Journal of Classification 18: 35-55 (2001)
DOI: 10.1007/s00357-001-0004-3

K-modes Clustering

Anil Chaturvedi

Kraft Foods

Paul E. Green

The Wharton School, University of Pennsylvania

J. Douglas Carroll

Rutgers University

Authors' Addresses: Anil Chaturvedi, GV529, 1 Kraft Court, Glenview, IL 60025, USA; Paul E. Green, Suite 1450, Dietrich Hall, Philadelphia, PA, 19104, USA; J. Douglas Carroll, Rutgers University, Graduate School of Management, MEC 125, 111 Washington Street, Newark NJ 07102-3027, USA.

Abstract: We present a nonparametric approach to deriving clusters from categorical (nominal scale) data using a new clustering procedure called K-modes, which is analogous to the traditional K-Means procedure (MacQueen 1967) for clustering interval scale data. Unlike most existing methods for clustering nominal scale data, the K-modes procedure explicitly optimizes a loss function based on the L_0 norm (defined as the limit of an L_p norm as p approaches zero).

In Monte Carlo simulations, both K-modes and latent class procedures (e.g., Goodman 1974) performed with equal efficiency in recovering a known underlying cluster structure. However, K-modes is an order of magnitude faster than the latent class procedure in speed and suffers from fewer problems of local optima than do latent class procedures. For data sets involving a large number of categorical variables, latent class procedures become computationally extremely slow and hence infeasible.

We conjecture that, although in some cases latent class procedures might perform better than K-modes, it could out-perform latent class procedures in other cases. Hence, we recommend that these two approaches be used as "complementary" procedures in performing cluster analysis. We also present an empirical comparison of K-modes and latent class, where the former method prevails.

Keywords: Categorical data; Cluster analysis; Groups; Modes; Latent class analysis

1. Introduction

This paper presents a simple procedure for clustering of nominal scale data. The procedure, which we call K-modes clustering, is analogous to MacQueen's (1967) K-means clustering procedure. Input data for K-means clustering procedures must generally have either interval or ratio scale properties. In contrast, researchers frequently need clusters based on nominal scale data. K-means procedures are generally inappropriate for such categorical data.

Five techniques commonly used for finding clusters from categorical data entail the following: (a) dummy code the categorical variables, compute intersubject distances from the dummy coded data, and use such hierarchical clustering procedures as single, complete, or average linkage on the derived intersubject distances; (b) dummy code the categorical variables, and use K-means on these dummy variables; (c) use correspondence analysis to derive spatial coordinates (e.g., see Carroll, Green, and Schaffer 1986) for each subject, and then use K-means on the derived spatial coordinates; (d) use latent class procedures (e.g., Goodman 1974) available for contingency table analysis; and (e) use Hartigan's Ditto Algorithm (1975, pp. 143-154) for categorical data.

The first three procedures have some drawbacks. In (a), one needs to select a distance measure from among many candidate choices. In addition,

most hierarchical clustering procedures are heuristic algorithms that do not explicitly optimize an overall measure of fit. Also, for large data sets (e.g., more than 4000 observations), most available computer programs for implementing these algorithms either abort unsuccessfully because of insufficient memory, or take an inordinately long time, even when mainframe computers are used.

The use of K-means clustering in (b) is not appropriate, because K-means minimizes an ordinary least-squares (OLS) fitting function, which is not valid for categorical data. Moreover, means are not appropriate measures of central tendency for categorical data. While the use of K-means on spatial coordinates derived from correspondence analysis, as in (c), does not violate any fundamental rules, defining clusters on spatial representations derived from such continuous models as correspondence analysis may be inappropriate. Arabie and Hubert (1994) argue that if cluster analysis is to be given the chance to reveal structure in the data, that structure should not first be "filtered" through an incompatible spatial model.

Although latent class techniques in (d) are theoretically sound and have been used extensively (Goodman 1974; Dillon and Mulani 1989; Ramaswamy, Chatterjee, and Cohen 1996), these techniques become computationally intense when the number of variables and/or the number of categories of these variables becomes large. Moreover, latent class procedures generally make assumptions of local independence and certain parametric assumptions about the nature of the data.

The Ditto Algorithm of Hartigan (1975, pp. 143-154) can also be used for clustering categorical data. Although this algorithm is designed especially for categorical data, the overall fit measure that it optimizes can deteriorate during some stages of the algorithm. Thus, the algorithm cannot even guarantee locally optimal solutions.

In this paper, we present a clustering procedure called K-modes, which (a) is nonparametric because it does not make any distributional assumptions about the data, (b) circumvents the need to define ad hoc distance measures on the categorical data to be clustered, (c) explicitly optimizes a "matching" metric (corresponding to the L_0-loss function that will be defined later), (d) is as fast as K-means clustering (an implementation of MacQueen's algorithm developed by the first author), (e) can handle the sizes of large data sets typically found in survey research applications, and (f) does not become computationally intense even when the number of categories or the number of variables to be clustered becomes very large.

In the following section, we describe the proposed general bilinear clustering model, and show that when estimated using an L_0-norm loss function, the model results in the K-modes clustering procedure.

2. The Bilinear Clustering Model

Assume that we have data on N categorical variables from M consumers. Let K be the number of clusters being sought. Then, the K-modes clustering model can be written as:

$$\mathbf{C}_{M\times N} = \mathbf{S}_{M\times K}\,\mathbf{W}_{K\times N} + \text{error}, \qquad (1)$$

where

C is a consumers × variables data matrix,
S is a *binary* indicator matrix for membership of the M consumers in K mutually exclusive, non-overlapping clusters (so that each row of **S** has exactly one element equal to one and the remaining elements are equal to zero), and
W is a matrix of "generalized centroids", defined in this case as modes of certain observations.

It should be noted that only **C**, the data matrix, is known in (1), whereas both **S** and **W** are unknown and must be estimated. This general bilinear model has been used in the past for clustering interval scale data (where **C** is continuous) by Mirkin (1990) and by Chaturvedi, Carroll, Green, and Rotondo (1997). For interval scale data, the matrix **S** can be a completely general binary matrix, not necessarily defining a partition. In this paper, we assume that the data matrix **C** is categorical. The matrix **W** will also have categorical or nominal scale valued elements, while **S** will be constrained to define a partition.

3. Parameter Estimation via an L_0 norm

As in Chaturvedi, Carroll, Green, and Rotondo (1997), we define the parameter estimation problem via minimizing an L_p–norm based loss function[1]

$$L_p = \sum_{m=1}^{M}\sum_{n=1}^{N}\left|c_{mn} - \hat{c}_{mn}\right|^p \quad ,$$

[1] The definition of the L_p-norm based loss functions presented in this paper does not use the power 1/p associated with general L_p metrics. Because the loss function is being minimized, this omission makes no difference in the resulting parameter estimate, since L_p^p, with $p > 0$, is an increasing monotonic function of L_p. For $p \geq 1$, L_p^p is a metric, as demonstrated by Carroll and Wish (1974, pp. 412-416). L_p, as defined here, is itself a metric for $0 < p \leq 1$. L_0 is the limiting case corresponding to the "counting metric", discussed in this paper.

where $\hat{c}_{mn}$ is the $(m, n)^{th}$ element of $\hat{\mathbf{C}} = \mathbf{SW}$, to estimate **S** and **W**, for positive values of $p \to 0$.

In the limiting case as $p \to 0$, the L_p norm-based loss function (hereafter referred to, for the sake of brevity, as the L_0 loss function), simply counts the number of mismatches in the matrices **C** and $\hat{\mathbf{C}}$. To be precise, mathematically, the L_p norm approaches a counting metric as $p \to 0$; the L_0 metric, is then defined as this limiting case. The L_0 loss function can be appropriate when the data are categorical, because counting is a permissible operation on categorical data. In this paper, we concentrate on estimating the model in (1) using the L_0 loss function, while **S** is constrained to be a partitioning matrix.

4. Estimation Procedure

The matrices **S** and **W** are estimated iteratively (estimating **S** given estimates of **W**, then revising the estimates of **W** given the new estimates of **S**) until the value of the L_0 loss function does not improve. The procedures for estimating **S** and **W** are given below:

To estimate **S**, the cluster membership, given estimates of **W**, consider the following illustrative case by first defining

$$\mathbf{C} = \begin{bmatrix} 1 & 5 & 0 & 3 \\ 2 & 6 & 1 & 3 \\ 3 & 6 & 0 & 3 \\ 2 & 7 & 0 & 4 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ s_{31} & s_{32} \\ s_{41} & s_{42} \end{bmatrix}, \quad \text{and } \mathbf{W} = \begin{bmatrix} 2 & 6 & 1 & 3 \\ 1 & 5 & 0 & 4 \end{bmatrix}.$$

The $(i, j)^{th}$ entry of matrix **W** corresponds to the category or nominal scale value defining the centroid for the i^{th} cluster and the j^{th} categorical variable. We wish to find the L_0-norm based estimates of **S**, where **C** = **SW** + error and **S** is either 0 or 1.

If we let

$$f_1 = (1 - 2s_{11} - 1s_{12})^0 + (5 - 6s_{11} - 5s_{12})^0 + (0 - 1s_{11} - 0s_{12})^0 + (3 - 3s_{11} - 4s_{12})^0,$$

$$f_2 = (2 - 2s_{21} - 1s_{22})^0 + (6 - 6s_{21} - 5s_{22})^0 + (1 - 1s_{21} - 0s_{22})^0 + (3 - 3s_{21} - 4s_{22})^0,$$

$$f_3 = (3 - 2s_{31} - 1s_{32})^0 + (6 - 6s_{31} - 5s_{32})^0 + (0 - 1s_{31} - 0s_{32})^0 + (3 - 3s_{31} - 4s_{32})^0,$$

and

$$f_4 = (2 - 2s_{41} - 1s_{42})^0 + (7 - 6s_{41} - 5s_{42})^0 + (0 - 1s_{41} - 0s_{42})^0 + (4 - 3s_{41} - 4s_{42})^0,$$

then the total mismatch L_0 loss function) is given by

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

Note that f_1 is a function only of s_{11} and s_{12}; f_2 is a function only of s_{21} and s_{22}; f_3 is a function only of s_{31} and s_{32}; and f_4 is a function only of s_{41} and s_{42}. Thus, F *is separable* with respect to parameters for each row of **S** (Chaturvedi and Carroll, 1994, used this row-wise separability property in their SINDCLUS procedure for fitting the INDCLUS model). To minimize F, one can separately minimize f_1 with respect to parameters for the first row, f_2 with respect to parameters for the second row, etc.

To minimize, say, f_1 with respect to parameters for the first row, [s_{11} s_{12}], one can evaluate f_1 explicitly at its two permissible values of row 1, [1 0] and [0 1], given the constraint of a partitioning solution. For the pattern[1 0], $f_1 = 3$, and for the pattern [0 1], $f_1 = 1$ (i.e., mismatch). Thus, $s_{11} = 0$ and $s_{12} = 1$ are the optimal estimates. The other rows of **S** can be determined using a similar procedure.

To estimate **W** (the cluster "generalized centroids"), given estimates of **S**, we first define

$$\mathbf{C} = \begin{bmatrix} 1 & 5 & 0 & 3 \\ 1 & 6 & 1 & 3 \\ 3 & 6 & 0 & 3 \\ 2 & 7 & 0 & 4 \\ 1 & 5 & 1 & 4 \\ 2 & 5 & 1 & 2 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \text{and } \mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \end{bmatrix}.$$

To determine the L_0 estimates of **W**, where **C** = **SW** + error, and **W** is categorical, we can see that the mode of (1,1,3) = 1 optimizes the L_0 loss function, and hence, $\hat{w}_{11} = 1$. Similarly $\hat{w}_{12}$ will be the mode of (5,6,6) = 6,..., and $\hat{w}_{24}$ will be the mode of (4,4,2) = 4.Thus, **W** is a matrix of modes of the variables for each cluster, and hence, we call this procedure K-modes clustering. We break ties arbitrarily in the current version of the software. We repeat the estimation of **S** given **W**, and **W** given **S** until the L_0 (or "counting") loss function does not improve. It should be noted that upon convergence, this procedure can yield locally optimal solutions (as can K-means, K-medians, and other related methods).

5. Certain Difficulties with K-modes

Two difficulties can arise when using K-modes for cluster analysis. First, the K-modes procedure can only guarantee a locally optimal solution. The best strategy when using K modes with real data is to use multiple random starting seeds, and choose the solution with the lowest L_0-norm value. Second, there are no statistically valid or reliable indices that can be used with K-modes to determine the "true" number of clusters in the data. This problem can possibly be circumvented by using K-modes in conjunction with latent class procedures, which have a variety of information-theoretic indices such as the Akaike Information Criterion (AIC) of Akaike (1973), Schwartz Information Criterion (Schwartz 1978), or Consistent AIC (Bozdogan 1987) for determining the "true" number of classes in the data. (For counterarguments, see McDonald 1989.)

6. A Monte Carlo Comparison of K-modes and Latent Class Analysis

We first conducted a Monte Carlo simulation to compare the performance of K-modes and latent class analysis procedures. Our primary interest was in comparing the degree of recovery of a known underlying cluster structure, the speed of execution of the program, and the severity of the local optimum problem. Consistent with these objectives, we systematically varied six critical factors that could potentially affect the performance of these clustering procedures in generating the artificial data: number of observations (600, 1200, 2400), number of clusters (2, 4, 6), number of categorical variables (7, 10, 15), number of levels for the categorical variables (3, 5, 7), amount of error (10%, 30%, 50%), and the ratio of large-to-small cluster sizes (50:50, 70:30, 90:10). This plan resulted in a 3^6 design. A 27-cell orthogonal design (1/27 fraction) that would enable unbiased estimation of the main effects was chosen. Three sets of data (replications) were generated within each cell, yielding a total of 81 data sets.

6.1 Data generation and analysis

Each observation (row) of the **S** matrix in (1) was assigned to a cluster at random. The number of observations (M) was divided among the number of clusters (K) so that exactly K/2 clusters each had large or small sizes. The modal level for each variable in each cluster, matrix **W** in (1), was generated at random from among the levels (3, 5, or 7) tested in this study. Error-free data were formed by multiplying **S** and **W**. A fixed percentage (10%, 30%, or 50%) of observations for each variable in each cluster was

then perturbed and assigned a non-modal level at random. Thus, after the error was added to the data via this perturbation, each variable in each cluster had approximately the same number of levels.

Two procedures were used to extract the underlying cluster structure from the artificially generated data sets: the K-modes procedure and an implementation of Goodman's (1974) latent class procedure[2], which we will refer to hereafter as LCA for the sake of brevity. Both K-modes and LCA were run from ten different random starts[3]. In the case of K-modes, the solution yielding the highest matches-accounted-for (MAF) criterion (which is analogous to the R^2 criterion for interval scale data), was chosen as the final solution. In the case of LCA, the solution yielding the highest likelihood was chosen as the final solution. For present purposes, the LCA solutions were converted to a discrete cluster solution by assigning each observation to the class for which the associated posterior probability was the highest. We used the corrected Rand index (Hubert and Arabie 1985, Equation 5) to assess the recovery of the underlying cluster structure (**S** matrix). The better the recovery of the true underlying cluster structure, the closer the corrected Rand index would be to 1.0.

6.2 Solution Recovery

A paired difference, one-tailed *t*-test indicates that the mean corrected Rand index is significantly higher (though only slightly) for K-modes than for LCA ($t = 2.35$, $p = 0.011$). This finding indicates that K-modes is only slightly superior to LCA in recovering a known underlying cluster structure. It can be safely assumed that the two procedures yielded almost identical results in recovering known cluster structure.

We also conducted several ANOVA's (regression via effects-coded dummy variables) to test the effect of the underlying factors on cluster structure recovery (i.e., corrected Rand index) of the K-modes and LCA procedures. Separate ANOVA's were performed using the corrected Rand indices derived from the K-modes and LCA procedures. A separate ANOVA was also performed on the difference of the corrected Rand indices derived using K-modes and LCA. The results of the analyses are given in Table 1. The independent factors have a significant impact on the corrected Rand index for both the K-modes and LCA procedures, as indicated by the significant F-values. Two factors, cluster-size ratio and replications, have no

---

[2] We thank Professor Abba Krieger for providing a FORTRAN implementation of Goodman's (1974) maximum likelihood procedure for fitting the latent class model.

[3] We would have used more than ten random starts for determining the best-fitting solutions, had doing so not required prohibitive computer time.

significant impact on the underlying cluster structure recovery across the two procedures.

The solution recovery using both K-modes and LCA has a strong negative relationship with the number of clusters and amount of error in the data, and a strong positive relationship with the number of levels and number of variables, as seen in Table 1. The mean corrected Rand index is 0.79 for K-modes, whereas the LCA procedure yields a slightly (though significantly) lower mean corrected Rand index (0.78). Table 1 suggests both K-modes and LCA perform similarly in recovery of solutions, but K-modes results in a slightly higher mean corrected Rand index; the ANOVA results using the difference in corrected Rand indices indicates a grand mean coefficient of 0.1, which is statistically significant.

6.3 Run Time

A paired difference, one-tailed *t*-test indicates that the run-time (for each run of K-modes and LCA, respectively) is significantly lower for K-modes than for the LCA procedure for the 81 data sets ($t = 5.61$, $p = 0.0001$). This result suggests that K-modes is significantly faster than LCA in speed of execution.

Three different ANOVA's were performed using (a) the run-time of K-modes, (b) the run-time of LCA, and (c) the ratio of the run-time for LCA to the run-time of K-modes. The results are given in Table 2. The first five factors have an important impact on the corrected Rand index for both the K-modes and LCA procedures, as indicated by the significant F-values. The last two factors, cluster-size ratio and replications, have no significant impact on run-time of the two procedures.

The run-times for both K-modes and LCA have a strong positive relationship with the number of observations and number of clusters and a strong negative relationship to the number of levels, as seen in Table 1. The mean run-time is 0.18 minutes for K-modes while the LCA procedure yields a much higher (by an order of magnitude) mean run-time of 2.02 minutes. The error level has a very strong positive relationship with run-time in the case of LCA, whereas in the case of K-modes, error level does not have any significant impact on run-time. Table 2 suggests that K-modes is superior to LCA in speed of execution.

Table 1
Corrected Rand Index: ANOVA Coefficients († implies significance at $\alpha = .05$)

Variable	Factor level	K-modes	LCA	Difference (K-modes-LCA)
Grand Mean		0.79†	0.78†	.01†
# Observations	600	0.00	-0.01	0.01†
	1200	-0.01	0.00	0.00
	2400	0.01	0.01	-0.01†
# Clusters	2	0.03†	0.03†	0.00
	4	0.02	0.02	0.00
	6	-0.05†	-0.05†	0.00
# Variables	7	-0.07†	-0.07†	0.00
	10	-0.02	-0.01	-0.01†
	15	0.09†	0.08†	0.01†
# Levels	3	-0.17†	-0.17†	0.00
	5	0.05†	0.05†	0.00
	7	0.12†	0.12†	0.00
Random error level	10%	0.20†	0.21†	-0.01†
	30%	0.10†	0.10†	0.00
	50%	-0.30†	-0.31†	0.01†
Cluster size ratio	50:50	-0.03	-0.02	0.00
	70:30	0.02	0.03	0.00
	90:10	0.01	-0.01	0.00
Replication	1	0.00	-0.01	0.00
	2	0.01	0.00	0.00
	3	-0.01	0.01	0.00
R^2		0.81	0.82	0.30
F-value		18.47†	19.74†	1.86†

Table 2.
Runtime (Minutes): ANOVA Coefficients († implies significance at α= 0.05)

Variable	Level	K-modes	LCA	Ratio of time for LCA over time for K-modes
Grand Mean		0.18 †	2.02 †	11.24 †
# Observations	600	-0.09 †	-1.22 †	0.61
	1200	-0.03 †	-0.02	0.45
	2400	0.12 †	1.24 †	-1.05
# Clusters	2	-0.07 †	-1.63 †	-4.08 †
	4	-0.02 †	-0.24	-0.28
	6	0.09 †	1.87 †	4.36 †
# Variables	7	-0.05 †	-0.66 †	1.71
	10	0.03 †	0.99 † g	1.37
	15	0.02 †	-0.33 †	-3.08 †
# Levels	3	0.05 †	1.86 †	6.53 †
	5	-0.03 †	-0.73 †	-0.54
	7	-0.02 †	-1.13 †	-5.99 †
Random error level	10%	0.00	-1.07 †	-6.89 †
	30%	0.01	-0.34 †	-2.88 †
	50%	-0.01	1.41 †	9.77 †
Cluster size ratio	50:50	0.00	-0.29	-2.40 †
	70:30	-0.01	0.31	-1.58
	90:10	0.01	-0.02	3.98 †
Replication	1	0.00	-0.04	-0.31
	2	0.00	0.03	0.41
	3	0.00	0.01	-0.10
R^2		0.84	0.73	0.69
F-value		22.75 †	11.72 †	9.65 †

6.4 Local Optima

For each of the 81 data sets, we computed the proportion of times (out of the ten solutions obtained from random starting seeds) that the "best" K-modes and LCA solutions were found. By "best" solutions for LCA and K-modes, we mean solutions with the maximum likelihood and the maximum Matches-Accounted-For (MAF), respectively. Thus, we kept a count of the number of times that the two procedures yielded their respective "best" solution for each of the 10 random starts.

This mean proportion for K-modes was 55.7%; i.e., on average, K-modes found its "best" solution for 5.57 of the ten different solutions. The mean proportion for LCA was 22.7%. A paired difference, one-tailed t-test indicates that the proportion of times that the "best" solution was found is significantly higher for K-modes than for LCA for the 81 data sets ($t = 7.18$, $p = 0.000$). This result suggests that K-modes may have fewer problems with locally optimal solutions than does LCA.

We also conducted a nonparametric sign test to assess the performance of K-modes and LCA. We converted the proportion of times that the two procedures found their respective "best" solutions for the 81 data sets to a sequence of 81 numbers that were either -1, 0, or +1 depending on whether the proportion for that data set was higher for LCA, equal for LCA and K-modes, or higher for K-modes, respectively. LCA yielded higher proportions for five data sets, LCA and K-modes were tied for 31 data sets, and K-modes yielded higher proportions for 45 data sets.

We conducted a sign test of the null hypothesis that K-modes did not yield higher proportions of "best" solutions than LCA, against the alternative hypothesis that K-modes yielded higher proportions than LCA. The sign test resulted in a p-value of 0.0001, indicating that K-modes did yield a significantly higher proportion of "best" solutions than LCA.

6.5 Summary of Monte Carlo Results

The simulation results indicate that (a) K-modes and LCA are equally good in recovering a known underlying cluster structure, (b) K-modes is an order of magnitude faster than LCA, and (c) K-modes has significantly less vulnerability to local optima than LCA does, when both procedures are started from random seeds.

7. Application of K-modes to Market Segmentation

We now demonstrate, using a sample data set, that K-modes can result in "better" clustering or market segmentation (in certain cases) than latent class procedures. We recognize that the comparative performance of K-modes and LCA could turn out differently depending on (a) the characteristics of data sets and (b) different criteria of performance evaluation used (e.g., interpretability of the segments, segment sizes, segment addressability, profitability of the segments, etc.). In this paper, we demonstrate the superior performance of K-modes to LCA using the criterion of segment addressability (segment addressability refers to the degree to which a clustering or segmentation solution can be related to variables that are (i) controllable by marketing managers, and (ii) that help marketing managers in finding or "targeting" the consumers accurately in the relevant clusters/segments). This criterion has been suggested for use as a cluster evaluation procedure by Helsen and Green (1991) and by Chaturvedi, Carroll, Green, and Rotondo (1997).

7.1 The Study

The XYZ corporation is a small computer software provider located in the Southwest US. To understand the personal computer (PC) market, XYZ acquired data from a large market research company in the US. The market research company had conducted a study to understand the usage and accessibility of PC's to the 98 million households in the US. A random sample of 2000 households in the US was surveyed for this purpose.

The survey gathered responses on such topics as: computer usage and ownership (e.g., PC ownership, use of PC from home and office, etc.), demographics, and the kinds of TV programs watched from home. Our task was to uncover market segments based on these data to help XYZ management (a) to understand the differential patterns of PC ownership/use and demographics across segments, and (b) to address these segments via various communication/advertising programs for marketing new PC based products and services.

A total of eight variables constituted the input data and are listed in Table 3. Five of the eight variables were demographic, while the rest concerned ownership and usage of PC's. We selected demographics and PC ownership/usage related variables to satisfy objective (a), and TV viewership variables to satisfy objective (b) regarding addressability of the segments.

7.2 Results Obtained via K-modes Clustering

We first used the K-modes procedure to obtain 2- through 5-cluster K-modes solutions. The criteria of interpretability and segment sizes were used to determine the number of clusters. We found the 3-cluster K-modes solution to be interpretable. To assess the fit of the 3-cluster K-modes solution, we computed the MAF statistic (which, again, is analogous to the R^2 statistic associated with OLS estimation). The MAF for the 3-cluster K-modes solution was 63.9% (Table 4), indicating a good fit to the data. A profile of the three clusters using the eight input variables is given in Table 5. The three clusters are interpreted very readily.

Cluster 1 (size 53.5%) corresponds to the "PC-novices" group. This segment has a very low penetration of PC's (only 14.2%) and a very low incidence of use of PC's either at work (14.4%) or at home (9.7%). Cluster 1 is composed of mostly older (44.3% are aged 50+) and less educated (60.2% have not graduated high school) people who are either retired (26.5%) or have blue-collar jobs (16.3%) with low incomes (48% have annual incomes $19K or less), and the head of household is a female (65.3%).

Cluster 2 (size 24.8%) corresponds to the "Use and like PC" households. Of these households, 96.2% own a PC, 89.7% use PC's at home, and 77% at work. This cluster comprises households whose heads usually are middle-aged (64.3% are aged 30-50), educated (64.5% have at least some college education), have white-collar jobs (70.1%), high-incomes (34.1% have annual incomes 50K$ or more), and male (66.7%).

Cluster 3 (size 21.7%) is the "Use PC only at work" segment. While a large majority of this cluster uses PC's at work (75.8%), a very low proportion actually owns PC's (5.1%), or uses one at home (4.6%). This cluster consists of households whose heads are usually middle-aged (61.7% between ages 30-50), have medium levels of education (77.4% have completed high school or have some college experience), white-collar professionals (77%), with moderate incomes (56.5% have incomes annual incomes between 20K-50K$), and male (74.7%).

For comparison, we also dummy-coded the 2000 x 8 data matrix, resulting in a 2000 x 30 data matrix of dummy variables. We used K-means clustering (PROC FASTCLUS in SAS) to extract a 3-cluster solution. The solution thus derived differed in segment membership when compared to the 3-cluster K-modes solution, resulting in a corrected Rand index (Hubert and Arabie 1985) of only 0.18. This inferior solution did not evince a clear interpretation; hence we discarded it.

Table 3. Variables Used for Determining Market Segments

Variable	Categories
Age of head of household	Missing:0-25 :25-30 :30-40 :40-50 :50+
Education of head of household	Not HS Grad:HS or Vocational: Some college:College graduate: Graduate school:Missing
Occupational class	White collar :Blue collar: Other employed: Student
	Retiree: Unemployed: Missing
Annual HH Income	< 10K:10-19K: 20-29K: 30-39K: 40-49K: 50-74K: 75-99K: 100K+
Gender of Head of household	Male:Female
Own a PC	No:Yes
Use computer at work	No:Yes
Use computer at home	No:Yes

Table 4. Percent Matches-Accounted-For (MAF)

Clustering procedure	MAF
3-cluster, single linkage	48.78
3-cluster, K-means	48.99
3-cluster, average linkage	58.45
3-cluster, complete linkage	59.60
3-class, LCA solution	60.73
2-cluster, K-modes	58.54
3-cluster, K-modes	63.90
4-cluster, K-modes	66.72
5-cluster, K-modes	68.08
6-cluster, K-modes	69.01
7-cluster, K-modes	70.39
8-cluster, K-modes	70.78

Table 5. The 3-cluster K-modes Solution

Variables	Levels	Segment 1	Segment 2	Segment 3
Age of Head of HH	Missing	1.8	2.2	1.6
	0-24.9	9.2	9.9	15.9
	25-29.9	10.7	10.1	12.9
	30-39.9	17.3	37.2	40.3
	40-49.9	16.7	27.1	21.4
	50+	44.3	13.5	7.8
Education of HH head	Not H.S. Graduate	60.2	11.1	12.2
	H.S. or Vocational	19.5	23.8	55.5
	Some College	12.0	35.4	21.9
	College Graduate	2.6	7.9	4.4
	Graduate School	4.9	21.2	5.5
	Missing	0.7	0.6	0.5
Occupation of head of HH	White Collar	44.4	70.1	77.0
	Blue Collar	16.3	17.6	12.7
	Other Employed	2.8	4.6	4.1
	Retiree	26.5	3.6	2.8
	Student	8.9	2.8	2.1
	Missing	1.1	1.2	1.4
Annual Household Income	Under 10K	29.4	5.1	5.3
	10K-19K	18.7	8.3	15.7
	20K-29K	14.3	12.9	18.0
	30K-39K	8.6	12.7	12.0
	40K-49K	4.9	26.9	26.5
	50K-74K	3.0	11.3	5.8
	75K-99K	2.9	8.7	2.5
	100K or more	18.2	14.1	14.3
Gender of Head of household	Male	34.7	66.7	74.7
	Female	65.3	33.3	25.3
Own a PC	No	85.8	3.8	94.9
	Yes	14.2	96.2	5.1
Use PC at Home	No	90.3	10.3	95.4
	Yes	9.7	89.7	4.6
Use PC at work	No	85.6	23	24.2
	Yes	14.4	77	75.8
Relative segment sizes		53.5%	24.8%	21.7%

We also tried to analyze the data using such hierarchical clustering procedures as average linkage, single linkage, and complete linkage on Euclidean distances derived from the dummy coded 2000 x 30 data. The average, complete, and single linkage procedures respectively yielded an MAF of only 58.45%, 59.6%, and 48.78%). None of these solutions was as easily interpretable as the 3-cluster K-modes solution. In fact, the sizes of the three clusters derived via the single-linkage procedure were (1998,1,1). This solution was not used for further analysis.

7.3 Results of Latent Class Analysis

For comparison, we also applied the LCA procedure to obtain market segments for the same categorical data set. We first obtained a 3-class LCA solution to compare with the 3-cluster K-modes solution. The former solution resulted in a MAF of 60.7% compared to 63.9% for the 3-cluster K-modes solution. Moreover, the 3-class LCA solution was quite different from the 3-cluster K-modes solution. Of the 2000 observations, 507 were in the off-diagonal portion of the 3 x 3 table formed using the two solutions, resulting in a corrected Rand index of only 0.32.

Because the LCA procedure employs a maximum likelihood approach to estimating the latent class model, we can choose the number of latent classes using the AIC, BIC, or CAIC information-theoretic criteria. We used the CAIC criterion for selecting the number of "true" classes in the data. In the present case, the CAIC statistic indicated a 7-class solution as the best fitting LCA solution. However, it resulted in relative class sizes of 3.4% for two of the classes, which are quite small and can be considered unstable. Assuming arguably that the CAIC is a pointer to picking out the "correct" number of clusters in the data, we also extracted a 7-cluster K-modes solution. The 7-cluster K-modes solution was also very different from the 7-class LCA solution (with a corrected Rand index of 0.29). The smallest cluster size for the 7-cluster K-modes solution was 7.8% (i.e., 145 of the 2000 observations).

At this stage, we had four candidate solutions for comparison: the 3- and 7-cluster K-modes solutions, and the 3- and 7-class LCA solutions. We decided to use the criterion of segment addressability to choose among these solutions.

7.4 Segment Addressability: Relationship to TV Viewership Variables

We also had data on twelve variables related to the kinds of programs/content watched on TV by the same 2000 households. We used the segment memberships derived from K-modes (3- and 7-cluster) and LCA (3-

and 7-class) to see how closely related each was to the set of these exogenous variables. We employed the p-value associated with the two-way cross-tabulation of each clustering with each exogenous variable to test the association. (Unlike χ^2, the p-value is not affected by the dimensions of the contingency table.) The smaller the p-value, of course, the more significant is the association. For each clustering solution, the average p-value across all the two-way tables involving that clustering versus each background variable provides a comparative descriptive measure across the different clustering solutions

Table 6 presents the distribution of the p-values for the various K-modes and LCA solutions with the twelve background variables. Both K-modes solutions (the 3- and 7-cluster K-mode solutions) have lower mean p-values across the twelve shopping variables, compared to the respective LCA solutions. For both the 3- and 7-cluster solutions, K-modes outperforms LCA. The best solution (with the lowest mean p-value) among the four solutions is the 7-cluster, K-modes solution.

We chose the 3-cluster K-modes solution as the superior solution because (a) it is more parsimonious than a 7-cluster K-modes solution, (b) it out-performs its counterpart, the 3-class LCA solution, (c) it performs almost as well as the 7-class LCA procedure according to mean p-value, and (d) it results in more equal segment sizes. These results suggest that the K-modes solutions have a greater degree of cluster validity than the counterpart LCA solutions for the data set under consideration.

8. Conclusions

We have presented a new approach to market segmentation using categorical data that is similar in spirit to the K-means (for interval or ratio scale data) and K-medians (for ordinal scale data) clustering procedures. The proposed approach is as good as latent class analysis in recovering a known underlying cluster structure, is considerably faster, and suffers fewer problems of local optima than does latent class analysis. Our approach can handle large data sets (involving many categorical variables) quite easily, unlike latent class procedures which become computationally cumbersome. Moreover, K-modes, unlike latent class procedures, can provide solutions even when the data have sparse marginal distributions. The current implemen-

Table 6. Cluster Validation: Distribution of p-values

Association of clustering solutions with TV-viewership variables				
Variable	3-cluster, K-modes	3-class, LCA	7-cluster, K-modes	7-class, LCA
Watch Sports	0.14	0.99	0.02	0.00
Comedy programs	0.08	0.00	0.00	0.00
Talk Shows	0.01	0.01	0.08	0.09
Music Channels	0.01	0.00	0.00	0.00
Community Cable	0.01	0.04	0.13	0.04
Games	0.00	0.00	0.00	0.00
Documentaries	0.00	0.00	0.00	0.00
C-SPAN	0.00	0.01	0.00	0.00
Home/Hobby Show	0.45	0.78	0.17	0.21
Finance shows	0.00	0.00	0.00	0.00
Health / Exercise / Medicine Shows	0.57	0.64	0.39	0.81
Home Shopping Channel	0.00	0.03	0.00	0.01
Mean association	0.11	0.21	0.07	0.10

tation of the K-modes program can handle up to 10,000 respondents and 100 variables with up to 50 categories each. [4]

A drawback of K-modes (as is true for most clustering procedures) is the lack of statistically valid/reliable indices for choosing the "correct" number of clusters. Latent class procedures, on the other hand, can arguably employ a variety of such information-theoretic approaches as AIC, BIC, CAIC, etc., because they use maximum likelihood estimation procedures.

We recommend that K-modes be used in conjunction with latent class procedures to uncover market segments based on categorical data. Although Monte Carlo simulations indicate that the two procedures are quite similar in

[4] Interested readers can obtain a copy of the FORTRAN program for the K-modes procedure from the first author.

recovering a known underlying structure, we find that these procedures yield very different solutions for empirical data. Until we have a theory determining the conditions in which K-modes would perform better (or worse) than latent class procedures, we recommend that both these techniques be used in parallel in real-life market segmentation applications, whenever possible. Of course, K-modes can also be viewed as an exploratory data analytic procedure for clustering, independent of LCA, or of other statistical methodologies.

References

AKAIKE, H. (1973), "Information Theory and Extensions of Maximum Likelihood Principle," in *2nd International Symposium of Information Theory*, Eds., B.N. Petrov and F. Csaki, Budapest: Academia Kiado, 267-281

ARABIE, P., and HUBERT, L. (1994), "Cluster Analysis in Marketing Research," in *Advanced Methods of Marketing Research*, Ed., R. P. Bagozzi, Oxford: Blackwell, 160-189.

BOZDOGAN, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions," *Psychometrika*, 52, 345-370.

CARROLL, J. D., GREEN, P. E., and SCHAFFER, C. M. (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research*, 22, 271-281.

CARROLL, J. D., and WISH, M. (1974), "Multidimensional Perceptual Models and Measurement Methods," in *Handbook of Perception* (Vol. 2), Eds., E. C. Carterette and M. P. Friedman, New York: Academic, 391-447. Reprinted (1984) in *Key Text on Multidimensional Scaling,* Eds., P. Davies and A. P.M. Coxon, Portsmouth, New York: Heinemann, 43-58.

CHATURVEDI, A. D. and CARROLL, J. D. (1994), "An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models," *Journal of Classification*, 11, 155-170.

CHATURVEDI, A. D., CARROLL, J. D., GREEN, P. E., and ROTONDO, J. A. (1997), "A Feature-Based Approach to Market Segmentation via Overlapping K-Centroids Clustering," *Journal of Marketing Research*, 34, 370-377.

DILLON, W. R., and MULANI, N. (1989), "LADI: A Latent Discriminant Model for Analyzing Market Research Data," *Journal of Marketing Research*, 26, 15-29.

GOODMAN, L. A. (1974), "Exploratory Latent Structure Analysis using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215-231.

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

HELSEN, K., and GREEN, P.E. (1991), "A Computational Study of Replicated Clustering with an Application to Marketing Research," *Decision Sciences*, 22, 1124-1141.

HUBERT, L., and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193-198.

MACQUEEN, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Eds., L. M. Le Cam and J. Neyman, Berkeley, CA: University of California Press, Vol. 1, 281-297.

MCDONALD, R. P. (1989), "An Index of Goodness-of-Fit Based on Noncentrality," *Journal of Classification*, 6, 97-103.

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

RAMASWAMY, V., CHATTERJEE, R., and COHEN, S. H. (1996), "Joint Segmentation on Distinct Interdependent Bases with Categorical Data," *Journal of Marketing Research*, 32, 337-351.

SCHWARTZ, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.