RADC-TR- 66-514
Interim Report

A COMPARISON OF SOME CLUSTER-SEEKING TECHNIQUES

Geoffrey H. Ball

Stanford Research Institute

TECHNICAL REPORT NO. RADC-TR-66-514
November 1966

Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

A COMPARISON OF SOME CLUSTER SEEKING TECHNIQUES

Geoffrey H. Ball

FOREWORD

This interim report was prepared by Mr. Geoffrey H. Ball of Stanford Research Institute, Menlo Park, California, under Contract AF30(602)-4196, Project 5581, Task 558104.

The author acknowledges the continuing encouragement and guidance of Dean F. Babcock, Stanford Research Institute, in this effort; considerable gratitude is also expressed to David J. Hall of SRI for the many conversations with him which greatly aided in the formulations in this report. The reporting period was from June 1966 to August 1966.

Mr. Charles A. Constantino (EMIID) was the Rome Air Development Center project engineer.

This report has been reviewed and is approved:

Approved: FRANK J. TOMAINI
Chief, Info Processing Branch

Approved: JAMES J. DIMEL, COLONEL, USAF
Chief, Intel and Info Processing Div.

ABSTRACT

Conventional multivariate statistics examine in considerable depth the significance of relationships existing in data as shown by the mean and the covariance matrix. Shortcomings of this approach are briefly discussed.

"Cluster-seeking techniques" are discussed as alternatives to conventional multivariate methods. Thirty variants of cluster-seeking techniques, the total number presently known to the author, are divided into seven categories: probabilistic, signal detection, clustering, clumping, eigenvalue, minimal mode seeking and miscellaneous. These larger classes are contrasted and, within each class, the techniques are summarized and compared.

A composite technique that combines the best features of the various approaches is proposed.

BLANK PAGE

TABLE OF CONTENTS

Section		Page
I	INTRODUCTION	1
II	STATISTICAL ANALYSIS OF SOCIAL SCIENCE DATA	3
III	CLUSTER-SEEKING TECHNIQUES	7
	1. Historical Background	7
	2. Definitions	7
IV	COMPARE AND RELATE CLUSTER-SEEKING METHODS	9
	1. The Use of Iterative Techniques	9
	2. The Use of Categorization Information to Cause Mode-Seeking	9
	3. Constraints on the Applicability of Cluster-Seeking Techniques Imposed by Various Factors	9
	4. Measures of Similarity Used by the Techniques	11
	5. Criteria for Adequacy of Clustering Used by the Techniques	11
	6. Lumping of Clusters	11
	7. The Nature of "Convergence" of Cluster-Seeking Techniques	15
	8. Shortcomings of Cluster-Seeking Techniques	16
V	MAJOR CLASSIFICATIONS OF CLUSTER-SEEKING TECHNIQUES	19
	1. Probabilistic Techniques	20
	2. Signal Detection	20
	3. Clustering Techniques	21
	4. Clumping Techniques	21
	5. Eigenvalue Techniques	21
	6. Minimal Mode Seeking	21
	7. Miscellaneous Techniques	22
VI	INDIVIDUAL TECHNIQUES	23
	1. Probabilistic Techniques	23
	2. Comments on Probabilistic Technique	23
	3. Signal Detection Techniques	24
	4. Comments on Signal Detection Technique	25
	5. Clustering Techniques	27
	6. Comments on Clustering Techniques	28
	7. Clumping Techniques	29
	8. Comments on Clumping Techniques	31
	9. Comments on Eigenvalue-Type Techniques	34
	10. Minimal Mode-Seeking Techniques	35
	11. Comments on Minimal Mode Seeking	36
	12. Miscellaneous Technique	37
	13. Comments on Miscellaneous Techniques	37

TABLE OF CONTENTS (Continued)

Section		Page
VII	COMPOSITE CLUSTER-SEEKING ALGORITHM	39
	1. Composite Technique (1965)	39
	2. Comments on Composite Technique	40
VIII	CONCLUSIONS	41
	REFERENCES	43

A COMPARISON OF SOME CLUSTER-SEEKING TECHNIQUES*

SECTION I

INTRODUCTION

The advent of the relatively inexpensive digital computer makes iterative cluster-seeking methods of analyzing complex multivariate data practical when they could not be seriously considered before. These "cluster seeking" techniques provide a way of viewing multivariate data that differs from factor analysis and discriminant analysis.

Cluster-seeking techniques are best suited to examining problems where the data is multi-modal. They provide a way of detecting isolated data points that are not "close" (relative to the data set) to any other points. These techniques can be used to show the relationship of a single data point to the entire set of data -- thus allowing an examination of the details in the data. Large numbers of data points can be structured and related to each other in the original high dimensional space.

We feel that the techniques described below provide a useful adjunct to other methods of analyzing multivariate data. We compare and describe below, a number of these methods reported in the literature.

We present an improved version of the ISODATA cluster-seeking technique that incorporates aspects of other techniques in ways that appear to overcome certain difficulties that arise in each of the techniques that have been suggested thus far. Finally, we speculate on the ramifications of a widespread use of these cluster-seeking techniques.

---

*Previously published under the title, DATA ANALYSIS IN THE SOCIAL SCIENCES: WHAT ABOUT THE DETAILS?, in the Proceedings of Fall Joint Computer Conference, 1965.

BLANK PAGE

SECTION II

STATISTICAL ANALYSIS OF SOCIAL SCIENCE DATA

In statistical data analysis, much use is made of the covariance and the correlation matrix. For example, the correlation matrix is used in a central way in principal components analysis, in factor analysis, and in canonical correlation analysis.

In Figure 1 we show three sets of data that, when plotted, appear to be very different. (The second data set consists of samples drawn from a normal distribution. No truncation of the data set should be inferred from the figure.) The interesting fact is that all three of these data sets give rise to the same covariance matrix and hence to the same correlation matrix. If the data points were specified by a different coordinate system, the covariance matrix would be modified in the same way for all the data sets. Hence we see that these data sets are indistinguishable if only the first and second moments and cross-moments are used to describe the data. The fact that these very different data sets lead to the same covariance matrix is rather unnerving when, for example, with the principal components analysis, it is realized that frequently no use is made of the original data except to abstract the means and the covariance matrix. It therefore seems reasonable to ask how the detailed structure of the data might be taken more accurately into account.

Before examining this question in some more detail, let us examine two other aspects of utilizing only the covariance matrix and the mean. Four factors seem particularly significant here:

1. The effect of erroneous data points, caused, for example, by card punching errors, can be rather considerable in a particular covariance matrix. Unless these erroneous points can be determined and removed from the data set, the statistics based on the entire data set will be considerably affected.

2. A second effect arises from the need to estimate the covariance matrix using a set of sample points. Dr. David Allais[3] has shown recently that the number of samples should roughly equal 10 times the number of dimensions in order to estimate adequately the covariance matrix.

3. More than one modal point in the data also causes the covariance matrix to be an inadequate description of the data. Thus, if the data is described as being the sum of two Gaussian distributions and a single overall covariance matrix is computed for the entire data set without taking the bi-modality of the probability distribution into account, then this covariance matrix depends critically on the distance between the means of these two modes. Intuitively this is rather unsatisfying as a description of the data.

4. Predominant subsets in the data can overwhelm subsets that occur less frequently, i.e., the significant but rare event, may not be singled out.

This is not to say that a principal components analysis or factor analysis does not have its place. The point seems rather to be that if these techniques are to be used, then the data ought to have characteristics that in some way satisfy assumptions of

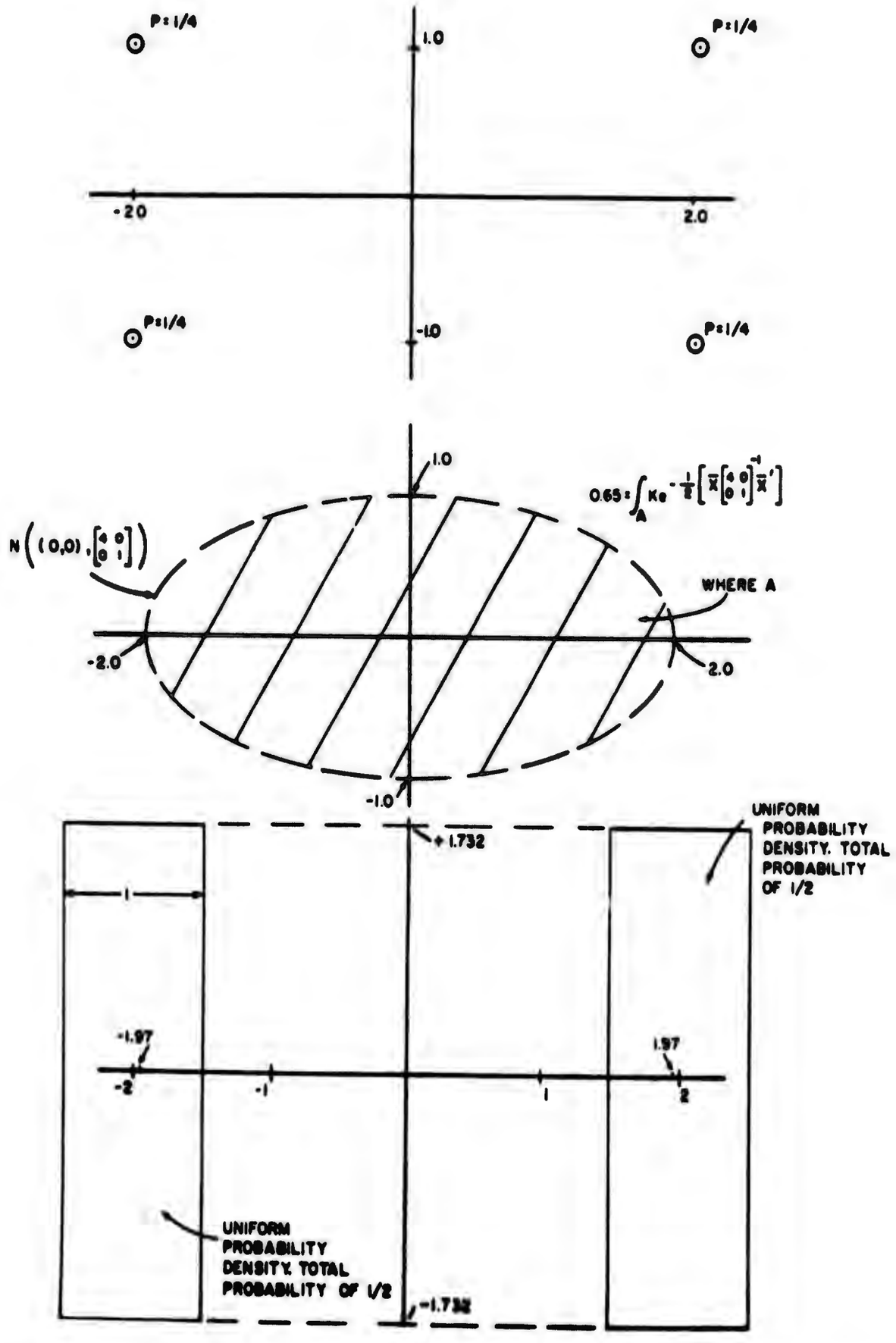

Figure 1. Three Distributions of Data Having the Same Covariance Matrix

Gaussian distributions; that is, the covariance matrix should be a good description of the characteristics of the data.

A useful description of all the data can be plotted using the axes found by factor analysis or principal components analysis. Even this may not be as meaningful as desired as can be seen by examining Figures 2a and 2b where we see that two quite different two-dimensional probability distributions both give rise to marginal distributions that are uniform along vertical and horizontal axes. We can see that this inability to distinguish between these different distributions can be resolved by using more axes to describe the data (as shown by the marginal distributions along the diagonal axes). Relating events on the different axes, particularly when the data are high dimensional is difficult, however. The major problem, then, is that two sets of data that are quite dissimilar can appear to be quite similar when viewed by data analysis techniques implicitly oriented toward Gaussian distributions.

It therefore appears to us that there is a need to treat local regions in the data space rather than projecting down on to a line or a plane from over the whole space. That is, there is a need to be concerned about the details. We feel that a set of techniques that have been developed primarily over the last five years provides a satisfactory direction for finding an answer to this need to examine the details. Much work remains to be done on these techniques, but it does appear that the particular point of view that they offer can be very helpful.

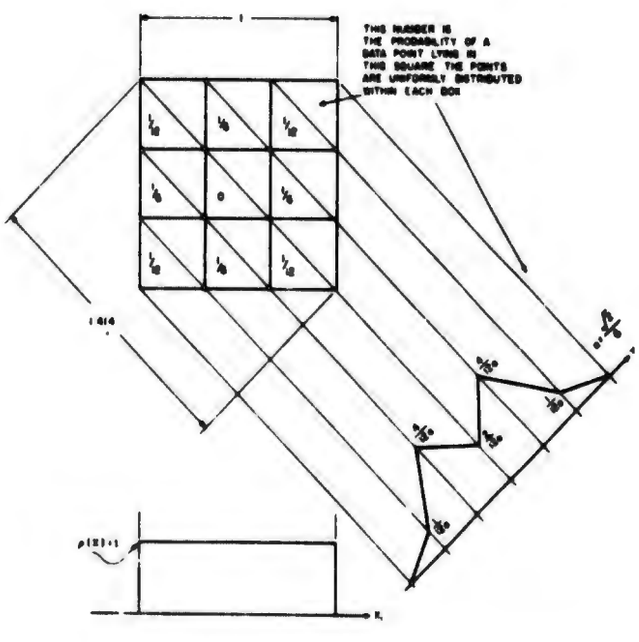

Figure 2a. A Nonuniform Two-Dimensional Distribution of Points Having Uniform Marginal Distributions

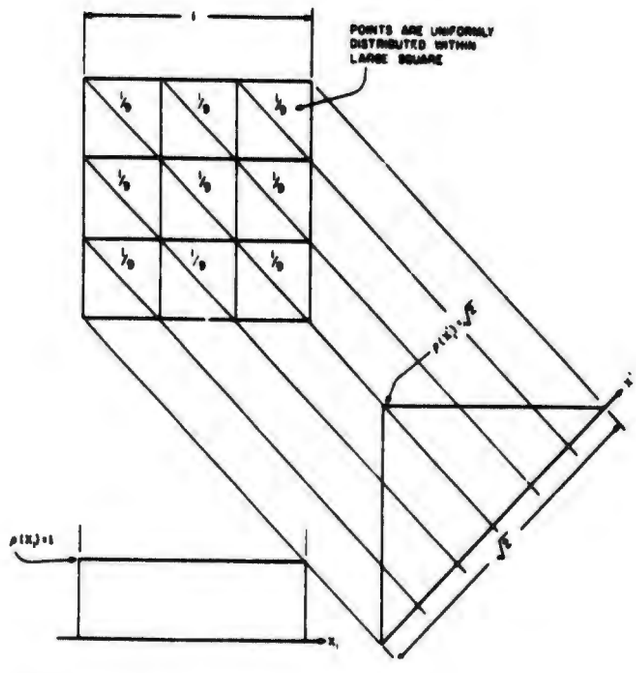

Figure 2b. A Uniform Two-Dimensional Distribution of Points Having Uniform Marginal Distributions

BLANK PAGE

SECTION III

CLUSTER-SEEKING TECHNIQUES

The essential characteristics of the techniques that we will be describing is the sorting of the set of data patterns into subsets, such that each subset contains data points that are as much "alike" as possible. The methods for arriving at the subsets differ in a variety of ways. We will, in the following paragraphs, attempt to describe the known existing techniques and to point out ways in which they are significantly different. We finally attempt to construct a technique that is a composite of the best features of some of these techniques and ISODATA.

1. Historical Background

In Table I we have arranged the papers by type (e.g., probabilistic, signal detection, etc.) and by date. We have made no attempt to determine how much the development of one technique depended on another. They are arranged in chronological order in a way that appears reasonable to us at this time, without attempting to accurately establish priorities. As far as we have been able to determine, nearly all the techniques with the exception of the factor analytic techniques and some lumping techniques originated after 1960. Since most of these techniques require a considerable amount of computation, it seems probable that only the advent of inexpensive digital computation has allowed these techniques to be developed.

2. Definitions

In order that certain terms to be used over and over again in this paper should not be confused, it seems important to define them:

a. Measurement — By measurement we mean a component of the pattern vector; that is, it is one of several numerical values (related to a property of the pattern) used to define a pattern — for example, one out of several answers on a questionnaire.

b. Pattern — By pattern we mean the collection of measurements considered to be a single entity in the clustering program for example, the answers to a set of questions on a questionnaire.

c. Parameter — By parameter we will not mean measurement in this paper. Rather, we will mean a number used to control the operation of one of the cluster-seeking techniques — for example, the threshold used to control lumping in the Ball-Hall technique[4] is a parameter.

d. Cluster — A cluster of patterns is, in our mind, a set of patterns contained in a high-dimensional space where the density of patterns is large compared to the density in the surrounding volumes. It is not yet a rigorously defined concept but rather one that depends on the nature of the data. Attempts are being made to define this more exactly.

e. Mode — A mode is a cluster of patterns that belong to a single class of patterns. (This definition varies from the more precise statistical definition of a mode as the most frequently observed value of a random variable.)

Table 1. Cluster-Seeking Techniques

Class of Technique	Year First Reported						
	Before 1960	1960	1961	1962	1963	1964	1965
Probabilistic				Daly		Fralick	
Signal Detection		Jakowatz et al	Brennan Glaser	Hinnich	Spilker	Turner Smith	
Clustering				Okajima et al Sebestyen Hyvarinen		Ball & Hall	
Clumping	Michener & Sokol (1957)	Rogers & Tanimoto Sawrey, Keller, & Conger	Needham Parker-Rhodes Abraham			Bonner Fortier & Solomon	
Eigenvalue I				Nunnally			
Eigenvalue II					Cooper	Mattson & Dammann	
Minimal Mode-Seeking					Firschein & Fischler Steinbuch & Piske		
Miscellaneous				Block, Knight & Rosenblatt	Bledsoe		

SECTION IV

COMPARE AND RELATE CLUSTER-SEEKING METHODS

In the following paragraphs we describe and discuss each of the aspects of cluster-seeking techniques that we consider particularly significant.

1. The Use of Iterative Techniques

It generally does appear that iterative techniques allow a more detailed examination of the data than do techniques that require the calculation of a single function of all the data (e.g., the covariance matrix). Iterative techniques can allow the examination of iteratively selected subsets of patterns where the selection of one subset depends on the results obtained from a previous selection. This examination increases the sensitivity of the methods to variations in patterns and the structure of the data without the assumptions that are usually necessary in non-iterative methods.

2. The Use of Categorization Information to Cause Mode-Seeking

It is possible to use classification information to determine the "modes" in the data. When this is done, it then becomes important that all the classes that are to be classified be represented in the data. In Figure 3 (upper half) we see that it is possible that two modes of Class 1 be sufficiently remote from Class 2 to be considered one mode by these methods. Yet later (as shown in Figure 3 (lower half)), when other classes are introduced, a new class, Class 3, can lie exactly on top of what was considered the "description" (the average point) of the previously determined mode of Class 2 and yet still be completely separable from Class 1.

The use of classification information can be extremely helpful in cases where the finding of the most economical cluster description is important, and where all the classes are initially available.

3. Constraints on the Applicability of Cluster-Seeking Techniques Imposed by Various Factors

Five factors primarily constrain the applicability of a given technique. These are:

- Computational complexity
- Memory requirement
- Sample size requirement
- Nature of the data
- Availability of categorization information

The computational complexity of iterative methods is usually determined primarily by the computations performed in the inner computational loop of the program. Another factor affecting the utility of the method in terms of computation is the ease with which the required computations can be performed by methods other than the conventional general-purpose digital computer. Distance calculations that require the computation

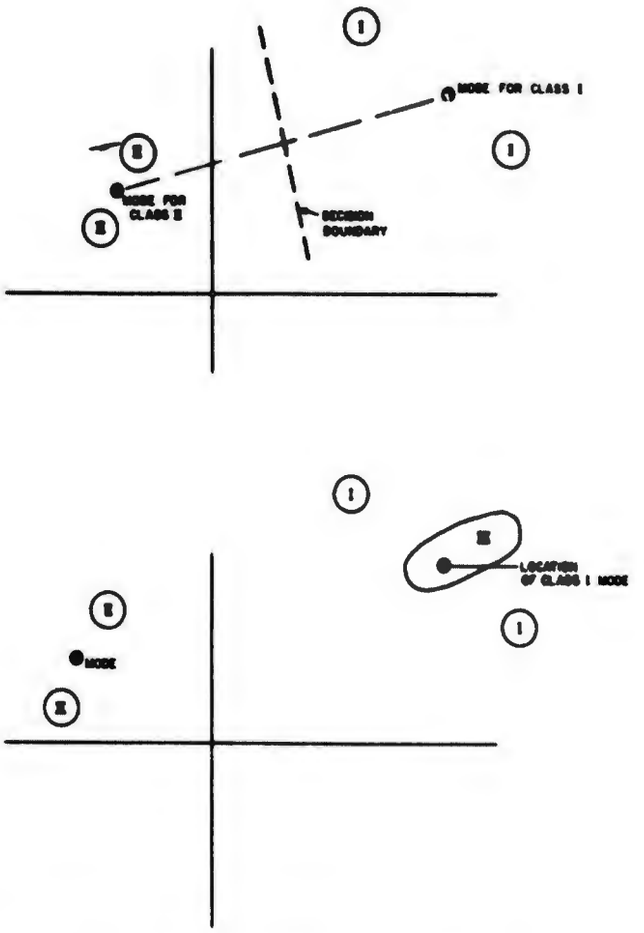

Figure 3. The Result of Introducing New Classes Later in Mode-Seeking Techniques

of either a correlation or Euclidean distance can be performed using optical correlation techniques.*

The memory required by a given technique may be so great that the technique is uneconomical. High-speed random access is costly (although disk files make fairly high-speed random access comparable in cost to tape) and it is primarily for this memory requirement that the various iterative techniques should be examined.

The sample size requirement relates to any quantities that must be estimated (in a statistical sense) from the set of sample patterns. Allais[3] showed that this requirement cannot be taken lightly. He shows that given N samples the estimation of a covariance matrix of dimension greater than N/10 usually increases the probability

---

* The calculation of Euclidean distance between vector P and vector M by using a correlation operation can be shown by the following argument:

The square of the Euclidean distance is $(P - M) \cdot (P - M)$, where P is the pattern and M is the mask, or weight vector. This expression can be expanded into three terms: $P \cdot P$, $M \cdot M$, and $P \cdot M$. If the mask, or weight vector, is not changed after each pattern, then $M \cdot M$ can be considered to be a constant over an entire iteration through all of the patterns. The quantity $P \cdot P$ can be calculated using an optical device that squares a picture in magnitude element by element. And finally the $P \cdot M$ term can be determined by usual optical correlation techniques.

of error for predictions based on that covariance matrix as compared with predictions using a covariance matrix of fewer dimensions. Small sample sizes seem to require that simpler quantities be estimated -- such as the means of clusters, or only the largest eigenvector of the covariance matrix.

The nature of the data is not easily specified. If the data is time-varying it will frequently be more effective (and possible necessary) to use a different technique from one suitable for stationary patterns. The degree to which the data exists in isolated clusters as opposed to being in "amoebic smears" (i.e., all in one contiguous mass of data) is also important. One of the important research tasks yet to be done is the establishment of better means of classifying data into broad types that suggest the technique that could be most successfully used to further analyze them.

Where categorization information is available and reliable, use of it to constrain the clustering seems advised.

4. Measures of Similarity Used by the Techniques

In relating techniques, consideration should be given to the measure of similarity that was or could be used. Correlation usually requires normalization of the pattern vectors with respect to magnitude. Euclidean distance does not. Euclidean distance is considerably affected by the scale factor associated with each pattern dimension -- particularly when these dimensions are not commensurate in units of measurement (e.g., feet vs inches or, worse yet, feet vs seconds).

The measures of similarity used in the papers reported on are given in Table 2. Additional measures of similarity are given on pages 129-130 of Sokal and Sneath.[40]

5. Criteria for Adequacy of Clustering Used by the Techniques

Some techniques depend critically on the criteria of clustering used. The clumping techniques described below are particularly sensitive to this, since only one iteration is performed and each decision as to the cluster with which a pattern is associated tends to be final.

In iterative techniques this dependence is reduced by allowing the algorithm several iterations in which to associate patterns into clusters. Hence changes can be made based on information obtained in the initial clusterings.

Table 3 gives a list of criteria for clustering used by various authors. In this table N_c represents the number of clusters.

6. Lumping of Clusters

The convergence of cluster seeking techniques is related to the way in which clusters are allowed to form. In particular, if clusters can be created, then some mechanism must be provided for either (1) tightly controlling the formation of new clusters, or (2) "lumping" them together. Without these mechanisms the number of clusters could grow until each pattern was in its own cluster, which would obviously not be very useful clustering. Therefore it is important to examine for each of the clustering techniques the way that the number of clusters is changed. (Note that

Table 2. Comparison of Measures of Similarity

Measures of Similarity

Dot Product	$P \cdot W = \sum_{i=1}^{D} p_i w_i = \lvert P \rvert \lvert W \rvert \cos (P, W)$	Block et al (1962) Steinbuch (1963) Mattson & Dammann (1965) Michener & Sokal (1957) Jakowatz et al (1960) Spilker et al (1963) Glaser (1961) Smith (1964)
Similarity ratio	where $R_{ij} = P_i \cdot P_j$ $S_{ij} = \frac{R_{ij}}{R_{ii} + R_{jj} - R_{ij}}$ when $p_{kl} = 0, 1.$ uses $d_{ij} = -\log S_{ij}$ as "distance."	Rogers & Tanimoto (1960)
Weighted Euclidean distance	$D_{\alpha\beta} = \sum_{i=1}^{D} k_i (\chi_{i\alpha} - \chi_{i\beta})^2$	Bonner (1962) Sebestyen (1962)
Unweighted Euclidean distance	$D_{\alpha\beta} = \sum_{i=1}^{D} (\chi_{i\alpha} - \chi_{i\beta})^2$	Ball & Hall (1964)
Measure for binary variables taking into account pairwise correlation	$S_{\alpha\beta} = \sum_{i=1}^{D} \sum_{j=1}^{D} r_{ij} [1 - \lvert \chi_{\alpha i} - \chi_{\beta i} \rvert] \cdot [1 - \lvert \chi_{\alpha j} - \chi_{\beta j} \rvert] \cdot [1 - 2 \lvert \chi_{\alpha i} - \chi_{\alpha j} \rvert]$ r_{ij} is correlation coefficient between measurements i and j	Bonner (1963)
Boolean "and"	$S_{ij} = \sum_{i=1}^{D} p_i \cap p_{jk}$ where $\cap$ denotes Boolean "and"	Needham (1961)
Weighted Boolean "and"	$S_{kl} = \sum_{j=1}^{D} \begin{cases} r_j, & \chi_{kj} = \chi_{lj} \neq 0 \\ 1, & \chi_{kj} \cdot \chi_{lj} = 0 \\ 0, & \text{otherwise} \end{cases}$ r_j is number of levels of j^{th} measurement (finite for his data)	Hyvarinen (1962)

Table 2. Comparison of Measures of Similarity (Continued)

Measures of Similarity		
Normalized correlation	$\frac{P_i \cdot P_j}{\sqrt{(P_i \cdot P_i)(P_j \cdot P_j)}}$ where $P_i \cdot P_j = \sum_{\ell=1}^{D} p_{i\ell} p_{j\ell}$	Okajima et al (1963) Nunnally (1962)
The following six measures of similarity are from Kochen (1963) p. 15, and refer primarily to information retrieval.	$S_{ij} = \frac{\log_{10} N(\lvert N \cdot n_{ij} - n_i n_j \rvert - \frac{N}{2})^2}{n_i n_j (N - n_i)(N - n_j)}$	Stiles (1961)
	$S_{ij} = \frac{n_{ij}}{n_j}$	Luhn (1959)
	$S_{ij} = \frac{n_{ij}}{N}$	Baxendale (1961)
	$S_{ij} = \frac{n_{ij}}{n_i + n_j - n_{ij}}$ and $-S_{ij}\log_2 S_{ij}$	(King-) Tanimoto (1960)
	$S_{ij} = N \cdot \frac{n_{ij}}{n_i \cdot n_j}$	Kochen & Wong (1962)
	$S_{ij} = -\log\left\{\frac{n_{ij} \cdot N}{(n_i n_j)}\right\}$ and $\frac{-n_i n_j}{N^2} \log\left\{\frac{n_{ij} \cdot N}{(n_i n_j)}\right\}$ where N is the total number of measurements and n_i is the number of ones in (binary) pattern N_{ij} is the number of measurements in which pattern i and pattern j are alike.	Abraham (1962)

Table 3. Criteria for Clustering

Entropy	$H = -\sum_{j=1}^{r} f(\chi_j) \log f(\chi_j)$ where $f(\chi_j)$ is discrete frequency distribution function.	Hyvarinen (1962) Sebestyen (1964)
Average distance from nearest cluster center	$\frac{1}{N} \sum_{i=1}^{N_c} N_i \, AVEDST_i$ where N_i is the number of patterns in the i^{th} cluster, $AVEDST_i$ is the average distance of patterns in the i^{th} cluster from the cluster mean, N_c is the number of clusters and N the total number of patterns.	Ball & Hall (1965)
Square (used for deviation from single signal)	$\sum_{j=1}^{D} (P_j - \overline{P}_j)^2$	Brennan (1961)
Value for a cluster	$I_{xx} - \frac{1}{N_R} \sum_{y=1}^{N_c} I_{xy}$ where $I_{xy} = \frac{1}{N_x N_y} \sum_{\alpha=1}^{\alpha=N_x} \sum_{\beta=1}^{\beta=N_y} S_{\alpha\beta}$ where N_x is the number of members in cluster y, $S_{\alpha\beta}$ is 1 if member α of cluster x is similar to member β of cluster y; or 0 if they are not similar. I_{xy} is the percentage of possible similarity "links" which are actually present between the members of cluster x and the members of cluster y.	Bonner (1964)

Table 3. Criteria for Clustering (Continued)

Coefficient of belongingness	$\dfrac{\sum_{k=1}^{N_c} \sum_{i,j \in C_k \text{ cluster } C_k} \rho_{ij} \begin{cases} \text{correll.} \\ \text{coeff.} \end{cases}}{\sum_{k=1}^{N_c} \sum \rho_{ij} \in C_K}$	Fortier & Solomon (1965)
Total entropy	$E_n[(d_{ij})] = -\frac{1}{2} \sum'_{ij} \frac{d_{ij}}{T_n[(d_{ij}]} \log_2 \left[\frac{d_{ij}}{T_n(d_{ij})}\right]$ where $T_n[(d_{ij})] = \frac{1}{2} \sum'_{ij} d_{ij}$ where Σ' means summation only of the finite terms after repeated rows and columns of the distance matrix have been removed and $d_{ij} \equiv -\log_2 S_{ij}$ where S_{ij} is the similarity ratio given by Tanimoto in Table 2.	Rogers & Tanimoto (1960)
Probability of error (when defined)	$= 1 - \dfrac{\Sigma \text{correct}}{\Sigma \text{total}}$	Firschein & Fischler (1963)

clusters can be "thrown away" if they become too small, and that this is one additional way of controlling the number of clusters.)

7. The Nature of "Convergence" of Cluster-Seeking Techniques

Considerable experimental evidence exists for the convergence of these techniques. Some analytic work has been done, notably in the probabilistic and signal detection classes of techniques. Nevertheless, we feel it fair to characterize the understanding of "convergence" for cluster-seeking techniques as minimal — particularly with respect to real, non-Gaussian data.

In a general way, it appears that if the data is indeed clustered then the final clusterings will tend to be unique. If, however, the data is "smeared" and "amoebic" then a greater variety of clusterings can exist. Finally, if the data is uniformly placed in data space then no real stable clusters are found — which is to us intuitively satisfying since no clusters really exist in the data.

8. Shortcomings of Cluster-Seeking Techniques

In examining the shortcomings of various clustering techniques, it seems important to ask what affects the way the data is clustered. In Figure 4 we show that for clustering techniques, the scaling of the various dimensions will undoubtedly affect the way that the patterns are clustered together.

It is, however, possible to normalize the various scales in a way that will lend to a uniqueness of scale with respect to a particular set of patterns. A straightforward way of doing this is to divide each of the measurements by the standard deviation of the marginal distribution for that particular dimension calculated for a given sample of data. Linear transformations of the data, that is, rotations and translations, have different marginal distributions than the same data not transformed. Therefore the scaling on these modified dimensions will be different. It seems probable that in some cases this difference will affect the way that a clustering occurs. Again, most of these statements are qualitative in the sense that if well-defined distinct clusters exist in the data, moderate scaling, rotation and translation probably will not affect the clustering greatly. If, however, as is frequently the case, the data is not tightly clustered but has clusters blended into each other, then these effects become more pronounced.

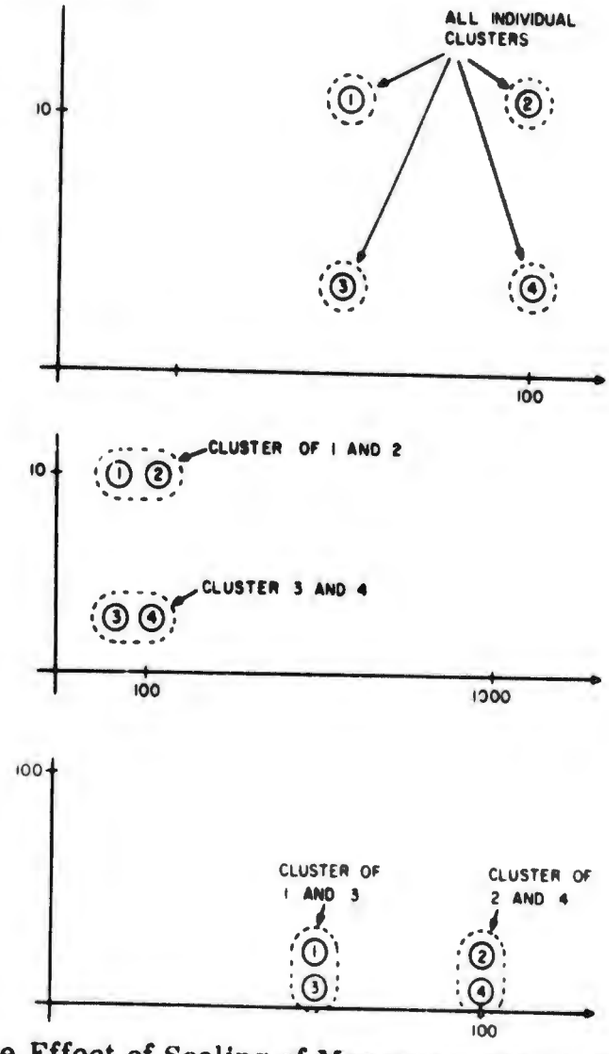

Figure 4. The Effect of Scaling of Measurements on Clustering

Clustering techniques that depend on a correlation measure of distance can be greatly affected by the translation of the data. For example, consider a set of data lying on a hypersphere, surrounding a central point. If the origin is moved outside of the hypersphere then data lying along a given angle from the origin will not consist of a unique part of a hypersphere, but rather of two parts of the hypersphere, and in fact, rather remote portions of the data will be clustered together, if only direction from the origin is used to determine pattern similarity (as is the case with correlation).

The distance measure used to determine similarity or closeness has a considerable effect on the way data is clustered together. For example, if Euclidean distance is modified by having constant multipliers times the various components, where the constant multipliers differ from cluster to cluster, it is then possible to have rather peculiarly shaped volumes assigned to the same cluster. In Figure 5 we see that it is possible to have a large dispersed cluster surrounding a small, rather compact cluster. The "distance" in Figure 5 is determined by the value of a Gaussian probability density. If the purpose of clustering is classification, then this peculiar shape of cluster may be acceptable. If, however, the purpose is to describe the data, then it seems that convex clusters would be most useful.

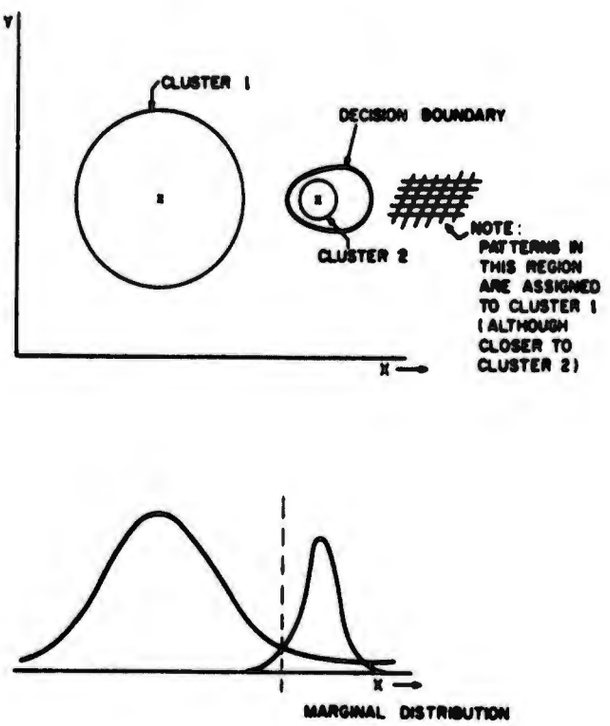

Figure 5. The Decision Boundaries That Result When the Same Measurement Is Differently Scaled for Different Clusters

BLANK PAGE

SECTION V

MAJOR CLASSIFICATIONS OF CLUSTER-SEEKING TECHNIQUES

The various cluster-seeking techniques have been broken down into seven categories:

- Probabilistic
- Signal detection
- Clustering
- Clumping
- Eigenvalue
- Minimal mode seeking
- Miscellaneous

The salient characteristics of each of these classes is described in the following paragraphs and summarized in Table 4.

Table 4. Techniques for Finding "Similar" Subsets in Data

Type	Proponents	Salient Characteristics
Probabilistic	Daly (1962) Fralick (1964)	Estimation of probability of occurrence of a pattern using decision theory and then using weighted combinations of patterns to estimate probability distributions.
Signal detection	Jakowatz, Shuey & White (1960) Glaser (1961) Spilker, Luby & Lawhorn (1963) Smith (1964)	Detection of presence of signal using energy detection, then estimation of parameters of matched filter (correlator).
Clustering	Okajima et al (1962) Sebestyen (1962) Hyvarinen (1962) Ball & Hall (1964)	Finding of minimum distance (or maximum correlation) between pattern and one "cluster center" out of a set of cluster centers. Iterative improvement of the position of these centers.
Clumping	Michener (1957) Sokal (1957) Rogers & Tanimoto (1960) Needham (1962) Sawrey, Keller & Conger (1962) Bonner (1964)	Use of closest pair of patterns to form nucleus for a clump of patterns. "Growing" of clump around this nucleus.

Table 4. Techniques for Finding "Similar" Subsets in Data (Continued)

Type	Proponents	Salient Characteristics
Eigenvalue I	Cooper (1964) Mattson & Dammann (1965)	Finding clusters by finding maximum eigenvalue of covariance matrix and splitting patterns on basis of correlation with corresponding eigenvector.
Eigenvalue II	Nunnally (1962)	Finding clusters by finding eigenvalues of distance matrix and then examining patterns with respect to this new basis, i.e., by seeking eigenvectors with which many patterns are highly correlated.
Minimal mode-seeking	Firschein & Fischler (1963) Steinbuch (1963)	Use of category information to provide impetus to formation of new modes for a category, i.e., incorrect categorization implies needs for new mode.
Miscellaneous	1) Block, Knight & Rosenblatt (1962)	Uses high probabilities of contiguous patterns (in a time sequence) being in the same class to provide a (somewhat noisy) classification of the patterns.
	2) Bledsoe	Seeks to find the set of hyperplanes passing through "corridors" in the data that have maximal average distance from the patterns.

1. Probabilistic Techniques

Probabilistic cluster-seeking techniques are primarily analytic studies in the sense that the main results are couched in analytic terms and were derived from decision theoretic considerations. Both papers describe experimental results obtained by implementing the algorithm implied by the decision theoretic mathematics.

Quite possibly this analytical work will provide considerable insight to other techniques in terms of convergence characteristics and imply ways that they might be modified. It also seems probable that the clustering techniques will suggest new formulations of the mathematics in terms of relaxing restrictions or simplifying calculations through the use of approximations.

2. Signal Detection

Signal detection cluster-seeking techniques grew out of a desire to detect unknown signals in noise. The final decision is based on correlation detection. The techniques

all suffer from the necessity of making an initial detection on the basis of the energy of the signal and then determining the direction of the signal in signal space. For the most part, these techniques consist of the detection of a single signal of unknown epoch in a noisy environment when the time of occurrence of this unknown signal is unknown.

3. Clustering Techniques

Clustering techniques can be characterized by the sorting of patterns by use of multiple cluster points. Tentative assignments of patterns are made to clusters and these assignments are improved until the means of the clusters "adequately" describe the data. The particular method chosen for modifying the description of this data is the primary distinction between these techniques.

4. Clumping Techniques

In these techniques a single pair of patterns usually the closest pair, is selected as a nucleus for a clump of patterns. Other patterns are assigned to this clump on the basis of their closeness to the pair of patterns, or to the mean of the pair of patterns. Generally speaking, these techniques require the calculation of all pairwise distances between all pairs of points and some of these distances must be recalculated after each new combination. In some of them, these distances must be stored in random access memories. The large amount of calculation caused by the calculation of all pairwise distances, even between pairs that are quite remote, makes these techniques less useful than they might otherwise be.

One application for which such techniques seem particularly valuable, however, is that of developing taxonomies. In these cases, one usually has a limited number of samples (with the possible exception of bacterial species) and the problem is tracing them back along an evolutionary tree, combining branches as clumps become close together. It appears that the same information can be obtained by clustering techniques as is obtained by clumping techniques.

5. Eigenvalue Techniques

Eigenvalue techniques are the only techniques we have characterized as non-iterative. Eigenvalue techniques usually depend on the estimation of the covariance matrix. For this reason these techniques tend to require a relatively large number of samples, particularly as the number of dimensions grows large. Large amount of core storage may be required to store the matrix analyzed. The factor analytic techniques of this sort content themselves with one calculation using all of the patterns and a diagonalization of the distance matrix. The later techniques of Mattson and Dammann[27] and Cooper and Cooper[12] use a calculation of only the largest eigenvalue and the corresponding eigenvector, and then subdivide the pattern set in order to proceed further, with a more detailed examination of each of the subsets.

6. Minimal Mode Seeking

These techniques require categorization information to work. A new mode is created only when patterns in one class are nearer to a mode of a different class. Pattern density in space, as such, is not used in cluster seeking.

7. Miscellaneous Techniques

Two techniques do not fit neatly into any of the above categories:

a. Tho technique of Block et al[7] utilizes a high probability of contiguous runs of patterns in a time sequence being from the same class to adjust the machine to a particular mode. This high probability of runs provides marginal teacher information in a probabilistic sense.

b. Bledsoe[6] seeks corridors in the data.

SECTION VI

INDIVIDUAL TECHNIQUES

In the following portion of the paper we describe briefly the known cluster-seeking techniques. We will use the following symbols in describing the difficulty of computation or the amount of storage needed.

N	The number of patterns.
D	The number of dimensions.
NROWS	The number of clusters the process finds.
C	The number of interval blocks the pattern space is broken up into.
ITER	The number of iterations required for the technique to describe the data satisfactorily.
N_T	The total number of patterns used when the patterns are repeatedly drawn from the same probability distributions.

1. Probabilistic Techniques

DALY:[15] Estimates probability of occurrence of all possible binary sequences. Updates estimates after each pattern sample. Computational complexity grows as $e^N \times D$. Convergence has been shown experimentally by computer simulation. Assumption is made that we are looking in a noisy environment, for a single signal that occurs only occasionally. System definitely limited by the exponential growth of the memory of the system as the number of samples increases.

FRALICK:[18] Recursively computes a posteriori probability density of distributional parameters by using a sequence of learning samples. Computation goes as $N \times D \times NROWS \times C \times ITER$. Convergence shown analytically for the case of detection of an unknown signal in noise. Appears to assume that the number of pattern classes is known, and appears willing to develop multimodal distributions in order to handle unexpected classes of patterns. Method limited by the need of storing its latest estimate of probability distribution of the samples, that is, the description of the probability density in terms of the frequency of "cells" in high dimensions or by assuming a particular form of distribution. In multivariate problems the storage of these distributions becomes quite serious and quite difficult. The amount of computation required to update all of the distributional parameters can be considerable unless simplifying assumptions are made regarding the nature of the underlying distribution.

2. Comments on Probabilistic Technique

In our opinion these techniques are more useful as guides to the development of practical techniques than as actual formulations of practical techniques themselves. Either the large amount of storage or the large amount of computation seems to imply that the techniques are outside of practical limits for high-dimensional multivariate data

Fralick[18] points out (p. 13) that in the case where the class a priori probabilities are all the same, the initial selection of the probability distributions for the various classes must be different, or "all computer branches will 'learn' the same thing, and the system as a whole will learn nothing."

A third paper related to this work is that by Patrick and Hancock.[33] This paper presents no new techniques that this author considers as mode-seeking, or learning-without-a-teacher techniques. It does closely examine the efficacy of various estimators that might be used under conditions of partial knowledge concerning probability distributions. The assumptions made in the paper seem to be so strong as to make the paper of relatively little practical significance. Many of the techniques discussed in this paper are for a single-dimensional problem. It does not appear that they generalize usefully to multivariate situations.

A pertinent quote from this paper[33] is as follows: "It is seen from the above discussion that the performance of different learning systems should be compared on the basis of what a priori information is required for their operation."

A second quote[18] relating to the same subject is: "we still require some a priori knowledge in order to be able to obtain the very first decision rule."

It does appear that the question of how much a priori knowledge is necessary to provide what might be called "learning" is one that should be looked at quite carefully.

We feel that the examination of this work motivated by decision theory may be very useful if related to the more ad hoc methods of clustering and finding the structure in data that will be discussed below. For example, in the ISODATA procedure if a pattern falls roughly equidistant from two cluster centers it is only added into the closest cluster center. Fralick's work seems to suggest that in this case it might be well to add portions of this pattern into both clusters.

3. Signal Detection Techniques

JAKOWATZ, SHUEY AND WHITE:[21] A sample of the input waveform is stored in the memory of a correlation detection device. When the dot product of the incoming waveform and the stored waveform in the correlator exceeds a threshold the waveform stored in memory is modified by adding in a certain fraction of the waveform stored in memory as it exists at the time the dot product is maximally greater than the threshold. The threshold grows with successful detection and decays with failure to detect. Parameters controlling the algorithm include the threshold level, the memory weighting factor, the threshold setting factor, the threshold decay factor, the filter length, and the bandwidth of the filter. It uses only one correlator. Computation goes as $N_T \times D$. Convergence shown analytically and experimentally.

This technique and those in this section of signal detection techniques are primarily used for signal detection, and as they are presently conceived, their utility outside of this area seems limited (with the possible exception of the proposals by Smith).

GLASER:[19] The incoming waveform is detected on the basis of its energy. When a signal has been detected because of its energy, it is set into the correlation detector. Future decisions are based on a mixed weighting of detection by the energy detector

and by the correlation detector. Detection threshold, memory weighting factor, and incoherent vs coherent detection factor are the parameters provided in the program. Computation goes as $N_T \times 2D$. Convergence shown analytically.

SPILKER, LUBY AND LAWHORN:[41] The incoming signal is quantized into three levels. Initially, if the waveform has energy greater than a threshold, then the correlation detector is modified. Future detections then depend on a mixed weighting of the energy detector and the correlation detector. Parameters provided are the detection threshold, the memory weighting factor, and the incoherent vs coherent detection factor. Uses only a single correlator. Computation goes as $N_T \times 2D$. Convergence shown analytically and experimentally.

SMITH:[38] Detects the waveform using energy detection. Energy detection causes the waveform to be stored in the correlator. Gradually correlation detection takes over. If multiple signals are to be detected, then one correlation is trained first and then a second correlator is trained after the first one is trained. Detection threshold, memory weighting factor, incoherent vs coherent detection factor are the parameters provided. A second cluster is formed if the first correlator does not detect a signal when the energy level threshold is exceeded. No mechanism is given for destroying clusters. The computation goes as $N_T \times 2D \times$ NROWS. The computations required by Smith's method are more complicated than Glaser's or Spilker's. It is not apparent that a great deal is gained by adding the complexity. The method of adding new clusters does not appear to be very effective in problems having a large number of clusters and it might be damagingly ineffective.

4. Comments on Signal Detection Technique

By using more memory to store possible patterns it might be possible for these techniques to avoid the difficulties caused by using energy detection to start the process going. That energy detection is hindering can be seen in the foregoing simple example, taken from Spilker et al. Spilker[41] (p. iii) mentions a definite thresholding effect on the performance of the filter at a -13 db. Calculations show that the marginal distributions for each component overlap to a very considerable extent (about 45 percent probability of error for decision made on the basis of any one component by itself). It can be shown, however, that if one is able to use a signal with a thousand components in it, then the probability of error when the problem is taken as a multivariate normal decision problem is only 0.0002, which is very small. When energy detection must be performed, the probability of error is considerably increased above this small number. By storing a number of signals a guess as to possible cluster locations can be made without using energy detection.

One point that Glaser[19] (p. 93) raises seems important enough to quote:

> For weak signals the component estimates will tend to be large and induce erratic behavior in the adaptor system. It appears that there may be a minimum signal strength which the signal system can adapt to, and this also may be a function of signal waveform. Signals abo this minimum will be successfully adapted to at a rate which decrease as signal strength increases. No calculations on this minimal adaptable signal have been made.

This hypothesis seems worth further investigation. It appears to this author that techniques that avoid the necessity of making threshold decisions based on total signal energy may significantly improve performance.

Brennan[9] (p. 70) quotes three factors as determining convergence for the system of Jakowatz, Shuey and White:

a. False alarm rate less than signal rate.

b. Probability of finding second signal greater than 1/2.

c. Best estimate of certain coefficients. These coefficients depend entirely on the memory factor γ.

. . . .

The final optimum which the filter can reach ... is dictated entirely by the memory factor.

He further (p. 72) comments that it is desirable for the filter to be not quite twice as long as the signal, and that the reduction in signal-to-noise level due to increasing W (W equals filter bandwidth) might then be well compensated by rapid convergence, or even making convergence possible (p. 73). He goes on to point out the importance of keeping the memory weighting factor low when the nature of the signal is not well known so that erroneous signals can be easily erased (p. 75).

Smith[38] uses the squared dot product between the filter and the incoming figure, as well as the dot product unsquared. It is difficult to see what effect this has on performance (except to make calculations more complicated). He does show an interesting example (on p. 21): "It can be seen that the lack of timing information makes it necessary to examine the paths of the signals throughout the space to be sure that there is no ambiguity in the decision regions."

Here he is pointing out that long signals entering into the correlator may appear to be a different short signal due to the limited number of sample points in the filter. Therefore it is necessary to examine translations of a signal to see if this type of ambiguity exists.

Smith has an interesting approach to detecting distinct multiple signals. He suggests, first, detection of one signal until an adequate representation of that signal has been obtained, and then the use of that representation to distinguish the first signal from a second signal. Clustering techniques indicate that this can be avoided by having more than one cluster point, or cluster vector, to adjust at a time.

It would appear that all the techniques are troubled by the epoch (or translation of the time origin) problem. Perhaps this also could be overcome by a greater use of storage (sufficient to store every shift of the signal during the time when the signal occupies a central position within the correlator). This suggests the idea of a double correlator, a correlator within a correlator, to determine when the signal is centrally located in the larger filter.

A paper making some interesting philosophical points, though following Jakowatz' technique, is that by Turner.[44]

5. Clustering Techniques

OKAJIMA, STARK, WHIPPLE AND YASUI:[31] Patterns are selected in random order, weighted, and compared with a set of "typical" patterns, using a normalized correlation measure of similarity. The typical pattern correlation that is maximum is compared with a threshold. If the threshold is exceeded, then the pattern is added into that filter. If the threshold is not exceeded, then a new filter is created. A smoothness of convergence parameter as well as a similarity threshold is used to control the program. Clusters are not destroyed in the process. Cluster points are the average of all of the patterns associated with that cluster point. Computation goes as D × N × NROWS × ITER. The clustering is dependent on the order in which the individual patterns are presented to the program. This appears to be quite an excellent technique.

SEBESTYEN[36] and SEBESTYEN AND EDIE:[37] A pattern is selected and compared with the existing cluster centers. The measure of similarity is a weighted Euclidean distance with a weighting by factors that depend on both the particular component and on the particular cluster. The minimum distance of the pattern from a cluster point is compared with two thresholds, one smaller than the other. If the smaller threshold is not exceeded, then the pattern is added into that particular cluster and a new mean computed. If the smaller is exceeded but the larger is not, then the pattern is rejected for the present time, to be used at a later stage in the process. If the pattern distance exceeds the second threshold, a new cluster is created. Parameters controlling the processing include the two thresholds, and a parameter that controls when rejected patterns are forced into the various clusters. Computation goes as D × N × NROWS × ITER.

This technique is specifically pointed toward providing as an output a probability distribution based on the sample data and the technique's complexity is increased by this goal.

HYVARINEN:[20] By using a typicality measure, a single pattern is selected as a starting point. Other patterns that are within a certain distance of this pattern are clustered with this pattern. All patterns that are clustered together are then removed from the total set of patterns. This reduced set of patterns is then examined for another most typical pattern on which to start another cluster. The threshold controlling similarity is a process parameter. Cluster centers are never modified after the initial selection of a typical pattern. Computation goes as D × NROWS × N. Computation of typicality of patterns may prove expensive for large pattern sets. It would appear that this technique requires data that is more structured than other clustering techniques of this type. The measure of typicality critically determines the effectiveness of this procedure.

BALL AND HALL:[4] Several patterns are selected as trial cluster points. All patterns are then clustered around these trial cluster points and an evaluation of the clustering is made. If the maximum marginal standard deviation for any cluster is too large, and if certain other conditions are satisfied, then the cluster is broken into two clusters by creating two cluster points out of the original cluster. A second allowable number of patterns in a cluster, the allowable maximum come closer together than a given threshold. The parameters that control a program are the minimum allowable size of a cluster, the allowable maximum standard deviation in the cluster, and the minimum distance between two cluster points. Clusters are formed by splitting

existing clusters. They are destroyed when two cluster points are lumped together or when a cluster gets too small. Trial cluster points are modified only after all patterns in the set have been clustered. The computation goes as D × N × NROWS × ITER. It can be generalized to include clustering about line segments and planar sections. Block diagrams of optical implementation have been developed.

6. Comments on Clustering Techniques

Hyvarinen's technique appears to have one major drawback: it is not iterative. If the measure of typicality used is thrown off by some peculiar structure of the data, then it appears possible that clusters could be obtained that would be quite lopsided and not as intuitively pleasing as ones obtained iteratively.

The techniques of Okajima et al and Sebestyen seem very similar in that in both of them distance from a nearest cluster point is compared with a threshold. If this threshold is exceeded, then a new cluster is formed. If it is not (with the exception of the guard ring in Sebestyen), then the pattern is added into that closest cluster. The similarity was stronger in the early forms of Adaptive Sample Set Construction (see Sebestyen[36]), than as modified in Sebestyen and Edie.[37] The later technique has many fine features that now allow it more truly to represent the data as an empirically derived probability distribution.

Sebestyen's technique differs from the other techniques in his use of a metric that is sensitive both to the individual cluster and to the dimension in which a measurement is being made. This should provide greater flexibility in its description of the data. If, however, one of the purposes of the technique is to provide a description of the data, then this may not be an advantage in that it may be more difficult for an individual attempting to visualize the structure of the data to bring into play these factors than it is for him to accept additional cluster points. (See discussion under "Shortcomings of Cluster-Seeking Techniques" above.)

The technique of Ball and Hall differs from the other clustering techniques in several ways. First no absolute threshold is used in the measure of similarity criterion of the technique. A pattern is assigned to the closest cluster point regardless of distance. (This has disadvantages in terms of wild shots that lie a great distance from any cluster center. However these rapidly become cluster centers by themselves and then no longer bother the process.) The criteria used to evaluate the "goodness" of the various clusters and in the splitting of the cluster appear to be conceptually different. The lumping in the Ball-Hall technique produces a similar result to that obtained by Sebestyen when he combines sample sets that are not contributing significantly to classification accuracy by their being distinct. In some sense the goals of the two techniques are different, in that Sebestyen's technique is to develop a probability density that can be used in making classifications, whereas the Ball-Hall technique is pointed toward an adequate representation of the data, initially, at least, without regard for the classification to be performed. A further distinction between these techniques is the sorting of all of the patterns by the Ball-Hall technique prior to the modification of any of the sorting criteria. This makes the technique independent of the sequence in which the patterns are presented.

One note of warning should be added regarding the use of metrics weighted with respect to cluster as well as component. For example, if distance were measured as in ISODATA without using a threshold, one would find that the decision boundaries no

longer had the simplicity and intuitive appeal of the perpendicular bisecting plane. Now it would be possible for a small dense cluster to be located in the interior of a larger cluster that was not so dense. Possibly this would be an advantage but it would make interpretation more difficult when the technique is used for data analysis.

Okajima et al provide a contribution in the use of their smoothness of convergence criteria. This is used to make it more and more difficult for a pattern to be added to a cluster as the cluster grows larger.

It is felt by the author that the Ball-Hall technique is so structured that it is better suited to generalizing to new types of clusterings, for example, clustering about line segments or planar sections.

One of the present weaknesses of the Ball-Hall technique is the weakness of the existing criteria for determining the adequacy of clustering. Hyvarinen's suggestion of using an information theoretic measure to determine the most typical pattern in order to choose a best starting place for a clustering is an interesting one.

Stark makes an important point (see Okajima, Stark et al,[31] p. 108) when he discusses how a modification of the metric of the classification space will cause a different grouping of data vectors to occur. This is one of the most vexing and apparently unsolved problems in clustering techniques having no externally supplied guidelines as to the relative importance of the various measurements. It appears possible that some rationale for the selection of scales could be developed by examining the changes in clustering obtained as a result of changing the scales of the various patterns.

Stark (p. 109) points out that modifications in the clustering caused by changes in order of the pattern presentation may provide some indication as to the structure of the data itself.

It is the opinion of the author that the clustering type of cluster-seeking techniques may well be the most profitable direction in which to extend out efforts. Much more analysis of these techniques has to be done. Possibly we will then discover that these techniques are controlled by a form of feedback supplied by the structure of the data itself.

7. Clumping Techniques

MICHENER AND SOKAL:[28] The nucleus of a cluster is established using those two patterns having the highest pairwise coefficient of correlation. Then patterns are added to this nucleus one at a time, always adding first the pattern with the highest average correlation with the members of the group. The limit of the groups could be found by decreases in the level of the average correlation. A "significant" drop was empirically determined. Correlations between small clusters were used to group these small clusters into larger clusters. Computations go as $D \times (N(N - 1)/2)$.

This technique and the two next techniques require a number of comparisons to be made and additional computations to be performed. The extent of these computations is difficult to determine without examining specific data sets. The calculation of all pairwise distances between all patterns is considered to be a relatively inefficient way to obtain the structure of the data. This technique is best suited to taxonomic applications.

ROGERS AND TANIMOTO:[34] A function related to information theoretic entropy is computed and minimized. This selects the point pattern nearest the centroid of the system of patterns. This most typical pattern is designated "a prime mode" and a clump of cases very similar to it are clustered around it. Patterns are added to this clump until the inhomogeneity measure suddenly takes a large jump in value, indicating that the last pattern added was not truly a member of this clump. Computations go as $D \times N(N - 1)/2$. Random access memory required is $N(N - 1)/2$. Patterns considered are binary in value.

NEEDHAM;[29] PARKER-RHODES;[32] ABRAHAM:[1, 2] The distance between all pairs of patterns is computed for the pattern set. A threshold related to the average distance between the patterns is computed and the similarity matrix (i.e., the matrix computing all pairwise distances between patterns) is thresholded at some level, reducing the matrix to a matrix of zeros and ones. The i^{th} row thus contains a list of all objects to which the i^{th} object is closer than the threshold. It is possible, working with these lists of objects, to obtain clusters of interrelated objects. One method of doing this is to take a list of patterns related to one pattern, to take the next pattern on the list, and to take the intersection between the two lists of these two patterns, etc., until the number of patterns at the intersection is satisfactory. In this way fundamental clumps can be constructed, which can then be used to build up other clumps around them; that is, these clumps can serve as nucleus clumps. One parameter of the program is the threshold above which patterns are declared related, and below which, not related. Computations go as $D \times N(N - 1)/2$. Random access memory required is approximately $N(N - 1)/2$. This technique seems primarily useful for taxonomic problems. The techniques proposed are suggested both intuitively and by the use of graph theory.

BONNER:[8] Two methods are proposed. The first computes a similarity matrix, giving the distance between all pairs of patterns, and then manipulates this similarity matrix. The second, which is the one we shall describe, as it appears to be more satisfactory, takes a "random" center pattern and builds a cluster around this by utilizing an arbitrary threshold. All members more similar to this center pattern than the threshold are considered to be in the crude cluster. The members of this crude cluster are then examined by more refined criteria to determine whether or not they should be considered as part of this cluster. The yardstick of cluster goodness used is a measure of the rarity of the cluster. The computation goes roughly as $D \times N$. Additional computations are required that depend on the character of the data. Program requires random access to all the set of patterns in order to improve the clusters obtained efficiently. Patterns used were binary. The use of a threshold around an arbitrary pattern seems like a useful way to restrict the number of measurements of distance that must be made between pairs of patterns, if the distance between pairs of patterns is to be measured.

FORTIER AND SOLOMON:[17] Computes the correlations between all pairs of patterns. The absolute value of the difference between this correlation squared and an arbitrary constant is stored in a "distance" matrix. The constant used should be related to the loss expected if two items are clustered that are not strongly related. The matrix is then summed over all clusters of pairs. That is, compute a number that is the sum of the "distance" between all pairs of patterns in a given cluster. Sum this number over all clusters. The attempt would then be made to maximize this final double summed number. Computation goes as $D \times N(N - 1)/2$ plus computations

necessary to maximize the sum of the differences between these various correlation coefficients and the threshold for each of the clusters formed. The techniques discussed in this paper seem particularly to bear out the combinatorial problems in clustering. They emphasize the necessity for making approximations and for iteratively computing the desired quantities rather than trying to compute it using all the possible information. These techniques seem quite limited with respect to the problems they can attack because of the restrictions on the number of clusters the techniques proposed can consider.

SAWREY, KELLER, AND CONGER:[35] The procedure selects small subsets of nearly identical patterns using the Euclidean distance between patterns. Based initially on a small number of basic profiles, highly homogeneous groups which are also dissimilar to each other are formed. To these homogeneous groups, the remaining profiles in the sample are then compared. As a result of these comparisons, additions are made to one or another of the groups by successively relaxing the criteria of group homogeneity. Profile groups are formed in which the group members are similar to each other but at the same time dissimilar from the members of all other groups. The paper gives quite a thorough analysis of the application of these clustering techniques to psychological data. The computations involved will depend quite definitely on the nature of the data though all pairwise distances must be computed.

8. Comments on Clumping Techniques

The techniques of Michener and Sokal[28] were developed for the specific purpose of obtaining a numerical taxonomy of a collection of animals, in this case, bees. The technique itself is highly oriented toward attaining this taxonomy, and as a result seems less useful for the more general problems of clustering than some of the other techniques. The comments found in a later book by Sokal and Sneath, Principles of Numerical Taxonomy,[40] are worth reading, however, and for those interested in approaching the formation of subsets by providing a nucleus pattern or patterns and clumping around them, this book is an excellent source of references. One quote from Michener and Sokal follows:

> By using a large number of characters and species (i.e., measurements and patterns) however, we feel that weighting becomes unnecessary because the magnitude of correlation coefficients calculated between species (distances between patterns) would be little affected by extreme weighting.

This seems to indicate they feel that in some sense the most valid general clustering one can obtain is by using a large number of descriptors of the pattern. (With limited sample sizes increasing the number of measurements may, however, lead to a decrease in "clusterability.") However, in the case of clusterings toward a particular end — for example, clustering weather measurements to obtain indications of high ceiling and low ceiling heights — the weighting of one or another measurement is then tied to the particular task for which the system is intended. For this reason, this particular comment seems most valid only when trying to develop a general taxonomical relationship between the various patterns.

The paper by Rogers and Tanimoto was one of the earliest papers in which a computer was a necessary and integral part of the process used. The program developed seems particularly significant in its attempt to allow the operator to

introduce new cases or attributes whenever these are found to be irrelevant to the clustering to be performed. A second interesting aspect of this paper is the use of information theoretic measures to determine the quality of the clustering obtained. If these measures from the clumping techniques were combined with a more efficient method of associating patterns together, such as is found in the clustering techniques, it appears that the resulting algorithms would be particularly powerful for sorting patterns into subsets.

The papers by Parker-Rhodes,[32] Needham,[29] Abraham[1,2] and Kochen[24] are all concerned primarily with the use of what might be called graph theory for the determination of clusters in a set of data. A common characteristic of these papers is the use of a similarity matrix as a sufficient description of the universe of patterns. As defined in these papers: "Broadly, members of a clump must be more like each other and less like non-members than elements of the universe picked at random" (Parker-Rhodes,[32] p. 9).

Much of the work in these papers is set-theoretic and attempts to determine some rather general concepts concerning what can be considered a clump and what is not a clump. The major problem of these techniques is that they all increase in computation at least as the square of the number of patterns which is a considerable limitation on large problems.

The paper by Bonner[8] contains a large number of insights into the clumping approach to obtaining subsets. Many of these comments are useful regardless of the technique class that is being used. Three algortihms are described in this paper. We have described only the third in any detail. Additional measures are provided in this paper for measuring the strength of internal clustering vs the strength or external interactions. Bonner suggests the following criteria for clustering:

> The clustering problem ... becomes the problem of finding sets of objects where the attributes are estimated to be independent within a set ... The philosophy of the clustering procedure to be followed involves finding sets of objects which do not come from this population. If none can be found, the original object set is judged to be one cluster. If any is found, it is removed from this set as a cluster.

Bonner then describes a statistic that can be used to determine the probability that the objects clustered together, if picked at random from a hypothetical population, are statistically significant.

His major conclusion regarding clustering techniques and factor analysis is an interesting one.

> The major point to be made is that clustering methods ... can be used for problems now done by factor analysis. It is not implied that such a cluster analysis should replace factor analysis, but that both methods applied to the same data should yield a deeper understanding than either method alone.

The paper by Fortier and Solomon[17] is notable in its spelling out of the difficulties involved in handling a clustering of a large number of patterns in high dimensions by direct methods. As they say (pp. 3-4):

> Theoretically speaking, we have a finite problem at hand for which there exists a solution: i) prepare a list of all possible partitions of a set of N points. ii) for each partition (set of clusters) calculate the B function, iii) choose the clustering configuration which corresponds to the optimal (large) B value. However, matters get out of hand rather quickly, for the amount of computation becomes inordinately large, even when N is moderate in size.

The paper contains many tables showing just how out of hand things can get if this point of view is taken, and some approximations are not made.

Two other papers worthy of mention are Cattell[10] and Kaskey.[22] These two papers deal primarily with the analysis related to clustering measurements as opposed to clustering patterns, and for this reason have not been included in detail. The paper by Cattell outlines some of the methods that were used prior to the advent of the electronic computer. The paper by Kaskey makes considerable use of the electronic computer and develops a significance test to be used on the correlations between various measurements. One of the limitations of both of these approaches seems to be the implicit assumption that the set of patterns being analyzed is in some sense normally distributed so that correlations have the desired significance. It seems quite probable that in cases in which a relatively large number of patterns is being used and in which a priori knowledge is not really adequate to divide these patterns into homogeneous subsets that this will not be true. It is therefore recommended that these procedures be amended to, first, cluster patterns into relatively homogeneous subsets, and secondly, to perform the clustering of the measurements. While the interpretation of the results of such an analysis would probably require new attitudes towards the information contained in the data, it does seem that the results obtained would be much more directly related to the physical phenomena underlying the data.

Cattell also notes in his article that Holzinger's B coefficient techniques might be used to relate the density within a cluster to the density of patterns immediately surrounding a cluster. Both Rogers and Tanimoto, and Michener and Sokal use the concept of adding patterns one at a time, until an inhomogeneity measure increases sharply over previous increases. This seems related to the same concept of learning how isolated a cluster is from the other patterns. In analyzing some real data, we have found that while isolated clusters do occur, an interesting and useful description of the data can be obtained by clustering (in the sense of relating together a group of similar patterns) even though each of these "clusters" is not really isolated from the surrounding patterns.

The general comment regarding clumping techniques is that they are in general non-iterative in the sense that there is a specific procedure that determines when a pattern is added to a clump. The use of a once-through, non-iterative procedure requires the selection of a stricter criteria of what constitutes a cluster than would be required if an iterative technique were used. As suggested above, it does appear that the combination of a good criteria of what constitutes a cluster with an iterative technique should provide an extremely powerful synthesis of clustering and clumping techniques.

9. Comments on Eigenvalue-Type Techniques

Nunnally[30] is in some sense intermediate between the clumping technique and the eigenvalue techniques. His technique is similar to the clumping techniques in that he uses the distance matrix, giving the pairwise distances between all of the patterns, and is similar to the eigenvalue techniques in that he uses the principal directions in the distance space to define the clusters, i.e., he uses the eigenvectors of the distance matrix.

This technique is particularly interesting in that it ties together the more classical approach of statistics, e.g., factor analysis, and the principal components analysis, to the techniques of clustering and clumping. Nunnally has several useful comments on the importance of considering what he calls level (signal amplitude), shape (which corresponds to direction in our pattern space) and dispersion (which is related to the variance of the pattern around its mean level); these comments are more directly related to psychology than to other types of patterns. Nunnally's work is in some ways reminiscent of the work of Bonner.

One quote from Nunnally seems particularly relevant:

> The decision to use profile (cluster) analysis is determined in part by preferences for methodologies, which are, in essence, wagers about the likely research payoff in the long run, from choosing one method of investigation rather than another. The reader can judge for himself whether the studies using comparisons of profiles (e.g., measures of "assumed similarity" in interpersonal perception) have borne the expected fruit.

With the availability of computers increasing and the cost decreasing, it seems very worthwhile to use both the more classical factor analytic or eigenvector methods and the methods of profile or clustering analysis on any given set of data. We feel that the two viewpoints are sufficiently different that they can each supply important information regarding the data.

The paper by Cooper and Cooper[13] appears to depend rather heavily on a number of assumptions regarding the nature of the data. While these assumptions are necessary for their analytic examination, and while they have suggested ways of getting around some of the assumptions that they have made, we feel that much work remains to be done before this technique will be of value when applied to problems in which little is known about the data. At this point it is worth noting the relationship between the technique of Mattson & Dammann[27] and that of Cooper and Cooper. In some sense it appears that Mattson and Dammann have taken the underlying philosophy of the Cooper and Cooper technique and extended it to problems where little is known about the structure of the data. This relaxation of assumptions makes the Mattson and Dammann technique particularly useful. The most prevalent assumption in the Cooper and Cooper paper is the one of the data being formed from two Gaussian distributions. In this respect it is reminiscent of the approach of Patrick and Hancock.[33]

An interesting quote from Cooper and Cooper[13] (p. 419) is the following:

> As is to be expected, non-supervised adaption cannot be uniquely achieved for arbitrary distributions. But where there is adequate probability structure of the problem, the partition can be unique. There are many cases for which this is possible. Of special importance is the two-category case where the distributions are translates of one another, but have general functional form, and further interest centers on the cases where the distributions are finitely parameterized.

While it is true that some of the clustering techniques that have been discussed do not arrive necessarily at a unique solution in a mathematically rigorous sense, it is nonetheless intuitively clear that in well-structured data the partitions achieved will be what might be called "epsilon-unique," that is, the partitions achieved will vary only slightly (for the same parameter settings of the clustering program). The most pertinent variable in determining the size of epsilon in this case is the inherent structure of the data. If the data exists in clusters that are compact and well isolated from each other, then the partitioning will be unique. If the data, on the other hand was drawn from a uniform distribution of random numbers, then in all probability the clustering will vary with every attempt at clustering. It has been our experience that most data lies somewhere between these two extremes. Perhaps one of the most useful aspects of the clustering techniques is their ability to indicate which extreme a particular set of data lies nearest.

The paper by Mattson and Dammann[27] is in the author's opinion an excellent example of combination of analytical and intuitive approaches. While the technique follows well-defined lines in terms of what is computed, nevertheless the concatenation of these various well defined mathematical functions arrives at a technique that is considerably more useful than any one of these techniques applied separately. It is worth noting that this technique aims toward synthesizing efficient threshold networks. The technique is, however, useful for much more than that, in that it makes possible an examination by the researcher of the nature of his data in a relatively limited region of the space. Perhaps the most important aspect of this technique is its use of direction in data, with the breaking up of the data into smaller subsets preventing heterogeneous subsets of the data from confusing the analysis.

Both Mattson and Dammann[27] and Cooper and Cooper[13] are in effect searching for valleys between high density regions in the data. A technique which will be described below (Bledsoe[6]), is a third technique that has at its core the projection of data onto a line and the searching for low density spots in the distribution of the patterns along this line.

10. Minimal Mode-Seeking Techniques

FIRSCHEIN AND FISCHLER:[16] Computes dot product of patterns with cluster centers (classes are partitioned into subclasses so that each member of a particular class is closer, in the sense of high correlation score, to the centroid of its one subclass than to the centroid of any other subclass); to classify an unknown vector, we

determine the closest subclass centroid to the vector and assign the unknown vector to that class. Patterns are divided into three classes:

1. Patterns having highest dot product with centroid of its own subclass
2. — with a centroid of another subclass which is in the same overall class as the pattern
3. — with the centroid of another subclass which is in another class than the one given.

New subclasses are formed by having that pattern having the lowest dot product with its own subclass centroid chosen to form a new subclass centroid. The procedure is iterated until allowable error rates are reached or some arbitrarily chosen number of system iterations has been made. Computation goes as N × NROWS × D × ITER. This technique and the following technique depend on the use of classification information in determining the subclasses. The technique may run into difficulty in cases where the pattern classes are overlapping unless some method is found for recognizing patterns that lie in the overlapping areas.

STEINBUCH AND PISKE:[42] Subclasses are formed if the distance between a pattern and a mode of its particular class is greater than a fixed threshold amount. The procedure is iterated until adequate separation is achieved, or other constraints are satisfied. Computation goes as D × N × NROWS × ITER.

11. Comments on Minimal Mode Seeking

Firschein and Fischler's technique[16] appears to be quite useful in cases in which pattern subclasses are linearly separable. However, from our experience at SRI it appears that modifications of the technique will be necessary when pattern classes are badly overlapped and intermixed in one another. That is, some method of storing patterns that are consistently causing difficulty should be used so that new subclasses would not be created for these patterns, if, for example, they lie right in the central part of another class.

A significant quote from this paper[16] (p. 138) is:

> Unlike previous procedures for subclass formation, this method does not require the specification of an arbitrary fixed distance as a criterion of membership in a subclass. Another advantage of the present technique is that it is not necessary to specify the required number of subclasses beforehand.

We consider these both to be significant aspects of this technique.

The paper by Steinbuch and Piske[42] we found difficult to read insofar as the recognition and clustering characteristics of the learning matrix are concerned. The article appears to be primarily a description of the learning matrix itself, rather than a description of the ways in which it operates, and its limitations. From our other experience it does appear that this technique would be useful, though some of the methods of normalization reduce its general applicability. It seems important to note that category information is required for this and for the Firschein and Fischler technique.

12. Miscellaneous Technique

BLOCK, KNIGHT AND ROSENBLATT:[7] The computational unit consists of two layers of summing networks followed by thresholds. The summing units on the first layer are in one-to-one correspondence with the summing units of the second layer and have a threshold of θ. The weightings on the connections between the first and second layer are incremented if, when the threshold of the first layer unit is exceeded at time T, and the threshold of the second layer unit is exceeding at time T + 1. All connections are decremented each unit of time. The size of this decrement in the thresholds is a control parameter of the program. The computation goes as D × N × NROW × ITER. Classification information is inferred from the sequence in which patterns are shown the machine. It is assumed that there is a high probability of continuous runs of patterns coming from the same class occurring in contiguous runs.

BLEDSOE:[6] An arbitrary plane passing through the patterns is selected. Distances from this plane are computed for all of the patterns. The average distance of the pattern from this plane is maximized by a series of iterative adjustments of the plane. This procedure is tried for several different initial starting points. The plane having the maximum average starting distance is selected as the best plane. All patterns are projected onto this plane and a second plane in D-1 dimensions that maximizes distance from all the patterns is sought. (The procedure is constrained to pass the plane in such a way that all the subsets formed by the plane are of approximately equal size.) Computations go approximately as N × D × ITER. The results given in this paper were not really adequate to determine the effectiveness of the technique.

13. Comments on Miscellaneous Techniques

The method of Block, Knight and Rosenblatt is an ingenious method of taking advantage of high probability of runs. It is commented in that paper that the techrique also has generalizing properties with regard to particular types of transformations. The paper referenced has a detailed analysis of the technique and its performance. The technique appears limited by the contiguous-run requirement. It is stated in the article that patterns need not necessarily look alike if they are contiguous in sequence for them to be classified together. In other words, this is in some sense not a clustering technique, but it is a self-organizing technique that does not require an explicit external teacher.

The procedure Bledsoe follows in his paper is essentially one of trying to learn the "best" corridor (in the sense of a hyper-plane) to the learning set. A method appears adequate to do this but it is difficult from the description to tell how generally useful this technique would be. It might provide useful preprocessing, for, say, alphanumeric character recognition, in that it would in some sense "optimally" divide up the pattern set.

It seems worthwhile to note that a new calculation is required each time the dividing plane is adjusted into a new position. This new calculation determines whether the plane is appropriately adjusted or not. It is not apparent from the paper that a hill-climbing technique is being employed except in a probabilistic way.

BLANK PAGE

SECTION VII

COMPOSITE CLUSTER-SEEKING ALGORITHM

The composite technique presented in this section is an attempt to improve the ISODATA algorithm by including the good points of other cluster-seeking techniques described. Though it seems unlikely that this one technique will be suitable for all types of data, we feel that it contains no obvious shortcomings.

1. Composite Technique (1965)

Compute the average of patterns and the average distance of all patterns from this average. Set a threshold θ_D = k x (average distance of all patterns from this overall average) where $0 \leq k$. All dimensions would be scaled at this time to cause each dimension to have a standard deviation of 1 for all patterns taken together. Until 3rd iteration minimum allowable cluster size is 1. The overall average of all patterns would be used as the initial cluster point.

START: Compute dimensions of all patterns from all existing cluster points. If minimum distance (measured using Euclidean distance) from a pattern to any existing cluster point exceeds θ_D, then create a new cluster point at the location of that pattern. (This should rapidly locate small isolated clusters.) Evaluate the goodness of the cluster, perhaps using a criterion similar to that used in the clumping techniques.

SPLIT: If clustering is not satisfactory, then "SPLIT" either those clusters having unsatisfactory cluster properties or those clusters having maximum standard deviations greater than a threshold (if the cluster also is of sufficient size and has sufficiently large average pattern distance from the cluster center). Return to (START) and repeat process of computing pattern distances from the new cluster centers.

LUMP: Combine (LUMP) cluster points that are closer than a given threshold. Return to (START) and repeat process using (SPLIT). At the end of this iteration all cluster points having fewer than a given number of patterns associated with them would be removed from the list of patterns (this should remove all wild shots). These patterns would be printed out.

Recompute new "standard deviations" for each dimension for reduced pattern set and rescale all patterns so that the normalized standard deviation in each dimension equals 1. These standard deviations could be the average using all patterns and computed about an overall minimum, or they could be the average standard deviation of patterns in each cluster about its own cluster center. This standard deviation might be further modified by taking into account the pairwise distances between cluster centers at each distance.

Lump and split until a criterion of adequacy of clustering is satisfied. The relative increase in measure of clustering from iteration to iteration could be used to determine when lumping and splitting should stop.

Computational difficulty is proportional to D × NROWS × N × ITER. Note: If all multiplications with elements in each cluster center can be done in parallel (optically) then this reduces to N × ITER.

Parameters that control the program are related to:

1. minimum allowable number of patterns in a cluster,
2. the allowable maximum standard deviation (or spread) of a cluster,
3. the minimum distances between two cluster points,
4. the fraction of the standard deviation used in the first iteration to control the initial creation of clusters and the discarding of wild shots,
5. the iteration-to-iteration difference required to imply convergence of clustering.

This technique could be generalized and probably implemented optically.

2. Comments on Composite Technique

It seems unlikely that any one algorithm will be ideal for all types of data and for all situations. It does seem possible, however, to describe types of data, and to determine the type of cluster-seeking algorithm that seems most useful for handling each type of data. A need exists for a preliminary, exploratory technique to tell which of the more specialized types of techniques to apply. We need to be able to broadly categorize types of data so as to decide how to process them further in more detail.

This suggested technique is an attempt to include in ISODA the best (noncontradictory) aspects of other techniques. The following are improvements over the original ISODATA program:

1. The modification relating to the initial cluster-creating procedure should cause more rapid convergence.
2. Discarding and identification of wild shots from pattern sets should improve efficiency considerably.
3. The use of good clustering criteria would allow realistic termination of the iterative procedure.
4. Scaling of dimensions so that some (yet to be determined) criterion is satisfied would remove from clustering some of the arbitrariness arising from the choice of the unit of measurement used in the various dimensions.

SECTION VIII

CONCLUSIONS

The main point we wish to make is that cluster-seeking techniques exist that can be used to organize data from the social sciences and other disciplines in a way that allows examination of the details, i.e., the individuals, in the data. Using these techniques a single data point can be related to an adequate description of the rest of the data. Cluster-seeking techniques can be effectively utilized by persons having relatively little formal mathematical and statistical training. These techniques have the further advantage that the effect of changing scale or measurements can be directly and easily related to its effect on the data. In other words, the relative arbitrariness of calling certain patterns "similar" becomes more apparent.

An important consideration in determining the value of the clustering obtained is the use to which the clustering is to be put. That is, is the clustering obtained from the data to be used to determine the significance of the data in a statistical sense, or is it to be used primarily as a descriptive organizing of the data for the researcher? It is important in criticizing or commenting on clustering techniques to keep this distinction clear. Some clustering criteria are not adequate to determine significance of the data; nevertheless, the clustering may provide a completely adequate description of the data that is very suggestive of new experiments to be performed or new interpretations of the data that then can be tested by more rigorous methods.

With respect to the specific cluster-seeking techniques, we regard the "clustering" techniques as providing the most efficient and easily interpreted clustering, except in taxonomic problems where clumping techniques (modified to be more efficient) appear best. The addition of more adequate cluster criteria to the "clustering" techniques seems a natural next step.

Much more work needs to be done on convergence of these techniques and on methods for interpreting and examining the resulting clusters.

It appears important to us not to rely entirely on some function of the data such as the covariance matrix. Rather it seems that for data analyses, <u>particularly where the data base is small</u>, working directly with the patterns themselves is required. Naturally the patterns must be organized into some comprehensible form. Clustering the data is one way to organize them.

Finally, the distinction to us, between classical statistics and clustering with regard to data analysis is the distinction between the point of view (statistical) that immediately generalizes from the specific (the patterns) to the general (the estimated distribution), and the point of view (clustering) that remains at the specific (the patterns) and has the capability of analyzing the individual patterns for local as well as global relationships.

Speculation Regarding the Effect on the World Around Us

We feel that computer-oriented techniques that can quickly organize data in a way that allows rapid analysis of the data will profoundly affect experimental science.

Starting with existing clustering techniques and using proposed peripheral computer programs, it will be possible for the experimental scientist to see on a display the data he is gathering as he gathers it. The potential value in such rapid feedback seems enormous when we think how rapidly we forget all of the details of an experimental situation. We at SRI consider ourselves to be working toward this eventuality which may have considerable effect on the world around us.

REFERENCES

1. C.T. Abraham, "Evaluation of Clusters on the Basis of Random Graph Theory," unpublished report, IBM, Yorktown Heights, N.Y.

2. ———, "A Note on a Measure of Similarity Used in the DICO Experiment", Appendix I, Quarterly Report 3, vol. 1, Contract AF 19(626)-10.

3. D.C. Allais, "The Selection of Measurements for Prediction," Stanford Elec. Lab., System Theory Div. Report, TR6103-9 (AD 456 770), Stanford, Calif. (Nov. 1964).

4. G.H. Ball and D.J. Hall, "ISODATA, A Novel Method of Data Analysis and Pattern Classification," Stanford Research Institute, Menlo Park, Calif. (Apr. 1965).

5. P. Baxendale, "An Empirical Model for Computer Indexing," Machine Indexings Progress and Problems, American University, Washington, D.C., Feb. 13-17, 1961, p. 267.

6. W.W. Bledsoe, "A Corridor-Projection Method for Determining Orthogonal Hyperplanes for Pattern Recognition," unpublished report, Panoramic Research Cor., Palo Alto, Calif. (1963).

7. H.D. Block, B.W. Knight and F. Rosenblatt, "The Perceptron: A Model for Brain Functioning, II," Reviews of Modern Physics, vol. 34, no. 1, pp. 135-142 (Jan. 1962).

8. R.E. Bonner, "On Some Clustering Techniques," IBM Journal of Res. and Dev., Jan. 1964.

9. E.J. Brennan, "An Analysis of the Adaptive Filter," General Electric Elec. Lab. Tech. Information Series Report R61 ELS-20, Syracuse, N.Y. (1961).

10. R. Cattell, "A Note on Correlation Clusters and Cluster Search Methods," Psychometrika, vol. 9, no. 3 (Sept. 1944).

11. D.B. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Raytheon Report, October 22, 1963.

12. ——— and P.W. Cooper, "Adaptive Pattern Recognition and Signal Detection without Supervision," IEEE International Convention Record, pt. 1, 1964, pp. 246-256.

13. ——— and ———, "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Information and Control, vol. 7, no. 3 (Sept. 1964).

14. R.F. Daly, "Adaptive Binary Detection," Stanford Elec. Lab. Tech. Report No. 2003-2, Stanford, Calif. (June 26, 1961).

15. ———, "The Adaptive Binary-Detection Problem on the Real Line," Stanford Elec. Lab. Report SEL-62-030, Stanford, Calif. (Feb. 1962).

16. O. Firschein and M. Fischler, "Automatic Subclass Determination for Pattern Recognition Applications," Trans. PGEC, EC-12, no. 2 (Apr. 1963).

17. J.J. Fortier and H. Solomon, "Clustering Procedures," Tech. Report 7, Dept. of Statistics, Stanford University (Mar. 20, 1964).

18. S.C. Fralick, "The Synthesis of Machines Which Learn Without a Teacher," Tech. Report No. 6103-8, Stanford University (Apr. 1964).

19. E.M. Glaser, "Signal Detection by Adaptive Filters," IRE Trans. On Info. Theory, vol. IT-7, no. 2 (Apr. 1961).

20. L. Hyvarinen, "Classification of Qualitative Data," British Info. Theory J., 1962 pp. 83-89.

21. C.V. Jakowatz, R.L. Shuey and G.M. White, "Adaptive Waveform Recognition," Information Theory, C. Cherry, ed., Butterworths, Washington, D.C., 1961.

22. G. Kaskey et al, "Cluster Formation and Diagnostic Significance in Psychiatric Symptom Evaluation," Proc. Fall Jt. Computer Conf., 1962, p. 285.

23. H. Kazmierczak and K. Steinbuch, "Adaptive Systems in Pattern Recognition," IEEE Trans. on Electronic Computers, vol. EC-12, no. 6 (Dec. 1963).

24. M. Kochen, "Techniques for Information Retrieval Research: State of the Art," presented at IBM World Trade Corporation Information Retrieval Symposium at Blaricum, Holland, Nov. 1962; to be published in the proceedings of the symposium.

25. ——— and E. Wong, "Concerning the Possibility of a Cooperative Information Exchange," IBM Journal of Research and Development, vol. 6, no. 2, pp. 270-271 (Apr. 1962).

26. H.P. Luhn, "Auto-Encoding of Documents for Information Retrieval System," Modern Trends in Documentation, M. Boaz, ed., Pergamon Press, New York 1959, pp. 45-58.

27. R.L. Mattson and J.E. Dammann, "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems," submitted for publication to IBM J. of Res. and Dev., Mar. 1965.

28. C.D. Michener and R.R. Sokal, "A Quantitative Approach to a Problem in Classification," Evolution, vol. 11, pp. 130-162 (June 1957).

29. R.M. Needham, "The Theory of Clumps, II," Report M. L. 139, Cambridge Language Research Unit, Cambridge, Eng. (Mar. 1961).

30. J. Nunnally, "The Analysis of Profile Data," Psychological Bulletin, vol. 59, no. 4, pp. 311-319, 1962.

31. M. Okajima, L. Stark, G. Whipple and S. Yasui, "Computer Pattern Recognition Techniques: Some Results with Real Electrocardiographic Data," IEEE Trans. on Bio-Medical Electronics, vol. BME-10, no. 3 (July 1963).

32. A.F. Parker-Rhodes, "Contributions to the Theory of Clumps," I.M.L. 138, Cambridge Language Research Unit, Cambridge, England (Mar. 1961).

33. E.A. Patrick and J.C. Hancock, "The Non-Supervised Learning of Probability Spaces and Recognition of Patterns," Tech. Report, Purdue Univ., Lafayette, Ind. (1965).

34. D.J. Rogers and T.T. Tanimoto, "A Computer Program for Classifying Plants," Science, vol. 132, Oct. 21, 1960.

35. W.L. Sawrey, L. Keller and J.J. Conger, "An Objective Method of Grouping Profiles by Distance Functions and its Relation to Factor Analysis," Educational and Psychological Measurement, vol. 20, no. 4 (1960).

36. G.S. Sebestyen, "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Trans. on Info. Theory, vol. IT-8, Sept. 1962.

37. ——— and J. Edie, "Pattern Recognition Research," Air Force Cambridge Res. Lab. Report 64-821 (AD 608 692), Bedford, Mass. (June 14, 1964).

38. J.W. Smith, "The Analysis of Multiple Signal Data," IEEE Trans. on Information Theory, vol. IT-10, no. 3 (July 1964).

39. R.R. Sokal and C.D. Michener, "A Statistical Method for Evaluating Systematic Relationships," University of Kansas Science Bulletin, Mar. 20, 1958.

40. ——— and P.H.A. Sneath, Principles of Numerical Taxonomy, W.H. Freeman and Co., San Francisco, 1963.

41 J.J. Spilker, Jr., D.D. Luby and R.D. Lawhorn, "Progress Report — Adaptive Binary Waveform Detection," Tech. Report 75, Communication Sciences Department, Philco Corp., Palo Alto, Calif. (Dec. 1963).

42. K. Steinbuch and U.A.W. Piske, "Learning Matrices and Their Applications," IEEE Trans. on Electronic Computers, vol. EC-12, no. 6 (Dec. 1963).

43. H.E. Stiles, "The Association Factor in Information Retrieval," Communications of the ACM, vol. 8, no. 2, pp. 271-279 (Apr. 1961).

44. R.D. Turner, "First-Order Experimental Concept Formation," Biological Prototypes and Synthetic Systems, E.E. Bernard and M. Kare, eds., Bionics Symposium 2, Ithaca, N.Y., Plenum Co., 1961.

SUPPLEMENTARY BIBLIOGRAPHY

B.M. Bass, "Iterative Inverse Factor Analysis — A Rapid Method for Clustering Persons," Psychometrika, vol. 22, no. 1, pp. 105-107 (Mar. 1957).

Eigenvalue I — Uses iterative method to factor analyze test scores in order to cluster people.

W.D. Fisher, "On Grouping for Maximum Homogeneity," Journal of American Stat. Assn., vol. 53^2, pp. 789-798 (Dec. 1958).

Clumping — On real line examines all possible partitions and selects that partition minimizing the weighted square distance from the cluster average point.

E.W. Forgy, "Detecting 'Natural' Clusters of Individuals," Western Psychological Association Meetings, Apr. 19, 1963, Santa Monica, Calif. (Can be obtained from author at Center for Health Sciences, U.C.L.A., Los Angeles, Calif.)

Miscellaneous.

E.W. Forgy, "Evaluation of Several Methods for Detecting Sample Mixtures from Different N-Dimensional Populations," American Psychology Association Meetings, Los Angeles, Calif., Sept. 9, 1964. (Available from author at Center for Health Sciences, U.C.L.A., Los Angeles, Calif.)

Gives results of five methods for clustering on several artificial problems. The five methods are:

1. Using the frequency histogram of interpoint distances.
2. Sokal and Sneath (1957).
3. Ward (1963) - See below for reference.
4. Forgy (1963).
5. Factor analysis — Q sort.

J.A. Gengerelli, "A Method for Detecting Subgroups in a Population and Specifying their Membership," Journal of Psychology, vol. 55, pp. 457-468 (1963).

Miscellaneous — Analysis of distribution of pairwise distances between patterns.

L.I. McQuitty, "Typal Analysis," Educational Psychological Meas., vol. 21, pp. 677-696 (1961).

Clumping.

P. Medgyessy, Decompositions of Superpositions of Distribution Functions, Publishing House of the Hungarian Academy of Sciences, Budapest, 1961.

Discussed how a "sum distribution" that is the sum of a known number of elementary distributions of known form can be used to find the

parameters of the elementary distributions. Though not directly providing a clustering method, it may provide some analytical insight into the analysis of clustering methods.

R.C. Tryon, Psychometrika, vol. 22, no. 3, pp. 241-260 (Sept. 1957).

Uses clustering to indicate factor analytic communalities. Gives references to Tryon's earlier work on clustering.

———, "Domain Sampling Formulation of Cluster and Factor Analysis," Psychometrika, vol, 24, no. 2, pp. 113-135 (June 1959).

Eigenvalue I — Discusses "Domain Sampling" as a viewpoint underlying his cluster analysis and factor analysis. Tends to look at cluster analysis as related to correlations between the measurements found over the entire data base.

J.H. Ward, Jr., "Hierarchical Grouping to Optimize an Objective Function," Journal of American Statistical Association, vol. 58, no. 301 (Mar. 1963).

Clumping — combines those two patterns that maximally increase an objective function. Primarily related to Taxonomic structures.

——— and Marion E. Hook, "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles," Educational and Psychological Measurement, vol. 23, no. 1 (1963).

See Ward above.

J.H. Wolfe, "A Computer Program for the Maximum Likelihood Analysis of Types," Tech. Bull., 65-15, U.S. Naval Personnel Research Activity, San Diego, Calif. 92152 (May 1965).

Miscellaneous — Using an initial rough clustering procedure, the clustering is refined using maximum likelihood methods and an assumed underlying mixture of multivariate normal distributions.

G. Young, "Factor Analysis and the Index of Clustering," Psychometrika, vol. 4, no. 3 (Sept. 1939).

Proposes the dispersion of the eigenvalues of the covariance matrix as an "index of clustering."

BLANK PAGE

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA · R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Stanford Research Institute
Menlo Park, Calif

2a REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b GROUP
N/A

3 REPORT TITLE

A Comparison of Some Cluster-Seeking Techniques

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Interim Report June 66 - August 66

5 AUTHOR(S) (Last name, first name, initial)

Ball, Geoffrey H.

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
November 1966	47	44

8a CONTRACT OR GRANT NO.	9a ORIGINATOR'S REPORT NUMBER(S)
AF30(602)-4196 b. PROJECT NO. 5581 c. Task # 558104 d	n/a 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report) RADC-TR-66-514

10 AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	Rome Air Development Center Information Processing Branch(EMIID) ~~Griffiss AFB, New York 13440~~

13 ABSTRACT

Conventional multivariate statistics examines in considerable depth the significance of relationships existing in data as shown by the mean and the covariance matrix. Shortcomings of this approach are briefly discussed.

"Cluster-seeking techniques" are discussed as alternatives to conventional multivariate methods. Thirty variants of cluster-seeking techniques, the total number presently known to the author, are divided into seven categories: probabilistic, signal detection, clustering, clumping, eigenvalue, minimal mode seeking and miscellaneous. These larger classes are contrasted and, within each class, the techniques are summarized and compared.

A composite technique that combines the best features of the various approaches is proposed.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Techniques, cluster-seeking Data analysis Social sciences						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional

Security Classification