WAGNER A. KAMAKURA*

A popular procedure for benefit segmentation based on conjoint experiments has been to estimate individual-level part worths and then form nonoverlapping clusters of consumers with similar estimates. Rather than using these estimates as the criteria for clustering, the least squares procedure discussed in the article attempts to group consumers into homogeneous segments so their stated preferences are explained maximally by their group-level preference functions. This procedure also provides a measure of the expected predictive accuracy that will help the researcher in choosing an adequate aggregation level.

A Least Squares Procedure for Benefit Segmentation with Conjoint Experiments

Conjoint analysis has been used extensively by marketing researchers for understanding consumers' preferences (Green and Srinivasan 1978). A comprehensive survey of its use in commercial research is presented by Cattin and Wittink (1982), who identify market segmentation as one of its major applications.

Conjoint analysis is especially helpful in the identification and understanding of benefit segments. With preference data from a sample of consumers, the objective in benefit segmentation is to identify groups of consumers having similar preferences that might be targeted more efficiently by specific marketing mixes (Currim 1981; Haley 1968; Wind 1978). The dual goals are (1) to form groups of consumers who share a common utility function and (2) to estimate the aggregate utility functions that would best explain the preferences stated by the members of each segment.

The most common approach to benefit segmentation based on conjoint experiments has been to use either a nonmetric procedure (e.g. LINMAP and MONANOVA), ordinary least squares, or binary/multinomial logit to estimate individual-level preference functions. A nonoverlapping clustering algorithm (e.g., Howard-Harris 1966; Johnson 1967; Ward 1963) then is used to form groups of individuals with similar estimated attribute weights (Currim 1981; Green, Wind, and Jain 1972; Huber and Moore 1979; Moore 1980). One assumes that groups with similar estimates are also homogeneous in terms of their preference structures and therefore represent benefit segments. However, the minimum-variance or minimum-distance clustering algorithms typically used in this two-stage segmentation procedure take the estimated weights as observed characteristics of the individuals being clustered, ignoring any error incurred in their estimation. Hence the cluster solution obtained through this process will represent homogeneous benefit segments to the extent that the estimated attribute weights are accurate, error-free proxies for the "true" unobservable utility functions of each consumer in the sample.

The distinction between the "true" and estimated utility functions is very important, given the potentially biasing errors likely to occur in the estimation of subject-level part worths. For example, in the least squares case, the regression estimates $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ will carry a bias that amounts to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, where $\mathbf{Y}$ are the preference ratings, $\mathbf{X}$ represents the attributes of the stimulus profiles, and $\mathbf{u}$ contains the measurement errors in the preference ratings (in addition to other errors such as model misspecification).

Errors in the estimated preference weights are particularly critical because fractional factorial designs often are used to form the stimulus profiles in a typical conjoint experiment (Green 1974; Green and Srinivasan 1978; Johnson 1974; Moore 1980). Such designs lead to few degrees of freedom in the estimation, thus making the estimates rather sensitive to the measurement error $\mathbf{u}$. Further, when the treatment levels of an orthogonal con-

*Wagner A Kamakura is Assistant Professor, Owen Graduate School of Management, Vanderbilt University

The author benefited from discussions with Fred Phillips, Raj Srivastava, Russ Winer, and Gary Russell on previous versions of the article. He is grateful to Arun Jain for providing the empirical data

Journal of Marketing Research
Vol XXV (May 1988), 157–67

joint design are coded as binary variables, the actual set predictors **X** are collinear. This collinearity observed between the levels of each treatment has direct implications for the reliability of the part worth estimates. "Our ability to interpret the coefficients declines the more persistent and severe the collinearity" (Judge et al. 1980, p. 453).

Because of the unreliability of subject-level part worth estimates, clustering consumers on the basis of these estimates may lead to misclassifications. Ideally, one would want to use a criterion that acknowledges the potential estimation errors. Econometricians, for example, pool regressions on the basis of their ability to explain the pooled data (sum of squared errors), rather than the similarity of the estimated regression coefficients (Chow 1960). Similarly, one would want to group consumers into benefit segments on the basis of the ability to explain their preferences with the segment-level (pooled) utility function, rather than the similarity of the potentially unreliable part worth estimates for each consumer.

Hagerty (1985) proposed a *Q*-factor analytical procedure that maximizes the predictive power of the segment-level functions. In this optimal-weighting estimation procedure, Hagerty allows each consumer to belong to every market segment and demonstrates that the optimum "participation" of each consumer in each of the segments is determined by the first eigenvectors of the correlation matrix among consumers (across their observed preference ratings).

As indicated by several authors (Cattel 1978; Stewart 1981; Ward 1963), factor analytical procedures lead to overlapping clusters that are not easily (if ever) identifiable. Stewart (1981) shows that the number of factors obtained in a *Q*-factor analysis of individual characteristics (i.e. preferences) is not indicative of the number of clusters. One may in fact have less/more clusters than factors and the identification of the homogeneous clusters is rather subjective and complex, especially when more than two factors are obtained.

Nevertheless, Hagerty's procedure has great merit because it attempts to group consumers into segments in a way that maximizes the understanding/explanation of the preferences by the whole sample of consumers. For practical purposes, however, managers need to be able to distinguish the benefit segments clearly so that their profiles can be drawn and specific marketing strategies can be targeted toward them. Evidently, there is a risk of oversimplification in forcing the segments to be mutually exclusive. In some instances a consumer might belong to more than one segment—for example, when multiple brands are consumed or in multiple usage situations. Therefore, if this simplification is forced into the segmentation scheme, one should be able to measure the loss of explanatory power incurred by forcing the same preference function on different consumers.

In the procedure described next, data from typical conjoint experiments are used to group consumers who respond similarly to the treatments being manipulated. The objective is to form groups and estimate group-level preference functions that best explain the preferences of each individual in the sample. The best *fit* to the observed preferences can be attained by prescribing one function for each individual. However, our objective is to pool the preference data from certain groups of individuals so that the *expected predictive power* of the estimated preference functions is at a desirable level. Hence, in contrast to the usual clustering based on the similarity of part worth estimates, the proposed procedure forms benefit segments in a way that maximizes the ability to explain each consumer's preferences with segment-level part worth estimates. The procedure also provides a measurement of the expected predictive accuracy, which can be used by the researcher to decide the number of benefit segments to use.

IDENTIFYING NONOVERLAPPING BENEFIT SEGMENTS

Problem Setting

Let us consider a simple conjoint experiment combining two attributes with two and three levels, respectively, in a full factorial design. Let the columns of **Y** contain the preference ratings for the six stimulus profiles by a sample of three individuals and **X** contain dummy variables representing the combinations of the two attributes:

$$\begin{array}{cc} \textit{Profile} & \\ 1 & \\ 2 & \\ 3 & \\ 4 & \mathbf{X} = \\ 5 & \\ 6 & \end{array} \begin{bmatrix} -1 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \quad \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3] = \begin{bmatrix} 3 & 2 & 4 \\ 3 & 3 & 5 \\ 5 & 4 & 3 \\ 4 & 5 & 1 \\ 1 & 2 & 4 \\ 1 & 5 & 3 \end{bmatrix}$$

The problem is how to group the individuals into segments that respond similarly to the attributes in **X**. As mentioned before, the usual approach is to estimate each individual-level function $\hat{\mathbf{y}}_i = \hat{\mathbf{c}}_i\mathbf{X}$ and then cluster the individuals according to their similarity in terms of the estimates $\hat{\mathbf{c}}_i$, assuming that the estimates are perfect representations of the "true" unobservable utilities. Our objective, in contrast, is to find a set of group-level functions that maximizes our ability to explain the observed preferences of each consumer in the sample.

A problem similar to the one outlined above, but in a different context, was addressed by Bottenberg and Christal (1961). To find the best aggregation of the in-

Table 1
BOTTENBERG-CHRISTAL ARRANGEMENT OF INDIVIDUAL DATA TO BE AGGREGATED

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Subject A	3	−1	−1	−1	0	0	0	0	0	0
	3	−1	1	0	0	0	0	0	0	0
	5	−1	0	1	0	0	0	0	0	0
	4	1	−1	−1	0	0	0	0	0	0
	1	1	1	0	0	0	0	0	0	0
	1	1	0	1	0	0	0	0	0	0
Subject B	2	0	0	0	−1	−1	−1	0	0	0
	3	0	0	0	−1	1	0	0	0	0
	4	0	0	0	−1	0	1	0	0	0
	5	0	0	0	1	−1	−1	0	0	0
	2	0	0	0	1	1	0	0	0	0
	5	0	0	0	1	0	1	0	0	0
Subject C	4	0	0	0	0	0	0	−1	−1	−1
	5	0	0	0	0	0	0	−1	1	0
	3	0	0	0	0	0	0	−1	0	1
	1	0	0	0	0	0	0	1	−1	−1
	4	0	0	0	0	0	0	1	1	0
	3	0	0	0	0	0	0	1	0	1

dividual data, these authors suggested that the data be arranged as shown in Table 1. If a multiple regression equation is estimated with the data as displayed, one obtains the individual-level preference function for subject A from the coefficients of $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, for subject B from $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$, and so on. Now suppose we want to identify the best way to form two benefit segments, that is, to choose among (A,B)C, A(B,C), and (A,C)B. The preference functions for the two clusters in the (A,B)C combination can be estimated by first creating three new variables, $\mathbf{z}_1 = (\mathbf{x}_1 + \mathbf{x}_4)$, $\mathbf{z}_2 = (\mathbf{x}_2 + \mathbf{x}_5)$, and $\mathbf{z}_3 = (\mathbf{x}_3 + \mathbf{x}_6)$, then running the multiple regression of $\mathbf{y}$ as a linear function of $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{x}_7$, $\mathbf{x}_8$, $\mathbf{x}_9$.

Bottenberg and Christal demonstrate how this procedure can be used along with a hierarchical clustering algorithm to solve more general problems. A hierarchical clustering algorithm would require that every pair of clusters be evaluated as described above, at each clustering level. This procedure would be tantamount to running tens of thousands of multiple regressions, depending on the sample size. However, the authors point out that because the total sum of squared errors (SSE) can be obtained by adding the SSE's from independent regressions for each group, one need only compute the SSE for the pair being linked at each stage of the clustering procedure. Further, they show that this SSE can be obtained from the regression results of each group being linked. Though reducing considerably the computational requirements, these simplifications still require knowledge of the regression results (albeit from a smaller set of predictors) from each of the two clusters being evaluated to form a new cluster. An application of Bottenberg and Christal's procedure in a marketing context is reported by Srivastava, Leone, and Shocker (1981).

A Benefit Segmentation Procedure

The benefit segmentation procedure described next, though also based on the least square criterion, approaches the problem in a distinct way. Instead of estimating the segment-level functions as parts of a single linear regression across all individuals as in the Bottenberg-Christal procedure, we specify the clustering problem as the simultaneous estimation of a set of segment-level preference functions. In other words, the procedure forms clusters by estimating "pooled" regressions for each segment. Bottenberg and Christal's procedure, in contrast, identifies the segments as subject-attribute interactions in a single-equation model. We also demonstrate that for our particular purpose (in which all individual-level regressions use exactly the same predictors $\mathbf{X}$), a simpler criterion that reduces the computation requirements even further can be used.

First let us define the segmentation problem in more general terms. Define S as the number of benefit segments under consideration, N as the sample size, K as the number of parameters to be estimated for each segment, J as the number of stimulus profiles used in the conjoint experiment, and:

$\mathbf{X}$ = $(J \times K)$ design matrix as defined before.
$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] = (J \times N)$ matrix containing the preference vectors for each of the N individuals.
$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s] = (K \times S)$ matrix containing the regression intercept and preference weights for the S segments.
$\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N] = (S \times N)$ matrix of boolean vectors $\mathbf{g}_i$ defining the cluster membership for each subject i. The S-vector $\mathbf{g}_i$ contains a value of 1 in the row corresponding to the segment assigned to the individual i and zero otherwise.
$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N] = (J \times N)$ matrix containing random errors.

Then, the preference functions for each of the N individuals can be written as

$$\mathbf{Y} = \mathbf{XBG} + \mathbf{E}. \tag{1}$$

This equation specifies one preference function for each individual in such a way that all members of a particular segment have the same pooled function. Our objective is to determine the segment membership (matrix $\mathbf{G}$) and to estimate the pooled preference functions (parameters in the $\mathbf{B}$ matrix) with a minimum expected prediction error.

Let us first assume that the segmentation matrix $\mathbf{G}$ is known; OLS estimates of the preference functions then are

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}. \tag{2}$$

Note that $(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}$ are the OLS coefficient estimates at the individual level. The pooled estimates are thus equal to a weighted sum of the individual-level estimates, with $\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}$ as the weights; our goal is to

identify the weighting scheme that yields the highest predictive power.[1] Given the expression for **B** in equation 2, we can compute the disturbances **E** as

$$\mathbf{E} = (\mathbf{Y} - \mathbf{XBG}) \tag{3}$$
$$= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\mathbf{G}.$$

Hagerty (1986) has shown the expected prediction error (MSE) can be computed fairly accurately as

$$D_m = (1 - R_k^2) + (1 - R_k^2)\, m/n + (R_k^2 - R_m^2) \tag{4}$$

where:

$(1 - R_k^2)$ = unavoidable error that will occur even under the perfect estimation of the correctly specified model (using k parameter estimates),

$(1 - R_k^2)\, m/n$ = expected prediction error due to errors in estimating a misspecified model with $m \leqq k$ parameters from a sample of n observations,

$(R_k^2 - R_m^2)$ = bias due to the use of a misspecified model, and

R_k^2 and R_m^2 = the "true" population R^2's for the correctly specified and unspecified model, respectively.

The "true" proportion of variance explained $(1 - R^2)$ can be estimated by using Theil's (1961) adjusted R^2. The expected predictive MSE thus can be written as

$$D_m = (1 - R_k^2)\frac{(n-1)m}{(n-k-1)n} + (1 - R_m^2)\frac{n-1}{n-m-1}, \tag{5}$$

where R_k^2 and R_m^2 are the fitted R^2's with the full (k parameters) and restricted ($m \leqq k$ parameters) models, respectively.

In our benefit segmentation application, the most comprehensive model is the one that estimates one preference function (part worths) for each individual consumer, requiring the estimation of $k = K \times N$ parameters based on $n = J \times N$ observations. A given cluster solution with S clusters will lead to a more parsimonious and potentially misspecified model requiring $m = K \times S$ parameters. The MSE for this cluster solution is given by

$$D_s = (1 - R_k^2)\frac{(J*N - 1)}{(N(J-K)-1)}\frac{(K*S)}{(J*N)} + (1 - R_s^2)\frac{J*N - 1}{J*N - K*S - 1} \tag{6}$$

where $(1 - R_k^2)$ and $(1 - R_s^2)$ are the fitted MSE obtained from N individual-level and S segment-level functions, respectively.

Notice that for a fixed number of clusters, the only term in equation 6 that can vary is the fitted mean square error for the segment-level model $(1 - R_s^2)$. Therefore, the allocation matrix **G** that maximizes R^2 *for a fixed number of segments* also maximizes predictive accuracy at the aggregate level. The other elements in equation 6 are important in determining the appropriate number of segments, as discussed subsequently.

Because the total variance of the observed preference ratings **Y** is not affected by the segmentation scheme, the lowest unavoidable prediction error at a given aggregation level can be attained by the allocation matrix **G** such that

$$\min_{\mathbf{G}} \operatorname{tr}\{\mathbf{E}'\mathbf{E}\}.$$

In Appendix A this objective is shown to be equivalent to

$$\max_{(\mathbf{G})} \operatorname{tr}\{\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\mathbf{G}\mathbf{D}\} \tag{7}$$

where:

$\mathbf{D} = \mathbf{Y}'\hat{\mathbf{Y}}$

$\hat{\mathbf{Y}}$ = estimated preference ratings based on individual-level part worths.

A typical element of matrix **D**, d_{ij}, contains the cross-product of the *observed* preferences ($\mathbf{y}_i$) for subject i and the *fitted* preferences $\hat{\mathbf{y}}_j = \mathbf{X}\hat{\mathbf{C}}_j$ for subject j. Because the design matrix **X** is the same for all subjects, $\hat{\mathbf{y}}_j$ is also the vector of *predicted* preferences for subject i based on the preference function estimated for subject j. Therefore, d_{ij} measures the ability to predict preferences for subject i by using subject j's estimates, that is, the *cross-validity* of subject j's preference function relating to the observed preferences for subject i. We easily see that the clustering objective in equation 7 is such that pairs with strong cross-validity (large d_{ij}'s) tend to be allocated to the same clusters.[2]

At this point we must acknowledge that, as presented in equation 7, our evaluation criterion does not seem simple to compute. However, the classification matrix **G** has some distinctive characteristics that simplify considerably the computation of $\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\mathbf{G}$ in equation 7. For example, consider the classification of individuals A, B, C, D, E, and F into segments (A,B), (C,E,F), and (D) represented by the matrix **G**.

[1]Note that this expression of the pooled estimates is similar to Hagerty's (1985) Q-factor analytical estimates, except that in his case the cluster assignment matrix contains the eigenvectors of the matrix of correlations among individuals. These eigenvectors indicate the contribution of each consumer to the estimation of each segment-level preference function, thus assuming that, to some extent, each consumer belongs to every market segment. Note also that if **G** is an identity matrix, one will obtain the individual-level coefficient estimates.

[2]For example, the decision to join element i to j rather than k depends on whether

$$(\mathbf{y}_i'\hat{\mathbf{y}}_i + \mathbf{y}_j'\hat{\mathbf{y}}_j) + \mathbf{y}_i'\hat{\mathbf{y}}_j + \mathbf{y}_j'\hat{\mathbf{y}}_i > (\mathbf{y}_i'\hat{\mathbf{y}}_i + \mathbf{y}_k'\hat{\mathbf{y}}_k) + \mathbf{y}_i'\hat{\mathbf{y}}_k + \mathbf{y}_k'\hat{\mathbf{y}}_i$$

$$\mathbf{y}_j'\hat{\mathbf{y}}_j + \mathbf{y}_i'\hat{\mathbf{y}}_j + \mathbf{y}_j'\hat{\mathbf{y}}_i > \mathbf{y}_k'\hat{\mathbf{y}}_k + \mathbf{y}_i'\hat{\mathbf{y}}_k + \mathbf{y}_k'\hat{\mathbf{y}}_i.$$

Therefore, the choice of pairwise linkage depends on the strengths of the sum of squares and cross-products within the pairs.

(8)

		Individuals					
		A	*B*	*C*	*D*	*E*	*F*
G =		1	1	0	0	0	0
		0	0	1	0	1	1
		0	0	0	1	0	0
$\mathbf{G}'(\mathbf{GG}')^{-1}\mathbf{G}$ =		.5	.5	0	0	0	0
		.5	.5	0	0	0	0
		0	0	.33	0	.33	.33
		0	0	0	1	0	0
		0	0	.33	0	.33	.33
		0	0	.33	0	.33	.33

One can easily see that matrix $\mathbf{G}'(\mathbf{GG}')^{-1}\mathbf{G}$ has a clear pattern and can be derived directly from **G**, with no need for extensive matrix operations. A particular element $\{i, j\}$ of this $N \times N$ matrix will be equal to zero if the subjects identified by i and j do not belong to the same segment or $1/n_s$ if the two subjects belong to the same segment s of size n_s.

Finding **G** that maximizes the objective above is a nonlinear integer programming problem; the solution for each consumer (i.e., each column $\mathbf{g}_i$ of matrix **G**) is a boolean ($S \times 1$) vector to be chosen from a set of S possible vectors (columns from an $S \times S$ identity matrix). In a typical segmentation study, this integer programming problem involves a very large number of variables (one for each consumer in the sample), with an almost infinite number of feasible solutions. Also, a different programming problem must be solved for each number of clusters (S). Because an optimal solution based on the nonlinear integer programming model is not practical, a heuristic clustering algorithm is used to define the classification matrix **G**.

Two major approaches can be used to form nonoverlapping clusters, an agglomerative hierarchical method and an iterative partitioning (Punj and Stewart 1983). Agglomerative algorithms (Johnson 1967; Ward 1963) impose a hierarchical structure on the segments formed at different levels of aggregation such that any misclassification in a lower level will carry on to higher aggregation levels. Nevertheless, if a highly aggregated solution is used (the case in most benefit segmentation applications), this misclassification is relevant only if it occurs across the final segments being used.

Iterative partitioning algorithms, besides allowing for a nonhierarchical structure, tend to perform better than hierarchical algorithms *if* prior knowledge is available on the starting points (e.g., "parent members") for the partitioning process (Punj and Stewart 1983). If "good" starting points are not known *a priori* (i.e., random starting points are used), iterative partitioning leads to poor results (Milligan 1980). Funkhouser (1983) found that "because of the mechanism of the Howard-Harris algorithm [*K*-means], cluster solutions of large bases of survey data apparently are very sensitive to minor and absolutely trivial differences between data sets." Most importantly, the popular partitioning algorithms (*K*-means) are based on the nearest-centroid criterion, which is not compatible with our predictive accuracy criterion. Application of the *K*-means algorithm with the clustering criterion in equation 7 would require the use of a costly and cumbersome "hill-climbing" search. However, the risk of misclassification would remain and no clear advantages over the agglomerative approach would be gained. On the basis of the foregoing considerations and for convenience, an agglomerative clustering algorithm was chosen.

The proposed algorithm may seem similar to the "clustering regression" algorithm developed by Spath (1985). However, a close look at Spath's algorithm shows that it forms clusters of $\{\mathbf{y}_i, \mathbf{x}_i\}$ sets, where i refers to one *independent* observation and the $(1 \times L)$ vector $\mathbf{x}_i$ represents the predictors for that particular observation. In a conjoint experiment, in contrast, we work with well-defined *sets of observations* (i.e., preference ratings and predictors for all stimuli, for each subject). In such cases, the dependent and predictor variables for one subject i are in fact a $(K \times 1)$ vector $\mathbf{y}_i$ and a $(K \times L)$ array $\mathbf{X}_i$, respectively, where K is the number of stimuli rated by the subject. In other words, the proposed benefit segmentation algorithm considers an additional dimension, representing the fixed (by design) grouping of observations within each subject.

Choosing the Adequate Aggregation Level

The clustering objective in equation 7, though convenient as a criterion for selecting pairwise linkages for a predetermined number of segments, may not have an intuitive meaning. Also, as discussed before, some adjustments should be made to the sum of squared errors to transform it into an effective measure of predictive accuracy. Therefore, after the best allocation matrix **G** is found for a predetermined number of segments, the predictive accuracy of the estimated preference functions can be assessed by

$$(9)\quad D_s = (1 - \mathrm{tr}\{\mathbf{D}\}/\mathrm{tr}\{\boldsymbol{\Sigma}\})\frac{(J*N-1)}{(N(\mathrm{J}-\mathrm{K})-1)}\frac{(K*S)}{(J*N)} + (1 - \mathrm{tr}\{\mathbf{G}'(\mathbf{GG}')^{-1}\mathbf{GD}\}/\mathrm{tr}\{\boldsymbol{\Sigma}\})\left[\frac{J*N-1}{J*N-K*S-1}\right]$$

where J, N, K, **D**, **G** are as defined before and $\boldsymbol{\Sigma}$ is the covariance matrix of the observed ratings **Y**.

The predictive accuracy index (PAI) computed in equation 9 provides an intuitive criterion for deciding how many benefit segments to retain. This index estimates the proportion of variance (of the preference ratings) expected to remain unexplained if the segment-level functions are used to predict the respondents' ratings for a holdout set of stimuli. This decision criterion makes the proposed clustering algorithm distinct from the usual two-stage approach. Using equation 9, the researcher will have an estimate of the *maximum predictive accuracy* to be expected at a given aggregation level. In contrast, the R^2 and statistical tests obtained from the minimum-vari-

ance and minimum-distance clustering algorithms commonly used in the two-stage approach are related to the between-groups and within-groups variances in the input variables (i.e., part worth estimates). The latter statistics are measures of *similarity* of the part worth *estimates* within each segment; the R^2 indicates how well the cluster averages fit the individual-level part worth estimates, ignoring estimation errors. Our criterion measures the ability to predict preference ratings for a holdout set.[3]

COMPARING THE PROPOSED AND TWO-STAGE PROCEDURES

As demonstrated before, the proposed segmentation algorithm seeks a matrix of weights $\mathbf{A} = \mathbf{G}'(\mathbf{GG}')^{-1}\mathbf{G}$ that will lead to the best expected predictive accuracy or, equivalently, that maximizes $\text{tr}\{\mathbf{AY}'\hat{\mathbf{Y}}\} = \text{tr}\{\mathbf{AY}'\mathbf{XC}\}$, where $\mathbf{C}$ is the matrix of subject-level OLS estimates. With further algebraic manipulation we also can show that $\text{tr}\{\mathbf{AY}'\hat{\mathbf{Y}}\} = \text{tr}\{\mathbf{AC}'\mathbf{X}'\mathbf{XC}\}$.[4] The minimum-variance algorithm used in the two-stage procedure, in contrast, seeks the set of weights that minimizes the total within-clusters variance of the OLS estimates, given by $\min \text{tr}\{(\mathbf{C}\text{-}\mathbf{CA})'(\mathbf{C}\text{-}\mathbf{CA})\} = \text{tr}\{\mathbf{C}'\mathbf{C}\} - \text{tr}\{\mathbf{AC}'\mathbf{C}\}$.

Because $\text{tr}\{C'C\}$ does not depend on the cluster solution $\mathbf{A}$, the ultimate objective in the two-stage procedure is to maximize $\text{tr}\{\mathbf{AC}'\mathbf{C}\}$. Whereas the two-stage procedure joins clusters with the highest sum of cross-products of part worth estimates ($c_i' c_j$), the proposed algorithm joins clusters with the highest cross-validity ($\mathbf{y}_i' \hat{\mathbf{y}}_j$). The two cluster solutions are identical only if the design matrix $\mathbf{X}$ is perfectly orthogonal. However, the degree of collinearity of the binary predictors in $\mathbf{X}$ is somewhat high, even when the conjoint design is orthogonal. In fact, the correlation among the levels of a treatment is always equal to .5, except for two-level treatments. In other words, with the exception of orthogonal designs with only two-level treatments, the proposed procedure is distinct from the two-stage procedure.

The degree of collinearity would be even more accentuated in nonorthogonal designs (sometimes used because of environmental correlations between attributes), where one would also find correlations between treatments. The practical implication is that collinearity leads to unreliable subject-level estimates $\mathbf{C}$, while not affecting the predictions $\hat{\mathbf{Y}}$. Because the two-stage procedure is based on $\mathbf{C}'\mathbf{C}$ and the proposed procedure uses $\mathbf{Y}'\hat{\mathbf{Y}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, this suggests another advantage of the latter over the former.

The preceding results indicate the criterion used by the proposed algorithm for identifying segments is distinct from the one used by the two-stage procedure. In the next section, we compare the predictive performance of these two procedures to show the superiority of the former.

Comparison Based on Synthetic Data

Six benefit segments containing five hypothetical consumers each were constructed with the following underlying utility functions.

$$(11) \quad U_j = W_0 X_1 + W_1 D_1 + W_2 D_2 + W_3 D_3 + W_4 D_4$$

Segment	W_0	W_1	W_2	W_3	W_4
A	−.3	0	1	0	1
B	−.1	0	3	0	3
C	−.3	1	0	0	1
D	−.3	0	1	1	1
E	−.3	0	−1	0	−1
F	.1	0	−3	0	−3

A Latin-square design was used to create nine stimulus profiles combining the continuous attribute X_1 in three levels (−10, 0, 10) and two nominal attributes in three levels each, represented by the effects-type-coded (Cohen and Cohen 1975) dummy variables D_1 through D_4.

Preference ratings for the nine stimuli were generated by adding a random, normally distributed disturbance to equation 11 under six different error conditions: 3%, 8%, 10%, 14%, 19%, and 25% error (as a proportion of the total variance in the preference ratings). The proposed segmentation algorithm and the clustering of subject-level OLS estimates (two-stage procedure) then were applied to these simulated preference ratings.

Figure 1 compares the fit of the two procedures. The curves show the minimum attainable error to be expected if the cluster solutions are used to predict preference ratings for a holdout sample. Obviously, the two procedures fit equally at the two extremes (aggregate and individual-level solutions). As expected, the proposed algorithm fits the estimation sample better than the clustering regression coefficients.

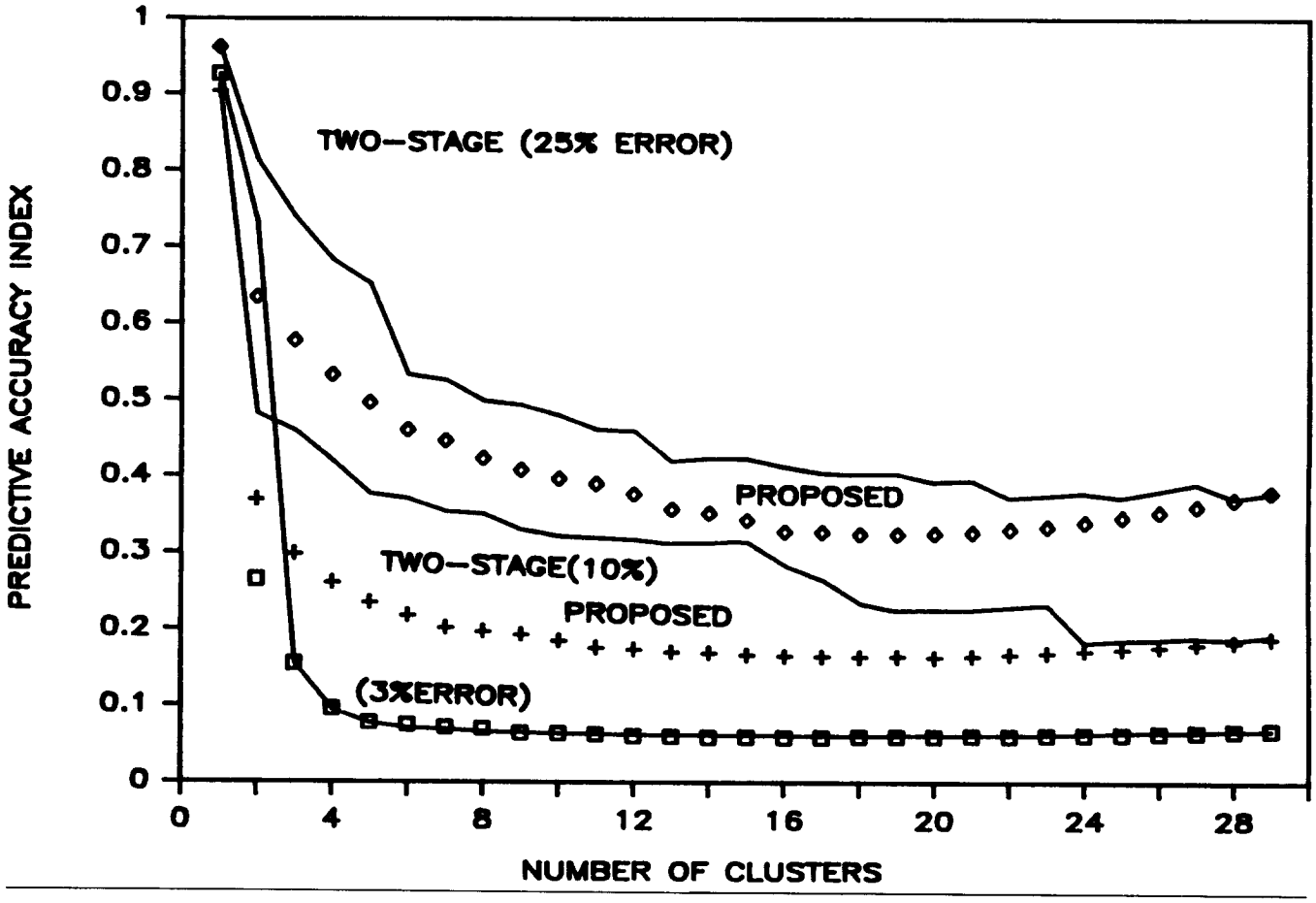
The plot of expected predictive accuracy for the proposed algorithm shows a substantial improvement in expected predictive fit between the aggregate solution and the two-cluster solution. Under the lower error condition (3%), Figure 1 shows an "elbow" at four clusters, rather than the six generated segments. This result is explained by the similarity of the "true" utility functions of segments A, C, and D. The same elbow rule leads to a smaller number of segments for the higher error conditions (e.g. 10%, 25%). This result can be explained by the fact that under a higher error variance, the distinctions among segments A, C, and D and between E and F become more diffused by the random error.

These results show only that the proposed procedure fits the estimation sample better than the two-stage procedure, which should be expected because it uses the

[3]The predictive accuracy index, as well as the similarity measures used in other procedures, is obtained through an iterative search at each clustering level. This process may capitalize on random data errors and overstate the true performance of the model in a way similar to the measures of fit in stepwise regression (McIntyre et al 1983).

[4]My thanks to Gary Russell and one anonynmous reviewer for pointing this out.

Figure 1
COMPARISON OF FIT OF TWO-STAGE AND PROPOSED PROCEDURES



degree of fit as the clustering criterion. Therefore, the "litmus test" is to compare the ability of the two procedures to predict preferences for a holdout set of stimuli (i.e., their predictive validity).

The segment-level utility functions estimated with the first (estimation) sample were used to predict preferences for a holdout set of nine stimuli, distinct from the estimation set. As shown in Table 2, the segmentation scheme and aggregate utility functions obtained with the proposed procedure have stronger predictive validity. The differences in fit to the holdout set between the two methods are substantial, especially in the high error conditions.

The preceding results show that, under the usual assumptions of least squares estimation, the aggregate utility functions estimated by the proposed procedure are closer to the "true" underlying utility functions of each individual than the ones obtained with the two-stage procedure. Next, these procedures are compared in the context of an actual conjoint experiment

Comparison Based on Empirical Data

The empirical data used in the comparison were part of a study undertaken by a major financial institution for redesigning its checking account services. Details of this study are reported elsewhere (Jain et al. 1979). For our purposes, it suffices to know that:

—Five determinant attributes were selected on the basis of an exploratory study and were conceptualized at three levels each. The attributes and their levels are described in Table 3.
—Twenty-seven (full-profile) verbal descriptions of hypothetical checking accounts were generated, forming a main-effects fractional factorial design of the five determinant attributes.
—A holdout set of eight profiles was created with different combinations of the same five attributes.
—Preference data for the 27 estimation sets, as well as the eight holdout profiles, were collected from a random sample of 105 consumers in a major northeastern metropolitan area.

Table 2
FIT (R^2) TO NINE HOLDOUT PROFILES

Error (%)	Model	Number of clusters 6	5	4	3	2	1
3	Proposed	88	.89	.89	.84	73	.07
	Two-stage	.89	90	89	84	.26	07
8	Proposed	73	74	76	72	.62	10
	Two-stage	.52	46	.30	25	.26	10
10	Proposed	.66	.69	.70	.66	58	.07
	Two-stage	.40	41	.39	40	41	.07
14	Proposed	.48	45	46	42	39	.06
	Two-stage	.32	28	.20	23	14	.06
19	Proposed	.40	41	.39	41	32	.06
	Two-stage	.28	.26	.19	19	.16	06
25	Proposed	15	.18	.20	22	24	03
	Two-stage	06	.01	.00	00	.00	03

Figure 2 is a plot of the PAI for the last 10 cluster solutions. It indicates that a single utility function would leave more than 61% of the variance unexplained if used to predict preferences for a holdout set of stimuli. Using two segments would lower the unexplained variance to 50% or more, a substantial improvement in accuracy. Segmenting the sample into three clusters would improve expected accuracy by a comparable amount, to 43% unexplained variance. Adding a fourth segment would reduce unexplained variance to 40%. The decision to use the three-cluster or four-cluster solution depends on the researcher's judgment of whether the 3% gain in predictive accuracy is worth the additional complexity in the model. Though the plot in Figure 2 and the similarity measures used in the two-stage procedure do not show a clear "elbow," a three-cluster solution was used because of the small gain in predictive accuracy with the four-cluster solution.

Table 3
CHECKING ACCOUNT ATTRIBUTES

Attribute name	Attribute levels
A Cost of checking account	1 15¢ a check
	2 $200 minimum balance in a checking or savings account *all* the time
	3 Absolutely free checking service
B. Type of bank	1. Commercial bank with headquarters outside City A
	2 Commercial bank with headquarters in City A
	3. Savings bank with headquarters in City A
C Accessibility to banking service	1. 15-minute drive from home
	2. 10-minute drive from home
	3. 5-minute drive from home
D Quality of service	1 Service is *less* friendly than average. Bank personnel are less likely to go out of their way to help you than those in most banks
	2. Service is *average* in friendliness Bank personnel are average in their willingness to go out of their way to help you in comparison with those in most banks.
	3. Service is *above average* in friendliness Bank personnel are more likely to go out of their way to help you than those in most banks
E. Hours	1 Weekdays 9:00am–4:00pm and evenings 5:00–8:00pm twice a week
	2 Weekdays 9:00am–4:00pm, evenings 5:00–8:00pm twice a week, and Saturdays 9:00am–12:00 noon
	3. Monday through Saturday 9:00am–10:00pm

The goodness of fit to the observed preferences for the three-cluster solution is the same (R^2 = .57) for both approaches, though there is some disagreement in the classification of the 105 consumers into the three segments. Twelve percent of the 105 consumers were allocated to different segments by the two procedures. The aggregate part worths for each segment are compared in Table 4.

In general, the part worth estimates from the two procedures lead to the same conclusions: segment A places the highest importance on service, B and C find cost to be most important, segment B assigns the lowest utility to the $200 minimum balance, and C dislikes most the 15¢ charge per check. Though the part worth estimates obtained through the two procedures are comparable, some discrepancies lead to small differences in fit to the eight holdout profiles. These differences are shown in the top rows of Table 5. Further analysis of the raw data indicated that one of the holdout stimuli (profile 8) had been ranked consistently as the most preferred, regardless of segment (by 83% of the sample). A similar pattern was found for profile 7, which had been ranked second or third by 87% of the sample. These unexpected results occurred because the two profiles had extremely favorable combinations of attributes, regardless of the segments, which could attenuate the differences in predictive fit between the two procedures being tested. Therefore, the predictive validity test was replicated after exclusion of profiles 7 and 8. The results are reported in Table 5.

Though the differences in predictive validity are not large for this particular application, the proposed procedure is consistently better than the two-stage procedure over different aggregation levels. Evidently, the differences in predictive validity depend on the nature of the holdout stimuli. For example, a hypothetical account offering free checks with above-average service at a branch within 5 minutes' driving time and open on weekdays, Saturdays, and evenings would be most preferred by any of the segments identified by either of the two procedures. Holdout stimuli such as this conceal the differences between the two procedures.

A pairwise comparison of fit between the two cluster solutions across all 105 subjects showed that the difference in predictive fit is statistically significant (at the .10 level) in all comparisons, except for the two-cluster solution.

Figure 2
PREDICTIVE ACCURACY INDEX: BENEFIT SEGMENTATION FOR BANK DATA

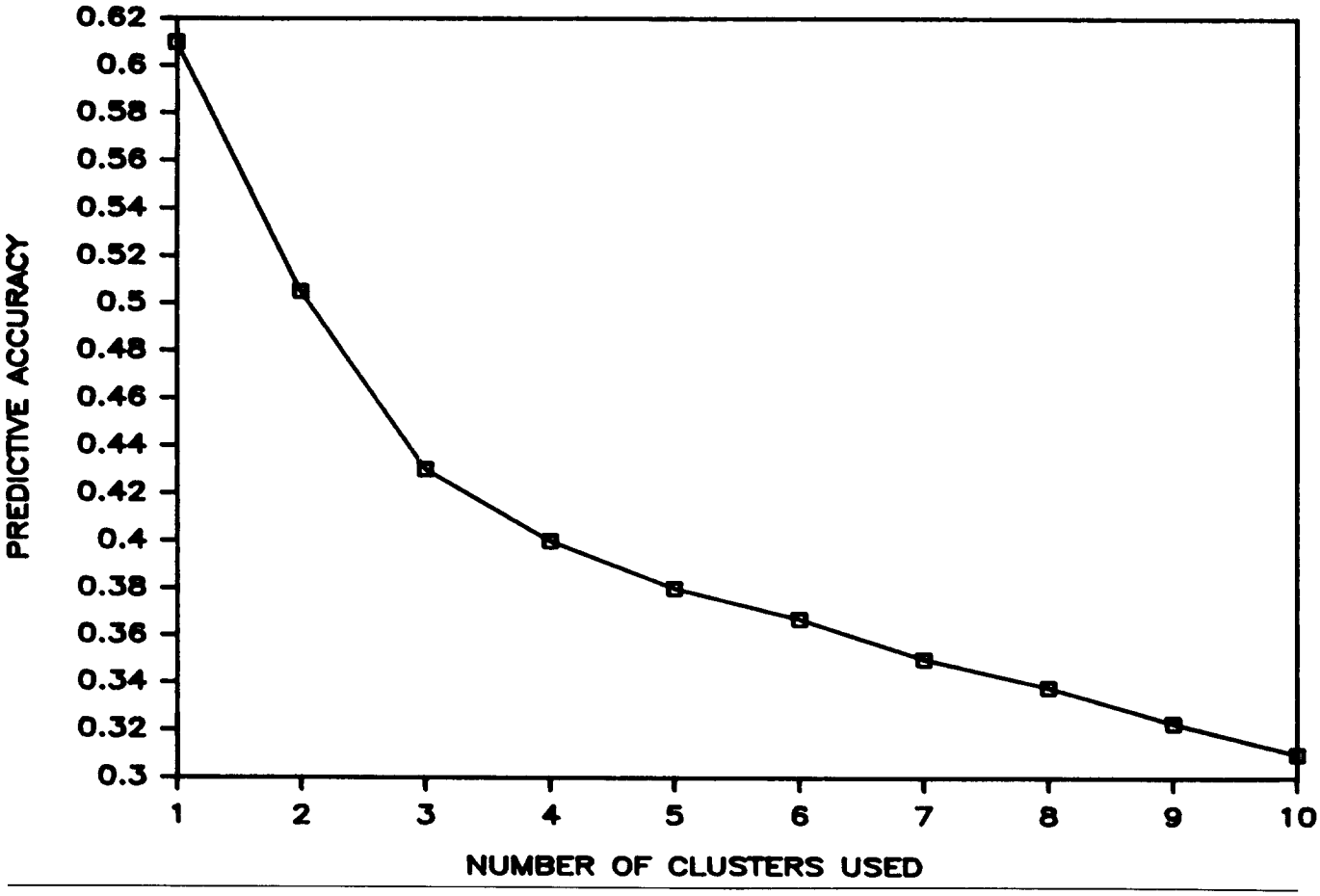


Table 4
AGGREGATE PART WORTH ESTIMATES

		Proposed procedure			*Two-stage procedure*		
		Seg A	*Seg. B*	*Seg. C*	*Seg A*	*Seg. B*	*Seg C*
Cost	15¢/check	− 67	29	−7 78	− 04	−.24	−7 51
	$200 bal.	−1 56	−7.40	0 00	−1 99	−7.63	− 20
	Free	2.23	7 11	7 78	2 03	7 87	7 71
Type	Bank out	−2 05	−1 18	− 63	−2 15	−1 14	− 48
	Bank in	1 32	45	60	1.50	34	.35
	S&L	73	73	03	65	80	.13
Access	15 min	−1 35	− 97	− 88	−1 12	−1 32	−.79
	10 min	01	− 26	− 21	− 03	− 21	− 25
	5 min.	1 34	1 23	1 09	1 15	1 53	1 04
Service	<avg.	−4 99	−2 48	−1 49	−4 69	−2.33	−1 91
	Average	84	64	29	71	.47	67
	>avg	4 15	1 84	1 20	3 98	1.86	1 24
Hours	Week + ev.	−.21	−.88	− 11	− 49	− 61	− 12
	Wk + Sat. + ev	59	06	.06	52	11	06
	Week + Sat	− 38	82	05	−.03	50	.06
Segment size		41	37	27	44	32	29

Table 5
FIT (R^2) TO THE HOLDOUT SET

Holdout set		Number of clusters 1	2	3	4	5
All eight profiles	Proposed	562	545	576	588	.585
	Two-stage	.562	528	.549	.554	.553
Excluding profile 8	Proposed	455	.445	.492	.502	.498
	Two-stage	455	.423	460	465	.462
Excluding profiles 7, 8	Proposed	340	356	.425	.434	.430
	Two-stage	.340	323	.392	.400	397

CONCLUSIONS

The proposed procedure for conjoint-based benefit segmentation approaches the aggregation problem in a different way than does the usual two-stage procedure. Rather than grouping consumers on the basis of their similarity in terms of part worth *estimates* (i.e., ignoring the estimation errors), we form segments that maximize the predictive validity of the segment-level part worth estimates—that is, they are adequate representations of the "true" underlying utility function of each member of the sample.

However, the ultimate objective is not maximum predictive power, which could be attained by other means such as Hagerty's (1985) optimal weighting procedure or a highly disaggregated solution. In benefit segmentation the manager seeks a balance between a meaningful and manageable set of segments (and their utility functions) and an acceptable predictive power. The aggregate part worth estimates and predictive accuracy index computed at each aggregation level can assist the manager in choosing an acceptable balance.

Because cluster identification and part worth estimates are obtained in a single stage, the proposed algorithm is also simpler to use. In fact, the experimental Turbo-Pascal implementation of the algorithm for IBM PC's (available upon request) automatically displays cluster membership, predictive validity index, part worth estimates, and plots of the utility functions for each benefit segment. The usual two-stage procedure, in contrast, requires the estimation of part worths for each individual and their input into a clustering algorithm. Despite the considerable number of pairwise linkages that must be evaluated in the hierarchical clustering algorithm, the computations can be performed within a feasible time (the analysis of the 105 individuals in the empirical illustration required approximately 74 seconds of CPU time on a VAX 8800 computer).

Because a hierarchical procedure was chosen to form the segments, combining two individuals or clusters to form a new cluster forces them to be in the same cluster in the latter stages of the algorithm. Any misclassification in an earlier stage of the algorithm will be carried on to the higher aggregation levels, as in any hierarchical heuristic. Nevertheless, such misclassification in the early stages becomes irrelevant if a highly aggregated solution is used and the earlier misclassification occurred within the same cluster. Further, the same risk of suboptimal solutions is present in other clustering heuristics (K-means, minimum variance) and in the usual two-stage procedure. A more adequate solution to this problem could be obtained from the nonlinear integer programming model in equation 8 for different aggregation levels, as indicated before. Unfortunately, for a typical segmentation study, the problem reaches unfeasible dimensions.

Finally, the hierarchical formation of segments implies that at each aggregation level a consumer can belong to only one segment. Though one may find situations in which overlapping segments would be most appropriate, the proposed procedure gives the researcher an indication of how inadequate the nonoverlapping solution is, estimating the loss of predictive power incurred by forcing consumers into "homogeneous" mutually exclusive segments.

APPENDIX
DERIVATION OF THE CLUSTERING CRITERION

To simplify the presentation, define the idempotent symmetric matrices (i.e., $\mathbf{Z'Z} = \mathbf{Z}$ and $\mathbf{A'A} = \mathbf{A}$):

$$\text{(A1)} \qquad \mathbf{Z} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$$

$$\mathbf{A} = \mathbf{G'}(\mathbf{GG'})^{-1}\mathbf{G}.$$

Then the sums of squared errors and cross-products are equal to

$$\text{(A2)} \qquad \mathbf{E'E} = (\mathbf{Y} - \mathbf{ZYA})'(\mathbf{Y} - \mathbf{ZYA})$$

$$= \mathbf{Y'Y} - 2\mathbf{A'Y'Z'Y} + \mathbf{A'Y'Z'YA},$$

and our objective can be rewritten as

$$\text{(A3)} \qquad \min_{(G)} [\text{tr}\{\mathbf{Y'Y}\} - 2\text{tr}\{\mathbf{A'Y'Z'Y}\} + \text{tr}\{\mathbf{A'Y'Z'YA}\}].$$

Because $\mathbf{Y'Y}$ does not depend on the allocation matrix $\mathbf{G}$ and because $\text{tr}\{\mathbf{A'Y'Z'YA}\} = \text{tr}\{\mathbf{A'Y'Z'Y}\}$, equation A3 can be simplified to[5]

$$\text{(A4)} \qquad \max_{(G)} \text{tr}\{\mathbf{A'Y'Z'Y}\} = \max_{(G)}$$

$$\text{tr}\{\mathbf{G'}(\mathbf{GG'})^{-1}\mathbf{GY'X}(\mathbf{X'X})^{-1}\mathbf{X'Y}\},$$

or finally

$$\text{(A5)} \qquad \max_{(G)} \text{tr}\{\mathbf{G'}(\mathbf{GG'})^{-1}\mathbf{GD}\}$$

where $\mathbf{D} = \mathbf{Y'}\hat{\mathbf{Y}}$.

REFERENCES

Bottenberg, Robert A. and Raymond E. Christal (1961), "An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency," Technical Note WADD-TN-61-30, Personnel Laboratory, Wright Air Development Division, Lackland Air Force Base.

[5] $\text{tr}\{\mathbf{A'Y'Z'YA}\} = \text{tr}\{\mathbf{AA'Y'Z'Y}\}$ and because $\mathbf{A}$ is idempotent, $\text{tr}\{\mathbf{A'Y'Z'YA}\} = \text{tr}\{\mathbf{A'Y'Z'Y}\}$.

Cattel, Raymond B. (1978), *The Scientific Use of Factor Analysis in the Behavioral and Life Sciences*. New York: Plenum Press

Cattin, Philippe and Dick R. Wittink (1982), "Commercial Use of Conjoint Analysis: A Survey," *Journal of Marketing*, 46 (Summer) 44–53

Chow, George C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28 (July), 591–605.

Cohen, Jacob and Patricia Cohen (1975), *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences* Hillsdale, NJ: Lawrence Erlbaum Associates.

Currim, Imram S. (1981), "Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models," *Journal of Marketing Research*, 18 (August), 301–9.

Funkhouser, G. Ray (1983), "A Note on the Reliability of Certain Clustering Algorithms," *Journal of Marketing Research*, 20 (February), 99–102.

Green, Paul E. (1974), "On the Design of Choice Experiments Involving Multifactor Alternatives," *Journal of Consumer Research*, 1 (September), 61–8.

——— and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103–23.

———, Yoram Wind, and Arun K. Jain (1972), "Preference Measurement of Item Collections," *Journal of Marketing Research*, 9 (November), 371–7.

Hagerty, Michael R. (1985), "Improving the Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis," *Journal of Marketing Research*, 22 (May), 168–84

——— (1986), "The Cost of Simplifying Preference Models," *Marketing Science*, 5 (Fall), 298–319.

Haley, Russell (1968), "Benefit Segmentation," *Journal of Marketing*, 32 (April), 30–5.

Howard, H. and B. Harris (1966), *A Hierarchial Grouping Routine IBM 360165 FORTRAN V Program*, University of Pennsylvania Computer Center.

Huber, Joel C. and William L. Moore (1979), "A Comparison of Alternative Ways to Aggregate Individual Conjoint Analyses," in *Educators' Conference Proceedings*, Neil E. Beckwith et al., eds. Chicago: American Marketing Association, 64–8.

Jain, Arun K., Franklin Acito, Naresh K. Malhotra, and Vijay Mahajan (1979), "A Comparison of the Internal Validity of Alternative Parameter Estimation Methods in Decompositional Multiattribute Preference Models," *Journal of Marketing Research*, 16 (August), 313–22.

Johnson, Richard M. (1974), "Trade-off Analysis of Consumer Values," *Journal of Marketing Research*, 11 (May), 121–7.

Johnson, Stephen C. (1967), "Hierarchial Clustering Schemes," *Psychometrika*, 32 (September), 241–54

Judge, G. G., W. E Griffiths, R. C Hill, and T. C. Lee (1980), *The Theory and Practice of Econometrics*. New York: John Wiley & Sons, Inc.

McIntyre, Shelby H., David B. Montgomery, V Srinivasan, and Barton A. Weitz (1983), "Evaluating the Statistical Significance of Models Developed by Stepwise Regression," *Journal of Marketing Research*, 20 (February), 1–11.

Milligan, G. W. (1980), "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika*, 45 (September), 325–42.

Moore, William L. (1980), "Levels of Aggregation in Conjoint Analysis: An Empirical Comparison," *Journal of Marketing Research*, 17 (November), 516–23.

Punj, Girish and David W Stewart (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 20 (May), 134–48.

Spath, Helmuth (1985), *Cluster Dissection and Analysis*. New York: Springer-Verlag

Srivastava, Rajendra K., Robert P. Leone, and Allan D. Shocker (1981), "Market Structure Analysis. Hierarchical Clustering of Products Based on Substitution-in-Use," *Journal of Marketing*, 45 (Summer), 38–48

Stewart, David W. (1981), "The Application and Misapplication of Factor Analysis in Marketing Research," *Journal of Marketing Research*, 17 (February), 51–62.

Theil, Henri (1961), *Principles of Econometrics*. New York: John Wiley & Sons, Inc.

Ward, Joe H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association* (March), 236–44.

Wind, Yoram (1978), "Issues and Advances in Segmentation Theory," *Journal of Marketing Research*, 15 (August), 317–37