Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254 (3684 citations in Google Scholar as of 1/1/2016).

One of the ten most-cited *Psychometrika* articles is by Stephen Johnson, who was a young computer scientist working at Bell Labs at the time the article was written. Interesting, this piece is Johnson's sole published work in the area of statistics and/or psychometrics. It was obviously influenced heavily by other Bell Labs personnel at the time, including several future Psychometric Society presidents. The introductory footnote to the article reads: "I am indebted to R. N. Shepard and J. D. Carroll for many stimulating discussions about this work, and for aid in preparing this paper."

From the late 1960s onward, Johnson was heavily involved with the Bell Labs development of the UNIX operating system and the C programming language, most notably developing the Portable C Compiler and other UNIX tools (for example, spell, lint, yacc). In the early 2000s, Johnson joined the Mathworks to contribute to the MATLAB programming language; he wrote, for example, the initial MATLAB compiler for turning a routine written in the MATLAB language into stand-alone and executable C code.

The Johnson paper in *Psychometrika* is nicely written and accessible, and provides for the first time in the literature the key concept of an ultrametric for characterizing all hierarchical clustering schemes. Johnson's inspiration for developing the correspondence between ultrametrics and hierarchical clustering schemes was from the work of Joe Ward, a personnel research psychologist working in the 1950s and 60s on occupational analyses for the United States Air Force. To quote Johnson (p. 242):

> The notion of a hierarchical clustering scheme, the central idea of this paper, was abstracted from examples given by Ward [1963]. We first consider such schemes, and develop a correspondence between hierarchical clustering schemes and a certain type of metric.

This important idea of an ultrametric will be reviewed below along with several other observations Johnson made over the course of his paper. We conclude with some discussion of the historical importance of the concept of an ultrametric, introduced some fifty years ago, and how it has guided the field of cluster analysis and various extensions up to the present.[1]

[1]Another 1967 paper by Jardine, Jardine, and Sibson also introduced the concept of an ultrametric.

To characterize more formally the basic problem posed by hierarchical clustering, suppose $S$ is a set of $n$ objects, $\{O_1, \ldots, O_n\}$, and between each pair of objects, $O_i$ and $O_j$, a symmetric proximity measure, $p_{ij}$, is available (that might possibly be constructed from a more basic object by variable data matrix). It is assumed that the proximity measure has a dissimilarity interpretation so larger proximity values correspond to more dissimilar objects. These proximity values are collected into an $n \times n$ proximity matrix, $\mathbf{P} = \{p_{ij}\}_{n \times n}$. Any hierarchical clustering strategy produces (using Johnson's terminology) a hierarchical clustering scheme; the latter is a sequence or hierarchy of partitions of $S$, denoted $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{n-1}$, from the information present in $\mathbf{P}$. In particular, the (disjoint) partition, $\mathcal{P}_0$, contains all objects in separate classes (Johnson's "weak" clustering); $\mathcal{P}_{n-1}$ (the conjoint partition) consists of one all-inclusive object class (Johnson's "strong" clustering); and $\mathcal{P}_{k+1}$ is defined from $\mathcal{P}_k$ by uniting a single pair of subsets in $\mathcal{P}_k$.

Generally, the two subsets chosen to unite in defining $\mathcal{P}_{k+1}$ from $\mathcal{P}_k$ are those that are "closest," with the characterization of this latter term specifying the particular hierarchical clustering method being used. We mention three of the most common options for this notion of closeness:

(a) complete-link: the maximum proximity value attained for pairs of objects within the union of two sets (thus, the maximum link [or the subset "diameter"] is minimized);

(b) single-link: the minimum proximity value attained for pairs of objects where the two objects from the pair belong to the separate classes (thus, we minimize the minimum link);

(c) average-link: the average proximity over pairs of objects defined across the separate classes (thus, the average link is minimized).

From the time of Johnson's 1967 paper (which mainly emphasized the single-link and complete-link criteria), it has been generally accepted that the complete-link criterion should be the default selection for the task of hierarchical clustering when done in the traditional agglomerative way that

But in contrast to Johnson (1967), Jardine, et al. (1967) is neither nicely written nor accessible. The intuitively pleasing correspondence between a hierarchical clustering scheme and an ultrametric, so apparent in Johnson's paper, is lost in Jardine, et al. by being buried under of lot of unnecessary topology and superfluous mathematical notation.

starts from $\mathcal{P}_0$ and proceeds step-by-step to $\mathcal{P}_{n-1}$. A reliance on the single-link criterion tends to produce "straggly" clusters that are not very internally homogenous or substantively interpretable. To quote Johnson (p. 252):

> ... the results of the Maximum Method [that is, complete-link] have appeared to be the more meaningful or interpretable. That is, the search for compact clusters (of small over-all "diameter") have proved more useful than the search for internally "connected" but potentially long chain-like clusters.

As noted by many users over the years, the average-link choice seems to produce results that are the same as or very similar to the complete-link criterion but relies on more information from the given proximities; the complete-link criterion (or for that matter, the single-link criterion) depend only on the rank order of the proximities. Given the Bell Labs context in which the Johnson paper was produced (with Kruskal and Shepard as senior colleagues to Johnson and the recent development of *nonmetric* multidimensional scaling) it shouldn't be surprising that Johnson emphasized only single-link and complete-link clustering.[2] To quote (p. 253):

> Sokal and Sneath [1963, p. 190] have pointed out that, in methods like our Minimum or Maximum Methods [that is, single- or complete-link], the merging of two clusters depends upon a single similarity value (viz., the least or greatest in the appropriate set). They suggest that, for greater robustness of the solution, it may sometimes be desirable to use some sort of average value instead. As we have already noted, *to base such a procedure upon averages of the more obvious types is to lose the invariance, sought here, under monotone transformation of the similarity values.* [emphasis added]

Johnson notes that the fortran program he wrote to carry out both single-link and complete-link clustering (called `hiclust.f`, and still available at `www.netlib.org`) could be easily modified to include other clustering criteria such as average-link. Again, we give a relevant quote (p. 254):

> Nevertheless, when this seems desirable, the methods described here can be (and, indeed, have been) modified to yield solutions intermediate between those obtained by these two extreme methods. J. D. Carroll (personal communication) has suggested an average method based upon medians which, of course, do have the desired property of monotone invariance. The main problem, in the case of medians, is the choice of an appropriate procedure for dealing with the ambiguities that tend to arise when two or more of the initial similarity estimates are tied.

[2] As we anticipate from a later discussion, the average-link criterion has some connections with rephrasing hierarchical clustering as a least-squares optimization task in which an ultrametric is fit to the given proximity values. The average proximities between the subsets united to form the hierarchy are the values fitted to the given proximities.

The raising of the problem of ties in the case of a median criterion (which also applies to the use of a complete-link criterion), and the ambiguities it might engender in the construction of a partition hierarchy as to what groups are formed depending on how ties are resolved, has a somewhat unfortunate history in the cluster analysis literature (at least to the present author's mind). First, in practice there seems to be no problem with tied proximities except in some very pathological and artificially made-up examples. But the small possibility of such an ambiguity has led some authors to conclude that the complete-link criterion is "inadmissible"; moreover, the only commonly "admissible" method is single-link because it satisfies a continuity condition in how a proximity measure is transformed into an ultrametric (see, for example, Jardine, Jardine, and Sibson, 1967). This is pure mathematical tyranny to recommend a generally less substantively interpretable procedure, such as single-link, over one that is typically much better, such as complete-link. Asserting an arbitrary continuity condition is no justification for the use of an inferior method.

As noted earlier, the seminal contribution of Johnson's paper and the reason for its continued popularity is the characterization of a hierarchical clustering scheme in terms of what is called an ultrametric. We now turn to a more formal definition.

Given the partition hierarchies from any of the three criteria mentioned (complete-, single-, or average-link), suppose the values for when the new subsets were formed (that is, the maximum, minimum, or average proximity between the united subsets) are placed into an $n \times n$ matrix, $\mathbf{U}$. In general, there are $n-1$ distinct nonzero values that define the levels at which the $n-1$ new subsets are formed in the hierarchy; thus, there are typically $n-1$ distinct nonzero values present in an appropriately row and column reordered matrix $\mathbf{U}$ that characterizes the identical blocks of matrix entries between subsets united in forming the hierarchy.[3]

[3]More explicitly, there exists an ordering (not unique) of the rows and simultaneously the columns of $\mathbf{U}$ that will give the reordered matrix, say $\mathbf{U}'$, the following properties:

1. $\mathbf{U}'$ can be partitioned as

$$\begin{bmatrix} \mathbf{U}'_{11} & \mathbf{U}'_{12} \\ \mathbf{U}'_{21} & \mathbf{U}'_{22} \end{bmatrix},$$

Based on a matrix such as $\mathbf{U}$, the partition hierarchy can be retrieved immediately along with the levels at which the new subsets were formed. In fact, any (strictly) monotone (that is, order preserving) transformation of the $n-1$ distinct values in such a matrix $\mathbf{U}$ would serve the same retrieval purposes. Thus, as one illustration, the $n-1$ distinct values in $\mathbf{U}$ could be replaced by the simple integers, $(1, 2, 3, \ldots, n-1)$, and the partitions of the hierarchy could still be reconstructed easily. Generally, a matrix $\mathbf{U}$ that can be used to retrieve a partition hierarchy in this way is called an ultrametric (matrix):

A matrix $\mathbf{U}$ is called an ultrametric (matrix) if for every triple of subscripts, $i, j,$ and $k$, $u_{ij} \leq \max(u_{ik}, u_{kj})$; or equivalently (and much more understandably), among the three terms, $u_{ij}, u_{ik},$ and $u_{kj}$, the largest two values are equal.

The last paragraph of Johnson's paper is very prescient about where his introduction of the ultrametric concept will lead. This last paragraph begins (p. 254):

> Finally, a different kind of possible extension will be briefly indicated here, in the form of a presently unsolved problem. In Section I we saw that the construction of an HCS is equivalent to finding a metric which satisfies the ultrametric inequality. Given a similarity measure $d$, we would in general like to find the closest metric $D$ which satisfies the ultrametric inequality — various measures of closeness could be used. ... To the author's knowledge, this problem is unsolved.[4]

So, as Johnson foreshadowed, it is now common in the literature to reformulate this hierarchical clustering task as that of locating a best-fitting ultrametric matrix, say, $\mathbf{U}^* = \{u^*_{ij}\}$, to the given proximity matrix, $\mathbf{P}$, such

---

where all the elements of $\mathbf{U}'_{12}$ and $\mathbf{U}'_{21}$ are equal to the single largest element of $\mathbf{U}$.

2. The submatrices $\mathbf{U}'_{11}$ and $\mathbf{U}'_{22}$ are partitionable as in 1.

3. The partitioning process can be repeated until all the resulting submatrices are of order 1.

In the graph theory literature, a matrix with properties 1 to 3 is said to be principally partitionable (see, for example, Shein and Frisch, 1969).

[4]This comment about being "unsolved" is somewhat of an understatement. The task of finding the "closest" ultrametric falls within the class of optimization problems known as NP-hard, which includes all the old combinatorial chestnuts, such as the traveling salesman problem. This was shown explicitly by Křivánek and Morávek in 1986. Although Johnson was a computer scientist, he cannot be faulted for not proving or knowing this in 1967. The notion of a problem being NP-hard was not even introduced into the computer science literature until the early 1970s.

that the least squares criterion

$$\sum_{i<j}(p_{ij} - u^*_{ij})^2 \ ,$$

is minimized. The approach can either be confirmatory (in which we look for the best-fitting ultrametric defined by some monotone transformation of the $n-1$ values making up a fixed and given ultrametric), or exploratory (where we merely look for the best-fitting ultrametric without any prior constraint as to form). In both cases, a convenient normalized loss measure is given by the variance-accounted-for (VAF):

$$1 - \frac{\sum_{i<j}(p_{ij} - u^*_{ij})^2}{\sum_{i<j}(p_{ij} - \bar{p})^2} \ ,$$

where $\bar{p}$ is the average off-diagonal proximity value in $\mathbf{P}$. Because of the NP hardness of locating a best-fitting ultrametric, the various computational methods to be mentioned below are all heuristic; there is no guarantee of identifying the closest ultrametric, even with the use of many random starts.

The R package 'clue' has a routine called `ls_fit_ultrametric` for finding an ultrametric for a given proximity matrix using a least-squares criterion. Three different methods are included:

(a) method "SUMT" implements the Sequential Unconstrained Minimization Technique of de Soete (1986);

(b) method "IP" implements the Iterative Projection approach of Hubert and Arabie (1995);

(c) method "IR" implements the Iterative Reduction approach suggested by Roux (1988).

Two MATLAB M-files (`ultrafit.m` and `ultrafnd.m`) discussed in the monograph by Hubert, Arabie, and Meulman (2005), can be used to carry out either a confirmatory or an exploratory fitting of an ultrametric to a given proximity matrix. These two M-files are available from

`cda.psych.uiuc.edu/srpm_mfiles`

A currently popular alternative to the use of a simple ultrametric in classification, and what can be considered an extension or generalization (and thus

part of the legacy of Johnson's original paper), is that of an additive tree.[5] Relaxing the earlier characterization of an ultrametric (which could be called a three-point condition), an $n \times n$ matrix, $\mathbf{D}$, can be called an additive-tree metric (matrix) if the three-point ultrametric inequality condition is replaced by the four-point condition:[6]

$d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{il} + d_{jk})$ for $1 \leq i, j, k, l \leq n$ (the additive-tree metric inequality). Or equivalently (and again, much more understandably), for any object quadruple $O_i$, $O_j$, $O_k$, and $O_l$, the largest two values among the sums $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$, and $d_{il} + d_{jk}$ are equal.

The four-point condition for an additive-tree metric was introduced into the psychometric literature by Sattah and Tversky (1977) (a fairly well-cited *Psychometrika* article, by the way), using earlier work by Buneman (1974). The R package 'clue' includes a routine `ls_fit_addtree` that implements the same least-squares optimization options as in `ls_fit_ultrametric` (that is, SUMT, IP, and IR), but now in the service of the least-squares fitting of an additive tree to a set of given proximities. Two MATLAB M-files that carry out a confirmatory or an exploratory fitting of additive trees (`atreefit.m` and `atreefnd.m`) are available at the web site listed previously.[7]

[5]We note that an ultrametric is a special case of an additive tree in which there is a spot (the "root") on the tree that is equidistant from all the terminal nodes.

[6]Biologists who work with phylogenetic trees (our additive trees) are fond of calling the four-point condition the "quartet inequality." From an embarrassing personal experience, however, I would advise not trying to be humorous when talking about ultrametrics before such a somber audience, and using an opening statement such as "to keep the musical motif going, I will talk about the trio inequality."

[7]The present author started to work in the area of cluster analysis with Frank Baker in the early 1970s. As part of this collaboration, Frank wrote a fortran program to carry out single-link and complete-link hierarchical clustering based on a computational algorithm of my design that relied on a preliminary ordering of the object pairs from most to least similar. The algorithm proceeded systematically through the ordering, checking and rechecking until all necessary links were present between two specific subsets that could then be united to form the next new subset (and thus, the next partition) in the hierarchy. This particular computational approach had proven to be a nice instructional demonstration when done "by hand"; but when computer-implemented with Frank's program, it seemed a bit slow for the complete-link criterion given all the exhaustive checking and rechecking that was necessary. Problems involving, say, twenty objects took several hours on a mainframe UNIVAC 1108 machine then installed at the Madison Academic Computing Center. Our strategy was quietly abandoned when we belatedly and with great embarrassment reread the following sentence in Johnson's paper (p. 253):

On an IBM 7094, the analysis is completed quite rapidly; in another application with 64 objects, solutions were obtained for both the Minimum and Maximum Methods in just 10.1 seconds.

## References


[1] Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory (B)*, *17*, 48–50.

[2] de Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika*, *48*, 621–626.

[3] de Soete, G. (1986). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters*, *2*, 133–137.

[4] Hubert, L., & 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.

[5] Hubert, L., Arabie, P., & Meulman, J. (2006). *The structural representation of proximity matrices with MATLAB*. Philadelphia: SIAM.

[6] Křivánek, M., & Morávek, J. (1986). NP-hard problems in hierarchical clustering. *Acta Informatica*, *23*, 311–137.

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

[8] Roux, M. (1988). Techniques of approximation for building two tree structures. In C. Hayashi, E. Diday, M. Jambu, and N. Ohsumi (Eds.), *Recent developments in clustering and data analysis* (pp. 151–170). New York: Academic Press.

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

[10] Schein, N. P., & Frisch, I. T. (1969). Vertex weighted trees with fewest relay vertices. *SIAM Journal of Applied Mathematics*, *10*, 31–38.

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

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