Adv Data Anal Classif (2008) 2:121–161
DOI 10.1007/s11634-008-0025-4

REGULAR ARTICLE

Fitting semiparametric clustering models to dissimilarity data

Maurizio Vichi

Received: 28 November 2007 / Revised: 7 June 2008 / Accepted: 31 July 2008 /
Published online: 18 September 2008

Abstract The cluster analysis problem of partitioning a set of objects from dissimilarity data is here handled with the statistical model-based approach of fitting the "closest" *classification matrix* to the observed dissimilarities. A classification matrix represents a clustering structure expressed in terms of dissimilarities. In cluster analysis there is a lack of methodologies widely used to directly partition a set of objects from dissimilarity data. In real applications, a hierarchical clustering algorithm is applied on dissimilarities and subsequently a partition is chosen by visual inspection of the dendrogram. Alternatively, a "tandem analysis" is used by first applying a Multidimensional Scaling (MDS) algorithm and then by using a partitioning algorithm such as k-means applied on the dimensions specified by the MDS. However, neither the hierarchical clustering algorithms nor the tandem analysis is specifically defined to solve the statistical problem of fitting the closest partition to the observed dissimilarities. This lack of appropriate methodologies motivates this paper, in particular, the introduction and the study of three new object partitioning models for dissimilarity data, their estimation via least-squares and the introduction of three new fast algorithms.

Keywords Cluster analysis · Partitions · Semiparametric clustering models · LS-estimation

Mathematics Subject Classification (2000) 62h30

M. Vichi (✉)
Department of Statistics, Probability and Applied Statistics,
"Sapienza", University of Rome, Rome, Italy
e-mail: maurizio.vichi@uniroma1.it

1 Introduction

When dissimilarity data are observed or computed on a finite set of multivariate objects, one of the most frequently applied statistical analyses for mining relevant information is unsupervised classification. Objects are clustered into classes with the property that those belonging to the same class have "small" observed pairwise dissimilarities and are perceived as similar to one another; while pairs of objects, belonging to different classes, have "large" observed dissimilarities and are recognized as separate and dissimilar. Such classes generally form partitions of objects and can be detected via clustering methods.

In real applications, analysts normally use two common alternative approaches for partitioning objects from dissimilarities. In the first, a hierarchical clustering technique is applied (Bock 1974; Hartigan 1975; Jain and Dubes 1988; Kaufman and Rousseeuw 1990; Gordon 1999), and subsequently a partition is chosen by visual inspection of the resulting dendrogram or by considering a cluster validity index for assessing the best partition from the dendrogram. However, hierarchical clustering algorithms are not specifically designed for the identification of a partition in the observed dissimilarities, since they detect a possibly optimal indexed hierarchy (i.e., a set of nested partitions) and thus a single partition may be chosen only as a byproduct of the hierarchical method. The second common approach uses a *tandem analysis* for dissimilarity data, by applying first a Multidimensional Scaling (MDS) method on dissimilarities and subsequently a partitioning algorithm such as k-means or Gaussian mixture model, that is applied to the (low-)dimensional point configuration resulting from MDS. However, in this case also there is no direct identification of a partition from the observed dissimilarities. This implies that the dimensions detected by the MDS for approximating dissimilarities via Euclidean distances are not necessarily the best ones for the following partitioning step for objects, because MDS optimizes a loss function that is not directly connected with the underlying clustering problem. The lack of appropriate and widely employed methodologies for partitioning objects directly from dissimilarity data motivates this work. Besides, we intend to propose new clustering methods following the model-based approach that has received attention, becoming popular and attractive in the case of the usual two-mode (objects $\times$ variables) data, i.e., when objects coordinates, (data vectors) are given (for reviews see Bock 1998; McLachlan and Peel 2000; Fraley and Raftery 2002).

Model-based clustering in the case of dissimilarity data is the statistical approach of fitting the "closest" *classification matrix* to data. A classification matrix is a clustering model for dissimilarities, i.e., a special dissimilarity matrix with a semiparametric form representing a partition described by two characteristics: *heterogeneity* for each class and *isolation* between classes (see the beginning of Sect. 3 for a complete description). The estimation of the parameters of the model (heterogeneities and isolations), in this paper, is based on the Least-Squares (LS) method, while Maximum Likelihood estimation is the topic of a following paper. This implies that the partitions obtained via model-based clustering have optimality properties connected with these two widely used statistical estimation methods and give a framework for putting cluster analysis on a principled statistical footing (Oh and Raftery 2007).

Specifically, three new clustering models for dissimilarity data are proposed here.

The first classification matrix hypothesizes the most parsimonious semiparametric model with equal heterogeneities for each class of the partition and equal isolations between classes; while the second relaxes these constraints allowing for different heterogeneities and isolations. The third classification matrix is equal to the second one with the additional property of estimating a supposing existing hierarchical relationship between classes.

It can be noted that the LS estimation of the first classification matrix where equal heterogeneity and equal isolation are assumed is linked, in the special case discussed in Sect. 4, to the well-known clique partitioning problem (Régnier 1965). Furthermore, it is useful to notice that when the classification matrix is a general ultrametric the semiparametric approach of fitting an ultrametric matrix to dissimilarity data in a least-squares sense has been discussed in Chandon et al. (1980), Carroll and Pruzansky (1975, 1980) and De Soete (1984); Hubert et al. (1997), even though these methodologies cannot be directly used to fit a partition as it is assumed in this paper. In particular, Chandon et al. (1980); Hubert et al. (1997) obtained optimal solutions for general indexed hierarchies for relatively small data sets by using branch-and-bound and dynamic programming methods. Carroll and Pruzansky (1975, 1980) and De Soete (1984) added to the least-square loss function a second function that penalizes the triplets of dissimilarities failing to satisfy the ultrametric inequality.

An outline of this paper is as follows. In Sect. 2 some basic notions necessary to describe the modeling approach to cluster analysis are discussed. Section 3 defines the modeling approach of clustering when the observed data are dissimilarities and three classification matrices defining three partitioning structures are described. Section 4 shows the semi-parametric LS estimation of the three proposed classification matrices together with details on the corresponding algorithms.

The proposed algorithms have been tested in a simulation study reported in Sect. 5, and with a data set analyzed in the literature, in Sect. 6. A final discussion follows in Sect. 7.

1.1 Motivating and illustrative examples

Before we discuss in detail the different partitioning models we wish to present synthetic experimental examples, including Monte Carlo simulations, showing how the models behave in comparison with some known clustering methods and approaches for partitioning objects from dissimilarity data.

Example (1). The first model (classification matrix) describes a clustering (partition), where equal heterogeneities α_1 for each class and equal isolations α_2 between classes are hypothesized (for details see Sect. 3.1). Rubin (1967) defines such partitions as *well-structured perfect in K clusters*. Note that when a partition is selected from a dendrogram at a fixed level (height) α_1, without preserving the levels below α_1, a well-structured perfect partition is implicitly chosen with heterogeneity α_1 and isolation α_2 equal to the highest height of the dendrogram. Therefore, the LS estimation of the first clustering model represents the model-based formal direct solution to the common practice of choosing a partition by visual inspection of the dendrogram obtained by an agglomerative hierarchical clustering method applied on the observed dissimilarity data.

An example can help to motivate and understand the use of the method proposed in this paper. A well-structured perfect partition of $n = 20$ objects into $K = 3$ clusters (C_1, C_2, C_3) has been generated with classes $C_1 = \{1, 2, 5, 10, 11, 12, 19, 20\}$, $C_2 = \{3, 7, 8, 9, 13, 16, 18\}$ and C_3=$\{4, 6, 14, 15, 17\}$. The heterogeneities for classes and isolations between classes have been fixed all equal to $\alpha_1 = 40$ and $\alpha_2 = 50$, respectively. The associated classification matrix has been perturbed by adding a left-truncated normal random error, like in Oh and Raftery 2007 ($\mu = 0, \sigma = 9$) to preserve non-negativity (i.e., $d_{il} = (\alpha_1 + N(0,9)$: $d_{il} \geq 0)$ if $(i, l) \in C_k$ for $k = 1, 2, 3$; otherwise $d_{il} = (\alpha_2 + N(0,9)$: $d_{il} \geq 0)$ if $i \in C_k$ and $l \in C_m$ for $k, m = 1, 2, 3$; $k \neq m$). The obtained dissimilarity matrix is represented by a heat map in Fig. 1a. The least-squares estimation of the well-structured perfect partition proposed in this paper gives the correct partition described above (represented in Fig. 1b as a dendrogram with only two different heights). Suppose that a researcher would try to find the well-structured perfect partition by cutting the dendrogram from the agglomerative group average linkage (UPGMA), method (see Fig. 1c). For 3 clusters the following partition is obtained: $C_1 = \{1, 3, 19\}, C_2 = \{7, 8, 9, 13, 16, 17, 18\}$ and

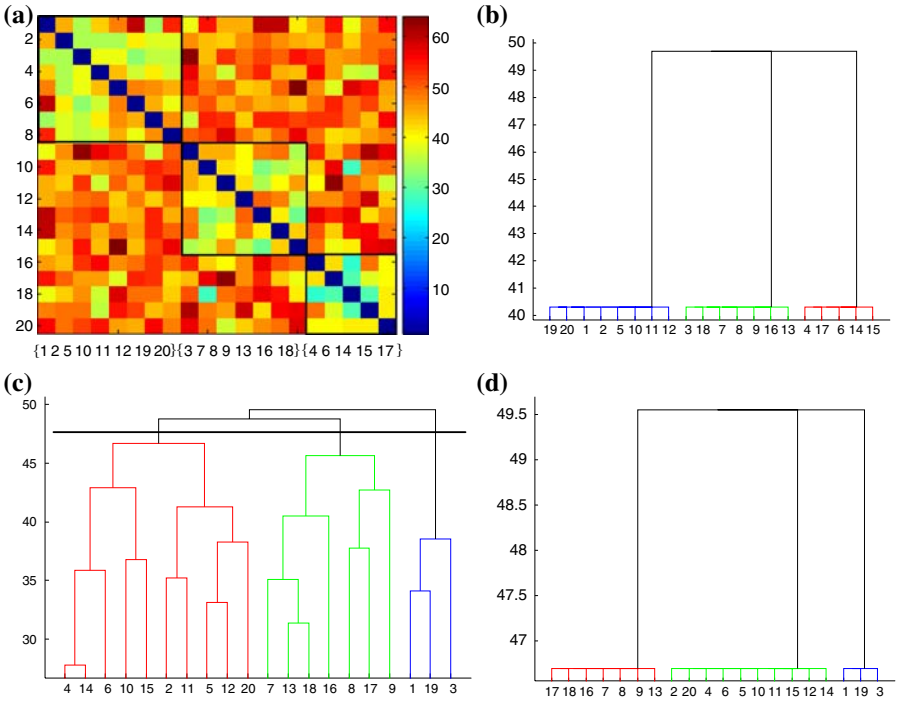


Fig. 1 **a** Heat map of the (20 × 20) dissimilarity matrix obtained by perturbing the classification matrix associated to a well-structured perfect partition in three clusters with heterogeneity within classes equal to $\alpha_1 = 40$ and isolation between classes equal to $\alpha_2 = 50$ (note that within highlighted squares, dissimilarities are generally *yellow* or *light blue* (*light grey* in print), while outside squares are generally *light* or *dark red* (*dark grey* in print). **b** Estimated least-squares well-structured partition. **c** Dendrogram from the UPGMA applied on dissimilarity matrix represented in **a**. **d** Partition obtained by cutting the dendrogram in **c** to obtain a partition in three clusters

$C_3 = \{2, 4, 5, 6, 10, 11, 12, 14, 15, 20\}$, (see Fig. 1d) which is quite different from the original one above. Therefore, the UPGMA fails to identify the correct well-structured perfect partition present in the dissimilarity data.

One may think that this is only limited to one artificial example, but by repeating the experiment with a Monte Carlo simulation on other 100 sampled dissimilarity matrices with a well-structured perfect partition generated as in the previous example (three classes with about equal size on average, and $\alpha_1 = 40$ and $\alpha_2 = 50$) plus left-truncated normal error, our least-squares estimation of the well-structured perfect partition has found exactly the true generated one in 85 cases, with a very satisfactory average Modified Rand index $MRand = 0.97$ (Hubert and Arabie 1985) between the generated well-structured perfect partition and the partition estimated by our method. Similarly, UPGMA was applied on the same 100 sampled dissimilarity matrices, but not surprisingly the true partition was found only 42 times with an average $MRand = 0.84$. Therefore, it can be concluded that the common practice of choosing a partition in the dendrogram from the UPGMA is not very appropriate to find a well-structured perfect partition in the dissimilarity data and in this situation it is better to use our proposed method.

Example (2). The second clustering model, discussed in this paper, relaxes the equality of heterogeneity for each class and isolation between classes, respectively, insofar as it allows for defining different heterogeneities and isolations so as to specify a *well-structured partition in K clusters* (Rubin 1967). The least-squares estimation of the well-structured partition in K clusters specifies a new algorithm called *square K-means* (see Sect. 4.3) that resembles the K-means algorithm (Ball and Hall 1967; MacQueen 1967) except that it can be applied on dissimilarity data.

Also in this case an example can help to understand this new clustering model proposed in the paper and motivate its usage. We generated a well-structured partition of $n = 20$ objects into $K = 3$ classes $C_1 = \{2, 3, 4, 6, 10, 12, 16\}$, $C_2 = \{1, 5, 8, 9, 11, 14, 15, 18\}$ and $C_3 = \{7, 13, 17, 18, 20\}$. The heterogeneities for classes C_1, C_2 and C_3 are chosen to be equal to ${}_{\mathrm{W}}d_{11} = 39.8$, ${}_{\mathrm{W}}d_{22} = 41.1$ and ${}_{\mathrm{W}}d_{33} = 43.7$, respectively, and the isolations between classes (C_1, C_2), (C_1, C_3) and (C_2, C_3) are equal to ${}_{\mathrm{B}}d_{12} = 52.2$, ${}_{\mathrm{B}}d_{13} = 51.5$, ${}_{\mathrm{B}}d_{23} = 48.8$, respectively. The associated classification matrix has been perturbed by a left-truncated normal random error ($\mu = 0, \sigma = 10$); it is illustrated by a heat map in Fig. 2a. The least-squares estimation of the well-structured partition proposed in this paper (via *square K-means*) reproduces the true partition described above (see Fig. 2b). Suppose an analyst would try to find the well-structured partition more simply by using the tandem analysis described above and frequently used in practice. First, the classical MDS (Torgerson 1958; Gower 1966) on the dissimilarity matrix in Fig. 2a is applied. This matrix is not Euclidean, in fact its doubly centered version is not positive definite, and has 3 negative eigenvalues. The optimal LS reconstruction and the corresponding configuration in $\Re^{17}$ is obtained by replacing the negative eigenvalues with zeros (Keller 1962; Mathar 1985). Subsequently, *K-means* has been applied on the 17 dimensions with positive eigenvalues which results in the partition $C_1 = \{2, 3, 4, 6, 10, 12, 16\}$, $C_2 = \{1, 5, 9, 11, 15, 19\}$ and $C_3 = \{7, \mathbf{8}, 13, \mathbf{14}, 17, 18, 20\}$. Obviously it has two misclassified objects in comparison with the generated well-structured one.

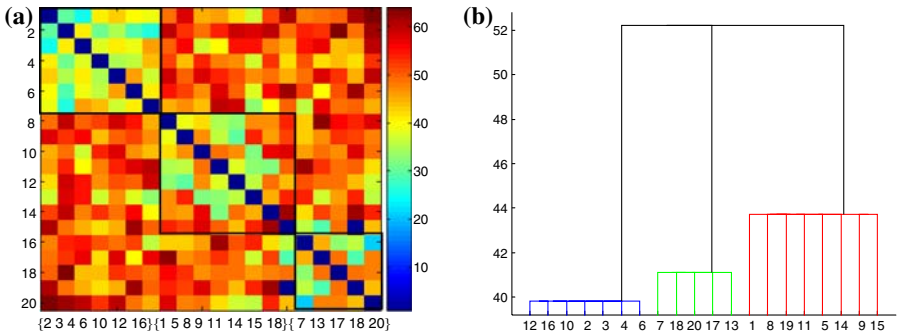


Fig. 2 **a** Heat map of the (20 × 20) dissimilarity matrix obtained by perturbing the classification matrix associated to a well-structured partition in three clusters with different heterogeneities for the three classes and different isolations between classes. **b** Square K-means partition in three classes represented as a dendrogram with four levels, the first three corresponding to the values of heterogeneity for the three clusters

By repeating the experiment, via Monte Carlo simulations, on other 100 dissimilarity matrices, generated as in the second example, our least-squares estimation of the well-structured partition via *square K-means* has found the correct partition 72 times with an average Modified Rand index $MRand = 0.92$. In addition, the tandem analysis was applied on the same 100 sampled dissimilarity matrices to identify the well-structured partition, but it was found only 62 times with an average $MRand = 0.88$. Therefore, the common practice to apply MDS followed by *K-means* is less appropriate to identify well-structured partitions than the *square K-means* proposed here. This is due to the fact that MDS in any case only attains approximations to the given dissimilarities, unless they are already Euclidean, and the dimensions found are not necessarily the best for *K-means* to detect the optimal partition in the data.

Example (3). Our third clustering model (classification matrix), is equal to the second one with the additional property to assume a hierarchical relationship between clusters (hierarchical isolation). In other terms, it is required that the isolations between classes satisfy the ultrametric inequality and specify a hierarchy into at most $2K - 1$ nested classes.

This third clustering model is called *hierarchical partition* and can be considered to construct a *parsimonious tree*, i.e. a hierarchy containing a limited number of internal nodes (Gordon 1999). Some algorithms have been proposed for seeking parsimonious trees directly from (dis)similarity data (Hartigan 1967; Sriram 1990) and by simplifying complete hierarchical classifications (Gordon 1987). Also in this case, an example may help to motivate and understand our third model for clustering objects from dissimilarity data. A hierarchical partition of $n = 20$ objects into $K = 3$ clusters has been generated $C_1 = \{1, 2, 3, 7, 10, 11, 12, 13\}$, $C_2 = \{8, 14, 19, 20\}$, $C_3 = \{4, 5, 6, 9, 15, 16, 17, 18\}$, with heterogeneities for classes C_1, C_2 and C_3 equal to ${}_{\mathrm{W}}d_{11} = 38.6$, ${}_{\mathrm{W}}d_{22} = 40.6$ and ${}_{\mathrm{W}}d_{33} = 39.5$, respectively and isolations between pairs of classes: (C_1, C_2), (C_1, C_3) and (C_2, C_3) equal to ${}_{\mathrm{B}}d_{12} = 47.6$, ${}_{\mathrm{B}}d_{13} = 48.4$, ${}_{\mathrm{B}}d_{23} = 48.4$, respectively. Since two isolations values are equal they satisfy the ultrametric inequality and specify a hierarchical relation between classes. The associated classification matrix has been perturbed by a left-truncated normal random

error ($\mu = 0, \sigma = 9$) and represented in Fig. 3a. The least-squares estimation of the hierarchical partition proposed in this paper, via a specific algorithm named *square hierarchical K-means* gives the same partition above described (see in Fig. 3b as a dendrogram with $2K - 1 = 5$ heights). A researcher could try to find the hierarchical partition by cutting the dendrogram of the group average linkage (UPGMA), (see Fig. 3c) at the height of the partition in three clusters and retaining the linkages above this height, thus obtaining: $C_1 = \{1, 2, 3, 7, 10, 11, 12, 13, 15\}, C_2 = \{19\}, C_3 = \{4, 5, 6, 8, 9, 14, 16, 17, 18, 20\}$ (see Fig. 3d, by considering also the heights above the cut of the dendrogram in Fig. 3c) which is different from the one above generated. Therefore, the UPGMA fails to identify the correct hierarchical partition given in the dissimilarity data.

By repeating, via Monte Carlo simulation, the experiment on other 100 sampled dissimilarity matrices, generated as in the previous example, our least-squares-estimation of the hierarchical partition has found the true one exactly 74 times, with a satisfactory average Modified Rand index $MRand = 0.95$. The UPGMA has been applied on the same 100 sampled dissimilarity matrices, but the given true partition has been found only 42 times with an average $MRand = 0.83$. Therefore, also in this hierarchical case, it can be concluded that the practice of choosing a partition in the dendrogram of the UPGMA retaining also the heights above the chosen cut is not appropriate to find

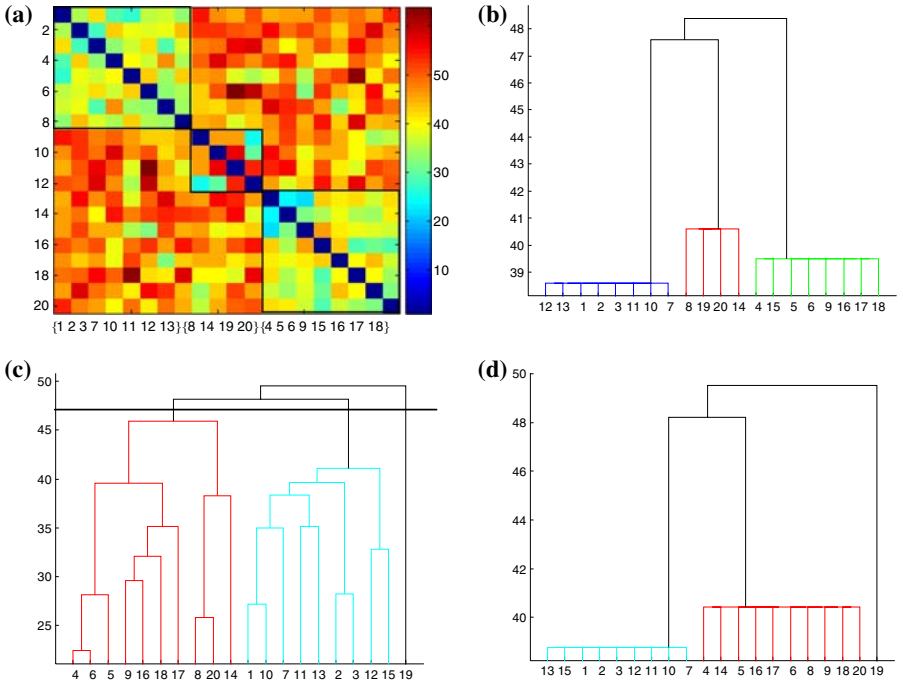


Fig. 3 **a** Heat map of the (20×20) dissimilarity obtained by perturbing the classification matrix associated to a hierarchical partition in three clusters with different heterogeneity within classes and hierarchical isolation between classes. **b** Estimated least-square hierarchical partition obtained via square hierarchical K-means. **c** Dendrogram of the UPGMA applied on dissimilarity matrix represented in **a**. **d** Partition obtained cutting the dendrogram in Fig. 1c to obtain a partition in three clusters

a hierarchical partition existing in the dissimilarity data, and even in this situation a more appropriate method is needed.

2 Basic clustering notions

For the convenience of the reader some basic notions necessary for the modeling approach to cluster analysis are briefly listed here.

A *partition* $C = \{C_1, \ldots, C_k, \ldots, C_K\}$ of the set of n objects $O = \{o_1, \ldots, o_n\}$ is a set of K disjoint non-empty subsets, such that their union is O itself. Each class C_k has a cardinality n_k.

The identification of the clustering structures in the data is strictly connected to the notion of dissimilarity between objects. In general, a dissimilarity between objects o_i, o_l is a function d satisfying the following properties: (i) $d_{il} \geq 0$, $(i, l = 1, \ldots, n)$; (ii) $d_{ii} = 0$ for $(i = 1, \ldots, n)$; (iii) $d_{il} = d_{li}$ for $(i, l = 1, \ldots, n)$. Several dissimilarity measures have been proposed in cluster analysis literature, but this problem does not concern this paper (for an extensive discussion, see for example Gordon 1999). A dissimilarity matrix $\mathbf{D}$ on O is a $(n \times n)$ matrix $\mathbf{D} = [d_{il}]$, whose elements represent dissimilarities between $o_i, o_l (i, l = 1, \ldots, n)$. The matrix $\mathbf{D}$ is a *metric* matrix, if its triplets satisfy the triangle inequality: (iv) $d_{il} \leq d_{ik} + d_{lk} \forall (i, l, k) \in O \times O \times O$.

Two well-known properties characterize the classes $C_k (k = 1, \ldots, K)$ of a partition: *isolation*and *heterogeneity*. The *isolation between classes* C_k and C_h is the extent to which C_k is dissimilar or separate from C_h. The isolation in this paper is evaluated as a function of the dissimilarities between all pairs of objects, where one belongs to C_k and the other to C_h. The *heterogeneity* or *lack of cohesion within a class* C_k is the extent to which, objects within C_k are dissimilar or separate one from the other. Similarly to isolation the heterogeneity of a class is here estimated as a function of dissimilarities between objects belonging to C_k. These concepts of heterogeneity and isolation are related to the internal cohesion and external isolation described by Cormack (1971).

For a partition C in K classes there are $K(K-1)$ measures of isolation for pairs of classes C_k and $C_h (k, h = 1, \ldots, K)$ that can be arranged in a square dissimilarity matrix $\mathbf{D}_\mathrm{B} = [{}_\mathrm{B}d_{kh}]$ of order K, where ${}_\mathrm{B}d_{kh} \geq 0$ denotes the measure of isolation between pairs of classes $(C_k, C_h) k, h = 1, \ldots, K$ of the partition C; hence, ${}_\mathrm{B}d_{kk} = 0$ for all k. For the same partition C in K classes there is a measure of heterogeneity for each class $C_k (k = 1, \ldots, K)$, that can be arranged on the main diagonal of a diagonal matrix $\mathbf{D}_\mathrm{W}$, where ${}_\mathrm{W}d_{kh}$=0 for all $k \neq h$ and ${}_\mathrm{W}d_{kk} \geq 0$ is the measure of the heterogeneity of class $C_k, k = 1, \ldots, K$. We are now in position to formulate the statistical modelling approach to cluster analysis.

3 Clustering models for dissimilarity data

Suppose that a dissimilarity matrix $\mathbf{D}$ has been observed or computed on pairs of objects belonging to the set O. Cluster analysis through matrix $\mathbf{D}$ is here handled by the statistical model-based approach of fitting an expected *clustering model* (e.g., a partition, hierarchy, pyramid, etc.) which is characterized by a specific dissimilarity

classification matrix $\mathbf{D}_c = [{}_c d_{il}]$ to the dissimilarity matrix $\mathbf{D}$. Formally, the clustering problem can be statistically specified by the following "error model"

$$\mathbf{D} = \mathbf{D}_c + \mathbf{E}, \tag{1}$$

where $\mathbf{E} = [e_{ij}]$ is a $(n \times n)$ random error matrix describing the part of the observed dissimilarity matrix $\mathbf{D}$ which is not explained by the classification matrix $\mathbf{D}_c$. It represents the extent to which an observed dissimilarity matrix $\mathbf{D}$ differs from its classification model represented by $\mathbf{D}_c$. In order to complete the model description, the random matrix $\mathbf{E}$ needs to be specified. Customary specifications are that the expected value of e_{il} is zero, i.e., $E(e_{il}) = 0$. Then $E(\mathbf{D}) = \mathbf{D}_c$ is the expected model for the dissimilarity data. In this paper, the semi-parametric least-squares estimation method will be adopted to estimate $\mathbf{D}_c$ (insofar, the distribution of $\mathbf{E}$ will not be specified).

The classification matrix $\mathbf{D}_c$ is a dissimilarity matrix that specifies the details of the clustering structure and generally satisfies some further constraints on the triplets ${}_c d_{il}, {}_c d_{ik}, {}_c d_{lk}$ of its elements. The constraints are necessary to guarantee that the classification matrix is associated to a specific clustering model (e.g., partition, covering, hierarchy, etc.).

3.1 Well-structures perfect partition: equal heterogeneity and isolation

The most parsimonious model for describing a clustering by using dissimilarities assumes equal heterogeneity for each class of C and equal isolation between pairs of classes (see Fig. 4). Suppose that value α_1 measures the heterogeneity of the objects within each class of C; while α_2 evaluates the isolation between classes. Of course, it is suitable that $\alpha_1 \leq \alpha_2$, because this implies that a function of the within cluster dissimilarities is smaller than a function of the between cluster dissimilarities. Rubin (1967) denotes data specifying such model a *well-structured perfect* partition and Fisher and Van Ness (1971) provide a form of admissibility for this type of partitioning model. Thus, this classification model will be called *well-structured perfect partition*.

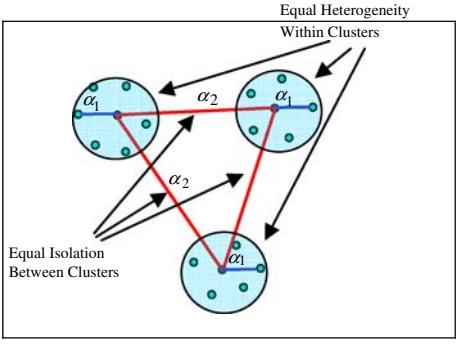


Fig. 4 Well-structured perfect partition with equal heterogeneity and isolation

The dissimilarity classification matrix $\mathbf{D}_c$, associated to a well-structured perfect partition is denoted by $\mathbf{P}$ and can be written as a function of α_1, α_2 and a $(n \times n)$ joint membership matrix $\mathbf{S}$ identifying the partition. $\mathbf{S}$ is a similarity matrix, where entries $s_{ii} = 1$, $(i = 1, \ldots, n)$, and s_{il}=1 (resp., 0), if the ith and the lth objects belong (resp., do not belong) to the same class of the partition C $(i, l = 1, \ldots, n)$.

Thus, formally, the semiparametric *well-structured perfect classification matrix* $\mathbf{P}$ is given by

$$\mathbf{D}_c = \mathbf{P} = \alpha_2(\mathbf{1}_n\mathbf{1}_n' - \mathbf{S}) + \alpha_1(\mathbf{S} - \mathbf{I}_n) = \alpha_2(\mathbf{1}_n\mathbf{1}_n' - \mathbf{M}\mathbf{M}') + \alpha_1(\mathbf{M}\mathbf{M}' - \mathbf{I}_n). \tag{2}$$

where $\mathbf{1}_n$ is a vector of n ones, $\mathbf{I}_n$ is the identity matrix of order n and $0 < \alpha_1 \leq \alpha_2$ and matrix $\mathbf{M}$ is a $(n \times K)$ membership matrix, binary and row-stochastic, i.e., with one nonzero element per row specifying a partition C of objects in K classes. In fact, $m_{ik} = 1$ if the ith object o_i belongs to the kth class C_k, $m_{ik} = 0$ otherwise.

In this model all the within clusters dissimilarities are supposed equal to α_1 while all the between clusters dissimilarities are hypothesized equal to α_2. Thus, $\mathbf{P}$ is characterized by $(\mathbf{M}, \alpha_1, \alpha_2)$.

Definition 1 (*Vicari and Vichi 2000*) A 2-*ultrametric matrix*, is an ultrametric matrix with off-diagonal elements that can assume one of at most 2 different values $0 < \alpha_1 \leq \alpha_2$. Formally, $\mathbf{P} = [p_{il}]$, $p_{il} \geq 0$, $p_{il} = p_{li}$, $p_{il} \leq \max(p_{ik}, p_{lk})$ $\forall(i, l, k)$ and $p_{il} \in \{0, \alpha_1, \alpha_2\}$ $\forall(i, l)$, with $0 < \alpha_1 \leq \alpha_2$.

Remark 1 Matrix $\mathbf{P}$ is a metric matrix and in particular a 2-*ultrametric matrix.*

Example 1 Given a partition $C = \{C_1 = \{1, 2, 3\}, C_2 = \{4\}, C_3 = \{5, 6\}\}$ of 6 objects in 3 clusters, the following $\mathbf{P}$ is associated to C by fixing $p_{ij} = \alpha_1$ if objects o_i and o_l belong to the same class or $p_{il} = \alpha_2$ if objects o_i and o_l belong to different classes. This is a 2-ultrametric matrix where all triplets satisfy ultrametric inequality.

$$\mathbf{P} = \begin{array}{|c|c|c|c|c|c|c|} \hline & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & \alpha_1 & \alpha_1 & \alpha_2 & \alpha_2 & \alpha_2 \\ \hline 2 & & 0 & \alpha_1 & \alpha_2 & \alpha_2 & \alpha_2 \\ \hline 3 & & & 0 & \alpha_2 & \alpha_2 & \alpha_2 \\ \hline 4 & & & & 0 & \alpha_2 & \alpha_2 \\ \hline 5 & & & & & 0 & \alpha_1 \\ \hline 6 & & & & & & 0 \\ \hline \end{array}$$

Any 2-ultrametric matrix can be represented by a parsimonious tree, i.e., a dendrogram with at most two levels (heights of the tree), briefly named 2-dendrogram. For example in Fig. 5 the partition for $n = 15$ objects in five clusters $C_1 = \{4, 13, 15\}$, $C_2 = \{2, 5, 9, 14\}$, $C_3 = \{7, 8, 11\}$, $C_4 = \{6, 10, 12\}$, $C_5 = \{1, 3\}$ is represented. The heterogeneity for each cluster is equal to 40.5 and the isolation between clusters is equal to 50.2.

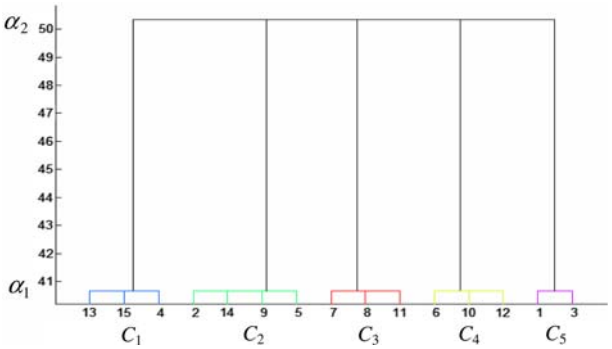


Fig. 5 Representation of a 2-dendrogram. Five clusters $(C_1, \ldots, C_5)$ form a partition C. The values $\alpha_1 = 40.5$ and $\alpha_2 = 50.2$

Lemma 1 *A well-structured perfect partition C of O with heterogeneity α_1 for each class and isolation α_2 for each pair of classes of C (with $\alpha_1 \leq \alpha_2$) has associated one-to-one a 2-ultrametric matrix* $\mathbf{P}$ *with values* $\{0, \alpha_1, \alpha_2 : \alpha_1 \leq \alpha_2\}$.

Proof In a well-structured perfect partition C with heterogeneity α_1 and isolation α_2, every element o_i of O belongs to only one class. Thus, any triplet (o_i, o_l, o_k) belongs, alternatively, to: (a) single set of C; (b) two distinct sets, e.g., two elements (o_i, o_l) belong to a set of C and o_k to another set of C; (c) three distinct sets of C. Cases (a), (b), and (c) determine only three types of distance triplets, respectively: $(\alpha_1, \alpha_1, \alpha_1)$, $(\alpha_1, \alpha_2, \alpha_2)$ and $(\alpha_2, \alpha_2, \alpha_2)$, which, all verify the ultrametric inequality. Thus, $\mathbf{P}$ is a 2-ultrametric matrix. Conversely, $\mathbf{P}$ obviously specifies a covering, (i.e., a set of classes not necessarily disjoint), because $p_{il} = \alpha_1$ is the distance between objects o_i and o_l belonging to the same class; while $p_{il} = \alpha_2$ is the distance between objects o_i and o_l belonging to different classes. Moreover, since $\mathbf{P}$ is an ultrametric matrix with values $\{0, \alpha_1, \alpha_2\}$ and no distance triplets of type $(\alpha_1, \alpha_1, \alpha_2)$ can be observed, (implying classes with elements in common), thus it follows that the covering is actually a partition C with heterogeneity α_1 and isolation α_2 □

Lemma 2 (Bijection between well-structured perfect partitions and 2-ultrametric matrices) *Let $\mathscr{C}$ be the set of well-structured perfect partitions of O with heterogeneity α_1 and isolation α_2 and let $\mathscr{D}_P$ be the set of 2-ultrametric matrices* $\mathbf{P}$ *with values* $\{0, \alpha_1, \alpha_2\}$ *then, there exists a bijection between $\mathscr{C}$ and $\mathscr{D}_P$.*

Proof Lemma 1 can be applied to any partition $C \in \mathscr{C}$ with heterogeneity α_1 and isolation α_2 and classification matrix $\mathbf{P} \in \mathscr{D}_P$ with values $\{0, \alpha_1, \alpha_2 : \alpha_1 \leq \alpha_2\}$. Thus, each element of $\mathscr{C}$ has associated one element of $\mathscr{D}_P$ and vice versa. □

Remark 2 It is worth to note that there is a dimensional restriction for the data vectors that represent the dissimilarity matrix corresponding to the well-structured perfect partition in a J dimensional space. The n objects representing a well-structured perfect partition must lie in a at least J dimensional space where

$$J \geq \max(n_k : k = 1, \ldots, K) - 1. \tag{3}$$

In fact, the n_k units of each class C_k need at least $J = n_k - 1$ dimensions to be represented with pairwise distance equal to α_1. Indeed, the n_k units represent vertices of a regular polytope if and only if they lie in (at least) a $n_k - 1$-dimensional space (e.g., three points are vertices of an equilateral triangle with side α_1, that lies in a two-dimensional space).

Remark 3 It is also useful to note that when a partition is selected from an arbitrary dendrogram, by cutting it at a fixed level α_1, and no other information is retained on the partitions below and above α_1, a well-structured perfect partition is implicitly chosen with heterogeneity α_1 and isolation α_2 equal to the maximum level (height) in the dendrogram.

Therefore, it is important to note that the LS estimation of a well-structured perfect partition directly from the observed dissimilarity data represents the model-based formal direct solution to the common practice of choosing a partition by visual inspection of the dendrogram associated to an agglomerative hierarchical clustering method applied on the observed dissimilarity data.

3.2 Well-structured partition: different heterogeneities and isolations

Let us now relax the hypothesis of equal heterogeneities for all classes and equal isolations between classes of the partition. A second more flexible classification matrix allows for specifying different heterogeneities and isolations (Fig. 6). This model has some similarities with the one described by Bock (1998) in section 2.1.4. Given a partition $C = \{C_1, \ldots, C_k, \ldots, C_K\}$ of O such that objects within C_k have heterogeneity ${}_{\mathrm{W}}d_{kk} > 0\,\forall k$, while objects between C_k and C_h have isolation ${}_{\mathrm{B}}d_{kh} > 0\,\forall k, h(k \neq h)$, a dissimilarity classification matrix $\mathbf{Q}$—defined as a function of the square matrices of order K, $\mathbf{D}_{\mathrm{B}} = [{}_{\mathrm{B}}d_{kh} > 0 : {}_{\mathrm{B}}d_{kk} = 0, h, k = 1, \ldots, K(k \neq h)]$, and $\mathbf{D}_{\mathrm{W}} = [{}_{\mathrm{W}}d_{kk} > 0 : {}_{\mathrm{W}}d_{kh} = 0, h, k = 1, \ldots, K(k \neq h)]$, and a membership matrix $\mathbf{M}$—can be

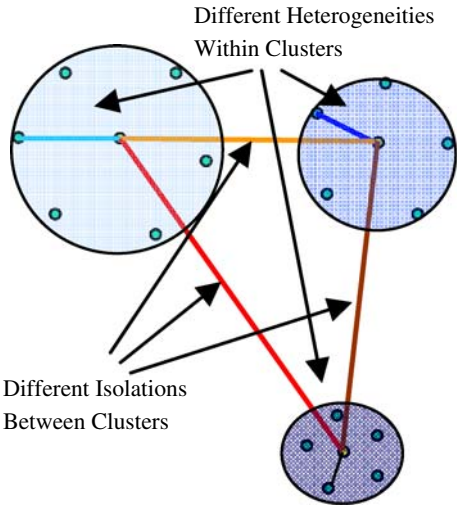


Fig. 6 Well-structured partition with different heterogeneity measures and different isolation measures

one-to-one associated to C. The semiparametric *well-structured classification matrix* $\mathbf{Q}$ has the form,

$$\mathbf{D}_c = \mathbf{Q} = \mathbf{M}\mathbf{D}_{\mathrm{B}}\mathbf{M}' + \mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}' - \mathrm{diag}(\mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}'). \tag{4}$$

Therefore, the classification matrix $\mathbf{Q}$ specifies a more flexible kind of partition, where both the expected clusters heterogeneities and the expected between clusters isolations may differ.

If we define $\mathbf{D}_{\mathrm{B}} = \alpha_2(\mathbf{1}_K\mathbf{1}'_K - \mathbf{I}_K)$, and $\mathbf{D}_{\mathrm{W}} = \alpha_1\mathbf{I}_K$, and $\alpha_1 \leq \alpha_2$, then matrix (4) coincides with matrix (2), i.e., $\mathbf{Q} = \mathbf{P}$ and therefore matrix $\mathbf{Q}$ has associated a well-structured perfect partition.

If $K = n$, i.e., there are n singletons, it follows that $\mathbf{D}_{\mathrm{B}} = \mathbf{D}$ and $\mathbf{D}_{\mathrm{W}} = \mathbf{0}$, and, trivially, model (4) perfectly fits data. Model (4) can be rewritten in a form that does not involve a classification matrix, but that can be still used to fit the dissimilarity data,

$$\mathbf{R} = \mathbf{M}(\mathbf{D}_{\mathrm{B}} + \mathbf{D}_{\mathrm{W}})\mathbf{M}'. \tag{5}$$

Matrix $\mathbf{R}$ is not a classification matrix because it is not a dissimilarity matrix. In fact, on the diagonal non null values representing within classes heterogeneities are reported.

However, as we will see in the following section, and we have anticipated in the introduction, fitting this matrix to dissimilarity data will allow for defining a fast clustering algorithm for fitting a partition to dissimilarity data. The classification matrix $\mathbf{Q}$ and the matrix $\mathbf{R}$ specify a *well-structured* partition in K clusters if the largest heterogeneity of a class of C is smaller than or equal to the smallest isolation between two classes of C, i.e., if

$$\max\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots K\} \leq \min\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (h \neq k)\}, \tag{6}$$

that is, within clusters heterogeneity values are not greater than between clusters isolation values. Condition (6) could be considered to be too restrictive; consequently, it can be replaced by less stringent conditions such as

$$\mathrm{mean}\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots K\} \leq \mathrm{mean}\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (k \neq h)\}; \tag{7}$$

$$\mathrm{median}\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots K\} \leq \mathrm{median}\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (k \neq h)\}. \tag{8}$$

3.3 Hierarchical partitioning model

Also the third classification matrix has associated a partition with the interesting and useful property to include the hierarchical relationship between classes as shown in Fig. 7. Thus, this classification matrix specifies a partition in K classes together with the hierarchical structure among classes and, this implies also the knowledge of the nested partitions in $K - 1, K - 2, \ldots, 1$ classes.

Formally a partition with hierarchical structure between clusters, that will be briefly called *hierarchical partition*, is specified by the set $C_H = \{C_k, (k = 1, \ldots, K), C_{K+1}, \ldots, C_{2K-1}\} = \{C, C_{K+1}, \ldots, C_{2K-1}\}$, formed by $2K - 1$ classes, where

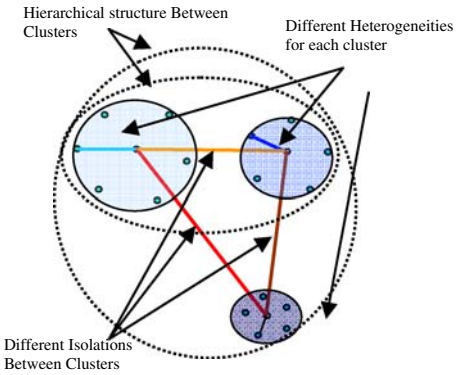


Fig. 7 Hierarchical partition with different heterogeneities within clusters and different isolations between clusters

the first $C_1, \ldots, C_k$ represent the subsets of a partition C with heterogeneities ${}_{\mathrm{W}}d_{kk}$ $(k = 1, \ldots, K)$ and the remaining $C_{K+1}, \ldots, C_{2K-1}$ are obtained by $K - 1$ pairwise possible amalgamations of subsets of C with isolations between classes ${}_{\mathrm{B}}d_{hk}(h, k = 1, \ldots, K; h \neq k)$, such that (6) is verified. Therefore, for each pair $C_h, C_m \in C_H \Rightarrow (C_m \cap C_h) \in \{C_m, C_h, \emptyset\}$, i.e., for each pair of classes, belonging to a hierarchical partition, either one is included in the other or they are disjoint. To define a hierarchical partition in terms of Eq. (4) it is necessary that:

(i) the two matrices $\mathbf{D}_{\mathrm{W}}$ and $\mathbf{D}_{\mathrm{B}}$ specify a well-structured partition in K clusters (i.e., satisfying (6));
(ii) matrix $\mathbf{D}_{\mathrm{B}}$ of order K is an ultrametric matrix (hence, it has at most $K - 1$ different off-diagonal values).

When conditions (i) and (ii) are satisfied the square matrix $\mathbf{Q}$ of order n is termed $(2K - 1)$-ultrametric matrix. It is formally defined as follows.

Definition 2 A $(2K - 1)$-*ultrametric matrix* is a square ultrametric matrix of order n, with off-diagonal elements that can assume one of at most $(2K - 1)$ different values: $0 <_{\mathrm{W}} d_{kk} \leq_{\mathrm{B}} d_{kh}$ $(k, h = 1, \ldots, K; h \neq k)$. In fact, formally: $\mathbf{Q} = [q_{il}]$, $q_{ii} = 0$, $q_{il} \geq 0$, $q_{il} = q_{li}$, $q_{il} \leq \max(q_{ik}, q_{lk})$ $\forall (i, l, k)$; furthermore $q_{il} \in \{0, {}_{\mathrm{W}}d_k, {}_{\mathrm{B}}d_{kh}\}$, with $0 <_{\mathrm{W}} d_{kk} \leq_{\mathrm{B}} d_{kh}$ $\forall (k, h : h \neq k)$.

Remark 4 The metric matrix $\mathbf{Q}$ satisfying (i) and (ii) is a $(2K - 1)$ *-ultrametric matrix.* Matrix $\mathbf{Q}$ includes the diagonal values of $\mathbf{D}_{\mathrm{W}}$ that, for the well-structured property (6), are the smallest K values identifying the first K levels of the ultrametric matrix $\mathbf{Q}$. Furthermore, matrix $\mathbf{D}_{\mathrm{B}}$ has $K - 1$ different values identifying the remaining levels of the ultrametric matrix. Therefore, $\mathbf{D}_{\mathrm{B}}$ controls the hierarchical relationship between classes of the partition.

Lemma 3 *A hierarchical partition in K classes* C_H *of O with heterogeneities* ${}_{\mathrm{W}}d_{kk}$ $(k = 1, \ldots, K)$ *and isolations* ${}_{\mathrm{B}}d_{kh}$ $(k, h = 1, \ldots, K,$ *with* $0 <_{\mathrm{W}} d_{kk} \leq_{\mathrm{B}} d_{kh})$ *is one-to-one associated to a* $(2K - 1)$*-ultrametric matrix* $\mathbf{Q}$ *with values* $\{0, {}_{\mathrm{W}}d_{kk}, {}_{\mathrm{B}}d_{kh}\}$.

Proof In a hierarchical partition C_H, every element o_i of O belongs to only one class because a partition C is included in C_H. Thus, any triplet (o_i, o_l, o_k) belongs, alternatively, to: (a) single set C_k of C in C_H; (b) two distinct sets; (c) three distinct sets

C_k, C_h, C_m of C in C_H. Cases (a), (b) and (c) determine only three types of distance triplets, respectively: $({}_{\mathrm{W}}d_{kk}, {}_{\mathrm{W}}d_{kk}, {}_{\mathrm{W}}d_{kk})$, $({}_{\mathrm{W}}d_{kk}, {}_{\mathrm{B}}d_{kh}, {}_{\mathrm{B}}d_{kh})$ and $({}_{\mathrm{B}}d_{kh}, {}_{\mathrm{B}}d_{km}, {}_{\mathrm{B}}d_{hm})$, all verifying the ultrametric inequality. In fact, the first triplet satisfies the ultrametric inequality because identifies an equilateral triangle. The second triplet is ultrametric because ${}_{\mathrm{W}}d_{kk} \leq_{\mathrm{B}} d_{kh}$, and the third triplet belongs to $\mathbf{D}_{\mathrm{B}}$ which is ultrametric by definition. Therefore, the $(2K-1)$-*ultrametric matrix* has K levels ${}_{\mathrm{W}}d_{kk}$ $(k = 1, \ldots, K)$ associated to the classes C_k of the partition C in C_H, while the remaining K-1 levels of the ultrametric matrix $\mathbf{D}_{\mathrm{B}}$ are associated to the classes $C_{K+1}, \ldots, C_{2K-1}$.

Conversely, each matrix $\mathbf{Q}$ has associated a hierarchical partition. In fact, $\mathbf{Q}$ specifies a covering, because $q_{il} = {}_{\mathrm{W}}d_{kk}$ is the distance between objects o_i and o_l when they belong to the same class C_k of C in C_H; while $q_{il} = {}_{\mathrm{B}}d_{kh}$ is the distance between objects o_i and o_l when they belong to different classes C_k and C_h of C in C_H. Moreover, since $\mathbf{Q}$ is an ultrametric matrix and no triplets of type $({}_{\mathrm{W}}d_{kk}, {}_{\mathrm{W}}d_{hh}, {}_{\mathrm{B}}d_{kh})$ can be observed, it follows that classes of the covering do not have elements in common and form a partition C in C_H. Furthermore, classes $C_{K+1}, \ldots, C_{2K-1}$ have a hierarchical structure which is specified by the ultrametric matrix $\mathbf{D}_{\mathrm{B}}$. □

Lemma 4 (Bijection between hierarchical partitions and $(2K-1)$-ultrametric matrices) *Let $\mathscr{C}_{\mathrm{H}}$ be the set of hierarchical partitions of O and let $\mathscr{D}_{PH}$ be the set of $(2K-1)$-ultrametric matrices $\mathbf{Q}$, then, there exists a bijection between $\mathscr{C}_H$ and $\mathscr{D}_{PH}$, i.e., each hierarchical partition has associated a matrix $\mathbf{Q}$ of the form (4) which satisfies* (i) *and* (ii)*, and vice versa.*

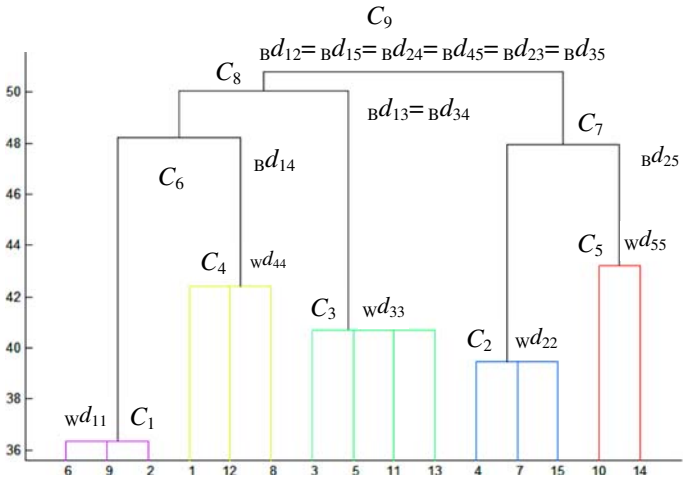
Proof Lemma 3 can be applied to any hierarchical partition $C_{\mathrm{H}} \in \mathscr{C}_{\mathrm{H}}$ with heterogeneity α_1 and isolation α_2 and classification matrix $\mathbf{Q} \in \mathscr{D}_{PH}$ with values $\{0, {}_{\mathrm{W}}d_k, {}_{\mathrm{B}}d_{kh}\}$. Thus it follows that each element of $\mathscr{C}_H$ has associated one element of $\mathscr{D}_{PH}$ and vice versa. □

Any $(2K-1)$-ultrametric matrix can be represented by a dendrogram with at most $(2K-1)$ levels (heights of the tree), briefly named $(2K-1)$-dendrogram (see Fig. 8 and Example 2).

Example 2 Given the partition $C = \{C_1 = \{1, 2, 4, 5, 6\}, C_2 = \{7, 9\}, C_3 = \{3, 8, 10\}\}$ of 10 objects in 3 classes with associated matrices $\mathbf{D}_{\mathrm{B}}$, $\mathbf{D}_{\mathrm{W}}$, and $\mathbf{M}$

$$\mathbf{D}_{\mathrm{B}} = \begin{array}{ccc} C_1 & C_2 & C_3 \\ 0 & 4 & 5 \\ & 0 & 5 \\ & & 0 \end{array} \quad \mathbf{D}_{\mathrm{W}} = \begin{array}{ccc} C_1 & C_2 & C_3 \\ 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{array} \quad \mathbf{M}' = \begin{array}{c|cccccccccc} & o_1 & o_2 & o_3 & o_4 & o_5 & o_6 & o_7 & o_8 & o_9 & o_{10} \\ C_1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ C_2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ C_3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array}.$$

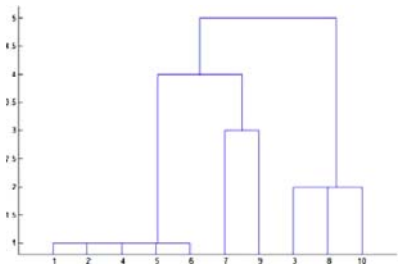
It can be observed that property (6) is satisfied; thus, C is a partition in 3 well-structured clusters. In fact, $4 = \min\{d_{kh} : h, k = 1, \ldots, 3\} > \max\{d_{kk} : k = 1, \ldots, 3\} = 3$. Furthermore, it can be verified that matrix $\mathbf{D}_{\mathrm{B}}$ is ultrametric, thus matrix $\mathbf{Q}$ is ultrametric. It has associated the following 5-dendrogram with 5 different nodes.



$$\mathbf{D}_B = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 & C_4 & C_5 \end{matrix} \\ \begin{bmatrix} 0 & {}_Bd_{12} & {}_Bd_{13} & {}_Bd_{14} & {}_Bd_{15} \\ {}_Bd_{12} & 0 & {}_Bd_{23} & {}_Bd_{24} & {}_Bd_{25} \\ {}_Bd_{13} & {}_Bd_{23} & 0 & {}_Bd_{34} & {}_Bd_{35} \\ {}_Bd_{14} & {}_Bd_{24} & {}_Bd_{34} & 0 & {}_Bd_{45} \\ {}_Bd_{15} & {}_Bd_{25} & {}_Bd_{35} & {}_Bd_{45} & 0 \end{bmatrix} \end{array} \quad \mathbf{D}_W = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 & C_4 & C_5 \end{matrix} \\ \begin{bmatrix} {}_Wd_{11} & 0 & 0 & 0 & 0 \\ 0 & {}_Wd_{22} & 0 & 0 & 0 \\ 0 & 0 & {}_Wd_{33} & 0 & 0 \\ 0 & 0 & 0 & {}_Wd_{44} & 0 \\ 0 & 0 & 0 & 0 & {}_Wd_{55} \end{bmatrix} \end{array}$$

Fig. 8 Representation of a $(2K-1)$-dendrogram when $K=5$, together with matrices $\mathbf{D}_W$ and $\mathbf{D}_B$. A 9-dendrogram is shown; the first five clusters $(C_1,\ldots,C_5)$ form a partition C; clusters $C_6=\{C_1,C_4\}$, $C_7=\{C_2,C_5\}$, $C_8=\{C_6,C_3\}$, $C_9=\{C_8,C_7\}$, specify the hierarchical structure of the partition

$$\mathbf{Q} = \begin{array}{c|cccccccccc} & o_1 & o_2 & o_3 & o_4 & o_5 & o_6 & o_7 & o_8 & o_9 & o_{10} \\ \hline o_1 & 0 & 1 & 5 & 1 & 1 & 1 & 4 & 5 & 4 & 5 \\ o_2 & 1 & 0 & 5 & 1 & 1 & 1 & 4 & 5 & 4 & 5 \\ o_3 & 5 & 5 & 0 & 5 & 5 & 5 & 5 & 2 & 5 & 2 \\ o_4 & 1 & 1 & 5 & 0 & 1 & 1 & 4 & 5 & 4 & 5 \\ o_5 & 1 & 1 & 5 & 1 & 0 & 1 & 4 & 5 & 4 & 5 \\ o_6 & 1 & 1 & 5 & 1 & 1 & 0 & 4 & 5 & 4 & 5 \\ o_7 & 4 & 4 & 5 & 4 & 4 & 4 & 0 & 5 & 3 & 5 \\ o_8 & 5 & 5 & 2 & 5 & 5 & 5 & 5 & 0 & 5 & 2 \\ o_9 & 4 & 4 & 5 & 4 & 4 & 4 & 3 & 5 & 0 & 5 \\ o_{10} & 5 & 5 & 2 & 5 & 5 & 5 & 5 & 2 & 5 & 0 \end{array}$$



Now it is necessary to estimate the unknown parameters of the three classification matrices introduced in Sects. 3.1, 3.2, and 3.3. A least-squares estimation will be adopted.

4 Least-squares estimation of the partitioning models and Algorithms

4.1 LS estimation of the well-structured perfect partition

The Least-Squares (LS) estimation of model (1), when the classification matrix is the 2-ultrametric matrix (2) identifying a well-structured perfect partition, is defined to be

the mathematical program (indicated by (P1)) of finding $\mathbf{P}$ minimizing the following quadratic constrained problem with respect to matrix $\mathbf{S}$, the heterogeneity α_1 and isolation α_2

$$F(\mathbf{S}, \alpha_1, \alpha_2) = ||\mathbf{D} - \mathbf{P}||^2 = ||\mathbf{D} - \alpha_2(\mathbf{1}_n\mathbf{1}_n' - \mathbf{S}) - \alpha_1(\mathbf{S} - \mathbf{I}_n)||^2 \rightarrow \min_{\mathbf{S},\alpha_1,\alpha_2}$$

subject to (P1)

$\mathbf{P}$ is 2-ultrametric matrix as in (2) with $0 < \alpha_1 \leq \alpha_2$ and

$\mathbf{S}$ is a joint membership matrix

(with $s_{ii} = 1$; $s_{il} = 1$ (resp., 0) if $o_i, o_l \in C_h$ (resp., $o_i, o_l \notin C_h$)$\forall i, l$).

Remark 5 If in Eq. (2), $\alpha_2 = 1$ and $\alpha_1 = 0$, then matrix $\mathbf{P}$ becomes: $\mathbf{P} = \mathbf{1}_n\mathbf{1}_n' - \mathbf{S}$. In this case, it is interesting to see that problem (P1) is equivalent to the linear 0/1-integer programming *clique-partitioning* problem (Régnier 1965),

$$\text{Min} \sum_{i=1}^{n-1} \sum_{l=i+1}^{n} b_{il} s_{il}$$

subject to (P1′)

$$s_{il} \in \{0, 1\}\ \forall (i, l);\ s_{ik} + s_{lk} - s_{il} \leq 1,\ s_{il} + s_{lk} - s_{ik} \leq 1,\ s_{ik} + s_{il} - s_{lk} \leq 1,\ \forall (i, l, k);$$

where $b_{il} = 2d_{il} - 1$. The equivalence is a consequence of the following remarks (Régnier 1965; Marcotorchino and Michaud 1982) due to $p_{il} = 1 - s_{il}$ and

(i) $$\sum_{1 \leq i < l \leq n} (d_{il} - p_{il})^2 = \sum_{1 \leq i < l \leq n} (d_{il} - 1)^2 + \sum_{1 \leq i < l \leq n} s_{il}\,(2d_{il} - 1)$$

$$= \text{const} + \sum_{1 \leq i < l \leq n} b_{il} s_{il}.$$

(ii) the $O(n^3)$ inequality constraints on $\mathbf{S}$ specify a joint membership matrix identifying a partition C of O.

If integer variables s_{ij} are replaced by continuous variables: $0 \leq s_{il} \leq 1, (1 \leq i < l \leq n)$ the relaxation of (P1′) is a linear programming problem whose solution, when all $s_{il} \in \{0, 1\}$, is also the solution of (P1′). In practice, it appears that the relaxation of (P1′) has often, but not invariably, a 0/1 solution (Grötschel and Wakabayashi 1989). There can be more than one optimal solution and different solutions can be identified by solving a series of linear programming problems in which different small random quantities are added to the right hand sides of inequalities in (P1′). If the solution that is obtained is not integral, more elaborate algorithms are required to provide heuristic solutions to the clique-partitioning problem; several algorithms of this type are reviewed by Hansen et al. (1994).

Including the second formula in (2) in problem (P1) and rewriting constraints on the membership matrix $\mathbf{M}$, problem (P1) needs to be minimized, with respect to α_1, α_2 and $\mathbf{M}$

$$F_1(\alpha_1, \alpha_2, \mathbf{M}) = ||\mathbf{D} - \alpha_2(\mathbf{1}_n\mathbf{1}_n' - \mathbf{M}\mathbf{M}') - \alpha_1(\mathbf{M}\mathbf{M}' - \mathbf{I}_n)||^2 \to \min_{\mathbf{M},\alpha_1,\alpha_2} \tag{9}$$

$$\text{subject to} \tag{P2}$$

$$0 < \alpha_1 \leq \alpha_2; \tag{10}$$

$$m_{ik} \in \{0, 1\} \quad (i = 1, \ldots, n;\ k = 1, \ldots, K), \tag{11}$$

$$\sum_{p=1}^{P} m_{ip} = 1 \quad (i = 1, \ldots, n). \tag{12}$$

Problem (P2) can be solved by considering a coordinate descent algorithm (see Zangwill 1969) that will be described in Sect. 4.5, which alternates between the updates of α_1, α_2 and $\mathbf{M}$ while decreasing the loss function (9). In this algorithm, to update α_1 and α_2, let us first rewrite function $F_1(\alpha_1, \alpha_2, \mathbf{M})$ as

$$\begin{aligned} F_1(\alpha_1, \alpha_2, \mathbf{M}) = tr(\mathbf{D}\mathbf{D}) + \alpha_2^2 tr(\mathbf{1}_n\mathbf{1}_n' - \mathbf{M}\mathbf{M}')^2 + \alpha_1^2 tr(\mathbf{M}\mathbf{M}' - \mathbf{I}_n)^2 \\ -2\alpha_2 tr(\mathbf{D}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{M}\mathbf{M}')) - 2\alpha_1 tr(\mathbf{D}(\mathbf{M}\mathbf{M}' - \mathbf{I}_n)), \end{aligned}$$

because it can be easily shown that $tr((\mathbf{1}_n\mathbf{1}_n' - \mathbf{M}\mathbf{M}')(\mathbf{M}\mathbf{M}' - \mathbf{I}_n)) = 0$.

4.1.1 Estimation of α_1 and α_2

By differentiating $F_1(\alpha_1, \alpha_2, \mathbf{M})$ with respect to α_1 and α_2, for a fixed $\hat{\mathbf{M}}$, we derive the two normal equations,

$$\begin{aligned} \partial F_1(\alpha_1, \alpha_2, \hat{\mathbf{M}})/\partial\alpha_1 &= \alpha_1 tr(\hat{\mathbf{M}}\hat{M}' - \mathbf{I}_n)^2 - tr(\mathbf{D}(\hat{\mathbf{M}}\hat{M}' - \mathbf{I}_n)) = 0 \\ \partial F_1(\alpha_1, \alpha_2, \hat{\mathbf{M}})/\partial\alpha_2 &= \alpha_2 tr(\mathbf{1}_n\mathbf{1}_n' - \hat{\mathbf{M}}\hat{\mathbf{M}}')^2 - tr(\mathbf{D}(\mathbf{1}_n\mathbf{1}_n' - \hat{\mathbf{M}}\hat{\mathbf{M}}')) = 0, \end{aligned} \tag{13}$$

which can be solved with respect to α_1 and α_2,

$$\alpha_1(\hat{\mathbf{M}}) = tr(\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}})/tr(\hat{\mathbf{M}}\hat{M}' - \mathbf{I}_n)^2 = \frac{2\sum_{k=1}^{K}\sum_{i,l\in C_k, i<l} d_{ij}}{\sum_{k=1}^{K} n_k^2 - n}, \tag{14}$$

$$\begin{aligned} \alpha_2(\hat{\mathbf{M}}) &= (\text{tr}(\mathbf{1}_n'\mathbf{D}\mathbf{1}_n) - tr(\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}))/\text{tr}(\mathbf{1}_n\mathbf{1}_{n'} - \hat{\mathbf{M}}\hat{M}')^2 \\ &= \frac{2\sum_{k=1}^{K-1}\sum_{h=k+1}^{K}\sum_{i\in C_k, l\in C_h, i<l} d_{il}}{n^2 - \sum_{k=1}^{K} n_k^2}, \end{aligned} \tag{15}$$

where n_k denotes the cardinality of C_k.

Such values minimize $F_1(\alpha_1, \alpha_2, \hat{M})$ without taking into account the constraint $0 < \alpha_1 \leq \alpha_2$ which is automatically satisfied (see Remark 6) if we start from a feasible solution (i.e., an initial (α_1, α_2) such that $0 < \alpha_1 \leq \alpha_2$).

It is important to note that the estimator of the heterogeneity α_1 is the arithmetic mean of the within class dissimilarities of the partition C, while the estimator of the isolation α_2 is the arithmetic mean of the between class dissimilarities of C. In fact, function (9) can be written as

$$F_1(\alpha_1, \alpha_2, \mathbf{M}) = \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i \neq l}}^{n} (d_{il} - \alpha_1)^2 \sum_{k=1}^{K} m_{ik} m_{lk} + \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i \neq l}}^{n} (d_{il} - \alpha_2)^2 \left(1 - \sum_{k=1}^{K} m_{ik} m_{lk} \right), \tag{16}$$

where the first part in the right hand side is the deviance of the dissimilarities of the objects belonging to a class of the partition C, while the second part is the deviance of the dissimilarities between objects belonging to different classes of the partition C.

4.1.2 Estimation of M

When $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are fixed, function (16) can be minimized row by row for each $\mathbf{m}_i$ of $\mathbf{M}$ when all the remaining rows are fixed, i.e. when $\mathbf{M} = [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i, \ldots, \hat{\mathbf{m}}_n]'$.

Thus, the ith unit belongs to the kth class, i.e., $m_{ik} = 1$, if function (16) reaches the minimum with respect to the assignment of the ith unit to any other vth class $v = 1, \ldots, K (v \neq k)$. Otherwise $m_{ik} = 0$. In formulas: for each row $\mathbf{m}_i$ of $\mathbf{M}$, $i = 1, \ldots, n$,

$$\begin{aligned} m_{ik} &= 1, \quad \text{if } F_1(\hat{\alpha}_1, \hat{\alpha}_2, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]') \\ &= \min\{F_1(\hat{\alpha}_1, \hat{\alpha}_2, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_v, \ldots, \hat{\mathbf{m}}_n]') : v = 1, \ldots, K (v \neq k)\} \\ m_{ik} &= 0 \quad \text{otherwise.} \end{aligned}$$

where $\mathbf{i}_v$ is the vth row of the identity matrix of order K.

4.2 LS estimation of the well-structured partition

The Least-Squares estimation of the parameters of the error model (1) when the classification matrix $\mathbf{Q}$ has the form (4) associated to a well-structured partition, defines an optimization quadratic problem minimizing the following function with respect to matrices $\mathbf{M}$, $\mathbf{D}_\mathrm{B}$ and $\mathbf{D}_\mathrm{W}$

$$F_2(\mathbf{M}, \mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{W}}) = ||\mathbf{D} - \mathbf{M}\mathbf{D}_{\mathrm{B}}\mathbf{M}' - \mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}' + \mathrm{diag}(\mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}')||^2$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{\substack{l=1 \\ i\neq l}}^{n}(d_{il} - {}_{\mathrm{W}}d_{kk})^2 m_{ik}m_{lk}$$
$$+ \sum_{h=1}^{K}\sum_{\substack{k=1 \\ k\neq h}}^{K}\sum_{i=1}^{n}\sum_{\substack{l=1 \\ i\neq l}}^{n}(d_{il} - {}_{\mathrm{B}}d_{hk})^2 m_{ik}m_{lh} \rightarrow \min_{\mathbf{M},\mathbf{D}_{\mathrm{B}},\mathbf{D}_{\mathrm{W}}} \quad (17)$$

subject to (P2)

$\mathbf{M}$ binary and row stochastic, i.e., with constraints (11) and (12), (18)

well-structured partition, i.e., $\max\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots, K\}$
$\leq \min\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (h \neq k)\}$, (19)

As noted before only constraints (18) are necessary to specify a partition of the objects and they are the only ones really needed. The constraint (19) can be omitted in the case the researcher does not necessarily require a partition with the well-structured property (6), as it is common practice with cluster analysis methodologies.

4.2.1 Estimation of $\mathbf{D}_{\mathrm{w}}$

By differentiating (17) with respect to ${}_{\mathrm{W}}d_{kk}(k = 1, \ldots, K)$ for a fixed $\hat{\mathbf{M}}$, and equating to zero, the estimators are

$${}_{\mathrm{W}}d_{kk} = \frac{\sum_{i=1}^{n}\sum_{l=1,i\neq l}^{n} d_{il}\hat{m}_{ik}\hat{m}_{lk}}{\sum_{i=1}^{n}\sum_{l=1,i\neq l}^{n}\hat{m}_{ik}\hat{m}_{lk}} = \frac{2\sum_{i,l\in C_k, i<l} d_{ij}}{n_k^2 - n_k}, \quad (k = 1, \ldots, K). \quad (20)$$

In fact, when $\hat{\mathbf{D}}_{\mathrm{B}}$ and $\hat{\mathbf{M}}$ are fixed, the residual matrix $\ddot{\mathbf{D}} = \mathbf{D} - \hat{\mathbf{M}}\,\hat{\mathbf{D}}_{\mathrm{B}}\,\hat{\mathbf{M}}'$ is known and function (17) can be written as

$$F_2(\hat{\mathbf{M}}, \hat{\mathbf{D}}_{\mathrm{B}}, \mathbf{D}_{\mathrm{W}}) = ||\ \ddot{\mathbf{D}} - \hat{\mathbf{M}}\mathbf{D}_{\mathrm{W}}\hat{\mathbf{M}}' + \mathrm{diag}(\hat{\mathbf{M}}\mathbf{D}_{\mathrm{W}}\hat{\mathbf{M}}')||^2, \quad (21)$$

which is minimized by

$$\mathbf{D}_{\mathrm{W}} = \mathrm{diag}(\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}})[(\mathbf{M}'\hat{\mathbf{M}})^2 - \mathbf{M}'\hat{\mathbf{M}})]^+, \quad (22)$$

where $\mathbf{A}^+$ denotes the Moore-Penrose inverse of matrix $\mathbf{A}$.

4.2.2 Estimation of $\mathbf{D}_{\mathrm{B}}$

By differentiating (17) with respect to ${}_{\mathrm{B}}d_{kh}$ $(k, h = 1, \ldots, K)$ for a fixed $\hat{\mathbf{M}}$ and equating to zero, the estimators are

$$ {}_{\mathrm{B}}d_{kh} = \frac{\sum_{i=1}^{n}\sum_{l=1, i\neq l}^{n} d_{il}\hat{m}_{ik}\hat{m}_{lh}}{\sum_{i=1}^{n}\sum_{l=1}^{n}\hat{m}_{ik}\hat{m}_{lh}} = \frac{2\sum_{i\in C_k, l\in C_h, i<l} d_{il}}{n_k n_h}, \quad (k, h = 1, \ldots, K). \tag{23} $$

In fact, when $\hat{\mathbf{D}}_{\mathrm{W}}$ and $\hat{\mathbf{M}}$ are fixed, matrix $\tilde{\mathbf{D}} = \mathbf{D} - \hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}' + \mathrm{diag}(\hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}')$ is known and function (17) can be written as

$$ F_2(\hat{\mathbf{M}}, \mathbf{D}_{\mathrm{B}}, \hat{\mathbf{D}}_{\mathrm{W}}) = ||\tilde{\mathbf{D}} - \hat{\mathbf{M}}\mathbf{D}_{\mathrm{B}}\hat{\mathbf{M}}'||^2. \tag{24} $$

The minimization of (24) is a Penrose multivariate regression problem with solution

$$ \begin{aligned} \mathbf{D}_{\mathrm{B}} &= (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\tilde{\mathbf{D}}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1} \\ &= (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\mathbf{M}'[\mathbf{D} - \hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\mathbf{M}' + \mathrm{diag}(\hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\mathbf{M}')]\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1} \\ &= (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1} - \mathrm{diag}((\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}), \end{aligned} \tag{25} $$

which corresponds, element by element, to the solution (23).

4.2.3 Estimation of M

When $\hat{\mathbf{D}}_{\mathrm{W}}$ and $\hat{\mathbf{D}}_{\mathrm{B}}$ are fixed, function (17) can be minimized row by row for each $\mathbf{m}_i$ of $\mathbf{M}$, when all remaining rows are fixed, i.e., when: $\mathbf{M} = [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i, \ldots, \hat{\mathbf{m}}_n]'$. Thus, the ith unit belongs to the kth class, i.e., m_{ik}=1, if function (17) reaches its minimum with respect to the assignment of the ith unit to any other vth class $v = 1, \ldots, K (v \neq k)$. Otherwise $m_{ik} = 0$. In formulas, for each row $\mathbf{m}_i$ of $\mathbf{M}$, $i = 1, \ldots, n$,

$$ \begin{aligned} &m_{ik} = 1, \text{ if } F_2(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]') \\ &\quad = \min\{F_2(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_v, \ldots, \hat{\mathbf{m}}_n]') : v = 1, \ldots, K (v \neq k)\} \\ &m_{ik} = 0 \text{ otherwise.} \end{aligned} $$

where $\mathbf{i}_v$ is the vth row of the identity matrix of order K.

4.3 The square K-means for dissimilarity data

The Least-Squares estimation of the parameters of the error model (1) when the matrix used for modelling the partition is $\mathbf{R}$, as given in (5), specifies the following quadratic

constrained problem to be minimized with respect to matrices $\mathbf{M}$, $\mathbf{D}_{\mathrm{B}}$ and $\mathbf{D}_{\mathrm{W}}$

$$\begin{aligned} F_3(\mathbf{M}, \mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{w}}) &= ||\mathbf{D} - \mathbf{M}(\mathbf{D}_{\mathrm{B}} + \mathbf{D}_{\mathrm{W}})\mathbf{M}'||^2 \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{l=1}^{n}(d_{il} - {}_{\mathbf{W}}d_{kk})^2 m_{ik}m_{lk} \\ &\quad + \sum_{h=1}^{K}\sum_{\substack{k=1 \\ k\neq h}}^{K}\sum_{i=1}^{n}\sum_{l=1}^{n}(d_{il} - {}_{\mathbf{B}}d_{hk})^2 m_{ik}m_{lh} \rightarrow \min_{\mathbf{M},\mathbf{D}_{\mathrm{B}},\mathbf{D}_{\mathrm{W}}} \end{aligned} \tag{26}$$

subject to (P2*)

$\mathbf{M}$ binary and row stochastic, i.e., constraints (11) and(12), (27)

well-structured partition,

i.e., $\max\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots, K\} \leq \min\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (h \neq k)\}$. (28)

Problem (P2*) is quite similar to problem (P2) with the distinction that the difference between the main diagonal entries of matrices $\mathbf{D}$ and $\mathbf{R}$ is estimated. Again, constraint (28) is imposed only if we are interested to obtain a well-structured partition.

4.3.1 Estimation of $\mathbf{D}_{\mathrm{B}} + \mathbf{D}_{\mathrm{W}}$

When $\hat{\mathbf{M}}$ is fixed the minimization of (26) is again a Penrose multivariate regression problem with solution

$$\mathbf{D}_{\mathrm{B}} + \mathbf{D}_{\mathrm{W}} = (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}. \tag{29}$$

By substituting (29) into (26) it remains to minimize,

$$F_3(\mathbf{M}) = ||\mathbf{D} - \hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'||^2 = ||\mathbf{D} - \mathbf{H}_{\mathrm{M}}\mathbf{D}\mathbf{H}_{\mathrm{M}}||^2, \tag{30}$$

where $\mathbf{H}_{\mathrm{M}} = \hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'$ is the $n \times n$ idempotent orthogonal projection matrix on the subspace spanned by the columns of $\hat{\mathbf{M}}$.

4.3.2 Estimation of M

The estimation of $\mathbf{M}$ can be achieved by a similar procedure as defined in Sect. 4.2.1 by using function (30). Thus, for each row $\mathbf{m}_i$ of $\mathbf{M}$, $i = 1, \ldots, n$,

$$\begin{aligned} m_{ik} &= 1, \text{if} \quad F_3(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]') \\ &= \min\{F_3(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_v, \ldots, \hat{\mathbf{m}}_n]') : v = 1, \ldots, K(v \neq k)\} \\ m_{ik} &= 0 \quad \text{otherwise.} \end{aligned}$$

where $\mathbf{i}_v$ is the vth row of the identity matrix of order K.

The algorithm shown in this section is named *square K-means* because it resembles the well-known technique for partitioning data (Ball and Hall 1967; MacQueen 1967) as explained below. The term "*square*" is used to distinguish our method, applied on a square dissimilarity matrix from the usual "*rectangular*" K-means applied on rectangular data matrices (units $\times$ variables).

In classical MDS (Torgerson 1958; Gower 1966), dissimilarities (squared euclidean) are converted into scalar products, by the transformation $\mathbf{S} = -1/2\mathbf{JDJ}$, where matrix $\mathbf{J} = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n'$ is the $n \times n$ centering operator. The scalar products, as given by $\mathbf{XX}'$, are reconstructed in a LS sense by J dimensions by using the $n \times J$ ($J \leq n-1$) matrix of objects coordinates $\mathbf{X} = \mathbf{VL}^{1/2}$, where the columns of the $n \times J$ matrix $\mathbf{V}$ are the orthonormal eigenvectors corresponding to the J largest positive eigenvalues which themselves are the diagonal entries of the $J \times J$ diagonal matrix $\mathbf{L}$. When $\mathbf{S}$ is not positive definite, the optimal J-dimensional LS reconstruction is obtained by replacing the negative eigenvalues with zeros (Keller 1962; Mathar 1985).

Thus, if $\mathbf{D}$ is a square Euclidean distance matrix then it can be converted into the scalar product matrix $\mathbf{XX}'$ or if $\mathbf{X}$ is directly observed then function (30) and the problem (P2*) can be rewritten

$$F_4(\mathbf{M}) = ||\mathbf{XX}' - \mathbf{H}_\mathbf{M}\mathbf{XX}'\mathbf{H}_\mathbf{M}||^2 = ||\mathbf{XX}' - \mathbf{M}\bar{\mathbf{X}}\bar{\mathbf{X}}'\mathbf{M}'||^2 \rightarrow \min_{\mathbf{M}} \tag{31}$$

subject to (P2**)

$\mathbf{M}$ binary and row stochastic, i.e., constraints (11) and(12), (32)

which represents the loss function of the problem called here *square k-means for scalar product matrices*, where $\bar{\mathbf{X}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}$ is the centroid matrix also obtained in the usual K-means algorithm for data matrices $\mathbf{X}$.

In fact, the K-means algorithm (Ball and Hall 1967; MacQueen 1967) has associated the error model

$$\mathbf{X} = \mathbf{M}\bar{\mathbf{X}} + \mathbf{E}_1, \tag{33}$$

where $\mathbf{E}_1$ is the matrix of error terms of dimension $n \times J$. The LS solution of the K-means model is obtained by minimizing the loss function $\text{tr}(\mathbf{E}_1'\mathbf{E}_1) = ||\mathbf{X} - \mathbf{M}\bar{\mathbf{X}}||^2$, subject to binary and row stochastic constraints on $\mathbf{M}$.

In terms of scalar products, the direct modeling formulation of the squared k-means is

$$\mathbf{XX}' = \mathbf{M}\bar{\mathbf{X}}\bar{\mathbf{X}}'\mathbf{M}' + \mathbf{E}_2, \tag{34}$$

where $\mathbf{E}_2$ is a square matrix of error terms of dimension n. The LS solution of the square k-means is given by minimizing problem (P2**) with loss function (31).

Thus, problem (P2*) can be considered *the squared k-means for dissimilarity matrices*, with error model: $\mathbf{D} = \mathbf{R} + \mathbf{E}_3$.

4.4 LS estimation of the hierarchical partition model—square hierarchical K-means

Here the same problem of the previous section has to be solved with the additional constraints that matrix $\mathbf{D}_B$ has to be ultrametric. Formally, it is necessary to define the following quadratic constrained problem to be minimized with respect to matrices $\mathbf{M}$, $\mathbf{D}_{\mathrm{B}}$ and $\mathbf{D}_{\mathrm{W}}$

$$F_2(\mathbf{M}, \mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{w}}) = ||\mathbf{D} - \mathbf{M}\mathbf{D}_{\mathrm{B}}\mathbf{M}' - \mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}' + \mathrm{diag}(\mathbf{M}\mathbf{D}_{\mathrm{W}}\mathbf{M}')||^2 \tag{35}$$

subject to (P3)

$\mathbf{M}$ binary and row stochastic, i.e., constraints (11) and (12), (36)

well-structured partition, i.e., $\max\{{}_{\mathrm{W}}d_{kk} : k = 1, \ldots, K\}$
$\leq \min\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (h \neq k)\}$, (37)

$\mathbf{D}_{\mathrm{B}}$ ultrametric. (38)

The ultrametricity constraint of matrix (38) actually implies the $O(K^3)$ constraints on the triplets of $\mathbf{D}_{\mathrm{B}}$,

$$\begin{aligned} {}_{\mathrm{B}}d_{kh} &\leq \max({}_{\mathrm{B}}d_{kq}, {}_{\mathrm{B}}d_{hq}), \\ {}_{\mathrm{B}}d_{hq} &\leq \max({}_{\mathrm{B}}d_{kh}, {}_{\mathrm{B}}d_{kq}), \quad k = 1, \ldots, K, \quad h = k, \ldots, K, \quad q = h, \ldots, K; \\ {}_{\mathrm{B}}d_{kq} &\leq \max({}_{\mathrm{B}}d_{kh}, {}_{\mathrm{B}}d_{hq}). \end{aligned} \tag{39}$$

Such constraints can be synthesized into a single (Carroll and Pruzansky 1980) one by requiring

$$\sum_{(k,h,q)\in\Gamma(\mathbf{D}_{\mathrm{B}})} ({}_{\mathrm{B}}d_{kq} - {}_{\mathrm{B}}d_{hq})^2 = 0, \tag{40}$$

where $\Gamma(\mathbf{D}_{\mathrm{B}}) = \{(k, h, q) : 1 \leq k, h, q \leq K, k \leq h \leq q : {}_{\mathrm{B}}d_{kh} \leq \min(u_{kq}, u_{hq})\}$, since for each triplet the largest two values must be equal and their squared difference null.

4.4.1 Estimation of D_{B}

When $\hat{\mathbf{D}}_{\mathrm{W}}$ and $\hat{\mathbf{M}}$ are fixed, matrix $\tilde{\mathbf{D}} = \mathbf{D} - \hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}' + \mathrm{diag}\,(\hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}')$ is known and function (35) can be written as

$$F_2(\mathbf{D}_{\mathrm{B}}) = ||\tilde{\mathbf{D}} - \hat{\mathbf{M}}\mathbf{D}_{\mathrm{B}}\hat{\mathbf{M}}'||^2 = ||(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\tilde{\mathbf{D}} \quad \hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1} - \mathbf{D}_{\mathrm{B}}||^2. \tag{41}$$

subject to (P3a)

$\mathbf{D}_{\mathrm{B}}$ ultrametric. (42)

This corresponds to find the closest, in the LS sense, ultrametric matrix to the true $\mathbf{D}_{\mathrm{B}}$ matrix.

The objective function (41) subject to ultrametric constraints (39) has been reformulated by Carroll and Pruzansky (1980) as an unconstrained problem by adding to

(41) the penalty function (40) multiplied by a penalty parameter. Then, a sequence of parameterized unconstrained optimizations, for increasing values of the penalty parameter, by using originally a gradient-based procedure, or the conjugate gradient (De Soete 1984) or later the truncated-Newton or the quasi-Newton (Vichi 1993) is solved. Vichi (1994) solved (P3a) more efficiently by rewriting constraints (39) and permuting indices (h, k, q) so as:

$$\begin{aligned} &{}_{\mathrm{B}}d_{hq} - {}_{\mathrm{B}}d_{kq} = 0, \\ &\text{for } {}_{\mathrm{B}}d_{hk} \leq \min({}_{\mathrm{B}}d_{hq}, {}_{\mathrm{B}}d_{kq}) h = 1, \ldots, K-2; \quad k = h+1, \ldots, K-1; \\ &q = k+1, \ldots, K, \end{aligned} \tag{43}$$

and using a sequential quadratic programming algorithm (SQP) (Powell 1983), or grouping the constraints into more than one quadratic as in (40) (Vichi 1996). This last approach allows for bounding correctly the feasible region and consequently reducing the computational complexity as found in several simulated and observed data. A reduction of time and space complexity for solving (P3a) can be achieved by recognizing matrix $\mathbf{D}_{\mathrm{B}}$ as ultrametric with $O(K^2 \log K)$ time complexity instead of $O(K^3)$. This is accomplished via the recognition procedure suggested by Bandelt (1990), for sequential algorithms and by Dahlhaus (1993) for parallel algorithms: sorting ${}_{\mathrm{B}}d_{hk}$ in $O(K^2 \log K)$ with heap-sort; taking the minimum ${}_{\mathrm{B}}d_{hk}$ of the ordered list; verifying if ${}_{\mathrm{B}}d_{hq} - {}_{\mathrm{B}}d_{kq} = 0$, for all $k \in \{1, \ldots, K\} - \{h, k\}$; deleting h; and proceeding in the ordered list with the smallest ${}_{\mathrm{B}}d_{kq}k$, $q \neq h$, until only two objects are left. Already for small data sets this procedure reduces significantly the time complexity as observed in our experiments.

When such heuristic algorithms are not at hand, a solution to problem (P3a) can be achieved by applying the hierarchical group average link clustering (UPGMA) on matrix

$$(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'\tilde{\mathbf{D}}\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}.$$

Since (P3a) has a closed non-convex feasible region (ultrametric cone) (Critchley and Fichet 1994) and it is known to be an NP-hard classification problem (Krivánek and Morávek 1986), its global minimum solution cannot be always guaranteed and the convergent sequence of ALS can be broken by a local minimum for (P3a). This problem is overcome by retaining for (P3a) only solutions where the objective function does not increase.

After describing the LS estimation of the parameters of the models proposed in Sect. 3 we are now in position to describe the steps of the necessary coordinate descent algorithms.

4.5 Algorithm for the well-structured perfect partition

The Least-Squares estimates of the loss function $F(\alpha_1, \alpha_2, \mathbf{M})$ are computed by a coordinate descent algorithm also known as alternating least-squares algorithm, where

parameters α_1, α_2 and matrix $\mathbf{M}$ are updated in turn by minimizing the loss function $F(\alpha_1, \alpha_2, \mathbf{M})$, conditionally upon the other fixed parameters. At each updating $F(\alpha_1, \alpha_2, \mathbf{M})$ does not increase and generally decreases, generating a sequence of solutions monotonically converging to a stationary point, which in the applications is usually at least a local minimum of the problem.

The ALS algorithm can be described by two basic steps (steps 1 and 2) that are sequentially repeated after an initialisation step and until a stopping rule is satisfied.

Step 0: Initialisation: a small non negative arbitrary convergence tolerance value (threshold) ε is chosen and an initial feasible partition $\mathbf{M}$ is given, i.e., such that $\alpha_1(\mathbf{M}) \leq \alpha_2(\mathbf{M})$. Note that initializing $\alpha_1(\mathbf{M}) \leq \alpha_2(\mathbf{M})$ guarantees to obtain a final solution with $\alpha_1(\mathbf{M}) \leq \alpha_2(\mathbf{M})$ as shown in Remark 6 below.

Step 1: Updating α_1 and α_2 for fixed $\hat{\mathbf{M}}$.

For the current matrix $\hat{\mathbf{M}}$, the update of α_1 and α_2, as described in Sect. 4, is given by

$$\alpha_1 = \mathrm{tr}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}}/\mathrm{tr}(\hat{\mathbf{M}}\hat{M}' - \mathbf{I})^2$$

$$\alpha_2 = (\mathrm{tr}\mathbf{1}'\mathbf{D}\mathbf{1} - \mathrm{tr}\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}})/\mathrm{tr}(\mathbf{1}\mathbf{1}' - \hat{\mathbf{M}}\hat{M}')^2.$$

Step 2: Updating $\mathbf{M}$ for fixed $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

For the current values $\hat{\alpha}_1$ and $\hat{\alpha}_2$, the updates of $\mathbf{M}$ is given by solving an assignment problem where, for the ith row of $\mathbf{M}$

$$m_{ik} = 1, \quad \text{if } F_1(\hat{\alpha}_1, \hat{\alpha}_2, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_\mathrm{i} = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]')$$

$$= \min\{F_1(\hat{\alpha}_1, \hat{\alpha}_2, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_v, \ldots, \hat{\mathbf{m}}_n]')\} : v = 1, \ldots, K(v \neq k)\}$$

$$m_{ik} = 0 \quad \text{otherwise.}$$

Therefore, for each row of $\mathbf{M}$ the unique element equal to 1 is set in the column that minimizes the loss function $F(\alpha_1, \alpha_2, \mathbf{M})$, for a fixed $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Passing to the following row, the loss function never increases and generally decreases.

Stopping rule. The function value $F(\alpha_1, \alpha_2, \mathbf{M})$ is computed for the current values of $\hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{\mathbf{M}}$. When such updated values have decreased considerably the function value (more than the threshold ε), α_1, α_2 and $\mathbf{M}$ are updated further according to Steps 1 and 2. Otherwise, the process is stopped. Since $\hat{\alpha}_1$ and $\hat{\alpha}_2$ can be directly included in $F(\alpha_1, \alpha_2, \mathbf{M})$ it remains to minimize F with respect to matrix $\mathbf{M}$, which is a discrete optimization problem. Thus the final solution is obtained when no changes are observed from one step to the following or when a maximum number of iterations has been reached.

Since the algorithm does not necessarily attain the global optimal solution of the optimization problem, it is advisable to start the algorithm from different random configurations and retain the best solution in terms of minimum function value among the different runs.

Remark 6 If the algorithm is started from a feasible solution (i.e., $\alpha_1 \leq \alpha_2$), the algorithm will produce new values α_1, α_2, such that $0 < \alpha_1 \leq \alpha_2$, according to formulas (14) and (15). This can be proved if we consider that the total deviance $T = ||\mathbf{D} - \bar{d}\mathbf{1}_n\mathbf{1}_n'||^2$ can be decomposed as

$$\begin{aligned} \mathrm{T} &= ||\mathbf{D} - \bar{d}\mathbf{1}_n\mathbf{1}_n'||^2 \\ &= F_1(\alpha_1, \alpha_2, \hat{\mathbf{M}}) + ||(\bar{d} - \alpha_1)(\hat{\mathbf{M}}\hat{\mathbf{M}}' - \mathbf{I})||^2 \\ &\quad + ||(\alpha_2 - \bar{d})(\mathbf{1}_n\mathbf{1}_n' - \hat{\mathbf{M}}\hat{\mathbf{M}}')||^2, \end{aligned}$$

where $\bar{d} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{l=1, i\neq l}^{n} d_{il}$, that is

$$\begin{aligned} \mathrm{T} &= \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i\neq l}}^{n} \left(d_{il} - \bar{d}\right)^2 \\ &= F_1(\alpha_1, \alpha_2, \hat{\mathbf{M}}) + \left(\alpha_1 - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i\neq l}}^{n} \sum_{k=1}^{K} \hat{m}_{ik}\hat{m}_{lk} \\ &\quad + \left(\alpha_2 - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i\neq l}}^{n} \left(1 - \sum_{k=1}^{K} \hat{m}_{ik}\hat{m}_{lk}\right). \end{aligned}$$

Now, T is constant, hence the minimization of $F_1(\alpha_1, \alpha_2, \hat{\mathbf{M}})$ with respect to α_1 and α_2 is equivalent to the maximization of

$$\begin{aligned} \mathrm{B} &= (\bar{d} - \alpha_1)^2 ||(\hat{\mathbf{M}}\hat{\mathbf{M}}' - \mathbf{I})||^2 + (\alpha_2 - \bar{d})^2 ||(\mathbf{1}_n\mathbf{1}_n' - \hat{\mathbf{M}}\hat{\mathbf{M}}')||^2 \\ &= (\bar{d} - \alpha_1)^2 \left(\sum_{k=1}^{K} n_k^2 - n\right) + (\alpha_2 - \bar{d})^2 \left(n^2 - \sum_{k=1}^{K} n_k^2\right). \end{aligned}$$

Furthermore,

$$\bar{d} = \frac{\alpha_1 \left(\sum_{k=1}^{K} n_k^2 - n\right) + \alpha_2 \left(n^2 - \sum_{k=1}^{K} n_k^2\right)}{n^2 - n}.$$

Hence, $\bar{d}$ is the weighted arithmetic mean of the current α_1 and α_2, and therefore $0 < \alpha_1 \leq \bar{d} \leq \alpha_2$, if $\alpha_1 \leq \alpha_2$. Now at the pth step, let us suppose that the updated alpha's fulfill $0 < \hat{\alpha}_1^{(p)} \leq \hat{\alpha}_2^{(p)}$ and that function (17) has value $\mathrm{B}^{(p)}$. Thus if:

$$\begin{aligned} B^{(p+1)} &= \left(\hat{\alpha}_1^{(p+1)} - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i\neq l}}^{n} \sum_{k=1}^{K} \hat{m}_{ik}^{(p+1)} \hat{m}_{lk}^{(p+1)} \\ &\quad + \left(\hat{\alpha}_2^{(p+1)} - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i\neq l}}^{n} \left(1 - \sum_{k=1}^{K} \hat{m}_{ik}^{(p+1)} \hat{m}_{lk}^{(p+1)}\right) \end{aligned}$$

$$B^{(p)} = \left(\hat{\alpha}_1^{(p)} - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i \neq l}}^{n} \sum_{k=1}^{K} \hat{m}_{ik}^{(p+1)} \hat{m}_{lk}^{(p+1)}$$
$$+ \left(\hat{\alpha}_2^{(p)} - \bar{d}\right)^2 \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i \neq l}}^{n} \left(1 - \sum_{k=1}^{K} \hat{m}_{ik}^{(p+1)} \hat{m}_{lk}^{(p+1)}\right)$$

the sum of the squared Euclidean distance between $\bar{d}$ and $\hat{\alpha}_1^{(p+1)}$ and between $\hat{\alpha}_2^{(p+1)}$ $\bar{d}$ does not decrease with respect to the previous step and always remains non-negative; it follows that $\hat{\alpha}_1^{(p+1)} \leq \hat{\alpha}_2^{(p+1)}$ still holds.

4.6 Algorithms for the well-structured partition and the hierarchical partition models

Even in this case, a coordinate descent algorithm can be described by two basic steps which are sequentially repeated, after the initialisation of the parameters and until a stopping rule is satisfied.

Step 0: Initialisation: a small convergence tolerance value ε is fixed and an initial feasible partition $\mathbf{M}$ is given, i.e., such that $\max\{{}_{\mathrm{W}}d_{kk}\ k = 1, \ldots K\} \leq \min\{{}_{\mathrm{B}}d_{kh} : h, k = 1, \ldots, K, (h \neq k)\}$. However, if we are not necessarily interested to find a well-structured partition, $\mathbf{M}$ can be randomly chosen.

Step 1: Updating $\mathbf{D}_{\mathrm{W}}$ and $\mathbf{D}_{\mathrm{B}}$.

For the current matrix $\hat{\mathbf{M}}$, the updates of $\mathbf{D}_{\mathrm{W}}$ and $\mathbf{D}_{\mathrm{B}}$ are given by

$$\mathbf{D}_{\mathrm{W}} = \mathrm{diag}(\hat{\mathbf{M}}'\mathbf{D}\hat{\mathbf{M}})[(\hat{\mathbf{M}}'\hat{\mathbf{M}})^2 - \hat{\mathbf{M}}'\hat{M})]^{+},$$
$$\mathbf{D}_{\mathrm{B}} = (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}'[\mathbf{D} - \hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}' + \mathrm{diag}(\hat{\mathbf{M}}\hat{\mathbf{D}}_{\mathrm{W}}\hat{\mathbf{M}}')]\hat{\mathbf{M}}(\hat{\mathbf{M}}'\hat{\mathrm{M}})^{-1}.$$

For the Square Hierarchical K-means algorithm the following step has to be added

Step 1′: Second Update of $\mathbf{D}_{\mathrm{B}}$ (Only for the Square hierarchical K-means).

The UPGMA algorithm is applied to $\mathbf{D}_{\mathrm{B}}$ to obtain a new $\mathbf{D}_{\mathrm{B}}$ satisfying the ultrametric inequalities.

Step 2: Updating $\mathbf{M}$

For the current matrices $\hat{\mathbf{D}}_{\mathrm{W}}$ and $\hat{\mathbf{D}}_{\mathrm{B}}$, the update of $\mathbf{M}$ is given by solving an assignment problem.

For each row $\mathbf{m}_i$ of $\mathbf{M}$, $i = 1, \ldots, n$,

$$m_{ik} = 1, \quad \text{if } F_2(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]')$$
$$= \min\{F_2(\hat{\mathbf{D}}_{\mathrm{W}}, \hat{\mathbf{D}}_{\mathrm{B}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_v, \ldots, \hat{\mathbf{m}}_n]') : v = 1, \ldots, K\}$$
$$m_{ik} = 0 \quad \text{otherwise.}$$

where $\mathbf{i}_v$ is the vth row of the identity matrix of order K.

Therefore, for each row ith of $\mathbf{M}$ the unique element equal to 1 is set in the column (class of the partition) that minimizes the loss function $F_2(\hat{\mathbf{D}}_{\mathrm{B}}, \hat{\mathbf{D}}_{\mathrm{W}}, [\hat{\mathbf{m}}_1, \ldots, \mathbf{m}_i = \mathbf{i}_k, \ldots, \hat{\mathbf{m}}_n]')$, for a fixed $\hat{\mathbf{D}}_{\mathrm{B}}, \hat{\mathbf{D}}_{\mathrm{W}}$. Passing to the next row, the loss function never increases and generally decreases.

Stopping rule. The function value $F_2(\mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{w}}, \mathbf{M})$ is computed for the current values of $\hat{\mathbf{D}}_{\mathrm{B}}, \hat{\mathbf{D}}_{\mathrm{W}}$ and $\hat{\mathbf{M}}$. When such updated values have decreased considerably the function value (more than an arbitrary small convergence tolerance value ε), $\mathbf{D}_B, \mathbf{D}_w$ and $\mathbf{M}$ are updated once more according to Steps 1 and 2.

Otherwise, the process is considered to have converged.

Remark 7 The following decomposition of the total deviance of the dissimilarities holds

$$\begin{aligned} T &= \sum_{i=1}^{n} \sum_{\substack{l=1 \\ i \neq l}}^{n} \left(d_{il} - \bar{d}\right)^2 \\ &= F(\hat{\mathbf{M}}, \mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{w}}) + \sum_{k=1}^{K} \left({}_{\mathrm{W}}d_{kk} - \bar{d}\right)^2 \left(n_k^2 - n_k\right) \\ &\quad + \sum_{k=1}^{K} \sum_{\substack{h=1 \\ k \neq h}}^{K} \left({}_{\mathrm{B}}d_{kh} - \bar{d}\right)^2 n_k n_h. \end{aligned}$$

where

$$\bar{d} = \frac{\sum_{k=1}^{K} {}_{\mathrm{W}}d_{kk} \left(n_k^2 - n_k\right) + \sum_{k=1}^{K} \sum_{k=1, k \neq h}^{K} {}_{\mathrm{B}}d_{kh} n_k n_h}{n^2 - n}.$$

Hence, $\bar{d}$ is the weighted arithmetic mean of the current ${}_{\mathrm{W}}d_{kk}$ and ${}_{\mathrm{B}}d_{kh}$, and therefore ${}_{\mathrm{W}}d_{kk} \leq \bar{d} \leq_{\mathrm{B}} d_{kh}$, if ${}_{\mathrm{W}}d_{kk} \leq_{\mathrm{B}} d_{kh}, \forall h, k$. Furthermore, since T is constant the minimization of $F_2(\hat{\mathbf{M}}, \mathbf{D}_{\mathrm{B}}, \mathbf{D}_{\mathrm{w}})$ with respect to ${}_{\mathrm{W}}d_{kk}, {}_{\mathrm{B}}d_{kh}$, is equivalent to the maximization of

$$B = \sum_{k=1}^{K} \left({}_{\mathrm{W}}d_{kk} - \bar{d}\right)^2 \left(n_k^2 - n_k\right) + \sum_{k=1}^{K} \sum_{\substack{h=1 \\ k \neq h}}^{K} \left({}_{\mathrm{B}}d_{kh} - \bar{d}\right)^2 n_k n_h. \tag{44}$$

If the solution at the pth step satisfies the well-structured partition property ${}_{\mathrm{W}}d_{kk}^{(p)} \leq {}_{\mathrm{B}}d_{kh}^{(p)} \forall h, k$, and function (44) has value $\mathrm{B}^{(p)}$, the algorithm produces, at $(p+1)$th step, new values ${}_{\mathrm{W}}d_{kk}^{(p+1)}, {}_{\mathrm{B}}d_{kh}^{(p+1)}$, still satisfying the well-structured partition property. In fact, $\mathrm{B}^{(p+1)} \geq B^{(p)}$ and therefore the sum of the squared Euclidean distance between $\bar{d}$ and ${}_{\mathrm{W}}d_{kk}^{(p+1)}$ and between ${}_{\mathrm{B}}d_{kh}^{(p+1)}$ and $\bar{d}$ does not decrease with respect to the previous step and always remains non negative; it follows that ${}_{\mathrm{W}}d_{kk}^{(p+1)} \leq {}_{\mathrm{B}}d_{kh}^{(p+1)}$ still holds.

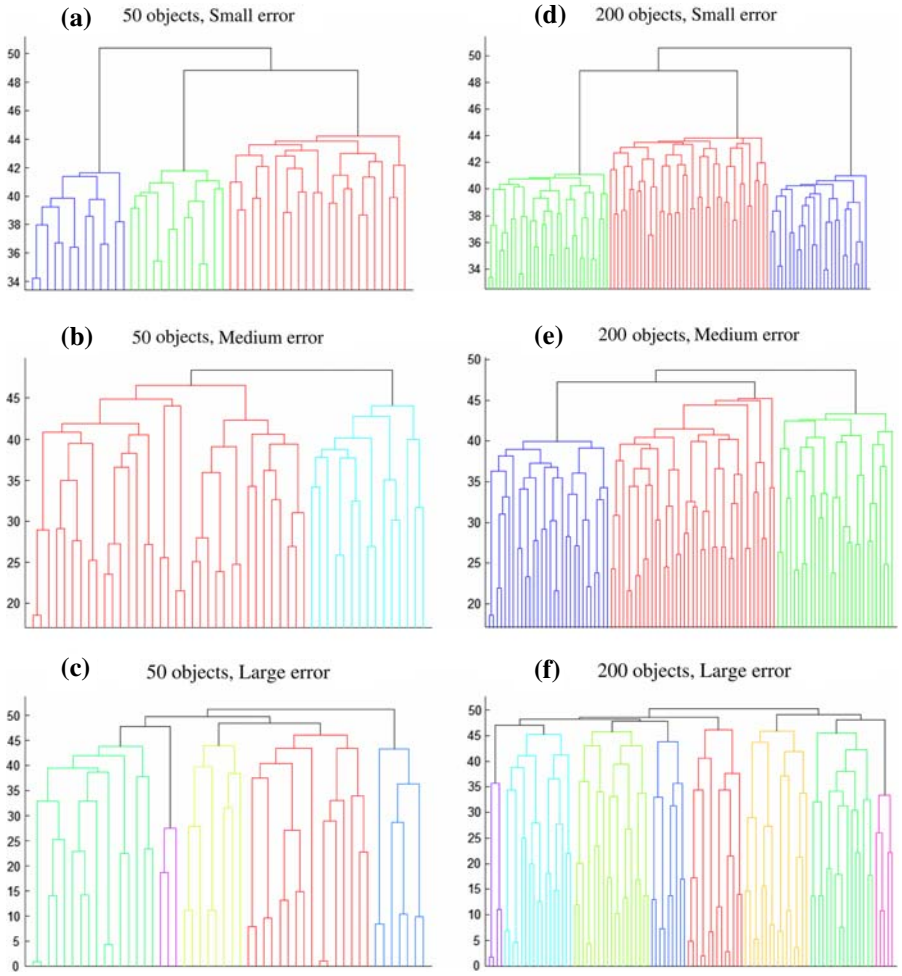


Fig. 9 Generated dendrogram with low, medium and high error levels. The cases **a**–**c** refers to $n = 50$ objects; the cases **d**–**f** refers to $n = 200$ objects

5 Simulation studies

In order to check the performance of the proposed algorithms, a simulation study has been considered to test the ability in recovering the true partitions and the sensitivity of the algorithms to the problem of local minima.

The dissimilarity matrices $\mathbf{D}$ have been generated according to the general error model (1), where the classification matrix $\mathbf{D}_c$ has one of the three semiparametric forms (2), (4) or (4) with $\mathbf{D}_B$ ultrametric, i.e., the well-structured perfect partition, the well-structured partition in K clusters, and the hierarchical partition, respectively.

The number of clusters has been fixed to $K = 3$ and our data sets have $n = 50$ or 200 objects with approximately equal-sized clusters.

Three error levels have been considered by adding to the model left truncated and centered normal random values. The low perturbation of the first level (0.2% of the Residual Sum of Squares (RSS)), can be observed in Fig. 9a and d where the well-structured partition in three separated clusters is still clearly observable. On the other hand, when the error level is high (8.5% RSS), the three clusters are hardly recognizable, as is shown in Fig. 9c and f. The proposed algorithms have been applied on the perturbed dissimilarity matrices to recover the clusters and, specifically, the original partitions.

Two starting point configurations have been proposed. The generated dissimilarity data values d_{il} are sorted in ascending order (considering only the upper triangle of the matrix, i.e., excluding $d_{ii} = 0$) and subsequently split into two sets d_W, d_B. The first set is formed by the $0.8 \times n(n-1)/4$ smallest dissimilarities, i.e., the 80% of those that potentially are the within clusters dissimilarities; while the second set is formed by the 80% of the largest dissimilarities that should represent the between classes dissimilarities.

In the first random starting procedure the elements of matrices $\mathbf{D}_W$ and $\mathbf{D}_B$ are sampled from normal distributions with mean equal to mean(d_W) and mean(d_B) for $\mathbf{D}_W$ and for $\mathbf{D}_B$, respectively and standard deviation equal to 2, for both cases.

In the second random starting procedure the elements of matrices $\mathbf{D}_W$ and $\mathbf{D}_B$ are sampled from d_W, d_B, respectively.

Combining, the two sizes of the dissimilarity matrices, the two different starting procedures and the three error levels, there are 12 different experimental situations. For each one, 100 data sets have been generated.

The performance of each algorithm has been evaluated by computing the following measures for each cell of the experiment (Tables 1, 2, 3, 4):

1. Average MRand($\mathbf{M}$,$\hat{\mathbf{M}}_t$); where MRand is the Modified Rand Index (Hubert and Arabie 1985) between the true $\mathbf{M}$ and the fitted matrix $\hat{\mathbf{M}}_t$ $t = 1, \ldots, 100$.
2. Percentage of times the fitted partition gives a loss function value smaller or equal to the loss function value of the true partition, and, in parentheses, the percentage of times the fitted partition is equal to the true partition (i.e., such that MRand($\mathbf{M}$,$\hat{\mathbf{M}}_t$) = 1).
3. Percentage of times where the fitted partition has the loss function value greater than the value of the true partition ("sure" local optima).
4. Average number of iterations of the algorithm.

The first two measures are computed to evaluate the goodness of the algorithms in recovering the true clustering. The third measure is introduced to study the local minima problem, because it represents the number of times the algorithm is trapped into a local minimum. Of course, this is only a lower bound of the true number. The fourth measure concerns the computational complexity.

All measures depend on the number of times the algorithm is run. To study the sensitivity of the performance of the algorithms to the number of random starts, we have considered 1, 5, 10, 20 and 30 random starts for each algorithm. They are nested, in the sense that the algorithm was run 30 times with different random starts and: the first solution, the best of first 5, 10, 20 and 30 solutions has been retained. In this way,

Table 1 Simulation study for 50 objects by fitting model (2) with a well-structured perfect partition in three clusters

First procedure for random start					Second procedure for random start				
# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations	# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations
Small error (0.2% RSS)									
1	1.0000	100 (100)	0	1.90	1	1.0000	100 (100)	0	1.76
5	1.0000	100 (100)	0	1.96	5	1.0000	100 (100)	0	1.76
10	1.0000	100 (100)	0	1.96	10	1.0000	100 (100)	0	1.76
20	1.0000	100 (100)	0	1.96	20	1.0000	100 (100)	0	1.76
30	1.0000	100 (100)	0	1.96	30	1.0000	100 (100)	0	1.76
Medium error (3.5% RSS)									
1	0.9988	100 (99)	0	2.57	1	0.9773	96 (96)	4	2.60
5	0.9988	100 (99)	0	2.57	5	0.9994	100 (99)	0	2.60
10	0.9988	100 (99)	0	2.57	10	0.9994	100 (99)	0	2.60
20	0.9988	100 (99)	0	2.57	20	0.9994	100 (99)	0	2.60
30	0.9988	100 (99)	0	2.57	30	0.9994	100 (99)	0	2.60
Large error (8.5% RSS)									
1	0.6151	68 (14)	32	4.66	1	0.5175	54 (12)	46	4.83
5	0.7722	99 (18)	1	5.26	5	0.7058	91 (16)	9	5.23
10	0.7781	100 (19)	0	5.02	10	0.7349	99 (16)	1	5.28
20	0.7908	100 (19)	0	5.14	20	0.7473	99 (16)	1	5.27
30	0.7927	100 (19)	0	5.05	30	0.7522	99 (16)	1	6.29

Table 2 Simulation study for 50 objects by fitting model (4) with the well-structured partition in three clusters

First procedure for random start					Second procedure for random start				
# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations	# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations
Small error (0.2% RSS)									
1	0.8931	76 (76)	24	1.83	1	0.9291	84 (84)	16	1.82
5	1.0000	100 (100)	0	1.87	5	1.0000	100 (100)	0	1.76
10	1.0000	100 (100)	0	1.87	10	1.0000	100 (100)	0	1.76
20	1.0000	100 (100)	0	1.87	20	1.0000	100 (100)	0	1.76
30	1.0000	100 (100)	0	1.87	30	1.0000	100 (100)	0	1.76
Medium error (3.5% RSS)									
1	0.8735	75 (71)	25	2.90	1	0.8400	68 (66)	32	3.00
5	0.9920	99 (96)	1	3.18	5	0.9919	99 (95)	1	3.12
10	0.9977	100 (97)	0	3.15	10	0.9940	100 (96)	0	3.10
20	0.9977	100 (97)	0	3.15	20	0.9940	100 (96)	0	3.10
30	0.9977	100 (97)	0	3.15	30	0.9940	100 (96)	0	3.10
Large error (8.5% RSS)									
1	0.5035	59 (6)	41	4.52	1	0.4529	45 (3)	55	5.67
5	0.7023	98 (10)	2	5.44	5	0.6157	89 (9)	11	5.97
10	0.7126	100 (10)	0	5.38	10	0.6703	99 (11)	1	6.44
20	0.7196	100 (11)	0	5.61	20	0.6744	99 (13)	1	6.37
30	0.7304	100 (11)	0	5.76	30	0.6888	100 (13)	0	6.29

Table 3 Simulation study for 50 objects by fitting (4) with partition C in well-structured three clusters and $\mathbf{D}_B$ ultrametric

First procedure for random start					Second procedure for random start				
# random starts	Average MRand	% partitions $\leq$ (=) true partitions	% sure local optima	Average # of iterations	# random starts	Average MRand	% partitions $\leq$ (=) true partitions	% sure local optima	Average # of iterations
Small error (0.2% RSS)									
1	0.9058	79 (79)	21	1.87	1	0.9273	84 (84)	16	1.87
5	1.0000	100 (100)	0	1.92	5	1.0000	100 (100)	0	1.81
10	1.0000	100 (100)	0	1.89	10	1.0000	100 (100)	0	1.81
20	1.0000	100 (100)	0	1.89	20	1.0000	100 (100)	0	1.81
30	1.0000	100 (100)	0	1.89	30	1.0000	100 (100)	0	1.81
Medium error (3.5% RSS)									
1	0.8908	78 (74)	22	2.93	1	0.8300	66 (64)	34	3.00
5	0.9972	100 (96)	0	3.18	5	0.9966	100 (97)	0	3.08
10	0.9977	100 (97)	0	3.19	10	0.9940	100 (97)	0	3.10
20	0.9977	100 (97)	0	3.19	20	0.9940	100 (97)	0	3.10
30	0.9977	100 (97)	0	3.19	30	0.9940	100 (97)	0	3.10
Large error (8.5% RSS)									
1	0.5185	55 (6)	45	4.54	1	0.4720	52 (6)	48	5.46
5	0.6959	95 (10)	5	5.22	5	0.6402	91 (14)	9	6.37
10	0.7029	99 (11)	1	5.35	10	0.6747	97 (15)	3	6.44
20	0.7112	100 (13)	0	5.69	20	0.6884	99 (15)	1	6.59
30	0.7246	100 (13)	0	5.81	30	0.6968	100 (15)	0	6.59

Table 4 Simulation study: results for 200 objects by fitting model (4) with the partition in well-structured three clusters

First procedure for random start					Second procedure for random start				
# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations	# random starts	Average MRand	% partitions ≤ (=) true partitions	% sure local optima	Average # of iterations
Small error (0.2% RSS) starting point 2									
1	0.9102	80 (80)	20	2.17	1	0.8822	74 (74)	26	2.07
5	0.9909	98 (98)	2	2.08	5	1.0000	100 (100)	0	2.03
10	1.0000	100 (100)	0	2.03	10	1.0000	100 (100)	0	2.03
20	1.0000	100 (100)	0	2.03	20	1.0000	100 (100)	0	2.03
30	1.0000	100 (100)	0	2.03	30	1.0000	100 (100)	0	2.03
Medium error (3.5% RSS)									
1	0.9233	85 (84)	15	3.27	1	0.8528	72 (72)	28	3.63
5	0.9998	100 (99)	0	2.82	5	0.9891	98 (98)	2	3.00
10	0.9998	100 (99)	0	2.82	10	0.9947	99 (97)	1	2.97
20	0.9998	100 (99)	0	2.82	20	0.9947	99 (97)	1	2.97
30	0.9998	100 (99)	0	2.82	30	0.9947	99 (97)	1	2.97
Large error (8.5% RSS)									
1	0.9496	91 (79)	9	4.87	1	0.8356	69 (60)	31	5.22
5	0.9974	100 (88)	5	4.26	5	0.9800	97 (85)	3	4.93
10	0.9974	100 (88)	5	4.26	10	0.9955	100 (88)	0	4.73
20	0.9974	100 (88)	5	4.26	20	0.9955	100 (88)	0	4.73
30	0.9974	100 (88)	5	4.26	30	0.9955	100 (88)	0	4.73

Table 5 Comparison of K-means, UPGMA, square K-means and square hierarchical K-means

Methods	MRand ($K = 6$)	MRand ($K = 7$)
K-Means	0.6335	0.6843
UPGMA with cut at K clusters	0.9024	0.7959
Square K-means	**0.9301**	0.7936
Square hierarchical K-means	0.7329	**0.8277**

MRand has been computed between the best partition in $K = 6$ and $K = 7$ classes and the expected partition in seven classes
The largest MRand is highlighted in bold

the comparability among results corresponding to algorithms with different numbers of random starts is guaranteed and some computational time saved.

Of course, the same starting configurations have been used for all the algorithms.

The simulation results for 50 objects are reported in Tables 1, 2 and 3 for the three classification matrices. The average performance measures show a significant improvement in the identification of the true partition when the solution is obtained from a large number of random initial solutions.

The first procedure for random initial starts seems to give generally better results with respect to the second one.

Table 4 shows the results of the simulation study for 200 objects with model (4) with the well-structured partition in 3 clusters. Similar considerations to the cases above follow, even for this size of the data.

6 Application on a real data set

The Richard Forsyth's zoological data set (UCI repository of machine learning databases, Asuncion and Newman 2007) refers to 101 animals characterized by 15 Boolean attributes evaluating the presence/absence of hair, feathers, eggs, milk, airbone, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize. The numeric variable for the number of legs has not considered in this application. The squared Euclidean distance was computed on the 15 binary variables.

An additional supplementary variable denoting which of seven different classes each animal belongs to, corresponding to mammals, birds, reptiles, fishes, amphibians, insects, mollusks and arthropods has been used to validate the results of the compared clustering methodologies. The applied methods are: K-means (directly on the data), the group average linkage (UPGMA) cutting its dendrogram to the height of K clusters, the square K-means, and the square hierarchical K-means. The MRand has been used to evaluate the similarity between each partition of the compared methods with the one defined by the supplementary variable.

The compared algorithms were run from 100 random starts (unless UPGMA) by setting the number of classes equal to 7, just as the number of groups of animals described by the supplementary variable. For each algorithm the best solution in terms of objective functions was retained. We also examined the partitions in 6 classes because reptiles and amphibians cannot be easily distinguished (also because the variable # of legs has not been considered). The partition in 6 classes which is the

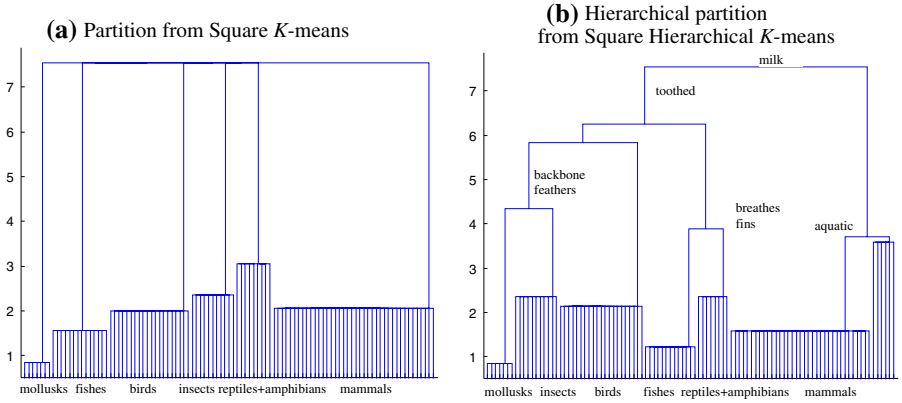


Fig. 10 Partition and hierarchical partition of the zoo data

most similar to the one of the supplementary variable is given by the *square K-means* (Table 5; Fig. 10a). There are six misclassified animals. The most similar partition in seven classes is given by the *square hierarchical K-means* because it takes into account the hierarchical evolutionary information included in the zoo data. In fact, by cutting the hierarchical partition in Fig. 10b at the level of two classes emerges the distinction between the 41 mammals and the 60 remaining oviparous animals which is detected by presence/absence of the variable "milk". The partition into 3 classes is obtained by splitting the largest class into two groups: 21 animals including most reptiles, amphibians and fishes, all toothed, mostly vertebrates, non-domestic, aquatic, with no hair, no-feathers, and 39 animals including birds, mollusks, insects, all non-toothed, non-fins, mostly non-venomous, terrestrial, with no hairs. Successively, mollusks-insects split from birds because for variables: "feathers", "backbone" and "tail". Then, reptiles+amphibians and fishes break into two classes according to the variables: "breathes" and "fins". Finally in the mammals cluster the six aquatic mammals (dolphin, mink, platypus, porpoise, seal and sealion) are detected.

7 Discussion

In this paper, the taxonomic problem to partition a finite set of objects, when a dissimilarity matrix $\mathbf{D}$ has been observed is formalized with the statistical modeling approach of fitting an expected *clustering model* (here a partition), specified by a *classification matrix* $\mathbf{D}_c$, to the observed $\mathbf{D}$. A classification matrix represents a clustering model expressed in terms of dissimilarities and characterized by heterogeneity or lack of cohesion for classes and isolation between classes. There is a lack of methodologies appropriate and widely employed to directly partition a set of objects from dissimilarity data, which motivates our study on three new partitioning models for dissimilarity data, their estimation via least-squares and the introduction of three new fast algorithms.

The first partitioning model specifies the *well-structured perfect partition in K classes*, which is characterized by equal heterogeneity and isolation within and bet-

ween classes, respectively. This partitioning model has been defined by Rubin (1967) and discussed by several authors (see Gordon 1999 for a review). It is implicitly used when one chooses a partition by cutting a dendrogram. In fact, the cutting level identifying the partition represents the value of heterogeneity within classes, while the isolation between classes is specified by the highest level of the dendrogram. In this paper we have shown that if a well-structured partition in K clusters is present in the dissimilarity data, the common practice of choosing a partition in the dendrogram frequently fails to recover this partition. On the other hand the appropriate algorithm performs better because it is specifically designed for detecting well-structured partitions.

The second classification model for dissimilarity data proposed in this paper allows for fixing different lacks of cohesion within classes and different isolations between classes, so to have that the largest heterogeneity is smaller than the smallest isolation. This is a well-structured partition in K classes.

We have verified if suitable standard procedures such as tandem analysis (i.e., MDS on $\mathbf{D}$ with the subsequent application of K-means on the dimensions of MDS) can detect a well-structured partition with the same performances of an appropriate method such as square K-means. This last technique is a new algorithm resembling K-means, but it can be used on dissimilarity data. The performances of square K-means are better than tandem analysis when dissimilarities are not Euclidean (as the simulated experiments have shown in Sect. 1.1) which is quite frequent in real applications. In the application on zoo data, K-means applied directly on the data has recovered the expected classification less properly in terms of MRand than square K-means (see Table 5).

The third classification matrix identifies a well-structured partition in K classes with the additional property to evaluate the hierarchical relationship between clusters and for this reason is called hierarchical partition. In fact, the well-structured partition is fitted to the dissimilarity data and the hierarchical structure between clusters is also required. The classification matrix associated to a hierarchical partition represents mainly a partition that in addition describes a particular hierarchy of nested partitions from K to 1 levels. This partition identifies a parsimonious tree. Even in this case we have shown that a suitable standard hierarchical method has lower performances with respect to the square hierarchical K-means, that represents the the algorithm for detecting hierarchical partitions.

It may be noted that the algorithms given to solve the quadratic constrained problems necessary to evaluate the LS estimated parameters of the three models are in the class of coordinate descent algorithms type (Zangwill 1969).

The initial partition used to start the algorithms can be chosen at random or according to a rational procedure. In any case, different starting partitions and parameter values should be considered to increase the chance to obtain the global optimum partition since the clustering problem of optimally partitioning a set of multivariate objects is known to be an NP-hard problem and therefore the optimal solution cannot be guaranteed.

The algorithms generally stop after a few iterations and therefore are computationally fast. However, according to the simulation study, the algorithms tend to be trapped into local minima at least in about 25% cases even when data have a well-

defined clustering structure (low error), rising to about 50% when data are no so well clustered (high error). According to the simulation study the impact of local minima is drastically reduced when at least 30 initial random starts are used; however, again it has to be noted that we may be confident, but not sure the global optimal solution has been found.

Two random procedures for starting within and between classes heterogeneities and isolations matrices $\mathbf{D}_W$ and $\mathbf{D}_B$ have been proposed and from a simulation study it has been observed that it is more useful to sample $\mathbf{D}_W$ and $\mathbf{D}_B$ from normal distributions with means equal to the 80% of the smallest and largest observed dissimilarities respectively and variance equal to 2.

The choice of K is an open problem even for a classical clustering problem and consequently also for our proposed methodologies; it deserves a further successive investigation.

When a well-structured perfect partition is fitted (model (2)) K can be fixed equal to the maximum number of classes are hypothesized in the data. In this case the best solution will be with the correct number of clusters K^* ($K^* \leq K$) for the given data, leaving $K - K^*$ empty classes. In fact, these last do not modify and in particular reduce the value of the loss function (9) used to fit a well structured perfect partition.

For the other two partitioning models the situation is different because increasing the number of classes in the fitted partition also the number of parameters to estimate in $\mathbf{D}_B$ and $\mathbf{D}_W$ increases and consequently the loss-function used to fit the model generally decreases. In this case criteria for assessing the number of clusters have to take into account both the loss-function values and the number of estimated parameters of the model and these last should in some way penalize the loss-function. Without such "information criteria", we suggest to use the final partition that gives a "reasonable" interpretability, by considering the smallest K that produces the strongest change of the loss function (this strategy of choice has been introduced by Cattell (1966) for choosing the number of factors, in a factorial analysis).

An important aspect of the cluster validation is the cluster stability (Hennig 2007), i.e., the propensity of a cluster to appear in the classification if the data set is changed in a non-essential way (addition of outliers, jittering, etc.) we will also deserve a further successive investigation on this relevant characteristic.

References

Asuncion A, Newman DJ (2007) UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html

Ball GH, Hall DJ (1967) A clustering technique for summarizing multivariate data. Behav Sci 12:153–5

Bandelt HJ (1990) Recognition of the tree metrics. SIAM J Discrete Math 3(1):1–6

Bock HH (1974) Automatische klassifikation. studia mathematica. Vandenhoek und Ruprecht, Göttingen

Bock HH (1998) Probabilistic aspects in Classification. In: Hayashi et al. (eds) Data science, classification and related methods. Springer, Heidelberg, pp 3–21

Carroll JD, Pruzansky S (1975) Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. Paper presented at U. S.-Japan Seminar on Theory, Methods and Applications of Multidimensional Scaling and Related Techniques, San Diego, August 20–24

Carroll JD, Pruzansky S (1980) Discrete and hybrid scaling models. In: Lantermann ED, Feger (eds) Similarity and choice. Huber, Bern, pp 108–139

Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1:245–276
Chandon JL, Lemaire J, Pouget J (1980) Construction de l'ultramétrique la plus proche d'une dissimilarité an sens des moindres carrés. Recherche Operationelle / Operations Research 14:157–70
Cormack RM (1971) A review of classification (with discussion). J R Stat Soc A 134:321–67
Critchley F, Fichet B (1994) The partial order by inclusion of the principal classes of dissimilarity on finite set, and some of the basic properties. In: Van Cutsem B (ed) Classification and dissimilarity analysis, Lecture Notes in Statistics. Springer, Berlin, pp 5–65
Dahlhaus E (1993) Fast parallel recognition of ultrametrics and tree metrics. SIAM J Discrete Math 6(4):523–532
De Soete G (1984) A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. Pattern Recognit Lett 2:133–7
Fisher L, Van Ness JW (1971) Admissible clustering procedures. Biometrika 58:91–104
Fraley C, Raftery AE (2002) Model based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631
Gordon AD (1987) Parsimonious trees. J Classif 4:85–101
Gordon AD (1999) Classification: methods for the exploratory analysis of multivariate data. Chapman and Hall, London
Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–38
Grötschel M, Wakabayashi Y (1989) A cutting plane algorithm for a clustering problem. Math Program 45:59–96
Hansen P, Jaumard B, Sanlaville E (1994) Partitioning problems in cluster analysis: a review of mathematical programming approaches, In: Diday E, Lechevallier Y, Schader M, Bertrand P Burtschy B (eds) New approaches in classification and data analysis. Springer, Berlin, pp 228–40
Hartigan JA (1967) Representation of similarity matrices by trees. J Am Stat Assoc 62:1140–58
Hartigan JA (1975) Clustering algorithms. Wiley, New York
Hennig C (2007) Cluster-wise assessment of cluster stability. Comput Stat Data Anal 52:258–271
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Hubert L, Arabie P, Meulman J (1997) Hierarchical clustering and the construction of (optimal) ultrametrics using in lecture Notes, Monograph Series, vol 31. Institute of Mathematical Statistics, Hayward, CA, pp 457–72
Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs, NJ
Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
Krivánek M, Morávek J (1986) NP-hard problems in hierarchical-tree clustering. Acta Inform 23:311–323
Keller JB (1962) Factorization of matrices by least squares. Biometrika 49:239–242
Marcotorchino F, Michaud P (1982) Agregation de similarites en classification automatique. Revue de Statistique Appliquée 30(2):21–44
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J (eds) Proceedings of the fifth berkeley symposium on mathematical statistics and probability, vol 1, Statistics. University of California Press, Berkeley, pp 281–297
Mathar R (1985) The best Euclidean fit to a given distance matrix in prescribed dimensions. Linear Algebra Appl 67:1–6
McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York
Oh M-S, Raftery AE (2007) Model-based clustering with dissimilarities: a bayesian approach. J Comput Graph Stat. A Technical Report is available on http://www.stat.washington.edu/raftery/Research/bayes.html (to appear)
Powell MJD (1983) Variable metric methods for constrained optimization. In: Bachem A et al (eds) Mathematical programming: the state of the art. Springer, Berlin, pp 288–311
Régnier S (1965) Sur quelques aspects mathématiques des problemes de classification automatique. International Computation Centre Bulletin 4, 175–91. Reprinted in Mathématiques et Sciences Humaines, 82, 13–29 (1982)
Rubin J (1967) Optimal classification into groups:an approach to solve the taxonomy problem. J Theor Biol 15:103–144
Sriram N (1990) Clique optimization: a method to construct parsimonious ultrametric trees from similarity data. J Classif 7:33–52
Torgerson WS (1958) Theory and methods of scaling. Wiley, New York

Vicari D, Vichi M (2000) Non-hierarchical classification structures. In: Gaul W, Opitz O, Schader M (eds) Data analysis, studies in classification data analysis and knowledge organization. Springer, Berlin, pp 51–66

Vichi M (1993) Un algoritmo dei minimi quadrati per interpolare un insieme di classificazioni gerarchiche con una classificazione consenso, Metron, vol 51, 3–4, 139–163

Vichi M (1994) Un algoritmo per il consenso tra classificazioni gerarchiche con l'ausilio di tecniche multiway. Proc Italian Stat Soc 37:261–268

Vichi M (1996) Computational complexity of one mode classification, In: Prat A (ed) Proceedings of COMPSTAT vol 96. Physica-Verlag

Zangwill WI (1969) Nonlinear programming: a unified approach. Prentice-Hall, Englewood Cliffs, NJ