PSYCHOMETRIKA—VOL. 54, NO. 4, 699–706
DECEMBER 1989

CLUSTERING N OBJECTS INTO K GROUPS UNDER OPTIMAL SCALING OF VARIABLES

STEF VAN BUUREN

DEPARTMENT OF PSYCHONOMY
UNIVERSITY OF UTRECHT

WILLEM J. HEISER

DEPARTMENT OF DATA THEORY
UNIVERSITY OF LEIDEN

We propose a method to reduce many categorical variables to one variable with k categories, or stated otherwise, to classify n objects into k groups. Objects are measured on a set of nominal, ordinal or numerical variables or any mix of these, and they are represented as n points in p-dimensional Euclidean space. Starting from homogeneity analysis, also called multiple correspondence analysis, the essential feature of our approach is that these object points are restricted to lie at only one of k locations. It follows that these k locations must be equal to the centroids of all objects belonging to the same group, which corresponds to a sum of squared distances clustering criterion. The problem is not only to estimate the group allocation, but also to obtain an optimal transformation of the data matrix. An alternating least squares algorithm and an example are given.

Key words: homogeneity analysis, cluster analysis, variable importance, GROUPALS.

Problem

Consider a data matrix H ($n \times m$) in which the rows correspond to n objects measured on m categorical variables. Let $\mathbf{c} = (c_1, \ldots, c_i, \ldots, c_n)'$ be an initially unknown vector of n integers ranging from 1 to k, and let k ($2 \leq k \leq n$) be a given number of groups. The problem is to estimate $\mathbf{c}$, that is, to sort each object into one of k groups, such that $\mathbf{c}$ preserves the differences among the profiles h_i ($i = 1, \ldots, n$) as closely as possible. In cluster analysis this problem is known as the set partitioning problem. Alternatively, it can also be viewed as a dimension reduction problem in which the goal is to reduce a large number of categorical variables to one categorical variable with k categories.

An important aspect of the problem is that the variables may be measured on nominal, ordinal or interval scales, or on any mix of these. In the field of cluster analysis a number of (dis)similarity coefficients has been proposed for mixed variables (e.g., Gower, 1971; Lance & Williams, 1967, 1968; Opitz, 1980). A different approach is to transform mixed data into numerical variables, so that the use of Euclidean metric is possible. We accomodate for mixed data by means of optimal scaling (Gifi, 1981; Young, 1981). The major difference with previous approaches is that we treat the transformation and clustering problem *simultaneously*.

A closely related issue is that of differential weighting of variables. In DeSarbo,

The authors thank Eveline Kroezen and Teije Euverman for their comments on a previous draft of this paper.

Requests for reprints should be sent to Stef van Buuren, Department of Psychonomy, Trans II 17.25, Heidelberglaan 2, 3584 CS Utrecht, THE NETHERLANDS.

0033-3123/89/1200-9106$00.75/0

Carroll, Clark and Green (1984) and De Soete, DeSarbo and Carroll (1985) techniques are discussed for simultaneously estimating both the cluster allocation and the variable importance. Our method can be considered as a special case of SYNCLUS (DeSarbo et al., 1984), since we will not consider a partitioning of the variables into sets. However, it is more general than SYNCLUS in that it allows a much wider class of data transformations. Furthermore, the present method does not require explicit calculation of the $(n \times n)$ inter object distances matrix.

The present method is a generalization of the sum of squared distances (SSQD) cluster analysis problem to the case of mixed measurement level variables. So, in principle it can be applied to any problem for which SSQD clustering has been proposed with the additional advantage that it provides a transformation of the data that is optimal with respect to the obtained cluster allocations. Theory and applications of SSQD clustering are discussed in Hartigan (1975), Späth (1985) and others. Our method can also be useful in detecting and matching shapes in binary data, for example, for the recognition of characters, although we have not systematically explored these possibilities. Another potential application area is in the field of latent class analysis (McCutcheon, 1987).

Method

We assume that the reader is familiar with homogeneity analysis, also known as multiple correspondence analysis or dual scaling. If not, one may consult for example van Rijckevorsel and de Leeuw (1988). We adopt their notation here.

Let $\mathbf{k} = (k_1, \ldots, k_j, \ldots, k_m)$ be the m-vector containing the number of categories of each variable, and let p denote the dimensionality of the analysis. Let each variable $h_j(j = 1, \ldots, m)$ be coded into an $(n \times k_j)$ indicator matrix G_j, and let the allocation vector $\mathbf{c}$ be coded into the $(n \times k)$ indicator matrix G_c. Furthermore, define X as a $(n \times p)$ matrix of object scores and define m $(k_j \times p)$ matrices Y_j of category quantifications. Homogeneity analysis then amounts to minimizing

$$\sigma(X; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} \operatorname{tr} (X - G_jY_j)'(X - G_jY_j) \tag{1}$$

over X and Y_j under appropriate normalization conditions. We deal with mixed measurement levels by restricting the class of data transformations, that is, we restrict Y_j. A systematic description of these types of restrictions can be found in de Leeuw (1984).

In this paper we introduce a restriction on the object scores X. Let Y be a $(k \times p)$ matrix of cluster points. We replace each point x_i, the i-th row of X, by a corresponding cluster point y_r, the r-th row of Y. This is equivalent to requiring $X = G_cY$. If v denotes the vector of the first k integers, then $\mathbf{c} = G_c\mathrm{v}$. Working with $X = G_cY$ instead implies that apart from allocation, we also aim for a scaling of the clusters in p-dimensional space. Now (1) can be written as

$$\sigma(G_c; Y; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} \operatorname{tr} (G_cY - G_jY_j)'(G_cY - G_jY_j). \tag{2}$$

We minimize (2) by alternating least squares. For fixed G_c and Y (2) can be minimized over Y_j by the procedures described in Gifi (1981). On the other hand, suppose that $Z = 1/m \; \Sigma \; G_jY_j$. Then by inserting the identity $G_cY = Z - (Z - G_cY)$ into (2) and noting that the cross product vanishes we find that the loss function can be split into additive components as follows:

$$\sigma(G_c;\ Y;\ Y_1,\ .\ .\ .\ ,\ Y_m) = \frac{1}{m}\sum_{j=1}^{m} \text{tr}\ (Z-G_jY_j)'(Z-G_jY_j) + \text{tr}\ (Z-G_cY)'(Z-G_cY). \qquad (3)$$

For fixed $Y_1, \ldots, Y_m$ the first component of (3) is constant, so it is only the second component that must be minimized over G_c and Y. In cluster analysis this problem is known as sum of squared distances (SSQD) clustering. It can be easily seen that for any allocation G_c the criterion is minimized over Y by setting $Y := (G_c'G_c)^{-1}G_c'Z$, that is, by setting the cluster points y_r equal to the cluster centroids in terms of Z. A number of procedures is known for minimizing the SSQD criterion over all possible allocations G_c. In the remainder we adopt the iterative K-means algorithm (Hartigan, 1975; Späth, 1985), because this algorithm is well studied, it is applicable to large data sets, and it has satisfactory performance characteristics (Milligan, 1980; Scheibler & Schneider, 1985). As a final step we set $X := G_cY$.

In order to avoid the trivial outcome where both $Y_1, \ldots, Y_m$ and X are zero, some normalization of (1) and (2) should be undertaken. In (1) we can use $X'X = \text{I}$, but in (2) this is inconvenient since we must simultaneously deal with two types of restrictions on X: the normalization and the clustering restriction, and this leads to computational complications. For the same reason, normalization of the category quantifications $Y_1, \ldots, Y_m$ is inconvenient.

A more attractive alternative is to apply a transfer of normalization procedure, also used by van der Burg and de Leeuw (1983) in a canonical correlation context. This can be done because the restrictions remain satisfied under linear transformations. The idea is that a normalization on X can be transferred to a normalization on $Y_1, \ldots, Y_m$, and vice versa, while *preserving* the loss. We will now demonstrate that this is possible for $\sigma(X;\ Y_1, \ldots, Y_m)$ in (1). A completely analogous result is true for (2). Suppose we have some solution with normalization $X'X = \text{I}$, then nonsingular transformation matrices P and Q can be found such that $\sigma(X;\ Y_1, \ldots, Y_m) = \sigma(XP;\ Y_1Q, \ldots, Y_mQ)$ with normalization $\Sigma\ (Y_jQ)'G_j'G_jY_jQ = \text{I}$ by using $P = K\Lambda$ and $Q = K\Lambda^{-1}$ from the eigenvalue decomposition $1/m\ \Sigma\ Y_j'G_j'G_jY_j = K\Lambda^2K'$. By substituting for X and Y_j and expanding the result we derive

$$\sigma(X;\ Y_1,\ .\ .\ .\ ,\ Y_m)=\sigma(XP;\ Y_1Q,\ .\ .\ .\ ,\ Y_mQ)=p+\text{tr}\ (\Lambda^2)-2\ \text{tr}\ X'\left(\frac{1}{m}\sum G_jY_j\right). \qquad (4)$$

Applying the procedure twice enables us to estimate G_c and Y under normalization $\Sigma\ (Y_jQ)'G_j'G_jY_jQ = \text{I}$ and $Y_1, \ldots, Y_m$ under normalization $X'X = \text{I}$.

Some comments must be made about the expected properties of the clusters that the SSQD criterion produces. First, as demonstrated theoretically by Binder (1978), anticipated in Bock (1972), and found empirically as well (Gordon, 1981, p. 52), the clusters tend to be of roughly equal size. If one has prior evidence that a data set strongly deviates from this type of clustering, one should hesitate to use our procedure. Second, Wishart (1969) notes that the SSQD criterion favors hyperspherically shaped clusters, even when the data clearly exhibit other (e.g., chaining) structures. Other criteria may be more appropriate in the latter case, although our use of data transformations will tend to alleviate this drawback. Third, Friedman and Rubin (1967) show that the criterion may give rise to different partitions if the data are linearly transformed. Indeed, in the present case, the SSQD criterion is not applied to the data itself, but to a subspace of the optimally scaled variables. Thus we search for the best partition over a potentially much wider class of transformations.

Algorithm

The method was implemented in a FORTRAN computer program called GROUPALS. The program takes the following algorithmic steps:

Step 1: Initialization. The user must supply the desired number of clusters k and the dimensionality of the solution p. Construct m indicator matrices G_j. Initialize X^0 with orthonormalized random numbers, and let G_c^0 be the indicator matrix of some initial partition. Define $D_j = G_j'G_j$. Set iteration counter $t = 1$.

Step 2: Quantification. Let $Y_j^t := D_j^{-1}G_j'X^{t-1}$ for $j = 1, \cdots, m$. This step minimizes (2) over Y_j for a given X^{t-1} and it corresponds simply to calculating the centroids of objects in the same category. Subsequently, level restrictions are carried out on the relevant quantifications Y_j^t, by projection.

Step 3: Transfer normalization to quantifications. Compute the eigenvalue decomposition of $1/m \ \Sigma \ Y_j^{t\prime}D_jY_j^t = K\Lambda^2K'$. Let $Z^t := 1/m \ \Sigma \ G_jY_j^tK\Lambda^{-1}$.

Step 4: Estimation of cluster allocations. Minimize the SSQD criterion $\text{tr}\,(Z^t - G_cY)'(Z^t - G_cY)$ over G_c and Y, given Z^t and G_c^{t-1}, by the K-means algorithm. This results in G_c^t and Y^t. Define $\bar{X}^t := G_c^tY^t$.

Step 5: Transfer normalization to object scores. Compute the eigenvalue decomposition of $\bar{X}^{t\prime}\bar{X}^t = L\Psi^2L'$. Let $X^t := \bar{X}^tL\Psi^{-1}$. Now $X^{t\prime}X^t = \text{I}$.

Step 6: Convergence test. Compute the value of loss function (1) and check whether the difference between the values at iterations t and $t - 1$ is smaller than some predetermined criterion value, or whether a maximum number of iterations has been reached. If so, stop. Otherwise, set $t := t + 1$, and go to Step 2.

If one uses rank-one restrictions (Gifi, 1981) the component loadings a_j^t should also be renormalized to insure that Step 2 always starts with a proper initialization. In this case we add $\bar{a}_j^t := a_j^tK\Lambda^{-1}$ after Step 3, and we add $a_j^t := \bar{a}_j^tL\Psi$ after Step 5. Now, the loss values will monotonically decrease, and so the algorithm converges to a minimum.

Test runs were carried out on an Amdahl V7B mainframe. For a data set with $n = 118$, $m = 7$, $k_j = 5$ ($j = 1, \ldots, m$), $p = 2$, all variables of nominal level, and for respectively $k = 3$ and $k = 15$, convergence occurs after about 0.09 respectively 0.32 seconds, excluding I/O operations. For ordinal levels, these figures are 0.13 and 0.48.

It is well known that the K-means algorithm does not guarantee the obtained allocation to be globally optimal. Using the above dataset with $k = 3$ and ordinal levels, we found 5 different solutions in 100 testruns. The losses are: 1.232 (38), 1.235 (39), 1.237 (5), 1.243 (17) and 1.258 (1). The bracketed figures indicate the frequency of the solutions. Assuming that 1.232 represents the globally optimal solution, the average number of misclassifications was found to be about 5% of the number of objects. The category quantifications were only slightly different in the 5 solutions.

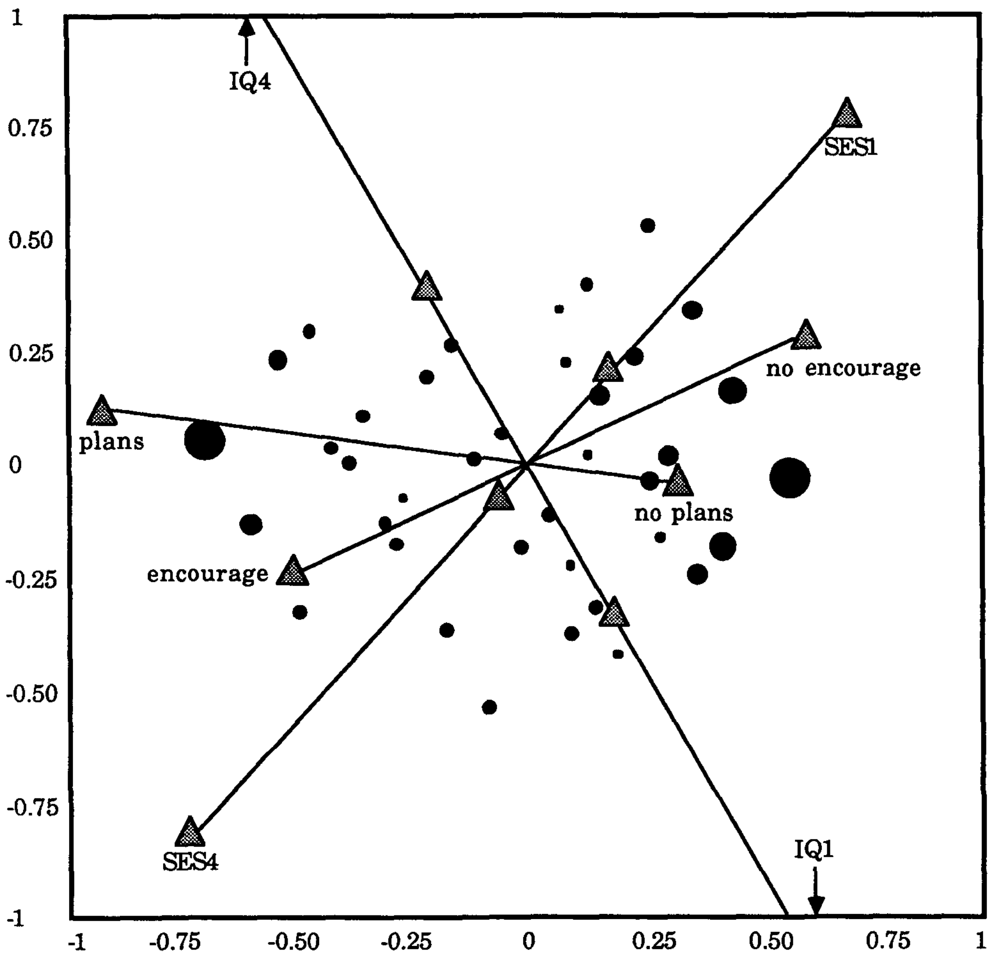

FIGURE 1.
Joint plot of the unrestricted solution.

Example

The method was applied to a subset of the data given by Fienberg (1980, p. 130) on 10318 high school seniors. We selected 98 cases by rounding off the entire frequency table divided by 100, and used four variables: intelligence (4 ordered categories), presence of college plans (2 categories), presence of parental encouragement (2 categories), and social economic status (4 ordered categories).

As to the choice of p and k, with $p < k$, two approaches are possible. In the first approach we choose p to be small, possibly aided by elbow or eigenvalue-greater-than-unity criteria, and we vary k over a number of interesting values. This approach is useful if one is interested in producing low-dimensional plots. In the other approach we try to use as much discriminatory information as possible by setting $p = k - 1$, provided that $p \leq \max(p)$. The maximum dimensionality is $\max(p) = p_1 + p_2$, where $p_1 = \Sigma\ (k_j - 1)$ for all variables with unrestricted Y_j and p_2 is the number of variables with a rank-one restriction on Y_j.

Using the first approach, the eigenvalues of the $p = 4$ solution for the *unrestricted* homogeneity analysis are 2.35, 0.74, 0.50, 0.41, so one- or two-dimensional solutions give reasonably accurate descriptions of the data. Figure 1 depicts the two-dimensional

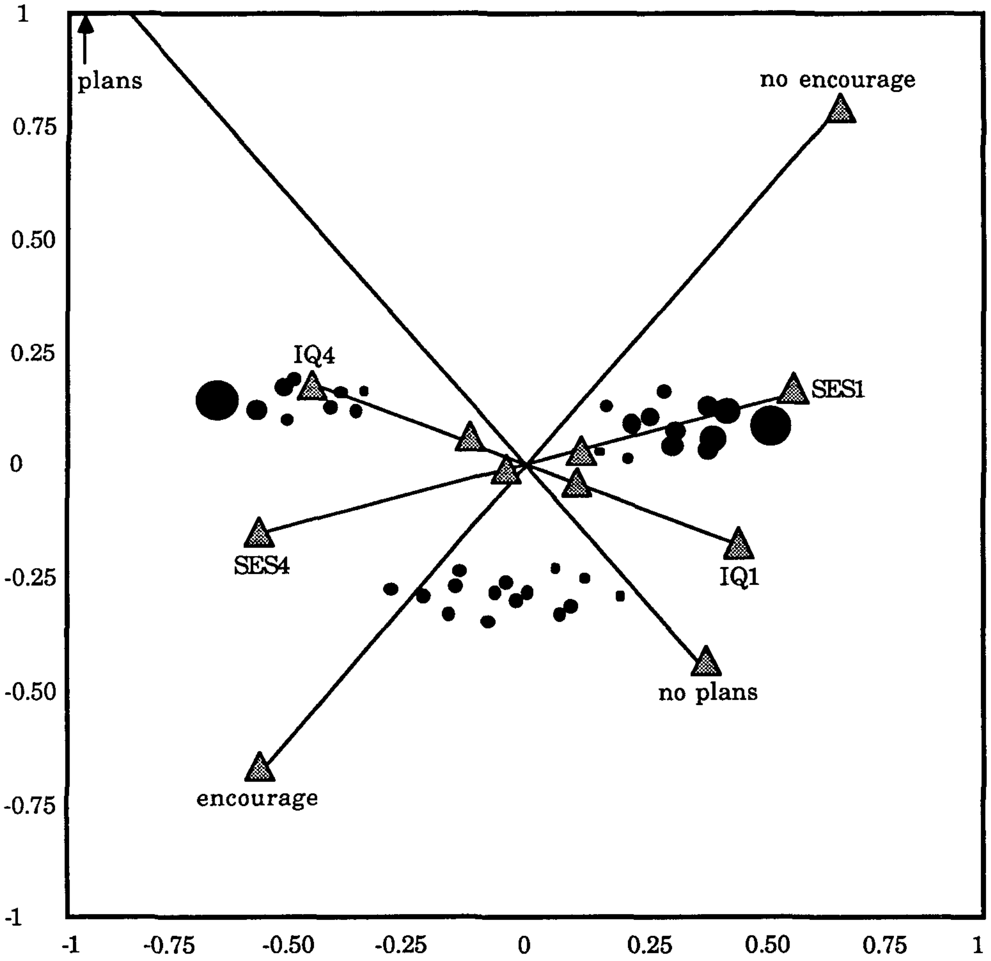

FIGURE 2.
Joint plot of the restricted solution, three groups.

solution. A dot indicates the position of an object, or profile, as given by the rows of X. The size of a dot corresponds with the number of objects on that location. We represent each optimally scaled category by a triangle. Because in this analysis a variable is either ordinal or binary, all categories of that variable are located on a line through the origin. The solution is normalized according to $\Sigma\, Y_j'G_j'G_jY_j = I$, and it satisfies the second centroid principle, that is, objects are located in the centroid of their category scores. The objects in Figure 1 form a bimodal cloud; in general, individuals on the left hand side are characterized by having college plans and high IQ scores, by obtaining parental encouragement, and by growing up in a moderate to high social economic environment. The reverse pattern can be found for persons on the right hand side of the plot. Deviations from these two dominant patterns make up the second dimension, where both IQ and SES account for the largest differences.

Suppose that we are interested in identifying a number of latent groups of objects from these data. Choosing $p = 2$ and $k = 3$ provides an attractive GROUPALS solution, with eigenvalues 2.09 and 0.46. Figure 2 shows the results. It shows three well separated and tight clusters. Going from left to right for Cluster 1, 2 and 3, the sizes are respectively 26, 27, and 45. It should be noted that all object points are plotted as if they were located in the category centroids (i.e., as given by Z in Step 3 of the algorithm), although their optimal positions as measured by (2) are the cluster means. The possibility to inspect Z, a low-dimensional continuous representation closest to the optimal cluster solution, is a major practical advantage of GROUPALS.

The main differences between the two solutions concern the second dimension. For the unrestricted solution, the variables IQ and SES contribute most to Dimension 2, but in the GROUPALS solution it is dominated by PLANS and ENCOURAGE. This demonstrates the fact that (nonmetric) PCA and clustering procedures may yield quite different results in terms of which variables are dominant in the reduced space. A closer inspection of the groups reveals that Cluster 1 is completely identified by the categories plans and encourage, Cluster 2 by no plans and encourage, and Cluster 3 by no plans and no encourage. Figure 2 nicely illustrates this if we project each group on either the ENCOURAGE or the PLANS axes. It turns out that the data set does not contain profiles with combined scores on plans and no encourage, so IQ and SES account for the entire within-groups variances. It is unlikely that we will find the same, optimal partitioning if we use the unrestricted object configuration as the starting point for a K-means analysis.

The resulting partition defines a latent categorical variable with k categories. This variable may be used in subsequent analyses, for example as in loglinear or discriminant analysis, and the optimally scaled categories may aid in its interpretation.

Discussion

Starting from homogeneity analysis, restricting objects to be located at one of k cluster points leads to a sum of squared distances criterion for estimating the unknown group allocations. If all variables are nominal, and if we replace the unknown group vector $\mathbf{c}$ by an observed variable, and then skip Step 4 of the algorithm, the solution becomes equivalent to the forced classification procedure of Nishisato (1984), in which one variable is made dominant by weighting. Thus GROUPALS can also be viewed as a generalization of forced classification to the case of mixed variables.

A problem of the current program is that it is likely to produce local optimal solutions, a property inherited from the combinatorial nature of the K-means algorithm. As a temporary fix, the program has an option for rapidly generating a large number of solutions, each beginning from a different starting partition. A more substantial alternative is to use mathematical programming techniques for finding the global optimum. Some work has been done in this area (Arthanari & Dodge, 1981; Littschwager & Wang, 1978), but we do not know whether these approaches are computationally feasible for the present problem. In practice, it appears that the locally optimal partitions do not differ to a great extent from the globally optimal one with respect to the obtained quantifications, component loadings and cluster means.

The present approach can be generalized in several ways. Missing data may be dealt with quite easily along the same lines as in homogeneity analysis, or by employing the K-means algorithm to estimate missing scores. It is also possible to extend the method to fuzzy clustering by dropping the restriction that G_c should be binary. However, it appears (Fisher, 1958; Gordon & Henderson, 1977) that the optimal fuzzy partition is necessarily mutually exclusive, so such an extension would require additional changes in the loss function, as in Bezdek (1981). Another generalization is to allow for spline transformations of the variables (Winsberg & Ramsay, 1982). This would make the method slightly more complicated, but on the other hand, such a procedure would not force the user to discard possibly relevant information by some discretization process. The method may also be generalized to problems with a partitioning of the variables into sets, and to problems with a constrained partition or with constrained cluster means, by introducing restrictions on respectively Y_j, G_c and Y.

References

Arthanari, T. S., & Dodge, Y. (1981). *Mathematical programming in statistics*. New York: Wiley.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, *65*, 31–38.

Bock, H. H. (1972). Statistische Modelle und Bayesische Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren [Statistical models and Bayesian problems for estimating an unknown classification of normally distributed random vectors]. *Metrika*, *18*, 120–132.

de Leeuw, J. (1984). The Gifi system of nonlinear multivariate analysis. In E. Diday, M. Jambu, L. Lebart, J. Pagès, & R. Tomassone (Eds.), *Data analysis and informatics III* (pp. 415–424). Amsterdam: North-Holland.

DeSarbo, W. S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, *49*, 57–78.

De Soete, G., DeSarbo, W. S., & Carroll, J. D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. *Journal of Classification*, *2*, 173–192.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, *53*, 789–798.

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

Gifi, A. (1981). *Nonlinear multivariate analysis*. Leiden: University of Leiden, Department of Data Theory.

Gordon, A. D. (1981). *Classification*. London: Chapman and Hall.

Gordon, A. D., & Henderson, J. T. (1977). An algorithm for Euclidean sum of squares classification. *Biometrics*, *33*, 355–362.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–872.

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

Lance, G. N., & Williams, W. T. (1967). Mixed data classificatory programs: I. Agglomerative systems. *Australian Computer Journal*, *1*, 15–20.

Lance, G. N., & Williams, W. T. (1968). Mixed data classificatory programs: II. Divisive systems. *Australian Computer Journal*, *1*, 82–85.

Littschwager, J. M., & Wong, C. (1978). Integer programming solution of a classification problem. *Management Science*, *24*, 151–165.

McCutcheon, A. L. (1987). *Latent class analysis*, Beverly Hills: Sage.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation of fifteen clustering algorithms. *Psychometrika*, *45*, 325–342.

Nishisato, S. (1984). Forced classification: A simple application of a quantification method. *Psychometrika*, *49*, 25–36.

Opitz, O. (1980). *Numerische taxonomie*. Stuttgart: Gustav Fisher.

Scheibler, D., & Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms. *Multivariate Behavioral Research*, *20*, 283–304.

Späth, H. (1985). *Cluster dissection and analysis*. Chichester: Ellis Horwood.

van der Burg, E., & de Leeuw, J. (1983). Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, *36*, 54–80.

van Rijckevorsel, J. L. A., & de Leeuw, J. (Eds.) (1988). *Component and correspondence analysis*. Chichester: Wiley.

Winsberg, S., & Ramsay, J. O. (1982). Monotone splines: A family of functions useful for data analysis. In H. Caussinus, P. Ettinger, & R. Tomassone (Eds.), *COMPSTAT 1982, Proceedings in Computational Statistics*, Vienna: Physica Verlag.

Wishart, D. (1969). Mode analysis. In A. J. Cole (Ed.), *Numerical taxonomy*. London: Academic Press.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, *46*, 357–388.

Manuscript received 5/15/87
Final version received 9/8/88