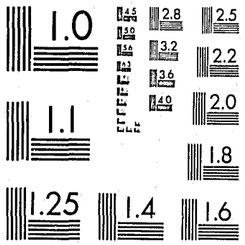


Nonhierarchical Cluster Analysis with Second Order Solutions

Norman Cliff
Robert Cudeck
Douglas McCormick
Linda Collins

Department of Psychology
University of Southern California
Los Angeles, California 90007

October, 1981

Funding for this research was provided by the National Institute of Justice Grant #79-NI-AX-0065.

Abstract

Cluster analysis has often been proposed as an alternative to factor analysis for the reduction of binary data. However, a review of the clustering literature fails to reveal a method that is well-suited to clustering binary variables. An ideal clustering method for binary variables is an efficient, robust method that does not impose a hierarchy or disjointness on the solution. The method should also allow the researcher a choice of several different measures of association. This paper introduces BINCLUS, a clustering procedure incorportating all of these features. Based on previous work on the recovery of ordinal information from binary data, BINCLUS is a flexible, heuristically-based, nonhierarchical clustering method tailor-made for binary data. BINCLUS clusters variables using matrices of Goodman-Kruskal gammas, Pearson r's (phi's), or the quality index q (Cliff, 1979). The clustering solution is presented in the form of a easily interpreted binary cluster membership matrix. In addition, BINCLUS provides a second-order solution that is especially useful when the first-order clusters are not clear-cut.

The method has been applied to extensive artifical data and several sets of empirical responses. It is highly successful in identifying clusters in both artifical and real data.

Nonhierarchical Cluster Analysis with Second Order Solutions

One of the main goals of a multivariate analysis is the resolution of a large number of manifest variables into a smaller number of underlying ones. Often these manifest variables are binary, but most of the widely used data analysis methods are based on linear models. Such models are inappropriate, to at least some degree, for the analysis of binary data. This paper describes a methodology for accomplishing meaningful data reduction with binary data for which linear methods are of questionable appropriateness.

A fundamental concept in multivariate methods is that relationships among objects or variables used to measure them can be described as distances among points in geometric space (Green & Carroll, 1976; Kruskal, 1978; Tatsuoka, 1971). For linear methods based upon covariances among variables, such as factor analysis, the nature of the hypothetical dimensions underlying the objects is studied explicitly. When a satisfying solution to the number and orientation of the dimensions is determined, the relationships among the objects can be examined in terms of their factor scores. These methods might be called "formal geometrical methods", for they produce precise solutions in geometrical terms. But they also make stringent assumptions about the distributions and metric of the variables and about the relations between

covariance and distance in the spatial models, require many cases to achieve a stable statistical solution, and are limited in the number of objects that can be examined practically. Furthermore, they are restricted to studying linear relationships among the variables, which perhaps is not always the correct model for depicting psychological phenomena (Armstrong, 1967). When distributional assumptions cannot be met, when the variables of interest are not interval-level, when the number of variables is large relative to the number of cases, or when other than linear relationships might be thought to exist in the data, an alternative method may be desired.

Such an alternative might be called an "informal geometrical method", and chief among these are the heterogeneous group of techniques known as cluster analysis (Blashfield & Aldenderfer, 1978; Cormack, 1971; Everitt, 1974; Hartigan, 1975; Sneath & Sokal, 1973; Spath, 1980). There are several nearly independent traditions in cluster analysis (Blashfield and Aldenderfer, 1978; Blashfield, 1980), and each has its own terminology and emphases. Our own focus is on the social science tradition, and that segment of it where the focus is on clustering the measures or variables, as opposed to clustering the objects or cases or individuals. Thus, in the discussion that follows, it is assumed that the objective of the cluster analysis is to group together variables that are somehow

behaving similarly or "measuring the same thing." The clustering of cases or individuals is a secondary consideration, and typically would take place subsequent to the clustering of variables. There is, of course, a formal duality in the data. Methods for clustering variables may be used to cluster cases, and vice versa. This does not mean that such role reversals are always advisable since models, measures of association or distance, etc., that are appropriate in one orientation of the data matrix may not be so in the other.

In contrast to formal methods, the informal clustering methods simply seek a way to group variables together into a small number of classes, subject to the condition that all variables within a class are maximally similar to each other, but dissimilar to variables in other classes. Clustering methods are multivariate techniques in that the objects are measured on a number of variables. They are geometrical in that the models at least allude to a multidimensional framework within which the objects or variables are located. However, the precise nature of these dimensions may be a secondary matter in cluster analysis. Instead, these methods merely attempt to describe which variables are related to each other, without formally defining the underlying geometric model.

It is this informality that is the strength as well as the weakness of cluster analysis. For instance, Fleiss and Zubin

(1969) correctly criticized the methods for the "handicap that they proceed from no mathematical or statistical model whatever." But in many instances, realistic data do not conform to a known mathematical model either, especially during early stages of theoretical development when exploratory analyses are emphasized. Thus to proceed with a formal method when the data might be suspected a priori to be inappropriate seems inadvisable. As an example, it is known that application of a standard method of factor analysis to dichotomous data may produce artifactual factors (Carroll, 1961) which are correlated with item difficulty, particularly when the items conform to a Guttman scale (Guttman, 1950). Although these so-called "difficulty factors" have not been studied extensively, they may occur in certain data because relationships among the variables are non-linear in the binary case (Bentler, 1970; Ferguson, 1941; Jöreskog, 1970; McDonald, 1969).

In this paper we discuss further the uses of cluster analysis as an alternative tool for studying relationships among variables. We briefly review the general trends in the methods proposed to date. Then we describe an approach to nonhierarchical cluster analysis that seems especially suited for qualitative data of the dichotomous sort, and will further outline a simple method for improving the solutions via a kind of second-order analysis. Apparently, the idea of second-order or super-ordinate

solutions has not been proposed before. Throughout the discussion we will emphasize the analysis of dichotomous data, although the general method is by no means limited to it. However, the approach to second-order analyses depends upon utilizing dichotomous cluster membership information from the first-order solution. Thus the ability to deal with binary responses is integral to the final results regardless of the character of the original data.

Nature of Clustering Methods

In order to discuss the proposed nonhierarchical method, it seems appropriate to provide a brief overview of other methods. The literature on cluster analysis is now quite extensive, and has developed in at least three distinct traditions (Blashfield, 1980). We will not attempt to be comprehensive in this review, but instead will emphasize general trends which have implications for the current presentation.

Elements of Cluster Analysis Models

The general purpose of cluster analysis is to group a large number of variables or objects into a small number of classes. Overwhelmingly, these classes have been viewed as hierarchical in nature and non-overlapping. A biological taxonomy is the archetypal solution, in which objects are grouped first into species, then species into genera, genera into families, and so forth, even when the emphasis is upon variables rather than organisms. There

is, of course, a hierarchical tradition in the factor-analytic literature (e.g., Cattell, 1978). Sometimes the hierarchy is depicted in a tree-like structure, such that highly related variables appear as outlying branches, which in turn join larger branches of more general organization at successively lower intersections. This model has become so pervasive that hierarchical analysis and clustering methods are now nearly synonymous terms (Blashfield & Aldenderfer, 1978).

There are other kinds of structures as well. In addition to hierarchical organizations, the underlying structure is sometimes viewed as not overlapping and nonhierarchical. In still others it is taken as overlapping and nonhierarchical. In applications in psychology it can be argued that these latter two structures are perhaps more appropriate as representations of psychological phenomena than is a hierarchy. Take, for example, the classification of abnormal behavior (Strauss, Bartko & Carpenter, 1973), which is frequently studied by means of cluster analysis. Although certain aspects of psycho-pathology may be fruitfully characterized like a biological taxonomy, it is by no means clear that the entire spectrum of behavioral disorders is organized in a hierarchy. So it is surprising that other models such as nonhierarchical approaches with overlapping classes have not seen wider application.

Measures of Association

Another component of clustering methods concerns the index of association chosen to represent similarity between the variables. Many indices have been tried (Lorr, 1976), but some are more frequently used than others. The popular alternatives are Euclidean distance and Pearson correlation, the former because it incorporates all the information in profile data (Cronbach & Gleser, 1953), the latter because it is ubiquitous in psychological research. In factor analysis, vector product indices like the correlation are most frequently employed because the model is itself a vector product type, and many procedures must begin with a Grammian matrix of proximity values. But in cluster analysis, no single association index appears to be as theoretically compelling as the correlation is in factor analysis.

Several thoughtful papers have discussed the implications of various association indices in the context of cluster analysis (Everitt, 1974; Fleiss & Zubin, 1969; Lorr, 1976). These writers correctly warn that the type of index used can have a major effect on the relationships which the analysis portrays. Most of these discussions assume that the data are scaled as ordinal or interval level variables. In the psychological literature, it seems that little regard has been paid to the matter of qualitative data, especially binary variables. Most of the common association indices were developed for other types of variables, and their application to ordinal dichotomous data, though

mathematically possible, has remained conceptually clouded at best (Carroll, 1961). This is unfortunate, for dichotomous ratings, such as test item scores and the like, are a fundamental source of information in the behavioral sciences. Ratings of this type have been used to describe such diverse aspects of behavior as the presence of a personality trait, the observation of a particular behavior in school, or a juvenile delinquent's arrest for a certain offense.

Clustering Algorithms

The third major aspect of cluster analysis is the algorithm. Blashfield and Aldenderfer (1978) identified two major approaches to cluster analysis algorithms, namely the agglomerative methods and the iterative partitioning methods. (The former is the classical approach that begins with individual objects and gradually joins them together into classes.) One type of partitioning method is hierarchical division. This method begins with the entire set of objects as a single cluster, which is then broken down into successively smaller clusters. The iterative partitioning methods start with a predetermined set of clusters, and then in a cyclic fashion attempt to re-assign objects to other clusters, thereby maximizing a global measure of cluster homogeneity.

There are several operational approaches to calculating the similarity of an object with the members already in the cluster. Single linkage methods define the similarity of an object to a

cluster as the distance between the object and the closest member. Complete linkage uses the distance between the object and the farthest member. Average linkage uses the arithmetic mean (or some other suitable average) of the distances between an object and all the other members of the cluster.

Blashfield and Aldenderfer (1978) reported that about 75% of the studies they reviewed in recent literature used a hierarchical agglomerative method. They suggested three reasons for the popularity of these methods: (1) hierarchical agglomerative methods have been available the longest; (2) researchers tend to use what has been previously used in their fields; (3) from recent empirical studies, more is known of these methods than any others. Unfortunately, none of these reasons is based upon a theoretical property of the hierarchical agglomerative model.

Ambiguities of Cluster Analysis

A major problem in cluster analysis concerns the decisions of the "correct" number of clusters in a sample of data. Inevitably, this is a trial-and-error process, with the outcome generally decided by the solution that appears most satisfactory to the individual researcher. The same kind of warnings applicable to the number of factors problem in factor analysis are also applicable here (Armstrong & Soelberg, 1968). Recently, investigators have begun to consider objective procedures for deciding the correct number of clusters in a hierarchical solution (i.e., Wainer & Schacht, 1978; McCormick, Cudeck, Cliff,

and Reynolds, Note 1) but thus far no objective guideline has found wide acceptance.

A second difficulty of clustering methods concerns identifying the final members of a cluster. As clusters are constructed, members added at early stages may sometimes turn out to be located on the periphery of the group defined at the final solution, and the composite may be cleaner if some initial members are deleted. Another facet of this problem concerns the decision of when to stop adding members to a cluster. Ideally, the index of within-cluster homogeneity will show a sharp decrease as elements from other clusters are added. With real data, however, the index of within-cluster homogeneity frequently does not show an abrupt change in value as non-members enter. No objective procedures to solve this problem have been forthcoming.

A third ambiguity has to do with selecting starting elements for those algorithms which depend upon them. Many iterative procedures are sensitive to the starting configuration. To the extent that the final groups differ according to the initial set of starting elements, the clustering method will be unsatisfactory.

The Ideal Method

The sections above present what appear to be the principle issues in cluster analysis. From all this one can summarize the desirable properties a clustering method might have. For one thing, the procedure should be tailored to fit the substantive nature of the data under examination, i.e., it should be

hierarchical, disjoint, etc. only when those properties are felt to characterize the data. This stipulation may seem obvious at first glance, but it does not seem to have been discussed very often. Second, the index of association chosen should reflect the special nature of the data which arise in a particular context. This means that the standard correlation or distance measure should at times be critically examined, and perhaps replaced by another that is more appropriate. Third, one would seek a method that is robust over starting configurations, and that will provide a means for detecting cluster boundaries. Fourth, it should be relatively efficient computationally. Finally, because much psychological data are dichotomous, the ideal method should be applicable to such information.

A General Procedure for Nonhierarchical Clustering

The present research was undertaken with the idea that available clustering methods were not well-suited to the data of primary interest to us. These data consist of dichotomous responses and the objective is to group the variables together into homogeneous subsets. Of course the method should generalize to other sorts of data, but this is not a primary motivation. In our data, the persons may, at some later stage, be assigned scores on the scales defined by the subsets of variables, but this is a secondary consideration. The subsets need not be distinct, although this is desirable and they need not be hierarchical, although it would be interesting if the solution indicated this.

In our approach, the general cluster analysis problem is one of identifying subsets of homogeneous items, rather than outlining a pure taxonomy in the traditional sense.

A clustering method has, then, three major aspects: an assumption concerning the type of structure, a method of measuring similarity, and an algorithm for assigning elements to clusters. All three of these are interrelated. Our goal here was to develop a method that is specifically tailored to the clustering of test or questionnaire items to which the responses are typically binary. The idea is to group into the same cluster those items that seem to measure the same underlying variable, an intent similar in many respects to factor analysis. Such clusters are sometimes felt to be overlapping, so the clusters need not be hierarchical or even disjoint, although this result could occur. Furthermore, most available indices of association have various defects when applied to binary data, so the present method is based on indices that are especially appropriate with dichotomous information. Finally, the assignment method should be effective and robust to error and starting position. The procedure discussed herein, termed BINCLUS, is a heuristically based method that is felt to have these features.

Association Indices

It seems intuitively obvious that as characteristics of the data to be analyzed vary, the type of association index used should change also. In most discussions of correlational methods

there is some consideration of the wide variety of association indices that are available and their properties. When the data are dichotomous the matter of an appropriate index is all the more important. In cluster analysis, the association index must serve the dual function of describing relationships between variables and describing within-cluster homogeneity during the analysis. This suggests the possibility that an index may be satisfactory as a measure of association between variables, while not optimum for describing within-cluster homogeneity. For example, the Kuder-Richardson 20 formula may increase monotonically with additional items, even if the average item covariance declines. Therefore, one may question its use in the second capacity.

Recently Cliff (1979) outlined a family of indices for assessing the quality of a set of dichotomous items which seems promising in the context of cluster analysis. These indices are based upon the ordinal information available in pairs of dichotomous items. For the sake of illustration, consider a situation where two persons, A and B, are administered a test consisting of only two items. Furthermore, suppose that A received a score of 1 (i.e., correct) and B a score of 0 on the first item. If A and B again receive scores of 1 and 0, respectively, on the second item, then the information provided by the second item is redundant, because the rank order it suggests for A and B is the same as that provided by the first item. However, if the scores earned by A and B are opposite on the second item, that is, A

receives a 0 and B receives a 1, the ordering information provided by the second item is contradictory to that provided by the first. Finally, consider a slightly different situation where both obtain the same score on the first item and differ only on the second item -- say both A and B get the first item correct, but only A obtains a 1 on the second item. In this case the second item contributes unique information, namely that A's performance is better than B's. Sometimes no ordering information is provided by a pair of items, for example when both A and B receive scores of 1 on both questions. In principle it is possible to consider all pairs of persons and pairs of items, and for each pair of items, find the number of person-person order relations of each type. In practice, this would be difficult if there were more than a few persons, but this is not necessary. It turns out that the number of relations of each kind with regard to items j and k can be deduced from the 2x2 table of responses to items j and k (Cliff, 1979; Cliff and Reynolds, Note 2).

We use r_{jk}, u_{jk}, and c_{jk} to denote the number of relations that are redundant, unique and contradictory in a pair of items, respectively. We let p_j and q_j denote the percentage of persons passing or failing an item, respectively. (Note that $u_{jk} \neq u_{kj}$ unless $p_j q_j = p_k q_k$.) Also, we let r.., u.. and c.. refer to the total number of relations of the three types on a whole collection of items for a sample of individuals.

It turns out that most of the well-known indices of

association can be expressed as a function of the redundant, unique and contradictory order information in a set of items. The critical differences between association indices most often are expressed in the way they utilize the unique information. In a recent paper, Cliff and Reynolds (Note 2) discuss these matters in detail, comparing the properties of many alternatives, and presenting a group of association measures called quality indices which use the three types of information in various weighted combinations.

Algorithm for the Primary Solution

As currently implemented, BINCLUS employs any of several different indices based on these quantities as the measure of cluster-belongingness. One of the simplest is the Goodman-Kruskal γ (Goodman & Kruskal, 1954), which in these terms is

$$\gamma_{jk} = \frac{r_{jk} - c_{jk}}{r_{jk} + c_{jk}} .$$

The major emphasis, however, is on the family of quality indices discussed by Cliff & Reynolds (Note 2) and Cliff (1979). These are defined in terms of a linear combination, t_{jk}, of the r_{jk}, u_{jk} and c_{jk} as compared to bench mark values for the combination. We define

$$t_{jk} = w_r r_{jk} + w_c c_{jk} + w_u(u_{jk} + u_{kj}) .$$

In a quality index, a very common statistical concept, the value of a quantity is compared to best t_b and worst t_w possible values. The quality index between the two items is

$$q_{jk} = \frac{t_{jk} - t_w}{t_b - t_w}$$

Here, the "best" value, t_b, is the value t_{jk} would have if items j and k were a perfect Guttman scale, while the "worst" value, t_w, is the value t_{jk} would have if the two items were completely independent. In both cases, the marginal proportions are held fixed at the observed values.

While the program will accept any combination of weights, a priori considerations and preliminary experimentation (McCormick, et al., Note 1) has led to focussing on $w_r = 1.0$; $w_c = -1.0$; $w_u = .25$. This form of q rewards consistency ($w_r = 1.0$), punishes contradiction ($w_c = -1.0$) and encourages the use of items that are different difficulty ($w_u = .25$) to form the cluster, rather than penalizing differences in difficulty as the phi coefficient and KR20 do. However, any other set of weights can also be used in q.

In addition to γ and the quality indices, the program also will use the Pearson r (phi coefficient) as the basis for clustering, and indeed a wide variety of indices based on r.., u.. and c.. .

Algorithm for the Primary Solution

The program uses the average proximity (average linkage) concept as the basis for constructing clusters, but with some variations that are felt to make it more effective for binary data. The measure of association can be selected from several alternatives,

namely phi, γ, KR_{20}, or some form of q as described above. If cluster C has v members then the v + 1st member is that one which has the highest average proximity to its current members. Let h_{kc} be the proximity of item k to cluster C, and h_{jk} is the proximity of k to j where j is one of the items already in cluster C; then no restriction of disjointness is placed on the assignment

$$h_{kc} = \frac{1}{v}\sum_{j=1}^{v} h_{jk} \quad . \qquad (1)$$

The procedure takes place exactly in this fashion in the case of phi (r), but an additional variation takes place in the case of γ and q so as to make the process more robust for these indices. Due to the variations in the amount of information shared by a pair of binary items, the index can be substantially affected by a few responses if the two items differ in difficulty or popularity. For example, in relating a .90-.10 item to a .10-.90 item, the percentage of individuals who both score 1 on the items can vary only from 0.00 to .10, and is .09 if the items are independent. Thus some values of γ or q are based on less information than others. The approach that is taken here is not to average the values of the indices directly, but to average the numerator and denominators separately. Let n_{jk} and d_{jk} be the numerator and denominator, respectively, of h_{jk}. Then, the actual form of h_{kc} that is used in the case of γ and q is

$$h'_{kc} = \frac{\sum_{j=1}^{v} n_{jk}}{\sum_{j=1}^{v} d_{jk}} \qquad (2)$$

Thus the proximity is less influenced by indices that are based on small amounts of information.

The user can select a subset of items that are to be the nuclei of the clusters, but the default is that all items begin a cluster. Then the program takes each cluster in turn and adds to it the item that has the maximum value of h_{kc} or h'_{kc}. There is no requirement of disjointness, but each time an item has been added to all clusters, the memberships are tested and any that are identical are merged for bookkeeping purposes, and the fact that they merged is recorded. Items are added to each cluster until some form of stopping rule decided upon by the user is satisfied.

Two matrices summarize the information about each cluster. One is the membership matrix M. Assuming that each of the p items starts a cluster, M is p-by-p whose element m_{jc} is the identifying number of the jth item to enter cluster C. The other matrix, H, is also p-by-p, and h_{jc} is the value of the proximity index when the jth member was added to cluster C.

An example using q as the proximity measure will clarify this porcedure. Table 1 contains n_{jk}, d_{jk} and q_{jk}. In Table 2 the membership martix M is shown for the clusters based upon the data in Table 1. Also shown is the H matrix which records the values of the item cluster index as each item is added. Consider cluster 1. As can be seen in the table, the item that produces the largest q when added to item 1 is number 4, where $q_{41} = .68$. Thus $m_{21} = 4$ and $h_{21} = .68$.

Insert Table 1 and 2 about here

To find the third variable for this cluster, we search all pairs of items for the highest h'_{kc}.

This is given by item 7:

$$h'_{71} = \frac{n_{71} + n_{74}}{d_{71} + d_{74}} = \frac{164.3 + 203.6}{445.7 + 398.0} = .436$$

$m_{31} = 7$, $h_{31} = .44$ are recorded in matrices M and H.

The fourth variable was found to be number 3, because

$$h'_{31} = \frac{n_{31} + n_{34} + n_{37}}{d_{31} + d_{34} + d_{37}} = \frac{-38.4 + 113.0 + (-46.8)}{296.0 + 426.2 + 237.9} = .029$$

is the largest third-level h'_{kc} for the first cluster. These data are recorded as $m_{41} = 3$ and $h_{41} = .03$. Additional variables are added to each cluster by continuing the sequence for all available objects, or until some sufficiently small value for h_{tc} has been recorded.

As can be seen in the first cluster, adding a fourth member produces a large drop in the within-cluster homogeneity index. This drop signifies that the objects that are most closely related in the cluster have been added, and that only those which are located farther away, presumably non-members, are left. However, it frequently happens in real data that no large gap between adjacent values will occur. In these cases, it is still possible to define the cluster members by using a somewhat arbitrary solution-wide cutoff point, c_o, which can be used to distinguish

the cluster boundaries. In the present artificial example, a value around .44 would serve. This gives the same three members to clusters 1, 4, and 7 and the same four to 2, 3, 6, and 8. Cluster 5 would be a singlet.

The procedure sketched above was found to be very effective at identifying clusters when true disjoint clusters exist, even when the data have a substantial amount of error (McCormick, et al, Note 1).

Permuting the Clusters

The primary clustering solution is essentially complete at this stage, but with real data the relationships among the clusters may not be obvious. In the present example several clusters contain identical members. The most satisfactory solutions are those where all elements of a cluster "pick each other" like this. However, in many instances with real data there turn out to be a number of clusters that are similar but not quite identical. In order to clarify further the nature of the clusters, we use the information in H and M to produce a final cluster membership matrix B, where $b_{jc} = 1$ if $h_{jc} \geq c_o$ and $b_{jc} = 0$ otherwise. (It is possible to set different cutoff values for different clusters, i.e., have c_{oc} instead of c_o, either in order to allow for a priori notions about cluster homogeneity of as a result of inspection of the h_{jc}, e.g., ending clusters where large gaps indicate the apparent boundaries of clusters.) The matrix B for the sample data is shown in the upper

section of Table 3, when c_o is set at .48, a natural "break" in the index values. Note that in the table the rows are defined in terms of the original order for the items and the structure is not very obvious.

Then to present a more visually useful form of the relationship among the clusters, the rows and columns of B are permuted so as to make similar ones adjacent. A variety of possible approaches to this are possible (see Wilkinson, Note 3 for an approach different from the one used here), and the present approach is a form of nearest neighbor ordering based on γ. It is similar in concept to the seriation procedure of Gelfand (1971). The γ's are computed among all pairs of rows of B, and the two having the largest value are placed next to each other. Call one the left member (l) and the other right (r). This is a two-member chain. Then, the one of the remaining p - 2 that is closest to r is found, and it is placed to the right of r unless it has a higher γ with l, in which case it is put to the left of l. In either case, this new member becomes one of the ends of the chain of three rows. Then, one of the p - 3 non-member rows is added to one end or the other of the chain in the same way. The process continues, adding rows to the ends of the chain until all the rows have been ordered. Then the process is repeated for columns. The process can be quite effective in arranging the data into a visually compelling form, as can be seen in Table 3 where this procedure has placed together the clusters with identical members (columnwise) and the items

with identical cluster memberships (rows).

Insert Table 3 about here

Here the nature of the two cluster solution is clearly visible. It is also apparent that the fifth item is a maverick, essentially independent of the other items, at least using .44 as the membership cutoff. This solution, in fact, perfectly recovers the true relationships among the objects, which were constructed to have just this one cluster of four elements, a second cluster of three, and one singleton, although as can be seen from Table 1 there was considerable noise. When the clusters are disjoint or nearly so, as they are in Table 3, the permuted pattern in B will assume a diagonal block appearance. It will have sections down the diagonal with 1's and sections in the off-diagonal with 0's. The blocks of 1's represent subsets of the objects that jointly select each other.

Second-order Solutions

Even with permutations of B designed to enhance the appearance of the clusters, many solutions are difficult to interpret. Sometimes the results display a rough block-diagonal pattern, but in many instances even these clusters can have "ragged" edges and can be difficult to understand. The permuted patterns can be vague enough to make final conclusions very tentative,

particularly when it is difficult to decide on a cutoff value. It is sometimes, then, helpful to perform a second-order analysis of the permuted information in B. In a second-order analysis, B is treated as a data matrix and itself clustered. The first cluster analysis usually will reveal that some variables are not clustered with any others in the data set. Item 5 in the example is of this kind. Before a second-order analysis is carried out, such objects can be deleted since it is known that they are unrelated to the rest of the set. The second-order analysis can then proceed with the number of elements reduced by the number of singletons in the first analysis.

The "variables" in the second-order analysis are no longer the original items, but rather are the clusters obtained from the first analysis. Inasmuch as the reduced form of B will be the input data for the analysis, the index of association chosen is always selected from among those suitable for binary information. We believe γ is best since it is not sensitive to difficulty. Then the same steps outlined above for a first-order analysis are repeated. Typically, the second-order analysis very clearly reveals which clusters are associated. In the first-order analysis, B is items-by-clusters, and a 1 denotes a row element that was selected by the cluster of the column. In a second-order analysis, the corresponding matrix of binary relations, B*, contains information about clusters-by-superclusters.

The B* corresponding to the two unique super-clusters of the second-order analysis of the data shown in Table 3 is presented in

the upper section of Table 4. Again, these results are interpreted as meaning that clusters 2, 3, 6, and 8 are all in the first super-cluster, while clusters 1, 4 and 7 are in the second.

Insert Table 4 about here

The final step is to relate the original objects to the structure of the super-clusters in B*. One approach to this is to construct a matrix P, with p_{ij} equal to the percentage of clusters in super-cluster j in which object i was a member. More explicitly, we write

$$P = B\, B^*\, E^{-1},$$

where

$$E = \text{diag}\,(B^{*\prime}\, B^*) .$$

Applications

BINCLUS has evolved over some period of time and several forms of it have been used in empirical studies. The artificial data studies of McCormick, et al. (Note 2) employed the basic procedures using indices of binary association to try to isolate clusters. The basic finding was that the method would identify clusters with high accuracy when a true disjoint cluster structure existed, including the identification of singletons, even when the amount of noise in the data was large. Studies using a variety of empirical response matrices have borne out McCormick,

et al.'s conclusions and have demonstrated the resistance of the method to difficulty artifacts provided an appropriate measure of correlation is used.

The empirical study of adjective checklist data by Zatkin, Cudeck, McCormick, and Cliff (Note 4) applied BINCLUS methods to data from a list of adjectives used to describe mothers of children at risk. This was the first study to employ the second-order aspect of the procedure. The first-order results for that study gave a large number of clusters, many of which were highly overlapping. The second-order procedure was quite successful in amalgamating these into a smaller number of relatively clear clusters. In that study, the results were compared to those from a number of widely used hierarchical clustering methods.

Cudeck, Cliff, Collins, and McCormick (Note 5) applied the methods to the criminal records of a large cohort of men, finding a situation somewhat similar to that of adjectives. That is, the primary solution gave a rather large number of fuzzy, overlapping clusters of crimes, whereas the second-order procedure was quite successful in amalgamating these into a small number of rather large, clear clusters. There was also a factor analysis of the same data, and the factor results were quite similar to the clustering. This was the first study to employ the re-ordering procedure that clarifies graphically the nature of the cluster structure.

One example of the utility of BINCLUS as a data-reduction

tool is the following analysis of binary indicators of social deviance. The items are primarily factual questions concerning the family background of the individual or descriptive items concerning the relationships among members of the family. The subjects were 265 of the individuals from the cohort of 9125 consecutive persons born at the Rigshospitalet, Copenhagen, 1959-61. The items are based on interviews of the individual and their parents in 1972 (For a fuller description see Gabrielli and Mednick (1980)). The data were made available by W. F. Gabrielli.

The results of a first-order BINCLUS analysis are given in Table 5, along with brief identifying phrases. The clustering, done on the basis of Goodman-Kruskal gammas between the items, results in a quite clear and striking cluster structure. In the upper left is a large cluster of items that might be called "broken-home" items, various types of departure from a stable two-parent family, along with various circumstances likely to be correlated with this. There is a second fair-sized cluster in the lower

---

Insert Table 5 about here

---

right; this consists entirely of items related to the father not having a normal, healthy role in the family constellation. There are also several small clusters that involve pairs and triplets of items that seem logically related.

The structure of the results seems quite clear, but the cluster membership matrix in Table 5 is typical of other results in that there is a certain amount of fuzziness in the clusters, a core of items that are consistently members of all the clusters along with some less consistent ones. That is, the borders of the clusters are usually diffuse. Sometimes the clusters overlap, but more often there is a blurring of the distinction between cluster members and singletons, rather like the variation in communalities of variables in a factor analysis.

Under these circumstances a second-order analysis can clarify the solution. Table 6 contains the second-order cluster membership matrix and the P matrix for these data. As a result of the second-order analysis it becomes much easier to see the contribution of various items to the clusters. Super-cluster 1 receives strong contributions from "family constellation": items; Superclusters 2 and 3, almost identical, seem to reflect home atmosphere and parental attitudes; Supercluster 4 is the "father's problems" items; and the small Supercluster 5 is a "mother employed fulltime" cluster. There is almost no overlap between the superclusters except for the two that are nearly identical. This kind of a result is typical of data where there seems to be a reasonable structure.

Insert Table 6 about here

Factor analysis of these data resulted in a much less substantively compelling solution. Although the first factor contained many of the same items as the "family constellation" cluster, only the items of moderate frequency of endorsement, i.e. between .4 and .6, had substantial loadings. In fact, loadings on the first rotated factor correlate .77 with frequency of endorsement, and the remaining factors are very small ones. BINCLUS analysis performed on a phi coefficient matrix yielded results similar to the factor analysis, leaving 19 items as singletons, as opposed to the 4 singletons left by the gamma solution. It seems that in this case when phi coefficients are used, frequency acts to break apart variables that may reflect the same underlying dimension.

Another example of the use of BINCLUS on test data is its application to some mathematics test items provided by Wise (see Wise, Note 6).[1] The results of the analyses are quite clear, identifying two major clusters and two items that overlap both. The clusters have a clear interpretation in terms of item content, as shown in Table 7. Super-cluster 1, including items 1-12, contains items tapping basic addition and subtraction and some relatively complex operations involving signed numbers. Supercluster 2, including items 11-16, taps advanced addition involving signed numbers.

Two slightly different BINCLUS analyses performed on these data using gamma coefficients illustrate the effect that changing

the cutpoint c_o can have on a clustering solution. Tables 7 and 8 contain the clustering solution using a cutpoint of .90; Tables 9 and 10 show the solution resulting from a very stringent cutpoint of .95 (the items are highly consistent). The higher cutpoint results in homogeneous clusters with very little between-cluster overlap. However, with this stringent cutpoint the P- matrix indicates that items 8 and 9 have relatively little identity with the first super-cluster and perhaps should be eliminated. On the other hand, the .90 solution results in slightly more heterogeneous clusters with considerably more overlap between them. Items 8 and 9 are now members of every cluster in Super-cluster 1.

The stimulus configuration derived from a multidimensional scaling analysis of the gamma matrix, shown in Figure 1, provides a graphical representation of the effect that changing cutpoint has on the cluster structure of these terms. Items 1-7 and 10 appear close together on the right side of the figure, with items 8 and 9 spread out toward the lower right. The .95 cutpoint forms two small homogeneous clusters, and items 8 and 9 are too distant from cluster 1 to belong in this solution. When the cutpoint is lowered to .90, i.e. more within-cluster heterogeneity is allowed, the size of the clusters is increased, and items 8, 9, 11, and 12 are all included in cluster 1.

Both factor analysis of these data (Wise, 1981) and BINCLUS analysis using phi coefficients as the measure of relationship

developed methods which have an explicit model for dichotomous responses (Christoffersson, 1975; Muthen, 1978) are practically limited in the number of variables they can treat, or require inordinate numbers of subjects for statistical estimation. Traditional approaches for item analysis can be applied once likely subsets have been defined, and factor analysis can be used to examine further the structure of the composite variables. But most popular methods for extracting subsets of items (Burisch, 1978; Hase & Goldberg, 1967) are not convincing with realistic data sets. This is all the more true when little previous work is available to guide the analysis, or when structural information among the items is desired. The present version of nonhierarchical clustering seems promising in this context, as witness the successful application by Zatkin, et al. (Note 4), Cudeck et al. (Note 5), and the biographical information and math examples described above.

A related problem to which this method may be applied concerns the issue of data reduction. Many prospective studies or other large-scale investigations collect massive amounts of information which is frequently qualitative or binary. Before a standard multivariate technique can be used to study relationships in the data, some method for reducing the information to a more manageable form must be undertaken. Often this is done on a rational basis which is arbitrary and prone to bias. Lorr (1976) among others has suggested cluster analysis for this purpose.

yield solutions where items 1-9 and items 11-16 form separate clusters, leaving item 10 as a singleton. Thus, the analyses using gamma and the analyses using phi coefficients are not in serious disagreement except in their treatment of item 10. That item 10 should be controversial is not surprising -- it has by far the most extreme difficulty of the item set. It seems likely that its failure to cluster with other items of more moderate difficulty when phi is used as the measure of relationship is attributable to a difficulty artifact. Since BINCLUS does not require the use of phi coefficients it was able to produce a cluster solution unaffected by difficulty.

A number of other studies are underway or contemplated, and doubtless the method will continue to evolve.

Discussion

To a certain extent, the utility of a method can be measured by the variety of applications for which it is appropriate. Thus it is appropriate to discuss the kinds of problems with which this scheme for nonhierarchical cluster analysis might be used. The first is one which is frequently found in psychology, namely constructing homogeneous sets of items from a heterogeneous pool, a problem that Napior (1972) terms multidimensional item analysis. It is becoming clear that the machinery of factor analysis, generally appropriate for this kind of problem when the data consist of continuous variables, is unsatisfactory with dichotomous items (Nunnally, 1967, Chapt. 8). Other recently

developed methods which have an explicit model for dichotomous responses (Christoffersson, 1975; Muthen, 1978) are practically limited in the number of variables they can treat, or require inordinate numbers of subjects for statistical estimation. Traditional approaches for item analysis can be applied once likely subsets have been defined, and factor analysis can be used to examine further the structure of the composite variables. But most popular methods for extracting subsets of items (Burisch, 1978; Hase & Goldberg, 1967) are not convincing with realistic data sets. This is all the more true when little previous work is available to guide the analysis, or when structural information among the items is desired. The present version of nonhierarchical clustering seems promising in this context, as witness the successful application by Zatkin, et al. (Note 4), Cudeck et al. (Note 5), and the biographical information and math examples described above.

A related problem to which this method may be applied concerns the issue of data reduction. Many prospective studies or other large-scale investigations collect massive amounts of information which is frequently qualitative or binary. Before a standard multivariate technique can be used to study relationships in the data, some method for reducing the information to a more manageable form must be undertaken. Often this is done on a rational basis which is arbitrary and prone to bias. Lorr (1976) among others has suggested cluster analysis for this purpose.

But as has been noted the most frequent kind of cluster analysis used is a hierarchical method, and in this context hierarchies are not generally expected, at least in the sense that hierarchy is meant in the cluster literature. On the other hand, tasks or items that arrange themselves in hierarchies -- as this term is used in the educational literature, e.g. Bart (Note 7) -- are eminently suited for analysis by these methods and indeed it was they that we have had in mind from the beginning. A non-hierarchical method with a provision for treating binary data seems well-suited for this problem. No prior information about the data is required, and so it is attractive in exploratory studies. Furthermore, it is efficient for a first pass through the data when information about possible subsets is desired.

Another source of potential applications are exploratory investigations which study structural aspects among objects without the benefit of a guiding hypothesis. It seems ill-advised to use a hierarchical clustering method if the structure itself is at issue since these methods always find a hierarchy. Similarly, it seems inappropriate to use a method which produces disjoint clusters if it is not hypothesized that such a structure is optimal for the data. Since a large percentage of investigations of this kind are exploratory in nature, it is important that a clustering method be selected that does not force a structure on the data before it is reasonable to do so.

In each of these kinds of applications, the idea of second-order solutions can be useful. Certainly in the case of data

reduction, a higher-order analysis would be valuable as a means of synthesizing findings from a complex analysis. Likewise in problems of multidimensional item analysis a second-order solution would reveal the extent to which the clusters overlap. This information would be useful in judging convergent or divergent association among scales defined by the clusters. Normally one assesses convergence or divergence at the level of aggregated quantitative variables. But a second-order clustering solution would provide this kind of information at the item level.

The method described here runs counter to current trends in psychometrics in that it is a collection of heuristics rather than a monolithic algorithm guaranteed to optimize some objective function such as maximum likelihood or some form of least squares. Two lines of defense are offered, one pragmatic and one philosophical. The pragmatic one is that the procedure has worked. McCormick, et al. (Note 2) found that even without the addition of the permutational and second-order features, it worked very well identifying clusters and separating them from singletons unless the data were very noisy. The permutation and second-order analyses were added when the method was applied to real data. It was found that the outlines of the clusters tended to be fuzzy and these procedures sharpened the definition of them. On the basis of the study by Milligan (1981), and the comparisons in Zatkin, et al. (Note 4), it appears unlikely that any other clustering procedure would work as well.

The philosophical defense has to do with the place of the objective function in data-fitting. As has been stated for many years (Guttman, 1971; Cliff, 1981), the solution one finds is influenced to a greater or lesser degree by the objective function that is tailored to the type of data analyzed. The clustering procedure is of the stepwise variety, adding the "currently best" item to the cluster. Although there is no guarantee that this will result in clusters which have the greatest possible homogeneity, it seems likely that this is hardly possible short of trying all possible combinations of items of each cluster size, i.e., 2^p clusters. Since 2^{60} = 1.15E18, this is impractical for any moderately large set of data. Thus we use a heuristic method here for reasons of economy. However, we attempt to make the process robust but using all possible (or at least very many) starting places.

Operating in conjunction with the provision, based on experience and beliefs concerning our sorts of data, that the clusters are not disjoint, the multiple starting positions tends to lead to numerous similar but non-identical clusters. Two heuristic methods of "purifying" the clusters are then added that attempt to cluster the clusters. The purpose of the first one is mainly graphical. It uses an ordering function to group the clusters into block-diagonal form, insofar as this is possible. Again, the basis of the procedure is heuristic. It is primarily an ordering procedure, and there is no guarantee that it finds

the best possible order, but it should be effective unless the data are highly noisy or somehow perverse. The second-order analysis has also a heuristic basis in the belief that the binary cluster-membership matrix can be meaningfully simplified by the application of the clustering procedure to it, using γ as the index of proximity, again. In their defense, it is asserted that the heuristics forming the basis for these procedures are intuitively sound and therefore preferable to more elegant methods that are rooted in more arbitrary objective functions.

The method runs counter to current trends in another way also. It assumes the intervention of the intelligent, substantively knowledgeable investigator at several points. First the investigator must choose an index, based on experience and beliefs concerning the nature of the data. Then there is the necessity of choosing cutoff values for the index in order to define the membership matrix. This can be expected to take place on partly substantive grounds, and we believe, justifiably so. Thus, the method is not expected to give good results unless the user is sophisticated, except that data with a strong cluster structure will impose itself quite strongly, regardless of the options chosen by the user.

Footnote

[1]These are the "adjusted" responses where the scoring of the 16 items was adjusted using Birenbaum and Tatsuoka's (Note 8) adjustment for use of incorrect problem-solving strategies.

Reference Notes

1. McCormick, D. J., Cudeck, R. A., Cliff, N., and Reynolds, T. J. Clustering binary items. Technical Report 81-1, Department of Psychology, University of Southern California.
2. Cliff, N. and Reynolds, T. J. Dominance relation as a basis for nonparametric test theory. Unpublished manuscript.
3. Wilkinson, L. Permuting a matrix to a simple pattern. Paper presented at the ASA meetings, August 1979, in Washington, D.C.
4. Zatkin, J. T., Cudeck, R. A., McCormick, D. J., and Cliff, N. A general method for non-hierarchical clustering based on binary relations and its comparison to other methods. Technical Report 80-1, Department of Psychology, University of Southern California.
5. Cudeck, R. A., Cliff, N., Collins, L. M., and McCormick, D. J. Patterns in the criminal behavior of young adults. Paper in preparation.
6. Wise, S. L. Some comparisons of four order-analytic methods and factor analysis for assessing dimensionality. Research Report 81-2, Computer-based Education Research Laboratory, University of Illinois.
7. Bart, W. The Ordering Analytic Approach to Hierarchical Analysis. A paper presented to the American Educational Research Association at the 1981 Annual Meeting in Los Angeles.

8. Birenbaum, M. and Tatsuoka, K. K. The use of information from wrong responses in measuring students' achievement. Research Report 80-1, Computer-based Education Laboratory, University of Illinois.

References

Armstrong, J. S. Derivation of theory by means of factor analysis, or Tom Swift and his electric factor analysis machine. American Statistican, 1967, 21, 17-21.

Armstrong, J. S. & Soelberg, P. On the interpretation of factor analysis. Psychological Bulletin, 1968, 70, 361-364.

Bentler, P. M. A comparison of monotonicity analysis with factor analysis. Educational and Psychological Measurement, 1970, 30, 241-249.

Blashfield, R. K. The growth of cluster analysis: Tryon, Ward and Johnson. Multivariate Behavioral Research, 1978, 13, 271-295.

Burisch, M. Construction strategies for multiscale personality inventories. Applied Psychological Measurement, 1978, 2, 97-111.

Carroll, J. B. The nature of data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Cattell, R. B. The scientific use of factor analysis in behavioral and life sciences. New York: Plenum, 1978.

Christofferson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Cliff, N. Test theory without true scores? Psychometrika, 1979, 44, 373-393.

Cormack, R. M. A review of classification. Journal of the Royal Statistical Society (Series A), 1971, 134, 321-367.

Cronbach, L. J. & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Everitt, B. Cluster analysis. London: Heinemann Educational Books, 1974.

Ferguson, G. A. The factorial interpretation of test difficulty. Psychometrika, 1941, 6, 323-329.

Fleiss, J. & Zubin On the methods and theory of clustering. Multivariate Behavioral Research, 1969, 4, 235-250.

Gabrielli, W. F., and Mednick, S. A. Sinistrality and delinquency. Journal of Abnormal Psychology, 1980, 89, 654-661.

Gelfand, A. E. Rapid seriation methods with archeological applications. IN F. R. Dodson, D. G. Kendall, and P. Tautu. Mathematical methods in the archeological and social sciences. Edinburgh: University of Edinburgh Press, 1971.

Green, P. E. & Carroll, J. D. Mathematical tools for applied multivariate analysis. New York: Academic, 1976.

Guttman, L. The principal components of scale analysis. IN S. S. Stouffer, et al, Measurement and prediction. Princeton: Princeton University Press, 1950, Chapt. 9.

Hartigan, J. A. Clustering algorithms. New York: Wiley, 1975.

Hase, H. D. & Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. Psychological Bulletin, 1967, 67, 231-248.

Joreskog, K. G. Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Kruskal, J. B. Bilinear methods. IN W. H. Kruskal & J. M. Tanur (Eds.), International encyclopedia of statistics. New York: MacMillan, 1978.

Lorr, M. Cluster and typological analysis. IN P. M. Bentler, D. J. Lettieri, & Austin, G. A. Data analysis strategies and designs for substance abuse research. Washington, D.C.: NIDA, 1976.

McDonald, R. P. The common factor analysis of multicategory data. British Journal of Mathematical and Statistical Psychology, 1969, 22, 165-175.

Muthen, B. Contributions to factor analysis of dichotomous variables. Psychometrika, 1978, 43, 551-560.

Napior, D. Nonmetric multidimensional techniques for summated rating. IN R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), Multidimensional scaling: theory and applications in the behavioral sciences. Volume 1. New York: Seminar, 1972.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Sneath, P. A. & Sokal, R. R. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: Freeman, 1973.

Spath, H. Cluster analysis algorithms for data reduction and classification of objects. New York: Wiley, 1980.

Strauss, J. S., Bartko, J. J. & Carpenter, W. T. The use of clustering techniques for the classification of psychiatric patients. British Journal of Psychiatry, 1973, 122, 531-540.

Tatsuoka, M. M. Multivariate analysis. New York: Wiley, 1971.

Wainer, H. & Schacht, S. Gapping. Psychometrika, 1978, 43, 203-212.

TABLE 1

QUALITY INDEX q FOR FICTITIOUS DATA WITH EIGHT OBJECTS.[1]

	Objects							
	1	2	3	4	5	6	7	8
	Numerators and Denominators							
1	0	43.0	38.4	337.4	-40.4	-8.1	164.3	-27.0
2	473.0	0	211.0	-18.0	10.0	358.5	-23.2	159.5
3	296.0	452.5	0	113.0	10.0	335.7	-46.8	93.1
4	500.2	699.6	426.2	0	-92.0	-57.4	203.6	-41.0
5	481.6	734.8	445.0	713.8	0	107.0	-10.0	78.4
6	467.1	738.4	457.3	689.9	724.4	0	28.7	163.3
7	445.7	377.2	237.9	398.0	383.8	372.6	0	71.9
8	175.5	264.5	300.1	250.0	260.4	267.1	141.5	0
	q Values							
1	1.00							
2	.09	1.00						
3	-.13	.47	1.00					
4	.68	-.03	.27	1.00				
5	-.08	.01	.02	-.13	1.00			
6	-.02	.49	.73	-.08	.15	1.00		
7	.37	.06	-.20	.51	-.03	.08	1.00	
8	-.15	.60	.31	-.16	.30	.61	.51	1.00

[1]Note: The upper triangular section contains numerators of q, while the lower triangular section contains the denominators.

TABLE 2

MEMBERSHIP MATRIX AND HISTORY

MATRIX FOR SAMPLE DATA.

			Items				
1	2	3	4	5	6	7	8

M Matrix

1	2	3	4	5	6	7	8
4	8	6	1	8	3	4	6
7	6	2	7	6	2	1	3
3	3	8	3	3	8	3	2
6	5	5	6	2	5	6	5
8	7	7	8	7	7	8	7
2	1	1	2	1	1	2	1
5	4	4	5	4	4	5	4

H Matrix

1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.68	.60	.73	.68	.30	.73	.51	.61
.44	.52	.48	.44	.27	.48	.53	.57
.03	.53	.50	.03	.37	.50	.03	.50
.15	.10	.10	.15	.34	.10	.15	.10
.23	.01	.01	.23	.01	.01	.23	.01
.24	.04	.04	.24	.04	.04	.24	.04
.02	.12	.12	.02	.12	.12	.02	.12

TABLE 3

BINARY MEMBERSHIP MATRIX AND

PERMUTED MEMBERSHIP MATRIX

Membership Matrix

	Cluster							
Item	1	2	3	4	5	6	7	8
1	1	0	0	1	0	0	1	0
2	0	1	1	0	0	1	0	1
3	0	1	1	0	0	1	0	1
4	1	0	0	1	0	0	1	0
5	0	0	0	0	1	0	0	0
6	0	1	1	0	0	1	0	1
7	1	0	0	1	0	0	1	0
8	0	1	1	0	0	1	0	1

Permuted Membership Matrix

	Cluster							
Item	2	3	6	8	1	4	7	5
2	1	1	1	1	0	0	0	0
3	1	1	1	1	0	0	0	0
6	1	1	1	1	0	0	0	0
8	1	1	1	1	0	0	0	0
1	0	0	0	0	1	1	1	0
4	0	0	0	0	1	1	1	0
7	0	0	0	0	1	1	1	0
5	0	0	0	0	0	0	0	1

TABLE 4

SUPER-CLUSTER SOLUTION

B*		1	0
		1	0
		1	0
		1	0
		0	1
		0	1
		0	1
P		1.00	0
		1.00	0
		1.00	0
		1.00	0
		0	1.00
		0	1.00
		0	1.00

TABLE 5

CLUSTER MEMBERSHIP MATRIX FOR BINCLUS

ANALYSIS OF DEVIANCE DATA

ITEM	CLUSTER	
	1111111111222222222233333333334444	
	1234567890123456789012345678901234567890123456789012	
20	000000000000000000000000000000000000000000000000	F ALCOHOLIC*
4	000000000111111111111111111000000000000000000	F DEAD
31	111100000111111111111111111100000000000000000	M CHAR DISORDER
5	111111111111111111111111111000000000000000000	BIOL PARENTS NOT MARRIED
6	111111111111111111111111111000000000000000000	BIOL PARENTS NOT TOGETHER
7	111111111111111111111111111000000000000000000	C HAS HAD >1 FAM CONSTELLATION
8	111111111111111111111111111000000000000000000	C NOT LIVING WITH PARENTS
9	111111111111111111111111111000000000000000000	F FIGURE NOT CHILD'S F
10	111111111111111111111111111000000000000000000	C <7 YEARS WITH PRESENT FAM
11	111111111111111111111111111000000000000000000	C <7 YEARS WITH 1 FAM
14	111111111111111111111111111000000000000000000	C HAS LIVED IN ORPHANAGE
15	111111111111111111111111111000000000000000000	C <7 YEARS WITH M
16	111111111111111111111111111000000000000000000	C <7 YEARS WITH F
19	111111111111111111111111111000000000000000000	M ALCOHOLIC
21	111111111111111111111111111000000000000000000	<2 ADULTS IN HOME
36	111111111111111111111111111000000000000000000	C NOT WITH OWN FAM
37	111111111111111111111111111000000000000000000	C NOT WITH BOTH PARENTS
2	000011111111111111111111100000000000000000000	C HAS M SUBSTITUTE
3	110000001110000000000000000000000000000000000	M DEAD
27	111100000000000000000000000000000000000000000	M PSYCHOTIC
43	111100000000000000000000000000000000000000000	M HOSPITALIZED FOR PYCH PROB
44	111100000000000000000000000000000000000000000	M SERIOUS PHYSICAL ILLNESS
17	111111110000000000000000000000000000000000000	C NOT ALWAYS WITH M FIRST YEAR
18	111111110000000000000000000000000000000000000	C NOT ALWAYS WITH M SECOND YEAR
1	000000000000000000000000000000001000000000000	M WORKS
30	000000000000000000000000000000111111111100	F NEUROTIC
28	000000001000000000000000000000111111111100	F PSYCHOTIC
24	000000000000000000000000000000011111111100	F DOES NOT LIKE C
26	000000000000000000000000000000011111111100	F IMMATURE
32	000000000000000000000000000000011111111100	F CHAR DISORDER
34	000000000000000000000000000000011111111100	F ANXIOUS
45	000000000000000000000000000000011111111100	F HOSPITALIZED FOR PSYCH PROB
46	000000000000000000000000000000011111111100	F SERIOUS PHYSICAL ILLNESS
40	000000001000000000000000000000000000000100	PARENTS QUARREL
39	000000001000000000000000000000000000000000	PARENTS PHYSICALLY FIGHT
12	000000000000000000000000000000000000000000011	C SPENT TIME IN WHOLE DAY CARE
41	000000000000000000000000000000000000000000011	M FULLTIME WORK WHILE C <5
22	000000000000000000000000000001110000000000000	INADEQUATE HOME ATMOSPHERE
23	000000000000000000000000000001110000000000000	M DOES NOT LIKE C
25	000000000000000000000000000001111000000000000	M IMMATURE
33	000000000000000000000000000001000000000000000	M ANXIOUS
29	000000000000000000000000000011000000000000000	M NEUROTIC

TABLE 5

(contd.)

UNUSED ITEMS (N = 4)

ITEM	LABEL
13	SPENT TIME IN HALF-DAY CARE
35	MOTHER HAS MISCELLANEOUS MENTAL PROBLEMS
38	FAMILY IS NOT TOGETHER REGULARLY ONCE A DAY
42	MOTHER HAS CHANGED EMPLOYMENT FREQUENTLY

*F = FATHER
M = MOTHER
C = CHILD
FAM = FAMILY

TABLE 6

SUPER-CLUSTER MEMBERSHIP MATRIX AND P MATRIX FOR DEVIANCE DATA

SUPERCLUSTER MEMBERSHIP MATRIX		P MATRIX					
CLUSTER	SUPERCLUSTER	ITEM	SUPER-CLUSTERS				
			1	2	3	4	5
1	10000	20	0.038	0.0	0.0	0.0	0.0
2	10000	4	0.654	0.0	0.0	0.0	0.0
3	10000	31	0.808	0.250	0.0	0.0	0.0
4	10000	5	1.000	0.0	0.0	0.0	0.0
5	10000	6	1.000	0.0	0.0	0.0	0.0
6	10000	7	1.000	0.0	0.0	0.0	0.0
7	10000	8	1.000	0.0	0.0	0.0	0.0
8	10000	9	1.000	0.0	0.0	0.0	0.0
9	10000	10	1.000	0.0	0.0	0.0	0.0
10	10000	11	1.000	0.0	0.0	0.0	0.0
11	10000	14	1.000	0.0	0.0	0.0	0.0
12	10000	15	1.000	0.0	0.0	0.0	0.0
13	10000	16	1.000	0.0	0.0	0.0	0.0
14	10000	19	1.000	0.0	0.0	0.0	0.0
15	10000	21	1.000	0.0	0.0	0.0	0.0
16	10000	36	1.000	0.0	0.0	0.0	0.0
17	10000	37	1.000	0.0	0.0	0.0	0.0
18	10000	2	0.769	0.0	0.0	0.0	0.0
19	10000	3	0.231	0.0	0.0	0.0	0.0
20	10000	27	0.154	0.0	0.0	0.0	0.0
21	10000	43	0.154	0.0	0.0	0.0	0.0
22	10000	44	0.154	0.0	0.0	0.0	0.0
23	10000	17	0.308	0.0	0.0	0.0	0.0
24	10000	18	0.308	0.0	0.0	0.0	0.0
25	10000	1	0.0	0.0	0.0	0.100	0.0
26	10000	30	0.0	0.0	0.0	1.000	0.0
27	01000	28	0.038	0.0	0.0	1.000	0.0
28	01100	24	0.0	0.0	0.0	0.900	0.0
29	01100	26	0.0	0.0	0.0	0.900	0.0
30	01100	32	0.0	0.0	0.0	0.900	0.0
31	00010	34	0.0	0.0	0.0	0.900	0.0
32	00010	45	0.0	0.0	0.0	0.900	0.0
33	00010	46	0.0	0.0	0.0	0.900	0.0
34	00010	40	0.038	0.0	0.0	0.100	0.0
35	00010	39	0.038	0.0	0.0	0.0	0.0
36	00010	12	0.0	0.0	0.0	0.0	1.000
37	00010	41	0.0	0.0	0.0	0.0	1.000
38	00010	22	0.0	0.750	1.000	0.0	0.0
39	00010	23	0.0	0.750	1.000	0.0	0.0
40	00010	25	0.0	1.000	1.000	0.0	0.0
41	00001	33	0.0	0.250	0.0	0.0	0.0
42	00001	29	0.038	0.250	0.0	0.0	0.0

TABLE 7

CLUSTER MEMBERSHIP MATRIX FOR MATH

ITEMS: C_o = .90

ITEM	CLUSTER	OPERATION	SAMPLE ITEM
	1111111		
	1234567890123456		
1	1111111111110000	Subtraction	1 - (-10) = 11
2	1111111111110000	Subtraction	9 - (-7) = 16
3	1111111111110000	Subtraction	-7 - 9 = 16
4	1111111111110000	Subtraction	-12 - 3 = -15
5	1111111111110000	Subtraction	-3 - 12 = -15
6	1111111111110000	Subtraction	-6 - (-8) = 2
7	1111111111110000	Subtraction	8 - 6 = 2
8	1111111111110000	Subtraction	2 - 11 = -9
9	1111111111110000	Addition	6 + 4 = 10
10	1111111111110000	Addition	-14 + (-5) = -19
11	1111111111111111	Addition	-5 + (-7) = -12
12	1111111111111111	Addition	-3 + 12 = 9
13	0000000000001111	Addition	-6 + 4 = 2
14	0000000000001111	Addition	12 + (-3) = 9
15	0000000000001111	Addition	3 + (-5) = 2

TABLE 8

SUPER-CLUSTER MEMBERSHIP MATRIX

AND P MATRIX FOR MATH ITEMS: C_o = .90

SUPER-CLUSTER MEMBERSHIP MATRIX

CLUSTER	SUPER-CLUSTER
1	10
2	10
3	10
4	10
5	10
6	10
7	10
8	10
9	10
10	10
11	10
12	10
13	01
14	01
15	01
16	01

P MATRIX

ITEM	SUPER-CLUSTERS	
	1	2
1	1.000	0.0
2	1.000	0.0
3	1.000	0.0
4	1.000	0.0
5	1.000	0.0
6	1.000	0.0
7	1.000	0.0
8	1.000	0.0
9	1.000	0.0
10	1.000	0.0
11	1.000	1.000
12	1.000	1.000
13	0.0	1.000
14	0.0	1.000
15	0.0	1.000
16	0.0	1.000

TABLE 9

CLUSTER MEMBERSHIP MATRIX FOR MATH

DATA: MINIMUM h'_{kc} = .95

ITEM	CLUSTER
	1111111
	1234567890123456
1	1111111111110000
2	1111111111110000
3	0011111110000000
4	0011111111110000
5	1111111111110000
6	1111111111110000
7	1111111111110000
8	0000000010000000
9	1100000000000000
10	1111111111110000
11	0000000001111111
12	0000000001111111
13	0000000000001111
14	0000000000001111
15	0000000000001111
16	0000000000001111

TABLE 10

SUPER-CLUSTER MEMBERSHIP MATRIX AND P MATRIX FOR MATH

DATA: MINIMUM h'_{kc} = .95

SUPER-CLUSTER MEMBERSHIP MATRIX

CLUSTER	SUPER-CLUSTER
1	10
2	10
3	10
4	10
5	10
6	10
7	10
8	10
9	10
10	10
11	10
12	10
13	01
14	01
15	01
16	01

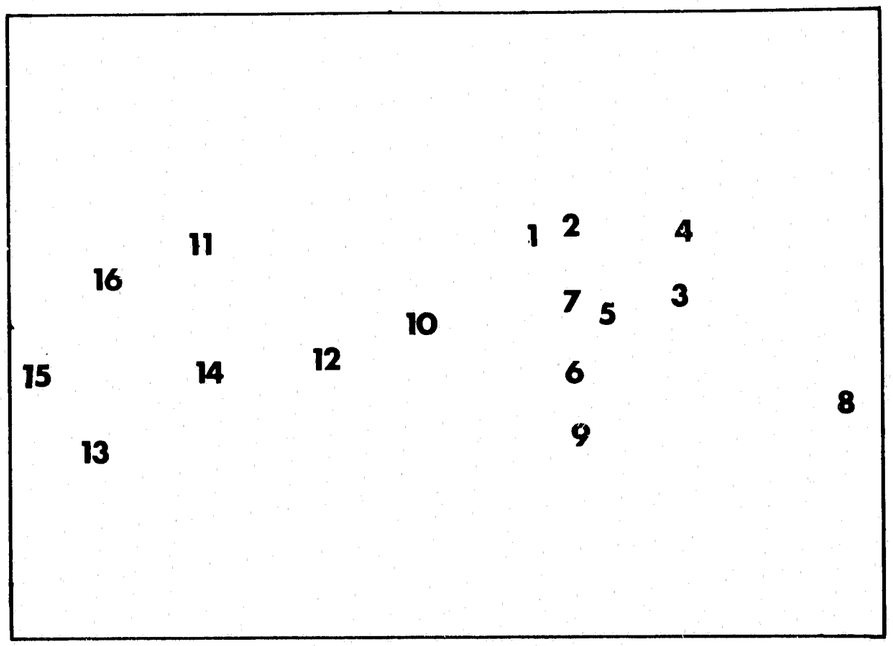
P MATRIX

ITEM	SUPER-CLUSTER 1	SUPER-CLUSTER 2
1	0	1.000
2	0	1.000
3	0	.583
4	0	.833
5	0	1.000
6	0	1.000
7	0	1.000
8	0	.083
9	0	.167
10	0	1.000
11	1.000	.250
12	1.000	.250
13	1.000	0
14	1.000	0
15	1.000	0
16	1.000	0

FIGURE 1

MULTIDIMENSIONAL SCALING OF

MATHEMATICAL ITEMS



END