PSYCHOMETRIKA—VOL. 51, NO. 3, 429–451
SEPTEMBER 1986

EXTENDED SIMILARITY TREES

JAMES E. CORTER

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

AMOS TVERSKY

STANFORD UNIVERSITY

Proximity data can be represented by an extended tree, which generalizes traditional trees by including marked segments that correspond to overlapping clusters. An extended tree is a graphical representation of the distinctive features model. A computer program (EXTREE) that constructs extended trees is described and applied to several sets of conceptual and perceptual proximity data.

Key words: proximities, nonhierarchical clustering, additive trees, feature models.

Introduction

Trees are commonly used to represent proximity relations that emerge, for instance, from studies of classification, similarity, and identification. Trees are employed to describe the data, explore their structure and model their generating process. They offer a convenient graphical display that is readily interpretable in terms of a hierarchy of clusters (Sokal & Sneath, 1963) or in terms of common and distinctive features (Tversky, 1977).

The simplest tree structure is the hierarchical clustering model (Jardine & Sibson, 1971; Johnson, 1967) based on the ultrametric inequality, which states that for any triple of points the two larger distances are equal. That is, any three points can be labeled x, y, z such that $d(x, z) = d(y, z) \geq d(x, y)$. This assumption gives rise to a tree in which all the endpoints (leaves) are equally distant from the root. The ultrametric tree is highly restrictive because any two elements of one cluster must be equally similar to any other element outside the cluster. This restriction is relaxed in the additive tree (e.g., Cunningham, 1978; Sattath & Tversky, 1977), where the leaves are not necessarily equidistant from the root. The additive tree provides greater flexibility than the ultrametric tree, but it too cannot accomodate (nonnested) overlapping clusters because any two clusters in a tree are either nested or disjoint. Throughout the paper we use the standard abbreviations (e.g., HICLUS, ADDTREE, ADCLUS) for scaling algorithms, and the unabbreviated forms (e.g., hierarchical clustering, additive tree, additive clustering) for the respective models.

This article describes a new representation of proximity relations, called an extended tree, which accommodates nonnested feature structures while maintaining the basic property of a tree that every pair of points is joined by a unique path. To motivate and

This research was supported in part by a National Science Foundation Pre-doctoral Fellowship to the first author.

Requests for reprints should be sent to James E. Corter, Box 41, Teachers College, Columbia University, New York, NY 10027. A magnetic tape containing both the EXTREE program described in the article and ADDTREE/P program for fitting additive trees can also be obtained from the above address. Requests for the program should be accompanied by a check for $25 made out to Teachers College, to cover the costs of the tape and postage.

0033-3123/86/0900-7004$00.75/0

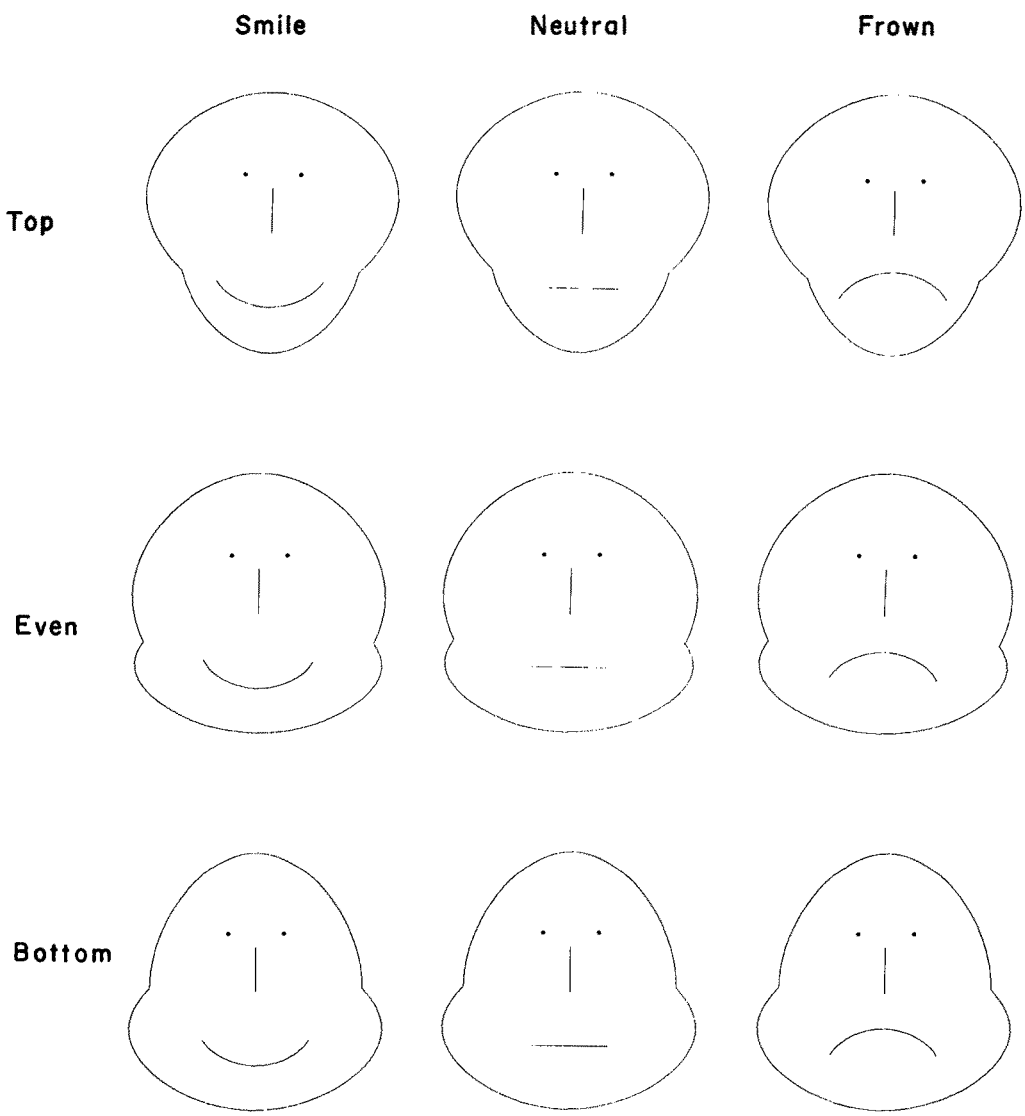


FIGURE 1
A factorially constructed set of schematic faces.

illustrate the proposed representation we begin with an analysis of the proximity between faces.

Similarity of Faces

Figure 1 presents nine schematic faces (see Chernoff, 1973) constructed factorially using three different face shapes (Top-Heavy, Even, Bottom-Heavy) and three different expressions (Smile, Neutral, Frown). The subjects ($N = 39$) were asked to rate the similarity between all 36 pairs of faces on a nine-point scale, where 1 denotes very low similarity and 9 denotes very high similarity. The average ratings appear below the diagonal in Table 1.

The rating data reflect the factorial structure of the stimulus set: Each face was judged most similar to the other faces having either the same shape or the same ex-

TABLE 1

Average Similarity Ratings (below diagonal) and Percentage of Confusions (above diagonal) Between the Faces of Figure 1.

	TS	TN	TF	ES	EN	EF	BS	BN	BF
TS	---	6.33	6.02	4.68	2.78	1.99	11.58	4.10	4.89
TN	7.23	---	6.12	2.11	8.19	1.55	4.11	14.60	3.33
TF	6.44	6.67	---	2.68	1.45	6.86	4.24	2.44	12.14
ES	5.23	2.97	2.56	---	7.99	7.96	11.02	3.66	4.23
EN	2.87	5.28	2.64	6.95	---	4.29	2.33	14.70	2.44
EF	2.46	2.69	5.46	6.18	6.62	---	4.99	2.31	12.92
BS	6.31	3.64	3.05	5.44	3.38	2.87	---	5.76	5.90
BN	3.49	6.03	3.21	3.28	5.13	3.26	6.79	---	4.99
BF	2.79	3.15	5.87	2.77	2.90	4.85	6.51	6.90	---

pression. Such a product structure with two nominal factors, however, is not readily represented by a two-dimensional solution. A two-way factorial design can be naturally embedded in the Euclidean plane only if each of the factors forms a unidimensional array. In general, however, proximity data generated by the product of two nominal factors, having n and m levels respectively, may require as many as $(n - 1)(m - 1)$ dimensions. Because neither the expressions nor the shapes of the faces in Figure 1 are exactly unidimensional, an adequate account of the data of Table 1 may require as many as four dimensions. Indeed, the multidimensional KYST solutions (Kruskal, Young & Seery, 1977) for these data in two, three, and four dimensions accounted for 56%, 69% and 99% of the variance, respectively.

An additive tree also cannot adequately represent a nominal factorial structure. The ADDTREE solution of the similarity between the faces is presented in Figure 2. Recall that the distance between objects in a tree is given by the (horizontal) length of the path that joins them; the vertical lines are introduced for graphical convenience. Figure 2 reveals three clusters corresponding to the three shapes. Once the faces are grouped by shape, however, the tree has no means of representing the fact that faces with the same expression are more similar to each other than faces with different expressions. Indeed,

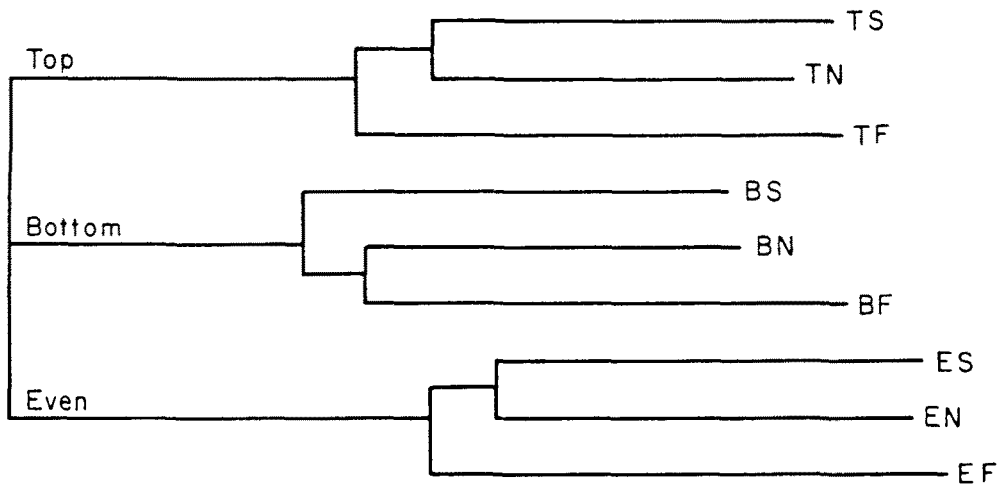


FIGURE 2
Additive tree (ADDTREE) representation of the similarity between the faces of Figure 1. Percentage of variance explained by the model, PV = 58%.

the additive tree accounted for only 58% of the variance, roughly the same as the two-dimensional solution.

A rooted additive tree can be interpreted as a feature tree, in which the length of each arc corresponds to the measure of the features shared by all the objects that follow from that arc (Tversky, 1977). For example, the length of the arc labeled *Top* in Figure 2 is the measure of the features shared by all three top-heavy faces. The terminal arcs of the tree correspond to the measure of the unique features of the respective objects, and the (horizontal) path-length distance between objects is the measure of their distinctive features. For example, the distance between *TS* and *TN* in Figure 2 is the sum of the measures of their unique features.

A feature structure is nested if any two clusters of objects (induced by these features) are either disjoint or one includes the other (see Tversky & Sattath, 1979). Nested feature structures are readily represented by an additive or an ultrametric tree.

In order to represent graphically feature structures that are not nested, we extend the tree model by introducing marked segments. A marked segment is an identified part of an arc which appears in more than one place in the tree and is used to denote a particular set of features. As in the simple tree (where all segments are unmarked) the distance between objects in an extended tree is given by the measure of their distinctive features. The difference between the simple and the extended tree is that marked segments representing features common to two objects do not enter into the computation of the path-length distance between the objects.

Figure 3 presents an extended tree representation of the factorial similarity data of Table 1, obtained using a computer program (EXTREE) described later in this article. The three marked segments, corresponding to the three facial expressions (Smile, Neutral, Frown) are denoted by *S*, *N*, *F*, respectively. The distance between *TS* and *BS*, then, is given by the horizontal length of the path between them, disregarding the two occurrences of the marked segment *S*. On the other hand, the marked segment *S* enters into the distance between *TS* and *BN* because they do not share the same expression, hence for these two faces *S* is a distinctive rather than a common feature. In this manner, the extended tree represents the pattern of similarity induced by the factorial structure: *TS* is closer to *BS* than to *BN*, whereas *BN* is closer to *TN* than to *TS*, and so on.

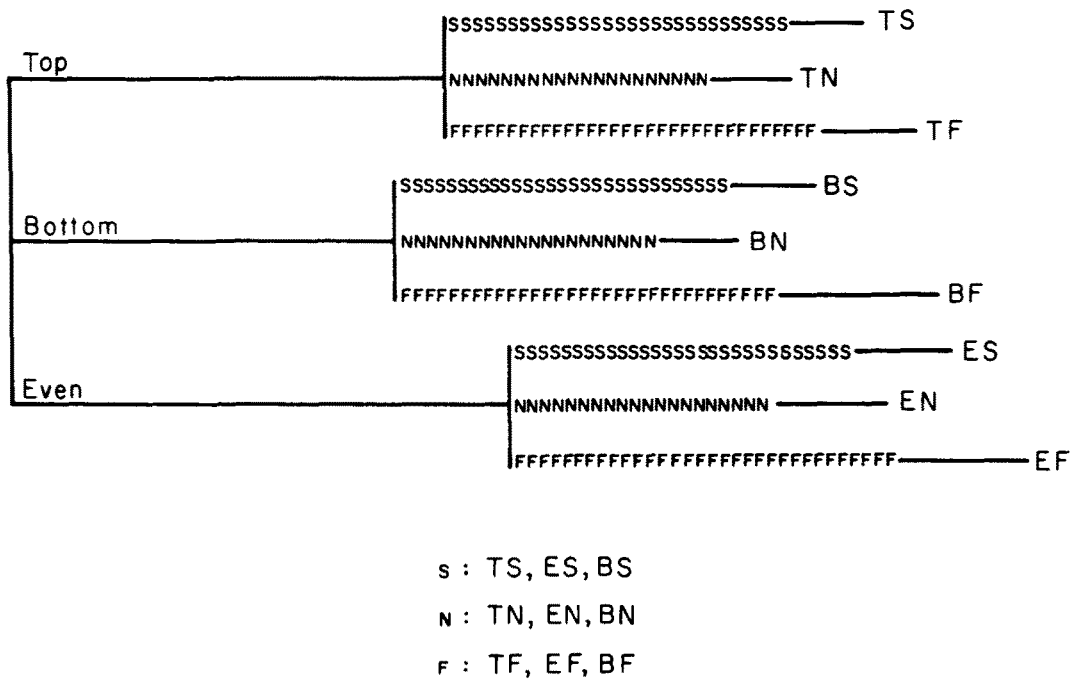


FIGURE 3
Extended tree (EXTREE) representation of the similarity between the faces of Figure 1. PV = 99%.

The EXTREE solution in Figure 3 accounted for 99% of the variance in the similarity ratings. The additive tree of Figure 2, in comparison, accounts for only 58% with the same number of free parameters (15). To achieve the same degree of goodness-of-fit with the spatial model, we need a four-dimensional solution, with more than twice as many free parameters. To test the stability of the EXTREE solution, we randomly divided the subjects into two equal groups and applied the EXTREE program to the data of the first (estimation) group. The solution was then correlated with the data of the second (validation) group. We repeated this procedure ten times. All ten solutions were essentially identical to the solution presented in Figure 3. Furthermore, the average correlation, across the ten splits, between the solution of the estimation group and the data of the validation group was .975.

Confusion Between Faces

The proximity between objects (e.g., faces) can be investigated by using direct judgments of similarity or by observing the confusion between them in a recall or an identification task. Naturally, the closer the objects the more likely they are to be confused. To illustrate the application of EXTREE to confusion data, we conducted a paired associates learning study in which the faces of Figure 1 were used as stimuli and nine common first names were used as responses. The assignment of names to faces was randomized separately for each subject. The subject was first told the name associated with each face. Faces were then presented in random order and the subject's task was to name each face. Feedback was given after each trial, and the task continued until the subject correctly named all nine faces twice in a row. The proximity of faces i and j was defined as the percentage of presentations of j incorrectly identified as i plus the percentage of presentations of i incorrectly identified as j. These values averaged across subjects ($N = 39$) are presented above the diagonal in Table 1.

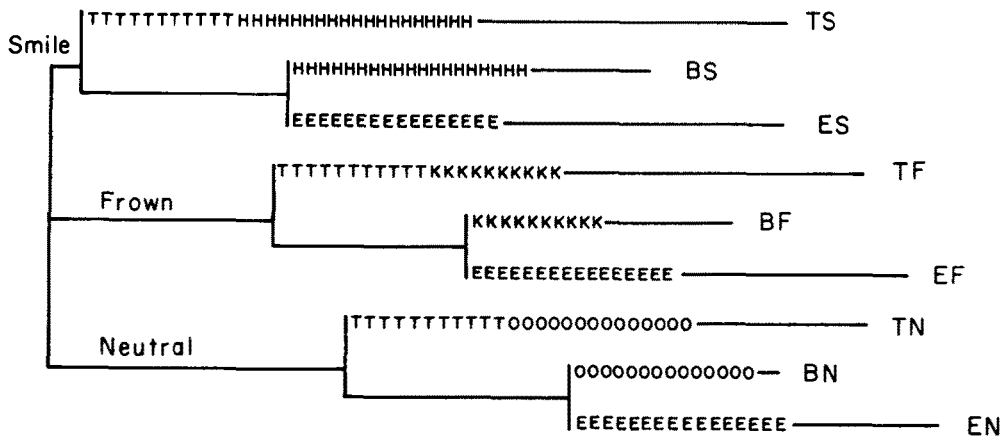


FIGURE 4
EXTREE solution for the confusion between the faces of Figure 1. PV = 98%.

The comparison of the two halves in Table 1 shows that the correlation between similarity and confusion is far from perfect ($r = .61$). Furthermore, the discrepancy between the two indices is highly systematic: Similarity ratings were influenced by shape more than by expression, whereas identification errors exhibited the opposite pattern. For each subject we computed the difference between the average similarity of the faces differing only in shape with the average similarity of the faces differing only in expression. Faces differing in shape were rated significantly less similar than those differing in expression ($t(38) = 14.14$, $p < .001$). In contrast, faces differing in expression were confused more often than faces differing in shape ($t(38) = 4.79$, $p < .001$). This pattern is reflected in the respective trees. Shape emerges as the primary classification in the trees (Figures 2 and 3) of the similarity ratings, whereas expression emerges as the primary classification in the trees derived from the confusion data (see Figure 4).

The Distinctive Features Model

The extended tree (see Figures 3 and 4) as well as the additive tree (Figure 2) can be viewed as instances of the distinctive features model, also called the symmetric difference metric (Restle, 1959). In this model each object x is characterized by a measurable collection of features, denoted X, and the distance between objects is given by the measure of their symmetric difference. Formally, consider a set of objects s, a set of features S, and a mapping that associates each object x in s with a set of features X in S. Both s and S are assumed to be finite. A dissimilarity function d satisfies the distinctive features model if

there exists an additive measure f on the subsets of S such that for all x, y in s

$$d(x, y) = f(X - Y) + f(Y - X) = \sum_{V \in X \Delta Y} f(V), \tag{1}$$

where $X - Y$ denotes the set of features that belong to X but not to Y, and $X \Delta Y = (X - Y) \cup (Y - X)$.

More specifically, consider the feature matrix $M = (m_{ij})$, $1 \leq i \leq k$, $1 \leq j \leq n$, where n is the number of objects in s and k is the number of features in S. The m_{ij} entry of M equals 1 if feature i belongs to object j, and 0 otherwise. For any pair of objects x, y in s, and any feature i in S, define

$$e_{i(x,y)} = \begin{cases} 1 & \text{if } M_{ix} \neq M_{iy} \\ 0 & \text{if } M_{ix} = M_{iy} \end{cases}. \tag{2}$$

The distinctive features model can now be expressed by

$$d(x, y) = \sum_{i=1}^{k} e_{i(x,y)} f_i, \tag{3}$$

where f_i is the measure of the i-th feature (Sattath & Tversky, 1985). In matrix form, $\mathbf{d} = \mathbf{Ef}$. In the example of Figure 3, $\mathbf{f}$ is a 15-component vector corresponding to 6 shared features (3 shape and 3 expression) and 9 unique features (one for each face).

It is not difficult to see that any set of distances generated by the distinctive features model can be represented as an extended tree. To demonstrate this proposition, construct a degenerate tree (i.e., a fan) having only one internal node. Set the length of the branch that corresponds to x equal to the measure, $f(X)$, of the set of features associated with x. The features that are unique to an object are represented by an unmarked segment. Any feature shared by two or more objects is represented by marked segments on the corresponding branches. It is easy to verify that in this representation the extended tree distance between x and y is the measure of their symmetric difference because any feature shared by x and y is marked and hence does not enter into the distance between them.

The EXTREE algorithm presented below first constructs an additive tree and then searches for additional clusters. It is not proposed as a general procedure for estimating the distinctive features model. Rather it is designed for an "imperfect" hierarchical structure that consists of a basic tree with a few nonnested clusters. In the next three sections we describe the algorithm, apply it to several data sets, and discuss the relation of the present development to other proximity models.

Fitting the Model

A program, EXTREE, has been written in the PASCAL language to fit the extended tree model. The model is fit in three stages. The first consists of fitting the best additive tree to the data. In the second stage the program estimates the weights of the marked features and selects a subset to be included in the model. The third stage consists of the simultaneous estimation of all parameters and the elimination of inconsequential features. Readers who are primarily interested in the applications of the model rather than in the algorithm may wish to skim or skip this section.

Fitting an Additive Tree

The first stage, fitting the best additive tree, is based on Sattath and Tversky's (1977) algorithm, as modified by Corter (1982). The input is a symmetric matrix of dissimilarities between all pairs of objects. A constant is added to all entries so that (a) all distances are

positive, (b) the triangle inequality is satisfied for all triples of objects, and (c) $d(x, y) + d(y, z) = d(x, z)$ for at least one triple x, y, z.

A dissimilarity matrix has an additive tree representation if every four objects can be labeled x, y, u, v, such that $d(x, y) + d(u, v) \leq d(x, u) + d(y, v) = d(x, v) + d(y, u)$ (Buneman, 1971). This condition is called the tree inequality or the four-point condition. If the data satisfy the tree inequality, the corresponding tree can be readily constructed. For fallible data, however, we need a procedure for identifying the tree structure that most closely approximates the data. We use a procedure based on neighbor scores (Sattath & Tversky, 1977), which provides an effective method for fitting additive trees to fallible data (Pruzansky, Tversky & Carroll, 1982).

For every quadruple of objects x, y, u, v we order the three sums: $d(x, y) + d(u, v)$, $d(x, u) + d(y, v)$, $d(x, v) + d(y, u)$. The object pairs corresponding to the smallest sum receive a score of 2, the object pairs corresponding to the intermediate sum receive a score of 1, and the object pairs corresponding to the largest sum receive a score of 0. The overall neighbor score for x and y is the sum of their scores across all quadruples including x and y. Using the matrix of neighbor scores, x and y are taken to be nearest neighbors and combined into a new cluster if the pair x, y has a larger neighbor score than any other pair. Cases of nonreciprocity (i.e., y is x's nearest neighbor, but x is v's nearest neighbor), and cases where y and v are tied as x's nearest neighbor are resolved by comparing $d(x, y) + d(v, Z)$ and $d(x, v) + d(y, Z)$, where $d(v, Z)$ is the average distance of v to all objects other than x, y, and v. We select y as x's nearest neighbor if the first sum is smaller than the second, and v otherwise.

Once the nearest-neighbor pairs are found, all such pairs are combined into clusters corresponding to branches of the tree. The distance between an object z and a new cluster comprised of x and y is defined as the average of $d(x, z)$ and $d(y, z)$. If x and y are themselves clusters the distances are weighted by the number of objects in each cluster. It can be shown that in an additive tree the length of the arcs joining the new node to its children, x and y, is a function of the distance between x and y and the distances from x and y to all other objects in the tree.

The process described above reduces the size of the matrix of distances by the number of clusters formed at that stage. From this new distance matrix we derive another matrix of neighbor scores, and new clusters are constructed. The procedure is repeated until only two or three clusters remain, at which point the topology of the (unrooted) tree is determined.

Because the rooting of the tree affects its interpretation, several options have been provided for the selection of the root. The default places the root on the arc joining the last two clusters. Another option selects the rooting that minimizes the variance of the distances from the leaves to the root. Finally, the user may choose the rooting by specifying the arc on which the root should be placed.

Selection of Marked Features

Note that the inequality part of the tree inequality is always satisfied for some labeling of a quadruple. However, the equality of the two larger sums may not hold, in which case the data cannot be perfectly represented by an additive tree. Consider four objects x, y, u, v for which

$$d(x, y) + d(u, v) < d(x, u) + d(y, v) < d(x, v) + d(y, u). \tag{4}$$

In Figure 5a this pattern of distances is represented by an unrooted extended tree with a marked segment, A, shared by x and u.

Introducing the marked segment A amounts to adding a parameter (θ). As shown in

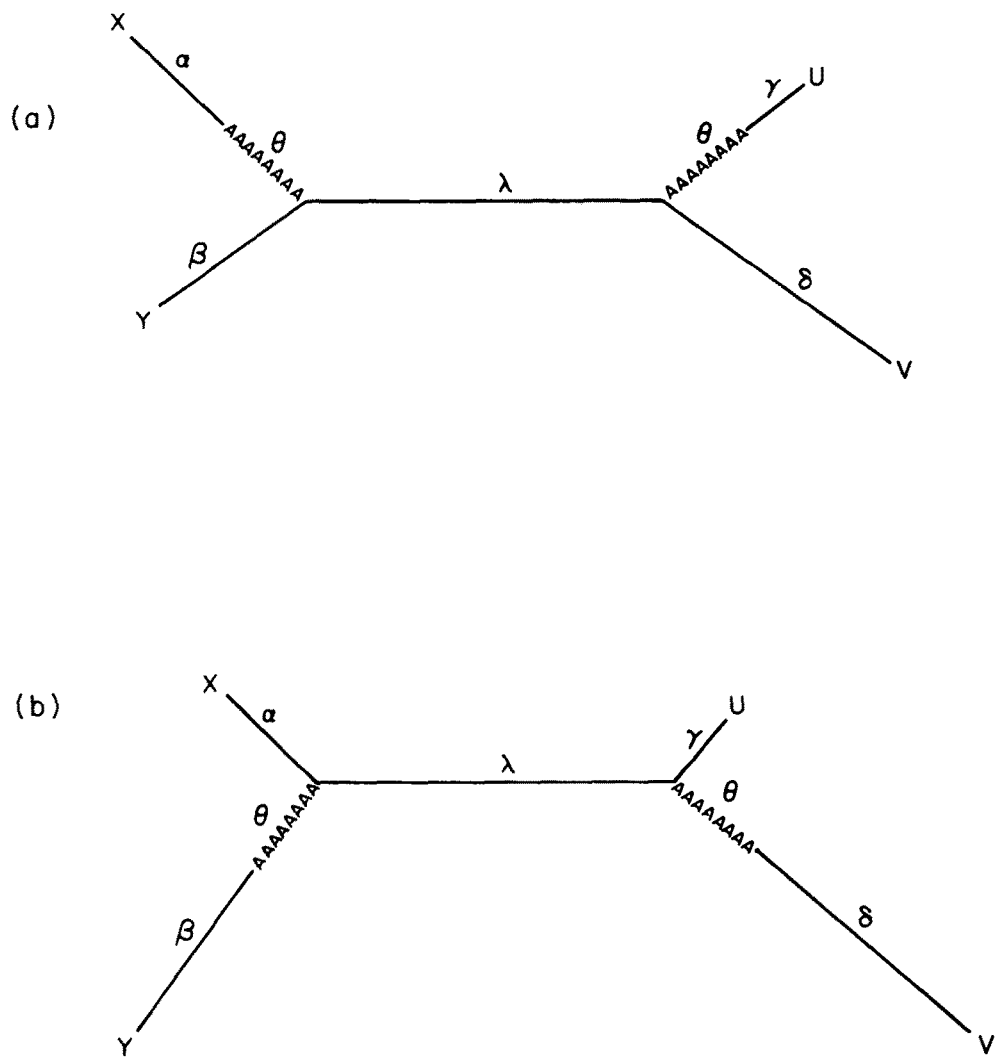


FIGURE 5
(a) An extended tree of four objects in unrooted form. (b) Alternative representation of the same structure.

Figure 5a, the distances among the four objects can be expressed by

$$\begin{aligned} d(x, y) &= \alpha + \theta + \beta, \\ d(u, v) &= \gamma + \theta + \delta, \\ d(x, u) &= \alpha + \lambda + \gamma, \\ d(y, v) &= \beta + \lambda + \delta, \\ d(x, v) &= \alpha + \theta + \lambda + \delta, \\ d(y, u) &= \beta + \lambda + \theta + \gamma. \end{aligned} \tag{5}$$

Note that the marked segment A represents a feature common to x and u, so θ does not

enter into the distance between x and u. Substituting these expressions in (4) yields

$$(\alpha + \theta + \beta) + (\gamma + \theta + \delta) < (\alpha + \lambda + \gamma) + (\beta + \lambda + \delta)$$
$$< (\alpha + \theta + \lambda + \delta) + (\beta + \lambda + \theta + \gamma),$$

or

$$\theta < \lambda < \theta + \lambda.$$

Note that if $\theta = 0$ the two largest sums in (4) are equal so the tree inequality holds.

It is easy to show that there exists a positive solution to (5), provided the triangle inequality is satisfied. For example,

$$\tfrac{1}{2}[d(x, y) + d(x, u) - d(y, u)]$$
$$= \tfrac{1}{2}[(2\alpha + \beta + \gamma + \lambda + \theta) - (\beta + \gamma + \lambda + \theta)],$$
$$= \alpha.$$

A similar argument applies to β, γ and δ.

The values of θ and λ are positive by (4), because

$$\tfrac{1}{2}[d(x, v) + d(y, u) - d(x, u) - d(y, v)]$$
$$= \tfrac{1}{2}[(\alpha + \beta + \gamma + \delta + 2\lambda + 2\theta) - (\alpha + \beta + \gamma + \delta + 2\lambda)]$$
$$= \theta,$$
$$\tfrac{1}{2}[d(x, v) + d(y, u) - d(x, y) - d(u, v)]$$
$$= \tfrac{1}{2}[(\alpha + \beta + \gamma + \delta + 2\lambda + 2\theta) - (\alpha + \beta + \gamma + \delta + 2\theta)]$$
$$= \lambda.$$

Although, as shown above, the solution to (5) is unique, the graphical representation in Figure 5a is not. Figure 5b presents an alternative representation in which marked segment A corresponds to a feature common to y and v, rather than to x and u. The two graphs are equivalent in the sense that they both represent the distances in (5). The non-uniqueness arises because the set of objects sharing the marked feature in the first graph (x, u) is the complement of the set of objects sharing the marked feature in the second graph (y, v). However, in either graph θ enters only into the distances between the two sets. This indeterminacy of the feature structure is a general characteristic of the distinctive features model, as shown by Sattath and Tversky (1985).

However, just as for an additive tree, the choice of a root for an extended tree restricts the placement of marked segments. For example, in Figure 6, the marked segment C has a common-feature interpretation, but a marked segment on the arcs labeled C? would not because the sets of objects "below" the two arcs in the rooted tree are not disjoint.

The placement of a marked feature may not be determined by the choice of a root for the additive tree, in which case the program represents the marked feature as common to the two branches that are closer in the data. In the example of Figure 5, $d(x, u) < d(y, v)$. Hence the marked segment A is represented as a feature common to x and u, as in Figure 5a. Use of this heuristic is largely a matter of convenience; the placement can also be specified by the user, and it does not affect the representation of the distances.

The program estimates the lengths of all possible marked segments in the tree. The

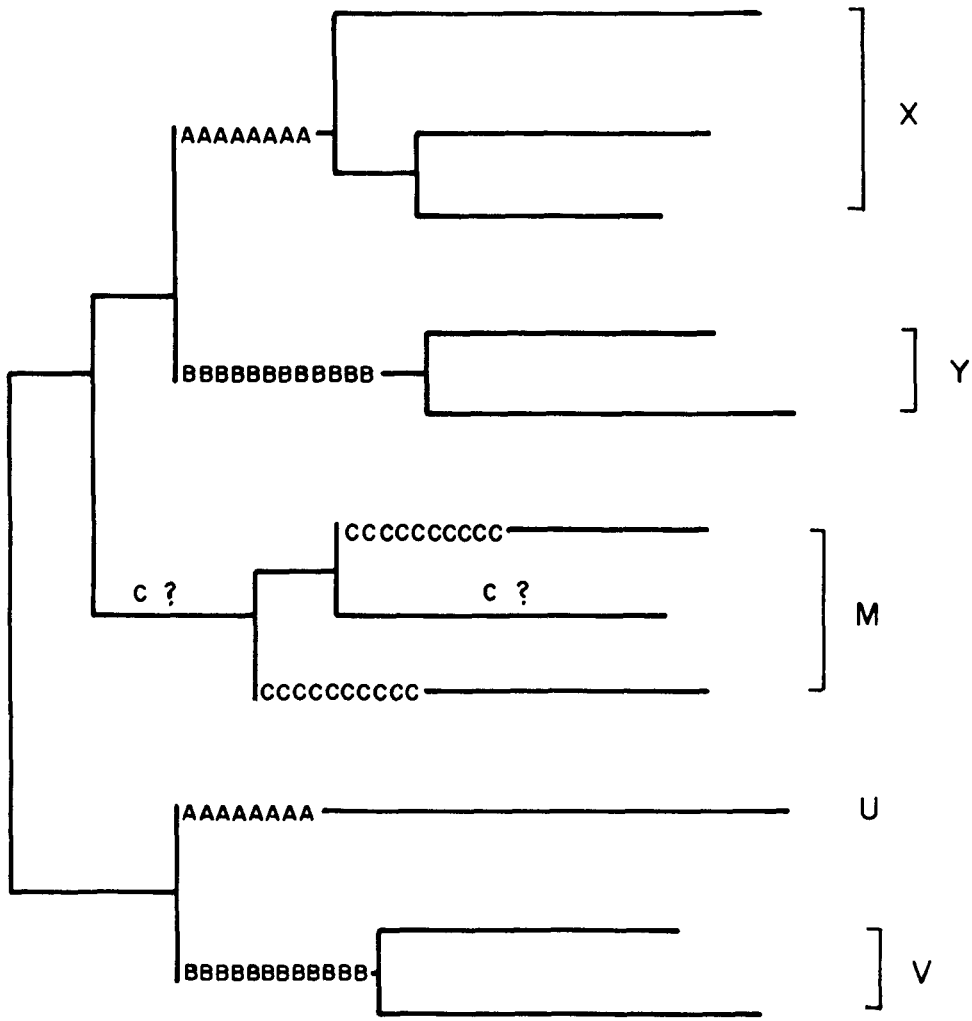


FIGURE 6
An extended tree of more than four objects.

estimate of the lengths of a segment shared by branches X and U is given by

$$\frac{1}{2N} \sum_{Q} [d(x, v) + d(y, u)) - (d(x, u) + d(y, v))],$$

where Q is the set of all quadruples of objects (x, y, u, v) in the tree with structure $((x, y)(u, v))$, where $x \in X$, $u \in U$, and N is the number of such quadruples. This is the least-squares estimate that would result if that feature alone were used to predict the residual distances from the tree.

Once the lengths of all marked features are estimated, we seek a relatively small subset of them that accounts for a large proportion of the residual variance in the distances, given the tree structure. Initially, the k features with largest estimates are selected, where k is a number specified by the user, or set by default to half the number of objects. However, some of these k features may be eliminated or combined. A redundant pattern of features arises when there exists a feature common to branch x and to the branch

consisting of u and v, as well as a feature common to x and u, and a feature common to x and v. In this case the smallest of the three features is eliminated. Also, the marked segments estimated so far represent features common to *pairs* of branches. A more parsimonious representation may often be achieved by constructing higher-order marked features. For example, if x and u share a feature, as do x and y, and y and u, then it is assumed that x, y and u all share a single feature, which can be represented by a single marked segment on the three branches. The search for higher order features is equivalent to the search for cliques in the general graph defined by the set of selected marked features (where the nodes correspond to branches in the extended tree, and the arcs are defined by the selected marked features). A similar rule was used by Shepard and Arabie (1979) to search for a parsimonious set of clusters.

Fitting of the Full Model

After the selection of marked features, the least-squares estimates of the parameters (i.e., arc lengths) are obtained by multiple regression. A matrix $\mathbf{P}$ is defined in which each row corresponds to a parameter of the model, and each column to one of the interobject distances. The p_{ij} entry of this matrix is 1 if the i-th feature is included in the j-th distance, and 0 otherwise. We then solve the matrix equation

$$\mathbf{E'Pf} = \mathbf{P'd},$$

where $\mathbf{f}$ is the vector of unknown parameters, $\mathbf{d}$ is the vector of observed distances, and $\mathbf{E}$ is the distance-feature matrix defined in (2).

The initial estimates of the parameters are revised by eliminating all negative estimates as well as positive estimates that do not exceed a preset threshold. The default value of the threshold is 2.5% of the largest observed distance. The solution is then reestimated without the eliminated parameters. These steps are repeated, if necessary, until no parameter estimates are less than the threshold.

Simulation Study

To test the effectiveness of the EXTREE algorithm, we conducted a small-scale simulation study. Because the program is designed for structures that are primarily—but not fully—hierarchical, we generated extended trees by selecting additive trees and adding to them a few nonnested (i.e., overlapping) clusters. The additive trees were generated as follows. First, an integer was assigned to each of n objects ($n = 10, 20$). An integer $1 \leq i < n$ was selected at random and objects i and $i + 1$ were grouped together. The process was repeated until all objects were clustered. A set of k overlapping clusters ($k = 2, 5$) was then selected at random. Each of these clusters included 2, 3 or 4 branches with probability 1/2, 1/3, and 1/6, respectively. Once the nested and overlapping clusters were identified, the weight of each cluster was drawn from a uniform distribution on the unit interval, and the distances between objects were generated according to the distinctive features model, (3).

For each of the four combinations of the number of objects (n) and the number of nonnested clusters (k), 25 data sets were generated and analyzed by EXTREE.

The correlations between the solutions and the data were perfect (average $r^2 = 1.00$) when $k = 2$ for both $n = 10$ and $n = 20$. When $k = 5$ the average r^2 reduces to .98 for $n = 10$, and to .99 for $n = 20$. These results suggest that the program is reasonably effective although there is room for improvement.

In a second condition, random normal error (with variance equal to 1/4 of the variance of the generated distances) was added. The average r^2 for the fallible data when $k = 2$ was .93 for $n = 10$ and .88 for $n = 20$. When $k = 5$ the average r^2 was .91 for $n = 10$

and .87 for $n = 20$. The decline in fit due to the present level of noise is roughly comparable in magnitude to that observed in a simulation study of ADDTREE and KYST (Pruzansky et al., 1982).

Applications

In this section we describe the application of EXTREE to several sets of data reported in the literature. The results of the analyses are summarized in Table 2. This table presents for each data set the number of parameters used in the ADDTREE and the EXTREE solutions (both marked and unmarked). It also reports the percentage of variance (PV) explained by each solution, and the F statistic used to test the (null) hypothesis that all the marked segments should equal zero. Because the standard stochastic assumptions that underlie the conventional F test (random sampling, independent error terms) are questionable in the present setting we adopt a nonstochastic interpretation of significance testing developed by Freedman and Lane.

Freedman and Lane (1983a, 1983b) showed that the traditional significance level associated, say, with an F test can be interpreted as the proportion of data sets obtained from the original one by permuting residuals for which the F statistic exceeds the observed value. In this interpretation the level of significance is a descriptive statistic; a small p value merely indicates an extreme data set. This approach is similar to the logic of randomization tests but it permutes observed residuals rather than unobservable (stochastic) disturbance terms.

The application of the Freedman-Lane approach to the present setting involves the following steps. First, the observed dissimilarity vector d is predicted (by $\hat{d}$) using only the unmarked segments, derived from ADDTREE. Second, a vector of residuals $r = d - \hat{d}$ is computed. Third, new data sets of the form $\hat{d} + r^*$ are constructed, where r^* is a permutation of r, (i.e., a rearrangement of the components of the residual vector r). If the marked segments added by EXTREE to the regression equation do not improve the prediction of d, the observed value of the F statistic should be comparable to the F values of the data sets generated by permuting the residuals. On the other hand, if only a very small fraction of the generated data sets give rise to higher F ratios, the contribution of the marked segments is probably not accidental. Table 2 shows that in all but one case (languages) the extended tree appears to fit the data significantly better than the additive tree. These applications are described in turn.

Identification of Digits

Keren and Baggen (1981) investigated the confusability of rectangular digits of the kind used in digital watches and calculators. The digits were presented on a tachistoscope for a very brief interval. Exposure time was adjusted to result in about 30% confusions for each subject. The frequency of confusing i with j plus the frequency of confusing j with i, pooled over the 8 subjects, was taken as a measure of proximity. The resulting distance matrix obtained by multiplying all entries by -1 and adding a positive constant was analyzed by EXTREE. The initial additive tree accounted for 76% of the variance, and this value was raised to 93% with the addition of only four marked segments. The obtained solution is presented in Figure 7.

The basic (unmarked) tree structure includes four binary clusters: (0, 8), (6, 5), (9, 3), (1, 7). Note that the two elements in each pair differ only in one line segment. Other digits that differ by one segment only, however, cannot be represented by the same additive tree. Some of these clusters emerge in the EXTREE solution. In particular, the marked segments E and C form the clusters (5, 9) and (6, 8), respectively. The remaining marked segments connect digits that differ in two segments ((1, 4) and (6, 0)).

TABLE 2

Summary of EXTREE Application to Seven Data Sets

Stimuli	Figure	Objects	ADDTREE		EXTREE			
			Parameters	PV	Parameters Unmarked	Parameters Marked	PV	F(df)
Faces (rating)	3	9	15	58%	12	3	99%	199.6** (3,17)
Faces (confusion)	4	9	15	76%	15	5	98%	37.0** (5,15)
Digits (visual)	7	10	17	76%	15	4	93%	13.9** (4,23)
Numbers (concept)	8	10	17	62%	17	5	90%	12.3** (5,22)
Languages	9	7	11	88%	11	3	94%	2.4 (3,6)
Furniture	10	20	37	83%	32	8	90%	13.1** (8,144)
Sports	11	20	37	82%	33	9	93%	25.7** (9,143)

PV - percentage of variance explained by the solution.
** - statistical significance at the .01 level.

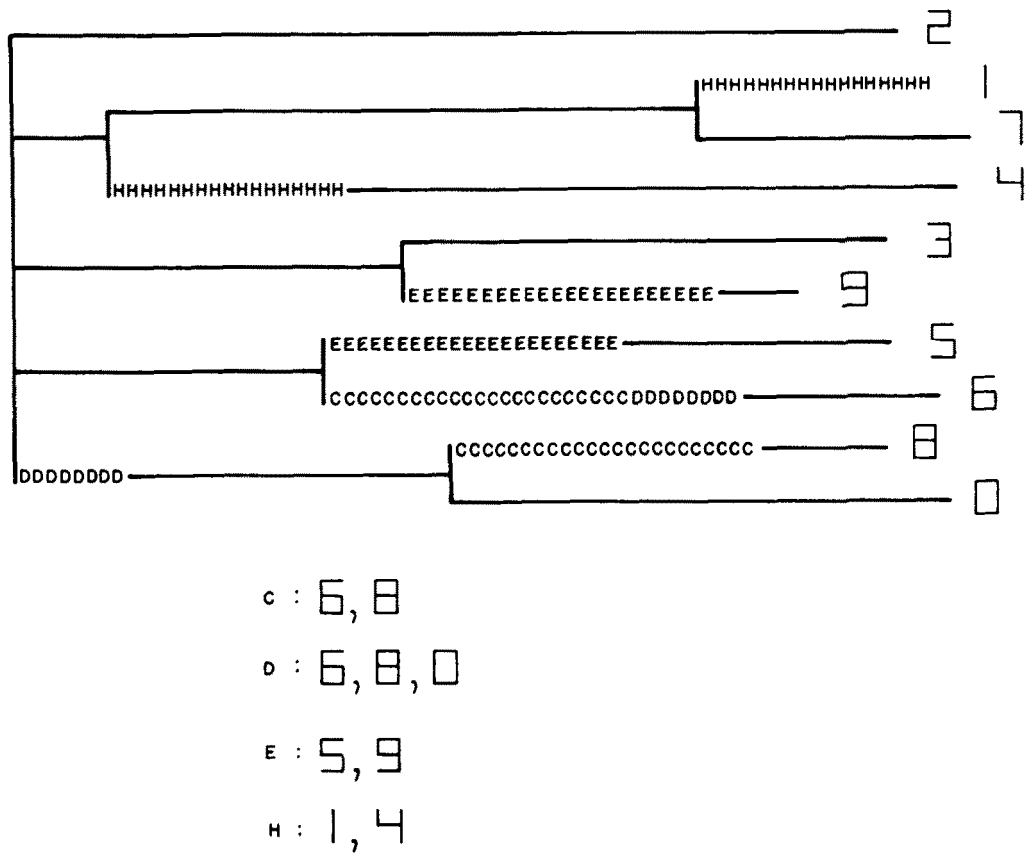


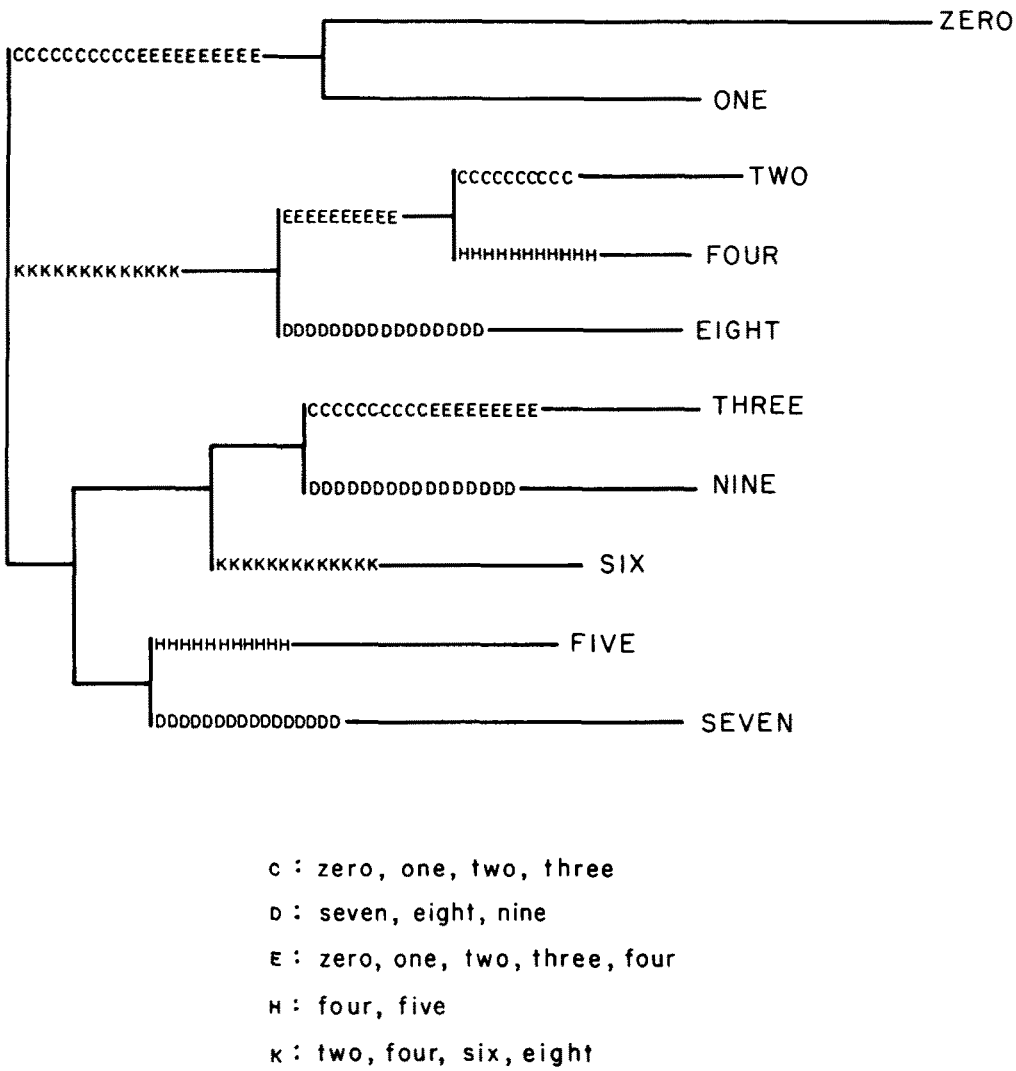
FIGURE 7
EXTREE solution for confusions between digits (Keren & Baggen, 1981). PV = 93%.

Similarity of Integers

Shepard, Kilpatric and Cunningham (1975) obtained ratings of similarity between all pairs of integers from 0 to 9, considered as abstract concepts. Shepard and Arabie (1979) presented an additive clustering (ADCLUS) solution of these ratings, pooled across subjects and several symbolic representations of the integers. Applying EXTREE to the same data accounted for 90% of the variance as compared with 62% of the variance accounted for by ADDTREE. The graphical solution, including five marked segments, is presented in Figure 8.

The basic (unmarked) tree consists of four distinct clusters: the additive and multiplicative identities (0, 1), powers of two (2, 4, 8), multiples of three (3, 9, 6), and the remaining primes (5, 7). The first four marked segments generate a different type of cluster based on proximity along the number line. Specifically, the marked features C, E, H and D capture, respectively, the following sets of consecutive integers: (0, 1, 2, 3), (0, 1, 2, 3, 4), (4, 5), and (7, 8, 9). Finally, K describes multiples of two, (2, 4, 6, 8), which could not be represented in the basic tree because 6 was clustered with the multiples of 3. With the inclusion of this feature, both the multiples and the powers of two and of three can be represented in the same structure.

These data also demonstrate how EXTREE forms higher level marked segments (C, D, E, and K) by combining pair-wise common features.



C : zero, one, two, three

D : seven, eight, nine

E : zero, one, two, three, four

H : four, five

K : two, four, six, eight

FIGURE 8
EXTREE solution for similarity of integers (Shepard et al., 1975). PV = 90%.

Proximity of Languages

Most contemporary European languages are thought to have developed from a common Indo-European proto-language by successive differentiation. A rooted tree offers a natural representation of this historical development. Due to cultural exchanges or geographical proximity, however, some languages have been influenced by others that are not necessarily genealogically related. Common forms of such influence include borrowing of terms, parallel constructions and phonological shifts. Because of the Norman invasion of England in the eleventh century, for example, many words in English derive from the Romance languages rather than from the Germanic languages to which English belongs. Similarly, Rumanian is generally classified as a Romance language on the basis of its morphology and history, but only a small part of its lexicon has Romance origins and the majority of its words have been borrowed from its geographical neighbors, especially the Slavic languages. Numerous examples of this kind have led some linguists to propose that

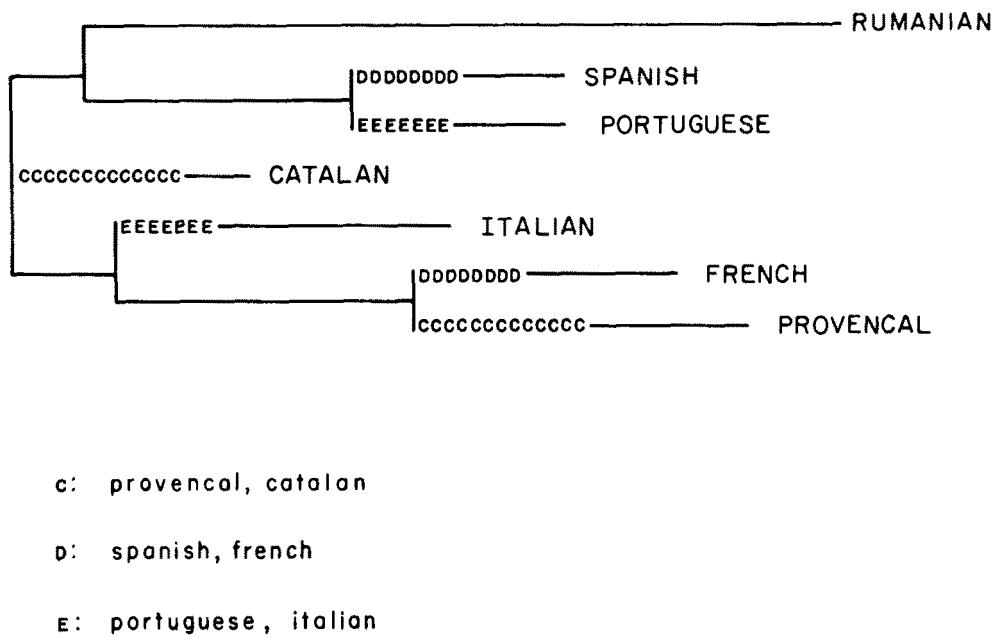


FIGURE 9
EXTREE solution for percentage of cognates between Romance languages (Tischler, 1973). PV = 94%.

language innovations spread out to geographically contiguous groups in a wavelike manner (see, e.g., Bynon, 1980).

The extended tree model provides a convenient way to represent such relations among languages, by augmenting the basic historical tree with marked segments that could represent the borrowing of terms and other cross-influences. To illustrate this application we present, in Figure 9, an extended tree solution for a proximity matrix compiled by Tischler (1973, p. 136), whose entries are the estimated percentage of cognates for pairs of Romance languages.

The basic (unmarked) tree consists of two major clusters (French, Provencal) and (Spanish, Portugese), which are joined, respectively, by Italian and Rumanian. Catalan, which is closest to Latin from which all these languages originated, appears as a separate branch near the root of the tree. The application of EXTREE to these data led to the addition of three marked features: D shared by Spanish and French, C shared by Catalan and Provencal, and E shared by Italian and Portugese. The addition of the marked segments increased the proportion of explained variance from 88% to 94% although the increase is not significant. Because the data base in this case is rather limited, the solution should be interpreted with care. The present example was introduced merely to illustrate the possibility of extending the traditional language tree to represent sources of influence that cut across the basic genealogical structure.

Proximity Between Instances of Furniture

The next two analyses are based on unpublished data collected by Mervis, Rips, Rosch, Shoben & Smith (1975), who asked subjects to rate the degree of relatedness between instances of natural categories (Rosch & Mervis, 1975). Figure 10 presents the EXTREE solution for judgments of proximity between instances of furniture. The ADDTREE solution accounted for 83% of the variance and the EXTREE solution, with eight marked segments, accounted for 90% of the variance.

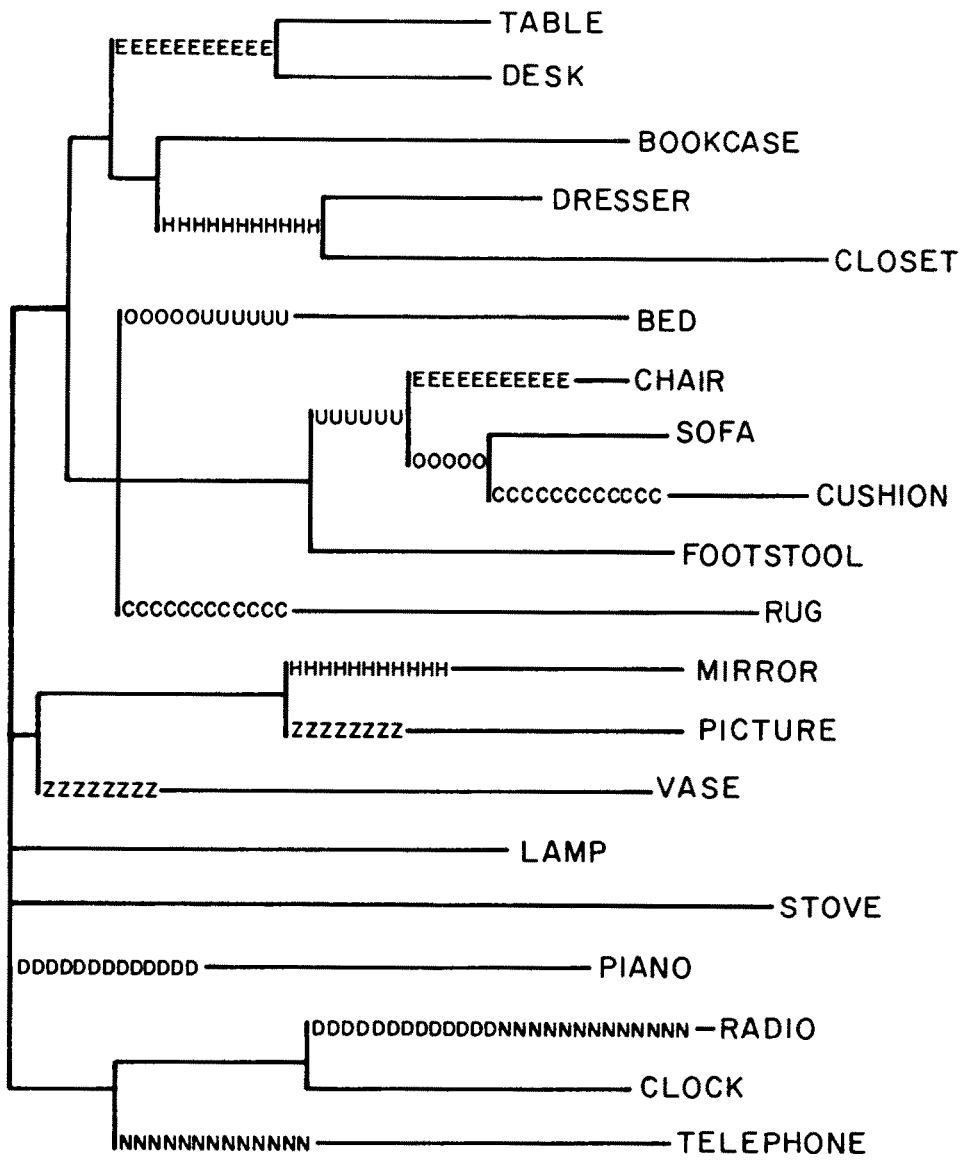


FIGURE 10
EXTREE solution for proximity between instances of furniture (Mervis et al., 1975). PV = 90%.

Both the unmarked and the marked segments seem readily interpretable. In general, the marked features tended to reflect natural associations more than taxonomic considerations. For example, the primary tree groups *Table* with *Desk*, and *Chair* with *Sofa* and *Cushion*. The secondary classification (induced by the marked segments) combined *Table* and *Desk* with *Chair* (E) and *Cushion* with *Rug* (C). Similarly, the primary classification combines *Mirror* with *Picture*, which have a similar shape and are both made of glass, whereas the secondary classification combines *Picture* with *Vase* (C), which are not perceptually similar but are both used for decoration.

Similarity of Sports

Mervis et al. (1975) also obtained ratings of relatedness between twenty sports. The EXTREE solution for these data is presented in Figure 11. The ADDTREE solution accounted for 82% of the variance as compared to 93% for EXTREE. The (unmarked) tree yielded two major clusters: spectator ball games (e.g., football, basketball), and outdoor and water sports (e.g., hiking, canoeing, skiing). In addition the data yielded a few binary clusters, such as tennis and ping pong, billiards and checkers, and boxing and fencing.

The marked features generated by the EXTREE solution nicely complement the basic classification by clustering related activities that cannot be expressed in the basic tree. For example, volleyball, tennis and ping pong (X) all involving passing a ball over a net. Another higher order feature (K) groups all the water sports: canoeing, surfing, skindiving and swimming. The other marked segments reflect natural associates: boxing is joined both with jump rope (H), which boxers use in training, and with football (E), which is also a contact sport.

Discussion

The extended tree has been developed in order to represent graphically both nested and overlapping clusters that emerge from the analysis of proximity data. The novel aspect of the proposed representation is the use of marked segments to describe nonnested features or overlapping clusters while preserving the graphical form of a tree where any two nodes are joined by a unique path. Although the simple additive tree and the extended tree are based on the distinctive features model, both common and distinctive features appear in the representation. This is a consequence of the fact that a given segment represents features that are distinctive in some comparisons and common in others.

It is instructive to compare the extended tree to other feature models of proximity data. Both the hierarchical clustering model and the additive tree assume a nested feature structure, or a nonoverlapping family of clusters. This restriction does not apply to the extended tree or to the additive clustering model of Shepard and Arabie (1979; see also Arabie & Carroll, 1980). In this model the distance between objects is inversely related to the measure of their common features, or equivalently to the sum of the weights of the clusters to which both objects belong.

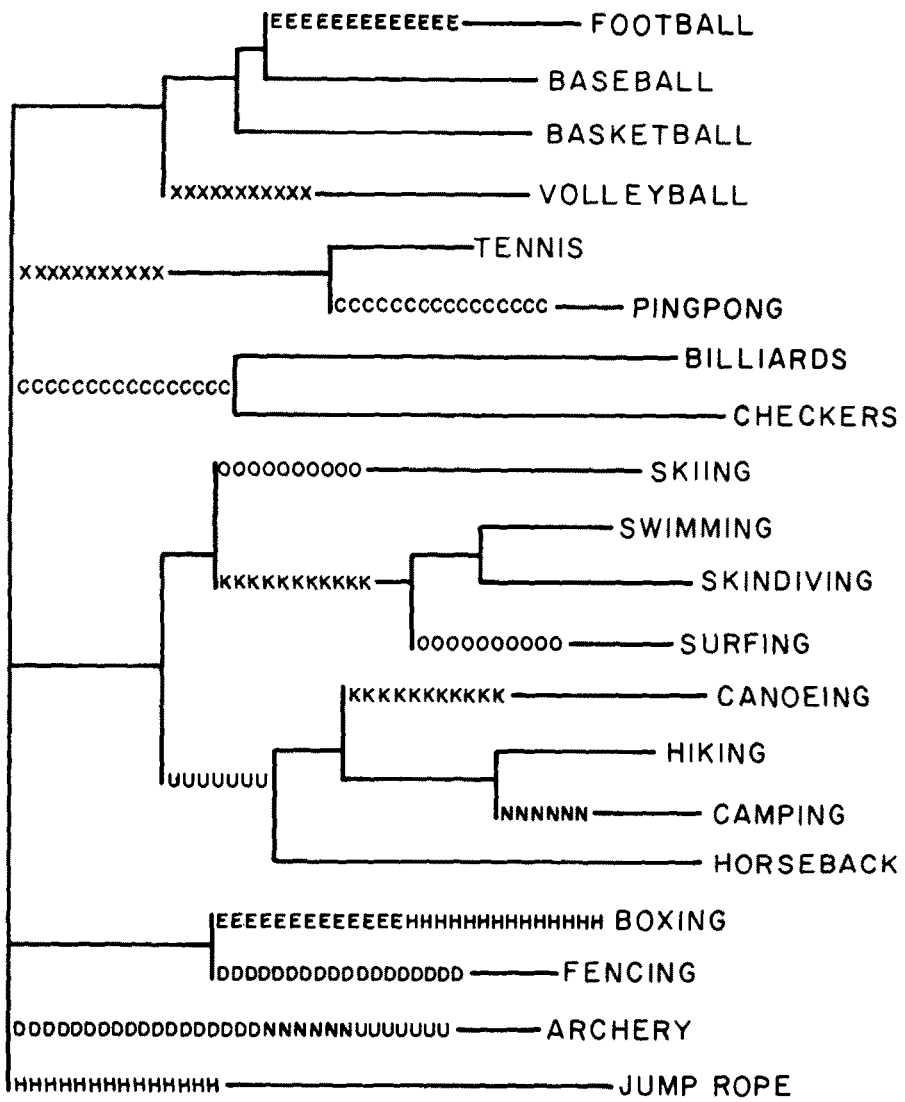
Formally, a dissimilarity function d satisfies the common features model if there exists an additive measure g defined on the subsets of the feature set S and a constant K, such that for all x, y in s

$$d(x, y) = K - g(X \cap Y) = K - \sum_{V \in X \cap Y} g(V). \tag{6}$$

In terms of the object-feature matrix $\mathbf{M} = (m_{ij})$, the common features model is expressed by

$$d(x, y) = K - \sum_{i=1}^{k} g_i m_{ix} m_{iy}, \tag{7}$$

where g_i is the measure of feature i, or equivalently the weight of cluster i.



c : pingpong, billiards, checkers

d : fencing, archery

e : football, boxing

h : boxing, jump rope

k : swimming, skindiving, surfing, canoeing

n : camping, archery

o : skiing, surfing

u : canoeing, hiking, camping, horseback, archery

x : volleyball, tennis, pingpong

FIGURE 11
EXTREE solution for proximity between sports (Mervis et al., 1975). PV = 93%.

The additive clustering model and the hierarchical clustering model are naturally expressed in terms of common features, whereas both the additive and the extended tree are expressed in terms of distinctive features. Table 3 summarizes the relations among the four models in terms of two facets: the feature structure (nested vs. nonnested) and the nature of the distance rule (common features vs. distinctive features). However, the hierarchical clustering model can be expressed in terms of either common or distinctive features. This follows from the fact that in all ultrametric tree all objects have the same measure (i.e., distance from the root), hence the measure of the common and of the distinctive features are linearly related (i.e., $f(X - Y) + f(Y - X) = K - 2g(X \cap Y)$). Note that all four models in Table 3 are special cases of the contrast model (Tversky, 1977).

We next discuss the relation between the distinctives features model (or equivalently, the extended tree), and the common features model (or equivalently, additive clustering). On the face of it, the models are quite different. In particular, the distinctive features model is a metric: it satisfies minimality ($d(x, x) = 0$ for all x in s) and the triangle inequality ($d(x, y) + d(y, z) \geq d(x, z)$). In contrast, the common features model generally does not obey minimality, and it need not satisfy the triangle inequality. Furthermore, given a particular object-feature matrix, the two models give rise to different dissimilarity orderings (Sattath & Tversky, 1985). To illustrate, consider a set of faces with a common frame F (including eyes, nose, and mouth) and three additive features: beard (X), glasses (Y), and mustache, (Z). In the common features model

$$d(FXY, FXZ) = K - g(F) - g(X) < K - g(F) = d(FZ, F),$$

but in the distinctive features model

$$d(FXY, FXZ) = f(Y) + f(Z) > f(Z) = d(FZ, F).$$

However, Sattath & Tversky (1985) showed that (excluding self-dissimilarities) the two models can be mapped into each other. That is, if there exists an additive measure f satisfying the distinctive features model, (1), relative to some feature matrix $\mathbf{M}$, then there exists an additive measure g satisfying (up to an additive constant) the common feature model relative to a different feature matrix $\mathbf{M}'$ and vice versa. The two measures f and g have the same numbers of free parameters, but they are not linearly related nor do they have the same support. (The support of a measure is the set of elements for which it is nonzero). Thus, data that have a common-features solution also have a distinctive-features solution, but the solutions are generally different because they need not include the same clusters.

From the perspective of scaling, then, EXTREE differs from the various ADCLUS programs in three respects. First, it is based on a different notion of distance that gives rise to different features or clusters. Second, it is based on a different algorithm for constructing clusters. Third, it yields a tree-like graphical representation of the proposed solution. The choice between an EXTREE representation and one obtained by an ADCLUS procedure, then, depends on the interpretability of the induced clusters, as well as on the usefulness of the display. The solutions cannot be compared in terms of goodness of fit, however, because the two representations necessarily account equally well for the observed dissimilarities.

The empirical examples discussed in the previous sections suggest contexts in which EXTREE could be usefully applied. EXTREE seems appropriate for the representation of nominal factorial structures, where neither a regular tree nor multidimensional scaling is very satisfactory. For example, the extended tree of Figure 3, with only three marked segments, provides an excellent amount of a 3×3 factorial design that requires a four-

TABLE 3

Classification of Additive Feature Models

		Feature Structure: Nested	Feature Structure: Non-nested
Distance Rule: Common Features		Hierarchical Clustering (Sokal & Sneath, 1963)	Additive Clustering (Shepard & Arabie, 1979)
Distance Rule: Distinctive Features		Additive Tree (Sattath & Tversky, 1977)	Extended Tree (Corter & Tversky, 1986)

dimensional solution. EXTREE is also well suited for describing perturbations of an inherently hierarchical structure. The Romance languages example (Figure 9), where the dominant taxonomic structure is perturbed by geograpical influences, is a case in point. Many classification systems are primarily but not perfectly hierarchical. The books in a college library, for example, are organized primarily in a hierarchical fashion according to disciplines and subdisciplines (e.g., social science, psychology, psychometrics) with some overlapping categories such as quantitative methods that include both psychometrics and econometrics.

The clusters fit by EXTREE are divided into the primary clusters, described by the unmarked segments, that constitutes the basic tree and the secondary clusters, described by the marked segments. In some applications the primary and the secondary clusters are conceptually distinct. For example, the primary clusters in Figure 8 refer to structural properties of the integers, and the secondary clusters describe the position of the integers along the number line. In other applications (e.g., the confusability of digits or the proximity of sports) there is no clear separation between the primary and the secondary clusters. Indeed, the user may sometimes wish to interchange some primary and secondary clusters in order to obtain a simpler or more interpretable configuration. In Figure 11, for example, *canoeing* is part of the branch of camping sports, and it is connected to the other water sports by a marked segment (K). However, we can interchange the primary and secondary clusters so that *canoeing* will be clustered with the other water sports and includes a marked segment shared with the other camping sports. The two graphs provide about the same fit to the data, hence one is free to choose the more readable or interpretable representation. The EXTREE program, therefore, should not be treated as a rigid canonical representation but rather as an interactive device for representing proximity data in a convenient graphical form.

References

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45*, 211–235.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, & P.

Tautu (Eds.), *Mathematics in the archeological and historical sciences*. Edinburgh: Edinburgh University Press.

Bynon, T. (1980). *Historical linguistics*. Cambridge: Cambridge University Press.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association, 68*, 361–368.

Corter, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behavior Research Methods and Instrumentation, 14*(3), 353–354.

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

Freedman, D., & Lane, D. (1983a). Significance testing in a nonstochastic setting. In P. Bickel, K. Doksum, & J. L. Hodges, Jr. (Eds.), *Lehmann Festschrift*. Belmont, CA: Wadsworth.

Freedman, D., & Lane, D. (1983b). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics, 1*, 292–298.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. London: Wiley.

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

Keren, G., & Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics, 29*, 234–246.

Kruskal, J. B., Young, F. W., & Seery, J. B. (1977). How to use KYST-2A, a very flexible program to do multidimensional scaling and unfolding. Unpublished manuscript, AT&T Bell laboratories.

Mervis, C. B., Rips, L. J., Rosch, E., Shoben, E. J., & Smith, E. E. (1975). Unpublished data.

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

Restle, F. (1959). A metric and an ordering on sets. *Psychometrika, 24*, 207–220.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319–345.

Sattath, S., & Tversky, A. (1985). On the relation between common and distinctive feature models. Unpublished manuscript, Stanford University.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87–123.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology, 7*, 82–138.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.

Tischler, J. (1973). *Glottochronologie und lexikostatistik*. Innsbruck: Innsbrucker Beitrage zur Sprachwissenschaft.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review, 86*, 542–573.

Manuscript received 7/16/84
Final version received 1/28/86