Journal of Classification 11:155-170 (1994)

An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models

Anil Chaturvedi
AT&T Bell Laboratories

J. Douglas Carroll
Rutgers University

Abstract: This paper presents a general approach for fitting the ADCLUS (Shepard and Arabie 1979; Arabie, Carroll, DeSarbo, and Wind 1981), INDCLUS (Carroll and Arabie 1983), and potentially a special case of the GENNCLUS (DeSarbo 1982) models. The proposed approach, based largely on a separability property observed for the least squares loss function being optimized, offers increased efficiency and other advantages over existing approaches like MAPCLUS (Arabie and Carroll 1980) for fitting the ADCLUS model, and the INDCLUS method for fitting the INDCLUS model. The new procedure (called "SINDCLUS") is applied to three sets of empirical data to demonstrate the effectiveness of the SINDCLUS methodology. Finally, some potentially useful extensions are discussed.

Keywords: Overlapping clustering; Cluster analysis; ADCLUS; INDCLUS; MAPCLUS; Separability.

Authors' Addresses: Anil Chaturvedi, Room 5A-127, AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ, 07974-0636; J. Douglas Carroll, Rutgers University, Faculty of Management, 180 University Avenue, Newark, NJ, 07102-1895.

1. Introduction

Shepard and Arabie (1979) first introduced ADCLUS as a model for overlapping clustering of one-mode, two-way data.[1] Subsequently, Arabie and Carroll (1980) provided a mathematical programming algorithm for fitting the ADCLUS model, called MAPCLUS. Carroll and Arabie (1983) proposed an individual differences generalization of the ADCLUS model, called the INDCLUS model, and also devised a generalization of the MAPCLUS procedure for fitting this three-way generalization of the ADCLUS model — called the INDCLUS *method*, thus providing a methodology for overlapping clustering of two-mode, three-way data. While the theoretical significance and conceptual elegance of overlapping clustering can be applied in many different substantive domains, as illustrated, for example, in Arabie, Carroll, DeSarbo, and Wind (1981) and Srivastava, Alpert, and Shocker (1984), widespread acceptance of overlapping clustering has been restricted because the existing algorithms — particularly MAPCLUS and INDCLUS, the most successful and widely used to date for fitting these models, cannot easily handle large data sets because the programs are extremely computationally intensive and therefore expensive to run.

In this paper, we present the SINDCLUS (Separability-based, Speedy INDCLUS) procedure for fitting both the ADCLUS and the INDCLUS models. SINDCLUS uses an iterative combinatorial optimization procedure for fitting both the ADCLUS and INDCLUS models, and as its name suggests, is considerably faster than the respective two- and three-way procedures, MAPCLUS and INDCLUS.

2. SINDCLUS: (Separability-based) Speedy INDCLUS

In this section, we describe the general approach of SINDCLUS for fitting the ADCLUS and the INDCLUS models.

---

1. Number of "ways" of a data array is the number of indices required to index a general entry in that array while the number of "modes" is the number of distinct classes of entities represented in the array. Thus, an objects × objects similarity matrix has two ways, i.e., rows and columns of the matrix, but only one mode, i.e., objects. An objects × objects × subjects array of data constitutes three-way, but two-mode data.

2.1. Description of the INDCLUS model

Assume that N objects are being clustered into R possibly overlapping clusters. The INDCLUS model (Carroll 1975, Carroll and Arabie 1983) is written as:

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{P}' + \mathbf{C}_k + \text{error}, \tag{1}$$

where:

$\mathbf{S}_k$ is an $(N \times N)$ similarity matrix for the k-th subject (or other source of data); $k = 1, \ldots, K$,

$\mathbf{W}_k$ is an $(R \times R)$ diagonal matrix of weights for the k-th subject (or other source of data); $k = 1, \ldots, K$,

$\mathbf{P}$ is an $(N \times R)$ *binary* indicator matrix defining the possibly overlapping clusters, and

$\mathbf{C}_k$ is an $(N \times N)$ matrix, all of whose entries are c_k, which can be thought of as a weight for a universal cluster denoted by an $N \times 1$ unit vector $\mathbf{1}$, all of whose components are 1.

The diagonal entries in the $(N \times N)$ matrix $\mathbf{S}_k$ are not defined. The estimation problem in INDCLUS is to determine the ordinary least squares (OLS) estimates of parameters $\mathbf{P}$, $\mathbf{W}_k$ and $\mathbf{C}_k$. The diagonal elements of the weight matrices $\mathbf{W}_k$ must be non-negative, and elements of $\mathbf{P}$ must be constrained to either 0 or 1. This is a nonlinear, mixed 0-1 integer programming problem.

Various techniques have been developed for fitting the ADCLUS and INDCLUS models, such as the penalty function approaches of the MAPCLUS and INDCLUS algorithms (Arabie and Carroll 1980; Carroll and Arabie 1983), the GENNCLUS approach of DeSarbo (1982), the Maximum Likelihood approach of Hiroshi Hojo (1983), and the Qualitative Factor Analysis (QFA) procedure of Mirkin (1987, 1989, 1990). Unlike the aforementioned procedures, the SINDCLUS procedure uses a property of separability in the INDCLUS model for determining the OLS estimates of the cluster membership parameters. This property of separability is described in the elementary binary least squares procedures below.

2.2. The Elementary Binary Least Squares Procedure

Consider the following illustrative problem of finding the least squares estimate of $\mathbf{x}$, where $\mathbf{L} = \mathbf{x}\mathbf{r}'$ + error, and $\mathbf{x}$ is a binary vector whose components are either 0 or 1:

$$\mathbf{L} = \begin{bmatrix} 7\,5\,3\,9 \\ 8\,6\,5\,1 \\ 9\,4\,2\,7 \\ 5\,3\,4\,6 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \mathbf{r}' = [1\,4\,9\,3]\,.$$

The problem can also be stated as

$$\begin{bmatrix} 7\,5\,3\,9 \\ 8\,6\,5\,1 \\ 9\,4\,2\,7 \\ 5\,3\,4\,6 \end{bmatrix} = \begin{bmatrix} 1x_1 & 4x_1 & 9x_1 & 3x_1 \\ 1x_2 & 4x_2 & 9x_2 & 3x_2 \\ 1x_3 & 4x_3 & 9x_3 & 3x_3 \\ 1x_4 & 4x_4 & 9x_4 & 3x_4 \end{bmatrix} + \text{error}\,.$$

If we let

$$f_1 = [7 - 1x_1]^2 + [5 - 4x_1]^2 + [3 - 9x_1]^2 + [9 - 3x_1]^2\ ,$$

$$f_2 = [8 - 1x_2]^2 + [6 - 4x_2]^2 + [5 - 9x_2]^2 + [1 - 3x_2]^2\ ,$$

$$f_3 = [9 - 1x_3]^2 + [4 - 4x_3]^2 + [2 - 9x_3]^2 + [7 - 3x_3]^2$$

and,

$$f_4 = [5 - 1x_4]^2 + [3 - 4x_4]^2 + [4 - 9x_4]^2 + [6 - 3x_4]^2\ ,$$

then the total sum of squared errors is given by

$$F = f_1 + f_2 + f_3 + f_4\,.$$

Note that f_1 is a function only of x_1; f_2 is a function only of x_2, etc. Thus, F is *separable* in x_1, x_2, x_3, and x_4. To minimize F, one can separately minimize f_1 with respect to x_1, f_2 w.r.t x_2, f_3 w.r.t x_3, and, f_4 w.r.t x_4. To minimize, say, f_1 w.r.t. x_1, one can easily evaluate f_1 at $x_1 = 1$ and $x_1 = 0$. The value of x_1 yielding a minimum of these two possible values is then chosen. Thus, for N (0-1) variables, only $2N$ function evaluations and N comparisons are needed, as compared to 2^N evaluations and comparisons for explicit enumeration. We use this elementary binary least squares procedure iteratively in estimating the ADCLUS and the INDCLUS models. It might be noted, in passing, that the separability property will hold if the elements of $\mathbf{x}$ were restricted to any finite set of integers, or even a finite set of arbitrary real numbers. This fact, could, if desired, be used to fit models entailing multivalued logic, or models in which coordinate values are restricted to a prespecified grid, as in quadratic assignment problems.

2.3. The SINDCLUS algorithm

The estimation problem in INDCLUS is to find the least squares estimates of $\mathbf{P}$, $\mathbf{W}_k$, and $\mathbf{C}_k (k = 1, \ldots, K)$ in the equation:

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{P}' + \mathbf{C}_k + \text{error}, \tag{2}$$

where $\mathbf{P}$, $\mathbf{W}_k$, and $\mathbf{C}_k$, are as defined in equation (1). If $\mathbf{S}_k$ were allowed to be a nonsymmetric $(N \times N)$ matrix (as in DeSarbo 1982), then the above equation can be generalized to

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{Q}' + \mathbf{C}_k + \text{error}, \tag{3}$$

where $\mathbf{Q}$ is an $(N \times R)$ matrix. In the *symmetric* case of INDCLUS, $\mathbf{P} = \mathbf{Q}$.

We use (3) to estimate the parameters for the symmetric case *without* imposing the constraint $\mathbf{P} = \mathbf{Q}$. Thus, we will first define the estimation problem for the nonsymmetric case, and then reduce it for the symmetric case. Defining the following notation:

$\mathbf{p}_i \equiv (N \times 1)$ binary vector for the i-th cluster,

$\mathbf{w}_i \equiv (K \times 1)$ vector of the weights for the i-th cluster,

$\mathbf{q}_i \equiv (N \times 1)$ binary vector for the i-th cluster (not necessarily $= \mathbf{p}_i$),

$\mathbf{P}_{(-i)} \equiv (N \times R)$ binary matrix including the universal cluster $\mathbf{1}$ but excluding the i-th cluster,

$\mathbf{W}_{k(-i)} \equiv (R \times R)$ weight matrix for the k-th subject or other source of data, including the weight for the universal cluster but excluding the weight for the i-th cluster, and

$\mathbf{Q}_{(-i)} \equiv (N \times R)$ matrix including the universal cluster $\mathbf{1}$ but excluding the i-th cluster.

We can rewrite (3) as

$$\mathbf{S}_k = \mathbf{p}_i w_{ki} \mathbf{q}_i' + \mathbf{P}_{(-i)}\mathbf{W}_{k(-i)}\mathbf{Q}'_{(-i)} + \text{error}. \tag{4}$$

If we have estimates of all but the i-th cluster, then we can define $\bar{\mathbf{S}}_k$ as

$$\bar{\mathbf{S}}_k = \mathbf{S}_k - \mathbf{P}_{(-i)}\mathbf{W}_{k(-i)}\mathbf{Q}'_{(-i)} \tag{5}$$

to get

$$\bar{\mathbf{S}}_k = \mathbf{p}_i w_{ki} \mathbf{q}_i' + \text{error}; \tag{6}$$

that is,

$$\bar{\mathbf{S}}_k = f(\textit{parameters} \text{ for } \textit{cluster } i) + \text{error}. \tag{7}$$

If we have K matrices $\mathbf{S}_k$ of order $(N \times N)$, we can use a procedure similar to the CANDECOMP-based algorithm called the INDSCAL method for fitting the INDSCAL model, formulated by Carroll and Chang (1970). Let us assume that we have the parameter estimates for all clusters, except for the i-th cluster. Let $\mathbf{T}_1$ be a $(K \times N^2)$ matrix where the k-th row has all the N^2 terms of the $N \times N$ matrix $\bar{\mathbf{S}}_k$, and $\mathbf{T}_2$ and $\mathbf{T}_3$ be $(N \times KN$ matrices). $\mathbf{T}_1$ has all

N^2 elements of $\bar{\mathbf{S}}_k$ in its k-th row. $\mathbf{T}_2$ is the supermatrix that has the j-th row of matrices $\mathbf{S}_1, \ldots, \mathbf{S}_k$ in the j-th row. Thus:

$$\mathbf{T}_2 = [\bar{\mathbf{S}}_1 \mid \bar{\mathbf{S}}_2 \mid \cdots \bar{\mathbf{S}}_k \mid \cdots \mid \bar{\mathbf{S}}_k] .$$

Similarly,

$$\mathbf{T}_3 = [\bar{\mathbf{S}}_1' \mid \bar{\mathbf{S}}_2' \mid \cdots \bar{\mathbf{S}}_k' \mid \cdots \mid \bar{\mathbf{S}}_K'] .$$

Assuming that estimates of $\mathbf{p}_i$ and $\mathbf{q}_i$ are known, the parameters for the i-th cluster are estimated by iterating the following three steps until at least a *local* optimum is reached.

2.3.1. Estimating $\mathbf{w}_i$ *conditionally:* Given current estimates, $\hat{\mathbf{p}}_i$ and $\hat{\mathbf{q}}_i$, of $\mathbf{p}_i$ and $\mathbf{q}_i$, let $\mathbf{g}_i$ be a vector of N^2 elements such that

$$\mathbf{g}_i = \hat{\mathbf{p}}_i \otimes \hat{\mathbf{q}}_i ,$$

where $\otimes$ is the Kronecker product. Then, using OLS regression to solve for the OLS estimate $\hat{\mathbf{w}}_i$ in $\mathbf{T}_1 = \mathbf{w}_i\mathbf{g}_i' +$ error, yields a closed form solution for $\hat{\mathbf{w}}_i$. Nonnegativity constraints are imposed easily by simply setting all negative weights to zero, as pointed out in the context of nonnegatively constrained (weighted) CANDECOMP, estimated by a "one-dimension-at-a-time" scheme analogous to the "one-cluster-at-a-time" approach utilized here, by Carroll, DeSoete, and Pruzansky (1989).

2.3.2. Estimating $\mathbf{p}_i$ *conditionally:* Given current estimates, $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{q}}_i$, of $\mathbf{w}_i$ and $\mathbf{q}_i$, let $\mathbf{h}_i$ be a vector of KN elements such that

$$\mathbf{h}_i = \hat{\mathbf{w}}_i \otimes \hat{\mathbf{q}}_i .$$

Then, by using the elementary binary least squares procedure in the equation $\mathbf{T}_2 = \mathbf{p}_i\mathbf{h}_i' +$ error, one can find $\hat{\mathbf{p}}_i$, the OLS estimates of $\mathbf{p}_i$.

2.3.3. Estimating $\mathbf{q}_i$ *conditionally:* Given current estimates, $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{p}}_i$, of $\mathbf{w}_i$ and $\mathbf{p}_i$, let $\mathbf{j}_i$ be a vector of KN elements such that $\mathbf{j}_i = \hat{\mathbf{w}}_i \otimes \hat{\mathbf{p}}_i$. Then, by using the elementary binary least squares procedure in the equation $\mathbf{T}_3 = \mathbf{q}_i\mathbf{j}_i' +$ error, one can find $\hat{\mathbf{q}}_i$, the OLS estimates of $\mathbf{q}_i$.

Use $\hat{\mathbf{q}}_i$ and repeat Steps [2.3.1], [2.3.2], and [2.3.3] until *no* improvement in fit results. In Steps [2.3.2] and [2.3.3] above, the $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ vectors cannot be all zero. Thus, each cluster must have at least one object in it. Since the matrices $\mathbf{S}_k$ do not have diagonals in the case of the INDCLUS model, Steps [2.3.1], [2.3.2], and [2.3.3] above need to be modified. We simply drop the corresponding columns from the $\mathbf{T}_1$ matrix and the $\mathbf{g}$ vector in Step

[2.3.1]. Similarly, we do not consider the diagonal elements in Steps [2.3.2] and [2.3.3] for fitting the INDCLUS model. For estimating the weights of the universal cluster, for which **p** and **q** are fixed, we just use Step [2.3.1], with $\mathbf{p} = \mathbf{q} = \mathbf{1}$.

While we do not impose the constraint that $\mathbf{P} = \mathbf{Q}$ in the estimation procedure, we have thus far found empirically that in the symmetric case of INDCLUS, the matrices **P** and **Q** are equal upon convergence at the global optimum.[2] Also, while the default option in SINDCLUS is to estimate the INDCLUS model by treating the diagonals as missing data, SINDCLUS also allows fitting of the INDCLUS model when diagonals are treated as non-missing. In general, one major advantage of SINDCLUS in addition to its greater computational efficiency is its ability to handle arbitrary patterns of missing data satisfactorily.

3. Empirical Applications of SINDCLUS

The SINDCLUS procedure was applied to three different data sets: The kinship data of Rosenberg and Kim (1975), as published in Arabie, Carroll, and DeSarbo (1987, pp. 62-63); Henley's (1969) data on perceived similarities of animals, and data on soft drinks (Chaturvedi 1993).

3.1. Application to Kinship Data of Rosenberg and Kim (1975)

Arabie, Carroll, and DeSarbo (1987) present an application of the INDCLUS model and method to the kinship data of Rosenberg and Kim (1975). We used the same data for analysis using the SINDCLUS procedure.

The fifteen most commonly used kinship terms — Aunt, Brother, Cousin, Daughter, Father, Granddaughter, Grandfather, Grandmother, Grandson, Mother, Nephew, Niece, Sister, Son, and Uncle, were printed on slips of paper for use in a sorting task (see Rosenberg 1982) by Rosenberg and Kim (1975). Eighty-five male and eighty-five female subjects were run in a condition where subjects gave (only) a single-sort of the fifteen terms. A different group of subjects (eighty males and eighty females) were told that after making their first sorts of the terms, they should give additional subjective partitioning(s) of these stimuli using "a different basis of meaning each

2. While in general, there will be many local optima for this problem, we conjecture that among the global optima, at least one will be symmetric, i.e., $P = Q$; in most cases, we conjecture there will be a single global optimum which is symmetric. While our empirical research strongly supports this proposition, we do not yet have a formal proof.

time." Rosenberg and Kim (1975) used only the data from the first and second sortings for this group of subjects. Thus, we have six *conditions* which will assume the role of our "subjects": females' single-sort, males' single-sort, females' first-sort, males' first-sort, females' second-sort, and males' second-sort. Again note that the subjects in the first two conditions were distinct from the subjects in the last four conditions.

Since the subjects' partitions of the stimuli comprise nominal scale data that do not immediately assume the form of a proximity matrix, some preprocessing is necessary to obtain such a matrix. If we form a stimulus × stimulus co-occurrence matrix for each experimental condition, with the (i,j)-th entry derived as the number of subjects who placed stimuli i and j in the same group, and subtract that entry from the total number of subjects contributing to the matrix, then we have what is called the S-measure (Arabie et al. 1987). As in Arabie et al. (1987), the six matrices constructed using the S-measure were analyzed using SINDCLUS via a matrix unconditional approach. A five-cluster solution explaining 81.1 percent variance was extracted. This optimal solution derived using the SINDCLUS procedure is identical to the solution presented by Arabie et al.(1987, pp. 61-64). The five-cluster solution and the importance weights are presented in Table 1.

The clusters are easily interpreted. In the order listed, the first two are sex-defined, the third is the collateral relatives, the fourth is the nuclear family, while the fifth consists of grandparents and grandchildren. The pattern of weights also offers an interesting interpretation. The statement of Rosenberg and Kim (1975) that subjects restricted to a single-sort ignore sex as a basis of organization is strongly supported by the relatively low weights for the sex-defined clusters in the first two columns (especially for female subjects) in Table 1. For the multiple-sort conditions, it is interesting to note that female subjects emphasized sex in the first sorting (given that the two relevant clusters have much higher weights), whereas male subjects waited until the second sorting to emphasize the salience of sex as a factor in sorting the kinship terms. Across all conditions, females' data were better fitted to the model than were males' data. Also, data from the first sort were better fitted than for the second sort for both females and males.

3.2. Application to Henley's Data

Dissimilarity ratings on all 435 distinct pairs of thirty animals — antelope, bear, beaver, camel, cat, chimpanzee, chipmunk, cow, deer, dog, donkey, elephant, fox, giraffe, goat, gorilla, horse, leopard, lion, monkey, mouse, pig, rabbit, raccoon, rat, sheep, squirrel, tiger, wolf, and zebra were gathered from 21 subjects by Henley (1969). At a second session, one week later, the subjects made the judgements again, this time with the pairs in the

Table 1 - SINDCLUS Solution for Rosenberg and Kim's (1975) data

Females' Single-Sort	Males' Single-Sort	Females' First Sort	Females' Second Sort	Males' First Sort	Males' Second Sort	Elements of Subset	Interpretation
.052	.143	.551	.241	.299	.295	Brother, father, grandfather, grandson, nephew, son, uncle	Male relatives, excluding cousin
.049	.146	.554	.246	.291	.306	Aunt, daughter, granddaughter, grandmother, mother, niece, sister	Female relatives, excluding cousin
.552	.397	.283	.373	.340	.237	Aunt, cousin, nephew, niece, uncle	Collateral (Romney & D'Andrade 1964) relatives
.478	.372	.206	.322	.241	.219	Brother, daughter, father, mother, sister, son	Nuclear family
.626	.449	.251	.385	.395	.253	Granddaughter, grandfather, grandmother, grandson	Direct ancestors and descendants 2 generations removed
.055	.075	.132	.158	.158	.207	Additive constants	
78.6%	68.8%	96.3%	78.9%	82.4%	71.7%	Variance accounted within condition	Overall VAF = 81.1%

Table 2 - VAF for Henley's Data

# Clusters	VAF
1	0.1946
2	0.3899
3	0.5604
4	0.6622
5	0.7346
6	0.7817

opposite (within-pair) order from that of the first session. A within-subject reliability was obtained from the correlation of first and second judgments.

Individual differences among the subjects were examined before their mean dissimilarity matrices were averaged for scaling, by a procedure similar to the Tucker and Messick (1963) points-of-view method, followed by the application of Johnson's (1967) complete-link clustering technique. The subjects were found to divide into three groups composed of eighteen, two, and one subjects. The dissimilarity matrices from all the eighteen subjects in the first group were averaged to give a single two-way mean dissimilarity matrix.[3] This mean dissimilarity matrix was then analyzed using the SINDCLUS procedure. Table 2 presents a summary of the variance accounted for by SINDCLUS. We chose the three- and the five-cluster solutions for reasons of interpretability. Tables 3 and 4 present them.

3.2.1. The three-cluster solution: The three-cluster solution accounted for 56.04% of the variance. The first cluster comprised of rodents and other burrowing mammals like beaver, chipmunk, mouse, rabbit, raccoon, rat, and squirrel. The second cluster, which included antelope, camel, cow, deer, donkey, giraffe, goat, horse, sheep, and zebra, can easily be interpreted as the cluster of ruminants and other ungulates. The third cluster includes the family of canines and felines cat, leopard, tiger, lion, dog, fox, and wolf.

3.2.2. The five-cluster solution: The five-cluster solution yielded a Variance Accounted For (VAF) of 73.46%. Three of the five clusters are identical to the clusters of the three-cluster solution, i.e., the rodents and other burrowing mammals cluster, the ruminants and other ungulates cluster, and the canines and felines cluster. The other two clusters were also interpretable. The fourth cluster comprises all the primates — chimpanzee, gorilla, and monkey. The fifth cluster includes antelope, beaver, cat, chimpanzee, chipmunk, deer, dog, donkey, fox, goat, horse, monkey, mouse, mouse, pig, rabbit, raccoon, rat, sheep, squirrel, and zebra. This cluster consists of all animals except camel, elephant, giraffe, tiger, wolf, fox, gorilla, leopard, lion, and cow. With the exception of cow, this cluster consists of animals that are neither big (unlike elephant, camel and giraffe) nor perceived to be dangerous (unlike tiger, wolf, fox, gorilla, leopard, and lion).

---

3. We thank Phipps Arabie for providing us with the Henley data for analysis via SINDCLUS.

Table 3: Three-cluster SINDCLUS solution for Henley's data

Cluster	Items in Cluster	Interpretation	Weight
a	Beaver, chipmunk, mouse, rabbit, raccoon, rat, squirrel	Rodents and other burrowing mammals	0.47
b	Antelope, camel, cow, deer, donkey, giraffe, goat, horse, sheep, zebra	Ruminants and other ungulates	0.31
c	Cat, dog, fox, leopard, lion, tiger, wolf	Canines and felines	0.39
d	All objects	Universal cluster	0.25

Table 4: Five-cluster SINDCLUS solution for Henley's data

Cluster	Items in Cluster	Interpretation	Weight
a	Beaver, chipmunk, mouse, rabbit, raccoon, rat, squirrel	Rodents and other burrowing mammals	0.40
b	Antelope, camel, cow, deer, donkey, giraffe, goat, horse, sheep, zebra	Ruminants and other ungulates	0.31
c	Cat, dog, fox, leopard, lion, tiger, wolf	Canines and felines	0.43
d	Antelope, beaver, cat, chimpanzee, chipmunk, deer, dog, donkey, fox, goat, horse, monkey, mouse, pig, rabbit, raccoon, rat, sheep, squirrel, zebra	Non-violent, medium/small sized mammals	0.13
e	Chimpanzee, gorilla, monkey	Primates	0.68
f	All objects	Universal cluster	0.19

Table 5 - VAF for soft drinks

# of clusters	VAF
One	23.83%
Two	48.95%
Three	61.90%
Four	68.23%
Five	71.75%
Six	73.44%
Seven	74.55%

Table 6 - SINDCLUS solution for soft drinks

Cluster	Items in Cluster	Interpretation
a	Coke, Diet Coke, Pepsi, Diet Pepsi, C&C Cola	Regular colas
b	Sprite, Diet Sprite, 7up, Diet 7up	Lemon/Lime
c	Orange Slice, Sunkist, Diet Sunkist	Orange
d	Coke, Diet Coke, Cherry Coke, Dr. Pepper, Pepsi, Diet Pepsi, C&C Cola	All colas
e	C&C Pineapple, Orange Slice, Sprite, Diet Sprite, Sunkist, Diet Sunkist, 7up, Diet 7up, Hawaiian Punch	All fruity beverages (non-colas)
f	Cherry Coke, Dr. Pepper, Hawaiian Punch	Fruity beverages

Table 7 - Subjects' weight matrix for soft drinks data

	a	*b*	*c*	*d*	*e*	*f*
Subject	**Colas**	**Lime**	**Orange**	**All colas**	**N-cola Fruit**	**Fruit**
1	.20	.33	.52	.03	.00	.00
2	.94	.96	.96	.02	.00	.65
3	.42	.79	.77	.00	.00	.00
4	.31	.61	.57	.27	.13	.32
5	.14	.59	.35	.36	.12	.08
6	.58	.82	.84	.02	.00	.01
7	.48	.80	.30	.31	.02	.25
8	.26	.28	.33	.37	.43	.24
9	.41	.79	.84	.11	.01	.00
10	.14	.38	.49	.18	.02	.00
11	.23	.22	.50	.38	.03	.07
12	.21	.57	.33	.18	.10	.00
13	.12	.57	.72	.47	.07	.03

3.3. Applications to soft drinks data

Chaturvedi (1993) presents proximity data on sixteen soft drinks. In the data collection task, pairwise dissimilarity ratings on sixteen soft drinks were collected from thirteen students at Rutgers University in 1992. The sixteen soft drinks chosen in the study were: Coke, Diet Coke, Cherry Coke, C&C Pineapple, Dr. Pepper, Pepsi, Diet Pepsi, Orange Slice, Sprite, Diet Sprite, Sunkist, Diet Sunkist, 7up, Diet 7up, C&C Cola, and Hawaiian Punch. The thirteen proximity matrices were then analyzed via the SINDCLUS procedure. A six-cluster solution was chosen for reasons of interpretability. Table 5 presents a summary of the VAF by SINDCLUS. Table 6 presents a summary description of the extracted clusters and their interpretation. Table 7 gives the weights of the thirteen subjects for the six clusters.

The six-cluster solution in Table 6 clearly separates the cola-based drinks (Cluster *d* corresponding to Coke, Diet Coke, Cherry Coke, Dr. Pepper, Pepsi, Diet Pepsi, and C&C Cola) from the non-cola based fruity beverages (Cluster *e* corresponding to C&C pineapple, orange Slice, Sprite, Diet Sprite, Sunkist, Diet Sunkist, 7up, Diet 7up, and Hawaiian Punch). The colas are further grouped into regular colas (without any fruit bases) corresponding to cluster a consisting of Coke, Diet Coke, Pepsi, Diet Pepsi, and C&C Cola.

The fruity beverages (Cluster *e*) has two distinct clusters wholly included in it: the lemon/lime based drinks like Sprite, Diet Sprite, 7up and Diet 7up corresponding to Cluster *b*, and the the orange drinks (Cluster *c*) like Orange Slice, Sunkist, and Diet Sunkist.

Cluster *f* has the fruity beverages like Cherry Coke, Dr. Pepper, and Hawaiian Punch. These drinks are close substitutes that compete strongly in the marketplace. While both Cherry Coke and Dr. Pepper are cola drinks, Hawaiian Punch is a non-carbonated drink. One might be surprised at the omission of C&C Cola from this set, but C&C Cola is a local brand not distributed in cans, vending machines, and convenience stores. It is only sold in bottles through supermarkets and grocery stores and has less than 0.4% market share. Thus, because of low awareness and availability in the market, it is not considered a threat to the fruity beverages.

4. Testing SINDCLUS With Error-Free Data

A study was designed to test the validity of the SINDCLUS procedure in attaining overall least-squares estimates of the ADCLUS and the INDCLUS models. Ten error-free random proximity matrices were generated with the number of objects (N) fixed at twelve, and the number of overlapping clusters (P) fixed at four. The two-way version of the SINDCLUS procedure was applied to these ten random data sets. Each data set was subjected to ten SINDCLUS runs, where the initial cluster configuration (matrices **P** and **Q** in (3)) for each of the ten runs was chosen randomly. The run accounting for the maximum variance was chosen as the final SINDCLUS solution.

For each of the 10 data sets, error-free global optima (VAF=100%) were obtained using the SINDCLUS procedure. The cluster memberships and the cluster weights were identical to the respective error-free cluster memberships and cluster weights that were used to generate the data. For each of the 10 data sets, even though the initial cluster configurations were chosen to be nonsymmetric (different **P** and **Q** matrices in (3)), the global optima were always symmetric, i.e., **P** was equal to **Q** in (3).

5. Conclusions

In this paper, we presented SINDCLUS — a new algorithm for fitting the ADCLUS and the INDCLUS models based on an approach analogous to the CANDECOMP procedure for fitting the INDSCAL model (Carroll and Chang 1970). SINDCLUS offers considerable savings in computer time over existing approaches like the MAPCLUS (Arabie and Carroll 1979) and INDCLUS (Carroll and Arabie 1983) algorithms for fitting the ADCLUS and the INDCLUS models, respectively. This advantage potentially expands the

range of applicability of the ADCLUS and the INDCLUS models to larger data sets. The SINDCLUS procedure can also be quite easily extended to fit some important generalizations of the ADCLUS and INDCLUS models.

Since the SINDCLUS procedure does not impose the constraint that $\mathbf{P} = \mathbf{Q}$ in (3), it can also be used to fit a special case of the GENNCLUS model. Because SINDCLUS utilizes a numerical procedure very similar to the Carroll and Chang (1970) CANDECOMP method used for fitting the INDSCAL model for three-way multidimensional scaling (and for multi-way components analysis), which we call "CANDCLUS," SINDCLUS, and the CANDCLUS model/method underlying it can be generalized to fit any general multi-way, multi-mode model for N-way and $M(\leq N)$-mode data, allowing binary parameters for any p of the M modes, and continuous parameters for the remaining $M - p$. The separability property described in Section 2.2 can be extended to allow straightforward conditional OLS estimation for each mode modeled by binary parameters. As a special case of this general model, the SINDCLUS/CANDCLUS approach can be used to fit a model in which all M modes are modeled via overlapping cluster structures.

References

ARABIE, P., and CARROLL, J. D. (1980), "MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model," *Psychometrika, 45*, 211-235.

ARABIE, P., CARROLL, J. D., and DESARBO, W. S. (1987), "Three-way Scaling and Clustering," Newbury Park, CA: Sage.

ARABIE, P., CARROLL, J. D., DESARBO, W. S., and WIND, J. (1981), "Overlapping Clustering: A New Method for Product Positioning," *Journal of Marketing Research, 18*, 310-17.

CARROLL, J. D., (1975), "Models for Individual Differences in Similarities," Paper presented at the Eighth Annual Mathematical Psychology Meeting, Purdue University, West Lafayette, IN.

CARROLL, J. D., and ARABIE, P. (1983), "An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm," *Psychometrika, 48*, 157-169.

CARROLL, J. D., and CHANG J. J. (1970), "Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of 'Eckart-Young' Decomposition," *Psychometrika*, 283-319.

CARROLL, J. D., DESOETE, G., and PRUZANSKY, S. (1989), "Fitting of the Latest Class Model via Iteratively Reweighted Least Squares CANDECOMP with Nonnegativity Constraints," in *Multiway Data Analysis*, Eds., R. Coppi and S. Bolasco, Amsterdam: North Holland, 463-472.

CHATURVEDI, A. D. (1993), "Perceived Product Uniqueness in Product Differentiation and Product Choice," *9316296*, Ann Arbor, MI: University Microfilms International.

DESARBO, W. S. (1982), "GENNCLUS: New Models for General Nonhierarchical Clustering Analysis," *Psychometrika, 47*, 449-475.

HENLEY, N. M. (1969), "A Psychological Study of the Semantics of Animal Terms," *Journal of Verbal Learning and Verbal Behavior, 8*, 176-184.

HOJO, H. (1983), "A Maximum Likelihood Method for Additive Clustering and its Applications," *Japanese Psychological Research, 25*, 4, 191-201.

JOHNSON, S. C. (1967), "Heirarchical Clustering Schemes," *Psychometrika, 32*, 241-254.

MIRKIN, B. G. (1987), "Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices," *Journal of Classification, 4*, 7-31.

MIRKIN, B. G. (1989), "Erratum," *Journal of Classification, 6*, 271-272.

MIRKIN, B. G. (1990), "A Sequential Fitting Procedure for Linear Data Analysis Models," *Journal of Classification, 7*, 167-195.

ROSENBERG, S., (1982), "The Method of Sorting in Multivariate Research with Applications Selected from Cognitive Psychology and Person Perception," in *Multivariate Applications in the Social Sciences*, Eds., N. Hirschberg and L. Humphreys, Hillsdale, NJ: Erlbaum, 117-142.

ROSENBERG, S., and KIM, M. P. (1975), "The Method of Sorting as a Data-gathering Procedure in Multivariate Research," *Multivariate Behavioral Research, 10*, 489-502.

SHEPARD, R. N., and ARABIE, P. (1979), "Additive Clustering Representation of Similarities as Combinations of Discrete Overlapping Properties," *Psychological Review, 86*, 87-123.

SRIVASTAVA, R. K., ALPERT, M. I., and SHOCKER, A. D. (1984), "A Customer-Oriented Approach for Determining Market Structures," *Journal of Marketing, 48*, 32-45.

TUCKER, L. R, and MESSICK, S. (1963), "An Individual Differences Model for Multidimensional Scaling," *Psychometrika, 28*, 333-367.