USING PRINCIPAL COMPONENT ANALYSIS PRIOR TO AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS

By

MARILYN ANN GAY SLOAN

Bachelor of Science
University of Oklahoma
Norman, Oklahoma
1966

Master of Arts
University of Oklahoma
Norman, Oklahoma
1973

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
May, 1983

USING PRINCIPAL COMPONENT ANALYSIS PRIOR TO AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS

Thesis Approved:

Thesis Adviser

P. L. Claypool

Lyle D. Broemeling

Kenneth E Case

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to extend my sincere thanks and appreciation to Dr. William D. Warde for suggesting the problem and for his guidance and assistance during the course of this research. I would also like to thank Dr. Lyle D. Broemeling, Dr. P. Larry Claypool, and Dr. Kenneth Case for serving on my advisory committee.

I would also like to express my sincere appreciation to my husband, Norman, and our two daughters, Mary Ann and Sara Sue, for their support, encouragement and sacrifices, especially in time, during this project; and to my parents, Mr. and Mrs. Harold M. Gay, for their interest and encouragement throughout my life.

I am indebted to Susan Bates and Barbara Caldwell for the professional typing of this manuscript.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Perspectives

During the past forty years a considerable amount of work has been done in the development of clustering techniques and factor analysis. A distinction which is often made between these two sets of techniques is that cluster analysis is concerned with the classification of individuals, while factor analytic techniques assess relationships between variables and could be considered to be concerned with the classification of these variables. Many computer programs have been developed to handle the large volume of data and the large matrices involved in both techniques. In education, psychology, agriculture, or other such fields in which clustering and factoring techniques are employed, it has become common in cases where a large number of variables are involved to reduce the number of variables by factor analysis before clustering the data values. Very little work has been done in studying the invariance of clustering methods to the transformations of factor analysis on variables prior to cluster analysis of data points.

In most analyses attention is focused on clustering either data units or variables alone, but not both together.

When data units are clustered, the usual practice is to choose one set of variables, a set of associated weights, and a similarity measure to be applied uniformly for the classification of all data units; but it may be that clusters are characterized by different orientations such as Figure 1a. In cluster 1 variable x_2 can vary widely as long as variable x_1 remains in a narrow range; the reverse relationship is true in cluster 2. The clusters have different descriptions in terms of the variables; therefore, the distance between a given data unit and each cluster centroid should be assessed using a different set of weights for each cluster. Using the same weights for all clusters implies a presumption that all clusters have approximately the same shape and orientation. On the other hand, if one knew enough about the problem to specify the unique weights for each cluster, there probably would not be much need for cluster analysis. Chernoff (1970) has explored the possibilities of constructing a continuing estimate of the shape and orientation of each cluster as data units are allocated and of using this information adaptively to define a unique distance measure for each cluster. Chernoff's work is directed specifically at extending MacQueen's k-means methods. Eddy (1968) and Rohlf (1970) also have considered ways of constructing a different distance measure for each cluster. All three of these discussions are somewhat exploratory in nature and describe potential developments rather than techniques presently suited to widespread use.

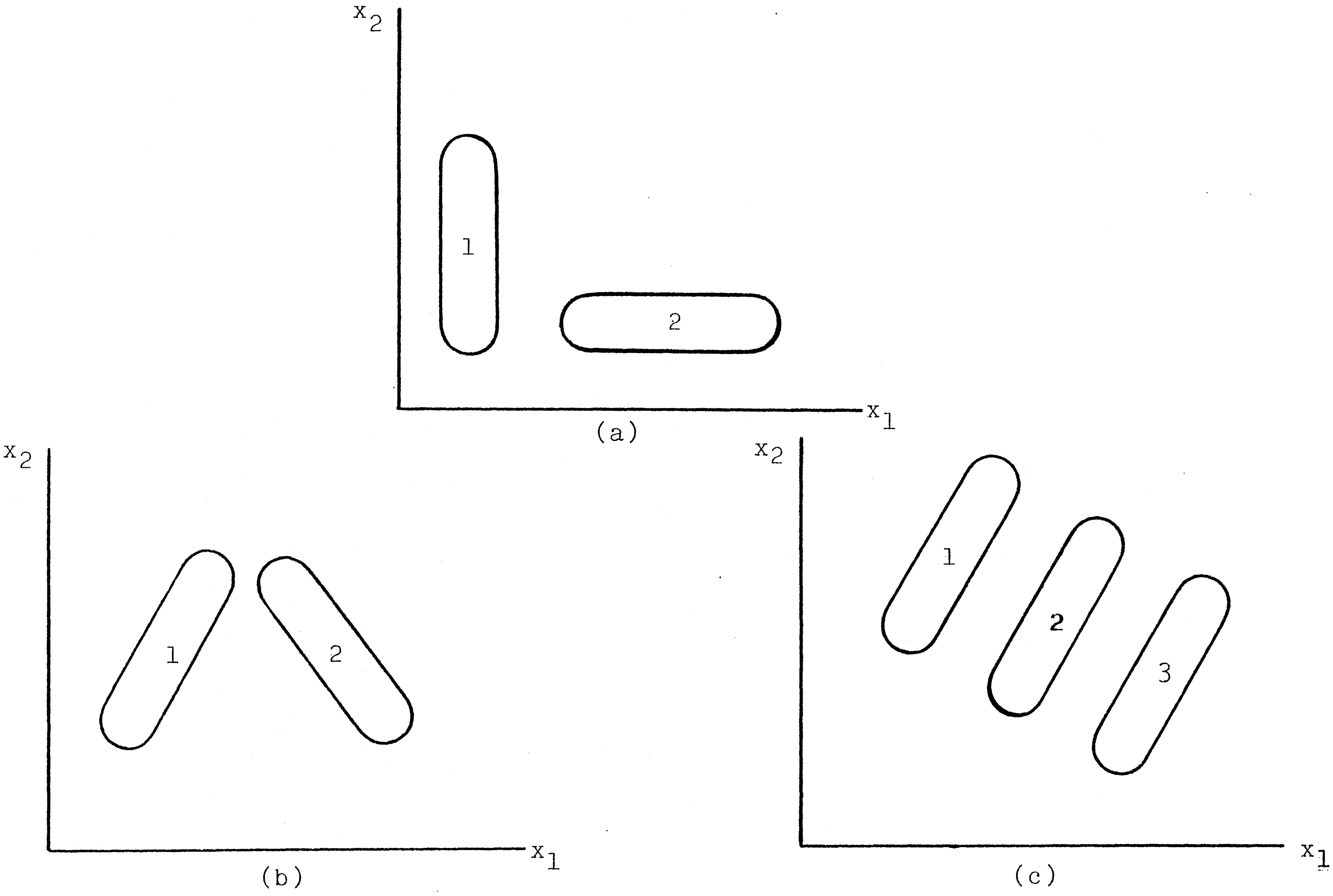

Figure 1. Cluster Orientations

When variables are clustered, there is an implicit assumption that all the elements in the data set share some essential characteristics so that they collectively and individually represent a single population. If the data set actually includes several different clusters of elements, then measures of association between variables will reflect a mixture of effects which may not be representative of the kind of association present within any of the clusters. Figures 1b and 1c illustrate how different mixtures of data unit clusters can conceal important within group relations.

In Figure 1b, variables x_1 and x_2 exhibit a very strong positive correlation in cluster 1 and an equally strong negative correlation in cluster 2; however, if all the elements are taken together in one undifferentiated mass, the computed correlation between the variables would be near zero. In Figure 1c, the relationship within each of the three clusters is one of strong positive correlation between the variables; however, if the three clusters are taken together, the observed relation is one of moderate negative correlation. In both of these cases, the data set as a whole exhibits an apparent relation between the variables which is totally deceptive; far more informative would be the joint knowledge of the cluster structure for the elements and the relations between variables within each cluster. The situations depicted in Figure 1 are easy to depict in two dimensions; but in a data set of 100 variables and 1000 elements, such situations may be difficult to map, even with the aid of systematic clustering methods.

> Such examples are strong evidence that any serious attempt to cluster variables should be preceded by an exploratory clustering of data units to assess the degree of homogeneity within the data set. These remarks also apply to Factor Analysis. (Anderberg, 1973, p. 188)

It appears that adequate clustering of a data set requires considerable insight into the relationships among variables, especially the manner in which the relationships vary from cluster to cluster. On the other hand, an informative cluster analysis of variables requires moderate homogeneity among elements, a requirement that can be satisfied most directly by undertaking a separate analysis for each distinct cluster of elements. Unfortunately, little prior knowledge about the classification of either variables or elements is available in most problems submitted for cluster analysis. Consequently, the task of clustering often seems to be a bootstrap problem in which the data clusters are needed to find the clusters of variables, but variable clusters are needed to find the element clusters, and neither set of clusters is known.

> A possible strategy for dealing with this situation is to undertake a sequential analysis in which elements are clustered at odd stages and variables at even stages until the two sets of clusters converge to a mutually harmonious classification of both variables and elements. The details of using such a strategy on real data remain to be developed. It may prove to be a formidable task to specify adequately these details for a batch process computer; however, it appears that an experienced and informed analyst could achieve a

> simultaneous analysis of both variables and elements (Anderberg, 1973, p. 189).

Several other authors have studied techniques in which each cluster of elements is constructed to have a unique interpretation in terms of variables. Litofsky (1969) and Dubin and Champoux (1970) present techniques based on the special properties of binary variables; Fisher (1968), Hartigan (1972), and Dubin (1971) propose new methods suitable for nominal and interval variables. The whole question of simultaneous clustering of variables and elements has only recently received serious study but offers considerable potential for increased effectiveness of cluster analysis.

There are three related criticisms of principal component analysis under various contexts: effectiveness, scale dependence, and criterion used in choosing the components. Terekhina (1973) gives an example which shows principal components to be much worse than the original variables in separating two subpopulations. These two subpopulations are different in both means and covariances. Mrachek (1972) considers the effect of uninformative variables on the ability of the single linkage and the complete linkage clustering algorithms to provide the correct clustering of a structured data set. This might be related to principal component analysis where the lower eigenvalue factors are uninformative. In other words, the loss of "information" in using only those principal components with relatively large eigenvalues may in fact not be a loss, but the elimination of uninformative components. The dependence of principal component analysis on the rather arbitrary choice of scale has been pointed out by many authors such as Dempster (1969), Kendall (1968), and Sneath and Sokal (1973). In particular Dempster (1969) remarks that the nature of the importance of the first few principal variables is not well-defined. He also indicates that the principal component corresponding to the smallest eigenvalue should be the only one of use in predicting some separate but scientifically important variable. His examples offer some support for his observations. The use of components other than those corresponding to the largest eigenvalues can also be seen in Bennett and Lewis (1978) and Dempster (1969) in the context of outlier detection.

Chang (1980) studied how the effectiveness of the principal component analysis is related to the parameters in the model if the data is a sample from a mixture of two multivariate normal distributions with a common covariance matrix. He has concluded that under some circumstances the most effective set of components is obtained by selecting those components wherein each individually contains relatively larger Mahalanobis distance between the two subpopulations. His equal weight method, rather than the correlation method, to determine principal components applies a scale transformation to the original four non-standardized variables by a diagonal matrix whose diagonal elements were proportional to $1/\sigma_i$.

In some studies, factor analysis has been used as a prelude to cluster analysis, but considerable caution should accompany any such usage. The analyst should confirm that the factors reflect the relationships among variables which are actually observed within the clusters of data elements. The most satisfactory strategy may be to alternate clustering and factor analysis until a harmonious set of clusters and factors is achieved.

Scope of This Study

The main objective of this dissertation involves the study of the effects of applying principal component analysis to variables prior to cluster analyzing observations of a non-supervised random sample from a mixture of normal distributions with a common covariance matrix. Attention is focused on some agglomerative hierarchical clustering techniques. The research is then extended to the study of random samples from multivariate multinomial populations with a common covariance matrix.

Chapter II contains a brief discussion of classification techniques and a general formulation for agglomerative clustering methods. A discussion of principal component analysis is contained in Chapter III. A comparative statistic is defined in Chapter IV. In Chapter V the design of the test procedure for the multivariate normal samples is discussed while the results of this procedure are discussed in Chapter VI. In Chapters VII and VIII discussions of the multivariate multinomial test procedures and results are presented.

CHAPTER II

PURPOSE AND DEVELOPMENT OF CLUSTERING TECHNIQUES

Classification and Cluster Analysis

The most commonly used term for techniques which seek to separate data into constituent groups is cluster analysis. Although several authors use the term cluster analysis for techniques which seek to group variables, such techniques are generally used for the grouping of the objects or individuals under investigation. Kendall and Stuart (1963) propose that the term cluster analysis be used for techniques which group variables and classification for techniques which group individuals. This can, however, lead to confusion since some authors use the term classification to describe techniques for assigning individuals to groups having *a priori* labels. (For example, discriminant analysis.)

Primitive components of set theory are element and set; parallel concepts in cluster analysis are the elements to be clustered and the set consisting of these elements. In general terms, the elements to be clustered have been called objects, individuals, patterns, and (by Sneath and Sokal, 1973) operational taxonomic units (OTUs). The elements to be clustered shall in this paper be referred to as data points, and each data point shall be represented by a 1 X p vector, X_i , where

$$X_i = (x_{i1} , x_{i2} , \dots , x_{ip})$$

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

$$X' = \left\{ X_1 , X_2 , \dots , X_N \right\} .$$

Obviously, the object space is embedded in Euclidean p-space. Thus, if E_p represents Euclidean p-space, then $X \subseteq E_p$.

A popular conceptualization of the object space is the data matrix which is formed by stacking the data points as rows of a matrix. Letting $X_{N,p}$ represent the data matrix, where N is the number of data points and p is the number of variables, then

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

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

Definition 1. A cluster, Y_k , is any nonempty subset of the object space. Symbolically, $Y_k \subseteq X$ which means that if $X_i \varepsilon Y_k$, then $X_i \varepsilon X$.

Thus, a cluster is simply a collection of data points.

Definition 2. A clustering, Y , is any partition of the object space. Symbolically, $Y = \left\{ Y_1, Y_2, \ldots, Y_K \right\}$ is a partition of X , if the following three conditions hold:

(i) For every $Y_k \varepsilon Y$, $Y_k \neq \phi$.

(ii) If $Y_k \varepsilon Y$, $Y_m \varepsilon Y$, and $Y_k \neq Y_m$, then $Y_k \bigcap Y_m = \phi$.

(iii) $\displaystyle\bigcup_{k=1}^{K} Y_k = X$.

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

A clustering of N data points can consist of K = 1, 2, ... , N clusters. The number of clusters contained in a clustering shall be termed the size of the clustering, and this designation will be incorporated into the general notation for a clustering by the use of a superscript. For example, if clustering Y contains K clusters, then Y^K denotes a clustering of size K . The set of all possible clusterings of the object shall be denoted by $\mathcal{Y}$. The fact that even for small values of N , the cardinality of $\mathcal{Y}$ is quite large has motivated the development of a multitude of clustering methods, not all of which are distinct. In very general terms, a clustering method consists of a criterion and a technique in which case 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).

Some of the preceding discussion was taken from DuBien (1976) and is included here for comprehension and completeness since the basic design of the test procedure as defined in Chapter V is an augmentation of DuBien's test procedure.

Objectives of Cluster Analysis Techniques

The goals of various users of clustering techniques are frequently dissimilar. Once this is realized it is easier to see why such a variety of clustering techniques exist. Ball (1971) lists seven possible uses of clustering techniques, these being as follows:

(i) Finding a true typology,

(ii) Model fitting,

(iii) Prediction based on groups,

(iv) Hypothesis testing,

(v) Data exploration,

(vi) Hypothesis generating,

(vii) Data reduction.

For example, in many fields the research worker is faced with a great bulk of observations which are quite intractable unless classified into manageable groups which, in some sense, can be treated as units. Clustering techniques can be used to perform this data reduction, reducing the information on the whole set of, say, N individuals to information about, say, g groups (where hopefully g is much smaller than N). In this way it may be possible to give a more concise and understandable account of the observations under consideration. In other words, simplification with minimal loss of information is sought.

Cluster analysis may also be used to generate hypotheses concerning the nature of the data. When it is used for this purpose, the hypothesis must be capable of being tested and any test must depend on new observations and cannot use the data from which the hypothesis was generated. As Williams and Dale (1965, p. 235) state: "Generation of the hypothesis may not be used as its own evidence."

In some cases clustering techniques may be useful in shedding light on previously made hypotheses. For example, in psychiatry there has long been controversy over the classification of depressed patients. The issues involved here have been reviewed on a number of occasions (Grinker et al., 1961; Kiloh, 1965; Mendels and Cochrane, 1968). Several attempts have been made to establish validity of classifying such patients into endogeneous and reactive or neurotic groups. Many various statistical techniques have been employed including factor analysis, principal component analysis, and multiple regression analysis, but more recently the problem has been tackled with some success by cluster analysis techniques (see Pilowsky et al., 1969, and Paykel, 1971).

In some investigation cluster analysis methods may be used to produce groups which form the basis of a classification scheme useful in later studies for predictive purposes of some kind. For example, a cluster analysis applied to data consisting of a sample of psychiatric patients may produce groups of patients who react differently when treated with some drug, thus enabling the investigator to decide whether a drug is suitable for a particular type of patient. Such a procedure is used by Paykel (1972) in an investigation of the usefulness of amitriptyline in the treatment of depression.

A General Formulation for Agglomerative Clustering Algorithms

In general the initial raw data collected by the investigator consist of an N X p matrix of measurements, say X , where

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

and in which x_{ij} is the score on the j-th variable for the i-th individual or entity. 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 . Thus, the properties and some examples of distance measures will be established before a general formulation for agglomerative clustering algorithms is given.

In very general terms, a measure of distance, d , on some arbitrary set, S , is a real-valued function on S X 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 point X_i and data point X_j , the hierarchy of properties for a measure of distance is depicted in Definitions 1, 2, and 3.

Definition 1. A semi-metric on the subject space, X , is a function

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

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.,

$$\forall X_i , X_j \in X , d_{ij} \geq 0$$

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

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

$$\forall X_i , X_j \in X , d_{ij} = d_{ji} .$$

Definition 2. 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.,

$$\forall X_i, X_j, X_k \varepsilon X,$$

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

Definition 3. 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.,

$$\forall X_i, X_j, X_k \varepsilon X,$$

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

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 derivations presented in this chapter, 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 ℓ_m. Since 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.

Agglomerative clustering methods are some of the oldest and most frequently used cluster methods. An agglomerative clustering method may be characterized as proceeding sequentially by joining pairs of clusters from the partition which consists of each data point grouped as a single cluster to the partition which consists of all data points grouped together in a single cluster (if no stopping rule is provided). An important concept in the definition of an agglomerative clustering method is a hierarchy.

Formal definitions for hierarchy and agglomerative clustering method are given as Definitions 4 and 5, respectively, which assume that there are N data points.

Definition 4. A hierarchy, H, on the object space is an ordered sequence of nested clusterings. Symbolically,

$$H : \quad Y^N , \; Y^{N-1} , \; \ldots , \; Y^2 , \; Y^1 ,$$

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

One useful visualization of a hierarchy is a tree diagram which is often called a dendrogram in cluster analysis applications. Summarizing, 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 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, $Y_j \varepsilon Y^K$ and they are deemed "closest", then $Y_i \bigcup Y_j \varepsilon Y^{K-1}$.

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

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) ; for future reference, the components of this triple have been carefully defined in this section. When, in general, a clustering method consists of a criterion and a technique, an agglomerative clustering method may be more specifically viewed as consisting of a measure of similarity or dissimilarity (usally a metric) 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 (metric, algorithm).

From this brief background, 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 necessary.

In the application of an agglomerative clustering method to a set of data, initially, the distance between each pair of data points, X_i and X_j, is computed using some measure of distance, d, which is at least semi-metric. Since d is at least semi-metric, the resultant set of distances may be denoted by

$$D = \left\{ d_{ij} \mid i < j \ , \ \ i = 1, 2, \ldots , \ \ N-1 \ , \ j = 2, 3, \ldots , N \right\} .$$

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 . However, 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 a data point X_i . Hence, the proces 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 of Y^N .

The role of the agglomerative clustering algorithm is to sequentially impose a measure of distance on each clustering, Y^K , $K = 1, 2, \ldots , N-1$, 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} (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.

To clarify the notation, since Y^K , $K = 1, 2, \ldots , N$, 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 $Y_i , Y_j \in Y^K$, $K = 1, 2, \ldots , N$. 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 joining of two clusters. As a further simplication of the notation, if Y_i , $Y_j \varepsilon Y^K$ join at distance d_{ij} to form Y^{K-1} , then $Y_{(ij)}$ will denote the new cluster, i.e.,

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

and d_{ij} shall be termed the joining distance for clustering Y^{K-1} .

For any clustering Y^K , if the distances d_{ij} , d_{ik} , and d_{jk} between pairs of clusters are obtained from some source (recursively from clustering Y^{K+1} , $K \neq N$) , then the distance between the new cluster $Y_{(ij)}$ and any other cluster $Y_k \varepsilon Y^K$ can be computed from the following formula:

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta\, d_{ij} + \gamma |d_{ik} - d_{jk}| , \quad (2.1)$$

where α_i , α_j , β , and γ are specified parameters, defining the particular member of the family of agglomerative clustering algorithms (Lance and Williams, 1966). Beginning with the initial distance matrix obtained by imposing d on X , Equation (2.1) is applied recursively to obtain each clustering in the hierarchy.

Equation (2.1) characterizes a family of agglomerative clustering algorithms so that for each choice of the parameter quadruple ($\alpha_i, \alpha_j, \beta , \gamma$) , a particular member of this family of agglomerative clustering algorithms is specified. This thesis will study the effect of applying principal component analysis in conjunction with eighteen members of this family. This recurrence formula makes the computer implementation of agglomerative methods relatively easy.

CHAPTER III

PRINCIPAL COMPONENT ANALYSIS

Factor Analysis and Principal Components

Factor analysis, like all statistics, is a branch of applied mathematics. Thus, it is used as a tool in the empirical sciences. One of the objectives of statistical theory is to provide a scientific law, or mathematical model, to explain the underlying behavior of the data. Some simple examples include: (1) a linear regression for the prediction of school success from three entrance examinations; (2) a mathematical curve, such as the normal distribution or one of the Pearson family of curves, for the explanation of an observed frequency distribution; (3) a Chi-square test of significance for the independence of such classifications as "treated or not treated with a certain serum," and "cured or not cured." Such laws make allowance for random variations of the observed data from the theoretically expected values. It is conceivable that any one of several, quite different, mathematical models may provide an equally good fit or explanation of a set of data.

Principal components are linear combinations of random variables which have special properties in terms of variances. For example, the first principal component is the normalized linear combination (i.e., the sum of squares of the coefficients being one) with maximum variance. In effect, transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system that has inherent statistical properties. The principal components turn out to be the characteristic vectors of the covariance matrix. Thus the study of principal components can be considered as putting into statistical terms the usual developments of characteristic roots and vectors (for positive semidefinite matrices).

From the point of view of statistical theory, the set of principal components yields a convenient set of coordinates, and the accompanying variances of the components characterize their statistical properties. In statistical practice, the method of principal components is used to find the linear combinations with large variances. In many exploratory studies the number of variables under consideration is too large to handle. Since it is the deviations in these studies which are of interest, one way of reducing the number of variables to be treated is to discard the linear combinations which have small variances and study only those with large variances.

Factor Analysis Model - Principal Components

A statistical study typically involves a group of individuals with some common attributes. The term "individual" is used here in a generic sense to stand for such objects or entities as persons, census tracts, businesses, etc. Measurements made on such individuals, or attributes of these entities, are designated simply as variables.

It is the object of factor analysis to represent a variable, z_j , in terms of several underlying factors, or hypothetical constructs. The simplest mathematical model for describing a variable in terms of several others is a linear one. However, there are still several alternatives within the linear framework, depending on the objective of the analysis. One distinction between two objectives can be made immediately, namely: (1) to extract the maximum variance,; and (2) to "best" reproduce the observed correlations.

An empirical method for the reduction of a large body of data so that a maximum of the variance is extracted was first proposed by Karl Pearson (1901) and fully developed as the method of Principal Components, or component analysis, by Harold Hotelling (1933). The model for component analysis is simply:

$$z_j = a_{j1} F_1 + a_{j2} F_2 + \dots + a_{jp} F_p \quad (j = 1,, 2, \dots, p) ,$$

where each of the p observed variables is described linearly in terms of p new uncorrelated components F_1, F_2, ... , F_p. When the point representation of a set of variables is employed, the loci of uniform frequency density are essentially concentric, similar and similarly situated ellipsoids. The axes of these ellipsoids correspond to the principal components. The method of component analysis, then, involves the rotation of coordinate axes to a new frame of reference in the total variable space -- an orthogonal transformation wherein each of the p original variables is describable in terms of the p new principal components.

An important feature of the new components is that they account, in turn, for a maximum amount of variance of the variables. More specifically, the first principal component is that linear combination of the original variables which contributes a maximum to the residual variance; and so on until the total variance is analyzed. The sum of the variances of all p principal components is equal to the sum of the variances of the original variables. For a practical problem only a few components might be retained, especially if they account for a large percentage of the total variance. However, all the components are needed to reproduce the correlations among the variables.

Since the method is so dependent on the total variance of the original variables, it is most suitable when all the variables are measured in the same units. Otherwise, by change of units or other linear transformations of the variables, the ellipsoids could be squeezed or stretched so that their axes (the principal components) would have no special meaning. Hence, it is customary to express the variables in standard form, i.e., to select the unit of measurement for each variable so that its sample variance is 1. Then the analysis is made on the correlation matrix, with the total variance equal to p.

The components obtained from S, the sample covariance matrix, and R, the correlation matrix, are in general not the same, nor is it possible to pass from one solution to the other by a simple scaling of the coefficients. Most applications of the technique have involved the correlation matrix, as if in keeping with the usage established by factor analysts. If the responses are widely different in magnitude (age in years, weight in kilograms, and biochemical excretions in a variety of units, to cite one plausible case), linear compounds of the original quantities would have little meaning, and the standardized variates and correlation matrix should be employed. Conversely, if the responses are reasonably commensurable, the covariance form has a greater statistical appeal, for the i-th principal component is that linear compound of the responses which explains the i-th largest portion of the total response variance, and maximization of such total variance of standard scores has a rather artificial quality (Anderson, 1971). Furthermore, as Anderson has shown, the sampling theory of components extracted from correlation matrices is exceedingly more complex than that of covariance-matrix components.

Suppose that the random variables $X_1, X_2, \ldots, X_p$ of interest have a certain multivariate distribution with mean vector $\underset{\sim}{\mu}$ and covariance matrix Σ. We assume, of course, that the elements of $\underset{\sim}{\mu}$ and Σ are finite. The rank of Σ is $r \leq p$, and the q largest characteristic roots

$$\lambda_1 > \lambda_2 \ldots > \lambda_q$$

of Σ are all distinct.

Definition 1. The j-th principal component of the sample of p-variate observations is the linear compound

$$Y_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p$$

whose coefficients are the elements of the characteristic vector of the sample correlation matrix R corresponding to the j-th largest characteristic root λ_j. If $\lambda_i \neq \lambda_j$, the coefficients of the i-th and j-th components are necessarily orthogonal; if $\lambda_i = \lambda_j$, the elements can be chosen to be orthogonal, although an infinity of such orthogonal vectors exists. The sum of the characteristic roots will be

$$\text{tr } R = p$$

and the proportion of the total "variance" in the scatter of dimensionless standard scores attributable to the j-th component will be λ_j / p. The sum of the squared correlations of the responses, $a_{ij} \sqrt{\lambda_j}$, on that component will of course be the component variance λ_j.

We have stated that one important use of the principal-component technique is that of summarizing most of the variation in a multivariate system in fewer variables. Unless the system is of less than full rank, some variance will always be unexplained if fewer than p components are taken to describe the system. In practice one usually knows from earlier studies, the subject-matter nature of the data, or even the pattern of the correlations in R that a certain minimum number of components with large and distinct variances should be extracted. Beyond that number, components might be computed until some arbitrarily large proportion (perhaps 75 percent or more) of the variances has been explained. If that proportion cannot be explained by the first four or five components, it is usually fruitless to persist in extracting vectors; for even if the later characteristic roots are sufficiently distinct to allow easy computation of the components, the interpretation of the components may be difficult if not impossible (Morrison, 1976).

CHAPTER IV

DEFINITION OF A COMPARATIVE TEST STATISTIC

Since the primary objective of this thesis is to compare results of clustering-principal component methods, a comparative statistic is required to quantify each comparison. Rand's (1969, 1971) c statistic is a very general and versatile statistic which may be used to compare our clustering results based on how the object space is partitioned. Essentially, c measures the similarity between clusterings derived from any source. However, if two clusterings are produced by the application of two different clustering methods to the same object space, then c is a measure of the similarity between the two clustering methods through their resultant clusterings.

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 unequivocably 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 (p. 847).

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

To facilitate the definition of the c statistic, Definition 1 concerning the similar assignment of point-pairs is given.

Definition 1. Given an object space X consisting of N data points, $X_1, X_2, \ldots, X_N$, and two clusterings of X, $Y = \{Y_1, Y_2, \ldots, Y_{K_1}\}$ and $Y' = \{Y'_1, Y'_2, \ldots, Y'_{K_2}\}$, 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) $\exists$ k and k' $\ni X_i, X_j \epsilon Y_k$ and $X_i, X_j \epsilon Y'_{k'}$;

(ii) $\exists$ k and k' $\ni X_i \epsilon Y_k, Y'_{k'}$, and $X_j \not\epsilon Y_k, Y'_{k'}$.

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 Defition 2.

Definition 2. Given an object space X consisting of N data points, $X_1, X_2, \ldots, X_N$, and two clusterings of X, $Y = \{Y_1, Y_2, \ldots, Y_{K_1}\}$ and $Y' = \{Y'_1, Y'_2, \ldots, Y'_{K_2}\}$, then the c statistic between Y and Y' is defined as follows :

$$c (Y , Y') = \frac{\sum_{i<j} n_{ij}}{\binom{N}{2}} , \qquad (4.1)$$

where

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

Hence, c is a measure of similarity on $\mathcal{Y}$, 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 X 2 contingency tables. To demonstrate the equivalence relationship between Rand's c statistic and the simple matching coefficient, a particular form of the sample 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 3.

Definition 3. A vector $V = (v_1, v_2, \ldots , v_n)$ is a binary vector if and only if for each $i = 1, 2, \ldots, n$, $v_i = 1$ or $v_i = 0$.

To compute the simple matching coefficient, it is necessary to define a match between two binary vectors as indicated in Definition 4.

Definition 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 $u_i = v_i$.

Definition 5. The simple matching coefficient between two binary vectors, U and V , of length n is given by

$$s (U , V) = m / n \quad ,$$

where m is the number of matches. 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 is 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 6.

Definition 6. The binary vector

$$U = (u_{12}, u_{13}, \dots , u_{1n}, u_{23}, \dots , u_{2n}, \dots , u_{n-1,n}),$$

is a binary representation of clustering $Y = \{Y_1, Y_2, \dots , Y_K\}$ if and only if for each $i < j$,

$$u_{ij} = \begin{cases} 1 \ , \text{ if } \exists \, k \ni X_i \, , X_j \in Y_k \\ 0 \ , \text{ otherwise.} \end{cases}$$

Therefore, if U is a binary representation of clustering Y, and V is a binary representation of clustering Y', then

$$s (U , V) = \frac{m}{n} = \frac{m}{\binom{N}{2}} = \frac{\sum\limits_{i<j} n_{ij}}{\binom{N}{2}} = c (Y , Y') \ .$$

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

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

(i) c is a measure of similarity with $0 \leq c \leq 1$;

(ii) $1 - c$ is a metric on the set of all possible clusterings of X ;

(iii) 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 $\mathcal{Y}$.

Since c is a random variable, under certain assumptions, c possesses a probability distribution. However, Rand (1969, p. 39) comments on the distribution of c as follows: "This is a complicated distribution, and analytic expression of it is not attempted here." 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 ; a clustering method maps X into $\mathcal{Y}$; and

$$c: \mathcal{Y} \times \mathcal{Y} \longrightarrow [0, 1] .$$

Recent studies have been done by DuBien and Warde (1983).

CHAPTER V

DESIGN OF TEST PROCEDURE

A Two Parameter Sub-Family of Agglomerative Clustering Algorithms

The design of the test procedure follows that suggested by DuBien (1976) and is augmented to include principal component techniques.

A two-parameter sub-family of agglomerative clustering algorithms may be derived from the four-parameter family discussed in section II.4 by placing a suitable set of constraints on the parameters given in Equation (2.1). If the constraints are given by

$$\alpha_i = \alpha_j = \alpha \quad ,$$

$$\alpha_i + \alpha_j + \beta = 1 \quad ,$$

then a member of the four parameter family of agglomerative clustering algorithms that has parameter values which satisfy the constraints can be represented by the ordered pair (β , γ).

Without loss of generality, it will be assumed that

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

Noting that the two constraints imply that

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

then equation (2.1) becomes

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

Since

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

then

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

Thus, Equation (3.2) characterizes a sub-family of agglomerative clustering algorithms which shall be referred to as the (β , γ) family, and each member of this sub-family shall be referred to as a (β , γ) algorithm. Consequently, it is possible to represent each member of the (β , γ) family of agglomerative clustering algorithms as a point in the (β , γ) Cartesian coordinate plane. It is also worth noting that single linkage, complete linkage, unweighted average linkage, and the flexible strategy given by Lance and Williams (1967) are members of the (β , γ) family of agglomerative clustering algorithms, namely, (0., -.5), (0., +.5), (0., 0.), (-.25, 0.), respectively.

The eighteen agglomerative clustering algorithms chosen for this study form natural groups of three or six algorithms. The rationale behind the choice of these algorithms is discussed by DuBien (1976) and DuBien and Warde (1979). Thus, the (β , γ) values which define the eighteen agglomerative clusterings are conveniently delineated in three groups of six algorithms as follows:

(1) $\beta = 0.0$ with $\gamma = -.5, -.25, \ldots, .75$;

(2) $\beta = -.25$ with $\gamma = -.5, -.25, \ldots, .75$;

(3) $\beta = -.50$ with $\gamma = -.5, -.25, \ldots, .75$.

In this study we will compare the effect of controlled structural changes within the data on the clusterings obtained from these clustering algorithms alone to the clusterings obtained from performing principal component analysis prior to applying the clustering algorithms.

Definition of Structural Parameters

A clustering method is purported to be a functional mechanism for finding or "retrieving" "natural" structure within data. Hence, the degree to which a clustering method retrieves known structure within generated data is an important characteristic of the clustering method. To quantify the retrieval ability of a clustering method, N data points are generated from K "well-separated" populations, and the clustering of size K which groups together data points which are generated from the same population is denoted by Y . In other words, Y represents the "true" structure of the population. If Y' denotes the clustering which results from applying a specific clustering method to the N data points and if Y'' denotes the clustering which results from applying principal component analysis and a specific clustering method to the N data points, then the values of c(Y, Y') , c(Y, Y'') , and c(Y', Y'') are measures of the "retrieval" ability of the clustering methods (subject to the random variation in the generated data).

For convenience, the important considerations 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 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 delineated as follows:

1. N , the number of data points in X ;
2. p , the number of variables defining each data point; 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 type of population or the probability distribution from which each of the K populations of data points are generated;
5. μ_k , k = 1, 2, ... , K , the mean vector for each population of data points;

6. Σ_k , $k = 1, 2, \ldots, K$, the variance-covariance structure for each population of data points;

7. δ_i , $i = 1, 2, \ldots, \binom{K}{2}$, the distance between each pair of the p variables of the population mean vectors;

8. The split or n_k , $k = 1, 2, \ldots, K$, the number of data points generated from each population of data points;

9. m , the number of principal components to be used.

10. η , the amount of "noise" in the variance-covariance matrix.

In any comparative study of clustering methods, some of the structural parameters in the set of possible structural parameters must remain fixed, and 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.

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 variable structural parameters. For the purposes of this study, the probability from which each of the K populations of data points was generated was fixed to be multivariate normal (MVN) . A brief discussion of the basic generating procedure used should suffice. For the purpose of efficient discussion, MVN populations with the same variance-covariance matrix will be termed "similar." MVN vectors may be generated from a population having a mean vector of zero and any specified positive definite, symmetric variance-covariance matrix by calling subroutine GGNRM from the IMSL catalogued programs. Generation from other similar MVN populations may be accomplished by adding a fixed constant vector to each vector generated from the GGNRM subroutine. This procedure simulates the generation of vectors from a MVN population with a mean vector equal to the fixed constant vector which was added to each of the generated vectors and the same variance-covariance matrix as was originally specified.

Because of the necessity to operate within certain cost constraints, 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:

(i) $N = 12$;

(ii) $p = 10$;

(iii) $K = 2$.

The choice of $N = 12$ was arbitrary subject to its divisibililty by two. N was later allowed to vary from 10 to 70. However, since the primary purpose of the comparative study was to investigate the effect of applying principal component analysis prior to clustering the data points, $p = 10$ was chosen so that we would have several variables combining to form more than one principal component. The choice of $K = 2$ was minimum for clustering into two populations.

The correlation matrix was chosen to have the following block diagonal structure:

$$\Sigma_k = \; = \begin{bmatrix}
1. & & & & & & & & & \\
\rho & 1. & & & & & & & & \\
\rho & \rho & 1. & & & & & & & \\
\rho & \rho & \rho & 1. & & & & & & \\
\eta & \eta & \eta & \eta & 1. & & & & & \\
\eta & \eta & \eta & \eta & \rho & 1. & & & & \\
\eta & \eta & \eta & \eta & \rho & \rho & 1. & & & \\
\eta & \eta & \eta & \eta & \eta & \eta & \eta & 1. & & \\
\eta & \eta & \eta & \eta & \eta & \eta & \eta & \rho & 1. & \\
\eta & \eta & \eta & \eta & \eta & \eta & \eta & \rho & \rho & 1.
\end{bmatrix}$$

This type of structure was chosen in order to produce three principal components of interest: one a combination of the first four variables; one a combination of the next three variables; and one a combination of the last three variables.

The number of principal components to be used was set by design of the variance-covariance matrix at three. In the computer program written to perform the computations needed for this study, the actual value of m was determined by choosing the principal components whose associated eigenvalues were greater than or equal to one. The eigenvalues for the first three components were the only values greater than or equal to one, and the first three principal components accounted for over 70 percent of the variance.

The three structural parameters subject to controlled variation in the comparative study were ρ , δ_i , and η . To facilitate the controlled change of the structure parameters δ_i , $i = 1, 2, \ldots , \binom{K}{2}$, it is apropos to quantify the distance between population mean vectors by a single structural parameter, $\delta * \sqrt{p}$; i.e., $\forall\, i = 1, 2, \ldots , \binom{K}{2}$, $\delta_i = \delta$. The settings for the structural parameter , the distance between each variable of the mean vectors, were set at $\delta = 1.$, $\delta = 1.5$, and $\delta = 2.0$; these three settings were deemed worthy of further consideration for populations separated by three to seven standard deviations as suggested by DuBien (1976). Less than a three standard deviation separation tended to cause difficulty in determining clusters, while more than a seven standard deviation separation tended to reproduce the known population clusters almost surely. The value of ρ was allowed to vary from .6 to .9 , in increments of .1, but did not vary within the diagonal blocks or between blocks. The amount of noise, η , was allowed to vary from .1 to .4 in increments of .1 . Taken together this yielded 16 combinations of (ρ, η) for study.

Test Procedure

One of the basic considerations in designing the comparative study was the choice of a logical running sequence which would produce each of the sets of results necessary to compare the clustering methods with the clustering methods after principal component analysis was employed with respect to their ability to "retrieve" the generated data structure. Each setting of the triple (ρ, δ, η) of variable structural parameters characterizes a different replication (rep) of the comparative study. For each setting of the triple (ρ, δ, η), the following sequence of steps was utilized to generate values of $c(Y, Y')$, $c(Y, Y'')$, and $c(Y', Y'')$ for the eighteen (β, γ) clustering algorithms chosen.

1. An object space X of data points was generated for the complete set of structural parameters;

2a. The Euclidean distance between each pair of data points in X was computed and stored in standard lower triangular matrix order by rows as the vector D;

2b. Principal component analysis was applied to X, the principal components whose corresponding eigenvalues were greater than or equal to one were chosen to transform the data points of X, and Euclidean distance between each pair of transformed data points was computed and stored in standard lower triangular matrix order by rows as the vector D1 ;

3a. Each of the eighteen agglomerative (β , γ) clustering algorithms was applied to D to produce a hierarchy, H_a , a = 1, 2, ... , 18 ;

3b. Each of eighteen agglomerative clustering algorithms was applied to D1 to produce a hierarchy, $H1_a$, a = 1, 2, ... , 18 ;

4. For each of the eighteen agglomerative clustering algorithms, the two-cluster clusterings, (Y') and (Y'') were chosen as the representative clusterings from H_a and $H1_a$, where a = 1, 2, ... , 18 ;

5. Each of the representative clusterings, $(Y')_a$ and $(Y'')_a$, a = 1, 2, ... , 18 , was compared by means of the c statistic to clustering Y of size two, which represents the "true" structure of the data.

6. Each of the representative clusterings, $(Y')_a$, a = 1, 2, ... , 18 , was compared by means of the c statistic to the representative clusterings, $(Y'')_a$, a = 1, 2, ... , 18 .

Thus, by means of the above sequence of steps, values of c(Y, Y') , c(Y, Y'') , and c(Y', Y'') were computed for each of the eighteen agglomerative clustering methods. For each setting of the triple (ρ , δ , n) , the above sequence of steps was replicated 100 times, and the following statistics were computed for each of the eighteen agglomerative clustering methods for each of the three comparisons:

1. $\overline{c}$, the sample mean of the c statistic for the sample of 100 reps;
2. s_c , the sample standard deviation for the 100 c values;
3. The % of the 100 clusterings which correspond exactly with the generated data structure, i.e., the number of times that c(Y1, Y2) was equal to one in the 100 reps.

Consequently, for each setting of the triple (ρ, δ, η) of variable structural parameters and for each of the eighteen agglomerative clustering methods, three triples ($\overline{c}$, s_c, %) result from 100 reps to quantify the "retrieval" ability for each of the agglomerative clustering methods alone and the "retrieval" ability of each of the agglomerative clustering methods after principal component analysis has been applied.

CHAPTER VI

DISCUSSION OF RESULTS FROM MULTIVARIATE NORMAL SAMPLES

Tables I-IX in the Appendix give the results from the comparative study of eighteen agglomerative clustering methods. Although eighteen methods were studied, only single link, group average, complete link, and three others, were summarized in order to save space. The results of the use of the other agglomerative clustering methods followed the same trend as these three.

In these nine tables, the results are given in the form of $\overline{c}$ computed over 100 reps for each setting of the triple variable structural parameters (ρ , δ , η) and for each of the six agglomerative clustering methods mentioned above. The three $\overline{c}$ values, $\overline{c}$ (Y, Y') , $\overline{c}$(Y, Y'') , $\overline{c}$(Y', Y'') are tabulated. Although Euclidean distance was used, an observed difference or similarity among the agglomerative clustering algorithms should be interpreted as a difference or similarity among the agglomerative clustering methods formed by combining the same algorithms with Euclidean distance. The results from the comparative study are also not independent of the fixed structural parameters which were specified in the previous chapter, but the results will be discussed in terms of the variable structural parameters. Thus, all results from the comparative study will be discussed in terms of changes in the variable structural parameters (ρ , δ , η) and changes in (β , γ) which defines the agglomerative clustering algorithm.

Tables I, II, and III in the Appendix display the results for the six algorithms for $\rho = .6$, $.7$, $.8$, $.9$, with $\delta = 1.0$, 1.5, 2.0 , and $\eta = 0.0$. The $\overline{c}$ values calculated show that, since there is essentially no difference between $\overline{c}(Y, Y')$ and $\overline{c}(Y, Y'')$, the difference between clustering methods was due to the agglomerative clustering algorithm chosen rather than to the use of principal component analysis prior to applying the clustering method. Applying principal component analysis prior to clustering produced clusterings comparable to those produced by the use of clustering alone as can be seen by the high $\overline{c}(Y', Y'')$ values in the ranges of .8 and .9 .

Tables IV-VII in the Appendix display the results for the six algorithms as in Tables I-III in the Appendix but rearranged in different order to show how the $\overline{c}$ values change when δ is allowed to vary as ρ is held constant. We see again that there is essentially no difference between $\overline{c}$ calculated from the clustering method and $\overline{c}$ calculated from the clustering method when principal component analysis is applied before clustering.

In Tables I-VII in the Appendix the noise was held constant at $\eta = 0.0$. Table VIII in the Appendix gives the results when ρ is held constant at .7 and δ is held constant at 1.5 but η is allowed to vary, $\eta = 0.0$, .1, .2, .3, .4 . Table IX shows the results when a small amount of noise is permitted, $\eta = 0.1$, δ is held constant at 1.5 but ρ is allowed to vary, $\rho = .6, .7, .8, .9$. Again we see that clusterings determined by applying principal component analysis prior to the clustering methods agree closely with those obtained by using clustering methods alone.

When N was allowed to vary from ten to seventy while the other structural parameters were held constant, very small decreases in the $\overline{c}$ values were observed, but a marked increase in the amount of computer time needed to execute the procedure was demonstrated.

Thus, we can see that under the design described in the previous chapter, essentially the same clusterings are retrieved whether principal component analysis is applied prior to applying the clustering algorithms or whether it is not. This is the desirable result since it is often necessary to apply principal component analysis initially to reduce the number of variables to be used in later computations.

CHAPTER VII

EXTENSION TO MULTIVARIATE MULTINOMIAL SAMPLES

Fundamental Concepts With Some Basic Definitions

The comparative study is now extended to the study of samples from multivariate multinomial distributions. Binomial variables are first considered; then a generalization to multinomial variables is made.

Before the test procedure is defined, it is necessary to offer a few definitions for distinction.

Definition 1. A Bernoulli trial is an experiment which has two possible outcomes, generally called success and failure. The sample space for a Bernoulli trial will in general be written $S = \{0, 1\}$, where 0 indicates "failure" and 1 indicates "success." Many different examples of Bernoulli trials can be cited: a flip of a single coin resulting in either a head or a tail, the flight of a missile (if we call it simply a success or not), performance of a student in a particular course (pass or fail), or performance of an athletic team (win or not win). Any chance mechanism whose outcomes are grouped into two classes can be looked at as being a Bernoulli trial.

A frequently used notation is

$$P(\{0\}) = p \quad ,$$
$$P(\{1\}) = q = 1 - p \ ;$$

the quantity p is of course free to take on any value in the interval from 0 to 1 , inclusive, for various types of Bernoulli trials.

When an experiment consists of n (a positive integer) repeated independent Bernoulli trials, the sample space for this experiment then is the Cartesian product of the sample spaces of the individual trials. Thus $S = S_1 \times S_2 \times S_3 \times \ldots \times S_n$ where $S_i = \{0, 1\}$, $i = 1, 2, \ldots, n$, and $p_i(\{1\}) = p$ for all i . The binomial random variable for this sample space is defined as follows.

Definition 2. Let X be the total number of successes in n repeated independent Bernoulli trials with probability p of success on a given trial. X is called the binomial random variable with parameters n and p .

The range of the random variable X is the integers 0, 1, 2, ..., n ; thus X is a discrete random variable and as such must have a probability function. The statement above that X has parameters n and p means that the probability function for X is completely specified if the values of n and p are known. This probability function is defined as follows.

Definition 3. If X is binomial with parameters n and 9 , then

$$p_X(x) = \binom{n}{x} p^x q^{n-x} \quad , x = 0, 1, \dots , n$$
$$= 0 , \quad \text{otherwise} .$$

The mean of X , μ_x , is equal to np , and the variance of X , $\sigma_x{}^2$ is equal to npq.

The Bernoulli random variable is actually a special case of the binomial with parameters $n = 1$ and p . then X is called the Bernoulli random variable with parameter p .

The Bernoulli random variable then is simply the number of successes we observe in a single Bernoulli trial and has probability function

$$\begin{aligned} p_X(x) &= p \; , \quad \text{for } x = 1 \\ &= q \; , \quad \text{for } x = 0 \\ &= 0 \; , \quad \text{otherwise.} \end{aligned}$$

It has mean $\mu_X = p$ and variance $\sigma_X{}^2 = pq$.

<u>Definition 4</u>. A multinomial trial, with parameters p_1, p_2, ... , p_k , is a trial which results in one of k possible outcomes (these outcomes are called classes). The probability of the i-th class occurring on a single trial is p_i , $i = 1, 2, \dots , k$; thus $0 \leq p_i \leq 1$, $i = 1, 2, \dots , k$, and $\sum_{i=1}^{k} p_i = 1$.

A single roll of a single die is a multinomial trial (with $k = 6$) since every roll results in one of the six faces being uppermost. The grade a student gets in a statistics course can be thought of as a multinomial trial with $k = 5$ (if the only grades he may receive are A, B, C, D, or F).

Clearly a multinomial is simply a generalization of a binomial trial, having an arbitrary k rather than just two possible outcomes. The multidimensional random variable is defined as follows:

Definition 5. Given an experiment which consists of n repeated, independent, multinomial trials with parameters p_i , $i = 1, 2, \ldots, k$, let X_i be the number of trials which result in outcomes in the i-th class, $i = 1, 2, \ldots, k$. $(X_1, X_2, \ldots, X_k)$ is called the multinomial random variable with parameters n , p_1 , p_2 , ... , p_k .

The reason for making a distinction between Bernoulli trial and binomial random variable or between a multinomial trial and multinomial random variables is to explain some difficulty that was found in generating samples from a multivariate multinomial distribution. In the data point represented by the 1 x p vector, X_i , where

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) ,$$

as described in Chapter II, it is desired that each component, x_{ij} be the result of the i-th Bernoulli or multinomial trial for the j-th characteristic. Currently, computer programs which generate "multivariate multinomial" data treat X_i as a multinomial variable. One is not able to randomly generate a multivariate observation in which each variable is an outcome of a multinomial trial. There is no correlation structure associated with the generation. It is necessary that one be able to impose a given correlation structure in order that principal component analysis may be incorporated into the test design.

Therefore, a multivariate multinomial was generated from a multivariate normal random variable with the desired correlation matrix as described in Chapter V with $\eta = 0.0$. First, a multivariate normal random variable is generated from a multivariate normal sample with mean 0 and given correlation matrix, as described in Chapter V. For the multivariate binomial, each variable is transformed to a Bernoulli random variable with parameter p by translating the normal z value for each variate to "1" if $p(x \leq z) \leq p$ and to "0" if $p(x \leq z) > p$. For two populations two different p values are chosen. The test procedure described in Chapter V is continued using the Euclidean metric and then several association coefficients which will be described in the next section.

To extend this procedure to a multivariate multinomial, each variate is chosen to be an outcome of a multinomial trial with parameters $p_1, p_2, \cdots, p_k$.

Association Coefficients

An association coefficient is a pair-function that measures the agreement between pairs of observations over an array of two-state or multistate characters. Many of these coefficients measure the numbers of actual agreement as compared with the number of theoretically possible ones. Characters coded in two or a few states are especially suitable for the computation of association coefficients, although even continuous characters can be coded to yield association coefficients.

In the most common model, association coefficients are computed with two-state characters, which are for convenience coded 0 or 1 . The 0 , 1 code can represent the presence or absence of a characteristic or property such as a bristle or a pigment; it may stand for the success or failure of a biochemical reaction; or it may be an arbitrary designation as in a structure having only two shapes, either rounded or pointed, where 0 might designate rounded, and 1 pointed. When character states are compared over pairs of rows in a conventional data matrix, the outcome can be summarized in a conventional 2 x 2 frequency table such as the one shown.

		Data Point j		
		1	0	
Data Point i	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	n = a + b + c + d

In the left upper quadrant of the figure is written the number of characters coded 1 in both data points, while in the right lower quadrant is written the number of characters coded 0 in both data points. The other two quadrants register the number of characters in which the two datapoints disagree, being coded 1 for data point j and 0 for data point i (or the converse).

The marginal totals are the sums of these frequencies, with n being reserved for the sum of the four frequencies, which equals the number of characters in the study. It is convenient to define m as the number of matches or agreements ($m = a + d$) , and to let u be the number of mismatches ($u = b + c$) , whence $m + u = n$.

In this comparative study, three association coefficients will be used in the test procedure: the coefficient of Jaccard, the Simple Matching Coefficient, and the Yule Coefficient.

The Coefficient of Jaccard (1908) is defined as

$$S_J = \frac{a}{a + u} = \frac{a}{a + b + c}$$

It is clear that $S_J \longrightarrow 0$ as $a/u \longrightarrow 0$, and that $S_J \longrightarrow 1$ as $u \longrightarrow 0$. The coefficient of Jaccard omits consideration of negative matches. Whether negative matches should be incorporated into a coefficient of association may occasion serious doubt. It may be argued that basing similarity between two species on the mutual absence of a certain character is improper. "The absence of wings, when observed among a group of distantly related organisms (such as camel, louse, and nematode), would surely be an absurd indication of similarity" (Sneath and Sokal, 1973). The coefficient of Jaccard is appropriate when negative matches are to be excluded.

The Simple Matching Coefficient is defined as

$$S_{SM} = \frac{m}{m + u} = \frac{a + d}{a + b + c + d}$$

This is one of the oldest and simplest coefficients, introduced to numerical taxonomy by Sokal and Michener (1958). From the formula it follows that $S_{SM} \longrightarrow 0$ as $m/u \longrightarrow 0$, and that $S_{SM} \longrightarrow 1$ as $u/m \longrightarrow 0$. In its complementary form, $1 - S_{SM}$, the simple matching coefficient is equal to the squared Euclidean distance based on unstandardized character states, which can take the value of 0 or 1 ; that is, $\sqrt{1 - S_{SM}} = d$.

The Yule Coefficient is defined as

$$S_Y = (ad - bc) / (ad + bc) .$$

Its numerator is the determinant of the 2 x 2 table and the limits of S_Y are from -1 to +1 . In the former case there are no matches at all, in the latter, matches are perfect. Other coefficients related to it, which are seldom used but described in greater detail by Sokal and Sneath (1963), include the well-known coefficient, $S_\phi = (ad - bc)/ [(a + b) (a + c) (c + d) (b + d)]^{1/2}$, which is the product moment correlation coefficient r for data coded 0 , 1 , and the coefficient of Hamann, $S_H = (m - u) / n = (a + d - b - c) / (a + b + c + d)$. "All these coefficients balance matches against mismatches, a concept that does not appear of special utility in the estimation of similarity" (Sneath & Sokal, 1973).

Multi-State Coding

The several states in qualitative multistate characters cannot necessarily be arrayed in some obvious order but still refer to a unit character on logical grounds. These characters are therefore often called unordered multistate characters. An example would be alternative color patterns of a given structure. One way of coding these is to use a separate symbol for each state; for example,

Color Structure	State
Red	0
Yellow	1
Blue	2

A match is scored if the same symbol occurs in two data points; otherwise, a mismatch is recorded.

Another method is to convert the qualitative multistate character into several new characters. The characters might then be coded as shown in the following chart.

Color of Structure	Two-State Characters		
	1	2	3
Red	1	0	0
Yellow	0	1	0
Blue	0	0	1

This is not an easy task inasmuch as the recoding has to be done in such a way that a positive score on one of the new characters does not automatically bring about negative scores on all other such characters derived from the same qualitative character. In practice it is commonly found that most qualitative multistate characters can be converted into several independent characters if a little thought is given to the problem.

By the method of additive coding, the multiple character states are coded as shown below:

Data Point	Multistate Character	Two-State Characters		
		1	2	3
a	state 0 (character undetectable)	0	0	0
b	state 1 (weak positive)	1	0	0
c	State 2 (moderate positive)	1	1	0
d	State 3 (strong positive)	1	1	1

In this way a multistate character i of m_i states is turned into $m_i - 1$ two-state characters. The scoring is termed additive because the state 3 , for instance, is expressed as the sum of the effects of the two-state characters 1, 2, and 3 .

In any of these methods of coding multi-state characters, two-state or binary data are produced. The procedure for binomial data is then applied to the binary codes of the multi-state data.

Since each of these coding methods transforms multinomial data to binomial data, it is sufficient for now to look at the effects of our procedure on binary data.

Extension to Principal Component Analysis

An important advance in ordination has come about through the principal coordinate analysis developed by Gower (1966). By this technique it is possible to compute principal components of any Euclidean distance matrix without being in possession of either the original data matrix or a variance-covariance matrix of the characters of the data points. Gower's method is also applicable to non-Euclidean distance and association coefficients.

The computational procedure applied to A , the matrix of association coefficients a_{ij} is summarized:

1. Form the association matrix A using one of the association coefficient methods;
2. Transform A to AT where α_{ij} , the element in row i and column j of matrix AT is given by

   $$\alpha_{ij} = a_{ij} - \overline{a}_i - \overline{a}_j + \overline{a} ,$$

   where a_{ij} is the ij-th element of matrix A , a_i is the mean value of the i-th row of A , a_j is the mean value of the j-th column of A , and $\overline{a}$

is the overall mean value of the elements of A ;

3. Find the eigenvalues and eigenvectors of AT .

The matrix of eigenvectors gives the coordinates of the data points on their coordinate axes.

Test Procedure

The procedure described in Chapter V with some enhancements was followed.

The computer programs were augmented to follow Gower's principal coordinate method. The number of variables was increased to 30 ; this number seemed to give better results without greatly decreasing the efficiency of the computer processing.

Instead of the spacing variable, δ , probability parameters p_1 and p_2 were used to separate the two clusters. Three different association coefficient methods, ac , were used.

Thus a quadruple variable structural parameter (ρ , p_1 , p_2 , ac) was defined. The value of ρ was allowed to vary from .5 to .9 . Three (p_1 , p_2) pairs were studied—(.3 , .7) , (.4 , .6) , and (.45 , .55) . The three association coefficient methods studied were S_J , S_{SM} , and S_Y .

CHAPTER VIII

DISCUSSION OF RESULTS FROM MULTIVARIATE MULTINOMIAL SAMPLES

Tables X-XIII and Figures 1-14 in the Appendix give the results from the comparative study of eighteen agglomerative methods. In these tables, the results are given in the form of $\overline{c}$ computed over 100 reps for each setting of the quadruple variable structural parameter (ρ , p_1 , p_2 , ac) .

Tables X and XI and Figures 1-6 in the Appendix show the results from using the association coefficient of Jaccard, S_J . It can easily be seen that the coefficient of Jaccard gave the best and most consistent results for members of the (β , γ) family with $\beta < 0$. Figure 13 shows the region in the (β , γ) plane (DuBien, 1976) where better cluster retrieval was obtained.

Using the Simple Matching Coefficient, S_{SM} , good results were obtained for members of the (β , γ) family with $\beta = 0$. The most frequently used clustering algorithms -- single linkage, average linkage, complete linkage -- are among these members of this family. This information is demonstrated in Tables XII and XIII and in Figures 7-12 in the Appendix. In general, with $\beta < 0$, better cluster retrieval was obtained by application of clustering alone.

Figure 14 compiles this information from both association coefficients, S_J and S_{SM}, to show which procedure (principal component/clustering or clustering alone) gave good results for each of the eighteen clustering algorithms.

The Yule coefficient, S_Y, gave unacceptable results due mainly to the fact that several of the coefficient values were undefined. This might have been caused by the way in which the samples were generated and might not be true in "real-life" data.

Changes in ρ, the dependency among the variables, or in p_1 and p_2, the binomial sample parameters, caused some changes in the $\overline{c}$ values, but no more than would be expected to occur due to the changes in the amount of separation of the populations.

The most important result is the fact that performing principal component analysis prior to performing cluster analysis very greatly improved the retrieval ability of the known clustering over the use of cluster analysis alone. Some (β , γ) pairs produced "good" $\overline{c}(Y, Y')$ values, but many (β , γ) pairs produced "good" $\overline{c}(Y, Y'')$ values. Another important observation is that there were no Y' clusterings that matched the Y clusterings exactly, but with most (β , γ) pairs there were many Y'' clusterings that matched the Y clusterings exactly. Thus we were better able to retrieve our known sample structure by using the principal component / cluster analysis procedure than by using the cluster analysis procedure alone when applied to the generated multivariate multinomial samples.

CHAPTER IX

GENERAL CONCLUSIONS AND POSSIBLE EXTENSIONS

The application of cluster analysis and principal component analysis to samples of data is a common practice. Often principal component analysis is applied to reduce the number of variables before performing a cluster analysis of a sample of data. In this study of the practice of applying principal component analysis prior to cluster analyzing a data sample, some important conclusions have been demonstrated.

In applying the procedure to multivariate normal data, the retrieval ability of the known clustering was improved slightly. No loss in retrieval ability was demonstrated for the eighteen members of the (β, γ) agglomerative clustering family. Therefore, for a large number of variables, it would be beneficial to reduce the number of variables before performing cluster anlaysis of a data sample in order to reduce processing expense.

The most important result was demonstrated when the procedure was applied to a multivariate multinomial sample. Retrieval ability of the known sample clustering was very greatly increased. Under the described conditions, application of principal component analysis prior to cluster analysis of a data set not only saves time and money in processing the sample, but also produces better clustering results.

As is true in most research, when attemting to solve the problem at hand, one often uncovers many more areas which need further study. This is surely true here. This study could be extended to include other clustering methods, other principal component methods, different measures of distance or association, and different ways of coding multi-state data, some of which are mentioned in this thesis. The computer programs which were written for this study were written in such a way that they easily could be enhanced to include some of these extensions to this study. If this procedure is to be applied in a professional setting, the computer programs should be converted to include array processing.

BIBLIOGRAPHY

Anderberg, Michael R.

1973 Cluster Analysis for Applications. New York: Academic Press.

Anderson, T. W.

1971 An Introduction to Multivariate Statistical Methods. New York: John Wiley & Sons.

Ball, Geoffrey H.

1965 "Data Analysis in the Social Sciences: What About the Details?" Fall Joint Computer Conference (In AFIPS Conference Proceedings), Vol. 27, p. 533-559.

Bennett, V. and T. Lewis.

1978 Outliers in Statistical Data. New York: John Wiley & Sons.

Chang, W. C.

1980 "On Using Principal Components Before Clustering a Two Populations Case." Presented to 1980 Annual National Meeting of ASA.

Chernoff, H.

1970 "Metric Considerations in Cluster Analysis." Tech. Rep. No. 67, AD 714810. Dept. of Statistics, Stanford University, Stanford, California.

Dempster, A. P.

1969 Elements of Continuous Multivariate Analysis. Reading, Mass: Addison Wesley Pub. Co.

DuBien, Janice L.

1976 "Comparative Techniques for the Evaluation of Clustering Methods." (Unpub. Ph.D. Thesis, Oklahoma State University, Stillwater, Oklahoma.)

DuBien, Janice L., and W. D. Warde.

1979 "A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms." Canadian Journ. of Stat., Vol. 7, No. 1, pp. 29-38.

1983 "Comparison of Agglomerative Clustering Methods With Respect to Noise." Accepted for publication, Communications in Statistics.

Dubin, R.

1971 "Typology Empirical Attributes: Multidimensional Typology Analysis (MTA)." TR-5, AD 728781, University of California, Irvine.

Dubin R. and Champoux, J. E.

1970 "Typology of Empirical Attributes: Dissimilarity Linkage Analysis (DLA)." Tech. Rep. 3, AD 708678, University of California, Irvine.

Eddy, R. P.

1968 "Class Membership Criteria and Pattern Recognition." Rep. 2524, AD 834208, Naval Ship Res. and Develop. Center of Washington, D.C.

Fisher, W. D.

1968 Clustering and Aggregation in Economics. Baltimore, Maryland: John Hopkins Press.

Grinker, R. R.; Miller, J.; Sabshin, M.; Nunn, R. and Nunnally, J. C.

1961 The Phenomena of Depression. New York: Paul B. Hoebes, Inc.

Gower, J. C.

1966 "Some Distance Properties of Latent Roots and Vector Methods Used in Multivariate Analysis." Biometrika, Vol. 53: pp. 325-338.

Hartigan, J. A.

1972 "Direct Clustering of a Data Matrix." JASA, Vol. 67, pp. 123-129.

Hotelling, H.

1933 "Analysis of a Complex of Statistical Variables into Principal Components." Journal of Educ. Psychology, Vol. 24, pp. 417-441, 498-520.

Jaccard, P.

1908 "Nouvelles Recherches Sur La Distribution Florale." Bull. Soc. Vaud. Sci. Nat., Vol. 44: pp. 223-270.

Johnson, Stephen C.

1967 "Hierarchical Clustering Schemes." Psychometrika, Vol. 32 (September), pp. 241-255.

Kendall, M. G.

1968 A Course in Multivariate Analysis. New York: Hofner.

Kendall, M. G. and Stuart, A.

1963 The Advanced Theory of Statistics, III. London: Griffin.

Kiloh, L. G.

1965 "The Differentiation of Depressive Syndromes." In Aspects of Depression Illness. (D. C. Maddison and G. M. Duncan, eds.), London: Livingstone.

Lance, G. N., and W. T. Williams.

1966 "A Generalized Sorting Strategy for Compter Classifications." Nature, Vol. 212, p. 218.

1967 "A Generalized Theory of Classificatory Sorting Strategies. I. Hierarchical Systems." The Computer Journal, Vol. 9, (February), pp. 373-380.

Litofsky, B.

1969 "Utility of Automatic Classification Systems for Information Storage Retrieval." Ph.D. Dissertation, University of Pennsylvania, Philadelphia, cited in Dissertation Abstracts 30, No. 7, 3264-B (1970). Also available from NTIS as AD 687140.

Mendels, J. and Cochrane, C.

1968 "The Nosology of Depression: The Endogeneous-Reactive Concept." Am. J. Psychia., Vol. 124 (May Suppl.) pp. 1-11.

Morrison, D. F.

1976 Multivariate Statistical Methods, 2nd edition, New York: McGraw-Hill, Inc.

Mrachek, Roger J.

1972 "Some Statistical Aspects of Clustering Procedures." (Unpub. M.S. thesis, Iowa State University.)

Paykel, E. S.

1971 "Classification of Depresssed Patients: A Cluster Analysis Derived Group." Br. J. Psychia., Vol. 118, p. 275.

1972 "Depressive Typology and Response to Amitriptyline." Br. J. Psychia., Vol. 120, pp. 147-156.

Pilowsky, I.: Levine, S. and Boulton D. M.

1969 "The Classification of Depression of Numerical Taxonomy." Br. J. Psychia., Vol. 115, pp. 937-945.

Pearson, K.

1901 "On Lines and Planes of Closest Fit to Systems of Points in Space." Phil. Mag. ser. 2, 6: pp. 559-72.

Rand, William Medden.

1969 "The Development of Objective Criteria for Evaluating Clustering Methods." (Unpub. Ph.D. Thesis, University of California at Los Angeles.)

1971 "Objective Criteria for the Evaluation of Clustering Methods." JASA, Vol. 66 (December), pp. 846-850.

Rohlf, F. J.

1970 "Adaptive Hierarchical Clustering Schemes." Syst. Zool., Vol. 19, No. 1, pp. 58-83.

Sneath, Peter H. A. and Robert R. Sokal.

1973 Numerical Taxonomy. San Francisco: W. H. Freeman & Co.

Sokal, R. R. and C. D. Michener.

1958 "A Statistical Method for Evaluating Systematic Relationships." University of Kansas Sci. Bull., Vol. 38, pp. 1409-1438.

Sokal, R. R. and P. H. A. Sneath.

1963 Principles of Numerical Taxonomy. San Francisco: W. H. Freeman and Company.

Terekhina, A. Y.

1973 "Methods of Multidimensional Data Scaling and Visualization - A Survey." Automation and Remote Control. Vol. 34, pp. 1109-1121.

Williams W. T. and M. B. Dale.

1965 "Fundamental Problems in Numerical Taxonomy." In Advances in Botanical Research (R. D., Preston, ed.), London: Academic Press.

TABLE I

A COMPARISON OF $\bar{c}$ ACROSS ρ FOR SIX ALGORITHMS ALONG $\eta = 0.0$ WHERE $\delta = 1.0$

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
.6	$\bar{c}(Y,Y')$	.49485	.55106	.60348	.62818	.63258	.63742
	$\bar{c}(Y,Y'')$	.51030	.55909	.60652	.63333	.62955	.62561
	$\bar{c}(Y',Y'')$	.85879	.83439	.86152	.84727	.85182	.79939
.7	$\bar{c}(Y,Y')$	.49788	.55258	.58652	.59318	.62106	.61348
	$\bar{c}(Y,Y'')$	.49545	.55758	.57227	.60970	.62455	.62803
	$\bar{c}(Y',Y'')$	..89758	.87924	.88636	.85500	.89621	.82727
.8	$\bar{c}(Y,Y')$	.49667	.54667	.57862	.57985	.59530	.60848
	$\bar{c}(Y,Y'')$	.49864	.54242	.58273	.57788	.60121	.60652
	$\bar{c}(Y',Y'')$	.93682	.92394	.91227	.90985	.88439	.83348
.9	$\bar{c}(Y,Y')$	.50485	.54636	.57212	.57288	.59379	.58682
	$\bar{c}(Y,Y'')$	.50303	.54333	.56273	.57273	.59788	.59030
	$\bar{c}(Y',Y'')$	.97667	.92121	.93515	.95379	.94348	.91318

TABLE II

A COMPARISON OF $\bar{c}$ ACROSS ρ FOR SIX ALGORITHMS ALONG η = 0.0 WHERE δ = 1.5

ρ		Single	(o, -.25)	Average	(0, .25)	Complete	(0, .75)
.6	$\bar{c}$(Y,Y')	.57848	.71500	.76091	.76955	.77955	.76879
	$\bar{c}$(Y,Y'')	.58485	.73258	.75000	.75985	.78061	.76970
	$\bar{c}$(Y',Y'')	..86788	.87788	.91091	.89455	.90227	.87818
.7	$\bar{c}$(Y,Y')	.56121	.68818	.73758	.74561	.76061	.74985
	$\bar{c}$(Y,Y'')	.56864	.67773	.73076	.75167	.74318	.74030
	$\bar{c}$(Y',Y'')	.91803	.87136	.89985	.94000	.91197	.86682
.8	$\bar{c}$(Y,Y')	.55530	.64682	.70955	.73652	.73242	.73318
	$\bar{c}$(Y,Y'')	.54394	.65682	.69136	.71727	.71924	.72061
	$\bar{c}$(Y',Y'')	.93439	.89576	.90788	.92864	.94500	.92652
.9	$\bar{c}$(Y,Y')	.52682	.61136	.70348	.71030	.70955	.71273
	$\bar{c}$(Y,Y'')	.52606	.60636	.68545	.69470	.71076	.70621
	$\bar{c}$(Y',Y'')	.96500	.94985	.94409	.95803	.94364	.92652

TABLE III

A COMPARISON OF $\bar{c}$ ACROSS ρ FOR SIX ALGORITHMS
ALONG $\eta = 0.0$ WHERE $\delta = 2.0$

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
.6	$\bar{c}(Y,Y')$	.70303	.84288	.88848	.89742	.90455	.89712
	$\bar{c}(Y,Y'')$	.71970	.85591	.88818	.89652	.91091	.88909
	$\bar{c}(Y',Y'')$	.87455	.93242	.85182	.95636	.95727	.93348
.7	$\bar{c}(Y,Y')$	.68909	.83682	.86621	.87833	.87727	.88561
	$\bar{c}(Y,Y'')$	.70955	.81652	.85515	.88273	.89697	.88076
	$\bar{c}(Y',Y'')$	.93076	.93758	.96045	.96227	.95545	.92970
.8	$\bar{c}(Y,Y')$	.68515	.79879	.84106	.85545	.86773	.85242
	$\bar{c}(Y,Y'')$	.68348	.79076	.84106	.85485	.86515	.85152
	$\bar{c}(Y',Y'')$	.93591	.91258	.92727	.97788	.96894	.92273
.9	$\bar{c}(Y,Y')$	.65333	.75652	.83136	.84439	.84030	.84106
	$\bar{c}(Y,Y'')$	.66030	.75076	.83061	.85197	.84545	.83545
	$\bar{c}(Y',Y'')$	.97061	.97455	.96258	.97970	.96848	.95348

TABLE IV

A COMPARISON OF $\bar{c}$ ACROSS δ FOR SIX ALGORITHMS ALONG $\eta = 0.0$ WHERE $\rho = .6$

δ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
1.0	$\bar{c}(Y,Y')$	.49858	.55106	.60348	.62818	.63258	.63742
	$\bar{c}(Y,Y'')$	.51030	.55909	.60652	.63333	.62955	.62561
	$\bar{c}(Y',Y'')$	.85879	.83439	.86152	.84727	.85182	.79939
1.5	$\bar{c}(Y,Y')$	.57848	.71500	.76091	.76955	.77955	.76879
	$\bar{c}(Y,Y'')$	.58485	.73258	.75000	.75985	.78061	.76970
	$\bar{c}(Y',Y'')$	.76788	.87788	.91091	.89455	.90227	.87818
2.0	$\bar{c}(Y,Y')$	.70303	.84288	.88848	.89742	.90455	.89712
	$\bar{c}(Y,Y'')$	.71970	.85591	.88818	.89652	.91091	.88909
	$\bar{c}(Y',Y'')$	.87455	.93242	.95182	.95363	.95727	.93348

TABLE V

A COMPARISON OF $\bar{c}$ ACROSS δ FOR SIX ALGORITHMS
ALONG $\eta = 0.0$ WHERE $\rho = .7$

δ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
1.0	$\bar{c}(Y,Y')$	.49788	.55258	.58652	.59318	.62106	.61348
	$\bar{c}(Y,Y'')$	.49545	.55758	.57227	.60970	.62455	.62803
	$\bar{c}(Y',Y'')$	.89758	.87924	.88636	.85500	.89621	.82727
1.5	$\bar{c}(Y,Y')$	.56121	.68818	.73758	.74561	.76061	.74985
	$\bar{c}(Y,Y'')$	.56864	.67773	.73076	.75167	.74138	.74030
	$\bar{c}(Y',Y'')$	.91803	.87136	.89985	.94000	.91197	.86682
2.0	$\bar{c}(Y,Y')$	.68909	.83682	.86621	.87833	.87727	.88561
	$\bar{c}(Y,Y'')$	.70955	.81652	.85515	.88273	.89697	.88076
	$\bar{c}(Y',Y'')$	.93076	.93758	.96045	.96227	.95545	.92970

TABLE VI

A COMPARISON OF $\bar{c}$ ACROSS δ FOR SIX ALGORITHMS
ALONG η = 0.0 WHERE ρ = .8

δ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
	$\bar{c}$(Y,Y')	.49667	.54667	.57682	.57985	.59530	.60848
1.0	$\bar{c}$(Y,Y'')	.49864	.54242	.58273	.57788	.60121	.60652
	$\bar{c}$(Y',Y'')	.93682	.92394	.91227	.90985	.88439	.83348
	$\bar{c}$(Y,Y')	.55530	.64682	.70955	.73652	.73242	.73318
1.5	$\bar{c}$(Y,Y'')	.54394	.65682	.69136	.71727	.71924	.72061
	$\bar{c}$(Y',Y'')	.93439	.89576	.90788	.92864	.94500	.92652
	$\bar{c}$(Y,Y')	.68515	.79879	.84106	.85545	.86773	.85242
2.0	$\bar{c}$(Y,Y'')	.68348	.79076	.84106	.85485	.86515	.85152
	$\bar{c}$(Y',Y'')	.93591	.91258	.92727	.97788	.96894	.92273

TABLE VII

A COMPARISON OF $\bar{c}$ ACROSS δ FOR SIX ALGORITHMS ALONG $\eta = 0.0$ WHERE $\rho = .9$

δ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
1.0	$\bar{c}(Y,Y')$	.50485	.54636	.57212	.57288	.59379	.58682
	$\bar{c}(Y,Y'')$	.50303	.54333	.56273	.57273	.59788	.59030
	$\bar{c}(Y',Y'')$	.97667	.92121	.93515	.95379	.94348	.91318
1.5	$\bar{c}(Y,Y')$	.52685	.61136	.70348	.71030	.70955	.71273
	$\bar{c}(Y,Y'')$	.52606	.60636	.68545	.69470	.71076	.70621
	$\bar{c}(Y',Y'')$	.96500	.94985	.94409	.95803	.94364	.92652
2.0	$\bar{c}(Y,Y')$	.65333	.75652	.83136	.84439	.84030	.84106
	$\bar{c}(Y,Y'')$	.66030	.75076	.83061	.85197	.84545	.83545
	$\bar{c}(Y',Y'')$	.97061	.97455	.96258	.97970	.96848	.95348

TABLE VIII

A COMPARISON OF $\bar{c}$ ACROSS η FOR SIX ALGORITHMS ALONG ρ = .7 WHERE δ = 1.5

η		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
0.0	$\bar{c}$(Y,Y')	.56121	.68818	.73758	.74561	.76061	.74985
	$\bar{c}$(Y,Y'')	.56864	.67773	.73076	.75167	.74318	.74030
	$\bar{c}$(Y',Y'')	.91803	.87136	.89985	.94000	.91197	.86682
0.1	$\bar{c}$(Y,Y'')	.55470	.66788	.70591	.73955	.74818	.73682
	$\bar{c}$(Y,Y'')	.55455	.65197	.71470	.72924	.74030	.71879
	$\bar{c}$(Y',Y'')	.90379	.87894	.91030	.90121	.92455	.85045
0.2	$\bar{c}$(Y,Y')	.55576	.64712	.68712	.71405	.72409	.72773
	$\bar{c}$(Y,Y'')	.54591	.66773	.70279	.72258	.72076	.72970
	$\bar{c}$(Y',Y'')	.91076	.87576	.91227	.89091	.91576	.87712
0.3	$\bar{c}$(Y,Y')	.54379	.62682	.64934	.68803	.70045	.71591
	$\bar{c}$(Y,Y'')	.54409	.65333	.67864	.69773	.71409	.71197
	$\bar{c}$(Y',Y'')	.92091	.97076	.90197	.89394	.89242	.86091
0.4	$\bar{c}$(Y,Y')	.53258	.61500	.64909	.68242	.69455	.69212
	$\bar{c}$(Y,Y'')	.53970	.62939	.66545	.67606	.68788	.69727
	$\bar{c}$(Y',Y'')	.91227	.89682	.90182	.86782	.92152	.86727

TABLE IX

A COMPARISON OF $\bar{c}$ ACROSS ρ FOR SIX ALGORITHMS ALONG $\eta = .1$ WHERE $\delta = 1.5$

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
.6	$\bar{c}(Y,Y')$	.56712	.70182	.73803	.75076	.76894	.76242
	$\bar{c}(Y,Y'')$	.59152	.71227	.73167	.77409	.76803	.76864
	$\bar{c}(Y',Y'')$	.85167	.86106	.88576	.88667	.91576	.85624
.7	$\bar{c}(Y,Y')$	.55470	.66788	.70591	.73955	.74818	.73682
	$\bar{c}(Y,Y'')$	.55455	.65197	.71470	.72924	.74030	.71879
	$\bar{c}(Y',Y'')$	.90379	.87894	.91030	.90121	.92455	.85045
.8	$\bar{c}(Y,Y')$	.54818	.64934	.69788	.72318	.72788	.71600
	$\bar{c}(Y,Y'')$	.54985	.65182	.68985	.71000	.70652	.72015
	$\bar{c}(Y',Y'')$	.93682	.90364	.90257	.92591	.94288	.89864
.9	$\bar{c}(Y,Y')$	.53485	.61697	.68679	.69848	.69803	.70455
	$\bar{c}(Y,Y'')$	.53364	.63167	.67848	.68455	.70333	.69879
	$\bar{c}(Y',Y'')$	.98636	.97015	.93758	.95121	.97106	.92121

TABLE X

A COMPARISON OF $\bar{c}$ ACROSS (p_1,p_2) FOR NINE ALGORITHMS ALONG $\rho = .6$ WHERE $ac = S_J$

p_1,p_2		Single		Complete		(-.25,0)		(-.5,-.5)		(-.5,.5)
			Average		(-.25,-.5)		(-.25,.5)		(-.5,0)	
		1	2	3	4	5	6	7	8	9
.3,.7	$\bar{c}(Y,Y')$	.48507	.48650	.51894	.46508	.48818	.56437	.48724	.55009	.59839
	$\bar{c}(Y,Y'')$	.52011	.51425	.54568	.71370	.91602	.89768	.94092	.99747	.88740
.4,.6	$\bar{c}(Y,Y')$	.48506	.48503	.49745	.48650	.49244	.51938	.49140	.52793	.53152
	$\bar{c}(Y,Y'')$	.51051	.52602	.56053	.78901	.99503	.95434	.97138	.99713	.96343
.45,.55	$\bar{c}(Y,Y')$	.48506	.48492	.48520	.49513	.50657	.49490	.48968	.51221	.51434
	$\bar{c}(Y,Y'')$	.51262	.52297	.56092	.81605	.99391	.93664	.98140	.99632	.97041

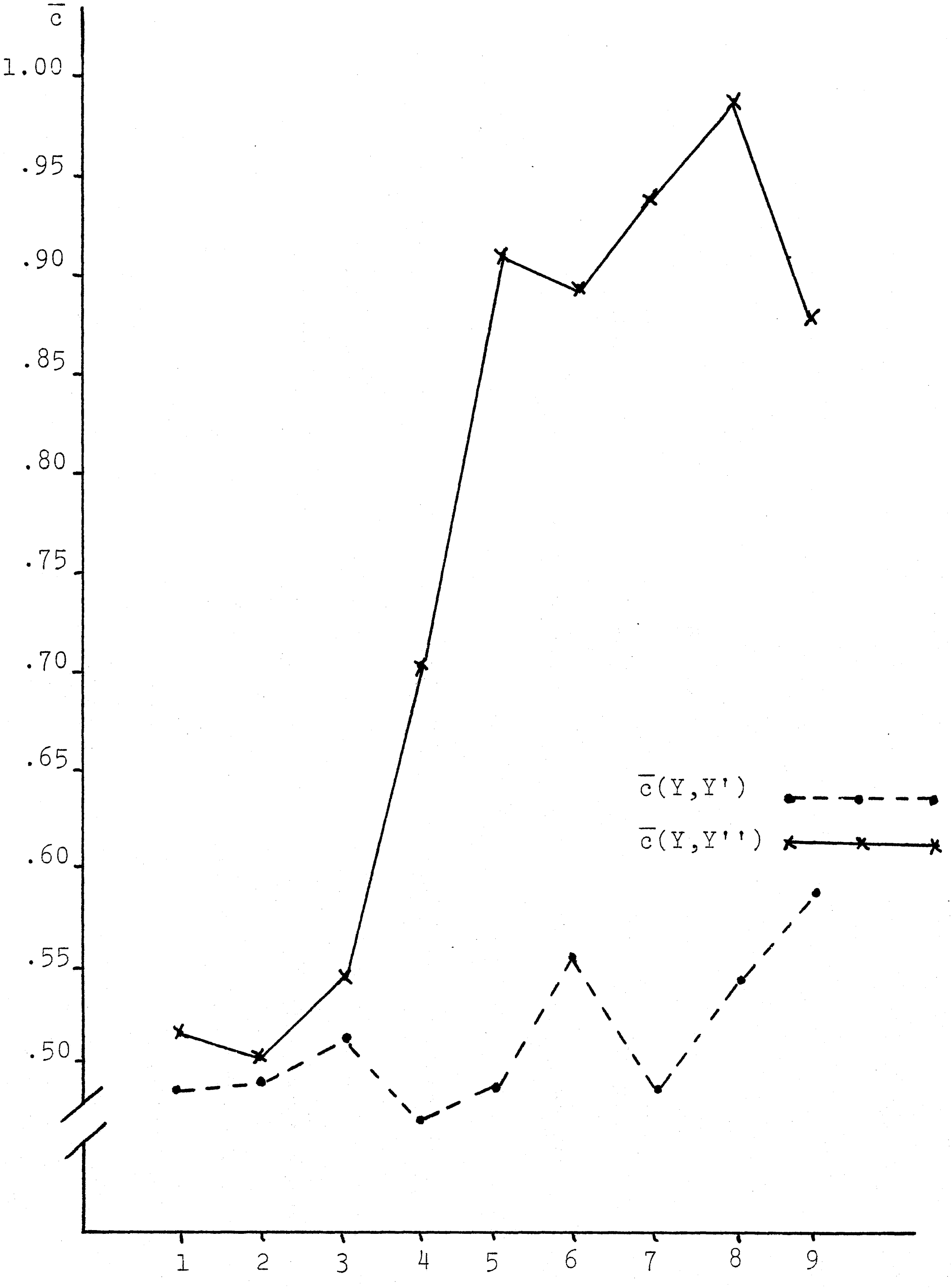

Figure 2. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE X with $(p_1,p_2)=(.3,.7)$

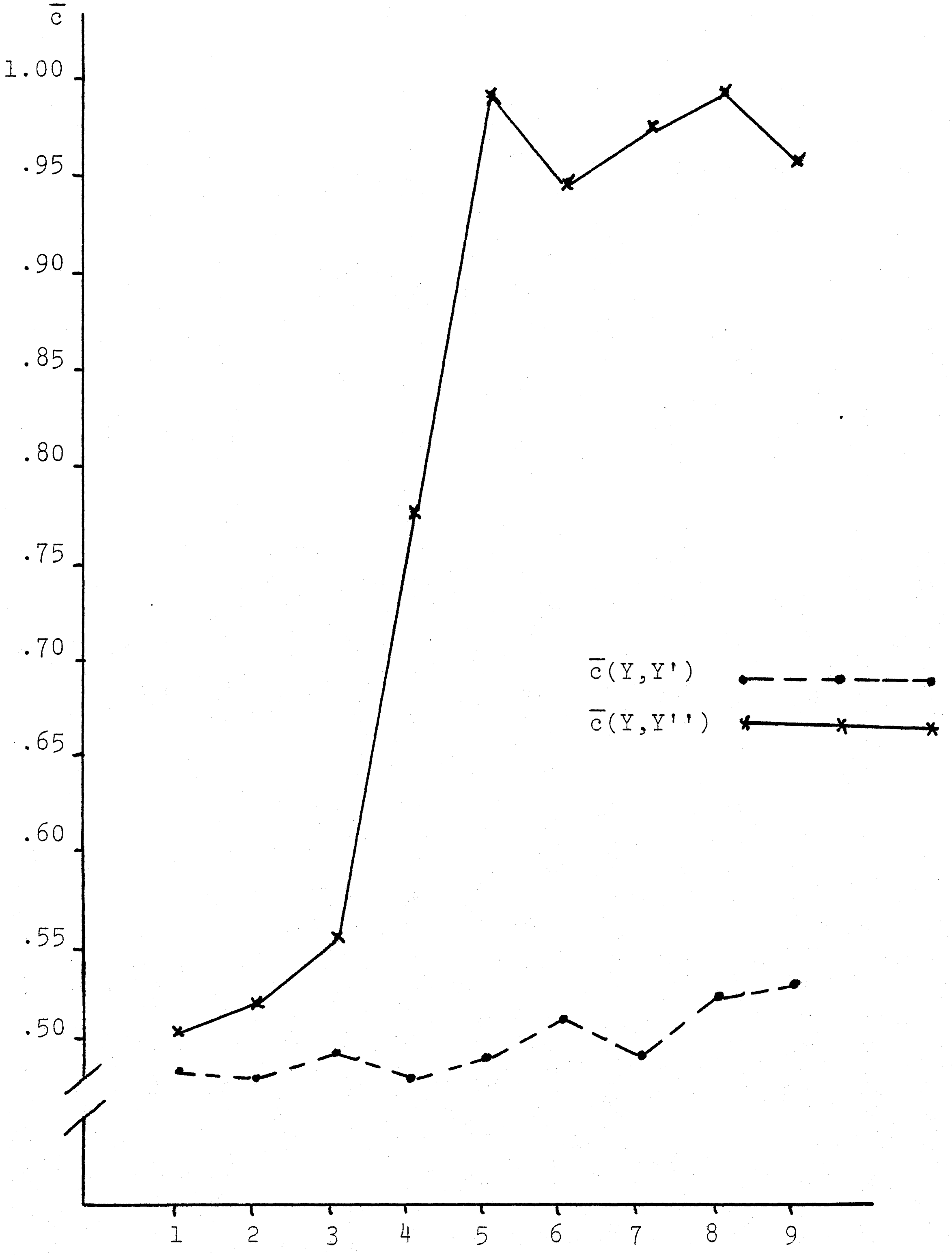

Figure 3. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE X with $(p_1,p_2)=(.4,.6)$

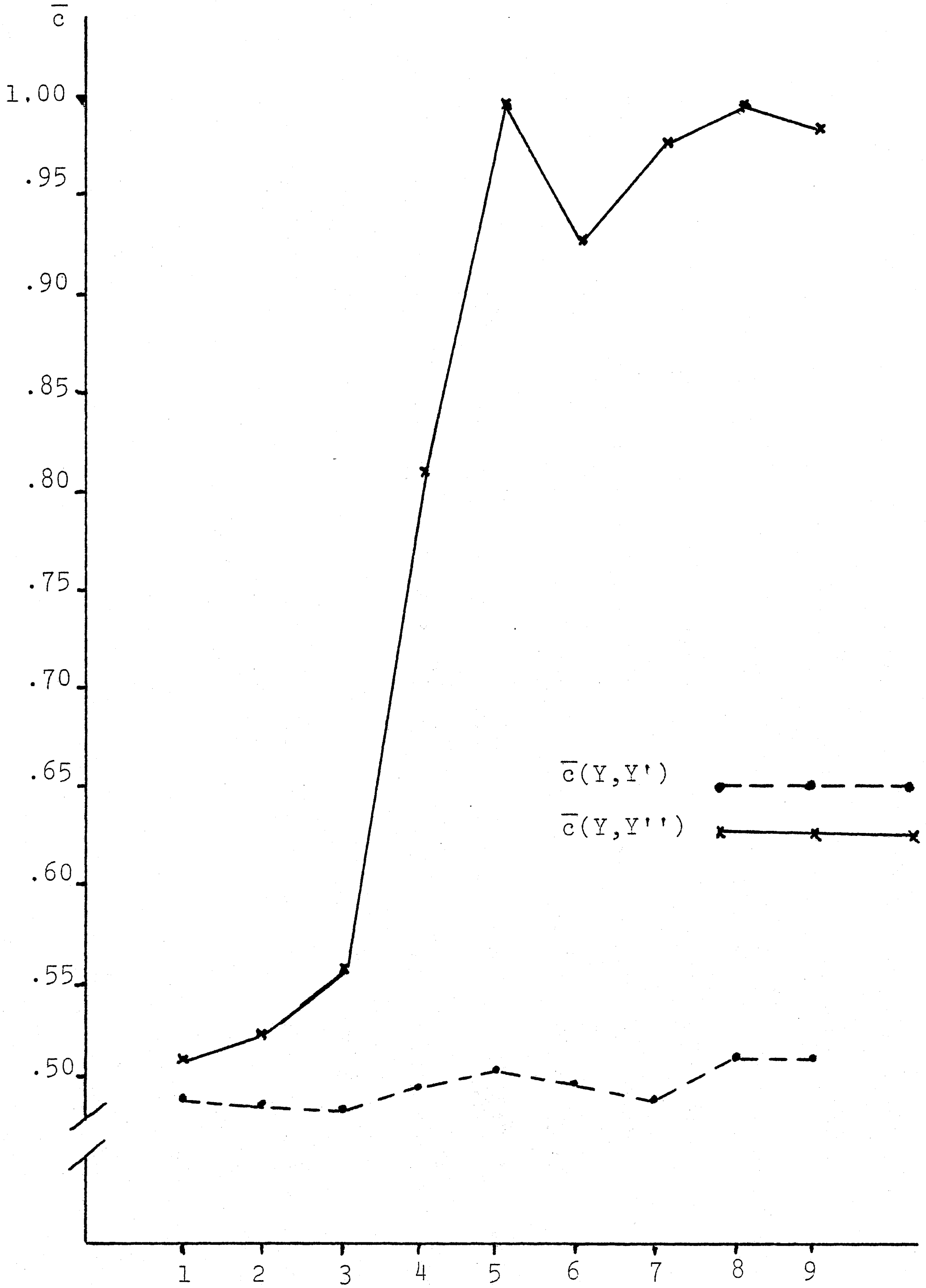

Figure 4. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE X with $(p_1,p_2)=(.45,.55)$

TABLE XI

A COMPARISON OF $\bar{c}$ ACROSS (p_1,p_2) FOR NINE ALGORITHMS ALONG $\rho = .8$ WHERE $ac = S_J$

p_1,p_2		Single		Complete		(-.25,0)		(-.5,-.5)		(-.5,.5)
			Average		(-.25,-.5)		(-.25,.5)		(-.5,0)	
		1	2	3	4	5	6	7	8	9
.3,.7	$\bar{c}(Y,Y')$	.48506	.48650	.51738	.46058	.48880	.56372	.49000	.54644	.59717
	$\bar{c}(Y,Y'')$	.52664	.51497	.53375	.74198	.93382	.90248	.92524	.99713	.91805
.4,.6	$\bar{c}(Y,Y')$	.48506	.48051	.49708	.48650	.49133	.52313	.48970	.52230	.53257
	$\bar{c}(Y,Y'')$	.52126	.51584	.55869	.75430	.99628	.94276	.96947	.99761	.95448
.45,.55	$\bar{c}(Y,Y')$	.48506	.48485	.49149	.48513	.49313	.50524	.49025	.51538	.51195
	$\bar{c}(Y,Y'')$	.52207	.51818	.57085	.78766	.99131	.93361	.97083	.99814	.95586

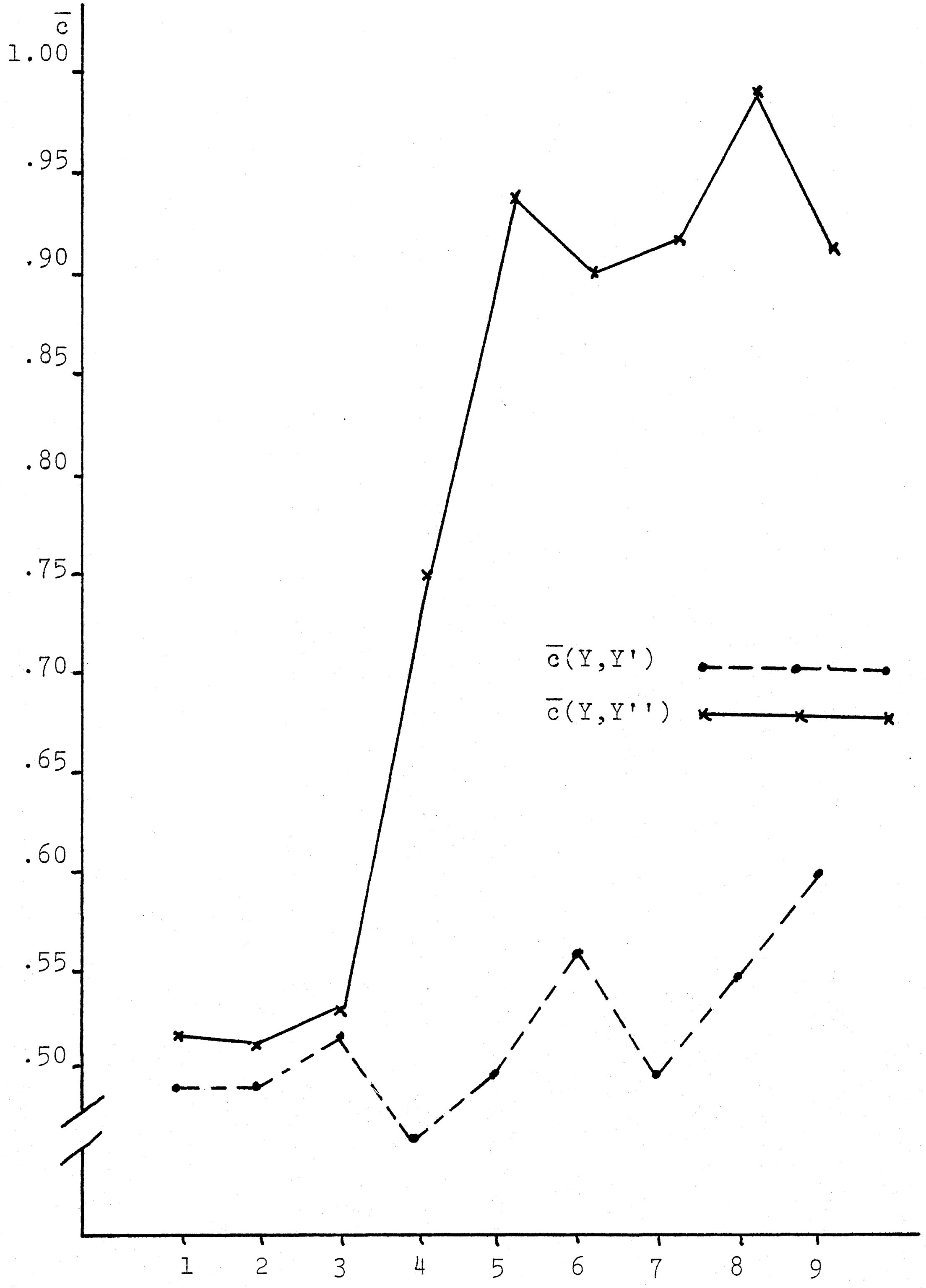

Figure 5. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XI with $(p_1,p_2)=(.3,.7)$

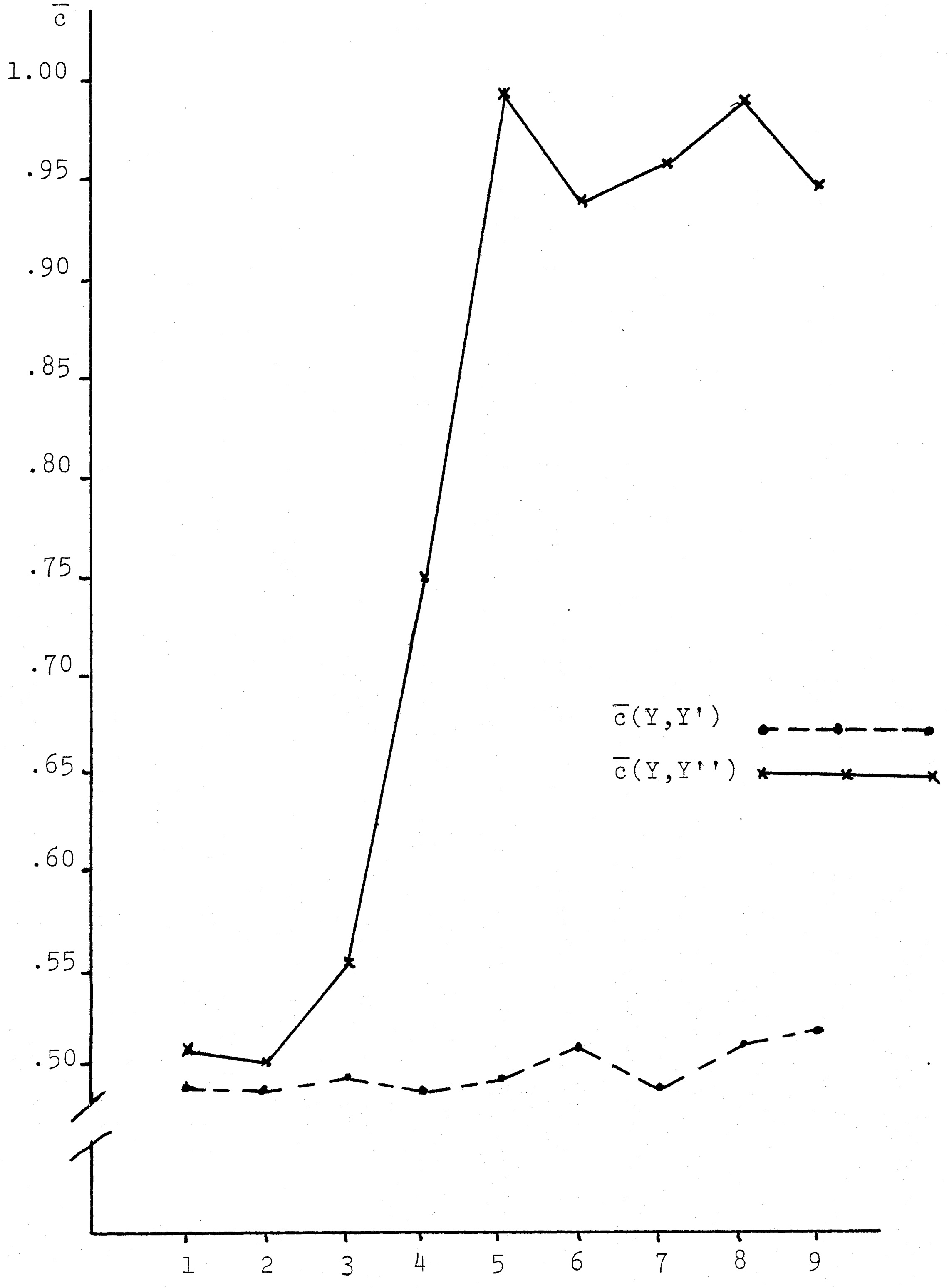

Figure 6. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XI with $(p_1,p_2)=(.4,.6)$

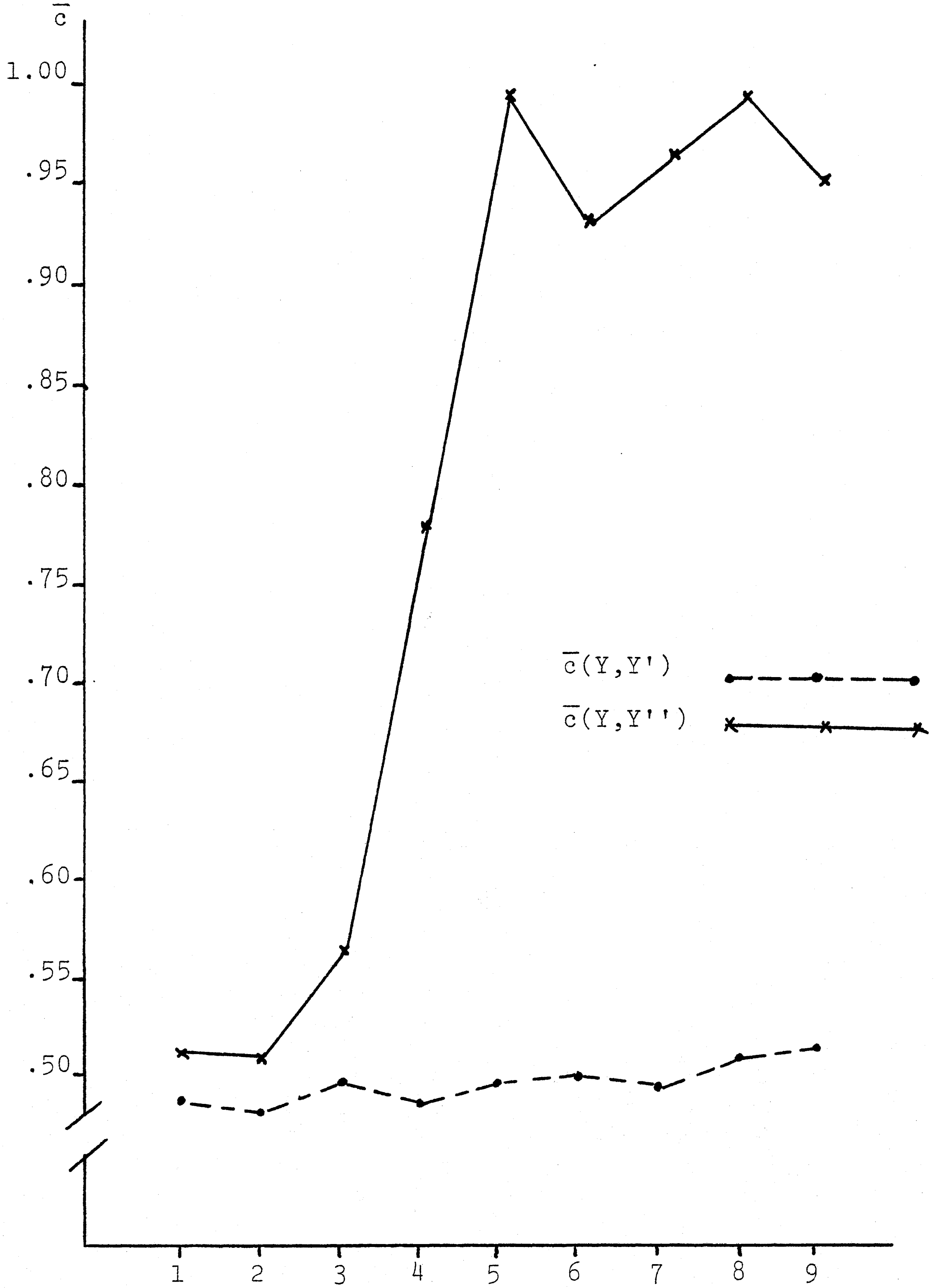

Figure 7. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XI with $(p_1,p_2)=(.45,.55)$

TABLE XII

A COMPARISON OF $\bar{c}$ ACROSS (p_1,p_2) OF NINE ALGORITHMS
ALONG $\rho = .6$ WHERE $ac = S_{SM}$

p_1,p_2		Single	Average	Complete	(-.25,-.5)	(-.25,0)	(-.25,.5)	(-.5,-.5)	(-.5,0)	(-.5,.5)
		1	2	3	4	5	6	7	8	9
.3,.7	$\bar{c}(Y,Y')$	.48560	.48513	.50032	.57069	.77648	.90924	.79591	.86609	.87083
	$\bar{c}(Y,Y'')$	.72041	.66368	.56605	.68779	.73699	.66260	.70108	.75789	.65366
.4,.6	$\bar{c}(Y,Y')$	.48506	.58540	.50439	.53766	.78154	.91811	.84448	.87069	.87136
	$\bar{c}(Y,Y'')$	.85163	.74818	.57275	.79782	.74076	.65445	.78037	.77103	.66437
.45,.55	$\bar{c}(Y,Y')$	.48506	.48575	.50310	.52699	.78563	.92611	.80462	.87069	.87379
	$\bar{c}(Y,Y'')$	.90032	.76531	.55508	.79768	.78198	.67315	.81547	.78405	.68216

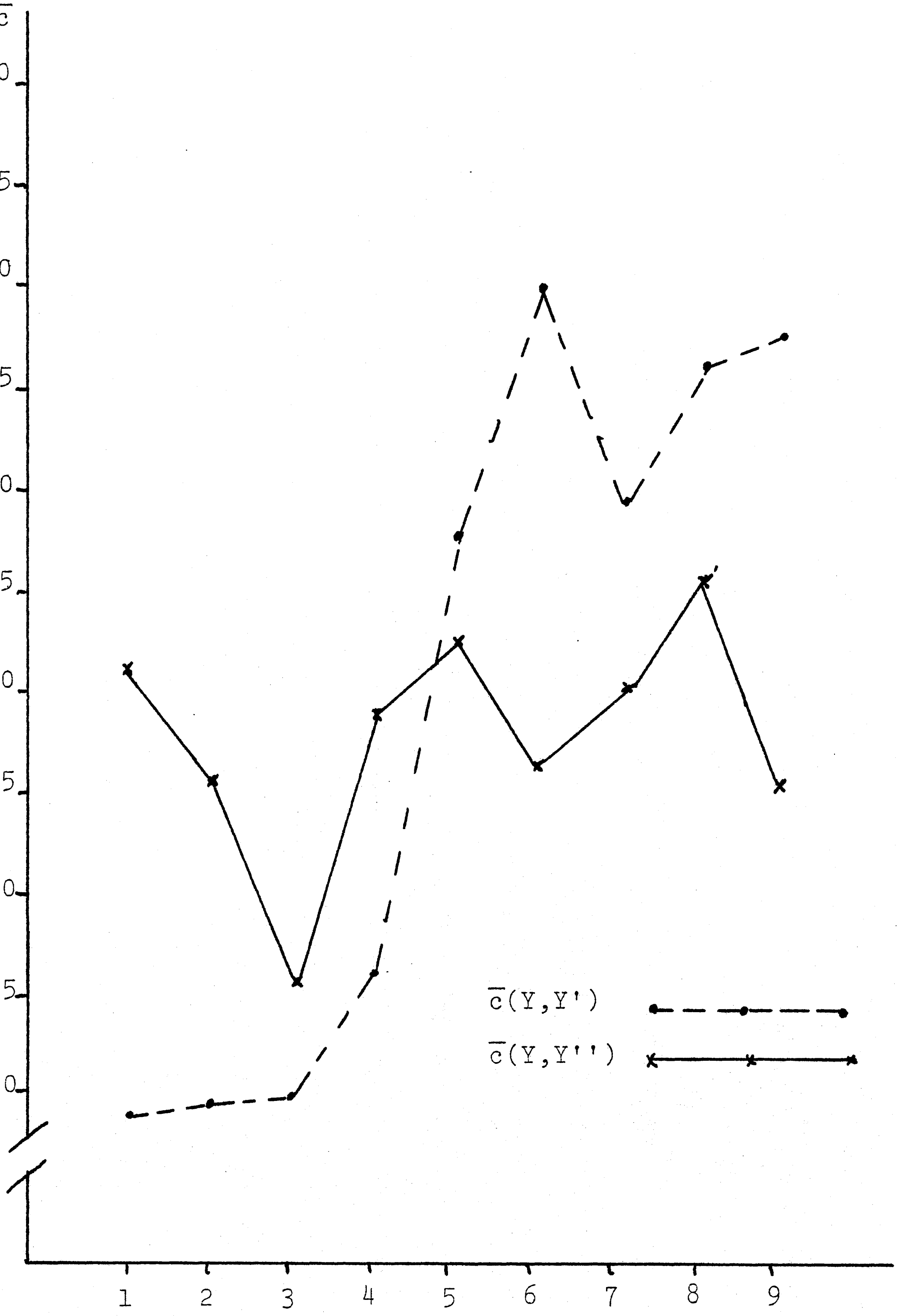

Figure 8. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XII with $(p_1,p_2)=(.3,.7)$

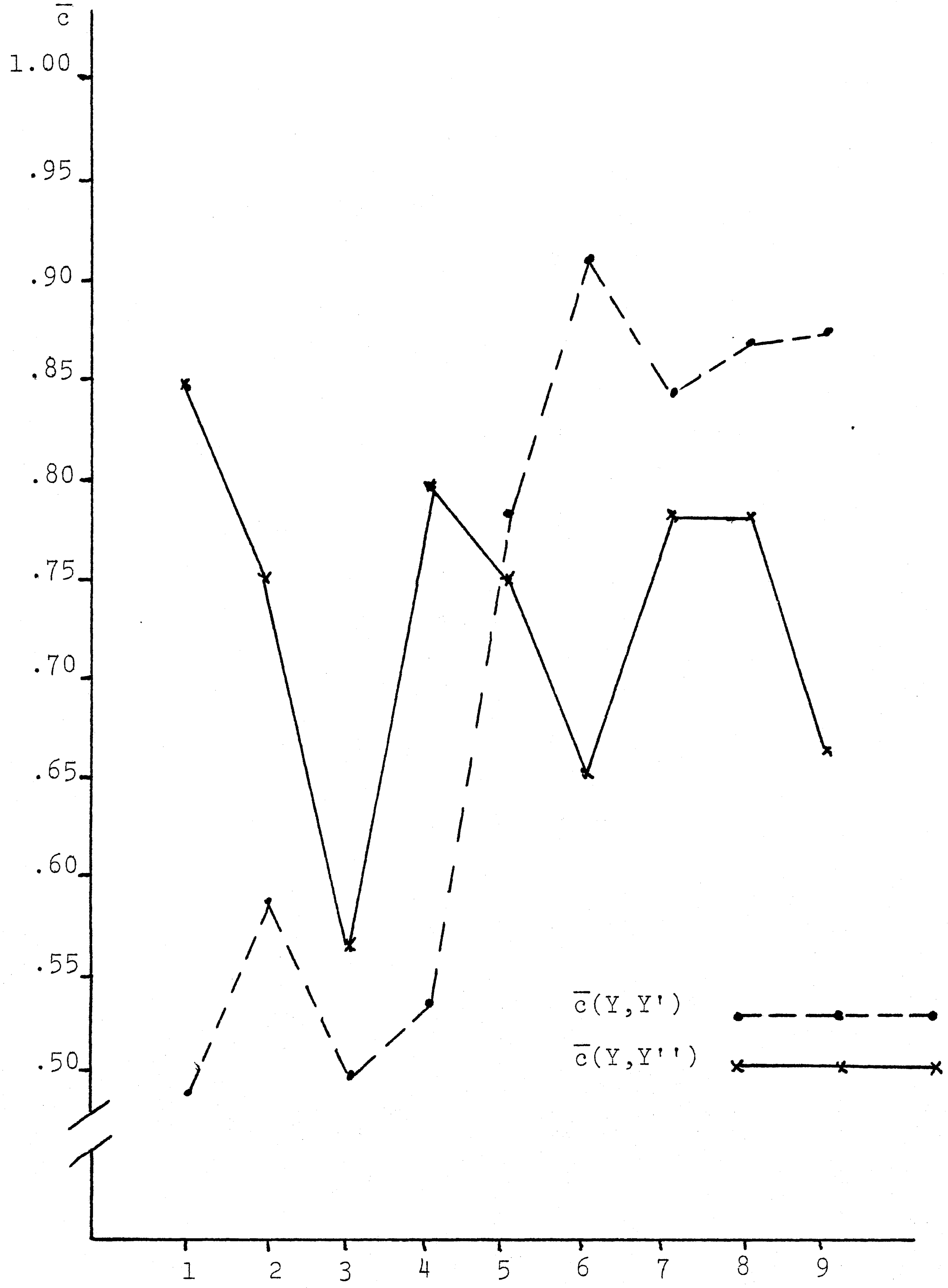

Figure 9. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XII with $(p_1,p_2)=(.4,.6)$

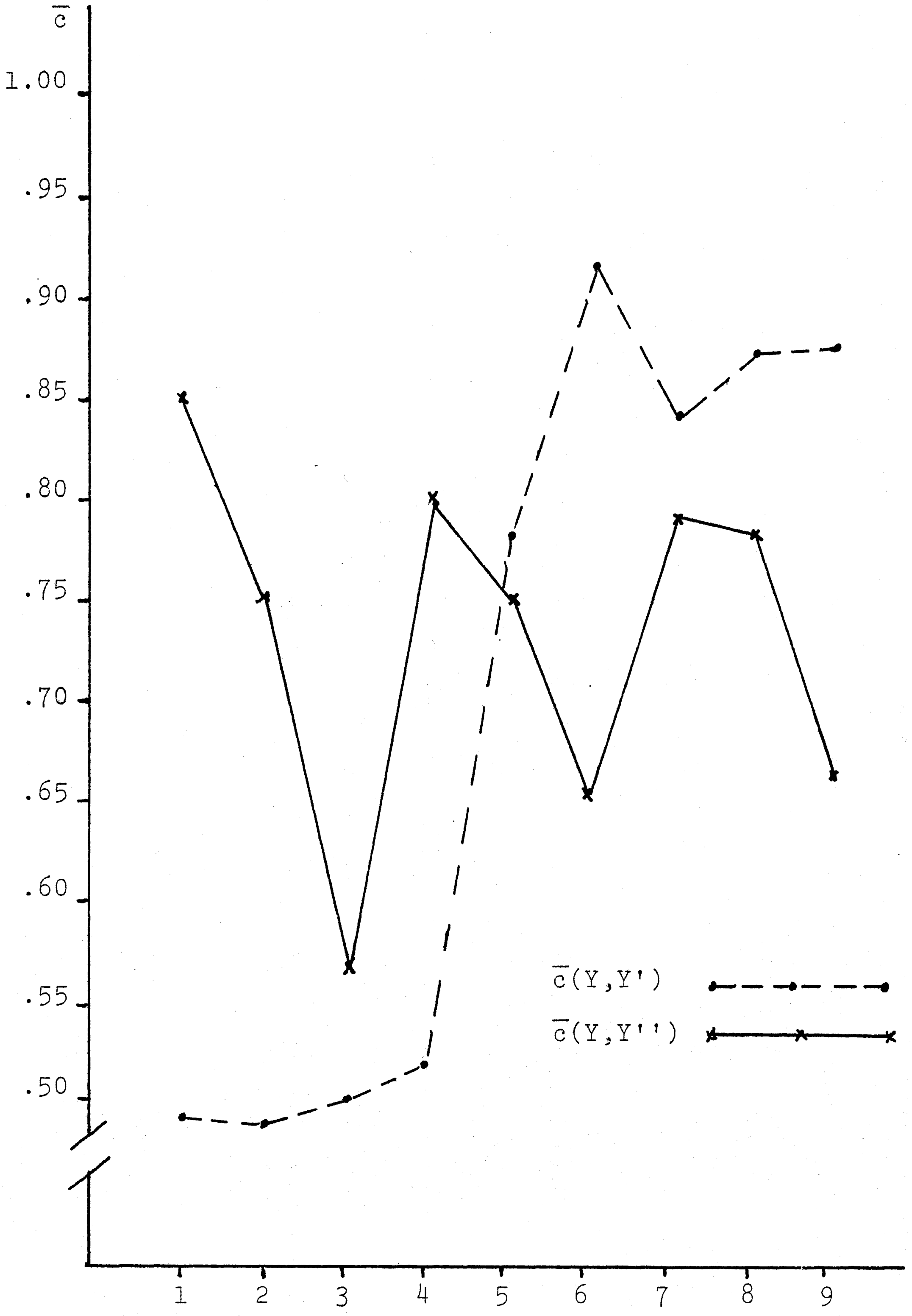

Figure 10. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XII with $(p_1,p_2)=(.45,.55)$

TABLE XIII

A COMPARISON OF $\bar{c}$ ACROSS (p_1,p_2) FOR NINE ALGORITHMS ALONG $\rho = .8$ WHERE $ac = S_{SM}$

p_1,p_2		Single	Average	Complete	(-.25,-.5)	(-.25,0)	(-.25,.5)	(-.5,-.5)	(-.5,0)	(-.5,.5)
		1	2	3	4	5	6	7	8	9
.3,.7	$\bar{c}(Y,Y')$	.48506	.48533	.50234	.58218	.77437	.90554	.79340	.86782	.87889
	$\bar{c}(Y,Y'')$	.71211	.65724	.58018	.70501	.73687	.63986	.67970	.73510	.63078
.4,.6	$\bar{c}(Y,Y')$	.48506	.48575	.49848	.53552	.78434	.91807	.79901	.87069	.87379
	$\bar{c}(Y,Y'')$	.87274	.79361	.54462	.84051	.77531	.68510	.78513	.78078	.64901
.45,.55	$\bar{c}(Y,Y')$	.48506	.48568	.49874	.53145	.79090	.92669	.80457	.87069	.87251
	$\bar{c}(Y,Y'')$	.87927	.75552	.54644	.80634	.74216	.64349	.78207	.77611	.65664

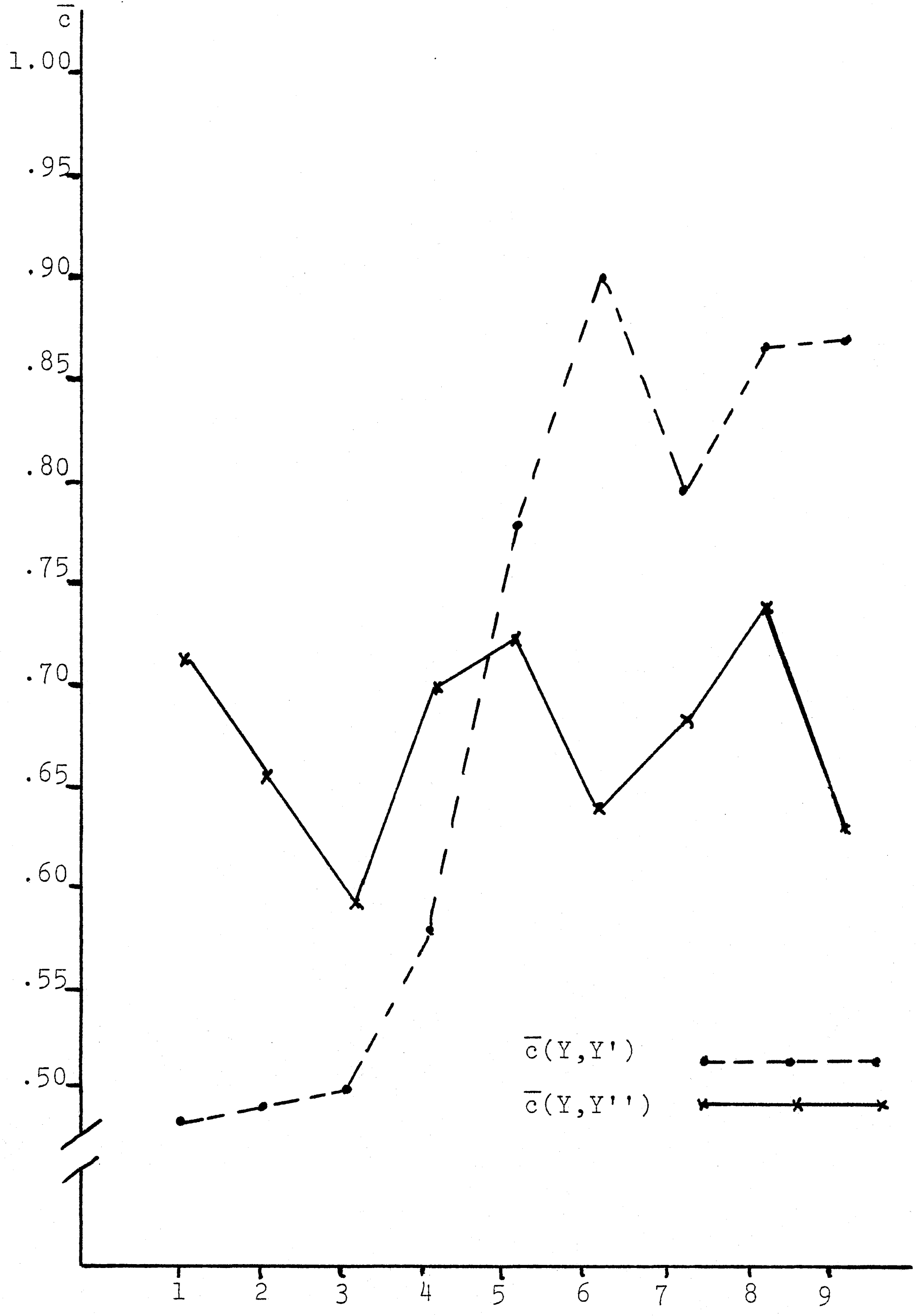

Figure 11. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XIII with $(p_1,p_2)=(.3,.7)$

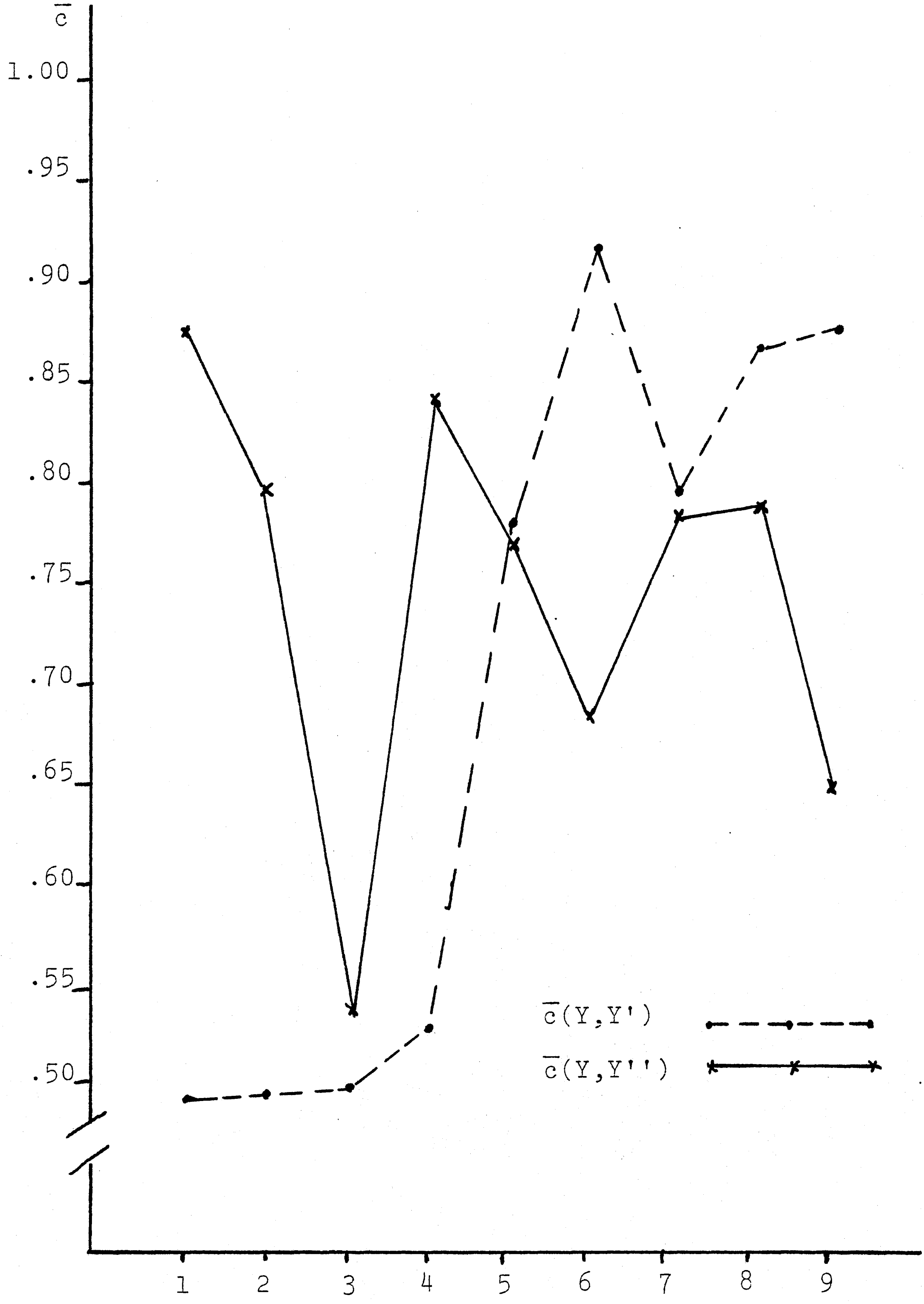

Figure 12. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XIII with (p_1 (p_1,p_2)=(.4,.6)

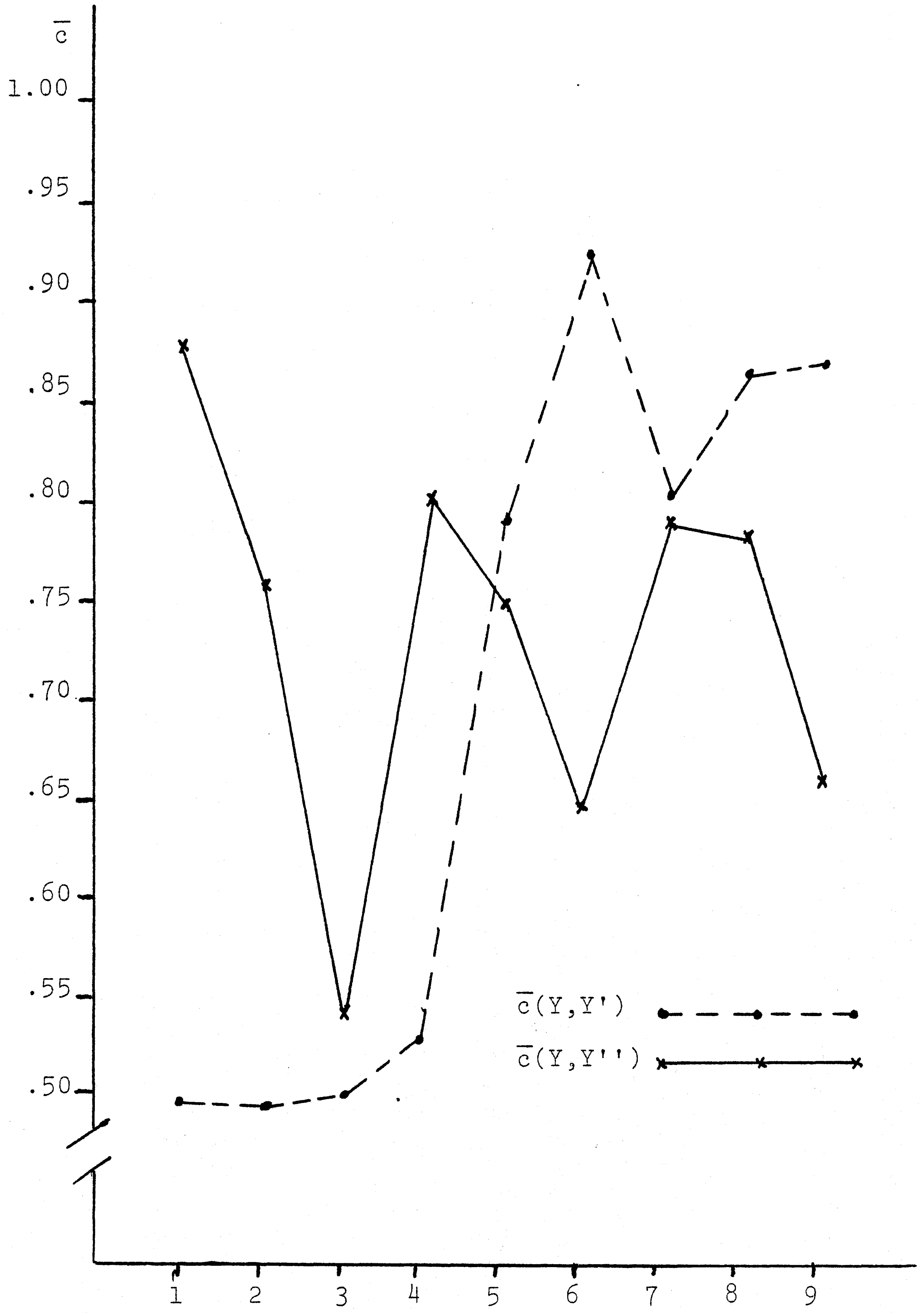

Figure 13. Using $\overline{c}$, a Graphical Representation Across Algorithms of TABLE XIII with $(p_1,p_2)=(.45,.55)$

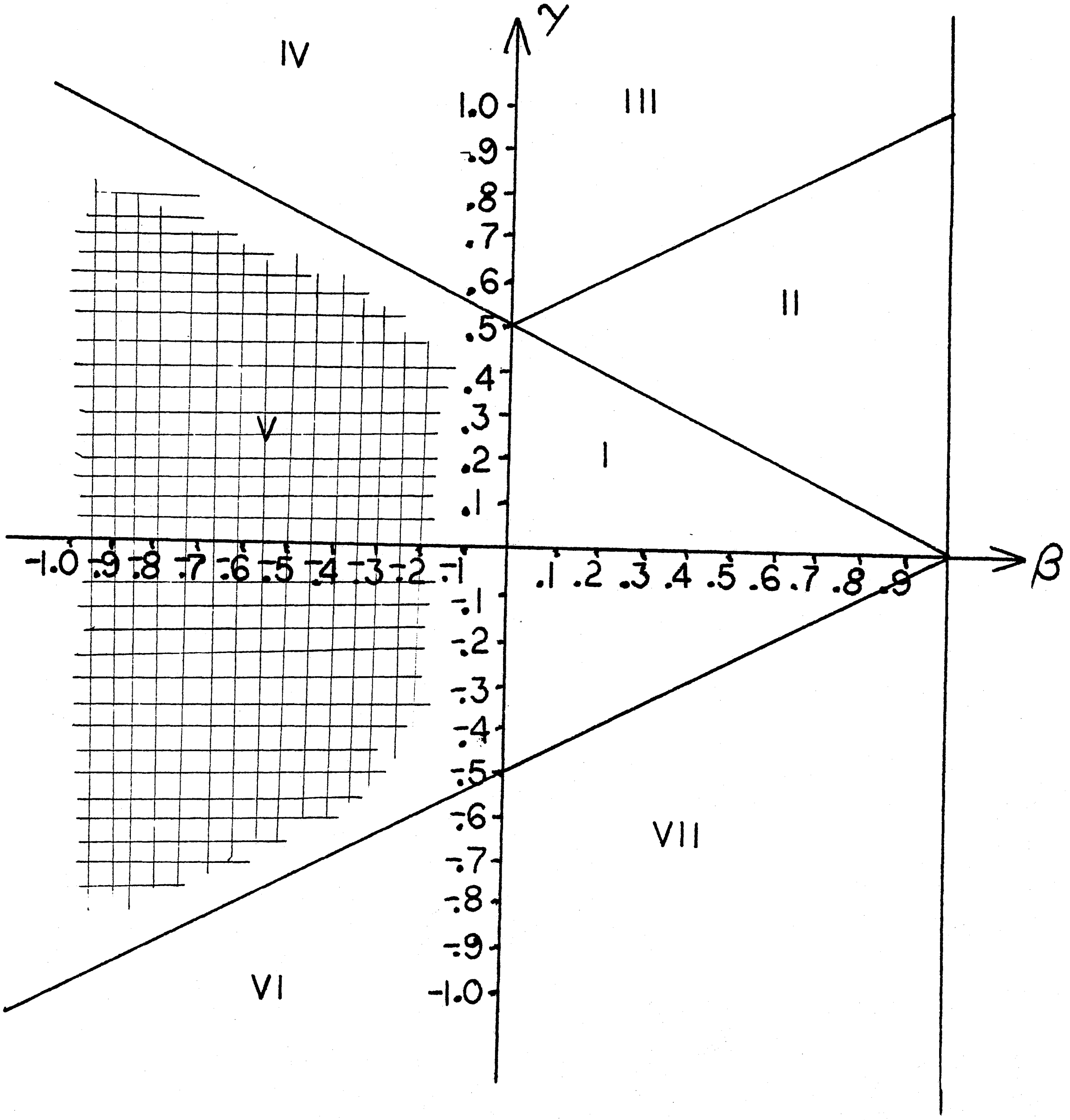

Figure 14. Range of (β , γ) Values for Which Principal Components/Clustering is Better When Using Association Coefficient S_J .

(β , γ)	S_J	S_{SM}
(0.0,-.50)	----	PC/C
(0.0,-.25)	----	PC/C
(0.0, 0.0)	----	PC/C
(0.0, .25)	----	PC/C
(0.0, .50)	----	PC/C
(0.0, .75)	----	PC/C
(-.25,-.50)	PC/C	PC/C
(-.25,-.25)	PC/C	PC/C
(-.25, 0.0)	PC/C	C
(-.25, .25)	PC/C	C
(-.25, .50)	PC/C	C
(-.25, .75)	PC/C	C
(-.50,-.50)	PC/C	C
(-.50,-.25)	PC/C	C
(-.50, 0.0)	PC/C	C
(-.50, .25)	PC/C	C
(-.50, .50)	PC/C	C
(-.50, .75)	PC/C	C

PC/C -- Principal Component/Clustering
C -- Clustering Alone
---- -- No Significant Difference

Figure 15. Choice of Procedure for Given Association Coefficient Over the Eighteen (β , γ) Clustering Algorithms.

VITA 2

Marilyn Ann Gay Sloan

Candidate for the Degree of

Doctor of Philosophy

Thesis: USING PRINCIPAL COMPONENT ANALYSIS PRIOR TO AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS .

Major Field: Statistics

Biographical:

Personal Data: Born in Oklahoma City, Oklahoma, June 15, 1946, the daughter of Mr. and Mrs. Harold M. Gay.

Education: Graduated from Norman High School, Norman, Oklahoma, in May, 1963; received the Bachelor of Science degree with a triple-major in chemistry, zoology and mathematics from the University of Oklahoma, Norman, Oklahoma, in August, 1966; received the Master of Arts degree with a major in mathematics from the University of Oklahoma, Norman, Oklahoma, in August, 1973; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in May, 1983.

Professional Experience: Graduate teaching assistant, Oklahoma University, from September, 1968, to June, 1973; Computer Systems Analyst, Standard Oil Co. (Ind), Tulsa, Okla., from June, 1973, to July, 1978; Computer Systems Analyst and Statistician, Cities Service Oil Co., Tulsa, Okla., from July, 1978, to July, 1981; Consultant in Computer Systems Analysis and Statistics, Tulsa, Okla., from July, 1981 to present.