RR 94-02

ITERATIVE PROJECTION STRATEGIES FOR THE LEAST-SQUARES FITTING OF PROXIMITY DATA

Lawrence Hubert
University of Illinois, Champaign
and
Department of Data Theory, University of Leiden

Phipps Arabie
Graduate Faculty of Management, Rutgers University, Newark

Abstract

A least-squares optimization strategy is first reviewed and applied to the task of fitting a given collection of symmetric proximity values defined between the objects from one set by a collection of reconstructed proximity values, satisfying a fixed set of constraints, generated from some specified graph-theoretic structure, such as an ultrametric or an additive tree, selected for representing the objects. Our method uses iterative projection onto closed convex sets defined by the collection of given constraints characterizing the structural representation specified, and in contrast to least-squares optimization methods that impose such constraints through the use of penalty functions, avoids the use of the latter, as well as the implementation of any gradient-based optimization technique. Secondly, just as various penalty-function/gradient-based optimization techniques have been turned into heuristic search strategies for such particular structures of interest as ultrametrics or additive trees, the use of iterative projection is suggested as a general heuristic search strategy for locating the best structural representations to impose in the first place, where the collection of constraints used may vary over the course of the optimization process. Our evaluation of the expected results uses several data sets previously analyzed in the literature. Finally, several other applications of iterative projection as a heuristic optimization technique are discussed, including the consideration of data beyond that of a single symmetric proximity matrix (for example, extensions to two-mode proximity matrices, i.e., between two distinct object sets, and to three-way proximity matrices either symmetric or not), and to representations based on sums of matrices where each is constrained separately to conform to some desired representational structure.

Key words: least-squares optimization, iterative projection, additive tree, ultrametric.

1. Introduction

Over the last several decades, numerous methods of data representation based on the use of various graph-theoretic structures have been developed for explaining the pattern of information potentially present in a single (or possibly, in a collection of) numerically given proximity matri(ces), each defined between pairs of objects from a single set (i.e., in the terminology of Tucker, 1964, a one-mode matrix), or in some cases, between the objects from several distinct sets (for example, see Carroll, 1976; Carroll, Clark, and DeSarbo, 1984; Carroll and Pruzansky, 1980; De Soete, 1983, 1984a, 1984b, 1984c; De Soete, Carroll, and DeSarbo, 1987; De Soete, DeSarbo, Furnas, and Carroll, 1984; Hutchinson, 1989; Klauer and Carroll, 1989, 1991). Typically, a specific class of graph-theoretic structures is assumed capable of representing the proximity information, and the proposed method seeks a member from the class producing a reconstructed set of proximities that are as close as possible to the original. The most prominent graph-theoretic structures to have been used are those usually referred to as ultrametrics and additive trees, but a number of other possibilities have also been considered, e.g., more general network models as in Klauer and Carroll (1989, 1991).

Although a variety of strategies have been proposed for locating good exemplars from whatever class of graph-theoretic structures is being considered, one approach has been to adopt a least-squares criterion in which the class exemplar is identified by attempting to minimize the sum of squared discrepancies between the original proximities and their reconstructions obtained through the use of the particular structure selected by the data analyst. Invariably, the least-squares optimization strategy implemented has been defined by the usual least-squares criterion but augmented by some collection of penalty functions that seek to impose whatever constraints are mandated by the structural representation being sought. Then, through the use of some unconstrained optimization scheme (e.g., steepest descent, conjugate gradients), an attempt is made to find both (a) the particular

constraints that should be imposed to define the specific structure from the class, and (b) the reconstructed proximities based on the structure finally identified. The resulting optimization strategy is heuristic in the sense that there is no guarantee of global optimality for the final structural representation identified even within the chosen graph-theoretic class, because the particular constraints defining the selected structure were located by a possibly reasonable but not verifiably optimal search strategy that was (implicitly) implemented in the course of the process of optimization.

The main purpose of the present paper is first to review and then eventually to extend heuristically a particular least-squares optimization strategy that in its non-heuristic form allows the reconstruction of a set of proximities based on a fixed collection of constraints implied by whatever specific graph-theoretic structure has been selected for their representation. The method discussed is relatively simple in design and uses successive (or iterative) projections onto closed convex sets defined by the collection of given constraints implied by the structural representation chosen, and thus, avoids the need for penalty terms; moreover, there is no explicit use of gradients in the attendant optimization strategy, and the strategy provides for the straightforward incorporation of a variety of different types of constraints that may be auxiliary to those generated from the given structural representation but nonetheless of interest to impose on the reconstruction. (For an extensive use of iterative projection methods in a different context than ours, the reader is referred to van der Lans, 1992, who applied iterative projection approaches to the nonlinear multivariate analysis of multiattribute preference data.)

As a least-squares optimization strategy (in a non-heuristic form), iterative projection assumes that whatever constraint set is to be applied is completely known prior to its application. However, just as the various penalty-function and gradient-optimization techniques have been turned into heuristic search strategies

for the particular structures of interest by allowing the collection of constraints to vary over the course of the optimization process, we attempt the same in the use of iterative projection. Thus, in addition to carrying out a least-squares task subject to given structural constraints, iterative projection will be considered as one possible heuristic search strategy (and an alternative to those heuristic methods that have been suggested in the literature and based exclusively on the use of some type of penalty functions) for locating the actual constraints to impose in the first instance, and therefore, to identify the general form of the structural representation sought.

The various least-squares optimization tasks entailing both the identification of the specific form of the structural representation to adopt and the subsequent least-squares fitting itself generally fall into the class of NP-hard problems (e.g., for ultrametric and additive trees, see Krivánek and Moravek, 1986; Krivánek, 1986; Day, 1987); thus, the best we can hope for is a heuristic extension of the iterative projection strategy leading to good but not necessarily optimal final structural representations within the general class of representations desired. As is standard with a reliance on such heuristic optimization methods, the use of multiple starting points will hopefully determine a set of local optima characterizing the better solutions attainable for a given data set. The presence of local optima in the use of any heuristic and combinatorially-based optimization strategy is unavoidable, given the NP-hardness of the basic optimization tasks of interest and the general inability of (partial) enumeration methods (when available) to be computationally feasible for use on even moderate-sized data sets. The number of and variation in the local optima observable for any specific situation will obviously depend on the given data, the structural representation sought, and the heuristic search strategy used. But whenever present, local optima may actually be diagnostic for the structure(s) potentially appropriate for characterizing a particular data set. Thus, their identification may even be valuable in explaining

the patterning of the data and/or in noting the difficulties with adopting a specific representational form to help discern underlying structure.

The organization of this paper is as follows: Section 2 reviews the least-squares optimization strategy, based on iterative projection onto closed convex sets defined by whatever constraints are of interest, and presents a number of examples of constraint sets for the particular graph-theoretic structures defined by ultrametric and additive trees, fitted to a single symmetric proximity matrix. Section 3 considers the use of iterative projection as a heuristic optimization strategy in the same context of locating ultrametric and additive trees based on a single symmetric proximity matrix, but where the specific constraint sets are not known a priori; results are discussed explicitly for two published data sets used elsewhere to illustrate various strategies of data representation. Finally, Section 4 discusses several other applications of iterative projection as a heuristic optimization technique, including the consideration of data beyond that of a (single) symmetric proximity matrix (for example, extensions to two-mode proximity matrices, i.e., between two distinct object sets, and to three-way proximity matrices either symmetric or not), and to representations based on sums of matrices where each is constrained separately to conform to some desired representational structure.

2. Least-Squares Optimization Using Iterative Projection

The general form of our least-squares optimization task can be given as follows (where R^n denotes the set of all $n \times 1$ real vectors, and n will eventually indicate the number of distinct proximity values to be fitted in a particular application):

For a given vector $\mathbf{p} \in R^n$, obtain the vector $\mathbf{x} \in R^n$, say $\mathbf{x}^*$, that minimizes

$$(\mathbf{p} - \mathbf{x})'(\mathbf{p} - \mathbf{x}) \ , \tag{1}$$

where $\mathbf{x} \in C_1 \cap C_2 \cap \ldots \cap C_K$, and each C_k for $1 \leq k \leq K$ is a closed convex subset of

R^n. Here, the term convex implies that if $\mathbf{x}, \mathbf{y} \in C_k$, then $a\mathbf{x} + (1-a)\mathbf{y} \in C_k$ for $0 \leq a \leq 1$, and closed implies (intuitively) that C_k contains all its boundary vectors. Generally, for the particular applications envisioned, the vector $\mathbf{p}$ will contain the given n proximities available between the pairs of objects from one set or possibly between two distinct sets, and the subsets $C_1,\ldots,C_K$ will define the individual constraints necessary for the specific structure being fitted so that the solution vector $\mathbf{x}^*$ gives the reconstructed proximities from that structure. (We might note that loss-functions more general than (1) could be considered having the forms $(\mathbf{p} - \mathbf{x})'\mathbf{Q}(\mathbf{p} - \mathbf{x})$ for a given $n \times n$ positive semi-definite matrix $\mathbf{Q}$, or $(\mathbf{p} - \mathbf{Fx})'\mathbf{Q}(\mathbf{p} - \mathbf{Fx})$ for some given $n \times m$ matrix $\mathbf{F}$, where m now denotes the size of the vector $\mathbf{x}$ that must be identified to produce the $n \times 1$ vector $\mathbf{Fx}$ fitted to $\mathbf{p}$; but in both cases, reductions to (1) are possible through the singular value decompositions of $\mathbf{Q}$ and $\mathbf{F}$, and appropriate redefinitions of the vectors and convex subsets involved. For our purposes, the simpler form of (1) will suffice. For examples of where such reductions have been carried out, see Dykstra, 1983, p. 840, or ten Berge, 1991.)

Assuming that a strategy is available for solving the projection subproblem:

minimize $(\mathbf{d} - \mathbf{x})'(\mathbf{d} - \mathbf{x})$ over $\mathbf{x} \in C_k$ for some given $\mathbf{d} \in R^n$ and k, $1 \leq k \leq K$,

the solution of the optimization task in (1) may be obtained by iterative projection as follows:

(i) starting with $\mathbf{f}_0^{(1)} = \mathbf{p}$, let $\mathbf{f}_k^{(1)}$ be the projection of $\mathbf{f}_{k-1}^{(1)}$ onto C_k, $1 \leq k \leq K$ (and let $\mathbf{e}_k^{(1)} = \mathbf{f}_k^{(1)} - \mathbf{f}_{k-1}^{(1)}$ be the vector of differences between $\mathbf{f}_{k-1}^{(1)}$ and its projection onto C_k);

(ii) for $t = 2,\ldots,$ and starting with $\mathbf{f}_0^{(t)} = \mathbf{f}_K^{(t-1)}$,

let $\mathbf{f}_k^{(t)}$ be the projection of $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$ onto C_k, $1 \leq k \leq K$, and $\mathbf{e}_k^{(t)} = \mathbf{f}_k^{(t)} - (\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)})$.

As $t \to \infty$, $\mathbf{f}_K^{(t)}$ converges to $\mathbf{x}^*$, the solution of (1).

In words, we start with $\mathbf{p}$ and successively project onto the sets C_k, for $1 \le k \le K$ in sequence, and cyclically reconsider the sets C_k until convergence, but each time a set C_k is reconsidered, the changes from the last projection, $\mathbf{e}_k^{(t-1)}$, are first subtracted from the current vector $\mathbf{f}_{k-1}^{(t)}$ before the projection onto C_k is carried out. When it is necessary to indicate explicitly that each projection onto C_k is preceded by a subtraction of the changes from the previous projection onto the same set, we will refer to the strategy as iterative projection <u>with augmentation</u>.

Proofs of the convergence of iterative projection with augmentation to the optimal solution $\mathbf{x}^*$ when the C_k are closed convex sets are given, for example, by Boyle and Dykstra (1986), Han (1988), and Gaffke and Mathar (1989); for closed convex cones ($\mathbf{x}, \mathbf{y} \in C_k$ implies $a\mathbf{x} + b\mathbf{y} \in C_k$ for $a, b \ge 0$), see Dykstra (1983); and for subspaces of R^n ($\mathbf{x}, \mathbf{y} \in C_k$ implies $a\mathbf{x} + b\mathbf{y} \in C_k$ for all a, b), see Wiener (1955) and von Neumann (1950). In this last case requiring subspaces, subtraction of the changes $\mathbf{e}_k^{(t-1)}$ is unnecessary to insure convergence to $\mathbf{x}^*$. In general, the use of iterative projection without augmentation, and thus, without the subtraction of the changes $\mathbf{e}_k^{(t-1)}$ prior to a projection onto C_k still leads to a vector within the closed convex set $C_1 \cap \ldots \cap C_K$ (see, for example, Cheney and Goldstein, 1959), but one that is not necessarily optimal with respect to the loss function in (1). Counterexamples are given by Han (1988).

Several strategies for constructing the sets $C_1, \ldots, C_K$ are possible, but for the present we consider only those definable as linear constraints, where

$$C_k = \{\mathbf{x} \in R^n \mid \mathbf{a}_k'\mathbf{x} \le b_k\} , \tag{2}$$

for some given $n \times 1$ vector $\mathbf{a}_k$ and scalar b_k. Given the form of C_k, the projection

subproblem of minimizing $(\mathbf{d} - \mathbf{x})'(\mathbf{d} - \mathbf{x})$ over $\mathbf{x} \in C_k$ for some fixed $\mathbf{d} \in R^n$ has the explicit solution (e.g., see Han, 1988):

$$\mathbf{x}^* = \mathbf{d} - (\mathbf{a}_k'\mathbf{a}_k)^{-1}[\max(\mathbf{a}_k'\mathbf{d} - b_k, 0)]\mathbf{a}_k \ . \tag{3}$$

This latter projection result is the basis for fitting the various kinds of fixed sets of constraints relevant to the reconstruction of the proximities based on a particular graph-theoretic structure. We give a few simple examples below of what the projection $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)'$ of $\mathbf{d} = (d_1, \ldots, d_n)'$ would be for several general categories of constraints (that would define a closed convex set, denoted generically as C) when the constraint is violated by $\mathbf{d}$ (otherwise, the projection $\mathbf{x}^*$ is just equal to $\mathbf{d}$).

(a) An order constraint on two entries: $x_u \leq x_v$ for some u, v —

$C = \{\mathbf{x} \in R^n \mid (0 \ldots 1 \ldots -1 \ldots 0)\mathbf{x} \leq 0\}$, where the 1 appears in position u and the −1 in position v.

$$\text{If } d_v > d_u \ , \ x_i^* = \begin{cases} d_i & \text{for } i \neq u \text{ or } v; \\ (d_u + d_v)/2 & \text{for } i = u \text{ or } v. \end{cases}$$

(b) An order constraint on the sum of two entries and a third: $x_u \leq x_v + x_w$ for some distinct u, v, w —

$C = \{\mathbf{x} \in R^n \mid (0 \ldots 1 \ldots -1 \ldots -1 \ldots 0)\mathbf{x} \leq 0\}$, where the 1 appears in position u and the −1's appear in positions v and w.

$$\text{If } d_u > d_v + d_w \ , \ x_i^* = \begin{cases} d_i & \text{for } i \neq u, v, \text{ or } w; \\ d_u - (1/3)(d_u - (d_v+d_w)) & \text{for } i = u; \\ d_v + (1/3)(d_u - (d_v+d_w)) & \text{for } i = v; \\ d_w + (1/3)(d_u - (d_v+d_w)) & \text{for } i = w. \end{cases}$$

(c) An order constraint on two sums: $x_u + x_v \leq x_u' + x_v'$ for some distinct

u, v, u′, v′ —

$C = \{\mathbf{x} \in R^n \mid (0\ldots1\ldots1\ldots-1\ldots-1\ldots0)\mathbf{x} \leq 0\}$, where the 1's appear in positions u and v and the −1's appear in positions u′ and v′.

If $d_u + d_v > d_u' + d_v'$,

$$x_i^* = \begin{cases} d_i & \text{for } i \neq u, v, u', \text{ or } v' \\ d_u - (1/4)(d_u+d_v - (d_u'+d_v')) & \text{for } i = u \\ d_v - (1/4)(d_u+d_v - (d_u'+d_v')) & \text{for } i = v \\ d_u' + (1/4)(d_u+d_v - (d_u'+d_v')) & \text{for } i = u' \\ d_v' + (1/4)(d_u+d_v - (d_u'+d_v')) & \text{for } i = v'. \end{cases}$$

(d) An upper bound constraint on one entry: $x_u \leq b$ for some u —

$C = \{\mathbf{x} \in R^n \mid (0\ldots1\ldots0)\mathbf{x} \leq b\}$, where the 1 appears in position u.

If $d_u > b$, $x_i^* = \begin{cases} d_i & \text{for } i \neq u; \\ b & \text{for } i = u. \end{cases}$

(A lower bound constraint would be handled analogously with the use of a single −1 in the constraint vector $\mathbf{a}$, i.e., for the constraint of $g \leq x_u$ for some u and constant g, and if $d_u < g$, then $x_i^* = d_i$ for $i \neq u$, and $= g$ for $i = u$.) In all cases of (a), (b), (c), and (d), a corresponding equality (as opposed to an inequality) constraint could be handled by merely imposing the inequality constraint in both directions, e.g., a constraint of $x_u = x_v$ would be separated into $x_u \leq x_v$ and $x_v \leq x_u$.

As one additional interpretative point regarding the choice of constraint sets of the form in (2), if each scalar b_k is 0 and the sum of entries in each $\mathbf{a}_k$ is also 0, $1 \leq k \leq K$ (as given, for example, in (a) and (c) above), the solution vector $\mathbf{x}^*$ minimizing the least-squares criterion in (1) also maximizes the square of the correlation between $\mathbf{p}$ and a vector $\mathbf{x} \in C_1 \cap \ldots \cap C_K$ (see Robertson, Wright, and

Dykstra, 1988, p. 379). Thus, in this case a natural "variance-accounted-for" measure (denoted as VAF) can be given simply as the square of the latter correlation, and we will do so below when appropriate. Moreover, for this restricted set of constraints, if $\mathbf{x}^*$ is the solution of (1) for given $\mathbf{p}$, then $a\mathbf{x}^* + b$ provides the solution when beginning with $a\mathbf{p} + b$ (for $a > 0$), i.e., under interval scale transformations of the entries in $\mathbf{p}$, the solution vector is identically transformed.

2.1 Some Illustrations Based on Ultrametrics and Additive Trees Fit to a Single Proximity Matrix

In the following examples, we consider for now only the case in which the available proximity data are on one set of N objects, $S = (O_1,\ldots,O_N)$, and are given in the form of an $N \times N$ symmetric matrix $P = \{P_{rs}\}$ (see Section 4 for extensions and alternatives). For convenience, the entries in P are assumed keyed as dissimilarities (i.e., large proximity values reflect dissimilar object pairs), and for O_r, $O_s \in S$, $P_{rs} \geq 0$ with equality holding only if $r = s$. The $n \times 1$ vector $\mathbf{p}$ to be fitted by the $n \times 1$ vector $\mathbf{x}^*$, where $\mathbf{x}^*$ belongs to the closed convex set $C_1 \cap \ldots \cap C_K$, is formed from the $n = N(N-1)/2$ (upper-triangular) entries in P. (Throughout the following discussion, the various stated results characterizing the conditions required for specific representations find much more complete presentations in Barthélemy and Guénouche, 1991 (Chapters 2 & 3), and the reader is referred to this comprehensive source for the necessary background information.)

Hierarchical classification (ultrametrics). A hierarchical classification of the object set S can be defined as a sequence of H+1 partitions of S, $\Pi_0,\Pi_1,\ldots,\Pi_H$, where (1) each Π_h contains a set of subsets of S that define a mutually exclusive and exhaustive partitioning of S, (2) Π_0 contains N classes each defined by a single object from S; Π_H contains a single class containing all the objects in S, (c) Π_{h+1}

is formed from Π_h, $0 \leq h \leq H-1$, by combining one or more classes in Π_h. The task of any agglomerative hierarchical clustering algorithm is to construct the partition sequence, $\Pi_0,\ldots,\Pi_H$, from the given proximity matrix $\mathbf{P}$, where an attempt is made through some optimization mechanism to place similar objects into common classes early in the sequence. As an alternative representational device connected with a given partition sequence, $\Pi_0,\ldots,\Pi_H$, we may define a class of $N \times N$ matrices, $\Xi(\Pi_0,\ldots,\Pi_H)$, where $\mathbf{U}_f = \{u^f_{rs}\} \in \Xi(\Pi_0,\ldots,\Pi_H)$ if $u^f_{rs} = \min\{ f(h) \mid O_r \text{ and } O_s \text{ appear in a common class in } \Pi_h\}$, for some (monotone nondecreasing) function $f(\cdot)$ (i.e., for $h \leq h'$, $f(h) \leq f(h')$, where $f(0) = 0$. Each of the matrices in $\Xi(\Pi_0,\ldots,\Pi_H)$ satisfies the four properties of an ultrametric: for O_r, $O_s \in S$, (1) $u^f_{rs} = u^f_{sr}$; (2) $u^f_{rs} \geq 0$ and $u_{rr} = 0$; (3) $u^f_{rs} = 0$ implies $O_r = O_s$; (4) $u^f_{rs} \leq \max\{u^f_{rz}, u^f_{zs}\}$ for $O_z \in S$ (or equivalently, for any object triple O_r, O_s, $O_z \in S$, the two larger values among u^f_{rs}, u^f_{rz}, and u^f_{zs} are equal). Given any matrix $\mathbf{U}_f \in \Xi(\Pi_0,\ldots,\Pi_H)$, a partition squence $\Pi_0,\ldots,\Pi_{h'},\ldots,\Pi_H$ may be retrieved, where object pairs O_r, $O_s \in S$ contained within a common class in $\Pi_{h'}$ are those for which $u^f_{rs} \leq f(h')$, and the partitions $\Pi_0,\ldots,\Pi_{h'},\ldots,\Pi_H$ are (at least) a subset of the original partitions used to define $\Xi(\Pi_0,\ldots,\Pi_H)$.

The least-squares task of fitting a vector $\mathbf{x}^*$ to the values in the proximity matrix $\mathbf{P}$ based on the constraints implied by a given partition hierarchy $\Pi_0,\ldots,\Pi_H$, is in effect the task of finding that member of $\Xi(\Pi_0,\ldots,\Pi_H)$ closest to $\mathbf{P}$ in a least-squares sense. Specifically, we wish to minimize $(\mathbf{p} - \mathbf{x})'(\mathbf{p} - \mathbf{x})$, where (a) $x_i \geq 0$ for $1 \leq i \leq n$; (b) for each distinct object triple and the three entries in $\mathbf{x}$ corresponding to the proximities between these three pairs of objects, say x_u, x_v, and x_w, four inequalities hold depending on the structure of $\Xi(\Pi_0,\ldots,\Pi_H)$, e.g., if x_v and x_w should be the two larger entries that are equal, then: $x_u \leq x_v$; $x_u \leq x_w$; $x_v \leq x_w$; $x_w \leq x_v$; and (c) for each pair of entries in $\mathbf{x}$ corresponding to the proximities for two object pairs not sharing a common object, say x_u and x_v, either

$x_u \leq x_v$ or $x_v \leq x_u$ depending on the structure of $\Xi(\Pi_0,\ldots,\Pi_H)$. Thus, for the constraints in (b), four closed convex sets are defined for each distinct object triple; for those in (a), n closed convex sets are defined; for those in (c) $N(N-1)(N-2)(N-3)/8$ closed convex sets are defined. The optimal solution vector $\mathbf{x}^*$ can be obtained by iterative projection onto the closed convex sets defined in (a), (b), and (c). (We note two points in connection with these sets of constraints: first, the constraints in (a) can actually be ignored because starting with $\mathbf{p}$, whose entries have been assumed positive, an iterative projection sequence using just the constraints in (b) and (c) converges to $\mathbf{x}^*$, whose entries must all be at least as large as the smallest entry in $\mathbf{p}$. Thus, the constraints in (a) are satisfied by $\mathbf{x}^*$, and moreover, since no zero values can occur in $\mathbf{x}^*$, condition (3) for an ultrametric automatically holds [i.e., zero values only correspond to identical objects]. Secondly, if the order constraints in (c) are not imposed, the optimal vector $\mathbf{x}^*$ will still be an ultrametric but one not necessarily consistent with the order in which new subsets were formed in the hierarchy $\Pi_0,\ldots,\Pi_H$. The vector $\mathbf{x}^*$ will correspond to an ultrametric defined by exactly the same subsets as in $\Pi_0,\ldots,\Pi_H$, but the order in which they are generated may be different.)

A numerical example. To illustrate the least-squares fit to a given proximity matrix $\mathbf{P}$ by a representative from a class $\Xi(\Pi_0,\ldots,\Pi_H)$, we consider a rather well-known proximity matrix from Rao (1952, p. 361), reproduced in the upper-triangular portion of Table 1 using the labels and abbreviations from Rao. The proximities are on 12 Indian castes and tribes and are squared Mahalanobis distances based on average group values for nine physical measurements (i.e., head length, head breadth, bizygomatic breadth, nasal height, nasal breadth, nasal depth, stature, sitting height, and frontal breadth). The fixed classification that we will consider is given below using the numerical labels attached to the 12 groups

from Table 1 (for now, we assume the hierarchical classification is fixed but its origin will become clear in Section 3):

Level	Partition (only those classes with more than one object are listed)
0	all objects separate
1	{10,11}
2	{10,11},{1,2}
3	{9,10,11},{1,2}
4	{9,10,11},{1,2},{7,8}
5	{9,10,11,12},{1,2},{7,8}
6	{9,10,11,12},{1,2},{7,8},{5,6}
7	{1,2,9,10,11,12},{7,8},{5,6}
8	{1,2,4,9,10,11,12},{7,8},{5,6}
9	{1,2,4,7,8,9,10,11,12},{5,6}
10	{1,2,4,5,6,7,8,9,10,11,12}
11	all objects together

The vector **p** containing the proximities is of size 66 × 1, and the solution vector **x*** is represented in the lower-triangular portion of Table 1, and was obtained by iterative projection starting from **p** and using the [(12×11×10)/6]×4 + [(12×11×10×9)/12]×3 = 2285 closed convex sets defined by the four constraints imposed for each distinct object triple and the three constraints imposed for each distinct object quadruple (the iterative process was terminated when the sum of the absolute values of the differences between the projections from one complete cycle through the constraints to the next was less than 10^{-4}; this latter value is the subthreshold change used generally throughout the paper). A VAF of 56.16% was obtained from the square of the correlation between **p** and **x***.

{Insert Table 1 and Figure 1 here}

It is possible to give a graph-theoretic interpretation, as in Figure 1, to the ultrametric in the lower-half of Table 1 that will lead naturally into our discussion below of a more general graph-theoretic representation for a proximity matrix **P** through what will be referred to as additive trees. Figure 1 includes 12 terminal nodes corresponding to the 12 original objects, and 10 internal nodes corresponding to each of the new groups formed (except for the complete set S at level 11) in the progression from Π_0 to Π_H in the hierarchical sequence. The numerical values (lengths) attached to each of the horizontal branches in Figure 1 (all vertical lines are for pictorial convenience only and are assumed to have length zero) serve to reconstruct the ultrametric in Table 1 by merely summing branch lengths on the unique path joining two terminal nodes. In general, a graph-theoretic structure such as in Figure 1, where proximities are approximated by the sum of branch lengths in the unique path joining two terminal nodes, is referred to as a representation by an additive tree. In the case of an ultrametric, the particular additive tree is restricted in the sense that there exists a location on the tree (often called the "root" and indicated by an open box in Figure 1) from which the sum of the branch lengths to each of the terminal nodes is constant.

Additive trees (ternary). The fitting of a proximity matrix **P** by a given additive tree structure obviously requires the initial specification of the particular tree to be used to obtain the constraints. For the moment, we consider only (ternary) trees like that given in Figure 1, defined by N terminal nodes (each with one branch attached) corresponding to the N objects in S and N-2 internal nodes (each with three branches attached), and thus, with a total of 2N-3 branches. (Having fewer internal nodes than N-2 necessitates the imposition of additional constraints, and a few examples will be presented below.) Given a specific ternary tree, say T_0, suppose we define a class of $N \times N$ matrices (each with an assumed zero main

diagonal), $\Gamma(T_0)$, where $T_{\mathcal{L}} = \{\tau^{\mathcal{L}}_{rs}\} \in \Gamma(T_0)$, if $\tau^{\mathcal{L}}_{rs}$ = sum of those values from the collection $\{\ell_1,\ldots,\ell_{2N-3}\}$ on the unique path in T_0 between O_r and O_s, where $\mathcal{L} = \{\ell_1,\ldots,\ell_{2N-3}\}$ is a collection of scalar values attached to each of the 2N–3 branches (which for now are unrestricted as to sign). Each of the matrices in $\Gamma(T_0)$ satisfies the following property: for each quadruple of distinct objects O_r, O_s, O_w, $O_z \in S$, among the three sums, $\tau^{\mathcal{L}}_{rs} + \tau^{\mathcal{L}}_{wz}$, $\tau^{\mathcal{L}}_{rw} + \tau^{\mathcal{L}}_{sz}$, and $\tau^{\mathcal{L}}_{rz} + \tau^{\mathcal{L}}_{sw}$, two are equal, and the same two sums are equal over all choices of $\ell_1,\ldots,\ell_{2N-3}$ (i.e., over all matrices in $\Gamma(T_0)$). Given a matrix $T_{\mathcal{L}} \in \Gamma(T_0)$, a ternary tree may be retrieved as well as the values for the scalars that would serve to reconstruct the matrix $T_{\mathcal{L}}$ (for example, see the constructive proof in Patrinos and Hakimi, 1972, or the method of closest predecessors in Barthélemy and Guénouche, 1991, pp. 70–73).

The least–squares task of fitting a vector $\mathbf{x}^*$ to the proximity vector $\mathbf{p}$ based on constraints implied by a ternary tree T_0 (and with no sign restriction as to the branch scalars $\ell_1,\ldots,\ell_{2N-3}$) reduces to finding that member of $\Gamma(T_0)$ closest to $\mathbf{P}$ in a least–squares sense. Explicitly, $(\mathbf{p} - \mathbf{x})'(\mathbf{p} - \mathbf{x})$ is minimized, where for each distinct object quadruple and the six corresponding entries in x, say x_u, x_v, x_w, x_u', x_v', and x_w', the structure of $\Gamma(T_0)$ implies that for some specific u, v, u′, and v′, the equality $x_u + x_v = x_u' + x_v'$ must hold. Thus, each distinct object quadruple defines two closed convex sets, $x_u + x_v \leq x_u' + x_v'$ and $x_u' + x_v' \leq x_u + x_v$, and the optimal solution vector $\mathbf{x}^*$ can be obtained by iterative projection.

The class of matrices $\Gamma(T_0)$ allows the scalars $\ell_1,\ldots,\ell_{2N-3}$ to be unrestricted in sign, and given the general difficulty associated with the possible substantive interpretation of what a negative branch scalar might mean in a particular representation, two subclasses of $\Gamma(T_0)$ are of interest for further restricting the optimal solution: to $\Gamma(T_0)_+$ where all branch scalars not attached to terminal nodes are nonnegative, and to $\Gamma(T_0)_{++}$ where all branch scalars are nonnegative. For each of these subclasses, the member matrices are characterized by a property that in

turn leads to a collection of constraints defined by closed convex sets; thus, optimal solution vectors $\mathbf{x}^*$ within these subclasses of $\Gamma(T_0)$ can again be obtained by iterative projection. Specifically, for $\Gamma(T_0)_+$:

for each quadruple of distinct objects, O_r , O_s , O_w , $O_z \in S$, among the three sums, $\overset{g}{\tau}_{rs} + \overset{g}{\tau}_{wz}$, $\overset{g}{\tau}_{rw} + \overset{g}{\tau}_{sz}$, and $\overset{g}{\tau}_{rz} + \overset{g}{\tau}_{sw}$, two of the sums are equal and not less than the third. (4)

for $\Gamma(T_0)_{++}$ (in addition to the property for $\Gamma(T_0)_+$):

for each triple of distinct objects O_r , O_s , $O_w \in S$, $\overset{g}{\tau}_{rs} \leq \overset{g}{\tau}_{rw} + \overset{g}{\tau}_{sw}$, $\overset{g}{\tau}_{rw} \leq \overset{g}{\tau}_{rs} + \overset{g}{\tau}_{sw}$, and $\overset{g}{\tau}_{sw} \leq \overset{g}{\tau}_{sr} + \overset{g}{\tau}_{rw}$ (i.e., the triangle inequalities hold). (5)

With $\Gamma(T_0)_{++}$ the usual interpretation of nonnegative branch lengths apply, where the $\overset{g}{\tau}_{rs}$ are merely constructed as the sum of nonnegative branch lengths in the unique path in T_0 between O_r and O_s.

Given the type of constraints imposed to obtain a solution within $\Gamma(T_0)$ or $\Gamma(T_0)_+$, where the use of positive linear transformations of $\mathbf{p}$ would be apparent in the same linear transformation of $\mathbf{x}^*$, the natural VAF measure defined as the square of the correlation between $\mathbf{p}$ and $\mathbf{x}^*$ is appropriate. For $\Gamma(T_0)_{++}$, such an invariance does not exist in general. However, for a given vector $\mathbf{p}$, there is always a sufficiently large positive constant, say B_p, such that beginning with $\mathbf{p} + b$, for $b \geq B_p$ as the vector of proximities to be fitted, an optimal solution within $\Gamma(T_0)_+$ must also be within $\Gamma(T_0)_{++}$ (i.e., the triangle inequalities will be automatically satisfied by the solution vector within $\Gamma(T_0)_+$). Since we generally consider a proximity measure with interval scale characteristics to be arbitrary up to a positive linear transformation, we will impose only those constraints necessary for a representation in $\Gamma(T_0)_+$, with the understanding that it may be necessary to add a constant to $\mathbf{p}$ to produce a representation that is also within $\Gamma(T_0)_{++}$.

Numerical example. To illustrate the least-squares fit to a given additive tree, we

again use the data of Table 1 and consider the specific form of the additive tree to be that in Figure 1. An optimal solution $\mathbf{x}^*$ within $\Gamma(T_0)_+$ based on iterative projection onto the $4\times([12\times11\times10\times9]/12) = 3960$ convex sets defined by the four inequalities derived for each distinct object quadruple is given in Table 2, and the additive tree with branch scalars attached is given in Figure 2. A substantial number of branches originally present in Figure 1 are now given zero scalar values and are therefore suppressed in Figure 2. In addition, there are three negative terminal branch lengths in Figure 2 (with values of −.030, −.048, and −.286); thus, an additive constant of at least .286 would have to be added to the entries of $\mathbf{p}$ to produce a solution within $\Gamma(T_0)_{++}$ as well. (The actual branch lengths in Figure 2 connecting terminal nodes are proportional to the given values but first augmented by an additive constant of .500.) Using $\mathbf{x}^*$ within $\Gamma(T_0)_+$, the VAF of 76.62% reflects over a 20% increase from the more restricted ultrametric representation given earlier and based upon the same ternary tree T_0.

{Table 2 and Figure 2 here}

Additive trees (nonternary). Completeness in this review requires mentioning the fitting of a proximity matrix $\mathbf{P}$ by a given additive tree defined by N terminal nodes but less than N−2 internal nodes, and, say, I branches. The iterative projection strategy would proceed much as for a ternary tree, but based on the structure of T, additional constraints will need to be imposed for certain object quadruples and/or object triples. As before, a class of $N \times N$ matrices (each with an assumed zero main diagonal) is defined and then denoted by $\Gamma(T)$, using a collection of I scalars (as yet unrestricted in sign), $L = \{\ell_1,\ldots,\ell_I\}$, attached to the I branches of T, and again the least-squares task is to find a member of $\Gamma(T)$ closest to $\mathbf{P}$. From a slightly different perspective, any additive tree, T, can be embedded in a ternary additive tree T_0' (not necessarily unique), where 2N−3−I of the latter branch lengths are zero. Thus, finding a member of $\Gamma(T)$ closest to $\mathbf{P}$ is equivalent to locating a

restricted member of $\Gamma(T_0')$ closest to $\mathbf{P}$, where that member of $\Gamma(T_0')$ is constructed from a scalar collection $\{\ell_1,\ldots,\ell_{2N-3}\}$ in which 2N−3−I values are forced to be zero.

To obtain the actual member of $\Gamma(T)$ closest to $\mathbf{P}$, we need to determine from the structure of T which additional constraints may be necessary to impose for each distinct object quadruple and for each distinct object triple. Specifically, if we consider a particular object quadruple O_r , O_s , O_w , $O_z \in S$ (and because some scalar attached to the branch between two internal nodes in T_0' must be 0), all three sums, $\overset{\mathbf{g}}{r}_{rs} + \overset{\mathbf{g}}{r}_{wz}$, $\overset{\mathbf{g}}{r}_{rw} + \overset{\mathbf{g}}{r}_{sz}$, and $\overset{\mathbf{g}}{r}_{rz} + \overset{\mathbf{g}}{r}_{sw}$, may be required to be equal. Similarly, if we consider a particular object triple O_r , O_s , $O_w \in S$ (and because some scalar attached to a branch ending at a terminal node in T_0' must be 0), among the three terms, $\overset{\mathbf{g}}{r}_{rs}$, $\overset{\mathbf{g}}{r}_{rw}$, and $\overset{\mathbf{g}}{r}_{sw}$, the sum of two may be required to equal the third. Further nonnegativity restrictions, if desired for $\Gamma(T)_+$ or $\Gamma(T)_{++}$, may be implemented analogously as for ternary additive trees. Unfortunately, if the structure of T is such that in certain object triples, the sum of two of the values in the optimal vector $\mathbf{x}^*$ is forced to equal a third, a general lack of invariance to linear transformations of $\mathbf{p}$ exists in the solution (and the device of using a sufficiently large additive constant as before will no longer suffice).

3. Iterative Projection as a Heuristic Optimization Strategy

The most obvious modification of iterative projection that would try to turn it into a heuristic optimization search strategy proceeds as follows:

Instead of defining a sequence $\mathbf{f}_k^{(t)}$ through the projection of $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$ onto a fixed closed convex set C_k, $1 \le k \le K$, constructed from the given structure to be fitted, the set C_k would be selected adaptively from a consideration of the entries in $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$, and thus, C_k could vary over t, the index for iterations. For example, in attempting to locate good approximating ultrametrics and additive trees, we might operationalize this procedure as follows:

Ultrametrics: Each C_k, $1 \leq k \leq K$, is associated with a triple of distinct objects, which could be any one of three closed convex sets defined by the corresponding entries x_u , x_v , and x_w , i.e., $x_u \leq x_v = x_w$, $x_v \leq x_u = x_w$, or $x_w \leq x_u = x_v$, where again u, v, and w refer to the positions in the vector **p** of the proximities between the pairs of three distinct objects. Then, depending the current values for the u, v, and w entries in $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$, one of the three possible sets for C_k is chosen, i.e., if among the u, v, and w entries in $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$, those for u and w are the largest, choose C_k defined by the condition $x_v \leq x_u = x_w$, and similarly for the other two patterns that might be observed. (We note that in the earlier discussion of finding a closest ultrametric for a given hierarchy, each distinct object triple corresponded to four convex sets defining the four inequalities necessary for the ultrametric inequality to hold for that particular triple; here, we merely use these four inequalities jointly [again giving a closed convex set] in characterizing the three possibilities present for a set C_k. Also, since we are searching for a best-fitting ultrametric and not merely fitting a given one, the constraints corresponding to proximities attached to distinct object quadruples can be ignored.)

Additive trees (within $\Gamma(T_0)_+$): Each C_k is now associated with a quadruple of distinct objects, which could be any one of three closed convex sets defined by the six corresponding entries x_u , x_v , x_w , $x_{u'}$, $x_{v'}$, and $x_{w'}$, where u, v, w, u′ , v′ , and w′ again refer to the positions of the proximities between the four distinct objects in the vector **p**. Each of the closed convex sets is defined by two of the pairwise sums being equal and no smaller than the third. And as for ultrametrics, the choice for C_k in the course of the iterative process is based on the current values for the u, v, w, u′ , v′ , and w′ entries $\mathbf{f}_{k-1}^{(t)} - \mathbf{e}_k^{(t-1)}$. In general, if such a heuristic iterative projection strategy is implemented, as for example in identifying ultrametrics and additive trees fitted to a given vector **p**, two obvious

questions arise: (a) does the sequence $\mathbf{f}_K^{(t)}$ eventually converge as $t \to \infty$? and (b) assuming that $\mathbf{f}_K^{(t)}$ converges, and because the particular assignment of object triples/quadruples to the sets C_k, $1 \leq k \leq K$, could lead to different graph-theoretic structures, how adequate according to the least-squares loss function are the graph-theoretic structures identified over different assignments of object triples/quadruples to the sets C_k, $1 \leq k \leq K$?

Without further modification, the sequence $\mathbf{f}_K^{(t)}$ may not necessarily converge, and because of the subtraction of the changes $\mathbf{e}_k^{(t-1)}$ prior to projection onto whatever C_k is identified, the sequence may eventually oscillate through a fixed collection of different vectors. If the changes $\mathbf{e}_k^{(t-1)}$ were not considered in the iterative process, however, and if $\mathbf{f}_k^{(t)}$ were merely the projection of $\mathbf{f}_{k-1}^{(t-1)}$ onto the set C_k selected in the same adaptive manner (heuristic iterative projection without augmentation), it is generally possible to show that $\mathbf{f}_K^{(t)}$ must converge, subject to usually minor regularity conditions on which set of vectors each of the C_k defines, and irrespective of how they were chosen by the iterative process. For example, suppose the zero vector is included in each C_k (as it is in the ultrametric and additive tree applications). Then, from the general properties of projections onto closed convex sets in a Hilbert space (e.g., see Robertson, Wright, and Dykstra, 1988, p. 375), we know that $\mathbf{f}_{k-1}^{(t)\prime}\mathbf{f}_{k-1}^{(t)} \geq \mathbf{f}_k^{(t)\prime}\mathbf{f}_k^{(t)}$ for $1 \leq k \leq K$ and $t \geq 1$, and thus, $\mathbf{f}_K^{(t)\prime}\mathbf{f}_K^{(t)}$ as $t \to \infty$ is a bounded (from below by zero) monotonic sequence, and is thus convergent. Furthermore, since $(\mathbf{f}_k^{(t)} - \mathbf{f}_{k-1}^{(t)})'(\mathbf{f}_k^{(t)} - \mathbf{f}_{k-1}^{(t)}) \leq \mathbf{f}_{k-1}^{(t)\prime}\mathbf{f}_{k-1}^{(t)} - \mathbf{f}_k^{(t)\prime}\mathbf{f}_k^{(t)}$ for $1 \leq k \leq K$ and $t \geq 1$, $\mathbf{f}_K^{(t)}$ must be a Cauchy sequence and thus convergent to a vector in R^n, say $\mathbf{f}_K^{(*)}$ (e.g., see Buck, 1965, p. 45). It is important to note that this latter vector $\mathbf{f}_K^{(*)}$, although belonging to $C_1 \cap \ldots \cap C_K$ using the collection of closed convex sets eventually identified at convergence of $\mathbf{f}_K^{(t)}$ for $t \to \infty$, may not necessarily be the optimal least-squares solution $\mathbf{x}^* \in C_1 \cap \ldots \cap C_K$. Thus, once the collection $C_1,\ldots,C_K$ is

identified though $f_X^{(*)}$, it will be necessary in general to obtain x^* through the use of iterative projection (with augmentation) onto the fixed collection of closed convex sets $C_1,\ldots,C_X$ starting with p.

Given the possible nonconvergence of $f_X^{(t)}$ using iterative projection with augmentation but the convergence of $f_X^{(t)}$ to $f_X^{(*)}$ using iterative projection without augmentation, one obvious heuristic search strategy would be to rely only on the latter, and obtain the distribution of possible local optima using the collection of those $f_X^{(*)}$ found over some number of different sequences in which the constraints implied by each of the distinct object triples/quadruples are considered. Another option, and one that generally seems to be empirically preferable based on our computational experiences (e.g., see the examples to follow), is to consider the sequences $f_X^{(t)}$ obtained by iterative projection with augmentation until (and if) an oscillation occurs among a collection of vectors; at this point, choose one of them randomly and proceed with iterative projection without augmentation until convergence. We give some numerical examples of the results obtainable with these two variations below (that we denote as iterative projection with augmentation [IPWA] and iterative projection without augmentation [IPWOA]), using the ultrametric and additive tree constraints on both the Rao data set of Table 1 and one from Shepard, Kilpatric, and Cunningham (1975) given in Table 3 on the dissimilarity of the first ten single-digit integers {0,1,...,9} considered as abstract concepts and obtained by averaging dissimilarity ratings between the integers over a number of subjects and conditions.

{Table 3 here}

3.1 Some Numerical Comparisons Using the Rao and Shepard et al. Proximity Matrices

Table 4 provides rather detailed frequency distributions for the values of the least-squares loss function obtained for the Rao and Shepard et al. proximity

matrices in the attempt to locate the best-fitting ultrametrics and additive trees. These distributions were generated over 100 random assignments of the distinct object triples/quadruples to the particular sets $C_1,\ldots,C_K$, and are given for the two heuristics of IPWA and IPWOA (where IPWA would be followed by iterative projection without augmentation if nonconvergence [i.e., an oscillation] is identified in its use; the number of instances of such nonconvergence are indicated in Table 4).

{Table 4 here}

One very general observation can be made from the results reported in Table 4 regarding IPWA and IPWOA. Over both data sets and structural representations, IPWA outperforms IPWOA in producing far fewer local optima, and they are generally biased toward the better (i.e., smaller) values of the loss function. This difference is not subtle, and even though IPWA may lead to nonconvergence and the eventual need to continue the iterative process without augmentation, we are adopting IPWA as the preferred heuristic optimization strategy. (We note that although the evidence presented for this preference is based only on these two data sets, this same advantage has been observed consistently by the authors over a much larger number of examples not explicitly reported here.)

Although the traditional goal in the development of any heuristic optimization method is to construct a procedure leading as consistently as possible to the absolute best solution for any chosen loss function, we would like to offer a different perspective on the presence of local optima and suggest that their presence may be reflective of some structure in the given data that is noteworthy but would be overlooked if attention were focused solely on the absolute best solution attainable. Thus, instead of viewing the presence of several salient local optima as an evil to be avoided, we consider them as possibly indicative of the patterning of the given data. More explicitly, we discuss briefly the range of structural representations by ultrametrics and additive trees that were identified

for the Rao and Shepard et al. proximity matrices.

For the Rao data, the best ultrametric found (37/100 for IPWA), according to the VAF of 56.155%, was given earlier in the example of fitting a given partition hierarchy to a proximity matrix; the other salient local optimum with a VAF of 49.555% (35/100 for IPWA) corresponds to the partition hierarchy given below:

Level	Partition (only classes with more than one object are listed)
0	all objects separate
1	{10,11}
2	{10,11},{1,2}
3	{9,10,11},{1,2}
4	{9,10,11},{1,2},{7,8}
5	{9,10,11,12},{1,2},{7,8}
6	{9,10,11,12},{1,2},{7,8},{5,6}
7	{9,10,11,12},{1,2},{7,8},{5,6},{3,4}
8	{1,2,9,10,11,12},{7,8},{5,6},{3,4}
9	{1,2,7,8,9,10,11,12},{5,6},{3,4}
10	{1,2,5,6,7,8,9,10,11,12},{3,4}
11	all objects together

In comparison to the best ultrametric identified, the partition sequence above is identical up to Level 6, at which point the next Level 7 includes the new group {3,4}, which remains as a separate pair until all objects are merged at Level 11. This latter hierarchy, and although obviously not the absolute best according to least-squares loss, is more consistent with the substantive discussion given by Rao on the structuring of the twelve groups (based on presumed genetic linkages) through the single partition {{9,10,11,12},{1,2},{7,8},{5,6},{3,4}}; the latter partition is identified at Level 7 in the hierarchy given above but not in the hierarchy associated with the best least-squares loss value. In fact, based on a comparison not recounted in detail here, all the partition hierarchies identified by IPWA can

be interpreted as being minor variations on these two basic hierarchies, which themselves differ primarily in the treatment of objects 3 and 4, e.g., the local optima identified can be seen as slightly different orders of object set merges leading to exactly the same Level 6 partition of {{9,10,11,12},{1,2},{7,8},{5,6},{3},{4}}; forming a next partition by merging three groups rather than two; or slightly different object set merges leading from the Level 6 partitions in both hierarchies to the last all-inclusive set. Concentrating solely on the best ultrametric identified and seeking a complete underlying partition hierarchy rather than, say, a single best explanatory partition, would cause the salience of the object pair {3,4} to be missed.

Considering the representation of the Rao data by additive trees, the three local minima identified with IPWA do not appear to reflect major substantive differences. Figure 3 gives the two additive trees corresponding to those with VAF of 89.338% and 89.324% (and identified 54/100 and 34/100, respectively, by IPWA). The third local optimum with a VAF of 89.321% (identified 12/100 with IPWA) has a form identical to that for the 89.324% solution except for the interchange of objects 10 and 11 (and corresponding minor variations in the branch lengths). For both additive trees in Figure 3 (and the third not shown), the deletion of the four branches between internal nodes with the largest scalars attached produces the partition {{9,10,11,12},{1,2},{7,8},{5,6},{3,4}}, which is consistent with the one used by Rao (and identified at Level 7 in the partition hierarchy given above), and which presumably reflects the salient genetic linkages among the groups.

{Figure 3 here}

Using IPWA to identify the best fitting ultrametric for the Shepard et al. dissimilarity data on the first nine integers, two salient local optima occurred with VAF of 49.410% and 47.812% (identified 27/100 and 65/100, respectively):

Level	Partition(VAF of 49.410%)	Partition(VAF of 47.812%)
0	all objects separate	all objects separate

1	{2,4}	{2,4}
2	{2,4},{3,9}	{2,4},{3,9}
3	{2,4},{3,6,9}	{2,4},{3,6,9}
4	{2,4,8},{3,6,9}	{2,4,8},{3,6,9}
5	{2,4,8},{3,6,9},{5,7}	{2,4,8},{3,6,9},{5,7}
6	{2,4,8},{3,5,6,7,9}	{2,4,8},{3,6,9},{5,7},{0,1}
7	{2,3,4,5,6,7,8,9}	{2,4,8},{3,5,6,7,9},{0,1}
8	{1,2,3,4,5,6,7,8,9}	{2,3,4,5,6,7,8,9},{0,1}
9	all objects together	all objects together

Much as for the Rao data, these two partition hierarchies are identical up to Level 5, and then differ primarily in the treatment of one object pair, {0,1}. In the hierarchy with greater VAF, that pair is ignored and these two objects are added sequentially to a larger subset at Levels 8 and 9; for the other salient local optima, the pair is formed at Level 6 and remains separate until all objects are merged at Level 9. Given the nice interpretation of the pair {0,1} as consisting of the identities (additive and multiplicative), in addition to the other subsets generally being structurally interpretable (e.g., powers of 2:{2,4,8}; multiples of 3:{3,6,9}; the nontrivial odd numbers that are not powers of 3:{5,7}), the lesser fitting partition of these two according to VAF could arguably be substantively preferable. Again, much as for the Rao data, the other four local optima identified were just minor variations on the two given above (e.g., merging three object sets early in the hierarchy before Level 5, and a sightly different merge order for the subsets both before and after Level 5).

For additive trees, IPWA identified two salient local optima with VAF of 63.591% and 62.486% (located 49/100 and 43/100, respectively), as presented in Figure 4. (The other four local optima are minor variations on these two and vary as to how two branches connecting two objects with a subset are sequenced and thus are not discussed further). The less well fitted additive tree in Figure 4(b) reflects the

same type of structural properties of the numbers generally indicated by the ultrametrics (e.g., powers of 2, multiples of 3, additive/multiplicative identities, odd numbers that are not powers of 3), but the better fitted additive tree reflects much more the magnitude of the numbers, which was not clearly identifiable in the ultrametric solutions. A detailed inspection of the residuals for the two solutions given in Figure 4 makes this interpretation fairly clear, i.e., the reconstructed proximities from Figure 4(a) fit the magnitude better than the structural properties of the integers, and conversely for Figure 4(b). Thus, in this case, the two salient local minima identify rather distinct aspects of the proximity data, and although one additive tree does have a higher VAF, from a substantive perspective the other one is noteworthy and reflects some of the overall patterning present in the initial proximities.

(Figure 4 here)

Incorporating bounds in IPWA. One flexible aspect in the use of IPWA as a heuristic optimization strategy is the possibility of easily placing additional restrictions on the reconstructed proximities merely through the incorporation of further constraint sets in the course of the iterative process. One general type of restriction that may be of interest to impose is to force an approximation to be either "from below", or "from above", in which case the reconstructed proximities from whatever structure is being considered must also be greater than or equal, or less than or equal, to the original proximities, respectively. Thus, the original proximities are used to form upper or lower bound constraints on the entries in an optimal solution vector through the incorporation of n additional convex sets defined for each proximity value in **p**.

There is one case, in particular, involving the incorporation of proximity upper bounds in the location of a best-fitting ultrametric, that has a very interesting connection to a familiar heuristic construction strategy for partition hierarchies,

called the single-link method of clustering. Moreover, this connection allows us to demonstrate the ability of IPWA to find an absolute best value of the least-squares loss function for a constrained structural representation because the single-link result itself provides the optimal solution. To be somewhat more specific, the single-link strategy can be characterized by how it forms the level k+1 partition from the level k partition, and specifically in how the new subset is formed at level k+1:

> The closeness of any two subsets at level k is considered to be the minimum proximity (dissimilarity) between a pair of objects chosen from the two subsets. The pair of subsets that are united to form the new subset at level k+1 minimizes this value.

The collection of minimal values corresponding to the formation of the subsets in the hierarchy defines a set of reconstructed proximities between the objects from the united subsets. If the minimum is not attained for a unique subset pair, one pair may be merged arbitrarily without loss of generality and any other pairs merged at successive levels. We denote as $\mathbf{f}_{sl}$ the vector of reconstructed proximities found by this single-link strategy (which is unique).

The vector $\mathbf{f}_{sl}$, besides satisfying the constraints of an ultrametric, is also subject to the bounds $\mathbf{p} \geq \mathbf{f}_{sl}$, where the notation indicates that all entries in $\mathbf{p}$ bear that given inequality relation to the entries in $\mathbf{f}_{sl}$. Moreover, the vector $\mathbf{f}_{sl}$ can be shown to provide an optimal least-squares approximation to $\mathbf{p}$ subject to this constraint. The proof (see, e.g., Barthélemy and Guénouche, 1991, p. 98) rests on the observation that the entries in any other vector that is both an ultrametric and bounded by $\mathbf{p}$ must also be bounded by $\mathbf{f}_{sl}$ ($\mathbf{f}_{sl}$ is called the subdominant ultrametric associated with $\mathbf{p}$). A major consequence of $\mathbf{f}_{sl}$ being identified as the (globally) optimal least-squares approximation to $\mathbf{p}$ subject to the additional upper-bound constraint is that we may evaluate the general effectiveness of IPWA in identifying

such an optimal solution. Although obviously computationally more expensive, if IPWA does well when the optimal solution is known, some confidence in its general ability to locate optimal solutions is obtained. To be explicit, the single-link partition hierarchies for the Rao and Shepard et al. proximity matrices are given below, along with the proximity values inducing the construction of each new subset at a particular level:

Shepard et al.			Rao		
Level	Partition	Proximity	Level	Partition	Proximity
0	all objects separate	.000	0	all objects separate	.000
1	{2,4}	.059	1	{10,11}	.120
2	{2,4,8}	.246	2	{10,11},{1,2}	.270
3	{2,4,8},{3,9}	.263	3	{9,10,11},{1,2}	.300
4	{1,2,4,8},{3,9}	.284	4	{9,10,11},{1,2},{7,8}	.400
5	{1,2,4,8},{3,6,9}	.296	5	{9,10,11,12},{1,2},{7,8}	.430
6	{1,2,3,4,6,8,9}	.346	6	{1,2,9,10,11,12},{7,8}	.780
7	{1,2,3,4,5,6,8,9}	.396	7	{1,2,7,8,9,10,11,12}	.900
8	{1,2,3,4,5,6,7,8,9}	.400	8	{1,2,7,8,9,10,11,12},{5,6}	1.150
9	all objects together	.421	9	{1,2,4,7,8,9,10,11,12},{5,6}	1.126
Loss value: 5.677			10	{1,2,3,4,7,8,9,10,11,12},{5,6}	1.320
			11	all objects together	1.750
			Loss value: 348.316		

Using IPWA over 100 random starts for the Rao proximity matrix and with the addition of the upper-bound constraints, the optimal single-link result was identified 75/100, along with 15 other local optima; for the Shepard et al. proximity matrix, the single-link result was generated 81/100, along with 8 other local optima. Thus, at least for these two data sets, IPWA is fairly effective in identifying the optimal solution (we might note that a similar consistency exists as well for the variety of other data sets we have experimented with).

4. Some Generalizations and Extensions

There are many extensions and applications of iterative projection, both in the fitting of a given vector **p** by some optimal vector **x*** subject to a fixed set of constraints, and in the subsequent attempt to use such an iterative strategy as a search heuristic to identify the best set of constraints for whatever particular representation is desired. We discuss a few of these possibilities briefly in the sections to follow and give several numerical illustrations.

4.1 Other Structural Representations for Symmetric Proximity Matrices.

Although we have emphasized the fitting of ultrametrics and additive trees to a given symmetric proximity matrix, other structures could also be analogously considered with the heuristic of IPWA whenever the constraint sets corresponding to a particular structure can be defined by the intersection of closed convex sets. To give an example of how other structural representations might be considered, one possible alternative to ultrametrics or additive trees in the representation of a symmetric proximity matrix would be by a k-ultrametric (see Jardine & Sibson, 1971, pp. 65-71) generalization of ultrametric structure. Specifically, a k-ultrametric corresponds to a sequence of H+1 sets of subsets of S, $\Delta_0,\ldots,\Delta_H$, where (a) each Δ_h contains a set of subsets of S that can overlap by at most k-1 objects, (b) Δ_0 contains N classes each defined by a single object from S; Δ_H contains a single class containing all the objects in S; (3) the subsets in Δ_{h+1} are formed from those in Δ_h, $0 \leq h \leq H-1$, by the union of one or more subsets in Δ_h. Thus, a 1-ultrametric corresponds to an ultrametric in the usual sense.

The defining condition for a k-ultrametric that in turn generates a set of linear inequality constraints for fitting a fixed k-ultrametric to a given proximity vector **p** is that, among all subsets of size k+2, the two largest pairwise proximities in the fitted vector must be equal. Consequently, for a 1-ultrametric

(and thus, for an ultrametric in the usual sense), among all object triples and the three corresponding fitted proximities, the larger two must be equal; for a 2-ultrametric and among all object quadruples and the six corresponding fitted proximities, the largest two must be equal, and so on.

To illustrate numerically the fitting of a k-ultrametric, we used IPWA to find the best-fitting 2-ultrametric for the Rao proximity data set. Based on 100 random assignments of the order of considering the constraints on all object quadruples, a total of 21 local optima were identified with one very salient exemplar located 64/100, with corresponding VAF of 71.992%. Four other local optima that were observed a collective nine times actually had somewhat better VAF (up to 72.256%) than the most salient local optimum, but these also had less apparent interpretative consistency with the presumed genetic linkages, i.e., the subset of Artisans, {9,10,11,12}, was not identified explicitly at any level (this finding again may reflect the possibility that an absolute best solution according to a given loss function may not be as substantively interpretable as one with a somewhat worse loss value). Table 5 presents the collection of subsets formed for the most salient 2-ultrametric along with the values for those object pairs fitted at each level and which serve to reconstruct the optimal solution vector for the given hierarchical collection of subsets. As can be seen at Level 13, for instance, the collection of subsets, {{1,2},{1,9},{2,9},{3,4},{4,9},{5,6},{7,8},{8,12},{9,10,11,12}}, is consistent with the presumed genetic linkages discussed previously in the context of the best-fitting ultrametrics through the single partition, {{1,2},{3,4},{5,6}, {7,8},{9,10,11,12}}, but it also includes an additional four object pairs, {1,9}, {2,9}, {4,9}, and {8,12}. The latter pairs correspond to relatively similar objects that are not found when solely considering mutually exclusive and exhaustive collections of subsets defining partitions (as in the fitting of 1-ultrametrics).

{Table 5 here}

4.2 Two-Mode Proximity Data

The proximity data considered thus far for obtaining some type of structural representation, such as say an ultrametric or an additive tree, have been assumed to be on one intact set of objects, $S = \{O_1,\ldots,O_N\}$, and complete in the sense that proximity values are present between all object pairs. Suppose now that the available proximity data are two-mode, that is between two distinct object sets, $S_A = \{O_{1A},\ldots,O_{N_A A}\}$ and $S_B = \{O_{1B},\ldots,O_{N_B B}\}$ containing N_A and N_B objects, respectively, and defined through an $N_A \times N_B$ proximity matrix $Q = \{Q_{rs}\}$, where again, for convenience, we assume that the entries in Q are keyed as dissimilarities, and for later purposes can be arrayed as an $(N_A \times N_B) \times 1$ vector q. What may be desirable is a joint structural representation of the set $S_A \cap S_B$ (considered as a single object set S containing $N_A + N_B$ (= N) objects), but one that is based only on the available proximities between the sets S_A and S_B.

Conditions have been proposed in the literature for when the entries in a matrix fitted to Q (or equivalently, fitted to the vector q) characterize an ultrametric or an additive tree representation. In particular, suppose a vector $\mathbf{x}$ is fitted to q through least-squares subject to the constraints that follow:

Ultrametric (Furnas, 1980):

for all distinct object quadruples, O_{rA} , O_{sA} , O_{rB} , O_{sB} , where O_{rA} , $O_{sA} \in S_A$ and O_{rB} , $O_{sB} \in S_B$, and considering the entries in $\mathbf{x}$ corresponding to the pairs, (O_{rA}, O_{rB}), (O_{rA}, O_{sB}), (O_{sA}, O_{rB}), and (O_{sA}, O_{sB}), say $x_{r_A r_B}$, $x_{r_A s_B}$, $x_{s_A r_B}$, $x_{s_A s_B}$, respectively, the largest two must be equal.

Additive trees (Brossier, 1987):

for all distinct object sextuples, O_{rA} , O_{sA} , O_{tA} , O_{rB} , O_{sB} , O_{tB} , where O_{rA} , O_{sA} , $O_{tA} \in S_A$ and O_{rB} , O_{sB} , $O_{tB} \in S_B$, and

considering the entries in $\mathbf{x}$ corresponding to the pairs,

(O_{rA}, O_{rB}), (O_{rA}, O_{sB}), (O_{rA}, O_{tB}), (O_{sA}, O_{rB}), (O_{sA}, O_{sB}),

(O_{sA}, O_{tB}), (O_{tA}, O_{rB}), (O_{tA}, O_{sB}), and (O_{tA}, O_{tB}), say

$x_{\substack{r\,r\\A\,B}}$, $x_{\substack{r\,s\\A\,B}}$, $x_{\substack{r\,t\\A\,B}}$, $x_{\substack{s\,r\\A\,B}}$, $x_{\substack{s\,s\\A\,B}}$, $x_{\substack{s\,t\\A\,B}}$, $x_{\substack{t\,r\\A\,B}}$, $x_{\substack{t\,s\\A\,B}}$,

$x_{\substack{t\,t\\A\,B}}$, respectively, the largest two of the following sums must be equal:

$$x_{\substack{r\,r\\A\,B}} + x_{\substack{s\,s\\A\,B}} + x_{\substack{t\,t\\A\,B}} ;$$

$$x_{\substack{r\,r\\A\,B}} + x_{\substack{s\,t\\A\,B}} + x_{\substack{t\,s\\A\,B}} ;$$

$$x_{\substack{r\,s\\A\,B}} + x_{\substack{s\,r\\A\,B}} + x_{\substack{t\,t\\A\,B}} ;$$

$$x_{\substack{r\,s\\A\,B}} + x_{\substack{s\,t\\A\,B}} + x_{\substack{t\,r\\A\,B}} ;$$

$$x_{\substack{r\,t\\A\,B}} + x_{\substack{s\,r\\A\,B}} + x_{\substack{t\,s\\A\,B}} ;$$

$$x_{\substack{r\,t\\A\,B}} + x_{\substack{s\,s\\A\,B}} + x_{\substack{t\,r\\A\,B}} .$$

In both the cases of ultrametric and additive trees for two-mode proximity data, the necessary constraints characterizing a solution are linear and define closed convex sets in which a solution vector must lie. Thus, the possible application of IPWA as a heuristic search strategy for the best-fitting solutions is fairly direct, and an example of an ultrametric fitted to a two-mode data matrix will be given below. We will not, however, give a comparable example as yet of fitting the additive tree constraints to such a proximity matrix; the (scratch) storage requirements necessitated by IPWA in directly using the additive tree constraints given above and keeping track of the various augmentations made in the course of carrying out its heuristic search can become rather onerous for moderate-sized data matrices, and for general use, an alternative approach to the fitting of additive trees that again uses IPWA is preferable because it avoids any major (scratch) storage difficulties.

This strategy will be reviewed in a section to follow, and at that point the

comparable additive tree representation will be given for the two-mode data matrix fitted below through ultrametric constraints. We might also note that the process of fitting two-mode proximity data by additive trees or ultrametrics using IPWA may generate a rather large number of distinct locally optimal solutions, particularly in contrast to the situation usually observed for symmetric proximity data. Although this abundance is not inevitably the case and depends on the particular data set being considered (as we have observed in some of our experimentation), it is true in the example below, for instance.

As an illustration of using IPWA to find structural representations for a given rectangular proximity matrix, we consider the ultrametric constraints given above and a data set originally collected by Price and Bouffard (1974), which has been reanalyzed by Eckes and Orlik (1993) among others, on the appropriateness of 15 behaviors in 15 situations. The data, given in Table 6, are averaged ratings (on a scale from 0 to 9) over 52 subjects; thus, higher values in the table reflect a higher appropriateness of the given row object (situation) for the given column object (behavior). Thus, if the proximities are considered dissimilarities and an ultrametric representation is of interest, the concern would be with grouping situations and inappropriate behaviors. Here, using 100 random assignments of order in imposing the ultrametric constraints, a very large number of local optima were observed using IPWA, with VAF ranging from 55.779% to 60.773%; in fact, only in a few instances were the same optima identical.

Of the various hierarchies constructed, some could be considered just minor variations of each other with slightly different orders in which objects were merged, but there were enough nontrivial differences observed among hierarchies to suggest that any attempt to represent the data by a single hierarchical structure would undoubtedly fail to account for some important structuring of the original situation by behavior proximity data, and the part not represented well would vary depending on the particular local optimum selected for study. We give the hierarchy

corresponding to the best VAF achieved of 60.773% in Table 7; the various clusters do make intuitive sense according to the original proximity data and the obvious interpretations of the behaviors and situations. For example, at Level 11 we have the two groupings of {class, family dinner, church, job interview, elevator; run, sleep, fight, belch, jump, shout} and {movies; argue}, which do indeed reflect clusters of situations for which the corresponding behaviors would be relatively inappropriate. However, selecting this particular hierarchy for interpretation and observing the residuals from the fitted matrix, we do not adequately account for some of the relatively extreme proximity values in the initial table, e.g., the rather inappropriate early pairing of class/write and date/kiss, and the late pairing of job interview/eat and church/laugh. Several other hierarchies that were observed do better for these specific pairs but then again do worse for some other pairs that are accounted for reasonably well with the hierarchy given in Table 7. We will come back to this point below when a multiple matrix representation for these data is considered.

{Tables 6 and 7 here}

4.3 Representations of Proximity Matrices (Symmetric or Two-Mode) as Sums of Matrices Subject to Structural Constraints

For fitting a given proximity vector by a single vector $\mathbf{x}$ subject to, say, ultrametric or additive tree constraints, one possible generalization uses a sum of T vectors, $x_1 + \ldots + x_T$, where each vector x_t, $1 \leq t \leq T$, would itself be subject to such constraints (for example, see De Soete et al., 1984; Carroll and Pruzansky, 1980). Formally, for a given vector $\mathbf{p} \in R^n$ (or replacing $\mathbf{p}$ by $\mathbf{q}$ if two-mode proximities are given), we wish to obtain the vector $\mathbf{x} \in R^n$, say $\mathbf{x}^*$, that minimizes

$$(\mathbf{p} - \mathbf{x})'(\mathbf{p} - \mathbf{x}) \ , \tag{6}$$

where $\mathbf{x} = \mathbf{x}_1 + \ldots + \mathbf{x}_T$ for $\mathbf{x}_t \in C_t$ and where each C_t , $1 \leq t \leq T$, is a closed convex subset of R^n, i.e., $\mathbf{x} \in C_1 + \ldots + C_T$, where $C_1 + \ldots + C_T$ denotes the direct sum of

$C_1,\ldots,C_T$. If the C_t are sufficiently well-formed (for example, if they are closed convex polyhedral cones defined through linear constraints as in (2), where the upper-bound scalar values for the linear inequality constraints are all zero — as they are when each C_t represents a fixed ultrametric or additive tree structure), a simple adaptation of iterative projection (with augmentation) can be used to solve the least-squares task in (6) because in this case, $C_1 + \ldots + C_T$ is itself a closed convex cone (see, for example, Dykstra, 1983, pp. 838–839). In fact, IPWA for a fixed collection of subsets $C_1,\ldots,C_T$ reduces in this case to what is now commonly referred to as an alternating least-squares strategy (e.g., see Carroll and Pruzansky, 1980).

To show how IPWA can be used to fit $\mathbf{p}$ by $\mathbf{x}^*$ within $C_1 + \ldots + C_T$, we first introduce the notion of a polar (closed convex) cone C^P corresponding to a given closed convex cone C defined as $C^P = \{\mathbf{y} \in R^n \mid \mathbf{x}'\mathbf{y} \leq 0 \text{ for } \mathbf{x} \in C\}$, and note that if $\mathbf{f}_c$ denotes the projection of $\mathbf{p}$ onto C, then (the "residual") vector $\mathbf{p} - \mathbf{f}_c$ is the projection of $\mathbf{p}$ onto C^P. Now, if $\mathbf{x}^*$ denotes the projection of $\mathbf{p}$ onto $C_1 + \ldots + C_T$, then ($\mathbf{y}^* = $) $\mathbf{p} - \mathbf{x}^*$ is the projection of $\mathbf{p}$ onto $(C_1 + \ldots + C_T)^P$, and the latter is equal to $C_1^P \cap \ldots \cap C_T^P$. Thus, if the projection $\mathbf{y}^*$ of $\mathbf{p}$ onto $C_1^P \cap \ldots \cap C_T^P$ is found though IPWA, then $\mathbf{x}^*$ may be obtained simply as $\mathbf{p} - \mathbf{y}^*$, and the component vectors comprising $\mathbf{x}^*$ identified in the course of the iterative process as the successive projections (at convergence) onto $C_1,\ldots,C_T$. Operationally, then, one starts with the vector $\mathbf{p}$, projects $\mathbf{p}$ onto C_1, and obtains the residual vector; the latter is projected onto C_2, and so on until the residual vector is obtained from the projection onto C_T. Now, considering the sets $C_1,\ldots,C_T$ again in order, the residual vector from the last projection onto C_T is first augmented by the projection onto C_1 from the previous stage and then reprojected onto C_1; the residual vector thus obtained is augmented by the previous projection onto C_2, and projected onto C_3, and so on, until convergence of the whole process is achieved by

continued recycling through the constraint sets.

Adopting IPWA as a heuristic search strategy for fitting a vector $\mathbf{x}^* \in C_1 + \ldots + C_T$ requires its implementation both (a) to locate reasonable sets of "complete" constraints, $C_1,\ldots,C_T$, that typically represent, say, the sets of constraints imposed by considering ultrametric or additive trees, and (b) to construct each of the individual constraint sets, C_t, $1 \leq t \leq T$. Because the process of using IPWA in the construction of $\mathbf{x}^*$ attempts to identify the constraint sets, C_t, in order from 1 to T, and does so through a heuristic mechanism during the iterative process, the (cumulative) VAF could actually decrease at some value of t, i.e., the heuristic search at that point failed to identify a better constraint set C_t than that already present at the previous stage when C_t was last considered. When such cases occur in our implementation of IPWA, the previous stage's constraint set C_t is automatically readopted, and the process proceeds to the consideration of C_{t+1}. This strategy guarantees a nondecreasing cumulative VAF that eventuates in the fitted vector $\mathbf{x}^*$.

To give a first illustration of representing a proximity matrix as a sum, we first recall that Table 7 presented the best ultrametric representation observed for the two-mode proximity data of Table 6 (with VAF of 60.773%), but we noted that the use of any single hierarchy would inevitably account for some of the initial proximities better than others. Given the possibility of now representing the data of Table 6 as a sum of ultrametric structures, we would expect that additional vectors in the sum would fit those proximities represented less well by the other vectors. Generally, based on our computational evaluations, this expectation was met, but the large number of distinct local optima observed in fitting a single ultrametric vector carries over to the representation of the proximity vector as a sum. In particular, the fitted proximities based on sums of ultrametrically constrained vectors (when using two or more component vectors) are very consistent across solutions, but the constituent terms of the sums vary, much as for the local

optima observed in the fitting of a single ultrametric vector.

As an illustration, a two-vector approximation to the data of Table 6 was obtained through the IPWA heuristic with both vectors having ultrametric constraints and with the process initialized with the hierarchy of Table 7 defining the first constraint set C_1. (To conserve space we do not list the complete hierarchies.) Upon convergence, the first constraint set C_1 (with some very minor variations) was the same as that in Table 7, but the fitted values now differ, given the additional presence of a second constraint set C_2 in defining the two-vector approximation. If the fitted proximities from the first vector are considered at and below a threshold value of 1.5 and those for the second at and below -1.00, two collections of subsets of situations and inappropriate behaviors are identified, with the first being identical to that obtained from level 11 in the hierarchy of Table 7:

For C_1 : {class, family dinner, church, job interview, elevator; run, sleep, fight, belch, jump, shout}, {movies; argue}

For C_2 : {class, job interview; kiss, mumble, cry}, {date; belch}, {bus; run, shout}, {family dinner, elevator; write}, {park, sidewalk, bar, restroom, own room, dorm lounge, football game; fight}, {church; talk, eat, laugh}, {movies; read}

The VAF by the two-vector approximation is 77.491%, an increase of about 17% over the single-vector approximation of Table 7. In particular, given the subsets identified above for C_2 at the threshold of -1.00, it is apparent that the second vector does provide additional sets of behaviors and relatively inappropriate situations not identifiable with just a single-vector approximation.

As a second illustration that uses the Shepard et al. data, and over 100 random assignments of the order of considering the constraints, six local optima were observed in the fitting of a two-vector representation where each vector was

required to satisfy the ultrametric constraints. The best local optimum according to a VAF of 80.009% was observed 23/100, four local optima were observed one time each, and the sixth local optimum (VAF of 78.130%) was identified 73/100. Considering the two salient local optima, the first vector component of each corresponded closely to the two salient local optima identified earlier in the single-vector ultrametric representations (which differ primarily in how the multiplicative and additive constants {0,1} were treated in the hierarchy). We give the result below corresponding to the best VAF of 80.009%; the second hierarchy can be generally interpreted as numerical magnitude (much as we identified in our earlier discussion of the salient local optima for a single additive tree representation), and the first corresponds more to the structural properties of the numbers.

	First Hierarchy		Second Hierarchy	
Level	Partition	Fitted Proximity	Partition	Fitted Proximity
0	all objects separate	.000	all objects separate	.000
1	{2,4}	.070	{0,1}	−.359
2	{2,4},{3,9}	.173	{0,1,2}	−.253
3	{2,4,8},{3,9}	.217	{0,1,2},{6,8}	−.201
4	{2,4,8},{3,6,9}	.277	{0,1,2,3},{6,8}	−.172
5	{2,4,8},{3,6,9},{5,7}	.310	{0,1,2,3},{4,5},{6,8}	−.142
6	{2,4,8},{3,5,6,7,9}	.462	{0,1,2,3},{4,5},{6,7,8}	−.098
7	{2,3,4,5,6,7,8,9}	.551	{0,1,2,3},{4,5},{6,7,8,9}	−.048
8	{1,2,3,4,5,6,7,8,9}	.595	{0,1,2,3,4,5},{6,7,8,9}	−.010
9	all objects together	.780	all objects together	.090

The decomposition of an additive tree. There is a connection between the representation of a (symmetric or two-mode) proximity vector through a single vector

$\mathbf{x}^*$ that satisfies the additive tree constraints, and a corresponding two-vector representation of the latter vector $\mathbf{x}^*$, say as $\mathbf{x}_1^* + \mathbf{x}_2^*$, where $\mathbf{x}_2^*$ satisfies the ultrametric constraints and $\mathbf{x}_1^*$ represents a particular form of an additive tree in which all the sums in the defining additive tree conditions must be equal (and not only the largest two). For this latter type of additive tree, the fitted values in $\mathbf{x}_1^*$ are representable as sums, say as $g_i + g_j$ for a symmetric proximity between objects O_i and O_j, or as $u_i + v_j$ for rectangular proximities between objects O_{iA} and O_{jB}. This decomposition has been shown in a number of sources (e.g., see Carroll, 1976; De Soete et al., 1984) and provides one mechanism for the fitting of single additive trees to either symmetric or rectangular proximity matrices by a two-vector decomposition. We note that in the course of the iterative strategy and using whatever residual vector is at hand, the fitting of the constrained additive tree can be done in closed form (i.e., for rectangular proximities, u_i (or v_j) can be given as the ith row (or jth column) mean of the residual matrix under consideration minus one-half the grand mean; for symmetric proximities, g_i can be given as the ith row sum excluding the diagonal entry, divided by N-2, minus the total off-diagonal sum divided by 2(N-1)(N-2); see Carroll and Pruzansky, 1980, and De Soete et al., 1984, for a further discussion).

To provide one numerical illustration of using the two-vector approach to the fitting of an additive tree, we first consider both the Rao and the Shepard et al. proximity matrices over 100 random considerations of the order of the constraints within C_1 and C_2, but where C_1 was consistently the restricted additive tree and C_2 imposed the ultrametric constraints. For the Rao data, only two different local optima were observed (about equally often) corresponding to the VAF values given earlier in Table 4 as 89.324% and 89.321%. For the Shepard et al. data, only the local optimum corresponding to a VAF of 62.486% was observed over the 100. Thus, in both cases the best VAF observed for the local optima in Table 4 was not identified,

presumably because of the manner in which the additive tree is heuristically constructed through the sequential consideration of two constraint sets. A reversal in the consideration of the sets C_1 and C_2, where C_2 was now identified as the restricted additive tree, and again over 100 random considerations of the order of the constraints within C_1 and C_2, consistently produced for the Shepard et al. data the same local optimum with a VAF of 62.486%, but for the Rao data, six different local optima were observed with one having VAF of 87.605% and identified 64/100, and which was not observed previously with IPWA. Again, the absolute best VAF results of Table 4 were not achieved, which in summary may suggest the greater ability of IPWA to identify the absolute best (single) additive tree representations when the additive tree constraints are searched for directly rather than indirectly through a two-vector decomposition.

As a second example of fitting additive tree constraints through a two-vector decomposition, the Price and Bouffard (1974) data from Table 6 are considered, again using 100 random assignments of the ordering of the constraints. A rather substantial number of local optima were observed almost comparable to the number of starting orders but with a very restricted range of VAF from 81.573% to 81.964%. Somewhat in contrast to the use of ultrametric constraints, however, the various additive trees identified could be considered rather minor variations of each other and reflected a slightly different local ordering of branches. We give the additive tree corresponding to the best VAF of 81.964% in Figure 5. The lengths of the branches attached to the 30 terminal nodes mirror extremely well the average row and column values in the original data of Table 6. In fact, the actual tree structure itself can be interpreted as an attempt to fit the situation/behavior "interactions" not reconstructible from the general "main effects" of situation and behavior as represented in the branch lengths attached to the terminal nodes.

{Figure 5 here}

4.4 Three-way Proximity Matrices, With Either Two or Three Modes.

One generalization to the fitting of a single (either two-mode or symmetric) proximity matrix by ultrametrics or additive trees that has already been suggested (for example, see Carroll et al., 1984) is the simultaneous consideration of, say, L (symmetric or two-mode) proximity matrices defined on the same object set(s). Typically, a common typology is sought for the L proximity matrices, but with an allowance for differential fitting of each proximity matrix based on the common set of constraints identified through some heuristic optimization strategy, e.g., different branch lengths for a common additive tree or different levels at which subsets are formed for a common ultrametric. Formally, in a least-squares context and for a common single representation vector, if $p_1,\ldots,p_L \in R^n$ denote the given proximity vectors (or replacing $p_1,\ldots,p_L$ by $q_1,\ldots,q_L$ throughout when two-mode proximity data are considered), we wish to obtain those vectors $x_1^*,\ldots,x_L^* \in R^n$ that minimize

$$\sum_l (p_l - x_l)'(p_l - x_l) \quad ,$$

where for $1 \leq l \leq L$, $x_l \in C_1 \cap \ldots \cap C_K$, and each C_k $(1 \leq k \leq K)$ is a closed convex subset of R^n. The collection of subsets $C_1,\ldots,C_K$ defines the specific common structural representation for all the initial proximity vectors, but each is fitted by a possibly different vector within $C_1 \cap \ldots \cap C_K$. If the closed convex subsets $C_1,\ldots,C_K$ are known, the vectors $x_1^*,\ldots,x_L^*$ can obviously be obtained by L separate applications of IPWA.

There are two possible approaches to extending IPWA as a heuristic optimization technique in the multiple proximity vector context where an actual identification of the subsets $C_1,\ldots,C_K$ is of interest. One strategy would iteratively cycle through the K subsets attempting to define better constraint sets sequentially by choosing C_k adaptively based on the current values (for example, through a simple aggregation) over all the relevant L vectors being modified in the iterative

process. That particular subset C_k would then be used for each of the L augmentations and reprojections. As a second alternative strategy, a two-stage procedure might be implemented in which the subsets $C_1,\ldots,C_K$ are first identified using IPWA from a single aggregate proximity vector, $\sum_l p_l$, with a subsequent construction of the vectors $\mathbf{x}_1^*,\ldots,\mathbf{x}_L^*$ by iterative projection onto the fixed sets $C_1,\ldots,C_K$. This latter approach is relatively simple to apply with the same computational program we have relied on in fitting a single proximity matrix, and the example that we report below uses it. (A more systematic evaluation of both approaches is warranted but left for the future.)

The example we use is from Carroll et al. (1984), and involves three (symmetric) proximity matrices (which appear as Table 11 in Carroll et al., 1984) among ten pain relievers [Anacin(An), Ascriptin(As), Bayer Aspirin(BAs), Bufferin(Bu), Cope(Co), Datril(Da), Excedrin(Ex), Hudson Aspirin(HAs), Tylenol(Ty), and Vanquish(Va)]; each matrix corresponds to a different malady (headache, fever, and muscle aches) and was obtained by computing Euclidean distances between pairs of brands (and within malady) based on a rating of intention to use by 61 MBA students. Based on fitting ultrametric constraints to the aggregate proximity matrix, and over 100 random assignments of the order of considering the constraints, only one local optimum was found (with a VAF of 76.747%). Fitting this common ultrametric to each of the three matrices resulted in the differentially fitted proximities indicated below next to the various partitions of the common hierarchy, and gave separate VAF measures of 76.589%, 53.657%, and 60.681%, respectively (Carroll et al. reported a single VAF of 60.11% as an aggregate measure over all three matrices). Given the graphical representation in Figures 8a-c in Carroll et al. (1984), our results appear essentially identical to theirs. We do note, however, that the fitted proximities given below reflect the monotonicity condition imposed by the common ultrametric according to when the subsets are formed (and thus, some of the fitted

proximity values are tied). The hierarchies given by Carroll et al. (1984), merely use the three-point conditions imposed on the object triples, and thus, the order of when the subsets are formed may vary from hierarchy to hierarchy. In brackets next to the tied fitted values given below are those that would be generated from the common ultrametric structure but not including the monotonicity condition for subset formation. Using the latter, the VAF measures in each case would increase slightly (to 76.7471%, 53.769%, and 60.690%, respectively), and the actual subsets listed in the hierarchies are modified to reflect the different object set merges that depend on the use of the fitted values in brackets.

		Fitted Proximities		
Level	Partition	Headache	Fever	Muscle Aches
0	all objects separate	0.000	0.000	0.000
1	{Co,Va}	20.930	21.085[21.930]	20.030
2	{Co,Va},{BAs,Bu}	21.140	21.085[20.240]	23.130
3	{Co,Va},{An,BAs,Bu}	22.635	24.775	24.190
4	{As,Co,Va},{An,BAs,Bu}	22.760	25.520	24.600
5	{As,Co,Va},{An,BAs,Bu}, {Da,Ty}	25.498[26.610]	27.490	27.647[27.760]
6	{As,Co,Va},{An,BAs,Bu,Ex}, {Da,Ty}	25.498[25.127]	29.527	27.647[27.740]
7	{As,Co,HAs,Va}, {An,BAs,Bu,Ex},{Da,Ty}	28.307	29.587	27.647[27.517]
8	{As,Co,HAs,Va}, {An,BAs,Bu,Da,Ex,Ty}	32.065	32.243	32.530
9	all objects together	37.406	34.651	33.464

5. A Final Note

The previous sections have attempted to demonstrate the heuristic use of iterative projection with augmentation to search for and fit a variety of structural

representations to a given set of (two-mode or symmetric) proximity data. To date, no systematic evaluation has been made between the use of IPWA and all the various implementations of penalty function optimization strategies that have been suggested in the literature for some of these same tasks, although such comparisons are probably warranted. At the present time, it appears that a systematic comparison between IPWA and the penalty function alternatives on an extensive set of common data bases, will need to await a concerted research endeavor that actually recodes the various penalty function approaches mentioned in the literature since programs for the latter are not generally available in the public domain.

Prior to engaging in this extended set of comparisons in a consistent manner, what informal numerical information we do have must come from using some of the published results involving the fit of particular structural representations achieved for specific data sets according to, say, VAF measures, and what can be obtained with IPWA. In general, these types of comparisons suggest that IPWA is extremely competitive in being able to find the better-fitting (VAF) representations, and either equals or exceeds those values available in the literature for specific data sets. As one somewhat representative example, an additive tree was fitted in De Soete et al. (1984), to an 8×11 rectangular proximity matrix involving word association frequencies to stimulus phrases concerning shampoos, with the resulting additive tree having a VAF of 84.2%. In using IPWA through a two-vector decomposition as in Section 4.3, and considering the constrained additive tree to be the first component in the decomposition, only one local optimum was observed over 100 random orderings of constraint consideration with a VAF of 84.013% (a result marginally worse than given in De Soete et al., 1984); however, considering the constrained additive tree to be the second component again produced only a single local optimum but one with a VAF of 84.459%, and thus, a result somewhat better than reported in De Soete et al. (1984). Both the local optima located with IPWA were similar to that given by De Soete et al. (1984), at

least according to the implicit groupings of the stimulus phrases and word associations, but varied somewhat in connections between the internal nodes.

To provide somewhat more theoretical reasons that IPWA may be an attractive alternative compared to the use of penalty function algorithms, IPWA is very simple in design and conceptually easy to understand in how it can be used to approach a particlar heuristic optimization task. There is no need to use gradients or to construct unique penalty terms in imposing constraints, and in turn, no requirement exists for a process of double iteration, i.e., for the gradient optimization and then in a second (outer) successive strengthening of the penalty parameters. Since the inner gradient optimization process for a penalty function approach appears to be of the same order of magnitude as IPWA itself, there is an expectation that IPWA would be computationally much faster given the availability of comparisons based on comparable code, but obviously, such conclusions must await the more extensive research endeavor alluded to above. We might also note that IPWA is very easily expandable to include a wide variety of other closed convex constraints that may involve auxillary bounds or functional relationships among the parameters of the structures being identified, or for that matter, to various nonlinear constraints, e.g., ellipsoidal restrictions of the following form (see Han, 1988, pp. 11-12):

$$(\mathbf{x} - \mathbf{a})'\mathbf{A}(\mathbf{x} - \mathbf{a}) \leq b \quad ,$$

where $\mathbf{a}$ is an $n \times 1$ vector, $\mathbf{A}$ is an $n \times n$ positive semi-definite matrix, and b is a scalar.

In any expanded research effort to evaluate IPWA and the penalty function approach, a number of Monte Carlo assessments might be of value to develop in detail, and specifically, to evalute the ability of the least-squares loss function to serve as a reasonable mechanism to effect estimation and recovery. These Monte Carlo assessments, however, should be focused on the use of the loss function itself, and possibly in comparison with others, e.g., the sum of absolute

differences, or the minimization of the maximum difference, assuming that algorithms for the latter could be developed. The topic of recovery and its assessment through a Monte Carlo study, although of some obvious interest to pursue, is distinct from the development of a specific least-squares algorithm, and it might even be argued that doing such a Monte Carlo analysis would be somewhat peripheral in the context of demonstrating or developing the use of a particular algorithm across a variety of contexts (as we tried to do in the current paper). Any such Monte Carlo analysis should be directed to the method level as opposed to the algorithm level (to adopt a distinction made by Jardine and Sibson, 1971, pp. 42–44, from the cluster analysis context), and in a sense, should be invariant to the particular least-squares algorithm used, although possibly not to different methods based on the use of different loss functions.

References

Barthélemy, J.-P., & Guénoche, A. (1991). Trees and proximity representations. Chichester: Wiley.

Boyle, J. P., & Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert space. In R. L. Dykstra, T. Robertson, & F. T. Wright (Eds.), Advances in order restricted statistical inference (vol. 37, Lecture Notes in Statistics) (pp. 28–47). Berlin: Springer-Verlag.

Brossier, G. (1987). Étude des matrices de proximité rectangulaires en vue de la classification [A study of rectangular proximity matrices from the point of view of classification]. Revue de Statistiques Appliquées, 35, (4), 43–68.

Buck, R. C. (1965). Advanced calculus. New York: McGraw-Hill.

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

Carroll, J. D., Clark, L. A., & DeSarbo, W. S. (1984). The representation of three-way proximity data by single and multiple tree structure models. Journal of Classification, 1, 25–75.

Carroll, J. D., & Pruzansky, S. (1980). Discrete and hybrid scaling models. In E. D. Lantermann & H. Feger (Eds.), Similarity and choice (pp. 108–139). Bern: Huber.

Cheney, W., & Goldstein, A. (1959). Proximity maps for convex sets. Proceedings of the American Mathematical Society, 10, 448–450.

Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. Bulletin of Mathematical Biology, 49, 461–467.

De Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. Psychometrika, 48, 621–626.

De Soete, G. (1984a). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. Pattern Recognition Letters, 2, 133–137.

De Soete, G. (1984b). Ultrametric tree representations of incomplete dissimilarity

data. Journal of Classification, 1, 235-242.

De Soete, G. (1984c). Additive tree representations of incomplete dissimilarity data. Quality & Quantity, 18, 387-393.

De Soete, G., Carroll, J. D., & DeSarbo, W. S. (1987). Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data. Journal of Classification, 4, 155-173.

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

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. Journal of the American Statistical Association, 78, 839-842.

Eckes, T., & Orlik, P. (1993). An error variance approach to two-mode hierarchical clustering. Journal of Classification, 10, 51-74.

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

Gaffke, N., & Mathar, R. (1989). A cyclic projection algorithm via duality. Metrika, 36, 29-54.

Han, S. P. (1988). A successive projection method. Mathematical Programming, 40, 1-14.

Hutchinson, J. W. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. Psychometrika, 54, 25-51.

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

Klauer, K. C., & Carroll, J. D. (1989). A mathematical programming approach to fitting general graphs. Journal of Classification, 6, 247-270.

Klauer, K. C., & Carroll, J. D. (1991). A comparison of two approaches to fitting directed graphs to nonsymmetric proximity measures. Journal of Classification, 8, 251-268.

Krivánek, M. (1986). On the computational complexity of clustering. In E. Diday,

Y. Escoufier, L. Lebart, J. P. Pagès, Y. Schektman, & R. Tomassone (Eds.), Data analysis and informatics, IV (pp. 89-96). Amsterdam: North-Holland.

Krivánek, M., & Moravek, J. (1986). NP-hard problems in hierarchical-tree clustering. Acta Informatica, 23, 311-323.

Patrinos, A. N., & Hakimi, S. L. (1972). The distance matrix of a graph and its tree realization. Quarterly of Applied Mathematics, 30, 255-269.

Price, R. H., & Bouffard, D. L. (1974). Behavioral appropriateness and situational constraint as dimensions of social behavior. Journal of Personality and Social Psychology, 30, 579-586.

Rao, C. R. (1952). Advanced statistical methods in biometric research. New York: Wiley.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). Order restricted statistical inference. New York: Wiley.

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

ten Berge, J. M. F. (1991). A general solution for a class of weakly constrained linear regression problems. Psychometrika, 56, 601-609.

Tucker, L. R (1964). The extension of factor analysis to three-dimensional matrices. In N. Frederiksen & H. Guilliksen (Eds.), Contributions to mathematical psychology (pp. 109-127). New York: Holt, Rinehart, and Winston.

van der Lans, I. A. (1992). Nonlinear MVA for multiattribute preference data. Leiden: DSWO Press.

von Neumann, J. (1950). Functional operators (Vol. II). Princeton, NJ: Princeton University Press.

Wiener, N. (1955). On factorization of matrices. Commentarii Mathematici Helvetici, 29, 97-111.

Table 1. A Proximity Matrix Using Squared Mahalanobis Distances From Rao (1952) on Twelve Indian Groups (Upper-Triangle); the Lower-Triangular Portion Provides the Least-Squares Approximation Based on the Fixed Partition Hierarchy Given in the Text.

	1	2	3	4	5	6	7	8	9	10	11	12
1: B1	x	.270	3.480	2.230	2.860	4.450	3.050	2.860	1.170	1.480	2.130	3.300
2: B2	.270	x	3.610	1.630	2.810	3.820	2.870	2.620	.780	1.030	1.470	2.720
3: C1	3.721	3.721	x	1.320	4.520	5.080	5.250	4.460	2.680	2.980	3.350	4.200
4: C2	1.865	1.865	3.721	x	2.110	2.430	4.680	3.740	1.260	1.530	1.670	2.870
5: D	2.889	2.889	3.721	2.889	x	1.150	3.840	2.470	2.910	2.410	2.310	2.660
6: Bh	2.889	2.889	3.721	2.889	1.150	x	5.020	3.160	2.530	2.230	1.750	2.240
7: Ch	2.601	2.601	3.721	2.601	2.889	2.889	x	.400	3.380	2.120	2.720	2.240
8: M	2.601	2.601	3.721	2.601	2.889	2.889	.400	x	2.450	1.340	1.450	.900
9: A1	1.760	1.760	3.721	1.865	2.889	2.889	2.601	2.601	x	.300	.490	1.520
10:A2	1.760	1.760	3.721	1.865	2.889	2.889	2.601	2.601	.395	x	.120	.580
11:A3	1.760	1.760	3.721	1.865	2.889	2.889	2.601	2.601	.395	.120	x	.430
12:A4	1.760	1.760	3.721	1.865	2.889	2.889	2.601	2.601	.843	.843	.843	x

1: Brahmin (Basti, B1)
2: Brahmin (Other, B2)
3: Bhatu (C1)
4: Nabru (C2)
5: Dom (D)
6: Bhil (Bh)
7: Chattri (Ch)
8: Muslim (M)
9: Ahir (A1)
10: Kurmi (A2)
11: Other Artisan (A3)
12: Kahar (A4)

Table 2. The Upper-Triangular Portion of the Matrix Below Provides the Least-Squares Approximation to the Upper-Triangular Proximity Data of Table 1 Based on the Constraints Implied by the Ternary Tree of Figure 1.

	1	2	3	4	5	6	7	8	9	10	11	12
1: B1	x	.270	3.932	2.386	3.111	3.492	3.726	2.754	1.921	1.586	1.763	2.340
2: B2		x	3.567	2.021	2.746	3.127	3.361	2.389	1.556	1.221	1.398	1.975
3: C1			x	3.536	4.261	4.642	4.876	3.904	3.071	2.736	2.913	3.490
4: C2				x	2.715	3.096	3.330	2.358	1.525	1.190	1.367	1.944
5: D					x	1.150	4.055	3.083	2.250	1.915	2.092	2.669
6: Bh						x	4.436	3.464	2.631	2.296	2.473	3.050
7: Ch							x	.400	2.865	2.530	2.707	3.284
8: M								x	1.893	1.558	1.735	2.312
9: A1									x	.275	.452	1.029
10:A2										x	.117	.694
11:A3											x	.871
12:A4												x

Table 3. A Proximity Matrix From Shepard et al. (1975) on the Dissimilarity Between the First Ten Single-Digit Integers {0,1,...,9} Considered as Abstract Concepts; Averaged Over Raters and Conditions.

Integer	0	1	2	3	4	5	6	7	8	9
0	x	.421	.584	.709	.684	.804	.788	.909	.821	.850
1		x	.284	.346	.646	.588	.758	.630	.791	.625
2			x	.354	.059	.671	.421	.796	.367	.808
3				x	.413	.429	.300	.592	.804	.263
4					x	.409	.388	.742	.246	.683
5						x	.396	.400	.671	.592
6							x	.417	.350	.296
7								x	.400	.459
8									x	.392
9										x

Table 4. Frequency Distributions For the Least-Squares Loss Values Over 100 Random Starting Assignments of Object Triples/Quadruples to the Sets $C_1,\ldots,C_K$; WA Refers to Iterative Projection With Augmentation, and WOA Refers to Iterative Projection Without Augmentation.

Data Set:	Shepard et al.					Rao				
Structure:	Ultrametric		Additive Tree			Ultrametric		Additive Tree		
	Value	WA	WOA	Value	WA	WOA	Value	WA	WOA	Value	WA	WOA
	.49410	27	3	.63591	49	-	.56155	37	4	.89338	54	8
	.49156	-	9	.63521	-	2	.56120	2	2	.89324	34	9
	.48854	-	3	.63472	1	1	.56008	7	2	.89323	-	1
	.48537	-	3	.63428]	-	(5)	.55972	1	1	.89321	12	14
	.48208	-	4	.62714]			.55915	1	-	.89318	-	7
	.48192	-	1	.62486	43	2	.55704	11	1	.89312	-	4
	.47812	65	23	.62461	-	1	.55668	1	1	.89302	-	3
	.47768	3	-	.62376	-	1	.54774	-	1	.89292	-	4
	.47529	-	6	.62342	-	3	.54605	-	1	.89278	-	3
	.47510	-	2	.62154	4	2	.54346	1	-	.89274	-	8
	.47387	-	1	.62125]	-	(3)	.54323	1	-	.89272	-	1
	.47149	-	1	.62081]			.54074]	-	(11)	.89266	-	1
	.46939	2	12	.61879	-	2	.49629]			.89256	-	4
	.46895	1	-	.61627	-	2	.49555	35	32	.89159	-	2
	.46595	-	4	.61611	-	3	.49407	1	4	.89151	-	4
	.46361	-	1	.61405	-	2	.49146	1	15	.89136	-	1
	.46284	-	4	.61384	-	1	.49104	1	-	.89104	-	2
	.46241]	-	(4)	.61382	-	2	.48998	-	2	.89090	-	1
	.44796]			.61371	-	2	.48492	-	1	.89026	-	1
	.44572	2	-	.61334	-	1	.48080	-	3	.88979	-	2
	.43288	-	3	.61301	1	-	.48004	-	2	.88323]	-	(10)
	.43210	-	2	.61233	-	1	.47765	-	3	.87605]		
	.43198	-	3	.61111	-	1	.47192	-	1	.87602	-	2
	.42828]	-	(6)	.61097	2	-	.46964	-	3	.87599]	-	(8)
	.41564]			.61044	-	1	.45836	-	1	.86109]		
	.41563	-	2	.60901	-	2	.45583	-	3			
	.39790]	-	(3)	.60859]	-	(4)	.45449]	-	(6)			
	.33572]			.60789]			.39928]					
				.60761	-	2						
				.60673]	-	(4)						
				.60370]								
				.60367	-	2						
				.60273]	-	(6)						
				.59606]								
				.59548	-	2						
				.59311	-	1						
				.59298	-	2						
				.59285	-	1						
				.59157	-	2						
				.59137	-	1						
				.59114	-	2						
				.59021]	-	(12)						
				.58689]								
				.58667	-	2						
				.58401]	-	(16)						
				.52390]								
Nonconvergent Starts		25	-		1	-		93	-		24	-
Total Local Minima		6	32		6	80		13	37		3	39

Note: A set of bracketed values indicates a range of values for the loss function and for which the number of local minima in that range were observed uniquely.

Table 5. The Collection of Hierarchically Organized Subsets with Overlap of at Most One Object at Each Level (a 2-Ultrametric) for the Rao Proximity Matrix; This Result Corresponds to the Most Salient Local Optimum Identified.

Level	Subsets (with more than one object)	Fitted Proximity
0	all objects separate	.000
1	{10,11}	.120
2	{10,11},{1,2}	.270
3	{10,11},{1,2},{9,10}	.300
4	{10,11},{1,2},{9,10},{7,8}	.400
5	{10,11},{1,2},{9,10},{7,8},{11,12}	.430
6	{10,11},{1,2},{9,10},{7,8},{11,12},{9,11}	.490
7	{10,11},{1,2},{9,10},{7,8},{11,12},{9,11},{2,9}	.780
8	{10,11},{1,2},{9,10},{7,8},{11,12},{9,11},{2,9},{8,12}	.900
9	{9,10,11,12},{1,2},{7,8},{2,9},{8,12}	1.050
10	{9,10,11,12},{1,2},{7,8},{2,9},{8,12},{5,6}	1.150
11	{9,10,11,12},{1,2},{7,8},{2,9},{8,12},{5,6},{1,9}	1.170
12	{9,10,11,12},{1,2},{7,8},{2,9},{8,12},{5,6},{1,9},{4,9}	1.260
13	{9,10,11,12},{1,2},{7,8},{2,9},{8,12},{5,6},{1,9},{4,9},{3,4}	1.320
14	{8,9,10,11,12},{1,2},{7,8},{2,9},{5,6},{1,9},{4,9},{3,4}	1.747
15	{8,9,10,11,12},{1,2},{7,8},{2,9},{5,6},{1,9},{4,9},{3,4},{6,11}	1.750
16	{8,9,10,11,12},{1,2,4},{7,8},{2,9},{5,6},{1,9},{4,9},{3,4},{6,11}	1.930
17	{8,9,10,11,12},{1,2,4},{7,8},{2,9},{5,6},{1,9},{4,9},{3,4},{6,11},{4,5}	2.110
18	{1,2,4,8,9,10,11,12},{7,8},{5,6},{3,4},{6,11},{4,5}	2.280
19	{1,2,4,5,6,7,8,9,10,11,12},{7,8},{3,4}	2.806
20	{1,2,4,5,6,7,8,9,10,11,12},{3,4}	3.324
21	all objects together	3.961

Table 6. Two-Mode Proximity Data From Price and Bouffard (1974) on Ratings of Behavior Appropriateness in Certain Situations; Higher Values Refer to Greater Appropriateness.

	Behavior														
Situation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2.52	6.21	2.10	8.17	4.23	3.60	3.62	7.27	1.21	1.77	5.33	1.79	2.21	6.23	1.94
2	5.00	8.56	8.73	3.62	7.79	3.77	3.12	2.88	3.58	2.23	4.50	4.42	3.04	8.00	3.79
3	1.44	8.08	4.27	4.87	5.48	7.04	5.17	7.17	1.52	2.15	4.17	3.12	3.08	7.10	3.00
4	2.56	8.52	4.92	2.58	8.44	2.29	2.54	3.96	1.67	2.50	3.25	2.29	3.21	7.13	1.96
5	7.94	8.42	7.71	7.00	8.13	5.63	5.40	7.77	3.06	5.00	5.06	7.42	5.21	8.10	6.92
6	1.38	3.29	2.38	2.85	1.38	1.77	3.52	3.58	0.62	1.42	1.92	1.71	3.13	2.60	1.33
7	1.94	8.46	1.08	4.85	1.73	0.75	1.31	2.48	1.04	1.21	1.83	1.48	1.37	5.88	1.65
8	5.58	8.19	4.75	3.38	4.83	1.46	4.96	4.81	1.46	2.81	4.08	3.54	3.71	7.40	4.88
9	2.46	4.98	6.21	2.73	7.48	4.08	4.13	1.73	1.37	2.58	1.71	2.31	7.15	7.94	2.42
10	1.96	8.25	5.17	5.38	7.67	2.90	6.21	4.71	1.90	5.04	4.31	3.75	3.44	8.23	4.13
11	1.63	7.40	4.79	3.04	5.10	1.31	5.12	4.48	1.58	2.54	2.58	2.12	3.48	6.77	1.73
12	2.83	7.25	2.81	3.46	2.35	2.83	5.04	4.75	1.77	5.12	3.48	3.65	4.79	5.90	3.52
13	6.15	8.58	8.52	8.29	7.94	8.85	7.67	8.58	4.25	6.81	7.52	6.73	8.00	8.17	6.44
14	4.40	7.88	6.54	7.73	7.19	6.08	5.50	8.56	2.40	4.00	4.88	4.58	3.88	7.75	3.60
15	4.12	8.08	5.08	4.56	8.04	2.98	5.23	3.69	2.04	3.85	4.98	7.12	4.31	7.90	7.94

Situations	Behaviors
1: class	1: run
2: date	2: talk
3: bus	3: kiss
4: family dinner	4: write
5: park	5: eat
6: church	6: sleep
7: job interview	7: mumble
8: sidewalk	8: read
9: movies	9: fight
10: bar	10: belch
11: elevator	11: argue
12: restroom	12: jump
13: own room	13: cry
14: dorm lounge	14: laugh
15: football game	15: shout

Table 7. Best Ultrametric Representation Observed for the Data in Table 6 (VAF = 60.773%). The Numerical Entries Before a Semicolon Refer to Row Objects (Situations) and Those After Refer to Column Objects (Behaviors).

Level	Subsets (with more than one object)	Fitted Proximity
0	all objects separate	.000
1	{6;9}	.620
2	{6;9},{7;6}	.750
3	{6;9},{7;6,10}	1.210
4	{6;9,15},{7;6,10}	1.330
5	{6,7;6,9,10,15}	1.470
6	{6,7;6,9,10,12,15}	1.595
7	{6,7;6,9,10,12,15},{11;1}	1.630
8	{6,7;6,9,10,12,15},{11;1},{9;11}	1.710
9	{6,7,11;1,6,9,10,12,15},{9;11}	1.800
10	{1,6,7,11;1,6,9,10,12,15},{9;11}	2.138
11	{1,4,6,7,11;1,6,9,10,12,15},{9,11}	2.212
12	{1,4,6,7,9,11;1,6,8,9,10,11,12,15}	2.751
13	{1,4,6,7,9,11;1,6,8,9,10,11,12,15},{12;3}	2.810
14	{1,3,4,6,7,9,11;1,6,7,8,9,10,11,12,15},{12;3}	3.279
15	{1,3,4,6,7,9,10,11;6,7,8,9,10,11,12,15},{12;3}	3.429
16	{1,3,4,6,7,8,9,10,11;6,7,8,9,10,11,12,15},{12;3}	3.440
17	{1,3,4,6,7,8,9,10,11;6,7,8,9,10,11,12,15},{12;3},{2;4}	3.620
18	{1,3,4,6,7,8,9,10,11,12;1,3,6,7,8,9,10,11,12,13,15},{2;4}	3.745
19	{1,2,3,4,6,7,8,9,10,11,12;1,3,4,6,7,8,9,10,11,12,13,15}	4.280
20	{1,2,3,4,6,7,8,9,10,11,12,15;1,3,4,6,7,8,9,10,11,12,13,15}	4.658
21	{1,2,3,4,6,7,8,9,10,11,12,14,15;1,3,4,6,7,8,9,10,11,12,13,15}	5.179
22	{1,2,3,4,6,7,8,9,10,11,12,14,15;1,3,4,5,6,7,8,9,10,11,12,13,15}	5.516
23	{1,2,3,4,5,6,7,8,9,10,11,12,14,15;1,3,4,5,6,7,8,9,10,11,12,13,15}	6.327
24	{1,2,3,4,5,6,7,8,9,10,11,12,14,15;1,3,4,5,6,7,8,9,10,11,12,13,14,15}	6.924
25	{1,2,3,4,5,6,7,8,9,10,11,12,14,15;1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}	7.398
26	all objects together	7.500

Figure Captions

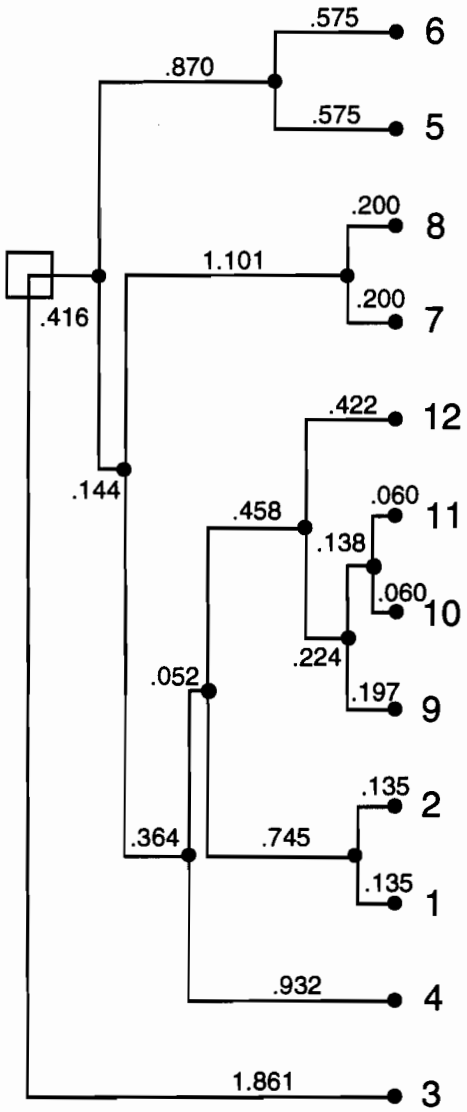
Figure 1. A graph-theoretic representation of the ultrametric given in the lower-triangular portion of Table 1. The sum of the indicated scalar values on the branches defining the unique path between two terminal nodes reconstructs the given ultrametric.

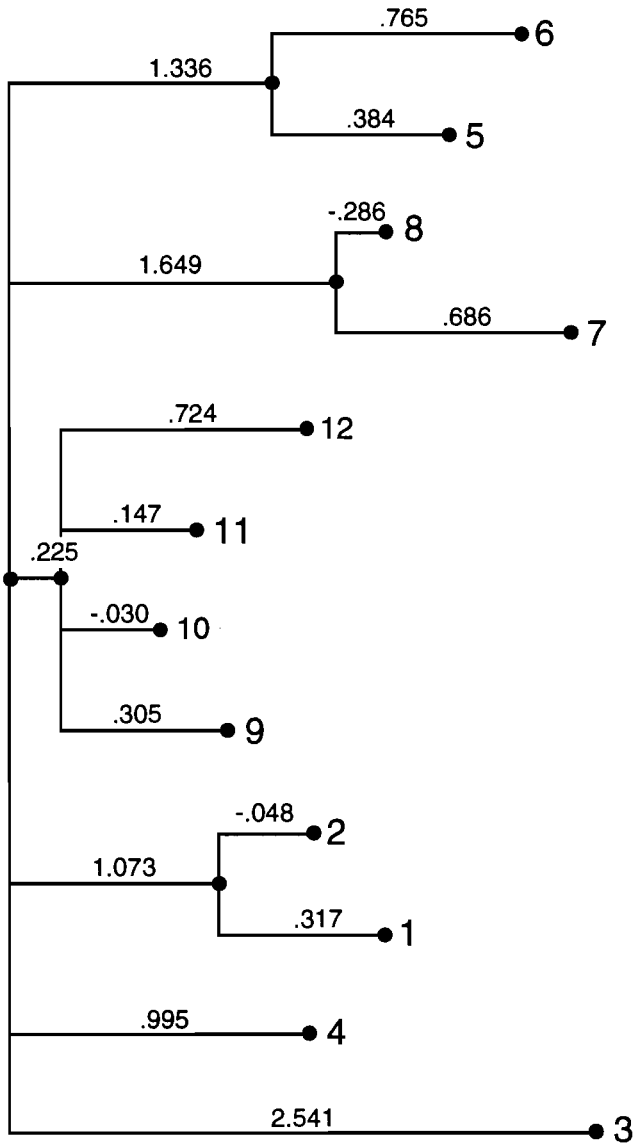
Figure 2. An explicit graph-theoretic representation of Table 2 providing the corresponding additive tree with branch scalars attached based on the constraints implied by the general form of the ternary tree of Figure 1. The actual branch lengths drawn in Figure 2 connecting terminal nodes are proportional to the given values but first augmented by an additive constant of .500.

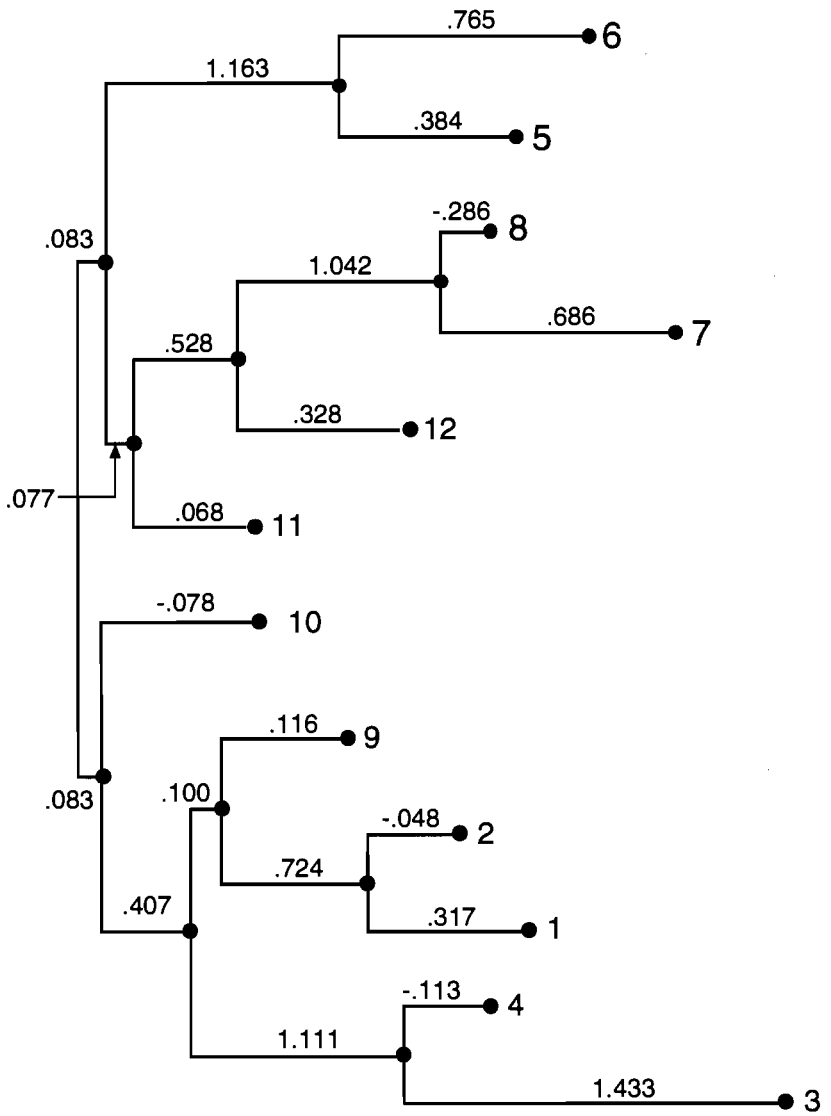
Figure 3. Two additive tree representations for the Rao data corresponding to the two most salient local optima identified with IPWA. Panel (a) corresponds to a VAF of 89.338% and panel (b) corresponds to a VAF of 89.324%. The actual branch lengths drawn in Figure 3 connecting terminal nodes are proportional to the given values but first augmented by an additive constant of .500.

Figure 4. Two additive tree representations for the Shepard et al. data corresponding to the two salient local optima identified with IPWA. Panel (a) corresponds to a VAF of 63.591% and panel (b) corresponds to a VAF of 62.486%.

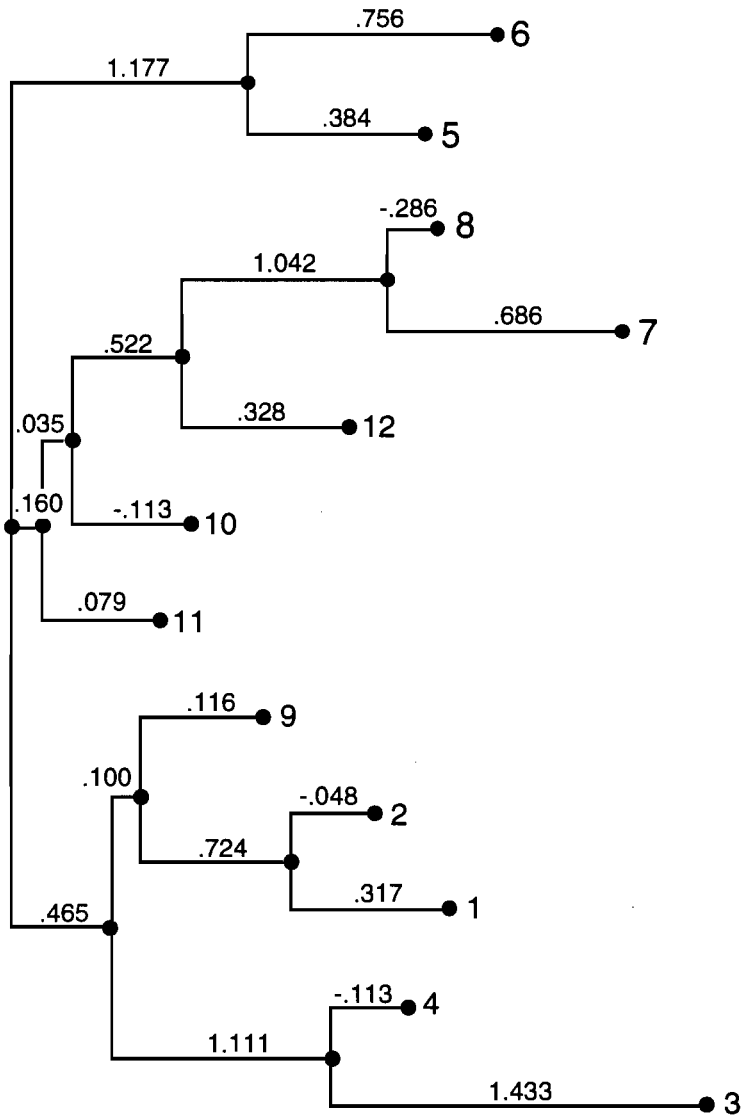
Figure 5. An additive tree representation for the data of Table 6 corresponding to the best VAF value identified with IPWA (VAF of 81.964%). The actual branch lengths drawn in Figure 5 connecting terminal nodes are proportional to the given values but first augmented by an additive constant of 1.5000. The numerical labels for the behaviors are underlined in the figure; those for situations are not underlined.


.575
6
.870
.575
5
.200
8
1.101
.200
7
.416
.422
12
.144
.060
11
.458
.138
.060
10
.224
.052
.197
9
.135
2
.364
.745
.135
1
.932
4
1.861
3

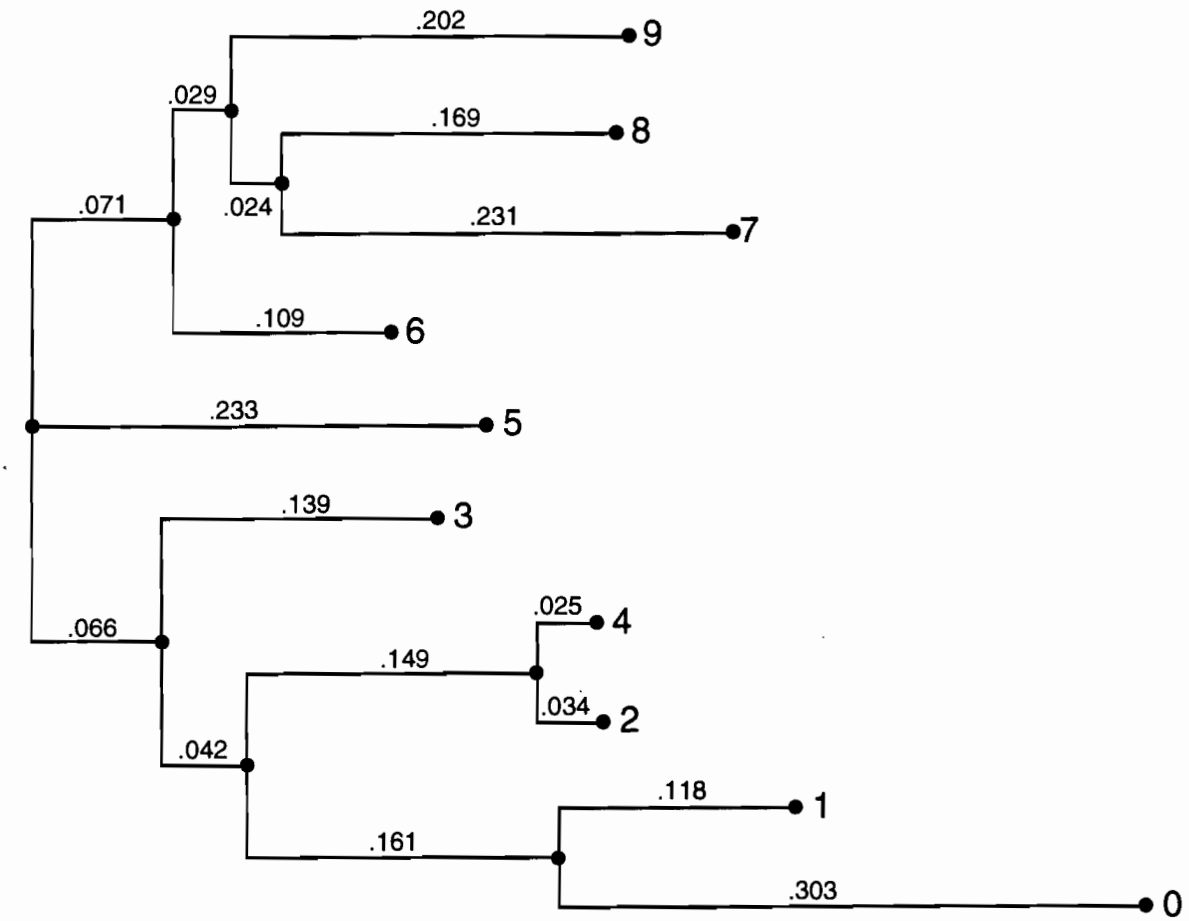

.765
6
1.336
.384
5
-.286
8
1.649
.686
7
.724
12
.147
11
.225
-.030
10
.305
9
-.048
2
1.073
.317
1
.995
4
2.541
3


.765
6
1.163
.384
5
-.286
8
.083
1.042
.686
7
.528
.328
12
.077
.068
11
-.078
10
.116
9
.083
.100
-.048
2
.724
.407
.317
1
-.113
4
1.111
1.433
3

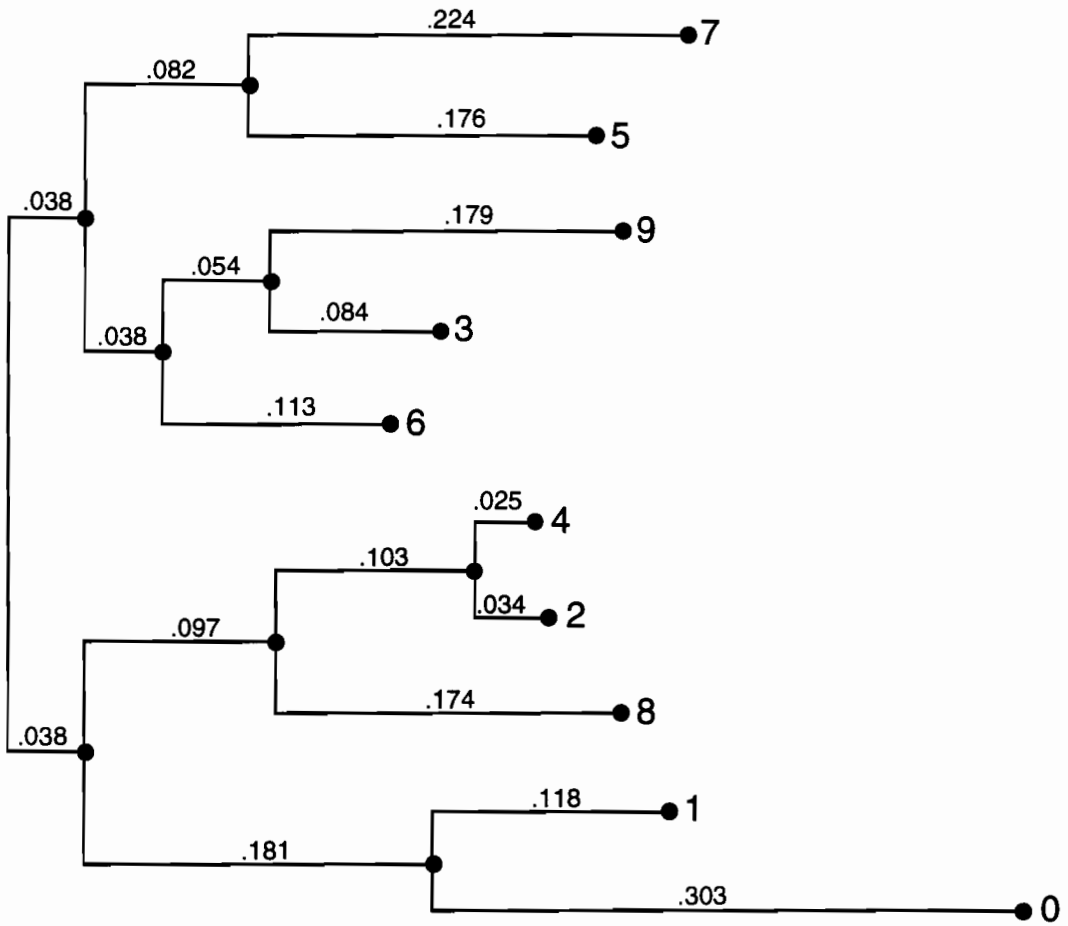
(a)


.756
6
1.177
.384
5
-.286
8
1.042
.686
7
.522
.328
12
.035
.160
-.113
10
.079
11
.116
9
.100
-.048
2
.724
.317
1
.465
-.113
4
1.111
1.433
3

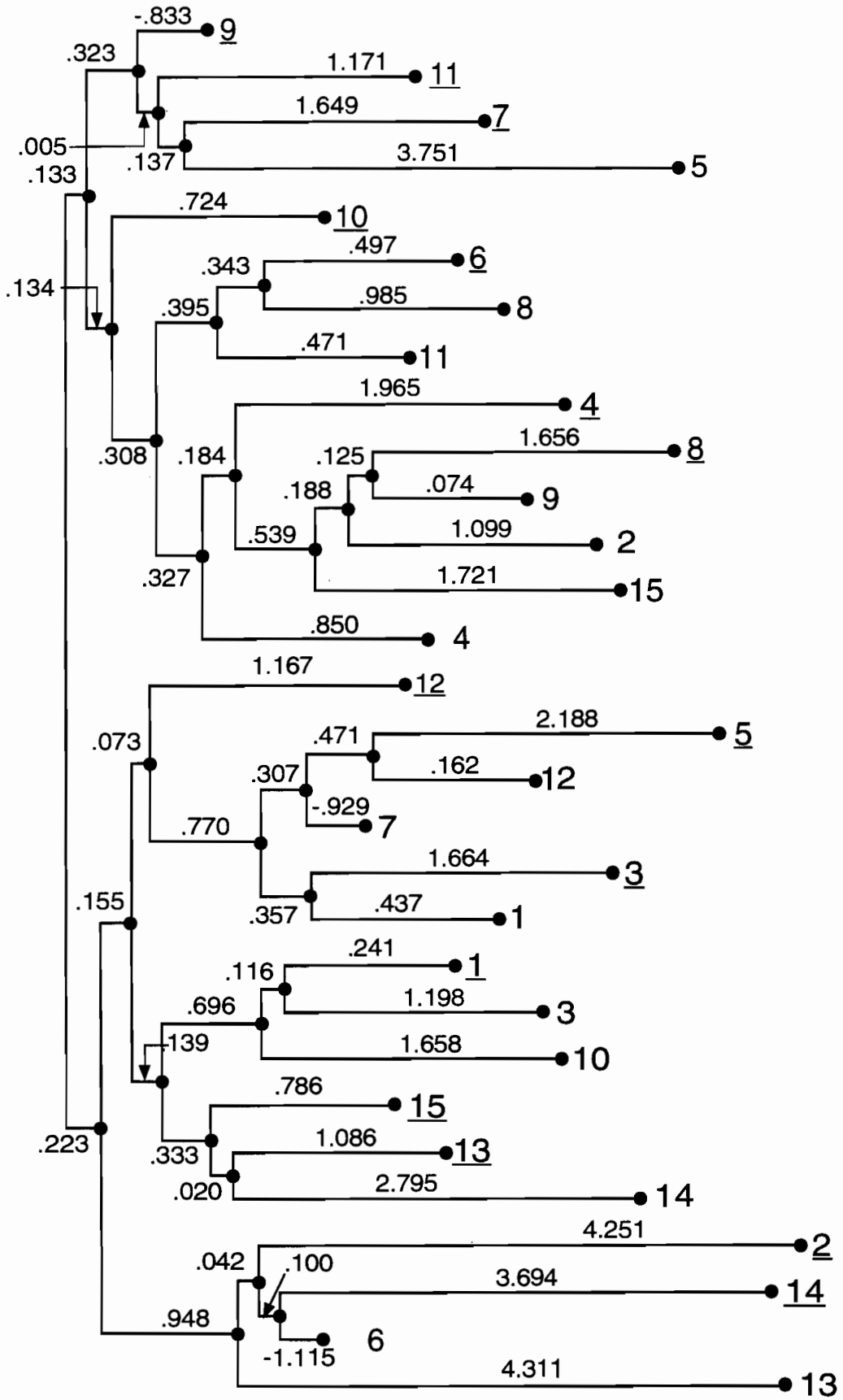
(b)


.202
9
.029
.169
8
.071
.024
.231
7
.109
6
.233
5
.139
3
.025
4
.066
.149
.034
2
.042
.118
1
.161
.303
0

(a)


.224
7
.082
.176
5
.038
.179
9
.054
.084
3
.038
.113
6
.025
4
.103
.034
2
.097
.174
8
.038
.118
1
.181
.303
0

(b)


-.833
9
.323
1.171
11
1.649
7
.005
.137
3.751
5
.133
.724
10
.497
.343
6
.134
.985
.395
8
.471
11
1.965
4
1.656
.308
.184
.125
8
.074
.188
9
1.099
.539
2
.327
1.721
15
.850
4
1.167
12
.073
2.188
.471
5
.162
.307
12
-.929
7
.770
1.664
3
.155
.357
.437
1
.241
.116
1
1.198
.696
3
.139
1.658
10
.786
15
.223
1.086
.333
13
2.795
.020
14
4.251
2
.042
.100
3.694
14
.948
6
-1.115
4.311
13