*$L_1$-Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

# Hierarchical clustering and the construction of (optimal) ultrametrics using $L_p$-norms

**Lawrence Hubert**

*University of Illinois, Champaign, USA*

**Phipps Arabie**

*Rutgers University, Newark, New Jersey, USA*

**Jacqueline Meulman**

*Leiden University, Leiden, The Netherlands*

*Abstract*: The classification task of hierarchical clustering can be characterized as one of constructing for an object set $S$ a sequence of successively less-refined partitions that attempts to represent the pattern of entries in a given symmetric proximity matrix defined between the objects. We discuss this process of constructing a partition hierarchy by the fitting through an $L_p$-norm (for $p = 1, 2, \text{or } \infty$) of a second symmetric matrix whose entries represent what is called an ultrametric and which can be used to induce a partition hierarchy. A dynamic programming strategy, and a heuristic extension for larger object sets, is suggested as the computational mechanism for carrying out the procedure of combinatorial search for the ultrametric that is the best-fitting according to the chosen $L_p$-norm. A numerical example is used to illustrate the complete fitting process that relies on a proximity matrix provided. A final extension is presented for the construction of best-fitting ultrametrics based on two-mode proximity data defined between distinct object sets.

*Key words*: Ultrametric, $\mathrm{L}_p$-norm, hierarchical clustering, dynamic programming, partitioning.

AMS subject classification: Primary 62H30, 92G30; secondary 90C39.

# 1 Introduction

One of the most studied data analysis topics in the field of classification is that of constructing a hierarchical clustering for an object set, $S = \{O_1, \ldots, O_n\}$, based on some given $n \times n$ symmetric proximity matrix $\mathbf{P} = \{p_{ij}\}$; an entry $p_{ij}$ $(= p_{ji} \geq 0$, and $p_{ii} = 0)$ is assumed to represent the dissimilarity of the objects $O_i$ and $O_j$, where larger values correspond to the more dissimilar objects. A hierarchical clustering of $S$ can be represented by a sequence of partitions, $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_T$, where $\mathcal{P}_1$ is the (disjoint) partition in which each object forms its separate class, $\mathcal{P}_T$ is the (conjoint) partition containing all objects in $S$ within a single class, and $\mathcal{P}_t$ is constructed by uniting two or more classes in $\mathcal{P}_{t-1}$. (Most commonly, only one pair of classes will be united in $\mathcal{P}_{t-1}$, so that $T = n$ and $\mathcal{P}_t$ therefore includes $n - t + 1$ classes.) The task of hierarchical clustering is typically carried out by a greedy optimization strategy, which begins with $\mathcal{P}_1$ and successively identifies $\mathcal{P}_t$ from $\mathcal{P}_{t-1}$, for $t \geq 2$, by minimizing some chosen measure of proximity between the subsets that could be united to form a new class in $\mathcal{P}_t$. Most commercially available statistical software packages (e.g., SYSTAT, SPSS, and SAS) implement their routines for hierarchical clustering in this manner and with various choices for how the proximity between subsets might be defined.

The present paper is concerned with this particular problem of constructing a partition hierarchy that is intended to represent the patterning of relationships present in the proximity matrix $\mathbf{P}$, but will do so indirectly by fitting a second matrix to $\mathbf{P} = \{p_{ij}\}$, denoted by $\mathbf{U} = \{u_{ij}\}$, minimizing an $L_p$-norm (for one of the usual values chosen for $p$ of 1, 2, or $\infty$). The entries in the fitted matrix $\mathbf{U}$ will satisfy a collection of linear inequality/equality constraints, characterizing what is called an ultrametric, that in turn can be used to retrieve a specific partition hierarchy for the object set $S$. The fitting task itself will be carried out through a recursive optimization strategy based on dynamic programming which for small object sets can provide globally optimal solutions. Later sections of the paper discuss the heuristic use of the same dynamic programming strategy for dealing with larger object sets, and an extension of the hierarchical clustering task for proximity matrices that only contain dissimilarity values between the objects from two distinct sets (i.e., two-mode proximity data).

## 2 Ultrametrics

A concept routinely encountered in formal discussions of hierarchical clustering is that of an ultrametric, which can be characterized by any nonnegative $n \times n$ symmetric dissimilarity matrix for the objects in $S$, denoted generically as $\mathbf{U} = \{u_{ij}\}$, where $u_{ij} = 0$ if and only if $i = j$ and the entries in $\mathbf{U}$ satisfy the ultrametric inequality: $u_{ij} \leq \max\{u_{ik}, u_{jk}\}$ for $1 \leq i, j, k \leq n$. An alternative characterization of this last inequality would be that for all distinct object triples, $O_i$, $O_j$, and $O_k$, the largest two dissimilarities among $u_{ij}$, $u_{ik}$, and $u_{jk}$ are equal and (therefore) not smaller than the third. Any ultrametric identifies a specific partition hierarchy, $\mathcal{P}_1, \ldots, \mathcal{P}_T$, where those object pairs defined between subsets united in $\mathcal{P}_{t-1}$ to form $\mathcal{P}_t$ all have a common ultrametric value that is not smaller than those for object pairs defined within these same subsets. Thus, the individual partitions in the sequence can be identified by increasing a threshold variable from zero and observing that $\mathcal{P}_t$ is associated with a particular threshold value where all dissimilarities within a class in $\mathcal{P}_t$ are less than or equal to this threshold and all dissimilarities between the classes in $\mathcal{P}_t$ are strictly greater. Conversely, the collection of all ultrametric matrices can be decomposed into equivalence classes where all members of an equivalence class induce the same partition hierarchy. If $\mathcal{P}_1, \ldots, \mathcal{P}_T$ denotes the specific partition hierarchy induced by all members of an equivalence class, we will refer to one particular member of this class as the *base ultrametric* defined by $\mathbf{U}^o = \{u_{ij}^o\}$, where $u_{ij}^o = \min\{t - 1 \mid \text{objects } O_i \text{ and } O_j \text{ appear within the same class in partition } \mathcal{P}_t\}$. All members of an equivalence class can be obtained from the entries for the base ultrametric by a strictly monotonic function that maps zero to zero. Moreover, since $\mathbf{U}^o$ contains $T - 1$ distinct positive values, each member of this equivalence class will also contain $T - 1$ distinct positive values, where the $(t-1)^{st}$ largest corresponds to partition $\mathcal{P}_t$ in the hierarchy and is implicitly associated with those object pairs that appear together for the first time within a subset in $\mathcal{P}_t$.

An ultrametric matrix is a convenient device for representing in matrix form the partition hierarchy it induces, and specifically, the integer-valued base ultrametric can serve as a direct way for generating the explicit set of linear inequality/equality constraints that any ultrametric within an equivalence class must satisfy. Thus, one could find a best-fitting ultrametric within an equivalence class by fitting $\{u_{ij}\}$ to the original proximity matrix $\{p_{ij}\}$ through, for example, an $L_p$-norm regression strategy that incorporates the linear inequality/equality constraints implied by the base ultrametric (e.g., those in Späth, 1991, Chapter 5). It is also possible to use the ultrametric notion more fundamentally as the basic mechanism for obtaining a partition hierarchy in the first place. Explicitly, we suggest here the development of hierarchical clustering methods by directly attempting to find a best-fitting ultrametric for **P** by optimizing a loss criterion defined by an $L_p$-norm between $\{p_{ij}\}$ and a (to be identified) ultrametric matrix $\{u_{ij}\}$. This usage of an $L_p$-norm is more general than what has been done thus far in the literature; the extant methods that attempt directly to obtain a best-fitting ultrametric have all adopted a least-squares criterion and some auxiliary search strategy for locating an appropriate set of constraints to impose (e.g., see Hartigan, 1967; Carroll and Pruzansky, 1980; De Soete, 1984; Chandon and De Soete, 1984; Hubert and Arabie, 1995).

To be specific, suppose for a given partition hierarchy, $\mathcal{P}_1, \ldots, \mathcal{P}_n$ (so $T = n$), we let $C_{t-1}^{(u)}$ and $C_{t-1}^{(v)}$ denote the two classes united in $\mathcal{P}_{t-1}$ to form $\mathcal{P}_t$, and specify $b_{t-1}$ to be some appropriate aggregate (or 'average') value of the proximities for object pairs between $C_{t-1}^{(u)}$ and $C_{t-1}^{(v)}$. Denoting the set of proximities between $C_{t-1}^{(u)}$ and $C_{t-1}^{(v)}$ as $B_{t-1}(u,v) \equiv \{p_{i'j'} \mid O_{i'} \in C_{t-1}^{(u)},\ O_{j'} \in C_{t-1}^{(v)}\}$, and depending on the $L_p$-norm chosen, this between-subset aggregate value will be variously defined as the median ($L_1$), the mean ($L_2$), or the average of the maximum and the minimum proximities ($L_\infty$) in the set $B_{t-1}(u,v)$. The loss functions based on an $L_p$-norm used to index the adequacy of a given partition hierarchy in producing an ultrametric fitted to **P** are for the

$L_1$-norm:

$$\sum_{t=2}^{n} \sum_{O_{i'} \in C_{t-1}^{(u)},\ O_{j'} \in C_{t-1}^{(v)}} \mid p_{i'j'} - b_{t-1} \mid,$$

where $b_{t-1}$ is the median proximity in the set $B_{t-1}(u,v)$;

$L_2$-norm:

$$\sum_{t=2}^{n} \sum_{O_{i'} \in C_{t-1}^{(u)},\ O_{j'} \in C_{t-1}^{(v)}} (p_{i'j'} - b_{t-1})^2,$$

where $b_{t-1}$ is the mean proximity in the set $B_{t-1}(u,v)$;

$L_\infty$-norm:

$$\sum_{t=2}^{n} \max_{O_{i'} \in C_{t-1}^{(u)},\ O_{j'} \in C_{t-1}^{(v)}} \mid p_{i'j'} - b_{t-1} \mid,$$

where $b_{t-1}$ is the average of the minimum and maximum proximities in the set $B_{t-1}(u,v)$.

For all three $L_p$-norms, an optimal ultrametric will be one for which the order constraint on the between-subset aggregate values holds: $b_1 \leq$

$b_2 \leq \cdots \leq b_{n-1}$, and the norm is minimized. For such an optimal solution, the between-subset aggregate values, $b_1, \ldots, b_{n-1}$, define the distinct entries in an (optimal) fitted ultrametric. (It might be noted that since less than $n-1$ distinct values could be identified if some of the between-subset aggregate values are tied, the search for an optimal ultrametric can assume without loss of generality that $T = n$, and for $t \geq 2$, only two classes are united within $\mathcal{P}_{t-1}$ to form $\mathcal{P}_t$. Also, as a technical convenience, we allow the possibility that some of the between-subset aggregate values may be identically zero when the proximities for calculating these are all zero. Although not technically an ultrametric since zero ultrametric values should not correspond to distinct objects, its structure would still satisfy the central ultrametric inequality for distinct object triples.)

# 3 A dynamic programming strategy for identifying (optimal) ultrametrics

The optimization task of constructing optimal ultrametrics fitted to a given proximity matrix $\mathbf{P}$ may be fairly easy to state, but the problem itself is a computationally very difficult one to solve. For both the $L_1$- and $L_2$-norm, for instance, the task has been shown to fall into the class of NP-hard problems (see Křivánek and Morávek, 1986; Křivánek, 1986; for a recent comprehensive review, see Day, 1996); thus, there is the usual expectation that for larger object sets, methods guaranteeing optimality would become computationally infeasible to implement. Keeping these computational difficulties in mind, along with the eventual necessity of moving to heuristic methods of solution for larger object sets, we will still begin with a strategy that can fit an optimal ultrametric to $\mathbf{P}$ for each of the three $L_p$-norms introduced in the last section. The approach suggested is based on dynamic programming and the construction of a recursive system that will eventually produce an optimal solution. There are some complications that arise in the use of a straightforward dynamic programming formulation because of the need to impose an order constraint on the successive between-subset aggregate values, and these difficulties will be addressed below in some detail. In addition, a strategy for heuristically extending the basic dynamic programming formulation is developed in the next subsection for dealing with large(r) object set sizes.

## 3.1 Identifying optimal ultrametrics

To implement a dynamic programming approach for locating an optimal ultrametric, we first define a collection of sets, $\Omega_1, \ldots, \Omega_n$, where $\Omega_k$ contains *all* partitions of the $n$ objects in $S$ into $n-k+1$ classes. For convenience, a member of $\Omega_k$ is denoted by $A_k$; thus, $\Omega_1$ contains the single partition $A_1$ that has $n$ classes in which each of the $n$ objects forms a separate class, and $\Omega_n$ contains the single partition $A_n$ that includes one class for all of the $n$ objects in $S$. We will say that a transition from $A_{k-1} \in \Omega_{k-1}$ to $A_k \in \Omega_k$ is *permissible* if the union of two classes in $A_{k-1}$ produces $A_k$, and if an admissibility criterion to be discussed shortly is satisfied (that would [hopefully] ensure that the sequence of between-subset aggregate values is nondecreasing). A function $\mathcal{F}(A_k)$ for $A_k \in \Omega_k$ is defined as the optimal value for the sum of the contributions for the chosen $L_p$-norm up to the partition $A_k$. Beginning with $\mathcal{F}(A_1) \equiv 0$ for $A_1 \in \Omega_1$, we construct $\mathcal{F}(A_k)$ recursively by

$$\mathcal{F}(A_k) = \min \{\mathcal{F}(A_{k-1}) + C(A_{k-1}, A_k)\},$$

where the minimum is taken over all $A_{k-1} \in \Omega_{k-1}$ for which a transition is permissible to $A_k \in \Omega_k$, and $C(A_{k-1}, A_k)$ is the incremental cost of transforming $A_{k-1}$ to $A_k$ characterized by the appropriate $L_p$-norm when that pair of subsets in $A_{k-1}$ is united to form $A_k$. (It is this latter independence of incremental cost from how $A_{k-1}$ was obtained that is crucial to proving the validity of the recursive process.) Finally, an optimal solution is identified by $\mathcal{F}(A_n)$ for the single entity $A_n \in \Omega_n$, and a partition hierarchy attaining this optimal value identified by working backwards through the recursion starting from $\Omega_n$ and proceeding to $\Omega_1$ and tracing the process of how $\mathcal{F}(A_n)$ was generated.

One unresolved issue needing discussion is the explicit imposition of some type of admissibility criterion for defining a permissible transition from $A_{k-1}$ to $A_k$ that could ensure a nondecreasing sequence of between-subset aggregate values. Unfortunately, the validity of the recursive process depends on the property that any proposed criterion for admissibility must only involve $A_{k-1}$ and $A_k$ and their relation to the matrix $\mathbf{P}$, and specifically *not* on how $A_{k-1}$ may have been arrived at. Thus, it is not possible to define admissibility directly by requiring the between-subset aggregate value that defines $A_k$ from $A_{k-1}$ to be greater than or equal to the last between-subset aggregate value that led to $A_{k-1}$ from $A_{k-2}$. What can be offered, however, are two (less-than-ideal) alternatives: (a) an admissibility criterion based only on $A_{k-1}$ and $A_k$ that may sometimes be too lenient and thus fail to ensure that the collection of between-subset aggregate values are nondecreasing for the (purportedly optimal) identified ultrametric, or (b) an admissibility criterion based only on $A_{k-1}$ and $A_k$ that may be too strict, and the (purportedly optimal) identified ultrametric could in fact not be the absolute best obtainable.

To be specific, the possibly too lenient criterion rests on the observation (made originally by Chandon, Lemaire, and Pouget, 1980, for the $L_2$-norm) that in an optimal ultrametric based on any of the three $L_p$-norms (with the notion of an aggregate value defined by the median, mean, or the average of the two extreme proximities), the nondecreasing constraint on the between-subset aggregate values, $b_1 \leq \cdots \leq b_{n-1}$, requires that $b_t$ be both greater than or equal to each such aggregate value calculated within a subset of $A_{k-1}$, and less than or equal to the aggregate value of all proximities between the subsets in $A_k$. Since these two conditions may be evaluated given only $A_{k-1}$ and $A_k$, they can be imposed in defining whether a transition from $A_{k-1}$ to $A_k$ is permissible. Alternately, the possibly too strict admissibility criterion would require that $b_t$ be less than or equal to any between-subset aggregate value calculated for the new subset formed in $A_k$ and some other subset present in $A_k$. This latter criterion would ensure that no (nontrivial) order inversions in the sequence of between-subset aggregate values would exist (a trivial inversion would be one in which an inversion may be present in the collection of between-subset aggregate values, but it can be removed by a simple reordering of when two disjoint subsets are formed).

The computer program relied on for the numerical examples in Section 4 allows the imposition of either of the two admissibility criteria discussed above. As a suggested analysis strategy, one would begin with the former (and possibly too lenient) admissibility criterion and if no nontrivial order inversions in the between-subset aggregate values are found, an optimal ultrametric has been identified. If nontrivial order inversions were present, the possibly too strict admissibility criterion could be adopted, and the then identified ultrametric presumed optimal (but with the caveat that it could be possible in some [rare] instances for an even better ultrametric to be generated). (For convenience of reference, the program we use is referred to by the acronym HPHI, for '*H*euristic *P*rogramming *HI*erarchical clustering', where the term 'heuristic' is included because of the extensions it includes for dealing with larger object sets, as discussed in the section to follow.)

## 3.2 Heuristic extensions for large(r) object sets

When the number of objects in $S$ is even moderate in size, the random access memory storage requirements necessary for a dynamic programming approach to constructing an optimal ultrametric can become quite large. Necessary for implementing the proposed recursive strategy is the availability of large arrays associated with the sets, $\Omega_1, \ldots, \Omega_n$, that contain for *all* partitions of $S$ the recursively-constructed values $\mathcal{F}(A_k)$ for $A_k \in \Omega_k$, as well as a mechanism for keeping track of what previous partitions in $\Omega_{k-1}$ led to these optimal values $\mathcal{F}(A_k)$.[1] For larger object sets, HPHI allows two options: (a) finding optimal ultrametrics for subsets of $S$, and (b) finding optimal ultrametrics when the basic objects to be hierarchically partitioned are themselves subsets of $S$. By the judicious and repeated use of these two options, we have been able to approach object sets with reasonably large sizes (and will do so for an object set of size 30 in the next section).

The analysis strategy we suggest begins by identifying [possibly through a heuristic mechanism] a partition of $S$, say $\mathcal{P}_e$, that is initially forced to be induced as part of the best-fitting ultrametric we construct. The classes of $\mathcal{P}_e$ are first treated as the basic objects on which an ultrametric is to be obtained, i.e., we begin with the classes of $\mathcal{P}_e$ and complete the identification of an optimal ultrametric from this point on. Secondly, each of the classes of $\mathcal{P}_e$ is then used to obtain a separate optimal ultrametric for the objects in that class. When these results are concatenated, an optimal ultrametric is identified, subject to the constraint that $\mathcal{P}_e$ is part of the partition hierarchy it induces. Obviously, if $\mathcal{P}_e$ is chosen appropriately to begin with, the concatenated results would be optimal for the complete object set $S$. A check on the choice of $\mathcal{P}_e$ (however it was obtained initially) can be carried out by using object classes identified within the subsets defining $\mathcal{P}_e$ as the basic units on which an optimal ultrametric is to be constructed and then completing the fitting from this point on. If $\mathcal{P}_e$ is retrieved as part of this latter process, some obviously increased confidence is obtained that the concatenated ultrametric may be the best we can find. If, on the other hand, $\mathcal{P}_e$ is not retrieved, we could then repeat this same strategy with whatever partition was observed (presumably for the same number of classes as contained in $\mathcal{P}_e$). This whole process could be carried out iteratively until convergence. Obviously, an absolute guarantee of optimality is not possible through this type of heuristic search, but the eventual stability achieved leads to an ultrametric that is usually very good (although not verifiably optimal). Throughout this discussion it is assumed that the subsets of objects for which separate optimal ultrametrics are generated, or the number of object classes to be used in obtaining an optimal ultrametric beginning from that point, are all of a size that could be handled optimally (i.e., some number in the lower teen's).

[1] Given the usual Pentium-level processors now commonly available and the amount of memory these systems typically contain, the program we have developed can deal (optimally) with object set sizes in the lower teen's, but even this requires the capability of Fortran90 to allocate very large arrays dynamically (and inform the user whether sufficient memory exists on the system to solve the problem of the size being requested).

## 4 A numerical illustration

To illustrate the construction of best-fitting ultrametrics based on the $L_p$-norm for a given proximity matrix, we use a data set originally collected by Arabie and Rips (1973) for a replication of a study initially conducted by Henley (1969) involving the subjectively-judged similarity of 30 animals. Fifty-three subjects assessed the similarity between all 435 animal pairs based on a scale from 1 (extremely dissimilar) to 10 (extremely similar). Table 1 provides the animal names and the summed ratings over the subjects subtracted from the maximum of 530 so the proximities would be keyed as dissimilarities. (We provide these data in Table 1 as a convenience to others who may wish to use this proximity matrix in their own methodological examples. Although these data have been analyzed elsewhere (see e.g., De Soete and Carroll, 1996), they have not been published explicitly.)

Based on the data of Table 1, the results are presented below for each of the three $L_p$-norms using the heuristic process of Section 3.1 for finding best-fitting ultrametrics. Specifically, a five-class partition, $\mathcal{P}_e$, of the object set $S$ was first identified heuristically (the greedy complete-link hierarchical clustering method was used up to the level of five classes). An (optimal) ultrametric was then found for each of the five classes within $\mathcal{P}_e$, and based on these separate ultrametrics, a collection of (smaller) object subsets identified and treated as the starting point from which to finish the identification of an ultrametric for the complete object set $S$. Based on this latter ultrametric, the object classes for the induced five-class partition were then considered as defining an initial partition, $\mathcal{P}_e$, and the whole procedure repeated. For all three $L_p$-norms, the latter five-class partitions were retrieved immediately. In all of these analyses, and as suggested in the last section, the admissibility criterion that may at times be too lenient (to ensure a strictly nondecreasing between-subset collection of aggregate values) was first used, and when nontrivial order inversions were observed (as they were for a few of the analyses carried out), the more strict admissibility condition was then adopted.

The results for both the $L_1$- and $L_2$-norm are very similar, and the same five-class partition was induced for the corresponding ultrametrics:

A: {bear (2), cat (5), dog (10), fox (13), leopard (18), lion (19), tiger (28), wolf (29)} — carnivorous feline/canine animals plus the omnivorous bear

B: {beaver (3), chipmunk (7), mouse (21), rabbit (23), raccoon (24), rat (25), squirrel (27)} — small rodent-like animals

C: {antelope (1), camel (4), cow (8), deer (9), donkey (11), elephant (12), giraffe (14), goat (15), horse (17), sheep (26), zebra (30)} — large

hoofed herbivores (ungulates)
D: {chimpanzee (6), gorilla (16), monkey (20)} — primates
E: {pig (22)} — *Suidae*

The (optimal) ultrametrics defined by the values of the between-subset aggregate values for the $L_1$- and $L_2$-norm constructed within each of the classes labeled above as A, B, C, and D are given below (we also present those for the $L_\infty$-norm in the case of the two classes labeled B and D that were also observed in the retrieved ultrametric using this latter norm). Within each class we also provide a summary measure of the discrepancy between the proximities and fitted values by giving the contribution each class has to the overall $L_p$-norm measure being minimized.

| *level* | *new class formed* | $L_1$ | $L_2$ | $L_\infty$ |
|---|---|---|---|---|
| A:8 | {2, 5, 10, 13, 18, 19, 28, 29}<br>an addition of omnivorous bear to the carnivorous feline plus canine classes | 282.0 | 278.0 | |
| A:7 | {5, 10, 13, 18, 19, 28, 29}<br>the union of the carnivorous feline and canine classes | 217.0 | 222.6 | |
| A:6 | {10, 13, 29}<br>canines | 104.5 | 104.5 | |
| A:5 | {5, 18, 19, 28}<br>felines | 57.0 | 61.3 | |
| A:4 | {13, 29}<br>nondomestic canines | 57.0 | 57.0 | |
| A:3 | {18, 19, 28}<br>nondomestic felines | 35.0 | 35.0 | |
| A:2 | {18, 28}<br>feline(subclass) | 24.0 | 24.0 | |
| A:1 | (all separate) | – | – | |
| | contribution to the norm measures: | 557.0 | 22,100. | |
| B:7 | {3, 7, 21, 23, 24, 25, 27} | 230.5 | 212.3 | 206.5 |
| B:6 | {3, 7, 23, 24, 27} | 197.5 | | |
| B:6 | {3, 23, 24}<br>somewhat larger animals | | 200.5 | 200.5 |
| B:5 | {3, 7, 24, 27} | 173.5 | | |
| B:5 | {7, 21, 25, 27}<br>very small animals | | 163.5 | 164.5 |
| B:4 | {3, 24} | 123.0 | 123.0 | 123.0 |
| B:3 | {7, 27} | 22.0 | 22.0 | 22.0 |

| | | | | |
|---|---|---|---|---|
| | long-bushy-tail animals | | | |
| B:2 | {21, 25} | 6.0 | 6.0 | 6.0 |
| | long-naked-tail animals | | | |
| B:1 | (all separate) | – | – | – |
| | contribution to the norm measures: | 458.0 | 18,500. | 79.5 |
| C:11 | {1, 4, 8, 9, 11, 12, 14, 15, 17, 26, 30} | 286.0 | 298.2 | |
| | the final addition of elephant | | | |
| C:10 | {1, 4, 8, 9, 11, 14, 15, 17, 26, 30} | 238.0 | 242.7 | |
| C:9 | {1, 4, 9, 11, 14, 17, 30} | 212.5 | 215.9 | |
| C:8 | {8, 15, 26} | 200.5 | 200.5 | |
| | farm animals | | | |
| C:7 | {4, 14} | 174.0 | 174.0 | |
| | African animals | | | |
| C:6 | {1, 9, 11, 17, 30} | 167.5 | 174.7 | |
| | horse-like animals | | | |
| C:5 | {15, 26} | 93.0 | 93.0 | |
| | farm animals (subclass) | | | |
| C:4 | {11, 17, 30} | 81.5 | 81.5 | |
| | equine | | | |
| C:3 | {1, 9} | 49.0 | 49.0 | |
| | deer-like animals | | | |
| C:2 | {11, 17} | 31.0 | 31.0 | |
| | domestic equine | | | |
| C:1 | (all separate) | – | – | |
| | contribution to the norm measures: | 149.6 | 298,200. | |
| D:3 | {6, 16, 20} | 49.5 | 49.5 | 49.5 |
| D:2 | {6, 20} | 26.0 | 26.0 | 26.0 |
| D:1 | (all separate) | – | – | |
| | contribution to the norm measures: | 19.0 | 200.0 | 9.5 |

Based on the five-classes, A, B, C, D, and E, the completions of a best-fitting ultrametric beginning from this point are given below for the $L_1$- and $L_2$-norm:

| *level* | *new class formed* | $L_1$ | $L_2$ |
|---|---|---|---|
| 5 | {A, B, C, D, E} | 389.0 | 382.1 |
| 4 | {A, C, D} | 375.0 | |
| 4 | {A, C, D, E} | | 373.1 |
| 3 | {E, B} | 354.0 | |
| 3 | {A, C, E} | | 357.5 |

| | | | |
|---|---|---|---|
| 2 | {A, C} | 323.0 | 323.2 |
| 1 | (all separate) | – | – |
| | contribution to the norm measures: | 10,494. | 598,600. |

For the $L_\infty$-norm, the five-class partition retrieved for the corresponding ultrametric differed slightly from that for the $L_1$- and $L_2$-norm and involved the placement of bear (2), elephant (12), and pig (22). Explicitly, the two classes previously labeled as B and D were again retrieved for the $L_\infty$-norm, but the three other classes varied slightly:

F: {antelope (1), camel (4), cow (8), deer (9), donkey (11), giraffe (14), goat (15), horse (17), pig (22), sheep (26), zebra (30)} — large hoofed herbivores including (appropriately) pig and excluding bear

G: {cat (5), dog (10), fox (13), leopard (18), lion (19), tiger (28), wolf (29)} — felines/canines only (excluding bear)

H: {bear (2), elephant (12)} — large animals

B: {beaver (3), chipmunk (7), mouse (21), rabbit (23), raccoon (24), rat (25), squirrel (27)} — small rodent-like animals

D: {chimpanzee (6), gorilla (16), monkey (20)} — primates

Using these latter five classes, the completion of a best-fitting ultrametric is given below for the $L_\infty$-norm; subsequently, the optimal ultrametrics within the classes labeled F, G, and H are given (those for the two classes B and D were provided previously along with the $L_1$- and $L_2$-norm results):

| *level* | *new class formed* | $L_\infty$ |
|---|---|---|
| 5 | {B, D, F, G, H} | 343.5 |
| 4 | {D, F, G, H} | 335.0 |
| 3 | {F, G, H} | 331.0 |
| 2 | {F, H} | 313.5 |
| 1 | (all separate) | – |
| | contribution to the norm measure: | 362.0 |
| F:11 | {1, 4, 8, 9, 11, 14, 15, 17, 22, 26, 30} | 294.5 |
| F:10 | {8, 15, 22, 26} | 272.5 |
| F:9 | {1, 4, 9, 11, 14, 17, 30} | 216.0 |
| F:8 | {8, 15, 26} | 200.5 |
| F:7 | {1, 9, 11, 17, 30} | 178.5 |
| F:6 | {4, 14} | 174.0 |
| F:5 | {15, 26} | 93.0 |
| F:4 | {11, 17, 30} | 81.5 |
| F:3 | {1, 9} | 49.0 |
| F:2 | {11, 17} | 31.0 |

| | | |
|---|---|---|
| F:1 | (all separate) | – |
| | contribution to the norm measure: | 292.5 |
| G:7 | $\{5, 10, 13, 18, 19, 28, 29\}$ | 232.0 |
| G:6 | $\{10, 13, 29\}$ | 104.5 |
| G:5 | $\{5, 18, 19, 28\}$ | 63.0 |
| G:4 | $\{13,29\}$ | 57.0 |
| G:3 | $\{18, 19, 28\}$ | 35.0 |
| G:2 | $\{18, 28\}$ | 24.0 |
| G:1 | (all separate) | – |
| | contribution to the norm measure: | 91.0 |
| H:2 | $\{2, 12\}$ | 305.0 |
| H:1 | (all separate) | – |
| | contribution to the norm measure: | 0.0 |

## 5 Constructing (optimal) ultrametrics for two-mode proximity data

The discussion of finding optimal ultrametrics has been restricted thus far to a single object set $S$ for which a symmetric $n \times n$ dissimilarity matrix $\mathbf{P}$ is available. A direct extension is possible, however, to the context of a (two-mode) $n_A \times n_B$ dissimilarity matrix $\mathbf{Q} = \{q_{ij}\}$ defined between the objects from two distinct sets, say $S_A = \{O_{r_1}, \ldots, O_{r_{n_A}}\}$ and $S_B = \{O_{c_1}, \ldots, O_{c_{n_B}}\}$, containing $n_A$ and $n_B$ objects respectively, and where $q_{ij}$ denotes a dissimilarity between the (row) object $O_{r_i}$ and the (column) object $O_{c_j}$. Specifically, a combined single object set $S$ is first constructed as $S \equiv S_A \cup S_B$ containing $n \equiv n_A + n_B$ objects, and the same dynamic programming strategy for locating an (optimal) ultrametric is now applied to the single set $S$ but with two modifications: (i) when considering the recursive process over the sets $\Omega_1, \ldots, \Omega_n$, a transition from $A_{k-1} \in \Omega_{k-1}$ to $A_k \in \Omega_k$ is not permissible whenever the new subset formed in $A_k$ would contain only objects from $S_A$ or from $S_B$; (ii) in generating the between-subset aggregate values and the contribution to the chosen norm measure for a transition from $A_{k-1}$ to $A_k$, only those proximities defined between the object sets $S_A$ and $S_B$ are considered. Based on this strategy, the between-subset aggregate values producing the fitted values for the proximities in $\mathbf{Q}$, denoted generically as $\mathbf{T} = \{t_{ij}\}$, will satisfy the two-set ultrametric inequality (e.g., see Furnas, 1980; De Soete, DeSarbo, Furnas, and Carroll, 1984a, 1984b): for $O_{r_i}, O_{r_{i'}} \in S_A$, and $O_{c_j}, O_{c_{j'}} \in S_B$, the largest two values among $t_{r_i c_j}, t_{r_i c_{j'}}, t_{r_{i'} c_j}$, and $t_{r_{i'} c_{j'}}$ are equal.

| *animal* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *name* | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| antelope (1) | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| bear (2) | 326 | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| beaver (3) | 378 | 333 | * | * | * | * | * | * | * | * | * | * | * | * | * |
| camel (4) | 225 | 334 | 410 | * | * | * | * | * | * | * | * | * | * | * | * |
| cat (5) | 340 | 341 | 266 | 384 | * | * | * | * | * | * | * | * | * | * | * |
| chimpanzee (6) | 380 | 366 | 354 | 403 | 355 | * | * | * | * | * | * | * | * | * | * |
| chipmunk (7) | 387 | 389 | 176 | 434 | 273 | 362 | * | * | * | * | * | * | * | * | * |
| cow (8) | 240 | 290 | 395 | 260 | 376 | 406 | 403 | * | * | * | * | * | * | * | * |
| deer (9) | 49 | 315 | 359 | 249 | 339 | 392 | 357 | 238 | * | * | * | * | * | * | * |
| dog (10) | 316 | 282 | 305 | 329 | 221 | 337 | 327 | 304 | 280 | * | * | * | * | * | * |
| donkey (11) | 206 | 341 | 385 | 199 | 365 | 377 | 404 | 216 | 228 | 316 | * | * | * | * | * |
| elephant (12) | 301 | 305 | 416 | 256 | 420 | 383 | 469 | 258 | 342 | 382 | 285 | * | * | * | * |
| fox (13) | 315 | 289 | 286 | 377 | 224 | 360 | 325 | 357 | 289 | 126 | 347 | 407 | * | * | * |
| giraffe (14) | 203 | 356 | 418 | 174 | 422 | 372 | 429 | 313 | 236 | 370 | 290 | 266 | 380 | * | * |
| goat (15) | 179 | 338 | 361 | 257 | 327 | 381 | 371 | 193 | 196 | 240 | 204 | 363 | 313 | 310 | * |
| gorilla (16) | 379 | 251 | 400 | 378 | 411 | 40 | 412 | 397 | 407 | 369 | 382 | 347 | 362 | 347 | 412 |
| horse (17) | 157 | 304 | 389 | 142 | 371 | 391 | 423 | 181 | 150 | 267 | 31 | 274 | 320 | 190 | 209 |
| | 374 | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| leopard (18) | 268 | 288 | 390 | 341 | 52 | 349 | 419 | 355 | 281 | 262 | 326 | 339 | 213 | 323 | 332 |
| | 335 | 285 | * | * | * | * | * | * | * | * | * | * | * | * | * |
| lion (19) | 276 | 257 | 381 | 327 | 75 | 365 | 435 | 322 | 296 | 263 | 323 | 307 | 213 | 331 | 325 |
| | 323 | 247 | 38 | * | * | * | * | * | * | * | * | * | * | * | * |
| monkey (20) | 383 | 358 | 352 | 403 | 359 | 26 | 310 | 419 | 381 | 334 | 377 | 381 | 357 | 388 | 372 |
| | 59 | 395 | 363 | 369 | * | * | * | * | * | * | * | * | * | * | * |
| mouse (21) | 434 | 436 | 261 | 439 | 336 | 388 | 164 | 416 | 386 | 349 | 412 | 473 | 372 | 453 | 386 |
| | 472 | 440 | 430 | 457 | 385 | * | * | * | * | * | * | * | * | * | * |
| pig (22) | 410 | 356 | 350 | 395 | 359 | 408 | 394 | 269 | 368 | 299 | 347 | 371 | 362 | 410 | 284 |
| | 404 | 344 | 400 | 389 | 401 | 375 | * | * | * | * | * | * | * | * | * |
| rabbit (23) | 321 | 394 | 207 | 407 | 271 | 378 | 201 | 400 | 323 | 297 | 390 | 435 | 301 | 420 | 340 |
| | 430 | 394 | 391 | 403 | 360 | 222 | 343 | * | * | * | * | * | * | * | * |
| raccoon (24) | 356 | 304 | 123 | 391 | 229 | 347 | 171 | 405 | 349 | 282 | 383 | 433 | 214 | 398 | 352 |
| | 397 | 383 | 356 | 371 | 307 | 248 | 344 | 194 | * | * | * | * | * | * | * |
| rat (25) | 422 | 406 | 245 | 440 | 295 | 401 | 155 | 431 | 405 | 353 | 421 | 465 | 358 | 448 | 382 |
| | 452 | 431 | 427 | 429 | 398 | 6 | 354 | 239 | 270 | * | * | * | * | * | * |
| sheep (26) | 233 | 335 | 355 | 296 | 314 | 390 | 384 | 208 | 230 | 263 | 239 | 350 | 333 | 341 | 93 |
| | 408 | 247 | 356 | 337 | 394 | 397 | 261 | 317 | 335 | 395 | * | * | * | * | * |
| squirrel (27) | 368 | 378 | 183 | 422 | 264 | 347 | 22 | 413 | 366 | 322 | 401 | 454 | 312 | 439 | 389 |
| | 438 | 413 | 410 | 409 | 313 | 161 | 385 | 188 | 143 | 174 | 364 | * | * | * | * |
| tiger (28) | 281 | 243 | 403 | 328 | 57 | 368 | 431 | 355 | 295 | 287 | 320 | 318 | 205 | 333 | 316 |
| | 320 | 297 | 24 | 32 | 354 | 445 | 415 | 412 | 371 | 430 | 348 | 415 | * | * | * |
| wolf (29) | 301 | 246 | 338 | 349 | 245 | 366 | 397 | 345 | 312 | 83 | 317 | 374 | 57 | 362 | 303 |
| | 339 | 279 | 180 | 177 | 382 | 417 | 383 | 367 | 292 | 374 | 310 | 377 | 181 | * | * |
| zebra (30) | 129 | 319 | 396 | 214 | 347 | 375 | 416 | 228 | 178 | 293 | 116 | 287 | 307 | 211 | 222 |
| | 367 | 47 | 244 | 252 | 378 | 437 | 370 | 384 | 377 | 431 | 258 | 413 | 240 | 290 | * |

Table 1: A lower-triangular dissimilarity matrix between thirty animals based on data collected by Arabie and Rips (1973).

## References

[1] Arabie, P., and Rips, L. (1973). A 3-way data set of similarities between Henley's animals. Unpublished manuscript, Stanford University.

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

[3] Chandon, J.-L., and De Soete, G. (1984). Fitting a least squares ultrametric to dissimilarity data: Approximation versus optimization. In *Data Analysis and Informatics,* Ed. E. Diday, Vol.3, pp. 213–221. Amsterdam: North-Holland.

[4] Chandon, J.-L., Lemaire, J., and Pouget, J. (1980). Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés. *RAIRO Recherche opérationelle* **14** 157–170.

[5] De Soete, G., DeSarbo, W. S., Furnas, G. W., and Carroll, J. D. (1984a). The representation of nonsymmetric rectangular proximity data by ultrametric and path length tree structures. *Psychometrika* **49** 289–310.

[6] De Soete, G., DeSarbo, W. S., Furnas, G. W., and Carroll, J. D. (1984b). Tree representations of rectangular proximity matrices. In *Trends in Mathematical Psychology,* Eds. R. Burggenhaut and V. Degreef, pp. 377–392. New York: North-Holland.

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

[8] De Soete, G., and Carroll, J. D. (1996). Tree and other network models for representing proximity data. In *Clustering and Classification,* Eds. P. Arabie, L. Hubert, and G. De Soete, pp. 157–197. Singapore: World Scientific.

[9] Day, W. H. E. (1996). Complexity theory: an introduction for practitioners of classification. In *Clustering and Classification,* Eds. P. Arabie, L. J. Hubert, and G. De Soete, pp. 199–233. River Edge, NJ: World Scientific.

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

[11] Hartigan, J. A. (1967). Representation of similarity matrices by trees. *J. Am. Statist. Assoc.* **62** 1140–1158.

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

[13] Hubert, L. J., and Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British*

*Journal of Mathematical and Statistical Psychology* **48** 281–317.

[14] Křivánek, M. (1986). On the computational complexity of clustering. In *Analyse des Données et Informatique,* Eds. E. Diday et al., Vol. 4, pp. 89–96. Amsterdam: North-Holland.

[15] Křivánek, M., and Morávek, J. (1986). NP-hard problems in hierarchical clustering. *Acta Informatica* **23** 311–323.

[16] Späth, H. (1991). *Mathematical Algorithms for Linear Regression.* New York: Academic Press.