PSYCHOMETRIKA—VOL. 60, NO. 1, 47–75
MARCH 1995

A PARAMETRIC PROCEDURE FOR ULTRAMETRIC TREE ESTIMATION FROM CONDITIONAL RANK ORDER PROXIMITY DATA

MARTIN R. YOUNG

STATISTICS DEPARTMENT
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN

WAYNE S. DESARBO

MARKETING AND STATISTICS DEPARTMENTS
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN

The psychometric and classification literatures have illustrated the fact that a wide class of discrete or network models (e.g., hierarchical or ultrametric trees) for the analysis of ordinal proximity data are plagued by potential degenerate solutions if estimated using traditional nonmetric procedures (i.e., procedures which optimize a STRESS-based criteria of fit and whose solutions are invariant under a monotone transformation of the input data). This paper proposes a new parametric, maximum likelihood based procedure for estimating ultrametric trees for the analysis of conditional rank order proximity data. We present the technical aspects of the model and the estimation algorithm. Some preliminary Monte Carlo results are discussed. A consumer psychology application is provided examining the similarity of fifteen types of snack/breakfast items. Finally, some directions for future research are provided.

Key words: hierarchical clustering, proximity data, conditional rank orders, maximum likelihood estimation, consumer psychology.

1. Introduction

An ultrametric or hierarchical tree is a rooted tree in which a nonnegative weight is assigned to each node such that (a) the terminal nodes have zero weight, (b) the root has the largest weight, and (c) the weights assigned to the nodes on the path from any terminal node to the root constitute a strictly increasing sequence (De Soete, 1984). The ultrametric tree distance between two nodes i and j, denoted as d_{ij}, is defined in such discrete representations as the maximum of the weights associated with the nodes on the path connecting nodes i and j. Such ultrametric trees have been quite useful for representing the discrete structure in proximity data since a hierarchical clustering is defined on the object set. Let $\underline{\Delta} = ((\delta_{ij}))$ be a square symmetric matrix containing the pairwise, nonnegative, observed dissimilarities between M objects; then an ultrametric tree Π is a representation of $\underline{\Delta}$ whenever its terminal nodes correspond in a one-to-one fashion with the M objects, and whenever for each (i, j) pair of objects, d_{ij}, the ultrametric distance between the two nodes corresponding to objects i and j, approximately equals δ_{ij}. If $d_{ij} = \delta_{ij}$ for all (i, j), then Π constitutes an exact ultrametric tree representation of $\underline{\Delta}$ (De Soete, 1984). Hartigan (1967), Jardine, Jardine, and Sibson (1967), and Johnson (1967) have all independently demonstrated that a necessary and

Requests for reprints should be sent to Martin R. Young, Statistics Department, School of Business Administration, University of Michigan, Ann Arbor, MI 48109-1234.

0033-3123/95/0300-93111$00.75/0

sufficient condition for the existence of an exact ultrametric tree representation is the ultrametric inequality, where if $\underline{\Delta}$ satisfies:

$$\delta_{ij} \leq \max(\delta_{ik}, \delta_{jk}) \tag{1}$$

for all i, j, and k triples, then an exact ultrametric tree representation can be uniquely constructed.

However, empirical metric proximity data rarely satisfy this strong property, and a wide variety of estimation heuristics have therefore been developed. Traditional hierarchical clustering methods employing various types of agglomerate or divisive rules (see Hartigan, 1975, for a review) such as single-linkage, complete-linkage, average-linkage, centroid-linkage, or median-linkage, are available in most statistical software packages. Unfortunately, different hierarchical clustering procedures applied to the same set of proximity data typically lead to different tree structures, as has been well documented in the classification literature (Dubes & Jain, 1979). Furthermore, there is rarely any a priori theory or guidelines as to which particular hierarchical clustering procedure to use for a specific social science research problem. In addition, with the exception of Ward's (1963) method, few of the commonly available hierarchical clustering procedures optimize any clear objective function (e.g., a least-squares fit to the $\underline{\Delta}$), and so the properties of the resulting ultrametric trees derived are unknown. For these reasons, a number of psychometricians have proposed new approaches for estimating a best fitting ultrametric tree representation for a given set of proximity data. For example, such an ultrametric tree representation can be constructed by minimizing the least-squares loss function:

$$Z = \sum_{i<j} (\delta_{ij} - d_{ij})^2. \tag{2}$$

Hartigan (1967) proposed a combinatorial optimization procedure in attempting to minimize Z in (2) by performing a number of local operations on the tree. Chandon, Lemaire, and Pouget (1980) developed a branch and bound estimation procedure that globally minimizes (2) in finding the best ultrametric tree. Note that both the Hartigan and Chandon, Lemaire, and Pouget procedures can only be applied to relatively small data sets ($M \leq 10$ or so) given the heavy numerical computation burden implicit in these combinatorial optimization approaches. Carroll and Pruzansky (1975, 1980) proposed a mathematical programming approach to estimating least-squares ultrametric trees from $\underline{\Delta}$ employing a penalty function approach to gradually enforce the "strong" ultrametric inequality constraints. This methodology models the discrete, constrained tree estimation problem by a sequence of continuous gradient-based unconstrained optimization iterates. De Soete (1984) later modified this penalty function approach by using a computationally more efficient penalty function, adopting an exact sequential unconstrained minimization framework, and applying a numerically more stable nonlinear minimization method for solving the unconstrained subproblem. DeSarbo, De Soete, Carroll, and Ramaswamy (1988) present an ultrametric tree estimation procedure for paired comparisons data in a stochastic model framework. DeSarbo, Manrai, and Burke (1990) discuss an extension to accomodate asymmetric proximity data (see also DeSoete et al., 1984a, 1984b). Barthélemy and Guénocke (1991) review these and other more recent approaches for estimating ultrametric trees from metric proximity data; see also DeSarbo, Manrai, & Manrai.

While there seemingly exists a plethora of ultrametric tree estimation procedures for the analysis of metric proximity data, or proximities created from the preprocessing of two-way dominance data, no known nonmetric procedure exists for ultrametric tree

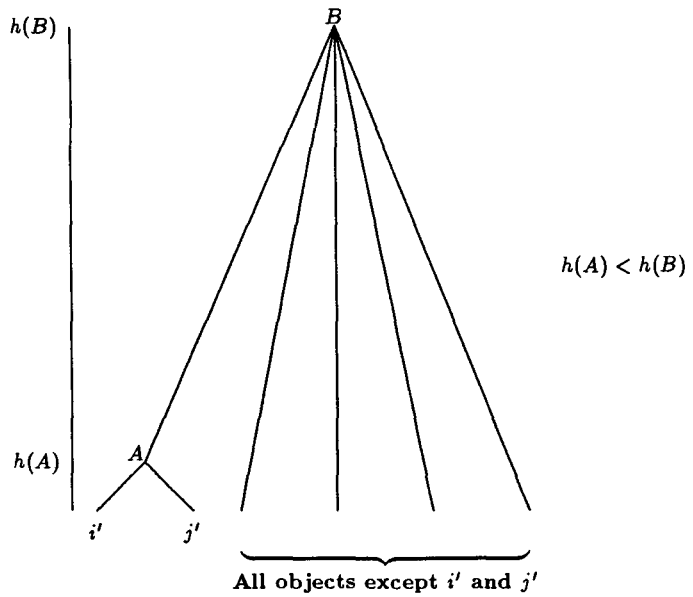

FIGURE 1.
Carroll's (1989) Degenerate Ultrametric Tree Illustration.

estimation from ordinal proximity data. Indeed, De Soete (1983a) was among the first to demonstrate that neither the quantitative (i.e., the magnitudes of the distances) nor the qualitative (i.e., the topological structure) properties of such tree representations (he examined additive trees—a more general case of ultrametric trees) are invariant under monotonic transformations of the data, and concluded that "it does not make very much sense to devise nonmetric algorithms for constructing additive-tree representations" (p. 476). Carroll (1989) demonstrated that a wide class of tree structure models for proximity data (both ultrametric and additive trees) are subject to intrinsic theoretical degenerate solutions if fit by standard nonmetric techniques whose solutions are invariant under a monotone transformation of the data, and which optimize a STRESS or STRESS-like criterion of fit. Assuming that the smallest dissimilarity is unique or untied, Carroll illustrated the degenerate solution derived by connecting the closest pair as terminal nodes in the ultrametric tree to a single internal node (A), which in turn attaches to a second higher internal node (B), to which all other terminal nodes are connected. Figure 1, taken from Carroll, illustrates the degenerate ultrametric tree that theoretically plagues fully nonmetric fitting procedures such that a (weak, but non-constant) monotone function of the data will agree precisely with the distances of the ultrametric tree. Winsberg and Carroll (1989) extend this argument to fully nonmetric procedures for fitting the INDSCAL (Carroll & Chang, 1970) multidimensional scaling model, and employ monotone splines as a way of avoiding potential degenerate solutions.

This paper proposes a parametric approach to estimating ultrametric tree structures from conditional rank order data that avoids the invariance and degeneracy problems associated with fully nonmetric procedures mentioned above. The next section describes the procedure by which the underlying model parameters—the tree configuration $((d_{ij}))$ and an associated scale parameter—are jointly estimated. Section 3 describes a small Monte Carlo study of the estimation procedure, the results of which show that the technique is able to accurately reconstruct tree configurations for moderate sized datasets. Section 4 describes the application of this procedure to conditional

rank order data collected on various breakfast/snack foods by Green and Rao (1972). It is shown that the method provides an accurate graphical summary of the structure in these data. Finally, section 5 offers some suggestions for future research extensions, including the incorporation of covariates and a latent class estimation framework.

2. Methodology

Multidimensional scaling (MDS) and nonspatial methods (e.g., ultrametric trees) are data analytic tools for graphically representing individuals' underlying perceptions of the relations among the stimuli within a proximity data set. Since the underlying perceptual structure in many domains is typically complex and multidimensional, one cannot expect experimental subjects to be able to describe the structure directly; rather, the structure must be inferred from responses to simpler queries. In the method of conditional rank orders, one of M stimuli is designated as a pivot stimulus and the subject is asked to select the most similar stimulus to this pivot from the $M - 1$ remaining stimuli. After this stimulus is selected or eliminated, the subject then selects the one of the remaining $M - 2$ stimuli which is most similar to the pivot stimulus. This process is continued until, in a *complete* conditional rank order, all $M - 1$ stimuli are rank ordered relative to this pivot. The pivot rotates to the next of the M stimuli and this task continues. In a complete rank order, all M stimuli will be used as a pivot. In *incomplete* conditional rank order data, one may decide to use only $G \leq M$ stimuli as pivots and/or order the first $T \leq M - 1$ closest stimuli. This format for data collection is referred to in Katahira (1990) as "pivot ordering". Dissimilarity data arising from conditional rank order tasks have been traditionally analyzed by MDS procedures such as those developed by Torgerson (1952), Shepard (1962), Kruskal (1964), Guttman (1968), Roskam (1970), Young (1974), etcetera. More recently, Takane and Carroll (1981) and Katahira have proposed parametric approaches to the nonmetric multidimensional scaling of such conditional rank orders. In these approaches, the nonmetric data are considered as incomplete data conveying only ordinal information about the distances. An unobserved metric process rendering complete information about distances is assumed to underlie the nonmetric data generation process. A likelihood function is specified for the observed nonmetric data which are related to the MDS distances based on some parametric assumptions about the underlying metric process (Takane & Carroll, 1981).

We extend a similar framework to ultrametric tree estimation. Suppose C subjects are presented with G pivot stimuli, and for each pivot stimulus, the subjects are asked to choose, in order, the T closest stimuli to the pivot, out of a possible set of M stimuli, where $T < M$. The pivot stimuli are chosen from among the total set of M stimuli. The object of multidimensional scaling procedures is to uncover, based on these responses, the set of true distances d_{ij} between stimuli i and j, $j = 1, \ldots, M$ and $i < j$. In Katahira (1990), and Takane and Carroll (1981), the true distances between stimuli, d_{ij}, are estimated, subject to the restriction that the distances form a Euclidean metric over a low dimensional space. In this paper, rather than force the distances to be Euclidean distances, we choose a different restriction—namely that the tree distances satisfy the ultrametric inequality:

$$d_{ij} \leq \max \{d_{ik}, d_{jk}\} \text{ for all distinct } i, j, k \text{ triples.} \tag{3}$$

A set of distances satisfying (3) uniquely determines a hierarchical tree representation. The ultrametric condition is not necessarily less restrictive than the Euclidean condition (see, e.g., Holman, 1972); however, in many applications, the discrete type of

representation proves to give a better fit to such data and offers a simpler interpretation than do competing spatial (Euclidean) models.

2.1. The Model

As in Takane and Carroll (1981) and Katahira (1990), we assume an additive model. That is, our assumption regarding conditional rank ordered responses is that stimulus j will be judged by a subject closer to pivot i than is stimulus k if $\delta_{ij} < \delta_{ik}$, where δ_{ij} is a latent distance related to the true distance d_{ij} by:

$$\delta_{ij} = \phi(d_{ij}) + \beta^{-1}\varepsilon_{ij}, \tag{4}$$

where $\phi(d_{ij})$ is a known monotone increasing function of d_{ij}, ε_{ij} is a deviate from the extreme value distribution:

$$f(\varepsilon) \propto \exp(\varepsilon - e^{\varepsilon}), \tag{5}$$

and β is a scale parameter reflective of the amount of "noise" in the dataset. The parameters to be estimated for an ultrametric tree model are the tree distances $\underline{\underline{D}} = ((d_{ij}))$ and the scale parameter β. As in Katahira (1990), we use a maximum likelihood estimation algorithm for estimating the tree distances and the scale parameter β.

2.2. The Likelihood Function

The stochastic modeling of rank ordered data has been discussed in Chapman and Staelin (1982), Keener and Waldman (1985), Hausman and Ruud (1987), and Fligner and Verducci (1988, 1993), where a variety of probabilistic models have been presented. To derive the likelihood function for conditional rank order data, it is helpful to first consider the case of simple rank order data. Suppose the following order relationship among quantities δ_i is observed:

$$\delta_1 < \delta_2 < \cdots < \delta_n. \tag{6}$$

Further suppose that the δ_i are related to unknown parameters d_i according to the conditional probability densities $f(\delta_i|d_i)$, with δ_i and δ_j independent given d_i and d_j. Then, the likelihood of observing the order relationship in (6), in terms of the unknown parameters d_i, is given by:

$$L(d_1, \ldots, d_n) = \text{Prob}\,(\delta_1 < \delta_2 < \cdots < \delta_n | d_1, \ldots, d_n) \tag{7}$$

$$= \int_{\delta_1 < \delta_2 < \cdots < \delta_n} f(\delta_1, \ldots, \delta_n | d_1, \ldots, d_n)\, d\delta_1 \cdots d\delta_n \tag{8}$$

$$= \int_{\delta_1 < \delta_2 < \cdots < \delta_n} f(\delta_1|d_1) \cdots f(\delta_n|d_n)\, d\delta_1 \cdots d\delta_n. \tag{9}$$

Equation (9) follows from the independence of the δ_i's.

Kalbfleisch and Prentice (1973) showed that if the conditional densities $f(\delta|d)$ have the so-called "proportional hazards" property:

$$h(\delta|d) = \frac{f(\delta|d)}{1 - F(\delta|d)} = h_0(\delta) \exp(\bar{\phi}(d)) \tag{10}$$

for some baseline hazard function $h_0(\delta)$ and some function $\bar{\phi}(d)$, then the marginal likelihood in (9) is given by:

$$\frac{\exp(\bar{\phi}(d_1))}{\sum_{i=1}^{n}\exp(\bar{\phi}(d_i))}\cdot\frac{\exp(\bar{\phi}(d_2))}{\sum_{i=2}^{n}\exp(\bar{\phi}(d_i))}\cdots\frac{\exp(\bar{\phi}(d_{n-1}))}{\sum_{i=n-1}^{n}\exp(\bar{\phi}(d_i))}. \tag{11}$$

One special case in which the proportional hazards assumption holds is when δ_i and d_i are related by the additive model:

$$\delta_i = \phi(d_i) + \beta^{-1}\varepsilon_i, \tag{12}$$

where ε_i has the extreme value distribution (5) and β is a constant scale factor. In this case, the conditional hazard rate is:

$$h(\delta|d) = \beta\exp(\beta(\delta - \phi(d))) \tag{13}$$

$$= h_0(\delta)\exp(\bar{\phi}(d)), \tag{14}$$

with:

$$h_0(\delta) = \beta\exp(\beta\delta) \tag{15}$$

$$\bar{\phi}(d) = -\beta\phi(d). \tag{16}$$

Several other models, including exponential and Weibull models, can also be shown to have proportional hazards (see, e.g., Lawless, 1982). Cox (1972) and Peto (1972) offer modifications of likelihood function (11) in the event that there are ties in the rank data.

For conditional rank order data, with C subjects ranking the T closest stimuli to G pivot stimuli, and the d_{ij}'s related to the δ_{ij}'s by equation (4), the probability of observing a given set of rank orderings R would be:

$$P(R|\underline{D},\beta) = \prod_{c=1}^{C}\prod_{g=1}^{G}\prod_{t=1}^{T}\frac{\exp(-\beta\phi(d_{gI_{cgt}}))}{\sum_{k\in K_{cgt}}\exp(-\beta\phi(d_{gk}))}, \tag{17}$$

where I_{cgt} is the stimulus judged by subject c as the t-th closest stimulus to pivot stimulus g, and K_{cgt} is the set of stimuli judged by subject c to be farther than stimulus I_{cgt} from pivot stimulus g. Equation (17) is just a generalization of (11) to the case of conditional rank order data with multiple subjects. From equation (17), the log likelihood for $\underline{D}$, given the observed rank order data R, is:

$$\mathcal{L}(\underline{D},\beta) = \log P(R|\underline{D},\beta) = \sum_{c=1}^{C}\sum_{g=1}^{G}\sum_{t=1}^{T}\left[-\beta\phi(d_{gI_{cgt}}) - \log\left(\sum_{k\in K_{cgt}}\exp(-\beta\phi(d_{gk}))\right)\right]. \tag{18}$$

The fact that the marginal likelihood function in (11) can be derived from much weaker assumptions than the additive model with extreme value errors suggests that the maximum likelihood estimation algorithm to be described may be fairly robust with respect to model misspecification.

To fit an ultrametric tree model to the observed rank orders R, the d_{ij}'s are chosen to maximize the log likelihood in (19), subject to the restrictions given in (3). The choice of an appropriate function $\phi(\cdot)$ is discussed in Katahira (1990). The form of $\phi(\cdot)$ determines the extent of error variance relative to the magnitude of a true dissimilarity d_{ij}. A logarithmic form implies that the relative error increases with the dissimilarity, an assumption which Ramsay (1977, 1982) supports. Further support for specifying $\phi(d) = \log d$ is provided by Abe (1993), who argues that the logarithm is the only function that leaves the configuration scale invariant; that is, only when $\phi(d) = \log d$ is the probability in (18) invariant with respect to the choice of the scale of the data. The function ϕ cannot be estimated from the data, as for example with splines, as there is an essential unidentifiability if the d_{ij}'s are also unknown, which is the case in this context. This is because $\phi(\cdot)$ and the d_{ij} appear in the likelihood function (19) only through the terms $\phi(d_{ij})$. Therefore, given some estimates $\hat{\phi}(\cdot)$ and $((\hat{d}_{ij}))$ of $\phi(\cdot)$ and $((d_{ij}))$, the alternative estimates $\hat{\phi}^*(\cdot)$ and $((\hat{d}^*_{ij}))$, with $\hat{\phi}^*(d) = \phi(s^{-1}(d))$ and $\hat{d}^*_{ij} = s(\hat{d}_{ij})$, will have an exactly equal likelihood, for any monotone function $s(\cdot)$. What is more, if the $((d_{ij}))$ satisfy the ultrametric inequality (3), then so will the $((\hat{d}_{ij}))$. Therefore, throughout this paper, we make the assumption that $\phi(d) = \log(d)$. Given this assumption, the log likelihood takes the form:

$$\mathcal{L}(\underline{D}, \beta) = \log P(R|\underline{D}, \beta) = \sum_{c=1}^{C} \sum_{g=1}^{G} \sum_{t=1}^{T} \left[-\beta \log d_{gI_{cgt}} - \log \left(\sum_{k \in K_{cgt}} d_{gk}^{-\beta} \right) \right]. \tag{19}$$

While in the foregoing it has been assumed that the parameter β is common to all respondents, Ramsay (1982) stresses the importance of considering differences in response variability across subjects. With our method, it is possible to consider the case in which each respondent c, $1 \leq c \leq C$, has a unique scale parameter β_c. It is also possible to a priori group the respondents according to some discrete covariate (e.g., gender) and assign a common scale parameter to each designated member within the same group.

2.3. The Estimation Algorithm

The unknown parameters of the model are the $((d_{ij}))$'s and β; to estimate these parameters, we employ a maximum likelihood procedure. However, there are a number of constraints that must be enforced by the optimization method. Since the $((d_{ij}))$'s are distances, they are constrained to be positive. Also, the scale of the $((d_{ij}))$'s must be fixed for reasons of identifiability; this can be achieved by constraining the sum of the $((d_{ij}))$'s to equal 1:

$$\sum_{i \neq j} d_{ij} = 1. \tag{20}$$

Finally the $((d_{ij}))$'s must obey the ultrametric inequality given in (3).

The scale parameter β must also be constrained. As a scale parameter, it naturally must be positive. On the other hand, given "error-less" data in which there are no inconsistencies in rankings, the maximum likelihood estimate of β is $+\infty$, a situation which leads to computational difficulties during numerical maximization. Thus, in practice, β must be constrained to be less than some large but finite constant, say B_0.

The bound constraints on β and on the d_{ij}, and the identifying constraint (20) are all linear constraints. Only the ultrametric inequality constraint is nonlinear. We choose

to handle the nonlinear constraint via an iterative penalty function approach (e.g., Ryan, 1974). Define the penalty function:

$$\alpha(\underline{D}) = \frac{1}{2} \sum_{(i,j,k)\in\Omega} (d_{ik} - d_{jk})^2, \tag{21}$$

where:

$$\Omega = \{(i, j, k) | d_{ij} \leq \min(d_{ik}, d_{jk}) \text{ and } d_{ik} \neq d_{jk}\}. \tag{22}$$

Ω is the set of triples (i, j, k) which violate the ultrametric inequality, and $\alpha(\underline{D})$ penalizes a tentative estimate of $\underline{D}$ for violation of the inequality. The maximum likelihood ultrametric tree is then found by the following sequence:

1. Choose initial estimates for $\underline{D}$ and β. The initial estimates for $\underline{D}$ will almost always be infeasible. Choose an initial value for a penalty parameter λ; say, $\lambda = 1$.
2. Solve the linearly constrained nonlinear optimization problem:

$$\max_{\underline{D},\beta} \mathcal{L}(\underline{D}, \beta) - \lambda \alpha(\underline{D}), \tag{23}$$

$$\text{subject to} \quad 0 \leq \beta \leq B_0 \tag{24}$$

$$0 \leq d_{ij}, j = 2, \ldots, M \text{ and } i < j \tag{25}$$

$$\sum_{i \neq j} d_{ij} = 1, \tag{26}$$

   via sequential quadratic programming (see Appendix).
3. If ultrametric constraint (3) is satisfied to within some specified tolerance, then quit. Else, let $\lambda \leftarrow 10\lambda$, and go back to Step 2.

The solution to the problem in Step 2 converges to the appropriate constrained optimum of the log-likelihood $\mathcal{L}(\underline{D}, \beta)$ as the penalty parameter λ tends toward infinity. That is, in the limit, the solution obtained will be the best fitting tree, subject to the ultrametric inequality constraints (3). Fiacco and McCormick (1968) describe the convergence properties of this approach to optimization subject to nonlinear constraints.

For the initial estimates of the $d_{ij}, j > i$, we take the ranks of the distances given i as the pivot stimulus, averaged over all subjects; the $((d_{ij}))$ are then scaled to satisfy $\sum_{i<j} d_{ij} = 1$. Because log-likelihood surfaces such as the one defined by (19) are typically multimodal, it is helpful to perform multiple analyses using random starts, to avoid being trapped in local optima. The user can also generate reasonable starting values for $\underline{D}$ using, say, existing hierarchical clustering methods on the averaged ranks.

2.4. Computing Derivatives

The derivatives of the log-likelihood function (19) can be computed using the following procedure:

- Set all $\partial\mathcal{L}/\partial d_{ij} = 0, i, j = 1, \ldots, M$ and $i < j$.
- For $c = 1$ to C
  For $g = 1$ to G
  For $t = 1$ to T

★ Let:

$$\frac{\partial \mathcal{L}}{\partial d_{gI_{cgt}}} = \frac{\partial \mathcal{L}}{\partial d_{gI_{cgt}}} - \frac{\beta}{d_{gI_{cgt}}}. \tag{27}$$

★ For $k \in K_{cgt}$, let:

$$\frac{\partial \mathcal{L}}{\partial d_{gk}} = \frac{\partial \mathcal{L}}{\partial d_{gk}} + \frac{\beta d_{gk}^{-\beta-1}}{\sum_{k' \in K_{cgt}} d_{gk'}^{-\beta}}. \tag{28}$$

The derivatives of the penalty function are given by:

$$\frac{\partial \alpha}{\partial d_{ik}} = \sum_{j \in U_{ik}} (d_{ik} - d_{jk}) - \sum_{j \in V_{ik}} (d_{jk} - d_{ik}), \tag{29}$$

where U_{ik} is the set of stimuli defined by:

$$U_{ik} = \{j : d_{ij} < d_{jk} < d_{ik},\ i \neq j,\ i \neq k,\ j \neq k\}, \tag{30}$$

and V_{ik} is the set of stimuli defined by:

$$V_{ik} = \{j : d_{ij} < d_{ik} < d_{jk},\ i \neq j,\ i \neq k,\ j \neq k\}. \tag{31}$$

If the d_{ij}'s satisfy the ultrametric inequality, then the sets U_{ik} and V_{ik} will be empty for all i and k.

The derivative of the log likelihood with respect to β is given by:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{c=1}^{C} \sum_{g=1}^{G} \sum_{t=1}^{T} \left[-\log d_{gI_{cgt}} - \frac{\sum_{k \in K_{cgt}} -(\log d_{gk})\, d_{gk}^{-\beta}}{\sum_{k \in K_{cgt}} d_{gk}^{-\beta}} \right]. \tag{32}$$

If one is modeling a separate scale parameter β_c for each subject, then the appropriate log likelihood derivative is:

$$\frac{\partial \mathcal{L}}{\partial \beta_c} = \sum_{g=1}^{G} \sum_{t=1}^{T} \left[-\log d_{gI_{cgt}} - \frac{\sum_{k \in K_{cgt}} -(\log d_{gk}) d_{gk}^{-\beta_c}}{\sum_{k \in K_{cgt}} d_{gk}^{-\beta_c}} \right]. \tag{33}$$

3. Simulation Analysis

A small-scale Monte-Carlo simulation was performed to assess the effectiveness of the estimation technique in recovering the true underlying ultrametric distance matrix. The maximum likelihood method described in this paper (ML) was compared to a least-squares (LS) procedure as in (De Soete, 1984) in which the distance matrix $((d_{ij}))$ was chosen to minimize the criterion:

$$\sum_{i<j} (d_{ij} - w_{ij})^2, \tag{34}$$

where $\underline{W} = ((w_{ij}))$ is a matrix formed by averaging the ranks of the distances over all subjects; that is, w_{ij} is the average, over all subjects, of the rank of the distance from

Table 1

Levels of Factors for Monte Carlo Experiment

Factor:		*Levels:*
C	(# of Subjects)	6, 12
M	(# of Stimuli)	8, 12
σ	(Error Standard Dev.)	1.0, 1.5
$\underline{D}$	(Tree Topology)	S=Symmetric, A=Asymmetric
T	(Depth of Ranking)	Full, Partial

j to i, given i as the pivot stimulus. This least-squares procedure is designed for metric distance data $((w_{ij}))$, whereas the conditional rank order data is in fact only ordinal data; the least-squares procedure is therefore expected to perform worse than our maximum likelihood procedure proposed in section 2, when applied to such ordinal data.

The synthetic conditional rank order data was simulated by generating latent distances δ_{ij} from the model:

$$\delta_{ij} = \log d_{ij} + \beta^{-1}\varepsilon_{ij},\ i, j = 1, \ldots, M, \tag{35}$$

with ε_{ij} being random deviates from the extreme value distribution (5), and then ranking the T closest δ_{ij}'s for each given pivot stimulus. This procedure was performed for each of C different hypothetical subjects. A factorial experimental design was employed with the following variables treated as independent factors: number of subjects, number of stimuli, error standard deviation, tree topology, and depth of ranking (partial or complete). A partial ranking is one in which the depth T is less than $M - 1$; for such simulations we used $T = M/2$. In the case of partial rank data, the ranks for all stimuli that were not listed among the T closest stimuli to the pivot stimulus were set to $T + 1$, for the purposes of computing the average rank matrix $\underline{W}$ used in the least-squares procedure, and of determining an initial estimate for the maximum likelihood procedure. Table 1 provides the different experimental levels for the factors. Note that the standard deviation σ of the random errors $\beta^{-1}\varepsilon_{ij}$ is determined by the rule $\sigma = \pi/\beta\sqrt{6}$ (e.g., Katahira, 1990).

The experimental design used in this study allows us to examine the effect of the size of the dataset, the nature of the true ultrametric tree, and the amount of noise in the dataset on the estimation accuracy for the two estimation algorithms. The true d_{ij}'s in (35) used to generate the data obeyed the ultrametric inequality. The true ultrametric tree for the case of 8 stimuli and symmetric topology is given in Figure 2, and the tree with asymmetric topology is shown in Figure 3. Ten simulation replications were performed for each of the 2^5 experimental combinations of factors, producing a total of 320 trials.

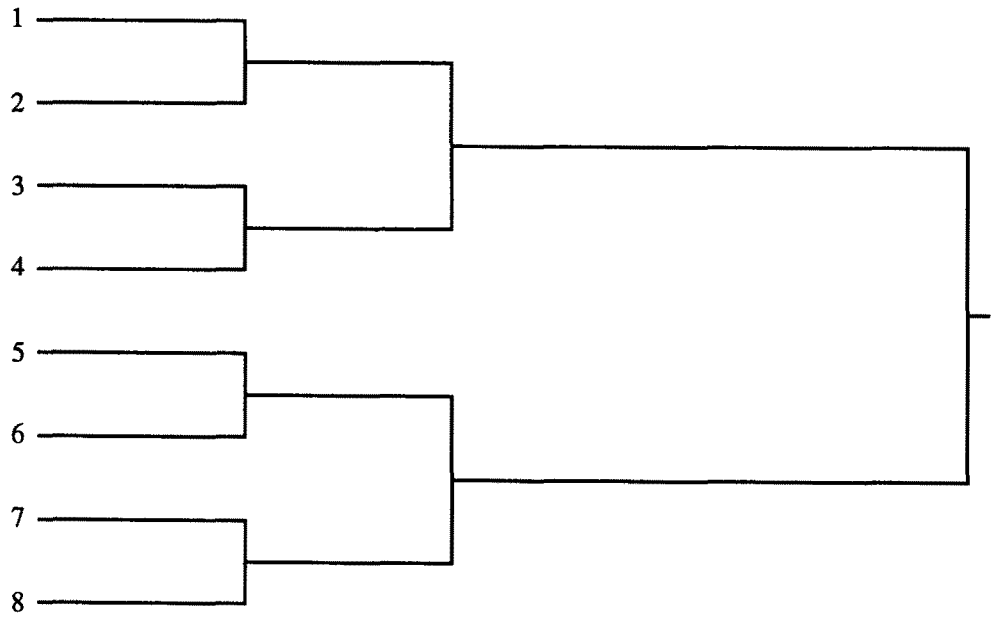

FIGURE 2.
True Ultrametric Tree Used for Simulation, Symmetric Topology.

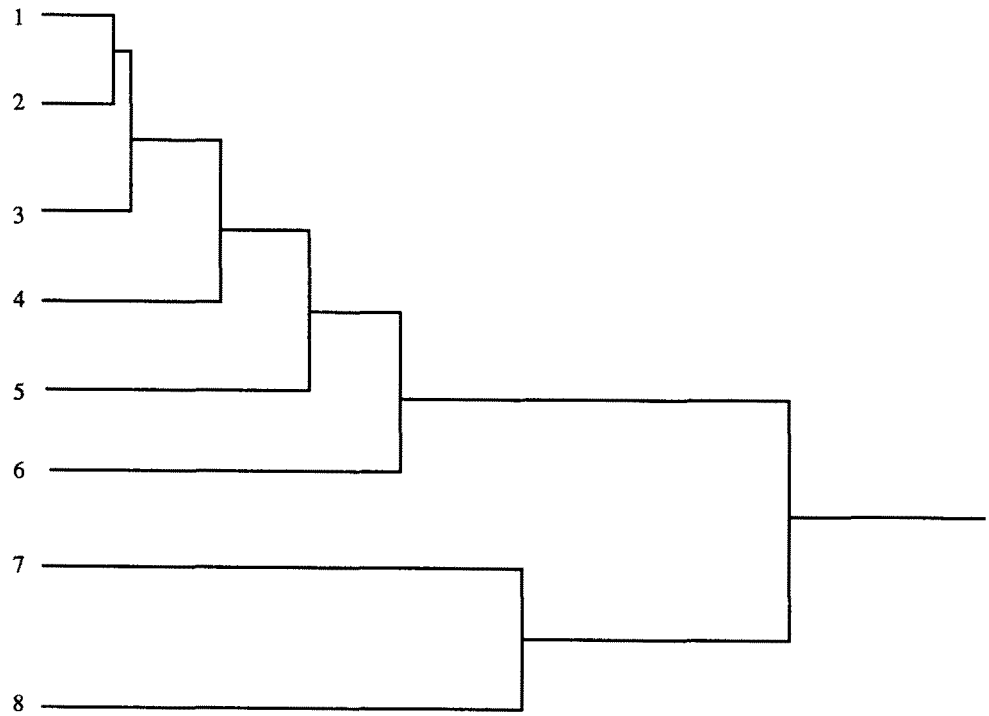

FIGURE 3.
True Ultrametric Tree Used for Simulation, Asymmetric Topology.

3.1. Measures of Performance

Two dependent measures of overall algorithm performance were used in this Monte Carlo analysis: a rank correlation measure, and a metric correlation measure. The rank correlation measure used in Katahira (1990) was computed as:

$$1 - 6 \sum_{i,j} \frac{(\hat{e}_{ij} - e_{ij})^2}{M(M^2 - 1)}, \tag{36}$$

where $\hat{e}_{ij}$ is the rank of the estimated distance $\hat{d}_{ij}$, and e_{ij} is the rank of the true distance d_{ij}. Because there are necessarily many ties in an ultrametric distance matrix, this measure seems to be inappropriate for this study. Therefore, Kendall's rank correlation measure τ was used (Lehmann, 1975). This measure, which is bounded between -1 and 1, quantifies the extent to which the ordering of the d_{ij}'s agrees with the ordering of the $\hat{d}_{ij}$'s. The metric measure of estimation accuracy used was the usual Pearson correlation ρ between the d_{ij}'s and the $\hat{d}_{ij}$'s. The computational burden for the two estimation procedures, least-squares and maximum likelihood, was measured in terms of CPU minutes. The simulation was performed on a computer with a 486/33MHz Intel processor, running under the MS-DOS operating system.

3.2. Simulation Results

Table 2 presents a summary table of the comparative results (means and standard deviations) of the proposed maximum likelihood procedure versus a least-squares methodology. As shown, our proposed ultrametric tree estimation procedure dominates the least-squares approach across all factor combinations in terms of the two correlation measures of recovery.

Overall, the maximum likelihood estimation procedure was successful in recovering the true tree configuration; the average of Pearson's ρ across all the simulation replications using the maximum likelihood method was .906 (SD = .13), while the average using the least-squares approach was .818 (SD = .176). The average of Kendall's τ using the maximum likelihood method was .907 (SD = .127), while the average using the least-squares method was .803 (SD = .162). The performance of the maximum likelihood and the least-squares procedures with respect to both the τ and ρ measures was compared using paired-t tests and sign tests applied to the 320 simulation replications. The values of ρ for the maximum likelihood procedure were significantly higher than those for the least-squares procedure ($p < .01$, t test, $p < .01$, sign test), as was the case for the rank correlation τ ($p < .01$, t test, $p < .01$, sign test). The improvement is most marked in those sets of trials for which the depth of ranking T is less than $M - 1$; that is, for those trials using only a partial order. For example, for the trials with $C = 12$, $M = 12$, $T = 6$, $\beta = 1.5$, symmetric topology ($\underline{D}$ = S), the average ρ for the maximum likelihood method was .97 (SD = .02), while that for the least-squares method was .82 (SD = .09). The ability to efficiently analyze partial order data, a form of data that may be particularly convenient to collect, appears to be a major advantage of the maximum likelihood procedure.

The least-squares approach is uniformly more computationally efficient than the proposed maximum likelihood approach. Yet, for moderately sized data sets, the maximum likelihood estimator seems to be reasonably efficient in terms of computational resources; for example, for the 10 trials performed with 6 subjects, 8 stimuli, full order, $\sigma = 1$, and symmetric true tree topology, the average CPU time required was .49 minutes. The average time increases to 7.94 minutes for the trials with equal parame-

Table 2

Results of Monte Carlo Experiment.

Output measures (ρ, τ, and CPU minutes) are displayed as

Means (Standard Deviations).

C	M	T	σ	D	ρ (MLE)	ρ (LS)	τ (MLE)	τ (LS)	CPU (MLE)	CPU (LS)
6	8	7	1.00	S	0.95 (0.10)	0.84 (0.25)	0.96 (0.05)	0.82 (0.23)	0.49 (0.11)	0.06 (0.01)
6	8	4	1.00	S	0.91 (0.14)	0.86 (0.18)	0.90 (0.11)	0.85 (0.15)	0.42 (0.07)	0.06 (0.01)
6	8	7	1.50	S	0.79 (0.29)	0.56 (0.40)	0.69 (0.35)	0.56 (0.38)	0.46 (0.06)	0.07 (0.02)
6	8	4	1.50	S	0.83 (0.16)	0.62 (0.21)	0.79 (0.13)	0.62 (0.25)	0.40 (0.07)	0.07 (0.02)
12	8	7	1.00	S	0.97 (0.09)	0.99 (0.02)	0.96 (0.07)	0.96 (0.02)	0.56 (0.06)	0.05 (0.01)
12	8	4	1.00	S	1.00 (0.01)	0.92 (0.16)	0.97 (0.05)	0.90 (0.13)	0.50 (0.04)	0.04 (0.02)
12	8	7	1.50	S	1.00 (0.01)	0.82 (0.18)	0.96 (0.05)	0.82 (0.18)	0.56 (0.08)	0.05 (0.01)
12	8	4	1.50	S	0.76 (0.43)	0.78 (0.17)	0.76 (0.39)	0.81 (0.15)	0.49 (0.06)	0.05 (0.01)
6	12	11	1.00	S	0.98 (0.01)	0.91 (0.10)	0.97 (0.03)	0.93 (0.07)	7.41 (1.47)	1.03 (0.46)
6	12	6	1.00	S	0.95 (0.07)	0.84 (0.12)	0.95 (0.04)	0.86 (0.06)	6.67 (1.24)	0.96 (0.43)
6	12	11	1.50	S	0.97 (0.02)	0.95 (0.04)	0.96 (0.05)	0.91 (0.06)	6.73 (1.24)	0.92 (0.41)
6	12	6	1.50	S	0.95 (0.08)	0.74 (0.19)	0.94 (0.04)	0.77 (0.14)	7.57 (1.16)	0.89 (0.27)
12	12	11	1.00	S	0.98 (0.01)	0.94 (0.09)	0.98 (0.02)	0.94 (0.07)	7.94 (1.36)	0.80 (0.44)
12	12	6	1.00	S	0.98 (0.01)	0.91 (0.11)	0.97 (0.02)	0.90 (0.06)	7.24 (1.56)	0.60 (0.21)
12	12	11	1.50	S	0.97 (0.03)	0.93 (0.10)	0.97 (0.03)	0.92 (0.07)	6.82 (1.07)	0.87 (0.27)
12	12	6	1.50	S	0.97 (0.02)	0.82 (0.09)	0.95 (0.05)	0.85 (0.07)	6.86 (1.81)	0.75 (0.14)
6	8	7	1.00	A	0.91 (0.06)	0.87 (0.05)	0.89 (0.06)	0.84 (0.04)	0.40 (0.05)	0.06 (0.01)
6	8	4	1.00	A	0.93 (0.07)	0.84 (0.05)	0.91 (0.06)	0.78 (0.05)	0.42 (0.05)	0.06 (0.01)
6	8	7	1.50	A	0.88 (0.08)	0.84 (0.10)	0.90 (0.08)	0.82 (0.08)	0.45 (0.06)	0.06 (0.01)
6	8	4	1.50	A	0.82 (0.11)	0.72 (0.28)	0.82 (0.12)	0.71 (0.15)	0.42 (0.06)	0.07 (0.02)
12	8	7	1.00	A	0.91 (0.06)	0.93 (0.04)	0.94 (0.03)	0.89 (0.02)	0.56 (0.10)	0.05 (0.01)
12	8	4	1.00	A	0.90 (0.06)	0.86 (0.06)	0.92 (0.03)	0.80 (0.04)	0.47 (0.08)	0.06 (0.01)
12	8	7	1.50	A	0.90 (0.06)	0.83 (0.10)	0.93 (0.04)	0.83 (0.05)	0.53 (0.09)	0.07 (0.01)
12	8	4	1.50	A	0.82 (0.15)	0.75 (0.15)	0.88 (0.07)	0.69 (0.16)	0.50 (0.07)	0.05 (0.01)
6	12	11	1.00	A	0.89 (0.07)	0.86 (0.06)	0.91 (0.02)	0.82 (0.06)	6.98 (1.54)	0.89 (0.30)
6	12	6	1.00	A	0.90 (0.04)	0.74 (0.07)	0.90 (0.04)	0.70 (0.06)	5.15 (0.89)	1.04 (0.31)
6	12	11	1.50	A	0.84 (0.07)	0.74 (0.12)	0.89 (0.05)	0.74 (0.11)	5.82 (0.88)	1.09 (0.31)
6	12	6	1.50	A	0.86 (0.05)	0.72 (0.06)	0.89 (0.04)	0.69 (0.06)	6.55 (1.67)	0.83 (0.30)
12	12	11	1.00	A	0.92 (0.02)	0.86 (0.08)	0.92 (0.01)	0.83 (0.07)	8.29 (1.34)	0.81 (0.23)
12	12	6	1.00	A	0.89 (0.03)	0.78 (0.07)	0.91 (0.03)	0.73 (0.06)	7.35 (0.63)	0.68 (0.17)
12	12	11	1.50	A	0.85 (0.07)	0.74 (0.17)	0.87 (0.05)	0.73 (0.13)	7.97 (1.85)	0.89 (0.21)
12	12	6	1.50	A	0.84 (0.11)	0.69 (0.12)	0.87 (0.09)	0.67 (0.09)	7.27 (1.29)	0.77 (0.23)

ters, but with 12 subjects and 12 stimuli. As expected, the execution time appears to be determined principally by the number of stimuli.

Our tentative conclusion from this Monte Carlo study is that the new maximum likelihood method has promise in accurately identifying ultrametric tree structures from conditional rank order data, and that the method may offer superior performance to methods which ignore the nonmetric aspect of such data. However, it should be noted that the benefit is at the expense of increased computational burden. Clearly, further testing is necessary.

4. Data Analysis—Green and Rao's Snack Data

The proposed maximum likelihood estimation technique of section 2 was applied to the conditional rank order similarity data of various snack or breakfast foods reported in Green and Rao (1972). In this dataset, 42 individuals (21 MBA students and their

Table 3

Stimuli in Green and Rao's Snack Food Dataset

English Muffin with Margarine	EMM
Hard Roll with Butter	HRB
Toast with Margarine	TMn
Toast with Butter	BT
Blueberry Muffin with Margarine	BMM
Corn Muffin with Butter	CMB
Toasted Popup	TP
Toast with Butter and Jelly	BTJ
Toast with Marmalade	TMd
Cinnamon Toast	CT
Cinnamon Bun	CB
Coffee Cake	CC
Danish Pastry	DP
Jelly Donut	JD
Glazed Donut	GD

wives) provided complete conditional rank order similarity responses for the 15 stimuli; the stimuli are listed in Table 3.

Green and Rao (1972) analyze the raw ordinal data by first preprocessing them with the TRICON procedure (Coombs, 1964), in order to convert the 42 sets of rank order responses into a single symmetric 15×15 matrix of "dissimilarities". This dissimilarity matrix was then subjected to various metric and non-metric scaling procedures, including TORSCA (Young & Torgerson, 1967), and Kruskal monotone scaling (Kruskal, 1964). The dissimilarity data, as well as the estimated 2-dimensional Euclidean distance matrices, were further analyzed by the hierarchical clustering technique of Johnson (1967). The results from the various scaling procedures were all deemed to be "quite similar". The tentative interpretations offered for the two dimensions in the scaling model were that the one dimension represented "sweetness" or "caloric content", while the other dimension represented a "toast/nontoast" dimension. When the distance matrix estimated by TORSCA was subjected to Johnson's cluster analysis procedure, the two principle clusters that emerged, as seen in Figure 4, were the sweeter

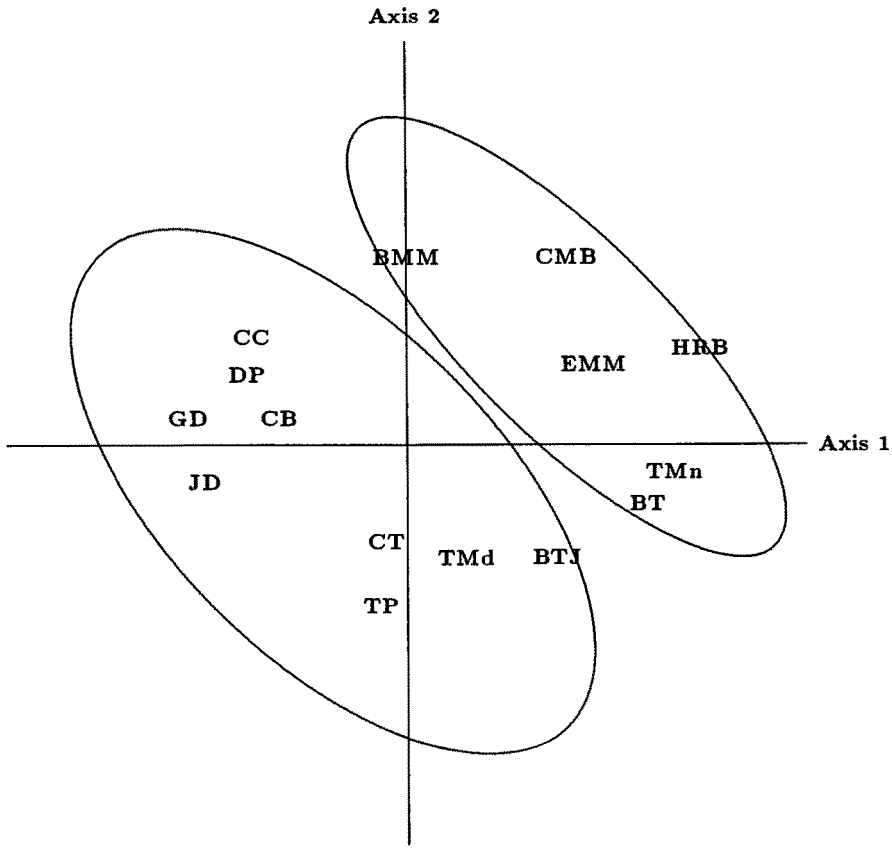

FIGURE 4.
Clustering in Two-Space Configuration from TORSCA 8 Analysis of Snack Data (Green and Rao, 1972).

items such as coffee cake, danish pastry, glazed donut, jelly donut, cinnamon bun, cinnamon toast, toast pop-up, toast with marmalade, and buttered toast with jelly, vs. the less sweet items including blueberry muffin with margarine, corn muffin with butter, english muffin with margarine, buttered hard roll, toast and margarine, and buttered toast.

Green and Rao (1972) note that there are possible shortcomings with their analysis. In particular, the TRICON conversion of the 42 conditional rank order (ordinal) data sets into a distance data matrix which is then treated as metric data may be questionable. Also, there is no weighting of subjects in this preprocessing, nor is there accounting for possible individual differences. Here, we reanalyze this data using the maximum likelihood procedure described in section 2. This procedure is especially designed to treat conditional rank order data, so there is no need for any problematic preprocessing. In addition, individual level scale parameters can be estimated to accommodate heterogeneity among the subjects.

The ultrametric trees were estimated using the proposed procedure under two different assumptions:

H_1: β is constant across subjects (Scale Homogeneity)

H_2: β is different for each subject (Scale Heterogeneity).

The ultrametric trees estimated under H_1 and H_2 are given in Figures 5 and 6; the topologies are nearly identical. Here too, we see evidence of the distinction of the sweet foods from the less sweet foods, plus some additional subtleties. For the sweet break-

fast foods, the one item that is always toasted, TP, is distinct from the other sweet foods. For the less sweet foods, the muffins, toasts, and hard roll seem to be separated, suggesting a "consistency" subclassification category. Finally, the estimated ultrametric tree depicts the hard roll and butter (HRB) as a very distinct stimulus, an aspect not recovered in any of the Green and Rao (1972) MDS or clustering solutions. To understand why HRB is estimated as a distinct stimulus, a display of the raw data was sought which might suggest the uniqueness of this particular breakfast food. Table 4 presents a matrix, the (j, k)-th element of which is the number of subjects out of 42 who chose element k (column food) as *the most distant* from element j (row food). In this table, the buttered hard roll (HRB) indeed appears to be quite distinct in that it is so commonly deemed to be the farthest stimulus from the other stimuli. While the HRB column might seem to suggest that some of the toast stimuli are not too distant from the buttered hard roll, comparison of the HRB column with the corresponding toast columns shows that there are considerable differences. For example, 21 of the subjects deemed buttered hard roll to be the farthest stimulus from danish pastry (DP), while only 3 deemed buttered toast (BT) to be the farthest from danish pastry. Thus, if the stimuli are truly perceived according to an ultrametric topology, the data suggest that the buttered hard roll must be positioned as separate from the toasts.

The log likelihood for model H_1 (scale homogeneity) was -14695, and that for H_2 (scale heterogeneity) was -14454; the increase of 241 is associated with 41 extra degrees of freedom. The estimates of β ranged from .0949 to 7.08. The usual assumptions for justifying likelihood ratio tests (LRT) and information criteria are not met, as the parameter space is bounded, with the MLE occurring on a boundary of the parameter space. However, using these techniques as rough heuristics for aiding model choice, we find support for the hypothesis of scale heterogeneity (H_2). In the case of the LRT, the apparent evidence in favor of heterogeneity can be seen to be significant at $p < .005$. The AIC measure of model goodness-of-fit (-2Log-likelihood $+ 2p$, where p is the number of free parameters) is difficult to calculate, since the number of *free* parameters in a distance matrix constrained to satisfy the ultrametric inequality is not clear. However, whatever this number is, it is clear that the AIC for model H_2 must be lower than that for H_1. Let p_d be the number of free parameters associated with the distance matrix. Then the AIC for H_1 is $-2(-14695) + 2(p_d + 1) = 29392 + 2p_d$. The AIC for H_2 is $-2(-14454) + 2(p_d + 42) = 28992 + 2p_d$. Thus, the AIC measure also supports the hypothesis of scale heterogeneity. Given the tendency for AIC to favor overparameterized models (Bozdogan, 1987), we also computed the Schwartz (1978) Bayesian Information Criterion (BIC $= -2 \log L + p_d(\log F)$) and the Consistent AIC measure (CAIC $= -2 \log L + p_d(\log F + 1)$), where F is the number of independent data observations (here, $CM(M - 2)$). We see that BIC$(H_1) = 29400 + 9.01p_d$ and BIC$(H_2) = 29286 + 9.01p_d$, supporting scale heterogeneity; CAIC$(H_1) = 29401 + 10.01p_d$ and CAIC$(H_2) = 29328 + 10.01p_d$, also supporting scale heterogeneity.

A relatively high β coefficient for a subject might be caused by at least two possible factors: (a) the subject's latent tree configuration may be different from the tree configuration common to the other members of the group, or (b) the subject may have a large amount of internal inconsistency in his or her rankings. To assess the cause of scale heterogeneity, we formed measures of agreement with the tree, and measures of internal consistency, for each of the 42 subjects. As a measure of agreement with the common tree $((d_{ij}))$, we used the quantity:

$$\gamma_c = \sum_{g=1}^{G} \sum_{t=1}^{T} \#\{t' \in K_{cgt} : d_{gI_{cgt}} > d_{gt'}\}. \tag{37}$$

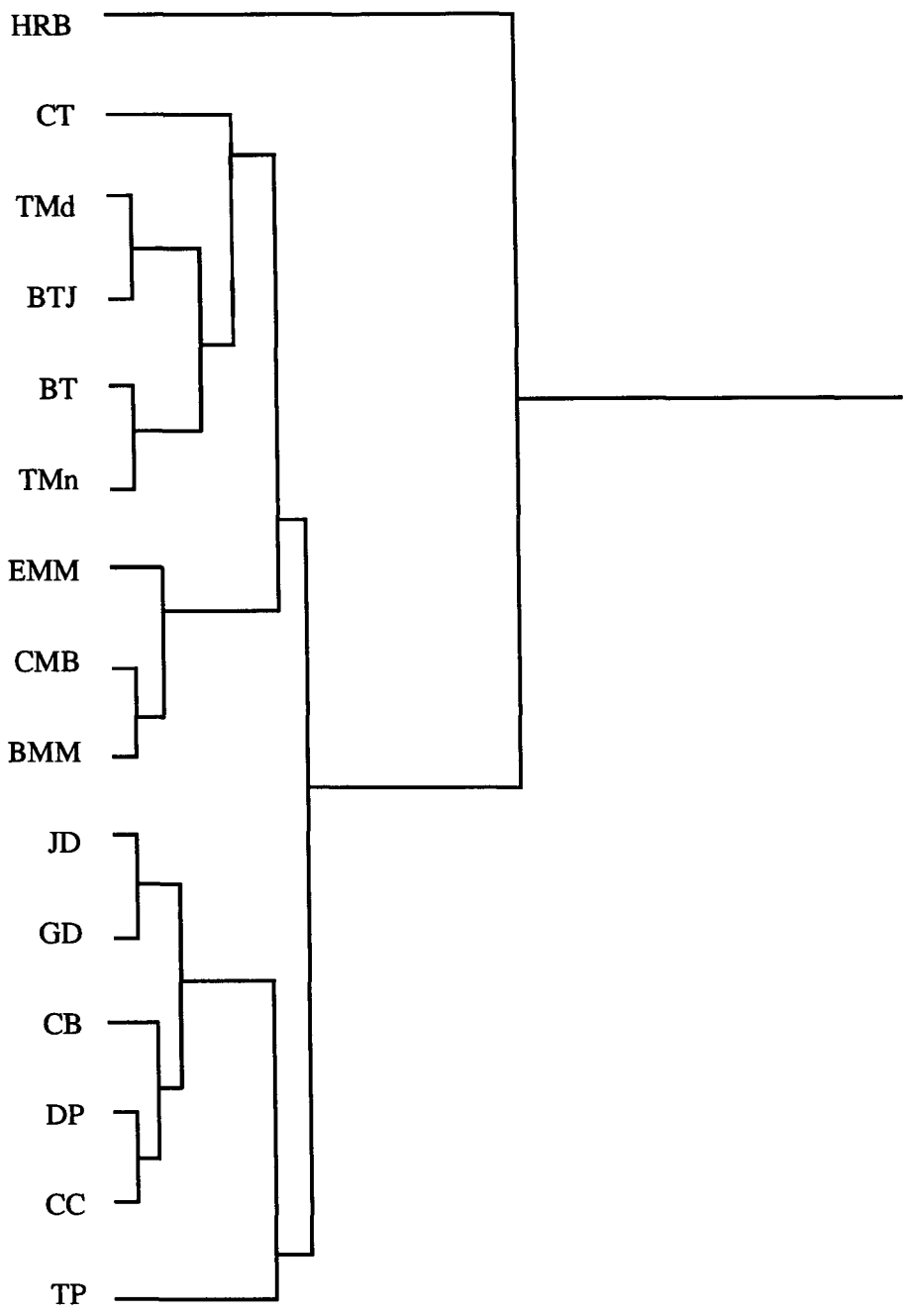

FIGURE 5.
Ultrametric Tree for Snack Data, Estimated Under H_1: Scale Homogeneity.

That is, γ_c is the number of instances in which subject c's rankings of distances between stimuli disagree with the rankings of the elements of the distance matrix $((d_{ij}))$. As a measure of internal consistency for subject c we used a count of intransitivities:

$$\psi_c = \#\{j, k, l: ((d_{kl}^{(c)} > d_{kj}^{(c)}), (d_{jk}^{(c)} > d_{jl}^{(c)}), \text{ and } (d_{lj}^{(c)} > d_{lk}^{(c)}))\}, \tag{38}$$

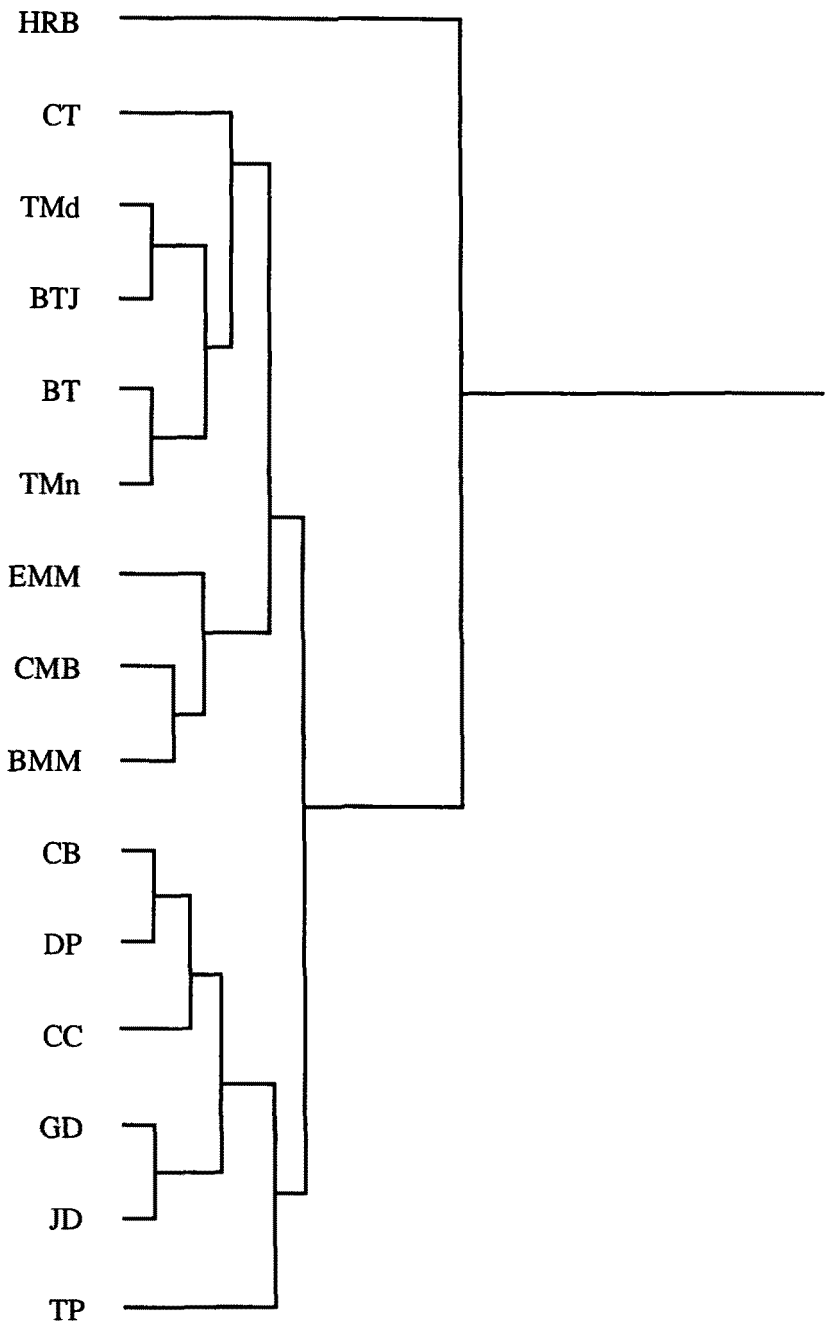

FIGURE 6.
Ultrametric Tree for Snack Data, Estimated Under H_2: Scale Heterogeneity.

where $d_{ij}^{(c)}$ is the implied distance from pivot stimulus i to stimulus j, based on the responses of subject c. Note that the actual distances $d_{ij}^{(c)}$ are not revealed, but their relative orders are, allowing for the calculation of ψ_c. The observed correlation between the individual β_c coefficients and the individual γ_c coefficients was .39. The correlation between the β_c coefficients and the internal consistency coefficients ψ_c was .45. The correlation between the ψ_c and the γ_c was .37. In a multiple regression, using

Table 4

Number of Subjects Choosing Column Stimulus as *the Farthest Stimulus* from the Row Stimulus

		1 TP	2 BT	3 EMM	4 JD	5 CT	6 BMM	7 HRB	8 TMd	9 BTJ	10 TMn	11 CB	12 DP	13 GD	14 CC	15 CMB
1	TP	0	3	2	3	0	1	18	1	0	4	1	2	0	2	5
2	BT	5	0	1	10	0	1	3	0	1	1	2	9	2	4	3
3	EMM	8	0	0	11	1	0	3	0	1	0	4	8	2	4	0
4	JD	4	5	4	0	0	0	18	1	0	4	1	1	0	1	3
5	CT	3	1	3	5	0	0	14	0	1	6	0	2	1	2	4
6	BMM	8	0	0	5	1	0	12	1	0	5	3	3	1	3	0
7	HRB	8	2	0	17	1	1	0	0	0	1	1	9	1	1	0
8	TMd	5	2	0	4	1	0	13	0	0	3	2	2	2	4	4
9	BTJ	6	2	0	4	1	0	14	0	0	1	3	3	3	4	1
10	TMn	9	0	0	9	0	1	6	1	0	0	3	6	2	4	1
11	CB	5	1	2	0	1	0	17	1	1	7	0	1	0	3	3
12	DP	4	3	3	0	1	0	21	0	1	6	1	0	0	0	2
13	GD	6	3	0	0	1	0	21	0	0	6	1	0	0	0	4
14	CC	3	5	1	1	1	0	20	0	1	9	1	0	0	0	0
15	CMB	6	2	1	6	1	0	8	0	3	3	3	4	2	3	0
	Avg.	5.7	2.1	1.2	5.4	0.7	0.3	18.4	0.4	0.6	4.0	1.9	3.6	1.1	2.5	2.1

the 42 estimated β_c's as the dependent measures and the respective γ_c's and ψ_c's as the independent measures, the relationship between the scale parameter β_c and the internal consistency measure ψ_c was seen to be significantly positive ($p < .02$), while the partial association of β_c with the measure γ_c was less significant ($p > .09$). These results suggest that the observed heterogeneity in the estimated β's for this dataset may be caused by individual differences in internal consistency.

As mentioned, one of the advantages of the proposed MLE based procedure is that ultrametric trees can be estimated with partial rank orders (i.e., with data of less than full depth). Maximum likelihood ultrametric trees were obtained by using partial order datasets, with depths of 4, 8, and 12. These partial datasets, then, consisted of the 4, 8, or 12 closest stimuli to a given pivot stimulus, for all subjects. Each one of the 15 stimuli was used as a pivot stimulus. The maximum likelihood ultrametric tree obtained using depth 4 is given in Figure 7. This tree appears to divide the stimuli into two main classes: "less sweet foods" (buttered corn muffin, blueberry muffin, english muffin with margarine, buttered hard roll, toast with margarine, buttered toast, toast with marmalade, and toast with butter and jelly), and "sweet foods" (toasted pop-up, cinnamon toast, jelly donut, glazed donut, cinnamon bun, danish pastry, and coffee cake). Here, cinnamon toast is portrayed in a different cluster compared to Figures 5 and 6. In addition, Figure 7 does not portray the HRB distinction. Thus, the partial order ultrametric tree does differ from the complete order tree in some respects. However, the ultrametric tree obtained from the partial order data does capture much of the essence of the complete order ultrametric tree, with just a fraction of the required data. Figures 8 and 9 display the ultrametric trees estimated using data depths of 8 and 12 respectively. These ultrametric trees have fairly similar structures to that in Figure 5; again, they do not portray the buttered hard roll as a separate cluster, and some interchanges are made with the less sweet foods.

An ultrametric tree was also fit to the dataset using a least-squares algorithm applied to the ranks averaged over all subjects. The resulting ultrametric tree is shown

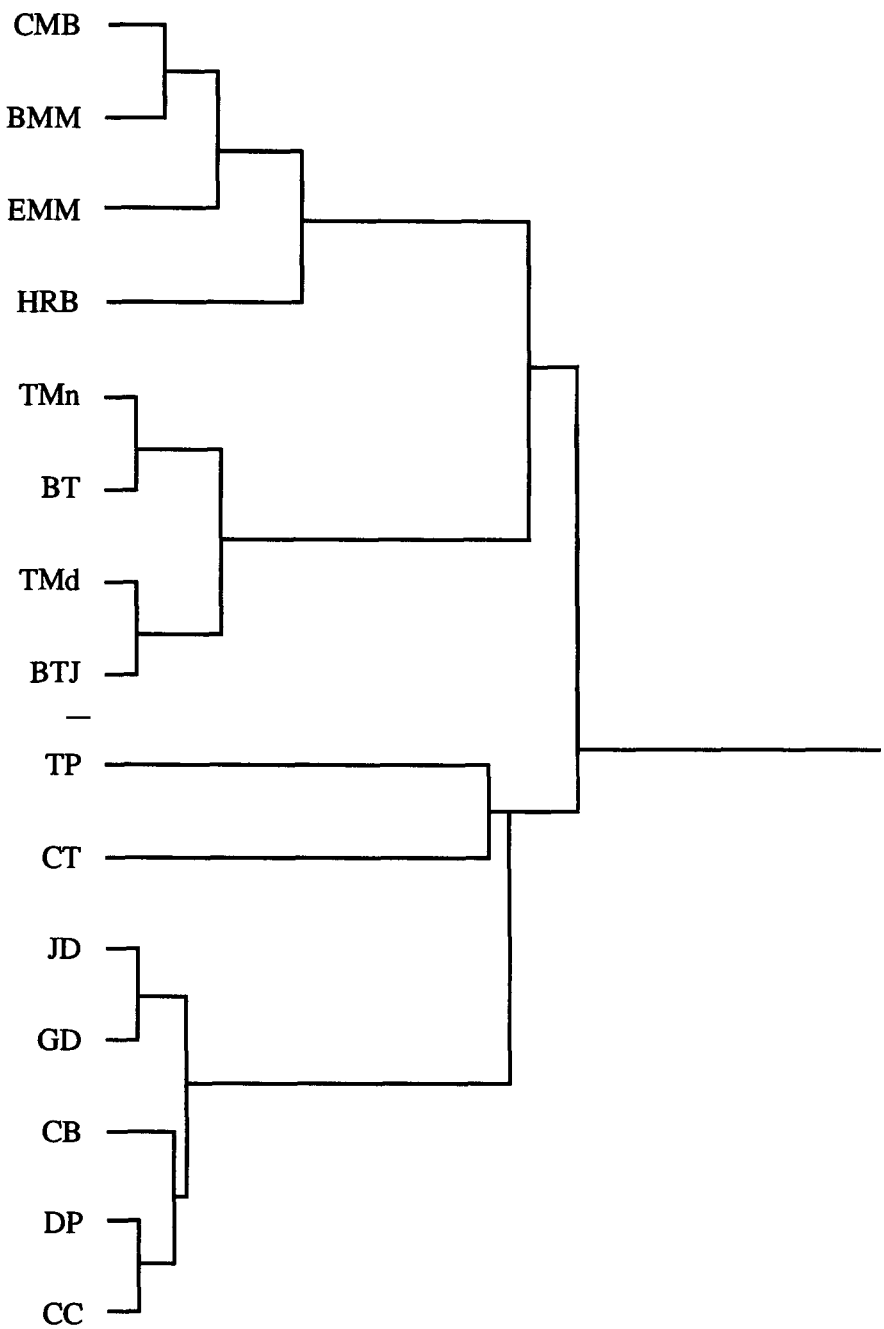

FIGURE 7.
Ultrametric Tree for Snack Data, Estimated Under H_1 with Depth 4 Data.

in Figure 10. Here, we see a number of anomalies. One, the buttered toast and jelly (BTJ) is portrayed as a distinct stimulus, but not the hard roll and butter (HRB). Yet, as shown in Table 4, there is no evidence to support this, as BTJ is one of the stimuli selected least frequently as "most distinct". Two, both cinnamon toast (CT) and buttered toast and jelly (BTJ) are classified with the sweeter breakfast/snack foods, and distinct from the other toasts. Finally, the muffins are separated. Thus, there seems to be little intuitive support for this solution.

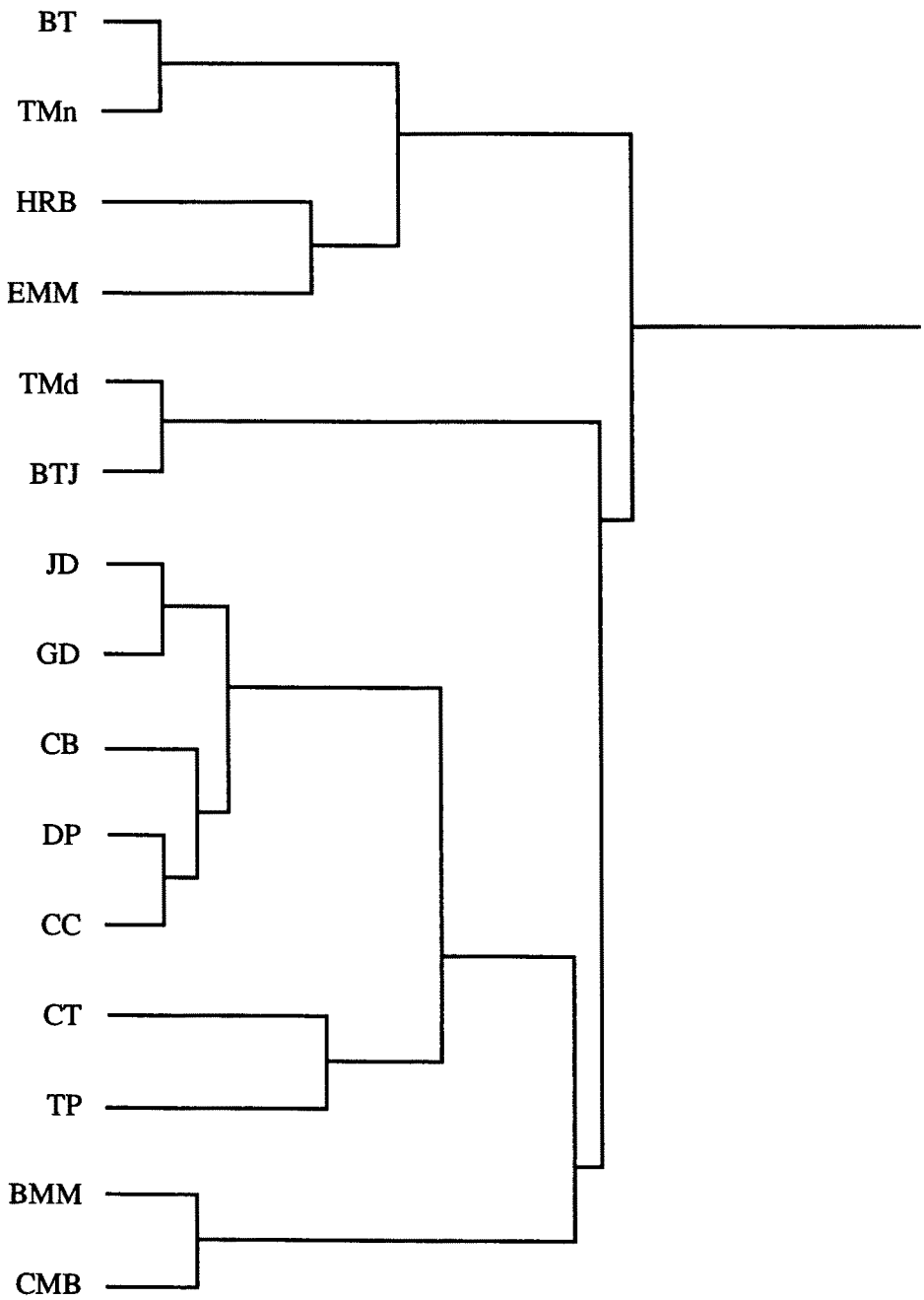

FIGURE 8.
Ultrametric Tree for Snack Data, Estimated Under H_1 with Depth 8 Data.

To statistically compare the least-squares ultrametric tree with that in Figure 5, the log likelihood was computed with the distance matrix $((d_{ij}))$ fixed at the least-squares solution, and β chosen to maximize the log-likelihood conditional on this fixed distance matrix. The log likelihood value was -15141; the data thus offers strong evidence for the ultrametric tree in Figure 5 over the least-squares tree, assuming the correctness of model (17). On the other hand, the sum-of-squares criterion for this least-squares tree

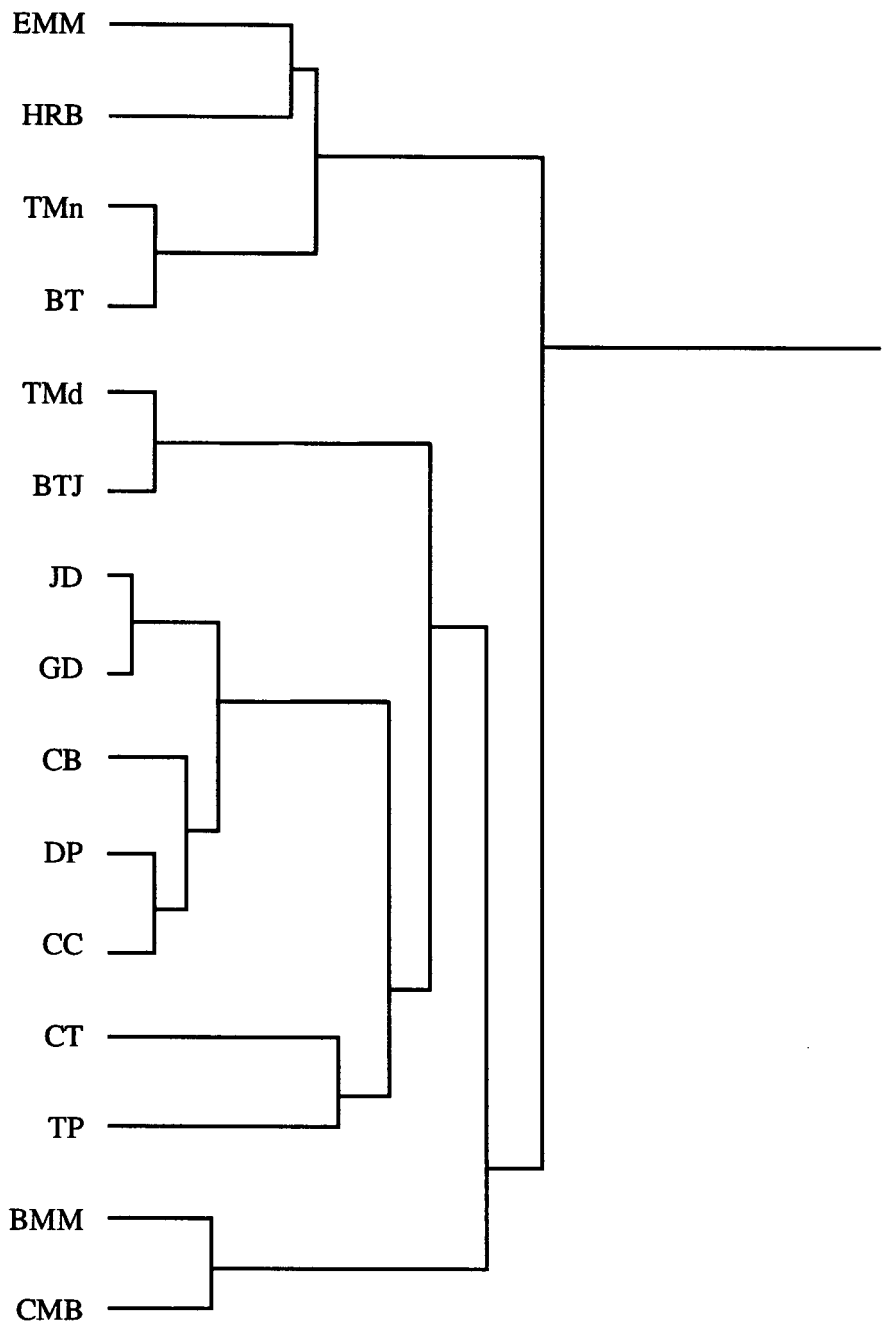

FIGURE 9.
Ultrametric Tree for Snack Data, Estimated Under H_1 with Depth 12 Data.

was 0.06, as opposed to 0.17 for the maximum likelihood tree. Each procedure, then, optimizes a different criterion and performs best with respect to its own maximand/minimand. However, the least-squares criterion makes little theoretical sense when dealing with ordinal scaled data. In addition, averaging ranks is not theoretically correct for such scale assumptions.

Finally, a two-dimensional Euclidean model was fit to the snack data using the

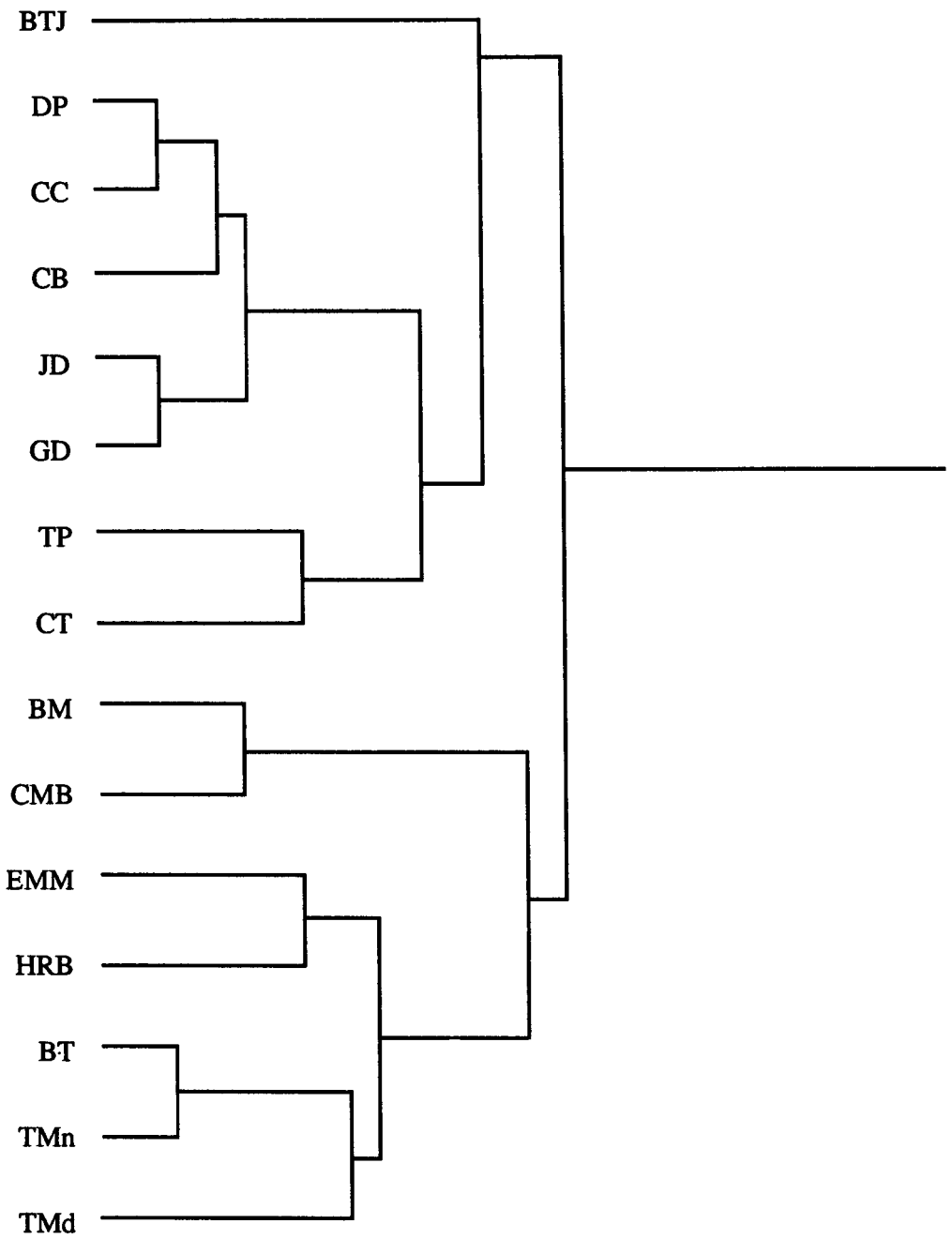

FIGURE 10.
Ultrametric Tree for Snack Data, Estimated by Least-Squares on Proximity Ranks.

nonmetric ALSCAL algorithm for row-conditional data (Young, 1987). This procedure jointly estimates a distance matrix $\underline{\underline{D}}$ and a monotone transformation $m(\underline{\Delta})$ of the observed dissimilarities $\underline{\Delta}$ by minimizing the stress measure:

$$\text{stress} = \left(\frac{\| m(\underline{\Delta}) - D^2 \|^2}{\| m(\underline{\Delta}) \|^2} \right)^{1/2}. \tag{39}$$

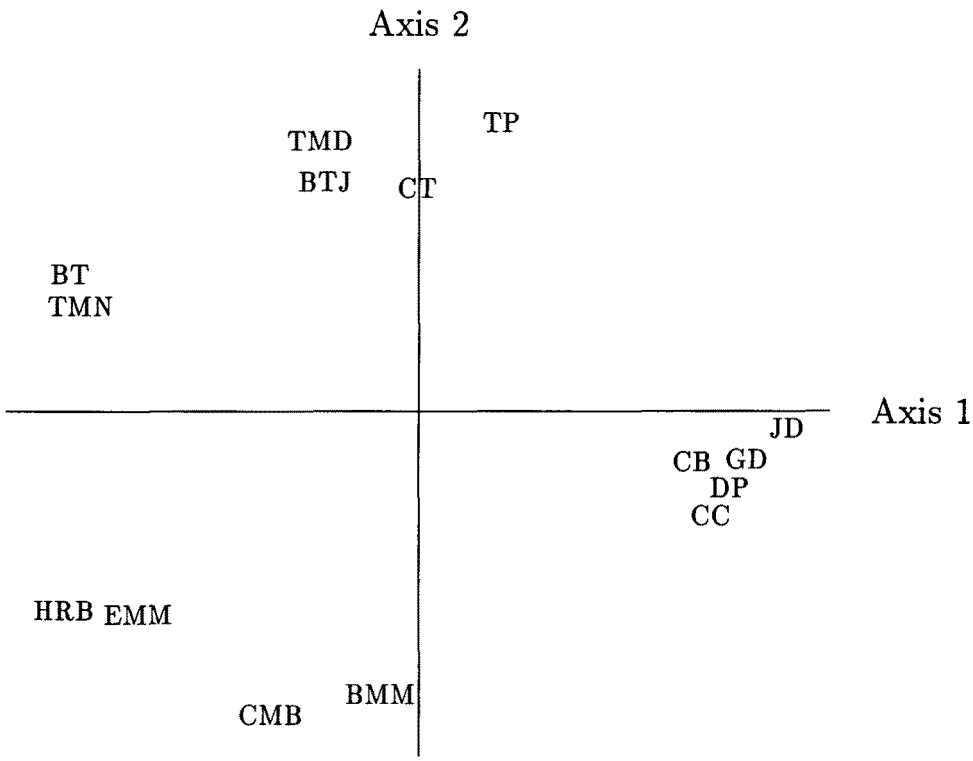

FIGURE 11.
Estimated ALSCAL Euclidean Configuration for Snack Data.

In the present context of conditional rank order data, the dissimilarities are just the ranks of the distances for each given subject and each given pivot stimulus. The term "row-conditional" refers to the fact that the ranks of distances are relative only to other values in the same row (same subject and same pivot stimulus), and not to values in other rows. The resulting two-dimensional configuration is given in Figure 11. Here, the vertical axis can be roughly interpreted as a sweetness dimension, while the horizontal dimension seems to represent a toast/non-toast dimension, quite similar to the MDS solutions reported in Green and Rao (1972). The discrete nature of the underlying representation is seen by the clumping of the stimuli in the space, suggesting that a tree or network may be a more suitable model for depicting the structure in the data.

The result obtained in these analyses, in which different scaling approaches lead to considerably different interpretations of a dataset, demonstrates the usefulness of having multiple models and estimation algorithms available for multivariate data analysis. Of course, it also alerts us to be concerned with the assumptions underlying a given model, and of the need to consider multiple alternatives when graphically representing proximity data.

5. Extensions

Section 2.1 described a model in which each individual's responses were determined by some common scale parameter β; the simple extension to individual specific scale parameters was also introduced. A middle ground between these two models is a *latent class* model, in which there are some small number, S, of different β's, $S \leq C$, and each individual's response is determined by one of the β_k, $k = 1, \ldots, S$. The likelihood in this case would simply be a finite mixture of likelihoods of the form (17).

Alternatively, one could consider a latent class model in which the different classes have different tree topologies as well as different scale parameters.

Another possible extension is the inclusion of covariates in the model. For example, it will often be the case that the level of interest in the stimuli, or the degree of experience with the stimuli, may affect the precision of the responses. The levels of interest and experience in turn may be measured by certain covariates. Determining the effects of covariates may be of interest in its own right; incorporating covariates may also be useful in providing more accurate fits of the true tree-configuration $\underline{D}$ via a relative down-weighting of inaccurate responses. Covariates could be included, for example, by modelling the scale parameter as $\beta_c = \exp(\underline{\theta}' \underline{x}_c)$, where $\underline{x}_c$ is the vector of covariates of respondent c, and $\underline{\theta}$ is a vector of parameters to be estimated. Alternatively, it is possible to imagine a response function in which the scale parameter depends not only on the respondent, but also on the particular stimuli being compared. Ramsay (1982) offers the model:

$$\beta_{ijc} = \tilde{\beta}_c \eta_{ij}^2, \tag{40}$$

with

$$\eta_{ij}^2 = \frac{a_i^2 + a_j^2}{2}, \sum_i a_i^2 = 1, \tag{41}$$

where β_{ijc} in (40) refers to the error magnitude associated with respondent c's evaluation of the distance between stimuli i and j. The parameter a_i in this model quantifies the inaccuracy associated with judging distances relative to stimulus i. This generalization could be fairly easily incorporated into the current procedure; all that would be required would be modifications to the routine evaluating the likelihood function and its gradient.

Finally, this tree estimation procedure can be easily extended to other tree topologies, such as additive trees (Carroll, 1976) or extended trees (Corter & Tversky, 1986). In addition, multiple trees and hybrid models involving combinations of trees and continuous Euclidean MDS spaces are also possible extensions that can be included in this stochastic framework.

APPENDIX: Sequential Quadratic Programming

Sequential quadratic programming (SQP) is a numerical technique for optimizing a smooth nonlinear function subject to linear constraints (Fletcher, 1987). The method is similar to the Newton or quasi-Newton optimization methods, in that both Newton methods and SQP methods make use of repeated quadratic approximations to the objective function. In the case of Newton's method, an objective function $f(\underline{x})$ is approximated by the Taylor series expansion:

$$f(\underline{x}) \approx Q(\underline{x}; \underline{x}_0) = f(\underline{x}_0) + \underline{g}(\underline{x}_0)'(\underline{x} - \underline{x}_0) + \frac{1}{2}(\underline{x} - \underline{x}_0)'\underline{H}(\underline{x}_0)(\underline{x} - \underline{x}_0), \tag{A1}$$

where $\underline{x}_0$ is some initial estimate of the optimum, $\underline{g}(\underline{x}_0)$ denotes the gradient of function $f(\underline{x})$ evaluated at $\underline{x} = \underline{x}_0$, and $\underline{H}(\underline{x}_0)$ denotes the Hessian of $f(\underline{x})$ evaluated at $\underline{x} = \underline{x}_0$. The quadratic function $Q(\underline{x}; \underline{x}_0)$ can be optimized analytically; this implies a new estimate for the optimal $\underline{x}$, namely:

$$\underline{x}_1 = \underline{x}_0 - \underline{H}^{-1}(\underline{x}_0)\underline{g}(\underline{x}_0). \tag{A2}$$

The function $f(\underline{x})$ is then expanded about $\underline{x}_1$, and the process is repeated; the general form of the Newton recursion is:

$$\underline{x}_{n+1} = \underline{x}_n - \underline{H}^{-1}(\underline{x}_n)\underline{g}(\underline{x}_n). \tag{A3}$$

The process is known to converge quadratically to the optimum for suitably continuous $f(\underline{x})$ (Fletcher, 1987, p. 46). In quasi-Newton schemes, the exact inverse of the Hessian $H^{-1}(\underline{x})$ is not used, but rather some computationally convenient approximation is employed; the so-called BFGS (Broyden, Fletcher, Goldfarb, and Shanno) scheme is one popular quasi-Newton method (Fletcher, 1987).

In sequential quadratic programming, a similar process of iterative approximation to the objective function is employed. Consider the linearly constrained nonlinear program:

$$\max f(\underline{x}) \tag{A4}$$

subject to:

$$\underline{A}_1\underline{x} = \underline{c}_1 \tag{A5}$$

$$\underline{A}_2\underline{x} \leq \underline{c}_2, \tag{A6}$$

where (A5) and (A6) represent equality and inequality constraints, respectively. An SQP procedure solves this problem by iteratively approximating $f(\underline{x})$ by a second order Taylor series expansion:

$$f(\underline{x}) \approx Q(\underline{x};\, \underline{x}_n) = f(\underline{x}_n) + \underline{g}(\underline{x}_n)'(\underline{x} - \underline{x}_n) + \frac{1}{2}(\underline{x} - \underline{x}_n)'\underline{H}(\underline{x}_n)(\underline{x} - \underline{x}_n), \tag{A7}$$

where $\underline{x}_n$ is the estimate of the optimal $\underline{x}$ at the nth iteration of the procedure, and then solving the problem:

$$\max Q(\underline{x};\, \underline{x}_n) \tag{A8}$$

subject to:

$$\underline{A}_1\underline{x} = \underline{c}_1 \tag{A9}$$

$$\underline{A}_2\underline{x} \leq \underline{c}_2. \tag{A10}$$

Problem (A8–A10) is a *quadratic program*—an optimization problem with quadratic objective function and linear constraints—and efficient methods exist for the solution of such problems; Fletcher (1987) describes a number of algorithms. The solution to problem (A8) is then treated as the new estimate of the optimum, $\underline{x}_{n+1}$, and a new approximation $Q(\underline{x};\, \underline{x}_{n+1})$ to the objective function is found; this process proceeds until convergence.

Our implementation of the maximum likelihood estimation procedure makes use of the FSQP sequential quadratic programming routine of Zhou and Tits (1993), which is based on Panier and Tits (1993). This SQP routine uses BFGS quasi-Newton approximations of the Hessian as the basis for the quadratic approximation of the objective function. The quadratic programming routine used to maximize the quadratic approximants is described in Schittkowski (1986), which is based on Powell (1983).

References

Abe, M. (1993). Issues in maximum likelihood multidimensional scaling (Working Paper). Chicago, IL: University of Illinois.

Barthélemy, J. P. & Guénoche A. (1991). *Trees and proximity representations*. New York: John Wiley & Sons.

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. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika, 41*, 439–63.

Carroll, J. D. (1989). Degenerate solutions in the non-metric fitting of a wide class of models for proximity data (Technical Memorandum). Murray Hill, NJ: Bell Laboratories.

Carroll, J. D., & Arabie P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*, 607–649.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 285–319.

Carroll, J. D., Clark, L., & DeSarbo, W. S. (1984). The representation of three-way proximities data by single and multiple tree structure models. *Journal of Classification, 1*, 25–74.

Carroll, J. D., & Pruzansky, S. (1975, August). *Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares*. Paper presented at U.S.-Japan Seminar of Multidimensional Scaling, University of California at San Diego, La Jolla, California.

Carroll, J. D., & Pruzansky S. (1980). Discrete and hybrid scaling models. In E. D. Lantermann & H. Feger, (Eds.), *Similarity and choice* (pp. 48–69). Bern: Hans Huber.

Chandon, J. L., Lemaire, J., & Pouget, J. (1980). Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés [The construction of ultrametric trees from dissimilarity matrices]. *R.A.I.R.O., Recherche Operationelle, 14*, 157–170.

Chapman, R., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research, 19*, 288–301.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons.

Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika, 51*, 429–451.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 187–202.

Critchlow, D. E., Filgner, M. A., & Verducci, J. S. Probability models on rankings. *Journal of Mathematical Psychology, 35*, 294–318.

Cunningham, J. P. (1974, August). Finding the optimal tree realization of a proximity matrix. Paper presented at the Mathematical Psychology Meetings, Ann Arbor, MI.

Cunningham, J. P. (1978). Free trees and bidirectional trees as a representation of psychological distance. *Journal of Mathematical Psychology, 17*, 165–188.

DeSarbo, W. S., De Soete, G., Carroll, J. D., & Ramaswamy, V. (1988). A new stochastic ultrametric tree unfolding methodology for assessing competitive market structure and deriving market segments. *Applied Stochastic Models & Data Analysis, 4*, 185–204.

DeSarbo, W. S., Manrai, A. K., & Burke, R. (1990). A non-spatial methodology for the analysis of a two-way proximity data incorporating the distance-density hypothesis. *Psychometrika, 55*, 229–253.

DeSarbo, W. S., Manrai, A., & Manrai, L. (in press). Mathematical programming approaches for the non-spatial assessment of competitive market structure: An integrated review of the marketing and psychometric literature. In G. Lilien & J. Eliashberg (Eds.), *Marketing models*. New York: Kluwer Pub.

De Soete, G. (1983a). Are nonmetric additive-tree representations of numerical proximity data meaningful? *Quality and Quantity, 13*, 475–478.

De Soete, G. (1983b). A least-squares algorithm for fitting trees to proximity data. *Psychometrika, 48*, 621–26.

De Soete, G. (1984). A least-squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters, 2*, 133–37.

De Soete, G., Carroll, J. D., & DeSarbo, W. S. (1987). Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data. *Journal of Classification, 4*, 155–74.

De Soete, G., DeSarbo, W. S., Furnas, G. W., & Carroll, J. D. (1984a). Tree representations of rectangular proximity matrices. In E. Degreef & J. Van Buggenhaut (Eds.), *Trends in mathematical psychology*. Amsterdam: North-Holland.

De Soete, G., DeSarbo, W. S., Furnas, G. W., & Carroll, J. D. (1984b). The estimation of ultrametric and path length trees from rectangular proximity data. *Psychometrika, 49*, 289–310.

Dobson, J. (1974). Unrooted trees for numerical taxonomy. *Journal of Applied Probability, 11*, 32–42.

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition, 11*, 235–254.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist, 106*, 645–668.

Fiacco, A. V., & McCormick, G. P. (1968). *Nonlinear programming*. New York: John Wiley & Sons.

Fletcher, R. (1987). *Practical methods of optimization* (2nd. ed.). New York: John Wiley & Sons.

Fligner, M. S., & Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association, 83*, 892–901.

Fligner, M. A., & Verducci, J. S. (1993). *Probability models and statistical analyses for ranking data*. New York: Springer-Verlag.

Furnas, G. W. (1980). *Objects and their features: The metric representation of two class data. Unpublished doctoral dissertation, Stanford University.*

Green, P. E., & Rao, V. R. (1972). *Multidimensional scaling*. Hinsdale, IL: Dryden Press.

Guttman, L. A. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika, 33*, 469–506.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association, 62*, 1140–1156.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hausman, J. A., & Ruud, P.A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics, 34*, 83–104.

Holman, E. W. (1972), The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika, 37*, 417–23.

Jardine, C. J., Jardine, N., & Sibson, R. (1967). The structure and construction of taxonomic hierarchies. *Mathematical BioScience, 1*, 173–79.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Kalbfleisch, J. D., & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika, 60*, 267–279.

Katahira, H. (1990). Perceptual mapping using ordered logit analysis. *Marketing Science, 9*(Winter, 1), 1–17.

Keener, R. W., & Waldman, D. M. (1985). Maximum likelihood regression of rank censored data. *Journal of the American Statistical Association, 80*, 385–392.

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

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden-Day.

Panier, E. R., & Tits, A. L. (1993). On combining feasibility, descent, and superlinear convergence in inequality constrained optimization. *Mathematical Programming, 59*, 261–276.

Peto, R. (1972). Discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B, 34*, 205–207.

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical Programming, 12*, 241–254.

Powell, M. J. D. (1983). ZQPCVX, A FORTRAN subroutine for convex programming (Report DAMTP/1983/NA17). Cambridge: University of Cambridge, England.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C*. New York: Cambridge University Press.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika, 47*, 3–24.

Ramaswamy, V., & DeSarbo, W. S. (1990). SCULPTRE: A new methodology for deriving and analyzing hierarchical product-market structures from panel data. *Journal of Marketing Research, 27*, 418–427.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42*, 241–266.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling (with discussion). *Journal of the Royal Statistical Society, Series A, 145*, 285–312.

Roskam, E. E. (1970). The methods of triads for multidimensional scaling. *Nederlands Tijdschrift Voor de Psychologie en haar grensgebieden, 25*, 404–417.

Ryan, D. M. (1974). Penalty and barrier functions. In P. E. Gill & W. Murray (Eds.), *Numerical methods for constrained optimization* (pp. 175–190). New York: Academic Press.

Schittkowski (1986). *QLD—A FORTRAN code for quadratic programming, User's Guide*. Bayreuth, Germany: Universität of Bayreuth, Mathematisches Institut.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I and II. *Psychometrika, 27*, 125–140, 219–246.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*, 390–398.

Takane, Y., & Carroll, J. D. (1981). Nonmetric maximum likelihood multidimensional scaling from directional rankings of similarities. *Psychometrika, 46*, 389–405.

Torgerson, W. W. (1952). Multidimensional scaling: Theory and method. *Psychometrika, 17*, 401–419.

Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review, 84*, 327–52.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of American Statistical Association, 58*, 236–244.

Winsberg, S., & Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via a restricted case of an extended INDSCAL model. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 405–414). Amsterdam: North Holland.

Young, F. W. (1975). Scaling replicated conditional rank-order data. *Sociological Methodology, 12*, 129–170.

Young, F. W. (1987). Data theory. In R. M. Hamer (Ed.), *Multi-dimensional Scaling: History, theory and applications* (pp. 43–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Young, F. W., & Torgerson, W. S. (1967). TORSCA: A FORTRAN IV program for Shepard-Kruskal multidimensional scaling analysis. *Behavioral Science, 12*, 498.

Zhou, J. L., & Tits, A. L. (1993). User's guide for FSQP Version 3.3: A FORTRAN code for solving constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality and linear constraints (Tech. Rep.). College Park, MD: University of Maryland, Department of Electrical Engineering.

Manuscript received 4/15/93
Final version received 11/9/93