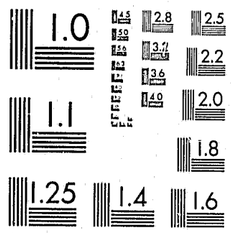


BINCLUS: NON-HIERARCHICAL CLUSTERING OF BINARY DATA

Norman Cliff
University of Southern California
Douglas J. McCormick
American Telephone and Telegraph
Judith Zatkin
University of Southern California
Robert Cudeck
University of Minnesota
Linda M. Collins
University of Southern California

This paper is the result of longterm collaboration among the authors on this research, and the order of authorship is to some extent arbitrary, not necessarily representing relative degrees of contribution to the work. The research was supported in part by the National Institute of Justice, Grant #79-NI-AX-0065.

Abstract

Available methods for grouping together variables into more or less homogeneous sets have some shortcomings when applied to binary data such as test items. This paper describes a form of cluster analysis designed to maximize within-set homogeneity which can be used as an alternative. The procedure is agglomerative and non-hierarchical, based on a particular form of average-link methods. The distance measure used is any one of several forms of association index that are felt to be appropriate for binary data. The procedure includes a type of "second-order" clustering designed to identify the most consistently forming clusters.

The procedure was extremely successful in identifying the clusters in artificial data and seemed very satisfactory in identifying an appropriate cluster solution in several sets of empirical data.

BINCLUS: NON-HIERARCHICAL CLUSTERING OF BINARY DATA

Introduction

Frequently, the motive of an exploratory factor analysis is the hope that it will be possible to group variables into approximately unidimensional subscales. Unfortunately those methods which serve well when applied to multivalued variables are of questionable utility when the variables are dichotomous.

The basic problem is that the linear factor model cannot be expected to hold very well when the observed scores are binary. This lack of fit of the model is reflected in the indices measuring association among the variables. Correlations, covariances and cross-products may allow sensible analysis of continuous measures for a variety of purposes, but their possible application to binary measures is uncertain because they are affected not only by agreement in the basic measurement of two variables but also the similarity in the frequency of endorsements. For example, suppose one has two identical continuous number series which are dichotomized differently for each series. One may obtain correlations which range from unity almost to zero simply by changing the value for each series below which all numbers become zeros and above which they become ones. Fairly modest differences in the two frequencies of 1's in

the two measures are known to result in substantial decreases in their correlation. (See, e.g., Carroll, 1961)

Consider the small correlation matrix in Table 1 and associated list of frequencies.

(Insert Table 1 Here)

The nature of the correlation coefficient obscures the fact that in actuality items a and b form a perfect scale, and so do c and d. That is, everyone who passes b also passes a, and everyone who passes d passes c. The correlations make it look as if a is related to c and b is related to d, but this is an artifact of their similarity in difficulty and the correlation between the two scales. Correspondingly, a factor analysis of these data will yield two factors, one for a and c, the easy items, and one for b and d, the hard items. The factor's exact nature will depend on the specific extraction and rotation option, but the factors above would be a typical solution using standard common factor analysis.

The second difficulty, if one intends to factor analyze binary data, is that if one abandons the questionable product moment indices, less conventional indices can no longer be interpreted as representing scalar products of vectors, and familiar concepts like variance-accounted-for no longer apply. Since the rationale for factor analysis is to account for the correlations or covariances, interpreted as scalar products, using a reduced number of underlying variables, the entire process becomes problematic. This can be especially true if the results are presented to an audience which attributes to them a conventional factor analytic interpretation.

Stated more mathematically, the factor model is

$$x_{ij} = \sum_{m=1}^{r} y_{im} a_{mj} + e_{ij} \qquad (1)$$

where:

x_{ij}= the value of the ith observation on the jth variable

y_{im} = the value of the ith observation on the mth common factor

a_{mj} = the regression coefficient of the mth common factor for predicting the jth variable

e_{ij} = the value of the ith observation on the jth unique factor.

This leads, under the usual assumptions of uncorrelated unique factors, to

$$R = A \Phi A' + U^2 \qquad (2)$$

where R is the observed correlation matrix, A is a matrix of coefficients from (1) and U^2 is the diagonal matrix containing the uniquenesses. If R is not a scalar product matrix derived from X, i.e., if it is simply a matrix of "measures of association", then (2) does not follow from (1), and R may be non-grammian. Thus the results of applying factor analytic methods may be misleading or difficult to interpret.

In short, factor analysis has severe shortcomings when applied to binary data. An alternative to factor analysis is needed for the purpose of grouping together binary measures into homogeneous subsets. While cluster analysis presents an alternative method of finding homogeneous subsets, the goals of available clustering methods seem not well suited to clustering the binary data of primary interest here. These data consist of an entities (persons) by variables matrix, where the scores are binary (dichotomous) and the object is to group together <u>variables</u> into homogeneous subsets. Thus, the overall goal we have in mind for the cluster analysis is more like that of factor analysis than like taxonomy.

The present paper describes an approach to cluster analysis developed for this sort of binary data. The clustering procedure maximizes the homogeneity of subsets through an agglomerative process that clusters variables measuring similar properties. The resulting subsets of variables need not be disjoint, and there is no attempt to make them hierarchical. The clustering is based upon matrices of association indices between pairs of variables, so the choice of a measure of association suitable for binary measures is crucial.

<u>Association Indices for Binary Variables</u>

<u>Problems with Traditional Indices</u>

A presumably general principle of clustering is that things that are "close" belong in the same cluster whereas things that are "distant" belong in separate clusters. The definitions and measures of "close" and "distant" vary, depending on the type of problem and the nature of the data. Our goal here is to put together <u>variables</u> that measure more or less the same thing, based on dichotomous responses. These ground rules mean that several common measures of distance may be unsuitable to varying degrees for this purpose. Euclidean distance is the most obvious. In the case of a persons-by-variables binary data matrix where the objective is to cluster variables, Euclidean distance, d_{jk}

is simply the number of persons who have different values on the two variables. As such, it is heavily influenced by the difficulty or popularity (p-value) of the two variables: $d_{jk} \geq |p_j - p_k|$ where p_j and p_k are the proportions of persons having a value of one on variables j and k. The Pearson correlation (phi) is affected similarly, albeit less stringently, as is well known, and the same is true of the inter-variable covariance. Thus, clustering procedures based on these indices may tend to group apart those variables that differ in p-value. Also, grouping variables so as to maximize Kuder-Richardson 20 reliability will show the same effect, because KR_{20} is a function of the average covariance.

Alternative measures of association

Cliff (1979, Cliff & Reynolds, Note 1) has provided a conceptual framework which may be used to describe any index applied to binary measures. This approach, an outgrowth of test theory, emphasizes the order relations created by variables among the people measured. If a variable orders two people, for example if one person misses a test item and the other answers it correctly, the order created by a second item can have three possible outcomes. First, the order provided by the second item can correspond to the order provided by the first item, thus providing redundant information concerning the person order. Second, it can contradict the first order by placing the persons in reverse order. Finally, the second item may provide no ordering

information at all if both persons miss or both persons answer it correctly. In this last instance the first item provides an ordering where the second item does not. This is called a unique relation. Unique relations may be unique to the first item but not present in the second (u_{jk}), as explained above. If the first item did not provide ordering information but the second does, then the other form of unique relationship (u_{kj}) exists. If more than one pair of people are given the items, the relationship between the two items j and k can be expressed as so many redundant pairs, (r_{jk}), so many contradictions (c_{jk}) and so many unique relations (u_{jk} and u_{kj}).

The r_{jk}, c_{jk}, and u_{jk} are readily obtainable from the usual 2 x 2 contingency table for two items shown in Table 2.

(Insert Table 2 Here)

Here, $r_{jk} = wz$, $c_{jk} = xy$, and $u_{jk} = wy + xz$. The sum of the three types of relations is $n_j(n - n_j)$, where n_j is the number of 1's on variable j. (i.e., the total is proportional to the item variance.) We let the symbols, r.., c.., and u.. stand for the total number of redundant,

contradictory and unique relations, summed across pairs of items.

Many commonly used indices of association can be expressed in terms of redundant, contradictory, and unique relations. When common indices are expressed in this way, it becomes clear how they all increase as the number of redundant relations increases and decrease with the number of contradictions. The differences between the indices lie mainly in their treatment of unique relations. For example, Equation (3) shows Pearson's r expressed as a function of r.., c.., and u...

$$\text{Pearson } r = \frac{R_{jk} - C_{jk}}{\sqrt{(R_{jk} + C_{jk} + U_{jk})(R_{jk} + C_{jk} + U_{kj})}} \qquad (3)$$

Where: U_{jk} = relations unique to item j and not found in k

U_{kj} = relations unique to item k and not found in j

The Pearson r extracts a mild penalty for unique relations by putting them in the denominator, that is, unique relations shrink Pearson r's. Equation 4 shows the formula for KR_{20}.

$$KR_{20} = \frac{x(R_{..} - C_{..})}{xR_{..} - (s - 2)C_{..} + U_{..}} \qquad (4)$$

Where: x = the number of items

KR_{20} is derived from Pearson's r and incorporates unique relations in a similar way. The fact that KR_{20} and r are both diminished by unique relations is another way of looking at their well known tendency to be reduced by differences in marginal distributions.

The Goodman-Kruskal gamma is expressed in Equation (5).

$$\text{Goodman-Kruskal Gamma} = \frac{R_{..} - C_{..}}{R_{..} + C_{..}} \qquad (5)$$

By ignoring unique relations, Gamma avoids the problem of limiting scales to variables having similar frequencies.

A different approach is to assign a positive weight to unique relations. One such possible index is q (Cliff, 1979), shown in Equation 6, where t in the equation is a combination of redundant, contradictory and unique relations weighted 1, -1 and .25.

$$q = \frac{T - T_c}{T_b - T_c} \qquad (6)$$

Where: $T = R_{..} - C_{..} + .25\,U_{..}$

T_c = the value of T given the observed marginals and random responses

T_b = the value of T given the observed marginals and a perfect Guttman scale

t_c is the value of t which would be expected with the observed marginals if the data were merely random. t_b is the value of t which would occur given the observed marginals and a perfect Guttman scale. Because it gives uniqueness a positive weight, q tends to favor sets of items which have different difficulty levels. Within-cluster homogeneity can be characterized in terms of the average value of any of these indices, to say nothing of many alternatives. Their main differences are seen here to be in the way "unique" ordinal relations are treated, and the bias of the more familiar correlation and KR_{20} measures is toward penalizing unique relations, whereas gamma treats them neutrally, and q rewards them.

The clustering procedure described here forms clusters of items so as to maximize within-cluster homogeneity in any of these forms, or in terms of any other simple function of r.., u.., and c... At this time, analyses have been performed using these four particular forms, Pearson r, gamma, KR_{20}, and q.

BINCLUS is a computer program that was written to cluster binary variables. The program uses the average linkage concept as the basis for constructing clusters, but with some variations that are felt to make it more effective. The measure of association can be phi, gamma, KR_{20}, or q as described above, or various other functions of r.., u.., and c...

Algorithm

In BINCLUS, each variable begins a cluster. (At the user's option, a subset can be designated as starting variables.) Then, to each cluster is added the variable that has the highest value of the proximity index with the variable that started the cluster. The process continues, each time adding to the cluster the variable with the highest proximity to the variables already in the cluster. That is, if cluster c has v members then the v + 1st member will be that one which has the highest average index of association with current members. Assignment of a new item to cluster c does not depend on whether or not that item already belongs to some other cluster. It is possible for an item to be placed in any number of clusters providing that item has a strong enough association with each of the

clusters. A part of the output is the cluster-membership matrix M, where m_{vc} is the index of the vth member added to cluster c.

The procedure takes place exactly in this fashion in the case of correlations, but in the case of gamma and q a modification is introduced to make the process more robust for these indices. The approach that is taken 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 the inter-variable proximity between variables j and k. Then, the actual form of h_{kc}, the average proximity of variable k to cluster c, that is used in the case of gamma and q is

$$h_{kc} = \frac{\sum_{j=1}^{v} n_{jk}}{\sum_{j=1}^{n} d_{jk}} \qquad (7)$$

where j refers to variables already in the cluster. The reason is that, due to the variations in the amount of information shared by a pair of binary variables, the index can be substantially affected by a few observations if the two variables differ in frequency. For example, in relating

a variable with 90% endorsement to a variable having only 10% endorsement, the percentage of individuals who have value 1 on both variables can vary only from 0.00 to .10, and is expected to be .09 even if the variables are independent. Thus some values of gamma or q are based on less information than others. The procedure of averaging numerators and denominators separately gives lower weight to indices based on less information.

Illustration of the Procedure

An example will be described in order to clarify this procedure. It uses q as the proximity measure. Table 3 contains n_{jk}, d_{jk} and q_{jk} from an 8-variable set.

(Insert Table 3 Here)

In Table 4 the membership matrix M is shown for the clusters based upon the q-matrix in Table 3.

(Insert Table 4 Here)

Also shown is the H matrix which records the values of the item-cluster index as each variable is added. Consider cluster 1. As can be seen in Table 3, the variable that produces the largest q when added to variable 1 is number 4, where $q_{41} = .68$. Thus $m_{21} = 4$ and $h_{21} = .68$.

To find the third variable for this cluster, we search all items for the highest h_{kc} (the highest "average" item-cluster q). This is given by variable 7:

$$h_{kc} = \frac{n_{71} + n_{74}}{d_{71} + d_{74}} = \frac{164.3 + 203.6}{445.7 + 398.0} = .436 \qquad (8)$$

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

The fourth variable was found to be number 3, because

$$h_{31} = \frac{n_{31} + n_{34} + n_{37}}{d_{31} + d_{34} + d_{37}} = \frac{-38.4 + 113.0 + (-46.8)}{296.0 + 426.2 + 237.9} = .029 \qquad (9)$$

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_{kc} has been recorded.

Deciding on cluster boundaries

A necessary feature of such a system is a way of deciding on the boundary of a cluster. One such rule is to look for a sudden drop in h_{kc} as some new item is chosen. As can be seen in the first cluster, adding a fourth member produces a large drop in the homogeneity index. This drop signifies that the objects that are most closely related in the cluster have already been added, and that only those which are located farther away, presumably non-members, are left. When such a drop in index values occurs, it marks a natural cluster boundary. However, it frequently happens in empirical 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 cutoff value, which can be used to define 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

singleton. These results are summarized in the third output matrix, the reordered binary membership matrix. In this p x p matrix V, v_{jc} = 1 means that item j is a member of cluster c, and v_{jc} = 0 means that it is not. The rows and columns are reordered so as to make clusters adjacent that have similar membership and make variables adjacent that belong to the same clusters. (An algorithm for doing this rearrangement is described in a later section.) The V matrix for these data is shown in Table 5.

(Insert Table 5 Here)

Simulation Studies

Data generation

Binclus has been evaluated using both simulated and empirical data. Data for the simulation studies were generated using the Birnbaum two parameter logistic item response model. The Birnbaum model describes item and person characteristics in reference to a single underlying trait or dimension. To create a multiple trait or multidimensional set of responses, the model was used repeatedly to generate subsets of the 24 items. A five-by-three design--five types of subject ability/item difficulty

distribution combinations by three levels of item discrimination--was used. Table 6 details the parameters used in each experimental condition.

(Insert Table 6 Here)

The three levels of item discrimination used were low (a=.5) , moderate (a=1.0) and high (a=3.0). The five subject ability/item difficulty combinations were produced by varying the distributional shape, mean, and standard deviation of the theta (ability) and b (difficulty) parameters. The normal data sets represented what most researchers would consider the least challenging case for a conventional analysis, that is, persons sampled from a normal distribution of ability taking items sampled from a normal distribution of difficulty. The low frequency condition was intended to simulate situations where marginal frequency in the item set is relatively homogeneous but extreme. The rectangular and mixed data sets simulated situations of highly variable marginal frequency. Both these conditions used rectangular distributions of item difficulty, but the rectangular condition used a rectangular

distribution of subject ability as well, while the mixed condition used a normal distribution of subject ability. The bimodal data sets included items of both very low or very high marginal frequency, with few items between. This was included as the classic example of a situation where product-moment indices should suffer from extreme differences in item distributions.

Each of the 15 conditions was replicated five times, resulting in 75 binary matrices. Each matrix contained one eight-item cluster, one six-item cluster, one four-item cluster, and six singleton items. Within each cluster, the person-items were sampled from one of the joint distributions described above. Each of the 500 simulated subjects had a true score for the ability underlying each cluster or singleton, and these abilities were uncorrelated across clusters.

Deciding on Cluster Boundaries in Artificial Data

Cluster boundaries in the BINCLUS analysis were decided upon in the following way. Each of the 18 items belonging in one of the three subsets started a cluster, as described above. Items were added to each cluster until the cluster contained the same number of items as the subset to which the starting item belonged. For example, if an item belonged to the eight-item subset, it was allowed to add seven items to the cluster it began. The result was eight

eight-item clusters, six six-item clusters, and four four-item clusters.

This set of clusters was finally reduced to one eight-item cluster, one six-item cluster, and one four-item cluster. This was done by comparing the cluster solutions obtained by the various starting items within a subset. If an item appeared in at least half of the cluster solutions, it was considered a part of the final cluster. For example, to determine the final membership of the eight-item cluster, any item appearing in four or more of these solutions was considered a part of the final eight-item cluster.

Judging Cluster Recovery: The Rand Statistic

The Rand statistic (1971) was computed to compare the known subset structure in each data set with the BINCLUS cluster solution. If one records for each pair of items in a data set whether they are placed together or apart by a clustering solution, the Rand statistic is merely the number of correct placements, that is, placements agreeing with the known subset structure, divided by the total number of item pairs. In most cases the number of item pairs belonging apart is high, and thus the Rand statistic often tends to be large. In fact, a Rand statistic of .82 in this case can be obtained by arbitrarily placing each item in a cluster by itself.

Results

A previous Monte Carlo study (McCormick, Cliff, Cudeck, & Reynolds, Note 2) had shown BINCLUS to be quite robust across association indices. Gamma was chosen for use in the simulations because it is relatively insensitive to differences in item response frequencies.

The resulting Rand statistics, averaged over replications, are shown in Table 7.

(Insert Table 7 Here)

The most striking finding is the excellent overall performance of BINCLUS, as evidenced by an average Rand statistic of .99. This is similar to the findings of McCormick et al. (Note 2) who obtained an overall average Rand statistic of .969 using much smaller sample sizes. BINCLUS recovered the subscale structure perfectly in all the data sets of moderate (a = 1.0) item discrimination. When item discrimination was high (a = 3.0), recovery was perfect in the Bimodal, Mixed, and Rectangular conditions, above .99 in the Normal condition, and .98 in the Low Frequency condition. BINCLUS performs slightly less well

when items are of low discrimination (a = .5). While recovery is perfect in the Normal condition and virtually perfect in the Mixed condition, BINCLUS fares somewhat less well in the Low frequency, Bimodal, and Rectangular conditions.

For purposes of comparison, common factors analysis was performed on each of the 75 data sets in each analysis, three factors were rotated to a Varimax solution. An item was considered part of a factor if it had its highest loading on that factor, provided that the loading was greater in absolute value than .2. The Rand statistics from this analysis also appear in Table 7. Clearly BINCLUS was much more successful at recovering the underlying subscale structure than factor analysis. This was particularly true in the Bimodal condition irrespective of item discrimination, and in the low discrimination condition overall.

An unknown degree of artificiality is introduced into the criterion by specifying in advance the cutpoints for the clusters as well as the number of factors to be rotated. This device is an attempt to circumvent the introduction of subjective methods which might bias the results and is not unique to this investigation (Milligan, 1980). This shortcoming might be overcome if data could be sent blind to factor analysts and users of BINCLUS who could then make

independent judgments to determine the number of factors and the items belonging to each. The experience of the authors is that not knowing the true cluster structure is less of a problem for BINCLUS than factor analysis.

Modifications

These analyses of artificial data showed that in its initial form BINCLUS was quite successful when applied to clusters that were generated according to the Birnbaum model, this being true even when the consistency was not high. Application to several sets of empirical data, however, showed that some refinement was desirable.

With empirical data, the clusters of variables tend to be much less clearly defined. Rather than being isolated clusters of points, each surrounded by a "moat" of empty space, there is more often a shading off from one location to the next. Thus the need was for some additions to the clustering procedure that would identify more clearly the central members of more diffuse clusters.

BINCLUS has two features that are designed to have this effect; one has been touched on earlier. This is the matrix rearrangement process that reorders the rows and columns of the binary membership matrix. It will be described in more detail here. The other is a kind of second order clustering that will also be described here.

Rearranging Output Matrix

In the artificial data, true cluster members had consecutive identification numbers, and therefore the final membership matrix, if correct, listed cluster-members next to each other. In real data, clusters were not likely to be as visually apparent as the matrix still appeared jumbled.

To present a more visually useful form of the relationship among the clusters, the rows and columns of a cluster membership matrix are permuted so as to make similar clusters adjacent. A variety of approaches to this problem are possible. The present approach is a form of nearest neighbor ordering based on gammas. It is similar in concept to the seriation procedure of Gelfand (1971).

The same procedure is applied independently to rows (variables) and columns (clusters). Gammas are computed among all pairs of variables or clusters; the two having the largest value are placed next to each other. Call one the left member (L) and the other right (R) member. This is a two-member chain. Then, the one element of the remaining p - 2 rows or columns that is closest (highest gamma) to R is found, and similarly the one closest to L. Call these R' and L', respectively. Then their gammas are compared, and R' is placed to the right of R, or L' is placed to the left of L, depending on which gamma is higher. In either case,

this new member becomes one of the ends of the chain of three elements. Then, one of the p - 3 remaining members is added to one end or the other of the chain in the same way. The process continues, adding members to the ends of the chain until all the members have been ordered. Then the variables by clusters matrix is transposed and the process is repeated. The process has been quite effective in arranging the data into a visually compelling form, as was seen in Table 5 where this procedure placed together the clusters with identical members (columnwise) and the items with identical cluster memberships (rows). Table 8 shows the process applied to a membership matrix whose variables were in random order.

(Insert Table 8 Here)

Here the nature of the two cluster solution is clearly visible, and it is also apparent that variable 5 is a maverick. When the clusters are disjoint or nearly so, as they are in Table 3, the permuted pattern will assume a block diagonal 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 designed to enhance the appearance of the clusters, some solutions proved difficult to interpret. Sometimes the results display a rough block-diagonal pattern, but in many instances even these clusters can have "ragged" edges, i.e., there are many clusters whose memberships differ slightly, depending on which variable was the starter. These solutions 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 for a cluster.

For this reason, we developed a "second order" analysis. In a second-order analysis, the cluster membership matrix resulting from the initial clustering is treated as a data matrix and itself clustered. Essentially, clusters whose membership differ only slightly are put together. The first order 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 variables can be deleted since it is known that they are unrelated to the rest of the set. The second-order analysis then proceeds with the

number of elements reduced by the number of singletons in the first analysis.

The columns in the data-matrix for a second-order analysis are no longer the original variables, but rather are the clusters obtained from the first analysis, and the rows are variables. Often, the second-order analysis very clearly reveals which clusters are similar. Its mode of operation is largely to "trim the edges" of clusters and also to delete clusters that do not form consistently from several starting members.

In the first-order analysis, the binary membership matrix V is variables-by-clusters, and a 1 denotes a variable that was a member of the cluster started by the column variable. In a second-order analysis, the corresponding membership matrix of binary relations is cluster-by-"superclusters". That is, $v^*_{cc^*} = 1$ implies that cluster c is a member of the supercluster started by cluster c^*. The final step in superclustering is the deletion of columns of V^* that duplicate other columns.

The complete V^* matrix for the Table 8 solution, shown in Table 9(a), is the same as the permuted membership matrix, but with the singleton deleted. The reduced supercluster membership matrix corresponding to the second order analysis of the data shown in Table 3 is presented in Table 9(b).

(Insert Table 9 Here)

The reduced V^* matrix is shown in section b of Table 9. 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.

The final step is to relate the original variables to the superclusters. One approach to this is to construct a matrix P, with p_{jc^*} equal to the proportion of clusters in super-cluster c* of which variable j was a member. This matrix is displayed in section c of Table 9.

Applications

Social Deviance Data

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)).

Results

The results of a first-order BINCLUS analysis are given in Table 10, along with brief phrases identifying the items.

(Insert Table 10 Here)

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 right; this consists entirely of items related to cases where the father does not have a normal, healthy role in the

---

*The authors are indebted to Sarnoff A. Mednick and W. F. Gabrielli, Jr. for making these data available.

family. 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 permuted cluster membership matrix in Table 10 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 items. Sometimes the clusters overlap, and often there is a blurring of the distinction between cluster members and singletons.

Under these circumstances a second-order analysis may clarify the solution. Table 11 contains the second-order cluster membership matrix and the P matrix for these data.

(Insert Table 11 Here)

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 the "mother employed fulltime" cluster. There is almost no overlap between the superclusters except for the two that are nearly

identical. These results are typical of data where there seems to be a reasonable structure.

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.

Criminal Records

Another set of data looked at is a very large file of criminal offenses committed by a birth cohort of 28,000 young men born in Copenhagen. The data (Witkin, Mednick, Schulsinger, Christiansen, Goodenough, Hirshhorn, Lundsteen, Owen, Phillip, Rubin, & Stocking, 1976) are unique, both for completeness and accuracy. They allowed us to test the notion that criminality is a unitary phenomenon against the competing hypothesis that criminality can be subdivided into distinct criminal specialties (Collins, Cliff, Cudeck, McCormick, & Zatkin, 1983). The initial data was a binary citizens by offenses matrix, where 1 means that the citizen had been arrested for the column offense.

Results

Table 12 shows the first order cluster histories for a selected sample of the 56 most common or serious crimes. (See Collins, et al. for identification of the offenses.)

(Insert Table 12 Here)

Ten of the more familiar crimes have been underlined, and one can see they are near the top of most clusters.

If a rather low index cutoff point near .15 is chosen, the binary membership matrix which results is that displayed in Table 13.

(Insert Table 13 here)

In the second order analysis, shown in Table 14, the pattern is even clearer.

(Insert Table 14 Here)

The items most important in the larger cluster, those which were circled, include burglary, forgery, fraud, robbery and receiving stolen goods, which are apparently core crimes in a scale or factor of general criminality. The small cluster in the corner consists of traffic offenses: speeding, illegally overtaking, failure to yield and negligent homicide.

Discussion

To a certain extent, the utility of a method can be measured by the variety of applications for which it is appropriate, and there seems to be a variety 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. The machinery of factor analysis, generally appropriate for this kind of problem when the data consist of continuous variables, is unsatisfactory with dichotomous items (Carroll, 1945; Gourlay, 1951; Guilford, 1941; Lord & Novick, 1968, p. 349; Wherry & Gaylord, 1944). Other recently developed methods which have an explicit model for dichotomous responses (Christoffersson, 1975; Muthen, 1978; Muthen & Christoffersson, 1982) are limited practically in the number of variables they can treat, or require very large numbers of subjects for statistical estimation, and are based on normality assumptions that are highly questionable.

Traditional approaches for item analysis can be applied once fairly homogeneous subsets have been defined, and factor analysis can be used to examine further the structure of the composite variables. But most of the 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. The present version of nonhierarchical clustering seems promising in this context, as witness the successful applications 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 an a priori basis which can be arbitrary, unrealistic, or prone to bias. Lorr (1976), among others, has suggested cluster analysis for this purpose. The most frequent kind of cluster analysis used is a hierarchical method, but 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 Guttman-scale sense in the educational literature e.g. Bart (Note 3)--are eminently suited for analysis by Binclus, and indeed it was they that we have had in mind from the beginning of this development. A nonhierarchical 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 the existence of possible subsets is desired.

Another source of potential applications are exploratory investigations which study structural aspects among variables 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. The artificial data experiments 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 empirical 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, 1982), 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 such a guarantee is hardly possible short of trying all possible combinations of items of each cluster size, i.e., 2^p clusters. Since, e.g., 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 by using all possible 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 tend 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 gammas as the index of association 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.

REFERENCE NOTES

1. Cliff, N. & Reynolds, T. J. Dominance relations as a basis for nonparametric test theory. Unpublished manuscript.

2. McCormick, D. J., Cliff, N., Cudeck, R. A., & Reynolds, T. J. Clustering binary items. Technical report 81-1, Department of Psychology, University of Southern California.

3. Bart, W. The ordering analytic approach to hierarchical analysis. Paper presented at the AERA conference, 1981, L.A.

4. Zatkin, J. T., Cudeck, R. A., McCormick, D. J., & Cliff, N. A method for non-hierarchical cluster analysis based on binary relations and a comparison with other clustering programs. Technical Report 81-2, Department of Psychology, University of Southern California.

REFERENCES

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

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10, 1-19.

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

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

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

Cliff, N. What is and isn't measurement. In G. Keren, Ed., Statistical and methodological issues in psychology and social sciences research. Hillsdale, New Jersey: Erlbaum, 1982.

Collins, L., Cliff, N., Cudeck, R., McCormick, D., Zatkin, J. Patterns of crime in a birth cohort, Multivariate Behavioral Research, in press.

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.

Gourlay, N. Difficulty factors arising from the use of tetrachoric correlations in factor analysis. British Journal of Psychology, Statistical Section, 1951, 4, 65-76.

Guilford, J. P. The difficulty of a test and its factor composition. Psychometrika, 1941, 6, 67-77.

Guttman, L. Measurement as structural theory. Psychometrika, 1971, 36, 329-347.

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

Lord, F. & Novick, M. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

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.

Milligan, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 1980, 45, 325-342.

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

Muthen, B. & Christoffersson, A. Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 1981, 46, 407-419.

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

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

Rand, W. M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, (33), 846-850.

Wherry, R. & Gaylord, T. Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty and constant error factors. Psychometrika, 1944, 9, 237-244.

Witkin, H. A., Mednick, S. A., Schulsinger, F., Christiansen, K. O., Goodenough, D. R., Philip, J., Rubin, D., & Stocking, M. Criminality, aggression, and intelligence among XYY and XXY men. Science, 1976, 193, 547-555.

Table 1

An Example of Item Difficulty Affecting Correlations

	a	b	c	d	Proportion Passing	Factors I	Factors II
a	1.00	.25	.38	.19	.80	a .63	.15
b		1.00	.19	.38	.20	b .15	.63
c			1.00	.25	.80	c .63	.15
d				1.00	.20	d .15	.63

Table 2

Contingency Table for Two Items

		Item k: Positive	Item k: Negative	
Item j	Positive	w	x	n_j
	Negative	y	z	$n-n_j$
		n_k	$n-n_k$	n

Table 3

Quality Index q for Fictitious Data with Eight Variables

		Objects							
		1	2	3	4	5	6	7	8
		Numerators and Denominators							
		n_{jk}							
d_{jk}	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

Table 4

Membership Matrix and History Matrix

for Fictitious Data with Eight Variables

	Variables							
	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 5

Binary Cluster Membership Matrix

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

Table 6

Description of Latent Trait Model Parameters Manipulated in Monte Carlo Study

Condition	Subject Ability		Item Difficulty		Item Discrimination		
	Shape		Shape		a = .5	a = 1	1 = 3
Normal	Normal	$\mu = 0$ $\theta = 1$	Normal	μ θ	0 1	0 1	0 1
Low Frequency	Normal	$\mu = 0$ $\theta = 1$	Normal	μ θ	2 1	2 1	1.5 1
Rectangular	Rectangular	$\mu = 0$ $\theta = 1$	Rectangular	μ θ	0 2	0 1.3	0 1
Bimodal	Normal	$\mu = 0$ $\theta = 1$	Bimodal	μ θ	± 2* .25	± 2* .25	± 1.5* .25
Mixed	Normal	$\mu = 0$ $\theta = 1$	Rectangular	μ θ	0 2.0	0 1.3	0 1

*The means of the two modes were at these two points; the standard deviations are the within-mode valus.

Table 7

Rand Statistics for Simulation Study

(including factor analysis)

	BINCLUS			Factor Analysis		
	a = .5	a = 1	1 = 3	a = .5	a = 1	a = 3
Normal	1.000	1.000	.997	.996	1.000	1.000
Low Frequency	.952	1.000	.978	.931	.991	.967
Rectangular	.965	1.000	1.000	.929	.996	1.000
Bimodal	.964	1.000	1.000	.932	.939	.911
Mixed	.978	1.000	1.000	.957	.997	.968

Table 8

Binary Membership Matrix and Permuted Membership Matrix

Membership Matrix in Random Order

	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

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 9

Super-Cluster Solution for Fictitious Data with Eight Variables

		Superclusters						
		2	3	6	8	1	4	7
Clusters	2	1	1	1	1	0	0	0
	3	1	1	1	1	0	0	0
	6	1	1	1	1	0	0	0
	8	1	1	1	1	0	0	0
	1	0	0	0	0	1	1	1
	4	0	0	0	0	1	1	1
	7	0	0	0	0	1	1	1

a. Complete V* matrix

	Superclusters	
	1	2
2	1	0
3	1	0
6	1	0
Clusters 8	1	0
1	0	1
4	0	1
7	0	1

b. Reduced V* matrix

	Superclusters	
Items	1	2
2	1.0	0
3	1.0	0
6	1.0	0
8	1.0	0
1	0	1.0
4	0	1.0
7	0	1.0

c. P matrix

Table 10

Cluster Membership Matrix for Binclus Analysis of Deviance Data

```
ITEM
                 11111111111222222222233333333334444
      12345678901234567890123456789012345678901234567890123456789012
 20   00000000000000000000000001000000000000000000 F ALCOHOLIC*
  4   00000000011111111111111111000000000000000000 F DEAD
 31   11110000011111111111111111100000000000000000 M CHAR DISORDER
  5   11111111111111111111111111000000000000000000 BIOL PARENTS NOT MARRIED
  6   11111111111111111111111111000000000000000000 BIOL PARENTS NOT TOGETHER
  7   11111111111111111111111111000000000000000000 C HAS HAD >1 FAM CONSTELLATIOI
  8   11111111111111111111111111000000000000000000 C NOT LIVING WITH PARENTS
  9   11111111111111111111111111000000000000000000 F FIGURE NOT CHILD'S F
 10   11111111111111111111111111000000000000000000 C <7 YEARS WITH PRESENT FAM
 11   11111111111111111111111111000000000000000000 C <7 YEARS WITH 1 FAM
 14   11111111111111111111111111000000000000000000 C HAS LIVED IN ORPHANAGE
 15   11111111111111111111111111000000000000000000 C <7 YEARS WITH M
 16   11111111111111111111111111000000000000000000 C <7 YEARS WITH F
 19   11111111111111111111111111000000000000000000 M ALCOHOLIC
 21   11111111111111111111111111000000000000000000 <2 ADULTS IN HOME
 36   11111111111111111111111111000000000000000000 C NOT WITH OWN FAM
 37   11111111111111111111111111000000000000000000 C NOT WITH BOTH PARENTS
  2   00001111111111111111111111000000000000000000 C HAS M SUBSTITUTE
  3   11000000111100000000000000000000000000000000 M DEAD
 27   11110000000000000000000000000000000000000000 M PSYCHOTIC
 43   11110000000000000000000000000000000000000000 M HOSPITALIZED FOR PSYCH PROB
 44   11110000000000000000000000000000000000000000 M SERIOUS PHYSICAL ILLNESS
 17   11111111000000000000000000000000000000000000 C NOT ALWAYS WITH M FIRST YEAF
 18   11111111000000000000000000000000000000000000 C NOT ALWAYS WITH M SECOND YEA
  1   00000000000000000000000000000100000000000000 M WORKS
 30   00000000000000000000000000000111111111100000 F NEUROTIC
 28   00000000100000000000000000000111111111100000 F PSYCHOTIC
 24   00000000000000000000000000000111111111100000 F DOES NOT LIKE C
 26   00000000000000000000000000000111111111100000 F IMMATURE
 32   00000000000000000000000000000111111111100000 F CHAR DISORDER
 34   00000000000000000000000000000111111111100000 F ANXIOUS
 45   00000000000000000000000000000111111111100000 F HOSPITALIZED FOR PSYCH PROB
 46   00000000000000000000000000000111111111100000 F SERIOUS PHYSICAL ILLNESS
 40   00000000100000000000000000000000000000000100 PARENTS QUARREL
 39   00000000100000000000000000000000000000000000 PARENTS PHYSICALLY FIGHT
 12   00000000000000000000000000000000000000000011 C SPENT TIME IN WHOLE DAY CARE
 41   00000000000000000000000000000000000000000011 M FULLTIME WORK WHILE C <5
 22   00000000000000000000000000011100000000000000 INADEQUATE HOME ATMOSPHERE
 23   00000000000000000000000000011100000000000000 M DOES NOT LIKE C
 25   00000000000000000000000000011100000000000000 M IMMATURE
 33   00000000000000000000000001000000000000000000 M ANXIOUS
 29   00000000000000000000000011000000000000000000 M NEUROTIC
```

Table 10

(cont.)

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 11

Super-Cluster Membership Matrix and P Matrix for Deviance Data

SUPERCLUSTER MEMBERSHIP MATRIX			P MATRIX				
			SUPERCLUSTERS				
CLUSTER	SUPERCLUSTER	ITEM	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 12

First Order History for Crimes

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
21	3	21	21	8	48	21	21	21	30	21	21	23	21	21	29	50	3	29	21	3	21	26	23	21	48	21	21
25	21	8	8	21	29	8	3	8	21	8	8	21	8	25	21	51	21	21	8	8	8	21	21	26	6	26	29
29	8	29	29	29	21	29	29	29	3	29	29	26	29	29	8	2	29	8	29	29	29	7	26	29	29	25	40
26	29	25	25	3	26	3	25	25	29	25	25	8	25	26	25	52	27	25	3	25	3	8	7	40	21	29	48
8	25	26	26	25	8	25	26	26	8	26	26	29	26	48	26	56	25	26	25	26	25	29	29	27	8	8	26
48	48	48	48	48	3	48	48	48	48	48	48	48	48	8	48	49	48	48	48	48	48	48	8	48	3	48	20
3	26	6	3	26	25	26	6	6	26	3	3	3	3	3	3	21	26	3	26	6	26	3	48	8	25	3	25
23	27	27	27	27	27	6	27	3	25	27	27	25	27	6	27	29	8	27	27	27	23	25	3	3	27	23	27
6	23	23	23	23	23	23	23	27	6	23	23	6	23	23	23	48	6	23	23	23	6	6	25	23	23	6	8
27	6	7	6	6	7	27	7	23	20	6	6	27	6	27	6	8	20	6	6	7	7	27	6	6	7	7	6
18	7	18	7	7	18	18	18	7	23	7	7	7	7	20	7	3	23	7	7	18	27	18	27	7	18	18	23
20	18	20	18	18	20	20	20	18	27	18	18	18	18	18	18	25	7	18	18	20	18	20	18	18	20	20	3
47	47	47	47	47	47	47	47	47	7	47	47	47	47	47	47	26	47	47	47	47	47	47	47	47	47	47	47
45	20	24	20	20	24	24	24	20	47	20	24	20	45	45	20	6	24	20	24	24	20	24	20	20	24	24	18
28	45	45	45	45	45	45	45	45	18	45	45	45	20	1	45	47	45	45	45	45	45	45	45	45	45	45	45
7	1	1	1	1	1	1	1	28	45	1	20	1	1	28	1	27	1	1	1	1	28	1	1	1	1	1	1
30	28	28	28	28	28	28	28	1	1	28	28	28	28	7	28	20	28	28	28	28	1	28	28	28	28	28	7
40	30	30	30	30	30	30	30	30	28	30	1	30	30	30	30	23	30	30	30	30	30	30	30	30	30	30	30
33	33	33	33	33	33	33	33	33	33	33	30	33	40	33	33	7	33	33	33	33	33	33	33	33	33	33	33
22	40	40	40	40	40	40	40	40	40	40	33	40	33	40	40	18	40	40	40	40	40	40	40	22	40	40	22
14	22	22	22	22	22	22	22	22	14	22	40	14	22	14	22	45	22	22	22	22	14	22	22	14	22	22	14
49	14	14	14	14	14	14	14	14	22	14	22	22	49	22	14	28	14	14	14	14	49	14	14	49	14	14	49
19	49	49	49	49	49	49	49	49	49	49	49	49	19	49	49	40	49	49	49	49	19	49	49	19	49	49	19
32	19	19	19	19	19	19	19	19	19	19	14	19	32	19	19	1	19	32	19	19	32	19	19	32	19	19	32
9	32	32	32	32	32	32	32	32	32	32	19	32	9	32	32	30	32	9	32	32	9	32	32	9	32	32	9
24	9	9	9	9	9	9	9	24	9	9	32	9	24	9	9	33	9	24	9	9	24	9	9	24	9	9	24
35	24	35	24	24	35	35	35	35	24	24	9	24	35	24	24	9	35	35	35	35	35	35	35	35	35	35	35
16	35	16	35	35	16	16	16	16	35	35	35	35	16	35	35	22	16	16	16	16	16	16	16	16	16	16	16
42	16	42	16	16	42	42	42	42	16	15	16	16	42	42	42	19	42	42	42	42	42	42	42	42	42	42	42
15	15	15	15	15	15	15	15	15	42	42	15	15	15	16	15	14	15	15	15	15	15	15	15	15	15	15	15
11	42	11	42	42	11	11	11	11	15	16	42	42	11	11	11	32	11	11	11	11	11	11	11	11	11	11	11
50	44	50	44	44	50	50	50	50	11	50	11	44	50	50	50	24	50	50	50	50	50	50	50	50	50	50	50
44	50	44	50	50	44	44	44	44	50	44	50	50	44	44	44	35	44	44	44	44	44	44	44	44	44	44	44
38	38	38	38	38	38	38	38	38	44	38	44	38	38	38	38	53	38	38	38	38	38	38	38	38	38	38	38
53	11	53	11	11	53	53	53	53	38	53	38	11	53	53	53	15	53	53	53	53	53	53	53	53	53	53	53
56	53	56	53	53	56	56	56	56	53	56	53	53	56	56	56	16	56	56	56	56	56	56	56	56	56	56	56
36	56	36	56	56	36	36	36	36	56	36	56	56	36	36	36	42	36	36	36	36	36	36	36	36	36	36	36
31	36	31	36	36	31	31	31	31	36	31	36	36	31	31	31	38	31	31	31	31	31	31	31	31	31	31	31
52	31	52	52	52	52	52	52	52	31	52	31	31	52	52	52	44	52	52	52	52	52	52	52	52	52	52	52
43	52	43	31	43	43	43	43	43	52	43	52	52	43	43	43	11	43	43	43	43	43	43	43	43	43	43	43
51	43	51	51	31	51	51	51	51	43	51	51	43	51	51	51	43	51	51	51	51	51	51	51	51	51	51	51
12	51	12	43	51	12	12	12	12	51	12	43	51	12	12	12	31	12	12	12	12	12	12	12	12	12	12	12
41	12	41	12	12	41	41	41	41	12	41	41	12	41	41	41	12	41	41	41	41	41	41	41	41	41	41	41
5	41	5	41	41	5	5	5	5	41	5	5	41	5	5	5	41	5	5	5	5	5	5	5	5	5	5	5
46	5	46	5	46	46	46	46	46	5	46	46	5	46	46	46	36	46	46	46	46	46	46	46	46	46	46	46
55	46	55	46	55	55	55	55	55	46	55	55	46	55	55	55	5	55	55	55	55	55	55	55	55	55	55	55
2	55	2	55	2	2	2	2	2	55	2	2	55	2	2	2	55	2	2	2	2	2	2	2	2	2	2	2
4	4	4	2	4	4	4	4	4	2	4	4	2	4	4	4	46	4	4	4	4	4	4	4	4	4	4	4
54	54	54	54	54	54	54	54	54	4	54	54	4	54	54	54	4	54	54	54	54	54	54	54	54	54	54	54
37	37	37	37	37	37	37	37	37	54	37	37	54	37	37	37	54	37	37	37	37	37	37	37	37	37	37	37
13	13	13	13	13	13	13	13	13	37	13	13	37	13	13	13	37	13	13	13	13	13	13	13	13	13	13	13
39	39	39	39	39	39	39	39	39	13	39	39	39	39	39	39	13	39	39	39	39	39	39	39	39	39	39	39
17	17	17	17	17	17	17	17	17	39	17	17	17	17	17	17	39	17	17	17	17	17	17	17	17	17	17	17
34	34	34	34	34	34	34	34	34	17	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
10	10	10	10	10	10	10	10	10	34	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Table 12 (cont.)

29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
40	10	3	26	21	21	21	21	40	21	21	21	21	21	8	45	19	47	56	26	47	17	39	56	21	50	2	52
21	21	2	21	26	8	29	8	21	25	8	8	8	8	21	21	21	56	46	6	6	51	50	17	29	17	3	17
8	3	21	29	29	29	40	29	25	29	3	29	29	29	29	3	3	49	49	29	46	2	17	51	40	51	21	51
25	29	29	25	25	3	25	25	29	40	29	25	25	25	25	25	25	21	21	21	21	52	52	50	25	52	25	50
26	8	25	48	48	25	48	26	48	8	25	26	26	26	26	29	29	29	29	8	29	56	56	49	48	56	29	49
48	48	48	6	6	48	26	48	26	48	48	48	48	48	48	27	27	48	48	3	48	49	49	21	26	49	48	21
3	26	8	8	8	26	8	3	8	26	26	3	3	3	3	48	48	6	6	25	8	21	21	29	8	21	8	29
27	25	26	3	3	27	3	27	3	3	27	27	27	27	27	26	26	8	8	27	26	29	29	48	3	29	26	48
23	6	23	27	23	23	23	23	23	23	23	23	23	23	23	8	8	25	25	23	25	48	48	8	23	48	23	8
6	20	6	23	27	6	6	6	6	6	6	6	6	6	6	18	18	26	26	7	3	8	8	26	6	8	6	26
7	23	7	7	7	7	27	7	27	27	7	7	7	7	7	20	20	27	27	18	27	3	6	6	27	26	7	6
18	27	27	47	47	18	47	18	47	18	18	18	18	18	18	23	23	23	23	20	23	25	25	25	47	25	27	25
47	7	47	18	18	47	7	47	18	47	47	47	47	47	47	6	6	40	40	47	40	26	26	3	18	6	18	3
20	47	18	20	20	20	18	24	20	20	20	20	20	20	20	47	1	7	7	24	7	6	3	47	20	47	47	47
45	18	45	45	45	45	45	45	45	45	45	45	45	45	45	1	47	18	18	45	18	47	47	20	45	27	20	20
1	45	20	1	1	1	20	20	1	1	1	1	1	1	1	28	28	3	3	1	45	27	27	27	1	20	45	27
28	1	28	28	28	28	1	28	28	28	28	28	28	28	28	7	7	45	45	28	20	20	20	23	28	23	28	23
30	28	1	30	30	30	28	1	7	30	30	30	30	30	30	30	30	20	20	30	28	23	23	7	7	3	1	7
33	33	30	40	40	33	30	30	30	7	33	33	33	33	33	33	33	28	28	33	1	7	7	18	30	7	30	18
22	40	40	33	22	40	33	33	33	33	40	22	40	40	40	40	40	1	1	40	30	18	18	45	33	18	33	45
14	14	33	14	14	22	22	40	14	22	22	14	22	22	22	22	22	30	30	22	33	45	45	40	14	45	40	40
49	22	14	22	49	14	14	22	22	14	14	49	14	14	14	14	14	33	33	14	22	28	28	28	22	28	14	28
19	49	22	49	19	49	32	14	49	49	49	19	49	49	49	49	49	22	22	49	14	40	40	1	49	1	22	1
32	19	49	19	32	19	9	49	19	19	19	32	19	19	19	19	32	14	14	19	19	1	1	30	19	30	49	30
9	32	19	9	9	32	49	19	32	32	32	9	32	32	32	32	9	19	19	32	32	30	30	33	32	40	19	33
24	9	32	24	24	9	19	32	9	9	9	24	9	9	9	9	24	32	32	9	9	33	33	9	9	33	32	9
35	24	9	35	35	24	24	9	24	24	24	35	24	24	24	24	35	9	9	35	24	9	9	22	24	9	9	22
16	35	24	16	16	35	16	35	35	35	35	16	35	35	35	35	16	24	24	16	35	22	22	19	35	22	24	19
42	16	35	42	42	16	42	16	16	16	16	42	16	16	16	42	42	35	35	42	16	19	19	14	16	19	35	14
15	42	16	15	15	15	15	15	42	15	15	15	15	15	15	15	15	15	15	15	15	14	14	32	42	14	16	32
11	15	15	11	11	42	11	42	15	42	42	11	42	11	42	16	11	42	42	11	42	32	32	24	15	32	15	24
50	11	42	50	50	44	50	50	50	11	44	50	44	50	44	50	50	16	16	50	50	24	24	35	50	24	42	35
44	50	50	44	44	50	44	44	44	50	50	44	50	44	50	38	44	50	50	44	44	35	35	53	56	35	50	53
38	44	44	38	38	38	38	38	38	44	38	38	38	38	38	11	38	44	44	38	38	53	53	16	44	53	44	16
53	38	38	53	53	11	53	11	11	53	11	53	11	53	11	53	53	38	38	53	56	15	15	15	38	15	38	15
56	53	11	56	56	53	56	53	53	56	53	56	53	56	53	56	56	53	53	56	53	16	16	42	11	16	56	42
36	56	53	36	36	56	36	56	56	36	56	36	56	36	56	36	36	11	11	36	11	42	42	38	36	42	11	38
31	36	56	31	31	36	31	31	36	31	36	31	36	31	36	31	31	36	36	31	36	38	38	11	31	38	53	11
52	31	36	52	52	31	52	52	52	52	31	52	31	52	52	52	52	52	52	52	52	44	44	44	52	44	36	44
43	52	52	43	43	52	43	43	43	43	52	43	52	43	31	43	43	43	43	43	43	11	11	43	43	11	52	43
51	43	43	51	51	51	51	51	51	51	43	51	51	51	51	51	51	31	31	51	31	43	43	31	51	43	31	31
12	51	51	12	12	43	12	12	31	12	51	12	43	12	12	12	12	51	51	12	51	31	31	12	12	31	51	12
41	12	12	41	41	12	41	41	12	41	12	41	12	41	41	41	41	12	12	41	12	12	12	41	41	12	43	41
5	41	41	5	5	41	5	5	41	5	41	5	5	5	5	5	5	41	41	5	41	41	41	36	5	36	12	36
46	5	5	46	46	5	46	46	5	46	5	46	46	46	46	46	46	5	5	46	5	36	36	5	46	41	41	5
55	46	46	55	55	46	55	55	46	55	46	55	55	55	55	55	55	55	55	55	55	5	5	55	55	5	5	55
2	55	55	2	2	55	2	2	55	2	55	2	2	2	2	2	2	2	2	2	2	55	55	2	2	55	46	2
4	2	4	4	4	2	4	4	2	4	2	4	4	4	4	4	4	4	4	4	4	46	46	46	4	46	4	46
54	4	54	54	54	4	54	54	4	54	4	54	54	54	54	54	54	54	54	54	54	4	2	4	54	2	54	4
37	54	37	37	37	54	37	37	54	37	54	37	37	37	37	37	37	37	37	37	37	54	4	54	37	4	37	54
13	37	13	13	13	37	13	13	13	13	37	13	13	13	13	13	13	13	13	13	13	37	54	37	13	37	13	37
39	13	39	39	39	13	39	39	39	39	13	39	39	39	39	39	39	39	39	39	39	13	37	13	39	13	39	13
17	39	17	17	17	39	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	39	13	39	17	39	17	39
34	17	34	34	34	17	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
10	34	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Table 13

First Order Membership

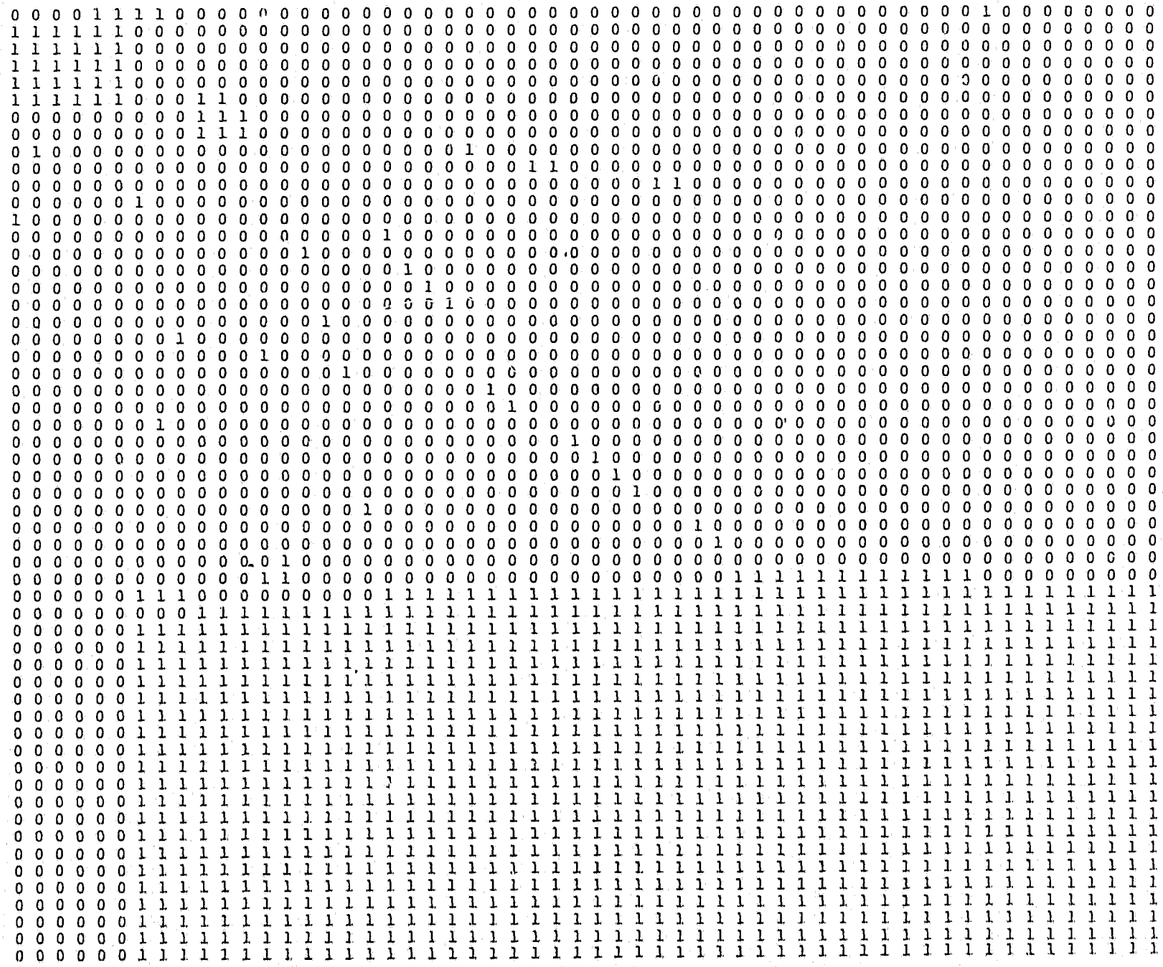
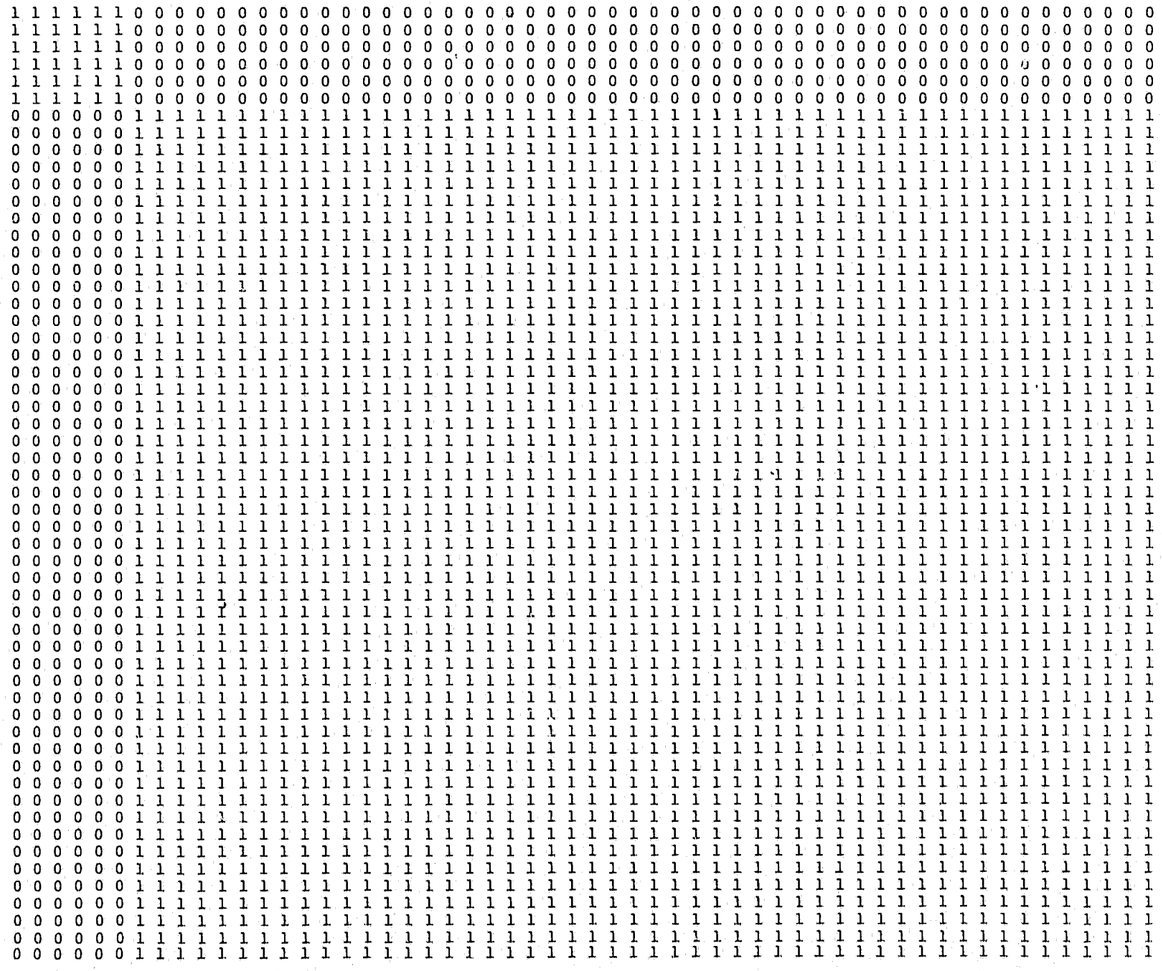


Table 14

Second Order Membership



END