the radius of the hypersphere.

3.5 The Computation of the Volume-Based Test

The implementation of the volume-based test to test the null hypothesis of uniformity against a general alternative requires the following steps.

(1) Place a point P in the sampling window,

(2) Compute $||Xi-P||=zi$, for each Xi, by specifying some distance

metric,

(3) Determine the sets

$$V(Xi)=\{\ Y\in R^K \mid ||P-Y|| \leq zi\ \} \bigcap S,\ i=1,2,\ldots,N$$

(4) measure the volume u(V(Xi)) for each Xi,

(5) test the uniformity of the set of volumes, {u(V(Xi))}.

These steps involve the following nontrivial operations: intersecting two sets and determining the volume of that intersection. All of these steps are treated in more detail below.

3.5.1 Intersections and Volume Measurement

Given two sets A and B in K-dimensional space, the degree of difficulty in computing the intersection of A and B and the volume of the intersection depends on both the shape of A and B and their representations. In fact, for arbitrary sets A and B finding their intersection and the corresponding volume is computationally unmanageable. If we limit ourselves to the case where A and B have simple parametric representations or where A and B are both convex polytopes, some results are available.

For instance, if A and B are convex polyhedra in 2 or 3 dimensions, Muller and Preparata [Mul78] have shown that it is possible to find their intersection in time proportional to mlogm, where m is the sum of the number of vertices of the two polyhedra. Using the method given in

[Coh79], it is theoretically possible to find the intersection of convex polyhedra in arbitrary dimensions (K), though the time required is proportional to the Kth power of the sum of the number of faces of the two polyhedra. Cohen and Hickey [Coh79] also give a procedure to compute volumes of arbitrary convex polytopes whose time requirement again grows exponentially with dimension.

There are two simple parametric representations of A and B in which we will be interested. The first is when A and B are hyperspheres and the second is when A and B are hyper-rectangles aligned with the coordinate axes of the space. In the latter case, finding the volume of the intersection of A and B is trivial. Appendix B gives an algorithm to compute the volume of the intersection of two hyperspheres.

3.5.2 The Choice of a Distance Metric

The volume-based test also needs a distance metric for calculating the distance from point P to the sample points. We will confine ourselves to the commonly used Euclidean and supremum metrics. The main reason for this is that the two metrics result in hyperspherical and hyper-rectangular volumes. In addition, if the same type of sampling window is used as is generated by the distance metric, then the computation of the volume of intersection is simplified. In the Euclidean case, if we confine ourselves to a spherical sampling window S, then the algorithm given in Appendix B can be used to compute the volume of the set

$$\{ Y \in R^K \mid ||P-Y||_2 \leq z_i \} \cap S.$$

When using the supremum metric, if only hyper-rectangular sampling windows aligned with the coordinate axes are considered, then volume computation is again trivial.

It should also be mentioned that it is advantageous to choose the supremum metric in the case when the sampling window is a convex polytope, since this is one of the few Minkowski metrics whose ball is a convex polytope. Algorithms to find the intersection of two convex polytopes [Coh79, Mul78] can then be applied. Finding the intersection of a hypersphere, generated by the Euclidean metric and a simple convex polytope such as a hypercube is extremely difficult.

3.5.3 Testing Univariate Uniformity

To apply the volume test, we must specify how we will test the uniformity of the set of volumes. There are many such tests against specified and general alternatives [Knu81, Cox65]. We choose to use one of the most widely used tests against unspecified alternatives, the Kolmogorov-Smirnov goodness-of-fit (KS) test [Con71]. This test was chosen both for its ubiquity and simplicity, as opposed to tests such as the scan test [Nau66] or tests based on Greenwood's statistic [Ste81], which might have been more powerful for departures from randomness expected in clustered data.

3.5.4 Placement of Point P

Finally, we need to discuss how best to place point P in the sampling window to apply the test. As we have said before, uniform data has the property that any placement of P generates a uniformly distributed set of volumes. However, in practice, we have the following considerations.

First, some placements of P may make the volumes easy to compute. For instance, if the sampling window is a hypersphere and Euclidean distance is used, placing P at the center of the sampling window allows us to avoid the computation of the spherical cap volumes. Placing P near the centroid of a convex polytope sampling window also yields computational advantages.

Second, since volumes expand around P, placing P in a region of high (or low) sample point density should yield higher power for clustered data. Thus one would expect higher power if P is placed near a cluster center. However, the proof of Theorem 1 implicitly assumes that P is independent of the set of sample points. This condition is not always satisfied, especially if the sampling window must be estimated from the data. While it is possible that placing P in the region of highest point density in a uniform sample may not greatly effect the distribution of the K-S statistic, we prefer to avoid this problem, perhaps by sacrificing power in clustered data. We have

experimented with placing P randomly in the sampling window, at the center of the sampling window, and at the mean of the data to be analyzed.

3.6 Summary

In this chapter, we have presented a theorem that generates a paradigm for various tests of K-dimensional randomness. This theorem is based on measuring volumes and thus directly captures information about the deviation of the sample density in a uniform sample. We have defined a test based on this volume paradigm that measures the density of the sample points about a single point P. We discussed the advantages and disadvantages of the volume-based test as well as computational details needed to perform the test. The performance of the volume-based test is given in Chapter 5.

CHAPTER 4

ESTIMATING THE SAMPLING WINDOW

4.1 Introduction

In exploratory pattern analysis one usually does not have any knowledge about the sampling window of the given data. Rather, we are simply given N points in K dimensions and told to analyze the structure of the points. If we want to assess the uniformity of the points, then we must either make some assumptions about the sampling window or estimate it.

Previous studies in assessing structure have overcome this crucial problem by assuming that the sampling window is known. We wish to relax this restriction as much as possible so that we can analyze real data sets, where the sampling window is usually not known. In this chapter, we will look at several ways to estimate the sampling window. One can argue that data used to estimate the sampling window should not then be tested for uniformity in the estimated window, for this may bias the test. However, due to the small size of the data sets common in Pattern Recognition, we are forced into using this methodology.

It is quite apparent that to use the volume-based test one must have precise knowledge of the sampling window since the sets whose volumes are to be measured are constrained to lie in the true sampling window. An error in estimating the sampling window may make uniform data appear as a single cluster in the center of the window. Previous studies [Smi81, Pan81, Cro80] have concluded that knowledge of the sampling window is required in virtually all tests of clustering tendency which have been proposed in the literature. Some of these studies have shown that, asymptotically, the only knowledge needed about the sampling window is its size and not its shape or location. The volume-based test's greater reliance on precise knowledge of the sampling window makes it a perfect vehicle for experimental studies of sampling window estimation procedures. The size and power of the test can be greatly effected by the estimation procedure used. We believe that any estimation procedure which works well for the volume-based test would necessarily be a good estimation procedure to use with other tests.

The need to have some knowledge of the sampling window is illustrated in Figure 2. However, the knowledge of the sampling window in Figures 2(a) and 2(b) comes to the forefront in different ways. The data in Figure 2(a) appears uniform, while in Figure 2(b), the data consists of two clusters. Therefore, to test the uniformity of the data in Figure 2(a) we need to know the set over which the density is non-zero. For the second data set, we need to distinguish regions of low density between the clusters from those regions outside the range of

the data. Intuitively, clustered data is composed of points of high density separated by less dense regions. When a region of low density is identified, the sampling window information is needed to distinguish when this region is outside the domain of interest versus the case when the region is between the clusters. A region between clusters could also lie outside the sampling window if this window was not convex. Thus we have placed the restriction that sampling windows be convex.

4.2 Estimation Procedures

The basic estimation problem is stated as follows. Given $\{Xi\}_{i=1}^{N}$ i.i.d. uniform over a convex set $S \subset R^K$, with $u(S)>0$, estimate S. We also require that S be compact. In other words, it is closed and bounded. Our approach will be to first simplify the problem by considering simple forms of the set S; we will then increase the difficulty of the problem until S is any compact convex set.

We will restrict ourselves to the following types of sampling windows. First, we consider the case when S is a hyper-rectangle aligned with the coordinate axes. This is the case that has been used for other studies when the sampling window was assumed known [Cro82, Pan81]. We give a procedure to estimate the hyper-rectangular sampling window from the given data. If the hyper-rectangle is not aligned with the coordinate axes, the estimation problem is very difficult and we are not able to treat that case. Another possible shape of the sampling window is a hypersphere. We give two methods of estimating a hypersphere from the given data. We next consider linear

transformations of the hyperspherical sampling window. That is we consider the case when S is a hyperellipse. We use the principal component transformation and the whitening transformation to estimate the parameters needed to map the true sampling window into a hypersphere. The transformed data is then tested in this simpler sampling window. Though a linear transformation exists which would transform any hyper-rectangle into an aligned hyper-rectangle, the principal component method would estimate it poorly. Finally, we consider the most general estimate of the sampling window based on the convex hull of the data.

4.3 Aligned Hyper-Rectangle

This type of sampling window has been used frequently in studies which assume that the sampling window is known. Its simplicity gives us an excellent estimation procedure.

An aligned hyper-rectangle S can be described by its range along each coordinate axes. That is, each coordinate has a minimum and maximum threshold which specifies the (K-1)-dimensional boundary flats defining two sides of the hyper-rectangle. We write $S=[ai,bi]_{i=1}^{K}$ to specify the aligned hyper-rectangle with range [ai,bi] along each coordinate. Thus the vector Y=(y1,y2,..,yK) is in S if and only if $yi \in [ai,bi]$ for each i=1,2,...,K. We derive an estimator of S under the hypothesis that the given data is uniformly distributed over S. The density function of Y can be written as

$$f(Y) = \prod_{i=1}^{K} \frac{I([a_i, b_i])(y_i)}{(b_i - a_i)}$$

where I(A) is the indicator function of the set A.

The independence of the coordinates allows us to treat the estimation problem along each axis in isolation. Thus we need to find an estimate for the endpoints of a one-dimensional uniform distribution. Let $\{Zi\}_{i=1}^{N}$ be a sample from a Uniform[a,b] density. Then the minimum variance unbiased (MVU) estimators for a and b are given below [Rao73].

$$\hat{a} = (N\ Z(1) - Z(N)) \ / \ (N - 1)$$

$$\hat{b} = (N\ Z(N) - Z(1)) \ / \ (N - 1)$$

where Z(1) and Z(N) are, respectively, the minimum and maximum order statistics of the {Zi}. Thus the MVU estimator of S is given by

$$\hat{S} = [\hat{ai}, \hat{bi}]_{i=1}^{K}$$

where ai and bi are the estimates of the end points along the ith coordinate.

The time taken to compute this estimator is O(KN), since, for each of the K coordinates, we must find the minimum and maximum values of the N sample points.

4.4 Hypersphere

The problem of estimating a hypersphere can be stated as follows. We have a set of i.i.d. random vectors {Xi} over a hyperspherical sampling window S, with radius r centered at vector c. We will denote this hypersphere as S(c,r). We wish to estimate the (K+1) scalar parameters in (c,r). We would like to find a MVU estimator as we did for the aligned hyper-rectangle. This does not appear possible and so we offer the following two estimates.

4.4.1 Unbiased Center

We assume that the density function of the sample is radially symmetric about the center of the hypersphere. This is true for the uniform density. In this case the center of the hypersphere is also the expected value of random vectors following this density. We know that an unbiased estimator for the expected value is the mean of the data and so this estimator is also unbiased for the K-dimensional parameter c. We choose to estimate the radius of the hypersphere, given our estimate for c, by the distance between c and the sample vector with the maximum distance from c. This is the minimum possible value of the radius for this choice of center. Thus the estimators of c and r are

$$\hat{c}_u = (N)^{-1} \sum Xi \quad \text{and}$$

$$\hat{r}_u = \max_i \; || \hat{c}_u - Xi ||_2 .$$

4.4.2 Smallest Hypersphere

Experimental evidence has shown that using the above estimator produces a window whose volume exceeds the true volume. We decided to obtain the smallest hypersphere which encloses the given set of points. Appendix C gives details of the algorithm [Elz72] used for computing the smallest hypersphere, $S(c^*,r^*)$. Since the true sampling window is a hypersphere, r^* is no larger than the true sampling window's radius. Experimentally we have found that r^* is closer to the true radius than $\hat{r}_u$ found in the previous section.

4.5 Hyperellipses

One method of estimating more complicated sampling windows is to transform the data into the simple cases of a hypersphere or aligned hyper-rectangle treated above. It is well-known that a linear transformation of uniform data preserves the uniformity of the data. That is, if the uniform density is defined over a set S and if T is a linear transformation, then the density induced on the image T(S) is also the uniform density. For any hyperellipse there exists a linear transformation which maps the hyperellipse into a hypersphere. Given these facts we choose to estimate a hyperelliptical sampling window in the following manner. First, we estimate the transformation T which carries the hyperellipse into a hypersphere and then we find the

smallest hypersphere enclosing the transformed data.

An estimate of the linear transformation T is based on the principal component (Karhunen-Loeve) transformation [Fuk72]. It operates as follows. The sample is first normalized to have a zero mean vector. The principal component transformation, based on the eigenvectors of the sample's covariance matrix, decorrelates the features. We then apply the whitening transformation so that each coordinate has unit variance. Since we use the sample mean vector and covariance matrix this transformation need not necessarily map the hyperellipsoidal sampling window into a hypersphere. This transformation has been used in a clustering tendency study with limited success by Cross [Cro80].

4.6 Compact Convex Sets

The most general sampling window is a compact convex set. We follow the exposition of Ripley and Rasson [Rip77b] for estimating this type of sampling window without presenting the details.

First we need some notation. Let H(A) denote the convex hull of a set A. Let X={Xi} be N i.i.d. uniform vectors over the compact convex sampling window S. We wish to find an estimate of S from the class of all compact convex sets of positive measure in K-dimensional space.

The joint density of the N points over a compact convex set S can be written as

$$f_S(X) = \frac{I[S](X)}{(u(S))^N}$$

where I[S] is the indicator function of S and u(S) is its volume. Since the convexity of set S implies that X is in S if and only if H(X) is in S, we have

$$f_S(X) = \frac{I[S](H(X))}{(u(S))^N}$$

Thus H(X) is both a sufficient statistic and the maximum likelihood estimate of S. Note, however, that the volume of H(X) is strictly less than the volume of S. We will use the convex hull of the data as the estimate of the sampling window and use the volume-based test to test those points strictly inside the hull for uniformity. We delete those points lying on the hull from consideration since they are obviously not random in H(X). However, conditioned on the fact that the remaining points lie in the convex hull, these interior points are uniformly distributed in H(X).

For the two-dimensional case, Ripley and Rasson show that

$$\frac{N}{M}(u(H(X)))$$

is an approximately unbiased estimator of the volume of S, where M is

the number of points strictly inside H(X). Appendix D gives details on computing the convex hull of a set of points. We also verify whether the above volume estimate remains unbiased in dimensions greater than two.

4.7 Summary

In this chapter we have looked at several ways of estimating the sampling window. In the case when the sampling window is restricted to be an aligned hyper-rectangle, we find a MVU estimator of the window for uniform data. When the sampling window is an arbitrary compact convex set, we find that the convex hull of the uniform data is the maximum likelihood estimate of the window. These two estimators have desirable properties for uniform data but they are reasonable estimates of sampling window for any data. For hyperspherical sampling windows, no best estimator emerges. We propose the heuristic of choosing the mean of the data, which is unbiased for uniform data, as the center of the hypersphere. We also provide the smallest hypersphere containing the data as an estimate of the sampling window.

We also propose a method of estimating a hyperellipsoidal sampling window. This estimator operates in two steps. First, it estimates the transformation needed to carry the hyperellipse into a hypersphere and second, it estimates the hypersphere using one of the estimators discussed above.

CHAPTER 5

PERFORMANCE OF THE VOLUME-BASED TEST

5.1 Introduction

In this chapter, we look at the performance of the volume-based test. This performance will be measured by Monte-Carlo simulation using various sampling window types. First, with known sampling windows, we wish to check both the size and the power of the test. Although the size of the test is guaranteed by Theorem 1, we wish to check if our implementation truely reflects the theoretical result. To check the power of the test, we study a number of clustered alternatives. For unknown sampling window, we investigate the estimation procedures given in Chapter 4.

The simulations reported here involve the following three parameters.

(1) N, the number of sample points,

(2) M, the number of Monte-Carlo trials, and

(3) K, the dimensionality of the space.

In addition, we vary the sampling window used, the distribution of the points, and the placement of P. The level of the Kolmogorov-Smirnov

test is set at 0.05, so that we expect a 5% rejection rate for uniform data. The results of the simulations are reported as the percent rejection of the null hypothesis of randomness at the .05 level of significance. To determine if the true size of the volume-based test is less than .05, we perform a binomial test [Con71] on the observed number of rejections for the null hypothesis (size<.05) against the alternative (size>.05). Likewise, to determine if there is a significant difference in the size or power of the volume-based test between two different data sets, we perform the Chi-squared test based on 2X2 contingency tables [Con71].

5.2 Known Sampling Windows

The sampling windows considered are hyperspheres and aligned hyper-rectangles. As we have said before, if a hypersphere is used as a sampling window then we compute distances from point P to the sample points using the Euclidean metric; for aligned hyper-rectangular sampling windows, we use the supremum metric.

5.2.1 Uniform Data

Here we determine the actual size of the volume-based test when the level of the K-S test is preset to 0.05. Table 1 shows the percent rejections of the null hypothesis for the volume-based test when uniform data is generated in the unit hypercube. Point P is placed randomly in the hypercube. These results show that the size of the test is 0.05 and

TABLE 1: Size of the Volume-Based Test for Uniform Data in Unit Hypercube. M=500, P=random

		N		
		50	100	200
K	2	4.3	5.3	4.7
	5	4.3	6.7	4.7
	10	4.3	6.0	4.7
	15	3.0	3.7	4.3

the size does not depend on dimensionality or sample size at the 0.02 level.

Tables 2 and 3 show similar results when the sampling window is changed to a hypersphere. A hypersphere of volume one is used for the results of Table 2 while Table 3 presents the results for a hypersphere of radius one. In both cases the number of patterns, N, is 200. Here we also study the effect of random placement of P in the sampling window versus choosing P as the center of the hypersphere. No significant differences (at the 0.02 level) are encountered between the entries in the two tables and between different placements of point P. In addition, no entry in these tables shows significant deviation (at the

TABLE 2: Size of the Volume-Based Test for Uniform Data in Unit Volume Hypersphere. M=500, N=200

		P=random	P=center
K	2	4.2	5.4
	5	3.8	4.6
	10	4.8	5.2
	15	3.8	4.4

TABLE 3: Size of the Volume-Based Test for Uniform Data in Unit Radius Hypersphere. M=500, N=200

		P=random	P=center
	2	4.6	5.2
	5	4.0	3.6
K	10	3.6	4.6
	15	4.0	3.4

0.05 level) from its expected value of 5.

The similarity between Tables 2 and 3 is expected. The only difference between them is in the volume of the sampling window, which is normalized by the volume-based test. We present both tables to confirm that the volume-based test does not lead to anomolies between these two windows, as has been observed with a distance-based test [Smi81].

We conclude from these tables that the volume-based test works as expected on uniform data. We now look at the power of the test.

5.2.2 Bilevel Density

The bilevel density [Smi81] generates a single cluster of high density in the middle of the unit hypercube. Formally, the N points are generated i.i.d. with density

$$f(X) = \begin{cases} h_0 \text{ if } X \text{ is in (Unit Hypercube} - W) \\ h_1 \text{ if } X \text{ is in } W \\ 0 \text{ otherwise} \end{cases}$$

where W is a hypercube of volume 1/9 centered inside the unit hypercube. Here hl is a parameter which may vary from 0 to 9, while ho depends on hl. Figure 7 shows a realization of the bilevel alternative in two dimensions with hl equal to 5. If hl is 9 all the points are in W. In the simulations, the parameter hl is varied from 1 (the null case) to 5 in steps of 1 for various values of N and K. Table 4 gives the results of the volume-based test when P is placed randomly in the unit hypercube.

TABLE 4: Power of the Volume-Based Test Against the Bilevel Density. M=100

		hl					
		1	2	3	4	5	
	2	4	8	35	48	80	
	5	4	4	17	41	71	
K	10	4	6	13	27	45	N=50
	15	3	9	15	15	42	
	2	5	14	51	81	96	
	5	7	10	40	70	80	
K	10	6	11	39	57	78	N=100
	15	4	7	23	38	62	
	2	5	29	79	96	100	
	5	5	21	68	92	100	
K	10	5	15	58	82	94	N=200
	15	4	8	39	68	88	

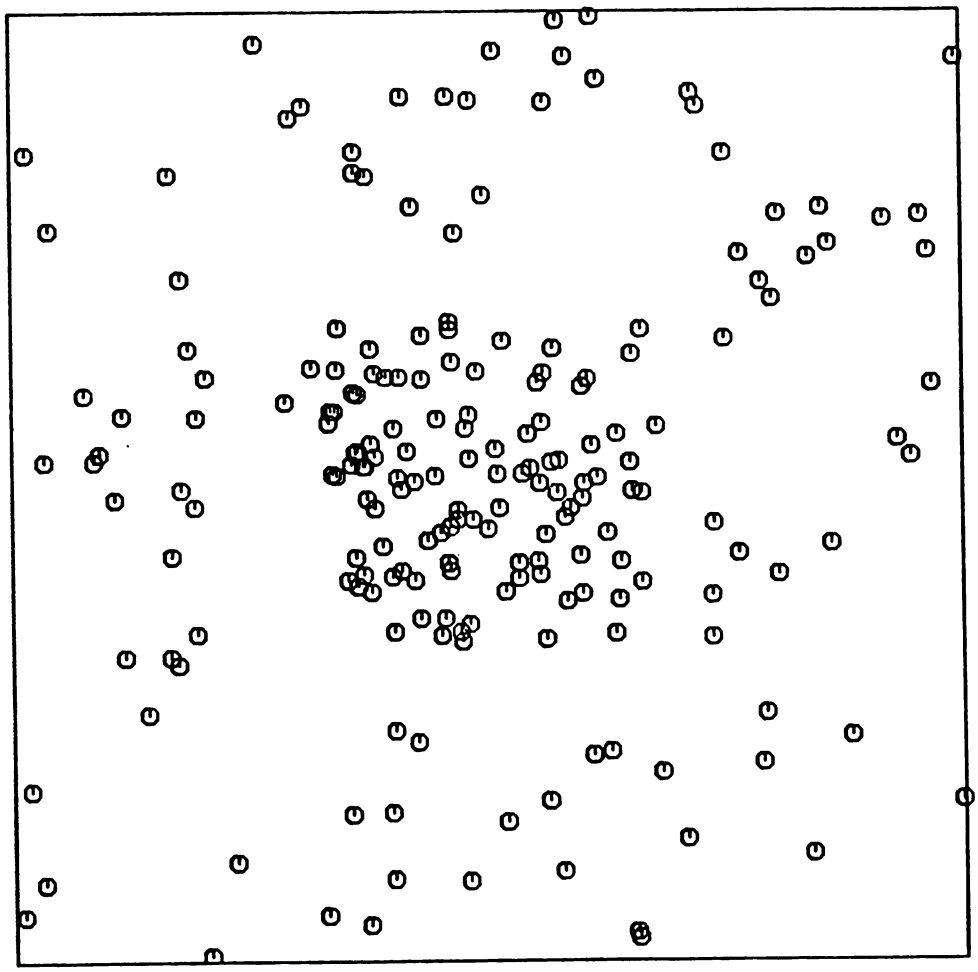


FIGURE 7: Realization of Points Following the Bilevel Density. Shown are 100 points with h1=5

We note the ubiquitous trends of increase in the power of the test as N and hl increase. These trends are expected. The decrease in power with increasing dimensionality is explained by the increasing side length of the hypercube W needed to maintain constant volume as dimensionality increases. Any sampling origin P which falls outside W encounters points in W 'sooner' (in terms of distance from P) in high dimensions than in low dimensions. If P is chosen as the center of the unit hypercube, then experiments show that this effect does not occur.

Using the bilevel density as a clustering alternative allows us to compare the volume-based test's performance to the theoretical power of a distance-based test described by Smith and Dubes [Smi81]. This test is based on a count of the number of interpoint distances which are below a given threshold. This threshold is defined by a parameter r. The theoretical power of this test can be computed by referring to asymptotic results, which are probably not valid for large r or small N. Figure 8 shows a graph of the power of the small distance test in various dimensionalities with r set to 1. Figure 9 is a graph of the data in Table 4 when N=200. Comparing the two figures, we note the higher power of the volume-based test against the bilevel alternative.

5.2.3 Neyman-Scott Clustering

This clustering alternative is the Neyman-Scott cluster process [Ney72] modified to generate N points over a sampling window. This

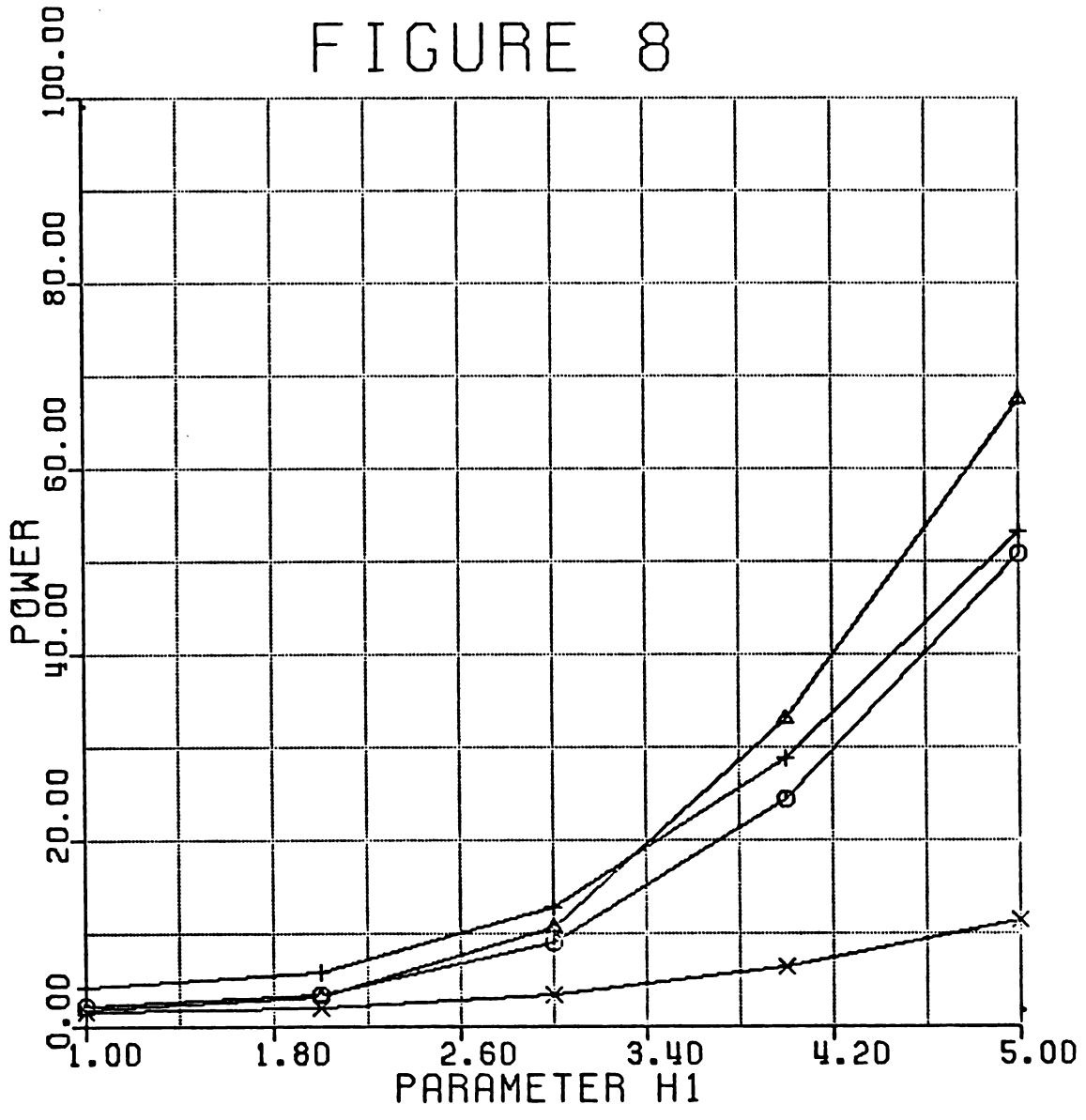


FIGURE 8: Power of the Small Distance Test Against the Bilevel Density
o....K=2
Δ....K=5
+....K=10
X....K=15

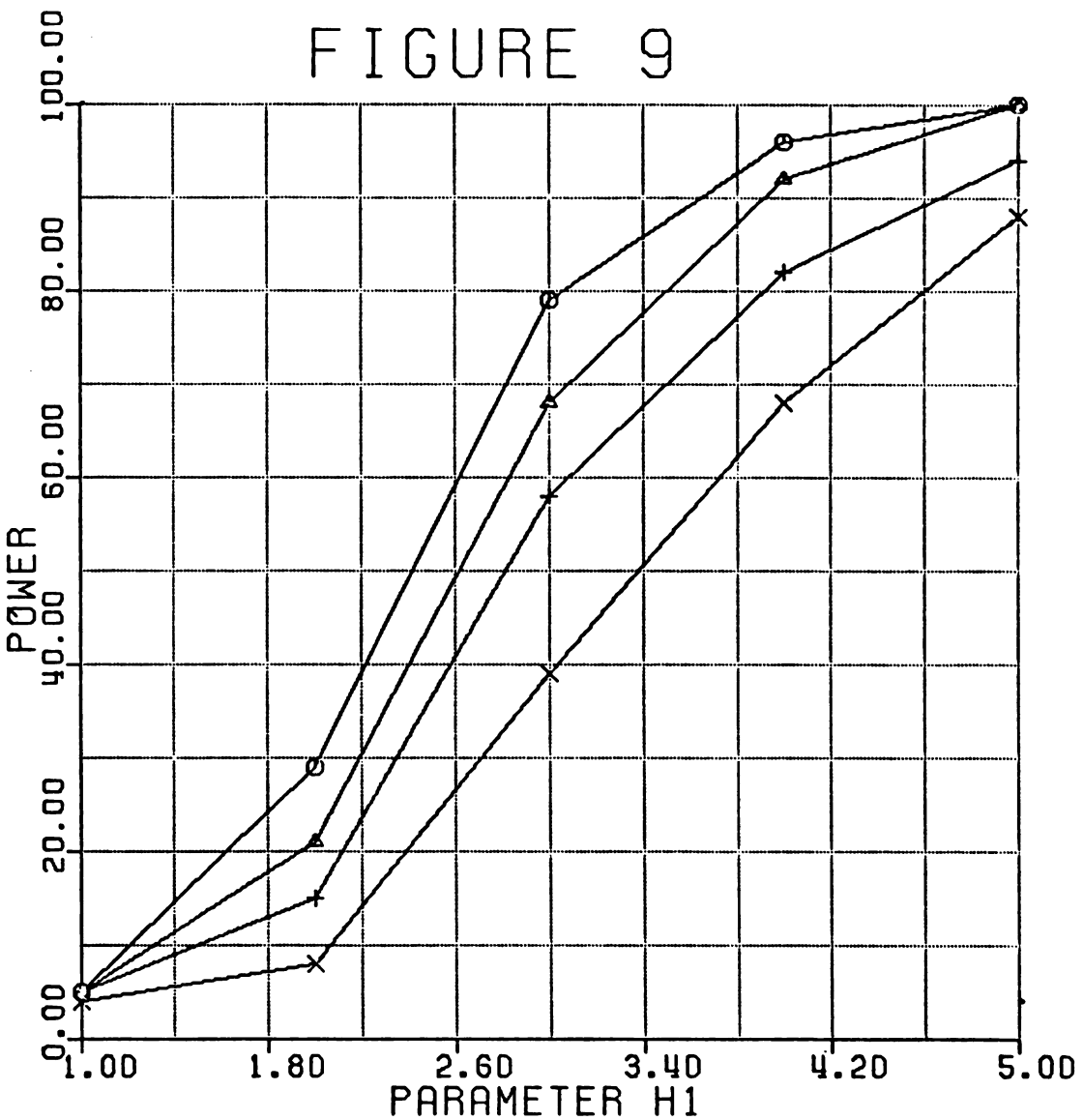


FIGURE 9: Power of the Volume-Based Test Against the Bilevel Density

o....K=2

Δ....K=5

+....K=10

X....K=15

TABLE 5: Power of the Volume-Based Test Against a Neyman-Scott Process (wrapped) in Unit Hypercube. M=500, N=200

μ	.01	.05	.1	.3	.5	
50	100.0	98.4	90.6	17.4	5.4	
8	92.2	78.4	55.4	8.4	4.8	K=2
1	35.4	25.8	17.6	6.0	5.2	
50	100.0	97.8	85.6	14.0	5.4	
8	92.4	69.2	46.6	8.4	6.2	K=5
1	29.8	19.4	16.4	4.8	4.4	

σ

process is characterized by two parameters: μ, the expected number of points per cluster and σ, the spread of each cluster. See Appendix A for a detailed description of the Neyman-Scott process and its generation.

TABLE 6: Power of the Volume-Based Test Against a Neyman-Scott Process (not wrapped) in Unit Hypercube. M=500, N=200

μ	.01	.05	.1	.2	.3	
50	100.0	99.0	92.8	68.6	34.6	
8	91.2	73.4	58.6	35.2	18.4	K=2
1	23.0	17.8	13.2	23.0	15.6	
50	100.0	98.6	92.0	65.8	39.0	
8	88.8	69.4	53.6	47.0	23.8	K=5
1	22.0	20.2	17.6	37.0	22.0	
50	100.0	98.2	87.0	61.4	40.6	
8	90.2	68.4	57.0	45.8	30.0	K=10
1	33.6	25.8	18.4	20.4	12.4	

σ

In Table 5 we report the power of the volume-based test for various values of μ and σ when the sampling window is the unit hypercube. In these tables, wrap around was used to generate the points while in Table 6 the same parameter values are studied, but without wrap around. Both Tables were generated with P chosen randomly in the unit cube. There is some effect of dimensionality on the power. For instance, Table 5 shows a significant (at the .001 level) change with dimensionality for $\mu=8$ and $\sigma=.05$. Table 6 shows a significant (at the .001 level) increase in power between the K=2 and K=10 cases for $\mu=8$ with $\sigma=.2,.3$, and $\mu=1$ with $\sigma=.01$. As expected, high power is achieved with a few tight clusters (μ large and σ small). The power falls off as μ is decreased and σ is increased. The only significant differences (at the .001 level) between data sets with wrap around and no wrap around occurs when σ is large; this is when most points get wrapped around. We see higher power in the no wrap around case since here cluster centers near the boundary of the sampling window generate clusters of higher density than those in the wrapped case.

It should be noted that, unlike most of the distance-based tests, the distances used in defining the volume-based test are not based on wrap around. The reason for using wrap around in data generation as well as in computing the test statistic has been to avoid edge effects which arise due to the assumption of a Poisson process as the null hypothesis. There are no edge effects in the volume-based test since the null hypothesis is that the data are uniformly distributed over the sampling window. Note also that no wrap around is possible for sampling

windows other than the hyper-rectangle. It is not reasonable to compare the powers of a test against the Neyman-Scott process with wrap around and without wrap around since they are 'different' data sets.

In the Neyman-Scott process with wrap around, we can compare the power results of the volume-based test to that of the Hopkins test reported by Cross and Jain [Cro82]. Table 7 gives these comparisons. We note consistently higher power of the volume-based test in two dimensions, while in 5 dimensions, with large μ and σ, the Hopkins test fairs better. Note the increasing power with dimensionality for the Hopkins test, while the volume-based test is fairly stable with respect to dimensionality. Table 8 shows the power comparison of the volume-based test to the power of a test based on the Cox-Lewis statistic. The powers for the Cox-Lewis statistic are taken from the (corrected) tables of Panayirci and Dubes [Pan81]. We note higher power for the volume-based test, except for the entries for K=5 with σ=.05 and .1.

Table 9 reports the power against the Neyman-Scott process when the sampling window is the unit volume hypersphere. All the parameters are the same as in Table 6 except that we have decreased the range of σ for large dimensionalities. There is no change between the hyperspherical and hypercubic sampling windows at the 0.001 significance level. We have also a study of the effect of varying the placement of P on the power. We note slightly higher power when P is placed randomly in the sphere except when σ is large.

TABLE 7: Comparison of the Power of the Hopkins and Volume Based Tests

Entries are Hopkins/Volume
N=200
M=500 for Volume-Based Test
M=100 for Hopkins test

K=2

μ	.01	.05	.1
8	87/ 92.2	59/ 78.4	23/ 55.4
1	15/ 35.4	7/ 25.8	5/ 17.6

K=5

μ	.01	.05	.1
8	87/ 92.4	96/ 69.2	77/ 46.6
1	22/ 29.8	20/ 19.4	4/ 16.4

σ

TABLE 8: Comparison of the Power of the Cox-Lewis and Volume-Based Tests

Entries are Cox-Lewis/Volume
N=200
M=500 for Volume-Based Test
M=100 for Cox-Lewis Test
The Neyman-Scott parameter μ=50

	.01	.05	.1	.3
K=2	100/100	92/ 98	32/ 91	3/ 17
K=5	100/100	100/ 98	99/ 86	1/ 14

σ

TABLE 9: Power of the Volume-Based Test Against a Neyman-Scott Process in Unit Volume Hypersphere for Different Placements of P Entries are P=random/P=center, M=100

	50	100/100	99/ 97	95/ 85	67/ 62	42/ 35	
μ	8	94/ 91	79/ 64	69/ 52	37/ 55	19/ 31	K=2
	1	34/ 34	29/ 25	21/ 19	16/ 23	7/ 13	
	50	100/100	100/ 92	96/ 75	63/ 82		
μ	8	94/ 85	78/ 58	66/ 51	35/ 77		K=5
	1	35/ 34	28/ 19	21/ 20	20/ 34		
	50	100/100	98/ 87	94/ 65			
μ	8	94/ 84	80/ 54	62/ 66			K=10
	1	32/ 28	29/ 22	24/ 32			
		.01	.05	.1	.3	.5	
				σ			

5.2.4 Other Types of Data

Two other types of alternatives need to be mentioned. The first arises from points following the multivariate normal density with identity covariance matrix. To perform the volume-based test on this data, we scale the data to fit into the unit hypersphere centered at the zero vector by dividing each data point by the maximum norm among the N samples. Performing the test with N=200 leads to 100% rejection of uniformity for normal data in all dimensions, both when P is placed randomly in the unit hypersphere and when P is placed at the origin. If we reduce the sample size to 50 then, with P placed at the origin, rejection rates of 92, 100 and 100 percent in 2, 5 and 10 dimensions,

respectively are obtained. However, under these conditions, if P is placed randomly, the rejection rates are only 33, 67 and 81 percent. This shows the increased power of the volume-based test when P is placed at the center of a cluster.

The other data ensemble is hardcore data. The generation procedure used to produce hardcore data is given in Appendix A. In these simulations, we do not use wrap around. The volume-based test shows no power against hardcore data if P is chosen randomly in the sampling window. This is expected, since the only difference between random data and hardcore data occurs for small volumes around P. We expect fewer points close to P in the hardcore case than in the random case due to the spacing imposed by the hard spheres. However, this effect is masked by the large volumes, where points generating these volumes do not have to be spatially adjacent. If P is chosen as the center of the sampling window, we do see power (62 and 100 percent rejections for $\rho = .1$ in 4 and 5 dimensions, respectively) against the hard core alternative. There is no power in two dimensions. This power in high dimensions against the hardcore process is due to the fact that many points are near the surface of the sampling window in high dimensions; the hard spheres around points near the surface take up less of the volume of the sampling window. This allows a greater density of hardcore points near the surface of the sampling window, a fact that is captured by the volume-based test. We can not compare the power of the volume-based test against the hardcore alternative to other studies since we did not use wrap around in generating the hardcore data.

5.3 Unknown Sampling Windows

In this section, we perform Monte-Carlo studies of the sampling window estimation procedures given in Chapter 4. We determine if the true size of the volume-based test performed over the estimated window is within the preset level of 0.05.

5.3.1 Estimator of an Aligned Hyper-Rectangle

Here we study the MVU estimate of an aligned hyper-rectangular sampling window given in Section 4.3. Table 10 gives the results of this study. We generate 200 points uniformly distributed in the unit hypercube and use the MVU estimate as the true sampling window in the volume-based test. None of the entries in the table are larger than 5. Thus we conclude that the MVU estimate is a good estimator of an aligned hyper-rectangular sampling window.

TABLE 10: Size of the Volume-Based Test with the MVU Estimator. M=400, N=200

	K					
	2	3	4	5	10	15
N 50	3	1	3	2	4	5
100	5	2	3	2	4	5
200	4	4	3	4	5	2

5.3.2 Estimator of a Hypersphere

In Section 4.4, we gave two estimation procedures for a hyperspherical sampling window. The first uses the sample mean as an estimate of the center of the hypersphere. The radius estimate is the distance from the farthest sample to the mean. We checked the performance of this estimator in recognizing uniformity with 200 points generated uniformly in the unit radius hypersphere. When P is chosen as the center of the estimated window we get 46 and 86 percent rejections of the null hypothesis in 2 and 5 dimensions, respectively. Analysis of these simulations showed that the estimated radius was too large and this made points in the center of the sampling window appear too dense.

The smallest hypersphere algorithm produced better results. Table 11 shows these results when P is chosen as the smallest hypersphere's center. To produce this table, the K+1 data points defining the smallest hypersphere are deleted from analysis by the volume-based test.

TABLE: 11: Size of the Volume-Based Test with the Smallest Hypersphere Estimate. M=500, N=200

	K					
	2	3	4	5	10	15
50	4	4	4	5	3	4
N 100	3	5	3	4	6	5
200	4	3	4	6	5	4

If these points are not deleted, the density of the points near the surface of the hypersphere becomes too large. The percent rejections increases to 11.3 in 10 dimensions if the surface points are not deleted. From Table 11, we conclude that the smallest hypersphere is a good estimator of a hyperspherical sampling window.

5.3.3 Estimator of a Hyperellipse

In Section 4.5, we gave a two-step procedure to estimate a hyperelliptical sampling window. This involved appyling both the principal component and the whitening transformation to the data and then testing the transformed data for uniformity in the smallest hypersphere enclosing it. The point P is chosen as the center of the smallest hypersphere. The details on the generation of uniform data over an ellipse are given in Appendix A, along with the parameters of the ellipses used.

Even in 2 dimensions, this estimation procedure is inaccurate. For 200 points uniformly distributed in an ellipse, we obtain 22% rejections of uniformity, while for 300 points we get 31% rejections. Table 12

TABLE 12: Effect of Transforming Uniform Data in a Circle
M=300

N	Original	Transformed
200	4.3	21.0
500	4.3	23.6
1000	2.8	33.3

studies the case of uniform data in a circle rather than in an ellipse. The principal component and whitening transformations appear to perturb the data enough to affect the volume-based test. Even though increasing N increases the accuracy of the estimated mean and covariance matrix and thus lessens the perturbation, it is not enough to counteract the decreasing range of acceptable values for the K-S test statistic. We conclude that the principal component transformation does not appear to be a viable way to change the shape of the sampling window, at least for the volume-based test.

5.3.4 Estimator of a Compact Convex Set

In Section 4.6 we described a procedure for estimating a compact convex sampling window by using the convex hull of the data. This is the most general form of the sampling window in situations where no prior information is available about the shape of the sampling window. Unfortunately, computing the convex hull of points in high dimensions is computationally difficult. More significantly, computing the volume of the intersection of the hull with hypercubes (about P) is burdensome, even in three dimensions. Finding the volume of the intersection set requires the computation of all the vertices of that set; the computation time grows exponentially with K. In two and three dimensions it is possible to find the intersection set in time proportional to m and mlogm, respectively, where m is the sum of the number of vertices in the two sets to be intersected [Mul78]. We, therefore, confine ourselves to two-dimensional data. Also, for computational efficiency, it is desirable to have P near the center of

the convex hull, and so we choose P as the mean of the data.

Using the convex hull as a sampling window we can now successfully treat the elliptical data used in the previous section. For 200 points generated uniformly in an ellipse in 2 dimensions, we have 4 rejections out of 100. Table 13 lists the results of using the convex hull estimator on various types of data. The points on the hull are removed from consideration in producing this table. We also compare the results of the various sampling window estimation procedures discussed in this chapter. We study both uniform data and Neyman-Scott clustered data with μ=50 and σ=.1 in a unit square and unit circle. The Neyman-Scott data is generated without wrap around.

From Table 13 we see that the convex hull estimator performs well. The size of the test using the convex hull data is well within the expected value of 5. For the clustered data over a unit square, there is a slight loss of power when the window must be estimated from the data versus when it is known. This loss of power does not seem to occur

TABLE 13: Comparison of Sampling Window Estimators
K=2, N=200, M=500, except convex hull where M=50

	Random Data		Neyman-Scott Data	
Estimator	Circle	Square	Circle	Square
Convex Hull	4.0	4.0	98.0	80.0
MVU Rectangle	---	4.0	----	81.2
Smallest Circle	4.4	---	95.8	----
Known Circle	5.2	---	94.8	----
Known Square	---	4.7	----	92.8

for the circular sampling window. The convex hull estimator performs as well as the two estimators geared to the special situations of circular and rectangular sampling windows.

5.4 Summary

This chapter has presented the size and power of the volume-based test. We have seen that the volume-based test can be set at a desired size. Our power studies have shown that the volume-based test is at least as powerful as other tests provided in the literature in most cases. We have studied the effect of different placements of P, and concluded that this can have an effect on the test. Our choice for P when analyzing real data would be the center of the sampling window.

We have found excellent estimators of hyperspherical and aligned hyper-rectangular sampling windows. However, if no prior information is available about the shape of the window, then the convex hull of the data is a good choice. Unfortunately, it is not computationally feasible to determine the convex hull and its volume in high dimensions ($K>3$). For this reason, the next chapter looks at a new test of uniformity of data.

CHAPTER 6

A MINIMAL SPANNING TREE BASED TEST

6.1 Introduction

We have seen that the volume-based test requires an accurate estimate of the true sampling window. If the true sampling window is a hypersphere or an aligned hyper-rectangle then efficient estimators are available to estimate the window. In situations where no prior knowledge is available about the shape of the sampling window, the convex hull of the data appears to be a reasonable estimate of the window. Unfortunately, computing the convex hull and its volume in high dimensions is not computationally feasible. Therefore, the applicability of the volume-based test is limited to data in low dimensions.

In this chapter we propose a test which does not explicitly require any knowledge of the true sampling window. We still assume that the convex hull of the data is a reasonable estimate of the sampling window, but we do not need to compute the convex hull or its volume. The idea of this test comes from a multivariate extension of the Wald-Wolfowitz runs test [Wal40] proposed by Friedman and Rafsky [Fri79]. 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 as follows. The given data which are to be tested for uniformity constitute one sample. We would like the other sample needed for the Friedman-Rafsky test to be 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 is uniformly distributed over the convex hull. One of the problems, of course, is in generating uniform data over the convex hull, since it is not computationally feasible to form the convex hull of high dimensional data. In Section 6.2 we present a heuristic that approximates the convex hull for the purpose of generating uniform points over it. Section 6.3 describes the Friedman-Rafsky test and the proposed MST-based test. Section 6.4 determines the size and power of the MST-based test for various sampling windows and data sets.

6.2 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. Our overall procedure will be as follows. We determine a (relatively simple) convex set containing the data. We generate uniformly distributed points over this set and retain those that fall in the convex hull of the data. This rejection technique would then result in a set of uniform points over the convex hull.

It is easy to reject points if the convex hull is known. However, we wish to find a rejection method that is less costly than explicitly determining the convex hull. Our 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 Y can be separated from the set {Xi} by a hyperplane. This follows from the definition of H(X) as the intersection of all convex subsets containing {Xi}. We can restate this property as follows. 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).n) > 0 for all i=1,2,..,N, where (v.w) is the inner product of vectors v and w. 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 would like to estimate n* from the given data. This, however, does not appear to be an easy problem to analyze and we resort to a heuristic. If the data are uniform over H(X), we expect to see points in the data set which are near the point Z. These points could be used to estimate n*. We choose to use the following estimator which takes a weighted average over all points in the data set.

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

This estimator is the sum over all i of the unit vectors from Y to Xi weighted by an amount inversely proportional to the 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 use of volume, rather than distance, is suggested by the volume-based test.

The procedure to compute a second sample of points uniformly distributed over (approximately) the convex hull of the data is as follows. We place a simple compact convex set around the data. This set is chosen as the MVU estimator of an aligned hyper-rectangle. The smallest hypersphere enclosing the data could also have been used here. We then consecutively generate points uniformly distributed over this set. We reject any point, Y, 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. This procedure continues until the desired number of points has been generated. If a point Y is in the convex hull of the data, we are guaranteed that this procedure will not reject Y. However, it is possible for a point Y outside the hull to be accepted. This procedure is less costly than computing the convex hull explicitly if the initial convex set used to enclose the data is not too large. With this method, the time taken to reject one point is proportional to N.

This procedure is demonstrated in Figure 10. One hundred uniform points are given inside a triangle 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

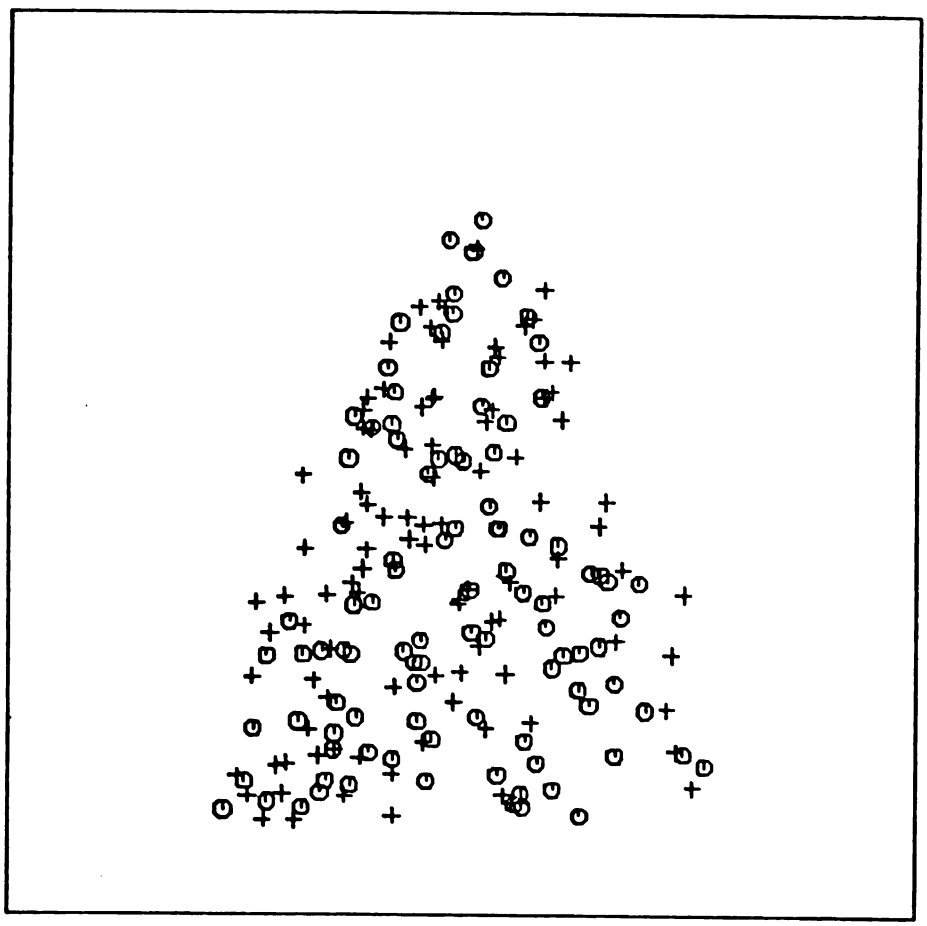


FIGURE 10: Points Generated by the Rejection Technique.
100 points "o" are generated uniformly in the triangle.
Another 100 points are shown as "+" after passing the
rejection procedure. A total of 406 points were generated
randomly over the unit square to obtain this second sample.

uniformly scattered like the original data in the triangle (denoted o).

It should be noted that this rejection procedure can be posed as a classical Pattern Recognition problem. We wish to find a linear discriminant function seperating the 'class' of points {Xi} from the 'class' containing one point Y. A number of algorithms are available to solve this problem [Dud73]. However, all of these algorithms are iterative in nature and most suffer from convergence problems. In addition, they do not offer any computational advantage over the rejection technique which we have presented.

6.3 Definition of the Test

We wish to test whether the original data and points uniformly distributed over a set which is approximately the convex hull of the data belong to the same population. That is, we wish to determine if the two samples have the same distribution function. Such a test would determine if the given data are uniformly distributed. Of course, one must keep in mind that the second sample is not independent of the first. This problem always exists whenever the sampling window is not known. We will see, however, that this only serves to make the proposed test more conservative.

Testing the equality of two univariate samples is a well-studied problem [Con71]. Classical univariate tests for general alternatives include the Wald-Wolfowitz runs test and the Kolmogorov-Smirnov two-sample test. Extension of the two-sample K-S test to a multivariate

situation is possible but is not practical [Hal73]. We explore the use of the multivariate extension of the Wald-Wolfowitz test recently proposed by Friedman and Rafsky [Fri79].

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 [And73, Zah71, Dub80]. The definition of an 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 [Har72]. For given data, if the set of distances between points has no ties then its MST is unique. Variants of Prim's algorithm [Pri57] are most widely used for forming an MST.

The MST extends to higher dimensions the concept of the one-dimensional sorted list needed to perform the Wald-Wolfowitz test. The Friedman-Rafsky test makes use of this fact in the following manner. The MST of the pooled samples is 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. We assume that the underlying distribution function of the samples is continuous so that T is unique with probability one. Note that $1 \leq T \leq (M+N-1)$. Under the null hypothesis that the two samples are from the same population,

Friedman and Rafsky show that

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

and

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

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 the realized graph, is asymptotically normal. That is,

$$\frac{T - E[T]}{VAR[T|C]} \Longrightarrow Z \text{ as } M,N \to \infty \text{ with } M/N \text{ bounded away from 0 and } \infty,$$

where Z is a random variable following the standard normal distribution. Friedman and Rafsky discuss the details of computing T and C. The most expensive part of using this test is in determining the MST which has computation time proportional to $(M+N)^2$. Bentley and Friedman [Ben78] present a MST algorithm whose expected run time is roughly proportional to $(M+N)\log(M+N)$.

In the context of our situation, 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, we expect the null hypothesis of the Friedman-Rafsky test to be true. In the case of clustered data, we expect many of the points labeled Y 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. We can thus

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

$$\frac{T - E[T]}{VAR[T|C]} < Z(\alpha)$$

where $Z(\alpha)$ is the α quantile of the standard normal distribution. We could, of course, perform the analogous upper tail test for uniformity against a hardcore alternative.

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 the MVU hyper-rectangle containing the data.

(2) Generate uniformly distributed points over this hyper-rectangle. Using the rejection technique, retain N of these points which lie in 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 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, we can replace steps 1 and 2 by

generating N points uniformly distributed over this window.

6.4 Performance of the MST-Based Test

In this section we analyze the performance of the MST-based test by simulation. We report only the rejection rates for a one-sided test against a clustering alternative. Since the distributional theory of the test statistic T is asymptotic, our results report the rejection rates of the test at both the .05 and .02 levels. The entries in the tables are (R(.05) , R(.02)), where $R(\alpha)$ is the percent rejections of the null hypothesis at the α level. The parameters K, N, and M in these simulations are the same as in experiments with the volume-based test reported in Chapter 5.

6.4.1 Uniform Data Over a Known Hypercube

Table 14 reports the results when a sample of uniform data in the unit hypercube is subjected to the MST-based test. The second uniform sample is also generated over this known sampling window. The results

TABLE 14: Size of the MST-Based Test for Uniform Data in Unit Hypercube. M=100

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

in the table provide a study of the effect of varying N and K. Since all the entries in the Table 14 are within their expected values (at the .05 level), we conclude 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.

6.4.2 Neyman-Scott Process with Known Sampling Window

Table 15 gives estimates of the power of the MST-based test for Neyman-Scott clustering alternatives. To generate this table, we assume that the hypercubic sampling window over which the data is generated is known. Further, to compare our results with previous studies, we use the wrap around paradigm, both for generating the data 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 increasing power

TABLE 15: Power of the MST-Based Test Against a Neyman-Scott Process (wrapped) in Unit Hypercube. N=200, M=100

		σ			
		.05	.1	.2	
μ	16	(100,100)	(86, 74)	(12, 7)	
	8	(100,100)	(56, 37)	(4, 1)	K=2
	1	(46, 28)	(11, 2)	(5, 1)	
μ	16	(100,100)	(100,100)	(46, 32)	
	8	(100,100)	(100,100)	(29, 18)	K=5
	1	(100,100)	(99, 94)	(15, 6)	

TABLE 16: 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)
M=100, N=200, μ=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
	.05	.1	.2
		σ	

*entries not provided by Cross and Jain [Cro82]

with dimensionality. Comparing these results with Table 7, which compared the powers of the volume-based and Hopkins tests for various parameters of Neyman-Scott clustering, we see that the MST-based test is the most powerful of these tests for these parameters. We can also compare the Hopkins test [Cro82] and the Cox-Lewis test [Pan81] to the MST-based when μ=16. Table 16 gives these comparisons. Again we see higher power for the MST-based test. The MST-based test gives significantly higher power (at the .001 level) against all other tests for K=2 with σ=.05 and .1 and for K=5 with σ=.2.

6.4.3 Other Data Types with Known Sampling Window

As with the volume-based test we can estimate the power of the MST-based test for a hardcore alternative. To use this alternative we must change the test's critical region to the 1-α upper tail of the normal distribution since we expect too few X-Y joins under a regular alternative. We see considerable power of the MST-based test against this alternative. With 100 Monte-Carlo trials, we have rejections of

(64,46), (100,100), and (100,100) in 2, 4 and 5 dimensions, respectively. The alternative studied is hardcore over the unit hypercube with out wrap around, with ρ=0.1, and with N=200. These results are much better than the volume-based test, especially for two-dimensional data. We will see, however, that for an unknown sampling window, our point generation procedure will not allow the MST-based test to be performed with hardcore data.

We also look at the power of the MST-based test for detecting a normal swarm of points. As in the volume-based test, the normally distributed points 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 over this hypersphere. As in the volume-based test when N=200, we obtain rejection rates of 100% for K=2, 5, and 10. However, when N is decreased to 50, our rejection rates (at the .05 level) become 50, 84, and 76 for 2, 5, and 10 dimensions, respectively. These are higher than the corresponding rates for the volume-based test with P placed randomly but lower than the rates with P placed at the center of the hypersphere.

6.4.4 Uniform Data in Unknown Sampling Windows

Here we determine the size of the MST-based test for unknown sampling windows. Table 17 reports the size estimates for uniform data in a unit hypercube, while Table 18 reports similar results for data uniform in a unit volume hypersphere. We note that all entries are within or below their expected values. It appear that as dimensionality

TABLE 17: Size of the MST-Based Test for Uniform Data in an Unknown Unit Hypercube. M=100

		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)

increases, the test becomes more conservative, i.e. the observed number of rejections of the null hypothesis is less than expected. This can be verified by noting that the mean of the MST statistic T increases as dimensionality increases. This arises from the fact that the volume of the convex hull of the data underestimates the volume of the true sampling window. Thus we are packing uniformly distributed points inside the convex hull which decreases the data point to data point (X-X) joins more than would be expected under the null hypothesis of the Friedman-Rafsky test. This makes a test against clustering possible, though a loss of power may result. However, this excludes using the MST-based test as a test of uniformity versus a hardcore alternative, since the proper size of the test can not be set with an unknown sampling window. The disadvantage in having no prior knowledge of the

TABLE 18: Size of the MST-Based Test for Uniform Data in an Unknown Unit Volume Hypersphere. M=100

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

sampling window can be seen by the time taken to perform the MST-based test in high dimensions for a hyperspherical sampling window. The major component of the computation time is in generating the second sample. For 100 uniform points in a hypersphere the times taken to perform the test are 25, 70, 189, and 1040 CPU seconds in 2, 3, 5, and 10 dimensions, respectively, for 100 Monte-Carlo simulations.

Another type of uniform data is that generated over a hyperellipse. Appendix A gives details on the parameters of the hyperellipses used. For 100 uniform points over hyperellipses in 2, 3, 4, and 5 dimensions we obtain rejection rates of (3,1), (5,1), (1,0), and (1,1), respectively. The number of Monte-Carlo simulations performed is 100. Also, for uniform data over a triangle in two dimensions, the rejection rate of the MST-based test is (6,3) with 100 Monte-Carlo trials. This triangle is formed by partitioning the unit square along one of the diagonals. These results lead us to believe that the rejection procedure used in generating the second uniform sample allows us to determine the level of significance of the MST-based test.

6.4.5 Neyman-Scott Process over Unknown Sampling Windows

We now study the power of the MST-based test against Neyman-Scott clustered data over an unknown sampling window. In Table 19, we use the Neyman-Scott cluster alternative over a hyper-rectangular sampling window with wrap around. Of course, since the sampling window is unknown, distances are not computed using wrap around. These results

TABLE 19: Power of the MST-Based Test Against the Neyman-Scott Process (wrapped) in an Unknown Unit Hypercube. M=100, N=200

μ	16	(100,100)	(78, 72)	(14, 6)	
	8	(100, 99)	(40, 30)	(11, 4)	K=2
	1	(47, 37)	(12, 8)	(3, 0)	
μ	16	(100,100)	(100,100)	(21, 14)	
	8	(100,100)	(100,100)	(6, 5)	K=5
	1	(100, 97)	(68, 51)	(3, 1)	
		.05	.1	.2	
			σ		

show that the power of the MST-based test with unknown sampling window is remarkably similar to the case when the sampling window is known, though a slight loss in power can be seen. To check the effect of sampling window on the power of the MST-based test, we repeat the above simulations, but for the Neyman-Scott process over the unit volume hypersphere. Here, however, wrap around can not be used. Table 20 reports these power estimates. We note an increase in the power as dimensionality increases, even though the test is also becoming more

TABLE 20: Power of the MST-Based Test Against the Neyman-Scott Process in an Unknown Hypersphere. M=100, N=200

μ	16	(100,100)	(100,100)	(76, 60)	
	8	(100,100)	(100, 96)	(38, 32)	K=2
	1	(90, 86)	(38, 28)	(4, 4)	
μ	16	(100,100)	(100,100)	(93, 92)	
	8	(100,100)	(100,100)	(43, 40)	K=5
	1	(96, 92)	(96, 92)	(10, 6)	
		.05	.1	.2	
			σ		

conservative. There is an increase in power from the unknown hypercube results, though this is probably due to the difference between the realizations of a no wrapped and a wrapped Neyman-Scott cluster process.

6.4.6 Experiments with Some Real Data

To demonstrate the applicability of the MST-based test in practical situations, we use this test to test for the presense of structure 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:

(1) IRIS....This is a well-known data set [Fis36] containing measurements on three species of iris (setosa, versicolor, and virginica). It consists of 50 patterns from each species on each of 4 features (sepal length, sepal width, petal length, and petal width). See Figure 11 for a projection of the IRIS data to two dimensions by the principal component method [Fuk72]. The axes are the eigenvectors corresponding to the two largest eigenvalues of the covariance matrix of the data.

(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 specie. Figure 12 shows this data projected to two dimensions by the principal component method.

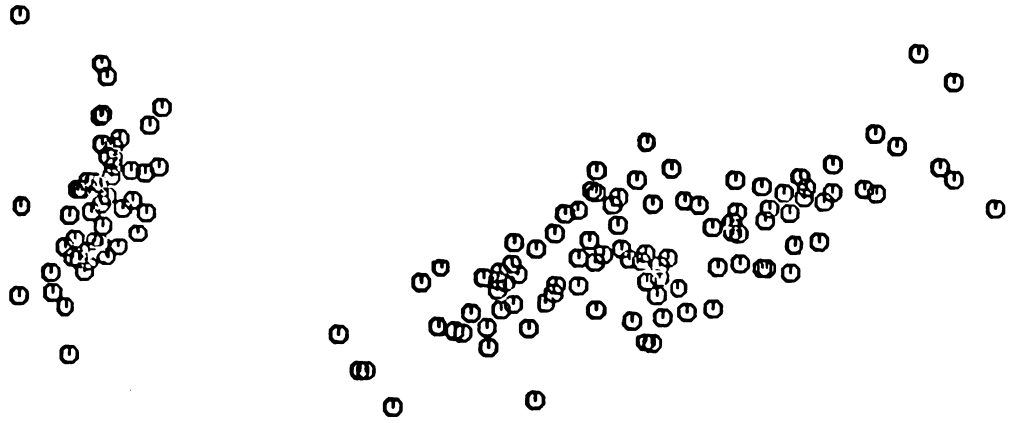


FIGURE 11: IRIS Data Projected by the Principal Component Method

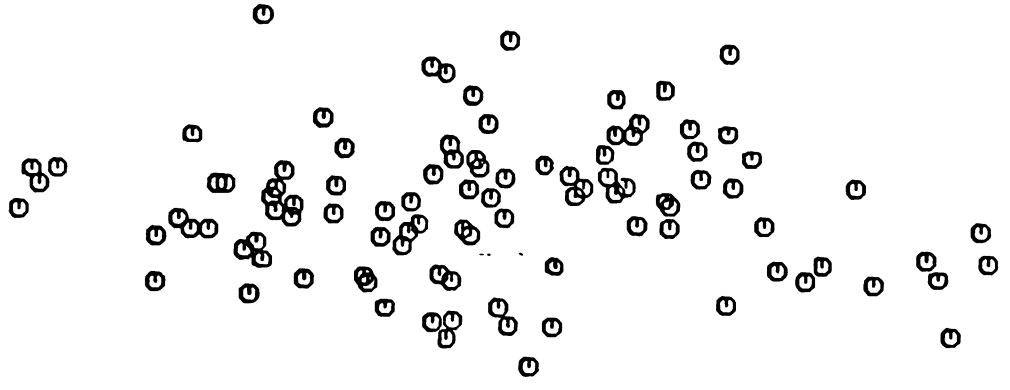


FIGURE 12: IRIS23 Data Projected by the Principal Component Method

(3) 80X....The 80X data set is derived from the Munson hand printed FORTRAN character set. Included are 15 patterns from each of the characters "8", "0", and "X". Each pattern consists of 8 feature measurements [Dub80]. Figure 13(a) shows the 80X data projected to two dimensions by principal component analysis while Figure 13(b) shows the 80X data projected to two dimensions by discriminant analysis [Fuk72].

(4) BCLUS....This data set, used by Bartlett [Bar64], consists of 100 patterns generated according to a Neyman-Scott cluster process over the unit square. Bartlett was able to show that this data was nonuniform with a spectral analysis technique. In our analysis of this data we assume that the sampling window is unknown. Figure 14(a) shows the original BCLUS data, while Figure 14(b) shows the BCLUS data after it has been subjected to the whitening transformation [Fuk72].

(5) SPEECH....This data set consists of patterns measured on 72 utterances from 8 Chinese speakers [He82]. Each pattern consists of 5 features measured from the pitch waveform. The principal component projection of this data to two dimensions is shown in Figure 15.

The data sets are tested in each of the following configurations.

(1) The original feature space.

(2) The patterns are transformed so that the data has 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 [Fuk72].

(3) The patterns are projected to two dimensions by the principal

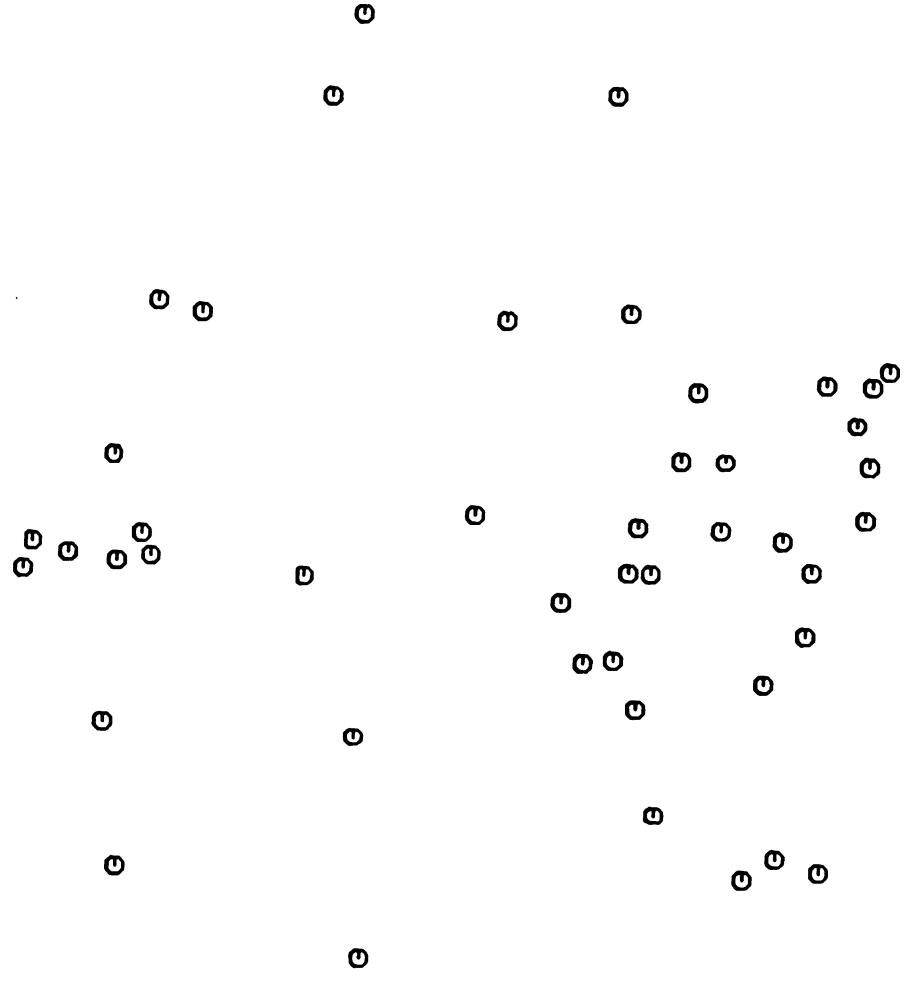
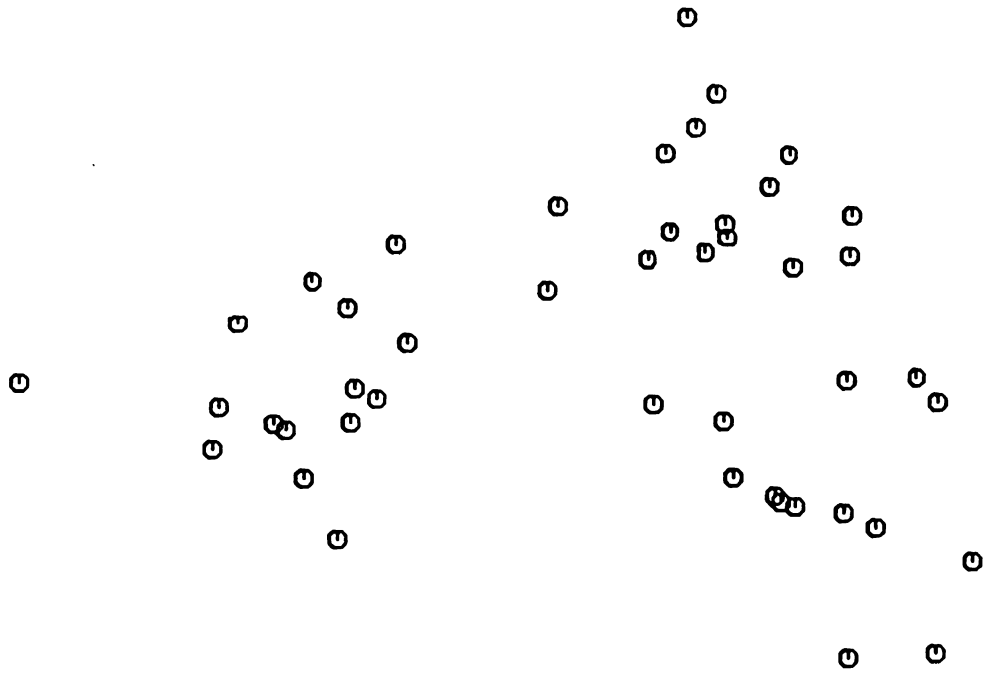


FIGURE 13: 80X Data
(a) Projected by the principal component method

FIGURE 13 (cont'd)



13(b) 80X data projected by discriminant analysis

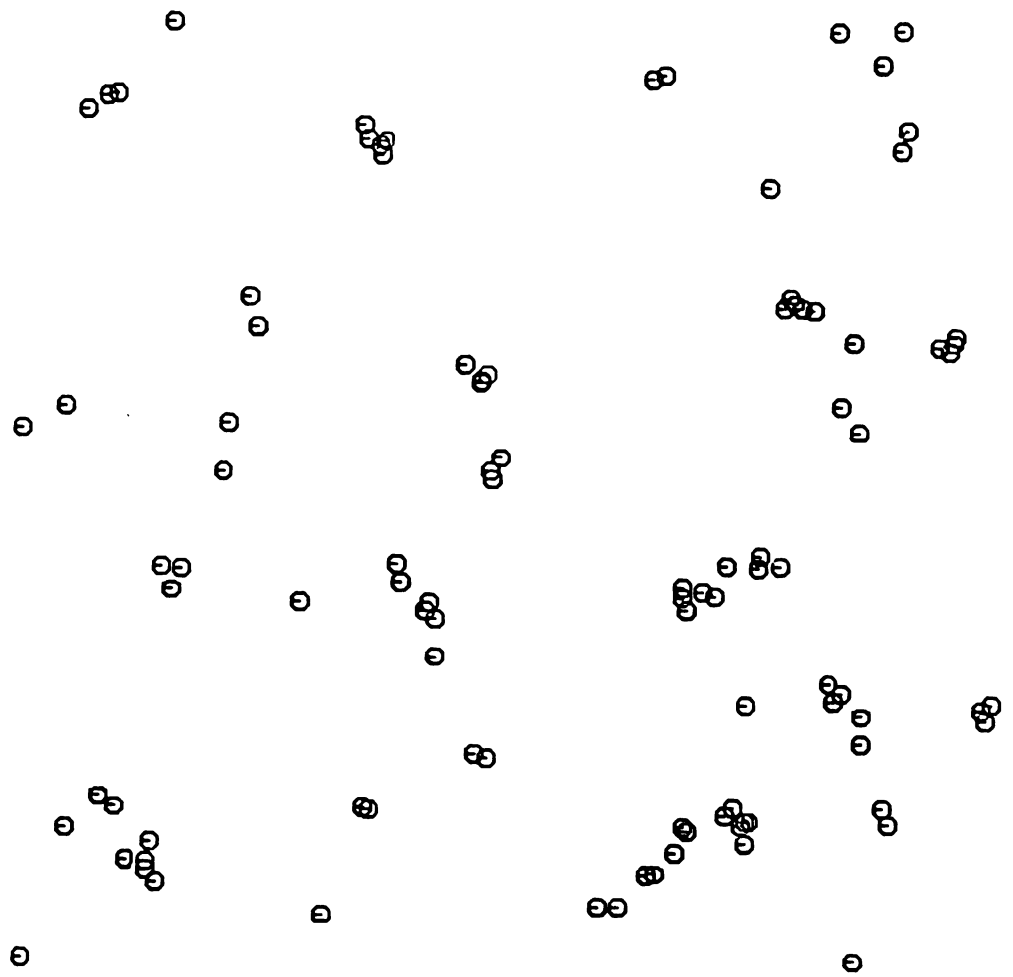
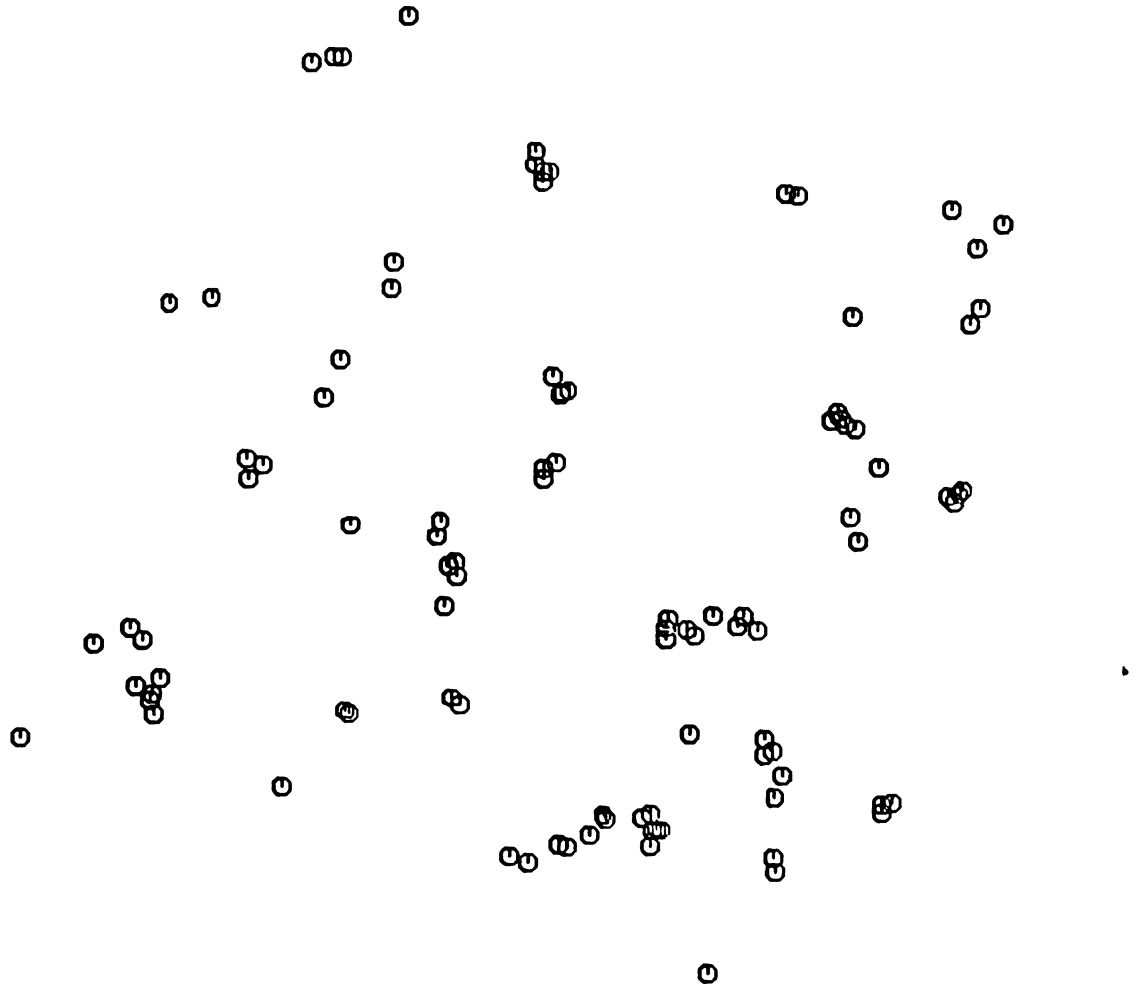


FIGURE 14: BCLUS Data
(a) original

FIGURE 14 (cont'd)



14(b) BCLUS data after whitening transformation

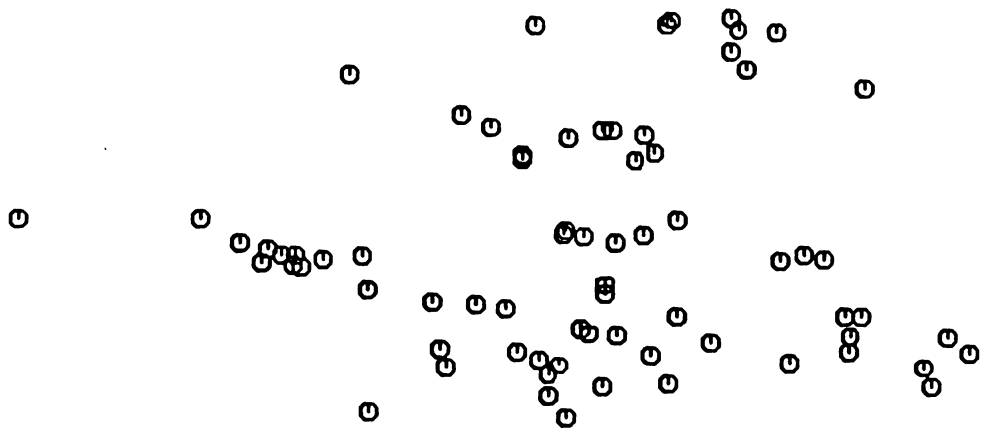


FIGURE 15: SPEECH Data Projected by the Principal Component Method

TABLE 21: 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 MVU or smallest hypersphere sampling window.

	Data Sets				
Configuration	IRIS	IRIS23	80X	BCLUS	SPEECH
Original	-11.08	-3.59	-1.90	-4.24	-4.50
	-12.91	-8.77	-4.64	-2.97	-6.33
Transformed	-5.42	-2.83	-.90	-4.94	-3.50
	-6.46	-5.45	-3.79	-6.13	-6.00
Projected	-7.27	-1.55	-1.68	-	-2.50
	-6.00	-2.12	-1.08	-	-1.83

component transformation.

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 needed for the MST-based test. Second, the MVU 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 21. Except in a few instances, the entries in the table are significant for rejecting the null hypothesis of uniformity at the .05 level (the .05 quantile of the normal distribution is -1.65). These exceptions are the transformed 80X data tested by using the rejection technique, the 80X data projected to two dimensions and tested in the smallest circle, and the IRIS23 data

set projected to two dimensions and tested by using the rejection technique. Even in these cases, the value of the test statistic indicates the presence of a slight clustering in the data. This can be confirmed by Figures 12 and 13(a) for the two projected data sets. There is a general trend in Table 21 which shows that the test statistic using the rejection technique is larger than when the best fitting sampling window is used, especially in high dimensions. This is expected since the high dimensional data does not usually fit very well inside the MVU 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.

The original 80X data appears clustered at approximately the .03 level. Of course, 45 patterns in 8 dimensions constitute a rather sparse data set and it is difficult to make a meaningful decision about the structure of such data. These data, after applying the whitening transformation, are accepted as uniform at the .05 level, which suggests that the whitening transformation can distort the structure present in the data. The representation of the 80X data in two dimensions produced by the principal component transformation has critical level approximately equal to .05 when using the rejection method. Even if the 80X data is projected to two dimensions using discriminant analysis, the critical level drops to only .03. This suggests that the clusters in the data are not compact or well-separated. This can be verified from Figure 13(b). Of course, if category information (class labels) is used, it may still be possible to determine simple decision boundaries

to seperate the three classes present in this data.

In contrast to the 80X data, there is evidence for strong clustering in the IRIS data. The values of the test statistic for all representations of the IRIS data have critical levels less than .001. This strong clustering can be seen in the two-dimensional representation of the IRIS data (Figure 11). By deleting the compact and well-separated class (setosa) from the IRIS data in the IRIS23 data, we can see an increase in the value of the test statistic.

The BCLUS and SPEECH data sets both show strong clustering tendency. The BCLUS data seems to appear slightly more clustered after the transformation.

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 is retested by both the rejection and the best fitting window methods, the test statistic values are .21 and -1.90, respectively. These new values suggest that this data are even less clustered than suggested previously. If the original IRIS data are retested we get about the same value of the test statistic for the rejection method, but the value using the MVU 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 level .05 say) then the interpretation requires caution. One solution is to perform the test

with various uniform 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 structure of real data.

6.5 Summary

In this chapter we have presented a new test for uniformity, called the MST-based test. The given data are tested against a second uniform sample which needs to be generated. The test statistic is derived from the MST of the pooled samples. If the sampling window is known, this second sample can be taken as uniform data over that sampling window. In this case, a Monte-Carlo study of the size and power of the MST-based test shows that the test performs very well. The power of the MST-based test is significantly higher (at the .001 level) compared to other tests against clustered data. If the sampling window is unknown, we present a point rejection procedure which places samples uniformly in a set which approximates the convex hull of the data. The size and power of the MST-based test using this rejection procedure are shown by a Monte-Carlo study. We conclude that the size of the MST-based test can be preset to a specified level of significance when testing against a clustered alternative. Due to the conservative nature of the MST-based test on the clustered alternative, we can not apply the test on a hardcore alternative in unknown sampling window. To demonstrate the applicability of the MST-based test we have applied it to some real data

sets.

CHAPTER 7

SUMMARY, DISCUSSION, AND FUTURE RESEARCH

7.1 Summary

Our goal is to differentiate data sets with structure from those with no structure. The structure we are most interested in is one of clustering or aggregation of points. We wish to use a statistical hypothesis test to make the decision. To do so, we define data with no structure as a set of independent points following the uniform distribution over a compact convex set in K-dimensional space, called the sampling window.

A careful review of the currently available tests for structure reveals three major deficiencies with these tests.

(1) Inapplicability to high dimensional data,

(2) The sampling window needs to be known, and

(3) Reliance on a Poisson process null hypothesis.

We feel that these limitations make the available tests inapplicable for most data in a Pattern Recognition environment.

The focus of this research has been on finding tests that would address some of the above problems. We presented a volume-based test which has the ability to compare density changes in the data to the changes expected for uniformly distributed points over a known sampling window in K-dimensional space.

To apply the volume-based test on data over an unknown sampling window, an accurate estimate of the true sampling window is needed. We presented a number of estimators when the shape of the window was known to be a hyper-rectangle, a hypersphere or a hyperellipse. For an arbitrary sampling window, the convex hull of the data appears to be an adequate estimator. The volume of the convex hull is not an unbiased estimator of the volume of the true sampling window.

We performed Monte-Carlo simulation to evaluate the volume-based test. For known sampling windows, we found that the size of the volume-based test could be fixed and that the power of the volume-based test against clustered data is comparable to that of other tests in the literature. We studied the proposed sampling window estimators by using the estimates as the true sampling window in the volume-based test. We found that we were able to set the size of the volume-based test using the MVU estimator for an aligned hyper-rectangular sampling window and the smallest hypersphere estimator for a hyperspherical sampling window. The attempt to transform points in a hyperelliptical sampling window into a hypersphere failed; the transformation did not produce uniform data in a hypersphere according to the volume-based test. In two

dimensions, the convex hull estimator yielded a test with determinable size and with power comparable to the test performed over a known sampling window. However, we found this estimator computationally infeasible to apply to high dimensional data.

The requirement that the volume and shape of the sampling window be precisely known limits the applicability of the volume-based test. To overcome this limitation, we developed a MST-based test which assumes only that the convex hull is a reasonable estimate of the sampling window. This MST-based test uses the Friedman-Rafsky multivariate extension of the Wald-Wolfowitz 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 sampling window of the given data. 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. This, however, violates an assumption of the Friedman-Rafsky test that the two samples be independent.

We found that if the sampling window is known then the size of the MST-based test could be determined, even for small sample size. The power against the clustered alternative is significantly higher (at the .001 level) than other tests. For unknown sampling windows, we found that the MST-based test was 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.

We conclude that the MST-based test is better than any other test given in the literature to determine if data have any clustering structure.

7.2 Discussion

The major contributions of this thesis have been in two areas. First, we have defined two new tests for uniformity of given data. For each of these tests, we have performed Monte-Carlo studies on the size and power against various alternatives. These tests appear powerful against clustered alternatives.

Second, this thesis provides the first study of tests for structure in data when the sampling window is unknown. We provide a number of sampling window estimators when the shape of the sampling window is known *a priori* to be an aligned hyper-rectangle or hypersphere. For a general sampling window, we have found that the convex hull of the data should be used as its estimator. Unfortunately, computing the convex hull of high dimensional data is computationally very demanding. To alleviate this problem, we developed a test that does not explicitly compute the convex hull of the data. This is the first study which has provided a test that can be used to detect clustering in data over an unknown sampling window.

We now list some advantages and disadvantages of each of the tests proposed in this thesis.

The volume-based test has the following advantages:

(1) The null hypothesis in the volume-based test is that the data are uniform over some compact convex set. This contrasts with other tests which use the null hypothesis of a Poisson process and eliminates the need for an 'edge correction' factor.

(2) There appears to be little effect of dimensionality on the size and power of the volume-based test.

(3) The volume-based test directly measures the density of the sample points as opposed to some distance-based tests which utilize only local information in the data.

The volume-based test has the following disadvantages:

(1) Using a single point P allows only one view of the data. Different placements of P may yield different views of the data. It is not clear how to utilize this information.

(2) The volume-based test is very sensitive to the size and shape of the sampling window. This limits its applicability to low dimensional data ($K<4$) when the convex hull is used as an estimate of the true sampling window.

(3) It is necessary to compute the volumes and the intersection of sets

in high dimensions, which is feasible only if the sets are simple.

The advantages of the MST-based test are:

(1) For known sampling window, the power of the MST-based test against clustered alternatives is significantly better than some well-known distance-based tests.

(2) The test does not require the exact shape and volume of the sampling window. Therefore, it can be easily applied to high dimensional data ($K \geq 4$) in an unknown sampling window.

The MST-based test has the following disadvantages:

(1) Generating the uniformly distributed sample over an unknown sampling window may require large amounts of computation time.

(2) The test against regularity can not be performed with unknown sampling window.

(3) The MST-based test could yield different results on the same data depending on the second sample which needs to be generated uniformly over the sampling window.

7.3 Future Research

Listed below are areas of investigation that extend the work in this thesis.

(1) 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 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. As an example, one would like to combine the results from different placements of point P in the volume-based test.

(2) For the MST-based test, a study of other methods of rejecting points outside the convex hull of the data should be undertaken. Our estimate of n^* was based on a heuristic which appears to give reasonable results. However, one may be able to derive a different estimator that does not lead to the conservative trend against clustered alternatives in high dimensions. This would also allow us to define a test against regularity in unknown sampling windows.

(3) The idea of using points generated over the convex hull of the data may now allow tests based on sampling origins (such as the Cox-Lewis test) to be used with data in unknown sampling windows. One would need to check if the null distribution of the statistic would still hold. Using these tests may have an advantage over the MST-based test since fewer sampling origins need to be generated. Also, the assumption that the sampling origins are uniformly distributed is usually unnecessary to apply these tests. This may allow the origins to be placed inside the convex hull by taking linear combinations of random pairs of data points.

(4) Our brief study of Ripley and Rasson's unbiased estimator of the volume of the sampling window for planar data indicated that it appears to be applicable to higher dimensional data. A proof of this extension would be satisfying.

(5) 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 or may have a regular structure. Therefore, it is necessary to look more closely at this non-uniform data. It would be exteremely useful to know the number of clusters present in the data. Perhaps the sequence of ordered volumes generated from the volume-based test would be helpful here. It is possible that knowing the number and location of significant 'holes' in the data would also be useful. A significant 'hole' could be defined as an unusually large subtree of Y points using the MST-based test.

(6) Our analysis of the computational complexity of various algorithms has assumed a serial model of computation. It may be possible to develope parallel algorithms to generate the convex hull and MST of the data, perform volume and set intersection computations, generate random samples, etc. An analysis of the computational requirements of the tests should be undertaken with this in mind.

APPENDICES

APPENDIX A

GENERATION OF RANDOM VARIABLES

This appendix describes the methods used to generate the random variables and data set ensembles used in this study.

A.1 Uniform Random Variates

All uniform random variables used in this study were generated by the 'Randomization by Shuffling' method (Algorithm M) described by Knuth [Knu81, p. 32]. The auxillary table size was set to 64 elements. The random number generator used to fill the table was the RANF generator provided in FORTRAN IV on the CYBER 170/750. The generator used to determine which element in the table to return at each call was a linear congruential generator with multiplier 16807 and modulus 2147483647. It is hoped that the use of this generator will avoid a problem of sequentially generated random samples used as K-dimensional point coordinates. Knuth [Knu81] has pointed out that such K-dimensional points tend to fall 'mainly in the planes' [p. 90 Knu81].

A.2 Normal Random Variables

The standard normal random variables used in this study were generated by the algorithm given in Kinderman and Ramage [Kin76]. This method uses a modest number of uniform deviates to produce a single

normal deviate by a rejection technique. It proceeds by decomposing the normal density into a number of regions where simpler density functions can be defined. Uniform deviates are then used to produce deviates following the appropriate component density. Samples from a multivariate normal distribution with diagonal covariance matrix were produced by using this method to generate each coordinate value independently.

A.3 Poisson Random Variables

Poisson random variables were generated from uniform deviates by using the algorithm given by Knuth [Knu81]. Essentially, we simulate a Poisson process on the line. We can produce a Poisson deviate, X, with mean r by generating independent exponential samples with mean 1/r, denoted Y1,Y2,...., stoppping as soon as Y1+Y2+...+Ym>1; then X <--(m-1). Simplifying this, we see that X can be obtained by generating one or more uniform deviates U1,U2,... until the product (U1)(U2)...(Um) < exp(-r), finally setting X <--(m-1). On the average, this procedure requires r+1 uniform deviates.

A.4 Uniform Random Vectors in a Hypersphere

Generating a sample from a uniform distribution in a K-dimensional aligned hyper-rectangle is a trivial combination of K one-dimensional uniform random variables, due to the independence among the coordinates. However, generating random vectors uniform in a hypersphere can be

computationally burdensome if a rejection technique is employed. To perform this generation efficiently, we use the method described by Pettis et. al. [Pet79], which uses K normal deviates and a uniform deviate to place a vector uniform in a hypersphere. A random vector following the normal distribution with zero mean and identity covariance matrix is generated. This vector is normalized to have unit length. Due to the lack of directionality in the normal density, this vector is uniformly distributed on the surface of the unit hypersphere. Finally, the length of the vector is made proportional to the Kth root of a uniform deviate, thus placing the vector uniformly inside the unit hypersphere.

Data uniformly distributed over a hyperellipse is derived from uniform hyperspherical data by applying a linear transformation. This transformation, T, is defined as follows. Transform each of the standard coordinate basis vectors, dj, j=1,...,K (the jth component of dj is 1, all other components are 0) into the orthogonal direction vectors as follows:

$T(d1) = (1,1,\ldots,1)$

$T(dj) = (yj(1),yj(2),\ldots,yj(K))$ $j=2,\ldots,K$ where

$yj(i) = 0$, for $i=1,\ldots,j-2$ (when $j\neq 2$)

$yj(i) = -(K-j+1)$ for $i=j-1$

$yj(i) = 1$ for $i=j,\ldots,K$

Further, these direction vectors are normalized so that

$$||T(dj)|| = \begin{cases} \text{modulo}(j,5) & \text{if } j \text{ is not a multiple of } 5 \\ 5 & \text{if } j \text{ is a multiple of } 5 \end{cases}$$

In two dimensions, this transformation takes the unit circle into an ellipse with major axis of length two on the y=-x line and minor axis of length one on the y=x line.

A.5 Neyman-Scott Ensembles

A Neyman-Scott cluster process [Ney72] is a stochastic point process representing a clustered alternative. It uses a Poisson field of points as cluster centers and generates daughter points around each cluster center with some specified distribution. To simulate this process over a bounded sampling window, S, we use the following steps.

First, three parameters of the process are specified:

(1) N, the number of points desired,

(2) μ, the average cluster size, and

(3) σ, the spread of each cluster.

Next, the following algorithm is applied.

(1) Select a point Y at random from S as a sample point. Y will also serve as a cluster center.

(2) Find the number of daughter points, L, to be placed about Y. Let M

be the number of points generated so far. Set L=min(N-M,X), where X is a Poisson random variate with mean μ.

(3) L points are then generated using the radially symmetric normal density with mean Y and covariance matrix $\sigma^2 I$, where I is the identity matrix. If wrap around is to occur, the point positions are taken modulo the sampling window; otherwise, if a point falls outside the window, it is rejected and new points are generated until one falls in the window.

(4) Steps (1) through (3) are repeated until N points have been generated.

To avoid edge effects under the null hypothesis of a Poisson process, previous studies have generated data from a Neyman-Scott process (as well as under other alternatives) using wrap around. The concept of wrap around may only be defined in the case of a hyper-rectangular sampling window. In the case of an aligned hyper-rectangle, each of the two faces of the hyper-rectangle associated with every coordinate axis are identified as being adjacent. To place an arbitrary point Y=(y1,y2,...,yK) in the aligned hyper-rectangular sampling window $S=[ai,bi]_{i=1}^{K}$, we change each of the coordinate values to yj<--modulo(yj,bj-aj)+aj, j=1,..,K. Distances can also be computed with wrap around as follows. Let X and Y be two points in $S=[ai,bi]_{i=1}^{K}$ with components xi and yi, respectively. Let Z be defined as the vector with components zj=min(|xj-yj|, (bj-aj)-|xj-yj|). The distance between X and Y is then ||Z||, using whatever norm is appropriate.

A.6 Hardcore Ensembles

Hardcore data sets used were generated under the SSI (Simple Sequential Inhibition) model of Diggle, Besag and Gleaves [Dig76]. This method places N points in the sampling window S consecutively according to the rule that the ith point is distributed uniformly over the set of all points in S at Euclidean distance of at least d from all previously located points. The parameter of the SSI process is its packing density,

$$\rho = N \cdot A_K \cdot (d/2)^K$$

where A_K is the volume of a unit hypersphere in K dimensions. The value of ρ, barring edge effects, is the proportion of S covered by N non-intersecting spheres of diameter d.

To implement the SSI process, we use a rejection technique that consecutively generates uniform points over S and checks if they satisfy the minimum distance criterion. This procedure uses an inordinate amount of computation if ρ is large.

APPENDIX B

THE VOLUME OF THE INTERSECTION OF TWO HYPERSPHERES

Let the two hyperspheres be denoted as S(c1,r1) and S(c2,r2), where S(c,r) stands for a sphere of radius r with center c. The following prodedure returns the volume of their intersection.

```
INTVOL( S(c1,r1) , S(c2,r2) )
  D <-- || c1-c2 ||2
  IF   (D >= r1+r2) THEN ; spheres don't intersect
       INTVOL <-- 0
       RETURN
  IF   (D+r2 < r1) THEN  ; sphere 2 is inside sphere 1
       INTVOL <-- Ak * (r2)**K
       RETURN
  IF   (D+r1 < r2) THEN  ; sphere 1 is inside sphere 2
       INTVOL <-- Ak * (r1)**K
       RETURN
                         ; need to find both spherical caps
```

$$B1 \leftarrow\!- \cos^{-1}\left[r_1^2 + d^2 - r_2^2 \,/\, 2 r_1 d \right]$$

$$B2 \leftarrow\!- \cos^{-1}\left[r_2^2 + d^2 - r_1^2 \,/\, 2 r_2 d \right]$$

```
  INTVOL <-- CAP( r1 , B1 )  +  CAP( r2 , B2 )
  RETURN
  END
```

Here Ak is the volume of a unit hypersphere in K dimensions,

$$Ak = \pi^{K/2} \,/\, \Gamma\left(\tfrac{K}{2}+1\right)$$

The function CAP(r,B) returns the volume of a spherical cap of a

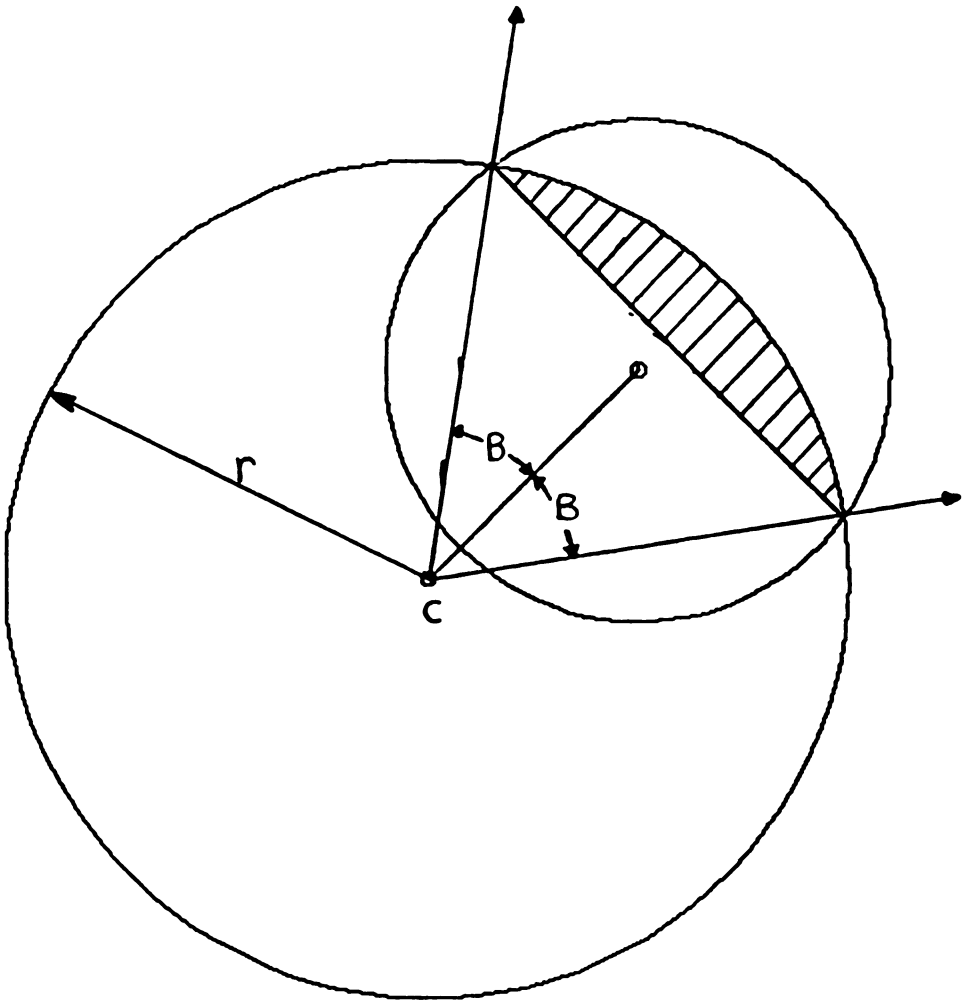


FIGURE 16: Definition of B to Compute Spherical Caps.
The spherical cap for S(c,r) is the shaded area

hypersphere of radius r, where B is the angle between the line generator of the cap and the symmetric axis of the cap. Figure 16 shows the definition of B in two dimensions. In our case, B1 is the angle between the line segment (c1,c2) and the line segment beginning at c1 and ending at any point on the surface of both spheres. B2 is similarly defined. The formula for a cap's volume is given by Panayirci and Dubes [Pan81] and can be rewritten as

$$\text{CAP}(r,B) = \frac{r^K}{2} A_K \, I_{\sin^2 B}\left[\tfrac{1}{2}(K+1), \tfrac{1}{2}\right]$$

$$= A'_K \, r^K \int_0^B \sin^K \gamma \, d\gamma$$

where $I_x[a,b]$ is the incomplete BETA function

and

$$A'_K = \frac{\pi^{\frac{K-1}{2}}}{\Gamma\left(\frac{K+1}{2}\right)}$$

APPENDIX C

COMPUTING THE SMALLEST HYPERSPHERE

The smallest hypersphere problem can be stated as follows. Given N points, {Xi}, in K dimensions, find the smallest hypersphere which contains all N points. That is, find a vector c such that $\max||Xi-c||$ is minimized over all c in R^K. Historically, the smallest hypersphere problem was first proposed for the planar case (K=2) in 1857 by J.J. Sylvester [Syl57]. Later, he gave a geometric solution attributed to Pierce [Syl60] which was rediscovered by Chrystal [Chr85]. A modern account of their technique is given by Rademacher and Toeplitz [Rad57]. For K>2, Lawson [Law65] gives an iterative algorithm that converges to the smallest hypersphere while Elzinga and Hearn [Elz72] were the first to exactly solve the problem. Their technique uses quadratic programming. Supowit [Sup81] gives an algorithm based on a grid heuristic.

We use Elzinga and Hearn's exact solution for the smallest hypersphere. We wish to solve the following primal problem:

Minimize r over all positive $r \in R$ and $c \in R^K$ such that

$$r^2 \geq (Xi-c)^t (Xi-c), \quad i=1,2,\ldots,N.$$

After rewriting this in terms of its Wolf dual [Elz72], we have the

following concave quadratic programming problem,

$$\underset{V}{MAX} \sum_i V_i (X_i^t X_i) - V^t (X^t X) V \quad \text{such that}$$

$$\sum_i V_i = 1 \quad \text{and} \quad V_i \geq 0 \,, \; i=1,2,\ldots,N$$

where Vi, i=1,2,...,N are Lagrange multipliers, V is the vector of Vi's and X is the matrix whose columns are the Xi. The smallest hypersphere is then specified as

$$C = \sum_i V_i X_i \quad \text{and} \quad r^2 = \sum_i V_i (X_i - C)^t (X_i - C)$$

Solving this problem by a simplex algorithm involves the introduction of an additional N nonnegative multipliers and an unconstrained multiplier. This means that the simplex algorithm will have basis size proportional to N. To reduce the complexity of the problem, we decompose the problem as follows. For any subset of K+2 points of {Xi}, we can find the smallest hypersphere containing these points by the simplex algorithm. If this hypersphere contains all N points, we have found the optimal hypersphere for {Xi}. If not, we use the fact that the smallest hypersphere is defined by only K+1 points. Thus the unused point can be eliminated from the set of K+2 points and another inserted from the points in {Xi} which lie outside the current hypersphere. Elzinga and Hearn prove that this procedure halts in a finite number of steps. The algorithm is as follows:

(1) Given {Xi}, select K+2 points from this set.

(2) Solve the quadratic programming dual subproblem for the K+2

points by the simplex method. One point, say Xj, will not be in the optimal basis.

(3) If the hypersphere defined by the solution contains all N points then stop. Otherwise, choose a point from {Xi} outside this hypersphere and replace Xj by this point. Go to (2).

Specification of methods to select the K+2 points in step (1) and to select the next point to be added in step (3) complete the algorithm. These two choices can greatly affect the computational requirements of the procedure. Our implementation chooses the K+2 points farthest from the mean of the data. The point in {Xi} farthest from the current hypersphere's center is added to the current basis in step (3).

The run times of our implementation of the smallest hypersphere problem are presented in Figure 17. Ten sets of both 100 and 200 points were generated at random in the unit hypersphere in 2,3,5,10 and 15 dimensions. The times plotted are the averages over the ten runs. We note the fairly linear behavior in computation time for small dimensions. Other experiments by us have also shown a linear behavior with respect to N.

FIGURE 17

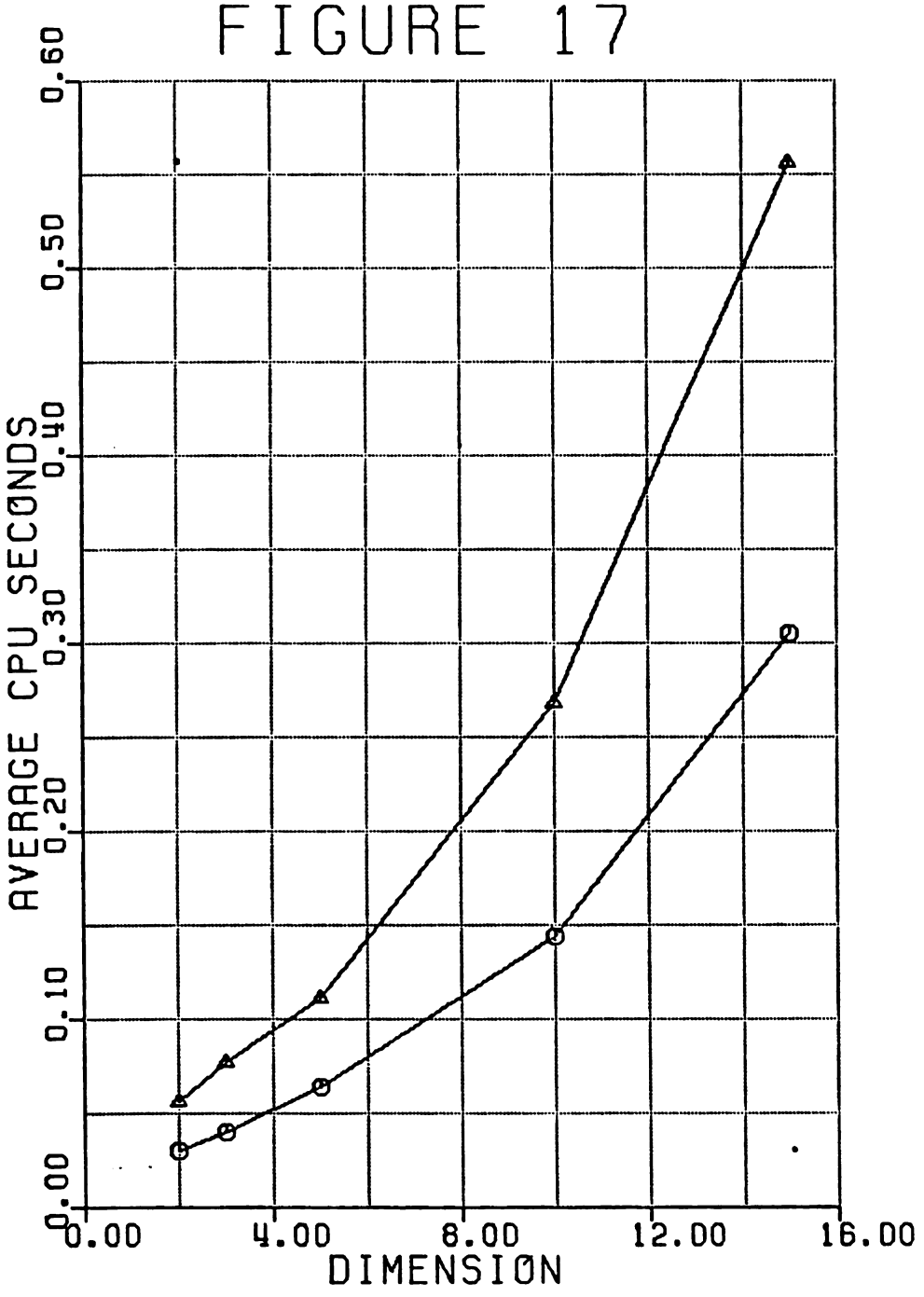


FIGURE 17: Run Times of the Smallest Hypersphere Algorithm. The times are averages over 10 runs of N points uniformly distributed in the unit hypersphere

⊙... N=100 Δ... N=200

APPENDIX D

THE CONVEX HULL OF A FINITE SET OF POINTS

Computing the convex hull of a set of N points is a well-studied problem. See Toussaint [Tou80] for a review. A number of algorithms have been proposed for the two-dimensional case, the best of which runs at minimum worst case time of O(NlogN). Some of these have expected time proportional to N. However, only a few algorithms are available for computing the convex hull in high dimensional spaces. For K=3, Preparata and Hong [Pre77] give an O(NlogN) algorithm. Chand and Kapur [Cha70] designed an algorithm based on the 'gift-wrapping' principle which can be used for any value of K. Toussaint [Tou80] mentions the time complexity of this algorithm is bounded below by

$$O\left[N^{\left\lceil \frac{K-1}{2} \right\rceil + 1} \right].$$

Devroye [Dev80], by using results on maximal vectors [Ben78b], showed that there exists a convex hull algorithm for general K that runs in O(N) expected time for certain classes of point distributions.

Since the Chand and Kapur algorithm is the only complete algorithm that has been published for high dimensional data, we choose to use it in our convex hull implementation. Their algorithm proceeds by 'gift-wrapping' the points; that is it works by finding one face of the hull, then finding its edges, and then pivoting this face about each edge until it forms a new face of the hull.

Due to inaccuracies in the published version of this algorithm, we give a corrected version of it here. Let X={Xi} be the set of N points in K dimesions. Its convex hull is found by calling the procedure CONVEX(X,K,N). This returns the faces of the convex hull in the form of a normal vector to each face and the set of data points on each face. The following is the terminology used in defining the procedure CONVEX. Let nj, j=1,...,K be global variables holding the surface normals at each recursion level of the algorithm. Let DIM(E) represent the dimension of set E, i.e. one minus the number of linearly independent points in E, and let |E| represent the number of points in set E. Let (e.v) represent the inner product of vectors e and v. The algorithm requires that N>K.

CONVEX(Y,M,L)

(1) Find a face (an M-1 dimensional flat)
NFACE <-- 0, NFACE is the current number of faces found.
(a) Let E be the set of point(s) of Y with the smallest first coordinate.
The hyperplane H passing through E with normal
$n_M = (1,0,\ldots,0)$
is a support hyperplane of Y.
(b) Let {vi} be the DIM(E) unit vectors spanned by flat E.
Solve the K-1 equations with K variables (some of which may be zero) for unit vector e
(e.vi) = 0 i=1,2,..,DIM(E)
(e.ni) = 0 i=K,K-1,..,M
(c) Find the next point(s) to be added to E by finding those points in Y which maximize

$$-(e.v_j)/(n_M.v_j)$$

where v_j is the unit vector from a point in E to the jth point of Y. Let λ/μ be this maximum value such that $\lambda^2+\mu^2=1$.
Update the normal to the hyperplane passing through E as
$n_M = \lambda n + \mu e$.
(d) If DIM(E) < M-1 go to (b)

(2) Store the faces and compute the edges
NFACE <-- NFACE + 1
FACE(NFACE) <-- E
IF |E| = M
We can compute the M edges of E. Each edge is defined by a subset of M-1 points. Store these edges in EDGES. If an edge is already stored, delete it.
ELSE (|E|>M)
CALL CONVEX(E,M-1,|E|)
Store the returned faces in EDGES. If an edge is already there, then delete it.

(3) Replace E by an edge from EDGES.
If there are none, then return FACES and end.
Let n be the normal to the face containing the edge in E.
Let Yo be a point in this face not in E.

(4) Go to step 1(b) with the additional constraint in step 1(b) on the solution vector e that (e.v)>0, where v is the unit vector from a point in E to Yo.

END

Table 22 gives some results from our implementation of the Chand and Kapur algorithm. We have used Cohen and Hickey's simplex decomposition algorithm [Coh77] to compute the volume of the hull. The table shows information about the convex hull for 100 points generated uniformly in the unit hypercube. We note the rapid increase in computer time, number of faces, and decrease in the volume of the convex hull as dimensionality increases. However, Ripley's unbiased volume estimator [Rip78] for planar data appears to operate quite well for higher dimensions, indicating that his results are probably valid in higher dimensions.

TABLE 22: Run Times to Compute the Convex Hull and its Volume
100 points at random in the unit hypercube
times are CPU seconds on a CYBER 170/750

K	Number of runs	Time/ run	Avg. number of faces	Avg. number points inside	Avg. Volume	Avg. unbiased Volume
2	10	.09	12	88	.87	.99
3	10	.86	56	68	.68	1.00
4	5	33.19	251	43	.46	1.10
5	1	>450	1046	27	*	*

*could not be run due to excessive computation time.

LIST OF REFERENCES

LIST OF REFERENCES

[Ahu81] N. Ahuja, 'Dot pattern processing using Voronoi polygons as neighborhoods', IEEE Trans. Pattern Anal. Machine Intell., Vol. 4, pp. 336-343, 1982.

[Ala76] V.S. Alagar, 'The distribution of the distance between two random points', J. Appl. Probl., Vol. 13, pp. 558-566, 1976.

[Ala81] V.S. Alagar and L.H. Thiel, 'Algorithms for detecting M-dimensional objects in N-dimensional spaces', IEEE Trans. Pattern Anal. Machine Intell., Vol. 3, pp. 245-256, 1981.

[And73] M.R. Anderberg, Cluster Analysis for Applications. New York: Academic Press, 1973.

[Bai78] T.A. Bailey, 'Cluster validity and intrinsic dimensionality', Ph.D. Thesis, Dept. of Comput. Science, Michigan State Univ., 1978.

[Bar64] M.S. Bartlett, 'The spectral analysis of two-dimensional point process', Biometrika, Vol. 51, pp. 299-311, 1964.

[Bar75] M.S. Bartlett, The Statistical Analysis of Spatial Pattern. London: Chapman and Hall, 1975.

[Ben78] J.L. Bentley and J.H. Friedman, 'Fast algorithms for constructing minimal spanning trees in coordinate spaces', IEEE Trans. on Comput., Vol. 27, pp. 97-105, 1978.

[Ben78b] J.L. Bentley et. al., 'On the average number of maxima in a set of vectors and applications', J. Assoc. Comput. Machinary, Vol. 25, pp. 536-543, 1978.

[Bis81] G. Biswas, et. al., 'Evaluation of projection algorithms', IEEE Trans. Pattern Anal. Machine Intell., Vol. 3, pp. 701-708, 1981.

[Bro75] D. Brown, 'A test of randomness of nest spacing', Wildfowl, Vol. 26, pp. 102-103, 1975.

[Bro78] D. Brown and P. Rothery, 'Randomness and local regularity of points in a plane', Biometrika, Vol. 65, pp. 115-122, 1978.

[Cha70] D.R. Chand and S.S. Kapur, 'An algorithm for convex polytopes', J. Assoc. Comput. Machinery, Vol. 17, pp. 78-86, 1970.

[Chr85] G. Chrystal, 'On the problem to construct the minimum circle enclosing n given points in the plane', Proc. Edinburgh Math. Soc., Vol. 3, pp. 30-33, 1885.

[Cla54] P.J. Clark and F.C. Evans, 'Distance to the nearest neighbor as

a measure of spatial pattern in biological populations', Ecology, Vol. 35, pp. 445-453, 1954.

[Coh79] J. Cohen and T. Hickey, 'Two algorithms for determining the volumes of convex polyhedra', J. Assoc. Comput. Machinery, Vol. 26, pp. 401-414, 1979.

[Con71] W.J. Conover, Practical Nonparametric Statistics. New York: Wiley, 1971.

[Con79] W.J. Conover, et. al., 'On a method for detecting clusters of possible uranium deposits', Technometrics, Vol. 21, pp. 277-282, 1979.

[Cox65] D.R. Cox and P.A.W. Lewis, The Statistical Analysis of Series of Events. London: Chapman and Hall, 1965.

[Cox80] D.R. Cox and V. Isham, Point Processes. London: Chapman and Hall, 1980.

[Cox76] T.F. Cox and T. Lewis, 'A conditioned distance ratio method for analyzing spatial patterns', Biometrika, Vol. 63, pp. 483-491, 1976.

[Cro80] G.R. Cross, 'Some approaches to measuring clustering tendency', Technical Report TR-80-03, Dept. of Comput. Science, Michigan State University, 1980.

[Cro82] G.R. Cross and A.K. Jain, 'Measurement of clustering tendency', in Proc. IFAC Symp. Digital Control, New Delhi, India, pp. 24-29, 1982.

[Dev80] L. Devroye, 'A note on finding convex hulls via maximal vectors', Information Processing Letters, Vol. 11, pp. 53-56, 1980.

[Dig76] P.J. Diggle, J. Besag, and J.T. Gleaves, 'Statistical analysis of spatial point processes by means of distance methods', Biometrics, Vol. 32, pp. 659-667, 1976.

[Dig79] P.J. Diggle, 'On parameter estimation and goodness-of-fit testing for spatial point patterns', Biometrics, Vol. 35, pp. 87-101, 1979.

[Dub76] R.C. Dubes and A.K. Jain, 'Clustering techniques: the user's dilemma', Pattern Recognition, Vol. 8, pp. 247-260, 1976.

[Dub79] R.C. Dubes and A.K. Jain, 'Validity studies in clustering methodologies', Pattern Recognition, Vol. 11, pp. 235-254, 1979.

[Dub80] 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 Press, pp. 113-228, 1980.

[Dud73] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[Elz72] D.J. Elzinga and D.W. Hearn, 'The minimum covering sphere problem', Management Science, Vol. 19, pp. 96-104, 1972.

[Eve74] B. Everitt, Cluster Analysis. London: Heinemann Educational Books, 1974.

[Eve79] B. Everitt, 'Unresolved problems in cluster analysis', Biometrics, Vol. 35, pp. 169-181, 1979.

[Fil71] S. Fillenbaum and A. Rapoport, Structures in the Subjective Lexicon. New York: Academic Press, 1971.

[Fis36] R.A. Fisher, 'The use of multiple measurements in taxonomic problems', Ann. Eugenics, Vol. 7, pp. 179-188, 1936.

[Fri79] J.H. Friedman and L.C. Rafsky, 'Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests', Ann. Stat., Vol. 7, pp. 697-717, 1979.

[Fu74] K.S. Fu, Syntactic Methods in Pattern Recognition. New York: Academic Press, 1974.

[Fuk72] K. Fukunaga, Introduction to Statistical Pattern Recognition. New York: Academic Press, 1972.

[Gna77] R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley, 1977.

[Gon78] R.C. Gonzalez and M.G. Thomason, Syntactic Pattern Recognition: An Introduction. Reading: Addison-Wesley, 1978.

[Gre76] U. Grenander, Lectures in Pattern Theory: Pattern Synthesis. New York: Springer-Verlag, 1976.

[Gre78] U. Grenander, Lectures in Pattern Theory: Pattern Analysis. New York: Springer-Verlag, 1978.

[Gri64] P. Grieg-Smith, Quantitative Plant Ecology. London: Buttersworths, 1964.

[Hal73] D.J. Hall et. al., 'Development of new pattern recognition methods', Stanford Research Institute, Final Report AD-772 614, SRI Project 1340, 1973.

[Ham50] J.M. Hammersley, 'The distribution of distances in a hypersphere', Ann. Math. Stat., Vol. 21, pp. 447-452, 1950.

[Har72] F. Harary, Graph Theory. Reading: Addison-Wesley, 1972.

[Har74] E.F. Harding and D.G. Kendall (eds.), Stochastic Geometry. London: Wiley, 1974.

[Har75] J.A. Hartigan, Clustering Algorithms. New York: Wiley, 1975.

[He82] Q. He and R.C. Dubes, 'An experiment in Chinese speaker identification', to appear in Proc. Int. Conf. Chinese Language Comput. Soc., Washington D.C., Sept. 22-23, 1982.

[Hin79] W.G.S. Hines and R.J.O. Hines, 'The Eberhardt statistic and the detection of nonrandomness of spatial point distributions', Biometrika, Vol. 66, pp. 73-79, 1979.

[Hop54] B. Hopkins, 'A new method for determining the type of distribution of plant individuals', Ann. Botany, Vol. 18, pp. 213-227, 1954.

[Ish81] V. Isham, 'An introduction to spatial point processes and markov random fields', Inter. Stat. Rev., Vol. 49, pp. 21-43, 1981.

[Kel76] F.P. Kelly and B.D. Ripley, 'A note on Strauss's model for clustering', Biometrika, Vol. 63, pp. 357-360, 1976.

[Kin71] K.J. Kinderman and R. Ramage, 'Computer generation of normal random variables', J. Amer. Stat. Assoc., Vol. 71, pp. 893-896, 1976.

[Knu81] D.E. Knuth, The Art of Computer Programming: Seminumerical Algorithms, Vol. 2 (2nd ed.). New York: Addison-Wesley, 1981.

[Law65] C.L. Lawson, 'The smallest covering cone or sphere', SIAM Review, Vol. 7, pp. 415-417, 1965.

[Lie77] A.M. Liebetrau and E.D. Rothman, 'A classification of spatial distributions based upon several cell sizes', Geo. Anal., Vol. 9, pp. 14-28, 1977.

[Lie77b] A.M. Liebetrau, 'Tests for randomness in two dimensions', Commun. Stat. Theory and Methods, Vol. A6, pp. 1367-1383, 1977.

[Lie78] A.M. Liebetrau, 'The weak convergence of a class of estimators of the variance function of a two-dimensional Poisson process', J. Appl. Probl., Vol. 15, pp. 433-439, 1978.

[Lin73] R.F. Ling, 'Probability theory of cluster analysis', J. Amer. Stat. Assoc., Vol. 68, pp. 159-164, 1973.

[Lin75] R.F. Ling, 'An exact probability distribution on the connectivity of random graphs', J. Math. Psychol., Vol. 12, pp. 90-98, 1975.

[Lor54] R.D. Lord, 'The distribution of distances in a hypersphere', Ann. Math. Stat., Vol. 25, pp. 794-798, 1954.

[Mat60] B. Matern, 'Spatial variation', Meddelanden fran Statens Skogsforningsinstitut, Vol. 49, 1960.

[Mea74] R. Mead, 'A test for spatial pattern at several scales using data from a grid of contiguous quadrats', Biometrics, Vol. 30, pp. 295-307, 1974.

[Moo74] D.J.H. Moore and D.J. Parker, 'Analysis of global pattern features', Pattern Recognition, Vol. 6, pp. 149-164, 1974.

[Mul78] D.E. Muller and F.P. Preparata, 'Finding the intersection of two convex polyhedra', Theo. Comput. Sci., Vol. 7, pp. 217-236, 1978.

[Nau65] J.I. Naus, 'Clustering of random points in two dimensions' Biometrika, Vol. 52, pp. 263-267, 1965.

[Nau66] J.I. Naus, 'A power comparison of two tests on non-random clustering', Technometrics, Vol. 8, pp. 493-517, 1966.

[Ney72] J. Neyman and E.L. Scott, 'Process of clustering and applications', in Stochastic Point Process, P.A.W. Lewis ed. New York: Wiley, pp. 646-681, 1972.

[Pan81] E. Panayirci and R.C. Dubes, 'A new statistic for assessing gross structure of multidimensional patterns', Technical Report TR-81-04, Dept. of Comput. Science, Michigan State Univ., 1981.

[Pav77] T. Pavlidis, Structural Pattern Recognition. New York: Springer-Verlag, 1977.

[Pie77] E.C. Pielou, Mathematical Ecology. New York: Wiley, 1977.

[Pre77] F.P. Preparata and S.J. Hung, 'Convex hulls of finite sets of points in two and three dimensions', Commun. Assoc. Comput. Machinery, Vol. 20, pp. 87-93, 1977.

[Pri57] R.C. Prim, 'Shortest connection networks and some generalizations', Bell System Tech. J., Vol. 36, pp. 1389-1401, 1957.

[Rad57] H. Rademacher and O. Toeplitz, The Enjoyment of Mathematics. Princeton: Princeton Univ. Press, 1957.

[Rao73] C.R. Rao, Linear Statistical Inference and Its Applications, (2nd ed.). New York: Wiley, 1973.

[Rip77] B.D. Ripley, 'Modelling spatial patterns', J. Roy. Stat. Soc., Vol. B-39, pp. 172-212, 1977.

[Rip77b] B.D. Ripley and J.P. Rasson, 'Finding the edge of a Poisson forest', J. Appl. Probl., Vol. 14, pp. 483-491, 1977.

[Rip78] B.D. Ripley and B.W. Silverman, 'Quick tests for spatial interaction', Biometrika, Vol. 65, no. 3, pp. 641-642, 1978.

[Rip79] B.D. Ripley, 'Tests of "randomness" for spatial patterns', J.

Roy. Stat. Soc., Vol. 41, pp. 368-374, 1979.

[Rip81] B.D. Ripley, Spatial Statistics. New York: Wiley, 1981.

[Rob68] F.D.K. Roberts, 'Random minimal trees', Biometrika, Vol. 55, pp. 255-258, 1968.

[Rog74] A. Rogers, Statistical Analysis of Spatial Dispersion. London: Pion, 1974.

[Sau77] R. Saunders and G.M. Funk, 'Poisson limits for a clustering model of Strauss', J. Appl. Probl., Vol. 14, pp. 776-784, 1977.

[Sil78] B.W. Silverman and T.C. Brown, 'Short distances, flat triangles and Poisson limits', J. Appl. Probl., Vol. 15, pp. 815-825, 1978.

[Smi81] S.P. Smith and R.C. Dubes, 'Clustering tendency using small interpoint distances', PRIP Lab Memo, Dept. of Comput. Science, Michigan State University, 1981.

[Ste81] M.A. Stephens, 'Further percentage points of Greenwood's statistic', J. Roy. Stat. Soc., Vol. A-144, pp. 364-366, 1981.

[Str75] D.J. Strauss, 'A model for clustering', Biometrika, Vol. 62, pp. 467-475, 1975.

[Sup81] K.J. Supowit, 'Topics in computational geometry', Ph.D. Thesis, Depart. of Comput. Science, Univ. of Illinois, Urbana-Champaign, 1981.

[Syl57] J.J. Sylvester, 'A question in the geometry of situation', Quarterly J. Pure Applied Math., Vol. 1, p. 79, 1857.

[Syl60] J.J. Sylvester, 'On Poncelet's approximate linear valuation of surd forms', Phil. Mag., Vol. 20, pp. 703-722, 1860.

[Tou80] G.T. Toussaint, 'Pattern recognition and geometrical complexity', in Proc. 5th Int. Conf. Pattern Recognition, Miami Beach, pp. 1324-1347, 1980.

[Tuk77] J.W. Tukey, Exploratory Data Analysis. Reading: Addison-Wesley, 1977.

[Wal40] A. Wald and J. Wolfowitz, 'On a test whether two samples are from the same population', Ann. Math. Stat., Vol. 11, pp. 147-162, 1940.

[Wal74] S.R. Wallenstein and J.I. Naus, 'Probabilities for size of largest clusters and smallest intervals', J. Amer. Stat. Assoc., Vol. 69, pp. 690-697, 1974.

[Zah71] C.T. Zahn, 'Graph theoretical methods for detecting and describing gestalt clusters', IEEE Trans. Comput., Vol. 20,

pp.68-86, 1971.