A MULTIMODAL DISTRIBUTION BASED CLUSTERING ALGORITHM

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
VASUDEVA ANANDA RAO
1971

This is to certify that the

thesis entitled

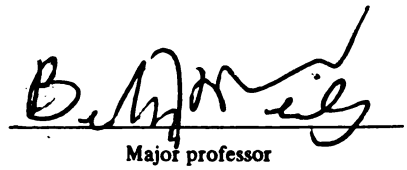
A MULTIMODAL DISTRIBUTION
BASED CLUSTERING ALGORITHM

presented by

VASUDEVA ANANDA RAO

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Computer Science



Major professor

Date December 8, 1971

ABSTRACT

A MULTIMODAL DISTRIBUTION BASED CLUSTERING ALGORITHM

By

Vasudeva Ananda Rao

A new mathematical model is proposed for the clustering problem encountered in data analysis and pattern recognition. The set of multivariate observations on the objects to be grouped is considered as having been generated by an unknown multivariate continuous probability distribution having one or more distinct modes. This distribution is not treated as a mixture of several source pdfs. The clustering problem is identified as that of (1) the estimation of the number and location of the modes of such a distribution and (2) the selection of a suitable distance measure to group the observations based on a 'similarity measure' defined as the distance of each observation from the modes. A practical method is proposed for estimating the number and location of the modes and for detecting the structure in the data in the form of clusters. This method does not require that the number of clusters desired be specified in advance.

For pattern classification in the absence of training patterns from each class, the clusters so detected are treated as the sets of training patterns for "learning" purposes. An algorithm is presented for classification of patterns from unknown

class using a minimum distance-to-mode decision rule. Apart from the Euclidean and Mahalanobis generalized distance measures, two other intuitively apealing distance measures are also discussed. Excellent numerical results have been obtained using test data.

A MULTIMODAL DISTRIBUTION BASED CLUSTERING ALGORITHM

By

Vasudeva Ananda Rao

A THESIS

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

DOCTOR OF PHILOSOPHY

Department of Computer Science

1971

TO

MY MOTHER

ACKNOWLEDGEMENTS

The people to whom one is indebted in acquiring an education are indeed numerous. It is hoped that these people will accept my sincere thanks.

I am highly grateful to Dr. B. Weinberg for his guidance and encouragement. He suggested the problem as a potentially fruitful area for research and devoted a generous part of his time for discussions leading to this thesis.

For their guidance and interest I wish to thank the other members of my doctoral committee: Dr. R.C. Dubes, Dr. C.V. Page, Dr. M. Rahimi of the Department of Computer Science and Dr. J.S. Frame of the Department of Mathematics.

For the financial support which made this work possible I wish to express my gratitude to: the United Nations Educational, Scientific and Cultural Organization, Paris; the Government of India, New Delhi; the Department of Technical Education, Government of Tamilnadu, Madras and the Department of Computer Science, Michigan State University.

Special thanks are due to my mother, wife and children for their patience and understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Motivation

1.1 Introduction:

In many diverse disciplines the scientist collects data in the form of p measurements or observations on each of N individuals or objects. He is interested in grouping these individuals into subgroups in such a manner that members of a subgroup are highly similar or associated and relatively unassociated with members not belonging to the subgroup. For example: in electrical engineering, one is interested in the detection of signals of unknown characteristics that recur frequently in a background of random noise; in medicine, to group electrocardiograms of EEGs into subgroups [D-1]; in psychology and sociology, to group people into types that may relate to treatment categories or behaviour categories; in information retrieval, to find classes of descriptors for articles and papers [G-1]; in numerical taxonomy, to group species of living organisms into hierarchic trees, etc. Such grouping is a very valuable tool in data analysis.

1.2 Pattern Recognition and Cluster Analysis:

One of the aims of pattern recognition study is to find meaningful descriptions to adequately characterize a set of data. The principal objective of cluster analysis as applied to pattern recognition is to gain more information about the structure of a

data set than is possible by more conventional methods such as factor analysis or principal components analysis [N-1]. In the statistical sense the pattern classification problem (2 class case) may be defined as a discriminatory problem as follows:

A vector random variable X of observed values x is distributed over some p-dimensional space according to distribution F or G. The problem is to determine which of the two distributions x came from.

A statistical approach to the above problem can be considered to fall into one of three subproblems:

(i) F and G are completely known.

(ii) F and G are known with the exception of some finite set of parameters.

(iii) F and G are unknown except possibly for assumptions about the existence of densities, continuity, symmetry, etc.

Subproblem (i) has been completely solved by the Neyman-Pearson lemma and results in a likelihood ratio (LR) test. This LR test yields optimum results in the sense of minimum probability of misclassifications.

In the formulation of the approach to subproblem (ii) one assumes the availability of training samples from each of the two distributions. It is also assumed that the forms of F and G are given but one or more parameters are unknown. The problem then reduces to one of parameter estimation or one can also employ learning strategies. Once these parameters have been estimated using the training samples the LR test is applied as though F and G were completely known.

The "parametric" approaches to subproblems (i) and (ii) appear reasonable provided the assumptions made are justifiable in practice in that the assumed parametric forms are good representations of the data. But when little a priori knowledge exists about the underlying probability distribution associated with each pattern class, these parametric approaches to the classification problem become questionable in the sense that bad results may be obtained. This conclusion has led researchers to require less stringent assumptions about the form of the data structure and this, in turn resulted in the emergence of a variety of nonparametric pattern recognition procedures as approaches to subproblem (iii). One such approach is to treat the pdf of the observations as a mixture of several unknown source pdfs and then try to identify these pdfs. Such an approach has been considered among others by Teicher [T-1, T-2, T-3], Yakowitz [Y-1, Y-2], Yakowitz and Spragins [Y-3], and Stanat [S-5]. Application and development of this technique under the name of "unsupervised learning" or "learning without a teacher" has been mainly done by Fralick [F-3], Spragins [S-4], Patrick [P-3], Hilborn and Lainiotis [H-1], Patrick and Costello [P-4] and Patrick and Hancock [P-2]. However it is known that the class of mixture distributions which have a unique solution for the parameters of the individual distributions constituting the mixture, is limited and whether or not it admits a unique solution depends on the identifiability of the mixture distribution [F-6, Y-1, Y-2, Y-3].

Cluster analysis is another type of approach which has been pursued for a solution to subproblem (iii). It is a non-parametric

technique to determine a type of structure describing a set of empirical data. A second way that cluster detection is applicable to pattern recognition is by providing an answer to the question whether or not a given set of features constitutes a good feature space in which to discriminate a given set of pattern classes [Z-1].

1.3 Literature Survey:

Clustering techniques have been used as long ago as 1939 by Tryon [T-4]. At present there are a number of proposed clustering procedures available. Widely used in numerical taxonomy [S-3] are agglomerative and divisive hierarchical clustering schemes. Here a small and fixed set of patterns is given and a matrix is computed whose (i,j)th entry is the association or similarity between the i-th and j-th patterns. The agglomerative procedures [L-2] generally link together the most similar patterns. Then the similarities between the groups of grouped patterns and the remaining groups (or patterns) are recomputed using the minimum, maximum, or mean similarity between the two groups. The procedure continues in this manner linking together the most similar patterns or groups. Michener and Sokal [M-4], Ward [W-1], and McQuitty [M-2] were early users of this scheme. In a recent paper Rohlf [R-2] describes sequential agglomerative hierarchical clustering schemes in particular detail and proposes several new methods.

The divisive procedures [L-2] begin with all patterns in the same group and splitting the group into the two most dissimilar groups. Edwards and Cavalli-Sforza [E-1] suggest dividing the points or patterns into two groups such that the sum of squared

distance between the sets is a maximum. They define this as a cluster and suggest an algorithm to find the two sets with the desired property. Because the total sum of squared distance is a constant for a given sample of points, maximizing the between-set sums of squared distance is equivalent to minimizing the within-set sum of squared distance. Their algorithm is to examine all $2^{N-1} - 1$ partitions of the N points and select the one which gives the minimum within-set sum of squared distance. Lance and Williams [L-1] suggest successively splitting the groups in a way which is expected to reduce the variance the greatest for the split groups. Mattson and Dammann [M-1] suggest successively splitting each group by thresholding the dominant eigenvector of the covariance matrix of that group. Wirth et al [W-4] suggest thresholding the association or similarity matrix and defining the components of the resulting graph as clusters. Thresholding is done successively from strict thresholds to more liberal thresholds. Jardine and Sibson [J-1] outline a theoretical framework within which the properties of classificatory systems, which operate on data in the form of a dissimilarity coefficient on a set of objects, may be discussed. Hierarchical clustering schemes are also discussed by Johnson [J-2].

The three superficially different hierarchical clustering schemes of Sokal and Michener [M-4], Edwards and Cavalli-Sforza [E-1], and Williams and Lambert [W-3] have been compared by Gower [G-2] and suggestions made for their improvement.

Most popular among the non-hierarchical clustering schemes have been those iterative schemes beginning with an arbitrary set

of all inclusive and mutually exclusive clusters and successively improving the set of clusters by transferring patterns from one cluster to another until no further improvement is available. Such methods use as an evaluation index, what has been called the "C Criterion" by Switzer [S-6]. As a means of evaluating any given partition of the sample, the within cluster distance or scatter [W-2] or between cluster distance or the ratio of total scatter to within cluster scatter is used [F-4]. To circumvent the computational time consuming difficulty involved in examining all possible partitions of the sample [F-2], Friedman and Rubin [F-4], use what they call a "hill climbing" algorithm. In principle their procedure attempts to examine only those partitions of the data for which the ratio of total scatter to within-group scatter is high. The logic of this technique can also be found in [F-5].

A technique for clustering which is very popular is the ISODATA (Iterative Self-Organizing Data Analysis Technique (A)) of Ball and Hall [B-1]. This technique clusters all of the data into distinct and independent groups. A computed mean or average response pattern is used to represent a group of patterns, and the iterative process creates new average response patterns to improve the accuracy of trial or existing average response patterns. The process also combines average response patterns that are so similar that their being separate fails to provide a significant amount of information about the structure of the response patterns. Each response pattern is put into that group for which the squared distance between it and the average response patterns (group mean)

is the least. ISODATA implements an intuitively appealing mathematical idea for clustering patterns. It is not itself a standard statistical procedure, such as analysis of covariance or factor analysis. The order of computations and the setting of the various thresholds in the algorithm were motivated by heuristic reasoning and the performance of the program with standard data sets [D-2].

Jones and Jackson [J-3] suggest an iterative technique where clusters are found one at a time. An initial pattern is picked to be the first pattern in the cluster. Patterns are successively transferred into and out of the cluster in a way which increases the within-cluster similarities and decreases the in-cluster to out-cluster similarities.

Graph-theoretical procedures have also been applied for clustering. Bonner [B-3] starts out by thresholding the association or similarity matrix and defining as "core clusters" the maximal complete subgraphs (cliques) of the resulting graph. Then the smaller core clusters are merged into large core clusters and largely overlapping core clusters are merged. Gotlieb and Kumar [G-1] discuss graph theoretical clustering methods useful in information retrieval applications. Zahn [Z-1] describes graph theoretical algorithms based on the minimal spanning tree of a graph and which are capable of detecting several kinds of cluster structure in arbitrary point sets. The concept of the minimal spanning tree of a graph (MST) for single linkage cluster analysis (SLCA) is also discussed by Gower and Ross [G-3]. They show that all the information required for the SLCA of a set of points is contained in the MST of the graph of these points. Augustson and

Minker [A-2] also analyze some graph theoretical clustering techniques as applied to information retrieval.

The statistical technique of Principal Components analysis used for clustering of multivariate data involves treating the number of observations (N) in p-dimensions as N points in an p-dimensional metric space and projecting these points onto a space of smaller dimensions with minimum loss of statistical information; that is, the inherent structure in the data is approximately preserved under the mapping. The primary interest is in mapping onto two or three dimensions since the resultant data configuration can be easily evaluated by human observations in three or less dimensions. Rao [R-1] discusses the application of this technique to clustering. Nunnally [N-3] describes the application of factor analysis procedures in clustering. Dubes [D-2] has an interesting report on cluster analysis and decision making with a correlation matrix wherein he describes minimum-average-distance clustering and mean-squared clustering with particular reference to Gaussian distributions.

A non-linear mapping technique useful in clustering multivariate data is discussed by Sammon [S-1]. The idea behind Sammon's method is similar to that of Kruskal [K-2] who discusses the technique as applied to non-metric hypothesis. An interesting new mathematical formulation of the clustering problem has been provided by Ruspini [R-3].

1.4 Contribution of the Thesis:

From the literature survey of the last section it is clear that, basically, cluster analysis techniques call for the identification of those regions of the observation space where the patterns

are most heavily concentrated, thereby establishing a structure in the data in the form of clusters. In this thesis, it is believed that, to detect such a structure, it is not absolutely necessary (1) to treat the underlying pdf of the observations as a mixture of several component pdfs and then (2) try to identify the individual source pdfs. It is sufficient to consider that the underlying pdf, is, in general, governed by a multivariate continuous distribution with one or more distinct modes. Since the mode of a pdf is that outcome which is likely to occur most often, the estimation of the number and location of the modes provides a clue to the structure of the data. Accordingly the method for clustering proposed in this thesis is based on the premise that such an estimation of the number and location of the modes of a multimodal multivariate pdf underlying a set of empirical data is a key to a solution to the clustering problem. Thereafter, the selection and application of a suitable distance measure determines the identity and membership of the clusters in a straightforward manner.

This intuitively appealing idea, not found in the literature, is presented and offered as a practical solution for the clustering problem. The performance of this method on data sets, both real and artificial, is found to strengthen the belief in such an approach. Specifically the contributions of this thesis are:

(i) A new mathematical model is proposed for the clustering problem. The set of multivariate observations, on the objects to be clustered, is assumed to have been generated by a multivariate continuous probability distribution whose density function has one or more distinct modes. This pdf is not treated as a mixture of

several source pdfs as is sometimes done in the classification problem [Y-1] of pattern recognition.

(ii) In his discussion on nonparametric decisions based on distance to modes, Nilsson [N-2] assumes the existence of a method to find good estimates for the modes of a probability density function, given the set of training samples. This thesis provides a practical method, guided by statistical theory, for the estimation of such modes. Once the modes have been estimated from a set of training samples, a starting point for the abstraction phase of the pattern recognition problem is available. The abstraction phase of the problem can then be completed using, for example, a piecewise linear machine, implementing a minimum distance classifier.

1.5 Organization of the Thesis:

Chapter II presents the proposed model for clustering and discusses the relevant mathematical preliminaries. Chapter III describes an algorithm for the estimation of the number and location of the modes of a multivariate, multimodal pdf. Chapter IV extends the mode seeking algorithm to clustering and summarizes the results obtained on a few real and simulated sets of data, using both supervised and unsupervised learning techniques. Conclusions and suggestions for future work are included in Chapter V.

CHAPTER II

Mathematical Preliminaries

2.1 Introduction:

In this chapter the theoretical preliminaries required for a solution to the clustering problem are discussed. The problem is stated in the framework of a simple mathematical model. The basic assumptions made are:

(i) that there exists a probability distribution which generates the multivariate observations to be clustered: and

(ii) that the probability density function (pdf) of this distribution is of the continuous type characterized by one or more distinct modes.

The word mode, as used here, denotes the location of a local maximum in the pdf. Assumption (ii) implies that the pdf need not necessarily be symmetric.

(iii) the number of observations is large compared with the number of dimensions of each observation.

(iv) the observations are independent.

2.2 A Mathematical Model for the Clustering Problem:

Let $x_1, x_2, \ldots, x_N$ denote the p-dimensional observations (features, measurements) of N objects. Assume these observations are values of a random variable, $X = (X_1, X_2, \ldots, X_p)$, having a cumulative distribution function (cdf) $F(x)$, and the corresponding

probability density function (pdf) $f(x)$. Let $f(x)$ be continuous and characterized by G distinct modes, that is,

$$f(x) = f(x_1,x_2,\ldots,x_p;\ M_1,M_2,\ldots,M_G)$$

where M_i, $i = 1,2,\ldots,G$, denotes the i-th distinct p-dimensional mode vector of $f(x)$.

Then the clustering problem is:

(i) Estimate G, the number of modes of $f(x)$;

(ii) Estimate the column vectors M_i, $i = 1,2,\ldots,G$; and

(iii) Define a distance measure $d(x_j,M_i)$, $j = 1,2,\ldots,N$; $i = 1,2,\ldots,G$, which partitions the set of observations into G groups, $\pi_1,\pi_2,\ldots,\pi_G$ such that

$x_j \in \pi_i$ if and only if

$$d(x_j,M_i) < d(x_j,M_k);\ i \neq k; \quad k = 1,2,\ldots,G.$$

It is immediately evident from the statement of the problem that what is mainly needed for the clustering procedure is a good and robust method to estimate the number and locations of the modes of a multivariate, multimodal pdf. It should be noted that the clustering procedure dictated by the above model does not require that the number of clusters or groups be specified a priori. Almost all existing techniques such as Friedman and Rubin's [F-4] require that the number of clusters to be formed be specified in advance. ISODATA [B-1] also calls for the value of a parameter to be specified to indicate the number of clusters desired by the user. The absence of the need to specify in advance, the number of modes desired, is claimed to be a distinct advantage of this model. It is also emphasized that the model does not treat the

pdf generating the observations as a mixture of several source pdfs.

2.3 Estimation of Modes and Mathematical Results:

The problem of directly estimating the mode of a univariate unimodal continuous pdf, f(x), has been considered by Grenander [G-4], Venter [V-1]. Other approaches to the estimation of modes appear as an extension of the related problem of estimating f(x) from a set of independent samples. For multiple mode determination as required by the model proposed here, density estimation in some manner is a necessity and this will be of major concern in the remaining sections of this chapter.

In order to estimate the density from a set of independent, identically distributed observations, various approaches have been proposed in the statistical literature . A potential function type of approach has been suggested by Aizerman and Braverman [A-1]. Kashyap and Blaydon [K-1] suggest a stochastic approximation type of approach. Other methods have been proposed by Chernoff [C-1], Murthy [M-6], Parzen [P-1]. In this thesis the univariate version of the estimator for the pdf proposed by Loftsgaarden and Quesenbury [L-3] is considered for mode estimation. The multivariate case is handled by a method motivated by Mattson and Dammann [M-1]. The p-dimensional observations are first projected on to the principal eigenvectors of the sample covariance matrix of the observations. For the univariate set of data so obtained, the modes are estimated, and then reprojected back into the original space.

Because the probability density function is, in general, assumed to be multimodal with distinct modes, an extension of the

method for mode estimation of unimodal distributions is considered. This is an adaptation of the method originally suggested by Fu and Henrichon [F-7].

Let $x_1, x_2, \ldots, x_N$ be independent observations on a p-dimensional random variable $X = (X_1, X_2, \ldots, X_p)$ with absolutely continuous cumulative distribution function (cdf), $F(x_1, x_2, \ldots, x_p)$, and the corresponding pdf, $f(x) = f(x_1, x_2, \ldots, x_p)$;

Let $z = (z_1, z_2, \ldots, z_p)$ be a point at which $f(x)$ is positive and continuous;

Let $d_z(r_N)$ denote the Euclidean distance from the point z to the r_N-th closest observation $\in \{x_i\}_{i=1}^N$ where

$\{r_N\}$ is a non-decreasing sequence of positive integers satisfying

$$\lim_{N\to\infty} r_N \to \infty$$

$$\lim_{N\to\infty} \frac{r_N}{N} = 0 \qquad \text{and}$$

$$\lim_{N\to\infty} N^{-\delta} r_N = \infty \quad \text{for some} \quad \delta > 0 \; .^{*}$$

Theorem 2.1: (Loftsgaarden and Quesenberry)

Under the above hypothesis, the estimate of $f(x)$ at the point z, given by the estimator

$$\hat{f}_N(z) = \left\{\frac{r_N - 1}{N}\right\} \cdot \left\{ p \cdot \Gamma\left(\frac{p}{2}\right) \Big/ 2 \, [d_z(r_N)]^p \, \pi^{\frac{p}{2}} \right\}$$

is consistent.

Proof:

Proof of the above theorem can be found in [L-3].

---

* Loftsgaarden and Quesenberry suggest the use of an integer closest to $\sqrt{N}$ as a value of r_N.

The following corollary to the above theorem is obtained in the univariate case $(p = 1)$.

A consistent estimate of $f(x)$ at z is given by

$$f_N(z) = \frac{(r_N-1)/N}{2\, d_z(r_N)} \quad . \tag{2.2}$$

Lemma 2.1: (Owen) [O-1]

Let $f(x)$ be a uniformly continuous probability density function and $\{f_N(x)\}$ be a sequence of estimates of $f(x)$.

Let M_1 be the value of x, assumed to be unique, where $\max_{x\in I_1} f(x)$ occurs and I_1 is an interval over which $f(x)$ is continuous.

Further let M_{1_N} be the value of x where $\max_{x\in I_1} f_N(x)$ occurs.

If $f_N(x)$ is a consistent estimate of $f(x)$, then M_{1_N} is a consistent estimate of M_1.

Proof of this lemma has been given by Owen.

Extension to Multiple Mode Estimation:

As before, let $x_1, x_2, \ldots, x_N$ be independent observations on a univariate random variable and $d_z(r_N)$ be defined as before. Let $f(x)$ be uniformly continuous and positive over an interval I_1 and let there exist a unique mode $M_1 \in I_1$.

Let a new estimate of the mode M_1 be formed as follows:

$$M'_{1_N} = x_{(k)}$$

where $x_{(k)}$ is the k-th order statistic, i.e., $x_{(k)}$ is the k-th smallest observation from the set of N observations and

$$k = \text{Arg}\left(\min_i \{d_{x_{(i)}}(r_N) \mid x_{(i)} \in I_1\}\right);$$

that is, the estimate M_{1_N}' of the mode is that order statistic contained in the interval I_1 which yields the minimum value of $d_z(r_N)$ as computed by the procedure of Loftsgaarden and Quesenberry.

Further let $f_N(x)$, for $x \in I_1$, be constructed in a stepwise fashion according to

$$f_N(x) = \frac{(r_N - 1)/N}{2\, d_{x_{(i)}}(r_N)}, \quad x_{(i)} \leq x \leq x_{(i+1)}.$$

Lemma 2.2: (Moore and Henrichon)

The estimate $f_N(x)$ converges uniformly, in probability, to $f(x)$, $x \in I_1$.

Proof: see [M-5].

From lemmas (1) and (2),

Theorem 2.2:

M_{1_N}' is a consistent estimate of the mode M_1; i.e.,

$$M_{1_N}' \to M_1 .$$

Proof:

By Lemma 2.2 we know that the estimate $f_N(x)$ constructed in the manner indicated converges in probability to $f(x)$, $x \in I_1$. Applying Lemma 2.1 and the fact that $\max_{x \in I_1} f_N(x)$ occurs at the value of $x_{(i)}$ where $d_{x_{(i)}}(r_N)$ is minimum, the theorem follows.

Q.E.D.

For estimation of the modes in the case of a multimodal, univariate pdf, consider the set of ℓ relative maxima (assumed unique), $\{m_i\}_{i=1}^{\ell}$ located at modes $\{M_i\}_{i=1}^{\ell}$ respectively, and such that $M_i \in I_i = (a_{2i-1}, a_{2i})$, $i = 1,2,\ldots,\ell$ where $a_1 < a_2 < \ldots < a_{2\ell}$. Further let the pdf $f(x)$ be uniformly continuous and positive over the intervals $\{I_i\}_{i=1}^{\ell}$, and the estimates M_{k_N} of M_k be determined by

$$M_{k_N} = x_{(j)}$$

where $j = \text{Arg}\left[\underset{i}{\text{Min}}\left\{h_{x_{(i)}}(r_N) \,\middle|\, x_{(i)} \in I_k\right\}\right]$.

Then, as an immediate consequence of Theorem 2.2,

Corollary:

The set of estimates $\{M_{i_N}\}_{i=1}^{\ell}$ converges in probability to the set $\{M_i\}_{i=1}^{\ell}$.

2.4 Chapter Summary:

In this chapter a new model for cluster analysis was proposed. The problem of cluster detection was identified with that of estimating the number of modes and the modes themselves of a multimodal pdf with distinct modes. The remainder of the chapter related to the theoretical aspects of mode estimation.

CHAPTER III

A Mode Seeking Algorithm

3.1 Introduction:

The goal of this chapter is to provide a computational procedure to estimate the number, G, and the p-dimensional vectors M_j, $j = 1,2,\ldots,G$, of the modes of a multimodal, multivariate pdf, $f(x)$, of Section 2.2. The mathematical results of the last chapter were directed towards the estimation of the modes of a univariate pdf because one of the problems encountered in a direct application of the mode seeking techniques to a multivariate problem consists of deciding how to store the boundaries which separate the observation space into regions containing only one mode. To appreciate this problem one has only to note that in the one dimensional case these boundaries are simply points on the real line. In two dimensions the corresponding boundaries are curvesand it is not easy to store arbitrary curves in a computer. The problem is even more difficult in a higher dimensional space. One method of circumventing this difficulty is to project the multivariate observations on to the principal eigenvectors of the sample covariance matrix [M-1] to get a univariate set of data; estimate the modes in this one dimensional space and finally transform back to the original p-dimensional space.

The direction of the eigenvector associated with the largest eigenvalue of the sample covariance matrix of the observations is the direction of maximum dispersion. This is the reason for projecting the observations on to the eigenvectors rather than projecting onto the coordinate axes of the p-dimensional observation space. If separation between the data does indeed exist, it should be more easily detected along the eigenvectors [H-2].

3.2 Mode Estimation Procedure for the Multivariate Case:

The proposed mode estimation procedure can be summarized as follows:

Step (i) Compute the principal eigenvectors associated with the sample covariance matrix determined from a set of independent observations.

Step (ii) Project the multivariate observations onto the first principal eigenvector to obtain a univariate data set.

Step (iii) Apply the procedure of Section 3.4 and estimate the number and locations of the local minima for this univariate set and the modes along this eigenvector.

If only one mode is found go to step (iv); else, go to step (v).

Step (iv) Using the next principal eigenvector repeat the procedure from step (ii). If all the eigenvectors have been processed and only one mode is found along each eigenvector, compute the location of the mode and transform back to the original space. Stop.

Step (v) Partition the observation space into regions by hyperplanes perpendicular to the eigenvector and passing

through the points of minima. Classify the observations into subsets corresponding to these regions.

Step (vi) For each new region so obtained repeat the procedure from step (i).

Figure 1 depicts the details of this procedure in flow chart form. An illustration of the procedure applied to a two dimensional data set is given in Figure 2.

3.3 Mathematical Details of the Algorithm:

Let $x_1^*, x_2^*, \ldots, x_N^*$ denote the N, p-dimensional column vectors corresponding to the p feature measurements on each object; that is,

$$x_i^* = \begin{bmatrix} x_{i1}^* \\ x_{i2}^* \\ \vdots \\ x_{ip}^* \end{bmatrix} \quad ; \quad i = 1,2,\ldots,N \; . \tag{3.3.1}$$

The asterisk indicates that these are raw measurements.

The j-th feature average, m_j is defined as the sample mean for the j-th feature:

$$m_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}^* \quad ; \quad j = 1,2,\ldots,p \tag{3.3.2}$$

and the vector of feature averages is

$$m = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{bmatrix} \; . \tag{3.3.3}$$

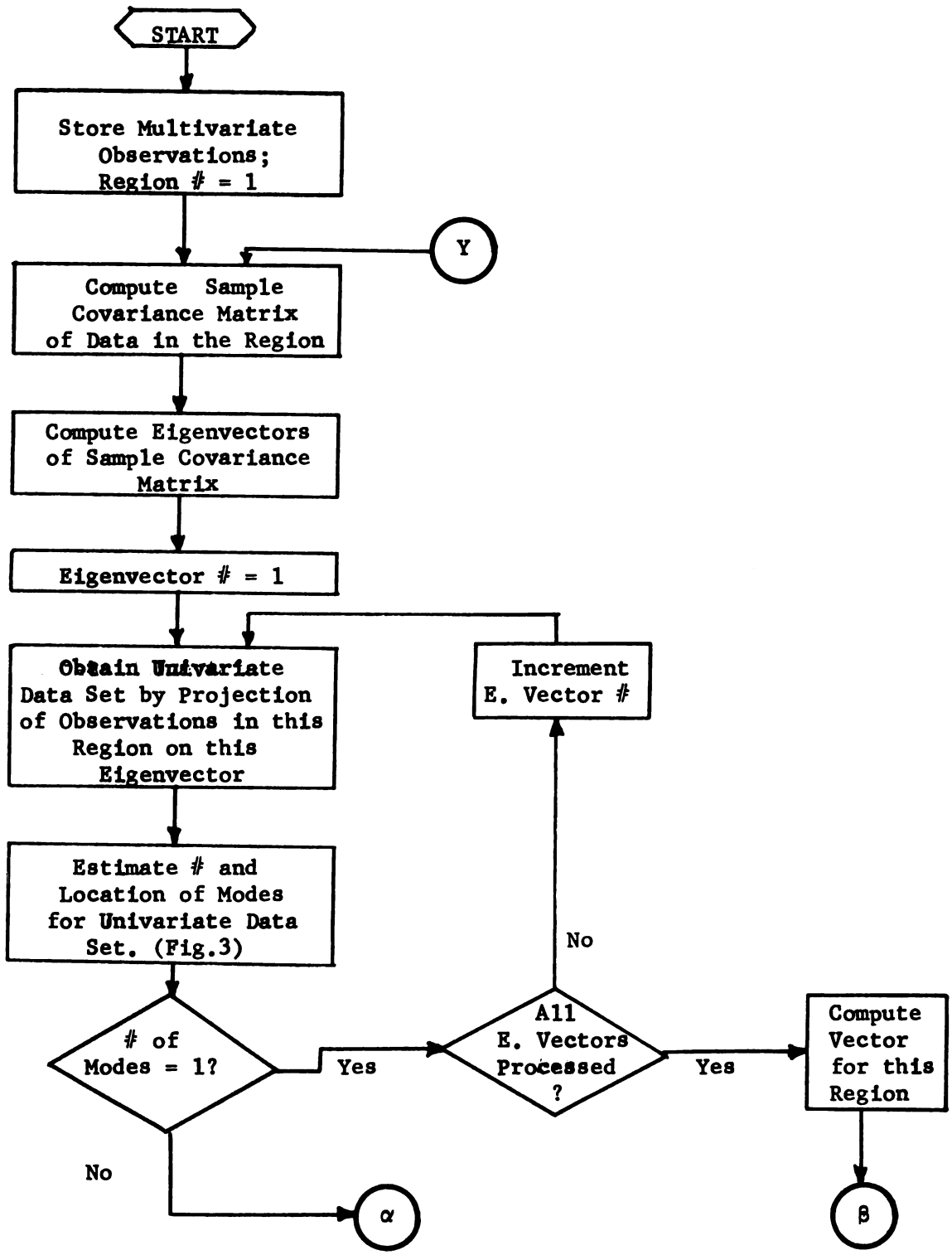


Figure 1. Flow chart for multivariate mode estimation

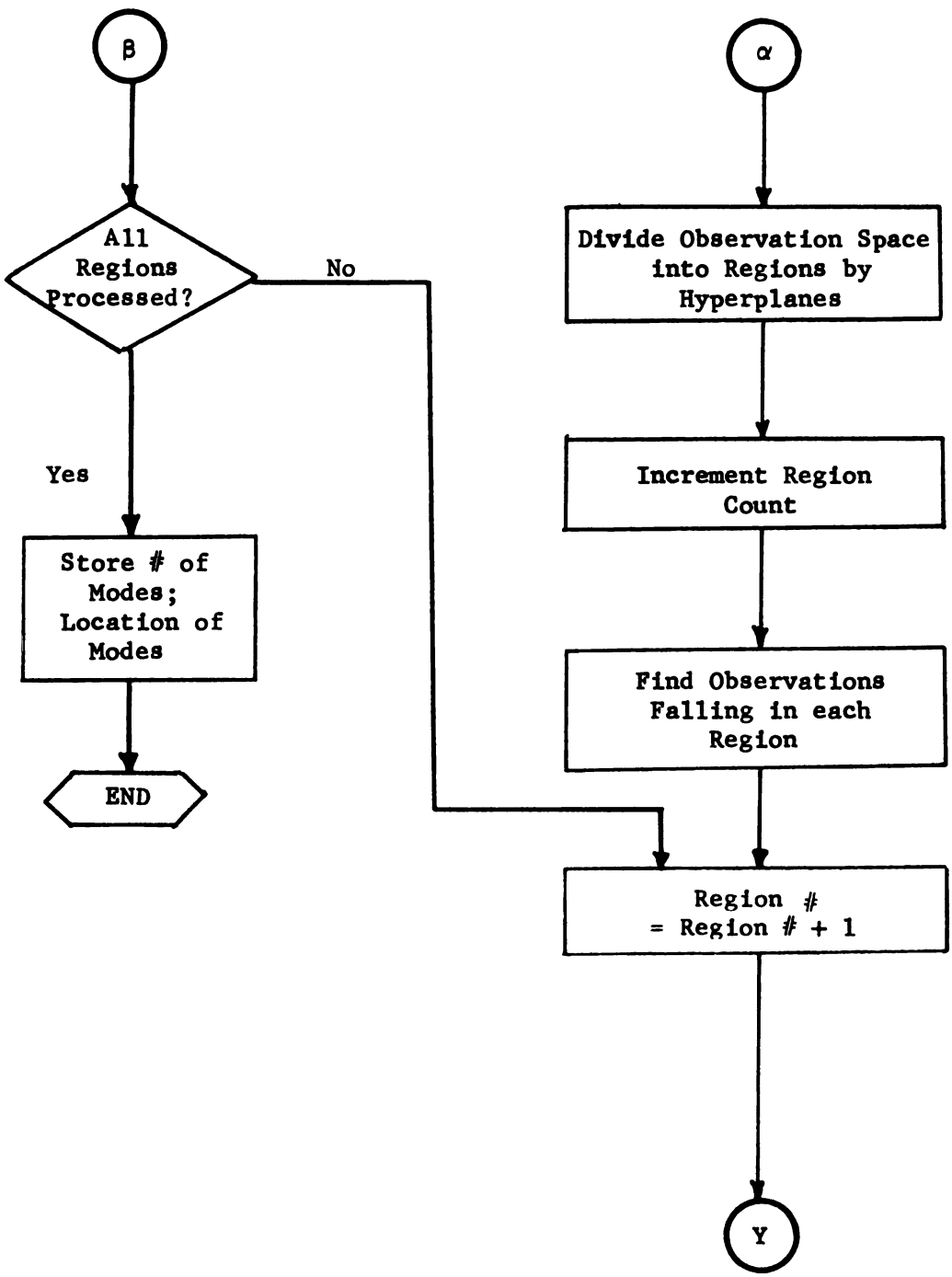
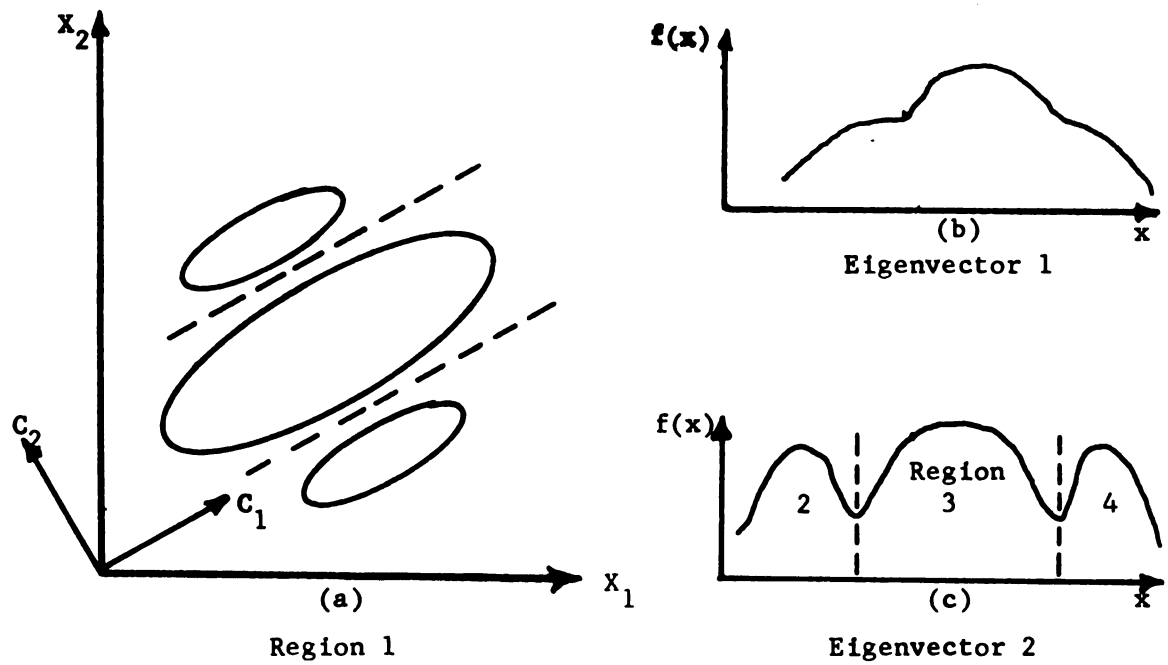
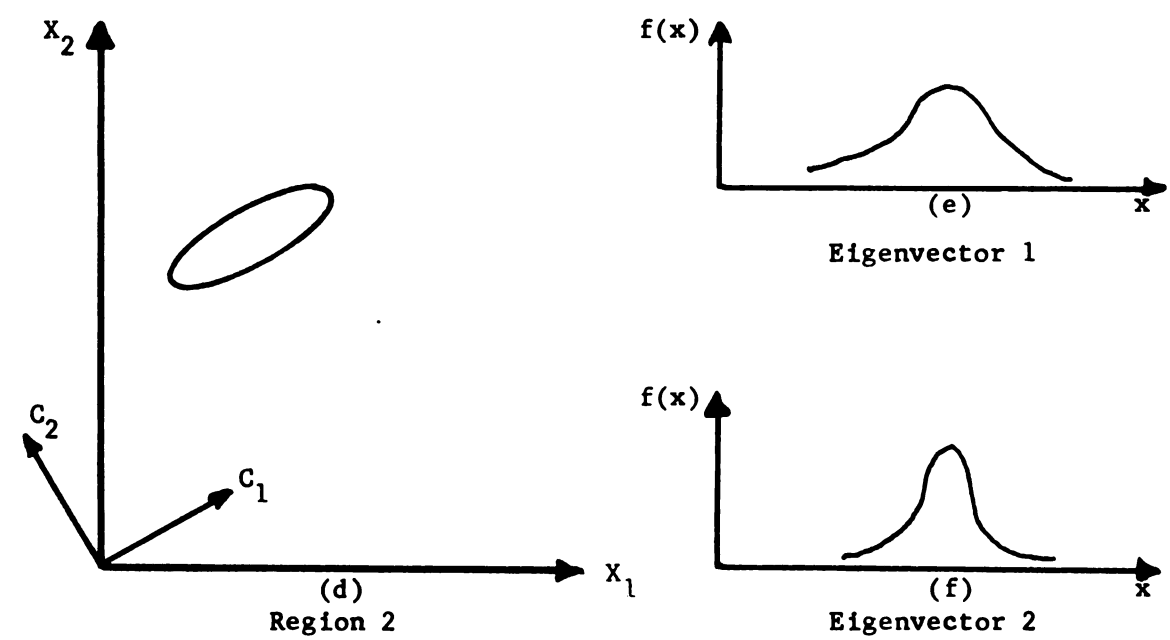


Figure 1 (cont'd.)



Processing of region 2



Regions 3 and 4 processed in a similar manner.

Note: Sketch intended only for visualizing the steps involved in the algorithm.

Figure 2. Two dimensional mode estimation example

Define the normalized measurements as

$$x_{ij} = x^{*}_{ij} - m_j \tag{3.3.4}$$

and the normalized observation vector is

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} . \tag{3.3.5}$$

These normalized vectors are arranged in the form of a $(N \times p)$ matrix, $[A]$:

$$[A] = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} . \tag{3.3.6}$$

The $(p \times p)$ sample covariance matrix, $[S]$, is:

$$[S] = \frac{1}{N-1} [A]^T [A] \tag{3.3.7}$$

and is assumed to be positive definite.

Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $[S]$ arranged in order such that

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p .$$

The first principal (column) eigenvector, C_1, of $[S]$ is the eigenvector corresponding to the eigenvalue λ_1. The remaining (column) eigenvectors are, respectively, $C_2, C_3, \ldots, C_p$, and the

$(p \times p)$ matrix of eigenvectors is

$$[C] = [C_1, C_2, \ldots, C_p] \ .$$

In step (i) these eigenvectors $C_1, C_2, \ldots, C_p$, of $[S]$ are computed treating the entire observation space as one region.

The univariate set of step (ii) is obtained by projecting the N observations onto C_1.

Let $V_{.k}$ represent the univariate data set obtained by projecting the observations onto the k-th eigenvector C_k. Then

$$V_{.k} = \begin{bmatrix} V_{1k} \\ V_{2k} \\ \vdots \\ V_{Nk} \end{bmatrix} = [A]C_k \ ; \quad k = 1,2,\ldots,p \ . \tag{3.3.8}$$

The p data sets so obtained can be arranged as the columns of a $(N \times p)$ matrix, $[V]$, given by,

$$[V] = [A][C] \ . \tag{3.3.9}$$

In step (iii) the mode seeking procedure of Section (3.4) is applied to the data set represented by the first column of $[V]$. If only one mode is detected for this set, the procedure is applied to the set given by the second column of $[V]$ and so on as mentioned in step (iv).

For the set of observations in each region of the observation space obtained in step (v), a sample covariance matrix $[S]$, a matrix of eigenvectors $[C]$ and a matrix $[V]$ is obtained.

Suppose that for the i-th region, only one mode is detected for each of the data sets corresponding to $V_{.1}^{(i)}, V_{.2}^{(i)}, \ldots, V_{.p}^{(i)}$. The components of the mode vector for the i-th region, in the transformed space, are the elements of some row of $[V^{(i)}]$, say, the j-th row.

The mode vector for the region, in the original observation space is therefore,

$$M^{(i)} = [C^{(i)^T}]^{-1} V_j^{(i)^T} + m^{(i)}$$

$$= [C^{(i)}] V_j^{(i)^T} + m^{(i)}; \text{ since } [C^{(i)}]^T [C^{(i)}] = [I] \quad (3.3.10)$$

where

$[C^{(i)}]$ is the matrix of eigenvectors for the i-th region;

$$V_j^{(i)} = [V_{j1}^{(i)} \quad V_{j2}^{(i)} \quad \ldots \quad V_{jp}^{(i)}] ;$$

and

$$m^{(i)} = \begin{bmatrix} m_1^{(i)} \\ \vdots \\ m_p^{(i)} \end{bmatrix}$$

is the vector of feature averages for the N_i observations in the i-th region.

3.4 Algorithm for Mode Estimation - The Univariate Case:

The theoretical formulation of the mode estimation problem for the univariate multimodal case, as presented in the last chapter, requires that those intervals on the real line where

distinct modes are assumed to exist, be known first. This is a condition which must be satisfied prior to the application of Theorem 2.2. Once these intervals have been identified, a consistent estimate of the mode of the underlying pdf of observations falling in this region, can be obtained in a straightforward manner. The problem therefore reduces to that of finding intervals $\{I_i\}_{i=1}^{G}$ on the real line. One method which suggests itself immediately is to actually obtain the pointwise density estimate of the univariate pdf using the estimate

$$f_N(z) = \frac{(r_N - 1)/N}{2\, d_z(r_N)}$$

and to plot these to identify the intervals containing the local extrema. The futility of this approach can be appreciated if one actually constructs such a plot. The local variations in the pointwise density estimate are too great to be of any use in isolating the intervals under consideration.

To overcome these difficulties Fu et al [F-7] suggest the following three "necessarily vague" guidelines:

(i) Some means of smoothing the pointwise approximation of the underlying density estimate is necessary;

(ii) Modes of an underlying density which are further apart from each other should be more readily detected; and

(iii) Modes which have associated with them a high probability mass should be more readily detected.

The approach adopted in this thesis is the construction of an equal bin-count histogram as a primary approximation to the underlying density. This histogram approach performs the

integrating effect of assumption (i), that is, local variations in the pointwise density approximation will tend to be smoothed. A discussion on three types of histograms useful for analyzing a set of empirical data can be found in Dubes [D-1]. The equal bin-count histogram approximation was motivated by this discussion. The construction of histograms for estimating the modes has also been suggested by Sebestyen and Edie [S-2]. Such an approach may also be useful to achieve the three objectives referred to above but their method is not used in this chapter and hence not discussed further.

In an equal bin-count type of histogram the widths of the various bins indicate the concentration of the probability mass in any interval. Bins with a smaller width imply a higher concentration just as in the equal bin width type of histogram a bin with more height is indicative of more observations lying in the interval specified by the bin width.

The first step in the algorithm to estimate the modes is the construction of a histogram. The number of bins desired is specified by the user. This number, NB, must be, preferably, selected such that the number of observations N, (sample size/ region count) is a multiple of NB. Once this number is specified the bin divisions are chosen to fulfill the condition that all bins are required to have the same, or as close to the same as possible, number of counts. The smaller the number of bins specified the greater will be the smoothing effect; that is, specifying a large number of observations per bin supresses the local variations in the histogram approximation to a greater

extent. The locations of the bin divisions carry the information required for further analysis.

An extrema seeking technique is then applied to this histogram in order to determine appropriate intervals in which to search for local maxima. The algorithm for selecting these intervals depends on a parameter, PARA, to be specified by the user. This parameter decides the value of a threshold which essentially determines how much local variation will be tolerated before a decision to specify an extremum interval is made. The algorithm first seeks bin divisions to be used in determining the intervals of existence of local minima between successive modes. The criterion used to store such an interval is that:

(i) there exist a bin to the left of the chosen bin such that the difference in their widths is greater than the threshold specified and

(ii) there exist a bin to the right satisfying a similar condition.

The threshold is calculated as PARA times the minimum bin width. From simulation studies it is found that the value of PARA in the range 1 to 10 gives good results.

The points at which relative minima exist are then chosen as the midpoints of these bins. The set of intervals $\{I_i\}_{i=1}^{G}$ in which modes are assumed to exist is now available. The techniques of the last chapter are now applied to estimate the mode in each interval.

A flow chart for the algorithm is given in Figure 3.

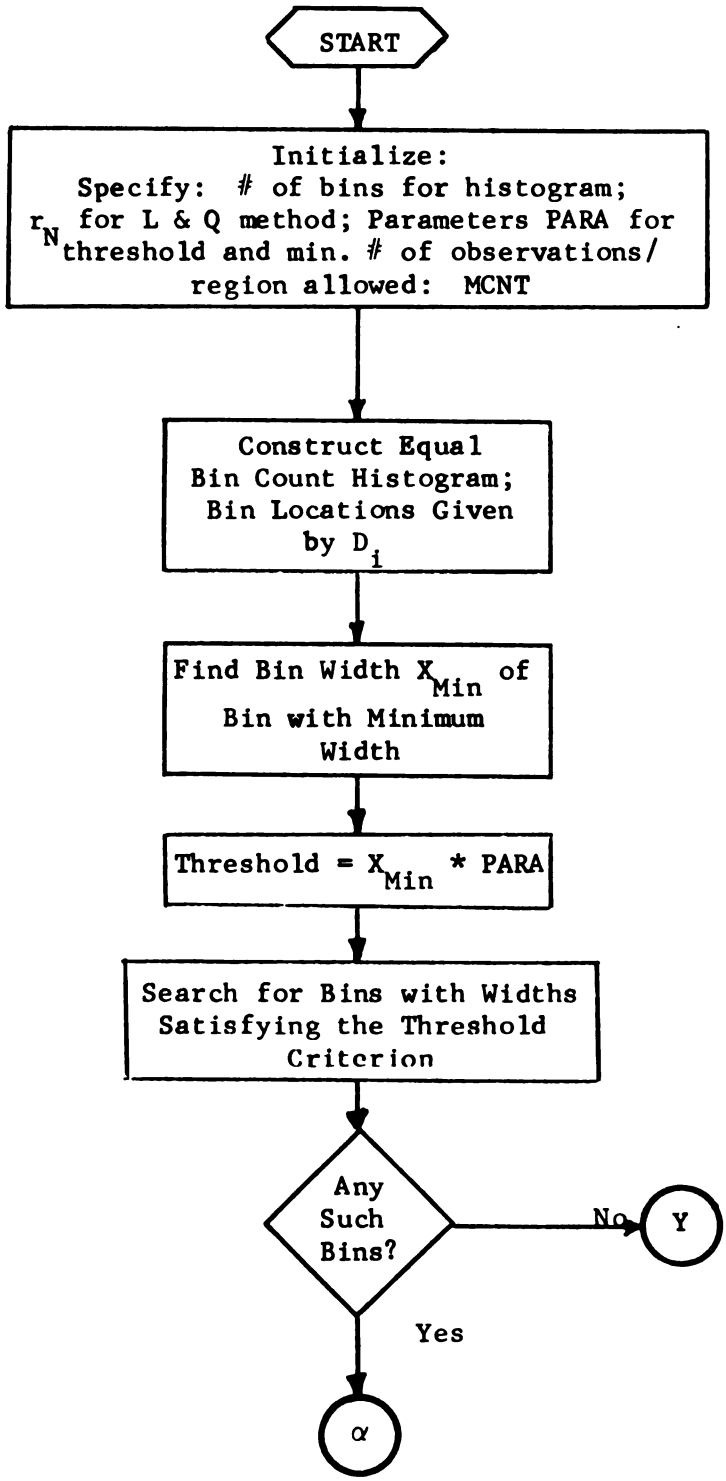


Figure 3. Flow chart for univariate mode estimation

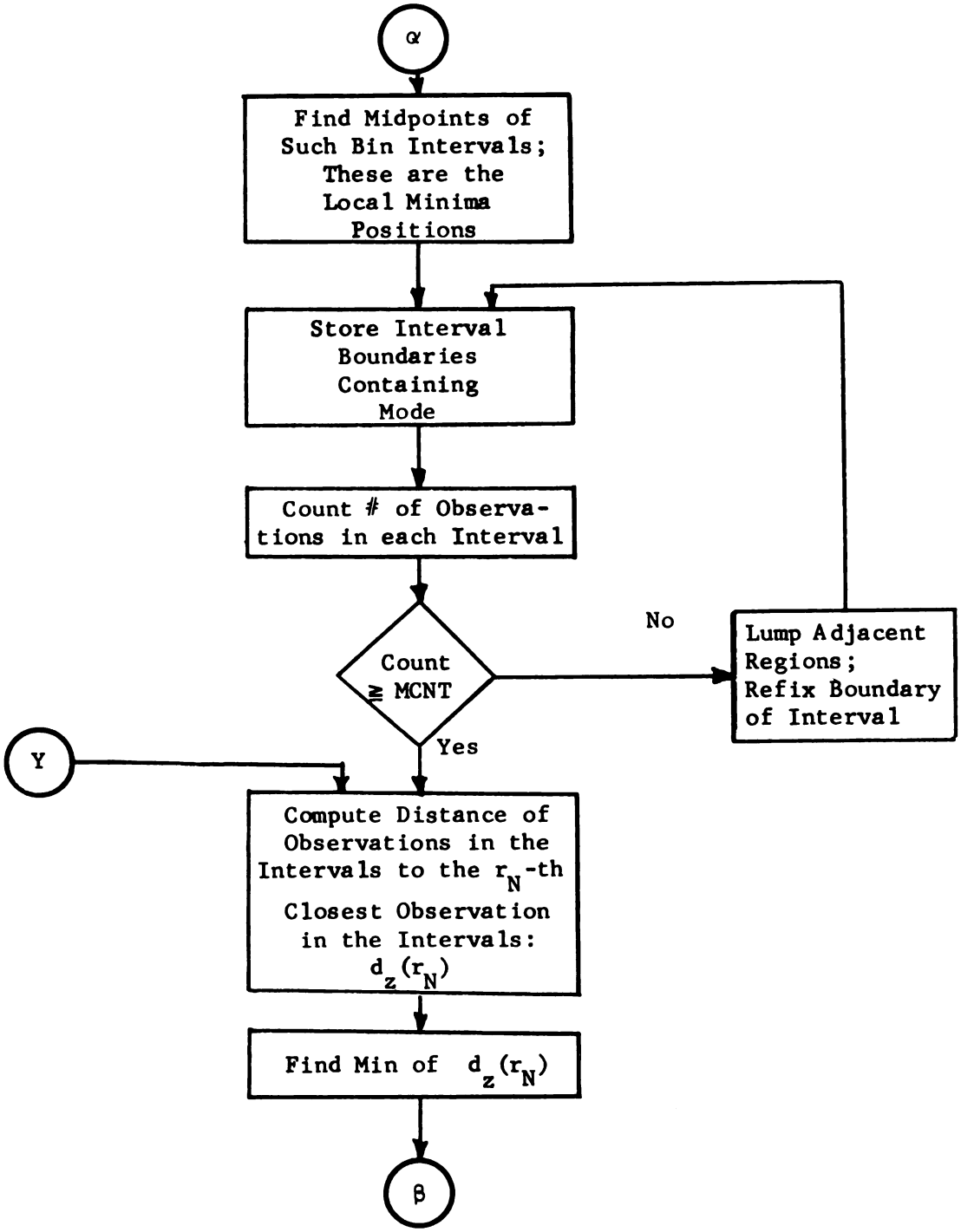


Figure 3 (cont'd.)

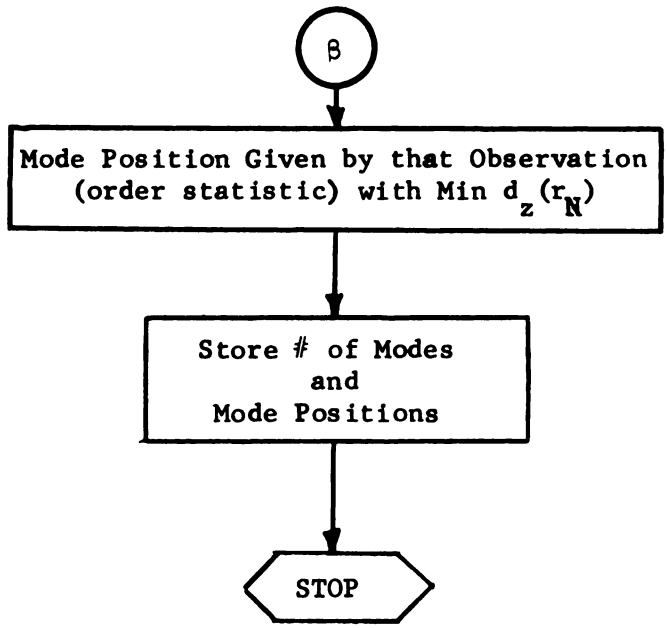


Figure 3 (cont'd.)

From the flow chart it is observed that the number of modes is controlled by the following parameters to be specified by the user:

(i) PARA, the parameter controlling the threshold to be used in finding the interval where the local minima of the pdf exist; and

(ii) MCNT, the minimum number of observations the user allows in any region. If any region is found to contain less number of observations than MCNT, then adjacent regions are lumped till this criterion is satisfied.

The estimate of the location of the mode in any interval is controlled by the value of r_N used in the expression for the pointwise density estimate derived by Loftsgaarden and Quesenberry. They suggest the use of an integer closest to $\sqrt{N}$, where N is the number of observations in a region. To use a more general value r_N is taken as the integer closest to PAR $\sqrt{N}$ where PAR is a third parameter to be specified by the user.

3.5 Results of Tests:

The algorithm presented in this chapter was tested on different sets of simulated as well as real data. The simulated sets of data were obtained by Monte Carlo methods. The results of such test runs are summarized in Tables 1 and 2. The time shown in each case refers to runs on a CDC 6500 digital computer.

Data set # 1. Simulated data generated by a mixture of two univariate Weibull pdfs given by

$$f(x) = 0.4 * W(x;\alpha_1,\beta_1,\delta_1) + 0.6 * W(x;\alpha_2,\beta_2,\delta_2)$$

TABLE 1. Results of Mode Estimation - Simulated Data

Data Set	Sample Size	# of Modes Found	Estimated Location of Modes		Theoretical Location of Modes		Computer CP time	Parameters used for the algorithm		
			Mode (1)	Mode (2)	Mode (1)	Mode (2)		PARA	MCNT	PAR
#1 Univariate: mixture of 2 Weibull pdfs	500	2	0.7132	2.7747	0.7070	2.7035	13.79 secs	7.0	100	1.00
#2 Bivariate: mixture of 2 Gaussian pdfs	300	2	-2.3396 -2.2151	-0.2595 0.3338	-2.023 -2.026	0.088 0.186	11.3 secs	10.0	75	1.00

where

$$W(x;\alpha_i,\beta_i,\delta_i) \triangleq \frac{\beta_i}{\alpha_i} * (x-\delta_i)^{\beta_i-1} * \exp\left[-\frac{1}{\alpha_i} * (x-\delta_i)^{\beta_i}\right], \text{ if } x \geq \delta_i$$

$$= 0, \text{ if } x < \delta_i \ ; \ i = 1,2 \ .$$

$$\alpha_1 = \alpha_2 = 1.0 \ ; \ \beta_1 = \beta_2 = 2.0 \ ; \ \delta_1 = 0.0 \ ; \ \delta_2 = 2.0 \ .$$

Data set # 2: Simulated Data:

Mixture of two dimensional Gaussian distributions given by

$$f(x_1,x_2) = 0.3 * N(x_1,x_2 \ ; \mu_1,\Sigma_1) + 0.7 * N(x_1,x_2;\mu_2,\Sigma_2)$$

where $$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ ; \quad \mu_2 = \begin{bmatrix} -2.0 \\ -2.0 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix} \ .$$

Referring to the results in Table 1, it should be noted that the estimated mode locations for both the sets of data are for a mixture of the two distributions. They are not necessarily the same as the mode locations of the component pdfs. The theorem given below shows that indeed such is the case except under special conditions. This theorem and the proof are given only to explain the results obtained for the simulated data sets. It is pointed out that in the model proposed in Chapter II, the pdf is not treated as a mixture of pdfs.

Theorem 3.4.1:

Let $f_1(x)$ and $f_2(x)$ be two continuous pdfs of the multivariate random variable $X = (X_1,X_2,\ldots,X_p)$ and let $f(x)$ be a

mixture of f_1 and f_2 given by

$$f(x) = p_1 f_1(x) + p_2 f_2(x) \; ; \; p_1 + p_2 = 1 \; .$$

If x_m is a mode of $f(x)$ and x_0 is a mode of $f_1(x)$, then the mode x_m is not the same as x_0 unless ∇f_2 is zero at $x = x_m$.

Proof:

At any mode of $f(x)$,

$$\nabla f(x) = p_1 \cdot \nabla f_1(x) + p_2 \cdot \nabla f_2(x) = 0 \; .$$

Since $\nabla f_1(x)\big|_{x=x_0} = 0$, locally, in a neighborhood of x_0, we have the expansions

$$\nabla f_1(x) = [B_1] \; (x - x_0) + \text{terms involving higher order terms},$$

$$\nabla f_2(x) = A_2 + [B_2] \; (x - x_0) + \text{terms involving higher order terms},$$

where

$$A_2 = \nabla f_2(x)\Big|_{x=x_0} \qquad \text{and}$$

$[B_1]$ and $[B_2]$ are square matrices depending on x_0. Retaining only the first order terms, the mode x_m occurs when

$$p_1[B_1] \; (x - x_0) + p_2 \, A_2 + p_2 \, [B_2] \; (x_m - x_0) = 0$$

from which

$$x_m = x_0 \cdot \; (p_1 \, [B_1] + p_2 \, [B_2])^{-1} \, A_2$$

$$= x_0 - [\frac{p_1}{p_2} \, [B_1] + [B_2]]^{-1} \, A_2$$

assuming that the indicated inverse exists.

If $A_2 = \nabla f_2(x) = 0$ at $x = x_0$, $x_m = x_0$. Q.E.D.

The last condition may be interpreted as either the two distributions are well separated in the observation space or a mode of one of the densities occurs where the other density is a constant.

The theoretical modes of the mixture pdf for the Weibull data are 0.7070 and 2.7035 respectively. Considering the fact that the estimated modes were for data produced by simulation, there is reasonably good agreement between the ideal and estimated mode locations.

The two artificial data sets were used only to illustrate the feasibility of the mode seeking algorithm. Applications to real data are considered next.

Data sets # 3, 4, 5 and 6:

The sets of data mentioned in Table 2 were selected from 150 four dimensional patterns used by Fisher in a classical paper [F-1]. Each pattern represents an iris and each feature represents a measurement on the iris. The patterns are divided into three species/classes called Setosa, versicolor and virginica with fifty patterns in each class.

In each of the tests all the four measurements were used for mode estimation. The parameter, PAR, to be specified, was varied for the different runs to demonstrate its effect on the estimate for the mode locations.

The iris data set has been used by many research workers to test the performance of the algorithms proposed for clustering

Table 2. Results of Mode Estimation - Real Data

Data Set		Sample Size	Parameters			# of Modes Found	Location of Modes			Computer CP time
#	Categories		PARA	PAR	MCNT		(1)	(2)	(3)	
3	IRIS Versicolor & Virginica	100	1.0	1.50	35	2	-	6.3914 2.7937 4.5138 1.3818	5.9423 2.9527 4.9438 1.8310	4.36 secs
4	IRIS Setosa & Versicolor	100	1.0	2.00	35	2	5.2086 3.5313 1.4566 0.2007	6.2338 2.4324 5.0690 1.5925	-	4.28 secs
5	IRIS Setosa & Virginica	100	1.0	1.00	35	2	5.1150 3.4420 1.5342 0.2029	-	7.3153 2.6078 7.4430 2.8999	3.47 secs
6	IRIS Setosa Versicolor Virginica	150	1.0	1.00	35	3	5.1150 3.4420 1.5342 0.2029	5.5501 2.4717 4.4728 1.3960	6.8392 2.5144 6.6541 2.5693	4.43 secs

as well as classification. It is known [Z-1] that the setosa species are well separated from the versicolor and virginica categories. The latter two are fairly 'mixed'. In two dimensions, this fact has been illustrated by Sammon [S-1] and Dubes [D-2].

The location of the modes given in Table 3 for data sets 5 and 6 (using the same sets of parameters) shows this very well. The mode location for the setosas remained the same whereas for the virginicas it changed when pooled with the versicolors. Results for data set 4 show the effect of changing the parameter, PAR, (used in computing r_N) on the location of the mode. Though not specifically shown in the table, for the set 4, with PAR = 1.0, the mode location for the setosas is found to be identical to the location estimated for sets 5 and 6. It is evident that for well separated groups the modes remain stationary and they are perturbed if the sets are fairly mixed.

3.6 Chapter Summary:

In this chapter a practical technique for the estimation of the number and location of the modes of a multivariate multimodal pdf is presented. The points to be noted are:

(i) The histogram approximation to the underlying pdf is used only to find the intervals in which to seek the location of the mode and not for the estimation of the modes themselves.

(ii) The equal bin count type of histogram rather than the equal bin width type is used for the approximation.

(iii) Pointwise density approximations are not computed for estimating the mode.

(iv) The mode estimating procedure did not require that the whole underlying density be absolutely continuous. What is required is that the density be absolutely continuous over an interval containing the pointwise mode estimate.

(v) A priori knowledge of the number of modes is not a requirement to start the procedure although this number is controlled by the user in specifying the parameters PARA and MCNT.

The method presented in this chapter is believed to be useful in a number of cluster seeking problems. Other methods for handling the multivariate problem directly (instead of obtaining first a univariate data set) may or may not be computationally faster, but such methods have not been considered in this chapter.

CHAPTER IV

Cluster Seeking and Pattern Classification

4.1 Introduction:

In this chapter two applications of the mode seeking algorithm are presented, namely the use of the estimated modes for clustering, and the utilization of the clusters so formed for pattern classification.

The underlying assumption behind the clustering procedure is that the modes are the prototype observations around which each of the observations or patterns tend to cluster. A cluster is defined to be the set of those observations surrounding a mode which is nearest to them in the sense of some metric. Hence the measure of "similarity" is the distance of an observation from each of the modes.

A general statement of the pattern classification problem involves the consideration of its three fundamental aspects, characterization, abstraction and generalization. Characterization involves the selection of the independent variables which characterize the different classes from which the patterns originated. Abstraction refers to the process of obtaining a decision rule for classifying a new pattern with vector x of unknown class. The decision rule is arrived at using all the available information. The ability of the decision rule or classifier to correctly

categorize the samples from unknown class is called generalization.

In the classification procedure proposed here, the clusters produced by the clustering algorithm are used as the sets of training samples from the G pattern classes. Assuming that there exists a pdf underlying each pattern class, the mode of the pdf is estimated using the mode seeking algorithm of Chapter III. Classification of any patterns with unknown class is achieved using a minimum distance classifier.

It should be noted that the modes used for the clustering algorithm are not necessarily identical to the modes of the clusters estimated during the classification phase referred to above, because the data sets are slightly different. If the data are initially well separated in the observation space, the two sets of modes tend to be identical.

As is usually done in most pattern classification studies, to test the "goodness" of the algorithms, tests are performed on data sets whose sources are known in advance. The measure of performance is stated in terms of the percentage or number of observations misclassified.

4.2 A Similarity Measure and an Algorithm for Clustering:

Let $x_1, x_2, \ldots, x_N$ be the vectors of the observations which are to be grouped using the measure of similarity defined by the Euclidean distance $d(x_i, M_j)$ between the observation vector x_i and the estimated mode vector M_j. The vectors M_j are assumed to be available as a result of the mode seeking algorithm.

Let $\pi_1, \pi_2, \ldots, \pi_G$ denote the G clusters each corresponding to an estimated mode with vector M_j.

Then the set of observation vectors $\{x_i\}_{i=1}^{N}$ is partitioned into the G clusters such that

$x_i \in \pi_j$ if and only if

$$d(x_i, M_j) < d(x_i, M_k) \; ; \; j \neq k, \quad \forall_k \; .$$

Even though the extension of the mode seeking algorithm for determining cluster membership is straightforward, it is given in flow chart form in Figure 4, for the sake of completeness.

4.3 Distance Measures for Classification:

In addition to the Euclidean and the Mahalanobis generalized distance functions, two other heuristically motivated measures are tried for classifying an unknown pattern X, using a minimum distance-to-mode classifier. The four measures are summarized below, where $M^{(i)}$ denotes the mode of the i-th cluster.

(i) Euclidean distance function:

$$d_E(x, M^{(i)}) = [(x - M^{(i)})^T (x - M^{(i)})]^{\frac{1}{2}} \; . \qquad (4.3.1)$$

(ii) Mahalanobis generalized D^2-measure:

$$d_M(x, M^{(i)}) = (x - M^{(i)})^T \, [S_W]^{-1} (x - M^{(i)}) \qquad (4.3.2)$$

where $[S_W]$ is defined as the sum of the sample covariance matrices of all the clusters.

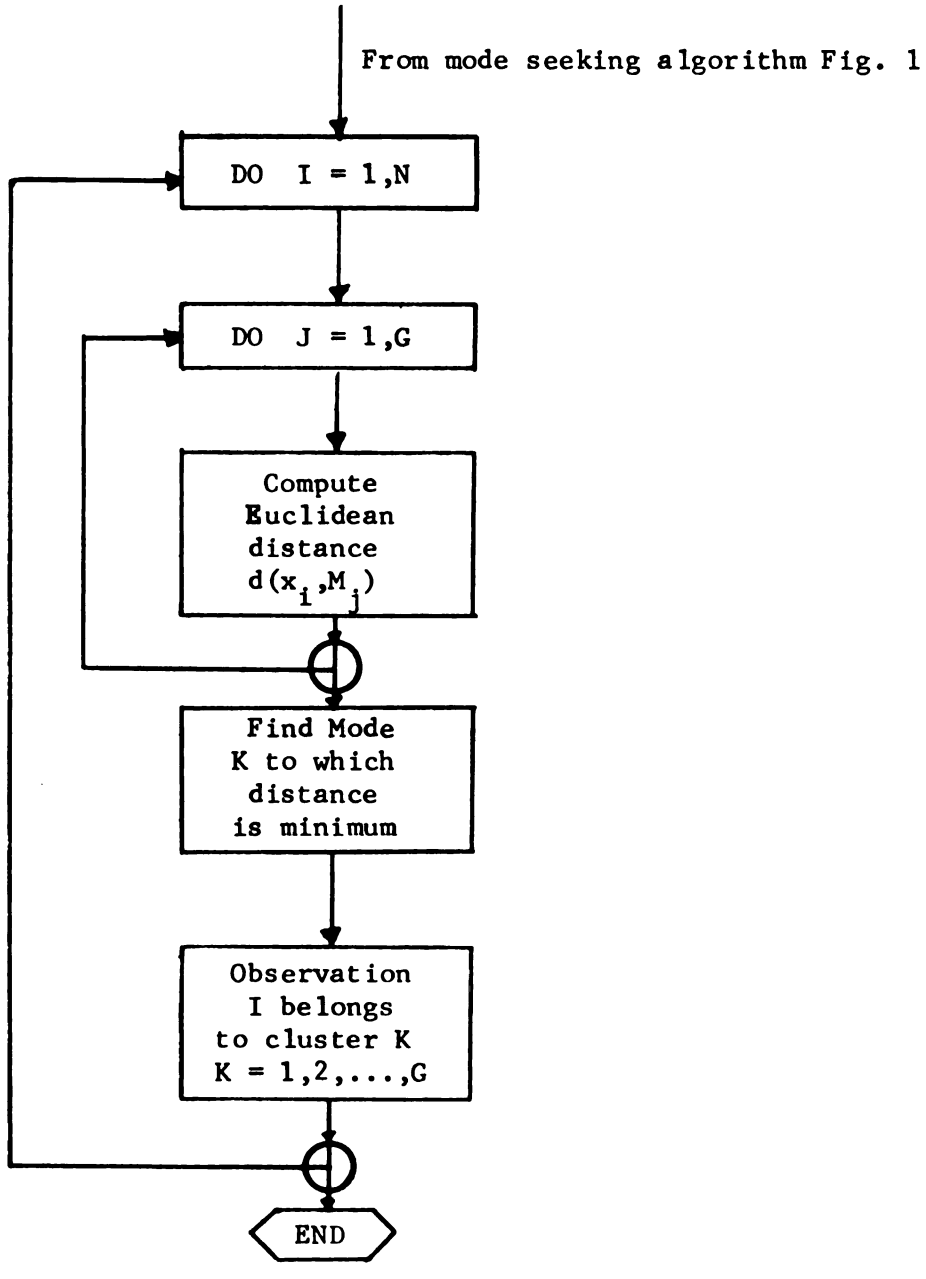


Figure 4. Flow chart for the clustering algorithm

(iii) "Symmetric" generalized distance measure:

$$d_s(x,M^{(i)}) = (x - M^{(i)})^T[S^{(i)}](x - M^{(i)}) \qquad (4.3.3)$$

where $[S^{(i)}]$ is a $(p \times p)$ matrix proportional to the inverse of the second-moment matrix of the observations in the i-th cluster, with respect to the mode, $M^{(i)}$, of the cluster. It is defined as

$$[S^{(i)}] = \frac{N_i}{p}\left[\sum_{j=1}^{N_i}(x_j^{*(i)} - M^{(i)})(x_j^{*(i)} - M^{(i)})^T\right]^{-1} \qquad (4.3.4)$$

where the asterisk denotes that raw measurements are implied and N_i is the number of observations in the i-th cluster.

(iv) 'Asymmetric' generalized distance measure:

$$d_{AS}(x,M^{(i)}) \triangleq \frac{q + \sqrt{q^2 + 4wr}}{2w} \qquad (4.3.5)$$

where:

with $a \triangleq x - M'^{(i)}$; $b \triangleq M^{(i)} - M'^{(i)}$;

$q = b^t[S_{AS}^{(i)}]a + a^t[S_{AS}^{(i)}]b - 2b^t[S_{AS}^{(i)}]b$;

$w = 1 - b^t[S_{AS}^{(i)}]b > 0$;

$r = (a - b)^t[S_{AS}^{(i)}](a - b) > 0$; and

$$[S_{AS}^{(i)}] \triangleq \frac{N_i}{p}\left[\sum_{j=1}^{p}(x_j^{*(i)} - M'^{(i)})(x_j^{*(i)} - M'^{(i)})^T\right]^{-1} \qquad (4.3.6)$$

$M'^{(i)}$ is a vector representing the center of a constant unit distance ellipse. A detailed derivation of the measures (iii) and (iv) is given in Appendix A.

The 'asymmetric' distance measure is motivated by the belief that if the pdf underlying the observations in any cluster is asymmetric in the sense that the mean is different from the mode, the distance measure chosen should reflect this skew. Because the dispersion of the observations is, in general, not the same in all directions, it is felt that the measure of distance used should reflect this.

4.4 A Classification Algorithm:

Based on the minimum distance-to-mode decision rule, an algorithm for classifying a pattern vector x, from unknown class, is given in Figure 5 in flow chart form.

4.5 Results:

The clustering algorithm and its application to pattern classification using the different distance measures are tried on the sets of data used in Chapter III. Another set of real data tested consisted of measurements of different types of grain. Ehrlich and Weinberg [E-2] discuss how grain shape may be described as precisely as needed by a Fourier series expansion of the radius about the center of mass utilizing co-ordinates of peripheral points. They also give illustrative examples to show that the shape variables easily discriminate grain differences arising from geographic, stratigraphic and process factors. In pattern recognition parlance this may be considered as their method of feature extraction. Professor Weinberg made available a set of eight feature measurements for each of the grains - Navy Bean,

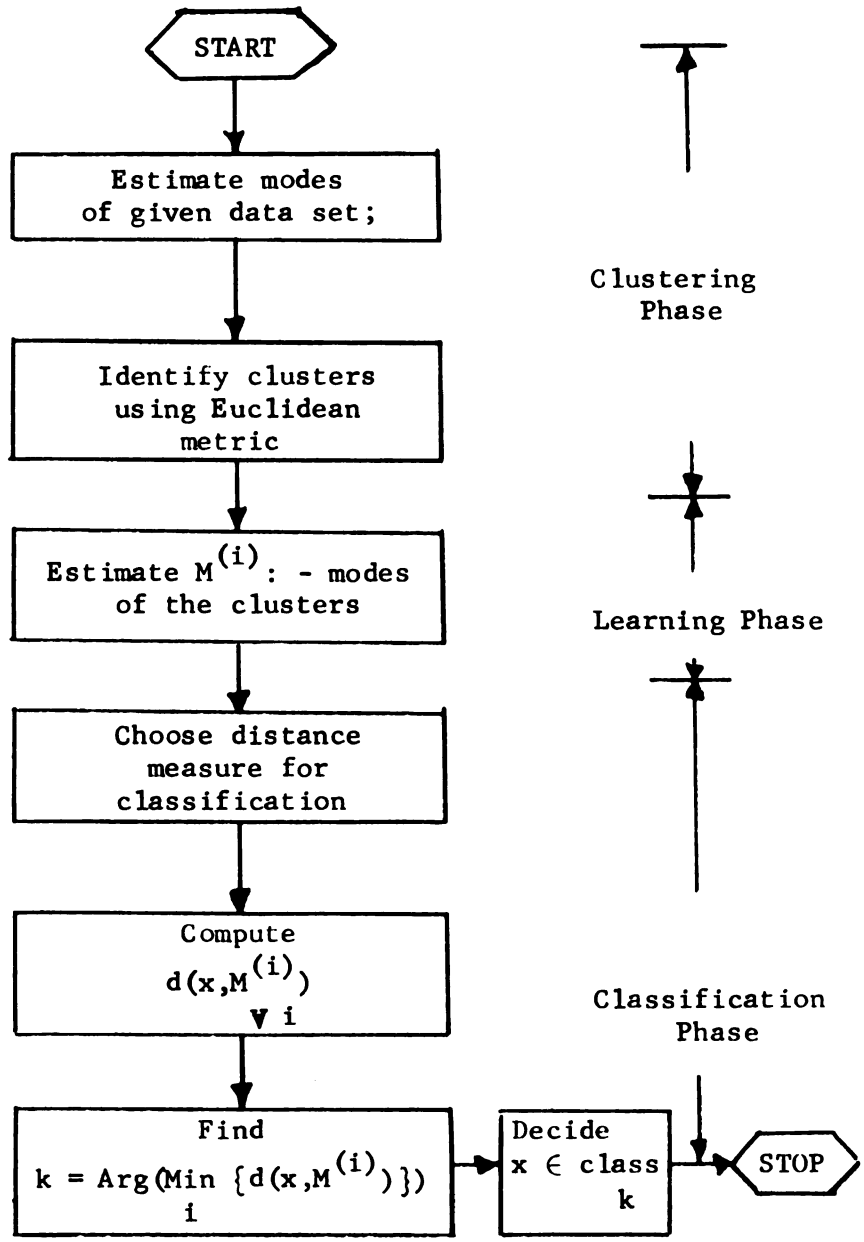


Figure 5. Flow chart for classification

Wheat and Oats. As the sample size of each variety of grain is small compared with the number of features only four out of the eight features are chosen for the test. Specifically, the available sample sizes are:

Navy Bean: 46

Oats : 50

Wheat : 42

The results of these tests are summarized in Table 3.

The mode seeking algorithm is also applied to estimate the modes of the pdf underlying the actual training sets as distinct from the sets identified by the clusters. For each data set two types of tests are conducted; first using all the samples from each training set and next using a subset of the set from each class. Both the tests gave encouraging results. The results of the former type of test are tabulated in Table 4.

An inspection of the results in Table 3 reveal that with one exception all the reclassifications based on the initial cluster membership are worse almost inversely proportional to the degree of the intuitive sophistication of each distance measure. The reason for this is not known and has not been investigated further.

4.6 Chapter Summary:

In this chapter a clustering procedure is discussed using the similarity measure defined by the distance of an observation point from the modes of the distribution which generated the set of observations. If the ultimate aim is clustering of a set of empirical data, the algorithm can be terminated at this stage.

Table 3. Results of clustering and classification algorithms

Data Set	Number of Dimensions	Cluster Modes			Number Misclassified by Cluster Algorithm *	No. Misclassified in classification			
		Mode (1)	Mode (2)	Mode (3)		Distance Measure			
						Euclid	Symmetric generalized	Asymmetric generalized	Mahalanobis
Oats Navy Bean Wheat	4	0.5546 0.0177 0.2195 0.0231	0.1615 0.0156 0.0123 0.0068	0.3042 0.0198 0.0564 0.0107	3/138	7/138	1/138	3/138	1/138
Iris Versicolor & Virginica	4	-	5.8720 2.7899 4.0673 1.2498	5.8761 2.7682 5.0027 1.7691	5/100	13/100	10/100	20/100	5/100
Iris Setosa Versicolor & Virginica	4	5.0964 3.4004 1.4975 0.2340	5.8720 2.7899 4.0673 1.2498	5.8761 2.7682 5.0027 1.7691	5/150	13/150	10/150	20/150	4/150
Bivariate Gaussian	2	-0.0755 0.2656	-2.0301 -2.0852	-	1/300	3/300	8/300	11/300	4/300

* see Figure 5, p. 47.

Table 4. Results of classification using training samples

Data Set	Number of Dimensions	Modes estimated using all categorized samples			Number Misclassified Distance Measure			
		Mode (1)	Mode (2)	Mode (3)	Euclid	Symmetric generalized	Asymmetric generalized	Mahalanobis
Oats Navy Bean Wheat	4	0.5546 0.0177 0.2195 0.0231	0.1611 0.0159 0.0095 0.0091	0.3031 0.0163 0.0588 0.0188	1/138	0	0	1/138
Iris Setosa Versicolor Virginica	4	5.0964 3.4004 1.4975 0.2324	5.8454 2.5682 4.0825 1.2155	6.4461 2.7430 5.5127 2.0456	10/150	2/150	3/150	3/150
Bivariate Gaussian	2	-0.0755 +0.2656	-2.0301 -2.0852	-	3/300	8/300	11/300	4/300

However if the aim is to detect a cluster structure for pattern classification analysis, the clusters generated could be used as sets of training patterns, their modes estimated and a pattern vector x from unknown class, classified using a minimum distance-to-mode classifier.

It is also demonstrated that if training subsets from known classes are availabe *a priori*, then the algorithm could be used for classification studies.

Different distance measures are discussed even though the two intuitively more sophisticated measures performed poorly on the data sets mentioned.

It may be recalled that in the mode seeking algorithm the observation space is partitioned into regions and the modes are estimated for observations in each such region. To explain why the observations in these regions are not treated as the final clusters, one has only to note that the modes could also be estimated using other procedures which do not call for such partitioning. The algorithm to seek clusters may also be used with such mode seeking methods.

CHAPTER V

Conclusions and Suggestions for Future Work

5.1 Thesis Summary:

In this thesis a new model for the clustering problem is presented. The formulation of the problem is based on assumptions which are not severely restrictive. The assumptions are of a nature similar to those generally made in most of the methods for the statistical analysis of the pattern recognition problem. The model is motivated by the belief that to find a cluster-structure in a set of empirical data, it is not necessary to treat the underlying pdf as a mixture of several source pdfs. The pdf need be treated, in its own right, as one governed by a multi-modal continuous distribution with one or more distinct modes. This leads to the identification of the clustering problem with that of estimating the modes.

An algorithm is given to estimate these modes and the modes estimated are used to detect the cluster membership. As no classified training samples from each pattern class are available *a priori*, the clusters so formed are taken to be the training sets for training a minimum distance classifier. However if training patterns from each class are available in advance, it is shown that the mode seeking algorithm could be used to complete the abstraction phase of the pattern classification problem.

Apart from the standard Euclidean and the Mahalanobis D^2-measure, two heuristically motivated distance measures are discussed. Even though the reasoning involved in the derivation of these new measures seems to be intuitively sound, the measures performed poorly contrary to expectation. However the Mahalanobis D^2-measure gives very good results when used in conjunction with the proposed model and strengthens the belief about the suitability of the model.

5.2 Conclusions:

The clustering procedure proposed is useful in a number of situations. However it is not hard to imagine examples where this approach may not give the desired results. One such situation is described by two concentric point sets in a plane. Ideally, one would like to obtain from a clustering algorithm applied to this set, exactly two clusters. However the scheme proposed here is likely to detect more than two clusters. It is doubtful whether any of the presently available clustering techniques will produce exactly two concentric clusters without rejecting any of the observations.

The method will certainly perform very well in situations where the ratio of the inter-cluster distance to intra-cluster distance is high; that is, in situations where the data are well separated into groups in the observation space. Such a performance is, of course, to be expected of any good clustering algorithm.

As implemented the algorithm is slightly more expensive as regards computer time involved, compared with the ISODATA, operating

on the same data sets. However this higher cost is compensated by the better performance produced at least in the case of the iris versicolor and virginica data sets. It is not claimed that the performance will be better in general but certainly the results are comparable to those of other popular algorithms currently in use.

The mode seeking algorithm is ideally suited for obtaining non-parametric decision rules based on distance to modes, given the training samples from each class.

In cases where the number of clusters present is not known in advance, the applicability of this procedure is quite clear. In fact this is a distinct advantage claimed for this method.

5.3 Suggestions for Future Work:

The two new distance measures used for classification are based on a heuristically sound principle, that the distance measure employed must in some way reflect the dispersion of the observations. The manner of implementation adopted here did not give encouraging results. The reasons for instability, that is, why with a distance measure as implemented, the results diverge instead of converging to form 'ideal' clusters (or remain stationary) are worth investigation. Other methods of implementing this notion may also be considered for future work. It is quite possible that better results may be obtained. Iterations performed with all the four distance measures show no further improvement in performance. Even with the classical Euclid and Mahalanobis distance measures, the iterative process diverges and the reason for this divergence requires further investigation.

Another area of future work is the estimation of the modes of a multivariate pdf directly from the Loftsgaarden and Quesenberry estimator of Chapter II, instead of first constructing a univariate data set. A stochastic approximation approach to estimate the pdf [K-1] may result in a computationally more economical method to solve the problem of mode estimation.

BIBLIOGRAPHY

BIBLIOGRAPHY

[A-1] Aizerman, M.A., Braverman, E.M., and L.I. Rozonoer, "Theoretical principles of potential function method in the problem of teaching automata to recognize classes", Automation and Remote Control, Vol. 25, No. 6, pp. 821-837, 1964.

[A-2] Augustson, J.G. and J. Minker, "An analysis of some graph-theoretical cluster techniques", Technical Report NGL-21-002-008, No. 70-106, University of Maryland, College Park, Maryland, 1970.

[B-1] Ball, G.H. and D.J. Hall, "ISODATA, a novel method of data analysis and pattern classification", Technical Report, Stanford Research Institute, Menlo Park, California, 1965.

[B-2] Bonner, R.E., "On some clustering techniques", I.B.M. Jour. Res. Dev., Vol. 8, pp. 22-32, 1969.

[C-1] Chernoff, H., "Estimation of the mode", Ann. of Inst. Stat. Math., Toyko, Vol. 16, pp. 31-41, 1964.

[D-1] Dubes, R.C., "Data reduction with grouping and Weibull models", Interim Scientific Report No. 7, Div. of Eng. Res., Michigan State University, East Lansing, Michigan, 1970.

[D-2] _____, "Information compression, structure analysis, and decision making with a correlation matrix", Interim Scientific Report No. 11, Div. of Eng. Res., Michigan State University, East Lansing, Michigan, 1970.

[E-1] Edwards, A.W.F. and L.T. Cavalli-Sforza, "A method for cluster analysis", Biometrics, Vol. 21, No. 2, pp. 362-375, 1965.

[E-2] Ehrlich, R. and B. Weinberg, "An exact method for characterization of grain shape", Jour. of Sedimentary Petrology, Vol. 40, No. 1, pp. 205-212, 1970.

[F-1] Fisher, R.A., "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 3, Part 2, pp. 179-188, 1936.

[F-2] Fortier, J. and H. Solomon, "Clustering Procedures", in Multivariate Analysis - I, Ed. P.R. Krishniah, Academic Press, New York, 1966.

[F-3] Fralick, S.C., "Learning to recognize patterns without a teacher", IEEE Trans., Inform. Theory, Vol. IT-13, pp. 57-64, 1967.

[F-4] Friedman, H.P. and J. Rubin, "On some invariant criteria for grouping data", Jour. of Amer. Stat. Assoc., Vol. 62, pp. 1159-1179, 1967.

[F-5] _____, "The Logic of the Statistical Methods", The Borderline Syndrome, Ch. 5, Basic Books, New York, 1968.

[F-6] Fu, K.S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

[F-7] Fu, K.S. and E.G. Henrichon, Jr., "On non-parametric methods for pattern recognition", Technical Report No. TR-EE 68-19, School of Electrical Engineering, Purdue University, Lafayette, Indiana, 1968.

[G-1] Gotlieb, C.C. and S. Kumar, "Semantic clustering of index terms", Jour. ACM, Vol. 15, pp. 493-513, 1968.

[G-2] Gower, J.C., "A comparison of some methods of cluster analysis", Biometrics, Vol. 23, pp. 623-637, 1967.

[G-3] Gower, J.C. and G.J.S. Ross, "Minimum spanning trees and single linkage cluster analysis", App. Statistics, Vol. 18, No. 1, pp. 54-64, 1969.

[G-4] Grenander, U., "Some direct estimates of the mode", Ann. of Math. Stat., Vol. 36, pp. 131-138, 1965.

[H-1] Hilborn, C.G. and D.G. Lainiotis, "Optimal unsupervised learning multicategory dependent hypotheses pattern recognition", IEEE Trans. Inform. Theory, Vol. IT-14, pp. 468-470, 1968.

[H-2] Hotelling, H., "Analysis of a complex of statistical variables into principal components", Jour. Educ. Psych., Vol. 24, pp. 417-441, 1933.

[J-1] Jardine, N. and R. Sibson, "The construction of hierarchic and non-hierarchic classifications", Comp. Jour., Vol. 11, pp. 177-184, 1968.

[J-2] Johnson, S.C., "Hierarchical clustering schemes", Psychometrika, Vol. 32, pp. 241-254, 1967.

[J-3] Jones, K.S. and D. Jackson, "Current approaches to classification and clump-finding at the Cambridge Language Research Unit", Comp. Jour., Vol. 10, pp. 29-37, 1967.

[K-1] Kashyap, R.L. and C.C. Blaydon, "Estimation of probability density and distribution functions", Technical Report TR-EE67-14, School of Electrical Engineering, Purdue University, Lafayette, Indiana, 1967.

[K-2] Kruskal, J.B., "Multidimensional scaling by optimizing goodness of fit to non-metric hypothesis", Psychometrika, Vol. 29, No. 2, pp. 115-129, 1964.

[L-1] Lance, G.N. and W.T. Williams, "Computer programs for monothetic classification (association analysis)", Comp. Jour., Vol. 8, pp. 246-249, 1965.

[L-2] _____, "A general theory of classificatory sorting strategies, Part I: hierarchical systems", Comp. Jour., Vol. 1, pp. 82-85, 1968.

[L-3] Loftsgaarden, D.O. and C.P. Quesenberry, "A non-parametric estimate of a multivariate density function", Ann. of Math. Stat., Vol. 36, pp. 1049-1051, 1965.

[M-1] Mattson, R.L. and J.E. Dammann, "A technique for determining and coding subclasses in pattern recognition problems", I.B.M. Jour. Res. Develop., pp. 294-302, 1965.

[M-2] McQuitty, L.L., "Hierarchical linkage analysis for the isolation of types", Educ. Psychol. Measurement, Vol. 20, No. 1, pp. 55-67, 1960.

[M-3] Mendel, J.M. and K.S. Fu, Adaptive, Learning and Pattern Recognition Systems, Theory and Applications, Academic Press, New York, 1970.

[M-4] Michener, C.D. and R.R. Sokal, "A qualitative approach to a problem in classification", Evolution, Vol. 11, pp. 130-162, 1957.

[M-5] Moore, D.S. and E.G. Henrichon, "Uniform consistency of some estimates of a density function", Purdue University Stat. Dept., No. 168, 1968.

[M-6] Murthy, V.K., "Estimation of probability density", Ann. of Math. Stat., Vol. 36, pp. 1027-1031, 1965.

[N-1] Nagy, G., "State of the art in pattern recognition", IEEE Proc., Vol. 56, No. 5, pp. 836-862, 1968.

[N-2] Nilsson, N.J., Learning Machines: Foundations of Trainable Pattern Classifying Systems, McGraw-Hill, New York, 1965.

[N-3] Nunnally, J., "The analysis of profile data", Psych. Bulletin, Vol. 59, No. 4, pp. 311-319, 1962.

[O-1] Owen, J., "The consistency of a non-parametric decision procedure", Engr. Note No. 334, Applied Res. Lab., Sylvania Electronic Systems, 1964. (Quoted in [F-7]).

[P-1] Parzen, E., "On estimation of a probability density function and mode", Ann. of Math. Stat., Vol. 33, pp. 1065-1076, 1962.

[P-2] Patrick, E.A. and J.C. Hancock, "Non-supervised sequential classification and recognition of patterns", IEEE Trans. Inform. Theory, Vol. IT-12, pp. 362-372, 1966.

[P-3] Patrick, E.A., "On a class of unsupervised estimation problem", IEEE Trans. Inform. Theory, Vol. IT-14, pp. 407-415, 1968.

[P-4] Patrick, E.A. and J.P. Costello, "Unsupervised estimation and processing of unknown signals", Technical Report TR-69-430, Rome Air Development Center, Rome, New York, 1970.

[R-1] Rao, C.R., "The use and interpretation of principal component analysis in applied research", Sankhya, The Indian Journal of Statistics, Series A, 26, pp. 329-358, 1965.

[R-2] Rohlf, F.J., "Adaptive hierarchical clustering schemes", Sys. Zoology, Vol. 19, No. 1, pp. 58-82, 1970.

[R-3] Ruspini, E.H., "A new approach to clustering", Inform. Contr., Vol. 15, pp. 22-32, 1969.

[S-1] Sammon, J.W., Jr., "A non-linear mapping for data structure analysis", IEEE Trans. Computers, Vol. C-18, No. 5, pp. 401-409, 1969.

[S-2] Sebestyen, G. and J. Edie, "An algorithm for non-parametric pattern recognition", IEEE Trans. Elec. Computers, Vol. EC-15, No. 6, pp. 908-915, 1966.

[S-3] Sokal, R.R. and P.H.A. Sneath, Principles of Numerical Taxonomy, W.H. Freeman, San Francisco, California, 1963.

[S-4] Spragins, J., "Learning without a teacher", IEEE Trans. Inform. Theory, Vol. IT-12, pp. 223-230, 1966.

[S-5] Stanat, D.F., "Unsupervised learning of mixtures of probability functions", in Pattern Recognition, Ed. L. Kanal, Thompson, Washington, D.C., 1966.

[S-6] Switzer, P., "Statistical techniques in clustering and pattern recognition", Technical Report No. 139, Department of Statistics, Stanford University, Stanford, California, 1968.

[T-1] Teicher, H., "Identifiability of mixtures of product measures", Ann. of Math. Stat., Vol. 38, pp. 1300-1302, 1967.

[T-2] _____, "Identifiability of finite mixtures", Ann. of Math. Stat., Vol. 34, pp. 1265-1269, 1963.

[T-3] _____, "Identifiability of mixtures", Ann. of Math. Stat., Vol. 32, pp. 244-248, 1961.

[T-4] Tryon, R.C., Cluster Analysis, Edwards, Ann Arbor, Michigan, 1939.

[V-1] Venter, J.H., "On estimation of the mode, Ann. of Math. Stat., Vol. 38, pp. 1446-1455, 1967.

[W-1] Ward, J.H., "Hierarchical grouping to optimize an objective function", Jour. Amer. Stat. Assn., Vol. 58, pp. 236-245, 1963.

[W-2] Wilks, S.S., Mathematical Statistics, Ch. 18, Wiley, New York, 1963.

[W-3] Williams, W.T. and J.M. Lambert, "Multivariate methods in plant ecology I. Association analysis in plant communities", J. Ecol., Vol. 47, pp. 83-89, 1959.

[W-4] Wirth, M., Estabrook, G., and D. Rogers, "A group theory model for systematic biology", Systematic Zoology, Vol. 15, No. 1, pp. 59-69, 1966.

[Y-1] Yakowitz, S.J., "Unsupervised learning and the identification of finite mixtures", IEEE Trans. Inform. Theory, Vol. IT-16, pp. 330-338, 1970.

[Y-2] _____, "A consistent estimator for the identification of finite mixtures", Ann. of Math. Stat., Vol. 40, No. 5, pp. 1728-1735, 1969.

[Y-3] Yakowitz, S.J. and J.D. Spragins, "On the identifiability of finite mixtures", Ann. of Math. Stat., Vol. 39, No. 1, pp. 209-214, 1968.

[Z-1] Zahn, C.T., "Graph-theoretical methods for detecting and describing Gestalt clusters", IEEE Trans. Computers, Vol. C-20, No. 1, pp. 68-86, 1971.

APPENDIX

Appendix A

This appendix discusses the motivation behind the use of the symmetrical and asymmetrical generalized distance measures referred to in Chapter IV and gives the details of the derivations leading to (4.3.3) thru (4.3.6).

I. Symmetric generalized distance:

A generalized second order distance metric, D^2, between two points X and Q in a p-dimensional space is

$$D^2 = (X - Q)^T[S] (X - Q) \qquad (A.1)$$

where $[S]$ is a $(p \times p)$ positive definite matrix. We seek a particular matrix $[S]$ which satisfies the condition that the average of the distances of the points in a cluster from their mode is 1. That is, a matrix $[S^{(i)}]$ is sought such that, for the i-th cluster,

$$\frac{1}{N_i} \sum_{j=1}^{N_i} D_j^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_j^{*(i)} - M^{(i)})^T[S^{(i)}](x_j^{*(i)} - M^{(i)}) = 1 . \qquad (A.2)$$

The existence of such a matrix is shown below:

Assume that the inverse of the matrix defined by

$$\sum_{j=1}^{N_i} (x_j^{*(i)} - M^{(i)}) (x_j^{*(i)} - M^{(i)})^T$$

exists. Then a matrix satisfying the condition (A.2) is given by

$$[S^{(i)}] = \frac{N_i}{p} \left[\sum_{j=1}^{N_i} (x_j^{*(i)} - M^{(i)})(x_j^{*(i)} - M^{(i)})^T \right]^{-1} . \qquad (A.3)$$

To prove this, we note,

$$\frac{1}{N_i} \sum_{j=1}^{N_i} D_j^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_j^{*(i)} - M^{(i)})^T [S^{(i)}] (x_j^{*(i)} - M^{(i)})$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \text{trace} \{ (x_j^{*(i)} - M^{(i)})^T [S^{(i)}] (x_j^{*(i)} - M^{(i)}) \}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \text{trace} \{ [S^{(i)}] (x_j^{*(i)} - M^{(i)})(x_j^{*(i)} - M^{(i)})^T \}$$

$$= \frac{1}{N_i} \text{trace} \{ \sum_{j=1}^{N_i} [S^{(i)}] (x_j^{*(i)} - M^{(i)})(x_j^{*(i)} - M^{(i)})^T \}$$

$$= \frac{1}{N_i} \text{trace} \{ [S^{(i)}] \sum_{j=1}^{N_i} (x_j^{*(i)} - M^{(i)})(x_j^{*(i)} - M^{(i)})^T \} .$$

Substituting for $[S^{(i)}]$ the expression given by (A.3)

$$\frac{1}{N_i} \sum_{j=1}^{N_i} D_j^2 = \frac{N_i}{N_i p} \text{trace} [I] = 1 .$$

The symmetric generalized distance measure between a pattern vector x and a mode $M^{(i)}$ of the i-th cluster is now defined as:

$$d_s(x, M^{(i)}) \triangleq (x - M^{(i)})^T [S^{(i)}] (x - M^{(i)}) . \qquad (A.4)$$

It is to be observed that $[S^{(i)}]$ is proportional to the inverse of the second moment matrix of the points in the i-th cluster, with respect to the mode, $M^{(i)}$, of the cluster and is a measure of the dispersion of that cluster.

The name "symmetric generalized distance" is used to distinguish it from the asymmetric measure discussed next.

II. Asymmetric generalized distance:

For defining the asymmetric generalized distance measure the ellipsoidal surfaces described by

$$(x - Q)^T [S] (x - Q) = 1$$

are centered at a new point which reflects the dispersion of the points in a cluster. The measures of the dispersion along each axis, both to the left and to the right of the mode $M^{(i)}$ are computed. The center of the ellipsoid is obtained by shifting $M^{(i)}$ by the average of these dispersions along each axis. To be more specific:

Let $x_{jk}^{*(i)}$ be the k-th component of the j-th pattern vector of the i-th cluster; $j = 1,2,\ldots,N_i$; $k = 1,2,\ldots,p$; and further let this component be designated

$$\ell x_{jk}^{*(i)} \quad \text{if} \quad x_{jk}^{*(i)} \leq M_k^{(i)} \tag{A.5}$$

and

$$r x_{jk}^{*(i)} \quad \text{if} \quad x_{jk}^{*(i)} > M_k^{(i)} \tag{A.6}$$

where $M_k^{(i)}$ is the k-th component of the mode vector $M^{(i)}$.

Also let $L^{(i)}$ and $R^{(i)}$ be the number of vectors from the i-th cluster satisfying the inequalities (A.5) and (A.6) respectively.

Define a measure of "spread to the left" and "spread to the right" along the k-th axis as:

$$\text{Spread to the left} \triangleq \ell\sigma_k^{(i)} \triangleq \left[\frac{1}{L^{(i)}} \sum_{n=1}^{L^{(i)}} (\ell x_{nk}^{*(i)} - M_k^{(i)})^2\right]^{\frac{1}{2}} \tag{A.7}$$

and

$$\text{spread to the right} \triangleq {}_r\sigma_k^{(i)}$$

$$\triangleq \left[\frac{1}{R^{(i)}} \sum_{n=1}^{R^{(i)}} ({}_r x_{nk}^{*(i)} - M_k^{(i)})^2\right]^{\frac{1}{2}} . \quad (A.8)$$

Associated with each estimated mode vector, $M^{(i)}$, define a new vector, $M_k'^{(i)}$, whose k-th component is given by

$$M_k'^{(i)} = M_k^{(i)} + [{}_\ell\sigma_k^{(i)} - {}_r\sigma_k^{(i)}]/2 \quad (A.9)$$

Referring to (A.1) we now seek a matrix $[S_{AS}^{(i)}]$ such that (A.2) holds except that instead of the vector $M^{(i)}$ for the i-th cluster, we use the new vector $M'^{(i)}$.

Proceeding as before a solution for $[S_{AS}^{(i)}]$ is

$$[S_{AS}^{(i)}] = \frac{N_i}{p} \left[\sum_{j=1}^{N_i} (x_j^{*(i)} - M'^{(i)})(x_j^{*(i)} - M'^{(i)})^T\right]^{-1} . \quad (A.10)$$

Consider the equation

$$(Y - M'^{(i)})^T [S_{AS}^{(i)}] (Y - M'^{(i)}) = 1 \quad (A.11)$$

which defines an ellipsoid centered at $M'^{(i)}$.

The unit of distance for computing the distance to modes is defined as the length of the vector from the mode $M^{(i)}$ to the surface of the ellipsoid, measured in the direction of the vector $(X - M^{(i)})$ where X is any point the distance to which from $M^{(i)}$ is required. It is assumed that $M^{(i)}$ lies interior to the ellipsoid. The distance measure is not valid if the assumption is not satisfied.

In two dimensions, the definition of "unit distance" is illustrated in Figure 6.

Denote the various vectors as in the figure,

$$\vec{a} = X - M'^{(i)} \; ; \; \vec{b} = M^{(i)} - M'^{(i)}$$

and $$\vec{c} = X - M^{(i)} .$$

In terms of the unit distance, $|c_1|$, the distance between $M^{(i)}$ and X, is given by the scalar, D, being the ratio,

$$d_{AS}(X, M^{(i)}) \triangleq D = \frac{|c|}{|c_1|} .$$

Let

$$[W] \triangleq [S_{AS}^{(i)}] .$$

We have

$$(X - M^{(i)}) = \vec{c} = (Y - M^{(i)}) \, D .$$

Substituting in (A.11) and arranging terms,

$$(X - M'^{(i)} \, D + M^{(i)} \, (D-1))^T [W] (X - M'^{(i)} \, D + M^{(i)} (D-1)) = D^2 \quad (A.12)$$

which is the equation of an ellipsoid centered at

$$D \, M'^{(i)} + M^{(i)} \, (1-D) .$$

It is to be noted that with a symmetric distribution, the mean and the mode are the same and hence $M'^{(i)} = M^{(i)}$ implying that the center of the ellipsoid is at the mode/mean.

Equation (A.12) is a quadratic in D and solving for D, we get

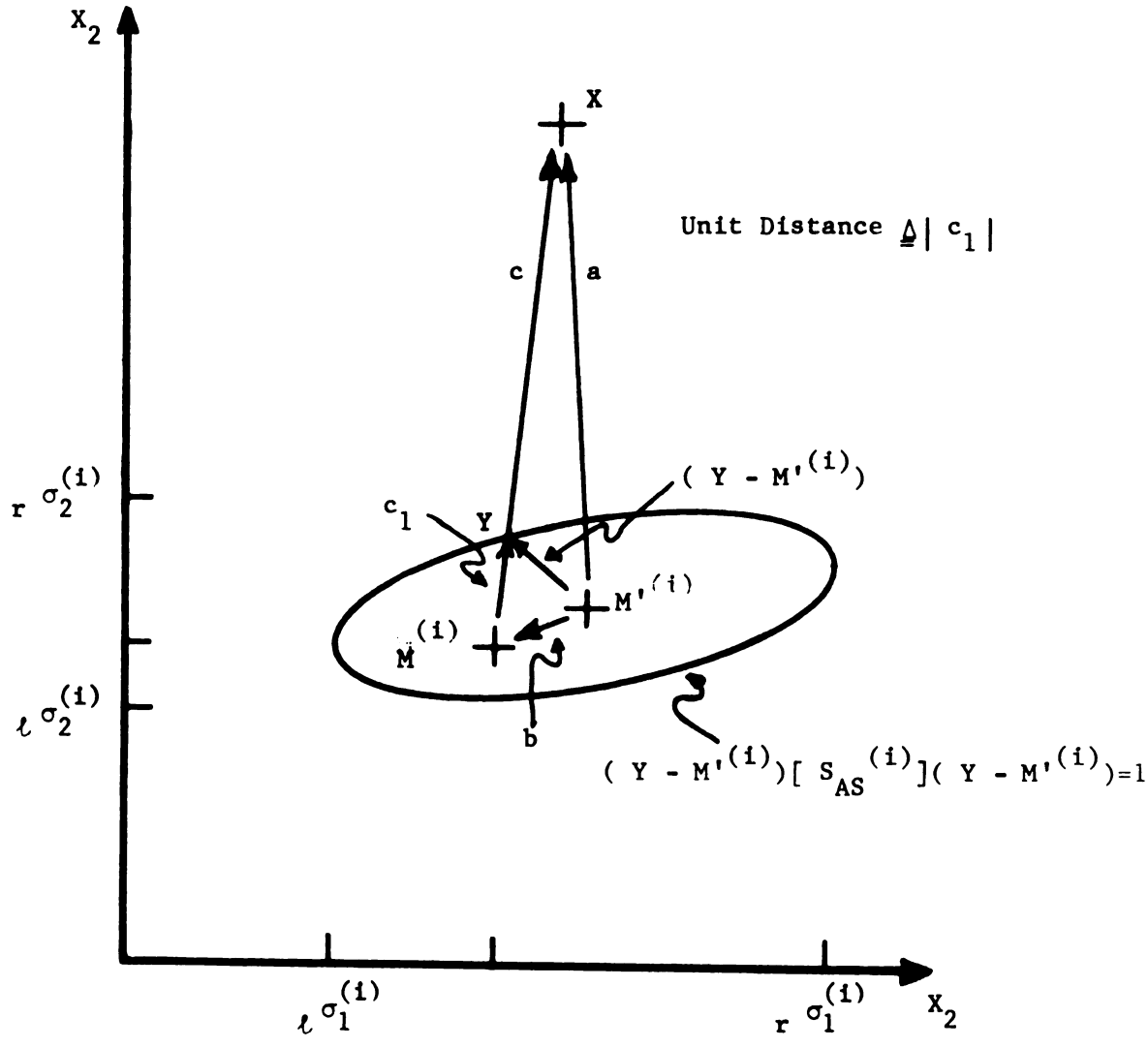


Figure 6. Explanation of the asymmetrical distance measure

$$D = \frac{q + \sqrt{q^2 + 4wr}}{2w}$$

where

$$q = b^t [W]a + a^t [W]b - 2b^t [W]b$$

$$w = 1 - b^t [W]b > 0$$

if the point $M^{(i)}$ is interior to the ellipsoid;

and $$r = (a-b)^t [W] (a-b) > 0$$

assuming $[W]$ to be positive definite.