PSYCHOMETRIKA—VOL. 49, NO. 1, 57–78
MARCH 1984

SYNTHESIZED CLUSTERING: A METHOD FOR AMALGAMATING ALTERNATIVE CLUSTERING BASES WITH DIFFERENTIAL WEIGHTING OF VARIABLES

WAYNE S. DESARBO
J. DOUGLAS CARROLL
LINDA A. CLARK

BELL LABORATORIES

PAUL E. GREEN

UNIVERSITY OF PENNSYLVANIA

In the application of clustering methods to real world data sets, two problems frequently arise: (a) how can the various contributory variables in a specific battery be weighted so as to enhance some cluster structure that may be present, and (b) how can various alternative batteries be combined to produce a single clustering that "best" incorporates each contributory set. A new method is proposed (*SYNCLUS*, *SYN*thesized *CLUS*tering) for dealing with these two problems.

Key words: Cluster Analysis, Variable Importance.

1. Introduction

In the application of clustering techniques to large scale empirical problems, the researcher often encounters two difficulties. First, in any given battery of variables, it appears to be the case that only a proper subset of the variables contribute in an important way to the resultant clustering. Indeed, the presence of additional variables on which the clusters are not distinguished may obscure the cluster structure. For example, in automobile marketing research applications, one often finds that certain attitudinal variables, such as those emphasizing styling and comfort versus design simplicity and high gas mileage, produce clusters of car owners with markedly different patterns of brand ownership. However, if extraneous attitudinal variables (describing general leisure-time interests or feature preferences) are added, the original structure may be completely obscured. Fowlkes (Note 2) has demonstrated this "masking" effect empirically with hierarchical clustering methods.

Second, in many applied problems—again those often arising in public opinion and marketing research—one often faces an embarrassment of riches: the availability of several competing batteries of variables on which to conduct the cluster analysis. While one could simply cluster on the basis of the full set of variables, the implicit weighting of the separate batteries is rarely in accord with the researcher's judgments regarding their relative importance. The judgments for this á priori weighting are often legitimately based on expert knowledge concerning the nature of the objects to be clustered.

This paper is concerned with both of these issues. A new clustering method, called SYNCLUS (*SYN*thesized *CLUS*tering) is proposed for dealing with these two problems.

We wish to thank Anne Freeny and Deborah Art for their computer assistance, and Ed Fowlkes for his helpful technical discussion. We would also like to acknowledge the insightful and helpful comments from the editor and reviewers.

Requests for reprints should be sent to Wayne DeSarbo, Bell Laboratories, Room 2C-256, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

0033-3123/84/0300-5049$00.75/0

We first discuss aspects of the literature bearing on the question of variable weighting so as to "optimize" the clustering qualities of a set of data. This is followed by a description of the model and algorithm underlying the SYNCLUS procedure. The algorithm is then applied to a synthetic data set with known cluster structure. An application concerning physician's attitudes and media preferences is discussed. We conclude the paper by discussing some potential applications of the method and further research needed to examine its performance under diverse sets of data conditions.

II. Literature Review

Much of the literature that is applicable to the problem of variables importance in clustering is obtained from classical multivariate statistics for given or known classifications. That is, once a clustering is obtained in K clusters or groups, one can employ such techniques as separate t tests, Hotelling's T^2, MANOVA, multiple discriminant analysis, stepwise discriminant analysis, etc. (Obviously, these techniques are inappropriate if the same data are used to develop the grouping.)

H. Friedman and Rubin (1967), in their discussion of various criteria for grouping data, employ such procedures for investigating relevant subsets of variables *after* their algorithm produces the desired clustering. They also discuss several graphical methods to aid in variable selection.

Kruskal (1972) attempts to find a linear transformation of the data which will reveal a "hidden cluster structure". He solves for the coefficients of this linear transformation by optimizing "an index of condensation", a measure of cluster compactness. Presumably, the coefficients of the optimal linear transformation provide insight into which variables should receive higher weights (or contribute more to the clustering) and which variables should receive lower weights. However, Kruskal (1972) admits to having difficulty in selecting an appropriate condensation index to optimize. Friedman and Tukey's (1974) "projection pursuit" approach also attempts to attain a similar aim, but in "real time" interaction with the user. It is also restricted to one or two dimensional projections of a higher dimensional space rather than a general linear transformation.

Sneath and Sokal (1973) discuss a number of related clustering techniques concerning weighted and adaptive clustering. One type of weighting in clustering they mention considers some dimensions of the clusters more important than others. Rohlf (1970) defines a generalized distance function which weights distances along axes of a hyperellipsoid inversely to the eigenvalues corresponding to each of these axes, and in this way evaluates distance along the principal axis of an ellipsoid as equivalent to the much smaller distances along the minor axes. Morrison (1967) develops a "Mahalanobis-like" distance measure which normalizes and compensates for different variances, intercorrelations, and variable importances.

More recently, Fowlkes, Gnanadesikan, and Kettenring (Note 3) have specifically addressed the variable importance problem in clustering. They are currently experimenting with three types of "stepwise" algorithms for variable selection and evaluating subsets selected in terms of three measures of "cluster strength". They are (a) forward selection (starting with a single variable and entering one variable at a time), (b) backward elimination (starting with all variables and eliminating one at a time), and (c) guided selection (uses the data to derive a subset or starting set of the original set of variables; then, variables are entered and deleted again to produce clusterings for a collection of subsets of variables). These are analogous to stepwise approaches to subset selection models in linear regression. In the present case, letting $\mathbf{B}$ and $\mathbf{W}$ be the usual sums of squares and cross-product matrices between clusters and pooled within clusters respec-

tively, three measures are examined:

1. $-\log \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$
2. Trace $(\mathbf{BW}^{-1})$
3. Maximum eigenvalue of $\mathbf{BW}^{-1}$.

This procedure is still in the experimental stages of development. It differs from the SYNCLUS methodology in that while SYNCLUS will utilize a differential weighting of variables (separately within each of one or more distinct sets of variables), the Fowlkes, Gnanadesikan, and Kettenring (Note 3) approach seeks a *selection* of a subset of variables (which can be viewed as defining the SYNCLUS differential weights to be zero or one within a single variable set).

Finally, Art, Gnanadesikan, and Kettenring (1982) have proposed a method for bootstrapping a metric from multivariate data that are to be clustered. The method exploits nearest neighbors for developing a metric that reflects possible differences in the scales of the initial variables as well as correlations among them which may indirectly provide some insights as to variable importance.

III. SYNCLUS

A. *Objectives*

The research objectives underlying the development of SYNCLUS are:

1. To provide an algorithm for K-means clustering (MacQueen, 1967) that can be directly applied to distances between objects, as well as to profile data (variables or characteristics of these objects) later converted to distances; and which can also be generalized to the case of three-way data (e.g., objects × objects × battery, or objects × variables × battery).

2. To provide a technique which, in addition to solving for a clustering of objects into K specified clusters (step 1 above), also renders numerical weights for the variables describing the objects—the weights indicating the variables' relative importance to the clustering; and,

3. To allow for the analysis of several different groups of variables (e.g., demographics, psychographics, product usage, etc.) where a priori known or believed group-level importance weights may be specified and processed in steps 1 and 2 above.

B. *The Model*

Let:

w_i^2 = the specified importance weight for the i-th battery or group of variables, normalized so that $\sum_{i=1}^{I} w_i^2 = 1; i = 1, 2, \ldots, I$.

$$e_{jk} = \begin{cases} 1 & \text{if object } j \text{ belongs to cluster } k \\ 0 & \text{otherwise;} \quad j, j' = 1, 2, \ldots, J; \; k = 1, 2, \ldots, K. \end{cases}$$

$y_{jt_i}^{(i)}$ = the t_i-th variable in the i-th battery describing object j; $t_i = 1, 2, \ldots, T_i$.

$v_{it_i}^2$ = the importance weight or square of the rescaling constant for the t_i-th variable in the i-th battery used in producing the clustering;

$d_{jj'}^{2(i)} = \sum_{t_i=1}^{T_i} v_{it_i}^2 (y_{jt_i}^{(i)} - y_{j't_i}^{(i)})^2$ = the weighted squared distance between objects j and j' de-

fined for the i-th battery;

$\delta_{jj'} = \alpha a^*_{jj'} + \beta$ (discussed in the Appendix);

where:

$$a^*_{jj'} = \begin{cases} \frac{1}{J_k} & \text{if objects } j \text{ and } j' \text{ are jointly members of cluster } k, \\ & \text{where } J_K = \text{the number of objects in cluster } k; \\ 0 & \text{if objects } j \text{ and } j' \text{ are in different clusters.} \end{cases}$$

(Note, α will generally be negative, and β positive and sufficiently large so that $\delta_{jj'} \geq 0$ for all j and j'; however, while there are no explicit constraints imposed to insure this, in practice $\alpha < 0$, $\beta > 0$, and $\delta_{jj'} \geq 0$). Then, we wish to solve for α, β, e_{jk}, and for $v^2_{it_i}$, given w_i^2, K, and $y^{(i)}_{jt_i}$, in order to minimize the following sum of squares:

$$Z^{*2} = \sum_{i=1}^{I} \sum_{j<j'}^{J} \sum^{J} w_i^2 (\delta_{jj'} - d^{2(i)}_{jj'})^2. \tag{1}$$

However, since both $\delta_{jj'}$ and $d^{2(i)}_{jj'}$ are to be estimated from the algorithm to follow, Z^{*2} in equation (1) can be trivially minimized by driving the $(\delta_{jj'} - d^{2(i)}_{jj'})^2$ term to zero by using smaller and smaller numbers for each of the entities. Because of this problem, we use an appropriate normalization factor producing a weighted mean-square, stress-like measure. One of these measures is:

$$Z_1^2 = \frac{\sum_{i=1}^{I} w_i^2 \sum_{j=1}^{J} \sum_{j'=1}^{J} (\delta_{jj'} - d^{2(i)}_{jj'})^2}{\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta^2_{jj'}}. \tag{2}$$

(It should be noted that the sum in both numerator and denominator of the expression on the right hand side of equation (2) is over all j and j'. Also, δ_{jj} is not defined to be zero, but is the same as $\delta_{jj'}$ for $j \neq j'$, where j' is in the same cluster as j. The reason for this, and for the particular definition of $\delta_{jj'}$, is discussed in the Appendix. As will be proved subsequently, $\delta_{jj'}$ as defined above, provides a measure of inter- and intra-cluster distance whose fit to the $d^2_{jj'}$ is optimized by the K-means approach to clustering.) Kruskal (1964a, b) discusses how such a stress-like measure avoids such degeneracies. Thus, we wish to find both the optimal clustering and appropriate scaling of the variables in minimizing Z_1^2 while *simultaneously* taking into account any a priori information on battery importance (w_i^2). Note, one can rewrite (2) as:

$$Z_1^2 = \sum_{i=1}^{I} w_i^2 Z^2_{1i}, \tag{3}$$

where,

$$Z^2_{1i} = \frac{\sum_{j=1}^{J} \sum_{j'=1}^{J} (\delta_{jj'} - d^{2(i)}_{jj'})^2}{\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta^2_{jj'}}. \tag{4}$$

Another possible definition of a weighted mean square stress would be of the form:

$$Z_2^2 = \sum_{i=1}^{I} w_i^2 Z^2_{2i}, \tag{5}$$

where:

$$Z_{2i}^2 = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J}(\delta_{jj'} - d_{jj'}^{2(i)})^2}{\sum_{j=1}^{J}\sum_{j'=1}^{J} d_{jj'}^{4(i)}}. \tag{6}$$

As can be seen by comparing equations (5) and (6) to (3) and (4), Z_2^2 differs from Z_1^2 only in the normalization factor in the squared stress term (Z_{1i}^2 and Z_{2i}^2, respectively) defined for the ith variable set. Based largely on results explicated in a paper by Kruskal and Carroll (1969), it can be shown that optimizing (minimizing) these two apparently different measures in fact leads to equivalent procedures, and that, in fact, the optimal solutions for the two measures are identical except for scaling (i.e., for the two different measures, the δ's and d^2's are simply rescaled by appropriate scaling constants). This has been demonstrated empirically by computing Z_1^2 and Z_2^2 and noting that both measures decrease monotonically and appear to approach a minimum simultaneously.

The equivalent problems of minimizing Z_1^2 or Z_2^2 are both equivalent to *maximizing* C^2, defined as:

$$C^2 = \sum_{i=1}^{I} w_i^2 C_i^2, \tag{7}$$

with C_i^2 defined as:

$$C_i^2 = \frac{\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}d_{jj'}^{2(i)}\right]^2}{\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}^2\sum_{j=1}^{J}\sum_{j'=1}^{J}d_{jj'}^{4(i)}}. \tag{8}$$

As can easily be seen, C_i^2 is simply the squared *uncentered* correlation between the δ's and the $d^{2(i)}$'s, and as such can be interpreted as "sums-of-squares accounted for" (by δ in d^2 or by d^2 in δ; in either case via a homogeneous linear regression, i.e., a linear regression without a constant term). C^2 is thus a weighted mean sum-of-squares accounted for. It can also easily be shown that, for the optimal values of the three indices (implying, for Z_1^2 and Z_2^2, the appropriate scaling for those indices) that

$$C_o^2 = 1 - Z_{1o}^2 = 1 - Z_{2o}^2 \tag{9}$$

(where the subscript "o" indicates the optimal values, corresponding to the optimal solutions-equivalent except for scaling). The analogous equations to (9) also obtain for each C_i^2, Z_{1i}^2 and Z_{2i}^2.

Since C^2, and each of its component terms C_i^2, is unaffected by the scaling of δ's and d^2's, we shall in fact use it henceforth as the preferred measure of goodness-of-fit. Since, however, optimizing the three measures is equivalent (except for scaling), we shall, in our description of the algorithm in the Appendix, shift, for reasons of mathematical convenience, from a discussion optimizing one to a discussion based on another.

C. *The Algorithm*

Appendix I presents a detailed description of the entire iterative SYNCLUS algorithm. The SYNCLUS algorithm is composed of seven phases. Not including the clustering, the number of parameters estimated ($v_{it_i}^2$) is equal to the total number of variables $T = \sum_i T_i$. Each phase is described in turn in the Appendix.

TABLE 1

Data Set for Monte Carlo Analysis

		Variable			
		X_1	X_2	X_3	X_4
Observation	1.	1.038	1.107	1.107	−0.896
	2.	−0.171	0.996	0.996	−0.112
	3.	0.635	0.885	0.885	1.568
	4.	−1.014	−0.885	1.107	1.680
	5.	−1.893	−0.996	0.996	0.299
	6.	−0.537	−1.107	0.885	−1.232
	7.	0.965	0.885	−1.107	1.269
	8.	−0.647	0.996	−0.996	−0.746
	9.	0.928	1.107	−0.885	0.112
	10.	0.672	−1.107	−1.107	−0.037
	11.	−1.197	−0.996	−0.996	−0.560
	12.	1.221	−0.885	−0.885	−1.344

IV. *Monte Carlo Results*

SYNCLUS has been run on numerous synthetic data structures where a known clustering existed. Random error was introduced and the SYNCLUS analysis was compared to results for ordinary K-means and average-link hierarchical clustering. A discussion of one set of data follows.

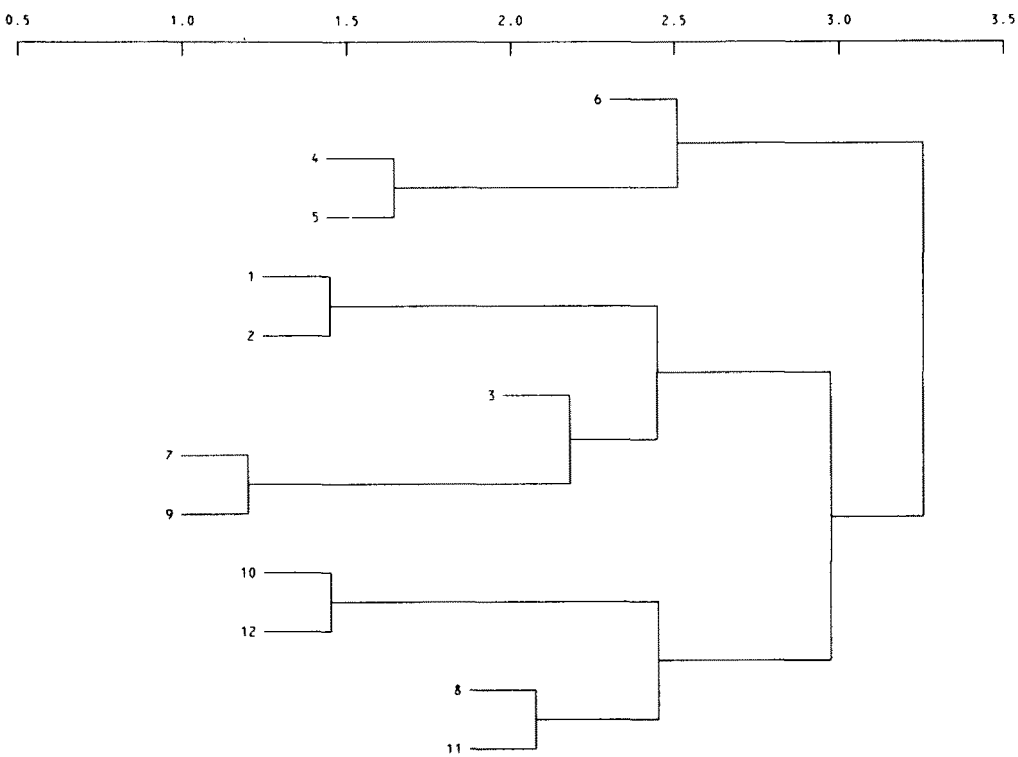

FIGURE 1.
Hierarchical Cluster Analysis Performed on X_1, X_2, X_3, and X_4 in Monte Carlo Analysis.

Table 1 lists the data used here to produce four clusters. Variables, X_2 and X_3 were used to generate four clusters (objects 1, 2, 3; objects 4, 5, 6; objects 7, 8, 9; objects 10, 11, 12), while X_1 and X_4 were random noise.

Figure 1 depicts the results of the hierarchical clustering analysis on all four variables, showing distortion in recovering the true clustering structure. Note that without X_1 and X_4, the hierarchical clustering algorithm works well in recovering the four clusters.

Table 2 presents the results of ordinary K-means for all four variables. Here too, the program has difficulty in recovering the cluster structure in Table 1. However, with X_1 and X_4 in the analysis, the K-means program did recover the four clusters.

Table 2 also presents the SYNCLUS results ($\mathbf{w}^2 = (.5, .5)$) for all four variables: X_1 and X_2 in one set and X_3 and X_4 in the other. SYNCLUS produces larger $v_{it_i}^2$ weights for X_2 and X_3 and smaller $v_{it_i}^2$ weights for X_1 and X_4. Also, the true cluster structure is recovered exactly.

Numerous other data sets involving different cluster shapes, number of clusters, number of variables, correlations between variables, number of batteries of variables, etc. were also run with similar results. In one case, SYNCLUS did not recover a known cluster structure when random noise was added to a set of variables that were responsible for the resulting clustering, although an appropriate weighting of the variables was obtained. Assuming that SYNCLUS had arrived at a local optimum solution, another run of

TABLE 2

K-means and SYNCLUS Results on X_1, X_2, X_3 and X_4 for Monte Carlo Analysis

K-means Results

		Four Cluster Solution	
No. of Clusters	Within-Groups Sum of Squares	Cluster #	Objects
2	33.11	1	1,2,8
3	22.57	2	3,7,9
4	17.17	3	4,5
5	12.39	4	6,10,11,12
6	9.26		

SYNCLUS Results

		Four Cluster Solution		
No. of Clusters	C^2	Cluster #	Objects	$v_{it_i}^2$
2	.621	1	1,2,3	X_1: 0.43
3	.644	2	4,5,6	X_2: 0.90
4	.651	3	7,8,9	X_3: 0.92
5	.645	4	10,11,12	X_4: 0.39

SYNCLUS was made in which the $E = \| ejk \|$ matrix was fixed on the first iteration of SYNCLUS, as generated by the known clustering. To our surprise, SYNCLUS moved from this a priori clustering and arrived at the same previous clustering, with a *higher* goodness of fit statistic! In this case, it appeared that the random noise included in the analysis might have legitimately provided a different clustering.

Clearly, more work in this area must be done. One of the obvious problems in such Monte Carlo work is how to limit the number of relevant factors to vary, experimentally, in order to investigate their impact on the effectiveness of SYNCLUS vs other competing techniques. For example, there are an infinite variety of cluster shapes one could investigate in many dimensions.

V. *Application*

A. *Study Description*

On a pilot basis, SYNCLUS was also applied to a set of real data, obtained from a recent marketing research study of South American physicians' attitudes toward a new antihypertensive drug. An initial sample of 160 respondents was obtained; all interviews were personally administered.

Among the data collected for each physician was a set of 13 judgments (X_1 to X_{13}) regarding the relative effectiveness of alternative promotional media, a set of 20 "life style" statements (X_{14} to X_{33}), and 7 demographic responses (X_{34} to X_{40}) for a total of 40 variables.

The 160 × 40 matrix was then partitioned, by columns, into three (media judgments, life style, and demographics) submatrices. For illustrative purposes, a random sample of $n = 40$ from the 160 subjects was selected for subsequent analysis. Each variable was standardized to have zero mean and unit variance prior to the various cluster analyses that follow. The object of the analysis was to examine the bases of possible market segmentation schemes.

B. *K-means Analysis*

A *K*-means analysis for all 40 variables was performed for two through seven clusters. Table 3 presents the total within cluster sum-of-squares for these clusterings. Note the relatively constant reduction in the error sums of squares for successive clusters indicating that two clusters may be sufficient in describing the structure of the data. In fact,

TABLE 3

K-means Statistics for Drug Study

		Two Cluster Solution	
Number of Clusters	Within-Groups Sum of Squares	Cluster #	Objects
2	1457.93	1	1,2,4,9,10,11,13,15,18,23,34,38
3	1366.07		
4	1264.39	2	3,5,6,7,8,12,14,16,17,19,20,21,22,24,25,26,
5	1157.54		27,28,29,30,31,32,33,35,36,37,39,40
6	1108.52		
7	1051.13		

using Hartigan's (1975, 1978) approximate F-test for K-means, one would stop at two clusters.

Table 3 also presents the two cluster solution. We calculated means and standard deviations (not shown) for the variables in the two cluster K-mean solution. By examining large mean differences with fairly low standard deviations (one could also perform a two-group discriminant analysis), one notes that variables X_4, X_6, X_7, X_{10}, X_{11}, X_{20}, X_{22}, X_{28}, and X_{32} possess large mean differences between the two clusters. Thus, Cluster 1 can be tentatively described as physicians agreeing with the following:

—Commercial scientific exhibits or displays are not effective (X_4);
—Direct mail is not effective (X_6);
—Educational materials for medical schools are not effective (X_7);
—Symposia are not effective (X_{10});
—Visual/audio cassettes/films are not effective (X_{11});
—Spends enough time with his family (X_{10});
—Does not only follow tried and tested medical practices (X_{22});
—Is not reluctant to experiment with new drugs (X_{28});
—Rarely considers the cost of drugs he prescribes (X_{32}).

Cluster 2 can be described with just the converse descriptions. Note that large mean differences were not present between the seven demographic variables. However, it is really unclear what variables account for cluster differences. It is also unclear what interpretation to give to these two clusters.

C. *SYNCLUS Results*

This same data set was then run through SYNCLUS ($\mathbf{w}^2 = (.33, .33, .33)$) where two through five clusters were obtained. Based upon the goodness of fit measures described earlier, two clusters also appeared to best represent the structure in the data. Table 4 presents the details of the two cluster SYNCLUS solution. Note, a different clustering is obtained as compared to the previous K-means solution. From a cursory inspection of the $v_{it_i}^2$ weights in Table 4 and the means and associated standard deviations (not shown), one can note that variables X_4, X_8, X_9, X_{11}, X_{13}, X_{30}, X_{39}, and X_{40} possess high within-set $v_{it_i}^2$ weights and have large mean differences between clusters. Given the I different regression phases to estimate $v_{it_i}^2$, these weights can only be compared numerically with variables in the same battery. Cluster 2 can be described as physicians tending to agree with:

—Commercial/scientific exhibits or displays are not effective (X_4);
—Journal ads are not effective (X_8);
—Sampling is not effective (X_9);
—Visual/Audio Cassettes or films are not effective (X_9);
—Physician's radio or T.V. network is not effective (X_{13});
—Once a day dosage is important for his patients (X_{30});
—He does not work in a hospital (X_{39});
—He tends to be a G.P. (X_{40}).

Accordingly, cluster one can be described in a converse manner.

The interpretation here is quite clear—Cluster 1 tends to be hospital based specialists (although it does contain some physicians who are not) with quite a different outlook towards detailing effectiveness than cluster two's general practitioners. It is interesting to note that cluster two contains *no* doctors working in a hospital (variable X_{39} for cluster two has zero variance).

TABLE 4

SYNCLUS Two-Cluster Solution For Drug Study

Two Cluster Solution								
Cluster #	Objects	$v_{it_i}^2$						
		$i=1$		$i=2$		$i=3$		
1	2,3,5,6,8,12,14,15,16,	X_1	.137	X_{14}	.116	X_{34}	.365
	17,19,20,21,22,23,24,26,	X_2	.104	X_{15}	.200	X_{35}	.359
	28,29,31,32,34,36,37,39,40	X_3	.298	X_{16}	.135	X_{36}	.253
		X_4	.373	X_{17}	.221	X_{37}	.161
		X_5	.169	X_{18}	.143	X_{38}	.385
		X_6	.111	X_{19}	.191	X_{39}	.528
2	1,4,7,9,10,11,13	X_7	.235	X_{20}	.266	X_{40}	.469
	18,25,27,30,33,35,38	X_8	.396	X_{21}	.170		
		X_9	.420	X_{22}	.139		
		X_{10}	.196	X_{23}	.445		
		X_{11}	.103	X_{24}	.174		
		X_{12}	.087	X_{25}	.269		
		X_{13}	.505	X_{26}	.208		
$C^2 = .808$				X_{27}	.010		
				X_{28}	.234		
				X_{29}	.313		
				X_{30}	.292		
				X_{31}	.243		
				X_{32}	.159		
				X_{33}	.193		

It is also interesting to note that, except for educational materials for medical schools, the specialist/hospital working cluster (1) is more receptive to all other forms of detailing as demonstrated by larger mean scores on media judgments 1–6 and 7–13. This may indicate that cluster one members are more receptive to manufacturers' marketing mix policies than Cluster 2's doctors.

Another important finding demonstrated in Table 4 is the fact that the psychographic variables, with the possible exception of X_{30}, do not differentiate cluster membership all that much. Rather, it is the media judgment and demographic variables that appear to be the most important batteries in determining this two-cluster solution.

These results are quite consistent with a separate analysis performed by Green and Goldberg (note 4) on the entire $N = 160$ data set.

C. *Comparisons*

Since K-means and SYNCLUS attempt to optimize two different objective functions, there is really no reason to expect to get the same clustering. To pursue this point further, two additional analyses were performed. In the first analysis, the SYNCLUS two-cluster solution was used as a starting configuration for K-means. The SYNCLUS solution produced a total error sums of squares of 1488.96—higher than the 1457.93 in Table 3. K-means then iterated on this configuration and resulted in a different solution—different, in fact, than the one presented in Table 3, with a higher error sums-of-squares = 1463.97. Similarly, the K-means solution in Table 3 was then used as a starting configuration for SYNCLUS. With equal $v_{it_i}^2$ weights on iteration one, it initially rendered a $C^2 = .779$, lower than the .803 in Table 4. SYNCLUS moved away from the K-means solution and converged to the one in Table 4. Note that SYNCLUS provided the same

K-means solution at iteration 1 where all the v_{iti}^2 were set equal. This is to be expected since the two procedures are equivalent under such equal weighting conditions. Again, since the two procedures seemingly try to optimize quite different objective functions, it is not surprising that different results were obtained.

VI. *Discussion*

On the assumption that SYNCLUS continues to show reasonable results when applied to synthetic and empirical data sets, a number of possible applications are suggested. We first describe some potential applications to problems in marketing research and segmentation, followed by a companion discussion of applications in psychology and other behavioral sciences.

A. *Marketing Research and Segmentation*

One of the most actively researched areas in marketing is market segmentation. By this is meant the delineation of groups of consumers who evince similar behavior within segments and different behavior across segments with respect to some set of marketing variables, such as brand preferences, product class consumption, and the like. A large number of different sets of variables—benefits sought, psychographics, demographics, self-concept measures—have been proposed as criteria for segmenting markets. Clustering methods represent a common technique for partitioning markets according to one or more of these batteries (Wind, 1982).

As described at the beginning of this paper, two of the main problems associated with market segmentation entail the weighting of variables within battery and the weighting of the batteries themselves, when the clustering is to be based on two or more data sets for the same individuals. Increasingly, marketing managers are requesting researchers to furnish a single "best" partitioning of the market, even though this usually entails amalgamation over several batteries of variables.

A second problem concerns the fact that most market segmentation studies involve a large number of variables—frequently in excess of 200—as candidates for clustering. Marketing researchers have tried to cope with this problem by conducting some preliminary factor analyses on the data and then clustering consumers on the basis of a relatively few factor scores, rather than the full set of original variables.

Factor scores are based on the intercorrelations of the original variables across the full set of consumers, however, and *not* upon the cluster structure that may exist with respect to the "objects" (i.e., consumers or other entities being clustered). It seems to us that one of the major advantages of SYNCLUS is its use in selecting variables (from some large candidate set) that are the most useful for revealing the inherent cluster structure.

In some sense, the variable weights obtained from SYNCLUS can be viewed as analogous to factor loadings, as obtained in R-type factor analysis (i.e., the factor analysis of variables over objects when the latter are treated simply as replications). For example, one could start with large batteries of candidate variables and use SYNCLUS to select subsets of variables which contribute most to the cluster structure.

Knowledge of the important segmenting variables is also important in its own right for decision making purposes. For example, if markets are best segmented on the basis of attitudes relative to product functionality (as opposed to image-type variables), this information should be useful in product research and advertising.

In addition to data routinely collected on consumer attitudes, demographics, and the like, marketing researchers often work with consumers' perceptions of competitive products on various attributes of interest. For example, new car models may be rated on style features, roominess, fuel economy, acceleration, anticipated trade-in value, availability of

repair service, and so on. It is not unusual to take profile data of this sort and find multidimensional scaling representations of the car models and the attributes in a common space.

SYNCLUS could be used in a complementary way to find a cluster representation of the new car models and a set of weights for the attributes that best delineate the underlying cluster structure. This type of analysis is analogous to the prevalent use of multiple discriminant analysis in which car models (or other sets of objects) are first grouped into segments—often by managerial judgment—and one then tries to find which attributes contribute most to among-group discrimination. In SYNCLUS, however, both the segments and the discriminating attributes are found simultaneously.

The "amalgamation" feature of SYNCLUS is also potentially important in market segmentation studies. For example, the research may test out the sensitivity of cluster structures and variable weights to alternative proposals for separate battery weighting (where different batteries assume the role of the "distinguished" battery which receives the highest weight). If one finds that individual variable importance is relatively insensitive to different sets of a priori specified battery weights, this finding would lend greater credence to the development of a single clustering of the data. On the other hand, if variable weights and the cluster structure are highly sensitive to the user-supplied battery weights, marketing managers might be well advised to consider alternative bases for segmenting their markets.

In sum, it seems to us that the concepts underlying SYNCLUS can be useful in both the screening of candidate segmentation variables and in exploring the robustness of cluster structures and variable weights to alternative battery weights. Moreover, this applicability covers not only the segmentation of consumers or industrial buyers but the clustering of brands or product varieties (e.g., soup flavors, cereal varieties) as well. This latter application appears particularly interesting in research dealing with consumer preferences for alternative bundles of items, such as liquor assortments, season programs of concerts, and the like.

B. *Behavioral Science Applications*

SYNCLUS can be extended to many types of behavioral science applications beyond marketing research. For example, SYNCLUS can be utilized as a classification device for classifying subjects who have taken a series of different psychological tests. SYNCLUS could reveal exactly which variable items in particular test batteries are most important in deriving the resultant classification scheme. This scenario can be extended to tests measuring personality, attitudes, learning, etc. For example, in clinical psychology, the various test batteries could be different personality tests, and SYNCLUS could derive a clustering and associated importance weights for each of the items within a specific battery. This could greatly aid the psychologist in rendering proper interpretation to the various clusters.

In educational psychology, one often takes measurements of student IQ or aptitude via a variety of different tests in order to predict student performance and classify the student population. There is ample psychological theory to also include variables describing the classroom scenario, the school, the teacher, the student's parents, etc. SYNCLUS could be used in such an application to derive a clustering and associated weights to indicate which variables in which batteries were most important in obtaining such a classification.

Indeed, there appear to be many possible psychological applications for SYNCLUS and the technique can be utilized for the case where there is only one battery of items ($I = 1$).

C. *Possible Algorithmic Enhancement*

Some areas of further research concerning different aspects of the algorithm are discussed. Several of these areas are currently being actively pursued.

1. *Phase III Modifications*

One could apply a branch and bound procedure (see Garfinkel and Nemhauser, 1972), to obtain an **E** that could maximize C^2, but for large K and J this would be prohibitively expensive computationally. An alternative, and one that is being explored currently, is a combinatorial optimization procedure (see DeSarbo, 1982) to maximize C^2 (or similar objective function) that would also allow for overlapping clusters.

2. *Integer, Non-negative* $v_{it_i}^2$

Whether or not one wishes to alter the exact nature of the objective function, there is the question of interpretability of the $v_{it_i}^2$ importance weights. First, there is no explicit constraint in the Phase V OLS procedure to insure that all the $v_{it_i}^2$ will be positive. How does one interpret a negative $v_{it_i}^2$ if one should arise? Secondly, how does one compare these weights? Clearly, the larger the weight (assumed to be positive), the more important the variable. But how does one compare one weight of 1.7362 vs another of 1.9937?

There is some justification for considering the imposition of positivity and/or integer constraints on the entire set of $v_{it_i}^2$. For example, one could constrain the $v_{it_i}^2$ to be in some small set of positive integers, such as $\{0, 1, 2, 3\}$. This would guarantee that all $v_{it_i}^2$ would be positive, and would also simplify interpretation considerably, especially if many $v_{it_i}^2$ were set to zero. Unfortunately, enforcing these constraints would complicate this section of the algorithm. Again, a branch and bound procedure or combinatorial optimization method as previously discussed would be necessary—either of which would significantly increase CPU time, especially for large T.

3. *Other Generalizations*

It would also be straightforward to generalize SYNCLUS to allow definition of the squared Euclidean distances for each battery of variables via a generalized metric involving a general quadratic form rather than a simple weighted Euclidean metric (in which the quadratic form is effectively constrained to be diagonal). This would be tantamount to allowing a general linear transformation of the variables in each data set (battery) followed by computation of the simple Euclidean metric on the transformed variables. (This interpretation assumes constraining the quadratic form matrices to be positive-definite or semi-definite, of course). DeSarbo and Mahajan (Note 1) are currently experimenting with such a model with respect to constructing a model and algorithm for constrained classification. Use of such a generalized Euclidean metric would have some of the spirit of the approaches of Kruskal (1972) and of Art, Gnanadesikan and Kettenring (1982) in which the general transformation of a *single* battery of variables is allowed.

A final algorithmic note is that the current approach to the definition of the $\delta_{jj'}$'s was chosen because of its particular relation to the K-means approach—a popular method for nonhierarchical clustering (and, in fact, an alternative formulation of K-means has been proposed here entailing optimizing an appropriate goodness-of-fit criterion defined in terms of these δ's). However, the δ's could be defined in a number of other ways on either nonhierarchical cluster structures (i.e., partitions into K mutually exclusive and exhaustive groups), hierarchical clustering structures (e.g., an ultrametric defined on a hierarchical tree structure) or on overlapping cluster structures such as the ADCLUS-MAPCLUS-INDCLUS-GENNCLUS type structure. The important thing is that however the δ's are defined, the clustering portion of the algorithm be directed at obtaining a least-squares fit of the δ's to the $\bar{d}^2$'s.

D. *Limitations of SYNCLUS*

There are a number of conceptual and computational limitations involving SYNCLUS that should be mentioned. One obvious limitation is that SYNCLUS only involves a rescaling of the variables to designate variable importance via the $v_{it_i}^2$. As mentioned earlier, one could more generally define $\delta_{jj'}$ as:

$$\delta_{jj'} = (\mathbf{e}_j - \mathbf{e}_{j'})\mathbf{A}^{-1}(\mathbf{e}_j - \mathbf{e}_{j'})' + c, \tag{10}$$

where:

$\mathbf{e}_j = j$-th row of $\mathbf{E}$,

$c =$ additive constant,

$\mathbf{A}^{-1} =$ real, symmetric matrix.

This approach is currently utilized in DeSarbo and Mahajan (Note 1) with a different algorithm for constrained classification.

Another limitation of SYNCLUS concerns its inability to accommodate overlapping clusters. Many applications, particularly in marketing are quite amenable to analysis via overlapping clusters, especially those involving market segmentation and product positioning (cf. Arabie, Carroll, DeSarbo, and Wind, 1981). The DeSarbo and Mahajan (Note 1) approach allows the flexibility of accommodating overlapping, nonoverlapping, or "fuzzy" clusters.

Finally, there are the computational limitations concerning the number of objects, number of batteries, and number of variables. In SYNCLUS, the real limitation concerns the number of objects to be classified since the various distance measures are taken across batteries and variables. The initial APL version of SYNCLUS could allow classifying perhaps up to 100 objects. A Fortran version of the program could more than double this bound (due to the computational expense implied by the interpretative nature of the APL language).

E. *More Extensive Monte Carlo Analysis*

Further, more extensive Monte Carlo work needs to be completed to examine in more depth such important issues as: (a) How sensitive are the results to the choice of w_i's and estimates of $v_{it_i}^2$?; (b) When is local optimality a serious problem?; (c) How does the user reliably select K, the number of clusters?; (d) How well does SYNCLUS recover different cluster shapes?; and, (e) How does SYNCLUS compare with other clustering methods?

Appendix
The SYNCLUS Algorithm

1. *Phase I: Input and Preprocessing*

The user must supply the J (objects) by $T = \sum_{i=1}^{I} T_i$ (total number of variables) matrix of profile data $\mathbf{Y}$, the vector of battery importance weights $\mathbf{w}^2$, and the number of clusters (K) for the analysis. Other control parameters including convergence criterion, maximum number of iterations, options for starting values of $v_{it_i}^2$, (either all equal or inversely proportional to the variance of $y_{t_i}^{(i)}$), and any preprocessing options must also be specified.

The SYNCLUS program allows one to cluster the raw data ($\mathbf{Y}$), column-centered data (which doesn't affect the computed distances between objects), column-standardized data, or orthogonalized data (employing standard singular value decomposition techniques).

2. *Phase II: Calculate distances*

Given starting values for $v_{it_i}^2$ and the desired preprocessing, the second phase of SYNCLUS calculates the three-way array of squared distances via:

$$d_{jj'}^{2(i)} = \sum_{t_i=1}^{T_i} v_{it_i}^2 (y_{jt_i}^{(i)} - y_{j't_i}^{(i)})^2, \tag{A-1}$$

one of the terms on the right hand side of equation (1). We then define:

$$\bar{d}_{jj'}^2 = \sum_{i=1}^{I} w_i^2 d_{jj'}^{2(i)}, \tag{A-2}$$

as the two-way matrix of averaged (weighted) squared distances to be used in the clustering phase to follow.

3. *Phase III: Generalized K-means*

MacQueen (1967) introduced his K-means clustering procedure as an iterative nonhierarchical clustering technique. Basically the K-means procedure starts with K "seed points," each of which defines the location of a single cluster. A sequence of points is sampled from some distribution, and each point is assigned to the group whose centroid it is closest to. After the points are allocated to clusters, the cluster centroids are adjusted, and the points reallocated. This procedure is iterated, and stops when there is no movement of a point from one cluster to another for any case. This procedure provides a heuristic for minimizing the error of the partition:

$$\begin{aligned} EP &= \sum_{k=1}^{K} \sum_{j_k \in k}^{J_k} \sum_{t=1}^{T} (y_{j_k t} - \bar{y}_{kt})^2 \\ &= \sum_{k=1}^{K} \sum_{j_k \in k}^{J_k} D_{j_k}^2 k, \end{aligned} \tag{A-3}$$

where:

$y_{j_k t}$ = the t-th coordinate of the j_k-th point;

$\bar{y}_{kt}$ = the t-th coordinate of the k-th cluster centroid:

$$= \frac{1}{J_k} \sum_{j_k=1}^{J_k} y_{j_k t};$$

J_k = the number of objects/points in the k-th cluster;

T = the total number of coordinates or variables:

$$= \sum_{i=1}^{I} T_i.$$

Hartigan (1975, 1978), and more recently Pollard (1980), discuss other relevant issues regarding K-means such as asymptotic distribution theory, approximate F-tests for testing for the number of clusters K, variable weighting, shapes of clusters, etc. We propose to generalize this two-way K-means procedure, which operates on the objects by variables profile data matrix (**Y**), to a three-way case. One important aspect of this is that our approach to K-means utilizes only the matrix of squared Euclidean distances between the objects or points, and not (explicitly, at least) the point coordinates. Späth (1980) provides a similar two-way K-means algorithm which uses distances instead of coordinates. Our approach to K-means clustering using distances is described below.

Given a single $J \times J$ matrix of squared Euclidean distances, we initially attempt to find a set of K (number of clusters) seed points. (In our current application, this matrix will be the matrix of weighted mean squared Euclidean distances defined in equation (A-2), but in fact the K-means algorithm described here could be applied to any matrix of squared Euclidean distances which we shall denote as $\mathbf{D}^2 = \| d_{jj'}^2 \|$.) In this case, we use the approach of choosing K of the actual points to define the seed points. This is done initially by searching $\mathbf{D}^2$ for the largest entry ($d_{j^*j'^*}$) and using the corresponding j^* and j'^* points or objects as the first two seeds. Then, given j^* and j'^*, we search for the point or object which maximizes the sum of the squared distances from the first two points, i.e., given j^* and j'^*, we find l^* such that $d_{j^*l^*}^2 + d_{j'^*l^*}^2$ is maximum over all other $J - 2$ objects. Once j^*, j'^*, and l^* are found, we iterate this procedure (if $K > 3$) until the total number of seed points equals K.

Once these K seed points have been selected, we then assign the remaining $J - K$ points to these K seeds/clusters simply by assigning each point (a) to the closest $\{\min d^2ab \; b = j^*, j'^* \cdots d_{ab}^2\}$ seed point. Because of potential problems that may arise in choosing K seed points in fewer than $K - 1$ dimensions, an option is provided to select the K seed points according to the following procedure. Having selected $L < K$ seed points, SYNCLUS selects the $L + 1$'st based on a max-min criterion defined as follows:

$$\underset{j_{L+1} \neq j_1 \cdots j_L}{\text{Max}} \left(\underset{l=1 \cdots L}{\text{Min}} \; d_{j_l j_{L+1}}^2 \right). \tag{A-4}$$

This provides an alternative starting clustering for the iterative procedure to follow. In our empirical work, however, we have found that the algorithm described earlier for generating seed points seems to lead to better solutions i.e., fewer local minima, speedier convergence, and otherwise superior performance of the algorithm. This may be because, in practice, we are in fact dealing with data in which the clusters are not embedded in such a small dimensional subspace.

With this starting clustering, each point is now reassigned to the cluster to whose centroid it is closest. We use equation (A-5) below to calculate the point-centroid squared distances without explicit use of coordinates:

$$D_{jk}^2 = \frac{1}{J_k} \sum_{j_k=1}^{J_k} d_{jj_k}^2 - D_k^2, \tag{A-5}$$

where:

$$D_k^2 = \frac{1}{2J_k^2} \sum_{j_k=1}^{J_k} \sum_{j'_k=1}^{J_k} d_{j_k j'_k}^2. \tag{A-6}$$

That is, object j is assigned to cluster k for minimum D_{jk}^2, $\forall k = 1, \ldots, K$. This is done simultaneously for $j = 1, \ldots, J$ and is repeated iteratively until *no* points change cluster membership, i.e., until $\sum_{j=1}^{J} \sum_{k=1}^{K} D_{jk}^2$ can not be further minimized. This phase thus renders the clustering to be used in this iteration: $\mathbf{E} = \| e_{jk} \|$.

Phase IV: Definition of Cluster Distances

As a secondary stage of this generalized K-means procedure, we define new distances $\delta_{jj'}$ in order to minimize Z_1^2 at this stage. The K-means solution can be shown to provide a least-squares fit to squared distances defined in terms of the data in a sense specified below.

Let us first define an adjacency matrix:

$$\mathbf{A} = \| a_{jj'} \|, \tag{A-7}$$

where:

$$a_{jj'} = \begin{cases} 1 & \text{if } e_{jk} = e_{j'k} = 1, \text{ for some } k, \\ 0 & \text{otherwise;} \end{cases}$$

(i.e., $a_{jj'} = 1$ iff j and j' are in the same cluster k). We then define a modified adjacency matrix:

$$\mathbf{A}^* = \| a^*_{jj'} \|, \tag{A-8}$$

where:

$$a^*_{jj'} = \frac{a_{jj'}}{\sqrt{n_j n_{j'}}} = \begin{cases} \dfrac{1}{J_k} & \text{if } j \text{ and } j' \text{ are both in cluster } k, \\ 0 & \text{if } j \text{ and } j' \text{ are in different clusters;} \end{cases}$$

n_j = the size of the cluster which contains object j.

Now, define:

$$\mathbf{U} = \| 1 \|, \tag{A-9}$$

where, $\mathbf{U}$, is simply a units matrix. We now estimate α and β in:

$$\boldsymbol{\Delta} = \alpha \mathbf{A}^* + \beta \mathbf{U}^* \cong \| d^2_{jj'} \|, \tag{A-10}$$

where estimates of α and β are obtained optimally by ordinary least-squares from the equation:

$$\begin{pmatrix} d^2_{11} \\ d^2_{12} \\ d^2_{13} \\ \vdots \\ d^2_{J-1,J} \\ d^2_{J,J} \end{pmatrix} \cong \begin{bmatrix} a^*_{11} & 1 \\ a^*_{12} & 1 \\ a^*_{13} & 1 \\ \vdots & \vdots \\ a^*_{J-1,J} & 1 \\ a^*_{J,J} & 1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{A-11}$$

with the solution (denoting the $J^2 \times 1$ matrix on the left as $\mathbf{D}^2$, the $J^2 \times 2$ matrix on the right as $\mathbf{X}$, and the vector containing α and β as $\mathbf{c}$) given by:

$$\hat{\mathbf{c}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^2 = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}. \tag{A-12}$$

After this estimation phase, $\boldsymbol{\Delta} = \| \delta_{jj'} \|$ is redefined as in equation (A-10), using the least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$ obtained as described above. We now prove that the K-means algorithm followed by this regression phase (at least locally) optimizes the fit of the δ's to the d^2's in a least-squares sense.

As discussed earlier, the generalized K-means algorithm (at least locally) optimizes (minimizes):

$$EP = \sum_{k=1}^{K} \sum_{j_k=1}^{J_k} D^2_{j_k k}. \tag{A-13}$$

(See equation (A-3) and following definitions). It can easily be shown that:

$$\sum_{j_k}^{J_k} D^2_{j_k k} = \frac{1}{2J_k} \sum_{j_k=1}^{J_k} \sum_{j'_k=1}^{J_k} d^2_{j_k j'_k}, \tag{A-14}$$

so generalized K-means minimizes:

$$EP = \frac{1}{2} \sum_{k=1}^{K} \sum_{j_k=1}^{J_k} \sum_{j'_k=1}^{J_k} \left(\frac{1}{J_k}\right) d^2_{j_k j'_k}. \tag{A-15}$$

Equation (A-15) can be written as:

$$EP = \frac{1}{2} \sum_{j=1}^{J} \sum_{j'=1}^{J} s_{jj'} d^2_{jj'}, \tag{A-16}$$

where:

$$s_{jj'} = \begin{cases} 0 & \text{if } j, j' \text{ are in distinct clusters,} \\ \dfrac{1}{J_k} & \text{if } j, j' \text{ are jointly in the } k\text{-th cluster.} \end{cases}$$

Minimizing EP for fixed d^2's is equivalent to maximizing the cross product:

$$S(\alpha, \beta) = \sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'}^{(\alpha,\beta)} d^2_{jj'} \qquad (\alpha < 0), \tag{A-17}$$

where:

$$\delta_{jj'}^{(\alpha,\beta)} = \alpha s_{jj'} + \beta,$$

for fixed α, β.

Again, assuming fixed α and β, we define a sum-of-squares function:

$$SS(\alpha, \beta) = \sum_{j=1}^{J} \sum_{j'=1}^{J} [\delta_{jj'}^{(\alpha,\beta)}]^2 \tag{A-18}$$

$$= \sum_{j=1}^{J} \sum_{j'=1}^{J} [\alpha s_{jj'} + \beta]^2 \tag{A-19}$$

$$= \sum_{j=1}^{J} \sum_{j'=1}^{J} [\alpha^2 s^2_{jj'} + 2\beta s_{jj'} + \beta^2] \tag{A-20}$$

$$= \alpha^2 \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} s^2_{jj'} \right] + 2\beta \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} s_{jj'} \right] + \beta^2 \tag{A-21}$$

$$= \alpha^2 \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} s^2_{jj'} \right] + 2\beta \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} s_{jj'} \right] + \beta^2. \tag{A-22}$$

It is straightforward to show that:

$$\sum_{j=1}^{J} \sum_{j'=1}^{J} s^2_{jj'} = K$$

$$\sum_{j=1}^{J} \sum_{j'=1}^{J} s_{jj'} = J.$$

So:

$$SS(\alpha, \beta) = \alpha^2 K + 2\beta J + \beta^2$$

$$= \text{constant (for fixed } \alpha, \beta, J, K). \tag{A-23}$$

Since for fixed $d^2_{jj'}$'s,

$$\sum_{j=1}^{J} \sum_{j'=1}^{J} (d^2_{jj'})^2 \equiv \sum_{j=1}^{J} \sum_{j'=1}^{J} d^4_{jj'}$$

is also constant, it follows that K-means *maximizes:*

$$C(\alpha, \beta) \equiv \frac{S(\alpha, \beta)}{\left\{SS(\alpha, \beta) \sum_{j=1}^{J} \sum_{j'=1}^{J} d_{jj'}^4\right\}^{1/2}} \tag{A-24}$$

$$= \frac{\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'}^{(\alpha, \beta)} d_{jj'}^2}{\left\{\sum_{j=1}^{J} \sum_{j'=1}^{J} [\delta_{jj'}^{(\alpha, \beta)}]^2 \sum_{j=1}^{J} \sum_{j'=1}^{J} d_{jj'}^4\right\}^{1/2}}. \tag{A-25}$$

It then follows that if we define $\delta_{jj'}$ to be that $\delta_{jj'}^{(\alpha, \beta)}$ maximizing $C(\alpha, \beta)$ over all α, β, then the resulting K clusters (defined via generalized K-means) and α and β *jointly* maximize:

$$C = \frac{\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'} d_{jj'}^2}{\left[\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'}^2 \sum_{j=1}^{J} \sum_{j'=1}^{J} d_{jj'}^4\right]^{1/2}} \tag{A-26}$$

(with $\delta_{jj'}$ of the form $\alpha s_{jj'} + \beta$).

This is equivalent (except for a rescaling of the $\delta_{jj'}$'s (see Kruskal and Carroll, 1969)) to minimizing:

$$\tilde{Z}_1^2 = \frac{\sum_{j=1}^{J} \sum_{j'=1}^{J} (\delta_{jj'} - d_{jj'}^2)^2}{\sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'}^2}. \tag{A-27}$$

In SYNCLUS we use generalized K-means to optimize $\tilde{Z}_1^2$, with

$$d_{jj'}^2 \equiv \bar{d}_{jj'}^2 = \sum_{i=1}^{I} w_i^2 [d_{jj'}^{(i)}]^2.$$

Note, however, that in all cases the summation is a double sum over **all** values of j and j' (including diagonals). In particular:

$$\delta_{jj} = \frac{\alpha}{n_j} + \beta \neq 0 \tag{A-28}$$

where n_j is the size of the cluster to which point j belongs (whereas, of course $\bar{d}_{jj'}^2 \equiv 0$, for all j). That is, for example the uncentered correlation C optimized by SYNCLUS, correlates δ and $\bar{d}^2$ over the entire $J \times J$ matrices Δ and $\bar{\mathbf{D}}^2$.

To prove that minimizing $\tilde{Z}_1^2(\delta, \bar{d}^2)$ over δ is equivalent to minimizing Z_1^2 over δ (with the $d_{jj'}^{(i)2}$'s fixed), we use the equivalence of optimizations of Z_1^2 and Z_2^2 (except for scaling). Thus, consider Z_2^2, now assumed to be optimized over the $v_{it_i}^2$'s (and thus over the class of permissible values of $d^{(i)2}$'s). From the Kruskal-Carroll (1969) results it follows that the $d_{jj}^{(i)2}$'s are so scaled that

$$\sum_{j=1}^{J} \sum_{j'=1}^{J} d_{jj'}^{(i)4} = \sum_{j=1}^{J} \sum_{j'=1}^{J} \delta_{jj'}^2 / \cos^2 \theta_i, \tag{A-29}$$

where:

$$\cos\theta_i = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}d_{jj'}^{(i)2}}{\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}^2\sum_{j=1}^{J}\sum_{j'=1}^{J}d_{jj'}^{(i)4}\right)^{1/2}}, \tag{A-30}$$

then:

$$Z_2^2 = \frac{1}{\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}^2}\sum_{i=1}^{I}w_i^{*2}\sum_{j=1}^{J}\sum_{j'=1}^{J}(\delta_{jj'} - d_{jj'}^{(i)2})^2, \tag{A-31}$$

where:

$$w_i^{*2} = w_i^2\cos^2\theta_i. \tag{A-32}$$

It can easily be shown that the *unconstrained* minimum (over $\delta_{jj'}$), of Z_2^2 as defined in equation (A-31) is obtained for

$$\delta_{jj'}^2 \propto \frac{\sum_{i=1}^{I}w_i^{*2}d_{jj'}^{(i)2}}{\sum_{i=1}^{I}w_i^{*2}}, \tag{A-33}$$

(with a constant of proportionality unimportant for present purposes). Let $\tilde{d}_{jj'}^{(i)2}$ represent values of $d_{jj'}^{(i)2}$, yielding an ordinary least squares fit to the $\delta_{jj'}$'s, so the values optimizing Z_2^2 are simply $\tilde{d}_{jj'}^{(i)2}/\cos^2\theta_i$, since

$$\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{d}_{jj'}^{(i)4} = \left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\delta_{jj'}^2\right)\cos\theta_i^2.$$

It then follows that

$$\frac{\sum_{i=1}^{I}w_i^{*2}d_{jj'}^{(i)2}}{\sum_{i=1}^{I}w_i^{*2}} = \frac{\sum_{i=1}^{I}w_i^2\tilde{d}_{jj'}^{(i)2}}{\sum_{i=1}^{I}w_i^{*2}}, \tag{A-34}$$

which (except for a scale factor, which is unimportant for present purposes) is just the weighted average of the $\tilde{d}_{jj'}^{(i)2}$'s. But in previous iterative cycles (which actually corresponds to Phase V, described below) the $d_{jj'}^{(i)2}$'s themselves were in fact calculated by precisely such an ordinary least-squares procedure. Therefore, the weighted mean of the $d_{jj'}^{(i)2}$'s (the $\bar{d}_{jj'}^2$'s calculated with the ordinally specified weights w_i^2) do indeed correspond to the *unconstrained* optimum values of $\delta_{jj'}$—that is, if we then fit *constrained* $\delta_{jj'}$'s to the $\bar{d}_{jj'}^2$'s, these values will constitute a constrained optimum for the problem at hand if and only if certain orthogonality conditions hold. (These orthogonality conditions state that the vector from the data vector to the unconstrained solution must be orthogonal to that from the unconstrained to the constrained solution.) These orthogonality conditions, in the present case, can be stated as

$$\sum_{i=1}^{I}w_i^2\sum_{j=1}^{J}\sum_{j'=1}^{J}(d_{jj'}^{(i)2} - \bar{d}_{jj'}^2)(\bar{d}_{jj'}^2 - \delta_{jj'}) = 0. \tag{A-35}$$

By straightforward algebra (and utilization of the definition of $\bar{d}^2_{jj'}$ as

$$\bar{d}^2_{jj'} \equiv \sum_{i=1}^{I} w_i^2 d^{(i)2}_{jj'} \tag{A-36}$$

as well as the side condition that $\sum_{i=1}^{I} w_i^2 = 1$), this orthogonality condition can easily be seen to hold. (Note that the scalar product in terms of which the required orthogonality condition is stated is a *weighted* scalar product, with weights corresponding to those in the weighted least squares problem being solved.)

Phase V: Solve for $v^2_{it_i}$

With the redefinition of $\Delta = \| \delta_{jj'} \|$ from the previous stage, we wish to estimate a set of optimal variable importance weights or rescalings $v^2_{it_i}$, $(i = 1, \ldots, I, t_i = 1, \ldots, T_i)$. Recall that:

$$d^{2(i)}_{jj'} = \sum_{t_i=1}^{T_i} v^2_{it_i}(y^{(i)}_{jt_i} - y^{(i)}_{j't_i})^2. \tag{A-37}$$

With Δ fixed, one can estimate the $v^2_{it_i}$ by a series of I regressions via:

$$\begin{pmatrix} \delta^2_{12} \\ \delta^2_{13} \\ \vdots \\ \delta^2_{J-1,J} \end{pmatrix} \cong \begin{bmatrix} (y^{(i)}_{11} - y^{(i)}_{21})^2 & (y^{(i)}_{1T_i} - y^{(i)}_{2T_i})^2 \\ (y^{(i)}_{11} - y^{(i)}_{31})^2 & (y^{(i)}_{1T_i} - y^{(i)}_{3T_i})^2 \\ \vdots & \vdots \\ (y^{(i)}_{J-1,1} - y^{(i)}_{J,1})^2 & (y^{(i)}_{J-1,T_i} - y^{(i)}_{J,T_i})^2 \end{bmatrix} \begin{bmatrix} v^2_{i1} \\ v^2_{i2} \\ \vdots \\ v^2_{iT_i} \end{bmatrix}; \tag{A-38}$$

or, denoting the $\binom{J}{2} \times T_i$ matrix of independent variables on the right by $\mathbf{X}^{(i)}$, we estimate the vector $\mathbf{v}_i^2$ containing the $v^2_{it_i}$'s via:

$$\hat{\mathbf{v}}_i^2 = (\mathbf{X}'^{(i)}\mathbf{X}^{(i)})^{-1}\mathbf{X}'^{(i)}\boldsymbol{\delta}. \tag{A-39}$$

(where $\boldsymbol{\delta}$ is the $\binom{J}{2}$ component vector on the left containing the $\delta_{jj'}$'s).

Phase VI: Test for Convergence

Once a complete major iteration (Phases II-V) has been completed, tests are performed to see if the iterative algorithm has converged or has exceeded a stipulated maximum number of iterations. Basically, we test to see if $(C^2_{IT} - C^2_{IT-1}) < \varepsilon$, where ε is some small constant (e.g., $\varepsilon = .001$), or if $IT \geq$ MAXIT, where MAXIT is some stipulated maximum number of iterations (e.g., $MAXIT = 100$). If either of these conditions is true, we go to Phase VII. If not, we set $IT = IT + 1$ and return to Phase II.

Phase VII: Output

At this last phase we print the final clustering matrix $\mathbf{E}$, the vectors of variable importance weights $\mathbf{v}_i^2$, Z_1^2, α, B, C^2, Z_2^2, Δ, $\bar{\mathbf{D}}^{2(\cdot)}$, $\mathbf{D}^{2(i)}$, for $i = 1, \ldots, I$. Options are also available to perform regression-type residual analyses via plots and listings.

REFERENCE NOTES

1. DeSarbo, W. S. and Mahajan, V. (1982). Constrained classification, *Working Paper*, Bell Laboratories, Murray Hill, N.J.
2. Fowlkes, E. (1981). Variable selection in clustering, *presented at Bell Laboratories Work Seminar*, Murray Hill, N.J.
3. Fowlkes, E., Gnanadesikan, R., and Kettenring, J. R. (1982). Variable selection in clustering, *Work in Progress*, Bell Laboratories, Murray Hill, N.J.
4. Green, P. E. and Goldberg, S. M. (1981). The beta drug company case, *Wharton-School Publication*, University of Pennsylvania.

REFERENCES

Arabie, P. A. and Carroll, J. D. (1980). *MAPCLUS:* A mathematical programming approach to fitting the *ADCLUS* model, *Psychometrika, 45*, 211–235.

Arabie, P., Carroll, J. D. DeSarbo, W. S., and Wind, Y. (1981). Overlapping clustering: A new methodology for product positioning, *Journal of Marketing Research, 18*, pp. 000–000.

Art, D., Gnanadesikan, R., and Kettenring, J. R. (in press). Data-based metrics for cluster analysis, *Utilitas Mathematica.*

Carroll, J. D. and Arabie, P. A. (1983). INDCLUS: An individual differences generalization of the ADCLUS Model and the MAPCLUS Algorithm, *Psychometrika, 48*, 157–169.

DeSarbo, W. S. (1982). GENNCLUS: New models for general nonhierarchical clustering analysis, *Psychometrika, 47*, 449–475.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data, *Journal of the American Statistical Association, 62*, 1159–1178.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers, C-23*, 881–890.

Garfinkle, R. S. and Nemhauser, G. L. (1972). *Integer Programming*, New York: J. Wiley and Sons.

Hartigan, J. A. (1975). *Clustering Algorithms*, New York: J. Wiley and Sons.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria, *Annals of Statistics*, Vol. 6, No. 1, 117–131.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika, 29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method, *Psychometrika, 29*, 115–129.

Kruskal, J. B. and Carroll, J. D. (1969). Geometrical models and badness-of-fit functions, in *Multivariate Analysis III*, edited by P. R. Krishnaiah, New York: Academic Press, 639–670.

Kruskal, J. B. (1972). Linear transformations of multivariate data to reveal clustering, in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, edited by Shepard, R. N., Romney, A. K., and Nerlove, S. B., New York: Seminar Press, 181–191.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, 231–297.

Morrison, D. G. (1967). Measurement problems in cluster analysis, *Management Science, 13*, 775–780.

Pollard, D. (1981). Strong consistency of K-means clustering, *Annals of Statistics*, Vol. 9, No. 1, 135–140.

Rohlf, F. J. (1970). Adaptive hierarchical clustering schemes, *Systematic Zoology, 19*, 58–82.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*, San Francisco: W. H. Freeman and Co.

Späth, H. (1980). *Cluster Analysis Algorithms*, Chichester, England: Ellis Horwood Ltd.

Shepard, R. N. and Arabie, P. (1979). Additive clustering: representation of similarities as combination of discrete overlapping properties, *Psychological Review, 86*, 87–123.

Wind, Y. (1982). *Product Policy: Concepts, Methods and Strategy*, Reading, Mass.: Addison-Wesley.

Manuscript received 11/22/82
First revision received 7/6/83
Final version received 10/6/83