

A COMPARATIVE STUDY TO PREDICT
THE NUMBER OF CLUSTERS
IN CLUSTER ANALYSIS

BY

SEONG-SAN CHAE

Bachelor of Science in Statistics
Chung-Ang University
Seoul, Korea
1983

Master of Science in Statistics
Iowa State University
Ames, Iowa
1985

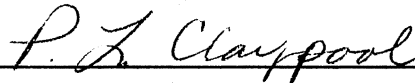
Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 1988

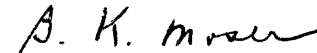
THESIS
1988D
C432C
COP. 2

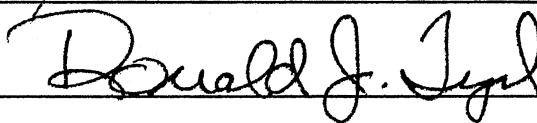
A COMPARATIVE STUDY TO PREDICT
THE NUMBER OF CLUSTERS
IN CLUSTER ANALYSIS

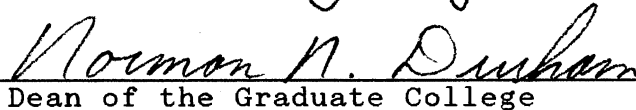
Thesis Approved:


Thesis Adviser








Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express my sincere thanks and appreciations to my adviser, Dr. William D. Warde, for suggesting the problem and for his guidance and assistance during the course of this study. I am grateful to Dr. L. Claypool, Dr. B. Moser, and Dr. R. Tyrl for serving on my graduate committee.

Sincere appreciations go to my parents for their support and encouragement throughout my life; and to my wife, Seonghae, for her sacrifice and understanding.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.	1
Perspective.	1
A Discussion of Basic Concepts and Definitions in Cluster Analysis.	3
A Discussion of Distance Functions used in Cluster Analysis	9
Scope of this Study.	13
II. A LITERATURE REVIEW ON DETERMINING THE NUMBER OF CLUSTERS.	16
General Reflection	16
Publications Related to Hypothesis Testing Methods.	18
Publications Related to Optimization Methods.	20
Publications Related to Mode or Density Estimation Methods	23
Publications Related to Agglomerative Clustering Methods	25
III. AGGLOMERATIVE CLUSTERING ALGORITHM.	28
The (β, π) Family of Agglomerative Clustering Algorithms.	28
Classification of the (β, π) Family of Agglomerative Clustering Algorithms.	33
IV. USE OF A COMPARATIVE STATISTIC TO PREDICT THE NUMBER OF CLUSTER	39
A Comparative Statistic.	39
Other Measures of Similarity	48
Rationale for the Use of C_k to Predict the number of clusters	52
V. DESIGN OF A COMPARATIVE STUDY AND RESULTS FROM THE MULTIVARIATE NORMAL SAMPLES.	57
Parameter Choice	57
A Discussion on the Design of the Comparative Study.	61

Chapter	page
Discussion of the Results from Multivariate Normal Samples.	70
VI. EXTENSION TO MULTIVARIATE LOGNORMAL SAMPLES . . .	81
Fundamental Concepts	81
Discussion of the Results from Multivariate Lognormal Samples	88
VII. GENERAL CONCLUSIONS AND POSSIBLE EXTENSIONS . . .	93
BIBLIOGRAPHY	106
APPENDIX	111

LIST OF TABLES

Table	Page
1. Parameter values, $(\alpha_i, \alpha_j, \beta, \pi)$, for Several Agglomerative Clustering Algorithms.	30
2. Percent Retrieval of True Population for all Algorithms with MVN	112
3. Retrieval Information of Nine Algorithms vs. Popn. with $\delta = 4.0$, $\rho = 0.0$, and 20-20-20 split for MVN	113
4. Retrieval Information of Nine Algorithms vs. Popn. with $\delta = 4.0$, $\rho = 0.0$, and 30-20-10 split for MVN	115
5. Retrieval Information of Nine Algorithms vs. Popn. with $\delta = 6.0$, $\rho = 0.0$, and 20-20-20 split for MVN	117
6. Retrieval Information of Nine Algorithms vs. Popn. with $\delta = 6.0$, $\rho = 0.0$, and 30-20-10 split for MVN	119
7. The % _s on Local Maximum for all possible pairs of the Nine Algorithms when $\delta = 4.0$ with MVN. . .	121
8. The % _s on Local Maximum for all possible pairs of the Nine Algorithms when $\delta = 6.0$ with MVN. . .	122
9. Agreement of Five Paired Clustering Algorithms with $\delta = 4.0$, $\rho = 0.0$, and 20-20-20 split for MVN	123
10. Agreement of Five Paired Clustering Algorithms with $\delta = 4.0$, $\rho = 0.0$, and 30-20-10 split for MVN	125
11. Agreement of Five Paired Clustering Algorithms with $\delta = 4.0$, $\Theta = 15^\circ$, and 20-20-20 split for MVLN.	127

Table	Page
12. Agreement of Five Paired Clustering Algorithms with $\delta = 4.0$, $\Theta = 15$, and 30-20-10 split for MVLN.	129
13. The $\%_s$ on Local Maximum for all possible pairs of the Nine Algorithms when $\delta = 4.0$ for MVLN. . .	131
14. The $\%_s$ on Local Maximum for all possible pairs of the Nine Algorithms when $\delta = 4.0$ for MVLN. . .	133
15. Percent Retrieval of True Population for all Algorithms with MVLN.	135
16. The $\%_s$ on Local Maximum for Four Pairs of ($-.5, .75$) with other algorithms for MVN and MVLN.	137

LIST OF FIGURES

Figure	Page
1. A Classification of the (β, π) Family of Agglomerative Clustering Algorithms.	38
2. An Example of the Structural Framework Developed for BVN	65
3. An Example of the Structural Framework Developed for BVLN	83
4. Retrieval Result of the Nine Algorithms for MVN with $\delta = 4.0$, $\rho = .0$, and 20-20-20 split.	114
5. Retrieval Result of the Nine Algorithms for MVN with $\delta = 4.0$, $\rho = .0$, and 30-20-10 split.	116
6. Retrieval Result of the Nine Algorithms for MVN with $\delta = 6.0$, $\rho = .0$, and 20-20-20 split.	118
7. Retrieval Result of the Nine Algorithms for MVN with $\delta = 6.0$, $\rho = .0$, and 30-20-10 split.	120
8. Retrieval Result of Five Paired Clustering Algorithms with $\delta = 4.0$, $\rho = .0$, and 20-20-20 split for MVN.	124
9. Retrieval Result of Five Paired Clustering Algorithms with $\delta = 4.0$, $\rho = .0$, and 30-20-10 split for MVN.	126
10. Retrieval Result of Five Paired Clustering Algorithms with $\delta = 4.0$, $\theta = 15.0$, and 20-20-20 split for MVLN	128
11. Retrieval Result of Five Paired Clustering Algorithms with $\delta = 4.0$, $\theta = 15.0$, and 30-20-10 split for MVLN	130

CHAPTER I

INTRODUCTION

Perspective

In partitioning N individuals to be clustered into k appropriate groups for a set of p -dimensional multivariate data, one may wish to find the best procedure to predict the number of distinct groups, K . If the number of groups is known a priori, discriminant analysis provides a solution to the problem of how well N individuals are classified into their own groups. Principal component analysis is a method of projecting points in multi-dimensional space into a space of fewer dimensions so that the maximum amount of information is retained.

Cluster analysis differs from these analyses and is a more primitive technique in which no assumptions are made concerning the number of groups or the group structure. If a very large body of data can be reduced to a relatively compact description, it may become the basis for further statistical research. Therefore, there is a need to organize or reorganize the data in search of a natural organizational structure.

Cluster analysis has developed in many diverse fields

including biology, psychiatry, ecology, psychology, sociology, engineering, and econometrics. In addition, some of the relevant research in cluster analysis is being published in the computer science and statistical journals, and a unifying framework for the development of the theoretical aspects of cluster analysis might be found among the statistical methods. Anderberg (1973) gives a more complete and organized listing with discussion of these points.

Because statistics is a body of methods purporting to make sense out of data, cluster analysis belongs among the descriptive statistical methods. As a descriptive method, cluster analysis possesses two noteworthy characteristics. First, it is an exploratory technique to be used in the initial stages of research which, hopefully, will precipitate hypotheses for further research. Second, it has as its goal, simplification through organization by revealing structure and relations in the data.

However, the number of problems associated with cluster analysis is bewildering because each of the several clustering methods can produce quite different groupings for the same data. The most common problem facing an investigator with a set of objects he would like to examine by clustering procedures is the choice of which procedure to use from the several clustering procedures.

DuBien (1976) gives a comparison of various agglomerative hierarchical clustering algorithms using

Rand's (1969, 1971) C statistic, and DuBien and Warde (1982) present the distribution of the C statistic. Also, DuBien and Warde (1987) offer an empirical investigation of the effect of correlated variables on the "retrieval" ability of a particular class of agglomerative clustering methods. In this paper, we are concerned with the problem of predicting the number of clusters in a given set of data when using only agglomerative hierarchical clustering methods. We will further examine the use of Rand's C as a comparative statistic.

A Discussion on Basic Concepts and Definitions in Cluster Analysis

The definitional problems associated with cluster analysis can be partially resolved by a mathematical approach. Using DuBien's (1976) notation to formalize the presentation, a general, set-theoretic framework will be established for cluster analysis.

In any field of application, the primitive concepts of cluster analysis are based on the elements to be clustered. They are referred to as data points and a cluster is an operationally determined collection of data points. Each data point shall be represented by a $p \times 1$ vector, X_i , where

$$X_i = [x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}]' .$$

The components, x_{ij} of X_i will be termed variables. The set of all elements to be clustered shall be called the object

space and symbolized by X . Letting N be the number of data points, then the object space X might be

$$X = [X_1, X_2, \dots, X_N].$$

Obviously, the object space is embedded in Euclidean p -space. Thus, if E_p represents Euclidean p -space, then $X \subseteq E_p$. Letting $X_{N,p}$ represent the data matrix, where N is the number of data points and p is the number of variables satisfying $N \geq p$, then

$$X_{N,p} = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{N1} & x_{N2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x_{Np} \end{bmatrix}$$

where x_{ij} represents the value of the measurement of the j -th variable on the i -th objects.

Having laid a set-theoretic foundation for discussing cluster analysis concepts, mathematical definitions for a cluster and a clustering can be given.

Definition 1.1. A cluster, Y_k , is any nonempty subset of the object space. Symbolically, $Y_k \subseteq X$ means that if $X_i \in Y_k$, then $X_i \in X$.

Thus, a cluster is simply a collection of data points. The number of data points contained in a cluster shall be termed the size of the cluster.

Definition 1.2. A clustering, Y , is any partition of the object space. Symbolically, $Y = [Y_1, Y_2, \dots, Y_K]$ is a

partition of X , if the following three conditions hold:

- (i) For every $Y_k \in Y$, $Y_k \neq \emptyset$.
- (ii) If $Y_k \in Y$, $Y_m \in Y$, and $Y_k \neq Y_m$, $k \neq m$,
then $Y_k \cap Y_m = \emptyset$.
- (iii) $\bigcup_{k=1}^K Y_k = X$.

Hence, a clustering is simply a special kind of collection of clusters.

A clustering of N data points can consist of $k = 1, 2, \dots, N$ clusters. The number of clusters contained in a clustering shall be termed the size of the clustering. If clustering Y contains K clusters, then Y^K denotes a clustering of size K , where $K = 1, 2, \dots, N$. The set of all possible clusterings of size K for an object space containing N data points will specify a population of clusterings as given in definition 1.3.

Definition 1.3. Let N be the number of data points. Two important populations of clusterings are defined as follows:

- (i) An $[N, K]$ -population of clusterings or an $[N, K]$ -population is defined to be the set of all possible clusterings of size K for X ;
- (ii) An $[N]$ -population of clusterings or an $[N]$ -population is defined to be the set of all possible clusterings of X .

For convenience, $Y^{[N, K]}$ shall be used to designate a clustering from the $[N, K]$ -population. Thus, an $[N]$ -population of clusterings may be obtained by merging the $[N,$

K]-populations for all $K = 1, 2, \dots, N$.

In general terms, a clustering method consists of a criterion and a technique in which the criterion assigns a numerical value to each clustering and the technique selects a subset of the set of all possible clusterings over which the criterion is optimized (providing only a local optimum). A problem is to classify the many clustering methods into a small number of different types. Noteworthy attempts at classifying and reviewing clustering methods appear in Sneath and Sokal (1973), Cormack (1971), Anderberg (1973), Everitt (1974), and DuBien (1976). However, no standard terminology has emerged for designating an entire family of similar clustering methods. Apparently, "hierarchical clustering scheme (HCS)" by Johnson (1967), "agglomerative hierarchical" given by Anderberg (1973) and Everitt (1974), "sequential, agglomerative, hierarchical" given by Norton (1975), and "sequential, agglomerative, hierarchic, nonoverlapping (SAHN)" given by Sneath and Sokal (1973) are all descriptors for the same class of clustering methods. This previously described class of clustering methods will be of primary importance in this paper, and these clustering methods shall be referred to simply as agglomerative clustering methods as used by DuBien (1976).

Agglomerative clustering methods are some of the oldest and most frequently used clustering methods. The method may be characterized as proceeding sequentially by joining pairs of clusters. It starts with the partition which consists of

each data point as a single cluster and proceeds until there is one cluster containing all data points. The investigator must decide at which stage in the analysis he wishes to stop because all agglomerative clustering procedures ultimately reduce the data to a single cluster. An important concept in the definition of an agglomerative clustering method is an hierarchy.

Assuming that there are N data points, formal definitions for an hierarchy and for agglomerative clustering methods are given as definitions 1.4 and 1.5, respectively.

Definition 1.4. A hierarchy, H , on the object space is an ordinal sequence of nested clusterings. Symbolically,

$$H : Y^N, Y^{N-1}, \dots, Y^2, Y^1,$$

$$\text{where } Y^N \subset Y^{N-1} \subset \dots \subset Y^2 \subset Y^1.$$

One useful visualization of a hierarchy is a tree-like diagram which is often called a dendrogram in cluster analysis applications. In summary, a hierarchy on the object space is a nested collection of clusterings (each consisting of a set of clusters) which may be aptly depicted by a dendrogram.

Definition 1.5. An agglomerative clustering method is any clustering method, M , which produces a hierarchy on the object space subject to the following constraints:

- (i) Y^N is the initial clustering;
- (ii) Clustering Y^{K-1} , $K \leq N$, is obtained from clustering Y^K by joining the two closest clusters

in clustering Y^K ; i.e., if $Y_i, Y_j \in Y^K$

and they are deemed closest, then $Y_i \cup Y_j \in Y^{K-1}$.

Thus, the application of an agglomerative clustering method to N data points results in a special kind of hierarchy, thereby imposing an hierarchical structure on the object space.

Based on the definitions given above, the resolution of a clustering problem by the application of an agglomerative clustering method to a data set can be described by the triple (X, H, M) . When a clustering method consists of a criterion and a technique, the agglomerative clustering method, M , may be more specifically viewed as consisting of a measure of similarity or dissimilarity (usually a measure of distance) and an algorithm (usually a form of linkage). The measure of similarity or dissimilarity explicates "close", initially; and the algorithm reevaluates the "closeness" of clusters after each join. As a further limitation, the agglomerative clustering methods of particular interest in this paper may be denoted by the pair (measure of distance, clustering algorithm).

At this point, the application of an agglomerative clustering method to a set of data requires that a measure of distance, d , be imposed on the object space, X . Hence, the properties and some examples of distance measures will be established, and then agglomerative clustering algorithms will be formalized in chapter III.

A Discussion of Distance Functions used in Cluster Analysis

In very general terms, a measure of distance, d , on some arbitrary set, S , is a real-valued function on $S * S$. In particular, some of the relevant properties which a measure of distance may possess will be given with respect to the object space, X . However, these properties may apply to an arbitrarily defined measure of distance on any set. Letting d_{ij} denote the distance between data points X_i and X_j , the properties for a measure of distance are described in definitions 1.6, 1.7, and 1.8.

Definition 1.6. A semi-metric on the object space, X , is a function,

$$d: X * X \longrightarrow R,$$

such that the following two properties hold for every pair of data points, X_i and X_j in X :

(i) d is a strictly positive function, i.e.,

$$d_{ij} \geq 0,$$

$$\text{and } d_{ij} = 0 \text{ iff } X_i = X_j;$$

(ii) d is a symmetric function, i.e.,

$$d_{ij} = d_{ji}.$$

Definition 1.7. A metric on the object space, X , is a semi-metric d such that the following third property also holds for every X_i , X_j , and X_k in X :

(iii) d satisfies the triangle inequality, i.e.,

$$d_{ik} \leq d_{ij} + d_{jk}.$$

Definition 1.8. An ultrametric (Johnson, 1967) on the object space, X , is a metric d such that the following fourth property also holds for every X_i , X_j , and X_k in X :

(iv) d satisfies the ultrametric inequality, i.e.,

$$d_{ik} \leq \max \{d_{ij}, d_{jk}\}.$$

The ultrametric inequality is a stronger property than the triangle inequality. Thus, if the ultrametric inequality holds for a measure of distance on X , then the triangle inequality necessarily holds for that measure of distance on X . It is also worth noting that an ultrametric measure of distance is invariant to all monotonic transformations of d . A metric measure of distance, however, is not, in general, invariant to monotonic transformations of the measure of distance because the triangle inequality is not preserved under all monotonic transformations of d . It should be noted that for the derivation presented in this study, only a semi-metric measure of distance is required as a basis for the initial distance matrix.

A well-known family of distance measures for which the metric properties hold is the family of Minkowski metrics. The m -th member of the family of Minkowski metrics will be designated by ℓ_m . Recalling that X_i is a p -component vector, if x_{iv} denotes the v -th component of data point X_i and x_{jv} denotes the v -th component of data point X_j , then the m -th Minkowski metric between data points X_i and X_j is computed by the following formula:

$$\ell_m(X_i, X_j) = \left[\sum_{v=1}^p |x_{iv} - x_{jv}|^m \right]^{1/m},$$

where $m \geq 1$.

Euclidean distance is a member of the family of Minkowski metrics, namely, ℓ_2 . However, squared Euclidean distance (in common use with some agglomerative clustering algorithms) is only a semi-metric measure of distance, since the triangle inequality is not preserved under the operation of squaring distances.

From this brief background on measures of distance, the general formulation for agglomerative clustering algorithms given by Lance and Williams (1966) can be presented in a notation consistent with the present development. First, however, with respect to an agglomerative clustering method, some subtle distinctions concerning the set on which d is a measure of distance are needed. In the application of an agglomerative clustering method to a set of data, the distance between each pair of data points, X_i and X_j , is initially computed using some measure of distance, d , which is at least a semi-metric. Since d is at least a semi-metric, the resultant set of distances may be denoted by

$$D = \{d_{ij} | i < j, i = 1, 2, \dots, N-1, j = 2, 3, \dots, N\}.$$

A convenient device for displaying D is the distance matrix $D_{N,N}$, where only the $N(N-1)/2$ upper triangular elements of $D_{N,N}$ are necessary. Therefore, d is a measure of distance on X . Also, the set of single-point clusters, Y^N , corresponds to X . Consequently, d is also a measure of

distance on Y^N , where an element of Y^N is a cluster, Y_i , corresponding to data point X_i . Hence, the process of clustering a set of data by means of an agglomerative clustering method is initiated by viewing the measure of distance on X as a measure of distance on Y^N ; and, thereby, D becomes the set of all distances between pairs of clusters in Y^N . As it is known, the results of clustering procedures depend on a metric among the pairs of objects. For the purpose of this study, the squared Euclidean distance, which is commonly used in agglomerative clustering methods, is considered as a measure of distance.

The role of the agglomerative clustering algorithm is to sequentially impose a measure of distance on each clustering, Y^K , $K = 1, 2, \dots, N-1$, in the hierarchy such that the measure of distance imposed on Y^K is functionally related to the measure of distance imposed on Y^{K+1} . In this sense, d is not the same measure of distance on Y^K and on Y^{K+1} (i.e., on two clusterings of different sizes). In fact, even when d is initially a metric, for some clustering in the hierarchy, d may not even be semi-metric, and this anomalous situation is well illustrated by DuBien (1976).

To clarify the notation, since Y^K , $K = 1, 2, \dots, N$, is a set of clusters, a measure of distance may be imposed on Y^K , and d_{ij} shall now be used to denote the distance between cluster Y_i and cluster Y_j , where $Y_i, Y_j \in Y^K$, $K = 1, 2, \dots, N$. This is not inconsistent since in the case of Y^N , X_i and Y_i correspond. Thus, the distance between data points is a

special case of the distance between clusters, and this distance between data points will be used to initiate a recursive algorithm for the recomputation of distance between clusters after each joining of two clusters. As a further simplification of the notation, if two clusters, $Y_i, Y_j \in Y^K$, join at distance d_{ij} to form a new cluster $(Y_i \cup Y_j)$ where $(Y_i \cup Y_j) \in Y^{K-1}$, then $Y_{(ij)}$ will denote the new cluster; i.e.,

$$Y_{(ij)} = Y_i \cup Y_j ,$$

and d_{ij} shall be termed the joining distance for clustering Y^{K-1} . It should be noted that the joining distance, d_{ij} , is always the smallest distance remaining in the set of all distances between clusters in clustering Y^K .

Further delineation of the particular agglomerative clustering methods of interest will be given in Chapter III.

Scope of This Study

Although cluster analysis has been widely used to create empirical classifications, the number of problems associated with cluster analysis, given a "real" set of data, is still enormous. These problems can be summarized as follows:

1. How should the variables be scaled ?
2. Which distance measure should be used ?
3. What clustering method should be used ?
4. How should the number of clusters, K , be specified or determined ?

Because of these problems, there are a large number of clustering methods which produce quite different results depending on which variable and which distance measures are used. A brief discussion of the first three problems can be found in DuBien (1976).

The problem of determining the correct number of clusters, K , is the main objective of this study. Attention is focused on the use of Rand's (1971) C statistic in conjunction with some agglomerative hierarchical clustering algorithms in predicting the number of clusters within the given set of data.

Because operational clustering methods search a subset of $Y^{[N]}$ or $Y^{[N,K]}$, which is usually defined as all of a certain type of rearrangement of a specific initial clustering, it is of interest to examine the behavior of the similarity measure in some of these situations. Ideally, this study will illustrate the remark made by Gordon (1981):

If the results of several different classification procedures agree closely, then one has more confidence in the reality of any group structure which is indicated.

A review of cluster analysis literature related to the problem of determining the number of clusters present in the data is given in Chapter II.

Chapter III contains the formulations of the nine agglomerative hierarchical clustering algorithms chosen for this study.

Because the comparative study presented in this paper

is limited to agglomerative hierarchical clustering procedures of the form (measure of distance, clustering algorithm), Chapter IV presents formulations of the comparative statistic used in Chapter V and summarizes a rationale for the use of the comparative statistic to predict the number of clusters present in the given object space, where agglomerative hierarchical clustering procedures are applied under the specific assumptions. Also, the mean and variance of the comparative statistic will be provided.

In Chapter V, the design of the comparative study for multivariate normal samples is presented while an empirical investigation of the effects of correlated variables and distance between population mean vectors on the "retrieval" ability of agglomerative clustering methods and the "agreement" between the pairs of them are discussed by examining the behavior of the comparative statistic. In addition, this investigation may provide useful information about the properties of different clustering methods.

In Chapter VI, the use of C_k is investigated for multivariate log-normal samples and a discussion on the results will be presented.

CHAPTER II

A LITERATURE REVIEW ON DETERMINING THE NUMBER OF CLUSTERS

General Reflection

An extensive number of clustering techniques have been developed for determining the number of clusters in a given set of data. These techniques themselves can be classified roughly into four general groups: hypothesis testing methods, optimization methods, mode or density estimation methods, and agglomerative clustering methods. First, however, a general review of the literature which is deemed to give significant contributions on the development of techniques in cluster analysis is given.

Cormack (1971), Anderberg (1973), Sneath and Sokal (1973), Everitt (1974), Duran and Odell (1974), Hartigan (1975), and Peck (1983) provide a comprehensive general review of clustering methods, including a classification of methods into broad general types and discussions of measures of similarity (or dissimilarity), clustering algorithms, clustering criteria, and clustering techniques. Despite the numerous attacks on the problem for determining the most probable number of clusters for a given set of data, it must

be said that no completely satisfactory solution is available in the realm of cluster analysis.

Mrachek (1972) and Norton (1975) necessarily make valuable contributions to the theoretical development of cluster analysis, and both of them are at least partially concerned with the problem of testing for the presence of structure in data.

DuBien (1976) provides a comparative study of agglomerative clustering methods which will guide the matching of clustering method with type of cluster generated. DuBien and Warde (1979) present an algebraic analysis of agglomerative clustering methods, which results in a graphic portrayal of these methods and a classification scheme for these methods based on the degree of distortion perpetrated on the object space by the methods in each group. Then DuBien and Warde (1982) derive some distributional results concerning a comparative statistic, Rand's (1971) C statistic. Further, DuBien and Warde (1987) present an empirical investigation of the effect of correlated variables on the "retrieval" ability of a particular class of agglomerative clustering methods.

However, critical comparisons of the recovery characteristics of the criteria have not been fully conducted until Milligan (1981), and Milligan and Cooper (1985). Milligan (1981, p187) mentioned:

It seems that the trend in the clustering literature has been for authors to continue to introduce new statistics while providing little

comparative information.

Hence, Milligan (1981) studied how to determine whether any criterion could indicate whether a given partition of the data recovered a significant portion of the true cluster structure or whether any structure exists at all. Further, Milligan and Cooper (1985) evaluated 30 stopping rules already in the clustering literature for determining the number of clusters on artificial data sets which contain either 2, 3, 4, or 5 distinct nonoverlapping clusters.

In this study the principal interest is to investigate the use of a comparative statistic for predicting the correct number of clusters, k , when applying specified agglomerative clustering algorithms to a given set of data. An extensive review of the literature is focused on the problem of determining the number of clusters within a given set of data.

Publications related to Hypothesis

Testing Methods

The journal articles by Wolfe (1970), Sneath (1977), Binder (1978), Lee (1979), Everitt (1981), and McLachlan (1987) are grouped in this category.

Wolfe (1970) presents a likelihood ratio criterion to test the hypothesis of k clusters against $k-1$ clusters. The process is modeled after the traditional Wilks' likelihood ratio (1938) criterion and is based on the assumption of multivariate normality. However, Binder (1978) has shown

that this test statistic is not asymptotically distributed as chi-square. Further, Everitt (1981) performed a Monte Carlo analysis of Wolfe's procedure in a mixture of normal distributions and found that Wolfe's formula for the degrees of freedom for the test appeared to be valid only for the cases where the sample size is about ten times larger than the number of dimensions. Also, McLachlan (1987) examined bootstrapping (Efron, 1982) the likelihood ratio statistic for testing the number of clusters $k = 1$ under the null hypothesis H_0 versus $k = 2$ under the alternative hypothesis H_a , and found that Wolfe's approximation may not be applicable in the unequal variance case.

Sneath (1977) describes a method for testing the distinctness of two clusters in Euclidean space based on a measure of overlap. It is assumed that the clusters are roughly hyperspherical multivariate normal. Then the suggested statistic measures the separation between clusters and not the overlap, and has a noncentral t-distribution under the null hypothesis. The test statistic, t_w , is compared to a critical score obtained from a noncentral t-distribution. The hypothesis of one cluster is rejected if t_w exceeds the critical score.

The ratio of the determinant of the total sum of squares and cross product matrix to the determinant of the pooled within group sum of squares matrix maximized over all possible partitions of the objects into k clusters, $\frac{|T|}{|W|}$,

was proposed by Friedman and Rubin (1967). Engleman and Hartigan (1969) derived a table of percentage points of a test for the presence of clusters in data, but their test for the presence of structure is limited to the univariate case. Norton (1975) comments that this test procedure is most appropriate for use with divisive algorithms which divide the observations into two groups in such a way as to maximize the ratio of between-group sum of squares to within-group sum of squares. In practice, use of this is limited since tables of percentage points have been generated for only a very limited number of sample sizes. Also Norton (1975) suggests the generalization of the above procedure to 3, 4, ..., $N-1$ cluster alternatives. However, these tests cannot be used yet because there are no percentage points available.

Lee (1979) generalized Engleman and Hartigan's work to the multivariate case. Criteria are considered for testing the hypothesis that the observations are a random sample from one multinormal population versus the alternative that the observations arise from two multinormal populations with different means and common variance-covariance matrix. For higher dimensions an approximation to the sampling distribution is also provided.

Publications related to

Optimization Methods

It is advantageous to use available prior information

to develop a criterion that when optimized produces a clustering with desirable properties.

A familiar objective function applicable in cluster analysis is the within-group sum of squares matrix (W) and the between group sum of squares matrix (B) with respect to the total sum of squares matrix (T). These are,

$$B = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})',$$

$$W = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)',$$

and

$$T = B + W = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})',$$

where k is the number of groups and $\sum_{i=1}^k n_i = N$. It seems natural to regard the optimal grouping of N objects into k clusters as that for which W is minimized or B is maximized. This criterion reflects a desire to find some minimum variance spherical clusters, and generally optimization methods in cluster analysis consider the functional relation among these sum of squares and crossproduct matrices. The $\text{trace}(W)$ criterion is one of the popular indices (Edwards and Cavalli-Sforza, 1965; Friedman and Rubin, 1967) used to determine the number of clusters by using the maximum difference scores since the criterion increases monotonically with solutions containing fewer clusters. However, a problem with the $\min\{\text{trace}(W)\}$ criterion is that the clusters produced are constrained to being hyper-spherical; in cases where the real clusters in

the data are of some other shape this may produce misleading solutions. Also, this method is transformation dependent as noted by Everitt (1979).

Calinski and Harabasz (1974) suggest the variance ratio criterion (VRC), $\{\text{trace}(B)/(k-1)\}/\{\text{trace}(W)/(N-k)\}$, where N and k are the total number of items and the number of clusters in the solution, respectively. Then the best number of groups in the data is indicated by an absolute or local maximum of the VRC. Similarly, the use of $\min(|W|)$, $\max\{\text{trace}(W^{-1}B)\}$, and $\log\{\max(|T|/|W|)\}$ for determining the number of clusters in the data was suggested by Friedman and Rubin (1967), and several further studies on similar methods have been conducted by Marriott (1971), Scott and Symons (1971), and Symons (1981).

Ratkowsky and Lance (1978) introduce a criterion, $\bar{C} / k^{.5}$, where the value for \bar{C} is equal to the average of the ratios of (SSB/SST) obtained from each dimension in the data, where SSB is the between group sum of squares and SST is the total sum of squares. Then the optimal number of groups is determined where this criterion presents its maximum value. However, they say that this criterion tends always to produce a small number of groups. Hill (1980) modifies this criterion, but recognizes that the modification has serious weaknesses in that his new criterion, \bar{C} , can continue to increase until it has split the cluster into its individual units. Ratkowsky (1984) proposes a new approach that uses the average similarity of

an individual with the members of its group. The similarity coefficient must give values that lie between 0.0 and 1.0. Then the optimum number of groups, k , is found if the criterion is maximum. However, this criterion sometimes produces a value less than the value obtained when N individuals are split into N singleton groups.

Krzanowski and Lai (1988) suggest a new criterion for determining the number of groups in a data set by using sum-of-squares clustering after observing the behavior of both Marriott's (1971) and their new criterions based on within-group sum-of-squares objective function trace (W). Then they give cautionary comment that their criterion should not be expected to yield optimum results if the use of the sum-of-squares objective function for a particular set of data is inappropriate. Moreover multiple local maxima of a criterion can occur frequently.

Publications related to Mode or Density Estimation Methods

From the reasoning that modes occur in the density, f , where points congregate, the number of modes of f , and therefore the number of clusters in the sample, can be estimated using this approach.

Wishart (1969) developed a method, hierarchical mode analysis, for moderate size data sets and outlined its proposed extension for large data sets. The procedure is to first detect whether the data is multi-modal. For the

univariate case one would construct a histogram and temporarily remove the low frequency (saddle) regions. Then a cluster can be associated with each modal region and the data falling in the saddle region can be assigned to their nearest mode. From each point on the projection, one tests whether n or more other objects lie within a distance threshold, R , which is determined by the user. Then the objects in a sphere are considered "dense", and "dense" objects are clustered together. The method of cluster analysis developed by Ling (1973) is similar to Wishart's mode analysis algorithm, but his method operates on the ranks of the distances instead of the distances themselves.

Silverman (1981) uses a kernel estimate of the density function for window width h based on univariate observations X_1, \dots, X_N , where the window width h controls the visual smoothness of the resulting density. Choices of the parameter h , and a method of choosing the window width when estimating a density is discussed by Silverman (1978). When h increases from 1 to N , the density estimate becomes smoother or less bumpy. Therefore, if the data are strongly bimodal, a large value of h will be needed to obtain a unimodal estimate.

Similarly, Wong and Schaack (1982) develop a procedure in univariate data by using the k -th nearest neighbor clustering algorithm (Wong and Lane, 1981) to provide a plot of the "estimated number of modes" against h by assuming that the clusters correspond to modes of the population

density function. The plot is a nonincreasing step function in h . It is expected that when the number of modes reaches the true number in the sample the plot will be stable over a large range of values of h . Otherwise, the smallest value of h yields k modes in the plot.

Further, Wong (1985) develops nonparametric procedures that are useful for testing the multimodality of f , where the clustered data are sampled from some general univariate distribution F with density function f . The test statistics based on Wong and Lane's (1981) k -th nearest neighbor clustering algorithm are proposed for testing multimodality by using a modified bootstrap method (Efron, 1982) to determine the number of clusters. Large values of criterion will reject the null hypothesis that the underlying density f has at most k modes, and suggest that f has more than k modes.

Publications related to Agglomerative Clustering Methods

Agglomerative clustering methods are perhaps the most popular of all the multitude of clustering methods, and the literature on them is enormous. Regardless of the method selected, the results may be displayed by a contour map or by a tree (dendrogram), a two dimensional graphical representation of the fusions or divisions of clusters at each successive level of the procedure. If the number of clusters is known, then the scientist uses the appropriate

stage of the algorithm to indicate which observations have been grouped together. If the number of clusters is unknown, then the number of clusters for the best representation of the data is determined subjectively. Some discussion and various applications of agglomerative clustering methods with stopping rules may be found in Lance and Williams (1967), Johnson (1967), and Baker and Hubert (1975). In general, the ideas presented in these articles are based on the measure of similarity between the clusters.

Since it is possible to measure the similarity between the clusters, the dendrogram may be drawn to scale to reflect the similarity between clusters that are grouped at a given level. A partition of a sample of N objects into k clusters is found by cutting the dendrogram at the $(N-k+1)$ -th level. Intuitively, if the similarity between objects clustered together at a given level is high, then one could conclude that the clustering is natural.

Lance and Williams (1967) have developed a formula which will compute the distance, $d_{(ij)k}$, between group k and group $(i \cup j)$ for many of the common linkage methods. This will be discussed in the next chapter, and the formula is given in equation (3.1). The number of clusters is taken to be k when the distance for $k-1$ is much smaller than it is for k clusters. However, their results have not been clearly interpreted by the lack of objective criteria.

Baker and Hubert (1975) discussed the problem of estimating a true partition at a certain level of hierarchy

by using the Goodman and Kruskal (1954) gamma coefficient. The maximum value across the hierarchy levels was used to indicate the correct hierarchy level.

Edelbrock (1979) suggested using the statistic kappa (Cohen, 1960) to assess the accuracy of clustering solution. The values of kappa range from -1.0 to 1.0, with larger values indicating larger agreement between the obtained clusters and the populations. In his study, the largest decrease of the value was generally observed from level $k+1$ to level k , where all elements had to be assessed. This result was also confirmed by Scheibler and Schneide (1985).

Rand (1971), and Fowlkes and Mallows (1983) developed two different measures, C and B_k , respectively, based on the proportion of object pairs from the same populations that are grouped together in the resultant clusterings for the agglomerative clustering algorithms. These statistics range from 0.0 and 1.0. Since these measures depend on similarity between two different clusterings generated by two different clustering algorithms performed on the same set of data, the values of measures would be 1.0 if the two clusterings correspond completely. Hence it could be thought that the number of clusters is k if the value of the similarity measure is close to 1.0.

Since the objective of this study is to investigate the use of the Rand's measure, C , on determining the number of clusters for a given set of data, more discussions about these two measures will be given later in Chapter IV.

CHAPTER III

AGGLOMERATIVE CLUSTERING ALGORITHM

The (β, π) Family of Agglomerative Clustering Algorithms

From Chapter I, the resolution of a clustering problem by the application of an agglomerative clustering method to a data set can be described by the triple (X, H, M) . The object space, X , and the clustering method, M , are elements of the parameter space which require specification, and the hierarchy, H , is the resultant sequence of clusterings for the specified pair (X, M) . X is essentially specified by N , the number of data points, and p , the dimension of the Euclidean space in which the object space is embedded. Thus the specification of the clustering methods, M , is required for the application of an agglomerative clustering method to a set of data points. Since the clustering method, M , is specified by the pair (measure of distance, clustering algorithm), all conclusions concerning the resultant hierarchy are dependent on these initial specifications.

The necessity of specifying both parameters (measure of distance, clustering algorithm) places a serious restriction on the generalizations which may be made from an empirical,

comparative investigation of agglomerative clustering methods. It is also possible that there is an interaction between the measure of distance and the clustering algorithm. At least, both members of the pair defining the agglomerative clustering method contribute to the process which produces the dendrogram, and varying either member of this pair may produce a different sequence of clusterings for a particular data set.

Using the notation in Chapter I, the general linear combinatorial strategy originally presented by Lance and Williams (1967) is given as Equation (3.1). For any clustering Y^k in the hierarchy, if the distances d_{ij} , d_{ik} , and d_{jk} between pairs of clusters Y_i , Y_j and Y_k are obtained from some source (e.g., recursively from clustering Y^{K+1} , $K \neq N$), then the distance between the new cluster $Y_{(ij)}$ and any other cluster $Y_k \in Y^K$ can be computed from the following formula:

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \pi |d_{ik} - d_{jk}|, \quad (3.1)$$

where d_{ij} denotes the distance between the clusters Y_i and Y_j with n_i and n_j elements, respectively, which have been combined to form a new cluster $Y_{(ij)}$, and α_i , α_j , β , and π are specified parameters defining the particular member of the family of agglomerative clustering algorithms.

Beginning with the initial distance matrix, D , obtained by imposing d on X , equation (3.1) is applied recursively to obtain each clustering in the hierarchy. Equation (3.1)

defines a four parameter family of agglomerative clustering algorithms, which contains an infinite number of distinct algorithms. Examples of some popular and widely used agglomerative clustering algorithms based on the choice of the parameters α_i , α_j , β , and π are given in table 1.

As shown in table 1, equation (3.1) characterizes a particular agglomerative clustering algorithm for the choices of the parameter quadruples $(\alpha_i, \alpha_j, \beta, \pi)$. Technically, the flexible strategies given in table 1 are specific examples of the flexible strategy, and the first set in the flexible strategy is the "best" flexible strategy

TABLE 1
PARAMETER VALUES, $(\alpha_i, \alpha_j, \beta, \pi)$, FOR SEVERAL
AGGLOMERATIVE CLUSTERING ALGORITHMS

Algorithm	Parameter		Values	
	α_i	α_j	β	π
Single Linkage	0.5	0.5	0.0	-0.5
Complete Linkage	0.5	0.5	0.0	0.5
Unweighted Average Linkage	0.5	0.5	0.0	0.0
Flexible Strategy	0.625	0.625	-0.25	0.0
	0.75	0.75	-0.5	0.0
Median	0.5	0.5	-0.25	0.0
Weighted Average Linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0.0	0.0
	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\alpha_i \alpha_j$	0.0
Centroid Linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\alpha_i \alpha_j$	0.0

according to Lance and Williams (1967). However, two quite different hierarchies may be derived from the same set of data, if two different agglomerative clustering algorithms are specified. Lance and Williams (1966) mentioned that the extent of clustering is not an inherent property of data, but rather a property of the user's desires about the shape of the clusters, which can be determined by varying the parameters. This implies that the application of an agglomerative clustering method to a given set of data might distort the result of the clustering procedure with respect to the properties of the sequence of distances, $d_{(ij)k}$.

DuBien (1976) has explored the properties of the sequence of distances, $d_{(ij)k}$, by placing a suitable set of constraints on the parameters given in equation (3.1) and deriving a two parameter family of agglomerative clustering algorithms from the four parameter family. It should be noted that the motivation for the two parameter family was the Lance and Williams' (1966) flexible strategy.

Letting

$$\begin{aligned}\alpha_i &= \alpha_j = \alpha, \\ \alpha_i + \alpha_i + \beta &= 1,\end{aligned}$$

some members of the four parameter family of agglomerative clustering algorithms can be represented by a two parameter sub-family, (β, π) , of agglomerative clustering algorithms. However, all algorithms in the four parameter set are not reduced to two parameter algorithms; e.g., weighted average and centroid linkages both have $\alpha_i \neq \alpha_j$, in general.

Without loss of generality, it will be assumed that

$$d_{ij} < d_{ik} < d_{jk} .$$

Then the two constraints used to define the two parameter family imply that

$$\alpha_i = \alpha_j = \frac{1 - \beta}{2} ,$$

and equation (3.1) becomes

$$d_{(ij)k} = \frac{1 - \beta}{2} d_{ik} + \frac{1 - \beta}{2} d_{jk} + \beta d_{ij} + \pi |d_{ik} - d_{jk}| .$$

Since $d_{ij} < d_{ik} < d_{jk}$, then

$$d_{(ij)k} = \frac{1 - \beta + 2\pi}{2} d_{jk} + \frac{1 - \beta - 2\pi}{2} d_{ik} + \beta d_{ij} . \quad (3.2)$$

Thus, equation (3.2) characterizes a sub-family of agglomerative clustering algorithms which shall be referred to as the (β, π) family, and each member of this sub-family shall be referred to as a (β, π) algorithm. Thus, a suitable set of constraints is necessary to make it possible to represent each member of the (β, π) family of agglomerative clustering algorithms as a point in the (β, π) Cartesian coordinate plane.

It is worth noting that single linkage (or nearest-neighbor), unweighted average linkage, complete linkage (or furthest-neighbor), and one of the family of flexible strategies given by Lance and Williams (1967) are members of the (β, π) family of agglomerative clustering algorithms, namely, $(0.0, -0.5)$, $(0.0, 0.0)$, $(0.0, 0.5)$, and $(-0.25, 0.0)$, respectively. The flexible strategy is actually a one parameter sub-family of agglomerative clustering algorithms

which may be derived from equation (3.1) by placing the following set of constraints on the four parameters (α_i , α_j , β , π):

$$\alpha_i + \alpha_j + \beta = 1;$$

$$\alpha_i = \alpha_j;$$

$$\beta < 1;$$

$$\pi = 0.$$

As a consequence, equation (3.1) becomes

$$d_{(ij)k} = \frac{1}{2}(1 - \beta)d_{jk} + \frac{1}{2}(1 - \beta)d_{ik} + \beta d_{ij}, \quad (3.3)$$

where $\beta < 1$.

Hence, equation (3.3) characterizes a sub-family of agglomerative clustering algorithms which is only dependent on the choice of the parameter β . Consequently, the flexible strategy could be referred to as the β family of agglomerative clustering algorithms, and each member of this sub-family could be referred to as a β algorithm. It is obvious that the β family is embedded in the (β, π) family of agglomerative clustering algorithms. A brief empirical study for the flexible strategy was presented by Lance and Williams (1967).

Classification of the (β, π) Family of Agglomerative Clustering Algorithms

At this point, it seems relevant to present the properties of the sequence of distances, $d_{(ij)k}$, as a means to exploring the amount of distortion which might result

from the application of an agglomerative clustering method to a set of data (DuBien and Warde, 1979). If

$$D_{(\beta, \pi)}^* = \{d_{(ij)k} \text{ at } (\beta, \pi) \mid d_{ij} < d_{ik} < d_{jk}\},$$

then the essential properties to consider for (β, π) algorithms are given by definitions 3.1, 3.2, 3.3, and 3.4.

Definition 3.1. A (β, π) algorithm is monotone increasing

iff $\forall d_{(ij)k} \in D_{(\beta, \pi)}^*, d_{(ij)k} > d_{ij}$.

Definition 3.2. A (β, π) algorithm is space-conserving

iff $\forall d_{(ij)k} \in D_{(\beta, \pi)}^*, d_{ik} < d_{(ij)k} < d_{jk}$.

Definition 3.3. A (β, π) algorithm is space-contracting

iff the g.l.b. (greatest lower bound) $(D_{(\beta, \pi)}^*) \leq d_{ik}$.

Definition 3.4. A (β, π) algorithm is space-dilating

iff the l.u.b. (least upper bound) $(D_{(\beta, \pi)}^*) \geq d_{jk}$.

Some further terminology related to the definitions is proposed to facilitate the classification of a (β, π) algorithm based on the range of $D_{(\beta, \pi)}^*$. If a (β, π) algorithm is not monotone increasing, then it shall be termed an extreme-space-contracting algorithm. If the range of $D_{(\beta, \pi)}^*$ is such that the associated (β, π) algorithm might be either space-contracting or space-dilating, then the (β, π) algorithm shall be termed a space-contract-dilating algorithm.

Although the terms space-conserving, space-contracting, and space-dilating were first given by Lance and Williams (1967), their characterizations of these concepts were only

intuitive in nature; and thus, their intuitive definitions for these concepts failed to yield a complete classification for the β family of agglomerative clustering algorithms based on the amount of distortion perpetrated on the object space by each β algorithm.

According to Lance and Williams (1967), a space-conserving algorithm preserves the spatial properties inherent in the original set of distances, and an algorithm which is not space-conserving is referred to as a space-distorting algorithm. They consider two different types of space-distorting algorithms, namely space-contracting and space-dilating algorithms. Intuitively, the application of a space-contracting algorithm to a set of distances implies that the new cluster moves closer to the old cluster upon formation. The application of a space-dilating algorithm to a set of distances implies that the new cluster moves further away from the old cluster upon formation. Thus, to make these intuitive concepts originated by Lance and Williams applicable to the problem of matching agglomerative clustering algorithms with the type of clusters generated, definitions 3.1, 3.2, 3.3, and 3.4 are tendered as mathematically rigorous interpretations for space-conserving, space-contracting, and space-dilating algorithms, respectively. Based on these definitions, the classification of the (β, π) family of agglomerative clustering algorithms and the investigation of the properties of $D_{(\beta, \pi)}^*$ over various regions of the (β, π)

plane are entirely presented by DuBien (1976), and DuBien and Warde (1979).

As a result, DuBien and Warde (1979) recommend the use of space-conserving and space-dilating (β, π) algorithms in conjunction with some measure of distance for clustering data set for which it is "semi-reasonable" to assume at least an interval scale of measurement for the variables comprising each data point. Space-dilating (β, π) algorithms should assist in picking up small distances between clusters of data points. Obviously, extreme-space-contracting (β, π) algorithms should not be used as clustering algorithms, and space-contract-dilating algorithms are too dependent on the relative magnitudes of d_{ij} , d_{ik} , and d_{jk} to be of general use as clustering algorithms. Space-contracting (β, π) algorithms should tend to minimize the distances between clusters of data points; and hence, some of these algorithms should be useful in indicating the existence of large distances between clusters of data points and might also be useful in identifying outliers in multivariate data.

On the basis of the rationale behind the choice of agglomerative clustering algorithms discussed by DuBien (1976) and DuBien and Warde (1979), only nine agglomerative clustering algorithms are chosen for the present study. The (β, π) values which define these nine agglomerative clustering algorithms are conveniently delineated in three groups of three algorithms as follows:

- (1) $\beta = 0.0$ with $\pi = -0.5, 0.0, 0.5$;
- (2) $\beta = -0.25$ with $\pi = -0.25, 0.0, 0.5$;
- (3) $\beta = -0.5$ with $\pi = 0.0, 0.25, 0.75$.

Figure 1 shows a classification of the (β, π) family of nine agglomerative clustering algorithms. It should be noted that single linkage, $(0.0, -.5)$, is the only space-contracting algorithm in this study; average linkage, $(0.0, 0.0)$, is the only space-conserving algorithm; and all of the other algorithms in the study are space-dilating algorithms as shown in DuBien and Warde (1979).

Since the primary objective of this present study is to investigate the use of a comparative statistic, Rand's (1969, 1971) C statistic, for predicting the correct number of clusters by applying agglomerative clustering procedures in a given set of data, a discussion related to this comparative statistic employed will be presented in the following chapter.

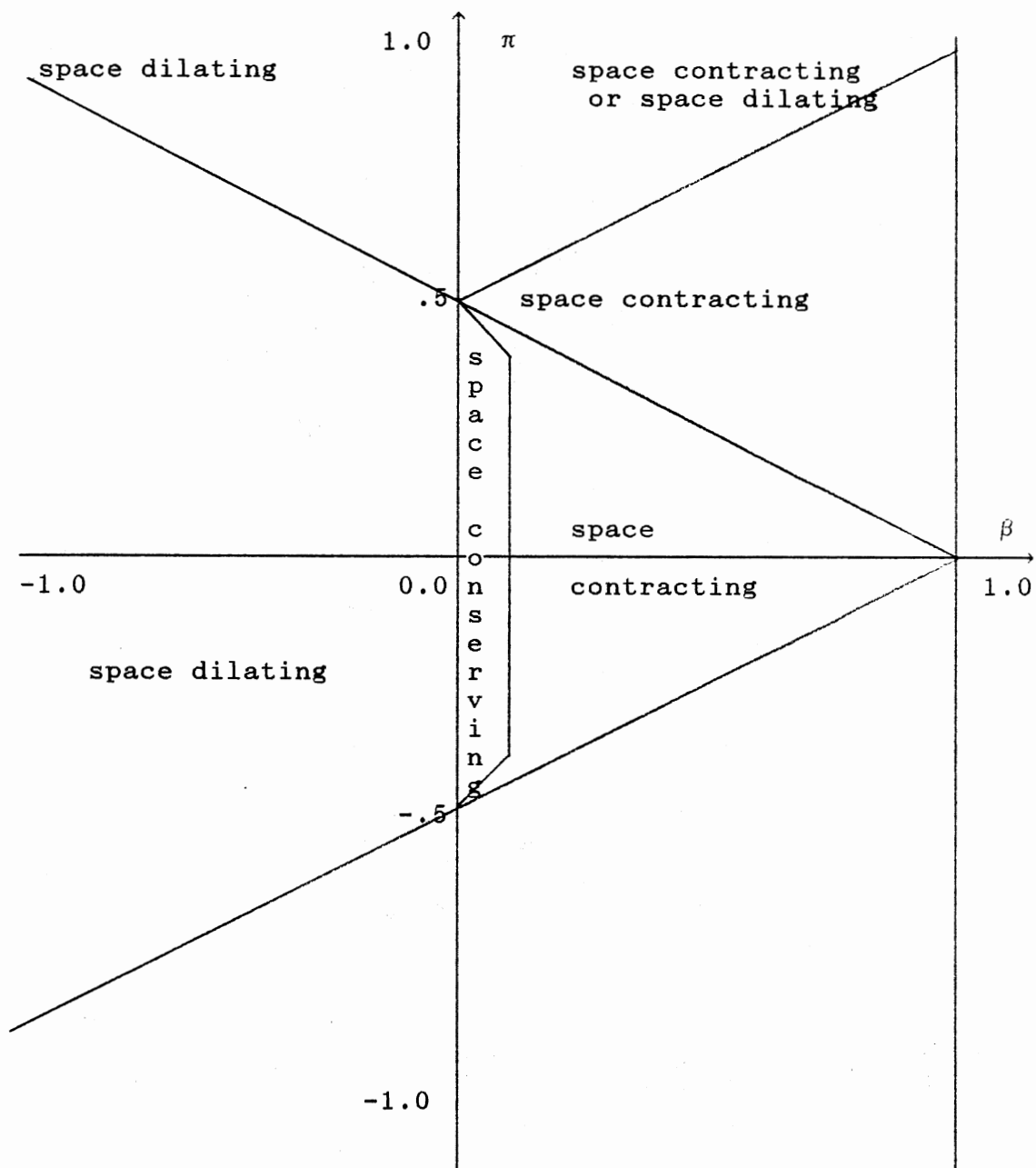


Figure 1. A Classification of the (β, π) family of Agglomerative Clustering Algorithms.

CHAPTER IV

USE OF A COMPARATIVE STATISTIC TO PREDICT THE NUMBER OF CLUSTER

A Comparative Statistic

The primary objective of this thesis is to investigate the use of a comparative statistic to predict the number of clusters in the object space. This will be achieved using a comparative statistic between the outcomes of different algorithms applied to the same data. The behavior of this comparative statistic will be used to determine the appropriate number of clusters.

Rand's (1969, 1971) C statistic is a very general and versatile statistic which may be used to compare clustering methods based on how they partition the object space. Further, C measures the similarity between two clusterings when clusterings from an $[N,K]$ -population of clusterings are produced by applying two different clustering methods to the same object space.

Rand (1971) makes the following three reasonable assumptions concerning the nature of a general clustering problem as a rationale for the development of the C statistic:

First, clustering is discrete in the sense that every point is unequivocally assigned to a specific cluster. Second, clusters are defined just as much by those points which they do not contain as by those points which they do contain. Third, all points are of equal importance in the determination of clusterings.

Thus, Rand (1971) points out that a basic unit of comparison between two clusterings is how pairs of points are clustered.

To facilitate the definition of the C statistic, definition 4.1 concerning the similar assignment of point-pairs is given.

Definition 4.1. Given an object space X consisting of N data points, X_1, X_2, \dots, X_N , and two clusterings of X , $Y = [Y_1, Y_2, \dots, Y_{K_1}]$ and $Y' = [Y'_1, Y'_2, \dots, Y'_{K_2}]$, then a similar assignment in clusterings Y and Y' of a pair of data points, X_i and X_j , results if and only if either of the following two conditions holds:

- (i) There exist k and g such that $X_i, X_j \in Y_k$ and $X_i, X_j \in Y'_g$;
- (ii) There exist k and g such that $X_i \in Y_k, Y'_g$, and $X_j \notin Y_k, Y'_g$.

Basically, if the elements of an individual point-pair are placed together in a cluster in each of two clusterings, or if they are assigned to different clusters in both clusterings, then a similar assignment of the point-pair has been made in the two clusterings. In essence, the C statistic gives a normalized count of the number of similar assignments of point-pairs between two clusterings as

designated in definition 4.2.

Definition 4.2. Given an object space X consisting of N data points, X_1, X_2, \dots, X_N , and two clusterings of X , $Y = [Y_1, Y_2, \dots, Y_{K_1}]$ and $Y' = [Y'_1, Y'_2, \dots, Y'_{K_2}]$, then the C statistic between Y and Y' is defined as follows:

$$C(Y, Y') = \frac{\sum_{i < j} m_{ij}}{\binom{N}{2}}, \quad (4.1)$$

where $i = 1, 2, \dots, N-1$, $j = 2, 3, \dots, N$, $i < j$ and

$$m_{ij} = \begin{cases} 1, & \text{if there is a similar assignment of} \\ & X_i \text{ and } X_j \text{ in } Y \text{ and } Y', \\ 0, & \text{otherwise.} \end{cases}$$

Hence, C is a measure of similarity on the set of all possible clusterings of X . Rand (1971) also gives a computational form for the C statistic, which is related to an incidence matrix concept. If the clusters within each clustering are arbitrarily numbered and n_{ij} represents the number of data points which are simultaneously in the i -th cluster of Y and the j -th cluster of Y' , then

$$C(Y, Y') = \frac{\binom{N}{2} - \frac{1}{2} \left(\sum_i (\sum_j n_{ij})^2 + \sum_j (\sum_i n_{ij})^2 \right) + \sum_{i,j} n_{ij}^2}{\binom{N}{2}}. \quad (4.2)$$

In this formulation, $C(Y, Y') = 1$ when the arbitrarily numbered clusters within each clustering correspond completely. Conceptually, $C(Y, Y') = 1$ when $K = 1$ or $K = N$ without justification, where K is the number of clusters for a given set of data. Thus, if two different clustering

algorithms are applied to the same set of data and the clusters within each clustering are similar, the values of $C(Y, Y')$ might be close to 1. Also, $C(Y, Y') = 0$ when the two clusterings have no similarities.

The C statistic has the following three fundamental properties as noted by Rand (1969, 1971):

1. C is a measure of similarity with $0 \leq C \leq 1$,
2. $1 - C$ is a measure of distance, being a metric on the set of all possible clusterings of X,
3. C is a random variable.

It should be noted that Rand (1969) provides a proof of the fact that $1 - C$ is a metric on \mathcal{Y} in his thesis, where \mathcal{Y} represents the set of all possible clusterings of X.

Another formulation of Rand's C statistic is worth noting. According to Anderberg (1973), the C statistic is equivalent to the simple matching coefficient. The simple matching coefficient, which was originally introduced to numerical taxonomy by Sokal and Michener (1958), is a binary measure of association based on 2×2 contingency tables. To demonstrate the equivalent relationship between Rand's C statistic and the simple matching coefficient, a particular form of the simple matching coefficient will be developed.

The simple matching coefficient may be used to assess the amount of agreement between any two binary vectors of the same length, where a binary vector is defined in definition 4.3.

Definition 4.3. A vector $V = (v_1, v_2, \dots, v_n)$ is a

binary vector if and only if for each $i = 1, 2, \dots, n$,
 $v_i = 1$ or $v_i = 0$.

To compute the simple matching coefficient, it is necessary to define a match between two binary vectors as defined in definition 4.4.

Definition 4.4. A match between the corresponding components of two binary vectors, $U = (u_1, u_2, \dots, u_n)$ and $V = (v_1, v_2, \dots, v_n)$, occurs if and only if $u_i = v_i$.

If the number of matches between two binary vectors of length n is denoted by m , then a definition for the simple matching coefficient is given by definition 4.5.

Definition 4.5. The simple matching coefficient between two binary vectors, U and V , of length n is given by

$$T(U, V) = \frac{m}{n} , \quad (4.3)$$

where m is the number of matches between the two binary vectors. Thus, the simple matching coefficient represents a normalized count of the number of matches between two binary vectors.

If a clustering can be represented as a binary vector, then a simple matching coefficient between clusterings can be computed. A binary representation of a clustering can be obtained by constructing a binary vector, U , consisting of $n = \binom{N}{2}$ components, where each component of U indicates whether a pair of data points are together or apart in the clustering. Letting X be an object space consisting of N data points, then a more precise formulization of a binary representation of a clustering is given in definition 4.6.

Definition 4.6. The binary vector, $U = (u_{12}, u_{13}, \dots, u_{1n}, u_{23}, \dots, u_{ij}, \dots, u_{n-1,n})$, is a binary representation of a clustering, $Y = (Y_1, Y_2, \dots, Y_k)$ if and only if for all $i = 1, 2, \dots, N-1, j = 2, 3, \dots, N, i < j$,

$$m_{ij} = \begin{cases} 1, & \text{if there is a cluster } Y_k \in Y \text{ such that} \\ & X_i, X_j \in Y_k, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, if U is a binary representation of clustering Y , V is a binary representation of clustering Y' and m is the number of matches between two binary vectors of length n , then

$$T(U, V) = \frac{m}{n} = \frac{m}{\binom{N}{2}} = \frac{\sum_{i < j} m_{ij}}{\binom{N}{2}} = C(Y, Y').$$

Consequently, Rand's (1969, 1971) C statistic is equivalent to the simple matching coefficient.

As noted previously, C possesses a probability distribution since C is a random variable under certain assumptions. However, as Rand (1969) notes, the distribution of C is complicated. Logically, part of the complication with respect to the distribution of C concerns the choice of the space on which initial distributional assumptions should be placed. Conceptually, X is a subset of Euclidean p -space with cardinality N for an $[N]$ -population of clusterings; a clustering method maps X into $Y^{[N]}$; and

$$C: Y^{[N]} * Y^{[N]} \longrightarrow [0, 1].$$

In studies by DuBien (1976), and DuBien and Warde (1982) on Rand's C statistic and its distribution under certain assumptions, a Stirling numbers of the second kind has been used. Stirling numbers of the second kind may be computed by the following formula:

$$S(N,K) = \frac{1}{K!} \sum_{j=0}^K \binom{K}{j} (-1)^j (K-j)^N, \quad (4.4)$$

where $K = 1, 2, \dots, N$.

Since Stirling numbers of the second kind are closely associated with the counting of clusterings, some results from Duran and Odell (1974) concerning these numbers are presented. By definition,

$$S(N,0) = 0;$$

and

$$S(N,N+\ell) = 0, \quad \text{if } \ell > 0.$$

A recursive relationship which is fundamental to the counting of clusterings is given as follows:

$$S(N+1,K) = K S(N,K) + S(N,K-1). \quad (4.5)$$

For an $[N,K]$ -population, the following fundamental results concerning clusterings and their binary representations facilitate the derivation of the distribution for C statistic.

- (1) The total number of binary representations in an $[N,K]$ -population corresponds to $Q = S(N,K)$;
- (2) The frequency of 1's on the (ij) -th component of the binary representations in the $[N,K]$ -population is a constant for all $i = 1, 2, \dots, N-1$, $j = 2, 3,$

..., N , $i < j$, and this constant is denoted by

$$Q_1 = S(N-1, K);$$

- (3) The frequency of pairs of 1's on the (ij) -th and (st) -th components of the binary representations in the $[N, K]$ -population is a constant for all $i, s = 1, 2, \dots, N-1, j, t = 2, 3, \dots, N, i < j, s < t$, and $i \neq s$ or $j \neq t$, and this constant is denoted by $Q_{11} = S(N-2, K)$.

The fundamental results establish the fact that Q_1 and Q_{11} are constants for all components and pairs of components, respectively, of the binary representations of the clusterings in an $[N, K]$ -population. Thus, the following additional notation for $[N, K]$ -populations follows directly from the fundamental results:

- (i) $Q_0 = Q - Q_1$ (the frequency of 0's);
- (ii) $f_1 = \frac{Q_1}{Q}$ (relative frequency of 1's);
- (iii) $f_0 = \frac{Q_0}{Q}$ (relative frequency of 0's);
- (iv) $Q_{10} = Q_{01} = Q_1 - Q_{11}$ (frequency of a 1-0 pair);
- (v) $Q_{00} = Q_0 - Q_{01}$ (frequency of pairs of 0's).

Further, two fundamental assumptions are assumed throughout the derivations:

- (1) The clusterings Y , in an $[N, K]$ -population of clusterings have a discrete uniform probability distribution; that is, for all $Y^{[N, K]}$,

$$P\{\text{choosing any particular } Y^{[N, K]}\} = \frac{1}{Q},$$

- (2) The two clusterings Y and Y' , with binary representations U and V , respectively, are

selected randomly with replacement from an
[N,K]-population of clusterings.

Consequently, if (Y, Y') represents an ordered pair of clusterings from an [N,K]-population, then under these assumptions,

$$P\{\text{choosing any particular ordered pair } (Y, Y')\} = \frac{1}{Q^2}.$$

Then, the mean and variance for the similarity between two clusterings drawn at random with replacement from an [N,K]-population of clusterings are given in (4.6) and (4.7), respectively. These are,

$$E(C) = f_1^2 + f_0^2 = \frac{1}{Q^2} [Q_1^2 + Q_0^2] \quad (4.6)$$

and

$$\text{VAR}(C) = \frac{E(C)[1-E(C)]}{\binom{N}{2}} + \frac{\binom{N}{2} - 1}{\binom{N}{2}} \{p_2 - [E(C)]^2\}, \quad (4.7)$$

$$\text{where } p_2 = \frac{1}{Q^2} (Q_{11}^2 + 2 Q_{10}^2 + Q_{00}^2).$$

These results also hold for [N]-populations of clusterings with slightly different values for Q , Q_1 , and Q_{11} as shown in DuBien and Warde (1982).

Since [N]-populations are obtained by merging all of the [N,K]-populations for $K = 1, 2, \dots, N$, then Q , Q_1 , and Q_{11} are obtained by adding the Stirling numbers of the second kind on K . Hence, for an [N]-population of clusterings,

$$Q = L_N = \sum_{K=1}^N S(N, K);$$

$$Q_1 = L_{N-1} = \sum_{K=1}^{N-1} S(N-1, K);$$

$$Q_{11} = L_{N-2} = \sum_{K=1}^{N-2} S(N-2, K).$$

Given these new values for Q , Q_1 , and Q_{11} , the previously derived formulas in (4.6), and (4.7) will hold for the mean and variance of C when the clusterings are randomly chosen with replacement from an $[N]$ -population of clusterings. Thus, the mean and variance of Rand's (1971) C statistic depend only on an appropriate set of the Stirling numbers of the second kind.

For the purpose of this study, the examination of the behavior of the similarity measure, C , for changing k is of interest in some situations. Thus, C will be represented as $C_k(Y, Y')$, which is the similarity measure between one clustering Y and another clustering Y' having the same number of clusters, k , resulting from different agglomerative clustering procedures applied to the same set of N data points, where $k = 1, 2, 3, \dots, N$. Also, it may be considered that the number of objects in each clusters within clusterings, Y and Y' , are different. However, the same number of clusters within two different types of clusterings generated by applying nine agglomerative clustering algorithms to a given set of data is assumed in this study.

Other Measures of Similarity

Several researchers have developed measures of similarity between hierarchical clusterings. Anderberg

(1973), and Hubert and Levin (1976) proposed measures that are functions of the incidence matrix, $[n_{ij}]$. In these measures, either they use one number to summarize the similarity between two hierarchical clusterings or they compare the clusterings for some fixed number of partitions.

Fowlkes and Mallows (1983) introduced the B_k statistic and tried to investigate the use of a sequence of measures, B_k , as the basis for a plotting procedure where $k = 2, 3, \dots, N-1$ and N is the number of objects. Further, they derived the mean and variance of B_k , under the assumption that the margins of the incidence matrix, $[n_{ij}]$, are fixed. Based on this formulation and by using the incidence matrix, $[n_{ij}]$, the B_k is calculated for each value of k , where $k = 2, 3, 4, \dots, N-1$. That is,

$$B_k = \frac{T_k}{[P_k \ Q_k]^{1/2}}$$

where

$$T_k = \sum_i^k \sum_j^k n_{ij}^2 - N,$$

$$P_k = \sum_i^k \left(\sum_j^k n_{ij} \right)^2 - N,$$

$$Q_k = \sum_j^k \left(\sum_i^k n_{ij} \right)^2 - N,$$

and N is the number of objects. The quantity n_{ij} is the number of objects in common between the i -th cluster in one clustering and j -th cluster in the other clustering, where $i, j = 1, 2, 3, \dots, k$ and $k = 2, 3, \dots, N-1$. Then, various properties of B_k have been investigated by means of

a series of Monte Carlo experiments.

They show that the B_k statistic has the following properties :

1. For each k , $0 \leq B_k \leq 1$;
2. $B_k = 1$, if $[n_{ij}]$ has exactly k nonempty cells, which happens when the k clusters within each clustering correspond completely;
3. $B_k = 0$, if each $n_{ij} = 0$ or 1 , so that every pair of objects that appear in the same cluster in one clustering are assigned to different clusters in other clustering; If $k = N$, $[n_{ij}]$ is a permutation matrix, and B_k is indeterminate, which is different from Rand's C_k .

Further, B_k has a probability distribution since it is considered to be a random variable under certain assumptions.

In their derivation of the mean and variance of B_k , Fowlkes and Mallows (1983) assumed that the margins of the incidence matrix, $[n_{ij}]$, namely, $(n_{.j}, n_{i.})$ were fixed. Then, the mean and variance for the B_k are;

$$E(B_k) = \frac{1}{2} \frac{[P_k Q_k]^{1/2}}{\binom{N}{2}},$$

$$\text{Var}(B_k) = \frac{1}{\binom{N}{2}} \left(1 + \frac{2}{(N-2)} \frac{P'_k Q'_k}{P_k Q_k} \right).$$

$$+ \frac{1}{2} \frac{(P_k' - 2 - 4 \frac{P_k'}{P_k}) (Q_k' - 2 - 4 \frac{Q_k'}{Q_k})}{(N - 2)(N - 3)} \Bigg\} - [E(B_k)]^2,$$

where

$$P_k' = \sum_{i=1}^k n_{i.} (n_{i.} - 1) (n_{i.} - 2),$$

$$Q_k' = \sum_{j=1}^k n_{.j} (n_{.j} - 1) (n_{.j} - 2).$$

Thus, the mean and variance of B_k depend only on the given assumptions, the fixed margins of the incidence matrix, $[n_{ij}]$. However, this assumption is only valid if the two clusterings are unrelated to each other.

Further, they defined the limits $E(B_k) \pm 2(\text{VAR}(B_k))^{1/2}$ and pointed out the defined limits give only an approximate indication of the significance of the similarity between two hierarchical clusterings, since successive values of B_k are correlated and the distribution of B_k is not normal.

On the other hand, Morey and Agresti (1984) suggested using the Rand statistic adjusted with respect to chance agreement for a pairing. They noted that the Rand's (1971) C_k can produce fairly large values even for randomly paired sets of partitions, since C_k does not take into account chance agreement. The adjustment factor, N_c , developed by Morey and Agresti (1984) gives the adjusted Rand statistic. The adjusted Rand statistic is given in the following form;

$$A_k = \frac{N_s - N_c}{\binom{N}{2} - N_c},$$

where

$$N_s = \binom{N}{2} - \frac{1}{2} \left(\sum_i n_{i.}^2 + \sum_j n_{.j}^2 \right) + \sum_{i,j} n_{ij}^2$$

and

$$N_c = \binom{N}{2} - \frac{1}{2} \left(\sum_i n_{i.}^2 + \sum_j n_{.j}^2 \right) + \frac{\sum_{i,j} n_{i.}^2 n_{.j}^2}{n^2}.$$

The properties of the adjusted Rand statistic could be summarized as follows:

1. For each k , $-1.0 \leq A_k \leq 1.0$;
2. $A_k = 1.0$, if the clusters within each clustering corresponds completely;
3. $A_k = 0.0$, if $N_s = N_c$ (i.e., for chance agreement).

Moreover, $A_k < 0.0$ when agreement of the clusters within each clusterings is less than that expected by chance. If $A_k > 0.0$, it represents the proportion of the maximum possible difference obtained between the probability of agreement and the probability of chance agreement. Note that the design of the adjustment factor is based on the logic of computing expected cell counts in a chi-square test for independence.

However, the components of a binary representation are not independent in terms of matches, since a clustering is a special type of structure as shown by DuBien and Warde (1982).

Rationale for the use of C_k to predict
the number of clusters

The difficulty of determining the number of clusters in

a set of data has been noted by many authors, including Friedman and Rubin (1967), Marriot (1971), Sneath and Sokal (1973), Hubert and Levin (1976), Ratkowsky and Lance (1978), Ratkowsky (1983), and Krzanowski and Lai (1988). They attempted to derive formal tests by optimizing some clustering criteria for determining the appropriate number of groups within clusterings.

An early attempt at its solution was made by Thorndike (1953), who plotted the average within-cluster distance against the number of groups. He suggests that a sudden marked flattening of the curve at any point indicates a distinctively "correct" value for k , since such a point will occur when the number of groups uniquely corresponds to the configuration of points and there is relatively little gain from further increase in k . Unfortunately, the derived curves by using artificial data provide little support for this intuitive notion.

In general, a plot of the criterion value against the number of clusters indicate the correct number to consider by showing a sharp increase (or decrease, depending on the criterion applied), at the correct number of clusters. However, the procedure has been found to be unsatisfactory, since the decision as to whether such plots contain the necessary "sharp step" is likely to be exceedingly subjective in practice (Everitt, 1979).

On the other hand, hierarchical clustering procedures have no clear indicators for the number of clusters. If

some indication of the correct number is required, an examination of the dendrogram or tree diagram for large changes between fusions would be useful. However, distinct clustering methods often produce quite different clusterings, even though they are applied to the same set of data depending on the structure within data. This implies that examination of the dendrograms or tree diagram given by agglomerative hierarchical clustering methods is not always helpful and may lead to misleading conclusions.

In their study, Fowlkes and Mallows (1983) suggest useful and interpretable methods for exploring the number of groups and comparing the results of clustering algorithms by using a similarity measure. The measure, B_k , and the plots, (k, B_k) , can be readily computed and displayed. They indicate that in comparing the original clustering of mixture data with the clustering of perturbed data, the (k, B_k) plots tend to peak at the k which is equal to the true number of clusters. This stimulates the consideration of a similar technique applying Rand's C_k for predicting the number of clusters present in a given set of data.

The two measures of similarity (Rand's (1971) C_k and Fowlkes and Mallows' (1982) B_k) between two hierarchical clusterings are somewhat similar in construction. They both depend on the incidence matrix, $[n_{ij}]$. Also, it is worth noting that C_k and B_k range from zero to one for every k . These similarity measures are equal to one when the k clusters in each clustering correspond completely or when k

$= 1$, and equal to zero when every pair of objects that appear in the same cluster in an initial clustering obtained by using an agglomerative clustering method are assigned to completely different clusters when another agglomerative clustering algorithm is used. However, $C_k = 1$ while B_k is indeterminate when $k = N$. In summary, both measures of similarity, B_k and C_k , have the following properties:

1. They depend on the matching matrix, $[n_{ij}]$;
2. They lie between 0.0 and 1.0;
3. They are 1.0 if the k clusters within each clustering correspond completely (except at $k = N$);
4. They are 0.0 if every pair of objects that appear in the same cluster in one clustering are assigned to different clusters in another clustering.

At this point, it is of interest to examine the behavior of the measure C_k for every k in some situations to predict the number of clusters for the given object space.

For the purpose of this study, three observations concerning the C_k statistic will suffice:

1. The closer C_k is to 1.0, the more similar are the two clusterings;
2. If $C_k(Y, Y') > C_k(Y, Y'')$, then Y and Y' are more similar than Y and Y'' ;
3. If $C_k(Y, Y') \geq C_{k-1}(Y, Y')$ and $C_k(Y, Y') > C_{k+1}(Y, Y')$, then C_k is the local maximum for given k for the two clusterings.

It is known that two distinct clustering methods often

produce two quite different clusterings from the same set of data, depending on the structure within the data. However, if the results of several different clustering procedures agree closely, then one may have more confidence in the reality of the common group structure which is indicated for the given set of data. In this sense, an investigation of the use of a comparative statistic in conjunction with several agglomerative clustering algorithms might provide useful information on determining the number of clusters within a given set of data. This will be accomplished by investigating the behavior of C_k , which is a similarity measure between the resultant clusterings produced by agglomerative clustering algorithms. Also, this study will provide useful information about the properties of different agglomerative clustering procedures by observing the effect of controlled structural parameters. The design of a comparative study will be discussed in the following chapter.

CHAPTER V

DESIGN OF A COMPARATIVE STUDY AND RESULTS FROM MULTIVARIATE NORMAL SAMPLES

Parameter Choice

The design of this comparative study follows that suggested by DuBien (1976) and is augmented to investigate the use of a comparative statistic in determining the number of clusters present within the given object space.

A clustering method is purported to be a functional mechanism for finding or retrieving the "natural" structure within data. Hence, the degree to which a clustering method "retrieves" the known structure within generated data is an important characteristic of the clustering method.

Moreover, if two different clustering methods are applied to the same set of data, the degree to which the two retrieved structures correspond to each other through their resultant clusterings is another characteristic to be considered in this comparative study. This characteristic could be thought of as the "agreement" between two clustering methods for any specific number of clusters for given set of data.

To quantify the "retrieval" ability of a clustering method and the "agreement" between the two clustering

methods, N data points are generated from K well-separated populations. Let Y represent the "true" structure of the data. Let Y' and Y'' denote the two different clusterings which result from applying two different clustering methods to the same N data points. Then $C_k(Y, Y')$, $k = 2, 3, \dots, K, \dots, N-1$, is a measure of the "retrieval" ability of the clustering method to the true structure generated, while $C_k(Y', Y'')$, $k = 2, 3, \dots, K, \dots, N-1$, is a measure of the "agreement" between the two clustering methods through their resultant clusterings (subject to the random variation in the generated data).

Further, "noise" in terms of the performance of a clustering method might be explained as interference with the ability of the clustering method to "retrieve" the true structure present in the data. The simulation of a particular type of "noise" by means of changing the correlation between variables embodies the essence of the idea in DuBien (1976), and DuBien and Warde (1987). For bivariate data, DuBien and Warde (1987, p. 1443) remark:

If ρ represents the population correlation between the two variables within a single population of data points, then the level of "noise" existent in this population to obscure the clustering of data points from this population into the same cluster is quantified by specification of a value for ρ . Thus, a specification of $\rho \neq 0.0$ implies that each variable within the single population of data points is semi-informative rather than completely informative or completely uninformative. It should also be noted that increasing ρ , $\rho \geq 0$, for an otherwise fixed population of data points causes the data points within this population to be systematically shifted from an approximately circular configuration to a more elliptical configuration.

In their study of the effect of increasing ρ , $\rho \geq 0$, on the "retrieval" ability of several agglomerative clustering methods, DuBien and Warde (1987) find out that the correlated variables affect the "retrieval" ability of different agglomerative clustering methods differently. They recommend three (β, π) algorithms, the flexible strategy at $(-0.25, 0.0)$, $(-0.25, 0.25)$, or $(-0.25, 0.5)$ for finding the unknown structure present in many data sets regardless of the amount of noise and the relative sizes of the clusters present in the data.

As an extension of DuBien and Warde's study, the effect of changing ρ , $\rho \geq 0.0$, on the "agreement" between the two clustering methods is investigated in this present study. If the results of several different clustering procedures agree closely, then we may have more confidence in the reality of any cluster structure which is indicated. Based on the "agreement" between the two clustering methods, we might be able to predict the number of clusters present in the data.

For convenience, the important consideration in any extensive, systematic comparison of clustering methods shall be termed structural parameters; a structural parameter is any variable which controls some aspect of the structure of the data. The data set of structural parameters for a comparative study of clustering methods should consist of all variable features within data which might affect the resultant clusterings. Some of the possible structural

parameters which require controlled change to make a comparative study "dynamic" are defined as follows:

1. N , the number of data points in X ;
2. p , the number of variables defining each data points; i.e., the dimensionality of the Euclidean p -space in which X is embedded;
3. K , the number of populations from which the data points are generated;
4. The types of population or the probability distribution from which each of the K populations of data points are generated;
5. μ_k , $k = 1, 2, \dots, K$, the mean vectors for each population of data points;
6. Σ_k , $k = 1, 2, \dots, K$, the variance-covariance structure for each population of data points;
7. δ_i , $i = 1, 2, \dots, \left(\frac{K}{2} \right)$, the distance between each pair of population mean vectors;
8. The relative location of the population mean vectors or the spatial configuration of the population mean vectors;
9. The split or n_k , $k = 1, 2, \dots, K$, the number of data points generated from each population of data points.

In any comparative study of clustering methods, some of the structural parameters in the set of possible structural parameters remain fixed. Then a few of the structural parameters of special interest may be extensively studied

over a range of meaningful settings for a fixed set of clustering methods. Since the primary objective of this comparative study is to investigate the use of the suggested comparative statistic, C_k , in determining the number of clusters within the populations of data points, several structural parameters should be considered for meaningful interpretations of the results of the comparative study and application to clustering methods. Therefore, the particular structural parameters of interest for the comparative study of nine agglomerative clustering methods are specified, and the fixed variable settings for these structural parameters are outlined in the next section.

A Discussion on the Design of the Comparative Study

In terms of the design of the comparative study, it is necessary to specify the setting for each of the fixed structural parameters and the range of settings for each of the structural parameters. For the purpose of this study, the probability distribution for each of the K populations of data points generated was fixed to be multivariate normal (MVN) with the same variance-covariance matrix. MVN vectors were generated from a population having a mean vector of zero with any specified positive definite, symmetric variance-covariance matrix. The subroutine GGNSM from the IMSL (International Mathematical and Statistical Library) catalogued programs was used to generate data.

In this study, the number of data points, the number of variables per data point and the number of MVN populations of data points in X were fixed at the following values:

$$1). \quad N = 60;$$

$$2). \quad p = 2;$$

$$3). \quad K = 3.$$

The choice of $N = 60$ was arbitrary. The choice of $p = 2$ was necessary to simplify the design of the comparative study and to enhance the interpretability of the results from the comparative study. One rationale for choosing $K = 3$ is to maintain the information content of the variables within a population of data points, and it is important to choose $K > p$. The choice $K = 3$ was also related to the choice of a potentially interesting spatial configuration for the population mean vectors.

To facilitate the controlled change of the structural parameters δ_i , $i = 1, 2, \dots, \binom{K}{2}$, it was well suited to quantify the distance between population mean vectors by a single structural parameter, δ_i ; i.e.,

$$\forall \delta_i = \delta \text{ for } i = 1, 2, \dots, \binom{K}{2}.$$

The number of populations, K , was fixed at three and the representation of the distance between the population mean vectors by a single structural parameter implies that the population mean vectors are equally spaced in the plane. Consequently, the spatial configuration for the population mean vectors was automatically fixed so that the three population mean vectors were always placed at the vertices

of an equilateral triangle. It should be noted that the specification of a value for distance, δ , in conjunction with the equilateral triangle configuration for the population mean vectors is sufficient with respect to locating the population vectors in Euclidean two-space since the actual location of the equilateral triangle in the plane does not affect the performance of an agglomerative clustering method. Therefore, N , p , K , the generating probability distribution, and the spatial configuration of the population mean vectors remained fixed at the previously mentioned settings throughout the comparative study.

The three structural parameters subject to controlled variation in the comparative study were ρ , δ , and split. The settings for the structural parameter δ , the distances between the mean vectors, were set at $\delta = 4.0$, and 6.0 . It has been demonstrated by other investigator (e.g., Everitt, 1974) that some clustering methods opt for equal sized clusters. Thus, a limited investigation of the robustness of nine agglomerative clustering methods to unequal sized clusters was attempted by contrasting the equal sized cluster setting for split, 20-20-20, with an unequal sized cluster setting for split, 30-20-10.

The variance-covariance structure for bivariate normal (BVN) populations of data points was one of interest in the comparative study. Since Everitt (1974) has demonstrated that some clustering methods opt for circular clusters, the structural parameter of interest in the variance-covariance

structure was ρ . Thus the data points forming the object space X were generated from three similar BVN populations with a specified value of ρ and unit variances; i.e.,

$$\forall k = 1, 2, 3, \Sigma_k = \Sigma = \begin{bmatrix} 1.0 & \rho \\ \rho & 1.0 \end{bmatrix},$$

where $\rho = 0.0, 0.4, \text{ and } 0.8$.

Consequently, the effect of correlated variables on the "retrieval" ability of the nine clustering methods and the "agreement" between two clustering algorithms consisting of a pair could be investigated by fixing all structural parameters except ρ which is systematically varied across its range of settings.

In Figure 2, the actual population mean vectors used in this study are portrayed for $\delta = 4.0$ and the equilateral triangle spatial configuration of population mean vectors. When the identity matrix is assumed for the variance-covariance matrix, Σ , the three circles represent the 2σ contours for each of BVN populations. Data points generated from this structural framework which, because of random variation, fall in the overlapping regions of the three circles are likely to be clustered with data points generated from a different BVN population than the one from which they were generated. This observation, of course, illustrates only one of the possible reasons that a clustering method fails to "retrieve" the exact structure as generated or two clustering algorithms consisting of a pair fail to agree through their resultant clusterings.

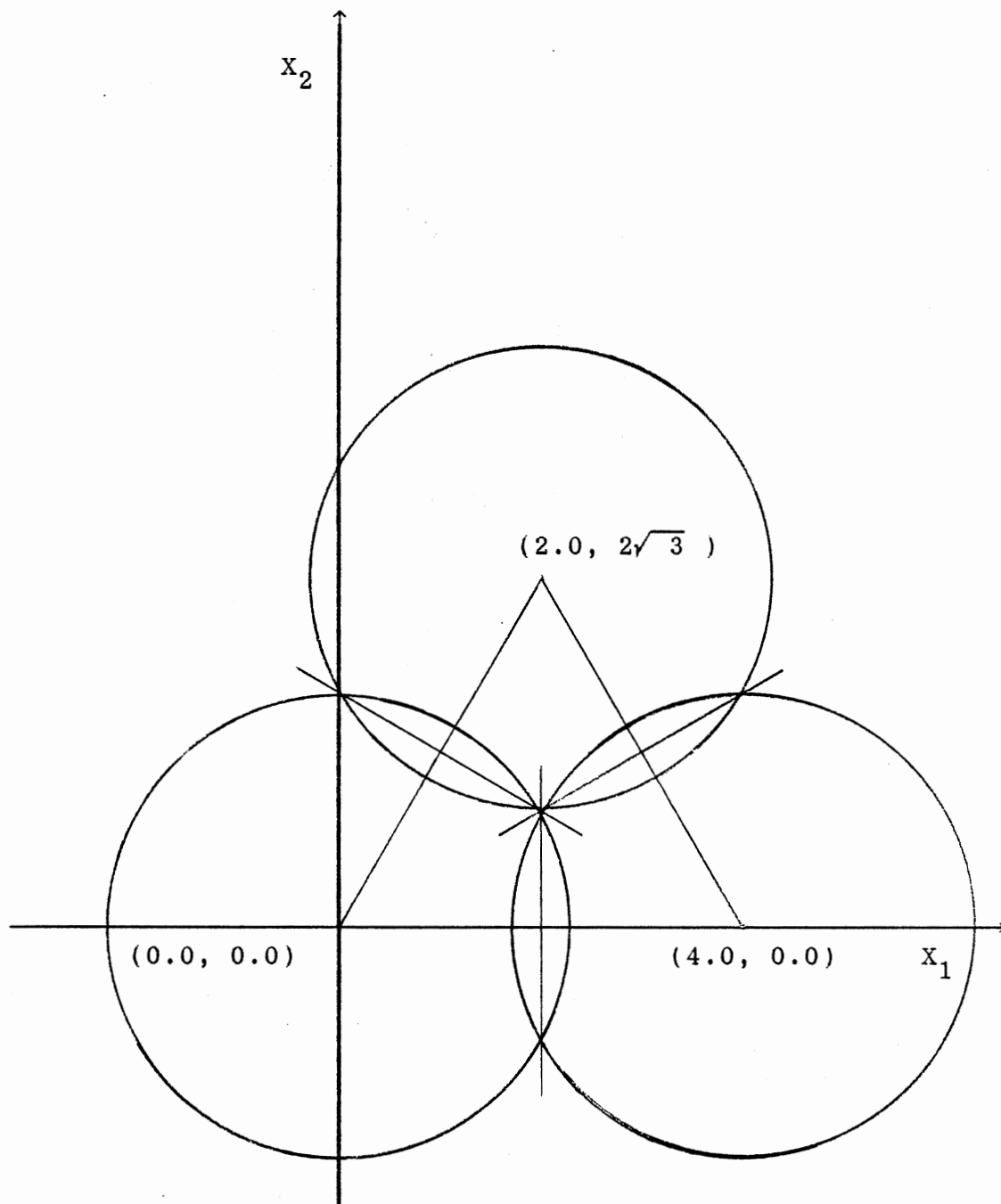


Figure 2. An Example of the Structural Framework Developed for BVN

A brief summary of data structure for the comparative study of agglomerative clustering methods may be outlined as follows:

$$X_i \sim \text{BVN}(\mu_k, \Sigma)$$

where: $i = 1, 2, \dots, 60$ with split into the $K = 3$ populations of either 20-20-20 or 30-20-10;

: μ_k , $k = 1, 2, 3$, is constrained by an equilateral triangle spatial configuration and $\delta = 4.0, 6.0$;

$$: \Sigma = \begin{bmatrix} 1.0 & \rho \\ \rho & 1.0 \end{bmatrix}, \rho = 0.0, 0.4, \text{ and } 0.8.$$

For this comparative study, the measure of distance was fixed to be squared Euclidean distance based on the claim by DuBien (1976) that the measure of distance is not as important in determining the resultant clusterings as the algorithm. The agglomerative clustering algorithms chosen for the comparative study were discussed by DuBien (1976); however, only nine agglomerative clustering algorithms are chosen in this study as explained in Chapter III.

For each setting of the triple $(\rho, \delta, \text{split})$, the following sequence of steps was utilized to generate values of C_k , $k = 2, 3, \dots, K, \dots, N-1$.

1. An object space X of data points is generated for the complete set of structural parameters;
2. The squared Euclidean distance between each pair of data points in X is computed and stored in standard

lower triangular matrix order by rows as the vector D ;

3. Each of the nine (β, π) agglomerative clustering algorithms is applied to D to produce a hierarchy, H_a , where $a = 1, 2, \dots, 9$;
4. For each of the nine agglomerative clustering algorithms, the k -cluster clusterings $(Y')_a$ and $(Y'')_a$ for $k = 2, 3, \dots, K, \dots, N-1$ are generated from hierarchy H_a ;
5. Each of the k -cluster clusterings for $k = 2, 3, \dots, K, \dots, N-1$, $(Y')_a$, $a = 1, 2, \dots, 9$, is compared by means of the $C_k(Y, Y')$ statistic to the true clustering Y of size three, which clustered together all data points generated from the given population of data points;
6. Both k -cluster clusterings, $(Y')_a$, $(Y'')_a$, $a = 1, 2, \dots, 9$, are compared by means of the $C_k(Y', Y'')$ statistic, where $k = 2, 3, \dots, K, \dots, N-1$;
7. By means of the above sequence of steps, a value $C_k(Y, Y')$ is computed for each algorithm, and $C_k(Y', Y'')$ is computed for each pair of 36 the possible pairs of agglomerative clustering algorithms in each replication for all $k = 2, 3, \dots, K, \dots, N-1$;
8. Then, the above sequence of steps is replicated 100 times for each setting of the triple $(\rho, \delta, \text{split})$ for all $k = 2, 3, \dots, K, \dots, N-1$;
9. \bar{C}_k , the sample mean, and S_C , the sample standard

deviation of C_k values, $k = 2, 3, \dots, N-1$, are obtained for the 100 replications;

10. The % of the replications which satisfy the conditions,

$$C_{k-1} \leq C_k \text{ and } C_{k+1} < C_k ,$$

for a known number of clusters, K , i.e., the number of times that C_k is a local maximum at given k , where $k = 2, 3, \dots, K, \dots, N-1$, is obtained for nine agglomerative clustering algorithms and all possible pairs of them;

11. $\bar{\%}$, the sample mean, and $S_{\bar{\%}}$, the sample standard error of $\bar{\%}$ values across all settings of the structural parameters ($\rho, \delta, \text{split}$) for nine algorithms and possible pairs of them.

Consequently, for each setting of the structural parameters, ($\rho, \delta, \text{split}$), the % resulting from 100 replications quantifies the "retrieval" ability of a clustering method, and the "agreement" between two clustering methods consisting of a pair. Specifically, the % obtained by $C_k(Y, Y')$ for each of the nine agglomerative clustering algorithms quantifies how well a clustering algorithm retrieves the known structure. The % calculated by $C_k(Y', Y'')$ for possible pairs of clustering algorithms quantifies how well two algorithms in each pair agree to each other through their resultant clusterings giving a local maximum at $k = 3$. At this point, the % calculated by

$C_k(Y', Y'')$ will be defined as $\%_s$ which is the number of times that two clustering algorithms estimate the number of clusters correctly.

Thus, the triple $(\bar{C}_k, S_c, \%)$ provides information on how well the comparative statistic, C_k , retrieves the "true" structure generated. The triple $(\bar{C}_k, S_c, \%_s)$, resulting from 100 replications provides a method for investigating the use of the comparative statistic when specific pairs of the nine agglomerative clustering methods are applied simultaneously for particular settings of the structural parameters. In addition, $(\bar{\%}, S_{\bar{\%}})$ and $(\bar{\%}_s, S_{\bar{\%}_s})$ provide information on how well the C_k "retrieves" the true structure and "estimates" the specified number of clusters, respectively, across all settings of the structural parameters.

At this point, the behavior of \bar{C}_k is observed for $3 \times 2 \times 2$ settings of $(\rho, \delta, \text{split})$ on the $\binom{9}{2}$ possible pairs of the nine agglomerative clustering algorithms. Then the pairs of clustering algorithms for specific settings of structural parameters will be chosen for further study as follows:

1. If the value of \bar{C}_k is close to 1.0 at $k = 3$ where the values of \bar{C}_k are considerably smaller for $k \neq 3$,
2. If a local maximum at $k = 3$ occurs frequently.

The results from the comparative study on the BVN population of data points are discussed in the following section.

Discussion of Results from Multivariate Normal Samples

Tables 2-10 in the Appendix give the results from the comparative study for the use of the comparative statistics, C_k , in predicting the number of clusters for BVN samples by applying nine agglomerative clustering algorithms.

In these tables, the results are computed over 100 replications for each setting of the structural parameters $(\rho, \delta, \text{split})$ and for the nine agglomerative clustering algorithms formed with squared Euclidean distance. An observed % will be interpreted as the "retrieval" ability of the nine agglomerative clustering methods. And a %_s will be interpreted as the "agreement" between two different clustering methods. An observed difference or similarity among the nine clustering methods will be discussed in terms of the algorithms defined by (β, π) . Also, it should be noted that the results from the comparative study are not independent of the structural settings, $(\rho, \delta, \text{split})$, which were specified in the previous sections. Thus, all results from the comparative study will be discussed in terms of changes in the structural parameters $(\rho, \delta, \text{split})$ and the ordered pair (β, π) . To enhance the interpretation of the results from the comparative study, figures 4-9 in the appendix portray the various behaviors of the comparative statistics, \bar{C}_k , $k = 2, 3, \dots, 10$, for the nine agglomerative clustering algorithms. Tables 2-10 and figures 4-7 given in

the appendix will be discussed in detail.

In tables 2-10, the % and %_s represent the number of times that a local maximum occurs at $k = 3$ over 100 replications. The performance of C_k is considered to be good with clustering algorithms when the % and %_s of local maxima are high and stable across the settings (ρ , δ , split) of the structural parameters. If any criterion is required, the hypothesis that a difference exists between %'s obtained for the nine agglomerative clustering algorithms, and between %'s obtained for the $\binom{9}{2}$ possible pairs of clusterings algorithms can be tested.

Let %[A] and %[B] be % values produced by algorithm A and B, respectively. Since %[A] and %[B] are the numbers of times that a local maximum occurs at $k = 3$ over 100 replications, %[A] and %[B] follow a binomial probability distribution, with parameters p_a and p_b , respectively. For large samples the point estimator of $(p_a - p_b)$, namely $(\hat{p}_a - \hat{p}_b)$, is approximately normally distributed, with a mean of $(p_a - p_b)$ and a standard deviation of

$$\sigma(\hat{p}_a - \hat{p}_b) = \sqrt{\frac{p_a q_a}{n_a} + \frac{p_b q_b}{n_b}} .$$

Then

$$z = \frac{(\hat{p}_a - \hat{p}_b) - (p_a - p_b)}{\sigma(\hat{p}_a - \hat{p}_b)} .$$

possesses a standard normal distribution. Hence z can be employed as a test statistic to test

$$H_0: p_a = p_b ,$$

when suitable approximations are used for p_a and p_b , which appear in $\sigma_{(\hat{p}_a - \hat{p}_b)}$. For this study, the maximum allowable standard deviation when $p_a = p_b = p = 0.5$ is used to test

$$H_0: p_a = p_b \text{ vs. } H_1: p_a > p_b ,$$

at the significance level $\alpha = 0.1$. A one-tailed test will be employed, because if a difference exists, we wish to detect $p_a > p_b$. Thus we will reject H_0 at $\alpha = 0.1$ if

$$\hat{p}_a - \hat{p}_b > z_{\alpha} \sqrt{\frac{2pq}{n}} = 1.28 \sqrt{\frac{0.5}{n}} ,$$

and conclude there exists a difference between p_a and p_b ; i.e., there is sufficient evidence to indicate that $\%[A]$ is higher than $\%[B]$. For the difference between $\%_s$'s, namely $(P_{[a,b]} - P_{[a',b']})$ for the comparison of the results from the $\binom{9}{2}$ pairs of agglomerative clustering algorithms, the same statistical test is applied.

Table 2 presents the results in terms of % for the comparison between the clusterings obtained by applying the nine agglomerative clustering algorithms and the population structure generated by $3*2*2$ settings of the parameters (ρ , δ , split). It should be noted that single linkage at $(.0, -.5)$ produces a smaller % than the other algorithms when $(\rho, \delta, \text{split})$ is fixed, except for the case when $\rho = .8$ and $\delta = 4.0$ with unequal sized clusters. However, single linkage is the only algorithm for which the number of local maxima at $k = 3$ increases if two variables are highly correlated for δ

= 4.0. Single linkage at (.0, -.5) performs better on the average for $\rho = 0.8$ than for any other values of ρ , as mentioned by DuBien and Warde (1987). In general, the retrieval abilities of all clustering algorithms except the algorithms lying along the line $\beta = 0.0$ are considered to be good for all settings of the structural parameters (ρ , δ , split) when the resultant clusterings are compared with the population structures generated. Specifically, the number of clusters with respect to the % across all settings (ρ , δ , split) is better predicted by employing the clustering algorithms defined by $\beta \leq -.25$ and $\pi \geq 0.0$ than any other clustering algorithms in (β, π) plane.

At this point, the changes of ρ across the structural parameters (δ , split) have little effect on predicting the number of clusters by using C_k with agglomerative clustering algorithms except single linkage. If the results are not significantly affected by the change in ρ , it is not necessary to observe the results for all settings of the correlation between the two variables. Therefore, the investigation on the behavior of C_k only for $\rho = 0.0$ will suffice for further study.

Tables 3-4 present the retrieval information for the nine agglomerative clustering algorithms by changing $k = 2, 3, \dots, 10$, in the form of $(\bar{C}_k, S_c, \%)$ for the two splits, 20-20-20 and 30-20-10, with fixed $\rho = 0.0$, $\delta = 4.0$. These results are graphically displayed in figures 4-5 in terms of \bar{C}_k across the number of clusters, $k = 2, 3, \dots, 10$, for the

nine agglomerative clustering algorithms. Regardless of the splits with fixed $\rho = 0.0$ and $\delta = 4.0$, the values \bar{C}_k at $k = 3$ are relatively large except for single linkage and thus the number of clusters is predicted correctly.

Tables 5-6 and figures 6-7 present the results when the distances among mean vectors are 6.0 for $\rho = 0.0$ with splits 20-20-20 and 30-20-10, respectively. All algorithms except for single linkage at $(.0, -.5)$ predict the number of clusters correctly giving local maximum values of \bar{C}_k at $k = 3$. It should be noted that the single linkage algorithm produces a uniformly larger S_c than the other algorithms.

Based on the results presented in tables 2-6 and figures 4-7, the use of the comparative statistic, C_k , is recommended in conjunction with the algorithms defined by $\beta \leq -.25$ and $\pi \geq 0.0$ in the (β, π) plane for all settings of the structural parameters $(\rho, \delta, \text{split})$.

Tables 7-8 represent the "agreement" between two clustering algorithms in each pair of $\binom{9}{2}$ possible pairs of the nine clustering algorithms relative to all settings $(\rho, \delta, \text{split})$ of the structural parameters. If the resultant clusterings produced by the pairs of agglomerative clustering algorithms are similar, we may have an indication of the natural grouping with any specific number of clusters, $k = K$, within the set of data. Based on comparison over the $\%_s$ of local maximum and the behavior of C_k , some conclusions and recommendations may be made on the use of C_k .

In Table 7, the $\%_s$ of local maxima at $k = 3$ increases as the correlation between the two variables increases from $\rho = .0$ to $\rho = .8$ when the clustering produced by the single linkage algorithm is compared with the clusterings produced by the other eight clustering algorithms for $\delta = 4.0$ and 20-20-20 split. Elsewhere, the $\%_s$'s of local maxima for all $\begin{pmatrix} 9 \\ 2 \end{pmatrix}$ pairs of clustering algorithms decrease, or at least remain constant as the correlation between the two variables increases. However, the $\bar{\%}_s$, which is the average of local maxima across all $\begin{pmatrix} 9 \\ 2 \end{pmatrix}$ possible pairs of clustering algorithms, increases from 39.4 to 45.4 for the 20-20-20 split while it decrease from 44.9 to 42.1 for the 30-20-10 split as the correlation increases from $\rho = 0.0$ to 0.8 with the distances among mean vectors fixed at $\delta = 4.0$. Also, in table 8 the $\bar{\%}_s$ decreases as the correlation increases regardless of the splits when $\delta = 6.0$.

As a general trend, the $\%_s$ of local maxima decreases as noise in the data increases, if the single linkage algorithm is not considered. The use of C_k to predict the number of clusters for data with highly correlated variables by applying single linkage algorithm in conjunction with the clustering algorithms defined by $\beta \leq -.25$ and $\pi \geq 0.0$ is recommended.

From now on, the single linkage algorithm will not be considered. We are interested in the general use of C_k rather than the extreme cases applying agglomerative clustering algorithms with squared Euclidean distance.

At this point, it should be mentioned that the investigation of the behavior of C_k for all $\binom{9}{2}$ possible pairs of agglomerative clustering algorithms is not necessary. We will concentrate on a smaller set of pairs of clustering algorithms which will suffice for the purpose of this study.

Since the patterns of the "agreement" between two clustering algorithms in each of the $\binom{9}{2}$ pairs for each structural setting of the parameters (ρ , split) are similar for increasing distances among mean vectors from $\delta = 4.0$ to $\delta = 6.0$, the distances among mean vectors will be fixed at $\delta = 4.0$. Also, the parameter ρ was used to simulate the effect of varying degrees of "noise" in the data on retrieval of the known structure; thus, the degree of information available in the data might be quantified by ρ . In fact, the $\%_s$'s of local maxima for all pairs of clustering algorithms decrease or remain stable on the average as the correlation increases for all settings (δ , split), if single linkage at $(.0, -.5)$ is not considered. Since the effect of correlation on the agreement between the agglomerative clustering algorithms is known (DuBien, 1976, and DuBien and Warde, 1987), it is reasonable to focus on the use of C_k in predicting the number of clusters imposed on the data after fixing the correlation at $\rho = 0.0$. The five pairs of algorithms for which the % retrieval of the true population is reasonably high for both algorithms (from table 2) and the averages of the agreement were largest

($\bar{\%}_s \geq 48.2$ from table 7) were subjectively chosen. In fact the pairs of algorithms were mainly selected if

$$P[(-.25, -.25), (-.5, .75)] - P[A, B] > 1.28 \sqrt{\frac{0.5}{n}},$$

where $n = 600$. Additionally, the pair $(.0, .5)$ vs. $(-.5, .75)$ was subjectively chosen. These are,

- (1) $(.0, .5)$ vs. $(-.5, .25)$,
- (2) $(-.25, -.25)$ vs. $(-.25, .5)$,
- (3) $(-.25, -.25)$ vs. $(-.5, .25)$,
- (4) $(-.25, -.25)$ vs. $(-.5, .75)$,
- (5) $(-.25, .0)$ vs. $(-.5, .75)$.

Hence, the use of C_k in determining the number of clusters within the set of data is investigated for the five specified pairs of agglomerative clustering algorithms across the splits. The value of \bar{C}_k is expected to be a local maximum at $k = 3$ if two clustering algorithms agree closely, since the structure of the clusters within the clusterings produced by each agglomerative clustering algorithm is expected to be very similar to the population structure.

Tables 9-10 present the behavior of \bar{C}_k , $k = 2, 3, \dots, 10$, in the form of $(\bar{C}_k, S_c, \%_s)$ with $\rho = 0.0$, $\delta = 4.0$, and two splits, 20-20-20 and 30-20-10 for the five specified pairs of algorithms. In addition, the agreement between two clustering algorithms for five pairs of algorithms is presented. To enhance the interpretation of the agreements, figures 8-9 portray the behavior of \bar{C}_k with respect to the

number of clusters, $k = 2, 3, \dots, 10$, for clusterings produced by the specified pairs of clustering algorithms. As shown in figures 8-9, the agreement between the resultant clusterings is not independent of the split. It should be noted that the 20-20-20 and the 30-20-10 splits were used to simulate data with equal sized cluster and unequal sized cluster, respectively. However, the size of the clusters is unknown in practice. With no prior information, the use of C_k is reasonable in conjunction with five pairs of agglomerative clustering algorithms chosen here regardless of the split. Specifically, it appears that the use of C_k with the pair $(-.5, .75)$ vs. $(.0, .5)$ is recommended to predict the number of clusters in the data set generated with $\delta = 4.0$, $\rho = 0.0$, irrespective of the split. In this case we might confirm that complete linkage at $(.0, .5)$ is at least one algorithm that works better with circular clusters than with stringy clusters.

The use of C_k with other pairs of clustering algorithms might be considered for the other settings of the structural parameters. With fixed $\delta = 4.0$, a general trend could be summarized as follows:

- 1). The pairs of single linkage at $(.0, -.5)$ with clustering algorithms defined with $\beta \leq -.25$ and $\pi \geq .0$ are recommended when ρ is close to 1.0 regardless of the split;
- 2). The pairs of algorithms $(.0, .5)$ vs. $(-.5, .75)$ and $(-.25, -.25)$ vs. $(-.5, .75)$ are better when ρ is

close to 0.0 for any split;

- 3). The pairs of algorithms $(-.25, .0)$ vs. $(-.5, .75)$ or $(-.25, -.25)$ vs. $(-.5, .75)$ are better for equal sized cluster, while the pairs of algorithms $(-.25, -.25)$ vs. $(-.5, .25)$ or $(-.25, -.25)$ vs. $(-.5, .75)$ are better for unequal sized cluster with any ρ ;
- 4). The pair of algorithms $(-.25, -.25)$ vs. $(-.5, .75)$ is recommended for all settings of the structural parameters (ρ, split) .

When the distance among mean vectors, δ , increases from 4.0 to 6.0, the agreement in clusterings increases across all settings of the structural parameters. However, the result is different from the result based on $\delta = 4.0$. That is,

- 5). The pairs of single linkage at $(.0, -.5)$ with other clustering algorithms defined by $\beta \leq 0.0$ and $\pi > 0.0$ in the (β, π) plane are not worse than the other pairs of algorithms for all the other settings of the structural parameters;
- 6). The pairs of average linkage at $(.0, .0)$ with other algorithms defined by $\beta \leq -.25$ and $\pi \geq .25$ in the (β, π) plane, or complete linkage at $(.0, .5)$ with $(-.5, .75)$ perform better when ρ is close to 0.0 regardless of the split;
- 7). When the distance among mean vectors is large, the pairs of algorithms $(.0, .0)$ vs. $(-.5, .75)$ or $(.0, .5)$ vs. $(-.5, .75)$ are recommended across all

settings of the structural parameters.

From the results obtained for $\delta = 4.0$ and 6.0 , the use of C_k in conjunction with the pairs of $(-.5, .75)$ with other clustering algorithms defined in the (β, π) plane performs better in predicting the number of clusters than other pairs of clustering algorithms across all settings of the structural parameters.

In the next chapter, the comparative study will be extended to the study on samples from multivariate lognormal distribution.

CHAPTER VI

EXTENSION TO MULTIVARIATE

LOGNORMAL SAMPLES

Fundamental Concepts

In the previous chapter the use of Rand's C_k was investigated to predict the number of clusters for data generated from multivariate normal distribution. However, the application of techniques developed on multivariate normal distributions is often limited.

In this chapter, the investigation of the use of Rand's C_k to determine the number of clusters by applying the agglomerative clustering algorithms chosen in the previous chapter is extended to a skewed distribution, the multivariate lognormal. At this point, it is necessary to obtain multivariate lognormal (MVN) data for the purpose of this study. Since a clustering method is used to find the natural structure present in data, the data structure generated should be reasonably well suited for the purpose of this study. The desire is to have multivariate lognormal data that has similar structure to that constructed for MVN in chapter V.

Let X_i be a random vector that follows $N_p(0, \Sigma)$ where

$X_i = [X_{i1}, X_{i2}, \dots, X_{ip}]'$ and set

$$Z_i = [Z_{i1}, Z_{i2}, \dots, Z_{ip}]'.$$

The transformation

$$Z_{ip} = m_i \exp(X_{ip}), \quad (6.1)$$

is applied to obtain a lognormal variate Z_{ip} having

$$E(Z_{ip}) = \xi_i = m_i \exp\left(\frac{\sigma_i^2}{2}\right),$$

$$\text{VAR}(Z_{ip}) = \lambda_i^2 = m_i^2 \exp(\sigma_i^2)(\exp(\sigma_i^2) - 1),$$

where $m_i, m_i > 0$, is the median.

Then the correlation ρ_{ij}^* between Z_i and Z_j with respect to the correlation ρ_{ij} in the $N_p(0, \Sigma)$ distribution is given by

$$\rho_{ij}^* = \frac{\exp(\rho_{ij}\sigma_i\sigma_j) - 1}{[\exp(\sigma_i^2) - 1]^{1/2} [\exp(\sigma_j^2) - 1]^{1/2}}.$$

Thus to obtain a specified correlation ρ_{ij}^* between Z_i and Z_j , the corresponding correlation ρ_{ij} is

$$\rho_{ij} = \frac{1}{\sigma_i\sigma_j} \ln\{1 + \rho_{ij}^*[\exp(\sigma_i^2) - 1]^{1/2} [\exp(\sigma_j^2) - 1]^{1/2}\}.$$

It is possible that particular ρ_{ij} 's violate $|\rho_{ij}| \leq 1$ or that the ρ_{ij} 's give a matrix Σ that is not positive definite (Johnson, 1987). In this study, the correlation ρ_{ij}^* is set to 0.0. Instead of investigating the effect of correlation (or, noise) between the two variables, the angle, Θ , used to set the spatial configuration of data points for each of the population median vectors was varied. In figure 3, the actual population median vectors and the angle, Θ , to set the equilateral triangle spatial configuration of population

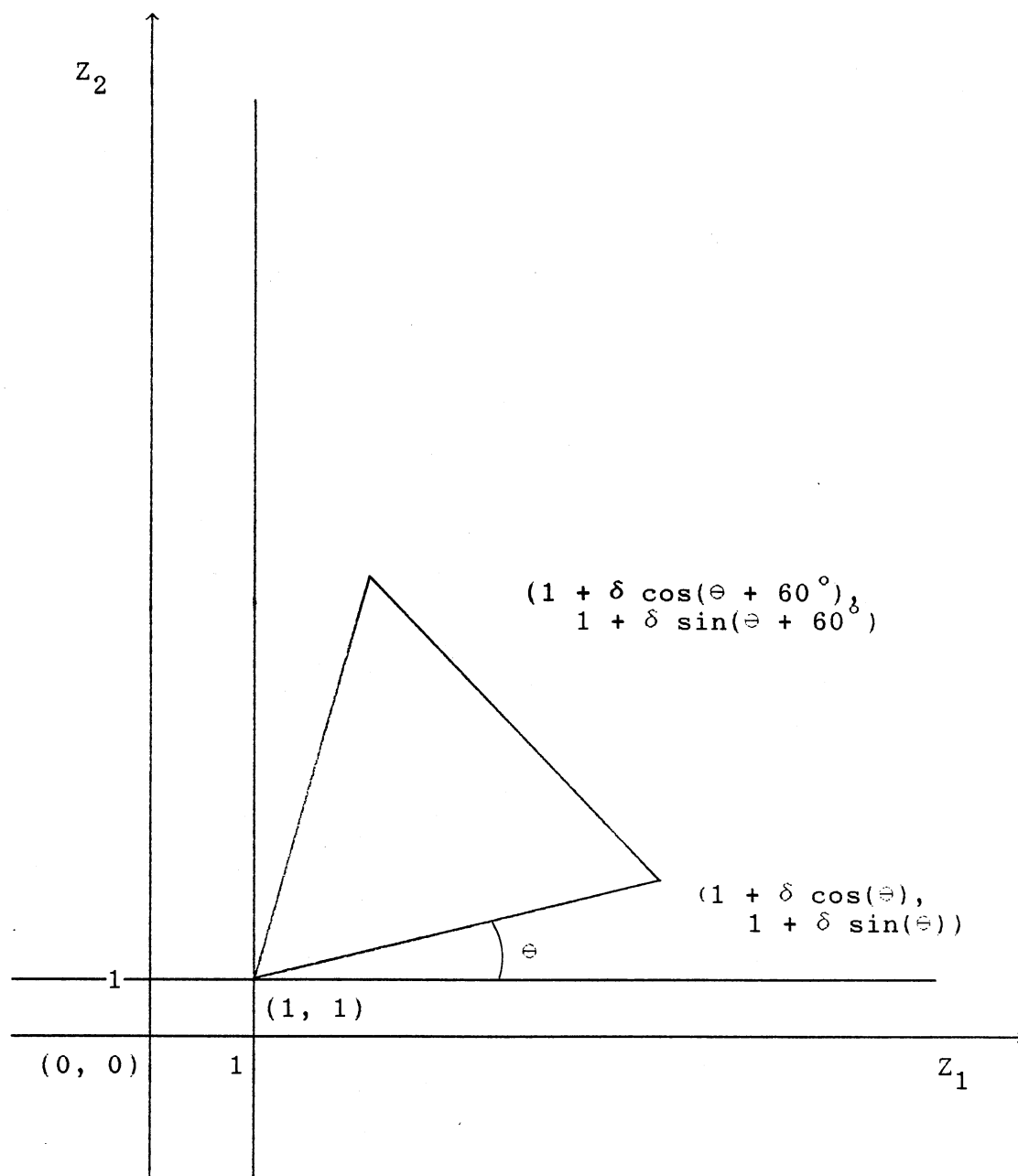


Figure 3. An Example of the Structural Framework Developed for BVLN.

median vectors used in this study are portrayed for $\delta = 4.0$. Difference in angle by rotating the equilateral triangle would be interpreted in terms of "noise" in the data structure generated from MVLN distribution since the shape of the data structure generated depends on the median vectors which are also dependent on the degree of rotation.

At this point, several structural parameters were considered for interpretation of the results of this study and applications of clustering methods. Some of the possible structural parameters on lognormal data are defined as follows:

1. N , the number of data points in Z ;
2. p , the number of variables defining each data point;
3. K , the number of populations from which the data points are generated;
4. m_k , $k = 1, 2, \dots, K$, the median vectors for each population of data points;
5. ψ_k , $k = 1, 2, \dots, K$, the variance-covariance structure for each population of data points;
6. δ , the distance between each pair of population median vectors;
7. The relative location of the population median vectors;
8. The split or n_k , $k = 1, 2, \dots, K$, the number of data points generated from each population of data points.

Since a similar data structure to that which was used for

the multivariate normal data is desired, the number of data points, the number of variables per data point, the number of MVLN populations, and the size of each population were fixed to be the same as in the MVN study. Thus this study is limited to bivariate lognormal distribution (BVLN) which could be extended to multivariate lognormal distribution (MVLN).

It should be mentioned that the mean vector, ξ , was considered to set the data points for each population with fixed median vector, m . However, a large number of the data points overlapped within the area below the fixed median vectors with skewed-right and long positive tail data regardless of ξ_i , where $\xi_i > m_i > 0$. Intuitively, the application of a clustering method was not reasonable even for large differences among the mean vectors. However, the use of the median vector to locate the data points for each population did not suffer from this problem.

Moreover, the variance depends on the median when σ^2 is fixed. The variance of Z_i increases rapidly as the median increases. A large portion of the data points which were generated with a large median always overlapped with another population generated with a small median because of the large difference in the variances. Even if the distance among the median vectors set for the different populations was large, the same type of data structure was obtained. At this point, a reasonable data structure for an application of clustering methods could not be obtained without

controlling the variance. The variance for a BVLN random variate Z_{ip} is

$$\lambda_i^2 = m_i^2 \exp(\sigma_i^2)(\exp(\sigma_i^2) - 1).$$

Let λ_i be 1.0 where the median m_i is specified for each population of data points. By solving the equation,

$$\sigma_i^2 + \ln[\exp(\sigma_i^2) - 1] + 2 \ln(m_i) = 0.0, \quad (6.2)$$

σ_i^2 was obtained to generate BVN with specified variance and hence a BVLN with variance 1.0 with specified median. Thus σ_i^2 decreases rapidly as the median increases. In addition, the shape of data structure generated for BVLN is close to normal (Johnson and Kotz, 1970) for any specified median if σ_i^2 is small, which in this study is a consequence of the choice of a large value for the median. Since the shapes of the distribution of the data points for each population differ from each other as a function of the median vectors, the size of the cluster (split) might effect the retrieval ability for unequal sized cluster.

Hence BVLN vectors for each population were generated by applying the transformation (6.1) to BVN vectors obtained from a population having a mean vector of zero with specified variance-covariance matrix by calling subroutine GGNSM in IMSL. Generation of other BVLN vectors with the same variance-covariance matrix might be accomplished by solving the equation (6.2) for fixed constant value of the median vector. Since the number of data points in each

population effect the retrieval ability of clustering algorithms, the number of data points is designated for each population generated at the median vectors shown in figure 3 as:

- 1). n_1 at $(1, 1)$,
- 2). n_2 at $(1 + \delta \cos(\theta), 1 + \delta \sin(\theta))$,
- 3). n_3 at $(1 + \delta \cos(\theta + 60), 1 + \delta \sin(\theta + 60))$.

The data structure for the comparative study on the use of C_k in this chapter may be outlined as follows:

$$Z_i \sim \text{BVLN}(m_k, \psi),$$

where $Z_i = [Z_{i1}, Z_{i2}, \dots, Z_{ip}]'$, $i = 1, 2, \dots, 60$,

with split into the $K = 3$ populations of

n_1 - n_2 - n_3 (i.e., 20-20-20, 30-20-10, 30-10-20, ..., 10-20-30);

: m_k , $k = 1, 2, 3$, is constrained by an equilateral triangle spatial configuration and the distance between each pair of population median vectors,

$$\delta = 4.0, 6.0;$$

$$: \psi = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix};$$

$$: \theta = 15^\circ, 30^\circ.$$

Using the sequence of steps explained in the previous chapter, Rand's C_k was computed on the data from BVLN by applying the five pairs of the agglomerative clustering algorithms chosen in chapter V in conjunction with squared Euclidean distance for all possible settings of the structural parameters $(\theta, \delta, \text{split})$. The results from

bivariate lognormal samples will be discussed in the following section.

Discussion on the Results from Multivariate Lognormal Samples

Tables 11-12 and figures 10-11 in the Appendix are provided to show the behavior of \bar{C}_k , $k = 2, 3, \dots, 10$, in the form of the measured statistic $(\bar{C}_k, S_c, \%_s)$ for the specific settings $(\Theta = 15^\circ, \delta = 4.0, \text{split})$ of the structural parameters over 100 replications for the five pairs of agglomerative clustering algorithms chosen in the previous chapter. Tables 13-14 present the agreement between two clustering algorithms in each pair of the $\begin{pmatrix} 9 \\ 2 \end{pmatrix}$ possible pairs of the nine agglomerative clustering algorithms. Additionally, table 15 provides the % retrieval of true population for the nine algorithms across all settings $(\Theta, \delta, \text{split})$ of the structural parameters.

As shown in table 11 and figure 10, Rand's C_k in conjunction with the specific pairs of agglomerative clustering algorithms is useful in predicting the number of clusters within the data generated from BVLN with $\Theta = 15^\circ$, $\delta = 4.0$, and 20-20-20 split. The local maxima occur at $k = 3$ in terms of \bar{C}_k for all pairs of algorithms except the pair $(.0, .5)$ vs. $(-.5, .75)$. Among the five pairs of algorithms, the use of C_k in conjunction with the pairs of algorithms $(-.5, .75)$ vs. $(-.25, .0)$ and $(-.5, .75)$ vs. $(-.25, -.25)$ are considered to be better than the others in

predicting the number of clusters within data generated with the structural setting $(15^\circ, 4.0, 20-20-20)$. Figure 10 portrays the behavior of the values \bar{C}_k , $k = 2, 3, \dots, 10$, within clusterings produced by the five pairs of agglomerative clustering algorithms.

Table 12 presents the behavior of \bar{C}_k , $k = 2, 3, \dots, 10$, in the form of $(\bar{C}_k, S_c, \%_s)$ for $\Theta = 15^\circ$, $\delta = 4.0$, and 30-20-10 split. The result is graphically displayed in figure 11 in terms of \bar{C}_k for the number of clusters within clusterings generated by the five pairs of agglomerative clustering algorithms. The result is similar to the result in table 11 and figure 10. The local maxima occur at $k = 3$ for all combinations of algorithms defined in the (β, π) plane. However, the use of C_k in conjunction with the pairs of algorithms $(-.5, .75)$ vs. $(.0, .5)$ and $(-.5, .75)$ vs. $(-.25, -.25)$ are considered to be better than the others in predicting the number of clusters in data set generated with the setting $(15^\circ, 4.0, 30-20-10)$. For another settings of the splits the behavior of \bar{C}_k was observed for five pairs of algorithms; however, the local maximum of \bar{C}_k at $k = 3$ was obtained regardless of various combinations of the split.

In addition, tables 13-14 provide the agreement for all possible $\binom{9}{2}$ pairs of the nine agglomerative clustering algorithms in the form of $\%_s$ for all settings of the structural parameters $(\Theta, \delta, \text{split})$. As with the MVN data, the "agreement" between agglomerative clustering algorithms is greatly affected by the changes in distance

among median vectors. The angle (or, noise) that changes the shape of the distribution of the data points generated affects the agreement of different agglomerative clustering algorithms differently with respect to the splits. With fixed δ , the agreement decreases as the angle increases from 15° to 30° for the 20-20-20 split, while the agreement varies with changes in the angle, Θ , for all unequal sized clusters through all pairs of clustering algorithms defined in the (β, π) plane. Thus, discussions of the results will be based on the effect of different splits ignoring the effect of angle differences.

Based on the results presented in table 13, the general trend for $\delta = 4.0$ with MVLN data is summarized as follows:

- 1). The pairs of single linkage at $(.0, -.5)$ with the other algorithms should not be used for samples from MVLN;
- 2). The pairs of $(-.5, .75)$ with clustering algorithms defined by $\beta = -.25$ in the (β, π) plane perform better with equal sized cluster;
- 3). The pairs of $(.0, .5)$ with the algorithms $(-.5, .75)$ and $(-.5, .25)$ are better with the 30-20-10 split;
- 4). The pairs of $(-.5, .75)$ with the algorithms $(-.5, .0)$, $(-.25, .0)$ and $(-.25, .5)$, and $(-.25, .0)$ vs. $(-.5, .25)$ are better with the 20-10-30 split;
- 5). The pairs of either $(-.5, .75)$ vs. $(-.25, .5)$ or $(-.25, .0)$ vs. $(-.25, .5)$ are better with the

20-30-10 split;

- 6). The pairs of $(.0, .5)$ with $(-.5, .0)$ and $(-.5, .25)$ or the pair $(-.5, .75)$ vs. $(-.25, -.25)$ are better with the 30-10-20 split;
- 7). The pairs of $(.0, .5)$ with clustering algorithms defined by $\beta < -.25$ and $\pi \geq 0.0$, and the pair of $(-.5, .75)$ vs. $(-.25, .0)$ are reasonable for the 10-20-30 split;
- 8). The pairs of $(.0, .5)$ with clustering algorithms defined by $\beta \leq -.25$ and $\pi > 0.0$, and the pairs of algorithms $(-.5, .75)$ vs. $(-.25, .0)$, $(-.5, .25)$ vs. $(-.25, -.25)$ and $(-.5, .25)$ vs. $(-.25, .5)$ are recommended for the 10-30-20 split;
- 9). The pairs of $(-.5, .75)$ with clustering algorithms defined by $\beta = -.25$ and $\pi \geq 0.0$ are better on the average across all settings of the structural parameters.

For fixed $\delta = 6.0$ in table 14, a brief discussion is given as follows:

- 10). The pairs of $(-.5, .75)$ with $(.0, .5)$, $(-.25, -.25)$, and $(-.25, .0)$ perform better with equal sized clusters;
- 11). For unequal sized clusters, the pairs of $(-.5, .75)$ with $(.0, .0)$, $(.0, .5)$ and $(-.25, -.25)$ cooperate well with C_k in predicting the number of clusters.

In general, the use of C_k with the pairs consisting of $(-.5,$

.75) with other clustering algorithms defined in the (β, π) plane except for single linkage at $(.0, -.5)$ are recommended for use in predicting the number of clusters within data generated for all settings of the structural parameters.

Moreover, the retrieval ability of the true population structure for nine agglomerative clustering algorithms in table 15 is considered. As with the results in table 2 for MVN data, the use of C_k with clustering algorithms defined by $\beta \leq -.25$ and $\pi \geq 0.0$ predict the number of clusters better than the other algorithms in the (β, π) plane.

General discussions of the results based on MVN and MVLN data will be given in the next chapter. However, we observe that C_k performs better with the pairs consisting of $(-.5, .75)$ with other clustering algorithms defined in the (β, π) plane than with combinations among the well-known clustering algorithms; i.e., single linkage, average linkage, complete linkage, and flexible linkage.

CHAPTER VII

GENERAL CONCLUSIONS AND POSSIBLE EXTENSIONS

The use of the comparative statistic, C_k , for predicting the number of clusters within a set of data applying agglomerative clustering algorithms is the objective of this research. This study was limited in its scope by controlling the structural parameters and choosing a finite number of agglomerative clustering methods. However, it is at least a basis for future comparative studies for determining the number of clusters by comparing the clusterings produced by clustering methods.

Observations and discussions from the comparative study on the use of C_k were made with respect to the agglomerative clustering algorithms defined by (β, π) and the settings of the structural parameters $(\rho, \delta, \text{split})$ for MVN data and $(\theta, \delta, \text{split})$ for MVLN data. Some general trends in the results were observable using the measured statistics $\%$ and $(\bar{C}_k, S_c, \%)$ for the "retrieval" ability on the true population for the nine agglomerative clustering algorithms. While the measured statistics $\%_s$ and $(\bar{C}_k, S_c, \%_s)$ were specified for the "agreement" between two clustering algorithms consisting of a pair for various structural parameters and in terms of

(β, π) which define the agglomerative clustering algorithms using squared Euclidean distance.

The values C_k are calculated by comparing the resultant clusterings produced by the nine agglomerative clustering algorithms with the population structure generated with specific settings of the structural parameters. The C_k has the local maximum if

$$C_k \geq C_{k-1} \quad \text{and} \quad C_k > C_{k+1} \quad \text{at } k = 3.$$

In the context of the % from the comparative study, the % is the number of times that a local maximum with respect to C_k occurs at $k = 3$ over 100 replications. A good prediction of the number of clusters for the data structure generated is the one that has a high value of % and remains stable for various settings of the structural parameters.

Let $\%[A]$ be a % value produced by algorithm A over n replications. Algorithm A will be termed "better" with respect to % than algorithm B iff

$$\forall \rho \text{ (or, } \ominus), \quad \%[A] > \%[B],$$

where the pair (δ, split) is fixed.

If any statistical criterion is required, the hypothesis that a difference exists between %'s can be tested. Since $\%[A]$ and $\%[B]$ are the number of times that a local maximum occurs at $k = 3$ over 100 replications, $\%[A]$ and $\%[B]$ follow a binomial probability distribution, with parameters p_a and p_b , respectively. For large samples the point estimator of $(p_a - p_b)$, namely $(\hat{p}_a - \hat{p}_b)$ is approximately normally

distributed with a mean of $(p_a - p_b)$ and a standard deviation of

$$\sigma(\hat{p}_a - \hat{p}_b) = \sqrt{\frac{p_a q_a}{n_a} + \frac{p_b q_b}{n_b}}.$$

Then

$$z = \frac{(\hat{p}_a - \hat{p}_b) - (p_a - p_b)}{\sigma(\hat{p}_a - \hat{p}_b)}$$

follows a standard normal distribution. Hence z can be used as a test statistic to test

$$H_0: p_a = p_b \quad \text{vs.} \quad H_1: p_a > p_b, \quad \alpha = 0.1,$$

and reject the null hypothesis in favor of the alternative hypothesis if

$$(\hat{p}_a - \hat{p}_b) > z_\alpha \sigma(\hat{p}_a - \hat{p}_b).$$

In fact, a one-tailed test is employed to detect $p_a > p_b$.

And $p_a = p_b = .5$ is used to obtain the maximum standard deviation, instead of using the best estimates of p_a and p_b to compute $\sigma(\hat{p}_a - \hat{p}_b)$. Thus, if

$$(\hat{p}_a - \hat{p}_b) > 1.28 \sqrt{\frac{0.5}{n}},$$

where n is the number of replications, then the use of C_k with algorithm A is preferred in predicting the number of clusters to the use of C_k with algorithm B for the specified structural settings. With given settings for the structural parameters and a metric of Euclidean distance, some general observations with respect to the structural parameters and the agglomerative clustering algorithms with (β, π) will be

offered with respect to the % for MVN data.

The single linkage algorithm, which is the only space-contracting algorithm included in the comparative study, was different from all of the other algorithms with respect to % for all settings (ρ , δ , split) used in this study. As discussed by DuBien and Warde (1987), the single linkage algorithm was the worst algorithm. An exception was when ρ was close to 1.0 for the 30-20-10 split. However, the single linkage algorithm was the only algorithm for which high ρ had a marked effect on its performance even if the number of local maxima for ρ close to 0.5 is smaller than the others. The performance of the single linkage algorithm improves as ρ increases for fixed distance, δ , among mean vectors and fixed splits. The observations concerning single linkage algorithm imply that space-contracting algorithms are worse at "retrieving" the population structure than either space-conserving or space-dilating algorithms when MVN data and squared Euclidean distance are employed.

The average linkage algorithm which is the only space-conserving algorithm, and the complete linkage algorithm which is one of the space-dilating algorithms, perform worse when ρ is close to 1.0 than when ρ is close to 0.0, regardless of the size of cluster (split) for fixed distance among mean vectors δ .

For any other agglomerative clustering algorithms except the algorithms with $\beta = 0.0$ in the (β, π) plane, the number of clusters for the population structure generated

was well predicted by C_k for all settings of the structural parameters $(\rho, \delta, \text{split})$.

The statistics $(\bar{C}_k, S_c, \%)$ from the comparative study were investigated for the behavior of C_k for specified settings $(\rho, \delta, \text{split})$. However, it was not necessary to observe the results for all possible settings of $(\rho, \delta, \text{split})$. The changes in ρ from 0.0 to 1.0 had little effect on the changes in the %'s of local maxima on the settings of (δ, split) for the nine agglomerative clustering algorithms except for single linkage algorithm. Thus, only the observations concerning the behavior of C_k for $\rho = 0.0$ with various settings (δ, split) of the structural parameters was provided in this study.

For a setting of (δ, split) , $\bar{C}_k[A]$ shall denote a \bar{C}_k value produced by algorithm A when the clustering is compared to the true population; $S_c[A]$ shall denote an S_c value produced by algorithm B; $\%[A]$ shall denote a % value produced by algorithm A over 100 replications. In predicting the number of clusters, algorithm A will work better with respect to \bar{C}_k, S_c , and % than algorithm B iff

- 1). $\%[A] > \%[B]$;
- 2). $\bar{C}_k[A] > \bar{C}_k[B]$ and

$$\exists \bar{C}_k[A] - \bar{C}_{k-1}[A] > \bar{C}_k[B] - \bar{C}_{k-1}[B]$$
 and $\bar{C}_k[A] - \bar{C}_{k+1}[A] > \bar{C}_k[B] - \bar{C}_{k+1}[B]$;
- 3). $S_c[A] \leq S_c[B]$,
 where \bar{C}_k is the local maximum at $k = 3$ and (δ, split) is fixed.

Thus, the performance of \bar{C}_k in predicting the number of clusters by applying the agglomerative clustering algorithms was investigated with respect to the measured statistics $(\bar{C}_k, S_c, \%)$ for all settings (δ, split) with fixed $\rho = 0.0$ for MVN data.

At this point, the use of C_k in predicting the number of clusters was considered by comparing the clusterings produced by the nine agglomerative clustering algorithms. If the clusterings agree closely, we may have more confidence in predicting the number of clusters by observing the comparative statistics C_k . The number of local maxima at $k = 3$ with respect to $\%_s$ was used to determine the performance of C_k in conjunction with the possible pairs of clustering algorithms for the settings $(\rho, \delta, \text{split})$. In fact, the $\%_s$ is the agreement between two clustering algorithms consisting of a pair. Hence in predicting the number of clusters by using C_k for the settings $(\rho, \delta, \text{split})$, the pair of clustering algorithms, A and B, that agree more closely with respect to $\%_s$ in terms of the clusterings than the pair clustering algorithms, A' and B', were chosen for further study iff

$$1). \%_s[A, B] > \%_s[A', B'],$$

where $\%_s[A, B]$ is the $\%_s$ of local maxima obtained for paired algorithms A and B;

$$2). \%[A], \%[B] \text{ are considered large for the settings } (\rho, \delta, \text{split}).$$

If any statistical criterion is required to test the

difference between %_s's, namely $p_{[a,b]}$ and $p_{[a',b']}$ for the comparison of the results from the possible $\binom{9}{2}$ pairs of agglomerative clustering algorithms, the same statistical test explained previously can be applied.

A few general observations with respect to the settings $(\rho, \delta, \text{split})$ can also be made. In general, as ρ increases, the %_s of correctly predicting $k = 3$ by C_k increases across $(\rho, \delta, \text{split})$; this observation is natural since the clusters become more distinct as the the population means move further apart. The changes in ρ give marked effects on the agreements among the clusterings produced by the clustering algorithms paired with the single linkage algorithm. As ρ increases from 0.0 to 1.0 for all settings (δ, split) of the structural parameters, the %_s increases rapidly for the clustering algorithms paired with single linkage algorithm. However, the %_s decreases or at least remains stable for the clustering algorithms paired with other algorithms when ρ increases from 0.0 to 1.0. Increasing δ from 4.0 to 6.0 causes an increase in %_s for all settings (ρ, split) . The two different splits with respect to %_s have little effect on the prediction of the number of clusters. Overall, ρ does not greatly affect the agreement between the agglomerative clustering algorithms with respect to %_s for the two splits with the effect becoming less for increasing δ , whenever single linkage algorithm is not considered.

For $\binom{9}{2}$ possible pairs of agglomerative clustering

algorithms on the settings $(\rho, \delta, \text{split})$ with MVN and $(\theta, \delta, \text{split})$ with MVLN data, the number of clusters present is predicted better with respect to $(\bar{C}_k, S_c, \%_s)$ for the pair of agglomerative clustering algorithms, A and B, than the pair of clustering algorithms, A' and B', iff

- 1). $\%_s[A, B] > \%_s[A', B']$;
- 2). $\bar{C}_k[A, B] > \bar{C}_k[A', B']$ and
 $\exists \bar{C}_k[A, B] - \bar{C}_{k-1}[A, B] > \bar{C}_k[A', B'] - \bar{C}_{k-1}[A', B']$ and
 $\bar{C}_k[A, B] - \bar{C}_{k+1}[A, B] > \bar{C}_k[A', B'] - \bar{C}_{k+1}[A', B']$;
- 3). $S_c[A, B] \leq S_c[A', B']$,
 where \bar{C}_k is the local maximum at $k = 3$
 and $(\rho, \delta, \text{split}), (\theta, \delta, \text{split})$ are fixed.

In terms of $\%_s$ for the specific settings of the structural parameters, the number of clusters is relatively well predicted by using C_k with the pair of clustering algorithms defined by $(-.5, .75)$ vs. $(.0, .5)$ for MVN data generated by the settings $(0.0, 4.0, \text{split})$. With MVLN data for settings $(15^\circ, 4.0, \text{split})$, the use of C_k with the pair of algorithms $(-.5, .75)$ vs. $(-.25, .0)$ for the 20-20-20 split and the pair of algorithms $(-.5, .75)$ vs. $(.0, .5)$ for the 30-20-10 split well predict the number of clusters giving a local maximum at $k = 3$. For the other specific settings of the structural parameters, the best pair of clustering algorithms might be found; i.e., the combinations of single linkage at $(.0, -.5)$ with other algorithms defined by $\beta \leq -.25, \pi \geq 0.0$ in the (β, π) plane when ρ is close to

1.0, $\delta = 4.0$ and 20-20-20 split for MVN data.

At this point, investigation on the general use of C_k with clustering algorithms when any prior information is unknown for given set of data was our objective. It was necessary to choose several pairs of clustering algorithms that cooperate with the comparative statistic, C_k , indicating the number of clusters $k = 3$ across all settings of the structural parameters.

In general, the pairs with $(-.5, .75)$ in the (β, π) plane performed better with respect to C_k than the other pairs of clustering algorithms across all settings of the structural parameters for both MVN and MVLN data. Also it was known that single linkage at $(.0, -.5)$ was the worst algorithm with MVLN, while recommended only when ρ is close to 1.0 with MVN data. And the pair of algorithms $(-.5, .75)$ vs. $(-.5, .25)$ is not considered because of the poor agreement between them through all settings of the structural parameters with both MVN and MVLN data. Moreover, the pairs of algorithms $(-.5, .75)$ with $(-.25, .5)$ and $(-.5, .0)$ in the (β, π) plane cooperated with the C_k only for MVLN. Thus these pairs will not be considered for further investigation on the general use of C_k .

Based on the $\%_s$ across all settings of the structural parameters for all possible $\begin{pmatrix} 9 \\ 2 \end{pmatrix}$ pairs of clustering algorithms, only four pairs of algorithms are subjectively chosen from table 7-8 and table 13-14. The four pairs of algorithms are,

- (1) $(-.5, .75)$ vs. $(-.25, -.25)$,
- (2) $(-.5, .75)$ vs. $(-.25, .0)$,
- (3) $(-.5, .75)$ vs. $(.0, .0)$,
- (4) $(-.5, .75)$ vs. $(.0, .5)$.

The results based on these four pairs of clustering algorithms are summarized in table 16.

As shown in table 16, the use of C_k with these four pairs of clustering algorithms defined in the (β, π) plane predict the number of clusters present in MVN and MVLN data generated with all settings of the structural parameters reasonably well. Some pairs of clustering algorithms indicate the number of clusters better than the others for specific settings of the structural parameters. However, we could recommend the use of C_k in conjunction with the pairs of clustering algorithms in predicting the number of clusters within set of data as follows:

- 1). With MVN data the pairs $(-.5, .75)$ vs. $(-.25, -.25)$ or $(-.5, .75)$ vs. $(-.25, .0)$ are recommended to be used with C_k for $\delta = 4.0$, while the pairs $(-.5, .75)$ vs. $(.0, .0)$ or $(-.5, .75)$ vs. $(.0, .5)$ are reasonable for $\delta = 6.0$;
- 2). With MVLN data the pairs $(-.5, .75)$ vs $(-.25, .0)$ or $(-.5, .75)$ vs. $(.0, .5)$ are better for all settings of the structural parameters (Θ, split) for $\delta = 4.0$, while the pairs $(-.5, .75)$ vs. $(.0, .0)$ or $(-.5, .75)$ vs. $(.0, .5)$ are recommended for $\delta = 6.0$.

Moreover, the % retrieval of the true population generated with the specific structural parameter for each clustering algorithm was considered from table 2 and table 15 for MVN and MVLN data, respectively. If both algorithms combined as a pair have high retrieval abilities for the true population, we will consider the pair to be the best among four pairs of algorithms for both MVN and MVLN data. In this way the structure of clusterings produced by the pair of clustering algorithms is also similar to the data structure generated. Thus, we conclude from the results of the comparative study that:

- 3). The use of C_k in predicting the number of clusters is recommended with the pair of algorithms, $(-.5, .75)$ vs. $(-.25, .0)$, defined in the (β, π) plane regardless of the characteristics of the given set of data.

This confirms that the flexible strategy at $(-.25, .0)$ recommended by DuBien and Warde (1987) is at least one algorithm for finding the unknown structure present in many data sets. Moreover, the pair of algorithms $(-.5, .75)$ vs. $(-.25, .0)$ generally performs better than any combinations of single, complete, and average linkage regardless of the degree of noise and the relative sizes of the clusters present in the data.

There are a lot of possible extensions for the comparative study on the use of C_k by applying agglomerative clustering algorithms formed with squared Euclidean

distance. In future comparative study on the behaviors of C_k with agglomerative clustering algorithms, at least a limited comparative investigation of the effect of correlated variables on the value of C_k should be attempted when $p \geq 3$ with a large value of N . The study on the behavior of C_k for various changes in split (the size of cluster) may be desired. Also it should include a large number of replications at each setting of the structural parameters. The populations of data points could be generated from probability distributions other than MVN and MVLN probability distributions, but the choices of the MVN and MVLN data structures for each of the populations seem reasonable. However, it would be worth attempting a limited comparative investigation on the use of C_k with clustering algorithms when each of MVN and MVLN populations of data points presented in X and Z , respectively, has a different variance-covariance matrix. Moreover, it would be desirable to investigate the use of C_k with agglomerative clustering algorithms for the data from the mixture of different distributional forms in multivariate settings of variables.

A great deal of flexibility in a limited extension of the comparative study on the use of C_k could be achieved by choosing different agglomerative clustering algorithms to pair with the agglomerative clustering algorithm $(-.5, .75)$ defined in the (β, π) plane. Since the use of C_k with the pairs of $(-.5, .75)$ with other clustering algorithms predicted the number clusters fairly well.

In conclusion, it appears from all the evidence on its performance that the C_k statistic used in conjunction with specified agglomerative clustering algorithms with squared Euclidean distance is a useful comparative statistic on determining the number of clusters present in the data. However, the performance of C_k is dependent on the characteristics of the data, the choices of agglomerative clustering algorithms and distance measures. Therefore, the results on the use of C_k should be examined critically to make sure they are meaningful.

BIBLIOGRAPHY

- Anderberg, Michael R. (1973). Cluster Analysis for Applications. New York: Academic Press.
- Baker, F. B. and Hubert, L. J. (1975). Measuring the Power of Hierarchical Cluster Analysis. JASA, 70, 31-38.
- Binder, D. A. (1978). Bayesian Cluster Analysis. Biometrika, 65, 31-38.
- Calinski, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. Commun. in Statistics, 3, 1-27.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37-46.
- Cormack, R. M. (1971). A review of Classification. Journal of the Royal Statistical Society, 134, 321-367.
- DuBien, J. L. (1976). Comparative Techniques for the Evaluation of Clustering Methods. (Unpub. Ph.D. thesis, Oklahoma State University.)
- DuBien, J. L. and Warde, W. D. (1979). A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms. Canadian Journal of Statistics, 7, 29-38.
- DuBien, J. L. and Warde, W.D. (1982). Some Distributional Results Concerning a Comparative Statistic used in Cluster Analysis. Paper Presented at the ASA Meeting, Detroit.
- DuBien, J. L. and Warde, W. D. (1987). A Comparison of Agglomerative Clustering Methods with respect to Noise. Commun. Statist. Theory Method, 16, 1433-1460.
- Duran, Benjamin S. and Odell, Patrick L. (1974). Cluster Analysis; A Survey. In econometrics (100). Managing Editors M. Beckman and H. P. Kunzi. Lecture Notes in Econ.

- Edelbrock C. (1979). Mixture Model Tests of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody. Multivariate Behavioral Research, 14, 367-384.
- Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). A Method for Cluster Analysis. Biometrics, 21, 362-376.
- Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics.
- Engelman, L. and Hartigan, J. A. (1969). Percentage Points of a Test for Clusters. JASA, 64, 1647-1649.
- Everitt, Brian. (1974). Cluster Analysis. New York: Halsted Press, Division of John Wiley & Sons.
- Everitt, Brian. (1979). Unresolved Problems in Cluster Analysis. Biometrics, 35, 169-182.
- Everitt, Brian. (1981). A Monte Carlo Investigation of the Likelihood Ratio Test for the Number of Components in a Mixture of Normal Distributions. Multivariate Behavioral Research, 16, 171-180.
- Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. JASA, 78, 553-584.
- Friedman, H. P. and Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. JASA, 62, 1159-1179.
- Goodman, L.A. and Kruskal, W. H. (1954). Measures of Association for Cross Classifications. JASA, 49, 732-764.
- Gordon, A. D. (1981). Classification. Methods for the Exploratory Analysis of Multivariate Data. New York: Chapman and Hall.
- Hartigan, J. A. (1975). Clustering Algorithms. New York: John Wiley & Sons.
- Hill, R. S. (1980). A stopping Rule for Partitioning Dendrograms. Botanical Gazette, 141, 321-324.
- Hubert, L. J. and Levin, J. R. (1976). Evaluating Object Set Partitions: Free-Sort Analysis and Some Generalizations. Journal of Verbal learning and Behavior, 15, 459-470.

- Johnson, Mark E. (1987). Multivariate Statistical Simulation. New York: John Wiley & Sons.
- Johnson, N. L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions 1, New York: John Wiley & Sons.
- Johnson, Stephen C. (1967). Hierarchical Clustering Scheme. Psychometrika, 32, 241-255.
- Krzanowski, W. J. and Lai, Y. T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. Biometrics, 44, 23-34.
- Lance, G. N. and Williams, W. T. (1966). A Generalized Sorting Strategy for Computer Classification. Nature, 212, 218.
- Lance, G. N. and Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems. The Computer Journal, 9, 373-380.
- Lee, K. L. (1979). Multivariate Tests for Clusters. JASA, 74, 708-714.
- Ling, R. F. (1973). A Probability Theory of Cluster Analysis. JASA, 68, 159-164.
- Marriott, F. H. C. (1971). Practical Problems in a Method of Cluster Analysis. Biometrics, 27, 501-514.
- McLachlan, G. J. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. Appl. Statist., 36, 318-324.
- Milligan, G. W. (1981). A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. Psychometrika, 46, 187-199.
- Milligan, G. W. and Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika, 50, 159-179.
- Morey, L. C. and Agresti, A. (1984). The Measurement of Classification Agreement: An Adjustment to the Rand Statistic for Chance Agreement. Educational and Psychological Measurement, 44, 33-37.
- Mrachek, Roger J. (1972). Some statistical Aspects of Clustering Procedure. (Unpub. M. S. thesis, Iowa State University.)

- Norton, James Michael. (1975). Some Statistical Procedure to Aid in the Evaluation of a Cluster Analysis. (Unpub. Ph.D. thesis, Oklahoma State University.)
- Peck, Roger Wayne. (1983). Confidence Bounds for the Number of Clusters in Cluster Analysis. (Unpub. Ph.D. thesis, University of Texas at Dallas.)
- Rand, William Medden. (1969). The Development of Object Criteria for Evaluating Clustering Methods. (Unpub. Ph.D. thesis, University of California at Los Angeles.)
- Rand, William Medden. (1971). Objective Criteria for the Evaluation of Clustering Methods. JASA, 66, 846-850.
- Ratkowsky, D. A. and Lance, G. N. (1978). A Criterion for Determining the Number of Groups in a Classification. Australian Computer Journal, 10, 115-117.
- Ratkowsky, D. A. (1984). A Stopping Rule and Clustering Method of Wide Applicability. Botanical Gazette, 145, 518-523.
- Scheibler, D. and Schneider, W. (1985). Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms: A Comparison of Hierarchical and Nonhierarchical Methods. Multivariate Behavioral Research, 20, 283-304.
- Scott, A. J. and Symons, Michael J. (1971). Clustering Methods Based on Likelihood Ratio Criteria. Biometrics, 27, 387-397.
- Silverman, B. W. (1978). Choosing the Window Width When Estimating a Density. Biometrika, 65, 1-11.
- Silverman, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. JRSS, B, 43, 97-99.
- Sneath, P. H. A. (1977). A Method for Testing the Distinctness of Clusters: A Test of the Disjunction of Two Clusters in Euclidean Space as Measured by Their Overlap. Mathematical Geology, 9, 123-143.
- Sneath, Peter H.A. and Sokal, R. R. (1973). Numerical Taxonomy. San Francisco: W. H. Freeman & Co.
- Sokal, R. R. and Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. Univ. Kansas Sci. Bull., 38, 1409-1438.
- Symons, M. J. (1981). Clustering Criteria and Multivariate Normal Mixtures. Biometrics, 37, 35-43.

- Thorndike, Robert L. (1953). Who Belongs in Family?. Psychometrika, 18, 267-276.
- Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Ann. Math. Stat., 9, 60-62.
- Wishart, David. (1969). Mode Analysis: A Generalization of Nearest Neighbour which Reduce Chaining Effects. In Numerical Taxonomy. Editor A. J. Cole. New York: Academic Press, 282-308.
- Wolfe, J. H. (1970). Pattern Clustering by Multivariate Mixture Analysis. Multivariate Behavioral Research, 5, 329-350.
- Wong, M. A. (1985). Bootstrap Testing Procedure for Investigating the Number of Subpopulations. J. Statist. Comput. Simul., 22, 99-112.
- Wong, M. A. and Lane, T. (1981). A k-th Nearest Neighbour Clustering Procedure. Proceedings of the 13th Interface Symposium on Statistics and Computer Science, W. F. Eddy(Ed.). Berlin: Springer-Verlag, 308-311.
- Wong, M. A. and Schaack, C. (1982). Using the k-th Nearest Neighbour Clustering Procedure to Determining the Number of Subpopulations. Proceedings of the Statistical Computing Section. American Statistical Association, 40-48.

APPENDIX

TABLE 2
PERCENT RETRIEVAL OF TRUE POPULATION FOR
ALL ALGORITHMS WITH MVN

(β, π)	split	20-20-20			30-20-10			$\bar{\%}$	$S_{\bar{\%}}$
	$\delta \rho$.0	.4	.8	.0	.4	.8		
(.0 , -.5)	4.0	13	9	36	19	14	62	25.5	8.26
	6.0	67	74	73	64	71	69	69.7	1.54
(.0 , .0)	4.0	63	68	59	72	65	50	62.8	3.13
	6.0	87	87	86	83	88	87	86.3	0.71
(.0 , .5)	4.0	72	73	56	76	73	50	66.7	4.42
	6.0	93	92	91	91	94	90	91.8	0.60
(-.25, -.25)	4.0	77	74	73	81	77	72	75.7	1.36
	6.0	96	94	92	95	95	88	93.3	1.20
(-.25, .0)	4.0	81	81	81	90	89	75	82.8	2.32
	6.0	98	99	98	96	96	93	96.7	0.88
(-.25, .5)	4.0	83	88	89	85	81	84	85.0	1.24
	6.0	93	94	98	96	97	95	95.5	0.76
(-.5 , .0)	4.0	85	84	91	84	86	69	83.2	3.03
	6.0	100	98	97	96	97	95	97.2	0.70
(-.5 , .25)	4.0	88	87	89	83	83	78	84.7	1.69
	6.0	100	100	100	96	99	97	98.7	0.71
(-.5 , .75)	4.0	79	98	88	78	83	72	83.0	3.71
	6.0	99	100	99	95	95	95	97.2	0.98
$\bar{\%}$	4.0	71.2	73.6	73.6	74.2	72.3	68.0	72.1	
	6.0	92.6	93.1	92.7	90.2	92.4	89.9	91.8	

TABLE 3

RETRIEVAL INFORMATION OF NINE ALGORITHMS VS. POPN. WITH
 $\delta = 4.0$, $\rho = 0.0$, AND 20-20-20 SPLIT FOR MVN

k		$\begin{pmatrix} .0 \\ .5 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .0 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .75 \end{pmatrix}$
2	\bar{C}_k	.3869	.6614	.6866	.7026	.7188	.7107	.7249	.7222	.7049
	S_c	.1345	.1131	.0813	.0606	.0431	.0551	.0359	.0380	.0565
3	\bar{C}_k	.4641	.8518	.8680	.8828	.8982	.8991	.8990	.8989	.8895
	S_c	.1734	.0949	.0852	.0711	.0535	.0551	.0552	.0517	.0610
4	\bar{C}_k	.5655	.8600	.8632	.8725	.8775	.8771	.8742	.8749	.8691
	S_c	.1869	.0625	.0544	.0512	.0388	.0395	.0388	.0372	.0401
5	\bar{C}_k	.6444	.8507	.8471	.8520	.8506	.8489	.8417	.8416	.8378
	S_c	.1896	.0465	.0415	.0374	.0326	.0323	.0318	.0294	.0300
6	\bar{C}_k	.6908	.8325	.8270	.8282	.8242	.8216	.8155	.8149	.8131
	S_c	.1756	.0392	.0331	.0333	.0278	.0265	.0249	.0243	.0243
7	\bar{C}_k	.7324	.8169	.8109	.8113	.8027	.7986	.7955	.7928	.7905
	S_c	.1579	.0333	.0270	.0276	.0226	.0222	.0200	.0196	.0210
8	\bar{C}_k	.7518	.8028	.7966	.7952	.7882	.7830	.7815	.7784	.7777
	S_c	.1484	.0274	.0222	.0252	.0179	.0178	.0156	.0158	.0162
9	\bar{C}_k	.7858	.7917	.7828	.7825	.7766	.7715	.7692	.7673	.7668
	S_c	.1173	.0245	.0202	.0213	.0175	.0150	.0143	.0148	.0127
10	\bar{C}_k	.7941	.7840	.7736	.7732	.7666	.7612	.7594	.7581	.7572
	S_c	.0987	.0210	.0183	.0186	.0162	.0122	.0130	.0127	.0118
%		13	63	72	77	81	83	85	88	79

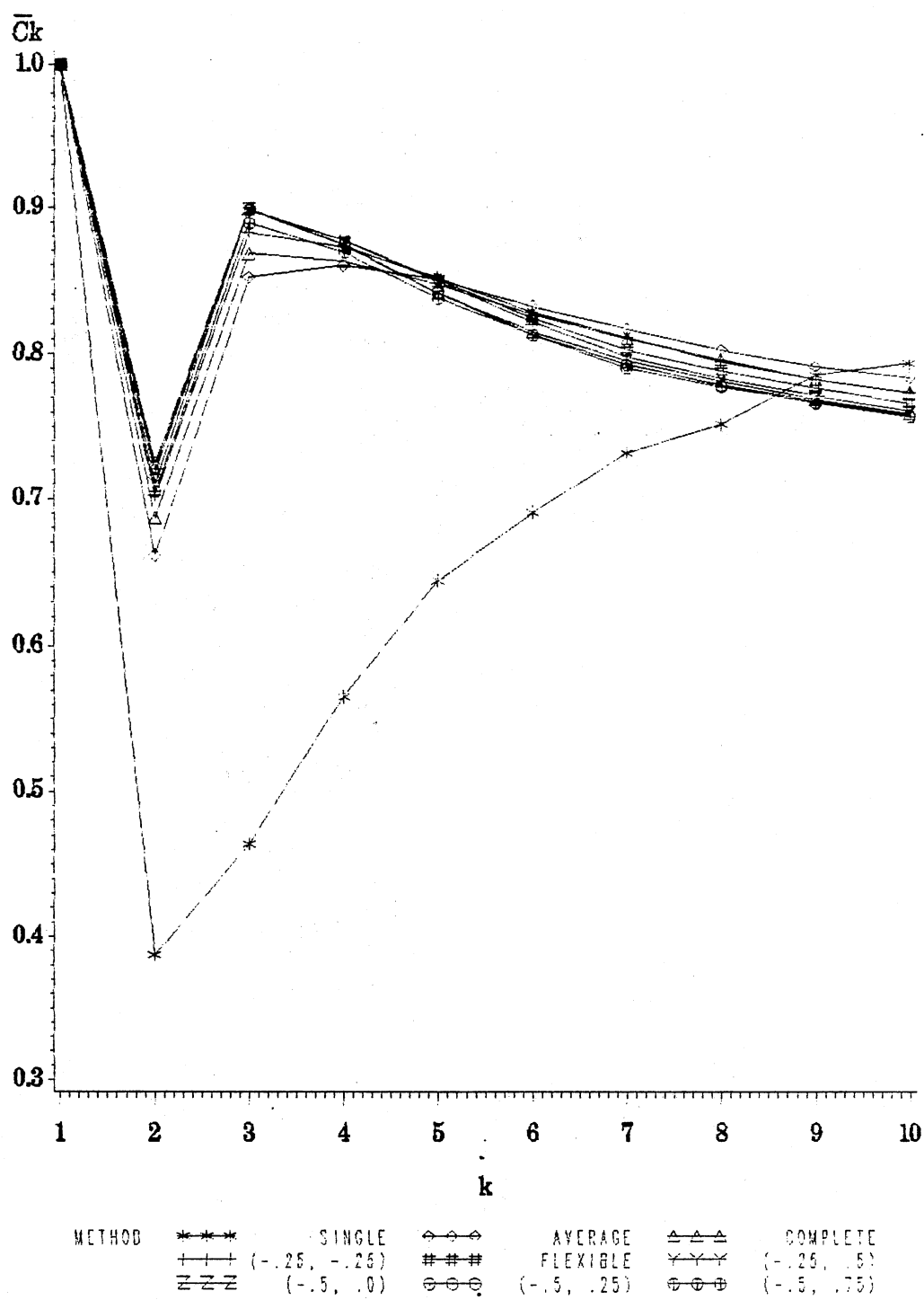


Figure 4. Retrieval result of the nine algorithms for MVN with $\delta = 4.0$, $\rho = .0$ and 20-20-20 split

TABLE 4

RETRIEVAL INFORMATION OF NINE ALGORITHMS VS. POPN. WITH
 $\delta = 4.0$, $\rho = 0.0$, AND 30-20-10 SPLIT FOR MVN

k		$\begin{pmatrix} .0 \\ -.5 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .0 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .75 \end{pmatrix}$
2	\bar{C}_k	.4575	.7215	.7417	.7703	.8008	.8002	.8002	.7984	.7896
	S_c	.1248	.1155	.1050	.0879	.0606	.0616	.0599	.0604	.0628
3	\bar{C}_k	.5439	.8744	.8838	.8898	.9098	.9059	.9016	.9032	.8840
	S_c	.1707	.0925	.0907	.0725	.0559	.0554	.0589	.0543	.0756
4	\bar{C}_k	.6282	.8562	.8504	.8400	.8438	.8400	.8353	.8371	.8253
	S_c	.1833	.0746	.0613	.0581	.0440	.0447	.0401	.0405	.0461
5	\bar{C}_k	.6953	.8247	.8184	.8080	.8022	.7934	.7893	.7904	.7857
	S_c	.1751	.0630	.0470	.0425	.0326	.0315	.0271	.0274	.0289
6	\bar{C}_k	.7401	.8008	.7840	.7825	.7733	.7643	.7620	.7593	.7580
	S_c	.1599	.0476	.0379	.0368	.0306	.0241	.0239	.0237	.0240
7	\bar{C}_k	.7706	.7823	.7657	.7607	.7563	.7467	.7424	.7411	.7396
	S_c	.1417	.0417	.0323	.0294	.0289	.0232	.0203	.0186	.0190
8	\bar{C}_k	.7930	.7678	.7473	.7475	.7387	.7312	.7261	.7245	.7238
	S_c	.1276	.0359	.0248	.0241	.0227	.0198	.0165	.0168	.0155
9	\bar{C}_k	.8042	.7524	.7338	.7340	.7252	.7185	.7143	.7139	.7100
	S_c	.1138	.0305	.0220	.0209	.0190	.0159	.0142	.0142	.0126
10	\bar{C}_k	.8140	.7398	.7240	.7252	.7155	.7090	.7054	.7042	.7018
	S_c	.0944	.0250	.0181	.0194	.0153	.0129	.0120	.0113	.0111
%		19	72	76	81	90	85	84	83	78

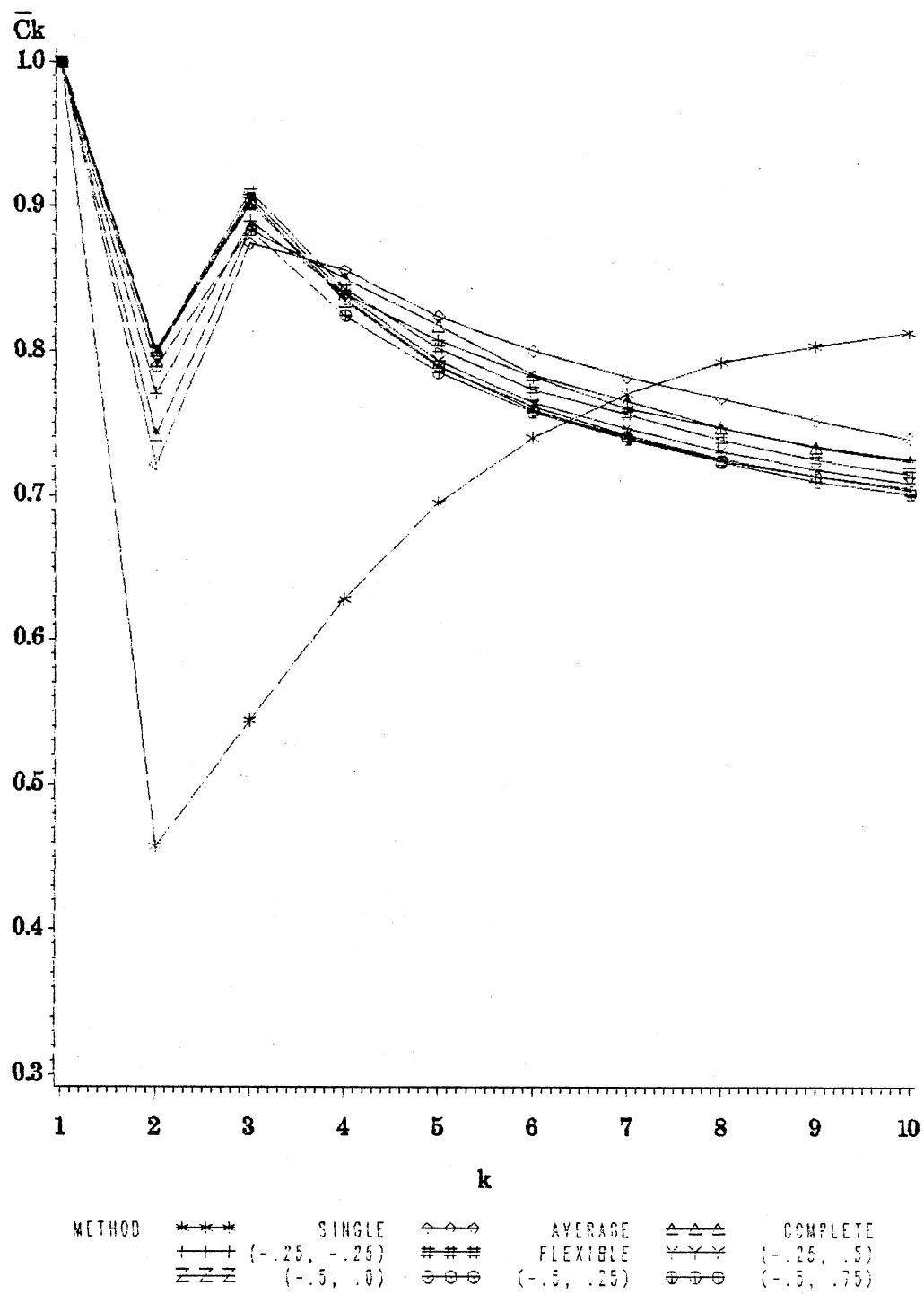


Figure 5. Retrieval result of the nine algorithms for MVN with $\delta = 4.0$, $\rho = .0$ and 30-20-10 split

TABLE 5

RETRIEVAL INFORMATION OF NINE ALGORITHMS VS. POPN. WITH
 $\delta = 6.0$, $\rho = 0.0$, AND 20-20-20 SPLIT FOR MVN

k		$\begin{pmatrix} .0 \\ -.5 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .0 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .75 \end{pmatrix}$
2	\bar{C}_k	.7452	.7628	.7642	.7671	.7675	.7652	.7688	.7689	.7643
	S_c	.1045	.0506	.0296	.0216	.0200	.0164	.0125	.0119	.0190
3	\bar{C}_k	.9267	.9864	.9858	.9869	.9894	.9853	.9894	.9898	.9833
	S_c	.1166	.0324	.0271	.0245	.0189	.0251	.0180	.0192	.0260
4	\bar{C}_k	.9561	.9636	.9502	.9488	.9484	.9444	.9413	.9415	.9361
	S_c	.0743	.0208	.0215	.0190	.0169	.0157	.0135	.0141	.0191
5	\bar{C}_k	.9677	.9286	.9096	.9084	.9060	.8991	.8930	.8946	.8903
	S_c	.0322	.0247	.0241	.0206	.0197	.0153	.0116	.0110	.0156
6	\bar{C}_k	.9597	.8990	.8749	.8736	.8649	.8571	.8489	.8488	.8457
	S_c	.0108	.0310	.0234	.0244	.0208	.0168	.0147	.0141	.0157
7	\bar{C}_k	.9456	.8727	.8479	.8457	.8391	.8317	.8265	.8246	.8223
	S_c	.0134	.0311	.0219	.0215	.0193	.0152	.0107	.0113	.0108
8	\bar{C}_k	.9299	.8502	.8268	.8276	.8179	.8101	.8067	.8038	.8036
	S_c	.0164	.0259	.0183	.0204	.0154	.0117	.0117	.0088	.0097
9	\bar{C}_k	.9147	.8315	.8097	.8112	.8021	.7947	.7898	.7886	.7871
	S_c	.0192	.0222	.0158	.0188	.0138	.0113	.0104	.0090	.0091
10	\bar{C}_k	.9003	.8163	.7965	.7991	.7892	.7814	.7780	.7773	.7748
	S_c	.0209	.0201	.0151	.0172	.0126	.0099	.0085	.0077	.0078
	%	67	87	93	96	98	93	100	100	99

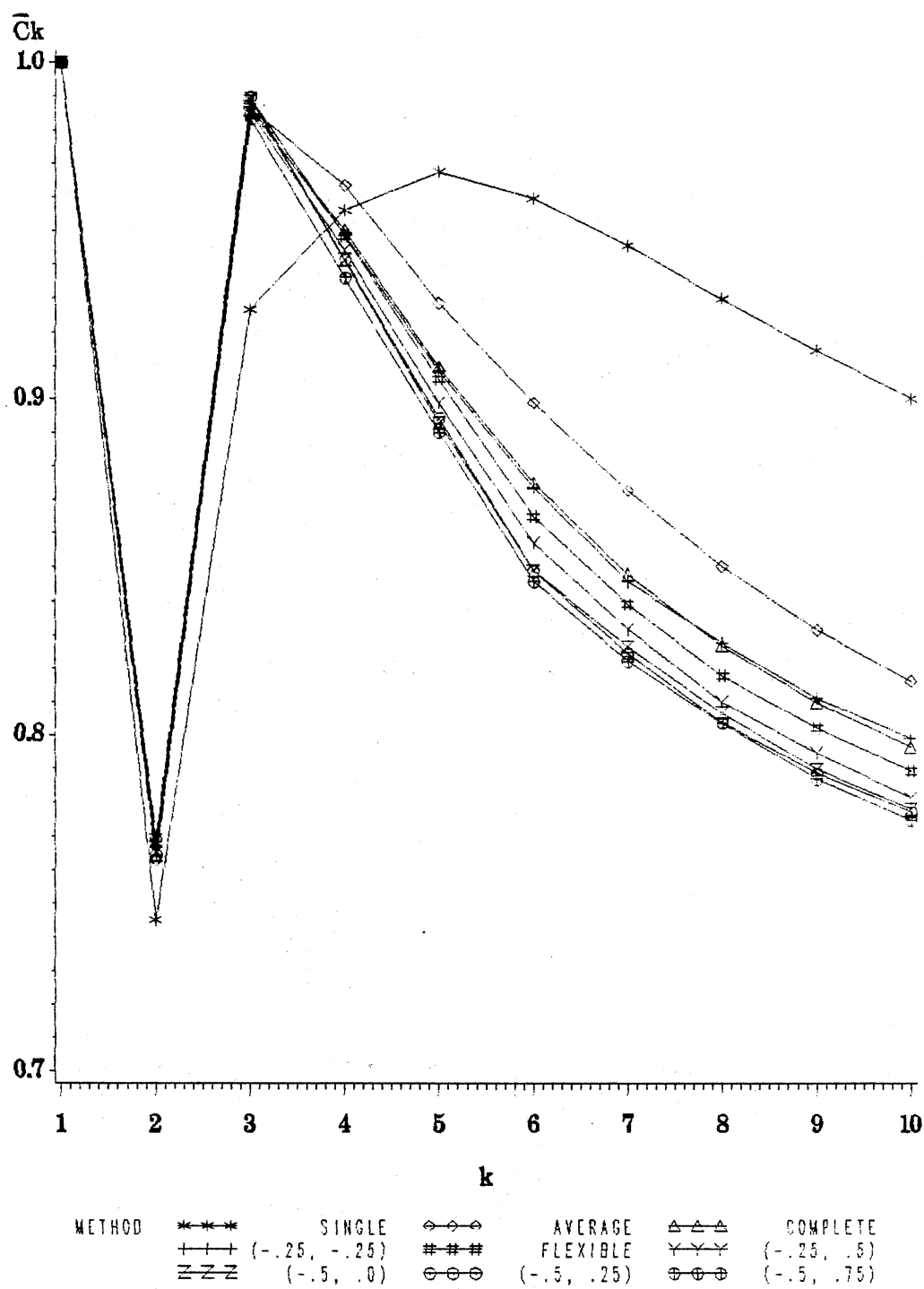


Figure 6. Retrieval result of the nine algorithms for MVN with $\delta = 6.0$, $\rho = .0$ and 20-20-20 split

TABLE 6

RETRIEVAL INFORMATION OF NINE ALGORITHMS VS. POPN. WITH
 $\delta = 6.0$, $\rho = 0.0$, AND 30-20-10 SPLIT FOR MVN

k		$\begin{pmatrix} .0 \\ -.5 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .0 \end{pmatrix}$	$\begin{pmatrix} .0 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .0 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ .75 \end{pmatrix}$
2	\bar{C}_k	.7309	.7945	.8187	.8416	.8690	.8680	.8730	.8737	.8612
	S_c	.1318	.1028	.0814	.0716	.0306	.0326	.0297	.0257	.0371
3	C_k	.9093	.9824	.9873	.9859	.9875	.9865	.9874	.9854	.9824
	S_c	.1370	.0352	.0221	.0267	.0223	.0251	.0238	.0245	.0272
4	\bar{C}_k	.9560	.9452	.9240	.9133	.9010	.8937	.8874	.8846	.8837
	S_c	.0869	.0450	.0378	.0411	.0304	.0284	.0236	.0223	.0278
5	\bar{C}_k	.9625	.9022	.8688	.8638	.8448	.8372	.8322	.8302	.8287
	S_c	.0584	.0547	.0434	.0328	.0266	.0222	.0180	.0172	.0151
6	\bar{C}_k	.9538	.8598	.8324	.8302	.8082	.7958	.7931	.7895	.7875
	S_c	.0501	.0477	.0393	.0323	.0259	.0206	.0158	.0143	.0138
7	\bar{C}_k	.9433	.8284	.8053	.8018	.7817	.7754	.7699	.7664	.7654
	S_c	.0419	.0381	.0331	.0281	.0195	.0170	.0121	.0110	.0101
8	\bar{C}_k	.9320	.8070	.7791	.7792	.7625	.7577	.7499	.7497	.7463
	S_c	.0370	.0376	.0225	.0246	.0157	.0163	.0105	.0101	.0089
9	\bar{C}_k	.9228	.7880	.7600	.7628	.7469	.7405	.7348	.7330	.7310
	S_c	.0257	.0347	.0207	.0197	.0151	.0129	.0100	.0093	.0077
10	\bar{C}_k	.9089	.7700	.7448	.7496	.7327	.7262	.7229	.7206	.7197
	S_c	.0241	.0297	.0171	.0185	.0116	.0094	.0084	.0077	.0075
%		64	83	91	95	96	96	96	96	95

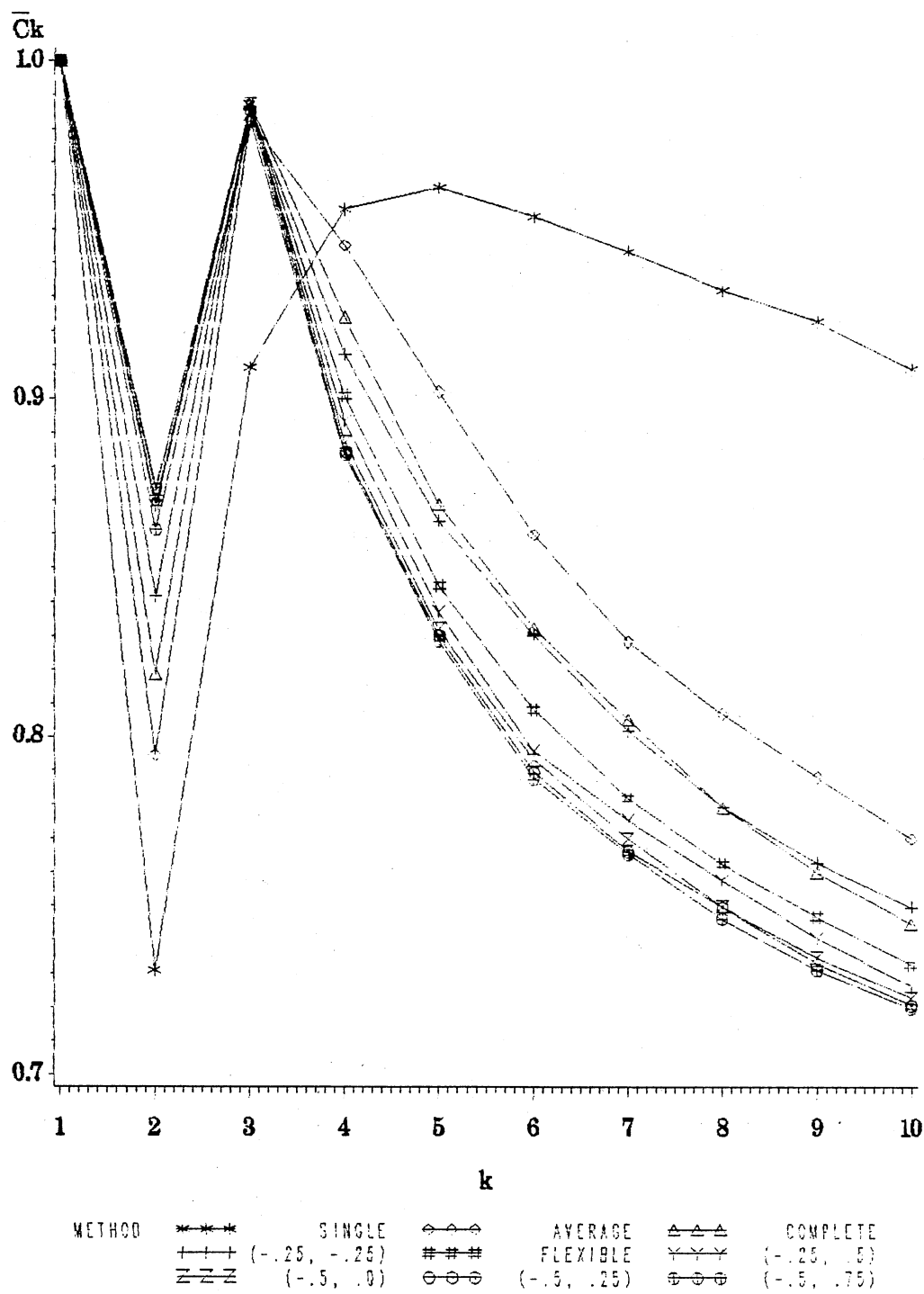


Figure 7. Retrieval result of the nine algorithms for MVN with $\delta = 6.0$, $\rho = .0$ and 30-20-10 split

TABLE 7

THE $\%_s$ ON LOCAL MAXIMUM FOR ALL POSSIBLE PAIRS OF
THE NINE ALGORITHMS WHEN $\delta = 4.0$ FOR MVN

(β, π)	split	20-20-20			30-20-10			$\bar{\%}_s$	$S_{\bar{\%}_s}$
	(β, π)	.0	.4	.8	.0	.4	.8		
$(.0, -.5)$	$(.0, .0)$	14	25	39	22	12	35	24.5	4.45
	$(.0, .5)$	20	22	44	25	24	34	28.2	3.73
	$(-.25, -.25)$	21	22	41	23	22	40	28.2	3.91
	$(.0, .0)$	23	24	63	34	25	44	35.5	6.40
	$(.0, .5)$	23	27	67	32	23	56	38.0	7.69
	$(-.5, .0)$	23	27	63	32	29	48	37.0	6.28
	$(.0, .25)$	21	26	65	34	26	59	38.5	7.66
	$(.0, .75)$	23	27	60	29	25	47	35.2	6.09
$(.0, .0)$	$(.0, .5)$	38	38	34	42	43	29	37.3	2.12
	$(-.25, -.25)$	28	29	33	47	33	32	33.7	2.80
	$(.0, .0)$	33	44	39	54	53	37	43.3	3.53
	$(.0, .5)$	46	48	47	54	52	40	47.8	2.01
	$(-.5, .0)$	42	46	40	54	54	40	46.0	2.68
	$(.0, .25)$	49	52	43	54	53	43	49.0	2.02
	$(.0, .75)$	46	48	48	54	45	44	47.5	1.45
$(.0, .5)$	$(-.25, -.25)$	39	34	36	48	47	35	39.8	2.52
	$(.0, .0)$	37	39	32	59	47	39	42.2	3.90
	$(.0, .5)$	42	56	38	57	55	40	48.0	3.62
	$(-.5, .0)$	51	52	36	53	48	40	46.7	2.87
	$(.0, .25)$	47	54	35	54	49	43	47.0	2.96
	$(.0, .75)$	58	51	35	60	48	37	48.2	4.25
$(-.25, -.25)$	$(-.25, .0)$	38	31	41	45	40	36	38.5	1.95
	$(.0, .5)$	50	48	50	55	51	49	50.5	0.99
	$(-.5, .0)$	49	42	46	47	46	48	46.3	0.99
	$(.0, .25)$	51	49	46	53	53	56	51.3	1.43
	$(.0, .75)$	57	54	51	53	51	53	53.2	0.91
$(-.25, .0)$	$(-.25, .5)$	48	49	50	51	44	45	47.8	1.14
	$(-.5, .0)$	41	35	37	31	38	38	36.7	1.38
	$(.0, .25)$	46	47	41	45	50	45	45.7	1.20
	$(.0, .75)$	52	53	57	50	48	50	51.7	1.28
$(-.25, .5)$	$(-.5, .0)$	49	40	45	53	40	41	44.7	2.20
	$(.0, .25)$	39	44	44	42	45	39	42.2	1.08
	$(.0, .75)$	48	49	56	47	35	41	46.0	2.94
$(-.5, .0)$	$(-.5, .25)$	34	37	33	37	37	25	33.8	1.90
	$(.0, .75)$	47	44	50	44	46	45	46.0	0.93
$(-.5, .25)$	$(-.5, .75)$	46	48	49	44	42	42	45.2	1.22
	$\bar{\%}_s$	39.4	40.6	45.4	44.9	41.1	42.1	42.3	

TABLE 8

THE $\%_s$ ON LOCAL MAXIMUM FOR ALL POSSIBLE PAIRS OF
THE NINE ALGORITHMS WHEN $\delta = 6.0$ FOR MVN

(β , π)	split	20-20-20			30-20-10			$\bar{\%}_s$	$S\bar{\%}_s$
	(β , π)	.0	.4	.8	.0	.4	.8		
(.0, -.5)	(.0 , .0)	54	58	66	49	52	53	55.3	2.44
	(, .5)	71	73	75	64	71	74	71.3	1.61
	(-.25, -.25)	71	63	68	64	64	60	65.0	1.59
	(, .0)	75	73	70	70	67	78	72.2	1.62
	(, .5)	77	72	76	71	75	82	75.5	1.61
	(-.5 , .0)	76	71	67	73	75	82	74.0	2.07
	(, .25)	78	74	76	72	77	85	77.0	1.83
	(, .75)	75	76	76	69	73	75	74.0	1.10
(.0, .0)	(.0 , .5)	61	54	52	65	64	61	59.5	2.17
	(-.25, -.25)	59	52	42	61	64	60	56.3	3.29
	(, .0)	63	64	53	80	70	72	67.0	3.76
	(, .5)	80	75	66	80	80	76	76.2	2.23
	(-.5 , .0)	72	70	58	86	75	78	73.2	3.80
	(, .25)	77	74	64	86	82	78	76.8	3.08
	(, .75)	91	78	70	84	85	74	80.3	3.17
(.0, .5)	(-.25, -.25)	64	57	54	72	59	61	61.2	2.57
	(, .0)	66	53	52	72	56	66	60.8	3.37
	(, .5)	71	73	64	77	71	74	71.7	1.78
	(-.5 , .0)	65	64	64	73	72	77	69.2	2.27
	(, .25)	75	73	61	81	78	78	74.3	2.89
	(, .75)	80	81	77	83	78	73	78.7	1.43
	(-.25, .0)	51	42	42	51	55	41	47.0	2.46
(-.25, -.25)	(, .5)	75	67	60	75	71	66	69.0	2.38
	(-.5 , .0)	53	51	48	70	70	60	58.7	3.93
	(, .25)	68	67	61	75	79	69	69.8	2.59
	(, .75)	79	77	74	78	78	68	75.7	1.69
	(-.25, .5)	73	57	65	66	59	58	63.0	2.52
(-.25, .0)	(-.5 , .0)	50	43	47	54	56	50	50.0	1.91
	(, .25)	63	55	59	66	66	61	61.7	1.74
	(, .75)	78	75	69	74	76	65	72.8	1.99
	(-.5 , .0)	64	52	59	66	56	51	58.0	2.52
(-.25, .5)	(, .25)	56	48	46	54	52	51	51.2	1.51
	(, .75)	71	67	64	68	59	59	64.7	2.01
	(-.5 , .25)	40	38	33	49	43	37	40.0	2.25
(-.5, .0)	(, .75)	70	65	64	76	70	58	67.2	2.54
	(-.5 , .75)	59	56	52	68	58	54	57.8	2.29
(-.5, .25)	$\bar{\%}_s$	68.1	63.6	60.9	70.1	67.7	65.7	66.0	

TABLE 9
 AGREEMENT OF FIVE PAIRED ALGORITHMS WITH $\delta = 4.0$,
 $\rho = 0.0$, AND 20-20-20 SPLIT FOR MVN

Paired Clustering Algorithms						
		$\begin{pmatrix} .0 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .0 \end{pmatrix}$
k		$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$
2	\bar{C}_k	.6689	.6548	.6831	.6470	.6718
	S_c	.1809	.1815	.1915	.1783	.1878
3	\bar{C}_k	.8814	.8999	.9112	.8918	.9094
	S_c	.1068	.0963	.0830	.0929	.0711
4	\bar{C}_k	.8743	.8992	.9043	.8886	.8912
	S_c	.0594	.0602	.0542	.0493	.0513
5	\bar{C}_k	.8746	.9038	.9047	.8861	.8899
	S_c	.0459	.0486	.0457	.0457	.0432
6	\bar{C}_k	.8849	.9054	.9121	.8890	.9014
	S_c	.0380	.0398	.0400	.0402	.0421
7	\bar{C}_k	.8982	.9111	.9148	.8988	.9086
	S_c	.0303	.0341	.0339	.0339	.0331
8	\bar{C}_k	.9097	.9228	.9272	.9102	.9206
	S_c	.0247	.0289	.0284	.0271	.0260
9	\bar{C}_k	.9195	.9307	.9351	.9193	.9289
	S_c	.0218	.0238	.0250	.0234	.0218
10	\bar{C}_k	.9284	.9394	.9425	.9297	.9392
	S_c	.0197	.0220	.0199	.0210	.0200
% _s		58	50	51	57	52

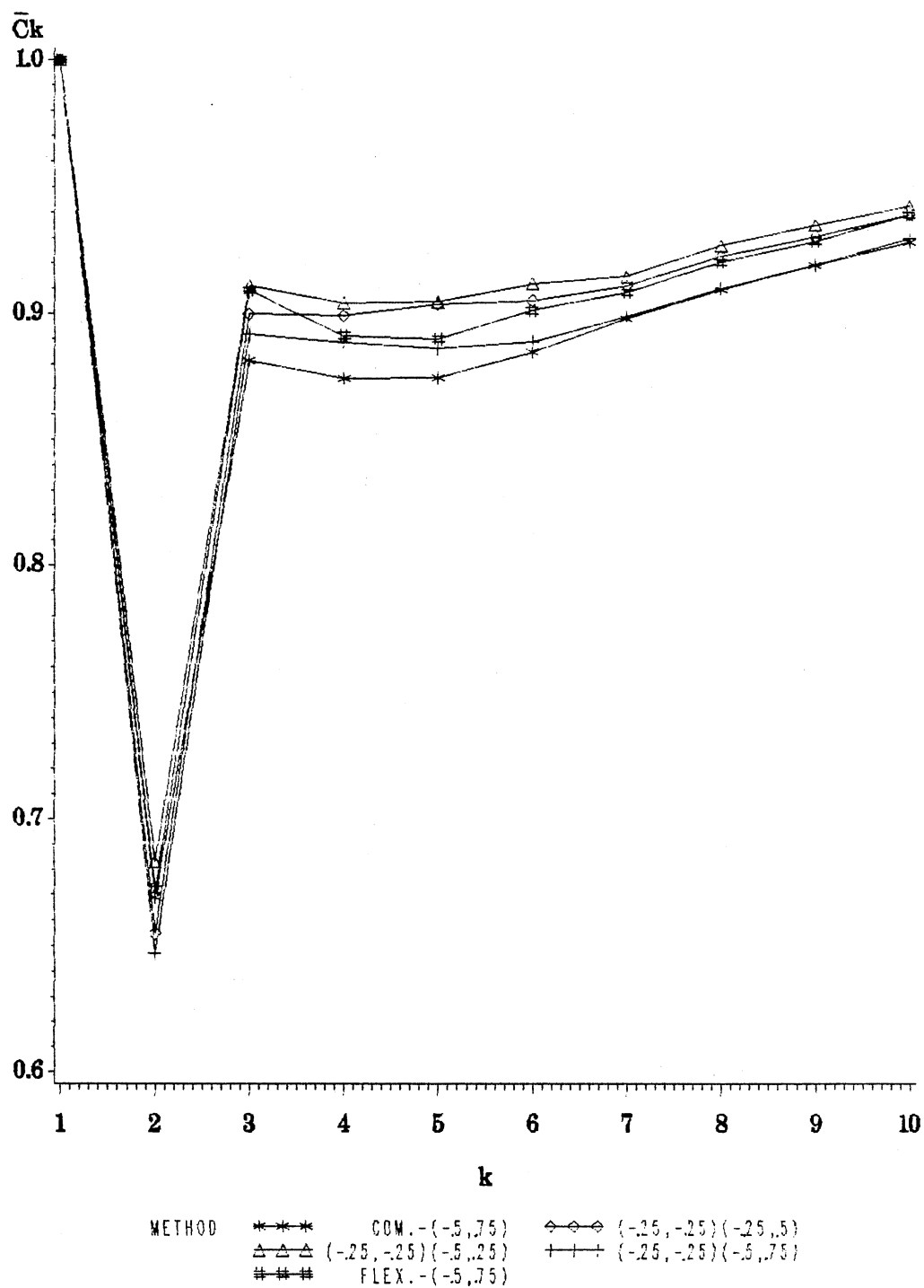


Figure 8. Retrieval result of five paired clustering algorithms with $\delta = 4.0$, $\rho = .0$ and 20-20-20 split for MVN

TABLE 10
 AGREEMENT OF FIVE PAIRED ALGORITHMS WITH $\delta = 4.0$,
 $\rho = 0.0$, AND 30-20-10 SPLIT FOR MVN

Paired Clustering Algorithms					
	$\begin{pmatrix} .0 \\ ; \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ ; \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ ; \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ ; \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ ; \\ .0 \end{pmatrix}$
k	$\begin{pmatrix} -.5 \\ ; \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ ; \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ ; \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ ; \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ ; \\ .75 \end{pmatrix}$
2 \bar{C}_k	.7408	.8060	.7935	.8045	.8430
S_c	.1773	.1712	.1748	.1779	.1552
3 \bar{C}_k	.8840	.9115	.9060	.8873	.9057
S_c	.1079	.0831	.0852	.0972	.0886
4 \bar{C}_k	.8566	.8755	.8749	.8567	.8791
S_c	.0604	.0580	.0582	.0591	.0628
5 \bar{C}_k	.8574	.8737	.8827	.8572	.8804
S_c	.0418	.0593	.0520	.0491	.0472
6 \bar{C}_k	.8791	.8918	.8910	.8787	.8959
S_c	.0422	.0508	.0494	.0439	.0401
7 \bar{C}_k	.8918	.9082	.9133	.8968	.9064
S_c	.0338	.0419	.0401	.0375	.0329
8 \bar{C}_k	.9081	.9176	.9212	.9058	.9191
S_c	.0301	.0331	.0332	.0308	.0282
9 \bar{C}_k	.9194	.9282	.9323	.9158	.9307
S_c	.0260	.0283	.0280	.0245	.0252
10 \bar{C}_k	.9282	.9345	.9394	.9246	.9392
S_c	.0215	.0244	.0258	.0217	.0214
% _s	60	55	53	53	50

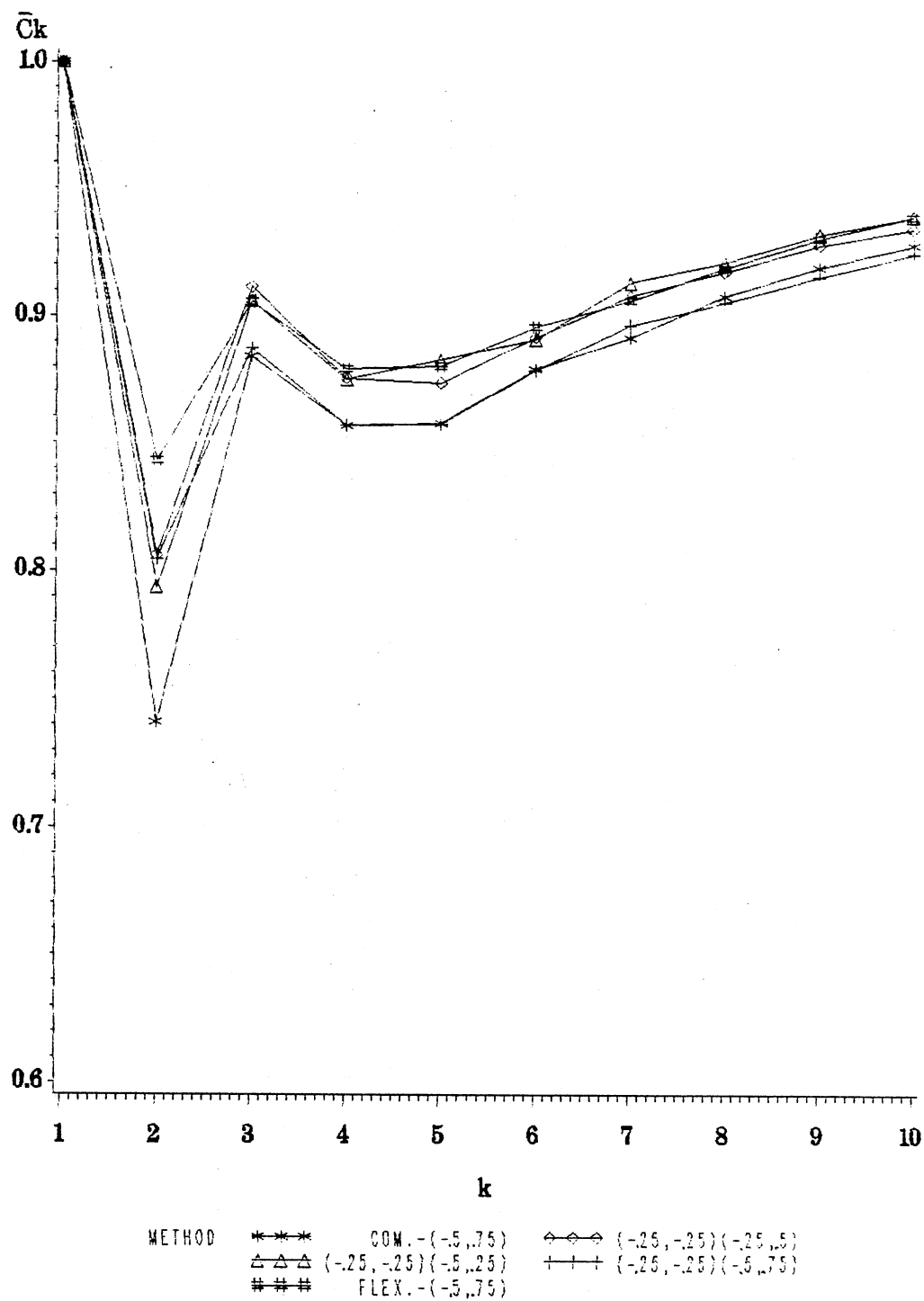


Figure 9. Retrieval result of five paired clustering algorithms with $\delta = 4.0$, $\rho = .0$ and 30-20-10 split for MVN

TAVLE 11

AGREEMENT OF FIVE PAIRED ALGORITHMS WITH $\delta = 4.0$,
 $\Theta = 15^\circ$, AND 20-20-20 SPLIT FOR MVLN

Paired Clustering Algorithms						
		$\begin{pmatrix} .0 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .0 \end{pmatrix}$
k		$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$
2	\bar{C}_k	.6836	.6641	.7054	.6698	.6938
	S_c	.1847	.1842	.1980	.1830	.1877
3	\bar{C}_k	.8659	.9053	.9183	.9012	.9196
	S_c	.1198	.0865	.0867	.0842	.0719
4	\bar{C}_k	.8871	.9032	.9100	.8909	.9021
	S_c	.0594	.0605	.0614	.0569	.0547
5	\bar{C}_k	.8878	.9072	.9102	.8939	.9047
	S_c	.0436	.0421	.0465	.0447	.0432
6	\bar{C}_k	.9000	.9140	.9192	.9076	.9137
	S_c	.0363	.0367	.0388	.0337	.0360
7	\bar{C}_k	.9075	.9258	.9262	.9106	.9172
	S_c	.0319	.0334	.0345	.0308	.0328
8	\bar{C}_k	.9137	.9293	.9291	.9164	.9222
	S_c	.0283	.0310	.0276	.0263	.0254
9	\bar{C}_k	.9209	.9331	.9371	.9238	.9303
	S_c	.0215	.0263	.0279	.0228	.0231
10	\bar{C}_k	.9275	.9389	.9410	.9301	.9383
	S_c	.0214	.0238	.0247	.0197	.0187
%s		48	43	36	53	60

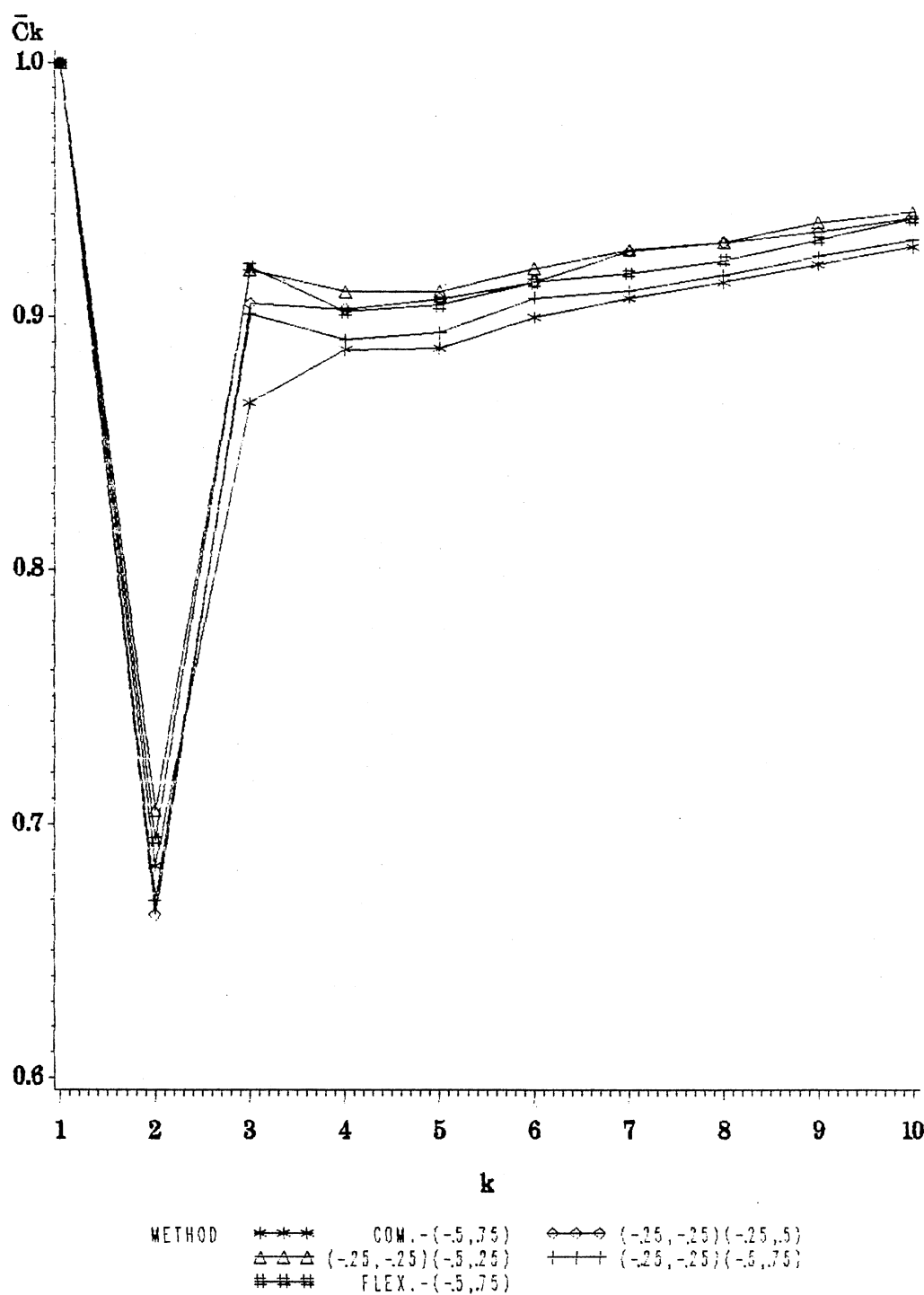


Figure 10. Retrieval result of the five paired clustering algorithms with $\delta = 4.0$, $\theta = 15.0$ and 20-20-20 split for MVLN

TABLE 12
 AGREEMENT OF FIVE PAIRED ALGORITHMS WITH $\delta = 4.0$,
 $\Theta = 15^\circ$, AND 30-20-10 SPLIT FOR MVLN

Paired Clustering Algorithms						
		$\begin{pmatrix} .0 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ -.25 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .0 \end{pmatrix}$
k		$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.25 \\ , \\ .5 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .25 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$	$\begin{pmatrix} -.5 \\ , \\ .75 \end{pmatrix}$
2	\bar{C}_k	.6525	.7537	.7557	.7326	.8193
	S_c	.1550	.1803	.1877	.1744	.1664
3	\bar{C}_k	.8789	.9172	.9198	.8945	.9118
	S_c	.1084	.0841	.0808	.0953	.0854
4	\bar{C}_k	.8691	.9048	.8995	.8768	.8921
	S_c	.0674	.0643	.0617	.0646	.0623
5	\bar{C}_k	.8635	.9052	.8925	.8692	.8853
	S_c	.0466	.0516	.0528	.0477	.0503
6	\bar{C}_k	.8666	.9029	.8979	.8761	.8926
	S_c	.0414	.0428	.0448	.0427	.0468
7	\bar{C}_k	.8767	.9029	.8988	.8861	.9017
	S_c	.0383	.0388	.0375	.0423	.0390
8	\bar{C}_k	.8819	.9147	.9102	.8899	.9043
	S_c	.0355	.0347	.0334	.0347	.0329
9	\bar{C}_k	.8945	.9168	.9152	.8985	.9157
	S_c	.0364	.0338	.0309	.0326	.0276
10	\bar{C}_k	.9130	.9252	.9229	.9118	.9290
	S_c	.0314	.0320	.0284	.0295	.0252
% _s		57	49	46	51	44

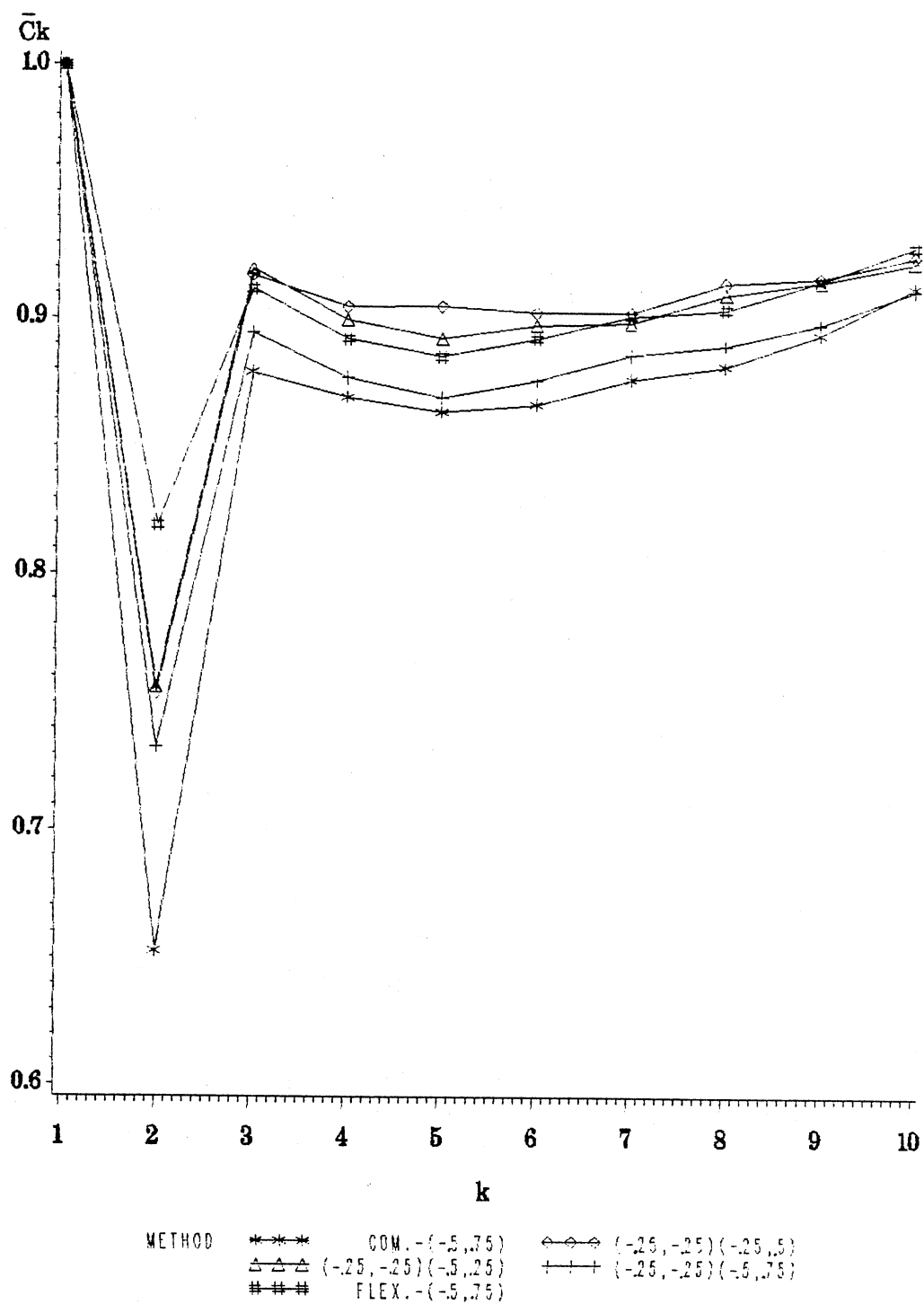


Figure 11. Retrieval result of the five paired clustering algorithms with $\delta = 4.0$, $\theta = 15.0$ and 30-20-10 split for MVLN

TABLE 13

THE %_s ON LOCAL MAXIMUM FOR ALL POSSIBLE PAIRS OF
THE NINE ALGORITHMS WHEN $\delta = 4.0$ WITH MVLN

(β, π)	split	20-20-20		30-20-10		20-10-30		20-30-10	
	$(\beta, \pi)^\ominus$	15	30	15	30	15	30	15	30
$(.0, -.5)$	$(.0, .0)$	20	17	20	15	17	14	26	25
	$(, .5)$	24	22	21	22	16	24	28	21
	$(-.25, -.25)$	24	27	19	20	19	24	34	32
	$(, .0)$	27	20	30	27	26	30	34	33
	$(, .5)$	22	26	32	30	25	27	36	37
	$(-.5, .0)$	27	21	31	31	34	31	38	37
	$(, .25)$	23	22	30	35	31	30	37	37
	$(, .75)$	27	29	28	33	29	22	37	29
$(.0, .0)$	$(.0, .5)$	29	32	27	36	26	27	30	36
	$(-.25, -.25)$	32	28	29	29	32	27	33	37
	$(, .0)$	32	30	30	34	33	35	44	34
	$(, .5)$	41	37	42	44	40	35	52	33
	$(-.5, .0)$	33	30	40	40	42	44	47	33
	$(, .25)$	34	35	41	45	46	44	47	37
	$(, .75)$	39	43	48	48	45	44	51	38
$(.0, .5)$	$(-.25, -.25)$	33	33	34	40	41	39	38	37
	$(, .0)$	34	31	44	41	36	46	36	35
	$(, .5)$	38	41	49	46	47	47	44	49
	$(-.5, .0)$	37	37	51	48	47	44	43	42
	$(, .25)$	42	39	53	58	47	51	44	52
	$(, .75)$	48	44	57	57	57	48	47	45
$(-.25, -.25)$	$(-.25, .0)$	32	26	28	32	34	30	32	27
	$(, .5)$	43	46	49	48	47	41	52	48
	$(-.5, .0)$	39	34	47	37	48	43	48	44
	$(, .25)$	36	41	46	50	50	52	53	50
	$(, .75)$	53	52	51	50	47	52	52	46
$(-.25, .0)$	$(-.25, .5)$	43	53	38	41	51	45	59	53
	$(-.5, .0)$	42	37	34	37	46	47	55	42
	$(, .25)$	35	45	40	46	52	54	51	53
	$(, .75)$	60	53	44	54	57	56	51	49
$(-.25, .5)$	$(-.5, .0)$	42	49	44	45	51	50	52	49
	$(, .25)$	45	42	43	43	50	48	51	52
	$(, .75)$	56	46	50	51	53	62	61	56
$(-.5, .0)$	$(-.5, .25)$	29	27	34	32	49	38	39	47
	$(, .75)$	52	52	51	42	57	55	52	48
$(-.5, .25)$	$(-.5, .75)$	53	46	39	49	46	51	44	56
	% _s	36.8	35.9	38.7	39.9	40.9	39.7	43.8	41.4

TABLE 13 (Continued)

(β, π)	split	30-10-20		10-20-30		10-30-20		$\bar{\%}_s$	$S_{\bar{\%}_s}$
	(β, π)	15	30	15	30	15	30		
$(.0, -.5)$	$(.0, .0)$	10	19	16	17	20	23	18.5	1.16
	$(, .5)$	10	19	28	25	34	32	23.3	1.66
	$(-.25, -.25)$	11	22	27	28	33	31	25.1	1.74
	$(, .0)$	18	21	36	34	44	39	29.9	1.98
	$(, .5)$	18	27	43	39	45	47	32.4	2.39
	$(-.5, .0)$	21	28	47	41	46	44	34.1	2.27
	$(, .25)$	22	26	48	40	45	48	33.9	2.44
	$(, .75)$	20	24	36	39	42	45	31.4	2.00
$(.0, .0)$	$(.0, .5)$	30	21	29	34	28	30	29.6	1.08
	$(-.25, -.25)$	25	26	30	38	33	39	31.3	1.18
	$(, .0)$	36	31	39	40	40	46	36.0	1.36
	$(, .5)$	45	36	46	47	44	43	41.8	1.40
	$(-.5, .0)$	48	36	42	58	49	46	42.0	2.01
	$(, .25)$	48	37	48	54	44	53	43.8	1.68
	$(, .75)$	46	33	39	48	47	42	43.6	1.33
$(.0, .5)$	$(-.25, -.25)$	31	34	32	40	35	42	36.4	0.98
	$(, .0)$	42	47	44	49	46	46	41.2	1.53
	$(, .5)$	54	49	53	48	55	52	48.0	1.28
	$(-.5, .0)$	53	51	54	53	53	56	47.8	1.68
	$(, .25)$	51	55	58	53	57	56	51.1	1.61
	$(, .75)$	50	48	53	53	57	54	51.3	1.26
$(-.25, -.25)$	$(-.25, .0)$	37	22	39	31	32	46	32.0	1.58
	$(, .5)$	52	44	51	44	48	53	47.6	0.98
	$(-.5, .0)$	44	40	43	50	46	47	43.6	1.24
	$(, .25)$	48	47	46	52	50	56	48.4	1.36
	$(, .75)$	50	50	49	53	49	51	50.4	0.56
$(-.25, .0)$	$(-.25, .5)$	43	40	48	45	48	53	47.1	1.62
	$(-.5, .0)$	40	31	50	50	41	48	42.9	1.82
	$(, .25)$	49	45	48	45	45	56	47.4	1.52
	$(, .75)$	49	47	50	56	58	58	53.0	1.28
$(-.25, .5)$	$(-.5, .0)$	41	41	56	45	49	49	47.4	1.19
	$(, .25)$	38	37	49	42	50	55	46.1	1.45
	$(, .75)$	50	41	45	60	53	50	52.4	1.65
$(-.5, .0)$	$(-.5, .25)$	41	35	46	39	39	42	38.4	1.75
	$(, .75)$	49	52	44	58	57	47	51.1	1.28
$(-.5, .25)$	$(-.5, .75)$	44	46	42	51	50	47	47.4	1.22
	$\bar{\%}_s$	37.9	36.3	43.2	43.9	44.8	46.4	40.7	

TABLE 14

THE %_s ON LOCAL MAXIMUM FOR ALL POSSIBLE PAIRS OF
THE NINE ALGORITHMS FOR $\delta = 6.0$ WITH MVLN.

(β, π)	split	20-20-20		30-20-10		20-10-30		20-30-10	
	$(\beta, \pi)^\ominus$	15	30	15	30	15	30	15	30
$(.0, -.5)$	$(.0, .0)$	45	40	30	43	44	30	37	40
	$(, .5)$	51	43	41	50	50	47	43	56
	$(-.25, -.25)$	46	50	46	42	46	46	43	50
	$(, .0)$	51	54	49	55	58	48	51	60
	$(, .5)$	53	52	53	61	58	52	56	64
	$(-.5, .0)$	58	57	54	64	59	54	57	62
	$(, .25)$	59	55	57	65	62	55	60	66
	$(, .75)$	67	54	54	62	58	57	58	69
$(.0, .0)$	$(.0, .5)$	50	43	49	55	49	43	41	54
	$(-.25, -.25)$	51	49	46	45	51	45	51	44
	$(, .0)$	66	49	64	51	67	55	65	61
	$(, .5)$	66	65	71	73	69	68	77	73
	$(-.5, .0)$	60	58	74	77	72	66	71	70
	$(, .25)$	64	67	77	78	79	72	71	69
	$(, .75)$	71	69	77	78	79	78	80	81
$(.0, .5)$	$(-.25, -.25)$	50	51	54	63	53	49	55	61
	$(, .0)$	53	47	63	59	64	49	51	68
	$(, .5)$	60	64	68	78	64	64	69	66
	$(-.5, .0)$	56	59	77	70	70	68	65	70
	$(, .25)$	65	67	80	77	76	76	69	69
	$(, .75)$	73	74	76	82	77	77	76	80
$(-.25, -.25)$	$(-.25, .0)$	31	41	39	35	43	38	50	50
	$(, .5)$	58	70	61	76	65	65	64	71
	$(-.5, .0)$	47	53	57	64	61	56	57	63
	$(, .25)$	58	67	66	74	68	65	58	65
	$(, .75)$	73	73	72	84	74	72	75	78
$(-.25, .0)$	$(-.25, .5)$	58	58	55	67	61	57	60	65
	$(-.5, .0)$	37	38	48	62	59	45	54	55
	$(, .25)$	53	56	58	72	71	55	53	59
	$(, .75)$	77	71	70	73	70	67	66	84
$(-.25, .5)$	$(-.5, .0)$	47	54	58	61	61	57	60	64
	$(, .25)$	43	42	48	56	56	56	55	60
	$(, .75)$	68	61	61	61	70	65	60	69
$(-.5, .0)$	$(-.5, .25)$	34	32	39	50	45	37	39	41
	$(, .75)$	74	60	61	67	70	61	63	71
$(-.5, .25)$	$(-.5, .75)$	62	53	51	59	65	53	61	70
	% _s	56.5	55.4	58.4	63.6	62.3	56.9	58.9	63.8

TABLE 14 (Continued)

(β, π)	split	30-10-20		10-20-30		10-30-20		$\bar{\%}_s$	$S_{\bar{\%}_s}$
	$(\beta, \pi)^\ominus$	15	30	15	30	15	30		
$(.0, -.5)$	$(.0, .0)$	41	31	44	30	45	47	40.2	1.60
	$(, .5)$	45	45	50	47	59	66	50.4	1.91
	$(-.25, -.25)$	46	53	46	46	58	63	50.1	1.68
	$(, .0)$	54	53	58	48	64	69	57.4	2.03
	$(, .5)$	58	52	58	52	68	73	60.3	2.16
	$(-.5, .0)$	61	56	59	54	68	74	62.6	1.96
	$(, .25)$	60	66	62	55	71	72	63.9	1.80
	$(, .75)$	59	63	58	57	70	75	63.4	1.88
$(.0, .0)$	$(.0, .5)$	44	41	49	43	47	59	48.2	1.47
	$(-.25, -.25)$	39	48	51	45	51	51	48.6	1.19
	$(, .0)$	53	57	67	55	60	58	59.1	1.64
	$(, .5)$	76	68	69	68	80	76	72.6	1.31
	$(-.5, .0)$	71	71	72	66	81	80	71.7	1.85
	$(, .25)$	77	78	79	72	81	74	74.4	1.40
	$(, .75)$	68	76	79	78	88	76	77.9	1.69
$(.0, .5)$	$(-.25, -.25)$	53	55	53	49	57	69	55.7	1.50
	$(, .0)$	51	55	64	49	55	68	56.8	1.80
	$(, .5)$	72	73	64	64	69	81	68.9	1.51
	$(-.5, .0)$	76	74	70	68	71	78	70.3	1.78
	$(, .25)$	82	78	76	76	73	80	74.5	1.40
	$(, .75)$	75	80	77	77	77	77	77.6	0.75
$(-.25, -.25)$	$(-.25, .0)$	37	39	43	38	49	48	41.7	1.57
	$(, .5)$	72	57	65	65	68	75	67.2	1.58
	$(-.5, .0)$	64	56	61	56	64	72	59.6	1.60
	$(, .25)$	72	68	68	65	72	73	67.6	1.33
	$(, .75)$	68	72	74	72	78	81	74.9	1.11
$(-.25, .0)$	$(-.25, .5)$	63	56	61	57	63	65	60.9	0.98
	$(-.5, .0)$	57	48	59	45	57	59	51.4	2.14
	$(, .25)$	70	61	71	55	63	68	61.7	1.84
	$(, .75)$	69	68	70	67	75	76	71.9	1.36
$(-.25, .5)$	$(-.5, .0)$	58	52	61	57	63	65	59.4	1.62
	$(, .25)$	55	51	56	56	55	56	53.8	1.51
	$(, .75)$	64	62	70	65	68	67	65.2	1.18
$(-.5, .0)$	$(-.5, .25)$	41	41	45	37	48	43	41.9	1.56
	$(, .75)$	63	59	70	61	71	67	66.3	1.50
$(-.5, .25)$	$(-.5, .75)$	60	55	65	53	63	58	60.6	1.91
	$\bar{\%}_s$	60.4	58.8	67.6	62.8	65.3	67.8	61.3	

TABLE 15
PERCENT RETRIEVAL OF TRUE POPULATION FOR ALL
ALGORITHMS WITH MVLN

(β, π)	split	20-20-20		30-20-10		20-10-30		20-30-10	
	$\delta \quad \ominus$	15	30	15	30	15	30	15	30
(.0 , -.5)	4.0	18	21	17	21	17	20	25	24
	6.0	48	50	43	43	47	41	53	45
(.0 , .0)	4.0	47	53	50	49	53	45	53	47
	6.0	76	72	75	75	77	72	77	72
(.0 , .5)	4.0	55	55	62	60	54	59	58	62
	6.0	79	78	84	84	80	71	83	86
(-.25, -.25)	4.0	73	63	65	68	66	68	71	69
	6.0	84	82	80	83	80	77	83	80
(-.25, .0)	4.0	76	73	65	66	74	74	78	77
	6.0	93	89	80	82	87	78	84	90
(-.25, .5)	4.0	81	79	74	76	76	74	75	82
	6.0	89	87	86	89	89	84	87	92
(-.5 , .0)	4.0	82	84	72	69	86	83	88	89
	6.0	93	94	86	90	90	89	90	98
(-.5 , .25)	4.0	82	80	76	71	86	85	81	88
	6.0	94	98	88	91	95	92	90	95
(-.5 , .75)	4.0	79	80	78	79	79	83	80	79
	6.0	96	91	86	81	93	90	94	95
$\bar{\%}$	4.0	65.3	65.3	62.1	62.1	65.7	65.7	67.7	68.6
	6.0	83.6	82.3	78.7	79.8	82.0	77.1	82.3	83.7

TABLE 15 (Continued)

(β, π)	split	30-10-20		10-20-30		10-30-20		$\bar{\%}$	$S_{\bar{\%}}$
	$\delta \quad \ominus$	15	30	15	30	15	30		
(.0 , -.5)	4.0	15	22	23	22	27	34	21.9	1.29
	6.0	45	46	54	44	53	52	47.4	1.15
(.0 , .0)	4.0	56	35	53	57	58	57	50.9	1.64
	6.0	67	72	85	76	82	83	75.8	1.31
(.0 , .5)	4.0	64	52	71	67	59	66	60.3	1.45
	6.0	82	79	79	86	83	91	81.8	1.26
(-.25, -.25)	4.0	61	59	68	68	66	73	67.0	1.09
	6.0	83	80	88	88	92	89	83.5	1.14
(-.25, .0)	4.0	73	61	78	82	74	82	73.8	1.64
	6.0	87	79	94	86	94	93	86.9	1.50
(-.25, .5)	4.0	67	62	77	83	80	88	76.7	1.75
	6.0	86	82	96	91	95	95	89.1	1.13
(-.5 , .0)	4.0	77	71	84	87	85	89	81.9	1.83
	6.0	92	82	95	94	95	97	91.8	1.16
(-.5 , .25)	4.0	75	68	83	86	83	90	81.0	1.72
	6.0	88	88	98	93	95	95	92.9	0.92
(-.5 , .75)	4.0	75	68	73	79	75	85	78.0	1.12
	6.0	82	87	96	94	97	92	91.0	1.38
$\bar{\%}$	4.0	62.6	55.3	67.8	70.1	67.4	73.8	65.7	
	6.0	79.1	77.2	87.2	83.6	87.3	87.4	82.2	

TABLE 16

THE $\%_s$ ON LOCAL MAXIMUM FOR FOUR PAIRS OF $(-.5, .75)$
WITH OTHER ALGORITHMS FOR MVN AND MVLN

	split		20-20-20			30-20-10			$\bar{\%}_s$	$S_{\bar{\%}_s}$
	(β, π)	$\delta \rho$.0	.4	.8	.0	.4	.8		
M V N	(.0 , .0)	4.0	46	48	48	54	45	44	47.5	1.45
		6.0	91	78	70	84	85	74	80.3	3.17
	(.0 , .5)	4.0	58	51	35	60	48	37	48.2	4.25
		6.0	80	81	77	83	78	73	78.7	1.43
M V L N	(-.25, -.25)	4.0	57	54	51	53	51	53	53.2	0.91
		6.0	79	77	74	78	78	68	75.7	1.69
	(-.25, .0)	4.0	52	53	57	50	48	50	51.7	1.28
		6.0	78	75	69	74	76	65	72.8	1.99
	split		20-20-20		30-20-10		20-10-30		20-30-10	
	(β, π)	$\delta \theta$	15	30	15	30	15	30	15	30
M V L N	(.0 , .0)	4.0	39	43	48	48	45	44	51	38
		6.0	71	69	77	78	79	78	80	81
	(.0 , .5)	4.0	48	44	57	57	57	48	47	45
		6.0	73	74	76	82	77	77	76	80
M V L N	(-.25, -.25)	4.0	53	52	51	50	47	52	52	56
		6.0	73	73	72	84	74	72	75	78
	(-.25, .0)	4.0	60	53	44	54	57	56	51	49
		6.0	77	71	70	73	70	67	66	84
	split		30-10-20		10-20-30		10-30-20		$\bar{\%}_s$	$S_{\bar{\%}_s}$
	(β, π)	$\delta \theta$	15	30	15	30	15	30		
M V L N	(.0 , .0)	4.0	46	33	39	48	47	42	43.6	1.33
		6.0	68	76	79	78	88	76	77.9	1.69
	(.0 , .5)	4.0	50	48	53	53	57	54	51.3	1.26
		6.0	75	80	77	77	77	77	77.6	0.75
M V L N	(-.25, -.25)	4.0	50	50	49	53	49	51	50.4	0.56
		6.0	68	72	74	72	78	81	74.9	1.11
	(-.25, .0)	4.0	49	47	50	56	58	58	53.0	1.28
		6.0	69	68	70	67	75	76	71.9	1.36

VITA 2

Seong-San Chae

Candidate for the Degree of

Doctor of Philosophy

Thesis: A COMPARATIVE STUDY TO PREDICT THE NUMBER OF
CLUSTERS IN CLUSTER ANALYSIS

Major Field: Statistics

Biographical:

Personal Data: Born in Jeon-Book, Korea, September 2,
1958, the son of Eui-Seock Chae and Soon-Bock Lee.

Education: Graduated from Han-Sung High School, Seoul,
Korea, in February, 1976; received Bachelor of
Science degree with a major in Statistics from
Chung-Ang University, Seoul, Korea, in August,
1983; received the Master of Science degree with
a major in Statistics from Iowa State University,
Ames, Iowa, in December, 1985; completed
requirements for the Doctor of Philosophy degree
at Oklahoma State University in December, 1988.

Professional Experience: Teaching assistant, Department
of Statistics, Oklahoma State University, August,
1986 to December, 1988.