PROGRAM ABSTRACTS/ALGORITHMS

Computer programs for fitting ultrametric and additive trees to proximity data by least squares methods

GEERT DE SOETE

University of Ghent, Ghent, Belgium

LSULT (Least Squares ULtrametric Tree analysis) and LSADT (Least Squares ADditive Tree analysis) are computer programs that implement the least squares algorithms proposed by De Soete (1983, 1984b) for fitting ultrametric and additive trees to proximity data. An ultrametric tree is a rooted tree in which a nonnegative weight is attached to each node such that: (1) the terminal nodes have zero weight, (2) the largest weight is assigned to the root, and (3) the weights attached to the nodes on the path from any terminal node to the root constitute a strictly increasing sequence. The distance between any two distinct nodes is defined as the maximum of the weights assigned to the nodes on the unique path connecting the two nodes. It is well-known that these distances satisfy the ultrametric inequality:

$$d_{ij} \leq \max(d_{ik}, d_{jk}).$$

An additive tree, on the other hand, is an unrooted tree in which a nonnegative weight is attached not to the nodes but to the links of the tree. The distance between any two distinct nodes is then defined as the sum of the weights attached to the links constituting the unique path that connects the two nodes. It can be shown that these distances satisfy the so-called additive inequality or four-point condition:

$$d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{il} + d_{jk}).$$

Ultrametric and additive trees are valuable alternatives to multidimensional scaling for representing psychological proximity data (see, e.g., Johnson, 1967, Pruzansky, Tversky, & Carroll, 1982, and Sattath & Tversky, 1977).

Description. Since a distance matrix D that perfectly satisfies the ultrametric (respectively additive) inequality defines a unique ultrametric (respectively additive) tree, it suffices, in order to construct a least squares ultra-

metric (respectively additive) tree representation of a set of proximity data, to solve for a distance matrix D that perfectly satisfies the ultrametric (respectively additive) inequality *and* that is closest to the data in the least squares sense. The algorithm solves for such a matrix by sequentially minimizing the function

$$F(D, r) = L(D) + rP(D) \quad (r > 0)$$

for a strictly increasing sequence of values of r . The objective function $F(D, r)$ consists of two parts: the loss function $L(D)$, which is the sum of squared deviations between the data and D , and $P(D)$, which is a penalty function that enforces the ultrametric (in LSULT) or the additive (in LSADT) inequality on D . For the subsequent minimizations of $F(D, r)$, Powell's (1977) conjugate gradient method with automatic restarts is used. Both LSULT and LSADT are capable of handling missing data. An extensive evaluation of the performance of LSULT and LSADT in the presence of missing data can be found in De Soete (1984a, 1984c).

LSADT differs from the additive tree-fitting program developed by Corter (1982) in that both the topology and the link weights of the constructed tree are optimal in least squares sense. Since Corter's program solves for the topology of the tree by means of an agglomerative method, the resulting topology is not necessarily least squares (although the weight estimates are least squares, given this topology). For a more thorough discussion of the relation between LSADT and other additive tree-fitting methods, the reader is referred to De Soete (1983), and a comparison between LSULT and a branch-and-bound algorithm for constructing a least squares ultrametric tree representation is reported in Chandon and De Soete (1984).

Input. The input to LSULT and LSADT consists of a descriptive job title, several run parameters, a variable data format, and a lower-diagonal proximity matrix. The data may be similarities or dissimilarities. Empty cells in the data matrix are flagged by a user-specified missing-data indicator.

Output. After the input data and (optionally) a brief history of the iterative estimation process are printed, the goodness-of-fit index, the least squares estimate of D , and a bottom-up merge list of the associated ultrametric or additive tree are output, followed by the matrix of residuals. Optionally, the least squares estimate of D is punched out or saved in a separate file.

Language and Computer. An attempt has been made to write the programs in a portable subset of FORTRAN IV, known as PFORT (Ryder, 1974). In order to further enhance (trans)portability, routines from the PORT

These programs were developed while the author was supported as Aangesteld Navorsers of the Belgian Nationaal Fonds voor Wetenschappelijk Onderzoek. He is indebted to J. Hoste for providing computer facilities at the Institute of Nuclear Sciences of the University of Ghent, Ghent, Belgium. The author's mailing address is: Department of Psychology, University of Ghent, Henri Dunantlaan 2, B-9000 Ghent, Belgium.

library (Fox, Hall, & Schryer, 1978b) are used for error handling, simulated dynamic memory allocation, and the specification of machine-dependent constants. These routines are in the public domain and can be found in Fox, Hall, and Schryer (1978a). Apart from the two routines in which the machine-dependent constants are initiated, the programs run unmodified on a VAX-11/780 (both under VMS and UNIX), a Cray I, a Honeywell 6000, and a Siemens 7551 computer.

Availability. A documented listing of the two programs may be obtained without charge by writing to Geert De Soete, Department of Psychology, University of Ghent, Henri Dunantlaan 2, B-9000 Ghent, Belgium. A tape copy can be obtained by sending a nine-track magnetic tape and a check covering the mailing costs to the author.

REFERENCES

- CHANDON, J. L., & DE SOETE, G. (1984). Fitting a least squares ultrametric to dissimilarity data: Approximation versus optimization. In E. Diday et al. (Eds.) *Data analysis and informatics III*. (pp. 213-221). Amsterdam: North-Holland.
- CORTER, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behavior Research Methods & Instrumentation*, **14**, 353-354.
- DE SOETE, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika*, **48**, 621-626.
- DE SOETE, G. (1984a). Additive tree representations of incomplete dissimilarity data. *Quality & Quantity*, **18**, 387-393.
- DE SOETE, G. (1984b). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters*, **2**, 133-137.
- DE SOETE, G. (1984c). Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification*, **1**, 235-242.
- FOX, P. A., HALL, A. D., & SCHRYER, N. L. (1978a). Algorithm 528. Framework for a portable library. *ACM Transactions on Mathematical Software*, **4**, 177-188.
- FOX, P. A., HALL, A. D., & SCHRYER, N. L. (1978b). The PORT mathematical subroutine library. *ACM Transactions on Mathematical Software*, **4**, 104-126.
- JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241-254.
- POWELL, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical Programming*, **12**, 241-254.
- PRUZANSKY, S., TVERSKY, A., & CARROLL, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, **47**, 3-24.
- RYDER, B. G. (1974). The PFORT verifier. *Software - Practice and Experience*, **4**, 359-377.
- SATTATH, S., & TVERSKY, A. (1977). Additive similarity trees. *Psychometrika*, **42**, 319-345.

(Revision accepted for publication December 11, 1984).