Optimal Variable Selection for Ultrametric and Additive Tree Clustering and *K*-means Partitioning

Vladimir Makarenkov[1, 2] and Pierre Legendre[3]

[1]*Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia*

[2] *Department of Informatics, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8. e-mail: makarenkov.vladimir@uqam.ca*

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

Abstract. The problem of identification of optimal weights of observed variables finds its applications in a variety of fields: In marketing, psychometrics, geography, and molecular biology, briefly, in any practical situation where a classification or clustering model is appropriate. Knowledge of optimal weights associated with observed variables allows to make decision about the importance of any given variable characterizing a group of given objects. In his pioneering paper, De Soete (1986) proposed a method for estimation of optimal variable weights intended for ultrametric or additive tree reconstruction. This paper extends De Soete's method to the case of *K*-means partitioning. We also discuss some new features and improvements to the algorithm proposed by De Soete (see also Makarenkov and Legendre, 2001). Monte Carlo simulations have been conducted using different error conditions. In all cases (i.e., ultrametric or additive trees, or *K*-means partitioning), the simulation results indicate that the optimal weighting procedure should be used to eliminate noisy variables that do not contribute relevant information to the classification structure. However, if the data involve error-perturbed variables that are relevant to the classification or outliers, it seems better to cluster or partition the entities by using variables with equal weights. A new computer program, OVW carries out improved algorithms for optimal variable weighting for ultrametric and additive tree clustering, and includes a new algorithm for optimal variable weighting for *K*-means partitioning.

1. Introduction

In two pioneering papers, De Soete (1986, 1988) proposed a numerical method for estimating optimal weights for variables intended for ultrametric or additive tree reconstruction. The present paper extends De Soete's method to least-squares *K*-means partitioning. We will also point out some properties of the algorithm proposed by De Soete that seem to have gone unnoticed; an understanding of these properties leads to improvements in the methods.

We carried out Monte Carlo studies for optimal variable weighting applied to additive tree reconstruction and *K*-means partitioning (the details on these studies are not reported here, see Makarenkov and Legendre, 2001). These studies conducted using different error conditions confirmed the ability of the method to identify and reduce the effect of 'noisy' variables. We did not test the ability of the method for recovering clusters in the framework of ultrametric clustering procedures because a Monte Carlo study had already been carried out and discussed in detail by Milligan (1989). Considering the complexity of the algorithms that we discuss, a computer program is made available to the scientific community to encourage researchers to use optimal variable weighting.

There is an appreciable literature about variable weighting. DeSarbo, Carroll, Clark, and Green (1984) described SYNCLUS, a program that solves for both variable weights and produces *K*-means clustering. Fowlkes, Gnanadesikan, and Kettenring (1988) also proposed a method, here called FGK, for selecting weights — in that case, binary (0 and 1) weights. These authors proposed a model that selects subsets of variables from the original data and produces binary weights for the variables; their procedure was applied to complete linkage hierarchical clustering.

In a later paper, Gnanadesikan, Kettenring, and Tsao (1995) compared Fowlkes et al.'s (1988) FGK procedure to De Soete's OVWTRE and to DeSarbo et al.'s (1984) SYNCLUS models. Gnanadesikan et al. (1995) determined that the FGK forward selection procedure performed reasonably well compared to its competitors. Subsequent to the FGK algorithm, Carmone, Kara, and Maxwell (1999) proposed a variable subset selection method based on Hubert and Arabie's (1985) adjusted Rand index. Their method was designed for partitioning using continuous variables. The procedure proposed by Carmone et al. (1999) in the context of partitioning clustering, called HINoV, was described as a heuristic method based upon the adjusted Rand statistics. These authors conducted a series of Monte Carlo simulations, using synthetic data with noise of various kinds added, including masking variables. The results indicated that variables selected using the HINoV procedure outperformed the all-variable cases in 70 out of 72 different computer runs. In contrast to the good results found by Carmone et al. (1999), in real data set analyses using HINoV, Green, Carmone, and Kim (1990) had earlier found mixed results in the ability of SYNCLUS to recover the correct variable weights.

2. Description of the Method

Given a rectangular (i.e., object-by-variable, or two-way, two-mode) data matrix **Y**, containing measurements of n objects on m variables, our algorithm computes weights $\mathbf{w} = \{w_1, w_2, \ldots w_m\}$ for the m variables such that the resulting matrix of predicted dissimilarities $\mathbf{D} = [d_{ij}]$ among objects, where

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

optimally satisfies *either* (a) the ultrametric or (b) the additive inequality, or (c) optimally corresponds to a *K*-means partition with a fixed number of groups K. Equation (1) is the weighted form of the familiar Euclidean distance formula. The weights are constrained to be nonnegative with their sum equal to one.

The ultrametric inequality which defines dendrograms (Hartigan 1967) is satisfied when:

$$d_{ij} \leq \max(d_{ik}, d_{jk}) \tag{2}$$

for all triplets i, j, and k, whereas the additive-tree inequality (four-point condition: Buneman 1974) is satisfied when:

$$d_{ij} + d_{kl} \leq \max (d_{ik} + d_{jl}, d_{il} + d_{jk}) \tag{3}$$

for all quadruplets i, j, k, and l. The K-means partitioning problem can be defined as follows: Find a partition of n objects into K groups, or clusters, such that the sum, over all groups, of the sums of within-group squared distances to the centroids is minimum.

For each of the three clustering problems, a particular loss function (L) is defined to compute optimal weights. In the ultrametric case (dendrograms), optimal weights are found by solving the optimization problem as described by De Soete (1986):

$$L_U(w_1, w_2, \ldots w_m) = \frac{\sum_{\mathbf{\Omega}_U} (d_{ik} - d_{jk})^2}{\sum_{i<j} d_{ij}^2} \quad \rightarrow \quad \min, \tag{4}$$

where $\mathbf{\Omega}_U = \{(i, j, k) \mid d_{ij} \leq \min(d_{ik}, d_{jk})$, and $d_{ik} \neq d_{jk}\}$ denotes the set of ordered triplets for which the distances violate the ultrametric inequality (De Soete 1986). The minimization is done subject to the following constraints:

$$w_1, w_2, \ldots, w_m \geq 0, \tag{5}$$

$$w_1 + w_2 + \ldots + w_m = 1. \tag{6}$$

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

$$L_A(w_1, w_2, \ldots w_m) = \frac{\sum_{\mathbf{\Omega}_A} (d_{ik} + d_{jl} - d_{il} - d_{jk})^2}{\sum_{i<j} d_{ij}^2} \quad \rightarrow \quad \min, \tag{7}$$

subject again to constraints (5) and (6); $\mathbf{\Omega}_A = \{(i, j, k, l) \mid (d_{ij} + d_{kl}) \leq \min(d_{ik} + d_{jl}, d_{il} + d_{jk})$, and $d_{ik} + d_{jl} \neq d_{il} + d_{jk}\}$ denotes the set of ordered quadruplets for which the distances violate the additive inequality (De Soete 1986).

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 clusters K:

$$L_P(w_1, w_2, \ldots w_m) = \sum_{k=1}^{K} \left[\sum_{i,j=1}^{n_k} d_{ij}^2 \right] / n_k \quad \rightarrow \quad \min, \tag{8}$$

subject to constraints (5) and (6); values d_{ij}^2 are the squared distances among objects in cluster k, and n_k is the number of objects in cluster k. The function L_P consists in the sum of the within-cluster sums of squared errors (the external sum in Equation 8), each one being computed as the mean of the squared distances among cluster's members (the internal sum in Equation 8).

We used the Polak-Ribière optimization procedure (see Press, Flannery, Teukolsky and Vetterling 1986, p. 303, and later editions, or Polak 1971, p. 53) to carry out the minimization of L_U, L_A and L_P. First, following De Soete (1986), we reduced the problem, which was originally formulated with constraints (5) and (6), to an unconstrained form, using the type of transformation of variables suggested by Gill, Murray, and Wright (1981, p. 270). The Polak-Ribière optimization method uses first partial derivatives of the functions L_U, L_A and L_P with

respect to the introduced weights. It has proved successful in applications to unconstrained minimization problems; see Press et al. (1986, p. 277, and later editions).

When optimal variable weights have been obtained using L_U or L_A, the dissimilarity matrix **D** among objects can be computed using Equation 1 and subjected to any of the existing ultrametric or additive-tree fitting procedures; see, for example, Arabie, Hubert, and De Soete (1996, pp. 65-199) for an overview of existing fitting algorithms. Alternatively, matrix **D** can be subjected to *K*-means partitioning if optimization has been carried out using loss function L_P. *K*-means partitioning can be computed from either a dissimilarity matrix or a rectangular data matrix; see for instance P. Legendre and L. Legendre (1998, p. 351). The latter option is the most commonly available in computer programs. There are two ways of passing the weights on to a *K*-means algorithm: (a) one can incorporate the weights into the calculation of distances and sums of squares in the *K*-means algorithm itself, as was done in the simulations conducted by Makarenkov and Legendre (2001), or (b), one can transform **D** into a rectangular object-by-variable matrix, preferably by metric scaling (also called principal coordinate analysis, Gower 1966), prior to *K*-means partitioning. Metric scaling is the only way of totally preserving the distance relationships among objects in the subsequent *K*-means procedure; nonmetric scaling would modify the distance relationships among objects.

The optimization methods described above may sometimes produce a local instead of a global minimum of L_U, L_A, or L_P. Hence, a good choice of initial weights is essential. While experimenting with our new program, we realized that making all weights equal to $1/m$ as an initial guess (where m is the number of variables), as implemented in the De Soete program OVWTRE, does not guarantee that the global minimum is always going to be reached. An interesting feature of our optimal variable weighting (OVW) program, compared to OVWTRE, is that it allows users to restart the optimization procedure any number of times, using different random initial configurations for the weights. As a consequence, OVW usually obtains better results than OVWTRE in the case of ultrametric clustering and additive tree reconstruction. Optimization for *K*-means partitioning, which is offered in program OVW, is not available in OVWTRE.

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 nonuniqueness may lead to different dissimilarity matrices **D**, from which different clustering hierarchies or additive trees can be inferred.

Moreover, in the optimization for additive tree reconstruction, degenerate solutions, which are trivial, represent a pervasive problem. Such solutions, which consist in giving a weight of 1 to any one of the variables and weights of 0 to all others, are frequently produced by De Soete's OVWTRE program. The theorem proved in Makarenkov and Legendre (2001) shows that any trivial solution of the type (1, 0, ..., 0), (0, 1, ..., 0), ..., or (0, 0, ..., 1) provides a perfect fit for the additive loss function L_A. In program OVW, we found a way of avoiding, where possible, this trivial solution which leads in most cases to a sub-optimal additive tree: users of the method can set a maximum value for the weight permitted for any single variable. This option effectively prevents obtaining a weight of 1 for a variable, which corresponds to a trivial solution. A numerical example in Section 3 shows how the program OVW works in practice.

An extensive Monte Carlo investigation of De Soete's variable weighting algorithm for hierarchical cluster analysis, based on results provided by De Soete's program OVWTRE, can

be found in Milligan (1989). The simulations reported in Makarenkov and Legendre (2001) focused rather on additive tree reconstruction and *K*-means partitioning.

3. Numerical Example

To demonstrate the effectiveness of the OVW program, we carried out computations on the synthetic data considered by De Soete (1986) to illustrate the usefulness of his weighting procedure for ultrametric trees. De Soete's data, reported in Table 1, possess a clear predefined structure; the first two variables perfectly determine the separation of the objects into clusters. The three clusters {1, 2, 3, 4}, {5, 6, 7, 8} and {9, 10, 11, 12} can easily be deduced from the first two variables which have a clear partitioning structure. The values in variables 3 and 4 are uniform random deviates, unrelated to the other variables and, thus, should not be taken into account when creating the cluster structure, which should be based solely on variables 1 and 2. We will apply to this data set the variable weighting algorithm designed for *additive* and for *ultrametric* clustering as implemented in OVW; note that in his paper, De Soete (1986) only applied the optimal variable weighting procedure for *ultrametric trees* to this data set.

Table 1. *Synthetic data used by De Soete (1986, Table 1) illustrating the application of his optimal variable weighting procedure for ultrametric trees.*

Objects	Variables			
	1	2	3	4
1	0.4082	0.000	0.0564	–0.0188
2	0.4082	0.000	0.7104	0.8879
3	0.4082	0.000	–0.5435	0.4931
4	0.4082	0.000	–0.0227	–0.6123
5	–0.2041	0.3536	0.6128	0.9475
6	–0.2041	0.3536	–0.7937	–0.7604
7	–0.2041	0.3536	–0.2072	–0.0368
8	–0.2041	0.3536	0.3818	0.1197
9	–0.2041	–0.3536	0.9152	0.3362
10	–0.2041	–0.3536	–0.6031	–0.9367
11	–0.2041	–0.3536	0.4861	0.2143
12	–0.2041	–0.3536	–0.3770	–0.0060

First consider the case of the additive tree clustering. Results were produced by OVW using the following options: (a) the optimization procedure was restarted 10 times with different initial estimates; (b) to avoid a trivial solution when a weight of 1 was assigned to a single variable, the maximum allowed weight of a single variable was set to 0.9 (in fact, to force the program to skip a trivial solution, we could choose any other value smaller than 1).

The following vector of optimal weights **w** was obtained: w_1=0.395, w_2=0.605, w_3=0.0, w_4=0.0; the value of the objective function L_A dropped from 0.329523 (when all weights were equal to 0.25) to 0.000007 (for the optimum weights). The correct additive tree structure effectively separating the three clusters could be found from the matrix of weighted distances provided by the program. For the same data set, De Soete's OVWTRE program failed to provide relevant results with the additive tree clustering option and produced only a trivial solution with w_1=0.0, w_2=0.0, w_3=1.0, w_4=0.0; the corresponding value of L_A was 0.

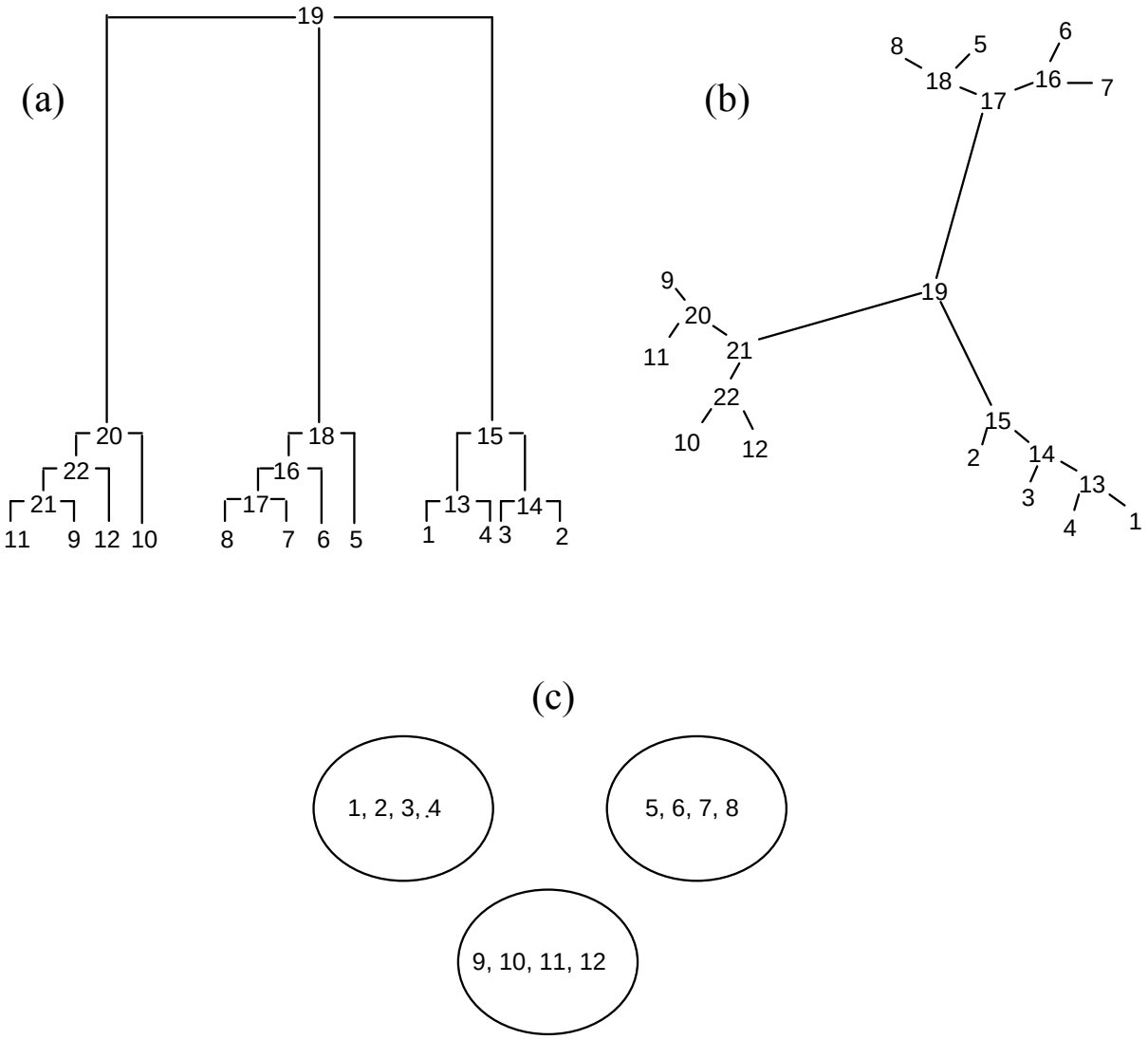

Figure 1. *Classification structures obtained for the Table 1 data using optimal weights computed by OVW. (a) Dendrogram from ultrametric clustering; (b) additive tree; (c) K-means partition.*

However, when OVWTRE was launched with the ultrametric clustering option, it was able to discover a good classification, finding the following set of optimal weights: w_1=0.558, w_2=0.439, w_3=0.000, w_4=0.003. Running the OVW program with the ultrametric clustering option provided a different set of optimal weights: w_1=0.708, w_2=0.292, w_3=0.000, w_4=0.000, which also led to the correct classification.

Finally, when OVW was run on the data from Table 1 using the *K*-means partitioning option, with a correct partition vector supplied to the program separating the 12 objects into 3

groups as (1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3), our *K*-means variable weighting procedure detected the 'noisy' variables in the data and assigned weights of zero to variables 3 and 4. The optimal weights assigned to variables 1 and 2 were respectively 0.906 and 0.094, after 10 starts of the optimization procedure using different initial random configurations for the weights, whereas the minimum value of the objective function L_P dropped from 1.815205 for all weights equal to 0.000000 for the optimal weights. When an incorrect classification vector (1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3) was supplied to OVW, the following weights were obtained for the four variables: w_1=0.909, w_2=0.091, w_3=0.000, w_4=0.000; the minimum value of the objective function L_P corresponding to the solution was 0.937442. This value, which is remote from 0, indicated that the classification vector supplied to the program was not optimal.

The classification structures obtained for the data of Table 1 using optimal weights computed by OVW are depicted in Figure 1. The dendrogram is represented in Part A, the additive tree in Part B, and the *K*-means clusters in Part C of the Figure. In the dendrogram and the additive tree, the interior nodes are numbered 13 to 22.

4. Discussion

The optimal weighting algorithm should be used prior to ultrametric or additive tree clustering, or *K*-means partitioning, if one assumes that the data may contain irrelevant or noisy variables. When the data mostly include error-perturbed variables or outliers, we suggest processing such data using equal weights. Equal or optimal OVW weights can be employed when the data are supposed to be free of errors. It is very difficult to handle error-perturbed data, which is the most complicated case of error condition. As for the outlier condition, we would like to suggest a new strategy which could be tested through simulations. If the data being analyzed are likely to contain more noisy objects than noisy variables, the following strategy could be employed: instead of assigning weights to the variables, weights can be associated with the objects. Using a weighting function that assigns weights of 0 or 1 to the objects would lead to a new objective function to be minimized for ultrametric and additive trees as well as for *K*-means partitioning. Such a strategy may allow one to detect noisy objects rather than noisy variables; weights of 0 would be assigned to the noisy objects. The resulting matrix of predicted dissimilarities $\mathbf{D} = [d_{ij}]$ among objects would be computed as follows:

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

where v_i and v_j are weights associated with the object i and j, respectively. Variants of the objective functions L_U, L_A and L_P should be considered: d_{ij} should be excluded from the objective function if v_i or v_j equal 0. A much more complicated model involving weights for both variables and objects may also be explored. Although the latter model would contain two sets of weights, it may allow one to reduce, at the same time, the effect of noisy variables and noisy objects or outliers. Investigation of weighting strategies implying weights for objects, or for both objects and variables, would constitute an interesting and relevant topic for future research. The OVW program including the methods discussed in this paper is freely available at the following URL: <http://www.fas.umontreal.ca/biol/casgrain/en/labo/ovw.html>.

References

1. *Buneman, P.* (1974), "A Note on the Metric Properties of Trees," Journal of Combinatorial Theory (B), 17, 48-50.

2. *Carmone, F. J., Kara, A., and Maxwell, S.* (1999), "HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables," Journal of Marketing Research, 36, 501-509.

3. *DeSarbo, W. S., Carroll, J. D., Clark, L. A., and Green, P. E.* (1984), "Synthesized Clustering: A Method for Amalgamating Clustering Bases with Differential Weighting of Variables", Psychometrika, 49, 57-78.

4. *De Soete, G.* (1986), Optimal Variable Weighting for Ultrametric and Additive Tree Clustering, Quality & Quantity, 20, 169-180.

5. *De Soete, G.* (1988), "OVWTRE: A Program for Optimal Variable Weighting for Ultrametric and Additive Tree Fitting," Journal of Classification, 5, 101-104.

6. *Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R.* (1988), "Variable Selection in Clustering," Journal of Classification, 5, 205-228.

7. *Gill, P. E., Murray, W., and Wright, M. H.* (1981), Practical Optimization, London: Academic Press.

8. *Gnanadesikan, R., Kettenring, J. R.., and Tsao, S. L.* (1995), "Weighting and Selection of Variables for Cluster Analysis," Journal of Classification, 5, 113-136.

9. *Gower, J. C.* (1966), "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis," Biometrika., 53, 325-338.

10. *Green, P. E., Carmone, F. J., and Kim, J.* (1990), "Variable Selection in Clustering," Journal of Classification, 7, 271-285.

11. *Hartigan, J. A.* (1967), "Representation of Similarity Matrices by Trees," Journal of the American Statistical Association, 62, 1140-1158.

12. *Hubert, L., and Arabie, P.* (1985), "Comparing Partitions," Journal of Classification, 2, 193-218.

13. *Makarenkov, V. and Legendre, P.* (2001), Optimal Variable Weighting for Ultrametric and Additive Trees and K-means Partitioning: Methods and Software, Jr. of Classification, 18, 245-271.

14. *Milligan, G. W.* (1989), "A Validation Study of a Variable Weighting Algorithm for Cluster Analysis," Journal of Classification, 6, 53-71.

15. *Polak, E.* (1971), Computational Methods in Optimization, New York: Academic Press.

16. *Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.* (1986), Numerical Recipes, The Art of Scientific Computing, Cambridge, England: Cambridge University Press.