Psychological Review

VOLUME 86 NUMBER 2 MARCH 1979

Additive Clustering: Representation of Similarities as Combinations of Discrete Overlapping Properties

Roger N. Shepard
Stanford University

Phipps Arabie
University of Minnesota

For the discovery and representation of structures in similarity data, we give the first full presentation of one alternative to the inherently continuous spatial models of multidimensional scaling and factor analysis and to the strictly hierarchical models of discrete clustering. We assume that the effective similarity of any two objects is a simple additive function of underlying weights associated with whatever properties are shared by both objects. Based on this model, we describe a method of additive clustering, ADCLUS, that is capable of estimating (a) which subsets of a given set of objects correspond to positively weighted properties and (b) the numerical values of those weights. The method subsumes hierarchical clustering as a special case and can be regarded as a discrete analogue of principal components analysis. Finally, we present illustrative applications to several diverse types of data, and we discuss the bearing of the results on current theoretical issues concerning distinctive features and the adequacy of purely hierarchical models.

In this article we describe and illustrate an approach to the problem of representing the structure underlying a set of measures of the similarities between the objects in all $n(n-1)/2$ pairs of n objects. The objects may be of any specifiable sort including, for example, physically presented stimuli, verbally defined concepts, individual persons, administrative units, or even biological species. The measures of similarity may be direct subjective measures of judged similarity, affinity, substitutability, or co-occurrence; direct objective measures of frequency of overt confusion, association, or disjunctive reaction time; or derived or computed measures of correlation or overlap. Assuming that these similarity measures are either inherently symmetric or have been symmetrized by the data analyst, we can conveniently display a complete set of $n(n-1)/2$ such measures in the triangular half of an $n \times n$ matrix in which the n rows and n columns correspond to the same n objects.

Statement of the Problem: Alternative Strategies and Models for Representing Similarity Data

The principal problem with which we are concerned is that of transforming such a matrix of numbers into a different form that meets two requirements. First, the transformed representation should preserve the essential information in the original matrix in the sense that the numerical values in that matrix can, to a satisfactory degree of approximation, be reconstructed from the transformed representation. Second, the underlying pattern or structure as a whole, which was only implicit in the original matrix of numbers, should be rendered manifest in a form that is suitable for substantive interpretation and quantification.

Any procedure for effecting such a transformation necessarily depends, whether implicitly or explicitly, either on some assumptions or on a model concerning the nature of the structure underlying the data. Nonmetric multidimensional scaling, originally developed by Shepard (1962a, 1962b) and Kruskal (1964a, 1964b) and subsequently one of the most widely used procedures for this purpose, assumes that the similarity measures are monotone decreasing functions of distances between the objects in a low-dimensional Euclidean (or, more generally, Minkowskian) coordinate space. A later extension, referred to as maximum variance nondimensional scaling (Cunningham & Shepard, 1974), assumes that the measures are monotonically related to distances in a completely general, coordinate-free metric space, that is, distances satisfying only the metric axioms: positivity, symmetry, and the triangle inequality. The quite different method of hierarchical clustering (as formalized by Hartigan, 1967; Jardine, Jardine, & Sibson, 1967; Johnson, 1967; Lance & Williams, 1967a, 1967b) assumes that the measures are monotonically related to distances in an ultrametric space in which the triangle inequality is constrained to take the more restrictive form of the ultrametric inequity. Finally, Ekman's (1954, 1963) proposed factor analysis of similarity measures, rather than assuming that the measures are related to distances of any of these sorts, assumes that they are related linearly to scalar products or to cosines of angles between vectors.

These and related existing methods have individually been found capable of generating useful and often illuminating representations for diverse sets of similarity data. Moreover, when applied to the same matrix of data, these alternative methods have often yielded different representations that are complementary. That is, the different types of resulting representations seem to bring out different aspects of the underlying structure in especially convenient or revealing ways. (See, e.g., Breiger, Boorman, & Arabie, 1975; Carroll, 1976; Carroll & Wish, 1974; Kruskal, 1977b; Shepard, 1972a; Shepard & Carroll, 1966.) Even so, the set of these currently available methods does not suffice to meet all of our needs for the representation of structure in similarity data (Shepard, 1974). Specifically, we see a need for a method based on a model that is more compellingly connected with the underlying substantive process and, at the same time, that is not restricted to representations that are either continuous or hierarchical. In this respect, our approach is broadly similar to the orientation taken by Tversky (1977) in his recent axiomatic approach to similarity judgments, by Carroll and his colleagues (see Carroll, 1976; Carroll & Pruzansky, Note 1, Note 2) in their development of hybrid models that combine continuous and discrete structures, and even by Guttman (1952) in his early factor-analytic approach to the fitting of discrete structures to correlation matrices.

Consideration of the Appropriateness of Continuous, Spatial Representations

The advent of numerical and mathematical methods for representing structure in data has often preceded the full explication of the

This research was supported by National Science Foundation Grants GB-31971X and BNS-75-02806 to Roger N. Shepard, GS-2689, SOC 76-24512, and SOC 76-24394 to Scott A. Boorman and Harrison C. White, and National Institute of Mental Health Grant MH-21747 to Richard C. Atkinson.

We are indebted to S. A. Boorman, J. P. Boyd, J. D. Carroll, J. P. Cunningham, E. W. Holman, L. J. Hubert, C. L. Krumhansl, J. B. Kruskal, P. R. Levitt, E. J. Shoben, S. D. Soli, E. A. C. Thomas, A. Tversky, M. Wish, and to two reviewers for helpful comments.

The research reported in this article was begun while both authors were at Stanford University, and much of the present article was written while the second author was a resident visitor at Bell Laboratories. The competent assistance of the staff of A. W. Talcott of the Technical Library at Bell Laboratories is gratefully acknowledged.

Requests for reprints should be sent to Roger N. Shepard, Department of Psychology, Stanford University, Stanford, California 94305 or to Phipps Arabie, Department of Psychology, Elliott Hall, University of Minnesota, Minneapolis, Minnesota 55455.

processes giving rise to those data—in psychology just as in other areas of behavioral science including, for example, price theory in economics (Koopmans, 1957). In nonmetric multidimensional scaling as well as in factor analysis, the availability of computer programs encouraged some attempts to fit data without an adequate commitment to the psychological interpretation of the results. For example, in order to achieve what was regarded as a sufficiently "good" fit, too many dimensions or too few objects were sometimes used, with the consequence that a convincing interpretation of the results could not be given and, in some cases, was not even attempted.

At a more fundamental level, some authors have questioned whether the very model on which multidimensional scaling is based has a plausible interpretation in terms of the processes giving rise to interstimulus similarity (e.g., see Beals, Krantz, & Tversky, 1968; Boyd, 1972; Tversky & Krantz, 1969). It is difficult in general to specify the particular nature of the processes involved, for example, in judgments of similarity so that the data are consistent with the constraints imposed by a particular metric, such as the constraints of a low-dimensional Euclidean space. This problem is especially challenging when we have no basis for assuming a particular functional form for the monotone relation between similarity and distance.

Some efforts have, of course, been made toward filling this lacuna with plausible psychological models. In the context of stimulus generalization, Cross (1965; Cross, Note 3) advanced an ingenious and elegant interpretation of three special cases of the Minkowski metric (viz., the city-block, Euclidean, and dominance metric), which can be extended to other judgmental tasks (Arabie, Kosslyn, & Nelson, 1975). Such interpretations have implications, moreover, for the distinction frequently made between the ways that information is combined from "analyzable" or "integral" dimensions in human information processing (Attneave, 1950; Garner, 1970; Shepard, 1964). More recently, the theoretical and experimental results of Rumelhart and Abrahamson (1973) have suggested that distance models can provide a reasonable description of judgmental processes in certain types of analogical reasoning by representing the solution of an A:B: :C:X analogy in terms of the completion of a parallelogram in a Euclidean scaling solution. And, in a salutary reversal of the usual order of development of multidimensional scaling methods, Carroll's highly successful method of individual differences scaling, INDSCAL (Carroll & Chang, 1970), began as a psychological model around which a computer program was then developed.

The psychological models put forward in these and other articles provide additional justification for the use of multidimensional scaling. At the same time, however, the emphasis on the finer details in such models brings into still sharper focus the question of whether such scaling techniques are applicable to all psychological domains and processes. Arabie and Boorman (1973), for instance, noted that the appropriateness of embedding an inherently discrete system in a continuous space had received little attention. It seems to us that the essentially continuous spatial representations obtained by multidimensional scaling or factor analysis are eminently reasonable for representing objects such as colors varying continuously in brightness, hue, and saturation (Ekman, 1954; Helm, 1964; Shepard, 1962b; Shepard & Carroll, 1966; Wish & Carroll, 1974) or such as closed curves or polygons varying continuously in size, shape, and orientation (Attneave, 1950; Shepard & Cermak, 1973; Shepard & Chipman, 1970). Such continuous spatial representations, however, may not fully and explicitly reveal the discrete or categorical nature of consonant phonemes (Jakobson, Fant, & Halle, 1963; Klatt, 1968; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), of kin and other category-specific terms (Boorman & Olivier, 1973; Romney & D'Andrade, 1964; Shoben, 1976), of social structures (Arabie & Boorman, 1973; Boorman & White, 1976; Breiger et al., 1975), or even, possibly, of continuously variable stimuli that are nevertheless psychologically "analyzable" (Shepard, 1964) into "separable" (Garner, 1970) dimensions or properties. (Also see Torgerson, 1958, p. 292.)

Consideration of the Appropriateness of Discrete, Hierarchical Representations

As an alternative to the spatial solutions from multidimensional scaling, clustering provides an explicitly discrete, categorical representation of structure. For most psychologists, the term clustering presently implies *hierarchical* clustering (Hartigan, 1967; Jardine et al., 1967; Johnson, 1967; Lance & Williams, 1967a, 1967b). Although the single-link and complete-link variants of hierarchical clustering existed well before publication of the articles cited (see Sibson, 1972, p. 319, for earlier publications), it was not until Johnson's (1967) introduction of these methods to the psychological community as diameter and connectedness methods that these methods came into wide use by American psychologists. (See Hartigan, 1975; Hubert, 1974b; or Johnson, 1967, for formal descriptions of the two methods.) Informally speaking, the method variously called the diameter, the compactness, the maximum, the furthest neighbor and, now most widely, the *complete-link* method seeks to minimize the largest between-object distance within each cluster. By contrast, the method called the connectedness, the minimum, the nearest neighbor and, now, the *single-link* method seeks to minimize the largest between-object link needed to achieve a connected path between all objects within each cluster. In common with other hierarchical schemes, the result of either method is a tree structure of strictly nested clusters. Because the clusters are hierarchically nested, they do not overlap at any single level of the hierarchy. That is, if A and B are any two distinct clusters, then the following possibilities are mutually exclusive and exhaustive: $A \subset B$; $B \subset A$; $A \cap B = \emptyset$.

From the standpoint of data analysis, such methods of hierarchical clustering uniformly yield an excessive number of clusters with respect both to substantive interpretation and statistical reduction of data, yet the final results are often presented in such a way as to conceal the *embarras de richesse*. To reduce the number of clusters to a suitably small set in an objective way, one must use one of several further procedures, one of which is to select only those clusters that are significant according to statistical assumptions appropriate to the clustering method used (Ling, 1972; Ling, Note 4). Another approach is to select from the hierarchical tree the level(s) at which the corresponding partition correlates most highly with the data (Baker & Hubert, 1975, 1976; Hubert, 1973; Rosenberg & Jones, 1972). In effect, the investigator is supplied with a tool for taking a slice or cross-section from the tree. Still another approach, appropriate when the data analyst is seeking ab initio only a single partition instead of a tree, is to use a clustering method that just produces a single partition (e.g., Diday, 1974; MacQueen, 1967). Such a partition constitutes a nonhierarchical and nonoverlapping representation of the data.

Often ignoring these and other possibilities, psychologists have routinely used the single- and complete-link methods, independent of substantive considerations. Moreover, as in the case of nonmetric multidimensional scaling, extensive use of hierarchical clustering preceded any serious examination of the method for its plausibility as a psychological model. One exception was Levelt (1967, 1970), who presented some original arguments as to why the single-link method is appropriate to the analysis of certain types of data, whereas the complete-link method is to be preferred in other cases. More recently, Holman (Note 5) provided an illuminating discussion of hierarchical structures as testable psychological models. In a specific substantive context, Friendly (1977) drew parallels between the graph-theoretic aspects of hierarchical clustering and recent results from the study of organization in human memory. Two significant articles that have emphasized the mathematical properties of the methods in relation to the typical uses of psychologists are Fisher and Van Ness (1971; see also Van Ness, 1973) and Holman (1972). Nonetheless, in all of these articles, the generality of the psychological postulates advocated is limited because tree structures are themselves quite restrictive.

The popularity of hierarchical clustering among experimental psychologists is probably a result, in part, of the early emphasis on hierarchical structure by cognitive psychologists. (See e.g., Levelt, 1970; Miller, Galanter, & Pribram, 1960; or for a view from a wider perspective, see the contrasting case offered by LaBerge, 1976.) In recent years, however, increased acquaintance with the relevant substantive areas has suggested that the requirement of having the clusters hierarchically nested seems to be unduly limiting. With regard to the perception of consonant phonemes, which we shall consider later, the force of such an assumption would be to imply that once all of the voiced consonants were grouped into one cluster (disjoint from the cluster of voiceless consonants), it would no longer be possible to group either all the stops or all the fricatives into one cluster. Since there are both voiced and voiceless stops and fricatives, the two groupings would have to overlap, in violation of a hierarchical structure. Similarly, with regard to the conception of semantic relationships, hierarchical constraints require that once "aunt" is grouped with "uncle" and "niece" with "nephew" on the basis of generation, it is no longer possible to group aunt with niece or uncle with nephew on the basis of sex (cf. Wexler & Romney, 1972). Finally, in the analysis of social structure, the hierarchical assumption becomes extremely limiting in that it excludes (to give only a few examples) corporate directorship interlock data (Levine, 1972), cliques in a social network (Arabie, 1977; Coleman, 1963), overlapping lobbies or political factions (Panning, Note 6), and even many formal organizational office structures.

Generally, the discrete psychological properties of objects overlap in arbitrary ways. Consider the following examples, taken from a study by Shepard, Kilpatric, and Cunningham (1975), of the judged similarities between one-digit numbers. When the numbers are judged with respect to the shapes of their Arabic numerals, the subset containing curved lines (2, 3, 5, 6, 8, 9) overlaps only partially with the subset containing an enclosed space (4, 6, 8, 9). When the numbers are judged instead as abstract concepts, the subset of odd numbers (1, 3, 5, 7, 9) overlaps only partially with the subset of multiples of three (3, 6, 9) or with any division between large and small numbers. Accordingly, any method intended to represent, as an explicit cluster, every subset of the objects having a shared property must provide for the emergence of clusters that cut across each other and that are, therefore, not purely disjoint or nested.

The Problem of Nonhierarchical Clustering

Although methods based on the assumption of overlapping and nonhierarchical clusters have been available for decades (e.g., Luce, 1950), only in recent years have the graph-theoretic and statistical approaches to clustering (e.g., Hartigan, 1975; Hubert, 1974b) given serious consideration to the possibilities of overlapping clustering in data analysis. Motivated by the limitations of the existing methods for the representation of structure in similarity data, we have been exploring a new method of nonhierarchical clustering. This method permits the explicit representation of discrete structure without entailing the undesirable limitation that this structure be strictly hierarchical. Moreover, we shall argue that the model on which this new method is based constitutes a rather more direct and plausible representation of the processes employed by the subjects in generating similarity data, at least of some types.

Unfortunately, the problem of finding the generally nonhierarchical clustering of a set of n objects that best accounts for the associated set of $n(n-1)/2$ similarity measures is beset by major difficulties both theoretical and practical. Ideally, we should have liked to base our method on a suitable, explicitly defined measure of the goodness of any proposed clustering (cf. Hartigan, 1967; Hubert, 1974a; Carroll & Pruzansky, Note 1, Note 2) —such as, for example, the fraction of variance of the original similarity data accounted for by the clustering. In principle, we could have then searched through the finite set of all distinct clusterings of the n objects and selected, as our final representation, one that achieves both suitable parsimony and goodness of fit according to our explicitly defined measure. In practice, however, such an approach is entirely beyond existing technology owing to the astronomical number of possible clusterings that would have to be considered for an n of any reasonable size. There are, in fact, $2^{(2^n-1)} - 1$ distinct ways in which the n objects might be grouped into possibly overlapping subsets.

To proceed, we seemed to need some powerful (though perhaps heuristic) device to reduce, to a manageably small subset of likely candidates, the set of possible clusterings that are to be examined. The situation here is quite analogous to that in single- and complete-link hierarchical clustering, in which the necessary reduction is achieved by considering only those relatively few clusterings that are strictly hierarchical. In our case, we selected the subset of candidate clusters from the maximal complete subgraphs and, for this reason, our approach to nonhierarchical clustering is more closely related to complete-link than to single-link methods of hierarchical clustering, as explained later.

Single Versus Complete Link as a Basis for Generalizing to Overlapping Clusters

Just as single- and complete-link methods have long represented two distinct orientations to hierarchical clustering (Hubert & Schultz, 1975; Lance & Williams, 1967a), the two methods also serve to orient different approaches to nonhierarchical clustering. Since there is an ongoing controversy as to which of the approaches is better, we briefly consider the comparative advantages and disadvantages of each.

The strongest advocacy of the single-link approach has come from Jardine and Sibson (1971), who pointed out that unlike the complete-link method, the single-link approach possesses the mathematical properties of being continuous and well defined. Also, as a point in favor of the single-link approach, Hartigan (1975) has cited the fact, noted by Gower and Ross (1969), that the partitioning it produces is isomorphic with that generated by the minimum spanning tree algorithms (Kruskal, 1956; Prim, 1957). Friendly (1977, p. 219) has further observed that the minimum spanning tree can be psychologically interpreted as a memory retrieval path of minimum overall cost or effort (cf. Boorman & Arabie, 1972, pp. 227–229 for a parallel substantive argument). In addition, Hartigan (1977a, 1977b) has pointed out that, in the restrictive case in which the objects are drawn from a one-dimensional space, the single-link method asymptotically splits the set of objects at the location of minimum density in the underlying one-dimensional space. Finally, single-link methods do not encounter the difficulty, which does arise for complete-link methods, that ties must be arbitrarily broken. Thus, as Hubert (1973) and Peay (1975) have observed, the complete-link method leads not to a unique tree structure or dendrogram but to a *class* of such structures, although the method can be modified in various ways to avoid the lack of uniqueness. (See Hubert, 1973; or the first article to describe complete-link clustering, Sørenson, 1948.)

Consistent with their espousal of the single-link method, Jardine and Sibson (1968a, 1968b, 1971) generalized that method to include "k-partitions," which allow a maximum of $k - 1$ stimuli to belong to overlapping clusters and thereby inhibit *chaining*. (See following discussion.) There have been few substantive applications of that method, perhaps owing to inefficiencies in the original implementation. Recent improvements in available programs (Cole & Wishart, 1970; Rohlf, 1975) may enhance the usefulness of Jardine–Sibson B_k clustering. However, Friendly (1977, pp. 214–215) has argued that this method shares with its ancestor, the single-link method, certain drawbacks to which we now turn.

The majority of psychologists who have been using clustering methods have in fact preferred complete-link representations. The main reason is probably that single-link clusterings often exhibit an excessive tendency toward a kind of chaining in which, at many levels of the resulting dendrogram, the only change in going to the next level is that a previously isolated object is attached to an ever more far-flung cluster. The completed dendrogram typically has only a relatively small number of such serpentine clusters. Partly because of this tendency, Lance and Williams (1967a) went so far as to denigrate the method as "obsolete" (p. 377). Although

Hartigan (1975) has countered that "sometimes clusters *are* far-flung sausage shapes with high densities of objects within each cluster" (p. 200), applications to psychological data have seldom yielded preferable results from the single-link approach. Rather, the substantively more satisfactory representations have generally been those obtained by the complete-link method, which is not susceptible to chaining and which tends to cluster the objects into several subsets having more nearly equal numbers of members. Still, we must be alert to the possibility that this ingratiating appearance is illusory, perhaps, as Hartigan (1977b) suggests "like a fortune teller who predicts only good news out of ignorance and a desire to please" (p. 445).

Moreover, the axiomatic basis that Jardine and Sibson offered for the single-link approach has found few supporters and has actually helped to engender the opposing "Australian school" (Boulton & Wallace, 1975; Lance & Williams, 1967a, 1967b; Williams, Lance, Dale, & Clifford, 1971). More importantly, despite Hartigan's (1977a, 1977b) suggestion that the results may to some extent have been an artifact of using rank correlation, we find support for the complete-link approach in Monte Carlo comparisons of the two methods (Baker, 1974; Baker & Hubert, 1975, 1976; Hubert, 1974a; Hubert & Baker, 1977a; Kuiper & Fisher, 1975). These studies have indicated that the results of the complete-link method generally correspond more closely with the matrices of input data. Indeed, using artificially constructed sets of data, Kuiper and Fisher (1975) found the single-link method superior to the complete-link method *only* in the situation described by Hartigan, and Baker (1974) found that even in that situation (as well as for the other conditions tested), the complete-link method was preferable in that it was less sensitive to noise added to the data. Finally, even if the preceding evidence were not sufficient, we would have an additional, compelling reason for our decision. The basic additive model that we are about to describe suggests in a very fundamental way that objects sharing a common property generally correspond to complete subgraphs—the graph-theoretic characterization of clusters from the complete-link method. Moreover, our generalization does not inherit what we see as a principal liability of complete link, namely, the need for arbitrary breaking of ties.

The Additive Model

The stated goal of representing data as overlapping subsets (or clusters—we use the two terms interchangeably) requires an explicit model if we are to evaluate the goodness of fit of such a representation to any particular set of data. Moreover, it is important to distinguish between the model and the various possibilities of its implementation via computer programs. We begin with a description of the model and then proceed with the details of the computational algorithm that we have devised as a first attempt for fitting the model. Although the following discussion is in terms of *similarities* between objects, the model can be fitted to any of the types of data mentioned at the beginning of this article, all of which fall under the rubric of "proximities" (Coombs, 1964; Shepard, 1972b). Although, as in some other formal discussions of interstimulus relationships (cf. Goodman, 1972; Tversky, 1977), the model seems most naturally formulated in terms of similarities, it can also be made to apply to dissimilarities.

In the model, each of the to-be-recovered, overlapping subsets of objects or stimuli corresponds to a discrete property shared by all and only those objects within that subset. It is therefore natural to assume that each such property contributes a fixed increment to the similarity between any two objects sharing that property—independently of the contributions of any and all other properties. We are thus led to the simple additive model

$$\hat{s}_{ij} = \sum_{k=1}^{m} w_k p_{ik} p_{jk}, \qquad (1)$$

where $\hat{s}_{ij}$ is the theoretically reconstructed similarity between objects i and j, w_k is a nonnegative weight representing the psychological salience of the property corresponding to subset k, and

$$p_{ik} = \begin{cases} 1, & \text{if object } i \text{ has property } k, \\ 0, & \text{otherwise.} \end{cases}$$

Notice that the product $p_{ik}p_{jk}$ is unity if and only if both objects i and j belong to subset k, and it becomes zero as soon as either stimulus falls outside that subset. Thus, $\hat{s}_{ij}$ is simply the sum of the weights, w_k, associated with just those subsets to which both objects belong.

Equation 1 can be reformulated in matrix notation as

$$\hat{\mathbf{S}} = \mathbf{PWP}', \tag{2}$$

where $\hat{\mathbf{S}}$ is an $n \times n$ symmetric matrix of reconstructed similarities $\hat{s}_{ij}$, $\mathbf{W}$ is an $m \times m$ diagonal matrix with the weights w_k ($k = 1, \ldots, m$) in the principal diagonal (and zeros elsewhere), and $\mathbf{P}$ is the $n \times m$ rectangular matrix of binary values p_{ik}. Here $\mathbf{P}'$ is the $m \times n$ matrix transpose of the matrix $\mathbf{P}$.

In this model, the entries $\hat{s}_{ij}$ in the computed matrix $\hat{\mathbf{S}}$ are to be fitted to the entries s_{ij} in the empirically obtained matrix $\mathbf{S}$ directly, rather than via a monotonic transformation as allowed in the case of nonmetric multidimensional scaling. By virtue of the metric nature of this approach, it becomes natural to adopt, as our measure of goodness of fit, the fraction of the total variance of the empirical measures, s_{ij} (ordinarily for $i \neq j$), that is accounted for by the theoretically reconstructed values, $\hat{s}_{ij}$. Specifically, if we let v denote the fraction of variance accounted for, then

$$v = 1 - \frac{\sum_{i>j} (s_{ij} - \hat{s}_{ij})^2}{\sum_{i>j} (s_{ij} - \bar{s})^2}, \tag{3}$$

where the summations are over all $n(n-1)/2$ unordered pairs of distinct objects i and j (with no missing data) and where $\bar{s}$ is the mean of all $n(n-1)/2$ observed similarity measures s_{ij}.

Given a matrix, $\mathbf{S}$, of empirical similarity data, s_{ij}, the problem becomes that of finding a clustering consisting of a suitably small number of clusters or subsets, together with their weights w_k, such that the theoretical similarity values, $\hat{s}_{ij}$, computed by means of Equations 1 or 2 account for a suitably large fraction of the variance of the empirical similarity measures, s_{ij}, as defined by Equation 3. In other words, given the $n \times n$ matrix, $\mathbf{S}$, we need to find a suitable $n \times m$ binary "property" matrix $\mathbf{P}$ and an associated $m \times m$ diagonal weight matrix $\mathbf{W}$ such that the variance accounted for is as large as possible for the m clusters. Our specification of variance accounted for here assumes that one of the m subsets is the complete set of objects; otherwise, an additive constant for the right side of Equation 3 also has to be fitted (Theil, 1971, p. 164; Hubert, Note 7). The additive constant is equal to the weight that would be computed for the complete set if included as a nonredundant $(m + 1)$st subset.

If it were not for the restriction that the entries, p_{ik}, of the property matrix $\mathbf{P}$ be binary, this model would be essentially identical to the standard factor-analytic model that Ekman (1954, 1963) proposed for the representation of structure in similarity data. Straightforward application of principal components analysis to the matrix of empirical similarity measures, $\mathbf{S}$, would then yield estimates of $\mathbf{P}$ (as a matrix of eigenvectors) and of $\mathbf{W}$ (as the associated diagonal matrix of eigenvalues). However, as is characteristic of factor-analytic solutions, the resulting solution would not be unique. Orthogonal transformations (corresponding to rotation of the axes in factor analysis) would still be permissible, since they would leave $\hat{\mathbf{S}}$ and, hence, the fit to the data unchanged. But such transformations would amount to continuous shifts in the compositions of the subsets or, more accurately, in the degrees to which each object partakes of each property.

By requiring that the possession of properties—like the membership in subsets—be all or none, we avoid the continuous rotational arbitrariness inherent in the standard factor-analytic model. But, at the same time, we replace a continuous numerical problem for which methods of solution (viz., computation of eigenvalues and vectors) are readily available with a discrete combinatorial problem for which methods, if they are to be feasible, seem to require some strategy for selecting only those subsets that are reasonable candidates for consideration. Fortunately, the model itself suggests such a strategy.

Restriction to Elevated Subsets as an "Initial Configuration"

As with iterative methods for fitting other complex models, the availability of a suitable initial configuration can be advantageous. The subsets could, of course, be specified a priori from substantive hypotheses (as we shall demonstrate in one application later), leaving only **W**, the weights matrix, in Equation 3 to be fitted by straightforward multiple linear regression. (For a related approach to assessing the psychological importance of features as subsets of stimuli, see Hubert & Baker, 1977b; Wang & Bilger, 1973.) However, in the more general case of exploratory data analysis, we would like the data themselves to suggest *both* the subsets (**P**) and the weights (**W**). The present strategy follows from substantive underpinnings of the model and is essential to the *implementation* described later, but it is not an integral part of the *model* in Equation 3. We introduce our strategy for an initial configuration of subsets immediately after the model because we hope that such an exposition will help clarify the substantive rationale of the model.

According to our model, if a particular subset, k, includes just the n_k of the n objects that share some salient property, then just the $n_k(n_k - 1)/2$ similarities between those n_k objects will contain the fixed additive value, w_k, representing the salience of that particular property. Conversely, then, if we find any subset, k, for which all $n_k(n_k - 1)/2$ similarity measures between the n_k objects are elevated, it is reasonable to suppose that this elevation may be a consequence of a salient property that is possessed by all of the objects in that subset. We say "may be" rather than "is" here because the elevation of all of the similarities in a subset might arise from some mixture of three quite different kinds of sources, only one of which is the unique possession of a single property by all and only the objects in that subset. Two other possible sources are a suitable combination of overlapping properties confined within that subset and a suitable combination of overlapping properties extending beyond that subset. We turn now to a consideration of each of these alternative sources and its implications.

First, in accordance with the stated supposition, it may indeed be that just the objects in the elevated subset in question possess a particular property. One example of this possibility is illustrated in Figure 1A, in which the subset consisting of the three objects a, b, and c is elevated above the set of all six objects, a–f, by virtue of a single property with an associated weight $w_1 = 2$. In this diagram (as in the other two in Figure 1), the similarity between any two objects is represented proportionately by the number of straight parallel lines in the bundle connecting those two objects in the diagram. Thus, of the three lines connecting any two objects in the subset [a,b,c], one is attributable to the membership of both objects in the total set [a,b,c,d,e,f] with weight $w_2 = 1$, and the remaining two are attributable to the membership of those same two objects in the subset [a,b,c] with weight $w_1 = 2$. It is the presence of the additional similarity (corresponding to the two extra lines) between all pairs of objects in the subset [a,b,c] that causes that subset to be elevated above the total set (in which many pairs are connected by no more than a single line).

Second, however, exactly the same pattern of similarities and hence the same elevation of the subset [a,b,c] could, according to the additive model, arise without the three objects a, b, and c having any one unique property in common. As is illustrated in Figure 1B, this would happen if each of the three pairs of objects in the subset [a,b,c] shared a *different* property, but with exactly the same weight ($w_1 = w_2 = w_3 = 2$). Indeed, for any subset, A, with which we have associated a unique property with weight w_A, we could use the additive model to generate exactly the same predicted similarities in terms of a larger number of properties as follows: (a) We partition Subset A into any two mutually exclusive and exhaustive Subsets B and C, we associate with each of these new subsets a new property with weight w_B ($= w_C$) $\leqslant w_A$, and we replace the weight for the property corresponding to the

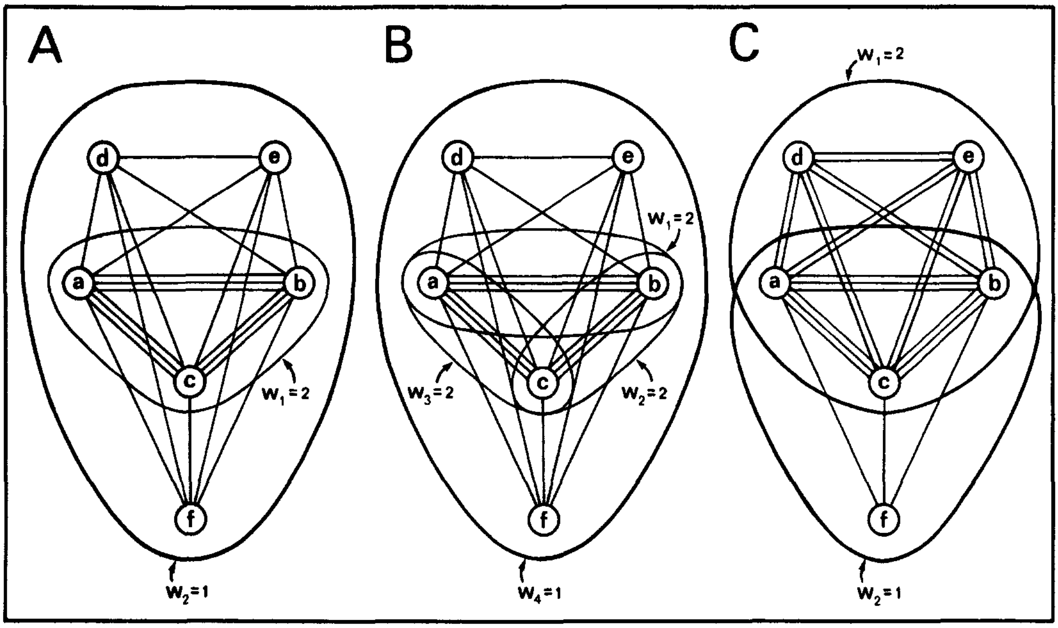

Figure 1. Schematic representation of three different patterns of shared properties that lead to elevation of the same subset of objects (a, b, c).

old subset, A, with the reduced weight $w'_A = w_A - w_B \geqslant 0$; and (b) for every pair of objects with one member in B and the other member in C, we associate a different new property having, however, the same weight, w_B. More generally, we can proceed, recursively, to further partition in the same way any of the smaller subsets obtained by this process.

Given any particular elevated subset, then, the additive model does not itself determine whether that subset corresponds to a single underlying property shared by all of the member objects or to some combination of properties corresponding to interlocking subsets of those objects. As a matter of inductive logic, however, the first alternative is surely to be preferred. It is a far simpler, more parsimonious, and hence more plausible hypothesis that a given pattern of similarities arose from a single property with a single associated weight than from a number of fragmental properties with overlaps and weights that interlock and compensate just so as to yield exactly the same pattern of similarities. Thus, to take the simplest case of an elevated subset consisting of only three members, as illustrated in Figure 1, we should reject the analysis indicated in Figure 1B, which requires three (equal) weights, in favor of the simpler representation depicted in Figure 1A, which requires only one. The larger the elevated subset is, the more compelling the arguments in terms of simplicity and parsimony will be.

It is on the basis of these considerations that we propose the strategy of considering only elevated subsets. Such a strategy will pick out the elevated subset [a,b,c] in the preferred interpretation of Figure 1A, but will not pick out the subsets [a,b], [b,c], and [a,c] in the nonpreferred interpretation, B, because these component subsets are not separately elevated above their union [a,b,c].

Of greatest interest in connection with the additive model is the third possible source of elevation of a subset; namely, overlap of properties extending beyond the elevated subset in question. A simple example is illustrated in Figure 1C. Although the overall pattern of similarities differs somewhat from the pattern common to Figure 1A and Figure 1B, it is again the case that the Subset [a,b,c] forms an elevated subset even though there is no single property uniquely shared by just those three objects. Now, however, the elevation comes about for the different reason that just those three objects are contained in the overlap of two larger subsets each of which does correspond to a single property. Specifically, Objects a, b, c, d, e in Figure 1C are all interconnected by double lines to represent the fact that these five objects all share a property with a weight $w_1 = 2$. Similarly, Objects a, b, c, and f are all interconnected by single lines to reflect the fact that these four objects all share another property with a lower weight $w_2 = 1$. The additive consequence, then, is that the three objects a, b, c, which belong to the intersection of the two subsets, share both properties with a combined weight of $w_1 + w_2 = 3$—as is represented by the bundles of three parallel lines interconnecting just these three objects. Here again, the fact that the similarities are elevated in a subset, such as that containing just a, b, and c, does not entail that the objects in that subset uniquely share some one particular property. In the simple illustrative example, we could conclude that the three objects a, b, and c uniquely share some third property only when the similarities among these objects are consistently greater than the sum of the average similarity of d and e to other objects in Subset 1 and the average similarity of f to other objects in Subset 2 (i.e., greater than the $w_1 + w_2$ represented by the sets of three parallel lines). Consequently, the following conclusion may be drawn: If we begin with all elevated subsets, we will generally begin with many subsets that do not correspond to individual properties but rather to the overlaps of two or more individual properties, and it will appear that we need more properties to account for the given similarity data than is actually the case.

The hope here is that the additive model will allow us to determine which of the elevated subsets can be eliminated because they are coextensive with the intersections of other subsets. In order to proceed, we need some formal definitions, beginning with the notion of the s-level, $s(\mathrm{A})$, of a subset A.

Definition 1. The *s-level* of any subset, A, is the lowest similarity measure between any two obects both contained within that subset, A. Formally,

$$s(A) \equiv \min_{i,j \in A} (s_{ij}).$$

The s-level thus corresponds to the α-level in Johnson's (1967) formulation of complete-link clustering. Our reliance on the ***minimum*** interstimulus similarity value within a subset or cluster is motivated by the idea that if a subset is elevated, then *all* pairwise similarities within that subset must likewise be elevated by the required amount. Thus, the ***minimum*** interstimulus similarity must reach a certain level if the entire subset is to be elevated. It is our selection of the minimum pairwise similarity value within a subset that aligns our method with the complete-link approach to clustering and that determines the associated maximal complete subgraph characterization of the clusters.

In speaking earlier of the desirability of limiting our consideration to elevated subsets, of course, we were speaking informally and without reference to the s-level just defined. As was illustrated in Figure 1B, a subset may be at a considerable elevation in the s-level sense without thereby standing out above the s-levels of other subsets. What is of importance, rather, is the comparative elevation of a subset *relative* to the elevation of a larger subsets containing that subset. Accordingly, we need to make precise the notion of a subset that is elevated in a relative sense:

Definition 2. A subset, A, is said to be *elevated* if and only if every larger subset containing Subset A has a lower s-level than the s-level of Subset A itself. Formally, A is elevated if and only if, for any subset B,

$$B \supset A \rightarrow s(B) < s(A).$$

A noteworthy property of this second definition is that it is based only on ordinal relations in the data. Thus, our set of admissible clusters (viz., those that are elevated in the sense of Definition 2) is invariant under monotone transformations of the data.

The set of all elevated subsets as thus defined has two significant properties. First, it can be obtained without using more than the merely ordinal properties of the similarity data and, so, is "nonmetric" in the sense in which that term is used in connection with multidimensional scaling. And second, according to the model, this reduced set of subsets should tend to include those relatively few of all possible $2^n - 1$ nonempty subsets that are likely to correspond in a one-to-one way to salient underlying properties. Nevertheless, the restriction to elevated subsets is *not* required by the basic model (Equation 1); it is no more than a heuristic device intended to reduce the magnitude of the numerical problem of fitting that model to data. In fact, other variants of the artificial examples illustrated in Figure 1 can be contrived in which subsets corresponding to positively weighted properties fail to emerge as elevated. Still, we believe the heuristic to be reasonable on two grounds: First, we expect subsets corresponding at least to relatively heavily weighted properties to emerge as elevated. Second, the results of applications to actual data using this heuristic have been encouraging, as we shall show. (In fact, they have been generally similar to those obtained by a subsequent numerical method, mentioned at the end of this article, not requiring any restriction to elevated subsets.)

Further Restriction to Subsets with Appreciable Positive Weights

The finding of all elevated subsets would not itself provide a complete solution to our problem. One reason is that we want, in addition, to obtain an estimate of the importance of each such subset, that is, an estimate of the weight of the corresponding property. A second reason is that the set of elevated subsets, though greatly reduced from the set of all possible subsets, is generally still too large for our final representation.

In fact, the number of elevated subsets generally exceeds the number of initially given similarity values, namely, $n(n-1)/2$. Such a fit is worse than trivial. Just as in multidimensional scaling and in factor analysis in which one can always achieve a trivially perfect fit to the data by including as many $n-1$ dimensions, in nonhierarchical clustering, one can always achieve a trivially perfect fit by including as many as $n(n-1)/2$ clusters, each one being simply a distinct pair of the n objects. Then we can account for 100% of the variance by simply but unparsimoniously equating the observed similarity for each pair with the weight of that two-element subset, just as was illustrated for three objects in Figure 1B. Here, as much as in multidimensional scaling or in factor analysis, we do not seek the representation with the best possible fit to the data; we seek, rather, a suitably *parsimonious* representation that still achieves an acceptably good fit to the data.

To account for the maximum fraction of the variance with an appropriately small number of subsets, we need to reduce the set of subsets to only those that according to the additive model, correspond directly to primitive properties. Thus, although the nonmetrically obtained list of elevated subsets can be represented in the form of an $n \times m$ binary property matrix, we need to eliminate many of its m columns before it can be accepted as the final matrix **P** for the purposes of Equation 2. Since the only basis for further reduction is the metric requirement of additivity, this reduction necessarily must be carried out in conjunction with the numerical estimation of the weights. Fortunately, the elimination of all but the elevated subsets at least makes feasible an iterative procedure for the estimation of the weights and, hence, the further elimination of subsets whose weights approach a positive value sufficiently close to zero.

Such an iterative process requires, however, some initial values for the weights, which can then be adjusted to optimize the explained fraction of the variance. To obtain an initial estimate of the weight for each elevated subset, we have found it reasonable and useful to introduce the following natural measure of the degree of relative elevation, which, for mnemonic convenience, we shall refer to as the rise, $r(\mathrm{A})$, of a subset, A.

Definition 3. The *rise* of any subset, A, is the extent to which the s-level of Subset A rises above the greatest s-level of any elevated subset containing Subset A. Formally,

$$r(A) \equiv \min_{B \supset A} [s(A) - s(B)].$$

After proposing this measure, we discovered that it has a fairly extensive history of being reinvented in the literature for single-link clustering (i.e., when the s-levels are based on the maximum similarity instead of the minimum employed in Definition 1). The measure apparently was first proposed by Estabrook (1966), who aptly referred to it as the *moat* of a cluster. A generalization of the measure is given by Jardine and Sibson (1968a, p. 477). With the data converted to ranks, the measure is also Ling's (1972; Ling, Note 4) "isolation index" [1] of a cluster. In our case, the rise is used only as a convenient starting value for each weight, prior to iterative adjustment. The rise plays no essential role in the model or in the final set of primitive weighted properties.

Additive Cluster Analysis: A Computer-Based Implementation

As we have developed it, the present implementation of our version of additive clustering, henceforth called ADCLUS, proceeds in two distinct phases. First, a nonmetric combinatorial algorithm is used to generate a complete list of all elevated subsets that is invariant under monotone transformations of the similarity data. In the second, metric phase, an iterative method (of a modified gradient type) is used to converge on approximately optimum weights for these elevated subsets and additionally to reduce the number of these subsets by eliminating any subset whose weight falls to a threshold level in which that subset is no longer contributing sufficiently to the fraction of variance accounted for. By setting the criterion for the elimination of subsets, we attempt to achieve the desired balance between parsimony and goodness of fit.

Construction of the Set of All Elevated Subsets

Unknown to us when we began this research, Constantinescu (1966) had already published a method that in effect yields the desired list of all elevated subsets for a symmetric matrix of similarities. Elsewhere (Arabie, 1977; Arabie & Shepard, Note 8; Shepard & Arabie, Note 10, Note 11) we have outlined the algorithm [2] that we devised independently and that also computes the rise of each elevated subset. Lawrence J. Hubert (Note 7) has pointed out that most of the algorithms devised for determining maximal complete subgraphs would also be serviceable for the same purpose. That is, after rank ordering the $n(n-1)/2$ similarities, one would find all of the maximal complete subgraphs at the largest of the pairwise s-levels and then proceed to do the same at each (successively smaller) distinct s-level. Indeed, since 1973, when we gave the first brief report of our method (Arabie & Shepard, Note 8), we have found that such algorithms have independently been invented by several workers in the behavioral sciences (e.g., Peay, 1974) and, especially, in computer science (e.g., Bron & Kerbosch, 1973; Johnston, 1976; Nieminen, 1975; Osteen, 1974). Since this graph-theoretic problem has been shown to be NP-complete (see Hansen & Delattre, 1978; Karp, 1972, 1976), it is likely that new algorithms will continue to emerge in the search for ever greater efficiency. For present purposes, however, the selection of any particular algorithm, although certainly affecting the efficiency and cost of the implementation, is irrelevant to the model or to the goodness of fit attained.

Two additional observations are in order concerning the list of elevated subsets. First, the algorithms cited above for extracting maximal complete subgraphs are monotone invariant and do not require that ties be arbitrarily broken. Thus, our clustering solutions do not inherit that liability from the complete-link method. Second, the elevated subset having the lowest s-level is the complete subset of all n stimuli corresponding to an overall additive constant implicitly present in the model of Equation 3.

[1] Ling (1973) also derived an exact distribution for his rank-based isolation index. Unfortunately, the derivation does not apply when the complete-link version (and replacing the rise [Definition 3] with the corresponding difference in ranks) is considered.

[2] We are indebted to C. L. Krumhansl (Note 18) for detecting an error in our algorithm, in that certain patterns of tied similarity values could lead to spurious subsets. Fortunately, the mistake is easily corrected and did not occur for any of the solutions presented later in this article.

Iterative Estimation of the Weights

We assume the given matrix of similarity data, **S**, to be approximately equal (i.e., except for random error) to the matrix $\hat{\mathbf{S}}$ in Equation 2. Thus, if we have a nonsingular matrix, **P**, corresponding to an appropriately chosen set of elevated subsets, we could in principle use multiple linear regression to solve analytically for an estimate of the remaining, unknown matrix in Equation 2, namely, the diagonal matrix, **W**, of desired weights for the chosen subsets (Carroll, Note 12; Cunningham, Note 13; Hubert, Note 7). For details see Shepard and Arabie (Note 10).

Unfortunately, this most direct approach to the estimation of the weights encounters three difficulties. First, regression requires taking the inverse of the $m \times m$ matrix ($\mathbf{P}'\mathbf{P}$). For one matrix of judged similarities between 30 animals (mentioned at the end of the present article), we found the number of elevated subsets, *m,* to be greater than 1,000. (See Moon & Moser, 1965, for upper bounds on the number of maximal complete subsets in a graph.) Inverting a matrix of that size (1,000 × 1,000) is impractical on many computers and expensive on machines for which the task is computationally feasible. Second, unless we have eliminated just the right subsets, the matrix in question is likely to be singular. Indeed, in the example just mentioned, the number of subsets (1,000) exceeds the corresponding maximum possible number of independent subsets ($30 \times 29/2 = 435$), and, so, the matrix will necessarily be singular. The occurrence of singularity would require the techniques of generalized inverses (Kruskal, 1975; Rao & Mitra, 1971). Third, the regression approach does not itself provide a very natural way to achieve the desired elimination. Of course, stepwise multiple regression (Beale, Kendall, & Mann, 1967; Hocking, 1976) could be used, but that procedure becomes very expensive when *m* is in the hundreds. Thus, although there are ways in which each of these difficulties might be overcome, we have chosen to circumvent all three by resorting, instead, to an entirely different, iterative method of optimization based on the gradient of the fraction of variance accounted for with respect to the weights.

Given any preliminary estimates of the weights for an appropriate set of subsets, the partial derivatives of the fraction of variance accounted for, v, with respect to each of the weights, w_k, will specify the small adjustments in these estimates that will result in the locally sharpest increase in the fraction v. We can think of any set of values for the m weights as a single point in an m-dimensional parameter space of the weights. Associated with every point in this space is a scalar (consisting of the single real-valued quantity, v) that specifies the total fraction of variance accounted for at that point, and a vector—the gradient of v (consisting of the m partial derivatives of v with respect to the weights)—that specifies the direction in which v increases most rapidly in that local region of the space. For the chosen appropriate set of subsets, moreover, the gradient of v (as well as v itself) is assumed to vary continuously from point to point in this space. Accordingly, by repeated adjustment of the weights in the direction of the gradient, v can be increased until a stationary point is attained in which the fraction of variance accounted for cannot be improved by any further local adjustments in the weights.

The components of the desired gradient, that is, the partial derivatives of v with respect to the weights, are given explicitly by

$$\frac{\partial v}{\partial w_k} = \frac{2}{d}\Big(\sum_{\substack{i>j \\ \in \text{set } k}} s_{ij} - \sum_{\substack{i>j \\ \in \text{set } k}} \hat{s}_{ij} \Big), \qquad (4)$$

where, as indicated, the summations are over all pairs for which both objects belong to subset k, and where

$$d = \sum_{i>j} (s_{ij} - \bar{s})^2$$

is the constant divisor, which needs to be calculated only once, at the outset.

To facilitate convergence, we use the rise, $r(\text{A})$, computed for each elevated subset, A, as the initial estimate of the corresponding weight w_k. Of course, to the extent that there is overlap of clusters in the true clustering, the rise values, which do not take account of the additivity of weights, will generally overestimate those weights—often quite markedly. (In Figure 1C, for example, the intersection of two subsets appeared as a third subset with a rise value of 1, even though to achieve additivity, the estimate for the weight of the third subset should eventually be adjusted down toward zero.) In order not to overestimate the weights too grossly on any given iteration, the estimated weights are normalized at the beginning of each iteration so that the mean of the predicted similarities, $\hat{s}_{ij}$, is equal to the mean ($\bar{s}$) of the obtained similarities, s_{ij}.

During each successive iteration, a small step is taken in the direction of the gradient by adding to each of the m current estimates of the weights, the corresponding component given in Equation 4, multiplied by a scalar step-size factor. This step-size factor is adaptively modified from iteration to iteration in the manner described by Kruskal (1964b, 1977a) to ensure that the adjustments in the weights become neither so large as to produce nonconvergent oscillation nor so small as to produce unduly slow convergence. In addition, we have found it helpful to weight the gradient $\partial v/\partial w_k$ (see Equation 4) by the reciprocal of the number of pairwise similarities in subset k, namely, $n_k(n_k - 1)/2$.

Approximation to the desired stationary point is assumed to have been achieved when the length of the gradient vector shrinks below a preset near-zero criterion, whereupon the iterative process is terminated. Our experience indicates that the stationary point thus approximated generally yields a substantively interpretable solution accounting for an acceptable fraction of the variance. And, although there is no guarantee that this maximum corresponds to a unique point (i.e., to a unique set of weights and surviving subsets), experience again suggests that if the number of subsets is properly reduced, the final representation will not be unduly arbitrary. We have, however, found that the parameters of the procedure for adjusting step size often require considerable modification for different sets of similarity data.

Further Reduction of the Base Set of Elevated Subsets

During the iterative adjustment of the weights, the requirement of additivity will tend to drive downward the initially overestimated weights associated with those elevated subsets that are also the intersections of other, partially overlapping elevated subsets. As a weight approaches zero, the contribution of that subset to the theoretically reconstructed similarities and, hence, to the total variance accounted for, v will also approach zero.

It is worth noting incidentally that in accordance with the restrictions specified for Equation 1, the program ensures that a weight can never become negative. That is, the ADCLUS representation yields a preferred direction or valence for each subset and thus avoids the arbitrariness of reflection of axes inherent in factor analysis and scaling. To elaborate, if $w_k < 0$, then $-w_k$ applied to the complement of subset k in Equation 2 will not account for the same variance that w_k (< 0) does for Subset k, even with appropriate adjustment of the additive constant. (This fact is readily apparent if one considers a subset having $n - 1$ stimuli and a negative weight. The singleton complement cannot account for any variance, irrespective of whatever weight is assigned to it.) A reinforcing theoretical point is made by Hubert and Baker (1977b, p. 86), who observed that if a subset is being substantively interpreted as representing an underlying feature, there is no a priori reason to suppose that the stimuli in the complement are highly interrelated simply because they lack the relevant property. To ensure that the fraction of variance accounted for has its usual meaning, we do, however, allow negative values for the weight of the complete set, corresponding to a fitted additive constant for the right side of Equation 2. If desired, the addition of an appropriate constant to each of the $n(n - 1)/2$ similarities in the transformed input matrix would render the weight of the complete set arbitrarily positive without changing the value of any of the remaining weights or the goodness of fit.

Elimination of subsets is accomplished during the iterative process by dropping a subset from consideration when its weight falls below a prespecified threshold for two consecutive iterations. The weight is only one of several measures of importance that could be used for this purpose. In particular, it would seem attractive to consider the fraction of variance accounted for by each subset, perhaps normalized according to the size of the subset. However, most schemes for separating out the variance accounted for by a set of nonindependent variables (the subsets in this case) are fraught with a host of problems, including *negative* fractions of variance accounted for. (See Green, Carroll, & DeSarbo, 1978.) Hence, using the weight as the measure of importance of a subset has various advantages in addition to its convenience.

The selection of a threshold for dropping a subset with a small weight has been largely determined by trial and error. In practice, we have usually been able to account for 80% or more of the variance with at most $2n$ subsets, loosely corresponding to the number of parameters fitted in a two-dimensional scaling solution. Once the final, reduced set of subsets has been determined, we have found it advantageous to sharpen the estimates of weights for these subsets together with the weight corresponding to the entire set (and, hence, improve the variance accounted for) by a final pass using multiple linear regression. For this reason, the final weights presented in the following section on applications include the additive constant (in accordance with a suggestion by Hubert, Note 7) and differ somewhat from those obtained earlier without the additive constant and this final pass (Arabie, 1977; Breiger et al., 1975; Shepard, 1974; Arabie & Shepard, Note 8; Shepard & Arabie, Note 10, Note 11).

Illustrative Applications

We now present a variety of applications of the current implementation of the ADCLUS model as it has been described here. The purposes are (a) to indicate the potential effectiveness of the additive model in discovering and representing the relative importance of properties underlying diverse sets of objects, (b) to contrast the nature of these results with those obtained by previous methods of multidimensional scaling and hierarchical clustering, and (c) to provide new evidence bearing on issues concerning the role of physical and semantic features in perception and judgment.

The final ADCLUS solutions presented here all achieve the two goals of (a) providing good fits to the data with relatively few estimated parameters (at least 80% of the variance of the $n(n-1)/2$ similarities accounted for with no more than about n weights in most cases) and (b) yielding subsets that are for the most part readily interpretable. It should, however, be noted that these final levels of goodness of fit and parsimony were often achieved only after many computer runs in which, by adjusting the control parameters governing the elimination of subsets and the choice of step size, we attempted to secure the best possible fit with a relatively small number of subsets. In part, the amount of computation and effort required to obtain these solutions may reflect a degree of nonuniqueness that is inherent in the additive model in which, for many matrices of data, somewhat different patterns of subsets and weights yield nearly equivalent goodnesses of fit. In part, the difficulty, which is in some ways analogous to (and more severe than) the local minimum problem in nonmetric multidimensional scaling, may also be a function of the particular numerical and heuristic methods used here to fit the ADCLUS model. If so, it may be possible to lessen the problem by using different numerical methods, such as the one mentioned at the end of this article.

For uniformity and convenience of comparison of estimated weights across sets of data, in all of the following analyses, the data have been linearly transformed into similarity estimates on the interval [0,1] before being submitted to ADCLUS. Also, for the purposes of graphical presentation of the results of ADCLUS as well as for purposes of direct comparison between these results and those of multidimensional scaling, we shall present spatial solutions for the same sets of data into which the ADCLUS representations are graphically embedded. The spatial solutions were obtained by nonmetric multidimensional scaling as introduced by Shepard (1962a, 1962b) and Kruskal (1964a, 1964b) using, except where specifically indicated, Kruskal's MDSCAL 5M program.

Table 1
Judged Similarities of the Abstract Concepts of the Integers 0 Through 9

Rank[a]	*s* level	Rise	Weight	Elements of subset	Interpretation of subset
1	.638	.064	.577	2 4 8	powers of two
2	.529	.249	.326	6 7 8 9	large numbers
3	.565	.285	.305	3 4 5 6	middle numbers
4	.653	.344	.299	1 2 3	small nonzero numbers
5	.717	.344	.277	3 6 9	multiples of three
6	.574	.192	.165	0 1	additive and multiplicative identities
7	.328	.132	.150	1 3 5 7 9	odd numbers
8	.579	.206	.138	5 6 7	moderately large numbers
9	.382	.118	.112	0 1 2	small numbers
10	.235	.093	.101	0 1 2 3 4	smallish numbers

Note. The data are from Shepard, Kilpatric, and Cunningham (1975). Variance accounted for = 83.1% with 10 subsets, plus additive constant (corresponding to the complete set of 10 numbers). Additive constant = .195.
[a] The ranks are assigned according to the weights (fourth column).

Judged Similarities Between the Concepts of the Numbers 0–9

Shepard et al. (1975) obtained judgments of the perceived similarities between all 10 of the integers 0–9 under each of several conditions. (In these conditions, the similarities were to be judged with respect to various concrete forms in which the numbers might be symbolically represented as well as with respect to the abstract concepts of those numbers themselves, regardless of any concrete symbolic presentation.) In addition to consequences of the results for the human engineering of telecommunication systems, the results are relevant to theories concerning the internal representation of numerical information generally and to processing models for short-term memory for, or comparison of digits, in particular (see Shepard et al., 1975). In the interest of space, we reanalyzed only the pooled data from those conditions in which the judgments were made with respect to the abstract concepts of the numbers themselves.

Table 1 presents the final results obtained when we applied ADCLUS to the judged similarities. The 10 recovered subsets together with the additive constant corresponding to the entire set accounted for 83.1% of the variance of the judged similarities. The recovered subsets are listed in order by weight, along with the computed *s*-level, rise, weight, and included members of each subset. Subsequently added, in the last column, is our suggested interpretation of each subset.

The interpretations appear to be quite clear. They are of two distinct types: those that depend on magnitude (large, middle, small) and those that depend on arithmetic properties (identities, multiples, powers). That these two kinds of properties cut across each other accounts in large part for the considerable overlap of these clusters. Thus the numbers 7 and 9 are grouped with the other large numbers (Subset 2, with $w = .326$) and also, with the odd numbers (Subset 7, with $w = .150$). In terms of the additive model, both of these two quite different properties, largeness and being odd, contribute to the relatively high perceived similarity of 7 and 9.

We have long found it useful to embed clustering representations within multidimensional scaling representations by drawing a closed curve around those points in the scaling solution that correspond to the members of each subset in the clustering solution (see, especially, Shepard, 1972a). In Figure 2 we have thus embedded the nonhierarchical clusters of Table 1 in the two-dimensional spatial solution given by Shepard et al. (1975, Figure 16). The extensive overlap of the obtained subsets is particularly evident when displayed in this spatial form in which the clusters representing magnitude and arithmetic properties tend to be vertically and horizontally elongated, respectively.

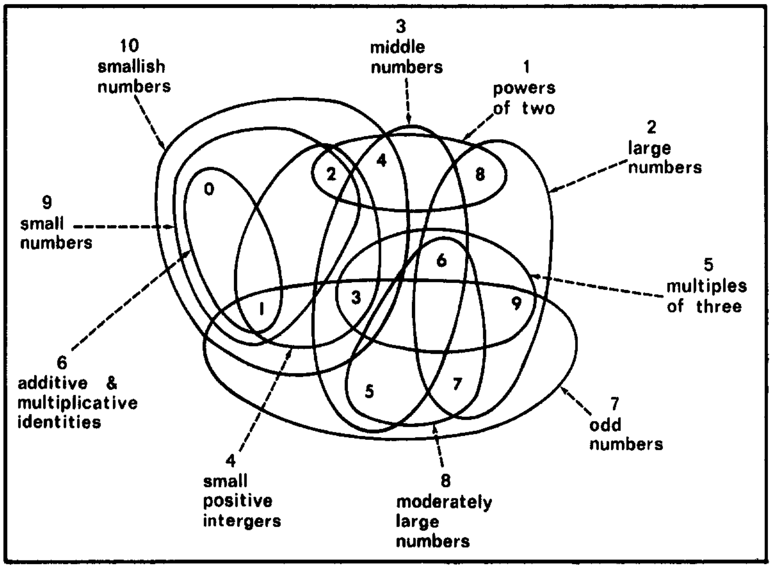

Figure 2. The 10 ADCLUS subsets obtained for the 10 integers, 0–9, studied by Shepard, Kilpatric, and Cunningham (1975), embedded in a two-dimensional scaling representation.

For purposes of comparison, we present in Figure 3 a representation of selected levels of complete-link hierarchical clustering of the same data embedded within the same spatial solution. Beyond the complete set of all 10 numbers, the only subsets that the hierarchical clustering yielded in common with the nonhierarchical are the three indicated in Figure 3 (viz., Subsets 1, 5, and 6). Moreover, the intersections of our ADCLUS solution with those of other hierarchical representations are similar and include the same three subsets for an average-link method (specifically, Sokal & Michener's, 1958, unweighted pair-group

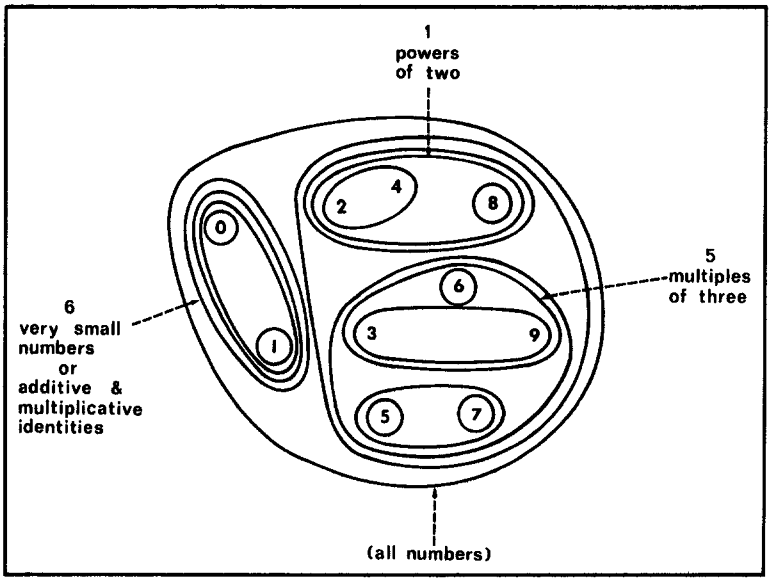

Figure 3. A strictly hierarchical clustering of the same integers embedded in the same two-dimensional scaling representation.

Table 2
Confusions Between 16 Consonant Phonemes

Rank	Weight	Elements of subset	Interpretation
1	.730	f θ	front voiceless fricatives
2	.575	d g	back voiced stops
3	.479	p t k	voiceless stops
4	.464	p k	voiceless stops, omitting t
5	.340	v ð	front voiced fricatives
6	.296	s θ	middle voiceless fricatives
7	.281	m n	nasals
8	.267	b v ð	front voiced consonants[a]
9	.197	s ʃ	back voiceless fricatives
10	.191	p f θ	front voiceless consonants[a]
11	.190	z ð	middle voiced fricatives
12	.156	d g z ʒ	back voiced consonants
13	.153	b v	front voiced consonants[a]
14	.114	v z ð	front and middle voiced fricatives
15	.081	g z ʒ	back voiced consonants
16	.009	z ʒ	back voiced fricatives

Note. The data are from Miller and Nicely (1955). Variance accounted for = 94.5% with 16 subsets, plus additive constant (corresponding to the complete set of 16 consonants). Additive constant = .057.
[a] Subset of questionable interpretation.

method using arithmetic averages) and just two of these subsets (1 and 5) for the single-link method. In a hierarchical model, an overall division between the identities 0 and 1 and the larger numbers 2–9 (as in Figure 3) is incompatible with and therefore precludes any grouping of 1 with 3 on the basis of magnitude (both small numbers, as in our Subset 4) or of arithmetic properties (both odd numbers, as in our Subset 7).

When the data are consistent with a strict hierarchical representation, ADCLUS can, by contrast, yield a set of strictly nested and nonoverlapping subsets. Although ADCLUS thus subsumes hierarchical clustering as a special case, we have never yet obtained a strictly hierarchical solution by applying ADCLUS to any nonartificial data. Holman (Note 5) has given possible reasons why hierarchical results may be an unrealistic expectation.

Confusions Between 16 Consonant Phonemes

In their efforts to explain the perception of speech, psychologists, phoneticians, and linguists have for many years sought a preferred set of discrete underlying perceptual features of consonant phonemes (Jakobson et al., 1963; Klatt, 1968; Liberman et al., 1967; Wickelgren, 1966). Among the most influential and useful of the empirical studies is Miller and Nicely's (1955) experimental investigation of subjects' errors of identification of 16 (English) consonant phonemes under different conditions of filtering and added noise.

The Miller–Nicely data have previously been analyzed through the use of various geometric and clustering models. The data from the so-called *flat noise* masking conditions have been studied by two-way multidimensional scaling methods (Arabie & Soli, 1979; Shepard, 1972a) as well as the extensive usage of complete-link hierarchical clustering (Shepard, 1972a, which also applied that method to two-way data from other experimental conditions). The three-way data (Conditions × Stimuli × Stimuli) have been analyzed using the Carroll–Chang INDSCAL model (Arabie & Soli, 1979; Carroll & Wish, 1974; Wish & Carroll, 1974).

The Miller–Nicely data and other confusion matrices (e.g., Wickelgren, 1966) have also been used for comparing the relative importance of discrete features in various proposed systems of critical features (e.g., Klatt, 1968; Miller & Nicely, 1955; Wang & Bilger, 1973; Wickelgren, 1966), often in an information-

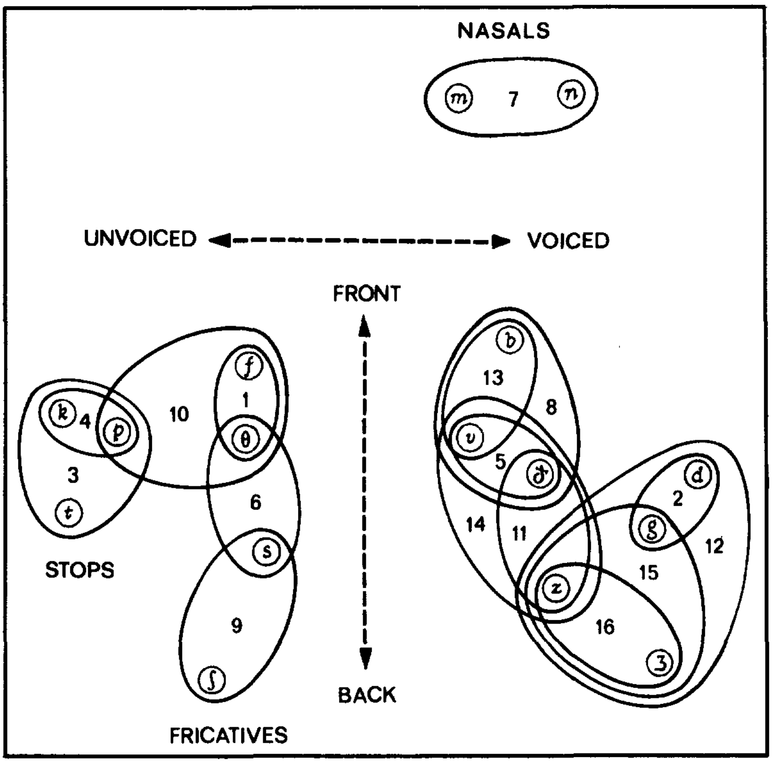

Figure 4. The 16 ADCLUS subsets obtained for the 16 consonant phonemes studied by Miller and Nicely (1955), embedded in a two-dimensional scaling representation.

theoretic context. Klatt (1968) proposed one iterative method for determining a set of "optimal binary features" Wang and Bilger (1973) compared several different proposed sets of features and, on finding no clear-cut superiority for any considered, reiterated Klatt's call for methods to seek optimal sets of features (Wang & Bilger, 1973, p. 1260). We believe that the ADCLUS model and the present implementation are in accord with that programmatic appeal for finding a discrete set of features that approximately optimize an explicit objective function.

Although brief interpretations of subsets are given in Table 2, further comments are in order. Subsets 2 through 4 illustrate some well-known findings from the literature on speech perception. Specifically, we find all the voiceless stops /p t k/ grouped together in the third subset. The fourth subset simply drops /t/ from the third. This grouping of /k p/ to form Subset 4 is not surprising, since both have relatively low frequency noise spectra at the time of burst release, unlike the corresponding high frequency for /t/ in the context of the vowel /a/ (Liberman, Delattre, & Cooper, 1952, p. 504). (Speech spectrograms for the 16 phonemes are given in Carroll & Wish, 1974, and are also reprinted in Arabie & Soli, 1979.)

In contrast to the grouping of the voiceless stops /p t k/ of Subset 3, the present analysis is consistent with that of Shepard (1972a) in showing that despite their phonetic similarity, the corresponding voiced stops /b d g/ do *not* group together. The stops /d g/, which both have the acoustic property of falling second formants, do form the second most heavily weighted subset. The remaining voiced stop /b/, however, has a rising second formant, which is presumably the reason that the phoneme separates from the other two voiced stops in the second cluster (Arabie & Soli, 1979).

The voiced stop /b/ groups instead with the voiced fricative /v/ in Cluster 13, and with /v ð/ as Cluster 8 (comprising /b v ð/). Al-

Table 3
Gibson's Distinctive Feature System for the Uppercase Roman Letters

	Straight lines				Curved lines				Redundancy		Discontinuity	
Elements of subset	1. Horizontal	2. Vertical	3. Diagonal \	4. Diagonal /	5. Closed	6. Open-vertical	7. Open-horizontal	8. Intersection	9. Cyclic change	10. Symmetry	11. Vertical	12. Horizontal
A	X		X	X				X		X	X	
B		X			X			X	X	X		
C							X			X		
D		X			X					X		
E	X	X						X	X	X		X
F	X	X						X			X	X
G	X						X					
H	X	X						X		X	X	
I		X								X	X	
J						X	X					
K		X	X	X				X		X	X	
L	X	X										X
M		X	X	X					X	X	X	
N		X		X							X	
O					X					X		
P		X			X			X			X	
Q				X	X			X				
R		X		X	X			X			X	
S							X		X			
T	X	X						X		X	X	X
U						X				X		
V			X	X						X		
W			X	X					X	X		
X			X	X				X		X		
Y		X	X	X						X	X	
Z	X		X									X

though that 8th cluster was not anticipated on the basis of previously proposed distinctive feature schemes, it did appear repeatedly in Shepard's (1972a) extensive complete-link analyses of independent sets of data from the different Miller–Nicely conditions. Two observations may help to explain this reliable departure from parallelism between the patterns for the voiced and voiceless consonants. The first is the fact, suggested to one of us by Harris Savin (see Shepard, 1972a, p. 104), that the voiced stops /b d g/ do not share, as do the corresponding voiceless stops /p t k/ the feature aspiration. The second is the fact that /b v ð/ all have rising or at least nonfalling second formants. Similarly, the 12th cluster, /d g z ʒ/ contains only consonants with rising second formants. The segregation of front /b v ð/ from back voiceless /d g ʒ z/ consonants is not closely paralleled by the voiceless consonants, perhaps because the second formant transitions are not as varied for the voiceless phonemes.

Quite apart from these specific substantive observations, the extensively overlapping, nonhierarchical nature of the clusters is again evident, especially when graphically embedded in a scaling solution (Figure 4, in which the two-dimensional solution obtained by Shepard, 1972a, is used for this purpose). Note, for example, that although the 16th cluster, which consists of the relatively back fricatives /z ʒ/ is a subset of the 12th cluster /d g z ʒ/ of back consonants with voiced falling second

Table 4
Confusions Between the 26 Uppercase Roman Letters

Subset no.	Rank by weight	Weight	Elements of subset	Subset no.	Rank by weight	Weight	Elements of subset
1	1.0	.685	E F	18	16.5	.208	U V
2	2.0	.542	M N W	19	19.0	.190	C U
3	3.0	.542	C G	20	20.0	.190	L T
4	4.0	.494	P R	21	22.5	.161	A V
5	5.0	.485	K X	22	22.5	.161	B E
6	6.5	.428	M W	23	22.5	.161	D H
7	6.5	.428	O Q	24	22.5	.161	L Z
8	8.0	.390	V Y	25	25.0	.143	D O
9	9.0	.351	E H	26	26.0	.132	X Y Z
10	10.0	.315	M N	27	27.0	.124	V X Y
11	11.5	.256	K N	28	28.0	.122	K V X Y
12	11.5	.256	C O	29	29.0	.113	H M N
13	13.0	.240	B G R	30	30.0	.113	C J U
14	14.0	.238	I L	31	31.0	.103	C G J Q
15	16.5	.208	D O Q	32	32.0	.095	H M
16	16.5	.208	I L T	33	33.0	.087	C G J S
17	16.5	.208	P Q	34	34.0	.077	K X Y

Note. The data are from Gibson, Osser, Schiff, and Smith (Note 14). Variance accounted for = 93.5% with 34 subsets, plus additive constant (corresponding to the complete set of 26 letters). Additive constant = .030.

formants, Cluster 16 /z ʒ/ is also attached, via /z/, to the relatively front fricatives /vð/ in Clusters 11 and 14. The strings of overlapping clusters connecting the four progressively further back voiceless fricatives /f θ s ʃ/, in parallel with the four progressively further back voiced fricatives /v ð z ʒ/, also form configurations that are excluded a priori by any strictly hierarchical representation.

The fact that these 16 subsets are generally interpretable and account for such a large portion of the variance (94.5%) in the original data suggests that further analyses with a different number of subsets and other sets of data could be useful in working toward a more satisfactory distinctive feature scheme for the English consonants.

Confusions Between the 26 Uppercase Letters

Gibson, Osser, Schiff, and Smith (Note 14) obtained confusions between all pairs of the 26 uppercase (Roman) letters using 4-year-old subjects in a matching task. The authors replicated their first experiment and reported the confusion matrices as Tables 1 and 2 of their report. Gibson et al. reported that the original 26 × 26 matrix from the first experiment was "quite symmetrical" (Note 14, p. 10), and the authors therefore symmetrized the matrix by summing the conjugate off-diagonal entries. The same procedure was applied to the matrix from the second experiment, and the two matrices were summed to form the data base of the analyses conducted by Gibson et al. as well as in the reanalysis that we now report.

Gibson et al. performed what they themselves considered a "crude" (Note 14, p. 15) correlational analysis using Gibson's hypothesized set of distinctive features, given here in Table 3. Their analysis assumed that all such features were of equal importance (i.e., equally weighted) and produced inconclusive results in attempting to reconstruct their confusion matrix.

We list in Table 4 the 34 subsets that we obtained by applying ADCLUS to the data of Gibson et al (Note 14). These subsets, together with their 34 weights and the additive constant, accounted for 93.5% of the variance. In Figure 5, these 34 subsets are embedded in a two-dimensional Euclidean MDSCAL solution. This spatial representation corresponds to the lowest stress (.296, with the primary approach to ties) from 20 different random initial configurations that were used to reduce the likelihood of a local minimum (Arabie,

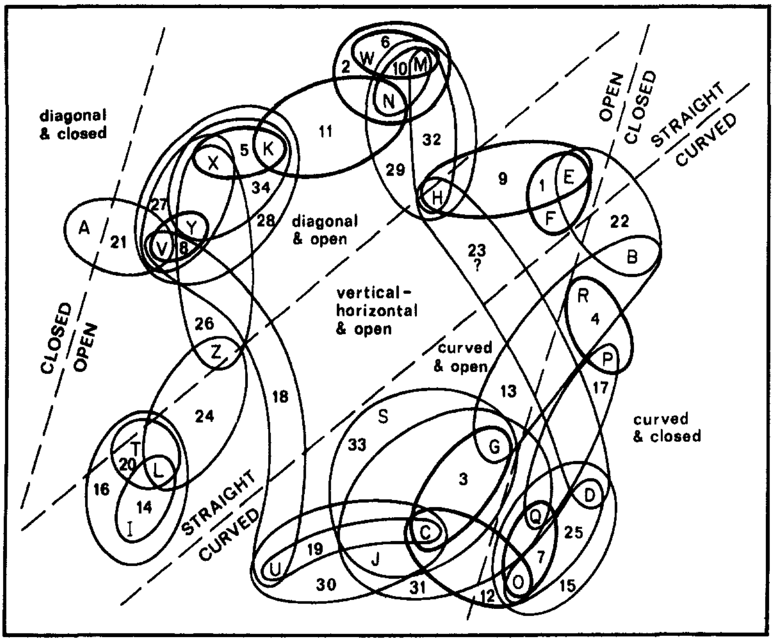

Figure 5. The 34 ADCLUS subsets obtained for the 26 uppercase letters studied by Gibson, Osser, Schiff, and Smith (Note 14), embedded in a two-dimensional scaling representation.

1973, 1978a, 1978b; Shepard, 1974). Several observations are in order concerning these clusters. First, the subsets are generally small; in fact, of the 12 with the largest weights, only 1 subset has as many as three letters. These small cardinalities of the subsets presumably reflect the sparsity of the matrix that they are designed to reconstruct. Second, nearly all of the subsets seem intuitively reasonable. Only the 23rd (D and H, indicated by a question mark in the figure) is puzzling. The rest of the subsets are generally in accord with previous proposals concerning the importance of such features as straightness and curvature, openness and closure. Indeed, as indicated by the dashed straight lines drawn in Figure 5, a global division of the spatial representation on the basis of these two pairs of contrasting features bears a close resemblance to the similar rectilinear division found by Shepard et al. (1975, Figure 9) for the perceived shapes of Arabic numerals. However, as in that study, it is also clear that more complex, configural properties are also significant, for example, in Subset 18 consisting of U and V.

As noted above, the model in Equation 2 can also be fitted by applying multiple linear regression to solve for **W** if the data analyst has supplied a hypothesized set of subsets (for the binary **P** matrix). For confusions between uppercase letters, such a priori subsets are readily available in the form of Gibson's system of distinctive features (Gibson, 1969; Gibson & Levin, 1975, pp. 15–20). These binary features are listed in Table 4 and have been considered in other statistical analyses (e.g., DeWald & Geyer, 1975; Hubert & Baker, 1977b) motivated by concerns similar to ours, but not using the ADCLUS model. (It should be emphasized that the features defining the **P** matrix must be binary; sets of proposed features, e.g., Geyer's in Geyer & DeWald, 1973, that have more than two discrete levels cannot readily be considered within the present ADCLUS framework.)

The use of regression for fitting the ADCLUS model using the 12 subsets corresponding to Gibson's 12 features accounted for only 19.4% of the variance. The resulting weights and additive constant for the features are listed in the middle column of Table 5. It is, of

Table 5
Weights for Gibson's Distinctive Features

	Estimated weights	
Features	12 subsets	24 subsets
Straight lines		
1. Horizontal	.016	.003
Complement		.004
2. Vertical	.038	.048
Complement		.012
3. Diagonal (/)	.085	.070
Complement		.048
4. Diagonal (\)	.073	.088
Complement		−.003
Curved lines		
5. Closed	.123	.107
Complement		.049
6. Open-vertical	.210	.138
Complement		.025
7. Open-horizontal	.250	.180
Complement		.008
8. Intersection	.004	.028
Complement		.027
Redundancy		
9. Cyclic change	.094	.113
Complement		.030
10. Symmetry	−.004	.001
Complement		.006
Discontinuity		
11. Vertical	−.009	−.021
Complement		.017
12. Horizontal	.090	.089
Complement		.020
Additive constant	.028	−.106
Variance accounted for	19.4%	23.8%

course, possible that Gibson's features are simply too inclusive (i.e., have too many elements in the subsets) for the sparsity of her confusion data. Because of this possibility and other considerations, we sought to resuscitate Gibson's proposed features by augmenting them with the complement subsets, in spite of our agreement with Hubert and Baker's (1977b) point about the limitations of the complement of a feature. Application of regression to the resulting 24 subsets yielded only a negligible improvement, accounting for 23.8% of the variance. The weights for the 24-subset analysis are listed in the far right column of Table 5.

Some comments are in order concerning the weights presented in Table 5. First, we note that some of the weights are slightly negative (e.g., for vertical discontinuity, the 11th of Gibson's subsets). Although the iterative implementation of ADCLUS, as described here, precluded negative weights, there is no inherent aspect of multiple linear regression to prevent negative weights from occurring, as they in fact do in Table 5. However, these weights are probably not statistically different from zero, indicating that the weights contribute little to the goodness-of-fit. (The absolute magnitudes of the negative weights are no larger than those occasionally encountered in our experience with Carroll and Chang's, 1970, INDSCAL, another model having no provision for the interpretation of negative weights.) Second, since there is little change in the variance accounted for in the 12- and 24-subset regression analyses, one might expect the pattern of weights for the 12 features common to both analyses to be consistent. If the fitted weights for the 12 features are rank-order correlated across the two analyses, this expectation is confirmed ($\tau_b = .909$).

The fact that Gibson's features yield such a disappointing goodness of fit for her data suggests that the features are in need of revision. The 19.4% and 23.8% variances accounted for (Table 5) cannot, of course, be directly juxtaposed with the 93.5% of the ADCLUS solution, since the latter used 34 subsets. It is naturally of interest to see how 12 ADCLUS-derived subsets would compare in goodness of fit with Gibson's 12 features. Toward this end, we took the 12 most heavily weighted subsets from the analysis presented in Table 3 and used regression to fit new weights for this reduced set of subsets. Such a selection of 12 subsets in no way ensures that they are statistically the best 12 subsets or even the best 12 of the 34; however, this method of selection is both easy and straightforward. Table 6 presents the newly fitted weights for the 12 subsets that along with an additive constant, accounted for 70.2% of the variance.

The 70.2% for the 12 subsets is, of course, considerably less than the 93.5% of variance accounted for by the 34 subsets, and the rank-order correlation of the patterns of weights (for the 12 subsets common to both analyses) between the analyses in Tables 3 and 6, although still considerable ($\tau_b = .651$), is smaller than the coefficient reported for Table 5. The drop in magnitude presumably corresponds to the similar decline in variance accounted for between the two representations. In absolute terms, although the 70.2% variance accounted for by the 12 ADCLUS-derived subsets may not be impressive, it is clearly superior to the 19.4% achieved by Gibson's features. From the standpoint of the ADCLUS model, we conclude (a) that Gibson's proposed set of distinctive features is in need of considerable revision and (b) that her stated goal of economy (viz., a small number of features), however laudable, has been achieved at the expense of an unsatisfactory account of her subjects' discrimination behavior.

Correlations for a Network of 14 Industrial Workers

In the examples considered thus far, the similarities used as input came from square matrices that were either symmetric ab initio or easily symmetrized. However, many types of data, for which investigators often wish to use two-way clustering and scaling methods, are nonsquare, rectangular matrices (e.g., Items × Attributes, instead of Stimuli × Stimuli). For such matrices, the data analyst typically must decide whether the rows or columns are to be represented and then derive some measure of profile similarity between all pairs of rows/columns such as product-moment correlations or Euclidean distances between rows/columns (see Cronbach & Gleser, 1953; Shepard, 1972b). The present example uses correlations between columns of a rectangular matrix of sociometric data. In the concluding section, we return to a consideration of the rationale for applying ADCLUS to matrices of correlation coefficients (and to the relation between ADCLUS and factor analysis). For now we merely note that in earlier attempts to obtain ADCLUS representations of data from (initially) rectangular matrices, straightforward application to computed correlations (Arabie, 1977; Breiger et al., 1975) worked surprisingly well.

The present data were collected by an observer who recorded social interactions between 14 industrial workers in the Western Electric Company's Hawthorne Works in Chicago between March 1931 and May 1932 under the auspices of a well-known industrial productivity study. The present description is necessarily abbreviated; extensive details of the study and data are given by the original investigators, Roethlisberger and Dickson (1939), and in a review by Homans (1950). The 14 workers were constructing banks of telephone switching apparatus and were classified as inspectors (I1, I3), wiremen (W1, W2, . . . W9), and soldermen (S1, S2, S4). These men worked in a specially designed room in which an observer external to the group recorded their interactions according to various types of social ties. In the present analysis, the five types of ties are "liking," playing "games" together, "antagonism," "helping," and disagreements over opening "windows." Roethlisberger and Dickson and Homans reported these data in the graphic form of sociograms, and Breiger et al. (1975, p. 345) presented them as binary sociomatrices. For the present analysis, the five nonsymmetric 14 × 14 matrices were "stacked" to form a 70 × 14 matrix, and product-moment correlations were computed between the columns (representing the recipients of actions, rather than the initiators). The resulting correlation matrix given in Breiger et al. (1975,

Table 6
Regression Weights for 12 ADCLUS-Derived Subsets Fitted to Data of Gibson, Osser, Schiff, and Smith

Subset no.	Rank by weight	Weight	Elements of subset
1	2.5	.664	E F
2	6	.521	M N W
3	1	.712	C G
4	7	.473	P R
5	2.5	.664	K X
6	8.5	.428	M W
7	4.5	.616	O Q
8	4.5	.616	V Y
9	10	.330	E H
10	8.5	.428	M N
11	11.5	.235	K N
12	11.5	.235	C O

Note. Variance accounted for = 70.2%, additive constant = .050.

Table 7
Social Categorization in the Bank Wiring Observation Room

Actor's identification	Roethlisberger–Dickson and Homans' assignment	Salient aspect of actor
W1	A	close friend of W3 (p. 460)
W3	A	most well-liked of group; leader (pp. 464–465)
W4	A	antagonistic to wiremen in clique B (p. 465)
S1	A	generally well liked (p. 480)
I1	A	socially adroit, popular (pp. 484–486)
W2	—[a]	unsociable "rate-buster" (p. 463)
W7	B	slow worker
W8	B	slow worker
W9	B	slow worker
S4	B	socially regarded as inferior (p. 483)
W6	—[b]	unsuccessful aspirant to leadership and popularity (p. 471)
W5	isolate[c]	"without doubt the most disliked wireman in the group" (p. 468)
S2	isolate[c]	socially impeded by a speech difficulty (p. 482)
I3	isolate[c]	extremely unpopular and eventually had to be transferred from the room (p. 487)

Note. Page numbers refer to Roethlisberger and Dickson (1939).
[a] Actor W2 was affiliated with Clique A but was not a stable member of it (Roethlisberger & Dickson, p. 510).
[b] "That he [W6] was not entirely accepted in clique B was shown in many ways, chief of which was the way in which clique B co-operated in resisting his attempts to dominate anyone in their group. Yet he participated in clique B much more than W2 did in clique A. It may be concluded that although W6 tended to participate in clique B, he was still in many ways an outsider" (Roethlisberger & Dickson, p. 509).
[c] "There were three individuals, I3, W5, and S2, who were clearly outside either clique" (Roethlisberger & Dickson, p. 510).

p. 343) [3] is used in the present analysis. Arguments have been made for why diagonal entries should not have been included in computing the correlations (Arabie, Boorman, & Levitt, 1978), but the data base of Breiger et al. (1975) is retained here to facilitate comparison with the present analysis.

The most prominent aspects of the social structure of the workers were noted by Roethlisberger and Dickson (1939), largely on the basis of the "helping" and "games" ties: two non-antagonistic cliques, A and B, and three social isolates. Table 7 summarizes Roethlisberger and Dickson's classification, with which Homans (1950) concurred. This basic structure is consonant with blockmodel and ADCLUS analyses in Breiger et al. (1975). The present ADCLUS solution is more parsimonious (i.e., approximately same goodness of fit, with fewer subsets) than the earlier one, owing to refinements in the implementation subsequent to the publication of Breiger et al.

Table 8 presents the 10 subsets that along with the additive constant, accounted for 89.0% of the variance. The 10 subsets are embedded in a two-dimensional Euclidean MDSCAL solution in Figure 6. The solution was obtained from the best of 20 different random initial configurations (stress = .126).

The 10 clusters are interpreted as falling into three general categories: (a) isolates, (b) Roethlisberger and Dickson Cliques A and B, plus subsets and modifications, and (c) the most popular worker (W3) and his associates. (These categories are not mutually exclusive.) Beginning with the first category, we note that the first, second, and sixth most heavily weighted clusters comprise the three social isolates (W5, S2, I3; see Table 7) and W2,

[3] The following corrections to misprints in Table III of Breiger, Boorman, and Arabie (1975) should be noted: $r(\text{W7, S1}) = .08$; $r(\text{W7, W4}) = .03$; $r(\text{S4, S2}) = -.05$; $r(\text{I3, W3}) = -.17$; $r(\text{I3, W6}) = .05$.

Table 8
ADCLUS Representation of 14 Workers at Hawthorne

Rank by weight	Weight	Members	Interpretation
1	.415	S2 I3	isolates
2	.385	W2 W5 I3	isolates plus unpopular "rate-buster" (W2)
3	.363	W6 W7 W8 W9 S4	Roethlisberger–Dickson Clique B and affiliate (W6)
4	.342	W1 W2 W3 W4 S1 I1	Roethlisberger–Dickson Clique A and affiliate (W2)
5	.302	W1 W3 W4 SI	leader (W3) and his closest friends (subset of Cluster 4)
6	.287	W2 I3	"rate-buster" (W2) and most unpopular member of group (I3); subset of Cluster 2
7	.270	W1 W2 W5 I1	questionable
8	.235	W1 W3	leader (W3) and reciprocated best friend; subset of Cluster 5.
9	.201	W4 W5 W6 S1 S2 W7 W8 W9 S4 I1 I3	W1, W2, W3 excluded from rest of group
10	.172	W6 W7 W8 W9 S2	modified Clique B

Note. The data are from Roethlisberger and Dickson (1939); see also Breiger, Boorman, and Arabie (1975). Eighty-nine percent of the variance accounted for with 10 subsets, plus an additive constant of .12.

who was not considered a "stable" member of clique A (Roethlisberger & Dickson, 1939, p. 510). Taking these three subsets in turn, we note that the first subset is S2, the solderman who suffered from a speech defect, and I3, the unpopular inspector who was eventually driven from the room. That same inspector, plus "rate buster" W2, and the most unpopular of the wiremen, W5, form the second cluster. The latter two members are also the entire membership of Cluster 6, which is therefore a proper subset of Cluster 2.

These subsets of isolates call for several methodological comments. First, there is no indication from Roethlisberger and Dickson (1939) that the four workers belonging to these subsets in any way formed coalitions, cliques, or other cohesive factions to oppose Cliques A and B: These men are grouped together because of the *consistency* of their stances vis-à-vis the social structure, rather than because of any social bonds or *connectivity* between these isolates. Specifically, these men were consistent in that they are generally ignored, avoided, or disliked by the rest of the workers. (For example, inspection of the data presented in Table V of Breiger et al., 1975, shows that the only form of recognition that S2 received, out of 65 possibilities, was W5's antagonism, and reciprocating that dislike was the only recorded social interaction that S2 initiated.) The fact that the three isolates are generally grouped together demonstrates that the ADCLUS representation is consistent with the blockmodeling emphasis of Breiger et al. in manifesting consistency or *structural equivalence* (Lorrain & White, 1971) within the social system while avoiding the "small group" research emphasis on connectivity.

Clusters 4 and 3 depict Roethlisberger and Dickson's (1939) Cliques A and B, respectively, including the tangential member of each (viz., W2 for A and W6 for B). Cluster 5 consists of the most popular member of the group (W3) and his closest friends (Roethlisberger & Dickson, 1939, p. 464) and is a subset of Cluster 4. A further refinement of this group of friends is given by Cluster 8, comprising the leader (W3) and his reciprocated closest friend W1 (Roethlisberger & Dickson, 1939, p. 460).

The remaining clusters (7, 9, 10) are problematic in that the most obvious interpretation for each would require the accretion or deletion of exactly one worker. For instance, Subset 7 would be a subset of clique A (Cluster 4), were it not for the inclusion of W5. Cluster 10 would be a subset of Clique B (Subset 3), except for the presence of S2.

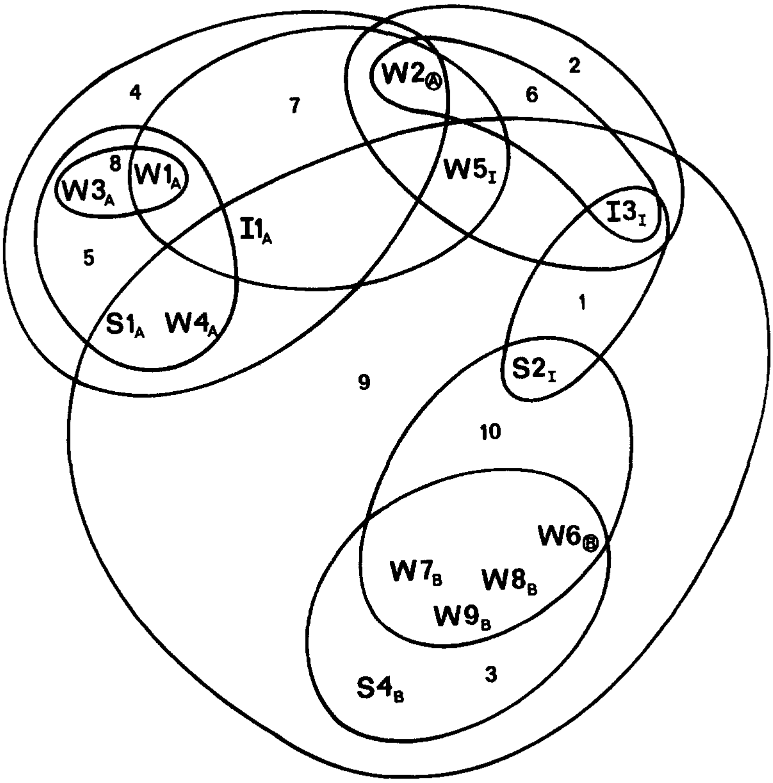

Figure 6. The 10 ADCLUS subsets obtained for the 14 Hawthorne workers studied by Roethlisberger and Dickson (1939), embedded in a two-dimensional scaling representation.

Finally, if W2 were added to Cluster 9, that cluster would be the complement of 8 and could be considered as relatively distant followers of the leader, W3, and his closest friend, W1. The fact that these clusters are "off" by one worker suggests that the model is harder to fit to these correlation data, at least for the subsets with comparatively lower weights. Nonetheless, the goodness of fit (89.0%) is adequate, and the interpretations of the first subsets could be given lengthy supporting citations from the source, Roethlisberger and Dickson (1939). Note also that the nonhierarchical representation does successfully accommodate the ambiguous status of W2 as being disliked (Clusters 2 and 6) but simultaneously being a satellite of Clique A (Clusters 4 and 7). This facility of ADCLUS for faithfully depicting seemingly contradictory aspects of social structure makes it a promising technique for the representation of increasingly complex and flexible models of social structure (cf. White, Boorman, & Breiger, 1976).

Co-occurrence in Sorting Names of Anatomical Parts

Miller (1969) collected data on the perceived similarity of the labels of 20 body parts, chosen on the basis of a rather clear hierarchy of anatomical inclusion, with some ambiguities. The task of each of the 50 subjects was to sort labels of these anatomical parts into groups on the basis of perceived similarity, and the number of subjects putting a pair of items into the same group is used as the measure of interstimulus similarity. The fact that such data (after a linear transformation) already satisfy the metric axioms (Miller, 1969) suggests that sorting data are "cleaner"—perhaps artifactually so—than many other forms of similarity data. For the purposes of the present reanalysis, we

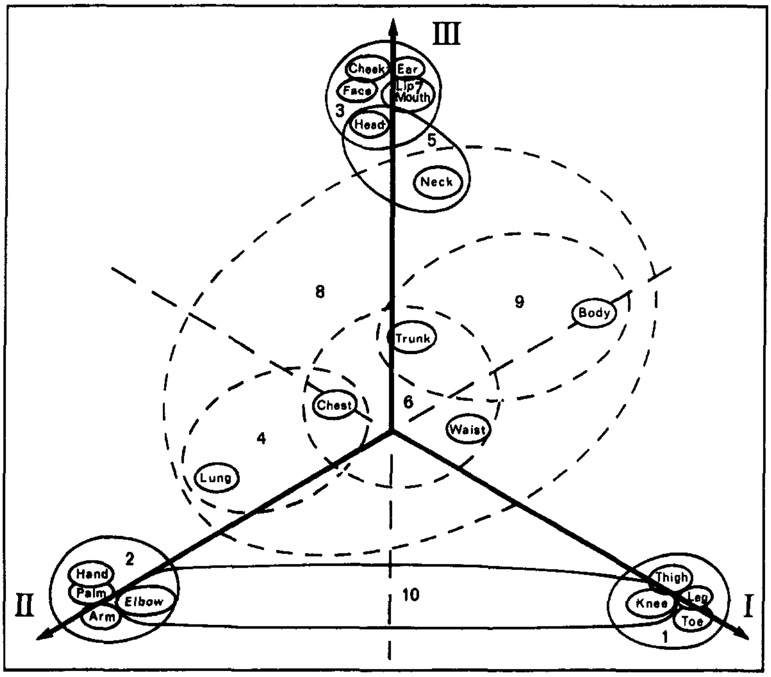

Figure 7. The 10 ADCLUS subsets obtained for the 20 body parts studied by Miller (see Carroll & Chang, 1973), embedded in a three-dimensional scaling representation.

started with Miller's data in the form in which they were presented by Carroll and Chang (1973).

Using ADCLUS we found 10 subsets and an additive constant that together accounted for 95.6% of the variance. These subsets, which are listed in Table 9, are to a large extent consonant with the results of Carroll and Chang's (1973) hierarchical analysis of these same data in that there are four clusters, corresponding to a leg group (Subset 1), arm group (Subset 2), head group (Subset 3), and a less tightly clustered trunk group (Subset 8).

In Figure 7, these same subsets are embedded in a two-dimensional projection of a three-dimensional MDSCAL configuration obtained using the so-called city-block variety of the Minkowski r-metric, $r = 1$, in place of the more usual Euclidean variety, $r = 2$ (see Kruskal, 1964a; Lew, 1978; Shepard, 1964, 1974). For the representation of these data, our choice of the city-block metric, which is based on a simple additive rule of combination very much like the one that underlies the ADCLUS model, had both theoretical and empirical reasons. Theoretically, if the city-block metric does apply, it has the advantage that the axes should be directly interpretable, without the need for any rotation. Empirically, although comparison of stress values between different metrics is open to question (Shepard, 1974), this measure of departure from good fit was appreciably better for the city-block solution (.019) than for the Euclidean (.046). Indeed, stress was lower for the city-block solution ($r = 1$) than for any of a considerable number of different r values tried for the Minkowski r-metrics obtained by the iterative approach of refined initial configurations utilized by Kruskal (1964a) and also independently used and described in detail by Arnold (1971). Finally, contrary to what

Table 9
Similarities for 20 Body Parts

Rank	Weight	Elements of subset
1	.820	knee leg thigh toe
2	.754	arm elbow hand palm
3	.722	ear cheek face head lip mouth
4	.433	chest lung
5	.348	head neck
6	.315	chest trunk waist
7	.231	lip mouth
8	.206	body chest lung neck trunk waist
9	.204	body trunk
10	.119	elbow knee

Note. The data are from Carroll and Chang (1973). Variance = 95.6% with 10 subsets, plus additive constant. Additive constant = .048.

would be expected with the rotationally arbitrary Euclidean metric, the axes of the obtained city-block solution did turn out to have direct interpretations to which we now turn.

Although the city-block metric differs from the Euclidean in its exclusion of rotations of axes (except through angles that are multiples of 90° and, so, correspond to reflections and permutations of axes), the two types of metric are alike in remaining invariant under translations, reflections, and permutations of axes. It is therefore significant that with a slight translation of the coordinate system only, the axes could be made to pass through the middle of the three most heavily weighted and compact clusters, namely, those corresponding to the leg, arm, and head (Axes I, II, and III in Figure 7). In Figure 7, we merely show the entire system of points and axes as it looks from a convenient viewing angle—not after a prohibited rotation of the axes with respect to the points. The three axes portrayed in Figure 7 are in fact orthogonal, with the solid darker extensions (which pierce Clusters 1, 2, and 3) pointing toward the viewer and with the lighter dashed extensions pointing back in depth. The position in depth of the remaining, much looser trunk cluster (Subset 8) can then be described as falling almost entirely in the rearmost octant of the space enclosed by the three planes defined by the three dashed extensions of the coordinate axes. As an indication of their location in depth, Subset 8 and the three smaller clusters 4, 6, and 9, which are proper subsets of Subset 8, are enclosed in broken curves. Notice incidentally that Subset 6 overlaps partially with both Subsets 4 and 9 and, for this reason, contraindicates a strictly hierarchical structure.

The most striking and interpretable violations of a hierarchical structure are provided by Subsets 5 and 10. Subset 5 forms an overlapping "neck" bridge between the head group (Subset 3) and the trunk group (Subset 8). Likewise, Subset 10 bridges between the two most heavily weighted but widely separated subsets, the leg group (Subset 1) and the arm group (Subset 2) by linking the two functionally analogous parts, the knee and the elbow. Thus, in addition to its faithful portrayal of the four major clusters of body parts, Subsets 1, 2, 3, and 8, found by other analyses, such as that of Carroll and Chang (1973) and the present city-block solution, the ADCLUS solution reveals interesting and interpretable departures from a purely hierarchical structure.

Some Relationships Between ADCLUS and Tversky's Contrast Model

In his seminal article on features of similarity, Tversky (1977) derived from a set of plausible axioms a general model in which the similarity between two objects, whose features separately constitute Sets A and B, is a weighted linear combination of a function, f, of the set of features common to the two objects and of the same function of the sets of features unique to each: $\theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$. He pointed out (p. 345) that the ADCLUS model, as we had already briefly outlined it in one of our preliminary reports (Shepard & Arabie, Note 10), can be regarded as one way of representing a limiting case of this general contrast model in which the objects are equally rich in features so that for any two objects A and B, $f(\mathrm{A}) = f(\mathrm{B})$.

An appealing aspect of Tversky's more general formulation is that it provides for the explanation of asymmetries in similarity data. For, if there are differences in the richness of the sets of features of two objects i and j, *and* if there are asymmetries in the way the two objects are presented such as to render $\alpha \neq \beta$, then it is also predicted that $s_{ij} \neq s_{ji}$. This fact is significant because the existence of such asymmetries in similarity data has been amply documented by Tversky (1977; and other articles cited there). Nevertheless, we believe it fair to say that except in cases in which the principal interest is in stimuli varying widely in prominence, familiarity, or complexity *and* in which an asymmetric method of presentation is used, the departures from symmetry of the resulting similarity data are generally sufficiently small relative to other kinds of variations in the data that the symmetrical ADCLUS model can, as in the illustrative applications, provide a useful first approximation to the discrete structure underlying the relations between the stimuli. In such cases, ADCLUS offers what, from the data-analytic standpoint, is the real advantage of an operational method for finding the implied underlying structure in a form not provided by Tversky's more abstract formalism: namely, as an explicit set of subsets and their associated weights. Indeed, by permitting explicit estimation of the weights of different sets of features under different contextual conditions, ADCLUS might provide for a structurally richer test of some of the predictions of Tversky's theory.

At the same time, we are tempted to think that the basic approach that we have taken to the representation of symmetric similarities as linear combinations of feature weights might be extendable to the more general case in which the similarities depend on the unique as well as on the common properties of the objects and, so, are not necessarily symmetric. Even as it stands, the ADCLUS model has the potential for accommodating variations in self-similarity that as Tversky also noted, are often found in the principal diagonal of a complete similarity matrix. For, although we have not as yet exploited this possibility, the sum of the weights for all subsets to which any single object belongs can vary in ways that depend on the relations of subset inclusion with other objects in a given set and also in other, independent ways.

Some Relationships Between ADCLUS and Factor Analysis

In psychology, factor analysis has often been used as a method of clustering. Axes are rotated in a high-dimensional space so that according to some specified criterion of "simple structure," each axis passes through a more or less well-defined cluster of the objects studied (much as was portrayed in Figure 7, here). Even so, the result of such a rotation is not itself an explicit, categorical specification of membership or nonmembership in discrete subsets. Traditionally, in psychological applications, such a specification requires an additional step of all-or-none assignment of each object to appropriate subsets, usually on the basis of the numerical values of the loadings on the rotated and possibly oblique factors. Notice, however, that regardless of whether such a categorical assignment is done on the basis of subjective judgment or some automatic clustering procedure tacked on following the rotation, the relationship of this final, discrete representation to the original correlational data is apt to be complex and difficult to specify (cf. Ling, Note 4).

Although it does not, of course, correspond to any one of the previously proposed methods of rotation in factor analysis, a possible advantage of ADCLUS for this same general purpose is that it provides an explicit final assignment of the n objects under study to discrete subsets, and it does so in such a way that those same discrete subsets are chosen on the basis of their ability to provide (through suitable weights) a direct account of the original data via the simple additive model. In this somewhat different sense, too, we can say that ADCLUS represents a kind of discrete factor analysis. Whereas in the standard factor-analytic model continuous variations in correlation are accounted for by continuous variations in factor loadings, in ADCLUS the analogues of the factor loadings (i.e., the assignments to subsets) are all or none; continuous variations in the data are explained, rather, by the continuously variable weights of these subsets. For purposes of classification of objects into clusters, it appears to us preferable to confine the continuous variation to the weights of the subsets and to ensure that the relations between objects and subsets will be of a discrete, all-or-none character.

Although ADCLUS has yielded readily interpretable results by direct application to product-moment correlations, as in the example presented in Table 5 and Figure 6, in some of our attempted applications, the obtained subsets have been so numerous, large, overlapping, and evenly weighted that the computation was judged excessively costly and the result insufficiently parsimonious. Perhaps some monotonic transformation of the correlations could yield data that are more appropriately decomposed into additive components (cf. Arabie & Soli, 1979). And one of us (R.N.S.) is exploring the possibility that under some circumstances, even the sometimes nonmonotonic transformation to r^2 might prove useful.

Incidentally, the possibility that some monotonic transformation might appropriately intervene between the formation of the additive combination of the underlying weights and the observed data suggests the following nonmetric generalization of the metric ADCLUS model:

$$\hat{s}_{ij} = f_{\text{mon}}(\sum_k w_k p_{ik} p_{jk}), \qquad (5)$$

where, for similarity data, f_{mon} would be the monotonically increasing function that optimizes goodness of fit.

Such a generalization would be analogous to the nonmetric generalization of linear factor analysis developed by Kruskal and Shepard (1974) in which the data are represented as monotonic functions of underlying linear combinations, or to the Carroll and Chang (Note 15) "quasi-nonmetric" version of INDSCAL. These two other programs exemplify some possibilities for incorporating the fitting of monotonic functions in the computer implementation of ADCLUS by means, for example, of the Miles-Kruskal algorithm for monotone regression (Kruskal, 1964b). So far, however, the numerical problems associated with the simple, metric version of the ADCLUS model have been sufficient that we have not attempted to implement a nonmetric generalization.

Prospects for a More Efficient and General Numerical Method

Although the ADCLUS method as implemented here appears to have been successful in analyzing the preceding sets of data, the analysis of matrices that are still larger and that contain a relatively greater proportion of distinct similarity values can become so costly that we have in some cases been unable to satisfy ourselves that the obtained representation is appropriately optimal or unique. In an analysis of similarity data for 30 animals, obtained from Arabie and Rips's (Note 16) replication of an experiment by Henley (1969), for example, consideration of total computing time led us to terminate the analysis before reducing the number of retained subsets below 40. Moreover, although it was necessary to include at least the first 25 of these subsets to ensure that each of the 30 animals was included in at least one of the subsets, the weights for a few of these subsets were estimated by subsequent regression to have values that, although small, were negative.

Nevertheless, the appropriateness of the ADCLUS model was still supported in that the subsets with the largest positive weights were strikingly coherent and interpretable. The first 5 subsets, for example, contained the following elements: monkey, chimpanzee, gorilla; rat, mouse; cat, lion, tiger, leopard; deer, antelope; and dog, wolf, fox; all of which correspond to prominent clusters recently obtained by Sattath and Tversky (1977) by fitting "additive similarity trees" to Henley's (1969) original data for these animals. Moreover, the clusters from an earlier analysis (presented in Note 8) were useful in Sternberg's (1977, chap. 10) investigations of subjects' reasoning processes in solving analogies problems based on the set of Henley's animals. Sternberg employed the clusters (without using their weights) to obtain a preliminary account of some results specific to the domain of these stimuli. However, his discussion correctly noted the potential difficulties of using an ADCLUS representation in which one or more of the stimuli is not covered by any of the clusters other than the complete set associated with the additive constant. Sternberg's observations on the compatibilities between the ADCLUS model and analogical reasoning have made us hopeful that with further advances in algorithms for implementing the model (see below), successful solutions for data such as the Arabie–Rips (Note 16) replication of Henley (1969) can be obtained.

The difficulties we encountered in analyzing the data from the replication of Henley's experiment together with more general considerations of computing time and amount of core required have motivated a continuing search for more efficient numerical methods for fitting the ADCLUS model. One alternative uses a "mathematical programming" approach combined with alternating least squares (Arabie & Carroll, Note 17). The data analyst must specify how many subsets are to be in the solution, but there is no restriction that the subsets be elevated in the sense of our Definition 2. In addition to what appears to be its greater computational efficiency, this new approach has the advantage of offering a facile generalization to the three-way case (see Carroll & Pruzansky, Note 2) in which individual subsets can have differing numerical weights for a common list of subsets. Therefore, although we take the results reported here as strongly supportive of the basic ADCLUS *model,* we have terminated further development of the particular numerical *method* and the associated computer program briefly outlined here in favor of the newer numerical method for fitting that same model to data. The full description of that newer method and of its performance in obtaining (similar) results to test cases will be reported elsewhere (Arabie & Carroll, Note 17).

Reference Notes

1. Carroll, J. D., & Pruzansky, S. Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. *Proceedings of the U.S.–Japan Seminar on Theory, Methods, and Applications of Multidimensional Scaling and Related Techniques.* University of California at San Diego, August 20–24, 1975.
2. Carroll, J. D., & Pruzansky, S. Handout for fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. *U.S.–Japan Seminar on Theory, Methods, and Applications of Multidimensional Scaling and Related Techniques.* University of California at San Diego, August 20–24, 1975.
3. Cross, D. V. *Multidimensional stimulus control of the discriminative response in experimental conditioning and psychophysics.* (Tech. Rep. 05613-4-F (78[d]), Ann Arbor: University of Michigan, 1965.
4. Ling, R. F. *Cluster analysis.* Ann Arbor, Michigan: University Microfilms, 1971. No. 71-22356.
5. Holman, E. W. *A test of the hierarchical clustering model for dissimilarity data.* Unpublished manuscript, University of California at Los Angeles, 1975.
6. Panning, W. H. *Committee factions in the U.S. House of Representatives.* Unpublished manuscript, Wesleyan University, 1977.
7. Hubert, L. J. *A note on additive cluster analysis.* Unpublished manuscript, University of Wisconsin-Madison, 1976.
8. Arabie, P., & Shepard, R. N. *Representation of similarities as additive combinations of discrete overlapping properties.* Paper presented at the Mathematical Psychology meeting in Montreal, August 1973.
9. Hubert, L. J. Personal communication, 1976.
10. Shepard, R. N., & Arabie, P. Additive cluster analysis of similarity data. *Proceedings of the U.S.–Japan Seminar on Theory, Methods, and Applications of Multidimensional Scaling and Related Techniques.* University of California at San Diego, August 20–24, 1975.
11. Shepard, R. N. & Arabie, P. *Handout for additive cluster analysis of similarity data.* Invited presentation at the joint meeting of the Psychometric Society and the Classification Society at the University of Iowa, Iowa City, April 25, 1975.
12. Carroll, J. D. Personal communication, 1973.
13. Cunningham, J. P. *Comments on the optimization procedure used in the Shepard-Arabie "nonhierarchical clustering" algorithm.* Unpublished memorandum, University of California at San Diego, 1973.
14. Gibson, E. J., Osser, H., Schiff, W., & Smith, J. An analysis of critical features of letters, tested by a confusion matrix. In, *A basic research program on reading.* (Cooperative Research Project No. 639) U.S. Office of Education, 1963.

15. Carroll, J. D., & Chang, J. J. *A "quasi-nonmetric" version of INDSCAL, a procedure for individual differences multidimensional scaling.* Paper presented at the meeting of the Psychometric Society, Stanford University, Stanford, California, March 21–22, 1970.

16. Arabie, P., & Rips, L. *A 3-way data set of similarities between Henley's 30 animals.* Unpublished manuscript, Stanford University, 1973.

17. Arabie, P., & Carroll, J. D. *MAPCLUS: A mathematical programming approach to fitting the ADCLUS model.* Manuscript submitted for publication.

18. Krumhansl, C. L. Personal communication, 1977.

References

Arabie, P. Concerning Monte Carlo evaluations of nonmetric multidimensional scaling algorithms. *Psychometrika,* 1973, *38,* 607–608.

Arabie, P. Clustering representations of group overlap. *Journal of Mathematical Sociology,* 1977, *5,* 113–128.

Arabie, P. Random versus rational strategies for initial configurations in nonmetric multidimensional scaling. *Psychometrika,* 1978, *43,* 111–113, (a)

Arabie, P. The difference between "several" and "single": A reply to Spence and Young. *Psychometrika,* 1978, *43,* 119. (b)

Arabie, P., & Boorman, S. A. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology,* 1973, *10,* 148–203.

Arabie, P., Boorman, S. A., & Levitt, P. R. Constructing block models: How and why. *Journal of Mathematical Psychology,* 1978, *17,* 21–63.

Arabie, P., Kosslyn, S. M., & Nelson, K. E. A multidimensional scaling study of visual memory in 5-year-olds and adults. *Journal of Experimental Child Psychology,* 1975, *19,* 327–345.

Arabie, P., & Soli, S. D. The interface between the type of regression and methods of collecting proximities data. In R. Golledge & J. N. Rayner (Eds.), *Multidimensional analysis of large data sets.* Minneapolis: University of Minnesota Press, 1979.

Arnold, J. B. A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology,* 1971, *90,* 349–372. (Monograph)

Attneave, F. Dimensions of similarity. *American Journal of Psychology,* 1950, *63,* 516–556.

Baker, F. B. Stability of two hierarchical grouping techniques, Case I: Sensitivity to data errors. *Journal of the American Statistical Association,* 1974, *69,* 440–445.

Baker, F. B., & Hubert, L. J. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association,* 1975, *70,* 31–38.

Baker, F. B., & Hubert, L. J. A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association,* 1976, *71,* 870–878.

Beale, E. M. L., Kendall, M. G., & Mann, D. W. The discarding of variables in multivariate analysis. *Biometrika,* 1967, *54,* 357–366.

Beals, R., Krantz, D. H., & Tversky, A. Foundations of multidimensional scaling. *Psychological Review,* 1968, *75,* 127–142.

Boorman, S. A., & Arabie, P. Structural measures and the method of sorting. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences.* (Vol. 1): *Theory.* New York: Seminar Press, 1972.

Boorman, S. A., & Olivier, D. C. Metrics on spaces of finite trees. *Journal of Mathematical Psychology,* 1973, *10,* 26–59.

Boorman, S. A., & White, H. C. Social structure from multiple networks: II. Role structures. *American Journal of Sociology,* 1976, *81,* 1384–1446.

Boulton, D. M., & Wallace, C. S. An information measure for single-link classification. *Computer Journal,* 1975, *18,* 3, 236–238.

Boyd, J. P. Information distance for discrete structures. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences,* (Vol. 1): *Theory.* New York: Seminar Press, 1972.

Breiger, R. L., Boorman, S. A., & Arabie, P. An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology,* 1975, *12,* 328–383.

Bron, C., & Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph [H]. *Communications of the ACM,* 1973, *16,* 575–577.

Carroll, J. D. Spatial, non-spatial, and hybrid models for scaling. *Psychometrika,* 1976, *41,* 439–463.

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition. *Psychometrika,* 1970, *35,* 283–319.

Carroll, J. D., & Chang, J. J. A method for fitting a class of hierarchical tree structure models to dissimilarities data, and its application to some "body parts" data of Miller's. *Proceedings of the 81st Annual Convention of the American Psychological Association,* 1973, *8,* 1097–1098. (Summary)

Carroll, J. D., & Wish, M. Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.) *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Cole, A. J., & Wishart, D. An improved algorithm for the Jardine–Sibson method of generating overlapping clusters. *Computer Journal,* 1970, *13,* 156–163.

Coleman, J. S. *The adolescent society.* New York: Free Press, 1963.

Constantinescu, P. The classification of a set of elements with respect to a set of properties. *Computer Journal,* 1966, *8,* 352–357.

Coombs, C. H. *A theory of data.* New York: Wiley, 1964.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. *Psychological Bulletin,* 1953, *50,* 456–473.

Cross, D. V. Metric properties of multidimensional stimulus generalization. In D. I. Mostofsky (Ed.), *Stimulus generalization.* Stanford, Calif.: Stanford University Press, 1965.

Cunningham, J. P., & Shepard, R. N. Monotone mapping of similarities into a general metric space. *Journal of Mathematical Psychology,* 1974, *11,* 335–363.

DeWald, C. G., & Geyer, L. H. An operations research approach to the modeling and analysis of different feature sets proposed for human perception of capital letters. *Computers & Operations Research,* 1975, *2,* 61–70.

Diday, E. Optimization in non-hierarchical clustering. *Pattern Recognition,* 1974, *6,* 17–33.

Ekman, G. Dimensions of color vision. *Journal of Psychology,* 1954, *38,* 467–474.

Ekman, G. A direct method for multidimensional ratio scaling. *Psychometrika,* 1963, *28,* 33–41.

Estabrook, G. F. A mathematical model in graph theory for biological classification. *Journal of Theoretical Biology,* 1966, *12,* 297–310.

Fisher, L., & Van Ness, J. W. Admissible clustering procedures. *Biometrika,* 1971, *58,* 91–104.

Friendly, M. L. In search of the M-gram: The structure of organization in free recall. *Cognitive Psychology,* 1977, *9,* 188–249.

Garner, W. R. The stimulus in information processing. *American Psychologist,* 1970, *25,* 350–358.

Geyer, L. H., & DeWald, C. G. Feature lists and confusion matrices. *Perception & Psychophysics,* 1973, *14,* 471–482.

Gibson, E. J. *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts, 1969.

Gibson, E. J., & Levin, H. *The psychology of reading.* Cambridge, Mass.: MIT Press, 1975.

Goodman, N. Seven strictures on similarity. *Problems and projects.* Indianapolis, Ind.: Bobbs-Merrill, 1972.

Gower, J. C., & Ross, G. J. S. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics,* 1969, *18,* 54–64.

Green, P. E., Carroll, J. D., & DeSarbo, W. S. A new measure of predictor variable importance in multiple regression. *Journal of Marketing Research,* 1978, *15,* 356–360.

Guttman, L. Multiple group methods for common-factor analysis: Their basis, computation, and interpretation. *Psychometrika,* 1952, *17,* 209–222.

Hansen, P., & Delattre, M. Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association,* 1978, *73,* 397–403.

Hartigan, J. A. Representation of similarity matrices by trees. *Journal of the American Statistical Association,* 1967, *62,* 1140–1158.

Hartigan, J. A. *Clustering algorithms.* New York: Wiley, 1975.

Hartigan, J. A. Distribution problems in clustering. In J. Van Ryzin (Ed.), *Classification and clustering.* New York: Academic Press, 1977. (a)

Hartigan, J. A. Clustering as modes. *First international symposium on data analysis and informatics* (Vol. 2). Rocquencourt, France: Institut de Recherché d'Informatique et d'Automatique, 1977. (b)

Helm, C. E. Multidimensional ratio scaling analysis of perceived color relations. *Journal of the Optical Society of America,* 1964, *54,* 256–262.

Henley, N. M. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior,* 1969, *8,* 176–184.

Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics,* 1976, *32,* 1–49.

Holman, E. W. The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika,* 1972, *37,* 417–423.

Homans, G. C. *The human group.* New York: Harcourt, Brace, 1950.

Hubert, L. J. Monotone invariant clustering procedures. *Psychometrika,* 1973, *38,* 47–62.

Hubert, L. J. Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association,* 1974, *69,* 698–704. (a)

Hubert, L. J. Some applications of graph theory to clustering. *Psychometrika,* 1974, *39,* 283–309. (b)

Hubert, L. J., & Baker, F. B. An empirical comparison of baseline models for goodness-of-fit in *r*-diameter hierarchical clustering. In J. Van Ryzin (Ed.), *Classification and clustering.* New York: Academic Press, 1977. (a)

Hubert, L. J., & Baker, F. B. Analyzing distinctive features. *Journal of Educational Statistics,* 1977, *2,* 79–98. (b)

Hubert, L. J., & Schultz, J. Hierarchical clustering and the concept of space distortion. *British Journal of Mathematical & Statistical Psychology,* 1975, *28,* 121–133.

Jakobson, R., Fant, G. M., & Halle, M. *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, Mass.: MIT Press, 1963.

Jardine, C. J., Jardine, N., & Sibson, R. The structure and construction of taxonomic hierarchies. *Mathematical Biosciences,* 1967, *1,* 173–179.

Jardine, N., & Sibson, R. The construction of hierarchic and nonhierarchic classifications. *Computer Journal,* 1968, *11,* 177–184. (a)

Jardine, N., & Sibson, R. A model for taxonomy. *Mathematical Biosciences,* 1968, *2,* 465–482. (b)

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

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

Johnston, H. C. Cliques of a graph—Variations on the Bron-Kerbosch algorithm. *International Journal of Computer and Information Sciences,* 1976, *5,* 209–238.

Karp, R. M. Reducibility among combinatorial problems. In R. E. Miller & J. W. Thatcher (Eds.), *Complexity of computer computations*. New York: Plenum, 1972.

Karp, R. M. On the probabilistic analysis of some combinatorial search algorithms. In J. F. Traub (Ed.), *Algorithms and complexity*. New York: Academic Press, 1976.

Klatt, D. H. Structure of confusions in short-term memory between English consonants. *Journal of the Acoustical Society of America*, 1968, *44*, 401-407.

Koopmans, T. C. *Three essays on the state of economic science*. New York: McGraw-Hill, 1957.

Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 1956, *7*, 48–50.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1–27. (a)

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

Kruskal, J. B. Multidimensional scaling and other methods for discovering structure. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (Vol. 3). New York: Wiley–Interscience, 1977. (a)

Kruskal, J. B. The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and clustering*. New York: Academic Press, 1977. (b)

Kruskal, J. B., & Shepard, R. N. A nonmetric variety of linear factor analysis. *Psychometrika*, 1974, *39*, 123–157.

Kruskal, W. H. The geometry of generalized inverses. *Journal of the Royal Statistical Society* (Series B), 1975, *37*, 272–283.

Kuiper, F. K., & Fisher, L. A Monte Carlo comparison of six clustering procedures. *Biometrics*, 1975, *31*, 777–783.

LaBerge, D. Perceptual learning and attention. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 4). Hillsdale, N.J.: Erlbaum, 1976.

Lance, G. N., & Williams, W. T. A general theory of classificatory sorting strategies: I. Hierarchical systems. *Computer Journal*, 1967, *9*, 373–380. (a)

Lance, G. N., & Williams, W. T. A general theory of classificatory sorting strategies: II. Clustering systems. *Computer Journal*, 1967, *10*, 271–277. (b)

Levelt, W. J. M. Psychological representations of syntactic structures. In T. G. Brewer & W. Weksel (Eds.), *The structure and psychology of language*. New York: Holt, Rinehart & Winston, 1967.

Levelt, W. J. M. Hierarchical chunking in sentence processing. *Perception & Psychophysics*, 1970, *8*, 99–103.

Levine, J. H. The sphere of influence. *American Sociological Review*, 1972, *37*, 14–27.

Lew, J. S. Some counterexamples in multidimensional scaling. *Journal of Mathematical Psychology*, 1978, *17*, 247–254.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431–461.

Liberman, A. M., Delattre, P., & Cooper, F. S. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497–516.

Ling, R. F. On the theory and construction of k-clusters. *Computer Journal*, 1972, *15*, 326–332.

Ling, R. F. A probability theory of cluster analysis. *Journal of the American Statistical Association*, 1973, *68*, 159–164.

Lorrain, F. P., & White, H. C. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1971, *1*, 49–80.

Luce, R. D. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 1950, *15*, 169–190.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. (Vol. 1). Berkeley: University of California Press, 1967.

Miller, G. A. A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 1969, *6*, 169–191.

Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the structure of behavior*. New York: Holt, 1960.

Miller, G., & Nicely, P. E. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1955, *27*, 338–352.

Moon, J. W., & Moser, L. On cliques in graphs. *Israel Journal of Mathematics*, 1965, *3*, 23–28.

Nieminen, J. On finding maximum compatibles. *Proceedings of the IEEE*, 1975, *63*, 729–730.

Osteen, R. E. Clique detection algorithms based on line addition and line removal. *SIAM Journal of Applied Mathematics*, 1974, *26*, 126–135.

Peay, E. R. Hierarchical clique structures. *Sociometry*, 1974, *37*, 54–65.

Peay, E. R. Nonmetric grouping: Clusters and cliques. *Psychometrika*, 1975, *40*, 297–313.

Prim, R. C. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 1957, *36*, 1389–1401.

Rao, C. R., & Mitra, S. K. *Generalized inverse of matrices and its applications*. New York: Wiley, 1971.

Roethlisberger, F. J., & Dickson, W. J. *Management and the worker*. Cambridge, Mass.: Harvard University Press, 1939.

Rohlf, F. J. A new approach to the computation of the Jardine–Sibson B_k clusters. *Computer Journal*, 1975, *18*, 164–168.

Romney, A. K., & D'Andrade, R. G. (Eds.). Cognitive aspects of English kin terms. *Transcultural studies in cognition. American Anthropologist Special Issue*, 1964, *66*(3, Pt. 2).

Rosenberg, S., & Jones, R. A method for investigating and representing a person's implicit theory of personality: Theodore Dreiser's view of people. *Journal of Personality and Social Psychology,* 1972, *22,* 372–386.

Rumelhart, D. E., & Abrahamson, A. A. A model for analogical reasoning. *Cognitive Psychology,* 1973, *5,* 1–28.

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

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

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

Shepard, R. N. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology,* 1964, *1,* 54–87.

Shepard, R. N. Psychological representation of speech sounds. In E. E. David, Jr., & P. B. Denes (Eds.), *Human communication: A unified view.* New York: McGraw-Hill, 1972. (a)

Shepard, R. N. A taxonomy of some principal types of data and of multidimensional methods for their analysis. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 1): *Theory.* New York: Seminar Press, 1972. (b)

Shepard, R. N. Representation of structure in similarity data: Problems and prospects. *Psychometrika,* 1974, *39,* 373–421.

Shepard, R. N., & Carroll, J. D. Parametric representation of nonlinear data structures. In P. R. Krishnaiah (Ed.), *Multivariate analysis: Proceedings of an international symposium.* New York: Academic Press, 1966.

Shepard, R. N., & Cermak, G. W. Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology,* 1973, *4,* 351–377.

Shepard, R. N., & Chipman, S. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology,* 1970, *1,* 1–17.

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

Shoben, E. J. The verification of semantic relations in a same-different paradigm: An asymmetry in semantic memory. *Journal of Verbal Learning and Verbal Behavior,* 1976, *15,* 365–379.

Sibson, R. Order invariant methods for data analysis. *Journal of the Royal Statistical Society* (Series B), 1972, *34,* 311–349.

Sokal, R. R., & Michener, C. D. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin,* 1958, *38,* 1409–1438.

Sørenson, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter,* 1948, *5,* 1–34.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, N.J.: Erlbaum, 1977.

Theil, H. *Principles of econometrics.* New York: Wiley, 1971.

Torgerson, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

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

Tversky, A., & Krantz, D. H. Similarity of schematic faces: A test of inter-dimensional additivity. *Perception & Psychophysics,* 1969, *5,* 124–128.

Van Ness, J. W. Admissible clustering procedures. *Biometrika,* 1973, *60,* 422–424.

Wang, M. D., & Bilger, R. C. Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America,* 1973, *54,* 1248–1266.

Wexler, K. N., & Romney, A. K. Individual variations in cognitive structures. In A. K. Romney, R. N. Shepard, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 2): *Applications.* New York: Seminar Press, 1972.

White, H. C., Boorman, S. A., & Breiger, R. L. Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology,* 1976, *81,* 730–780.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America,* 1966, *39,* 388–398.

Williams, W. T., Lance, G. N., Dale, M. B., & Clifford, H. T. Controversy concerning the criteria for taxonometric strategies. *Computer Journal,* 1971, *14,* 162–165.

Wish, M., & Carroll, J. D. Applications of "INDSCAL" to studies of human perception and judgment. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception.* (Vol. 2). New York: Academic Press, 1974.

Received June 28, 1978
Revision received December 15, 1978 ■