1975

A survey of cluster analysis and its admissible procedures.

Michael D. Baillargeon
University of Windsor

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

A SURVEY OF CLUSTER ANALYSIS

AND

ITS ADMISSIBLE PROCEDURES

BY

M. D. BAILLARGEON

A THESIS

PRESENTED TO THE FACULTY OF GRADUATE STUDIES
THROUGH THE DEPARTMENT OF MATHEMATICS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

FACULTY OF GRADUATE STUDIES
THE UNIVERSITY OF WINDSOR
WINDSOR, ONTARIO

July, 1975

ABSTRACT

The techniques of cluster analysis are among the most widely used and little understood techniques for the statistical analysis of data. A mathematical basis for these techniques remains to be constructed. This paper describes the work done up to the present in the construction of such a basis. In the course of this discription a survey of the more common cluster analytic techniques is given and the admissibility of thes techniques under criteria suggested by G.N. Lance and W.T. Williams is given.

TABLE OF CONTENTS

A SURVEY OF CLUSTER ANALYSIS AND ITS ADMISSIBLE PROCEDURES

I: INTRODUCTION

We are, here, concerned with certain admissible techniques for the analysis of multivariate data, which attempt to solve the following problem:

> Given a sample of N objects or individuals, each of which is measured on each of p variables, devise a classification scheme for grouping the objects into g classes. The number of classes and the characteristics of the classes to be determined. [9]

The most commonly used term for techniques which seek to separate data into consituent groups is *cluster analysis*.

Ideal data for such an analysis would yield obvious clusters that, at least in small-scale cases, could be picked-out without the need for complicated mathematical techniques and without a precise definition of the term 'cluster'. In two or three dimensions the data could be examined visually and any clusters present identified.

In practice, however, things are not so simple and , consequently, there has been a great proliferation of clustering techniques. The importance of clustering techniques in such diverse fields as psychology, zoology, biology, botany, sociology, artifical intelligence and information retrival has added to the proliferation of clustering techniques.

This paper attacks the problem of choosing a clustering procedure from among the myriad proposed. Too often in practice, not enough is known about *a priori* conditions, the possible losses involved, etc.,

to determine a best procedure.

This type of perplexing problem is solved in decision theory by restricting attention to admissible decision rules. This approach eliminates obvious bad rules, though additional information often is necessary to select, from among the admissible rules, the best. The suggestion is, therefore, to formulate some optimum properties that a reasonable procedure should satisfy and call a procedure satisfying them admissible. Obviously, requiring admissibility eliminates only bad clustering procedures but does not attempt to determine the best method.

Let A denote some property which would be satisfied by any reasonable clustering procedure either in general or when used in a special application. Any procedure which satisfies A is called A-admissible.

The attempt to present a formal structure within which admissible clustering procedures may be identified and studied leads us to consider in some detail some of the more common clustering procedures and their attendant problems.

We restate the original problem in order to present the original model upon which all that follows is based.

> Let Y be a finite subset of R^p with N elements. The problem is now to determine a covering of Y which has some optimum properties (admissibility) for the purpose of classification. In some clustering techniques not all members of the cover are considered.

II: THE ASPECTS OF CLUSTER ANALYSIS

The ideal strategy for the development of a data simplification technique is as follows:

First, formulate a precise mathematical characterization of the data and the kind of representation desired so that the simplification methods may be treated as transformations from a structure of one kind to that of another.

Next, formulate criteria of adequacy of the operations including invariant and covariant representation, preservation of structure and optimality conditions.

Then the existence of appropriate methods, their uniqueness and mathematical properties must be ascertained.

Finally, efficient algorithms must be found. The existence and uniqueness of appropriate methods may be unconstructive or may specify a computationaly unfeasible algorithm.

Unfortunately, the development of clustering methods has followed an inverse sequence. In the 1940's and 1950's a variety of clustering alorithms were proposed. Over the years it was gradually realized that some superfically different methods implement the same method, and that different methods differ widely in their properties. The recent work of Jardine and Sibson [16,17] and Lance and Williams [23] has resulted in the construction of a general theory within which such methods may be analyzed.

The general theory assumes that it is reasonable to seek clusters in data, these being subsets of Y characterized by the possession of the properties of coherence and isolation. This assumption and experience

lead to the following dichotomies for clustering methods:

Clustering procedures can be divided, immediately, into two classes:

i) those methods which work directly with the data.

ii) those methods which work with a dissimilarity coefficient.

In what what follows, we will investigate both of these classes.

We also have the following division of clustering methods into disjoint classes:

1. Sequential versus Simultaneous Methods.

Most clustering methods are sequential. A recursive sequence of operations is applied to the set Y producing a sequence of covers for Y. Simultaneous techniques are rare and only one example will be given.

2. Agglomerative versus Divisive Methods.

Staring with a cover of t sets, agglomerative techniques group these into covers having successively fewer sets, arriving eventually at a single set containing all the elements of Y. By contrast, divisive techniques commence with a cover of t sets, subdividing these sets into covers having successively more sets, arriving eventually at a cover of N sets.

3. Nonoverlapping versus Overlapping Methods.

In a nonoverlapping method, any two members of a cover are disjoint.

The goals of a cluster analysis may be approached by several methods. The availability of sufficient information makes it desirable to formulate explicit criteria for comparing solutions. Computational difficulties

may prevent the finding of a global optimization method which finds an optimum and may force the use of a local optimization method which finds an optimum over some restricted class of solutions even in the presence of some criterion. The end result may not have any optimality properties that can be stated.

III: THE STRUCTURE OF CLUSTER ANALYSIS

In order to formalize the concept of clustering with a dissimilarity coefficient, we need the following notation. In what follows:

$\Sigma(P)$ will denote the set of all reflexive, symmetric relations on a set P.

$E(P)$ will denote the set of all equivalence relations on a set p.

From graph theory, we have the following definition :

Definition 1: Let $r \in \Sigma(P)$. A maximal linked set for r, denoted ML(r), is a set $S \subseteq P$ such that

i) $S \times S \subseteq r$;

ii) $(\forall a \notin S)(\exists b \in S)\ (a,b) \notin r$.

It is convienient to think of this in grahical terms. If we draw a graph whose vertices are the elements of P, and whose edges link just those pairs of elements of P contained in r, then the ML-sets are the vertex sets of maximal complete subgraphs. ML-sets have appeared in literature [21] under various other names:

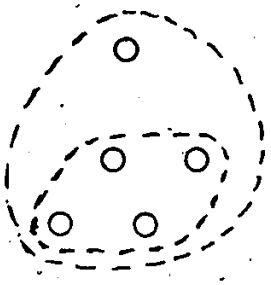
Kuhn's clumps,

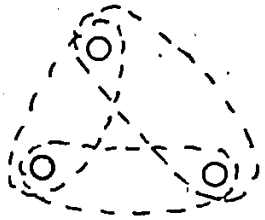
cliques,

maximal cliques,

maximal complete subgraphs.

It should be noted that the definition of an ML-set implies that its elements form a coherent unit isolated from other elements in P. This property will be used later to supply a definition for a cluster.





Systems of clusters such as those shown above will not be admissible within this framework.

Definition 2: Let $Y \subseteq R^p$, $0 < |Y| = N < \infty$. A dissimilarity coefficient, DC, on Y is a function $d: Y \times Y \to R$ such that

i) $(\forall a,b \in Y)\ d(a,b) \geq 0$,

ii) $d(a,a) = 0$,

iii) $d(a,b) = d(b,a)$

Definition 3: A DC is said to be even if

$(\forall a,b,c \in Y)\ d(a,b) = 0 \to d(a,c) = d(b,c)$.

Definition 4: A DC is said to be definite if

$d(a,b) \to a = b$.

Definition 5: A DC is said to satisfy the ultrametric inequality if

$(\forall a,b,c \in Y)\ d(a,b) \leq \max\{ d(a,c), d(b,c)\}$.

Notation:

$C(Y)$ = the set of all DC's on Y.

$C'(Y)$ = the set of all definite DC's on Y.

$M(Y)$ = the set of all DC's on Y satisfying the triangle inequality.

$U(Y)$ = the set of all DC's on Y satisfying definition 5.

$M'(Y) = M(Y) \cap C'(Y)$ = the set of all metrics on Y.

$U'(Y) = U(Y) \cap C'(Y)$ = the set of all ultrametrics on Y.

Definition 6: Let $d,d' \in C(Y)$. Then d is said to be dominated by d', denoted $d' >> d$, if

$$(\forall a,b \in Y)\ d(a,b) \leq d'(a,b).$$

Definition 7: A set $X \subseteq C(Y)$ is said to be bounded if

$$(\forall d' \in C(Y))(\exists d \in X)\ d < d'.$$

We define $\sup X$ by $\sup X(a,b) = \sup\{ d(a,b) \mid d \in X\}$.

A bounded set $X \subseteq C(Y)$ will also be called sup-closed.

Definition 8: A numerically agglomerative clustering (NSAC) is a function $c:[0,\infty) \to \Sigma(Y)$ such that

i) $(\forall h,h' \in [0,\infty))(0 \leq h \leq h')\ c(h) \subseteq c(h')$,

ii) $c(h)$ is eventually $Y \times Y$,

iii) $(\forall h \geq 0)(\exists \delta > 0)\ c(h+\delta) = c(h)$.

Definition 9: A NSAC is said to be definite if $c(0)$ is the equality relation on Y.

Definition 10: A numerically stratified divise clustering (NSDC) is a function $c:[0,\infty) \to \Sigma(Y)$ such that

i) $(\forall h,h' \in [0,\infty))(0 \leq h \leq h')\ c(h) \supseteq c(h')$,

ii) $c(h)$ is eventually the equality relation on Y,

iii) $(\forall h \geq 0)(\exists \delta > 0)\ c(h+\delta) = c(h)$.

Definition 11: AN NSDC is called definite if $c(0) = Y\ Y$.

Definition 12: An NSAC (NSDC) is called hierarchical if

$$(\forall h \geq 0)\ c(h) \in E(Y).$$

Definition 13: A definite hierarchical NSAC (NSDC) is called a dendogram.

Definition 14: Let c be an NSAC(NSDC). A cluster for c at level h is the elements of $ML(c(h))$.

Theorem 1: There exists a 1-1, onto, correspondence between the set C(Y) and the set of all NSAC's on Y (denoted NSAC(Y)).

Proof: Let T be defined by

$$(Td)h = \{ (a,b) \mid d(a,b) \leq h \} \qquad (d \in C(Y), h \in [0,\infty)) ;$$

$$(T^{-1}c)(a,b) = \inf \{ h \mid (a,b) \in c(h) \} \qquad (c \in NSAC(Y), (a,b) \in Y \times Y).$$

It is easy to show Td is a NSAC, $T^{-1}c \in C(Y)$ and T is one-to-one. To show that T is onto requires the use of iii) in definition 6. This result is due to Jardine and Sibson [16,17].

Theorem 2: There exists a 1-1, onto, correspondence between the set C(Y) and the set of all NSDC's on Y.

Proof: Let T be defined by

$$(Td)h = \{ (a,b) \mid d(a,b) \leq 1/h \} \qquad (d \in C(Y), h \in (0,\infty)) ;$$

$$(T^{-1}c)(a,b) = \sup \{ h \mid (a,b) \in c(1/h) \} \qquad (h \in (0,\infty), c \in NSDC(Y)).$$

Corollary: There exists a 1-1, onto, correspondence between the NSAC's and the NSDC's on Y.

This corollary explains why the division between the agglomerative and devisive clustering techniques is artifical. More agglomerative algorithms exist because they are easier to implement.

T induces a 1-1 correspondence between C'(Y) and the definite NSAC's (NSDC's) and between U(Y) and the hierarchical NSAC's (NSDC's). The possibility of identifying the NSAC's (NSDC's) on Y with the DC's on Y is fundamental to all that follows. If we select some subset of the NSAC's as the kind of output we would like a cluster method to have, then this set will correspond to some subset $Z \subset C(Y)$. Hence a cluster method may be regarded as a function $D: C(Y) \to Z$. Mathematically this is an agreeable situation, attention can be focused to objects of one kind and

constraints imposed on a cluster method can be formulated neatly. The output from a cluster method is a list of elements of ML(c(h)) for each splitting level of h of c. A splitting level is a level h' such that $c(h) \neq c(h')$ if $h < h'$; h=0 is a splitting level if and only if the NSAC is definite.

Definition 15: With each chosen Z one method $D:C(Y) \to Z$ will be associated, namely the method for which D(d) is the maximal element of Z dominated by d - the Z-subdominant of d. Such methods will be called subdominant methods.

IV: ADMISSIBILITY CONDITIONS

The first condition would be desirable in almost any situation. It has been studied elsewhere where it is called the comparison property [11]. It is based on a 1-1 transformation of the data.

Let $C_1=\{x_1,\ldots,x_{j_1}\},\ldots,C_k=\{x_{j_{k-1}},\ldots,x_N\}$ be a clustering of Y. Let $y_1,\ldots,y_N$ be any reordering of the points of Y and define

$$C_1'=\{y_1,\ldots,y_{j_1}\},\ldots,C_k'=\{x_{j_{k-1}},\ldots,y_N\}.$$

We call $C_1',\ldots,C_k'$ an image of $C_1,\ldots,C_k$.

A clustering $C_1,\ldots,C_k$ is said to be *image admissible* if it does not have an image which is uniformly better in the sense that

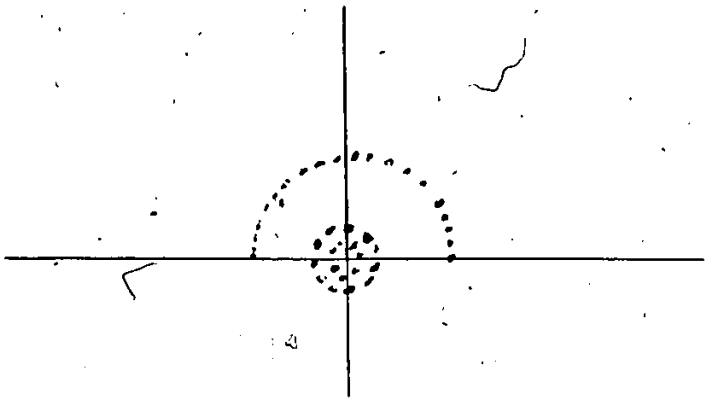
i) $d(x_i,x_j)\geq d(y_i,y_j)$ when the i'th and j'th points are in the same cluster,

ii) $d(x_i,x_j)\leq d(y_i,y_j)$ when the i'th and j'th points are in different clusters,

with strict inequality holding for at least one pair of indices.

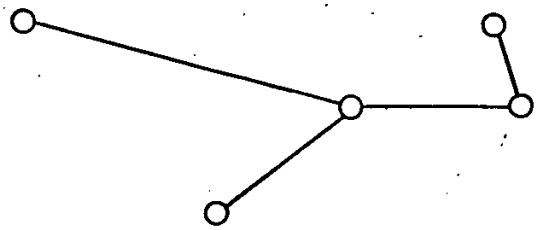
A second form of admissibility arose from the feeling that one cluster should not 'cut through another'. One way to prevent such clusterings is to require convex admissibility. If the set Y resides in a linear space then a clustering $C_1,\ldots,C_k$ is said to be *convex admissible* if the k convex hulls of $C_1,\ldots,C_k$ do not intersect.

Convex admissibility, while useful in many situations, does not seem very universal since it eliminates many reasonable clusterings. For example, let C_1 be a cluster in the plane tightly grouped on the perimeter of the entire half circle with radius 4 and center (0,0) lying above the x-axis and let C_2 be a cluster covering the disk with radius 1 about (0,0) as in the following diagram:



Then the clustering C_1, C_2 is not convex admissible but appears reasonable.

Given any set of points $A \subset R^2$ we define the linkage L_A of A as follows: perform a nearest neighbour clustering on A (the exact technigue will be described latter). As two sets are grouped draw a straight line connecting any pair of points one in the first set and the other in the second, which are closest. The linkage L_A is the network of lines formed when all of A is so connected. An example of a linkage is given by the following diagram:



For the case $Y \subset R^2$ a clustering $C_1, \ldots, C_k$ is called *connected admissible* if $L_{C_1}, \ldots, L_{C_k}$ are pairwise disjoint.

Clustering in R^2 is very common in multivariate analysis, especially due to the practice of reducing many dimensions to two dimensions to get

approximate scatter diagrams. As with all admissibility conditions, connected admissibility prevents a certain type of bad behaviour but is by no means a panacea.

There are several ways of defining well-structured data [11] but in general the data is so arranged as to make the correct clustering obvious. A procedure is well-structured admissible if it gives the correct clustering whenever it is confronted with well-structured data.

A clustering method is said to be *well-structured (exact tree) admissible* if it has an exact tree structure; i.e. if one can reconstruct the dissimilarity matrix of the original data from knowledge of the tree (NSAC) alone.

Theorem 3: A necessary and sufficient condition that a cluster method be well-structured (exact tree) admissible is that the dissimilarity coefficient satisfy the ultrametric inequality. (This result is claimed by Johnson [18] without formal proof).

Proof: First note that the definition of well-structured (exact tree) admissibility requires that the cluster method produce a NSAC (tree) whose splitting levels (nodes) are a known function of the original dissimilarities.

Let $d \in U(Y)$ it is then easy to see how such a tree can be constructed.

Let c be such an NSAC. Then its splitting levels $h_1 < h_2 < \ldots < h_k$ are a known function, f say, of the original distances. Also we have

$$h_i < h_j \rightarrow d_i < d_j,$$

so that f is strictly increasing.

Then $f^{-1}(T^{-1}c)(a,b)=f^{-1}[\inf\{ h \mid (a,b)\ c(h)\}]$ is its corresponding DC. Use of i) in definition 6 shows that $f^{-1}(T^{-1}c)$ satisfies the ultrametric inequality.

A cluster method is said to be *well-structured (k-group) admissible* if there exists a clustering $C_1,\ldots,C_k$ such that all within-cluster distances are smaller then all between-cluster distances.

A cluster method is said to be *well-structured (perfect) admissible* if the clustering is such that the dissimilarity between any two objects in the same group is the same value s_1 and the dissimilarity between two objects in different groups is the same value s_2 $(s_1 < s_2)$.

We now turn to more specialized conditions which are not generally useful but which are important in several applications. Among these are the various forms of proportion admissibility.

Proportion admissibility attempts to give an analytical description of the idea important in some applications that the geometrical aspects of the clusters are more important than the density of points in the clusters.

A procedure is said to be *point proportion admissible* if after we duplicate one or more points any number of times and reapply the procedure the boundaries of the clusters in Y are not changed.

A procedure giving clusters $C_1,\ldots,C_k$ at some stage is said to be *cluster proportion admissible* at that stage if after duplicating each cluster an arbitrary number of times; i.e. each point within the same cluster is duplicated the same number of times, it yields clusters having the same boundaries at that stage.

If a clustering results in k clusters, and all points in any one of these clusters, say C_j, are removed from Y the procedure is *cluster omission admissible* if when applied to the subset $M-C_j$ to get k-1 clusters it yields the original clusters except for C_j.

In many applications it is hard to assign a scale to variables and the only relevant information in the data is the order of the variables or ranks. For such problems it would be comforting to know that monotone transformations of the dissimilarities do not change the clustering. In other words, one can arbitrarily assign distances as long as their rankings are preserved.

A procedure is *monotone admissible* if a montone transformation applied to each element of the dissimilarity matrix does not change the resulting clustering. By a monotone transformation we mean a strictly increasing function f with $f(0)=0$.

Objects should be clustered in response to some observed structure in the data and not be clustered unnecessarily. Thatis, the resultant NSAC, or DC in Z, should in some sense be the best fit to the original data. With this in mind a clustering method DC is *optimal admissible* if

$$d^*\varepsilon Z, D(d) \leq d' \leq d \rightarrow d' = D(d)$$

where

$$D(d) = \sup\{ d' \mid d'\varepsilon Z, d' << d\}.$$

This ensures that D(d) is a local optimum. To ensure global optimality the following condition is imposed on Z;

if X is a bounded subset of Z, then the DC, $\sup X \varepsilon Z$.

Theorem 4: Monotone and optimal admissibility imply that $Z-\{0\}$ is path connected.

Proof: Let $d', d \in Z$ and let

$$d_x(a,b) = \max\{ xd(a,b), (1-x)d'(a,b)\} \qquad (x \in [0,1]).$$

Then $d_0 = d'$, $d_1 = d$. By monotone admissibility xd and $(1-x)d'$ are in Z for all $x \in (0,1)$. By optimal admissibility then,

$$(x \in (1,0)) \cdot d_x \in Z,$$

$$(x \in (1,0))\ d, d' \neq 0 \rightarrow d_x \neq 0,$$

and since the function sending x to d_x is continuous, it is a path in $Z-\{0\}$ from d to d'.

V: THE ADMISSIBILITY OF SOME CLUSTERING PROCEDURES

5.1: Lance and Williams combinatorial clustering procedures. [23]

The algorithm for these techniques is as follows:

i) fuse those groups having the smallest dissimilarity coefficients.

ii) calculate

$$d_{k(ij)}=\alpha_i d_{ki}+\alpha_j d_{kj}+\beta d_{ij}+\gamma |d_{ki}-d_{kj}|,$$

where d_{ij} is the distance between groups i and j and α,β and γ are parameters whose values determine which particular combinatorial technique is to be employeed.

iii) go to i).

The effect of the algorithm is to generate a sequence $D_0,D_1,\ldots,D_k$ of dissimilarity matrices of decreasing order.

The following constraints can be imposed on α,β and γ to ensure that $d_{k(ij)}$ is a DC and to determine the type of ML-sets resulting from the procedure:

$$\alpha_1=\alpha_2 \; ; \; \beta<1 \; ; \; \gamma=0$$

$$\alpha_1=\alpha_2 \; ; \; \alpha_1+\alpha_2+\beta=1 \; ; \; \gamma=0$$

The following general admissibility conditions hold for all combinatorial techniques:

A) Combinatorial clustering procedures are image admissible.

Proof: Let $C_1,\ldots,C_k$ be a clustering resulting from a combinatorial procedure. Let its splitting level be $h=h_s$. Let $x_1,\ldots,x_N$ be the objects to be clustered and let $d(x_i,x_j)$ be the dissimilarity between x_i and x_j. Suppose f is a 1-1 map of $x_1,\ldots,x_N$ onto itself such that

i) if x_i and x_j are in the same C_k, then

$$d\{f(x_i),f(x_j)\} \leq d(y_i,y_j) \ ;$$

ii) if x_i and x_j are in different clusters, then

$$d\{f(x_i),f(x_j)\} \geq d(y_i,y_j).$$

To show image admissibility, we need to show that all the inequalities in i) and ii) are equalities.

Let the splitting levels of h up to h be listed as

$$h_0 < h_1 < \ldots < h_s = h.$$

We show by induction on the levels of h_i that if at level h_i the clusters are $C_1(i),\ldots,C_{k_i}(i)$ then

$$\{f(C_1(i)),\ldots,f(C_{k_i}(i))\} = \{C_1(i),\ldots,C_{k_i}\}.$$

For h_0 this is clear. Suppose the result is true for h_i $(i < s)$ By i) for two clusters $C_e(i),C_j(i)$ if $d_{ej} \leq h_{i+1}$, then

$$d_{f(C_e),f(C_j)} \leq h_{i+1}$$

since $C_e(i)$ and $C_j(i)$ belong to the same cluster at level h_{i+1}. This proves that $f(C_e)$ and $f(C_j)$ are clustered together at level h_{i+1}; i.e. clusters at level h_{i+1} map onto themselves. This result holds at level s.

From i) and the fact that

$$\sum_{j=1}^{k(s)} \sum_{\substack{\text{pairs} \\ \text{in} \\ C_j}} d(x_i,x_j) = \sum_{j=1}^{k(s)} \sum_{\substack{\text{pairs} \\ \text{in} \\ C_j}} [d[f(x_i),f(x_j)]]$$

we see that $d(x_i,x_j)=d[f(x_i),f(x_j)]$. A similar agruement holds for x_i and x_j in different clusters by using ii), thus completing the proof.

B) Combinatorial clustering procedures are well-structured (k-group) admissible.

Proof: Let $C_1,\ldots,C_k$ be a clustering resulting from a combinatorial procedure. Let its splitting level be $h=h_s$, and let the splitting levels of h up to h be listed as

$$h_0 < h_1 < \ldots < h_s = h.$$

We show by induction on the levels of h_i that if at level h_i the clusters are $C_1(i),\ldots,C_{k_i}(i)$ then

$$d(x_k,x_l) \leq d_{C_e,C_f}$$

where x_k and x_l are both in cluster $C_h(i)$, say, and $C_e(i)$, $C_f(i)$ are distinct clusters.

For h_0, we have that $(\forall i,j)\ d_{C_i,C_j} \geq h_0$ and if $d(x_i,x_j)=h_0$,

x_i and x_j are fused into cluster $C_1(0)$, say, and the result holds.

Suppose the result holds for h_i $(i < s)$. Then two clusters C_e, C_j are fused if $d_{ej} = h_{i+1}$ and $h_i < h_{i+1}$ by the definition of splitting levels.

Also h_{i+1} is the minimum distance between all pairs of groups at level h_{i+1}; i.e., necessarily all within-cluster distances are less than all between-cluster distances as required.

The following results appear obvious:

C) Combinatorial procedures are cluster proportion admissible.

D) Combinatorial procedures are cluster omission admissible.

E) Combinatorial procedures are not convex admissible.

This is a conjecture. A general counter-example to convex admissibility remains to be determined. Counter-examples for particular combinatorial procedures are presented in what follows.

5.2: Some important particular combinatorial procedures.

5.2.1: Nearest Neighbour

Groups initially consist of single objects which are fused according to the distance between their nearest members, the groups with the smallest distances being fused. Each fusion decreases by one the number of groups. In this case, then, the distance between their closest members.

Accordingly, in the general algorithm:

$$\alpha_i=\alpha_j=\tfrac{1}{2} \; ; \; \beta=0 \; ; \; \gamma=-\tfrac{1}{2}$$

$$\text{or} \quad d_{k(ij)}=\min(d_{ki},d_{kj}).$$

This is, also, called the single-link or minimum method.

5.2.2: Furthest Neighbour

This method is exactly the opposite of the nearest neighbour method, in that the distance between groups is, here, defined as the distance between their most remote pair of members.

Accordingly,

$$\alpha_i=\alpha_j=\tfrac{1}{2} \; ; \; \beta=0 \; ; \; \gamma=\tfrac{1}{2}$$

so that

$$d_{k(ij)}=\max(d_{ki},d_{kj}).$$

This is also called the complete linkage or maximum method.

5.2.3: Centroid

Here, clusters are depicted to lie in Euclidean space, and are replaced when formed by the coordinates of their centroids. The distance between groups is defined to be the distance between group centroids. The method is then to fuse groups according to this distance, groups with the smallest distances between them being fused first.

Accordingly,

$$\alpha_i=n_i/(n_i+n_j) \quad ; \quad \alpha_j=n_j/(n_i+n_j)$$

$$\beta=-\alpha_i\alpha_j \quad ; \quad \gamma=0$$

5.2.4: Median

This technique was developed to overcome a disadvantage of the centroid method where if the sizes of the two groups to be fused are very different the centroid of the new group will be very close to that of the larger group and may remain within that group: the characteristics of the smaller group are then virtually lost. The proposed strategy is based on the following: If we represent the centroids of the groups to be fused by (i) and (j), then the distance of the centroid of the third group (h) from the group formed by the fusion of (i) and (j) lies along the median of the triangle defined by (i), (j) and (h).

Geometrical considerations lead to the following constraints:

$$\alpha_i = \alpha_j = \tfrac{1}{2} \quad ; \quad \beta = -\tfrac{1}{4} \quad ; \quad \gamma = 0$$

5.2.5: Group Average

This method defines the distance between groups as the average of the distances between all pairs of objects in the two groups and can be used provided the concept of an average measure is acceptable. In this context Lance and Williams [23] point out that the concept of an average correlation coefficient is not entirely acceptable, and suggest a more satisfactory method might be achieved by setting

$$d_{ij} = \cos\left(\frac{1}{n_i n_j} \sum_{i,j} \cos^{-1} s_{ij} \right) ,$$

where d_{ij} is the distance between groups i and j, n_i and n_j are the numbers in these groups, and s_{ij} represents some inter-object measure.

The constraints for the Group Average method are

$$\alpha_i = n_i/(n_i+n_j) \quad ; \quad \alpha_j = n_j/(n_i+n_j)$$

$$\beta = \gamma = 0$$

5.2.6: Ward's Method

At any stage of an analysis the loss of information which results from the grouping of individuals into clusters can be measured by the total sum of squared deviations of every point from the mean of the cluster to which it belongs. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined.

$$\alpha_i = (n_i+n_k)/(n_i+n_j+n_k) \ ;$$

$$\alpha_j = (n_j+n_k)/(n_i+n_j+n_k) \ ;$$

$$\beta = -n_k/(n_i+n_j+n_k) \ ;$$

$$\gamma = 0$$

5.2.7: An example of combinatorial clustering

We will consider the clustering of the points given below. These coordinates correspond to points regularly spaced in an H-configuation, having had a small random noise added to each coordinate. The dissimilarity matrix of Euclidean distances among these points was used as input to the above six combinatorial techniques. This matrix is given in Appendix II.

I	X(I)	Y(I)	I	X(I)	Y(I)
1	2.964	0.489	19	-3.981	3.512
2	-2.957	-2.529	20	2.976	2.482
3	4.039	3.474	21	-2.965	2.490
4	-4.017	-2.503	22	-3.993	2.452
5	1.021	0.527	23	4.011	-0.467
6	-3.018	-3.486	24	3.962	2.458
7	-0.022	0.480	25	1.986	0.492
8	4.027	-2.550	26	-3.991	-0.518
9	-4.049	1.469	27	4.044	1.471
10	4.026	-1.509	28	-3.012	1.526

I	X(I)	Y(I)	I	X(I)	Y(I)
11	-4.014	-3.541	29	-3.012	-1.474
12	-2.005	0.495	30	3.039	3.546
13	2.973	-0.509	31	-0.994	0.458
14	3.045	1.493	32	-3.012	3.464
15	-3.965	0.518	33	-2.975	-0.549
16	2.951	-2.492	34	2.985	-1.494
17	3.997	-3.491	35	-2.952	0.486
18	3.003	-3.530	36	4.016	0.545
			37	-3.984	-1.505

The next three pages consist of the resulting clusterings at each stage oa a Nearest neighbour analysis (clusters of single points are ommited). Page 28 contains what is usually called in literature a dendogram. In actuality it is a graphical representation of a dendogram constructed to aid in interpretability. The presence of a few points located so as to form a bridge between the arms of the H results in the production of two large elongated clusters. This behaviour is often called the 'chaining effect', and is sometimes considered to be a defect of this method. In the case where the clusters are compact and well separated, the obvious clusters are found without evidence of this effect. In some other cases, to the extent that the results are very sensitive to noise or to slight changes in the position of the data points, this is certainly a valid criticism of the method. However, this very tendency to form chains can be advantageous if the clusters are elongated or posses elongated limbs.

NEAREST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
1	0.429	(29,33)	36
2	0.430	(27,36),(29,33)	35
3	0.444	(8,17),(27,36),(29,33)	34
4	0.449	(8,17),(12,35),(27,36),(29,33)	33
5	0.456	(8,17),(9,15),(12,35),(27,36),(29,33),	32
6	0.460	(2,6),(8,17),(9,15),(12,35),(27,36),(29,33)	31
7	0.466	(2,6),(8,17),(9,15),(12,35),(21,28),(27,36), (29,33)	30
8	0.467	(2,6),(5,25),(8,17),(9,15),(12,35),(21,28), (27,36),(29,33)	29
9	0.471	(2,6),(5,25),(8,17),(9,15),(12,35),(19,32), (21,28),(27,36),(29,33)	28
10	0.473	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (19,32),(21,28),(27,36),(29,33)	27
11	0.473	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (19,32),(21,28),(27,36),(29,33,37)	26
12	0.476	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (19,21,28,32),(27,36),(29,33,37)	25
13	0.479	(1,5,25),(2,6),(7,31),(8,17),(9,15),(12,35), (19,21,28,32),(27,36),(29,33,37)	24
14	0.485	(1,5,25),(2,6),(7,31),(8,17),(9,15,22), (12,35),(19,21,28,32),(27,36),(29,33,37)	23
15	0.486	(1,5,25),(2,6),(7,31),(8,17),(9,15,22), (12,35),(13,34),(19,21,28,32),(27,36), (29,33,37)	22

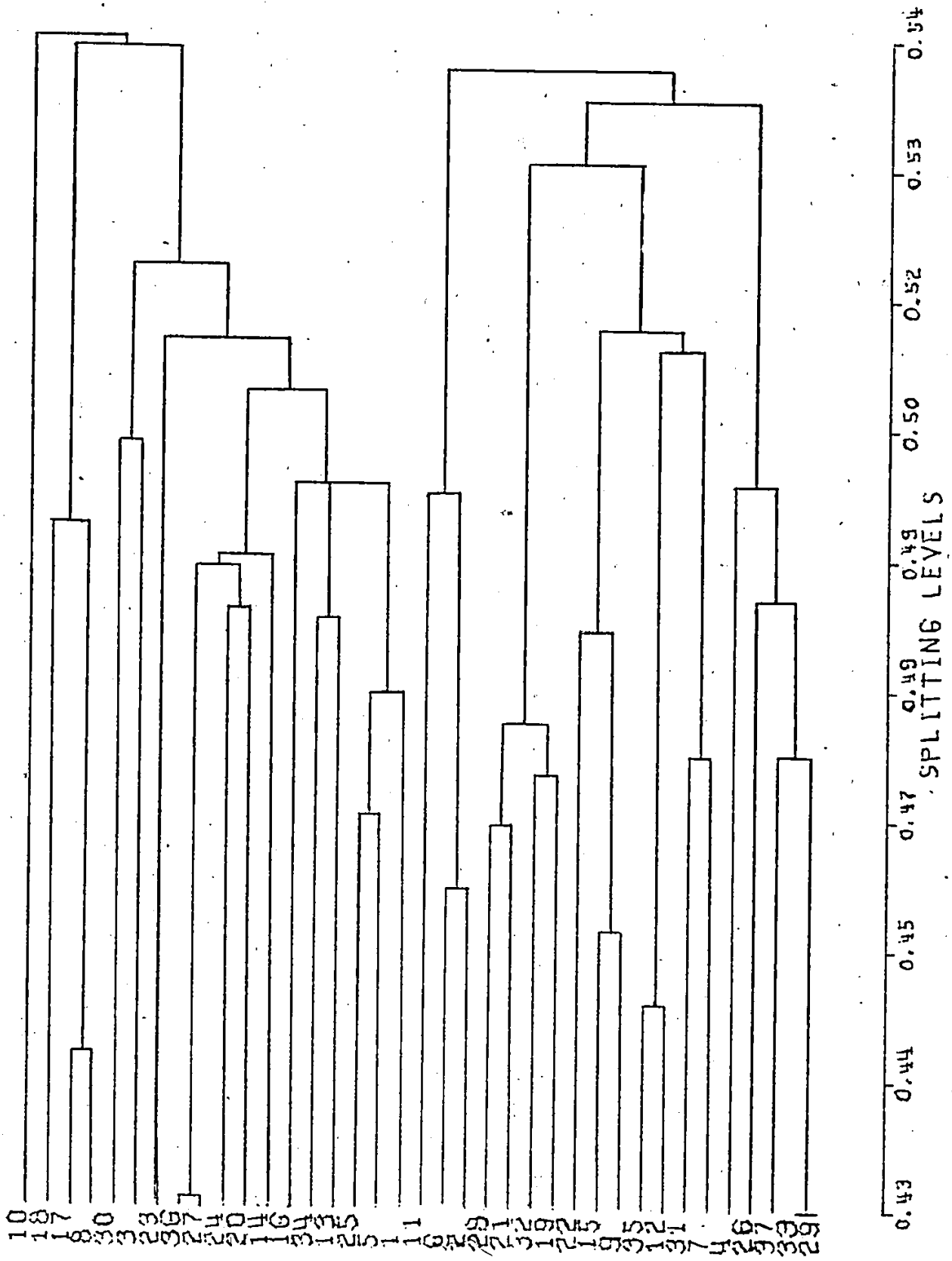
NEAREST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
16	0.487	(1,5,25),(2,6),(7,31),(8,17),(9,15,22),(12,35),(13,34),(19,21,28,32),(20,24),(27,36),(29,33,37)	21
17	0.488	(1,5,25),(2,6),(7,31),(8,17),(9,15,22),(12,35),(13,34),(19,21,28,32),(20,24),(26,29,33,37),(27,36)	20
18	0.491	(1,5,25),(2,6),(7,31),(8,17),(9,15,22),(12,35),(13,34),(19,21,28,32),(20,24,27,36),(26,29,33,37)	19
19	0.492	(1,5,25),(2,6),(7,31),(8,17),(9,15,22),(12,35),(13,34),(14,20,24,27,36),(19,21,28,32),(26,29,33,37)	18
20	0.495	(1,5,25),(2,6),(7,31),(8,17,18),(9,15,22),(12,35),(13,34),(14,20,24,27,36),(19,21,28,32),(26,29,33,37)	17
21	0.498	(1,5,25),(2,6,11),(7,31),(8,17,18),(9,15,22),(12,35),(13,34),(14,20,24,27,36),(19,21,28,32),(26,29,33,37)	16
22	0.499	(1,5,25,13,34),(2,6,11),(7,31),(8,17,18),(9,15,22),(12,35),(14,20,24,27,36),(19,21,28,32),(26,29,33,37)	15
23	0.499	(1,5,13,25,34),(2,6,11),(4,26,29,33,37),(7,31),(8,17,18),(9,15,22),(12,35),(14,20,24,27,36),(19,21,28,32)	14
24	0.499	(1,5,13,16,25,34),(2,6,11),(4,26,29,33,37),(7,31),(8,17,18),(9,15,22),(12,35),(14,20,24,27,36),(19,21,28,32)	13
25	0.503	(1,5,13,16,25,34),(2,6,11),(3,30),(4,26,29,33,37),(7,31),(8,17,18),(9,15,22),(12,35),(14,20,24,27,36),(19,21,28,32)	12

NEAREST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
26	0.508	(1,5,13,14,16,25,20,24,27,36,34), (2,6,11),(3,30),(4,26,29,33,37), (7,31),(8,17,18),(9,15,22),(12,35), (19,21,28,32)	11
27	0.512	(1,5,13,14,16,20,24,25,27,34,36), (2,6,11),(3,30),(4,26,29,33,37), (7,12,31,35),(8,17,18),(9,15,22), (19,21,28,32)	10
28	0.513	(1,5,13,14,16,20,23,24,25,27,34,36), (2,6,11),(3,30),(4,26,29,33,37), (7,12,31,35),(8,17,18),(9,15,22), (19,21,28,32)	9
29	0.514	(1,5,13,14,16,20,23,24,25,27,34,36), (2,6,11),(3,30),(4,26,29,33,37), (7,9,12,15,22,31,35),(8,17,18), (19,21,28,32)	8
30	0.520	(1,3,5,13,14,16,20,23,24,25,27,30,34,36), (2,6,11),(4,26,29,33,37), (7,9,12,15,22,31,35),(8,17,18), (19,21,28,32)	7
31	0.530	(1,3,5,13,14,16,20,23,24,25,27,30,34,36), (2,6,11),(4,26,29,33,37), (7,9,12,15,19,22,28,31,32,35),(8,17,18)	6
32	0.536	(1,3,5,13,14,16,20,23,24,25,27,30,34,36), (2,6,11),(4,7,9,12,15,19,22,26,28,29,31, 32,33,35,37),(8,17,18)	5
33	0.539	(1,3,5,13,14,16,20,23,24,25,27,30,34,36), (2,4,6,9,11,12,15,19,22,26,28,29,31,32,33, 35,37),(8,17,18)	4
34	0.541	(1,3,5,8,13,14,16,17,18,20,23,24,25,27,30, 34,36),(2,4,6,9,11,12,15,19,22,26,28,29,31, 32,33,35,37)	3
35	0.542	(1,3,5,8,13,14,16,17,18,10,20,23,24,25,27,30, 34,36),(2,4,6,9,11,12,15,19,22,26,28,29,31, 32,33,35,37)	2

NEAREST NEIGHBOUR DENDOGRAM GRAPH FOR THE H-CONFIGURATION



Application of the Furthest Neighbour technique to the H data produces the clusterings reported on pages 30-32 and the dendogram graph on page 33. Application of this technique can be thought of as producing a graph in which edges connect all nodes in a cluster. When the nearest clusters are merged, the graph is changed by adding edges between every pair of nodes in the two clusters. If we define the diameter of a cluster as the largest distance between points in the cluster, then the distance between two clusters is merely the diameter of their union. If we define the diameter of a partition as the largest diameter for clusters in the partition, then each iteration increase the diameter of a partition by as little as possible. This is advantageous when the true clusters are compact and roughly equal in size. However, when this is not the case, as happens with two elongated clusters, say, the resulting groupings can be meaningless. This is another example of impossing structure on data rather than finding structure in it.

For the purpose of completeness we include the clustering at each stage for the Centroid, Median, Group Average and Ward's method and their respective dendogram graphs. These methods represent a varying degree of compromise between the 'chaining effect' of the Nearest Neighbour technique and the compact clusterings resulting from the Furthest Neighbour technique. Also the global optimum search techniques Centroid, Median and Group Average all reverse their steps in the region of stages 16,15,14 inorder to form a new clustering at a splitting level lower then that in stage 17.

FURTHEST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
		Stages 1-10 are exactly the same as those for the Nearest Nearest Neithbour technique.	
11	0.486	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (13,34),(19,32),(21,28),(27,36),(29,33)	26
12	0.487	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (13,34),(19,32),(21,28),(27,36),(29,33), (20,24)	25
13	0.488	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35), (13,34),(19,32),(20,24),(21,28),(26,37), (27,36),(29,33)	24
14	0.503	(2,6),(3,30),(5,25),(7,31),(8,17),(9,15), (12,35),(13,34),(19,32),(20,24),(21,28), (26,37),(27,36),(29,33)	23
15	0.508	(1,14)(2,6),(5,25),(7,31),(8,17),(9,15), (3,30),(12,35),(13,34),(19,32),(20,24), (21,28),(26,37),(27,36),(29,33)	22
16	0.539	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31), (9,15),(12,35),(13,34),(19,32),(20,24), (21,28),(26,37),(27,36),(29,33)	21
17	0.541	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31), (9,15),(12,35),(13,34),(16,18),(19,32), (20,24),(21,28),(26,37),(29,33),(27,36)	20
18	0.543	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31), (9,15),(10,23),(12,35),(13,34),(16,18), (19,32),(20,24),(21,28),(26,37),(27,36), (29,33)	19
19	0.910	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31), (9,15),(10,23),(12,35),(13,34),(16,18), (19,32),(20,24),(21,22,28),(26,37),(27,36), (29,33)	18

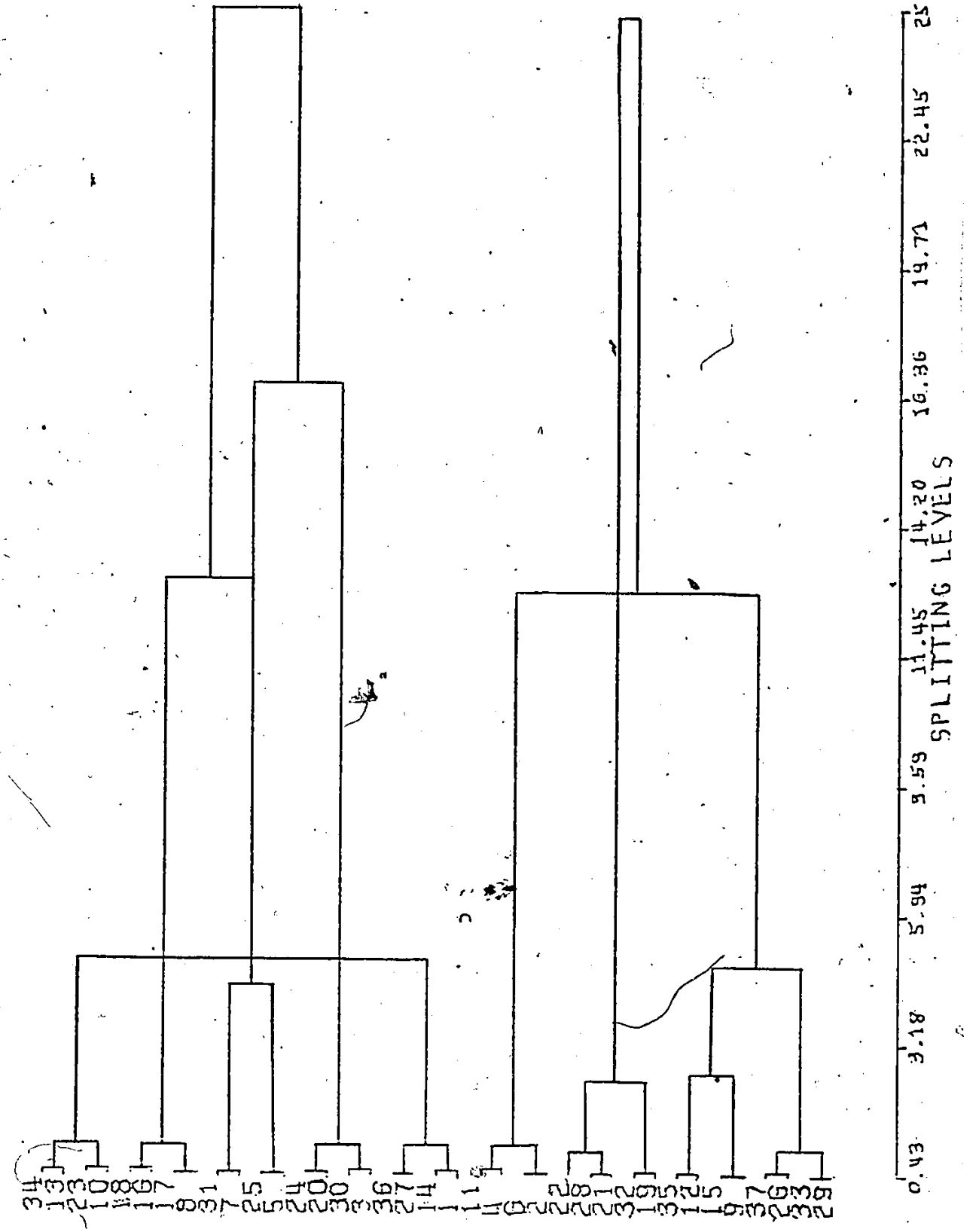
FURTHEST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
20	0.967	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37),(27,36),(8,17)	17
21	1.047	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(8,16,17,18),(9,15),(10,23),(12,35),(13,34),(19,32),(20,24),(21,22,28),(26,29,33,37),(27,36)	16
22	1.055	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(20,24),(21,22,28),(26,29,33,37),(27,36)	15
23	1.058	(1,14),(2,6),(3,20,24,30),(4,11),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(21,22,28),(26,29,33,37),(27,36)	14
24	1.066	(1,14,27,36),(2,6),(3,20,24,30),(4,11),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(21,22,28),(26,29,33,37)	13
25	1.071	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(21,22,28),(26,29,33,37)	12
26	2.442	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,21,22,28,32),(26,29,33,37)	11
27	2.564	(1,14,27,36),(2,4,6,11),(3,20,24,30)(5,25),(7,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,28,32),(26,29,33,37)	10
28	4.441	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,7,25,31),(9,12,15,35),(10,13,23,34),(19,21,22,28,32),(26,29,33,37)	9

FURTHEST NEIGHBOUR CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
29	4.869	(1,14,27,36),(2,4,6,11),(3,20,24,30), (8,16,17,18),(9,12,15,26,29,33,35,37), (10,13,23,34),(19,21,22,28,32),(5,7,25,31)	8
30	4.988	(1,10,13,14,23,27,34,36),(2,4,6,11), (3,20,24,30),(8,16,17,18), (9,12,25,26,29,33,35,37),(5,7,25,31), (19,21,22,28)	7
31	12.808	(1,10,13,14,23,27,34,36), (2,4,6,9,11,12,15,26,29,33,35,37), (3,20,24,30),(5,7,25,31),(8,16,17,18), (19,21,22,26)	6
32	13.047	(1,8,10,13,14,16,17,18,23,27,34,36), (2,4,6,9,11,12,15,26,29,33,35,37), (3,20,24,30),(5,7,25,31),(19,21,22,26)	5
33	17.214	(1,8,10,13,14,16,17,18,23,27,34,36), (2,4,6,9,11,12,15,26,29,33,35,37), (3,5,7,20,24,25,30,31),(19,21,22,26)	4
34	25.037	(1,8,10,13,14,16,17,18,23,27,34,36), (2,4,6,9,11,15,19,21,22,25,26,29,33,35,37), (3,5,7,20,24,25,30,31)	3
35	25.219	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,27, 30,31,34,36), (2,4,6,9,11,15,19,21,22,25,26,29,33,35,37)	2

FURTHEST NEIGHBOUR DENDOGRAM GRAPH FOR THE H-CONFIGURATION



CENTROID CLUSTER ANALYSIS FOR THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
		Stages 1-13 are exactly the same as those for the Furthest Neighbour technique.	
14	0.495	(2,6),(5,25),(7,31),(8,17),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37),(27,36)	23
15	0.503	(2,6),(3,30),(7,31),(5,25),(8,17),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37),(27,36)	22
16	0.508	(1,14),(2,6),(5,25),(8,17),(7,31),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37),(27,36),(3,30)	21
17	0.526	(1,14,27,36),(2,6),(5,25),(7,31),(8,17),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37),(3,30)	20
18	0.539	(1,14,27,36),(2,6),(3,30),(4,11),(5,25),(7,31),(8,17),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37)	19
19	0.528	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,17),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37)	18
20	0.541	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,17),(9,15),(12,35),(13,34),(16,18),(19,32),(20,24),(21,28),(26,29,33,37)	17
21	0.536	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,16,27,18),(9,15),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37)	16
22	0.543	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,23),(12,35),(13,34),(19,32),(20,24),(21,28),(26,29,33,37)	15

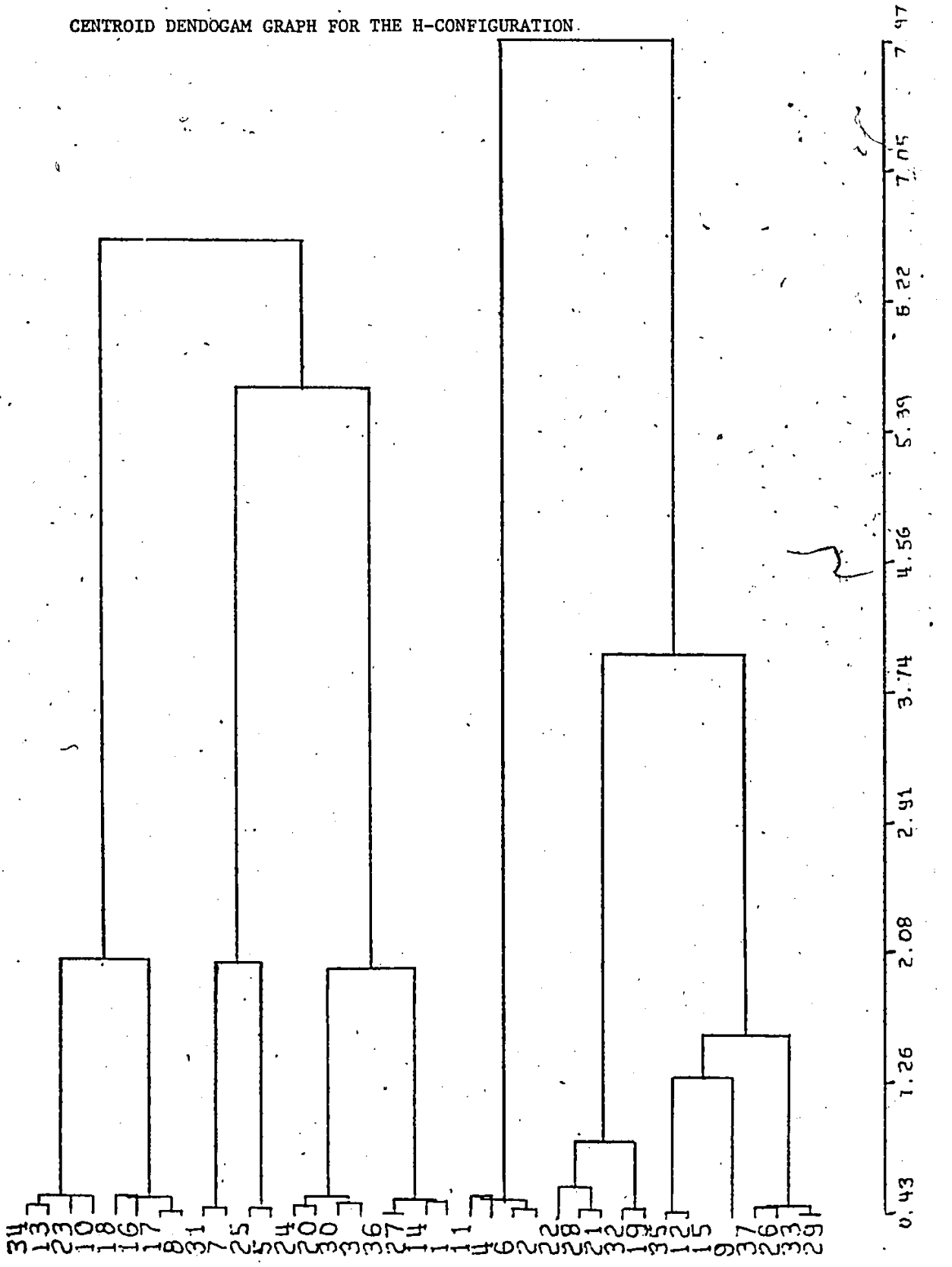
CENTROID CLUSTER ANALYSIS FOR THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
23	0.541	(1,14,27,36),(2,4,6,11),(3,30),(5,25), (8,16,17,18),(9,15),(10,13,23,34),(12,35), (19,32),(20,24),(21,28),(26,29,33,37),(7,31)	14
24	0.544	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25), (8,16,17,18),(9,15),(10,13,23,34),(12,35), (19,32),(21,28),(26,29,33,37),(7,31)	13
25	0.604	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25), (8,16,17,18),(9,15),(10,13,23,34),(12,35), (19,32),(21,22,28),(26,29,33,37),(7,31)	12
26	0.903	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25), (8,16,17,18),(9,15),(10,13,23,34),(12,35), (19,21,22,28,32),(26,29,33,37),(7,31)	11
27	1.295	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25), (8,16,17,18),(9,12,15,35),(10,13,23,34), (19,21,22,28,32),(26,29,33,37),(7,31)	10
28	1.569	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25), (8,16,17,18),(9,12,15,35,26,29,33,37), (10,13,23,34),(19,21,22,28,32),(7,31)	9
29	1.982	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,25), (8,16,17,18),(9,12,15,35,26,29,33,37), (10,13,23,34),(19,21,22,28,32),(7,31)	8
30	2.024	(1,3,14,20,24,27,30,36),(2,4,6,11), (5,7,25,31),(8,16,17,18), (9,12,15,35,26,29,33,37),(10,13,23,34), (19,21,22,28,32)	7
31	2.043	(1,3,14,20,24,27,30,36),(2,4,6,11) (5,7,25,31),(8,10,13,16,17,18,23,34), (9,12,15,35,26,29,33,37),(19,21,22,28,32)	6
32	3.987	(1,3,14,20,24,27,30,36),(2,4,6,11), (5,7,25,31),(8,10,13,16,17,18,23,34), (9,12,15,19,21,22,26,28,29,32,33,35,37)	5

CENTROID CLUSTER ANALYSIS FOR THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
33	5.672	(1,3,5,7,14,20,24,25,27,30,31,36), (2,4,6,11),(8,10,13,16,17,18,23,34), (9,12,15,19,21,22,26,28,29,32,33,35,37)	4
34	6.609	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25, 27,30,31,34,36),(2,4,6,11), (9,12,15,19,21,22,26,28,29,32,33,35,37)	3
35	7.873	(1,3,5,7,8,10,13,14,16,17,18,20,24,23,25, 27,30,31,34,36),(2,4,6,9,11,12,15,19,21,22, 26,28,29,32,33,35,37)	2

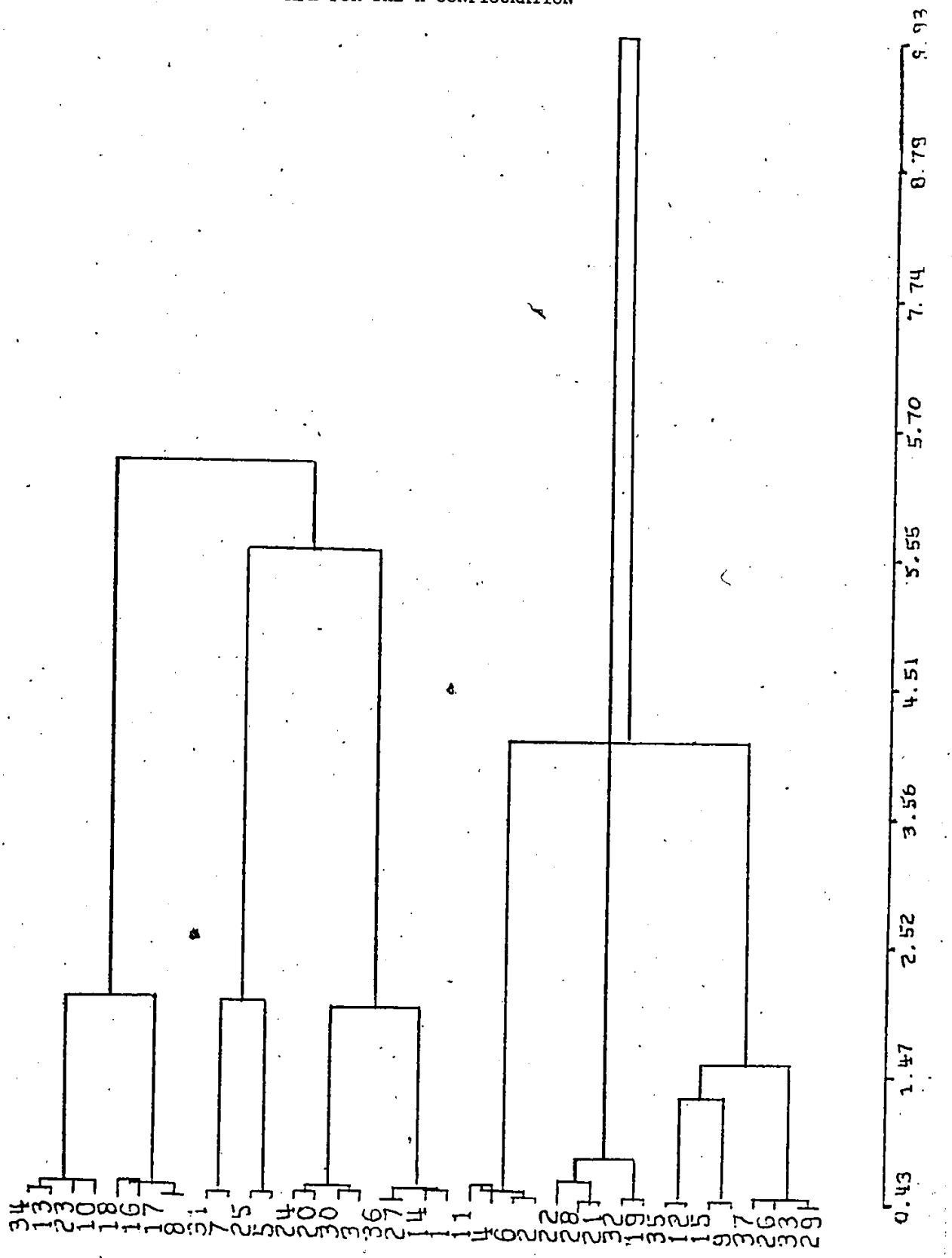
CENTROID DENDOGAM GRAPH FOR THE H-CONFIGURATION.



MEDIAN CLUSTER ANALYSIS FOR THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
		Stages 1-25 are exactly the same as those for the Centroid method.	
26	0.792	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,21,22,28,32),(26,29,33,37)	11
27	1.295	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,28,32),(26,29,33,37)	10
28	1.569	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,12,15,26,29,33,35,37),(10,13,23,34),(19,21,22,28,32)	9
29	1.982	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,25),(7,31),(8,16,17,18),(9,12,15,26,29,33,35,37),(10,13,23,34),(19,21,22,28,32)	8
30	2.024	(1,3,14,20,24,27,20,36),(2,4,6,11),(5,7,25,31),(8,16,17,18),(9,12,15,26,29,33,35,37),(10,13,23,34),(19,21,22,28,32)	7
31	2.043	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,7,25,31),(8,10,13,16,17,18,23,34),(9,12,15,26,29,33,35,37),(19,21,22,28,32)	6
32	4.157	(1,3,14,20,24,27,30,36),(2,4,6,9,11,12,15,26,29,33,35,37),(5,7,25,31),(8,10,13,16,17,18,23,34),(19,21,22,28,32)	5
33	5.672	(1,3,5,7,14,20,24,25,27,30,31,36),(2,4,6,9,11,12,15,26,29,33,35,37),(8,10,13,16,17,18,23,34),(19,21,22,28,32)	4
34	6.386	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,27,30,31,34,36),(2,4,6,9,11,12,15,26,29,33,35,37),(19,21,22,28,32)	3
35	9.831	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,27,30,31,34,36),(2,4,6,9,11,12,15,19,21,22,26,28,29,32,33,35,37)	2

MEDIAN DENDOGRAM GRAPH FOR THE H-CONFIGURATION



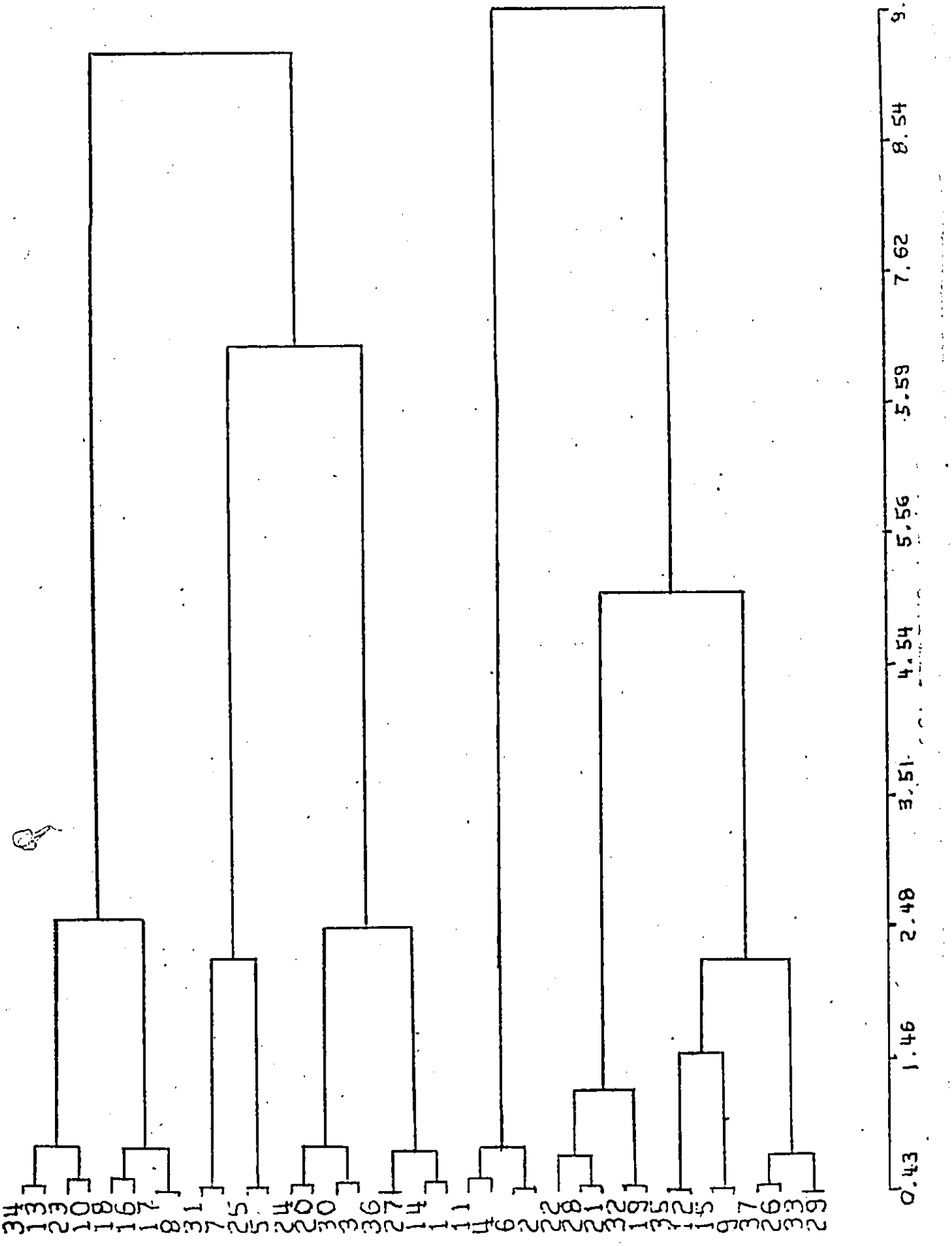
GROUP AVERAGE CLUSTER ANALYSIS OF THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
		Stages 1-18 are exactly the same as those for the Furthest Neigbour technique.	
19	0.720	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,37),(27,26),(29,33),(8,17)	18
20	0.723	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37),(27,36),(8,17)	17
21	0.761	(1,14,27,36),(2,6),(3,30),(4,11),(5,25),(7,31),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37),(8,17)	16
22	0.778	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37),(8,17)	15
23	0.782	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,23),(12,35),(13,34),(19,32),(20,24),(21,22,28),(26,29,33,37)	14
24	0.791	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,23),(12,35),(13,34),(19,32),(21,22,28),(26,29,33,37)	13
25	0.798	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(21,22,28),(26,29,33,37)	12
26	1.232	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,21,22,28,32),(26,29,33,37)	11
27	1.521	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,28,32),(26,29,33,37)	10

GROUP AVERAGE CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
28	2.243	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,12,15,26,29,33,34,37),(10,13,23,34),(19,21,22,28,32)	9
29	2.259	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,7,25,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,28,32)	8
30	2.490	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,7,25,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,28,32)	7
31	2.563	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,7,25,31),(8,10,13,16,17,18,23,34),(9,12,15,35),(19,21,22,28,32)	6
32	5.107	(1,3,14,20,24,27,30,36),(2,4,6,11),(5,7,25,31),(8,10,13,16,17,18,23,34),(9,12,15,19,21,22,28,32,35)	5
33	7.045	(1,3,5,7,14,20,24,25,30,31,36),(2,4,6,11),(8,10,13,16,17,18,23,34),(9,12,15,19,21,22,28,32,35)	4
34	9.348	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,30,31,32,35,36),(2,4,6,11),(9,12,15,19,21,22,28,32,35)	3
35	9,672	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,30,31,32,35,36),(2,4,6,9,11,12,15,19,21,22,28,32,35)	2

GROUP AVERAGE DENDOGRAM GRAPH FOR THE H-CONFIGURATION



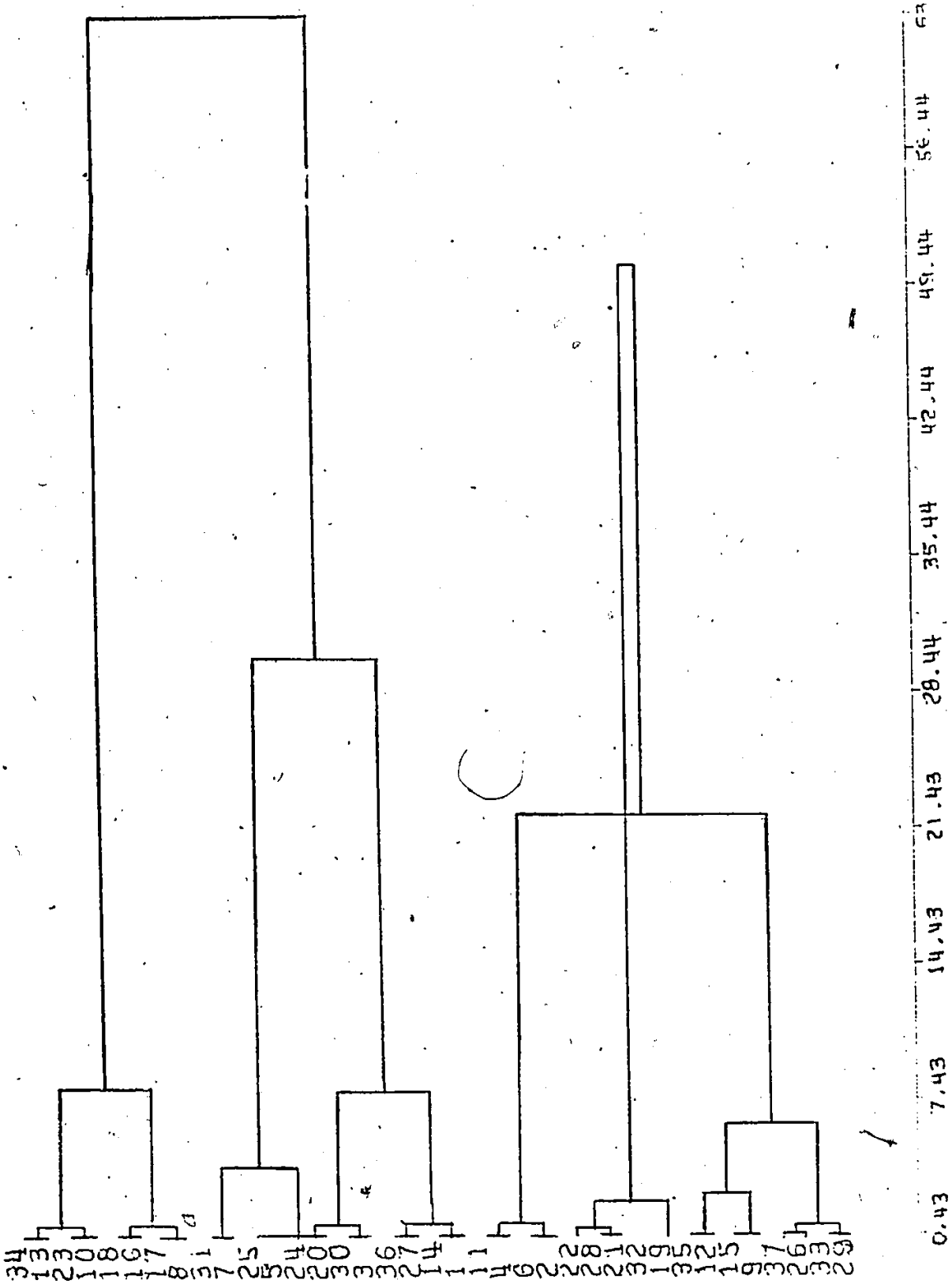
WARD'S METHOD CLUSTER ANALYSIS OF THE H-CONFIGURATION

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
		Stages 1-18 are exactly the same as those for the Furthest Neighbour technique.	
19	0.805	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(8,17),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,37),(29,33),(27,36)	18
20	0.989	(1,14),(2,6),(3,30),(4,11),(5,25),(7,31),(8,17),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37),(27,36)	17
21	1.052	(1,14,27,36),(2,6),(3,30),(4,11),(5,25),(7,31),(8,17),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37)	16
22	1.056	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,17),(9,15),(10,23),(12,35),(13,34),(16,18),(19,32),(20,24),(21,22,28),(26,29,33,37)	15
23	1.072	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,23),(12,35),(13,34),(19,32),(20,24),(21,22,28),(26,29,33,37)	14
24	1.081	(1,14,27,36),(2,4,6,11),(3,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(20,24),(21,22,28),(26,29,33,37)	13
25	1.087	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(7,31),(8,16,17,18),(9,15),(10,13,23,34),(12,35),(19,32),(21,22,28),(26,29,33,37)	12
26	2.166	(1,14,27,36),(2,4,6,11),(3,20,24,30),(5,25),(12,35),(19,21,22,28,32),(26,29,33,37),(7,31),(8,16,17,18),(9,15),(10,13,23,34)	11
27	2.590	(1,14,27,36),(2,4,6,11),(3,20,24,20),(5,25),(7,31),(8,16,17,18),(9,12,15,35),(10,13,23,34),(19,21,22,2832),(26,29,33,37)	10

WARD'S METHOD CLUSTER ANALYSIS OF THE H-CONFIGURATION (CONT.)

STAGE	SPLITTING LEVEL	CLUSTERS	NUMBER OF CLUSTERS
28	4.048	(1,14,27,36),(2,4,6,11),(3,20,24,30), (5,7,25,31),(8,16,17,18),(9,12,15,35), (10,13,23,34),(19,21,22,28,32),(26,29,33,37)	9
29	6.273	(1,14,27,36),(2,4,6,11),(3,20,24,30), (5,7,25,31),(8,16,17,18), (9,12,15,26,29,33,35,37),(10,13,23,34), (19,21,22,28,32)	8
30	7.925	(1,3,14,20,24,27,30,36),(2,4,6,11), (5,7,25,31),(8,16,17,18,), (9,12,15,26,29,33,35,37),(10,13,23,34), (19,21,22,28,32)	7
31	8.169	(1,3,14,20,24,27,30,36),(2,4,6,11), (5,7,25,31),(8,10,13,16,17,18,23,34), (9,12,15,26,29,33,35,37),(19,21,22,28,32)	6
32	22.167	(1,3,14,20,24,27,30,36), (2,4,6,9,11,12,15,26,29,33,35,37), (5,7,25,31),(8,10,13,16,17,18,23,34), (19,21,22,28,32)	5
33	30.251	(1,3,5,7,14,20,24,25,27,30,31,36), (2,4,6,9,11,12,15,26,29,33,35,37), (8,10,13,16,17,18,23,34),(19,21,22,28,32)	4
34	50.526	(1,3,5,7,14,20,24,25,27,30,31,36), (2,4,6,9,11,12,15,19,21,22,26,28,29,32, 33,35,37),(8,10,13,16,17,18,23,34)	3
35	63.446	(1,3,5,7,8,10,13,14,16,17,18,20,23,24,25,27, 30,31,34,36),(2,4,6,9,11,12,15,19,21,22,26, 28,29,32,33,35,37)	2

WARD'S METHOD DENDOGRAM GRAPH FOR THE H-CONFIGURATION



5.2.8: Proofs and Counterexamples for Particular Combinatorial Procedures

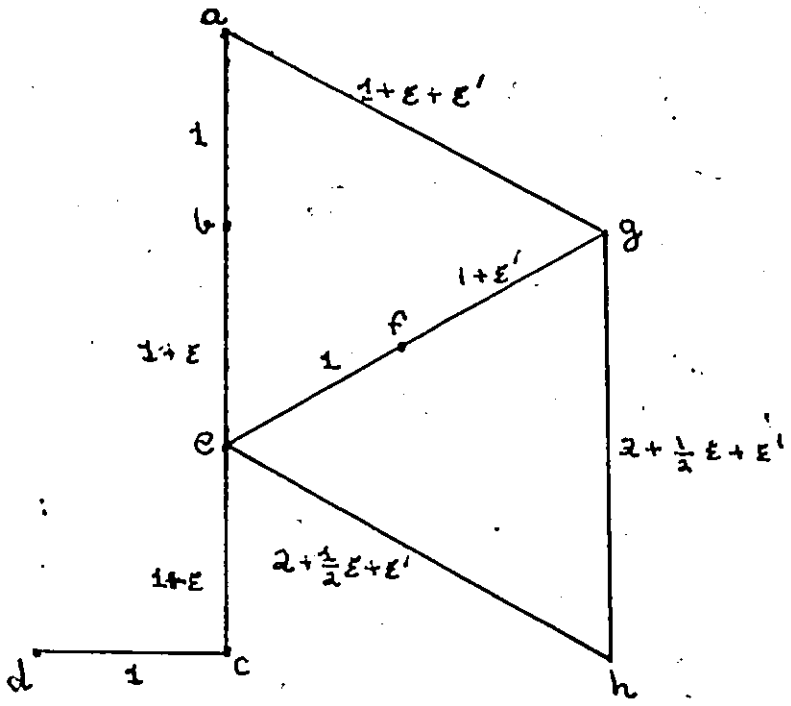
A) The Nearest Neighbour and Furthest Neighbour procedures are well-structured (exact tree) admissible.

This is clear, since at each step in the analysis the dissimilarity coefficient is a ultrametric. This fact also quarantees the next property.

B) The Nearest Neighbour and Furthest Neighbour procedures are monotone admissible.

C) None of these procedures, with the exception of Nearest Neighbour, are connected admissible.

For the case of Furthest Neighbour this may be shown by considering the following diagram:



In this case for ε sufficiently small we get { a,b,c,d} and { e,f,g,h} at one stage. The dissimilarity messure in this case being the city block metric . Similar examples exist for the other techniques.

D) Centroid, Group Average and Ward's Method are not point proportion admissible.

This results from the fact that at each stage the inter-group distances depend on the number of elements in the groups.

E) Centroid, Group Average and Ward's Method are not monotone admissible.

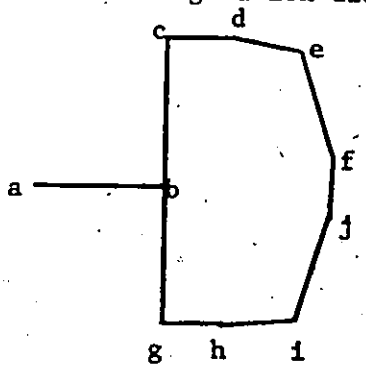
We give the counterexample for the Centroid method. Consider the points $a=[-\{(1+\epsilon)^2-\frac{1}{4}\}^{\frac{1}{2}},\frac{1}{2}]$, $b=[-\{(1+\epsilon)^2-\frac{1}{4}\}^{\frac{1}{2}},-\frac{1}{2}]$, $c=(0,0)$, and $d=(1+\frac{1}{2}\epsilon,0)$. There are four inter-point distances 1, $1+\frac{1}{2}\epsilon$, and $[2+3\epsilon+\epsilon^2+(2+\epsilon)\{(1+\epsilon)^2-\frac{1}{4}\}^{\frac{1}{2}}]^{\frac{1}{2}}$. Map this into 1, $1+\frac{1}{2}\epsilon$, $(1+(1+\epsilon)^2)^{\frac{1}{2}}$ and $[(2+3/2\times\epsilon)^2+\frac{1}{4}]^{\frac{1}{2}}$, respectively. Then the clustering changes from (a,b,c) and (d) to (a,b) and (c,d). Other counter-examples exist for the other methods named.

F) Nearest Neighbour and Furthest Neighbour procedures are not convex admissible.

Consider the following set of points

a=(0,0)	f=(2.2,0.2)
b=(1,0)	g=(1,-1.1)
c=(1,1.1)	h=(1.5,-1.1)
d=(1.5,1.1)	i=(2.0,-1)
e=(2.0,1)	j=(2.2,-0.2)

These points form the configuration diagrammed below:



Then the clustering change from (a,b,c) and (d) to (a,b).

At one stage in the Nearest Neighbour procedure, we get (a,b) and (c,d,e,f,g,h,i,j) and at one stage in the Furthest Neighbour procedure we get (a,b,f,i,j) and c,d,e,g,h). Proving the result.

5.3: *Ling's (k,r)-cluster procedure* [25,26]

This is a nonoverlapping generalization of the Nearest Neighbour cluster procedure and one would expect it to have similar properties.

Definition 16: Given (Y,D), where D is a matrix of dissimilarity coefficients, a nonempty $X \subseteq Y$ is r-connected if for each pair of elements (x,y) in X, there exists an r-chain of X connecting x and y; i.e., there exists a sequence $x=x_1,x_2,\ldots,x_m=y$ in X such that

$$d(x_i,x_{i+1}) \leq r \text{ for } i=1,2,\ldots,m-1.$$

Definition 17: Given (Y,D) a nonempty $X \subseteq Y$ is (k,r)-bonded, where k is a positive integer if

$$(\forall x \in X)(\exists \text{ a k-element subset } T \ X \mid x \notin T)$$
$$(\forall t \in T)\ d(x,t) \leq r.$$

Definition 18: Given (Y,D) a noempty $X \subseteq Y$ is (k,r)-connected if X is (k,r)-bonded and r-connected.

Definition 19: A subset $S \subseteq Y$ is maximal with respect to some property P if S satisfies P and is not a proper subset of any set in X that stisfies P.

Definition 20: Given (Y,D,k) a subset $X \subseteq Y$ is a (k,r)-cluster if r is the minimum value of s for which X is (k,s)-connected for some s and X is maximal (k,r)-connected.

Definition 21: X Y is a k-cluster if X is a (k,r)-cluster for some r.

If C is a (k,r)-cluster and C' is a (k,r')-cluster with $r \leq r'$, then either $C \cap C' = \emptyset$ or $C \subset C'$. The parameter r can be interpreted as the 'time of birth' of a (k,r)-cluster. The 'isolation index' or 'survival time' of a cluster is defined as follows:

Definition 22: Given (Y,D,k) and a (k,r)-cluster C≠Y, the isolation index of C, denoted $\iota(C)$, is r'-r, where r' is the parameter of the smallest k-cluster properly containing C.

The algorithm used to implement the method is as follows:

i) Replace the dissimilarity matrix D by its corresponding matrix Δ of ranks. Let $r=\binom{K+1}{2}$, the smallest rank that can give rise to a (k,r)-cluster.

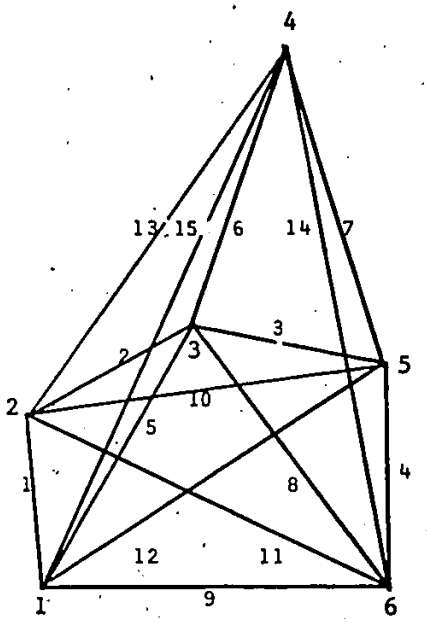
ii) Find B_r, the maximal (k,r)-bonded set of Y.

iii) If $B_r \neq \phi$, decompose B_r into r-connected components (k-clusters) $S_1,\ldots,S_m$, $m \geq 1$.

iv) If a new cluster is found, save all relevant information (cluster size and element identification) about this cluster, and if some previously found cluster C, whose isolation index is not yet defined, is a proper subset of the new cluster, then $\iota(C)$ is computed.

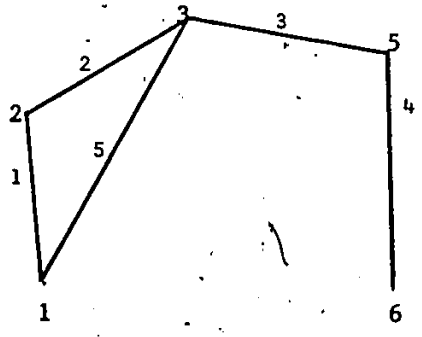
v) Increase r by 1 and repeat steps ii)-iv) until the largest cluster X is reached.

In most clustering methods, one generally gives (n-1) cluster candidates in the process of clustering n points. One distinctive feature of the k-clustering method, as a consequence of the definition of a (k,r)-cluster, is that often the number of possible k-clusters is considerably less than (n-1). This phenomenon and an illustration of some clustering definitions are proved for k=2 and 3 by the following example:

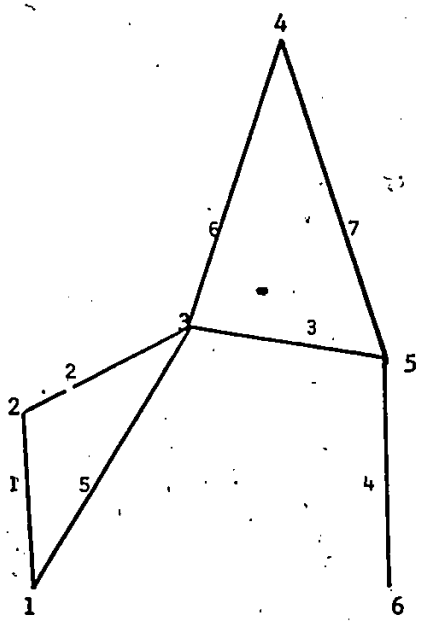


Δ-matrix

	1	2	3	4	5	6
1						
2	1					
3	5	2				
4	15	13	6			
5	12	10	3	7		
6	9	11	8	14	4	

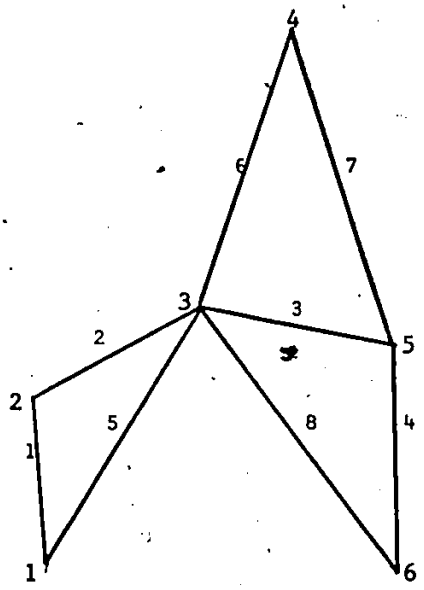


k=2,r=5
I=(1,2,3,5,6)
II=(1,2,3)
III=(1,2,3)

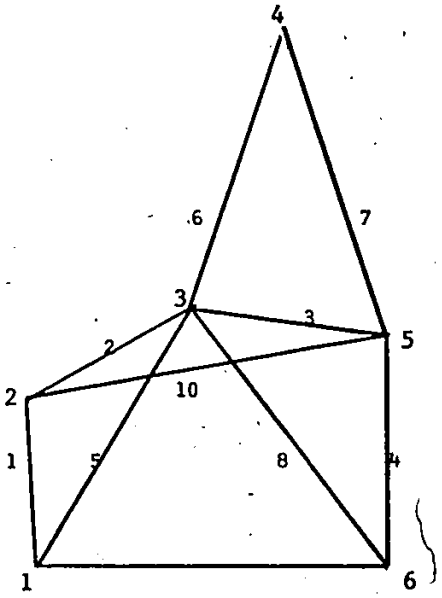


k=2,r=7
I=(1,2,3,4,5,6)
II=(1,2,3),(3,4,5),
(1,2,3,4,5)
III=(1,2,3,4,5)

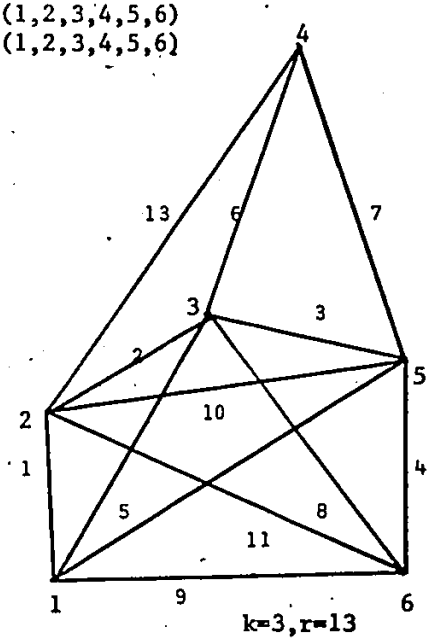
I. maximal r-connected set
II. bonded sets
III. cluster



k=2,r=8
I=(1,2,3,4,5,6)
II=(1,2,3),(3,4,5),(3,5,6),
(3,4,5,6),(1,2,3,4,5),(1,2,3,5,6),
(1,2,3,4,5,6)
III=(1,2,3,4,5,6)



k=3,r=10
I=(1,2,3,4,5,6)
II=(1,2,3,5,6)
III=(1,2,3,5,6)



k=3,r=13
I=(1,2,3,4,5,6)
II=(1,2,3,5),(1,2,3,6),(1,2,5,6),(2,3,4,5),
(2,3,5,6),(1,2,3,4,5),(1,2,3,5,6),
(2,3,4,5,6),(1,2,3,4,5,6)
III=(1,2,3,4,5,6)

5.3.1: The (k,r)-clustering of the H-configuration.

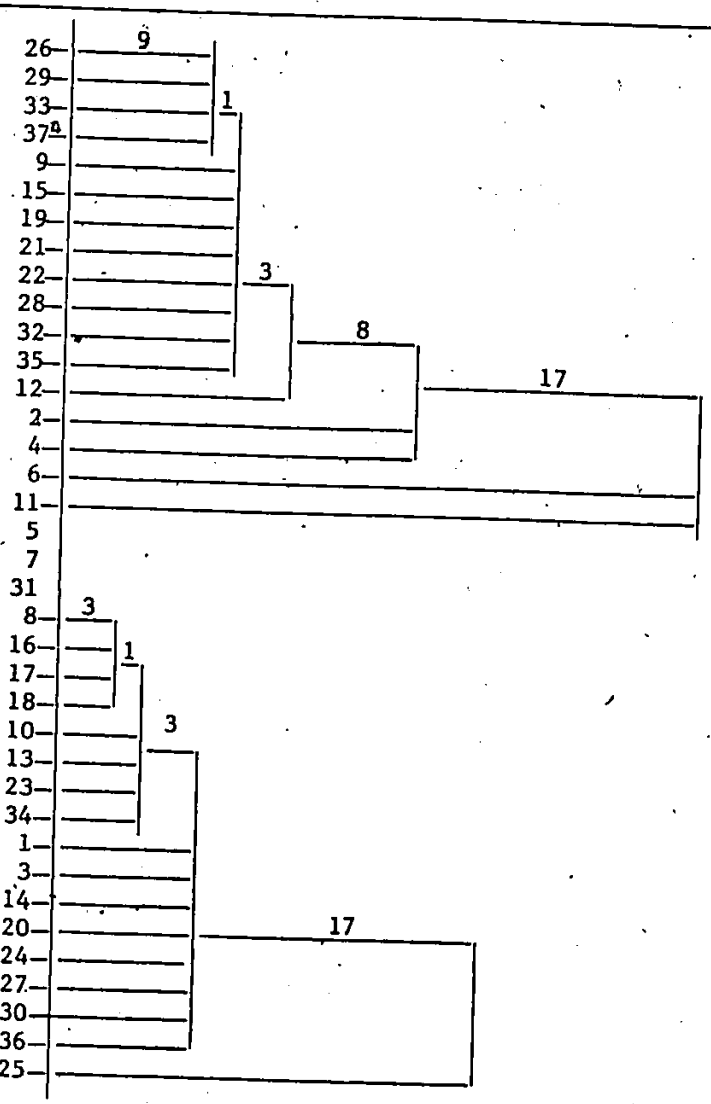
HISTORY OF CLUSTERING
BOND SIZE=3

CLUSTER	MODE	SIZE	ISOLATION INDICES	CLUSTERS
1	56	4	9.	26,29,33,37
2	65	12	1.	9,15,19,21,22,26,28,29,32, 33,35,37
3	66	13	3.	9,12,15,19,21,22,26,28,29, 32 33 35,37
4	68	4	3.	8,16,17,18
5	69	15	8.	2,4,9,12,15,19,21,22,26,28, 29,32,33,35,37
6	71	8	1.	8,10,13,16,17,18,23,34
7	72	16	3.	1,3,8,10,13,14,16,17,18,20, 23,24,27,30,34,36
8	75	17	19.	1,3,8,10,13,14,16,17,18,20, 23,24,25,27,30,34,36
9	77	17	17.	2,4,6,9,11,12,15,19,21,22, 26,28,29,32,33,35,37
10	94	37	——	1,2,3,4,5,6,7,8,9,10,11,12, 13,14,15,16,17,18,19,20,21, 22,23,24,25,26,27,28,29,30, 31,32,33,34,35,36,37

The dendogram graph for this analysis appears on the following page. The clusters that emerge towards the end of the combinatorial processes are on the whole unatural. The configuration of points while it does not suggest any particular set of points to be clustered, it does suggest symmetry. The lack of symmetry in the combinatorial results tends to indicate the sensitivity of the algorithms to small local perturbations. The distance matrix of ranks was clustered by Ling's method using k=3. Examination of the isolation indices reveals, that clusters 8 and 9 may reasonably be considered as 'real'. Under the probability model studied in Ling[26] , which is not quite appropriate for the data under

consideration but nevertheless offers a rough guide to the assessment of the clustering indices, these clusters are seen to be statistically significant. The symmetry of the configuration, which was not brought out by the other methods examined, is easily seen here.

THE (K,R)-CLUSTERING DENDOGRAM GRAPH FOR THE H-CONFIGURATION



5.3.2: The admissibility properties of (k,r)-clustering.

A) The (k,r)-clustering procedure is image admissible.

This is true since the algorithm works only with the ranks of the DC's which are unaffected by any one-to-one map of the data points into themselves. This consideration also shows the following to hold:

B) The (k,r)-clustering procedure is monotone admissible.

C) The (k,r)-clustering procedure is not convex admissible.

This seems reasonable due to this technique's relationship with the Nearest Neighbour method. But a counter-example appears hard to find.

D) The (k,r)-clustering procedure is connected admissible.

The entire point in developing this technique was to supply this connected property.

E) The (k,r)-clustering procedure is not well-structured (exact tree) admissible.

Again the procedure works only with the ranks of the DC's so that it is not possible to work backwards to obtain the original dissimilarity matrix.

F) The (k,r)-clustering procedure is well-structured (k-group) admissible.

This and the following appear obvious.

G) The (k,r)-clustering procedure is not well-structured (perfect) admissible.

H) The (k,r)-clustering procedure is point proportion admissible.

This holds if a suitable method is used to break tied ranks.

I) The (k,r)-clustering procedure is cluster proportion admissible.

J) The (k,r)-clustering procedure is cluster ommission admissible.

5.4: Jardine and Sibson's B_k, fine, clustering procedures. [17]

This is an overlapping technique based on an absolute restriction, namely that the overlap between distinct ML-sets at the same level shall not contain more than k-1 elements of Y. Thus, B_1 is the single link method and as k increases the methods B_k allow progressively more overlap between the ML-sets until when k=N-1 the overlap restriction becomes vacuous and B_k=I for $k \geq N-1$.

If r is a symmetric reflexive relation on P ($r \in \Sigma(P)$) and $a,b,c \in P$, then r is transitive if

$$[(a,c) \in r \cap (c,b) \in r] \rightarrow (a,b) \in r.$$

The first condition to be imposed is obtained from this ordinary transitivity relation by replacing the single element c by the k-element set S, which must be completely linked. On the left hand side of the implication the condition $\{a\} \times S \in r$ corresponds to $(a,c) \in r$ and $S \times \{b\} \in r$ to $(c,b) \in r$. So that

Definition 23: Let $r \in \Sigma(Y)$. r is (weakly) k-transitive, (denoted T_k), if whenever $S \subseteq Y$, $|S|=k$, $a,b \in Y$,

then $[\{a\} \times S \cup S \times S \cup S \times \{b\}] \subseteq r \rightarrow (a,b) \in r$.

The effect of this condition is easy to see in terms of ML-sets. Suppose that S_1, S_2 are distinct ML-sets for r. Then both $S_1 - S_2$ and $S_2 - S_1$ are non-empty. There exists $a \in S_1 - S_2$, $b \in S_2 - S_1$ such that $(a,b) \in r$, because if this where not so S_1 S_2 would be an ML-set. It follows that $|S_1 \cap S_2| < k$ if r satistfies definition 23, that is, the overlap between distinct ML-sets contains at most k-1 elements. Then the left-hand side of the implication in definition 23 is true if and only if $\{a\} \times S$ and $S \times \{b\}$ lie wholly inside the ML-sets for r. If they lie in the same ML-set, then $(a,b) \in r$ and the condition is satisfied. If they lie wholly

in distinct ML-sets the S must lie in the intersection of these two ML-sets and so must contain at most k-1 elements, and again the condition is satisfied. If the left-hand side of the implication is false, then the condition is satisfied anyhow. Thus the k-transitivity condition is precisely the condition to apply to a symmetric reflexive relation to ensure that the overlap between distinct ML-sets contains at most k-1 elements.

The associated set of NSAC's consists of those NSAC's for which c(h) is k-transitive for all h. The set of k-transitive symmetric reflexive relations on a set P will be denoted by $J_k(P)$. An NSAC such that $c[0,\infty) \subseteq J_k(P)$ will be called a (fine) k-dendogram. $C_k(P)$ is the set of DC's identified with the set of k-dendograms under the correspondence T in theorem 1, but this a very inconvenient way of characterizing it; a condition analogous to the ultrametric inequality is desired if possible. This leads to the following definition:

Definition 24: Let $d \in C(Y)$. d is (weakly) k-ultrametric

if whenever $S \subseteq Y, |S|=k$, $a,b \in Y$

then $d(a,b) \leq \max\{ d(c,e) \mid c \in \{ e\} \cup \{a,b\}, e \in S\}$.

If whenever $R \subseteq P$, we write

$$\text{diam}(d,R) = \max\{ d(X,Y) \mid X,Y \in R\},$$

then the k-ultametric inequality can be written as

$$d(a,b) \leq \max\{ \text{diam}(e, S \cup a), \text{diam}(e, S \cup b)\}.$$

it is easily seen that d is k-ultrametric if and only if $Td(h) \in J_k(Y)$.

There is a very simple condition which is equivalent to the k-ultrametric inequality. The condition is that on every (k+2)-element subset of Y, the largest dissimilarity value on the subset should occur more than once in the subset.

Thus if $d(x,z)=\text{diam}(d,R)$ where $x,z\in R$, $|R|=k+2$,

$$(\exists x',z'\in R\,|\,\{x,z\}\neq\{x',z'\})\ d(x,z)=d(x',z').$$

If this condition is taken to characterize $C_k(Y)$, then it is immediately obvious that $C_k(Y)$ is sup-closed and, consequently, gives rise to a well-defined, sub-dominate, method $C(Y)\to C_k(Y)$: this is the fine clustering method B_k.

The implications

$$T=T_1\to T_2\to\ldots\to T_{N-2}\to T_{N-1}=NN$$

where NN is the vacuous condition, give rise to the containment relations

$$E(Y)=J_1(Y)\subset J_2(Y)\subset\ldots\subset J_{N-2}\subset J_{N-1}=\Sigma(Y)$$

and so to

$$U(Y)=C_1(Y)\subset C_2(Y)\subset\ldots\subset C_{N-2}(Y)\subset C_{N-1}(Y)=C(Y)$$

It follows that B_1 is the single link method, $B_{N-2}=I$ and $B_k\cdot B_j=B_k$ for $k\leq j$. As k increases $B_k(d)$ yields progressively more and more information about d, until when $k\geq N-1$, $B_k(d)=d$ and complete information is recovered. The relation $B_k\cdot B_j=B_k$ for $k\leq j$, is computationally valuable for in the process of finding $B_k(d)$ from $B_{k+1}(d)$ - we do not need to go back to d itself.

This definition of B_k as the subdominant method associated with $C_k(Y)$ is by no means the simplest, because, B_k has special properties which derive essentially from its relatedness to an absolute restriction on overlap. For each $h\in[0,\infty)$, $[TB_k(d)](h)\in J_k(Y)$ and, because of the restriction, $J_k(Y)$ is fixed and independent of d and h.

It is possible to write

$$[TB_k(d)](h)=\gamma_k[Td(h)]$$

where γ_k is a function from (Y) to $J_k(Y)$ which maps each element of

$\Sigma(Y)$ to the smallest element of $J_k(Y)$ containg it. If $D:C(Y)\to Z$ is a cluster method, and if there exists a function $\gamma:\Sigma(Y)\to\Sigma(Y)$ such that

$$[TD(d)](h)=\gamma\,[Td(h)]$$

for all $d\in C(Y)$, $h\geq 0$, then D is called a uniform cluster method. Thus the methods B_k are uniform cluster methods, and all we need know is the function γ_k: $\Sigma(Y)\to J_k(Y)$. In terms of ML-sets, if r is a relation, then the sets in ML(r) will in general have arbitrary overlap. Add pairs to r so that whenever S_1 and S_2 are in ML(r) and $|S_1\cap S_2|\geq k$, S_1 and S_2 are replaced by S_1 S_2 in $ML(r^*)$. Continue until $ML(r^*)$ contains no pairs S_1,S_2 with $|S_1\cap S_2|\geq k$. Then r^* is $\gamma_k(r)$.

This suggests a method of generating uniform cluster methods which are also subdominant. Let $J\subseteq\Sigma(Y)$, $\gamma:\Sigma(Y)\to J$ which associates with each element of $\Sigma(Y)$ the smallest member of J containing it. Then if Z_J is the set of DC's corresponding to those NSAC's for which $c[0,\infty)\subseteq J$, the subdominant method D defined by Z_J is given by

$$[TD(d)](h)=\gamma\,[(Td)(h)].$$

In order for to be well-defined some conditions must be placed on J, these conditions are given in Jardine and Sibson [17,chapter 10].

A) The B_k, fine, clustering procedure is image admissible.

This appears clear.

B) The B_k, fine, clustering procedure is not convex admissible.

No overlapping clustering procedure can fulfill this requirement.

C) The B_k, fine, clustering procedure is not connected admissible.

This again is due to the fact that clusters may overlap.

D) The B_k, fine, clustering procedure is not well-structured (exact tree) admissible.

The generating DC may not be truly ultrametric.

E) The B_k, fine, clustering procedure is not well-structured (k-group) admissible.

Do to the fact of overlap, at certain stages, some within-cluster distances may be equal to some between-cluster distances. It also follows that:

F) The B_k, fine, clustering procedure is not well-structured (perfect) admissible.

G) The B_k, fine, clustering procedure is point proportion admissible.

H) The B_k, fine, clustering procedure is cluster propotion admissible.

I) The B_k, fine, clustering procedure is not cluster omission admissible.

Removal of all elements of one cluster also causes removal of some elements from other clusters and may change the cluster boundaries.

J) The B_k, fine, clustering procedure is monotone admissible.

This is a flat cluster method and hence monotone admissible by definition see Jardine and Sibson [17,chapter 10].

K) The B_k, fine, clustering procedure is optimal admissible.

This follows because the B_k clustering procedures are subdominate methods. Note that this result also applies to B_1, the Nearest Neighbour procedure, and the (k,r)-clustering procedure which are also members of this family of uniform cluster procedures, the family of flat cluster methods.

L) The B_k, fine, clustering procedure, the (k,r)-cluster procedure and the Nearest Neighbour procedure are path connected.

This follows by theorem 4 and is implied by the fact that they belong to the same family of uniform cluster methods.

5.5: Jardine and Sibson's B_k^c, coarse, cluster procedure.

These are overlapping techniques closely related to the B_k methods are are defined in terms of absolute restrictions on relations, but these restrictions do not have the simple interpretation in terms of overlap between ML-sets as do the latter methods.

The strong k-transitivity condition ST_k on which coase k-clustering is based, is similar to that of weak k-transitivity and follows:

Defintion 25: Let $r \in (Y)$. r is (strongly) k-transitive

if whenever $S \subseteq Y$, $|S|=k$, $a,b \in Y$,

then, $[\{a\} \times S \cup S \times \{b\}]\ r \rightarrow (a,b) \in r$.

A coase k-dendogram is an NSAC such that $c[0,\infty) \subseteq J_k^c(Y)$, the set of strongly k-transitive relations on Y. The corresponding set of DC's is $C_k^c(Y)$, DC's satisfying the strongly k-ultrametric inequality.

Defintion 26: Let $d \in C(Y)$. d is (strongly) k-ultrametric

if whenever $S \subseteq Y$, $|S|=k$, $a,b \in Y$

then $d(a,b) \leq \max\{d(t,e) \mid c \in \{a,b\}, e \in S\}$.

$C_k^c(Y)$ is readily seen to be sup-closed. The resultant sub-dominant method is the coarse k-clustering method $B_k^c : C(Y) \rightarrow C_k^c(Y)$.

We have the following diagrams of implications and containments:

$$
\begin{array}{ccccccccc}
T & = & T_1 & \rightarrow & T_2 & \rightarrow \ldots \rightarrow & T_{N-2} & \rightarrow & T_{N-1} = NN \\
\Updownarrow & & \Updownarrow & & \Updownarrow & & \Updownarrow & & \Updownarrow \\
T & = & ST_1 & \rightarrow & ST_2 & \rightarrow \ldots \rightarrow & ST_{N-2} & \rightarrow & ST_{N-1} = NN
\end{array}
$$

$$
\begin{array}{ccccccccc}
E(Y) & = & J_1(Y) & \subset & J_2(Y) & \subset \ldots \subset & J_{N-2} & \subset & J_{N-1} = (Y) \\
\| & & \| & & \cup & & \cup & & \| \\
E(Y) & = & J_1^c(Y) & \subset & J_2^c(Y) & \subset \ldots \subset & J_{N-2}^c & \subset & J_{N-1}^c = (Y) \\
\cup & & \cup & & \cup & & \cup & & \\
U(Y) & = & C_1(Y) & \subset & C_2(Y) & \subset \ldots \subset & C_{N-2} & \subset & C_{N-1} = C(Y) \\
\| & & \| & & \cup & & \cup & & \| \\
U(Y) & = & C_1^c(Y) & \subset & C_2^c(Y) & \subset \ldots \subset & C_{N-2}^c & \subset & C_{N-1}^c = C(Y)
\end{array}
$$

Like the B_k methods, the B_k^c methods are uniform cluster methods.

Coarse k-clustering lacks the convenient interpretation in terms of overlap between ML-sets available for fine k-clustering. Certainly strong k-transitivity implies transitivity with the convers holding only for k=1 or k N-1. So that the ML-sets for a strongly k-transitive relation have overlap of at most k-1 elements of Y, but are not characterized by this property. Coarse k-clustering may be regarde as allowing overlaps of the same kind as the fine k-clustering but making less efficient use of them.

This inefficiency makes this a bad clustering procedure relative to the B_k methods although they satisfy exactly the same admissibility conditions.

5.6: Jardine and Sibson's C_u (u-diametric) clustering procedures.

This type C cluster method is based on an internal restriction for controlling overlap in terms of overlap diameter in relation to the current level. The methods C_u are subdominant methods associated with the restriction that the diameter of the overlap of distinct ML-sets at level h will be at most uh, where u is a constant..

Definition 27: Let $d \in C(Y)$. d is u-diametric

if whenever $\phi \neq S \subset Y$, $a,b \in Y$,

then setting

$$l = \max\{\mathrm{diam}(d, S \cup \{a\}), \mathrm{diam}(d, S \cup \{b\})\},$$

we have

$$\mathrm{diam}(d,S) > ul \rightarrow d(a,b) \leq l. \qquad (u \geq 0)$$

If d is u-diametric, then the NSAC Td has the property that if S_1, S_2 are distinct ML-sets at level h, then

$$\mathrm{diam}(d, S_1 \cap S_2) \leq uh:$$

it follows that $E_u(Y)$, the set of u-diametric DC's, is sup-closed. So that there is a sub-dominant method $C_u : C(Y) \rightarrow E_u(Y)$; this method is called u-diametric clustering. If $u \geq 1$, the condition $\mathrm{diam}(d,S) > ul$ cannot be satisfied, so for $u \geq 1$, $E_u(Y) = C(Y)$ and $C_u = I$.

If d is definite, then $(\forall S \subset Y)\, |S| > 1 \rightarrow \mathrm{diam}(d,S) > 0$. Thus if u=0, $|S| > 1 \rightarrow d(a,b) \leq \max\{\mathrm{diam}(d, S \cup \{a\}), \mathrm{diam}(d, S \cup \{b\})\}$, and d is weakly 2-transitive. So that on the set of definite DC's, the methods C_0 and B_2 are the same. However, they are not equivalent on the set of nondefinite DC's. If two objects have small or zero DC's, it may be inferred that duplicates have been chosen. This will upset the type B methods which count objects, but not the type C methods based on diameter. However, there is a price to pay for this advantage. The type C

methods are not uniform, and since the DC's employed must have more than ordinal significance, thsy are not monotone admissible.

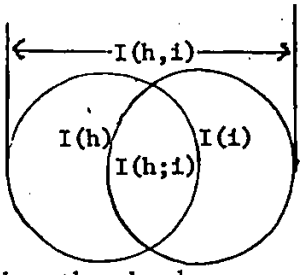
Just as B_k is nested, so C_u is nested for values of u which may be taken to be in the range [0,1] and $C_u \cdot C_v = C_u$ if $u \leq v$. Note that C_u is not the single link method for any value of u.

The type C methods satisfy the same admissibility requirements as the type B methods except for monotone admissibility.

5.7: Orloci's Information Theory Method [29]

It must be absolutely clear that the term 'information' is here used in a strictly technical sense. It is conceived as a physical property of events related to probability. In accordance with this definitio n a rare event conveys more information than a common event. Information as a technical term thus is conceptually related closely to surprisal than to either knowledge or informativeness in ordinary speech.

For reasons of convenience, our collection of objects will be conceived of as a set of individual freguencies, each according to the problem at hand, representing an individual, or a character. Information theory offers several distinct functions suitable for the definition of structure in such a collection. These functions include total information and mutual information and joint information. In the following diagram I(h) and I(i) indicate the total information within the frequency distributions labelled X_h and X_i respectively.



I(h,i) and I(h;i) on the other hand, represent the joint information and mutual information between X_h and X_i. The collection consists of r frequency distributions which are either independent, or related, as the case may be. Each distribution is conceived as a union of several subsets of frequency classes, X_{hA}, X_{hB} and so forth. The union may take place in such a manner that the frequencies are pooled between classes possessing

equal class values or identical class symbols. However, pooling of frequencies is not justified if the subsets are qualitatively disjoint.

A much used measure for total information is Shannon's [22] entropy function. This function for the h'th frequency distribution X_h is

$$I(h) = -\sum_{j=1}^{n_h} f_{hj} \ln p_{hj} = N \ln N - \sum f_{hj} \ln f_{hj}. \qquad 1)$$

where the f's are class frequencies, of which there are n_h, the p's are independent *a posteriori* probabilities, and N is the total number of observations so that

$$N = \sum f_{hj} \quad \text{and} \quad p_{hj} = f_{hj}/N \ . \qquad 2)$$

If X_{hA} and X_{hB} are subsets of classes such that

$$X_{hA} X_{hB} = X_{hAB}$$

then

$$I(h)_{AB} \geq I(h)_A + I(h)_B. \qquad 3)$$

The total information conveyed jointly by two subsets in h, is never less than the pooled information conveyed separately by the subsets. This relation will be used to find optimum unions, or subdivisions, in the collection.

The r I's may be pooled to derive an overall measure of joint information:

$$I(1,2,\ldots,r) = \sum_{j=1}^{r} I(j).$$

This is valid only if the r distributions are independent. If they are not, the pooled value of information exceeds the true joint information by an amount equal to the mutual information shared between the r distributions.

It is possible to represent the r frequency distributions in an r-dimensional contingency table. In such a table, original class frequencies appear as r sets of marginal totals, and the values in the body of the table specify the frequencies of the joint observations made on the r entities simultaneously. When r=2, we have the common case of paired comparisons and the joint information of X_h and X_i is

$$I(h,i) = -\sum_{j=1}^{n_h}\sum_{k=1}^{n_i} f_{jk} \ln p_{jk} = N \ln N - \sum\sum f_{jk} \ln f_{jk}. \quad 4)$$

A similar expression can be found for the joint information $I(1,2,\ldots,r)$ in the case of any r dimensional table.

If we visualize the contingency table as containing the relationship between subsets X_{hA} and X_{iA} and the relationship X_{hB} and X_{iB} is represented by a second table, the the following relations hold:

$$I(h,i)_{AB} \geq I(h,i)_A + I(h,i)_B$$

$$I(h,i)_{AB} \leq I(h)_{AB} + I(i)_{AB} \quad 5)$$

or

$$I(h,i)_A \leq I(h)_A + I(i)_A. \quad 6)$$

The quantity $I(h,i)_A$ corresponds to a contingency table relating X_{hA} and X_{iA}. Similarly $I(h,i)_B$ corresponds to a table relating X_{hB} and X_{iB}. In this context, $I(h,i)_{AB}$ corresponds to a table derived from the two tables corresponding to subsets with or without pooling frequencies, depending on the problem at hand.

Inequality 5 plays an important role in selecting optimal fusions or subdisions in cluster analysis. It also implies that the value of the joint information in $A \cup B$ cannot be less than a quantity obtained by pooling the joint information corresponding to the subsets. Inequality 6

is discussed in detail by Khinchin and Kullbach [22]. Note that the joint information as given in equation 4, differs from the pooled information, given in equation 3, by a correction for mutual information.

The information possessed in common between two frequency distributions X_h and X_i, called mutual information, can be expressed in the following terms:

$$I(h;i) = \sum_{j=1}^{n_h} \sum_{k=1}^{n_i} f_{jk} \ln \frac{N f_{jk}}{f_{hj} f_{ik}}$$

$$= \sum\sum f_{jk} \ln f_{jk} + N \ln N - \sum_{j=1}^{n_h} f_{hj} \ln f_{hj} - \sum_{k=1}^{n_i} f_{ik} \ln f_{ik} \qquad 7)$$

Equation 7 is called the error or independence component of the discrimination information [22]. In the case of an r-dimensional contingency table the simplest expression for the overall mutual information is

$$I(1;2;\ldots;r) = \sum_{h=1}^{r} I(h) - I(1,2,\ldots,r)$$

For N sufficiently large, the mutual information is asymptotically distributed as χ^2 with $(n_h-1)(n_i-1)$ degrees of freedom in the case of a two dimensional table, or $n_1 n_2 \ldots n_r - n_1 - n_2 - \ldots - n_r + r - 1$ degrees of freedom in the case of an r-dimensional contingency table.

The following relations are of importance:

$$I(h;i)_{AB} \geq I(h;i)_A + I(h;i)_B \qquad 8)$$

and

$$I(h;i) \leq I(q) \text{ where } I(q) = \min (I(h), I(i)).$$

similar expressions can be easily derived for the r-dimensional case. Inequality 8 implies that the mutual information in A B cannot be less then the quantity obtained by pooling the mutual information corresponding to the subsets.

The relative relatedness of two frequency distributions can be measured by Rajski's coherence coefficient which for X_h and X_i is

$$R(h;i) + (1-d^2(h;i))^{\frac{1}{2}} \qquad 9)$$

where $d(h;i) = 1 - \frac{I(h;i)}{I(h,i)}$ is called Rajski's metric. R(h;i) varies between zero and unity, indicating respectively the degree of relatedness from none to perfect. The probability corresponding to 2I(h,i) is called the relative measure of relatedness.

As a computational example of the proceding consider the following data indicating the performance of 2 species in 28 stands of vegetation evaluated in accordance with an arbitrary scale.

SUBSET A

SPECIES	1	2	3	4	5	6	7	8	9	10	11	12	13
h	3	3	3	3	3	3	3	0	+	0	0	0	3
i	1	+	1	+	+	+	1	2	2	2	2	2	1

SUBSET B

SPECIES	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
h	3	3	3	3	3	2	4	3	3	3	3	3	+	4	2
i	+	+	+	+	1	0	1	1	2	0	0	0	0	+	0

CLASS VALUES

CLASS VALUES	+	0	1	2	X_h
+	0	1	0	1	2
0	0	0	0	3	3
2	0	2	0	0	2
3	8	3	6	2	19
4	1	0	1	0	2
X_i	9	6	7	6	28

I(h)=29.9027
I(i)=38.4041
d(h;i)= 0.7922

I(h,i)=56.5515
I(h;i)=11.7554
R(h;i)= 0.6103

SUBSET A

CLASS VALUES

	+	1	2	X_{hA}
+	0	0	1	1
0	0	0	3	3
3	4	4	1	9
X_{iA}	4	4	5	13

SUBSET B

CLASS VALUES

	+	0	1	2	X_{hB}
+	0	1	0	0	1
2	0	2	0	0	2
3	4	3	2	1	10
4	1	0	1	0	2
X_{iB}	5	6	3	1	15

$I(h)_A$=10.2735
$I(i)_A$=14.2067
$I(h,i)_A$=18.9581
$I(h;i)_A$= 5.5221
$d(h;i)_A$= 0.7088
$R(h;i)_A$= 0.7054

$I(h)_B$=14.8228
$I(i)_B$=18.5272
$I(h,i)_B$=29.0022
$I(h;i)_B$= 4.3428
$d(h;i)_B$= 0.8503
$R(h;i)_B$= 0.5263

The clustering techniques based on information rely on the heterogeniety information given by

$$\Delta I_{AB} = I(1;2)_{AB} - I(1;2)_A - I(1;2)_B \qquad 11)$$

where the I's represent the mutual information between two freguency distributions X_1 and X_2 representing the row and column classifications in a table A or B in A∪B. If the goal is the classification of the rows of X, A is fused with B if ΔI_{AB} is minimum. The probability is defined by the asymptotic relation

$$2\Delta I_{AB} = \chi^2 \text{ at } (r_A+r_B-1)(c_A+c_B-1)-(r_A-1)(c_A-1)-(r_B-1)(c_B-1)$$

or r-1 when $r_A=r_B$ degrees of freedom. The symbols r_A, c_A, r_B, c_B indicate the number of rows and columns in table A and B respectively.

Basic data are usually not given as frequencies. It is nevertheless possible to summerize most types of data in frequency tables and then manipulate these tables to derive a measure of heterogenity between subjects. There are two alternative avenues of approach to follow in this connection. The first makes use of the joint information and is as follows

$$\Delta I_{AB}=I(1,2,\ldots,r)_{AB}-I(1,2,\ldots,r)_{A}-I(1,2,\ldots,r)_{B}.$$

The second utilizes an expression similar to the error component of discrimination information as applied to r-dimensional frequency tables and is

$$I_{AB}=I(1;2;\ldots;r)_{AB}-I(1;2;\ldots;r)_{A}-I(1;2;\ldots;r)_{B}$$

where the different I's represent the joint information between r rows, or the mutual information, within the disjoint subset A or B of the columns of X and in $A \cup B$.

A) Information theory clustering procedures are image admissible.

A one-to-one map of the data onto itself cannot effect the frequency tables on which the method is based.

B) Information theory clustering procedures are not convex admissible.

C) Information theory clustering procedures are not connected admissible.

D) Information theory clustering procedures are not well-stuctured (exact tree) admissible.

The information statistics are not by nature ultrametrics.

E) Information theory clustering procedures are well-structured (k-group) admissible.

F) Information theory clustering procedures are not well-structured (perfect) admissible.

The following results hold because of the method's dependence on group size. A relative measure of infomation including the average information or entrophy

$$H(h)=I(h)/n_h ,$$

the probabilities corresponding to information at given degrees of freedom, or other measures such as Rajski's metric or the coherence could be used to get around this problem.

G) Information theory clustering procedures are not point proportion admissible.

H) Information theory clustering procedures are not cluster proportion admissible.

I) Information theory clustering procedures are cluster omission admissible.

J) Information theory clustering procedures are not monotone admissible.

5.8: Unsupervised Bayesian Estimation.

5.8.1: Introductory Remarks.

An outline of the theoretical development of this section, due to Patrick (31), is given in APPENDIX I. This is the only clustering technique which, in the classical sense, has a statistical appearance.

Emphasizing mixtures, Patrick and Costello (34) showed that the Bayes minimum-conditional-risk solution involves the information function

$$\eta(\underline{b},\underline{b}^*) = \int \ln h(\underline{x}|\underline{b})\, h(\underline{x}|\underline{b}^*)\, d\underline{x}$$

where $h(\underline{x}|\underline{b}^*)$ is the true density and $h(\underline{x}|\underline{b})$ is a mixture probability characterized by $\underline{b}$. They showed that an estimate of $\eta(\underline{b})$,

$$\hat{\eta}(\underline{b}) = \frac{1}{n} \sum_{s=1}^{n} \ln h(\underline{x}_s\ \underline{b})$$

should be evaluated at every $\underline{b}$ in the parameter space. The Bayes solution with mean square error loss then uses $\hat{\eta}(\underline{b})$ indirectly to weight $\underline{b}$ to form the average estimator $(\underline{b}_n)_s$ which is a Bayes estimator. This averaging property of the Bayes approach can be contrasted with a stochastic-approximation approach (based on some starting value $(\underline{b})_0$ which searches for the maximum of $\eta(\underline{b})$ with respect to $\underline{b}$. Stochastic approximation is starting-point dependent, whereas Bayes can 'average out' the starting points. The 'quasi-Bayes' approach developed later, was developed to incorporate the desirable Bayes averaging effect with the desirable stochastic-approximation property of reduced complexity.

Properties of mixtures were first considered in statistical literature and applied to the unsupervised estimation problem by engineers. The problem of unsupervised estimation is to resolve an unknown mixture into the underlying catagories or, equivalently, to find the indicies (parameter vectors) and weights (mixing parameters) that express the

unknown mixture density as a linear combination of density functions.

Maximum-likelihood-estimator equations for the two-catagory problem where F is Gaussian were developed by Cooper and Cooper [7] for the case of a single unknown parameter. Patrick [30;31] extended this to more than one-unknown parameter. These results were for M=2 and known (M is the number of catagories). The multicatagory problem and M unknown has a numerical solution developed by Wolfe [52].

Stochastic-approximation algorithms which seek the maximum of the average likelihood function are defined by Sakrison [43] for a normalized problem. If $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ have density $h(\underline{x}|\underline{b}^*)$, a maximum-likelihood estimator $\hat{\underline{b}}$ for $\underline{b}^*$ is a solution of

$$\begin{aligned}\hat{\underline{b}} &= \arg\{\max_{\underline{b}} \ln f_n(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n|\underline{b})\}\\ &= \arg\{\max_{\underline{b}} \ln \prod_{s=1}^{n} h(\underline{x}_s|\underline{b})\}\\ &= \arg\{\max_{\underline{b}} \sum_{s=1}^{n} \ln h(\underline{x}_s|\underline{b})\}\\ &= \arg\{\max_{\underline{b}} n\hat{\underline{l}}(\underline{b})\}\end{aligned}$$

IN the decision-directed approach, *a priori*, there are M mean vectors covarianve matrices, and *a priori* catagory propabilities

$$(\underline{m}_i),\ (\underline{\Sigma}_i),\ (P_i), \qquad i=1,2,\ldots,M.$$

When a sample $\underline{x}_1$ is received it is assumed to be from catergory i if

$$\ln \frac{(P_i)_0}{|(\underline{\Sigma}_i)_0|^{\frac{1}{2}}} - (\underline{x}_1-(\underline{m}_i)_0)'[(\underline{\Sigma}_i)_0]^{-1}(\underline{x}_1-(\underline{m}_i)_0)$$

is larger than for any one catagory. The $\underline{x}_1$ is used to update $(\underline{m}_i)_0$, $(\underline{\Sigma}_i)_0$ to $(\underline{m}_i)_1$, $(\underline{\Sigma}_i)_1$. Also $(P_i)_0$ can be updated. Thus, under certain conditions, the decision-directed approach may be considered a good tracking procedure.

A minimum-norm-square-error estimator for $\underline{b}^*$ leads to the definition of a criteria $\underline{\Gamma}(\underline{b})$,

$$\begin{aligned}\underline{\Gamma}(\underline{b}) &= 2E[h(\underline{x}|\underline{b}) - ||h(\underline{x}|\underline{b})||]^2 \\ &= 2\int h(\underline{x}|\underline{b})\, h(\underline{x}|\underline{b}^*)d\underline{x} - \int h^2(\underline{x}|\underline{b})d\underline{x} \\ &= ||h(\underline{x}|\underline{b}^*)||^2 - \underline{e}^2(\underline{b})\end{aligned}$$

where $\underline{e}(\underline{b}) \triangleq ||h(\underline{x}|\underline{b}) - h(\underline{x}|\underline{b}^*)||^2$.

An estimator called 'quasi-Bayes' is now developed using both criterion $\underline{\eta}(\underline{b})$ and $\underline{\Gamma}(\underline{b})$.

Assume that $\underline{b}^*$ is contained in a known bounded set B' so that the i'th component b_i^* of $\underline{b}^*$ lies in the known interval $[\alpha_i, \beta_i]$. The set B' is decomposed into V cells formed by equally subdividing the interval $[\alpha_i, \beta_i]$ into V_i subintervals, where $\prod V_i = V$. After the n'th observation $\underline{x}_n$ of $\underline{x}$ is obtained, V s.a.(stochastic approximation) estimates $(\underline{b}^r)_{n+1}$, r=1,2,...,V are computed, the starting points $(\underline{b}^r)_0$, r=1,2,...,V being the centres of the cells in B'.

The i'th component $(b_i^r)_{n+1}$ is computed by means of

$$(b_i^r)_{n+1} = (b_i^r)_n + \frac{a_n}{c_n}(y^i_{r,2n} - y^i_{r,2n-1}), \quad r=1,2,\ldots,V,$$

where $\{a_n\}$ and $\{c_n\}$ are infinite sequences satisfying

$$\lim_{n\to\infty} c_n = 0$$

$$\sum_{n=1}^{\infty} a_n = 0$$

$$\sum_{n=1}^{\infty} a_n c_n < \infty$$

$$\sum_{n=1}^{\infty} a_n^2 c_n^{-2} < \infty$$

and $y^i_{r,2n}$ and $y^i_{r,2n-1}$ are noisy measurements of the regression function:

$$y^i_{r,2n} = \ln h(\underline{x}|(\underline{b}^r)_n + c_n \underline{e}_i),$$

$$y^i_{r,2n-1} = \ln h(\underline{x}|(\underline{b}^r)_n - c_n \underline{e}_i).$$

The vector $\underline{e}_i$ is a column vector having 1 in the i'th row and zero's elsewhere. If $(b^r_i)_{n+1}$ falls outside the r'th cell; i.e., if

$$|(b^r_i)_{n+1} - (b^r_i)_0| > \tfrac{1}{2}(\beta_i - \alpha_i)/V_i,$$

then $(b^r_i)_{n+1}$ is moved to the nearer end point

$$(b^r_i)_0 \pm \tfrac{1}{2}(\beta_i - \alpha_i)/V_i.$$

So that $(b^r_i)_{n+1}$ will differ in absolute value from its starting point $(b^r_i)_0$ by at most $\tfrac{1}{2}(\beta_i - \alpha_i)/V_i$.

Having incremented the Vs.a. estimates, next form the 'quasi-Bayes' estimator $(\underline{b})_{n+1}$ as

$$(\underline{b})_{n+1} = \sum_{r=1}^{V} (\underline{b}^r)_{n+1} p((\underline{b}^r)_0 | \underline{\dot{x}}_n),$$

where $p((\underline{b}^r)_0 | \underline{\dot{x}}_n)$ is the *a posteriori* probability mass. Thus, $(\underline{b})_{n+1}$ is a weighted average of the V s.a. estimates, the weighting coefficeint of the r'th s.a. estimate being the Bayes *a posteriori* probability mass at the centre of the r'th cell. Forming V s.a. estimates in this way effectively divides the interval of search for b^*_i by V_i. For n sufficiently large, $(\underline{b})_{n+1}$ can be made to differ from $\underline{b}^m$ by as little as required. It may be concluded, Appendix I, that the 'quasi-Bayes' estimator singles out the s.a. estimator whose starting position is the best, in the sense

that the corresponding density function is 'closest' in the norm-square-error sense to the true density.

Of importance in the Bayesian clustering techniques are the empirical and histogram distribution and density of the mixture. The empirical distribution of the mixture from N samples is defined as

$$C_n = \frac{1}{N}\sum_{k=1}^{N} \chi_{(\underline{x}_k,\infty)}(\underline{x})$$

and the empirical density is

$$c_n = \frac{1}{N}\sum_{k=1}^{N} \delta(\underline{x}-\underline{x}_k),$$

where χ and δ are the usual characteristic and delta functions respectively.

Defining N^* ordered regions $\{I_k\}_{k=1}^{N^*}$ on a bounded portion of the observation space V_L, where $\cup I_k = V_L$, $\mu(I_k \cap I_j)=0$, $k\neq j$ (μ denotes Lebesque measure), the histogram estimate of the mixture density is

$$c_n(\underline{x}) = \sum_{k=1}^{N} \chi_{I_k}(\underline{x}) a_{kN}$$

where $a_{kN} = \left(\frac{N-1}{N}\right) a_{kN-1} + \frac{1}{N}\chi_{I_k}(\underline{x})$, $k=1,2,\ldots,N^*$.

The distribution histogram estimate is

$$C_n(\underline{x}) = \sum_{k=1}^{N^*} \chi_{I_k}(\underline{x}) \sum_{t=1}^{k} a_{tN}.$$

This will be generically refered to as the histogram mixture functions.

5.8.2: Bayesian Clustering Techniques.

1) Clustering Technique derived from $\underline{n}(\underline{b})$.

In this section $\underline{n}(\underline{b})$ is investigated under the ssumption that the catagories are Gaussian and 'separated'. This approach presumes a search over admissible partitions and thus admissible $\underline{b}$ given n samples. Advantages and disadvantages of this approach are summarized as follows:

Advantages:

1) A criterion such as $\underline{n}(\underline{b})$ evaluates the goodness of the clustering. The use of a criterion may be necessary when the number of catagories M is unknown.

2) The partition based on n samples is updated with the n+1'st sample without the need to store the first n samples.

3) *A priori* knowledge about P_i and the other catagory parameters can be utilized.

Disadvantages:

1) The procedure may not work well when the 'separable' criterion is violated.

2) The procedure does not provide for assuming (if true) that the catagories are widly separated; consequently, the partition adjustment procedure may be more complex then necessary for the problem concerned.

For an symptotic minimum risk solution it is necessary to find the parameter vector $\underline{b}\epsilon B^{M'}$, with M nonzero mixing parameters M<M', that maximizes

$$\underline{n}(\underline{b}) = \int \ln h(\underline{x}|\underline{b})\, h(\underline{x})\, d\underline{x}$$

$$= \int \ln\Big[\sum_{k=1}^{M} f(\underline{x}|\underline{b}_k)P_k\Big]h(\underline{x})\, d\underline{x}$$

The sample space is partitioned into M disjoint regions defined by

$$S^k \triangleq \{ \underline{x} | f(\underline{x}|\underline{b}_k)P_k > f(\underline{x}|\underline{b}_j)P_j, \quad j \neq k\}, \quad k=1,2,\ldots,M,$$

given a parameter vector $\underline{b}$. It is assumed that over each partioned set the class is Gaussian having mean vector $\underline{m}_k$, covariance matrix $\underline{\Sigma}_k$, with the density truncated at the partition boundary. The true mixture $h(\underline{x})$ is assumed bounded.

Under these assumptions $\underline{\eta}(\underline{b})$ can be expanded to

$$\underline{\eta}(\underline{b}) = \sum_{k=1}^{M} \int_{S^k} \ln[f(\underline{x}|\underline{b}_k)P_k]h(\underline{x})\, d\underline{x}$$

$$= \sum_{k=1}^{M} \left\{ \int_{S^k} h(x)dx \right\}$$

$$\times \left\{ \ln P_k + \ln\left(\frac{1}{(2\pi)^{L/2}|\Sigma_k|^{\frac{1}{2}}}\right) \right.$$

$$\left. - \frac{1}{2} \cdot \frac{\int_{S^k} (\underline{x}-\underline{m}_k)' \underline{\Sigma}_k^{-1} (\underline{x}-\underline{m}_k)\, h(\underline{x})\, d\underline{x}}{\int_{S^k} h(\underline{x})d\underline{x}} \right\}.$$

Ignoring for now the defintion of S_k, take a fixed set of regions characterized by independent parameters. It can be shown by taking partial derivatives with respect to each parameter under the constraint $\sum_{k=1}^{M} P_k = 1$ that for this fixed partition $\underline{\eta}(\underline{b})$ is maximized if the parameters are defined as

$$P_k = \int_{S_k} h(\underline{x})d\underline{x},$$

$$\underline{m}_k = \int_{S_k} \underline{x}h(\underline{x})d\underline{x} \; / \; \int_{S_k} h(\underline{x})d\underline{x},$$

$$\underline{\Sigma}_k = \int_{S_k} (\underline{x}-\underline{m}_k)(\underline{x}-\underline{m}_k)' h(\underline{x})d\underline{x} \; / \; \int_{S_k} h(\underline{x})d\underline{x}, \quad k=1,2,\ldots,M.$$

Maximizing $\underline{n}(\underline{b})$ is equivalent to finding the partition and value of M which maximizes

$$\underline{n}(\underline{b}) = \sum_{k=1}^{M} P_k \ln\left(\frac{1}{(2\pi)^{L/2}|\underline{\Sigma}_k|^{\frac{1}{2}}}\right), - \frac{L}{2}$$

This equation illustrates the advantage of the separable assumption. If the complexity-reducing assumption $\Sigma_k = \sigma_k^2 I$ is made, it can be shown by taking partial derivatives, that

$$\sigma_k^2 = \int_{S^k} ||\underline{x}-\underline{m}_k||^2 h(\underline{x})d\underline{x} \; / \int_{S^k} h(\underline{x})d\underline{x}, \; k=1,2,\ldots,M.$$

The following rules define the manner in which the statistics of two fused classes are updated. It is assumed classes r and s are fused into j.

Case A: Two Clusters:

1) $n^j = n^r + n^s$,

2) $\underline{m}_j = (n^r\underline{m} + n^s\underline{m})/\, n^j$,

3) $C_j = (n^r C_r + n^s C_s)/\, n^j$;

Case B: A cluster (class r) and an isolated point (class s):

1) $n^j = n^r + 1$,

2) $\underline{m}_j = (n^r\underline{m}_r + \underline{m}_s)/\, n^j$,

3) $C_j = (n^r C_r + C_s)/n^j$;

Case C: Two isolated points:

1) $n^j = 2$,

2) $\underline{m}_j = \frac{1}{2}(\underline{m}_r + \underline{m}_s)$,

3) $C_j = \frac{1}{2}(C_r + C_s)$;

Case D: A cluster (class r) and the n'th sample:

1) $n^r = n^r + 1$,

2) $\underline{m}_r = (n^r\underline{m}_r + \underline{x}_n)/n^r$,

3) $C_r = (n^r C_r + \underline{x}_n \underline{x}_n')/n^r$;

Case E: An isolated point (class r) and the n'th sample:

$$1)\ n^r = 2,$$

$$2)\ \underline{m}_r = (\underline{m}_r + \underline{x}_n) / 2,$$

$$3)\ C_r = (C_r + \underline{x}_n \underline{x}_n') / 2 .$$

The algoithm is started by using the first M' samples as M' isolated points. Then given a sample $\underline{x}_n$, there are two types of possible actions:

i) Assign $\underline{x}_n$ to one of M' classes (M' possible actions),

ii) Combine two of the M' classes and make x_n a new isolated class $\binom{M'}{2}$ possible actions.

The $\underline{n}(\underline{b})$ criterion is calculated for each of the $M' + \binom{M'}{2}$ possible actions. The update corresponding to the action which maximizes $\underline{n}(\underline{b})$ is then performed. Effectively, the favourable action produces the largest estimated $\underline{n}(\underline{b})$ by accepting classes having large estimated values of

$$P_k \ln P_k |\Sigma_k|^{\frac{1}{2}}.$$

SPECIAL CASE: $\Sigma_k = \sigma^2 I$ and $P_k = 1/M$, where M is known.

Under these conditions

$$\underline{n}(\underline{b}) = \ln\left(\frac{1}{M(2\pi)^{L/2}\sigma^L} - \frac{1}{2\sigma^2} \sum_{k=1}^{M} \int_{S_k} ||\underline{x}-\underline{m}_k||^2 h(\underline{x}) d\underline{x} \right),$$

and maximizing $\underline{n}(\underline{b})$ is eqivalent to finding

$$\min_{\underline{m}_k} \sum_{k=1}^{M} \int_{S_k} ||\underline{x}-\underline{m}_k||^2 h(\underline{x}) d\underline{x}.$$

This criterion, or the slight generalization for P_k not identical, is relevant to the class of decision-directed algorithms. It is extremely easy to implement an algorithm to asymptotically minimize risk under this

a priori assumption. One disadvantage of this criterion is that it cannot be used to determine M if this knowledge is not available. This drawback is due to the fact that it does not incorporate a cost for adding additional catagories. Parallel processing for M=1,2,...,M' can produce more classes than are really there. For example, let $h(\underline{x})$ be a one-dimensional Gaussian distribution with mean zero and variance σ^2. If is is assumed that M=2, the solution generated by the above criterion is a partition through $\underline{x}=0$. Denoting the variances on either side of this partition by σ_i^2, i=1,2, the strict inequality $\sigma_1^2+\sigma_2^2<\sigma^2$ holds. So even though there is only one class the criterion is less for M=2 than for M=1 and the criterion's use to determine M fails. For certain applications, knowledge of M may not be an unreasonable assumption, and the criterion yields on of the simplest unsupervised estimation algorithms in existence [39] .

2) Clustering Utilizing "Portable Magnifying Glass" (Cluster Map or Gravity technique).

Suppose $\underline{x}$ has a density

$$h(\underline{x}) = \sum_{i=1}^{M} P_i f(\underline{x} | \underline{m}_i, \Sigma_i).$$

If the covariance matrix of the test function is carfully chosen and the M catagories reasonably separate, it can be assumed that the test function $t(\underline{x}|\underline{x}_s, \underline{\Phi})$ "singles out" the d'th member if $f(\underline{x}|\underline{m}_d, \Sigma_d)$ is the dominant cluster near $\underline{x}_s$,

$$t(\underline{x}|\underline{x}_s, \underline{\Phi})h(\underline{x}) = t(\underline{x}|\underline{x}_s, \underline{\Phi})P_d f(\underline{x}|\underline{m}_d, \Sigma_d).$$

Under the above assumptions, $t(\underline{x})h(\underline{x})$ is itself Gaussian: denote this as a Gaussian function with mean vector $\underline{\gamma}_s$ and covariance matrix $\underline{C}_s$. The procedure for estimating the parameters $\underline{m}_d$ and Σ_d characterizing the d'th cluster is as follows:

1) Estimate $\underline{\gamma}_s$ and $\underline{C}_s$ utilizing a method to be described, after selecting a point s.
2) Supply P_d and $\underline{\Phi}$.
3) Calculate $\underline{m}_d$ and Σ_d in terms of $\underline{\gamma}_s$, $\underline{C}_s$, $\underline{\Phi}$, P_d and $\underline{x}_s$.
4) repeat steps 1 and 3 using another point $\underline{x}_s$

Using a theorem by Miller, we have that

$$\Sigma_d = (\underline{C}_s^{-1} - \underline{\Phi}^{-1})^{-1},$$

$$\underline{m}_d = (\Sigma_d \underline{\Phi}^{-1} + I)(\underline{\gamma}_s - \underline{x}_s) + \underline{x}_s.$$

Moments of $t(\underline{x}|\underline{x}_s, \underline{\Phi})\hat{h}(\underline{x})$ are estimated next, where $h(\underline{x})$ is

$$h(\underline{x}) = \frac{1}{n} \sum_{s=1}^{n} \delta(\underline{x} - \underline{x}_s).$$

The moments are

$$\hat{\underline{a}}_s = \frac{1}{n}\sum_{j=1}^{n} t(\underline{x}_j|\underline{x}_s,\underline{\Phi}),$$

$$\hat{\underline{a}}_{su} = \frac{1}{n}\sum_{j=1}^{n} x_{ju}\, t(\underline{x}_j|\underline{x}_s,\underline{\Phi}), \quad u=1,2,\ldots,L,$$

$$\hat{\underline{a}}_{suv} = \frac{1}{n}\sum_{j=1}^{n} (x_{ju}-\hat{a}_{su})(x_{jv}-\hat{a}_{sv})\, t(\underline{x}_j|\underline{x}_s,\underline{\Phi}), \quad u,v=1,2,\ldots,L,$$

and then

$$\hat{\underline{\gamma}}_s = (\hat{a}_{sj}), \quad j=1,2,\ldots,L,$$
$$\hat{\underline{C}}_s = (\hat{a}_{suv}), \quad u,v=1,2,\ldots,L,$$
$$\hat{\underline{\Sigma}}_{sd} = (\hat{\underline{C}}_s - \underline{\Phi}^{-1})^{-1},$$
$$\hat{\underline{m}}_{sd} = (\hat{\underline{\Sigma}}_{sd}\underline{\Phi}^{-1}+1)(\hat{\underline{\gamma}}_s - \underline{x}_s) + \underline{x}_s.$$

A set of parameters $(\hat{\underline{\Sigma}}_{sd}, \underline{m}_{sd})$ may be associated with sample $\underline{x}_s$, $s=1,2,\ldots,n$. If the assumption of separability is reasonably well satisfied the set of points $(\hat{\underline{\Sigma}}_{sd}, \underline{m}_{sd})$, $s=1,2,\ldots,n$, may be expected to form clusters in the parameter space.

For some special applications, it may suffice to assume $\underline{\Sigma}_i = \underline{\Sigma}$ for all catagories $i=1,2,\ldots,M$, where $\underline{\Sigma}$ is known and supplied as *a priori* knowledge, Then it remains to estimate means. In this case

$$\hat{\underline{m}}_{sd} = (\underline{\Sigma}\underline{\Phi}^{-1}+I)(\hat{\underline{\gamma}}_s - \underline{x}_s) + \underline{x}_s.$$

If $\underline{\Sigma}$ is chosen to be an a^{-1} multiple of $\underline{\Phi}$, then

$$\hat{\underline{m}}_{sd} = -a\underline{x}_s + (1+a)\,\hat{\underline{\gamma}}_s.$$

If $\underline{\Sigma}_d$ is unknown, it may be reasonable to employ the following estimation procedure. Let $\underline{\Sigma}_{sd}$ be an *a priori* guess for $\underline{\Sigma}_d$ with confidence of n_a samples, then,

$$\underline{\Sigma}_{sd} = \frac{n_a}{n_d+n_a}\,\underline{\Sigma}_{ad} + \frac{n_d}{n_d+n_a}\,(\hat{C}_s - \underline{\Phi}^{-1})^{-1},$$

and n_d is the number of samples within some distance of $\underline{x}_s$.

Advantages and disadvantages of the cluster-map approach are the following:

Advantages:

1) Search for an optimum partition is not the objective of the procedure. This is an advantage from a time and complexity viewpoint but a possible disadvantage.

2) Clusters are displayed in the parameter space. By successive applications of the mapping, these clusters get tighter.

Disadvantages:

1) A criterion such as $\underline{\eta}(\underline{b})$ for evaluating cluster quality is not utilized. This could be a disadvantage if it is necessary to know the precise number of clusters, which is a difficult problem.

2) The catagory covariance matrix must be provided by interaction.

3) Means and covariances are not directly estimated but displayed as clusters in the parameter space. A subsequent mapping is required to extract means and covariances or to display the cluster in one-, two- or three-dimensional space.

4) To obtain the parameter estimates of $(\hat{\underline{\Sigma}}_{sd}, \hat{\underline{m}}_{sd})$ corresponding to the individual sample $\underline{x}$, requires processing all n samples. Even if the test function $t(x|x_s$ is truncated, it must be determined which samples $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ are within the influence of the test function.

3) Chain map

A relatively simple mapping of clusters in an L-dimensional space to a lower-dimensional space is described now and called a chain map.

Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be n samples. Then:

1) Arbitrarily chose any one of the n vectors, say $\underline{x}_1$

2) Locate the nearest sample to $\underline{x}_1$, say $\underline{x}_2$ using Euclidean distance. Plot the distance between $\underline{x}_1$ and $\underline{x}_2$, denoted d_{12}, along the y_2 axis.

3) Continue this process, producing the 'chain' and plot distances between elements in the chain.

The chain map can be very-educational when viewed on a computer-output display device. When catagories are well separated with the respective catagories 'tightly clustered', the cluster can be identified as those samples between large jumps in the mapped space. If the catagories are not tightly clustered, then there may be frequent 'small jumps' in the display. This is why it may be advantageous to apply a 'cluster tightener' such as a cluster map prior to applying chain map.

Caution should be exercised in using a chain map because it does not directly provide for a distance measure between two vectors other than Euclidean distance. In some applications it is useful to compute the global variances using the mixture of data from all M catagories and then measure the distance between x and y as

$$\sum_{i=1}^{L} \frac{(x_i - y_i)^2}{\sigma^2_i} .$$

3) Clustering technique derived from $\Gamma(\underline{b})$.

Let $h(\underline{x})$ be a mixture of functions from the Gaussian family

$$h(\underline{x}|\underline{b}^*) = \sum_{i=1}^{M} P_i^* N(\underline{x}|\underline{m}_i, \underline{\Sigma}_i^*).$$

the parameter space B is the set of points $(\underline{b}_i)$, where $\underline{b}_i=(\underline{m}_i, \underline{\Sigma}_i)$. Constraints are

$$0 \leq P_i \geq 1$$

$$\sum_{i=1}^{M} P_i = 1.$$

It follows that

$$||h(\underline{x}|\underline{b})||^2 = \sum_{i=1}^{M} \sum_{j=1}^{M} P_i P_j c_{ij}$$

where

$$c_{ij} = \int N(\underline{x}|\underline{m}_i, \underline{\Sigma}_i) N(\underline{x}|\underline{m}_j, \underline{\Sigma}_j) d\underline{x}.$$

By completing the square, integration yields

$$c_{ij} = (2\pi)^{-L/2} |\underline{\Sigma}_i|^{-\frac{1}{2}} |\underline{\Sigma}_j|^{-\frac{1}{2}} |\underline{\Sigma}_i^{-1} + \underline{\Sigma}_j^{-1}|^{-\frac{1}{2}}$$

$$\times \exp\left(-\frac{1}{2} \left[(\underline{m}_i - \underline{m}_{ij})' \underline{\Sigma}_i^{-1} (\underline{m}_i - \underline{m}_{ij}) + (\underline{m}_j - \underline{m}_{ij})' \underline{\Sigma}_i^{-1} (\underline{m}_j - \underline{m}_{ij}) \right] \right),$$

where $\underline{m}_{ij} = (\underline{\Sigma}_i^{-1} + \underline{\Sigma}_j^{-1})^{-1} (\underline{\Sigma}_i^{-1} \underline{m}_i + \underline{\Sigma}_j^{-1} \underline{m}_j)$.

When $\underline{\Sigma}_i = \underline{\Sigma}_j = \underline{\Sigma}$, we have

$$c_{ij} = 2^{-\frac{1}{2}} (2\pi)^{-L/2} |\underline{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\underline{m}_i - \underline{m}_j)' \underline{\Sigma}^{-1} (\underline{m}_i - \underline{m}_j) \right).$$

Now, $||h(\underline{x}\ \underline{b})||^2$ simplifies when a high signal-to-noise ratio prevails. A high signal-to-noise ratio is defined as

$$(\underline{m}_i - \underline{m}_j)' (\underline{\Sigma}_i^{-1} + \underline{\Sigma}_j^{-1}) (\underline{m}_i - \underline{m}_j) > \alpha >> 1.$$

Then, the class-conditional d.f.'s are 'quasi-ortogonal'; i.e.,

$$c_{ij} = \int N(\underline{x}|\underline{m}_i, \underline{\Sigma}_i) N(\underline{x}|\underline{m}_j, \underline{\Sigma}_j) d\underline{x} \approx 0, \quad i \neq j.$$

Accordingly

$$||h(\underline{x}|\underline{b})||^2 \leq \sum_{i=1}^{M} P_i c_{ii}.$$

Thus, when the classes are widely separated, the joint density $q(\underline{b}|\underline{x}_n)$ on the parameter factors, in an approximate sense, into the M d.f.'s on the parameters for each class, the approximation improving as increases.

A clustering algorithm is obtained as follows. Let

$$\underline{r}(\underline{b}) = \frac{1}{n} \sum_{i=1}^{n} h(\underline{x}_s|\underline{b}) = \frac{1}{n} \sum_{s=1}^{n} \sum_{i=1}^{M} P_i f(\underline{x}_s|\underline{b}_i)$$

If F is Gaussian with Σ diagonal

$$\underline{r}(\underline{b}) = \sum_{i=1}^{M} \frac{1}{n} \sum_{s=1}^{n} \exp\left(-\frac{1}{2} \sum_{r=1}^{L} \frac{(x_{sr}-m_{ir})}{\sigma_{ir}} \right),$$

where m_{ir} is the r'th component of $\underline{m}$ and σ_{ir}^2 is the r'th component along the diagonal of $\underline{\Sigma}$.

Suppose $\underline{x}_s$ is a currently unclassified sample; measure the distance

$$\exp\left(-\frac{1}{2} \sum_{r=1}^{L} \frac{(x_{sr}-m_{ir})}{\sigma_{ir}^2}\right), \quad i=1,2,\ldots,M,$$

and classify $\underline{x}_s$ as in the class i having smallest distance. Then update the mean vector $\underline{m}_i$ and the covariance matrix $\underline{\Sigma}_i$ using this sample. Note a decision directed flavour in this approach.

4) Continuity Map

It is desirable that the measure d_{ij} of the dissimilarity between two vectors $\underline{x}_i$ and $\underline{x}_j$ increase with an increase in disrepancy between the two vector components $x_{i,k}$ and $x_{j,k}$ for any k. Two possible distances are the squared Euclidean distance and the 'city-block' distance.

It is desirable that the mapping from V_L to V_l be one-to-one and continuous. The one-to-one property assumes that a sample point in V_L will not map to more than one sample point in V_l. The continuous property assumes that samples 'close' in V_L are 'close' in V_l. Unfortunately a bicontinuous map from V_L to V_l, $l < L$, is, in general, impossible.

A map suggested by Shepard and Carrol attempts to obtain continuity as follows: Let the distance between $\underline{x}_i$ and $\underline{x}_j$ be defined as

$$d_{ij}^2 = \sum_{k=1}^{L} (x_{ik}-x_{jk})^2$$

and the distance between the two corresponding mapped samples in V_l be defined as

$$D_{ij}^2 = \sum_{s=1}^{l} (y_{is}-y_{js})^2.$$

A measure of continuity, considering $\underline{x}$ as a function of $\underline{y}$, in the vicinity of $\underline{y}_i$ and $\underline{y}_j$, is

$$\delta_{ij}^2 = \frac{d_{ij}}{D_{ij}}.$$

If the mapping of the samples from V_L to V_l could be achieved maintaining $\delta_{ij}^2 = 1$ for all i,j, then the properties of the catagories or clusters would not be lost. Such a juggling for all pairs seems a difficult task; nevertheless, Shepard and Carroll [44] have proposed a measure

$$\delta^2 = \sum_{i \neq j} \sum \frac{d_{ij}^2}{D_{ij}^2} w_{ij},$$

where the weight w_{ij} decreases monotonically with increasing multidimensional distances; for example

$$w_{ij} = 1/D_{ij} \quad \text{or} \quad w_{ij} = 1/d_{ij}.$$

Perhaps a better weight would be

$$w_{ij} = 1 \qquad d_{ij} < T \text{ and } D_{ij} \leq T$$
$$= 0 \qquad \text{otherwise,}$$

where T is an a priori threshold.

The object is to minimize δ^2 by adjusting the locations of points $\underline{x}_i$ in V_ℓ. Obviously a solution is to make all D_{ij} arbitrarily large. To eliminate this possibility, Shepard and Carroll suggest dividing δ^2 by

$$\sum\sum_{i \neq j} 1/D_{ij}^2$$

to obtain

$$K = \sum\sum_{i \neq j} \frac{d_{ij}^2}{D_{ij}^4} \cdot \left(\sum\sum_{i \neq j} (D_{ij}^2)^{-1}\right)^2$$

as the measure to be minimized.

5.8.3: The admissibility of the Bayesian clustering techniques.

A) All the Bayesian techniques are image admissible.

This appears clear. Reordering the sample space will not effect the calculation of the required statistics.

B) All Bayesian techniques are not convex nor connected admissible.

Bayesian techniques can produce overlapping clusters . This is especially true if the separable assumption does not hold.

C) The condition, 'well-structured exact tree', does not apply to the majority of the Bayesian techniques which do not work with a dissimilarity coefficient.

D) Bayesian techniques based on $\eta(b)$ and $\Gamma(b)$ are not well-structured (k-group) admissible.

This again results from possible overlapping clusters.

E) Techniques based on $\eta(b)$ and $\Gamma(b)$ are neither point proportion admissible or cluster omission admissible.

Duplication of one or more points or the omission of any one cluster, will adversely effect the calculation of a posteriori probabilities at the k-1'th stage. Since these probabilities are a priori in the k'th stage a different clustering may then arise.

F) Techniques based on $\eta(b)$ and $\Gamma(b)$ are cluster proportion admissible.

Duplicating eac cluster does not effect the a posteriori probabilities and the calculation of the required statistics.

G) Techniques based on $\eta(b)$ and $\Gamma(b)$ are not monotone admissible.

The required statistics are not monotone invariant.

H) It is not easy to see how optimal admissibility may be extended to the Bayesian techniques.

VI: DISCUSSION AND REMARKS ON FURTHER STUDY

Because of their conceptual simplicity, the combinatorial clustering procedures are the best known and most frequently used methods. But all the combinatorial procedures tend to be overly sensitive to 'mavericks' or 'sports' or 'outliers' or 'wildshots' and so, before, a combinatorial cluster analysis is applied some attempt should be made to remove them. Methods for the detection and removal of outliners is given in [1] and are usually based on principal components analysis.

Also, the combinatorial procedures are overly sensitive to noise in the structure. In the case of the data in the H patern, the symmetry of the pattern is nearly entirely lost due to noise. The (k,r)-method seems rather robust against such permutations and clearly shows the structure's symmetry The closely related B_k, fine, clustering, not given here, also indicates this symmetry while adding a new dimension: It seems to see the structure as consisting of two vertical bars in one plane overlying a horizontal bar in another.

It must be admitted that one of the biggest deficiencies of cluster analysis is the lack of rigorous tests for the presence of clusters and for testing the significance of those that are found. It would also be extremely useful, to have a test to tell us when it was likely to be profitable to make a cluster analysis, as opposed to some other method.

The general idea of the 'clusteriness' of a set of points is related to the concept of entropy. The analogy is not to close, because entropy measures disorder, and both a regularly spaced distribution and a clustered distribution are ordered and thus have low entropy. Also, it is not widely recognized that the degree of clustering depends on the scale of the observations. This is also true in principal component analysis on the

covariance matrix where the first principal component may only represent a size factor.

Perhaps the most difficult problem is to set up satisfactory null hypothesis. A random distribution of objects in hyperspace seems to be the most generally useful hypothesis, and this is different from the common assumption of multivariate statistics that a multivariate normal distribution is appropriate.

Partial answers to these problems are provided by the information theory and Bayesian approach. This is particularly true of the former where we have a robust, nonparametric, statistic of known distribution on which to base tests. For the Bayesian solution, the criteria $\underline{n}(b)$ and $\underline{\Gamma}(b)$, while useful, have unknown properties of robustness, and their distributions are at present unknown. In this area, the determination of tests of significance is the most wide open area for further study.

The minimum and maximum combinatorial procedures represent the two extremes in measuring the distance between clusters. In attempting to extend the concept of power to clustering methods, Baker and Hubert [2] studied these two methods and found that they are differently sensitive to particular partitions of objects imbedded in the dissimilarity values. These results indicate that the minimum procedure is likely to reject the randomness hypothesis and estimate the 'true' partition better when the 'true' partition includes a single large subset.

From a theoretical point of view, the use of partitions of objects at a specific level appears promising since it reduces the power and estimation problem to a tractable single level. Although the combinatorial problems are considerable, conceivably it would be possible to work towards a concept of power for the complete dendogram by combining the results obtained for individual levels. This area of computer simulation to compare the power

of the combinatorial procedures appears promising.

After running any method of cluster analysis, it is usually helpful to obtain some graphical representation of the groups found. One way is to find the canonical variates, and plot the groups in the space of the first two, or in combinations of any two, of these variates. A further possible way of obtaining a two-dimensional mapping after clustering, is to compute an inter-group distance matrix, and use this as input to a multi-dimensional scaling technique such as that of Kruskal [19]. Euclidean distance could be used, but perhaps better would be either Mahalanobis D^2, or Sibson's and Jardine's [17] extension of this to the case of unequal within group variance-covariance matrices.

This brings up to point of the relationship between cluster analysis and multi-dimensional scaling. The B_k, fine, clustering methods and the u-diametric method tend to yield nice low dimensional structures when the resulting groups are displayed through the use of a multidimensional scaling technique. In the Bayesian Continuity map approach, the distance measure δ^2 is the square of the stress used as a criterion in multidimensional scaling procedures. An investigation of the relationship of cluster analysis to ordination techniques would prove fruitful but would be difficult because of the topological and differential geometry involved.

Application of the admissibility criteria to the clustering procedures reported here, the minimum method and the (k,r)-method look good. This agrees with previous results reported by Jardine and Sibson. It becomes apparent that these conditions are not entirely satisfactory. First, the overlapping techniques are treated unfairly which suggests that a different set of criteria be developed to handle this particular case. Second, the conditions provided no globally optimal best method. Although the application of other

criteria tend to suggest that the family of uniform clustering techniques are globally optimal among combinatorial procedures with probable local optimality achieved by the (k,r)-method by virtue of its indifference to noise.

The fitting of mixtures of multivariate normal distributions using such programs as Normap and Mormix, developed by Wolfe, [53, 54] may be extremely useful in many situations and the sequence of likelihood ratio tests for the number of groups which attend these methods is possibly the best procedure available given the assumption of normality. The method also has the considerable advantage in that it does not rely on an arbitrary choice of dissimilarity or distance measure. Unfortunately, because of the large number of parameters to be estimated, these methods ideally require large sets of data, and, in general, they consume large amounts of computer time. Also, the development of the theory of mixed non-normal distributions is still primitive. Moreover, the problem of local maxima of the likelihood function arises, and several runs using different initial estimates of the parameters are desirable. All other clustering techniques are potential sources of initial estimates.

Cluster analysis is potentially a very useful technique, but it requires care in its application, because of many associated problems. In many of the applications that have been reported in the literature, the authors have either ignored or have been unaware of these problems and, consequently, few results of lasting value can be pointed to. Hopefully, future users of the techniques will adopt a more cautious approach and, in addition, remember that, along with most statistical techniques, classification procedures are essentially descriptive techniques for multivariate data, and solutions given should lead to a proper re-examination of the data matrix rather than a mere acceptance of the clusters produced.

In this paper, an attempt has been made to review techniques of cluster analysis, and to describe and illustrate problems associated with them. Because of the ever-growing volume of relevant literature, any review of the field is likely to be out of date before it is started. However, it is hoped that this paper will serve some purpose, if only to dissuade people from using uncritically the nearest clustering program available.

APPENDIX I

AN OVERVIEW OF THE THEORY OF UNSUPERVISED BAYESIAN ESTIMATION

i) Convergence Theorems.

Unsupervised estimation arises in classes of problems including nonstationary class probalilities, statistically dependent measurement vaectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, and unknown synchronization (for waveforms, and, sometimes, in the case of images, called unknown registration). The general problems were formulated in a Bayesian minimum-conditional-risk framework by Patrick 28 , with the mixture concept empahasized. Combined with similar work by Hilborn and Lainiotis , this provides a precise formal definition of the problem.

Implicit in the solution of unsupervised estimation problems is the concept of identifiability - that there should be a one-to-one mapping or relationship between a set of mixing parameters and the resulting mixtures. Teicher's work on finite mixtures was reduced by Yakowitz and Spragins 57 to a sufficiency theorem that a necessary and sufficient condition for the identifiability of a class of mixtures is the linear independence of the density functions in each finite mixture. They showed also that a large number of parametric families (including Gaussian) are identifiable.

The Bayes estimator for the true parameter $\underline{b}^*$ characterizing $h(\underline{x})$ computes the a posteriori density of each point $\underline{b}^k$ in $B^{M'}$ using Bayes theorem,

$$p(\underline{b}^k|\underline{x}_1) = \frac{[\sum_{i=1}^{M_k} f(\underline{x}_1|\underline{b}_i^k)P_i^k]\, p_0(\underline{b}^k)}{\sum [\text{numerator}]}, \qquad 1)$$

$$k=1,2,\ldots,V, \quad \text{all } \underline{b}^k \in B^{M'}$$

given one sample $\underline{x}_1$, an a priori density $\{p_0(\underline{b}^k)\}_{k=1}^{V}$ and the familty F of densities. For a given sequence of n sample $\dot{\underline{x}}_n$, an a priori density, and the family F, the a posteriori density is

$$p(\underline{b}^k|\dot{\underline{x}}_n) = \frac{\left[\sum_{i=1}^{M_k} f(\dot{\underline{x}}_n|\underline{b}_i^k)P_i^k\right] p(\underline{b}^k|\dot{\underline{x}}_{n-1})}{\sum [\text{numerator}]}$$

$$k=1,2,\ldots,V$$

Samples $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ are assumed parameter conditionally independent and the prior probabilities P_i^k are assumed independent of the samples.

Implementation of the above, requires that $B^{M'}$ be a finite set of V vector points $\{\underline{b}^k\}$. For a quadratic loss function and this discritized parameter space, the Bayes estimator is

$$(\underline{b})_n = \sum_{k=1}^{V} \underline{b}^k p(\underline{b}^k|\dot{\underline{x}}_n)$$

with $p(\underline{b}^k|\dot{\underline{x}}_n)$ calculated as above.

Denoting the true mixture density of $h(\underline{x})$, we define

$$\underline{n}(\underline{b}^k) = E[\ln(h(\underline{x}|\underline{b}^k))] = \int [\ln(h(\underline{x}|\underline{b}^k))]h(\underline{x})d\underline{x} .$$

We now show how the convergence properties of $p(\underline{b}^k|\underline{x}_n)$ and $(\underline{b})_n$ depend on $\underline{n}(\underline{b}^k)$, which is a measure of the projection of $\ln(h(\underline{x}|\underline{b}^k))$ onto $h(\underline{x})$.

The convergence properties of the Bayes estimator on a finite $B^{M'}$ were discovered by Patrick [29] and are listed below. The following assumptions are necessary:

I) $h(\underline{x}_n|\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_{n-1},\underline{b}) = h(\underline{x}_n|\underline{b})$,

II) There exists an integer $s > 1$ such that

$$E[|\ln(h(\underline{x}|\underline{b}^k))|^s] < \infty \quad \text{for all } \underline{b}^k \in B^{M'}$$

III) The probability measures corresponding to $h(\underline{x}|\underline{b}^k)$ are absolutely continuous with respect to the Lebesque measure μ .

IV) $\mu[x \mid |h(\underline{x}|\underline{b}^k) - h(\underline{x}|\underline{b}^j)| > \theta\ ] > 0$ for all $\underline{b}^k, \underline{b}^j$ $j \neq k$.

V) $h(\underline{x}|\underline{b}^k)$ contains the true mixture $h(\underline{x})$.

VI) The a priori probabilities $p_0(\underline{b}^k)$ are nonzero.

Theorem I:1: Conditions III, IV and V imply

a) $\underline{b}^* = \arg\left[\max_{\underline{b}^k \epsilon B^{M'}} \eta(\underline{b}^k)\right]$ is unique.

In addition conditions I, II and VI imply

b) $p\left[\lim_{n\to\infty} (\underline{b})_n = \underline{b}^*\right] = 1$,

c) $(c < \infty)(n)\ E[||(\underline{b})_n - \underline{b}^*||^2] \leq cn^{-S^*/2}$.

Corollary: If condition I in the previous theorem is satisfied and if in addition

$$(\underline{b}^k \epsilon B^{M'})(c > \infty) \quad \sup_{x \epsilon B} |\ln(h(\underline{x}|\underline{b}^k))| \leq c,$$

$$R = \max_k ||\underline{b}^k - \underline{b}^*||$$

then

$$E[||(\underline{b})_n - \underline{b}^*||] \leq R^2(V-1)2p^{n^{\frac{1}{2}}}.$$

Let $\underline{x}$ be an p-dimensional observation having denisty $h(\underline{x}|\underline{b}^*)$. Individual observations $\underline{x}$ are denoted $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ and are parameter individually independent and identically distributed.

Assumptions:

I) The function $\eta(\underline{b}) = E[\ln(h(\underline{x}|\underline{b}))]$ is uniquely maximum at $\underline{b} = \underline{b}^*$; i.e. $\underline{b} \neq \underline{b}^*$ $\eta(\underline{b}) < \eta(\underline{b}^*)$.

II) A practical bound on B is available; i.e., it is known that $\underline{b} \epsilon B' - B$ and that for all $\underline{b} \epsilon B'$,

$$||\underline{b} - \underline{b}^*|| \leq R < \infty.$$

III) $E\,(\ln(h(\underline{x}|\underline{b})) - \underline{\eta}(\underline{b}))^2 < \infty, \quad \underline{b}$.

Theorem I:2: Let ϵ be an arbitrarily small positive constant and let G_ϵ denote an ϵ neighbourhood of $\underline{b}^*$. If for the family H, conditions I to III hold, then for n sufficiently large,

$$(\delta > 0)(k) \qquad \int_{B - G_\epsilon} p(\underline{b}|\underline{\dot{x}}_n)\,d\underline{b} < k \exp(-n2\delta).$$

Thus the 'time constant' of the convergence rate is dependent upon the magnitude of the gradient of $\underline{\eta}(\underline{b})$ near $\underline{b}^*$; the rate is enhanced the more sharply peaked the function $\underline{\eta}$ is at $\underline{b}^*$. This result also shows that the rate depends, through k, upon the tightness of the bound on B.

Corollary: Let $||\underline{b} - \underline{b}^*|| < R$, then for the Bayes estimator

$$(\underline{b})_n = \int_B \underline{b}\,p(\underline{b}|\underline{\dot{x}}_n)\,d\underline{b} ,$$

$$\sigma_n^2 = E\,||(\underline{b})_n - \underline{b}^*||^2 < kR(R+\epsilon)\exp(-n2\delta)+\epsilon^2.$$

The Bayes estimator $(\underline{b})_n$ minizing average risk on B for a quadratic loss function is defined as

$$(\underline{b})_n = \sum_{r=1}^{V} \underline{b}^r p(\underline{b}^r|\underline{\dot{x}}_n) ,$$

where $p(\underline{b}^r|\underline{\dot{x}}_n)$ is the a posteriori probability mass on $\underline{b}^r$ computed as

$$p(\underline{b}^r|\underline{\dot{x}}_n) = \frac{\prod_{k=1}^{n} h(\underline{x}_k|\underline{b}^r)}{\sum_{r=1}^{V}\prod_{s=1}^{n} h(\underline{x}_s|\underline{b}^r)} , \quad r=1,2,\ldots,V.$$

The following set of results show that $p(\underline{b}^r|\underline{\dot{x}}_n)$, $r \neq m$, diminishes to zero exponentially for large enough n with probability zero. Then, that the Bayes estimator on a finite set is asymptotically superefficient; i.e., it has variance smaller than $O(1/n)$. An estimator can be superefficient only

on a set of parameters of Lebesque zero.

Theorem I:3: If condition III holds, then the a posteriori probability mass $p(\underline{b}^r|\underline{\dot{x}}_n)$ is bounded for n sufficiently large by

$$p(\underline{b}^r|\underline{\dot{x}}_n) < \exp(-n\delta_r), \qquad r \neq m,$$

with probability 1, where $\delta_r = \frac{1}{3}[\underline{n}(\underline{b}^m) - \underline{n}(\underline{b}^r)]$.

Corollary: For $n \geq \max_r\{\delta_r\}$, the mean norm-square-error

$$\sigma_n^2 \triangleq E\,[||(\underline{b})_n - \underline{b}^m||^2],$$

is bounded above by

$$\sigma_n^2 < \sum_{j\neq m} \sum_{k\neq m} (\underline{b}^i - \underline{b}^m)\cdot(\underline{b}^k - \underline{b}^m)\exp(-n\delta_j)$$

$$\sigma_n^2 < C\exp(-n\delta'),$$

with probability 1, where C is a positive constant and

$$\delta' = < \min_{j\neq m}\{\delta_j\}$$

This last implies that the Bayes estimator converges in L^2 to the point in B^V at which $\underline{n}$ is greatest. However, for some family H and some $\underline{b}^*$, $\underline{n}$ may be multimodal and $\underline{b}^m$ may not be close to $\underline{b}^*$.

The form of the Bayes a posteriori probability suggests that there may be other product-type functions which can be utilized similar to that of Bayes. One such function results in the so-called minimum-norm-square-error estimator, denoted $(\underline{a})_n$, is defined on B' as

$$(\underline{a})_n = \int_{B'} \underline{b}\; q(\underline{b}|\underline{\dot{x}}_n)\,d\underline{b},$$

where

$$q(\underline{b}|\underline{\dot{x}}_n) = \frac{\sum_{s=1}^{n} 2h(\underline{x}_s|\underline{b}) - ||(\underline{x}\;\underline{b})||^2}{\int_{B'}[\text{numerator}]\,d\underline{b}}$$

and the norm is defined as

$$||f(\underline{x})||^2 = \int f^2(\underline{x})d\underline{x} .$$

The norm-square-error between $h(\underline{x}|\underline{b})$ and the true density function $h(\underline{x}|\underline{b}^*)$ is defined to be

$$e^2(\underline{b}) = ||h(\underline{x}|\underline{b}) - h(\underline{x}|\underline{b}^*)||^2$$
$$= ||h(\underline{x}\ \underline{b})||^2 - 2E[\ h(\underline{x}|\underline{b})\ ] + ||\ h(\underline{x}|\underline{b}^*)||^2.$$

This can be described as the expected fractional squared error,

$$e^2(\underline{b}) = \int [\ \frac{h(x|b) - h(x|b^*)}{h(x|b^*)}\ ]\ h(\underline{x}|\underline{b}^*)d\underline{x}$$

$$= E\ [\ \frac{h(x|b) - h(x|b^*)}{h(x|b)}\ ]$$

We define $\Gamma(\underline{b})$ as

$$\Gamma(\underline{b}) \triangleq 2E[h(\underline{x}|\underline{b})] - ||h(\underline{x}|\underline{b})||^2$$
$$= ||h(\underline{x}|\underline{b}^*)||^2 - e^2(\underline{b}) .$$

Theorem I:4: Let G_ε denote an ε neighbourhood of $\underline{b}^*$. If $h(\underline{x}|\underline{b})$ has finite means for all $\underline{b} \varepsilon B'$, then with probability 1, for n sufficiently large,

$$\int_{B'-G_\varepsilon} q(\underline{b}|\underline{x}_n)d\underline{b} < K\exp[-\tfrac{1}{2}(\ e^2(\underline{b}') - e^2(\underline{b}^\varepsilon))\ ],$$

where $\underline{b}' \varepsilon B'-G_\varepsilon$, $\underline{b}^\varepsilon \varepsilon G_\varepsilon$ and K is a finit number.

Theorem I:5: For n sufficiently large,

$$p(\underline{b}^r|\underline{x}_n) < \exp[\ -n[e^2(\underline{b}^r) - e^2(\underline{b}^m)\ ]],\quad r \neq m$$

with probability 1.

Corollary: Define the expected norm-square-error in $(\underline{a})_n$ by $\sigma^2(n)$; i.e.,

$$\sigma^2(n) \triangleq E\,||(\underline{a})_n - \underline{b}^m||^2.$$

For n sufficiently large,

$$\sigma^2(n) < C\exp\{-n\,(e^2(\underline{b}') - e^2(\underline{b}^m))\},$$

with probability 1, where $0 < C < \infty$, $\underline{b}' \in \{\underline{b}^r\}_{r\neq m}$.

This shows that the convergence rate depends on both how close $h(\underline{x}|\underline{b}^r)$, $r\neq m$, is to $h(\underline{x}|\underline{b}^*)$ rlative to $h(\underline{x}|\underline{b}^m)$, and also upon how close $h(\underline{x}|\underline{b}^*)$, in the norm-square sense. That is, the probability mass at $\underline{b}^r$, $r\neq m$, will diminish to zero faster the closer $\underline{b}^m$ is to $\underline{b}^*$.

Theoretical results that compare the convergence rate using $\Gamma(\underline{b})$ versus the rate using $\eta(\underline{b})$ are not available. Although such results would be useful it must be remembered that pragmatic assumptions may lead to estimators that make such teoretical results unnecessary.

The regression function Γ has appeal over η because the former involves $h(\underline{x}|\underline{b})$ rather than $\ln(h(\underline{x}|\underline{b}))$ and yields better results in situations under a high signal-to-noise assumption.

The average norm-square-error in $(\underline{b})_n$ is

$$\sigma^2(n) \triangleq E||(\underline{b})_n - \underline{b}^*||^2$$

$$= \sum_{r=1}^{V} \sum_{t=1}^{V} E[(\underline{b}^r)_n - \underline{b}^*]'[(\underline{b}^t)_n - \underline{b}^*]$$

$$p((\underline{b}^r)_0|\underline{x}_n)\,p((\underline{b}^t)_n|\underline{x}_n)$$

This yields

$$\sigma^2(n) = \sum_{\substack{r=1 \\ r,t\neq m}}^{V} \sum_{t=1}^{V} E[(\underline{b}^r)_n - (\underline{b}^m)_n]'[(\underline{b}^t)_n - (\underline{b}^m)_n]$$

$$p((\underline{b}^r)_0|\underline{x}_n)\,p((\underline{b}^t)_0|\underline{x}_n)$$

$$+2E[(\underline{b}^m)_0 - \underline{b}^*]' \sum_{r\neq m}^{V} [(\underline{b}^r)_n - (\underline{b}^m)_n]$$

$$\times p((\underline{b}^r)_0|\underline{x}_n)$$

$$+ E||(\underline{b}^m)_n - \underline{b}^*||^2$$

Since the parameter set is bounded,

$$||(\underline{b}^r)_n - (\underline{b}^m)_n|| \leq R < \infty .$$

By the Schwarz inequality and letting

$$\sigma^2(n) \triangleq E\,||(\underline{b}^m)_n - \underline{b}^*||^2 ,$$

we have

$$\sigma^2(n) < (V-1)\,R^2 \sum_{r\neq m}^{V} E[p((\underline{b}^r)_0|\underline{x}_n)]$$

$$+ 2R\sigma_m^2(n)\; E \sum_{r\neq m}^{V}[p((\underline{b}^r)_0|\underline{x}_n)] + \sigma_m^2(n)$$

$$< (V-1)\,R^2 \sum_{r\neq m}^{V} E[p((\underline{b}^r)_0|\underline{x}_n)]$$

$$+ 2R\sigma_m^2(n) \sum_{r\neq m}^{V} E[p((\underline{b}^r)_0|\underline{x}_n)] + \sigma_m^2(n).$$

It can be shown that, for large n, $p((\underline{b}^r)_0|\underline{x}_n) < 0\ \exp(-nd_r)$, a.e., whereas by the Rao-Cramer lower bound, $\sigma^2_m(n) > O(n^{-1})$, so that for large n, $\sigma^2(n) \simeq \sigma^2_m(n)$. That is, the performance of the quasi-Bayes estimator is asymptotically indistinguishable from that of the s.a. estimator, which is best in the sense of having the starting point at which the regression function is the greatest.

An alternative way of implementing the averaging technique is to use $\Gamma(\underline{b})$, and then form the average with $q((\underline{b}^r)_0|\underline{x}_n)$, $1 \leq r \leq V$, as the weighting

coefficients. Specifically, the i'th component of $(\underline{b}^r)_{n+1}$ is computed as

$$(b_i^r)_{n+1} = (b_i^r)_n + \frac{a_n}{c_n}(u^i_{r,2n} - u^i_{r,2n-1}),$$

$u^i_{r,2n}$ and $u^i_{r,2n-1}$ being noisy measurements of the regression function $\underline{\Gamma}$;i.e.,

$$u^i_{r,2n} = 2h(\underline{x}_n | (\underline{b}^r)_n + c_{n-1}\underline{e}_i) - ||h(\cdot|(\underline{b}^r)_n + c_{n-1}\underline{e}_i)||^2,$$

$$u^i_{r,2n-1} = 2h(\underline{x}_n | (\underline{b}^r)_n - c_{n-1}\underline{e}_i) - ||h(\cdot|(\underline{b}^r)_n - c_{n-1}\underline{e}_i)||^2.$$

The average is then computed as

$$(\underline{b})_{n+1} = \sum_{r=1}^{V} (\underline{b}^r)_{n+1} q((\underline{b}^r)_0 | \underline{x}_n).$$

It may be concluded the average estimator singles out the s.a. estimator whose starting point is the best, in the sense that the corresponding density function is 'closest' in the norm-square-error sense to the true density.

The merit of the Bayes approach is that it shows $p(\underline{b}|\underline{x}_n)$ must be computed for each point $\underline{b}$ in the parameter space. Because this a posteriori density is computed at each point $\underline{b}$, the Bayes estimator $(\underline{b})_n = \int \underline{b}\, p(\underline{b}|\underline{x}_n)dP$ need not be staring point-dependent. Stochastic approximation is, on the other hand, starting point-dependent; given n iterations in the stochastic approximation algorithm, the estimator's performance may be expected to be poor.

Denote the estimate of $\underline{\Gamma}(\underline{b})$ based on $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ by

$$\underline{\Gamma}_n(\underline{b}) = \frac{1}{n} \sum_{s=1}^{N} \{2h(\underline{x}_s | \underline{b}) - h(\cdot|\underline{b})\}$$

$$= \frac{2}{n} \sum_{s=1}^{n} h(\underline{x}_s|\underline{b}) - h(\cdot|\underline{b}).$$

Let e be some prespecified allowable error in the parameter estimate and let c and d be positive constants with c smaller than e. Denote by $(\underline{b})_{n,k}$ the k'th recursive estimate of $\underline{b}^*$ based on n observations. Then

$$(\underline{b}_j)_{n,k+1} = (\underline{b}_j)_{n,k} + \frac{d}{c}\underline{\Gamma}((\underline{b})_{n,k} + c\underline{e}_j) - \underline{\Gamma}((\underline{b})_{n,k})$$

The algorithm increments the estimator in the direction of the gradient

by an amount proportional to the gradient. The iterations on k terminate when

$$||(\underline{b})_{n,k+1} - (\underline{b})_{n,k}|| << e.$$

For large L, it may be impossible to search over all possible values of $\underline{b}$. The algorithm may have application as a follow-up to clustering where the starting vector-point $(\underline{b}_j)_{n,1}$ has been determined using a clustering algorithm.

It must be emphasized that the estimator resulting from a stochastic hill climb is generally not as good as the Bayes estimator. At best, the estimator $(\underline{b})_n$, is the solution maximizing $\Gamma_n(\underline{b})$ of $\eta_n(\underline{b})$. The Bayes estimator evaluates

$$\eta_n(\underline{b}) = \frac{1}{n} \sum_{s=1}^{n} \ln(h(\underline{x}_s|\underline{b})$$

for each point $\underline{b}$ in the parameter space and then takes an average.

ii) A Class of Minimum- Integral-Square-Distance Algorithms

The class of minimum-integral-square-difference algorithms requires the restriction of A to a finite set $\{\underline{\alpha}\}_{k=1}^{N}$ of N vector points.

We are interested in unbiased estimators of the mixing parameters of F which are contained in $L^1 L^2$ and which are optimum in the sense that they minimize the squared norm of the difference between the estimated mixture function us ing a finite family of functions and either the hitogram mixture funtion or the empirical mixture function. The unbiased vector estimator

$$P_n = (P_n(\underline{\alpha}^1), P_n(\underline{\alpha}^2), \ldots, P_n(\underline{\alpha}^n)),$$

which maximizes

$$\iint |c_n(\underline{x}) \sum_{i=1}^{n} f(\underline{x}|\underline{\alpha}^i) P_n(\underline{\alpha}^i)||^2 d\underline{x}, \qquad 1)$$

is an example. Robbins' functions will be used in two simple stochastic approximation algorithms that estimate mixing parameters.

Although the family of distributions is not in $L^1 L^2$,

$$(\gamma_0 > 0)(\gamma > \gamma_0, F^* = \{F^*(\underline{x}|\underline{\alpha}^i)\}_{i=1}^{n}),$$

$$F^*(\underline{x}|\underline{\alpha}^i) = F(\underline{x}|\underline{\alpha}^i), \qquad ||x|| < \gamma$$
$$= 0, \qquad \text{otherwise}$$
$$x \varepsilon V_L, \; i=1,2,\ldots,n$$

is in $L^1 L^2$ and composed of linearly independent functions.

Taking the set of regions $\{I_j\}_{j=1}^{n^*}$ define a new family of functions

$$d(\underline{x}|\underline{\alpha}^i) = \sum_{j=1}^{n^*} \chi_{I_j}(\underline{x}) b_{ij}, \qquad i=1,2,\ldots,n,$$

where $$b_{ij} = \int_{I_j} f(\underline{x}|\underline{\alpha}^i) d\underline{x}, \qquad i=1,2,\ldots,n; j=1,2,\ldots,n^*.$$

The corresponding family of cumulative densities may be defined as

$$D(\underline{x}|\underline{\alpha}^i) = \sum_{t=1}^{n^*} \chi_{I_t}(\underline{x}) \sum_{j=1}^{t} b_{ij}, \quad i=1,2,\ldots,n$$

The functions $\{D(\underline{x}|\underline{\alpha}^i)\}_{i=1}^{n}$ are not in $L^1\ L^2$. But, we can retain their important properties in the cases of interest by requiring that

$$I_{n^*} = \{\underline{x} \mid ||x|| < \gamma, \ \underline{x}\epsilon V_L\},$$

$$b_{in^*} = 0, \quad i=1,2,\ldots,n.$$

The resulting family is in $L^1\ L^2$.

The functions $\{f(\underline{x}|\underline{\alpha}^i)\}_{i=1}^{M'}$ from F span an M'-dimensional subspace U in $L^1\ L^2$. Let $U_j^{\perp}$ be the orthogonal complement of this subspace. Define the function

$$\phi^i(\underline{x}) = \frac{f^{\perp}(\underline{x}|\underline{\alpha}^i)}{||f^{\perp}(\underline{x}|\underline{\alpha}^i)||^2}, \quad i=1,2,\ldots,M'.$$

These functions are called 'Robbin's functions' and have the property that

$$\int \phi^j(\underline{x})\ f(\underline{x}|\underline{\alpha}^i)\ d\underline{x} = \delta_{ij}.$$

Assuming that the mixture density function $h(\underline{x})$ contains only functions in $\{f(\underline{x}|\underline{\alpha}^i)\}$,

$$\int \phi^j(\underline{x})\ h(\underline{x})d\underline{x} = \int \phi^t(\underline{x}) \sum_{j=1}^{M'} f(\underline{x}|\underline{\alpha}^j)\ P(\underline{\alpha}^j)\ d\underline{x}$$

$$= P(\underline{\alpha}^i), \quad i=1,2,\ldots,M'.$$

Let $\underline{\alpha}^{i(j,k)}$ be one of the parameter points $\underline{\alpha}^1, \underline{\alpha}^2, \ldots, \underline{\alpha}^n$. If each subset is of size M', there are $\binom{n}{M'}$ subsets of parameters, and k denotes the k'th subset. The last equation suggests the following algorithm for estimating $P(\underline{\alpha}^{i(j,k)})$, the j'th mixing parameter for the k'th subset.

ALGORITHM 1:

$$P_n(\underline{\alpha}^{i(j,k)}) = \int \phi^{i(j,k)}(\underline{x})\ c_n(\underline{x})\ d\underline{x}, \quad i=1,2,\ldots,M'.$$

where c_n is the empirical probability density function of the samples. This

suggests the basis of the second algorithm due to Robbins.

$$P_n(\underline{\alpha}^{i(j,k)}) = \frac{1}{n} \sum_{s=1}^{n} \phi^{i(j,k)}(\underline{x}_s), \quad i=1,2,\ldots,M',$$

or in terms of the last estimate of $P(\underline{\alpha}^{i(j,k)})$,

ALGORITHM 2:

$$P_n(\underline{\alpha}^{i(j,k)}) = \frac{n-1}{n} P_{n-1}(\underline{\alpha}^{i(j,k)}) + \frac{1}{n} \phi^{i(j,k)}(\underline{x}_n), \quad i=1,2,\ldots,M'.$$

We now show that the systems resulting from algorithms 1 or 2 minimize 1), as required. Taking the k'th subfamily of F as before, the estimated mixture density can be factored relative to $\{f(\underline{x}|\underline{\alpha}^{i(j,k)})\}$ into $c'_{n_k}(\underline{x}) + c^{\perp}_{n_k}(\underline{x})$.

Then the integral mean square difference becomes,

$$\int [c_n(\underline{x}) - \sum_{j=1}^{MJ} f(\underline{x}|\underline{\alpha}^{i(j,k)})]^2 d\underline{x}$$

$$= \int [c'_{n_k}(\underline{x}) + c^{\perp}_{n_k}(\underline{x}) - \sum_{j=1}^{M'} f(\underline{x}|\underline{\alpha}^{i(j,k)})]^2 \, d\underline{x}$$

$$= \int ||c^{\perp}_{n_k}(\underline{x})||^2 \, d\underline{x}.$$

And the integral square difference is minimized relative to the $\{f(\underline{x}|\underline{\alpha}^{i(j,k)})\}$ subfamilies.

The second algorithm maximizes the weighted likelihood function over $B^{Ma'}$. Define the estimator

$$(\hat{\underline{b}})_{n,a} = \arg \left[\max_{b_k} \left[\Pi_{j=1}^{n} h(\underline{x}_j|\underline{b}^k) P_0(\underline{b}^k) \right] \middle| \, \underline{b}^k \in B^{Ma'} \quad B^{M'} \right]$$

where $P_0(\cdot)$ is the a priori parameter density function on $B^{M'}$. In what follows

let

$$E_{\underline{b}^*}[g(\underline{x})] \triangleq \int g(\underline{x})h(\underline{x}|\underline{b}^*)\,d\underline{x},$$

where $h(\underline{x}|\underline{b}^*)$ is the true mixture function.

Theorem I:6: If

i) $B^{M'}$ is closed and $P_0(\underline{b})$ is continuous on $B^{M'}$ with $P_0(\underline{b}) > 0$,

ii) $h(\underline{x}|\underline{b})$ is jointly measurable $[\mu]$ in $\underline{x},\underline{b}$.

iii) the first and second-order partials of $\ln H(\underline{x}|\underline{b})$ with respect to the components Θ_k of $\underline{b}$ exist and are continuous,

iv) $E_{\underline{b}^*}[\sup[\partial^2 \ln h(\underline{x}|\underline{b})/\partial\Theta_i] \mid ||\underline{b}-\underline{b}^*|| < \varepsilon, \underline{b} \in B^{M'}] < \infty$ for some $\varepsilon > 0$; conditions iii) and iv) imply that

$$E_{\underline{b}^*}\left[\left.\frac{\partial \ln h(\underline{x}|\underline{b})}{\partial\Theta_i\,\partial\Theta_j}\right|_{\underline{b}^*}\right] = 0,$$

$$c_{ij}(\underline{b}^*) = -E_{\underline{b}^*}\left[\left.\frac{\partial^2 \ln h(\underline{x}|\underline{b})}{\partial\Theta_i\,\partial\Theta_j}\right|_{\underline{b}^*}\right]$$

$$= E_{\underline{b}^*}\left[\left.\frac{\partial \ln h}{\partial\Theta_i}\,\frac{\partial \ln h}{\partial\Theta_j}\right|_{\underline{b}^*}\right]$$

v) $c(\underline{b}^*)$ is positive definite,

vi) $E_{\underline{b}^*}[\sup[\ln h(\underline{x}|\underline{b}) - \ln h(\underline{x}|\underline{b}^*)] \mid ||\underline{b}-\underline{b}^*|| > \varepsilon, \underline{b} \in B^{M'}] < \infty$ for some $\varepsilon > 0$,

vii) a bound M' on the number of active classes is known,

viii) $\{h(\underline{x}|\underline{b}^k) \mid \underline{b}^k \in B^{M'}\}$ is identifiable,

then

$$P\left[\lim_{\substack{n\to\infty \\ a\to\infty}} (\underline{b})_{n,a} = \underline{b}^*\right] = 1.$$

iii) The Construction of Robbins' Functions:

Let M'=n and denote the Robbins' functions by $\{ \phi^k \}$. From their definition, the ϕ^k functions are linear combinations of $\{ f(\underline{x}|\underline{\alpha}^i) \}_{i=1}^n$,

$$\phi^k(\underline{x}) = \sum_{i=1}^{n} a_{ki} f(\underline{x}|\underline{\alpha}^i). \qquad 1)$$

Let Q denote an n×n matrix and let

$$q_{ij} = \int f(\underline{x}|\underline{\alpha}^i) f(\underline{x}|\underline{\alpha}^j) d\underline{x} .$$

The by orthogonality, $AQ = I$

$$A = Q^{-1}.$$

The row vectors of A give the coefficients for i).

If

$$P_n = [\frac{1}{n} \sum_{s=1}^{n} \phi^1(\underline{x}_s), \ldots, \frac{1}{n} \sum_{s=1}^{n} \phi^n(\underline{x}_s)]'$$

is the vector of mixing parameter estimates based on n samples and

$$P_0 = [P(\underline{\alpha}^1), \ldots, P(\underline{\alpha}^n)]$$

is the vector of true mixing parameters, the mean-square error is

$$E[||P_n - P_0||^2] = \frac{1}{n} \sum_{k=1}^{n} E [\phi^k(\underline{x}) - P(\underline{\alpha}^k)] ,$$

using the results $E[\phi^k(\underline{x})] = P(\underline{\alpha}^k)$ and $\phi^k(\underline{x}_s)$ is statistically independent of $\phi^k(\underline{x}_j)$ for $j \neq s$.

For many problems representing the parameter space with the finite set $\{ \underline{\alpha}^i \}$ may be an approximation. Nevertheless, with probability 1, the values obtained by sequential approximation of the algorithm caonverge at a rate O(1/n) although not necessarily to the true parameter. Let $\sum_{i=1}^{M} f_i(\underline{x}) P_i$ denote the true mixture. Then the estimator for $P(\underline{\alpha}^{i(j,k)})$ has a limit,

$$\int \phi^{i(j,k)}(\underline{x}) \sum_{s=1}^{M} f_s(\underline{x}) P_s d\underline{x} = \sum_{s=1}^{M} P_s \int \phi^{i(j,k)}(\underline{x}) f_s(\underline{x}) d\underline{x}.$$

And if $h(\underline{x}) \notin \text{span} \{ f(\underline{x}|\underline{\alpha}^{i(j,k)}) \}$, then, with probability 1,

$$\lim_{n} \sum_{i=1}^{n} P_n(\underline{\alpha}^i) < 1.$$

APPENDIX II

THE DISSIMILARITY MATRIX FOR THE H-CONFIGURATION

—	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.0.0												
2	22.0833	0.0											
3	5.0329	42.4900	0.0										
4	28.8362	0.5611	50.3038	0.0									
5	1.8883	12.5818	8.8966	17.2761	0.0								
6	25.7924	0.4598	49.1214	0.9811	16.2088	0.0							
7	4.4581	8.8341	12.7279	12.4252	0.5450	12.3526	0.0						
8	5.1827	24.3883	18.1443	32.3460	9.2520	25.2540	12.7876	0.0					
9	25.0713	8.5882	34.7179	7.8889	13.2961	12.8075	8.5974	40.6870	0.0				
10	2.5599	24.9013	12.4152	32.8309	6.5877	26.7632	10.1712	0.5418	37.0370	0.0			
11	32.4667	1.0707	57.0305	0.5387	20.9499	0.4975	16.0522	32.8199	12.5507	34.3853	0.0		
12	12.3455	5.0254	22.7022	6.5161	4.5788	8.4373	1.9663	22.8285	2.5633	20.1945	10.1627	0.0	
13	0.4980	19.6226	8.5003	26.4111	2.4418	22.3773	4.9741	2.6551	26.6105	1.0544	29.0056	12.8942	0.0
14	0.5073	26.1002	2.4562	32.9128	2.5149	30.7752	5.2163	8.6383	25.1627	4.9872	37.5853	13.2492	2.0066
15	24.0059	5.1501	36.4010	4.5645	12.4301	8.4644	7.7743	36.6432	0.4557	33.9824	8.2389	1.9211	24.5953
16	4.4433	17.4529	18.3884	24.2696	6.4195	18.3085	8.8357	0.5806	32.3447	1.0610	24.8058	16.7420	1.9664
17	8.4537	24.6418	24.2565	32.5921	12.5004	24.6051	15.9606	0.4432	44.6698	1.9646	32.0893	25.9561	4.9704
18	8.0769	18.2618	25.0646	25.1605	10.1938	18.1272	12.6153	1.0045	37.3603	2.5655	24.6192	20.6403	4.5637
19	28.6858	18.7711	32.1600	18.0907	16.9651	24.9497	12.4333	50.4379	2.0892	44.6612	24.8729	6.5084	32.2633
20	1.9861	30.1553	1.0570	36.8691	3.8220	35.7725	6.4980	13.2128	25.1884	8.5153	42.5683	14.3792	4.4739
21	19.5785	12.5952	25.0121	13.0173	9,8708	17.8577	6.3507	37.1448	1.1087	32.4330	18.7367	2.4508	22.1269
22	26.1266	12.9418	32.7787	12.2763	14.4229	18.1052	9.8288	44.6702	0.4847	39.9969	17.9582	3.8910	28.6463
23	1.0051	26.4024	7.7661	34.2890	4.9641	29.2606	8.5808	2.1696	34.3558	0.5430	36.9250	18.5588	0.5396
24	2.4365	36.3713	0.5190	44.1300	6.1891	42.0257	9.8924	12.5421	32.5771	7.8706	49.8023	19.7292	4.8906
25	0.4782	16.7798	6.5536	22.4970	0.4662	20.4322	2.0161	6.7097	18.6879	4.0828	26.1325	7.9640	0.9881
26	24.6930	2.5566	40.2085	1.9704	13.1061	4.8779	8.3745	34.2086	1.9758	32.6272	4.5695	2.4852	24.2487
27	1.0654	32.5070	2.0060	40.3781	5.0148	37.2218	8.7572	8.0844	32.7483	4.4404	45.0257	18.7715	2.5337
28	18.3940	8.2230	26.7556	8.6204	8.6315	12.5601	5.0171	33.0806	0.5393	29.3723	13.3392	1.0385	19.9807
29	19.7830	0.5580	37.0996	1.0334	10.1345	2.0241	6.3791	25.3526	4.8683	24.7673	2.6382	2.4455	18.3757
30	4.6754	36.4288	0.5026	43.1817	6.5933	43.0681	9.3850	19.0687	27.2768	13.2636	49.9852	17.3752	8.2237
31	7.8334	6.3878	17.2137	8.9500	2.0325	9.8258	0.4726	17.1292	5.1776	14.5347	12.5562	0.5117	8.3361
32	22.2816	17.9595	24.8583	18.3065	12.4455	24.1512	8.9222	42.8578	2.5277	37.1321	25.0370	4.9145	25.8025
33	18.1746	1.9604	32.6903	2.4500	8.5629	4.3139	4.8895	26.5160	2.6129	24.9678	5.0158	1.0154	17.6901
34	1.9664	18.1893	12.8960	25.0160	3.9709	20.0020	6.4694	1.1004	29.1283	0.5420	26.5882	14.4281	0.4852
35	17.4995	4.5451	28.9011	5.0331	7.8932	7.8906	4.2925	28.9619	1.0848	26.3362	8.6723	0.4484	18.0478
36	0.5542	29.0361	4.2898	36.9016	4.4852	32.8630	8.1548	4.7896	32.9490	2.1095	40.5881	18.1275	1.0994
37	26.1254	1.0517	44.5795	0.4985	14.5895	2.4288	9.8188	32.6340	4.4244	32.0800	2.0731	3.9582	24.6959

	14	15	16	17	18	19	20	21	22	23	24	25	26
14	0.0												
15	25.0453	0.0											
16	7.9445	28.4456	0.0										
17	12.8733	39.7327	1.0461	0.0									
18	12.6161	32.4696	0.5401	0.4948	0.0								
19	26.7205	4.4821	42.0503	56.3452	49.1830	0.0							
20	0.4914	26.0174	12.3706	18.3596	18.0724	24.7304	0.0						
21	18.5570	2.4444	29.9097	42.1209	35.9287	1.0384	17.6478	0.0					
22	25.2265	1.8706	36.3311	49.5797	42.3641	0.5619	24.2839	0.5291	0.0				
23	2.3874	32.2934	2.6121	4.5724	5.1990	39.8522	4.8839	28.7042	36.2923	0.0			
24	0.8861	33.3004	12.7623	17.6959	18.3879	32.1011	0.4864	23.9922	31.6410	4.2790	0.0		
25	1.0617	17.7075	4.9177	9.9542	8.6054	22.3627	2.4701	14.2522	19.7950	2.5102	3.8849	0.0	
26	26.7747	0.5370	26.0440	36.3234	28.9941	8.1205	28.7695	5.0504	4.4104	32.0173	36.0534	18.3723	0.0
27	0.4992	32.5261	8.4500	12.3118	13.0468	34.2831	1.0814	25.0822	32.7778	1.8785	0.4904	2.5969	34.2587
28	18.3441	0.9621	25.8508	37.1481	30.8716	2.4416	18.3850	0.4658	0.9099	26.6473	24.7526	13.0246	2.5682
29	22.7451	2.4381	18.2968	26.5972	20.2037	12.8996	25.7530	7.8578	8.1879	25.1683	32.0486	14.4226	0.9362
30	2.1074	29.1124	18.2326	25.2186	25.0355	24.6408	0.5680	18.5816	25.3229	8.5245	1.0178	5.2179	32.9685
31	8.6924	4.4152	12.1328	20.2523	15.9401	9.1245	9.9287	4.0069	6.4850	12.9528	14.2810	4.4408	4.9673
32	20.2860	4.7936	35.5156	48.7490	42.5481	0.4706	18.4102	0.4754	0.9933	32.3875	24.8243	16.9064	8.4074
33	20.2051	1.0593	19.4463	28.6320	22.3114	8.7519	22.3007	4.6178	5.0212	24.4054	28.5820	12.8476	0.5166
34	4.4629	26.1753	0.4986	2.5061	2.0728	36.7926	7.9043	25.6374	32.1317	1.0537	8.2864	2.4711	24.8085
35	18.4890	0.5136	21.8569	32.0526	25.7951	5.1078	19.5626	2.0081	2.4744	24.6958	25.8461	12.1919	1.0438
36	0.9203	31.8485	5.1788	8.1448	8.8159	36.3775	2.4168	26.2587	33.8904	0.5121	1.8312	2.0619	32.6210
37	29.1974	2.0464	24.5346	33.8203	26.4594	12.5851	32.1689	8.4992	7.8290	32.4987	39.4221	19.8144	0.4871

	27	28	29	30	31	32	33	34	35	36	37
27	0.0										
28	24.8951	0.0									
29	29.2301	4.5000	0.0								
30	2.6578	20.3475	30.9075	0.0							
31	13.2038	2.6065	3.9025	12.9004	0.0						
32	26.8796	1.8779	12.1919	18.3106	6.5542	0.0					
33	26.6734	2.1535	0.4285	26.4686	2.4692	8.0528	0.0				
34	4.9564	22.5422	17.9822	12.7022	9.8214	30.2729	18.2073	0.0			
35	24.9571	0.5426	1.9226	22.6278	1.9173	4.4360	0.5359	19.5842	0.0		
36	0.4291	25.1776	26.7346	4.9803	12.5538	28.9565	25.0354	2.6102	24.2782	0.0	
37	36.6527	5.0659	0.4729	37.4175	6.3967	12.8179	0.9660	24.2835	2.5145	34.1012	0.0

BIBLIOGRAHY

[1] Anscombe,F.J. Rejection of Outliers. *Technometrics*,2,123-146,1970

[2] Baker,F.B. and Hubert, L.J. Measuring the Power of Hierarchical Cluster Analysis. *JASA*,70,31-38,1975

[3] Ball, G.H. Data Analysis in the Social Sciences: What about the Details? In *Proceedings of the Fall Joint Computer Conference, Stanford.* New York: Macmillan, 533-559,1965

[4] Classification Analysis Stanford Research Institute, SRI project 5533, 1971

[5] Bolshev, L.N. Cluster Analysis. *Bull. I.S.I.*, 43, 411-425,1969

[6] Bonnèr, R.E. On Some Clustering Techniques. *I.B.M. J. Res. Dev.*, 8, 22-32, 1964

[7] Cooper, D.B. and Cooper, P.W. Nonsupervised Adaptive Signal Detection and Pattern Recognition. *Information and Control*;7, 416-444, 1964

[8] Day, N.E. Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, 56, 463-474, 1969

[9] Everitt, B. Cluster Analysis. London: Heineman Educational Books: 1974

[10] Fisher, L. and Van Ness, J. Admissible Clustering Procedures *Biometrika*, 58, 91-104, 1971

[11] Friedman, H.P. and Rubin, J. On Some Invariant Criteria for Grouping Data. *JASA*,62, 1159-1178, 1967

[12] Gitman, I. and Levine, M.D. An Algorithm for Detecting Unimodal Fuzzy Sets and its Application as a Clustering Technique. *IEEE Trans. Comp.*, C19, 583-593, 1970

[13] Gower, J.C. Multivariate Analysis and Multidimensional Geometry. *The Statistician*, 17, 13-25, 1967

[14] A General Coefficient of Similarity and Some of its Properties. *Biometrics*, 27, 857-872, 1971

[15] Gower, J.C. and Ross, C.J.S. Minimum Spanning Tress and Single Linkage Analysis. *Appl. Statist.*, 18, 54-56, 1969

[16] Jardine, J. Towards a General Theory of Clustering, *Biometrics*, 25,609-610, 1969

[17] Jardine, J. and Sibson, R. Mathematical Taxonomy. New york: John Wiley and Sons:1971

[18] Johnson, S.C. Hierarchical Clustering Schemes. *Psychometrika*, 32, 241-254, 1967

[19] Kruskal, J.B. Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika*, 29, 115-119, 1964

[20] Kruskal, J.B. and Carroll, J.D. Geometrical Modes and Goodness-of-fit Functions. In Multivariate analysis (P.R. Krishnaiah, ed.), II, 639-671. New York: Academic Press: 1969

[21] Kuhns, J. Work Correlations and Automatic Indexing. *Res. Rep.*, *Ramo Woolbridge Corp.*, California: 1957

[22] Kullback, S. Information Theory and Statistics. New York: John Wiley and Sons: 1959

[23] Lance, G.N. and Williams, W.T. A General Theory of Classification Sorting Strategies. I. Hierarchical Systems. *Comput. J.*, 9, 373-380, 1967

[24] A General Theory of Classificatory Sorting Strategies. II. Clustering Systems. *Comput. J.*, 10, 271-276

[25] Ling, R.F. On the Theory and Construction of k-clusters. *Comput. J.*, 15, 326-332, 1972

[26] A Probability Theory of Cluster Analysis. *JASA*, 68, 159-164, 1974

[27] Mahalanobis, P.C. On the Generalized Distance in Statistics. *Proc. Natl. Inst. Sci. (India)*, 12, 49-55, 1936

[28] Patrick, E.A. Learning Probability Spaces for Classification and Recognition of Patterns with or without Supervision. Ph.D. Thesis, Purdue University, Lafayette, Ind., 1965

[29] Asymptotic Distribution of Maximum Likelihood Estimators for a Nonsupervised Adaptive Reciever. *IEEE Intern. Conference Record*, Philadelphia, 1966

[30] On a class of Unsupervised Estimation Problems, *IEEE Trans. Inf. Theory*, IT-14, 407-415, 1968

[31] Concepts of an estimation System, Adaptive System and a Network of Adaptive Estimation Systems. *IEEE Trans. System science and Cybernetics*, 1, 79-85, 1969

[32] Fundamentals of Pattern Recognition. Englewood Cliffs, N.J. : Prentice-Hall Inc.: 1972

[33] Patrick, E.A. and Carayannopolus, Codes for Unsupervised Estimation of Source and Binary Channel Probabilities. *Information and Control*, 14, 358-375, 1970

[34] Patrick, E.A. and Costello. Asymptotic Probability of Error using Two Decision-directed Estimators for Two Unknown Means. *IEEE Trans. Inf. Theory*, IT-14, 160-162, 1968

[35] Unsupervised Estimation and Processing of Unknown Signals. *Purdue School of Electrical Engineering, Tech. Rep.* EE 69-18, 1969

[36] On Unsupervised Estimation Algorithms. *IEEE Trans. Inf. Theory*, IT-16, 556-559, 1970

[37] Patrick, E.A. and Hancock, J.C. The Unsupervised Learning of Probability Spaces and Recognition of Patterns. *IEEE Intern. Convention Record*, 1965

[38] Interactive Computation of Aposteriori Probability for M-ary Nonsupervised Adaptation. *IEEE Trans. Inf. Theory*, IT-12, 483-484, 1966

[39] Patrick, E.A. and Liporace, L. Unsupervised Estimation of Parametric Mixtures. *Purdue School of Electrical Engineering Tech. Rep.*, EE 70-31, 1970

[40] Patrick, E.A., Costello, J.P. and Monds, F.C. Decision Directed Estimation of a Two Class Decision Boundary. *IEEE Trans. Computers*, C-19, 197-205, 1970

[41] Rao, M.R. Cluster Analysis and Mathematical Programming. *JASA*, 66, 622-626, 1971

[42] Rubin, J. Optimal Classification into Groups: an Approach for Solving the Taxonomy Problem. *J. Theor. Biol.*, 15, 103-144, 1967

[43] Sakrison, D.F. Stochastic Approximation, a Recursive Method for Solving Regression Problems. In Advances in Communication Systems (A.V. Balakrishnan, ed.), 2, 51-106, New York: Academic Press: 1966

[44] Shepard, R.N. and Carroll, J.D. Parametric Representation of Nonlinear Data Structures. In Multivariate Analysis (P.R. Krishnaianh, ed.). New York: Academic Press: 1966

[45] Sibson, R. A Model for Taxonomy II. *Math. Biosci.*, 6, 405-430, 1970

[46] Some Observations on a Paper by Lance and Williams. *Comput. J.*, 14, 156-157, 1971

[47] Sneath, P.H.A. A Comparison of Different Clustering Mehtods as Applied to Radomly Spaced Points. *Classification Soc. Bull.*, 1, 2-18, 1966

[48] Evaluation of Clustering Methods. In Numerical Taxonomy (A.J. Cole, ed.). London: Academic Press: 1969

[49] Sokal, R.R. and Sneath, P.H.A. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman and Company: 1963

[50] Switzer,P. Statistical Techniques in Clustering and Pattern Recognition. Department of Statistics, Stanford Univ., TR139

[51] Ward, J.H. Hierarchical Grouping to OPtimize an Objective Function. *JASA*, 58, 236-244, 1963

[52] Wolfe, J.H. A Computer Program for the Maximum Analysis of Types. *U.S. Naval Personnel Research Activity, Tech. Bull.*, 65-15, 1965

[53] NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. *U.S. Naval Personnel Research Activity, Tech. Memo.* SRM 68-2, 1967

[54] Pattern Clustering by Multivariate Mixture Analysis. *U.S. Naval Personnel Research Activity, Res. Memo* SRM 69-12, 1969

[55] A Monte-Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. *Naval Personnel and Training Research Laboratory, Tech. Bull.* STB 72-2, 1971

[56] Yakowitz, S. A Consistent Estimator for the Identification of finite Mixtures. *Ann. Math. Statistics*, 40, 1728-1735, 1969

[57] Yakowitz, S. and Spragins, J. A Characterization Theorem on the Identifiability of Finite Mixtures. *Ann. Math. Statistics*, 39, 209-214, 1968

VITA AUCTORIS

Graduated St. Annes High School, Tecumseh, Ont. 1963

Received B.A. from The University of Windsor 1972