Program OVW: A Program for Optimal Variable Weighting for Ultrametric and Additive Tree Clustering, as well as *K*-means Partitioning

Vladimir Makarenkov and Pierre Legendre

May 2001

Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec H3C 3J7, Canada
e-mail: **makarenv@magellan.umontreal.ca** and **pierre.legendre@umontreal.ca**

What does program OVW do?

Program OVW performs optimal variable weighting for ultrametric and additive tree clustering as well as for *K*-means partitioning, following the method proposed by De Soete (1986, 1988), and Makarenkov and Legendre (2001). The new program, which is available free of charge to academic users, provides some improvements and extra options, compared to De Soete's (1988) program OVWTRE which only implemented fitting to the first two families of clustering methods mentioned above.

Given a rectangular data matrix **Y**, containing measurements of *n* objects on *m* variables, the algorithm computes variable weights $\mathbf{w} = \{w_1, w_2, \ldots w_m\}$ such that the resulting matrix of dissimilarities among objects $\mathbf{D} = [d_{ij}]$:

$$d_{ij} = \left[\sum_{p=1}^{m} w_p (y_{ip} - y_{jp})^2 \right]^{1/2}$$

optimally satisfies either the ultrametric or the additive inequality, or optimally corresponds to a *K*-means partition with fixed number of groups *K*. The weights are constrained to be nonnegative and their sum is equal to one.

The ultrametric inequality is fulfilled when:

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

for all *i*, *j*, and *k*, while the additive inequality is verified when:

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

for all *i*, *j*, *k*, and *l*. In the same way, the *K*-means partitioning problem can be defined as follows: determine a partition of *n* objects into *K* groups, or clusters, such that the sum, over all groups, of the squared within-group residuals is minimal.

For each of the three clustering problems, a particular function to be minimized should be defined to compute optimal weights. In the ultrametric case, the optimal weights are computed by solving the optimization problem described by De Soete (1986):

$$L_U\,(w_1, w_2, \ldots w_m) = \frac{\sum_{\Omega_U}(d_{ik} - d_{jk})^2}{\sum_{i<j} d_{ij}^2} \qquad \rightarrow \qquad \min$$

where $\Omega_U = \{(i, j, k) \mid d_{ij} \leq \min(d_{ik}, d_{jk})$, and $d_{ik} \neq d_{jk}\}$. The minimization is done subject to the constraints:

$$w_1, w_2, \ldots, w_m \geq 0, \qquad (1)$$

$$w_1 + w_2 + \ldots + w_m = 1. \qquad (2)$$

In the additive case, the optimization problem is formulated as in De Soete (1986):

$$L_A\,(w_1, w_2, \ldots w_m) = \frac{\sum_{\Omega_A}(d_{ik} + d_{jl} - d_{il} - d_{jk})^2}{\sum_{i<j} d_{ij}^2} \qquad \rightarrow \qquad \min$$

subject to constraints (1) and (2); $\Omega_A = \{(i, j, k, l) \mid (d_{ij} + d_{kl}) \leq \min\,(d_{ik} + d_{jl}, d_{il} + d_{jk})$, and $d_{ik} + d_{jl} \neq d_{il} + d_{jk}\}$.

In the case of *K*-means partitioning, the minimization problem can be formulated as follows for a partition of *n* objects into a fixed number of groups *K*:

$$L_P\,(w_1, w_2, \ldots w_m) = \sum_{k=1}^{K}\left[\sum_{i,j=1}^{n_k} d_{ij}^2\right]/n_k \qquad \rightarrow \qquad \min$$

subject to constraints (1) and (2); n_k is the number of objects in group number k.

We used the Polak-Ribiere optimization method (see Press et al. 1986 and later editions, or Polak 1971) to carry out the minimization of L_u, L_a and L_p. Once the optimal variable weights for L_u or L_a are obtained, the resulting dissimilarities **D** among objects can be subjected to any of the existing ultrametric or additive-tree fitting procedures; see Arabie et al. (1996) for an overview of these methods. For hierarchical clustering methods, an extensive Monte-Carlo investigation based on the results provided by De Soete's program OVWTRE can be found in Milligan (1989).

Sometimes, the method described above may produce a local instead of a global minimum for L_u, L_a or L_p. Hence, a good choice of initial weights is essential. While experimenting with the program, we realized that an initial guess where all the weights are equal to 1/(*number of variables*), as implemented in program OVWTRE, cannot guaranty that the global minimum is always reached. An important detail not reported in De Soete (1986, 1988) is that the global minimum of L_a or L_u can sometimes be reached with several different sets of optimal weights **w**. This may lead to different dissimilarity matrices **D**, from which different clustering hierarchies or additive trees can be inferred.

An interesting feature of program OVW, compared to OVWTRE, is that it allows users to restart the optimization procedure any number of times, using different random initial guesses for the weights. As a consequence, OVW usually obtains better results than OVWTRE in the case of ultrametric and additive clustering; optimization for *K*-means partitioning is not available in OVWTRE.

In the optimization for additive clustering, a degenerate trivial solution may be obtained by program OVWTRE. It consists of giving a weight of 1 to any one variable and weights of 0 to all other variables. It is easy to see that any single quantitative variable satisfies the additive inequality which defines an additive tree. This is why assigning a weight of 1 to any one of m quantitative variables always guarantees a perfect fit to an additive-tree distance. In program OVW, we provide a way to avoid this trivial solution which overshadows the effect of all the other variables and often leads to a sub-optimal additive tree: the user can introduce a maximum value for the weight of any single variable, in order to avoid a trivial solution. A numerical example, below, illustrates how program OVW works in practice.

Input files

The input data file is an ASCII text file which contains a data matrix **Y**(*n* x *m*) as well as the parameters *n* and *m*.. If the *K*-means partitioning option is selected, a vector of group assignment for each object has to be provided.

Matrix of **Y** contains measurements of *n* objects on *m* variables.

The data file is organized as follows:

- First, a line with the two parameters *n* and *m*, separated by one or more spaces.

  n is the number of objects, or rows in the matrix **Y**.

  m is the number of variables (columns) in the matrix **Y**.

- The data matrix **Y** follow. A row of data can take as many successive physical lines as needed. Values in the same line are separated by one or more spaces; the number of spaces does not matter.

- If the *K*-means partitioning option is selected, each row *i* of **Y** is followed by an integer that defines the group's number for the object associated with the row *i*.

Options of the program

The program can carry out one of the three following analyses:

- Ultrametric clustering.

- Additive tree clustering.

- *K*-means partitioning.

Output file

The output is produced either on the monitor, or in a separate output file. The output consists

- Dissimilarity matrix **D**(*n*x*n*) obtained from **Y** using optimal weights

- Vector of optimal weights **w**(*m*) obtained using the Polak-Ribiere minimization procedure.

- Minimum value of the objective loss function being minimized.
- Number of iterations in the Polak-Ribiere minimization procedure leading to the final results.

Disclaimer and Availability

Program OVW is freeware. It is available via Internet on the WWWeb page of the Laboratory of Numerical Ecology at Université de Montréal: <http://www.fas.umontreal.ca/biol/legendre/> or <http://www.fas.umontreal.ca/biol/casgrain/en/labo/ovw.html>. This program has been developed as part of a university-based research programme. Users who encounter problems with this program may report it to the authors who will be happy to help solve them. Researchers may use this program for scientific purposes, but the source code remains the property of Vladimir Makarenkov and Pierre Legendre (© 1999). Publications should give proper credit to the method by referring to the original papers. Users of program OVW may refer to the paper by Makarenkov and Legendre (2001) or to the user's manual as follows:

MAKARENKOV, V., and LEGENDRE, P. (2001), "OVW (Optimal variable weighting for ultrametric and additive tree clustering)," Département de sciences biologiques, Université de Montréal. 6 pp.

Dimensionality and running time

There are no limitations to the size of matrix **Y**(*n*x*m*) in the program. The only existing limitation is the size of the random access memory (RAM) of the user's computer. However, the Polak and Ribière optimization procedure uses the matrix of partial derivatives (d_{ij} par $\mathbf{w}_k$). This intermediate matrix, which is repeatedly computed by the program, requires $O(m \times n^2)$ bytes for storage. For example, for an input matrix **Y** of size (100 x 100), the program requires about 4 MB of memory only to store the auxiliary matrix of partial derivatives. There are also some other auxiliary matrices and vectors occupying a substantial, but not so huge, amount of RAM. As to the running time, during the simulations involving a matrix **Y** with 300 objects and 166 variables, the program ran during approximately 4.5 hours on a Power Macintosh 604 at 350 MHz with 80 MB of RAM before providing a solution for the *K*-means partitioning problem; the optimization procedure was run only once for this problem.

Technical notes

The program is distributed in a variety of formats:

- C source code for Macintosh and for Windows (files in the folder Source), which can be compiled using a C/C++ compiler.
- Compiled versions of the program for Win32 compatible computers (OVW.exe). The executable file is a Win32 "**console**" executable, not DOS executables. Therefore it cannot run under plain DOS, nor in a DOS window under Windows 3.x, only in Windows 95/98 or Windows NT consoles.
- Compiled version for PowerPC processors for Macintosh (file OVW_PPC).
- C source code for different versions of UNIX (files in the folder Source-files) as well as the corresponding Make file.

References

ARABIE, P., HUBERT, L.J., and DE SOETE, G. (Eds.), *Clustering and Classification*, River Edge, New Jersey: World Scientific Publ. Co.

POLAK, E. (1971), "*Computational methods in optimization*," New York: Academic Press.

DE SOETE, G. (1986), "Optimal variable weighting for ultrametric and additive tree clustering," *Quality&Quantity*, 20, 169-180.

DE SOETE, G. (1988), "OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting," *Journal of Classification*, 5, 101-104.

MAKARENKOV, V., and LEGENDRE, P. (2001). "Optimal variable weighting for ultrametric and additive trees and K-means partitioning: methods and software," *Journal of Classification*, to appear in 18, 2.

MAKARENKOV, V., and LECLERC, B. (1999), "An Algorithm for the fitting of a tree metric according to a weighted least-squares criterion", *Journal of Classification*, 16, 1, pp.3-26.

MILLIGAN, G.W. (1989), "A validation study of a variable weighting algorithm for cluster analysis," *Journal of Classification*, 6, 53-71.

PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A., and VETTERLING, W. T. (1986), *Numerical Recipes, The Art of Scientific Computing*, Cambridge: Cambridge University Press.