Journal of Classification 26:121-154 (2009)
DOI: 10.1007/s00357-009-9033-0

Structural Classification Analysis of Three-Way Dissimilarity Data

Donatella Vicari

Sapienza University of Rome

Maurizio Vichi

Sapienza University of Rome

Abstract: The paper presents a methodology for classifying three-way dissimilarity data, which are reconstructed by a small number of consensus classifications of the objects each defined by a sum of two order constrained distance matrices, so as to identify both a partition and an indexed hierarchy.

Specifically, the dissimilarity matrices are partitioned in homogeneous classes and, within each class, a partition and an indexed hierarchy are simultaneously fitted.

The model proposed is mathematically formalized as a constrained mixed-integer quadratic problem to be fitted in the least-squares sense and an alternating least-squares algorithm is proposed which is computationally efficient.

Two applications of the methodology are also described together with an extensive simulation to investigate the performance of the algorithm.

Keywords: Dissimilarity; Three-Way data; Classification; Hierarchy; Partition.

Authors' Address: Department of Statistics, Probability and Applied Statistics, Sapienza University of Rome, Italy, e-mails: donatella.vicari@uniroma1.it; maurizio.vichi@uniroma1.it

Published online 6 June 2009

1. Introduction

Cluster analysis methodologies usually detect a unique classification model[1] (e.g., covering, partition, hierarchy, fuzzy partition) within the observed dissimilarity data, chosen in the class of classification models of the same type. The aim is to reconstruct the original observed dissimilarity matrix by a new one which is constrained to identify (one-to-one) a classification structure. However, it may be unrealistic that a single classification model is representative of the taxonomic information of the data. Several classification structures often coexist, because objects are generally described by several variables which can specify different classification models (e.g., each categorical variable induces a partition and a group of variables may specify nested partitions of objects). Multiple classification structures may exist and one desires to synthesize such clusterings into a single “consensus”; more frequently, such multiple classification structures are not observable and need to be detected (Gordon 1999, chapter 4.4; for high-dimensional data, see Strehl and Ghosh 2002; Fern and Brodley 2003; Fred and Jain 2003). In other terms, the observed dissimilarity matrix may be thought of as a “combination” of dissimilarity matrices corresponding to variables or subsets of variables related to similar contexts (e.g., economic, demographic, social, psychological) and each of such dissimilarity matrices may be properly reconstructed by a different classification model. According to this perspective, a clustering algorithm should identify not only a unique classification structure, but the most relevant ones possibly present in the observed dissimilarity matrix. Several interesting papers have faced this approach in the case of two-way one-mode dissimilarity matrices. Hubert and Arabie (1994) propose to approximate the observed dissimilarity matrix through a sum of a small number of symmetric order-constrained matrices in the class of the Robinson matrices. Hubert, Arabie and Meulman (1998) fitted more order-constrained matrices such as strongly-(anti)-Robinson matrices to permit a representation of the fitted values which has the nice and useful property to improve and enrich the interpretability of the results by graphical displays.

As a further development within this approach, it could be assumed that several classification models of different types may be identified in a dissimilarity matrix. In particular, in this paper we will suppose that a partition and an indexed hierarchy can properly reconstruct the most relevant taxonomic information present in the data. We focus on these two kinds of structures since they are the most widely used and easy to be interpreted and, for reasons of parsimony, we impose to fit only one single partition

[1] Classification models and classification structures are hereafter used inter-changeably.

and one single indexed hierarchy to the observed dissimilarity data. However, straightforward modifications of the algorithm here proposed enable fitting other models with several partitions and/or several indexed hierarchies.

Of course, along with this approach it is necessary to verify whether such classification models are appropriate to describe and reconstruct the taxonomy present in the observed data, by evaluating the fit of the models to the data and the validity of the results.

In this paper we will also suppose that K dissimilarity matrices between pairs of the same set of I objects are observed so as to form a three-way dissimilarity array $\mathbf{D}=\{\mathbf{D}_1,\ldots, \mathbf{D}_K\}$. Large data sets of this type may be complex to analyze and specific methodologies are necessary to extract relevant information. Classification methodologies for three-way dissimilarity data assume, as for the two-way case, a unique common underlying classification structure in the data. However, since the taxonomy may change from occasion to occasion (e.g., preferences towards products given by different customers), the hypothesis that a single classification model can reconstruct the whole information in **D** appears even more unrealistic than in the two-way case.

To overcome this problem Carroll and Arabie (1983) introduced the INDCLUS model to extract overlapping clusters from a set of symmetric proximity matrices. Specifically, the model assumes that the objects are placed in a common (possibly overlapping) set of classes which are weighted differently by different observers (occasions).

Later, POP (Partitions Of Partitions) and PARSCLA (PARtition and least-Squares Consensus cLassifications Analysis) have been proposed by Gordon and Vichi (1998) and Vichi (1999), where a partition (termed "*secondary*") of the observed dissimilarity matrices forming **D** is found and a consensus classification model (termed "*primary*"), which can be a partition or a hierarchy, for each class of the secondary partition is also detected.

This paper can be placed in the same framework, by including the idea to find a secondary partition of $\mathbf{D}=\{\mathbf{D}_1,\ldots, \mathbf{D}_K\}$ in classes of homogeneous dissimilarity matrices and, within each class, primary classifications of the objects. Specifically, within each of such classes of occasions, two constrained distance matrices are fitted corresponding to partitions and indexed hierarchies of the objects.

The novelty lies in the fact that the primary classification may itself exist of a combination of classification structures (i.e. hierarchies and partitions, instead of a hierarchy or a partition).

Without loss of generality and for the sake of simplicity and clarity, we fully discuss the case of fitting a single partition (**P**) and a single indexed hierarchy (**U**), since a more general model can be easily fitted by a slight modification of the algorithm proposed. In fact, the algorithm is based on successive residualizations of the given three-way data matrix: within each class of occasions, one matrix is fitted, obtaining the residual dissimilarities from it and then the second matrix is fitted to these residuals. The general case with several partitions and/or indexed hierarchies follows straightforwardly.

The simultaneous fitting of **P** and **U** is motivated by the underlying hypothesis that the variables characterizing dissimilarities can be assumed to be partitioned into two subsets not related to each other: one describing the "evolutionary" features of the objects, which motivates the use of the indexed hierarchy, and the remaining subset of variables inducing homogeneous clusters which motivates the use of a partition (see Gordon 1999 or Kaufman and Rousseeuw 2005).

For example, in biological studies on animals or vegetation two taxonomies may be reasonably expected: the first generated by evolutionary characteristics specifying an evolutionary tree of the animals and the second by experimental variables identifying a partition of the animals or plants (e.g., with respect to a reaction to a pharmacological agent).

In Section 5.1 an illustrative application on zoological data, taken from the UCI Machine Learning repository (http://www.ics.uci.edu/~mlearn/MLRepository.html), is presented to show the coexistence of two different taxonomies in the data.

The proposed model is formalized in a statistically model-free least-squares context, by a quadratic minimization mixed-integer problem with four sets of constraints. The first two sets guarantee the identification of a partition of **D** in homogeneous classes of dissimilarity matrices and the remaining two sets impose hierarchical and partitioning order constraints on triplets of dissimilarities to detect a consensus partition and a consensus indexed hierarchy within each homogeneous class.

An outline of the material in this paper is as follows. Section 2 describes the model for fitting the sum of a hierarchy and a partition to the three-way data set **D**. The model is estimated by solving a constrained least-squares problem. Section 3 extends the model to detect a "secondary" partition of the set of dissimilarity matrices forming **D** and, simultaneously, both a consensus indexed hierarchy and a consensus partition within each class of such a secondary partition of **D**. A numerical algorithm, based on alternating least-squares strategy, is also described in Section 4. The models discussed in Sections 2 and 3 are illustrated through the application of the algorithm on two well-known data sets which are analyzed in Section 5. In Section 6 the results of an extensive simulation study are also included to evaluate the performances of the algorithm. Finally, Section 7 includes comments and a discussion of the proposed methodology.

2. Fitting a Single Partition and a Single Indexed Hierarchy to Three-Way Dissimilarity Data

For the convenience of the reader, the basic terminology used is listed here:

I, K number of objects and occasions, respectively;

C number of classes of the *secondary* partition of $\mathbf{D}$, i.e. number of the classes of occasions ($C \leq I$);

$O = \{o_1,\ldots, o_I\}$ set of I objects to be classified;

$\mathbf{D}=[d_{ijk}]$ $(I \times I \times K)$ array specifying dissimilarities between (o_i,o_j), $(i,j=1,\ldots,I)$ at the k^{th} occasion ($k=1,\ldots,K$); the array $\mathbf{D}$ can be written as a set of K $(I \times I)$ dissimilarity matrices i.e., $\mathbf{D}=[\mathbf{D}_1,\ldots,\mathbf{D}_K]$, where $\mathbf{D}_k=[d_{ijk}\!: i,j=1,\ldots,I]$;

w_k weight of the k^{th} occasion or dissimilarity data set ($k=1,\ldots,K$).

We consider here only the standard case where the diagonal values of the dissimilarity and distance matrices are equal to zero.

There are different classification models that are commonly adopted for a set O of objects: a *hierarchical partition*, which is a set of disjoint or nested subsets of O uniquely associated with an *ultrametric* matrix $\mathbf{U}$ (i.e. $\mathbf{U}=[u_{il}]$, $u_{il} \geq 0$, $u_{il} = u_{li}$, $u_{il} \leq \max(u_{ij}, u_{jl})$, $\forall\ (i,l,j) \in O$) or a *non-hierarchical partition* (see Gordon 1999). The latter can be uniquely associated to a 2-*ultrametric* matrix $\mathbf{P}$, i.e., a constrained ultrametric matrix with at most two different off-diagonal values: $\mathbf{P}=[p_{il}]$, $p_{il} \geq 0$, $p_{il} = p_{li}$, $p_{il} \leq \max(p_{ij}, p_{lj})$ $\forall\ (i,l,j) \in O$ and $p_{il} \in \{0, a, b\}$ where $a \leq b$ (Vicari and Vichi 2000). Therefore, the non-hierarchical partition can be regarded as a constrained hierarchy where only two levels of aggregation are allowed.

The two classification structures (*hierarchies* and *partitions*) considered for a given set O, define cones, i.e. subsets closed under non-negative scalar multiplication. In particular, the set of ultrametrics $\mathcal{U}$ is a non-convex closed cone. Obviously, the set of 2-ultrametrics $_p\mathcal{U}$ (i.e., the set of the ultrametric matrices with at most two different off-diagonal entries) is included in $\mathcal{U}$, i.e., $_p\mathcal{U} \subseteq \mathcal{U}$.

In this paper we consider that the taxonomic structure of $\mathbf{D}$ is specified both by a *secondary* partition of the K occasions into C classes and by a total of $2C$ (specifically, C hierarchies and C partitions) consensus *primary* classifications of the I objects.

We are now in position to state the model for reconstructing the classification structure of $\mathbf{D}$. In this Section we consider the simplest model where only one class of the secondary partition (C=1) is specified and

M+L matrices $\mathbf{P}_m$ and $\mathbf{U}_l$ are supposed to reconstruct the set of dissimilarity matrices according to the following model

$$\mathbf{D}_k = \sum_{m=1}^{M} \mathbf{P}_m + \sum_{l=1}^{L} \mathbf{U}_l + \mathbf{E}_k \qquad (k=1,...,K), \tag{1}$$

where $\mathbf{E}_k$ is a square matrix of error components.

For the sake of simplicity, we will fully discuss only the simplest case where M=L=1, i.e. the model

$$\mathbf{D}_k = \mathbf{P} + \mathbf{U} + \mathbf{E}_k \qquad (k=1,...,K), \tag{2}$$

where only two matrices **P** and **U** are assumed to reconstruct the set of dissimilarity matrices.

The presentation of the model (2) with respect to the more general model (1) does not imply lack of generality, because its extension is straight-forward and conversely the notation would become too burdensome.

The three terms in (2) explain portions of the original dissimilarity matrix $\mathbf{D}_k$. The first term pertains to what can be explained by a consensus partition of O, the second term by a consensus indexed hierarchy of O, the third term by neither **P** nor **U**. However, there might be redundancy in the model, due to the inclusion between the two classification structures above examined. In particular, if $\mathbf{D}_k$ has the structure of a (non-hierarchical) partition, i.e. $\mathbf{D}_k \in {}_p\mathcal{U}$, it follows that the fitted $\hat{\mathbf{P}}$ and $\hat{\mathbf{U}}$ identify the same classification model, because of the mentioned inclusion. Thus, the first term is subsumed under the other term. When $\mathbf{D}_k \in \mathcal{U}$, but not to ${}_p\mathcal{U}$, the two structures are different and there is no redundancy.

The problem of fitting the two order-constrained distance matrices **P** and **U** in (2) can be formalized by the following weighted quadratic constrained minimization problem with respect to **P** and **U**

$$\min \sum_{k=1}^{K} \| \mathbf{D}_k - \mathbf{P} - \mathbf{U} \|^2 w_k \qquad \text{[P1]}$$

subject to

$$\mathbf{P} \in {}_p\mathcal{U}$$

$$\mathbf{U} \in \mathcal{U},$$

where w_k is the positive weight of matrix $\mathbf{D}_k$. Suitable choices for possible weights depend on the particular applications and may be, for example, frequencies, subjective weights or the reciprocals of the sum-of-squares of the individual proximity matrices.

By adding and subtracting the weighted mean of the dissimilarity matrices, the objective function of [P1] can be rewritten

$$\sum_{k=1}^{K} \| \mathbf{D}_k - \mathbf{P} - \mathbf{U} \|^2 w_k = \sum_{k=1}^{K} \left\| \mathbf{D}_k - \sum_{k=1}^{K} \mathbf{D}_k \frac{w_k}{\sum_{k=1}^{K} w_k} \right\|^2 w_k + \left\| \sum_{k=1}^{K} \mathbf{D}_k \frac{w_k}{\sum_{k=1}^{K} w_k} - \mathbf{P} - \mathbf{U} \right\|^2 \sum_{k=1}^{K} w_k,$$

where only the second term on the right hand side depends on **P** and **U** and has to be minimized.

Problem [P1] can be directly solved by a Sequential Quadratic Programming (SQP) algorithm (Powell 1983). However, a direct solution of the problem poses computational complexity problems and it is particularly useful to study an alternative coordinate descent algorithm developed specifically for this method. An alternating least-squares algorithm will be discussed in Section 4 which solves problem [P1].

The methodology here presented will be termed *structural classification analysis*.

3. Fitting Secondary Partition, Consensus Partitions and Indexed Hierarchies to Three-Way Data

Model (2) is correctly formulated under the hypothesis that a *single* partition and a *single* indexed hierarchy are sufficient to synthesize the taxonomic structure in the three-way array **D**. However, as discussed in the introduction, the assumption that a single classification can describe a complex taxonomic structure may be often unrealistic, especially for three-way data, where different taxonomic information frequently corresponds to different occasions.

Therefore, a more flexible scheme is based on the idea that for some occasions of the three-way observed dissimilarity data, the taxonomic structure changes systematically, but in other occasions it differs only for some errors (e.g., sampling or measurement). Under this second hypothesis, it is useful to partition the set of dissimilarity matrices $\{\mathbf{D}_1,\ldots,\mathbf{D}_K\}$ into C disjoint homogeneous classes with similar taxonomic structure and identify both consensus indexed hierarchies and consensus partitions within each class. From now on, this problem will be discussed.

Given the three-way array **D**, a partition of the set of the K dissimilarity matrices $\{\mathbf{D}_1,\ldots,\mathbf{D}_K\}$ in C disjoint classes is determined, so that each class is synthesized by both a set of M consensus partitions and a set of L consensus hierarchies, each identified by an order-constrained distance matrix. The problem is mathematically formalized according to the following model

$$\mathbf{D}_k=\sum_{m=1}^{M}\mathbf{P}_{mc}+\sum_{l=1}^{L}\mathbf{U}_{lc}+\mathbf{E}_k,\quad \mathbf{D}_k\in\mathcal{G}_c,\qquad c=1,...,C;\qquad k=1,...,K, \tag{3}$$

where $\mathbf{E}_k$ is a square matrix of error components.

According to model (3), each matrix $\mathbf{D}_k$ belongs to a class $\mathcal{G}_c$ of the *secondary* partition of **D** and the dissimilarities in $\mathbf{D}_k$ are supposed to be reconstructed (at least approximately) by a number of distance matrices $\mathbf{P}_{mc}$ and $\mathbf{U}_{mc}$ with order constraints on the triplets of objects, specifying M partitions and L hierarchies, respectively.

As in Section 2, for the sake of clarity we will fully discuss the case M=L=1, i.e. the model

$$\mathbf{D}_k=\ \mathbf{P}_c+\mathbf{U}_c+\mathbf{E}_k,\quad \mathbf{D}_k\in\ \mathcal{G}_c,\quad c=1,...,C;\quad k=1,...,K, \tag{4}$$

where each matrix $\mathbf{D}_k$ belonging to a class $\mathcal{G}_c$ of the *secondary* partition of **D** is assumed to be reconstructed (at least approximately) by two distance matrices $\mathbf{P}_c$ and $\mathbf{U}_c$ with order constraints on the triplets of objects, identifying a partition and a hierarchy, respectively. The generalization to model (3) is straightforward.

Let $\mathbf{V}$=$[v_{kc}]$ be the (KxC) binary matrix specifying the secondary partition of **D**, where v_{kc}=1, if matrix $\mathbf{D}_k$ belongs to class $\mathcal{G}_c$, and v_{kc}=0, otherwise.

The primary classification structures $\mathbf{P}_c$=$[p_{ijc}]$ and $\mathbf{U}_c$=$[u_{ijc}]$ and the secondary partition $\mathbf{V}$=$[v_{kc}]$ in model (4) can be estimated according to the following least-squares fitting problem

$$\min\sum_{k=1}^{K}\sum_{c=1}^{C}\|\mathbf{D}_k-\mathbf{P}_c-\mathbf{U}_c\|^2\ w_k v_{kc} \qquad \text{[P2]}$$

subject to

$$(a)\quad v_{kc}\in\{0,\ 1\}\qquad\qquad k=1,\ldots,K,\quad c=1,...,C;$$

$$\sum_{c=1}^{C}v_{kc}=1\qquad\qquad k=1,\ldots,K;$$

(b) $u_{ijc} \leq \max(u_{ilc}, u_{jlc})$

$u_{jlc} \leq \max(u_{ijc}, u_{ilc})$ $\quad i=1,\ldots,I, j=i,\ldots,I,\ l=j,\ldots,I;\ c=1,\ldots,C;$

$u_{ilc} \leq \max(u_{ijc}, u_{jlc})$

(c) $p_{ijc} \in \{0, a_c, b_c\}$ $\quad i,j=1,\ldots,I,\quad c=1,\ldots,C;$

$p_{ijc} \leq \max(p_{ilc}, p_{jlc})$

$p_{jlc} \leq \max(p_{ijc}, p_{ilc})$ $\quad i=1,\ldots,I, j=i,\ldots,I,\ l=j,\ldots,I;\ c=1,\ldots,C;$

$p_{ilc} \leq \max(p_{ijc}, p_{jlc})$

where:

- the first two sets of constraints *(a)* guarantee that a partition of the set $\{\mathbf{D}_1,\ldots,\mathbf{D}_K\}$ into K disjoint classes is defined;
- the second set of triplets of constraints *(b)* ensures that $\mathbf{U}_c$ is an ultrametric matrix identifying an indexed hierarchy;
- the third set of constraints *(c)* guarantees that $\mathbf{P}_c$ is a 2-ultrametric matrix, identifying a hierarchy with only two levels (i.e., a partition).

Now, matrix $\mathbf{P}_c$ can be rewritten

$$\mathbf{P}_c = b_c\,(\mathbf{11}' - \mathbf{M}_c\mathbf{M}'_c) + a_c\,(\mathbf{M}_c\mathbf{M}'_c - \mathbf{I}) \qquad (5)$$

where $\mathbf{1}$ is a vector of I ones, $\mathbf{I}$ is the (I x I) identity matrix and $\mathbf{M}_c=[m_{ilc}]$ is a (I x C_c) matrix of binary values m_{ilc} specifying the membership of object o_i to the l^{th} group of the c^{th} consensus partition into C_c groups of objects (c=1,…,C). Substituting (5) into problem [P2], we have

$$\min \sum_{k=1}^{K} \sum_{c=1}^{C} \|\mathbf{D}_k - [b_c(\mathbf{11}' - \mathbf{M}_c\mathbf{M}_c') + a_c(\mathbf{M}_c\mathbf{M}_c' - \mathbf{I})] - \mathbf{U}_c\|^2\, w_k v_{kc}$$

[P2']

subject to

(a) $v_{kc} \in \{0, 1\}$ $\quad k=1,\ldots,K,\quad c=1,\ldots,C;$

$\sum_{c=1}^{C} v_{kc} = 1$ $\quad k=1,\ldots,K;$

(b) $\sum_{(i,j,l)\in\Gamma(\mathbf{U}_c)} (u_{ilc} - u_{jlc})^2 = 0$ $\quad c=1,\ldots,C;$

$$\begin{array}{lll}
(c) & b_c \geq a_c > 0 & c=1,...,C; \\
 & m_{ilc} \in \{0, 1\} & i=1,\ldots,I, \;\; l=1,\ldots,C_c, \;\; c=1,...,C; \\
 & \sum_{l=1}^{C_c} m_{ilc} = 1 & i=1,\ldots,I, \;\; c=1,...,C,
\end{array}$$

where $\Gamma(\mathbf{U}_c) = \{(i,j,l) : 1 \leq i,j,l \leq I, i \leq j \leq l : u_{ijc} \leq \min(u_{ilc}, u_{jlc})\}$, $c=1,\ldots,C$, that is the set of the triplets of objects having the two largest values equal (or, equivalently, identifying acute isosceles triangles).

It can be noted that problem [P2’] has a much smaller number of constraints than [P2]. In practice the $O(I^3)$ constraints on the triplets of $\mathbf{U}_c$ are synthesized by only one constraint since for each triplet the largest two values must be equal and their squared difference is null. Furthermore, the $O(I^3)$ constraints pertaining to matrix $\mathbf{P}_c$ are reduced to $O(IC_c)$, because we can properly reformulate constraints (c) in terms of the binary membership matrices $\mathbf{M}_c$, according to (5).

Problem [P2’] could be solved using a Sequential Quadratic Programming algorithm, even though it is not specific to solve mixed integer quadratic constrained problems; moreover, it quickly becomes infeasible and therefore, it is worthwhile to develop an alternative coordinate descent algorithm to solve problem [P2’], as described in the following Section.

4. An Alternating Least-Squares Algorithm

Problem [P2’] can be solved using a coordinate descent algorithm also known as Alternating Least-Squares algorithm, which splits the discrete and continuous optimization problems into parts easier to be solved.

The algorithm here proposed detects a partition of occasions and, within each of such classes, it fits two constrained distance matrices. After a first step where the partition of occasions is determined, the algorithm is based on successive residualizations of the given three-way data matrix: within each class of occasions, one matrix is fitted, obtaining the residual dissimilarities from it and then the second matrix is fitted to these residuals. The two steps are alternated and iterated until convergence.

Briefly, three steps are repeated in turn as follows:

a) The partition of occasions (secondary classification) is determined (*Updating of matrix* $\mathbf{V}$);

b) Within each class c of occasions, the hierarchy is determined (*Updating of matrix* $\mathbf{U}_c$);

c) Within each class c of occasions, the partition is determined (*Updating of matrix* $\mathbf{P}_c$).

Steps *b*) and *c)* are performed on the residuals of the data matrix from the previous step.

The generalization of the algorithm to the case where more than two classification structures are assumed in the model (1) is straightforward. It is simply carried out by adding supplementary steps where each of the constrained distance matrices is estimated on the residuals from the previous ones.

The algorithm fully described below has been implemented in MATLAB and it is available upon request to the authors.

Given initial values $\hat{\mathbf{V}}, \hat{\mathbf{M}}_c, \hat{a}_c, \hat{b}_c$ (c=1,..,C) the objective function of problem [P2']

$F(\mathbf{V}, \mathbf{M}_c, \mathbf{U}_c, a_c, b_c) =$

$$\sum_{k=1}^{K}\sum_{c=1}^{C} \|\mathbf{D}_k - [b_c(\mathbf{11}' - \mathbf{M}_c\mathbf{M}_c') + a_c(\mathbf{M}_c\mathbf{M}_c' - \mathbf{I})] - \mathbf{U}_c\|^2 \, w_k v_{kc}$$

can be minimized with respect to:

d) $\mathbf{U}_1,\ldots,\mathbf{U}_C$, given the current $\hat{\mathbf{V}}, \hat{\mathbf{M}}_c, \hat{a}_c, \hat{b}_c$ (c=1,..,C);

e) $\mathbf{M}_1,\ldots,\mathbf{M}_C$, given the current $\hat{\mathbf{V}}, \hat{a}_c, \hat{b}_c, \hat{\mathbf{U}}_c$ (c=1,…,C);

f) $a_1,\ldots,a_C, b_1,\ldots, b_C$, given the current $\hat{\mathbf{V}}$, $\hat{\mathbf{U}}_c, \hat{\mathbf{M}}_c$ (c=1,…,C);

g) $\mathbf{V}$, given $\hat{\mathbf{U}}_c, \hat{\mathbf{M}}_c, \hat{a}_c, \hat{b}_c$ (c=1,..,C).

The four steps are repeated in turn. At each step the objective function $F(\mathbf{V}, \mathbf{M}_c, \mathbf{U}_c, a_c, b_c, c=1,\ldots,C)$ does not increase and generally decreases. The process continues until it monotonically converges to a stationary point which turns out to be at least a local minimum of the problem [P2'], being F a function bounded below.

A more formal description of this sequential algorithm is given on the next page.

Let us now describe how the different sub-problems are solved.

Sub-problem a

In order to simplify the notation, let

$$\hat{\mathbf{P}}_c = \hat{b}_c(\mathbf{11}' - \hat{\mathbf{M}}_c\hat{\mathbf{M}}_c') + \hat{a}_c(\hat{\mathbf{M}}_c\hat{\mathbf{M}}_c' - \mathbf{I})$$

be the current estimate of matrix **P** and

$$\mathbf{R}_c = \sum_{k=1}^{K}\left[\mathbf{D}_k - \hat{\mathbf{P}}_c\right]\frac{w_k\hat{v}_{kc}}{\sum_{k=1}^{K} w_k\hat{v}_{kc}}$$

denote the mean residual dissimilarities from $\hat{\mathbf{P}}_c$.

Alternating Least-Squares Algorithm

Initialization Initial estimates of problem [P2']: $\hat{\mathbf{V}}$, $\hat{\mathbf{M}}_c$, $\hat{a}_c$, $\hat{b}_c$ (c=1,..,C) are given.

Step a Given the current estimates $\hat{\mathbf{V}}$, $\hat{\mathbf{M}}_c$, $\hat{a}_c$, $\hat{b}_c$ (c=1,..,C), new values $\mathbf{U}_c$ (c=1,..,C), are estimated by solving the following quadratic sub-problem

$$\min \sum_{k=1}^{K} \| \mathbf{D}_k - [\hat{b}_c(\mathbf{11'} - \hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'}) + \hat{a}_c(\hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'} - \mathbf{I})] - \mathbf{U}_c \|^2 \; w_k \hat{v}_{kc}$$

subject to

$\mathbf{U}_c \in \mathcal{U}$ c=1,..,C; [P2*a*]

Step b Given the current estimates $\hat{\mathbf{V}}$, $\hat{\mathbf{U}}_c$, $\hat{a}_c$, $\hat{b}_c$ (c=1,..,C), new values $\mathbf{M}_c$ (c=1,..,C) are estimated by solving the following quadratic sub-problem

$$\min \sum_{k=1}^{K} \| \mathbf{D}_k - [\hat{b}_c(\mathbf{11'} - \mathbf{M}_c\mathbf{M}_c^{'}) + \hat{a}_c(\mathbf{M}_c\mathbf{M}_c^{'} - \mathbf{I})] - \hat{\mathbf{U}}_c \|^2 \; w_k \hat{v}_{kc}$$

subject to

$m_{ilc} \in \{ 0, 1 \}$ i=1,…,I, l=1,…,C_c, c=1,...,C

$\sum_{l=1}^{C_c} m_{ilc} = 1$ i=1,…,I, c=1,...,C; [P2*b*]

Step c Given the current estimates $\hat{\mathbf{V}}$, $\hat{\mathbf{U}}_c$, $\hat{\mathbf{M}}_c$ (c=1,…,C), new values a_c and b_c are estimated by solving the following sub-problem

$$\min \sum_{k=1}^{K} \| \mathbf{D}_k - [b_c(\mathbf{11'} - \hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'}) + a_c(\hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'} - \mathbf{I})] - \hat{\mathbf{U}}_c \|^2 \; w_k \hat{v}_{kc}$$

subject to

$b_c \geq a_c > 0$ c=1,...,C; [P2*c*]

Step d Given the current estimates $\hat{\mathbf{U}}_c$, $\hat{\mathbf{M}}_c$, $\hat{a}_c$, $\hat{b}_c$ (c=1,…,C), a new $\mathbf{V}$ is estimated by solving the following sub-problem

$$\min \sum_{k=1}^{K} \| \mathbf{D}_k - [\hat{b}_c(\mathbf{11'} - \hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'}) + \hat{a}_c(\hat{\mathbf{M}}_c\hat{\mathbf{M}}_c^{'} - \mathbf{I})] - \hat{\mathbf{U}}_c \|^2 \; w_k v_{kc}$$

subject to

$v_{kc} \in \{0, 1\}$ k=1,…,K, c=1,...,C

$\sum_{c=1}^{C} v_{kc} = 1$ k=1,…,K. [P2*d*]

Stopping rule The objective function value is computed for the current values of $\hat{\mathbf{V}}, \hat{\mathbf{M}}_c, \hat{\mathbf{U}}_c, \hat{a}_c, \hat{b}_c,\ (c = 1,...,C)$. When the objective function has not decreased considerably with respect to a small convergence tolerance value, the process has converged. Otherwise, steps *a-d* are repeated in turn.

Note that the objective function of the problem [P2a] can be written

$$\sum_{k=1}^{K} \| \mathbf{D}_k - \hat{\mathbf{P}}_c - \mathbf{U}_c \|^2 w_k \hat{v}_{kc} \; = \; \left\| \mathbf{R}_c - \mathbf{U}_c \right\|^2 \left(\sum_{k=1}^{K} w_k \hat{v}_{kc} \right) + \tag{6}$$

$$\sum_{k=1}^{K} \left\| (\mathbf{D}_k - \hat{\mathbf{P}}_c) - \mathbf{R}_c \right\|^2 w_k \hat{v}_{kc} \;,$$

where only the first term of the right hand side depends on $\mathbf{U}_c$.

Thus, expanding the expression of $\mathbf{R}_c$, it remains to optimize the following problem equivalent to [P2a]

$$\text{Min} \left\| \sum_{k=1}^{K} \left[\mathbf{D}_k - [\hat{b}_c(\mathbf{11}' - \hat{\mathbf{M}}_c \hat{\mathbf{M}}_c') + \hat{a}_c(\hat{\mathbf{M}}_c \hat{\mathbf{M}}_c' - \mathbf{I})] \right] \frac{w_k \hat{v}_{kc}}{\sum_{k=1}^{K} w_k \hat{v}_{kc}} - \mathbf{U}_c \right\|^2 \quad [\text{P2}a']$$

subject to

$\mathbf{U}_c \in \mathcal{U}$ $\qquad\qquad$ c=1,..,C,

which can be solved through the algorithm by DeSoete (1984). However, this algorithm becomes computationally unfeasible when the number of objects to be classified is large (more than 100). Thus, for larger problems an alternative solution for [P2a’] has to be found and it is provided as follows.

Let $\mathbf{U}_c$ be written: $\mathbf{U}_c = \overline{\mathbf{U}}_c + h(\mathbf{11}' - \mathbf{I})$, where $\overline{\mathbf{U}}_c$ is a *pseudo*-ultrametric matrix that can have negative off-diagonal entries still satisfying the ultrametric inequalities and h is a non-negative constant chosen to guarantee the non-negativity of $\mathbf{U}_c$.

Therefore, the objective function of [P2a’] can be written

$$\left\| \mathbf{R}_c - \mathbf{U}_c \right\|^2 = \left\| \mathbf{R}_c - \overline{\mathbf{U}}_c \right\|^2 + h^2 I(I-1) - 2h\mathbf{1}'(\mathbf{R}_c - \overline{\mathbf{U}}_c)\mathbf{1} \tag{7}$$

where $\mathbf{1}$ is a vector of I ones and the terms involving $tr(\overline{\mathbf{U}}_c)$ and $tr(\mathbf{R}_c)$ vanish, because the diagonal values of all the dissimilarity and distance matrices are zero.

If $\left\| \mathbf{R}_c - \overline{\mathbf{U}}_c \right\|^2$ is minimized with respect to $\overline{\mathbf{U}}_c$ by using the group average link clustering (UPGMA), the last term of (7) vanishes because the sum of the residual dissimilarities in $\mathbf{R}_c$ equals the sum of the fitted ultrametric values (the UPGMA is based on the average values of dissimilarities between the two clusters being merged).

Therefore, without loss of generality, problem [P2*a*'] can be equivalently solved by carrying out the UPGMA clustering on matrix $\mathbf{R}_c$, so obtaining the optimal pseudo-ultrametric $\hat{\overline{\mathbf{U}}}_c$. Finally, the optimal $\hat{\mathbf{U}}_c$ is obtained by adding the minimum non-negative constant to the off-diagonal entries of $\hat{\overline{\mathbf{U}}}_c$, so that the non-negativity constraint on $\hat{\mathbf{U}}_c$ is still satisfied.

Since [P2 *a*] has a non-convex feasible region (ultrametric cone) and it is known to be an NP-hard classification problem, the global minimum solution for [P2] or [P2'] cannot be always guaranteed and the convergent sequence of ALS can be broken by a local minimum for [P2*a*]. This problem is overcome by retaining for [P2*a*] only solutions where the objective function does not increase.

Sub-problem b

The objective function of [P2*b*] can be written as above for the problem [P2*a*] in terms of mean residual dissimilarities from $\hat{\mathbf{U}}$ (say $\mathbf{R}_c^*$) and solved with respect to $\mathbf{M}_c$ (c=1,…,C)

$$F(\mathbf{M}_c; \hat{\mathbf{V}}, \hat{\mathbf{U}}_c, \hat{a}_c, \hat{b}_c) = \left\| \mathbf{R}_c^* - [\hat{b}_c(\mathbf{11}' - \mathbf{M}_c\mathbf{M}_c') + \hat{a}_c(\mathbf{M}_c\mathbf{M}_c' - \mathbf{I})] \right\|^2 .$$

Thus, problem [P2*b*] can be equivalently written as follows

$$\min F(\mathbf{M}_c ; \hat{\mathbf{V}}, \hat{\mathbf{U}}_c, \hat{a}_c, \hat{b}_c) \qquad \text{[P2}b\text{']}$$

subject to

$$m_{ilc} \in \{0, 1\} \qquad i=1,\dots,I, \quad l=1,\dots,C_c, \quad c=1,\dots,C$$

$$\sum_{l=1}^{C_c} m_{ilc} = 1 \qquad i=1,\dots,I \quad c=1,\dots,C.$$

This problem can be solved sequentially for the i^{th} row (i=1,…,I) of $\mathbf{M}_c$, by setting

$$m_{ilc}=1 \qquad \text{if } F([m_{ilc}], \hat{\mathbf{V}}, \hat{\mathbf{U}}_c, \hat{a}_c, \hat{b}_c)$$

$$= \min\{F([m_{itc}], \hat{\mathbf{V}}, \hat{\mathbf{U}}_c, \hat{a}_c, \hat{b}_c) : t=1,\dots,C_c\}$$

$$m_{ilc}=0 \qquad \text{otherwise.}$$

It has to be observed that in this way it is not guaranteed to find the optimal solution of the NP-hard problem [P2*b*']. However, the objective function never decreases, thus maintaining the monotonicity property of the algorithm.

Sub-problem c

Problem [P2*c*], rewritten as problem [P2*b*’] and solved with respect to a_c and b_c, simply reduces to an ordinary linear regression problem with two predictors. The least-squares estimators in closed form are

$$\hat{a}_c = \frac{1}{\sum_{i_c=1}^{C_c} I_{i_c}^2 - I} \sum_{k=1}^{K} trace((\mathbf{D}_k - \hat{\mathbf{U}}_c)\hat{\mathbf{M}}_c\hat{\mathbf{M}}_c') \frac{w_k \hat{v}_{kc}}{\sum_{k=1}^{K} w_k \hat{v}_{kc}} \qquad c=1,...,C$$

$$\hat{b}_c = \frac{1}{I^2 - \sum_{i_c=1}^{C_c} I_{i_c}^2} \sum_{k=1}^{K} trace((\mathbf{D}_k - \hat{\mathbf{U}}_c)(\mathbf{11}' - \hat{\mathbf{M}}_c\hat{\mathbf{M}}_c')) \frac{w_k \hat{v}_{kc}}{\sum_{c=1}^{K} w_k \hat{v}_{kc}}$$

$$c=1,...,C$$

where I_{i_c} denotes the size of the group i_c of the c^{th} consensus partition into C_c groups.

It is worth noticing that the estimates $\hat{a}_c$ and $\hat{b}_c$, given the current estimates of the other parameters, represent, respectively, the average of the *within*-class and *between*-class residual dissimilarities from $\mathbf{U}_c$ related to the partition identified by $\mathbf{M}_c$. When the partition is well-defined, (i.e. with clusters compact and separated) the constraint $\hat{a}_c \leq \hat{b}_c$ holds. Therefore, it is not necessary to impose this constraint because it is sufficient to start from a feasible partition where $\hat{a}_c \leq \hat{b}_c$, i.e. the average of the within-class dissimilarities is not greater than the average of the between-class dissimilarities. In this way, since in the following step *b*) a partition is defined by the new $\hat{\mathbf{M}}_c$ minimizing the objective function, then at the next updating of the levels of fusion, the new values $\hat{a}_c$ and $\hat{b}_c$ will still reflect the compactness and separation of the groups and consequently the constraint $\hat{a}_c \leq \hat{b}_c$ (c=1,…,C).

Since even in this case the non-negativity constraint is not imposed on a_c, an appropriate constant can be finally added at the optimal solution $\hat{\mathbf{P}}_c$, if necessary, as in problem [P2*a*’].

Sub-problem d

Problem [P2*d*] can be solved as an assignment problem for each dissimilarity matrix by assigning $\mathbf{D}_k$ (k=1,..,K) to the class $\mathcal{G}_c$ where

v_{kc}=1 if $\| \mathbf{D}_k - \hat{\mathbf{P}}_c - \hat{\mathbf{U}}_c \|^2 w_k = \min\left\{ \| \mathbf{D}_k - \hat{\mathbf{P}}_l - \hat{\mathbf{U}}_l \|^2 w_k : l = 1,...,C \right\}$
v_{kc}=0 otherwise.

It has to be noted that, since this assignment problem is solved sequentially for the different rows of **V**, it is not guaranteed to find the global optimal solution, but the monotonicity property of the algorithm is assured.

The final feasible (non-negative) solutions for $\hat{\mathbf{U}}_c$ and $\hat{\mathbf{P}}_c$ (c=1,...,C) can be obtained from the optimal fitted matrices $\hat{\bar{\mathbf{U}}}_c$ and $\hat{\bar{\mathbf{P}}}_c$ as explained in detail above in the sub-problems *a* and *b* by generally increasing the objective value. Note that when only one of the two fitted matrices has some negative entries, infinite solutions can be obtained, giving the same value of the objective function. In this case, without loss of generality, let $\hat{\bar{\mathbf{U}}}_c$ be the optimal pseudo–ultrametric matrix, while $\hat{\bar{\mathbf{P}}}_c$ be the optimal 2-ultrametric matrix with positive entries for some class c of the secondary partition. Furthermore, let -h<0 and $\hat{a}_c$>0 be the minimum off-diagonal entries of $\hat{\bar{\mathbf{U}}}_c$ and $\hat{\bar{\mathbf{P}}}_c$, respectively. When $h \leq \hat{a}_c$, the optimal objective function value

$$F^* = \sum_{c=1}^{C}\sum_{k=1}^{K} \| \mathbf{D}_k - \hat{\bar{\mathbf{P}}}_c - \hat{\bar{\mathbf{U}}}_c \|^2 \, w_k \hat{v}_{kc}$$

does not change if we consider the ultrametric and 2-ultrametric matrices with non-negative entries,

$$\hat{\mathbf{U}}_c = \hat{\bar{\mathbf{U}}}_c + \alpha(\mathbf{11}' - \mathbf{I}) \qquad \text{and} \qquad \hat{\mathbf{P}}_c = \hat{\bar{\mathbf{P}}}_c - \alpha(\mathbf{11}' - \mathbf{I})$$

for any α such that $h \leq \alpha \leq \hat{a}_c$, determining infinite solutions corresponding to the infinite values of α.

This indeterminacy of the solution can be fruitfully exploited by choosing the value for α yielding the matrices $\hat{\mathbf{U}}_c$ and $\hat{\mathbf{P}}_c$ which together best account for portions of the original dissimilarities

$$\max_{\alpha} f(\alpha) = \max_{\alpha} \left(\frac{\| \hat{\mathbf{P}}_c(\alpha) \|^2 \sum_{k=1}^{K} w_k \hat{v}_{kc}}{\sum_{k=1}^{K} \| \mathbf{D}_k \|^2 \, w_k \hat{v}_{kc}} + \frac{\| \hat{\mathbf{U}}_c(\alpha) \|^2 \sum_{k=1}^{K} w_k \hat{v}_{kc}}{\sum_{k=1}^{K} \| \mathbf{D}_k \|^2 \, w_k \hat{v}_{kc}} \right), \qquad (8)$$

subject to $h \leq \alpha \leq \hat{a}_c$.

The maximum is attained for $\hat{\alpha} = \hat{a}_c$ or $\hat{\alpha} = h$, being the function (8) a parabola in α limited from above by its constrained maximum which corresponds to $f(h)$ or $f(\hat{a}_c)$.

5. Applications

5.1 Analysis of Zoological Data

The well-known Richard Forsyth's (artificial) zoological dataset (UCI repository of machine learning databases, Asuncion and Newman 2007) was analyzed by fitting model (2) in the relevant case of a two-way data matrix (K=1). This illustrative application aims to show the relevance of fitting multiple classification structures to a single proximity matrix, without considering for a moment about multiple proximity matrices. This issue will be considered in the second application.

The dataset includes 101 animals characterized by 15 Boolean attributes, measuring some peculiar characteristics in terms of presence/absence (hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize) and one numeric-valued variable for the number of legs (set of values: {0,2,4,5,6,8}). The latter has been transformed into 5 binary variables of presence/absence of 2, 4, 5, 6, 8 legs, respectively, to avoid a different weight in determining the classification.

A supplementary variable not used in the analysis is also available in the database, denoting which of 7 different classes each animal belongs to, corresponding to 7 well-known classes (1=mammals, 2=birds, 3= reptiles, 4=fishes, 5=amphibians, 6=insects, 7=molluscs and arthropods). This variable was used to better interpret the results of the analysis.

Such a dataset is appropriate to illustrate our structural classification analysis. In fact, different taxonomies (both hierarchical and non-hierarchical) coexist based on different characteristics of the animals.

The zoology data set contributed by Richard Forsyth has been used extensively (see UCI website for papers citing this data set), usually in a classification or pattern recognition context even if it is acknowledged that we are not aware of any correct structure (Koivisto and Sood 2004), probably because the variables included are not sufficient to exclusively classify the animals. So, the 7 classes given are assumed to be the clustering by biological experts.

McKenzie and Forsyth (1995) who firstly analyzed the data set note that none of the methods were able to classify the reptiles correctly. Reptiles tended to be misclassified as amphibians and fishes.

Wang, Chaudhari and Patra (2004) analyzed the data set by an unsupervised algorithm after transforming the numeric variable related to the number of legs into binary variables (as we did). They note that, for example, *porpoise* and *dolphin* are clustered with terrestrial animals by biology experts, but with other oceanic animals by their method due to the value of attributes (hair, feathers, aquatic, fins, tail, leg, etc). The reptiles *Pitviper, slowworm, tuatara* are mixed up with some amphibians *toad, newt, frog* also because of their attributes.

We firstly analyzed the data set by standard clustering methods.

In the solution of *k*-means with *c*=6 clusters, in the largest group of mammals, *porpoise* and *dolphin* are not included but they are clustered together with one reptile *(seasnake*) and all the fishes. Reptiles and Amphibians are put together in the same group.

In the average linkage solution, different criteria seem to lead the hierarchy: the presence of milk, backbone and eggs, but their effects are not completely separate. In fact at level of two clusters we can find the gross division between mammals and non-mammals, and at level of three clusters the non mammals are divided in vertebrates and invertebrates. Then birds are split from the rest of oviparous animals.

The algorithm proposed here was run on the matrix of the squared Euclidean distances computed from the 20 binary variables.

To investigate the choice of the number of groups of the (primary) partition, the algorithm was run by setting the maximum number of groups equal to 8 (which is greater than the number of known classes of the supplementary variable) and retaining the best solution in terms of objective function value, obtained from 3000 random starts of the algorithm. Since the best solution was found having 2 empty groups, we analyzed as effective solution the one with *C*=6 non-empty groups.

For the optimal fitted matrices $\hat{\mathbf{P}}$ and $\hat{\mathbf{U}}$, the proportions of the sum of squares of the original dissimilarities accounted for are computed:

$$\frac{\|\hat{\mathbf{P}}\|^2}{\|\mathbf{D}\|^2}=0.8709\ , \quad \frac{\|\hat{\mathbf{U}}\|^2}{\|\mathbf{D}\|^2}=0.1188, \quad 1-\frac{\|\mathbf{D}-\hat{\mathbf{P}}-\hat{\mathbf{U}}\|^2}{\|\mathbf{D}\|^2}=0.9699.$$

In order to help the reader in the interpretation of the results, we have evaluated which of the 20 binary variables were more discriminating by a pseudo-*F* index:

$$F_j=\frac{\mathbf{x}_j\,'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}\,'\mathbf{x}_j}{\mathbf{x}_j\,'\mathbf{x}_j-\mathbf{x}_j\,'\mathbf{B}(\mathbf{B}\,'\mathbf{B})^{-1}\mathbf{B}\,'\mathbf{x}_j}\,\frac{I-C-1}{C} \qquad (j=1,...,20)$$

where $\mathbf{x}_j$ is the vector of the *j*-th variable and **B** is generally a binary membership matrix defining a partition. In this case, when considering the resulting optimal partition $\mathbf{B}=\hat{\mathbf{M}}$, otherwise, in evaluating the optimal hierarchy, **B** is the membership matrix corresponding to the partition

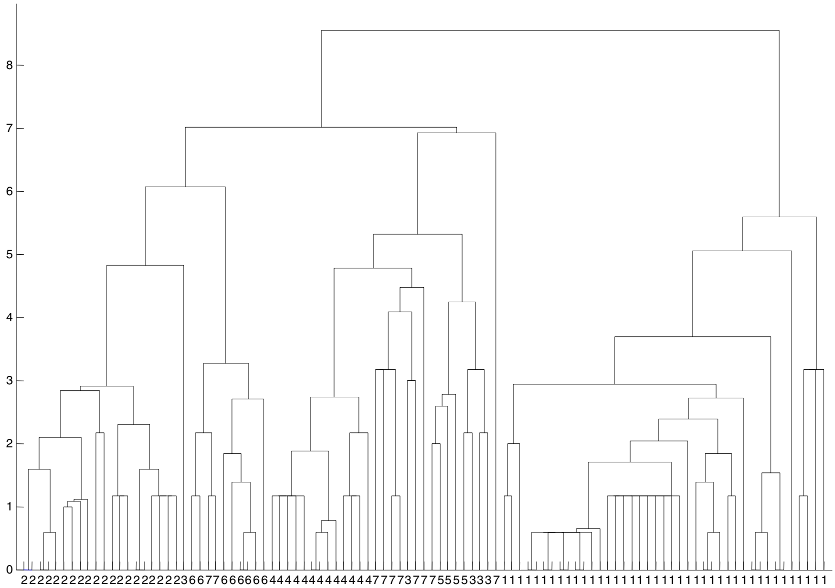


Figure 1. Optimal hierarchy of the Zoological data (objects are labeled by the categories of the supplementary variable).

obtained by cutting the dendrogram at a given height. The pseudo-*F* index accounts for the between-to-within ratio variability due to the partition and allows to detect which variables are more discriminant in terms of both separation between groups and cohesion within groups.

Figure 1 displays the dendrogram of the optimal hierarchy resulting from the algorithm.

By cutting the dendrogram at level of two classes emerges the distinction between the 41 mammals and the remaining 60 animals. In fact, the Pseudo-*F* index detects “milk” as the most relevant variable.

The partition into 3 groups (Table 1a) is detected by splitting the largest class of 60 animals into 2 classes: 31 oviparous breathed non-toothed animals (including birds, insects, one reptile {*tortoise*}, and two terrestrial molluscs {*slug*, *worm*}) and 29 mostly oviparous non-breathed toothed animals (including all fishes, amphibians, reptiles and most of the molluscs and arthropods), being such partition mainly discriminated by the variables “breathes” and “eggs”.

The partition into 4 classes separates the {*scorpion*} from the class of 29 animals above: in fact, it is the only non-oviparous (but “milk absent”) animal.

Table 1. Zoological data.

Table 1a: Partitions in 3, 4 and 5 classes derived from the optimal hierarchy.

Partition in 3 classes from the optimal hierarchy								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*Size*
Class 1	41	0	0	0	0	0	0	41
Class 2	0	20	1	0	0	8	2	31
Class 3	0	0	4	13	4	0	8	29

Partition in 4 classes from the optimal hierarchy by splitting Class 3								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*size*
Class 3a	0	0	4	13	4	0	7	28
Class 3b	0	0	0	0	0	0	1	1

Partition in 5 classes from the optimal hierarchy by splitting Class 2								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*size*
Class 2a	0	20	1	0	0	0	0	21
Class 2b	0	0	0	0	0	8	2	10

Partition in 6 classes from the optimal hierarchy by splitting Class 1								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*size*
Class 1a	37	0	0	0	0	0	0	37
Class 1b	4	0	0	0	0	4	0	4

Partition in 7 classes from the optimal hierarchy by splitting Class 3a								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*size*
Class 3a1	0	0	3	0	4	0	1	8
Class 3b2	0	0	1	13	0	0	6	20

Table 1b: Optimal (primary) partition in 6 classes

Optimal primary partition in 6 classes								
	Mammals	*Birds*	*Reptiles*	*Fishes*	*Amphibian*	*Insects*	*Molluscs et al.*	*Class*
	1	2	3	4	5	6	7	*size*
Class 1	6	0	0	1	0	0	0	7
Class 2	2	0	0	0	0	0	0	2
Class 3	31	4	5	12	4	0	0	56
Class 4	2	16	0	0	0	0	0	18
Class 5	0	0	0	0	0	6	10	16
Class 6	0	0	0	0	0	2	0	2
	Backbone	*Tail*	*Toothed*	*Airborne*	*Domestic*			*Class*
								size
Class 1	7	6	7	0	7			7
Class 2	2	0	2	0	1			2
Class 3	56	50	50	0	1			56
Class 4	18	18	2	18	3			18
Class 5	0	1	0	4	0			16
Class 6	0	0	0	2	1			2

Going down along the branches of the dendrogram at the level of 5 groups the new two classes derive by splitting the class of 31 animals into the class of birds (plus {*tortoise*}) and the class of insects (plus {*slug*, *worm*}); here the separation is accounted for by the presence/absence of both feathers and backbone as pointed out by the Pseudo-*F* index, too.

Actually, the hierarchy reflects the well-known taxonomy of the animals in agreement with the supplementary information in the dataset and provides nested partitions based on the zoological evolutionary tree, well interpretable with respect to some discriminating binary variables.

The analysis of the optimal partition into 6 classes (Table 1b) reflects mainly the backbone-related classification of the animals into vertebrates and invertebrates with the structural features specific of these classes:

Class 1 contains 7 tailed (except one) toothed non-flying vertebrates;
Class 2 includes 2 (non-tailed) primates;
Class 3 is formed by 56 mostly tailed, mostly toothed non-flying vertebrates;
Class 4 has 18 tailed non-toothed flying vertebrates;
Class 5 includes 16 non-tailed (except one), mostly non-flying invertebrates
Class 6 is formed by 2 flying invertebrates.

Such a partition is well characterized by several features of the animals (presence/absence of backbone, tail, teeth, wings) as pointed out by the Pseudo-*F* index.

It is important to note that such a taxonomy is based on features of the animals which are different from the ones underlying the hierarchy.

5.2 Analysis of the Kinship Terms Data

As an application of the structural classification analysis of three-way dissimilarity data the well-known data set analyzed by Rosenberg and Kim (1975) has been used, where 85 students were asked to sort the following 15 kinship terms into categories “on the basis of some aspect of meaning”: grandfather (GrF), grandmother (GrM), grandson (GrS), granddaughter (GrD), brother (Bro), sister (Sis), father (Fat), mother (Mot), son (Son), daughter (Dau), nephew (Nep), niece (Nie), uncle (Unc), aunt (Aun) and cousin (Cou).

The students actually described only 39 different partitions of the 15 kinship terms, 11 of such partitions being provided more than once. Therefore the frequency of each of these K=39 different “categories” has been used as weight w_k (k=1,…,39).

For each of the 39 categories a dissimilarity matrix was defined by considering a binary matrix $\mathbf{D}_k$ (k=1,…,39) where d_{ijk}=1 (respectively, 0) if

o_i and o_j do not belong (respectively belong) to the same observed class of kinship terms.

5.2.1 Fitting a Single Partition and a Single Indexed Hierarchy to the Kinship Terms Data

Model (2) was fitted to the Kinship Terms Data to find both a partition and an indexed hierarchy able to describe the taxonomy present in the data. As far as the choice of the partition is concerned, the algorithm was run by varying the number of groups from 2 to 7 and the best solution was retained over 100 random starts of the algorithm. The minimum value of the objective function was achieved by requiring 6 groups but, since 3 of them were empty, actually the proper groups of the partition were the remaining three (the corresponding objective function value divided by the squared Euclidean norm of matrix **D** is equal to 0.1321).

The optimal fitted matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{P}}$ are obtained according to (8) by considering the solutions which best account for the sum of the portions of the original dissimilarities:

$$\frac{\|\hat{\mathbf{P}}\|^2 \sum_{k=1}^{K} w_k}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.0059 \quad \text{and} \quad \frac{\|\hat{\mathbf{U}}\|^2 \sum_{k=1}^{K} w_k}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.7678 \ ,$$

determining a total fit of

$$1 - \frac{\sum_{k=1}^{K} \|\mathbf{D}_k - \hat{\mathbf{P}} - \hat{\mathbf{U}}\|^2 w_k}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.8679 \ .$$

In Figures 2(a)[2] and 2(b) the consensus partition (represented as a 2-dendrogram) and the consensus hierarchy are displayed, respectively. They represent the best (least-squares) reconstruction of the dissimilarity data by two order-constrained matrices, i.e., a 2-ultrametric and an ultrametric matrix.

The partition (Figure 2(a)) identifies two gender-related classes, plus a singleton cluster formed by the gender-ambiguous term "cousin",

Male terms	(GrF, GrS, Bro, Fat, Son, Nep, Unc),
Female terms	(GrM, GrD, Sis, Mot, Dau, Nie, Aun),
Neutral term	(Cou).

[2] To enhance the clarity of the graphical display, the off-diagonal elements of the matrix **P** have been augmented of 0.01 to have the first level of fusion slightly greater than 0.

The dendrogram in Figure 2(b) detects, at a high level of fusion 3 classes:

Nuclear family terms	(Mot, Fat, Bro, Sis, Son, Dau);
Grands (ie, ± 2 generations from ego)	(GrF, GrM, GrS, GrD);
Collateral kinship terms	(Aun, Cou, Nie, Nep, Unc),

but at a low level is also evident a gender effect due to pairs of terms denoting equivalent relationship, but opposite gender (e.g., Mot/Fat).

It can be observed that the interpretation of the results derives directly from the graphical representations (Figure 2).

In the papers by Gordon and Vichi (1998, 2001) the same data set was also analyzed to detect hard and fuzzy consensus partitions of the kinship terms, respectively. In the first analysis, the solution found in case of only one class of subjects was: (GrF, GrM), (GrS, GrD), (Bro, Sis), (Fat, Mot, Son, Dau), (Nep, Nie), (Unc, Aun, Cou), where both the gender effect and the closeness of the relationship are present, but not completely distinguishable.

In the second paper, where a fuzzy consensus partition was fitted to the same data set, the hard partition, which the consensus partition is closest to, comprises the same three classes: "grands", "nuclear family" and "collateral relatives", as here in the dendrogram of Figure 2(b). But, in that paper it was also noted that "…the methods of analysis also show a "gender" effect essentially because differences in the membership functions were found for terms related to different gender".

Hubert and Arabie (1994) also analyzed the same data set by fitting two models which reconstruct the dissimilarities as approximate sums of two Robinson and two circular Robinson matrices, respectively. By cutting the two hierarchical structures they find overlapping clusters mostly similar with the ones we obtain, showing the complexity of the data set and the necessity to fit more than one structure to deeply understand the nature of the data. Their resulting structures are quite difficult to be interpreted but highlight the different ways of judgements of the subjects which can be better analyzed only when the subjects are partitioned, too.

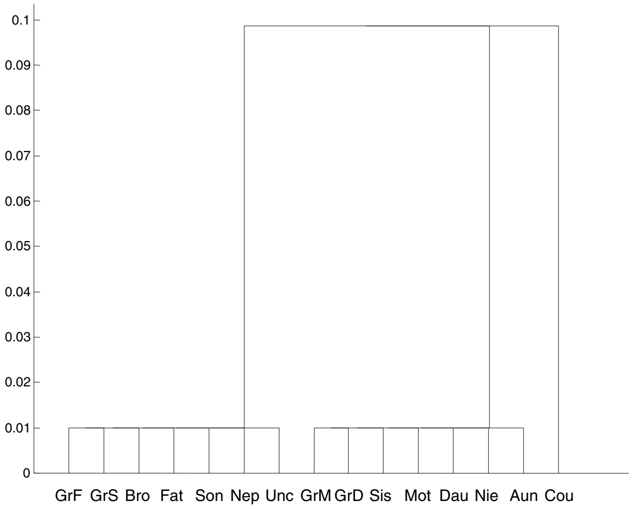
Here, the proposed methodology manages to split the two effects (gender and degree of kinship) through the two classification structures, emphasizing different aspects present in the data set.

5.2.2. Secondary Partition, Consensus Partitions and Indexed Hierarchies for the Kinship Terms Data

In the Kinship Terms Data several authors have observed that different criteria for classifying terms have been defined by the 85 subjects.

Carroll and Arabie (1983), fitting INDCLUS model to the three-way data, detected that more than one criterion was being used and summarized

a) Optimal partition



b) Optimal hierarchy

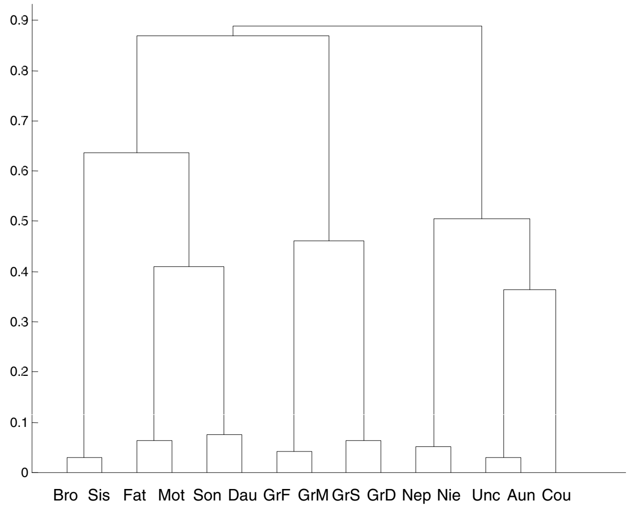


Figure 2. Rosenberg and Kim Kinship Terms Data.

relationships between kinship terms implied by the different criteria. However this work did not directly identify which subjects were using each criterion. This information is provided by the methodology proposed here where the most relevant classification structures are found.

Problem [P2'] has been solved by using the ALS algorithm illustrated in the previous Section and choosing C=3 classes for the secondary partition of the set $\{\mathbf{D}_1,\ldots,\mathbf{D}_{39}\}$. This choice is consistent with the solution given in Gordon and Vichi (1998, 2001), because consensus classifications (partitions and hierarchies) in a larger number of classes displayed similar patterns. To investigate the choice of the numbers of groups of the primary partitions, at first the algorithm was run 300 times by requiring 8 groups within each class of occasions, being 8 the maximum number of groups indicated by the students. Since the best solution obtained had empty groups in some classes of the secondary partition (the non-empty groups were 3, 7 and 8, respectively), the algorithm was run again with different numbers of groups by taking into account these results and the ones obtained by Gordon and Vichi (1998, 2001). The algorithm was run 100 times for each different choice of the number of groups (C_c=3,5,8, C_c=3,4,8, C_c=3,5,4, C_c=3,5,3) and the best solution in terms of objective function was retained which corresponds to C_c =3,5,3 with the objective function value divided by the square Euclidean norm of matrix **D** equal to 0.0741.

The optimal fitted matrices $\hat{\mathbf{U}}_c$ and $\hat{\mathbf{P}}_c$ (c=1,2,3) are obtained according to (8) by considering the solutions which best account for the sum of the portions of the original dissimilarities:

$$\frac{\sum_{c=1}^{C} \| \hat{\mathbf{P}}_c \|^2 \sum_{k=1}^{K} w_k \hat{v}_{kc}}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.0283 \text{ and } \frac{\sum_{c=1}^{C} \|\mathbf{U}_c\|^2 \sum_{k=1}^{K} w_k \hat{v}_{kc}}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.6829\,,$$

determining a total fit of $1-\frac{\sum_{c=1}^{C}\sum_{k=1}^{K} \| \mathbf{D}_k - \hat{\mathbf{P}}_c - \hat{\mathbf{U}}_c \|^2 w_k \hat{v}_{kc}}{\sum_{k=1}^{K} \|\mathbf{D}_k\|^2 w_k} = 0.9259$.

The assignment of the 39 different original partitions to the three classes as obtained by the best solution is reported in Table 2.

It has to be noted that the algorithm frequently stops at local optima and to enhance the probability to obtain the optimal solution, the analysis

Table 2. Assignments of the 85 subjects into three classes whose consensus hierarchies and partitions are given in Figure 3.

Class	Partition(frequency>1)	Number of subjects
I	11, 12, 14, 15, 16, 17, 18(2), 19, 21(2), 22, 23, 24, 25(9), 26(7), 27, 30, 31, 32, 33, 37, 39	37
II	1, 2, 3(12), 4, 5, 6, 7, 8(2), 9(5), 10(5), 13(5), 20(2), 28, 29, 35, 36, 38	42
III	34(6)	6

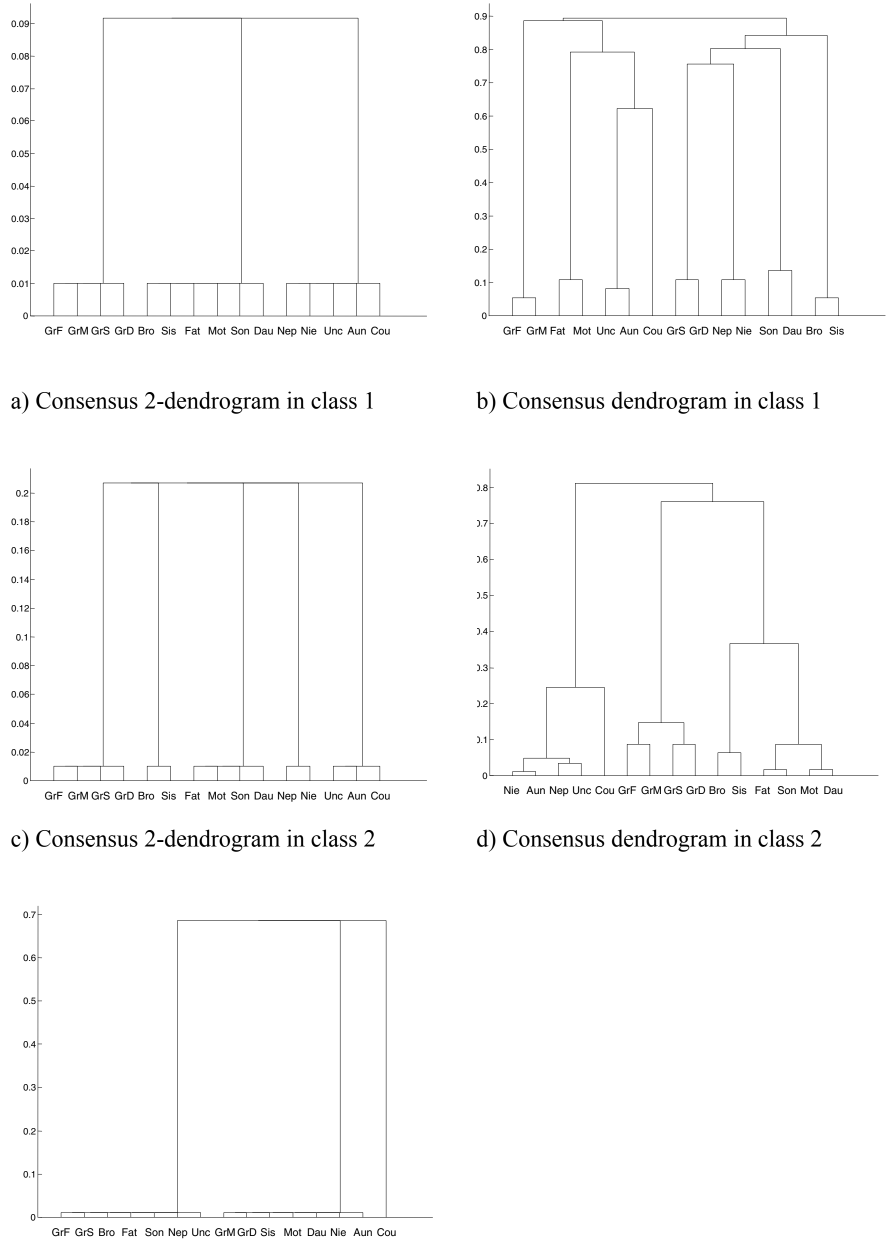
was repeated starting from different initial random solutions. In practice a detailed study was carried out to investigate the possibility that the solution was locally rather than globally optimal. The algorithm was run from other 100 different randomly generated values of **V** and it still converged to the same solution (the starting values for $\mathbf{M}_c$ (c=1,2,3) were computed as the solutions of the *k*-means algorithm within each class of the random secondary partition).

The partition in Table 2 coincides with the secondary partition provided by the algorithm proposed by Gordon and Vichi (1998) where only one consensus partition for each class of the secondary partition was fitted.

Figure 3(a)[3] displays the 2-dendrogram corresponding to the consensus partition into 3 groups of the kinship terms within the *first* class of the secondary partition. It refers to the same partition of the terms in Nuclear family terms (Mot, Fat, Bro, Sis, Son, Dau), Grands (GrF, GrM, GrS, GrD) and Collateral kinship terms (Aun, Cou, Nie, Nep, Unc), already found for C=1 in Section 4. But examining the original 37 partitions provided by the students belonging to this class, it can be observed that 9 of them identify exactly the partition of Figure 3(a) and the remaining students considered different criteria to cluster the terms. This information is subsumed in the dendrogram of Figure 3(b), which depicts the optimal consensus hierarchy fitted in the first class of the secondary partition. At a low level of fusion a partition corresponding to a strong gender effect due to opposite parallel terms (gender-dyads) is shown. At a higher level of fusion such gender-dyads are grouped together according to a gender-related criterion related to generation (senior and junior relatives).

Gender seems the most relevant criterion that induced a few of the 85 students to classify the kinship terms. However, as shown further, this is not the only one considered by all the subjects.

[3] To enhance the clarity of the graphical displays of Figure 3, the off-diagonal elements of the ultrametric and 2-ultrametric matrices have been augmented of 0.01, to have the first levels of fusions slightly greater than 0.



a) Consensus 2-dendrogram in class 1

b) Consensus dendrogram in class 1

c) Consensus 2-dendrogram in class 2

d) Consensus dendrogram in class 2

e) Consensus 2-dendrogram in class 3

Figure 3. Rosenberg and Kim kinship terms data: consensus dendrograms associated to the 3 classes of the secondary partition of the dissimilarity matrices $\{\mathbf{D}_1,\ldots,\mathbf{D}_{39}\}$.

The dendrogram in Figure 3(b) has at low levels of fusion the same topology of the consensus dendrogram (found for C=1) in Figure 2(b), but the higher part of the tree reveals a different pattern followed by the students in partitioning the terms.

Note also that the consensus partition of the 2-dendrogram in Figure 3(a) is dissimilar to any other partition present in the dendrogram in Figure 3(b).

Figures 3(c) and 3(d) refer to the *second* class of the secondary partition which includes 42 students, but only 17 different partitions provided by them. The 2-dendrogram in Figure 3(c) shows again a classification based on the relationship (Grands, Core Nuclear family, Core Collateral relatives) with the difference that two parallel-gender groups are also indicated: (Bro, Sis) and (Nep, Nie) that these 42 students perceived distinct from the whole Nuclear and Collateral groups, respectively. The dendrogram in Figure 3(d), which depicts the optimal hierarchy in the second class of subjects, mainly identifies dyads different from those in Figure 3(b), because they are formed by the terms that within the same gender reflect a type of reciprocity where one type of kin must exist to define the second (e.g., Aun/Nie; Fat/Son).

The *third* class of dissimilarity matrices is a singleton formed by matrix $\mathbf{D}_{34}$, which has frequency 6. Since in this application each dissimilarity matrix defines a partition by construction (each student was asked to provide a partition), in this case $\mathbf{D}_{34} \in_p \mathcal{U}$ (cone of the 2-ultrametrics) and the consensus hierarchy collapses into a partition defined by matrix $\mathbf{D}_{34}$ itself. Such partition is reported as a 2-dendrogram in Figure 3(e). In this case the consensus partition and hierarchy are coincident and it is sufficient to show only a unique consensus classification.

Figure 3(e) shows the same gender-related consensus partition found for C=1 in Section 4, discriminating between female and male kinship terms, and leaving alone the gender ambiguous term “Cousin”.

The results confirm the presence of underlying dimensions (Meulman and Heiser 2004) related to gender, generation and degree of separation of each of the terms but it is more evident here how these aspects play roles within each class of subjects.

6. A Simulation Study

A Monte Carlo experiment was performed to test how well the algorithm for the structural classification analysis of three-way dissimilarity data performs. A number of data sets (K = 20,50) of dissimilarities pertaining to I=20 objects were generated as in model (4) by setting C=2. Each occasion was drawn from one of two different multivariate normal distributions, according to random binary membership matrices (matrix $\mathbf{V}$).

Such two multinormal distributions had mean vectors φ_c (c=1,2) and covariance matrices $h\mathrm{diag}(\varphi_c)^2$, ($h$=0.5, 1, 1.5, 2) where the constant h allowed to set different error levels in terms of the coefficients of variation of each dimension. The mean vector φ_g was set by half vectorizing (i.e. by vectorizing only the lower triangular part of) matrix $\mathbf{\Phi}_c = \mathbf{P}_c + \mathbf{U}_c$ (c=1,2) where $\mathbf{P}_c$ and $\mathbf{U}_c$ (c=1,2) were generated as follows. In order to obtain well separated groups of objects in defining $\mathbf{P}_c$, a data matrix was randomly generated following the procedure proposed by Milligan and Cooper (1985). The membership matrix $\mathbf{M}_c$ and two levels a_c and b_c forming $\mathbf{P}_c$ were derived by setting C_1=2 and C_2=3 groups of objects for c=1, 2, respectively[4].

The same generating process of Milligan and Cooper was used to have a data matrix with well-separated nested groups of objects on which the group average link clustering was applied providing the ultrametric matrix $\mathbf{U}_c$.

The model was fitted to each dissimilarity data set, by using as input the true underlying number of classes (C=2) for the secondary partition and the true numbers of groups (C_1=2 and C_2=3, respectively) for the consensus primary partitions of objects.

In each analysis, we chose as initialization for the binary membership matrix of the secondary partition $\mathbf{V}$, the solution of the k-means algorithm (McQueen 1967), applied to the (KXI(I-1)/2) matrix whose rows are the vectors containing the generated dissimilarities The starting values for $\mathbf{M}_c$ (c=1,2) were computed as the solutions of the k-means algorithm within each class of the random starting secondary partition. Consequently, the computation of the starting values for a_c and b_c (c=1,2) follow.

The performance of the method has been evaluated by using the following measures.

- $MRand(\mathbf{V}, \hat{\mathbf{V}})$: Modified Rand Index (Hubert and Arabie 1985) between true and fitted membership matrices of the secondary partition;
- $MRand(\mathbf{M}_c, \hat{\mathbf{M}}_c)$: Modified Rand Index between true and fitted membership matrices of the primary partitions;
- $Coph(\mathbf{U}_c, \hat{\mathbf{U}}_c)$: Cophenetic Coefficient (Sokal and Rohlf 1962) between true and fitted ultrametric matrices of the primary classifications;
- VAF (Variance Accounted For) between true and fitted dissimilarity values.

The number of iterations before convergence has also been recorded.

[4] In particular, (I x 5) data matrices were generated following the generating process by Milligan and Cooper (1985) by setting the parameters rd=0.01 and rad=2.

The best solution in terms of objective function in a number of different runs of the algorithm was retained to prevent from falling in local optima due to the starting solutions.

A preliminary simulation was performed just to investigate the influence of the starting solution on the recovery of the true classification structures in case of no error (h=0 in the covariance matrices). The numbers of occasions and objects were set to 50 and 20, respectively and the averages of the performance measures were computed in 100 replications, by retaining in each replication the best solution in 5, 10, 15, 20, 50 runs of the algorithm, respectively.

Table 3 displays the results of the first simulation experiment: the averages of the performance measures show a significant improvement when the best solution is retained in an increasing number of runs. The improvement is particularly evident in the recovery of the true primary partitions measured in terms of MRand index. Furthermore, from a detailed analysis of all the results it turns out that when the number of runs is 50, actually only one case over 100 fails in recovering the true structure. A very good performance is reached already when the optimal solution is retained over 20 runs of the algorithm. Thus, in the following simulation this value has been considered to retain the best solution when the algorithm has been run with different error levels and number of occasions.

In the second simulation experiment for each of the 2 (numbers of occasions) x 4 (error levels) factors, 100 data sets were constructed, giving a total of 800 data sets. The algorithm ran 16000 times in total.

Table 4 displays the outcomes of the simulation experiment: for all conditions, the averages of the indices over all replications are given, which exhibit a good performance of the algorithm, even when the error level is quite high. The performance is better, as expected, when the number of occasions is higher, since a large amount of information is available to fit the model.

7. Discussion

Dissimilarity data observed or computed on a set of I multivariate objects are frequently analyzed by cluster analysis techniques that fit a unique theoretical structure of classification (e.g., partition, hierarchy, covering) to the dissimilarity matrix.

It can be noted that each single variable describing the objects determines a dissimilarity matrix which is a part of the whole dissimilarity matrix relative to all variables. However, a dissimilarity matrix associated to a single variable may induce a classification different from the one obtained on the entire set of variables and, consequently, a classification of a set of multivariate objects can be seen as a consensus classification of those obtained by different variables describing the same objects.

Table 3. Results of the first simulation experiment. The averages are computed over 100 runs of the algorithm.

Number of runs to retain the best solution	*Average MRand* (*Secondary Partition*)	*Average MRand* (*Primary Partition*) *in Class* 1	*Average MRand* (*Primary Partition*) *in Class* 2	*Average Cophenetic Coefficient* (*Primary Hierarchy*) *in Class* 1	*Average Cophenetic Coefficient* (*Primary Hierarchy*) *in Class* 2	*Average VAF*	*Average number of iterations*
5	1	0.9057	0.8892	0.9762	0.9655	0.9838	22.74
10	1	0.9512	0.9565	0.9906	0.9818	0.9843	23.53
15	1	0.9834	0.9683	0.9942	0.9866	0.9848	21.75
20	1	0.9930	0.9784	0.9970	0.9898	0.9842	21.47
50	1	0.9974	0.9960	0.9987	0.9977	0.9846	22.16
Number of runs to retain the best solution	*Number of times when the measures are equal to* >0.99						
5	100	81	72	81	73	18	
10	100	91	89	91	89	15	
15	100	97	92	97	92	26	
20	100	99	95	99	95	21	
50	100	99	99	99	99	18	
Number of runs to retain the best solution	*Number of times when the measures are* > 0.98						
5	100	81	72	81	73	82	
10	100	91	89	91	89	76	
15	100	97	92	97	92	76	
20	100	99	95	99	95	74	
50	100	99	99	99	99	75	

Table 4. Results of the second simulation experiment. The averages are computed over 100 runs of the algorithm.

Number of Occasions	*Error Levels*	*Average MRand* (*Secondary Partition*)	*Average MRand* (*Primary Partition*) *in Class* 1	*Average MRand* (*Primary Partition*) *in Class* 2	*Average Cophenetic Coefficient* (*Primary Hierarchy*) *in Class* 1	*Average Cophenetic Coefficient* (*Primary Hierarchy*) *in Class* 2	*Average VAF*	*Average number of iterations*
20	0.5	0.9537	0.8980	0.7002	0.9600	0.8690	0.6642	17.22
	1	0.6193	0.7281	0.7538	0.8282	0.3769	0.4936	14.90
	1.5	0.5193	0.6373	0.6662	0.6199	0.3101	0.3931	12.77
	2	0.4594	0.6158	0.6309	0.4904	0.2646	0.3322	11.33
50	0.5	0.9836	0.9868	0.9650	0.9951	0.9832	0.6588	22.33
	1	0.8887	0.9263	0.8135	0.9772	0.9116	0.4984	20.87
	1.5	0.7148	0.8604	0.7654	0.9460	0.6305	0.3982	18.51
	2	0.5929	0.7543	0.7855	0.8782	0.3870	0.3290	15.41

Moreover, variables describing the objects explain different aspects and frequently define very different classifications of the objects when taken alone. In this situation a single theoretical classification is not sufficient to describe the taxonomic information present in the data and we suppose that the most relevant classification structures have to be detected.

Furthermore, since we deal with more complex data (three-way dissimilarity data) a more flexible technique of classification is required to account also for the heterogeneity along the different occasions (Gordon and Vichi 1998; Vichi 1999), by partitioning the set of dissimilarity matrices $\{\mathbf{D}_1,\ldots,\mathbf{D}_K\}$ into disjoint classes with similar classification structure that can be properly summarized by a consensus classification.

Here we suppose that such a consensus classification is defined by the sum of two order-constrained distance matrices representing, respectively, a hierarchy and a partition.

The model proposed detects a partition of occasions and, within each of such classes, it fits two constrained distance matrices. The algorithm is based on successive residualizations of the given three-way data matrix: within each class of occasions, one matrix is fitted, obtaining the residual dissimilarities from it and then the second matrix is fitted to these residuals. The two steps are alternated and iterated until convergence.

Obviously, within each class of the secondary partition, a sum of more than two order-constrained matrices (even of the same type) could be fitted to the three-way dissimilarity matrix. The ALS algorithm here proposed could be easily generalized by considering several constrained distance matrices corresponding to different classification structures and taking into account, even in this case, the possible occurrence of nested structures.

Since the proposed algorithm overcomes the efficiency problems of a standard quadratic problem by relaxing the positiveness-constraint on the matrices **P** and **U**, the ALS procedure can be generally used in other contexts where efficient alternative approaches are not available.

The proposed methodology can be applied even for large data sets, (in terms of both objects and occasions) because the coordinate descent algorithm presented in this paper includes: the partitioning step *d* (step *b*) for the occasions (objects), which is solved as an assignment problem in linear time (and two simple linear regressions); the hierarchical step *a*, which is solved by carrying out the UPGMA clustering in time complexity $O(I^2\log(I))$.

However, when the number of objects is large, even in the simplest case of a two-way dissimilarity matrix, that is, a three-way dissimilarity matrix with a single occasion, the resulting indexed hierarchy often becomes hard to interpret (even though recently more and more used as for example in a microarray data context).

References

ASUNCION, A., and NEWMAN, D.J. (2007), *UCI Machine Learning Repository*, http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, School of Information and Computer Science.

CARROLL, J.D., and ARABIE, P. (1983), “An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm”, *Psychometrika, 48*, 157-169.

DE SOETE, G. (1984), “A Least Squares Algorithm for Fitting an Ultrametric Tree to Dissimilarity Matrix”, *Pattern Recognition Letters*, *2*, 133-137.

FERN, X.Z., and BRODLEY, C.E. (2003), “Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach”, in *Proceedings of the 20th International Conference on Machine Learning*, *ICML*, Washington D.C., pp.186-193.

FRED, A.L.N., and JAIN, A.K. (2003), “Robust Data Clustering”, in *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *CVPR*, USA.

GORDON, A.D. (1999), *Classification* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

GORDON, A.D., and VICHI, M. (1998), “Partitions of Partitions”, *Journal of Classification*, *15*, 265-285.

GORDON, A.D., and VICHI, M. (2001), “Fuzzy Partition Models for Fitting a Set of Partitions”, *Psychometrika*, *66*(2), 229-248.

HUBERT, L., and ARABIE, P. (1985), “Comparing Partitions”, *Journal of Classification*, *2*, 193-218.

HUBERT, L., and ARABIE, P. (1994), “The Analysis of Proximity Matrices through Sums of Matrices Having (Anti-)Robinson Forms”, *British Journal of Mathematical and Statistical Psychology*, *47*, 1-40.

HUBERT, L., ARABIE, P., and MEULMAN, J. (1998), "Graph-Theoretic Representations for Proximity Matrices through Strongly-Anti-Robinson or Circular Strongly-Anti-Robinson Matrices", *Psychometrika*, *63*(4), 341-358.

KAUFMAN, L., and ROUSSEEUW, P.J. (2005), *Finding Groups in Data. An Introduction to Cluster Analysis*, New York: John Wiley & Sons.

KOIVISTO, M., and SOOD, K. (2004), "Exact Bayesian Structure Discovery in Bayesian Networks", *Journal of Machine Learning Research*, *5*, 549-573.

MACQUEEN, J.B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1 Statistics*, eds. L.M. Le Cam and J. Neyman, Berkeley: University of California Press, pp. 281-297.

MCKENZIE, D.P., and FORSYTH, R.S. (1995), "Classification by Similarity: An Overview of Statistical Methods of Case-based Reasoning", *Computers in Human Behavior*, *11*(2), 273-288.

MEULMAN, J.J., and HEISER, W.J. (2004), *SPSS Categories 13.0*, Chicago: SPSS Inc.

MILLIGAN, G.W., and COOPER, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, *50*, 159-179.

POWELL, M.J.D. (1983), "Variable Metric Methods for Constrained Optimization", in: *Mathematical Programming: The State of Art*, eds. A. Bachem, M. Grotschel, B. Korte, New York: Springer-Verlag, pp. 288-311.

ROSENBERG, S., and KIM, M.P. (1975), "The Method of Sorting as Data-Gathering Procedure in Multivariate Research", *Multivariate Behavioral Research*, *10*, 489-502.

SOKAL, R.R., and ROHLF, F.J. (1962), "The Comparison of Dendrograms by Objective Methods", *Taxon*, *11*, 33-40.

STREHL, A., and GHOSH, J. (2002), "Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions", *Journal of Machine Learning Research*, *3*, 583-618.

VICARI, D., and VICHI, M. (2000), "Non-Hierarchical Classification Structures", in *Data Analysis*, eds. W. Gaul, O. Opitz, and M. Schader, Heidelberg-Berlin: Springer-Verlag, pp. 51-65.

VICHI, M. (1999), "One Mode Classification of a Three-Way Data Matrix", *Journal of Classification*, *16*, 27-44.

WANG, D., CHAUDHARI, N.S., and PATRA, J.C. (2004), "A Constructive Unsupervised Learning Algorithm for Clustering Binary Patterns", in *Proceedings of the International Joint Conference on Neural Networks, IJCNN-04*, *2*, Budapest, Hungary (IEEE Cat. No. 04CH37541C), (ISBN: 0-7803-8360-5), pp. 1381-1386.