SAPIENZA
UNIVERSITÀ DI ROMA

Fuzzy clustering for complex data structures

Scuola di Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXXVI Ciclo

Candidate

Ilaria Bombelli
ID number 1696658

Thesis Advisors

Prof. Maurizio Vichi
Prof.ssa Maria Brigida Ferraro

Academic Year 2022/2023

Thesis defended on January 23, 2024
in front of a Board of Examiners composed by:

Prof. Antonio Di Crescenzo (chairman)

Prof. Alfonso Iodice D'Enza

Prof. Gianluca Mastrantonio

Fuzzy clustering for complex data structures
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: January 2024

Author's email: ilaria.bombelli@uniroma1.it

Abstract

Multidimensional phenomena are often represented by complex data structures. With the rapid growth of data availability and complexity, new methodologies are needed to handle these kind of data. Among complex data structures, deep interest has been devoted to three-dimensional data and network data, since many applications can be represented as such. Among methodological techniques, cluster analysis is one of the most popular and successful techniques for data exploration and characterization. However, existing methodologies for describing and analyzing such complex data use a hard approach to clustering, even though many applications show the need to use a fuzzy approach, as it allows for better interpretation of results and greater closeness of results to reality.

What is proposed in this thesis are new methodologies for applying fuzzy clustering to complex data structures, such as three-way data and network data. The fuzzy approach to clustering proves extremely useful in the simulations and real-world applications which will be discussed through the chapters.

The first chapter introduces the notions of complex data structures and positions the problem, highlighting the rationale behind the proposed methodologies through theoretical discussions and real-world practical examples. The second chapter provides the reader with terminology used throughout the thesis and definitions of basic concepts. From the third to the sixth chapter, four different research works are presented.

The first work introduces the notions of three-way three-mode data, as a data array made up by different units-by-variables matrix, each of which refers to a specific occasion (usually time); by applying hierarchical clustering techniques to each units-by-variables data matrix, a set of hierarchies (dendrograms) is obtained. The new methodology proposes to obtain a fuzzy partition of the set of hierarchies and simultaneously, within each class of the partition, identify a *consensus hierarchy*.

The second work can be considered as an extension of the previous one. Given a set of hierarchies, the proposed new methodology makes it possible to obtain a fuzzy partition of them, and within each class of the partition, identify a *parsimonious* consensus dendrogram. The notion of parsimonious is extensively commented and discussed in the corresponding chapter. However, here it is important to recall that a parsimonious dendrogram is useful for getting a clear and direct idea of how units aggregate into clusters, highlighting only the most important aggregations and deleting misleading ones.

The third work introduces a new methodological proposal to obtain a fuzzy partition of a three-way three-mode data array with corresponding consensus matrices for each class in the partition and simultaneously reduce the dimension of the variables in the consensus matrices by applying a disjoint second-order factor analysis. The motivation and theoretical background are discussed in the corresponding chapter.

Finally, the last work focuses on how to apply different fuzzy clustering techniques to a set of networks. In particular, the main issue that arises in this kind of problem concerns how to represent networks so that they can be given as input to the clustering algorithms. Several representations of networks involving probability distributions and graph embedding techniques are presented and discussed.

The last chapter summarizes the main contents of the thesis, recalling the methodological proposals, emphasizing their relevance and contribution, especially their strength when applied to real scenarios. Finally, the necessity of using a fuzzy approach to clustering and its main advantage are emphasized.

Contents

List of Figures

List of Tables

Chapter 1

Motivation and introduction

Statistics is the grammar of science.

Karl Pearson

Multivariate collective phenomena are usually described by means of a set of statistical units characterized by a set of variables observed on several successive occasions that frequently represent time. This is the most complete way to statistically describe phenomena under observation, because units can be represented realistically in their complexity with a large number of variables, without naive simplifications based on a few characteristics, and even the histories of units can be compared over time in the same descriptive analysis in order to show convergences or divergences over the considered time period. The data structure that allows this complete statistical analysis is a *three-way three-mode data array*, where modes are *units, variables, and occasions* and the term way refers to a dimension of the data, i.e. *rows, columns, layers*.

When phenomena are economic, social, or demographic and occasions represent times, the investigator may be interested in studying "stable" subsets of occasions (e.g., years) where units do not change much in their cross-sectional relations; that is, their pairwise dissimilarity structure does not change substantially and variables do not differentiate consistently their covariance structure. In other terms, in these subsets of occasions, the multivariate unit-by-variable data matrices are perceived as similar to each other. Thus, in this situation, it is natural to identify clusters of similar occasions, frequently corresponding to different historical periods. Then, once the clusters of years are detected, as usually done in the partitioning methods of Cluster Analysis, K *centroids*, one for each cluster, are identified to summarize the elements belonging to each specific cluster. In this case, since we are dealing with three-way data structures, the K centroids represent K *consensus matrices* summarizing the unit-by-variable matrices belonging to the clusters. In other terms, each consensus represents the closest data matrix -in a least-squares sense- to the matrices belonging to the same cluster.

For a given occasion, it is also possible to observe, instead of the complete units-by-variables data matrix, a hierarchy of the set of units, obtained by applying hierarchical clustering on the original units-by-variables properly transformed in its corresponding distance matrix or are directly observed. In such situations, the investigator would need to obtain a single *consensus hierarchy* of the original set of hierarchies, which can be defined as the *closest* hierarchy to the given set of hierarchies.

For example, units (or objects) may correspond to a set of countries whose

macroeconomic performance is compared over a series of years, by acquiring a hierarchy of classes (a dendrogram) for each year with the property that countries, in the same class in that year, are perceived as similar to one another. Thus, a consensus hierarchy would identify similar clusters and similar agglomerations of clusters in the different hierarchies.

Alternatively, such data may occur from data-gathering techniques, such as data cards, in psychometric studies (Rosenberg and Kim, 1975; Whaley and Longoria, 2009), or products in marketing applications, where individuals (customers) are required to sort similar items into clusters they perceived as homogeneous and then asked to aggregate the clusters to obtain a hierarchy. Consequently, the consensus hierarchy of the hierarchies defined by the customers would identify the closest hierarchy to those observed.

In addition, a consensus may be needed when several different analyses of the same set of objects are carried out. For example, hierarchical clustering methods applied on the dissimilarity matrix computed on the same units-by-variables matrix can vary and be different depending on the type of measure used to compute pairwise dissimilarity or on the clustering criterion adopted. In effect, each decision that is taken (e.g. choice of type of dissimilarity measure) involves a model for the clusters that may bias the results of an analysis towards the assumptions of the model. For this reason, investigators often carry out several different analyses of the same set of objects, each implicitly incorporating a different set of assumptions that are considered to be reasonable. A consensus classification may be considered an *ensemble* classification estimating the *true classification*, that is, the classification less likely to be biased towards the models corresponding to the separate analyses and more likely to reflect the underlying structure of the data.

Finally, hierarchical classifications may be obtained by application of an agglomerative or divisive algorithm separately to the same set of multivariate objects observed on different occasions using a set of variables forming a three-way data set, or panel data. A consensus hierarchy provides a way of simplifying this information and obtaining an overall view of the relationships within the set of objects. Therefore, identifying a consensus hierarchy allows us to fine similar clusters that have been observed in the different hierarchies. In general, this implies that the assumption of a single (hierarchical) clustering for the same objects can be too strict. This has motivated the emerging area of *multi-clustering* (Muller et al., 2012) for which a consensus hierarchical clustering may be required to highlight similar clusters and similar agglomerations of clusters in different hierarchies.

Nevertheless, there are several situations in which obtaining a single consensus hierarchy is too simplistic, misleading and naïve because several differences may be observed among the set given hierarchies and consequently, more than one consensus hierarchy could be required to synthesize the initial hierarchies.

For example, the macroeconomic performances of different countries, after a period of stability, may change under the effect of an economic shock. Thus, after a period in which hierarchical relations remain similar, the relationships between the countries may change, and the new relationships may remain stable for a successive number of years. Therefore, for each period of stability, a different consensus hierarchy may be required. In the case of data gathering, individuals or customers might use two or more different criteria to sort the set of items, thereby producing different hierarchical relations to be classified into homogeneous classes. In multi-clustering there might be more than one general agreement among the different clustering views. As a results, a reduced number of consensus classifications may be sufficient to synthesize the different clustering views.

As the three-way three-mode data array is an important and useful complex data

structure used to represent multivariate collective phenomena, so are the *networks*. Indeed, they are a powerful model for describing problems in different scientific fields, such as biology, informatics, and social sciences. Given a network, an intuitive and naïve analysis consists in applying clustering techniques to detect cluster of entities, represented by the *nodes* in the networks. Being this task well-known and widely explored in literature, a more interesting, novel and challenging task consists in identify groups of homogeneous networks, given a set of networks: this need is highly demonstrated by the usage of networks representation to describe many problems pertaining to multiple disciplines. For example, in biology networks can represent metabolites and the clustering will allow to identify groups of patients with similar patologies (Manipur et al., 2020a); in transportation, networks may represent airline companies and the need to use clustering is demonstrated by the need of identifying groups of airline companies characterized by similar flight route (Carpi et al., 2019); in trade, networks can represent imported/exported products in the trade market and the identification of clusters of products allows to find which products are characterized by similar trade behaviour (Tantardini et al., 2019).

In this brief introduction, I hope the reader already got the scent of the potential of the application of clustering techniques to complex data structures (units-by-variables matrices, hierarchies, networks) to group them into K clusters and to identify K *consensuses*, referred as *consensus matrices* or *consensus hierarchies* or *network centroids*, respectively. However, what is proposed in this dissertation is to apply *fuzzy* clustering algorithms to such kind of data structures. In order to stress the motivation behind the usage of a fuzzy approach and to make the reader see first-hand the strength and advantages of such approach, it worthy recalling the example of hierarchies describing the macroeconomic outlook of the countries in several years. Imagine, for example, that during a period of stability, the hierarchies of the macroeconomic outlook for different countries remain more or less equal. Thus, each hierarchy is assigned to the class of the stable years with the membership degree almost equal to one. Suppose an economic shock occurs that changes drastically the macroeconomic outlook for each country and, consequently, the hierarchical relations among countries. Then, after a period of instability, suppose that a new stable period is observed. In that case, new hierarchical relations among countries might be noted, and these relations will remain stable in this second period. Thus, each hierarchy of countries will have a membership degree for this second class of years almost equal to one. However, during the years when the shock occurred, it is realistic to suppose that some countries will have features similar to those observed before the shock, while others will have aspects of the stable period after the shock. A hard secondary partition of the primary hierarchies would assign the hierarchy to one or the other of these two classes of years, whereas a fuzzy partition would be able to indicate the uncertainty of the years of the shock by specifying its membership degrees for each of the classes. The range of values that can be taken by membership degrees is seen as an advantage of fuzzy clustering methods over hard clustering methods.

1.1 Content of the thesis

This section gives a brief summary of the dissertation structure. At first, basic terminology and notations used through the paper is provided to make the reader able to easily understand and to clarify some used terminology (Chapter 2). Then, in the following chapters of this dissertation (Chapters 3-6), I include most of the scientific papers belonging to my whole scientific production. As discussed above,

what brings them together is the application of fuzzy clustering techniques to complex data structure. Then, what differentiate between them is obviously the kind of complex data structure which is object of analysis and the methodological proposals presented to make further and deeper analysis.

Chapter 3 presents a novel methodology, called PARoDENo3WD (PARtition of DENdrograms of a 3-Way Data array), to study multidimensional phenomena. The proposed methodology allows to obtain a secondary fuzzy partition of the primary hierarchies, where hierarchies belonging to the same class are perceived as similar and each class is associated to a consensus hierarchy.
The contents of Chapter 3 were developed with Prof. Maurizio Vichi and Prof.ssa Maria Brigida Ferraro, and are reported in a paper which is published in *Information Sciences* (Bombelli, Ferraro, and Vichi, 2023).

Chapter 4 provides a new methodology aiming at fitting a fuzzy partition to a set of hierchies of the same set of objects and at identifying parsimonious consensus hierarchies. The methodological and theoretical background related to parsimonious hierarchy is fully described in Section 4.2.
The contents of Chapter 4 were developed with Prof. Maurizio Vichi and they are reported in a paper submitted to *Statistics and Computing* and is currently under review.

Chapter 5 provides a new methodology which, given a set of units-by-variables matrices, aims at obtaining a simultaneous reduction of the dimensions of the occasions and the variables. This is done by obtaining a fuzzy partition of the matrices to reduce the dimension of the occasions and by applying a Second-order Disjoint Factor Analysis to the consensus matrices identifies for each class of the partition in order to reduce the dimension of the variables. The theoretical background related to the Second-order Disjoint Factor Analysis is provided in Section 5.2.
The contents of Chapter 5 were developed with Prof. Maurizio Vichi and they are reported in a paper submitted to the *Journal of Computational and Graphical Statistics* and is currently under review.

Finally, Chapter 6 is fully devoted on the application of fuzzy clustering techniques to networks. Particularly, an important focus is given on how to represent networks ensambles for fuzzy clustering.
The contents of Chapter 6 were developed with Prof.ssa Maria Brigida Ferraro, Prof. Mario Rosario Guarracino and Dr. Ichcha Manipur and most of the chapter is reported in a paper which is published in *Data Mining and Knowledge Discovery* (Bombelli et al., 2023).

All the models presented in Chapter 3-5 have been implemented mainly by using a MATLAB routine (MATLAB, 2021), while the research presented in Chapter 6 has been developed by using mainly the R software (R Core Team, 2022).

Chapter 2

Notation and basic concept

For the convenience of the reader, the notation and the basic concepts used in this dissertation are listed here. It is worth noticing that Chapter 6 does not strictly follows this notation, as networks notation is introduced and discussed in the related chapter.

List of Symbols

N, J, H, K	number of observations, variables, occasions (layers), clusters of primary hierarchies (or matrices in Chapter 5), respectively;
$\mathcal{I} \equiv \{1, \ldots, N\}$	the set of indices identifying units;
$\mathcal{J} \equiv \{1, \ldots, J\}$	the set of indices identifying variables;
$\mathcal{H} \equiv \{1, \ldots, H\}$	the set of indices identifying occasions (layers);
$\mathcal{K} \equiv \{1, \ldots, K\}$	the set of indices identifying clusters of primary hierarchies (or matrices in Chapter 5);
μ_{hk}	membership of h-th primary hierarchy (or matrix in Chapter 5) in the k-th cluster, for $k \in \mathcal{K}$, for $h \in \mathcal{H}$. For a given occasion, the sum of the membership values for all clusters is one; moreover, memberships can be hard, i.e. $\mu_{hk} \in \{0, 1\}$, or fuzzy i.e. $\mu_{hk} \in [0, 1]$;
m	the fuzziness parameter or fuzzifier that controls how fuzzy the classes of the partition are;
$\mathbf{X} = [x_{ijh}]$	an $(N \times J \times H)$ three-way data matrix, formed by data matrices $[\mathbf{X}_1, ..., \mathbf{X}_H]$, where value x_{ijh} is the observation on the i-th unit (row), on the j-th variable (column) on the h-th occasion (layer);
$\mathbf{D} = [u_{ilh}]$	the $(N \times N \times H)$ three-way matrix formed by dissimilarity matrices $[\mathbf{D}_1, ..., \mathbf{D}_H]$, where value d_{ilh} is the dissimilarity between the i-th unit (row) and the l-th unit (column) on the h-th occasion (layer);
$\mathbf{U} = [u_{ilh}]$	the $(N \times N \times H)$ three-way ultrametric matrix formed by ultrametric matrices $[\mathbf{U}_1, ..., \mathbf{U}_H]$, where value u_{ilh} is the ultrametric distance between the i-th unit (row) and the l-th unit (column) on the h-th occasion (layer);

$\mathbf{U}^* = [u^*_{ilk}]$ the $(N \times N \times K)$ three-way consensus matrix formed by ultrametric matrices associated with the K consensuses $[\mathbf{U}^*_1, ..., \mathbf{U}^*_K]$, where value u^*_{ilh} is the consensus ultrametric distance between the i-th unit (row) and the l-th unit (column) on the k-th consensus (layer).

Definition 1 (Dissimilarity matrix)**.** A *dissimilarity matrix* $\mathbf{D} = [d_{ij} : i, j \in \mathcal{I}]$, is an $N \times N$ square matrix where elements satisfy the following properties: 1. non-negativity: $d_{ij} \geq 0,\ \forall\ i, j \in \mathcal{I}$; 2. null diagonal: $d_{ii} = 0,\ \forall\ i \in \mathcal{I}$; 3. symmetry: $d_{ij} = d_{ji},\ \forall\ i, j \in \mathcal{I}$.

Definition 2 (Distance matrix)**.** A *distance matrix* is a dissimilarity matrix, in which triplets of units satisfy properties 1-3, and also the following property: 4. triangular inequality: $d_{il} \leq d_{ij} + d_{lj},\ \forall\ i, j, l \in \mathcal{I}$.

Definition 3 (Ultrametric matrix)**.** An *ultrametric matrix* is a distance matrix, which triplets satisfy also the following property: 5. ultrametric inequality: $d_{il} \leq \max\{d_{ij}, d_{lj}\},\ \forall\ i, j, l \in \mathcal{I}$.

2.1 Three-way three-mode data

Formally, a *three-way three-mode data array (matrix)* is denoted by $\mathbf{X} = [x_{ijh}]$ of size $(N \times J \times H)$, where value x_{ijh} is the observation on the i-th unit (row), of the j-th variable (column), on the h-th occasion (layer). Therefore, the three-way array is the data structure used to organize multidimensional phenomena, with J variables measured on the same set of N individuals, in H different occasions. $\mathbf{X} = [x_{ijh}]$ has three modes: units (rows), variables (columns), and occasions (times, layers). The term 'way' refers to a dimension of the data, while the word 'mode' is reserved for the methods or models used to analyze the data (Kroonenberg, 2008). The array $\mathbf{X}$ may be seen as a set of multivariate data matrices unit-by-variable, as many as the number of occasions, representing a multi-view cross-sectional observation of the phenomenon. On the other hand, $\mathbf{X}$ may be seen as a set of multivariate time series matrices variable-by-time (when occasions are times), as many as the units, representing the multivariate histories of the units to be compared over time.

2.2 Hierarchical clustering

Hierarchical clustering methods are popular procedures, which can be either *agglomerative* or *divisive*, for yielding a hierarchy of partitioned units (Hartigan, 1975; Gordon, 1999; Vichi, Cavicchia, and Groenen, 2022). They start from dissimilarity data between pairs of N objects and produce a nested set of $N - 1$ partitions. The most commonly used hierarchical clustering methods are agglomerative where pairs of objects or clusters are merged into larger ones. The starting point is the set of N singleton clusters, and after fixing a distance between clusters function of the dissimilarities between units, the algorithm progressively agglomerates the closest clusters into larger and larger ones until obtaining the whole set of units (*bottom-up procedure*). In contrast, divisive clustering methods split the whole set of units or its successively divided clusters until all clusters are singletons, by splitting each time the most dissimilar (isolated) pair of clusters according to a fixed dissimilarity between clusters (*top-down procedure*). In the simulation studies, as well as in the applications on real data presented and discussed throughout the dissertation, the

agglomerative procedure was considered for hierarchical clustering rather than the divisive one. In fact, among others, there are two main advantages: first, outliers can be better handled by agglomerative hierarchical clustering than by its divisive counterpart, since in the former procedure outliers can be absorbed into larger clusters, whereas the latter procedure can create clusters around outliers, leading to suboptimal cluster results; second, the interpretability of the results is better and clearer using a bottom-up procedure, since the entire merging process can be observed, rather than the splitting process, and the choice of the number of clusters can be based on the desired level of detail and granularity. Well-known hierarchical clustering methods include the single linkage (Florek et al., 1951), the average linkage (Sokal, 1958), the complete linkage (McQuitty, 1960), and Ward's method (Ward Jr, 1963), where, for the first three methods the dissimilarity between clusters is fixed as the minimum, the mean, and the maximum, of the dissimilarities between units (one in one cluster and the other in the other cluster), while for the fourth method, the dissimilarity between clusters is the increase in the deviance when the two clusters agglomerate into one. It is evident that the choice of linkage method to be used in hierarchical clustering influences the structure of the hierarchy obtained. In simulations and real applications, the average linkage method (UPGMA) was mainly used. Note that each hierarchical clustering method produces as a solution an N-tree, a hierarchy (dendrogram), and an ultrametric matrix.

2.3 N-tree

An *N-tree* $\boldsymbol{T}_h = \{\{i\}, (i \in \mathcal{I}), I_{1,h}, I_{2,h}, \ldots, I_{N-1,h}, \mathcal{I}\}$ is an unordered rooted tree with labeled leaves representing units (singletons) and internal nodes clusters of units. Generally, the N-trees are binary, i.e., have exactly $N-1$ internal nodes and each node has at most two descendants. More precisely, the N-tree is described as a set of subsets of $\mathcal{I}$, with $I_{l,h}$ the generic l-th subset of $\mathcal{I}$ taken in the h-th occasion: $\mathcal{I} \in \boldsymbol{T}_h$; $\emptyset \notin \boldsymbol{T}_h$; $\{i\} \in \boldsymbol{T}_h$; if $I_{i,h}, I_{l,h} \in \boldsymbol{T}_h \Rightarrow (I_{i,h} \cap I_{l,h}) \in (I_{i,h}, I_{l,h}, \emptyset)$. Thus, the N-tree, for the h-th occasion, is given by the N trivial clusters (leaves) $\{i\}$, $(i \in \mathcal{I})$ and the $N-1$ clusters of units (internal nodes), which are disjoint or one included in the other, obtained by the $N-1$ steps of fusion (agglomeration or division) performed by a hierarchical algorithm (hence, the last cluster is $\mathcal{I}$, i.e., the root). Thus, an N-tree specifies only the subsets belonging to a hierarchical classification.

2.4 Dendrogram (Hierarchy)

A *Dendrogram (hierarchy)* is a diagram that shows the hierarchical relationships between clusters of units. A vertically orientated dendrogram has on the x-axis the units (singletons, leaves of the tree) and on the y-axis the level of fusion (agglomeration or division) between clusters. The dendrogram is a valued N-tree, in which a non-decreasing level of fusion is associated with each internal node and the root as the size of the clusters increases. Formally, the primary dendrogram of occasion h-th is, $\boldsymbol{\delta}_h = \{\delta(I_{1,h}), \delta(I_{2,h}), \ldots, \delta(I_{N-1,h})\}$ and $\delta(I_{l,h})$ is the value of fusion determining $I_{l,h}$, such that if $\delta(I_{l,h}) \leq \delta(I_{i,h})$, implies: $I_{l,h} \subseteq I_{i,h}$ if $I_{l,h} \cap I_{i,h} \neq \emptyset$; otherwise $l \leq i$ if $I_{l,h} \cap I_{i,h} = \emptyset$.
The set of H dendrograms is denoted by $\Delta = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_H]$.
Johnson (1967) has proved a bijection between the set of ultrametric matrices and the set of dendrograms (hierarchies), thus, to each dendrogram $\boldsymbol{\delta}_h$ there corresponds

an ultrametric matrix $\mathbf{U}_h$, and vice versa. Hence, the set $\Delta = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_H]$ corresponds to the set $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_H]$. In this dissertation, the primary hierarchies are supposed observed with associated ultrametric matrices $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_H$. When hierarchies are not directly observed, the data array $\mathbf{X} = [x_{ijh} : i \in \mathcal{I}, j \in \mathcal{J}, h \in \mathcal{H}]$ is supposed given. Then, a fixed hierarchical clustering algorithm (Gordon, 1999) is applied to each dissimilarity matrix $\mathbf{D}_h$ related to the data matrix $\mathbf{X}_h$, $h = 1, ..., H$, by choosing a dissimilarity measure between multivariate objects.
Each hierarchical classification applied on the data matrix $\boldsymbol{X}_h$ has associated: (i) N-tree, (ii) Hierarchy (Dendrogram), (iii) Ultrametric matrix.

2.5 Cluster Analysis

Cluster Analysis or briefly clustering includes a large set of unsupervised methodologies and their associated algorithms that allow the grouping of the units into clusters, with the property that intra-cluster units are perceived as similar, and inter-cluster units are seen as different. In particular, when clustering regards a partitioning problem (Gordon, 1999), units are grouped into, say, K clusters, and one centroid for each cluster is identified to summarize the existing characteristics of units within the cluster. It is important to remember that two clustering approaches are possible in the partitioning problem: those *hard* and *fuzzy*. According to the former, units are assigned exclusively to a single cluster, and a *hard partition* is defined in this way. In the latter, the assignment is more flexible and allows each unit to belong to all clusters, but with some degree of membership. This approach allows the introduction of a form of "uncertainty" into the specification of clusters, which helps the researcher to deal with situations where some units clearly have characteristics of several clusters and therefore cannot be supposed to belong to a single cluster only. Specifically, a membership degree matrix $[\mu_{ik}]_{i=1,\ldots,N,k=1,\ldots,K}$, is defined with size $N \times K$, having the property of being row-stochastic, i.e. being such that $\sum_{k=1}^{K} \mu_{ik} = 1, \ \forall\ i = 1, \ldots, N$. In detail, when a hard approach is used, then $\mu_{ik} \in \{0, 1\}, \ \forall\ i = 1, \ldots, N, \ k = 1, \ldots, K$; instead, when a fuzzy approach is used, then $\mu_{ik} \in [0, 1], \ \forall\ i = 1, \ldots, N, \ k = 1, \ldots, K$.

2.6 Fuzzy Clustering

In some practical situations objects do not have a clear assignment to a cluster, but, unfortunately, if a hard/standard approach is used, each object is only assigned to one cluster. To overcome this drawback, the fuzzy approach to cluster analysis was introduced. It allows each object to be assigned to all clusters with certain membership degrees varying in the unit interval: $\mu_{hk} \in [0, 1]$ and $\sum_{k=1}^{K} \mu_{hk} = 1, \ \forall\ h \in \mathcal{H}$. The most known and used fuzzy clustering method is the fuzzy c-means by Bezdek (Bezdek, 1981), the fuzzy generalization of the k-means algorithm (MacQueen, 1967b). It is a fuzzy algorithm that starts from a units-by-variables data matrix and consists of clustering N units into K clusters, allowing each unit to belong to more than one cluster. It returns as output the clusters' prototypes (centroids) and a membership matrix $[\mu_{hk}]$, i.e. an $(N \times K)$ matrix with the generic element μ_{hk} satisfying a) $\mu_{hk} \in [0, 1] \ \forall\ h \in \mathcal{H}, \ \forall\ k \in \mathcal{K}$ and b) $\sum_{k=1}^{K} \mu_{hk} = 1, \ \forall\ h \in \mathcal{H}$ and indicating the extent to which each unit belongs to the corresponding cluster.

Chapter 3

Consensus and fuzzy partition of dendrograms from a three-way dissimilarity array

3.1 Introduction

This chapter addresses the problem of obtaining partitions of the set of hierarchical partitions of objects. The hierarchical partitions to be partitioned will be referred to as *primary hierarchies* (dendrograms). A fuzzy partition of a set of primary hierarchies will be, instead, referred to as a *secondary fuzzy partition*. The methodology described in this chapter aims to obtain a secondary fuzzy partition of the set of primary hierarchies into classes for which primary hierarchies with a relevant membership degree for the same class are perceived as similar to one another. Each class will have an associated *consensus hierarchy*, which serves as a summary of the set of primary hierarchies belonging to the class. The new methodology is named PARtionon of DENdrograms of a 3-Way Data array, or PARoDENo3WD. A flowchart describing the methodology and clarifying the process is displayed in Figure 3.1. The secondary partition is fuzzy because it can describe 'uncertainties' in the observed set of primary hierarchies and provides further information: for each class the membership degrees can show which primary hierarchies are more strongly associated with it and which hierarchies have only a weak association. Therefore, each hierarchy contributes to the definition of all classes according to different membership degrees.

For the technical formulation of the methodology proposed in this chapter, the reader may refer to Chapter 2.
An outline of the material in this chapter is as follows. Section 3.2 reviews the literature. In Section 3.3 the proposed methodology is fully described and detailed. Section 3.4 includes an extended simulation study for the evaluation of the performance of the new methodology. Section 3.5 provides the application of the proposed methodology to a real dataset. Finally, in Section 3.6 general remarks and specific considerations are given.

3.2 Related literature

The proposed methodology for clustering three-way data has some features that are similar and different with respect to past approaches. However, all methodologies

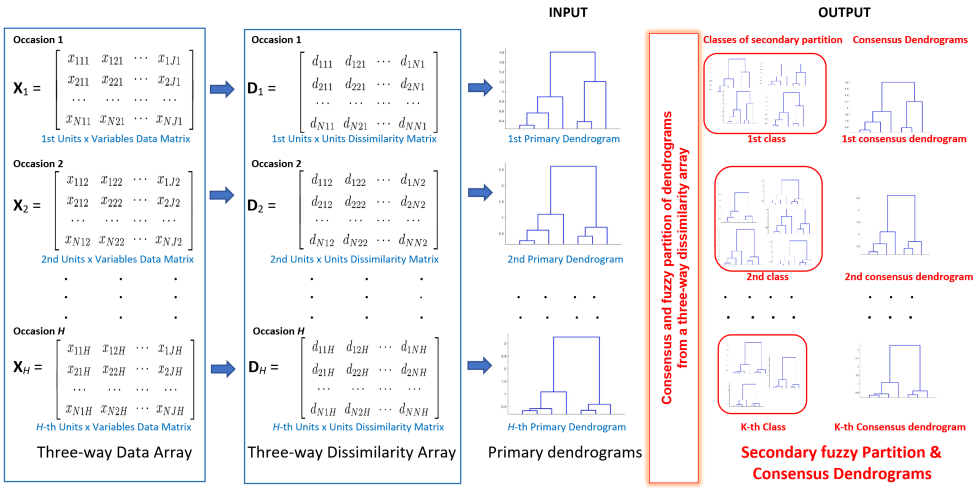


Figure 3.1. Flowchart describing the proposed methodology PARoDENo3WD

proposed in the literature either use a hard approach only, or do not aim to identify different primary consensuses. More precisely, they define different consensus data matrices, i.e., views of the data, and a hard secondary partition for occasions. In a maximum likelihood framework, Cappozzo, Alessandro, Michael, et al. (2021) applied a clustering technique to perform a hard partition of the occasions. Similarly, Cariou, Alexandre-Gouabau, and Wilderjans (2021) proposed a constrained algorithm to perform a hard partition of the occasions according to units and variables. A different problem is addressed by Schoonees, Groenen, and van de Velden (2021) who proposed a model to simultaneously partition the three modes in a hard way. Bocci and Vicari (2019) recommended another clustering algorithm in a three-way two-mode data framework. In the neuroscience framework, Durieux and Wilderjans (2019) transformed the three-way data into a two-way symmetric similarity matrix, then applied classical hard clustering algorithms, such as hierarchical clustering and Partitioning Around Medoids, to the corresponding dissimilarity matrix.

Among the most recent research works focused on three-way data structure, Yağ and Altan (2022) proposed to detect plant diseases using an optimization algorithm with low computational complexity to analyze a set of images (i.e. a set of 256×256 pixels matrices). In addition, it is worth mentioning Abu Arqub, Singh, and Alhodaly (2021) and Abu Arqub et al. (2021), who deeply investigated fuzziness in the mathematical field.

It must be noted that in many recent papers, the term 'multi-view data' refers to the notion of 'multi-way data arrays' and frequently 'three-way data'. The multi-view clustering is associated with the multi-view data: it aims to cluster the dataset with multiple views. A recent review of these methods is given in Fu et al. (2020). Multi-view clustering can be categorized into three typologies based on the approach used in the clustering process (Yin et al., 2015). Algorithms of the first typology find a unique consensus subspace (low-rank data matrix) of the different views, then cluster the data using a cluster analysis algorithm directly on the low-rank data matrix or on the (dis)similarity matrix associated with it. It is worth mentioning Yu et al. (2020) who focused on clustering multi-view data of high dimension with an active three-way clustering method; in the same context, Chao et al. (2019) proposed to use a multi-view co-cluster analysis that aims to partition objects into consistent clusters across the views. Their paper proposed a method to cope with missingness

problem, which is less sensitive to imputation uncertainty. In addition, Yang et al. (2022) proposed a multi-view robust clustering method to be used when the information is not complete, namely when either the assumption of view consistency or the assumption of instance completeness does not hold. In the framework of incomplete multiview clustering, Wen et al. (2022) provided a detailed review of the existing contributions on this topic.

The second type of multi-view clustering integrates the multi-view data into the clustering process. A well-known example is the co-EM algorithm (Bickel and Scheffer, 2004). The third type of multi-view clustering is called multi-view ensemble clustering, where the final clustering result is derived from the integration of the different clustering views of the data (Hussain, Mushtaq, and Halim, 2014). Another proposal on this topic is the contribution of El Hajjar, Dornaika, and Abdallah (2022), who proposed multi-view clustering, by integrating two embeddings to overcome the limitation of the standard multi-view spectral clustering.

It is worth observing that all proposed multi-view clustering methods have attracted more and more attention because they are known to improve the single-view clustering performance. The main difference between the proposed methodology PARoDENo3WD and the multi-view clustering is that the former allows improvement of the single-view clustering, but also partitions the different clustering views into classes with the property that clustering views with a relevant membership degree for the same class are perceived as similar to one another.

3.3 PARtition of DENdrograms of 3-Way Data array (PARoDENo3WD)

A data set pertaining to the same sets of units and variables, observed on different occasions (i.e., a set of multivariate data matrices) may be arranged into a three-way array $\mathbf{X} = [x_{ijh}],\ i \in \mathcal{I},\ j \in \mathcal{J},\ h \in \mathcal{H}$ with three modes: units (rows), variables (columns), and occasions (times, layers). These data can be considered as the result of the observation, on N units, of J variables repeated for H occasions. The term 'way' refers to a dimension of the data, while the word 'mode' is reserved for the methods or models used to analyze the data (Kroonenberg, 2008). For an introductory discussion on multi-way data analysis, the reader may refer to Coppi and Bolasco (1989). In econometrics, when the dimension occasion is the time, the three-way arrays are referred to as balanced panel data (Diggle et al., 2002). In computer science, these data are referred to as data cubes and much research has been concentrated on the definition of a series of data exchange formats, support storage, and transmission, such as MDX, to allow data cube interoperability (Friedrich et al., 2021; Madaan and Gosain, 2022).

In this chapter, a three-way array of dissimilarity data is assumed to be observed $\mathbf{D} = [D_{ilh} :\ i, l \in \mathcal{I},\ h \in \mathcal{H}]$. Each dissimilarity matrix $\mathbf{D}_h$ included in $\mathbf{D}$ can be also obtained by computing a dissimilarity measure between each pair of units in $\mathbf{X}_h$. The three-way ultrametric matrix $\mathbf{U} = [u_{ilh} :\ i, l \in \mathcal{I}\ h \in \mathcal{H}]$, could be assumed to have been observed as the result of a data-gathering process as discussed in the introduction. The ultrametric matrices $\mathbf{U}_1, \ldots, \mathbf{U}_H$ could be also computed by applying a hierarchical clustering algorithm to the H dissimilarity matrices $\mathbf{D}_1, \ldots, \mathbf{D}_H$. Indeed, u_{ilh} measures the difference between i-th and l-th units in the h-th primary classification, indicating the value of fusion of the smallest subset containing both units.

We are now in a position to state the new proposed methodology.

Given the three-way ultrametric matrix $\mathbf{U}$ formed by ultrametric matrices

$[\mathbf{U}_1, ..., \mathbf{U}_H]$, the problem of finding a fuzzy secondary partition into K clusters of the H primary dendrograms and of identifying within each class of the secondary partition a consensus ultrametric matrix is mathematically formulated, according to the following quadratic constraint optimization problem with respect to continuous variables $\mathbf{U}_k^*$, μ_{hk} and m:

$$
\begin{cases}
\text{minimize} \sum_{k=1}^{K}\sum_{h=1}^{H} \|\mathbf{U}_h - \mathbf{U}_k^*\|^2 \mu_{hk}^m = \\
\text{minimize} \sum_{i=1}^{N}\sum_{l=1}^{N}\sum_{k=1}^{K}\sum_{h=1}^{H} (u_{ilh} - u_{ilk}^*)^2 \mu_{hk}^m \\
\text{s.t.} \\
\sum_{k=1}^{K} \mu_{hk} = 1 & \text{for } h \in \mathcal{H} \\
\mu_{hk} \in [0,1] & \text{for } h \in \mathcal{H},\ k \in \mathcal{K} \\
u_{ilk}^* \leq \max\{u_{ipk}^*, u_{lpk}^*\} \\
u_{ipk}^* \leq \max\{u_{ilk}^*, u_{lpk}^*\} \\
u_{lpk}^* \leq \max\{u_{ipk}^*, u_{ilk}^*\} & \text{for } i = 1, \ldots, N-2, \\
& \quad l = i+1, \ldots, N-1, \\
& \quad p = l+1, \ldots, N
\end{cases}
\tag{3.P1}
$$

The number K of clusters of the secondary partition is a parameter of (3.P1) to be fixed *a priori*. The simulation study in Section 3.4 fully discusses the performances of different methods for choosing K. The first two constraints guarantee that $\mathbf{U}$, and therefore Δ, is fuzzily partitioned. The symbol μ_{hk} is the membership degree of h-th primary hierarchy in the k-th secondary consensus hierarchy, for $k \in \mathcal{K}$, for $h \in \mathcal{H}$. The membership degrees can assume values between 0 and 1, i.e. $\mu_{hk} \in [0,1]$ and, for a given primary hierarchy, the sum of the membership degrees for all consensus hierarchies is equal to one. The fuzziness value, or *fuzzifier*, m is the second and last parameter of (3.P1) to be fixed *a priori*. It controls how fuzzy the partition of the primary hierarchies tends to be. For $m \to \infty$ the clustering tends to be maximally fuzzified, leading to the same constant membership degrees $\frac{1}{K}$; when $m \to 1$ the membership degree tends to be either 0 or 1 and the fuzzy approach becomes the classical hard one. Many investigators have carried out analyses using $m = 2$. We have adopted the choice of setting $m = 2$ as widely used in the literature. Clearly, it may happen that the choices of the two parameters K and m are not independent, and this aspect will be recalled in further developments. The last three constraints guarantee that each matrix $\mathbf{U}_k^*$ is ultrametric. The matrix $\mathbf{U}_k^*$ is the k-th Least Squares Secondary Consensus Dendrogram (k-LSSCD).

Remark 1. *The ultrametricity of $\mathbf{U}_k^*$ requires that $\mathcal{O}(N^3)$ triplets in $\mathbf{U}_k^*$ satisfy the ultrametric inequality $u_{il}^* \leq \max\{u_{ij}^*, u_{lj}^*\},\ \forall\ i,j,l \in \mathcal{I}$. An equivalent condition is that every triple of objects $i,j,l \in \mathcal{I}$ possesses the property that the two largest values in the set $\{u_{il}^*, u_{ij}^*, u_{lj}^*\}$ are equal.* ■

From Remark 1 it follows that the $\mathcal{O}(N^3)$ constraints guaranteeing the ultra-

metricity of the consensus matrices can be also expressed synthetically as follows:

$$\sum_{i=1}^{N-2} \sum_{\substack{l=i+1 \\ u^*_{ilk} \leq \min\{u^*_{ipk}, u^*_{lpk}\}}}^{N-1} \sum_{p=l+1}^{N} (u^*_{ipk} - u^*_{lpk})^2 = 0 \qquad \text{for } k \in \mathcal{K} \tag{3.1}$$

In other words, equation 3.1 holds because an ultrametric matrix is a Euclidean matrix where triplets of values represent the edges of equilateral or acute isosceles triangles of a Euclidean space. Thus, in each triplet, the largest two values must be equal and their squared difference equal to zero. Therefore, when $\mathbf{U}^*_k$ is not ultrametric constraint (3.1) forces it to be so.

Remark 2. *From Remark 1, problem (3.P1) can be stated synthetically as follows, by including Equation (3.1):*

$$\begin{cases} \textit{minimize} \sum_{k=1}^{K} \sum_{h=1}^{H} \|\mathbf{U}_h - \mathbf{U}^*_k\|^2 \mu^m_{hk} = \\ \textit{minimize} \sum_{i=1}^{N} \sum_{l=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H} (u_{ilh} - u^*_{ilk})^2 \mu^m_{hk} \\ \textit{s.t.} \\ \sum_{k=1}^{K} \mu_{hk} = 1 & \textit{for } h \in \mathcal{H} \\ \mu_{hk} \in [0,1] & \textit{for } h \in \mathcal{H},\ k \in \mathcal{K} \\ \sum_{i=1}^{N-2} \sum_{\substack{l=i+1 \\ u^*_{ilk} \leq \min\{u^*_{ipk}, u^*_{lpk}\}}}^{N-1} \sum_{p=l+1}^{N} (u^*_{ipk} - u^*_{lpk})^2 = 0 & \textit{for } k \in \mathcal{K} \end{cases} \tag{3.P2}$$

■

In problem (3.P1), the H ultrametric matrices can be replaced by H dissimilarity matrices, named *primary dissimilarity matrices* and Problem (3.P1) can be formulated as follows:

$$\begin{cases} \text{minimize} \sum_{h=1}^{H} \sum_{k=1}^{K} \|\mathbf{D}_h - \mathbf{U}^*_k\|^2 \mu^m_{hk} = \\ \text{minimize} \sum_{i=1}^{N} \sum_{l=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H} (d_{ilh} - u^*_{ilk})^2 \mu^m_{hk} \\ \text{s.t.} \\ \sum_{k=1}^{K} \mu_{hk} = 1 & \text{for } h \in \mathcal{H} \\ \mu_{hk} \in [0,1] & \text{for } h \in \mathcal{H},\ k \in \mathcal{K} \\ u^*_{ilk} \leq \max\{u^*_{ipk}, u^*_{lpk}\} & \\ u^*_{ipk} \leq \max\{u^*_{ilk}, u^*_{lpk}\} & \\ u^*_{lpk} \leq \max\{u^*_{ipk}, u^*_{ilk}\} & \text{for } i = 1, \ldots, N-2, \\ & \quad l = i+1, \ldots, N-1, \\ & \quad p = l+1, \ldots, N \end{cases} \tag{3.P3}$$

In (3.P3) the first two constraints guarantee that the set of matrices $\mathbf{D}$ is soft partitioned; the following three triplets of constraints guarantee that $\mathbf{U}_k^*$ is ultrametric.

3.3.1 Least-Squares Estimation

In order to solve (3.P3), it is worth noticing that it can be decomposed into two minimization sub-problems:

(A) Given $\hat{\mu}_{hk}$, the partial minimization of the objective function of (3.P3) with respect to secondary consensus matrices $\mathbf{U}_k^*$ is:

$$\begin{cases} \text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m = \\ \text{minimize} \sum_{i=1}^{N}\sum_{l=1}^{N}\sum_{k=1}^{K}\sum_{h=1}^{H} (d_{ilh} - u_{ilk}^*)^2 \hat{\mu}_{hk}^m \\ \text{s.t.} \\ u_{ilk}^* \leq \max\{u_{ipk}^*, u_{lpk}^*\} \\ u_{ipk}^* \leq \max\{u_{ilk}^*, u_{lpk}^*\} \\ u_{lpk}^* \leq \max\{u_{ipk}^*, u_{ilk}^*\} \qquad \text{for } i = 1,\ldots,N-2, \\ \qquad\qquad\qquad\qquad\quad l = i+1,\ldots,N-1, \\ \qquad\qquad\qquad\qquad\quad p = l+1,\ldots,N \end{cases} \tag{3.P4}$$

The solution to this sub-problem (A) can be found by using the Sequential Quadratic Programming (SQP) algorithm (Powell, 1983). SQP is an iterative method able to solve constrained non-linear optimization problems. SQP methods solve a sequence of optimization subproblems, each of which optimizes a quadratic model of the objective subject to a linearization of the constraints (Wright and Nocedal, 2006).

Remark 3. *It is worth noting that SQP can solve Problem (3.P4) with $\mathcal{O}(N^3)$ constraints on the ultrametricity of matrix $\mathbf{U}_k^*$ or the problem with the same objective function of Problem (3.P4) by considering just the single constraint on the ultrametricity of matrix $\mathbf{U}_k^*$ (Equation (3.1)).* ■

Remark 4. *The unconstrained least square solution of (3.P4) is given by $\bar{\mathbf{U}}_k$, for $k \in \mathcal{K}$, where*

$$\bar{\mathbf{U}}_k = \frac{1}{\sum_{h=1}^{H} \hat{\mu}_{hk}^m} \sum_{h=1}^{H} \hat{\mu}_{hk}^m \mathbf{D}_h, \tag{3.2}$$

that is, the arithmetic mean matrix of $\mathbf{D}_h$, for $h \in \mathcal{H}$, weighted by $\hat{\mu}_{hk}^m$. ■

Proof. It is needed to show that the solution in Equation 3.2 is a minimum for the objective function of Problem (3.P4), which is the following:

$$\mathcal{F} = \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m \tag{3.3}$$

To prove it, we show that $\bar{\mathbf{U}}_k$ expressed in Equation 3.2 is a critical point, i.e. $\frac{d\mathcal{F}}{d\mathbf{U}_k^*}$ is null if and only if Equation 3.2 holds. Then, it is needed to show that $\bar{\mathbf{U}}_k$ expressed in Equation 3.2 is a minimum for $\mathcal{F}$, that $\frac{d^2\mathcal{F}}{d\mathbf{U}_k^{*2}}\Big|_{\mathbf{U}_k^*=\bar{\mathbf{U}}_k}$ is larger than 0.
The first derivative of $\mathcal{F}$ w.r.t. $\mathbf{U}_k^*$ is

$$\frac{d\mathcal{F}}{d\mathbf{U}_k^*} = \sum_{h=1}^{H} \hat{\mu}_{hk}^m \frac{d}{d\mathbf{U}_k^*} tr[(\mathbf{D}_h - \mathbf{U}_k^*)'(\mathbf{D}_h - \mathbf{U}_k^*)] \tag{3.4}$$

Focusing on $tr[(\mathbf{D}_h - \mathbf{U}_k^*)'(\mathbf{D}_h - \mathbf{U}_k^*)]$, we have that:

$$\begin{aligned} & tr[(\mathbf{D}_h - \mathbf{U}_k^*)'(\mathbf{D}_h - \mathbf{U}_k^*)] = \\ &= tr[\mathbf{D}_h'\mathbf{D}_h - \mathbf{D}_h'\mathbf{U}_k^* - \mathbf{U}_k^{*\prime}\mathbf{D}_h + \mathbf{U}_k^{*\prime}\mathbf{U}_k^*] = \\ &= tr[\mathbf{D}_h'\mathbf{D}_h] - 2tr[\mathbf{U}_k^{*\prime}\mathbf{D}_h] + tr[\mathbf{U}_k^{*\prime}\mathbf{U}_k^*] \end{aligned} \tag{3.5}$$

where the last equality is given by the fact that $tr(\mathbf{AB}') = tr(\mathbf{A}'\mathbf{B})$. Moreover, by recalling some properties of the derivatives of the traces (see, for example, Petersen, Pedersen, et al. (2008)), we have:

$$\begin{aligned} & \frac{d}{d\mathbf{U}_k^*}(tr[\mathbf{D}_h'\mathbf{D}_h] - 2tr[\mathbf{U}_k^{*\prime}\mathbf{D}_h] + tr[\mathbf{U}_k^{*\prime}\mathbf{U}_k^*]) = \\ &= \frac{d}{d\mathbf{U}_k^*} tr[\mathbf{D}_h'\mathbf{D}_h] - 2\frac{d}{d\mathbf{U}_k^*} tr[\mathbf{U}_k^{*\prime}\mathbf{D}_h] + \frac{d}{d\mathbf{U}_k^*} tr[\mathbf{U}_k^{*\prime}\mathbf{U}_k^*] = \\ &= 0 - 2\mathbf{D}_h + 2\mathbf{U}_k^* \end{aligned} \tag{3.6}$$

where the last equality is given by the fact that $\frac{d}{d\mathbf{A}} tr(\mathbf{A}'\mathbf{B}) = \mathbf{B}$ and $\frac{d}{d\mathbf{A}} tr(\mathbf{A}'\mathbf{A}) = 2\mathbf{A}$.
Then, Equation 3.4 becomes:

$$\frac{d\mathcal{F}}{d\mathbf{U}_k^*} = \sum_{h=1}^{H} \hat{\mu}_{hk}^m(-2\mathbf{D}_h + 2\mathbf{U}_k^*) = -2\sum_{h=1}^{H} \hat{\mu}_{hk}^m(\mathbf{D}_h - \mathbf{U}_k^*) \tag{3.7}$$

Now, we set Equation 3.4 equal to 0 and we obtain

$$\begin{aligned} & \frac{d\mathcal{F}}{d\mathbf{U}_k^*} = 0 \iff -2\sum_{h=1}^{H} \hat{\mu}_{hk}^m(\mathbf{D}_h - \mathbf{U}_k^*) = 0 \iff \\ & -2\sum_{h=1}^{H} \hat{\mu}_{hk}^m\mathbf{D}_h = -2\sum_{h=1}^{H} \hat{\mu}_{hk}^m\mathbf{U}_k^* \iff \sum_{h=1}^{H} \hat{\mu}_{hk}^m\mathbf{D}_h = \sum_{h=1}^{H} \hat{\mu}_{hk}^m\mathbf{U}_k^* \end{aligned} \tag{3.8}$$

Therefore, the critical point is:

$$\hat{\mathbf{U}}_k^* = \frac{1}{\sum_{h=1}^{H} \hat{\mu}_{hk}^m} \sum_{h=1}^{H} \hat{\mu}_{hk}^m\mathbf{D}_h \tag{3.9}$$

The second derivative of $\mathcal{F}$ w.r.t. $\mathbf{U}_k^*$ is

$$\frac{d^2\mathcal{F}}{d\mathbf{U}_k^{*2}} = -2\sum_{h=1}^{H} \hat{\mu}_{hk}^m(-1) = 2\sum_{h=1}^{H} \hat{\mu}_{hk}^m \tag{3.10}$$

which is always larger than 0. Therefore, $\frac{d^2\mathcal{F}}{d\mathbf{U}_k^{*2}}\Big|_{\mathbf{U}_k^*=\hat{\mathbf{U}}_k^*} > 0$.

Thus,

$$\hat{\mathbf{U}}_k^* = \frac{1}{\sum_{h=1}^{H}\hat{\mu}_{hk}^m}\sum_{h=1}^{H}\hat{\mu}_{hk}^m\mathbf{D}_h = \bar{\mathbf{U}}_k \tag{3.11}$$

is a minimum of $\mathcal{F}$. □

Typically, matrices $\bar{\mathbf{U}}_k$ are not ultrametric. However, using $\bar{\mathbf{U}}_k$ as initial values of (3.P4) the SQP algorithm generally stops at a stationary point in a few iterations, because the constrained solution generally needs only a few steps. Problem (3.P4) is equivalent to the following problem with respect to $\mathbf{U}_k^*$:

$$\begin{cases} \text{minimize } \sum_{k=1}^{K}\|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2\sum_{h=1}^{H}\hat{\mu}_{hk}^m = \\ \text{minimize } \sum_{i=1}^{N}\sum_{l=1}^{N}\sum_{k=1}^{K}(\bar{u}_{ilk} - u_{ilk}^*)^2\sum_{h=1}^{H}\hat{\mu}_{hk}^m \\ \text{s.t.} \\ u_{ilk}^* \le \max\{u_{ipk}^*, u_{lpk}^*\} \\ u_{ipk}^* \le \max\{u_{ilk}^*, u_{lpk}^*\} \\ u_{lpk}^* \le \max\{u_{ipk}^*, u_{ilk}^*\} \qquad \text{for } i = 1,\dots,N-2, \\ \qquad\qquad\qquad\qquad\quad l = i+1,\dots,N-1, \\ \qquad\qquad\qquad\qquad\quad p = l+1,\dots,N, \end{cases} \tag{3.P5}$$

To prove the equality between problems (3.P4) and (3.P5), we prove that the minimization of

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\|\mathbf{D}_h - \mathbf{U}_k^*\|^2\hat{\mu}_{hk}^m \tag{3.12}$$

w.r.t. $\mathbf{U}_k^*$ under the ultrametricity constraints is equivalent to the minimization of

$$\sum_{k=1}^{K}\|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2\sum_{h=1}^{H}\hat{\mu}_{hk}^m \tag{3.13}$$

w.r.t. $\mathbf{U}_k^*$ under the ultrametricity constraints. It has to be observed that the following decomposition holds:

$$\|\mathbf{D}_h - \mathbf{U}_k^*\|^2 = \|\mathbf{D}_h - \bar{\mathbf{U}}_k\|^2 + \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \tag{3.14}$$

Proof.

$$\begin{aligned} \|\mathbf{D}_h - \mathbf{U}_k^*\|^2 &= \|\mathbf{D}_h - \bar{\mathbf{U}}_k + \bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 = \\ &= tr[((\mathbf{D}_h - \bar{\mathbf{U}}_k) + (\bar{\mathbf{U}}_k - \mathbf{U}_k^*))'((\mathbf{D}_h - \bar{\mathbf{U}}_k) + (\bar{\mathbf{U}}_k - \mathbf{U}_k^*))] = \\ &= tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\mathbf{D}_h - \bar{\mathbf{U}}_k)] + tr[(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)] + \\ &+ tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)] + tr[(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)'(\mathbf{D}_h - \bar{\mathbf{U}}_k)] = \\ &= tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\mathbf{D}_h - \bar{\mathbf{U}}_k)] + tr[(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)] + \\ &+ 2tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)] = \\ &= \|\mathbf{D}_h - \bar{\mathbf{U}}_k\|^2 + \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 + 2tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)] \end{aligned} \tag{3.15}$$

The second-last equality is given by the fact that $tr[\mathbf{A}'\mathbf{B}] = tr[\mathbf{B}'\mathbf{A}]$. Note that $tr[(\mathbf{D}_h - \bar{\mathbf{U}}_k)'(\bar{\mathbf{U}}_k - \mathbf{U}_k^*)]$ is 0, as $\bar{\mathbf{U}}_\mathbf{k}$ is the weighted arithmetic mean matrix of matrices $\mathbf{D}_h$. □

Therefore, we have that

$$\text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m, \tag{3.16}$$

is equivalent to

$$\text{minimize } (\sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \bar{\mathbf{U}}_k\|^2 \hat{\mu}_{hk}^m + \sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m), \tag{3.17}$$

which is equivalent to

$$\text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \bar{\mathbf{U}}_k\|^2 \hat{\mu}_{hk}^m + \text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m \tag{3.18}$$

because, given $\hat{\mu}_{hk}$, the two minimization problems are independent.
Moreover, we observe that $\sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \bar{\mathbf{U}}_k\|^2 \hat{\mu}_{hk}^m$ is already minimized being $\bar{\mathbf{U}}_k$ the minimum for the unconstrained version of Problem (3.P4).
Therefore, the solution of Equation 3.16 is equivalent to the solution of the minimization of $\sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \hat{\mu}_{hk}^m$, which can be also written as $\sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \sum_{h=1}^{H} \hat{\mu}_{hk}^m$. Given $\hat{\mu}_{hk}$, then, the minimization of $\sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - \mathbf{U}_k^*\|^2 \sum_{h=1}^{H} \hat{\mu}_{hk}^m$ can be obtained by solving K separate independent minimization problems with respect to $\mathbf{U}_k^*$ under the ultrametricity constraints. This holds because for each $k \in \mathcal{K}$, $\sum_{h=1}^{H} \hat{\mu}_{hk}^m$ is a constant in the objective function. Problem (3.P5) can be solved by using SQP. An alternative way to optimize (3.P5) is to apply the average linkage method (UPGMA) on matrices $\bar{\mathbf{U}}_k$, for $k \in \mathcal{K}$, because UPGMA is well-known to find a LS solution to (3.P5). In fact, (3.P5) transforms the dissimilarity matrix $\bar{\mathbf{U}}_k$ into the closest ultrametric matrix.

(B) Given $\hat{\mathbf{U}}_k^*$, the partial minimization of the objective function of (3.P3) with respect to the membership degree μ_{hk} is:

$$\begin{cases} \text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{D}_h - \hat{\mathbf{U}}_k^*\|^2 \mu_{hk}^m & \text{(3.P6)} \\ \text{s.t.} & \\ \sum_{k=1}^{K} \mu_{hk} = 1 & \text{for } h \in \mathcal{H} \\ \mu_{hk} \in [0,1] & \text{for } h \in \mathcal{H},\ k \in \mathcal{K} \end{cases}$$

The minimization of sub-problem (B) is obtained by solving it using the first-order conditions for stationarity. Indeed, the stationary point can be found by considering the Lagrangian function

$$\mathcal{L} = \sum_{h=1}^{H}\sum_{k=1}^{K} \left\|\mathbf{D}_h - \hat{\mathbf{U}}_k^*\right\|^2 \mu_{hk}^m - \sum_{h=1}^{H} \lambda_h \left(\sum_{k=1}^{K} \mu_{hk} - 1\right), \tag{3.19}$$

where the solution with respect to μ_{hk} is

$$\mu_{hk} = \frac{1}{\sum_{k'=1}^{K} \left(d_{hk}/d_{hk'}\right)^{\frac{2}{m-1}}}, \qquad \text{for } h \in \mathcal{H},\ k \in \mathcal{K}. \tag{3.20}$$

where $d_{lp} = tr[(\mathbf{D}_l - \hat{\mathbf{U}}_p^*)'(\mathbf{D}_l - \hat{\mathbf{U}}_p^*)]$.

Proof. To prove the solution stated in Equation 3.20, the Lagrangian function in Equation 3.19 has to be derived with respect to μ_{hk} and with respect to λ_h. Then, set the two derivatives equal to zero. In the following, for the sake of brevity and simplicity, we let d_{hk} denote $\left\|\mathbf{D}_h - \hat{\mathbf{U}}_k^*\right\| = tr[(\mathbf{D}_h - \hat{\mathbf{U}}_k^*)'(\mathbf{D}_h - \hat{\mathbf{U}}_k^*)]$.
The first derivative with respect to μ_{hk} is

$$\frac{d}{d\mu_{hk}}\mathcal{L} = d_{hk}^2 m \mu_{hk}^{m-1} - \lambda_h \tag{3.21}$$

The first derivative with respect to λ_h is

$$\frac{d}{d\lambda_h}\mathcal{L} = \sum_{k=1}^{K} \mu_{hk} - 1 \tag{3.22}$$

By setting equal to zero Equation 3.21, we obtain:

$$\frac{d}{d\mu_{hk}}\mathcal{L} = d_{hk}^2 m \mu_{hk}^{m-1} - \lambda_h = 0 \iff \mu_{hk} = \left(\frac{\lambda_h}{m d_{hk}^2}\right)^{\frac{1}{m-1}} \tag{3.23}$$

By setting equal to zero Equation 3.22, we obtain:

$$\frac{d}{d\lambda_h}\mathcal{L} = \sum_{k=1}^{K} \mu_{hk} - 1 = 0 \iff \sum_{k=1}^{K} \mu_{hk} = 1 \tag{3.24}$$

By inserting Equation 3.24 into Equation 3.23:

$$\sum_{k'=1}^{K} \mu_{hk'} = \sum_{k'=1}^{K} \left(\frac{\lambda_h}{m d_{hk'}^2}\right)^{\frac{1}{m-1}} = 1 \iff \left(\frac{\lambda_h}{m}\right)^{\frac{1}{m-1}} \sum_{k'=1}^{K} \left(\frac{1}{d_{hk'}^2}\right)^{\frac{1}{m-1}} = 1 \tag{3.25}$$

$$\left(\frac{\lambda_h}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{k'=1}^{K} \left(\frac{1}{d_{hk'}^2}\right)^{\frac{1}{m-1}}} \tag{3.26}$$

By inserting Equation 3.26 into Equation 3.23,

$$\begin{aligned} \mu_{hk} = \left(\frac{\lambda_h}{m d_{hk}^2}\right)^{\frac{1}{m-1}} &\iff \mu_{hk} = \frac{1}{\sum_{k'=1}^{K} \left(\frac{d_{hk}^2}{d_{hk'}^2}\right)^{\frac{1}{m-1}}} \iff \\ &\iff \mu_{hk} = \frac{1}{\sum_{k'=1}^{K} \left(\frac{d_{hk}}{d_{hk'}}\right)^{\frac{2}{m-1}}} \end{aligned} \tag{3.27}$$

Equation 3.27 proves the solution stated in Equation 3.20. □

After the solution of sub-problems (A) and (B), the objective function generally reduces w.r.t. the previous iteration, or at least does not increase. Thus, the two steps (A) and (B) are reiterated. Since the objective function is bounded below by zero, after some iterations the algorithm stops at a stationary point that is not guaranteed to be the global minimum of the problem. For this reason, the algorithm is recommended to be run from several initial starting points to improve the chance to identify the global optimal solution. The algorithm can now be presented.

ALGORITHM for (3.P3):

0. Initialization
   Set $t = 0$; $\epsilon > 0$ convergence constant; random initialization from a uniform distribution of the membership degree matrix $[\hat{\mu}_{hk}]$, $h \in \mathcal{H}$, $k \in \mathcal{K}$

1. Do $t = t + 1$

2. Given $\hat{\mu}_{hk}$, solve sub-problem (A)
   Compute $\bar{\mathbf{U}}_k$, for $k \in \mathcal{K}$ as follows

$$\bar{\mathbf{U}}_k = \frac{1}{\sum_{h=1}^{H} \hat{\mu}_{hk}^m} \sum_{h=1}^{H} \hat{\mu}_{hk}^m \mathbf{D}_h \tag{3.28}$$

   Solve (3.P5) by SQP algorithm.

3. Given $\hat{\mathbf{U}}_k$, solve sub-problem (B)
   The solution of (3.P6) is given by:

$$\mu_{hk} = \frac{1}{\sum_{j=1}^{K} \left(d_{hk}/d_{hj}\right)^{\frac{2}{m-1}}}, \qquad \text{for } h \in \mathcal{H},\ k \in \mathcal{K}. \tag{3.29}$$

   where $d_{lp} = tr[(\mathbf{D}_l - \hat{\mathbf{U}}_p^{(t)})'(\mathbf{D}_l - \hat{\mathbf{U}}_p^{(t)})]$.

4. Stopping rule
   Repeat steps 1-3 until the difference between the objective functions of (3.P3) at iteration t and $t-1$ is greater than ϵ.

Note that the algorithm generally stops after a few iterations. The optimal solution cannot be guaranteed; therefore, the researcher is advised to repeat the analysis from several random starts and retain the best solution.

3.4 Simulation study

To assess the performance of the proposed methodology, a large simulation study has been developed. Dissimilarity matrices have been simulated according to the following model:

$$\mathbf{D}_h = \mathbf{U}_k + \mathbf{E}_h \qquad \mathbf{D}_h \in \mathcal{G}_k \qquad \text{for } k \in \mathcal{K},\ h \in \mathcal{H}. \tag{3.30}$$

where $\mathbf{E_h}$, $h \in \mathcal{H}$ is a matrix of errors generated from a normal distribution and then symmetrized. The parametric uncertainty on the errors is introduced to assess the performance of the new methodology under different scenarios. However,

note that the proposed technique uses the LS estimation method and, therefore, it does not require knowledge of the parametric distribution of the data. Thus, each dissimilarity matrix $\mathbf{D}_h$ belonging to class $\mathcal{G}_k$ of the secondary partition is assumed to be reconstructed by the ultrametric matrix $\mathbf{U}_k$ that identifies a hierarchy (dendrogram).

Matlab software (MATLAB, 2021) was used to perform the analysis and to plot the results.

The simulation study is organized into four scenarios with two levels of errors. For each scenario and error level, 200 three-way matrices have been generated for a total of 1800 samples (for Scenario 0 only one error was considered and for Scenario 1 a total of 400 three-way matrices are generated for each level of error). Scenario 0 studies the number of random starts necessary to reduce the final local minimum occurrences. Scenario 1 assesses whether the proposed methodology recognizes the existing underlying hard partition, if each dissimilarity matrix is generated by a single consensus $\mathbf{U}_k$. Scenario 2 aims to assess whether the proposed methodology distinguishes between dissimilarity matrices characterized by a single consensus and those generated by more than one consensus, thus producing a fuzzy partition. Scenario 3 discusses the choice of the optimal number of clusters. Additional details on the settings of the four scenarios are reported in Sections 3.4.1, 3.4.2, 3.4.3, and 3.4.4.

The K ultrametric matrices $\mathbf{U}_k$ necessary to generate dissimilarity matrices $\mathbf{D}_h$ have been defined according to the Absenteeism at work Data Set, downloaded from the UCI Machine Learning repository, to use a realistic situation for the simulation study. The dendrograms associated with the K ultrametric consensus matrices are reported in Figure 3.2: each dendrogram refers to a season and describes the hierarchical classification of 19 employees in a workspace according to some personal characteristics and variables related to their absenteeism behavior. Then, adding the error matrices to the consensus $\mathbf{U}_k$, the dissimilarity matrices $\mathbf{D}_h$ will be generated in such a way that still contain $\mathbf{U}_k$, but not exactly (unless the error is so high to overwhelm the ultrametric structure) and represent new members of the cluster $\mathcal{G}_k$.

3.4.1 Scenario 0: assessment of random starts

The new methodology does not guarantee the identification of the global optimal solution. This is expected since the partitioning problem is known to be NP-hard (Křivánek and Morávek, 1986).

Thus, before presenting the results, it was necessary to focus on a preliminary Scenario, i.e. Scenario 0, where the choice of the number of random starts (*RndStarts*) is discussed. Indeed, in this Scenario, the new algorithm was run 200 times with a high level of error under the conditions of hard secondary partition of primary dissimilarity matrices (for details, see Scenario 1 in Section 3.4.2), by letting the number of *RndStarts* assume the values $[1, 2, 3, 5, 10, 20, 30, 40, 50]$.

3.4.2 Scenario 1: hard secondary partition of primary dissimilarity matrices

In the first scenario of the simulation study, for each of the 4 consensus ultrametric matrices, 3 dissimilarity matrices have been generated, to obtain 12 dissimilarity matrices. More formally,

$$\mathbf{D}_h = \mathbf{U}_k + \mathbf{E}_h \qquad \mathbf{D}_h \in \mathcal{G}_k \qquad \text{for } k = 1, \ldots, 4,\ h = 1, \ldots, 12. \tag{3.31}$$

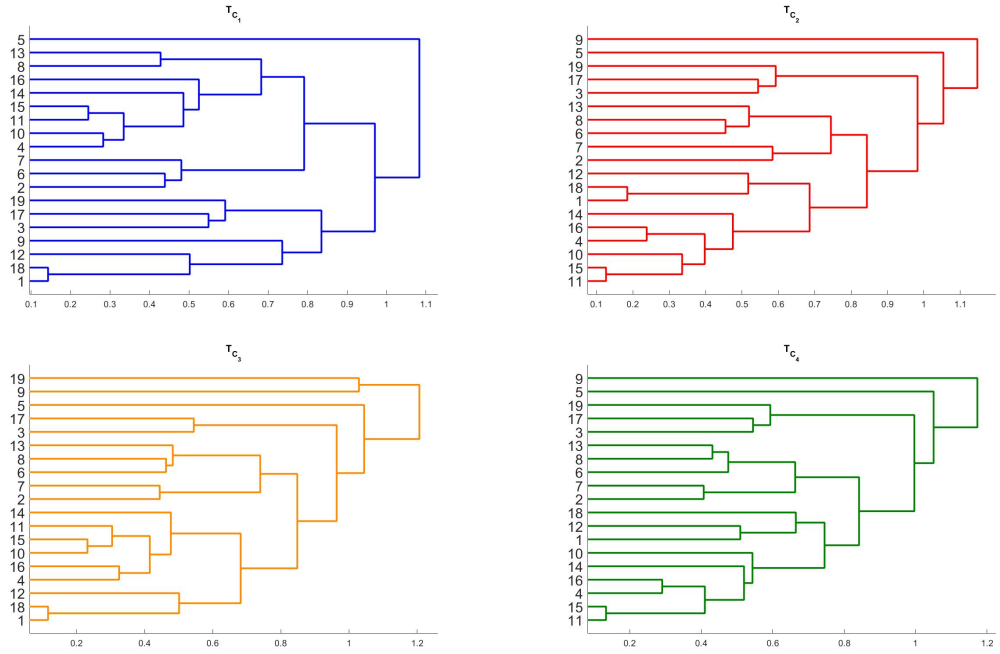


Figure 3.2. Consensus dendrograms of Absenteeism Data

Thus, from this generation process, the true hard secondary partition of the 12 matrices is known. Indeed, each dissimilarity matrix is generated by a unique consensus and, therefore, a hard secondary partition of the 12 matrices is expected, especially when the added error $\mathbf{E}_h$ is 'small'. On the other hand, an increasing '*fuzzification*' of the partition is foreseen with an increasing error $\mathbf{E}_h$ which will reduce the similarity of each dissimilarity matrix to the corresponding consensus.

Therefore, in this first scenario of the simulation study, 200 samples of three-way dissimilarity matrices are generated with size $(19 \times 19 \times 12)$, for two error levels, hence, for a total of 400 samples.

Note that the low error allows us to obtain 100% of ARI equal to 1, while the high error must allow the model to still hold in a majority of cases. Indeed, ARI is equal to 1 for 62% of cases, but ARI values show quite high variability, reaching a minimum value of 0.45.

3.4.3 Scenario 2: fuzzy secondary partition of primary dissimilarity matrices

The second scenario of the simulation study considers a mixed situation where some dissimilarity matrices are generated by a unique consensus matrix, as before, while some others are generated by the average of two consensuses.

More precisely, considering the ultrametric consensus matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, for each of the 2 consensus ultrametric matrices, 3 dissimilarity matrices were generated to obtain 6 dissimilarity matrices, according to the model, as follows:

$$\mathbf{D}_h = \mathbf{U}_k + \mathbf{E}_h \qquad \mathbf{D}_h \in \mathcal{G}_k \qquad \text{for } k = 1, 2,\ h = 1, \ldots, 6. \tag{3.32}$$

Moreover, 3 additional dissimilarity matrices were generated by

$$\mathbf{D}_l = \frac{\mathbf{U}_1 + \mathbf{U}_2}{2} + \mathbf{E}_l, \; l = 7, 8, 9, \tag{3.33}$$

where $\frac{\mathbf{U}_1+\mathbf{U}_2}{2}$ is the average consensus of the two original ultrametric matrices, $\mathbf{U}_1$ and $\mathbf{U}_2$. Therefore, in this case, the new methodology is applied to a total of $H = 9$ dissimilarity matrices.

Thus, for the first six generated matrices a (nearly) hard membership to the unique generating consensus is expected, while for the other three generated matrices a fuzzy membership degree to the consensuses included in the average is supposed. To assess whether the methodology recognizes the fuzzy nature of the last three dissimilarity matrices, a 'Detection of fuzziness' index is included and fully detailed in Section 3.4.5.

Therefore, in this second scenario of the simulation study, 200 samples of three-way dissimilarity matrices are generated with size $(19 \times 19 \times 9)$, with two error levels, hence, for a total of additional 400 samples. Also in this case two levels of errors are used: low error leads to a 100% of ARI equal to 1, while high error gives a 54% of ARI equal to 1. Moreover, the ARI distribution associated with the high error is highly variable, and, in the remaining 56% of cases, ARI takes values mainly between 0.6 and 0.9, but also reaches a minimum at level 0, as is observed in the corresponding boxplot.

Note that, in both Scenarios 1 and 2, the correct number of clusters is known *a priori* from how the data are generated from the model. Indeed, in Scenario 1, the data were produced from $K = 4$ original ultrametric matrices (dendrograms), while in Scenario 2, $K = 2$ ultrametric matrices were used. However, the choice of the number of clusters is an important issue that deserves careful consideration, because K is generally unknown *a priori*. Thus, Scenario 3 has been additionally considered in the simulation study to assess the performance of the new methodology using different well-known methods for choosing the number of clusters of the partition, as described in Section 3.4.4.

3.4.4 Scenario 3: assessment of K

This additional scenario was realized to assess the optimal number of clusters: indeed, even if the number of clusters is known *a priori* in the simulation study, in the applications generally K is not known. Thus, the choice of K was investigated using several indices. The first measure is the fuzzy version of the Silhouette index (Campello and Hruschka, 2006a); the pseudo F index (Caliński and Harabasz, 1974) and the Xie-Beni index (Xie and Beni, 1991) were then considered. The dissimilarity matrices were generated under both Scenario 1 and 2. In each Scenario, and for each level of error, the number of iterations was set equal to 200. At each iteration, when the dissimilarity matrices were generated under the conditions of Scenario 1, K is let to vary in the interval $[2, 6]$, while, under the conditions of Scenario 2, it is let to vary in the interval $[2, 4]$. Recall that the expected K was equal to 4 and 2, according to the generation process of Scenario 1 and Scenario 2, respectively. For each sample, the new algorithm was applied and the Fuzzy Silhouette, the pseudo F, and the Xie-Beni indices were computed. This was repeated for the different values of K. So, in the end, for each index, 5 and 3 (under Scenarios 1 and 2, respectively) final distributions of 200 values are obtained, each distribution corresponding to one specific K and each value corresponding to one specific sample. The procedure was repeated for both low and high levels of error.

3.4.5 Performance evaluation

The results of the simulation study on the performances of the new method have been evaluated using the following measures:

- ARI: Adjusted Rand Index (Hubert and Arabie, 1985). The ARI index measures the similarity between the hard partitions defined by the generated matrices when MAP ($\mathbf{D}_h$ is assigned to the cluster with maximal membership degree) is used for both the true partition and the partition provided by the new algorithm. It usually ranges in the interval $[0, 1]$: it is equal to 0 when the reference partition is compared to a partition that would be obtained just by chance and it is equal to 1 when the two partitions perfectly match. It can also yield negative values, meaning that the agreement between the two partitions is less than what is expected by chance (Wagner and Wagner, 2007).

- FUZZY ARI: Fuzzy Adjusted Rand Index (Campello, 2007). This is a fuzzy version of the ARI and compares the true (hard) partition and the fuzzy one obtained. The closer it is to 1, the more similar the two partitions.

- FUZZY RI: Fuzzy Rand Index (Campello, 2007). This is a fuzzy version of the RI index (Rand, 1971), ranging in the interval $[0, 1]$ and measuring the similarity between the true (hard) and the obtained fuzzy partitions: the closer it is to 1, the more similar the two partitions.

- RMSE: Root Mean Square Error. The RMSE is computed to quantify the difference between the original (generated) ultrametric matrices and those obtained by the algorithm. The RMSE is defined as the square root of the Mean Square Error between the K consensus ultrametrics provided by the new algorithm, and the K 'true' original ones. The lower the RMSE, the better the results.

- *Detection of fuzziness* index for Scenario 2. It is defined as the percentage of times based on 200 simulations the algorithm finds as membership degrees of the last three dissimilarity matrices two values similar to one another, i.e. close to 0.5, 0.5. The index provides a measure to assess how many times the algorithm is able to recognize the fuzzy nature of the last three dissimilarity matrices; when it is equal to 100%, the algorithm always softly assigns the dissimilarity matrices to clusters.

- Percentage of local minima: at each iteration, the value of the objective function for the generated true partition was computed and compared with the one that the algorithm returns as the optimal value. If the resulting objective function is higher than the true one, a local minimum has been found and underfitting occurred.

Moreover, to assess if the assignment is substantially *hard* or *fuzzy* as expected in Scenarios 1 and 2, the mean membership matrix is considered, i.e. the matrix obtained by averaging the obtained membership matrices over all the simulation's iterations.

3.4.6 Results of the simulation study under Scenario 0

The results are reported in Table 3.1. The percentage of local minima always decreases with the increase of the *RndStarts* until reaching 0, when the number of

random starts is set equal to 40. Thus, in each simulation study of Scenarios 1 and 2, the selected number of random starts for the whole simulation study was set to $RndStarts = 40$.

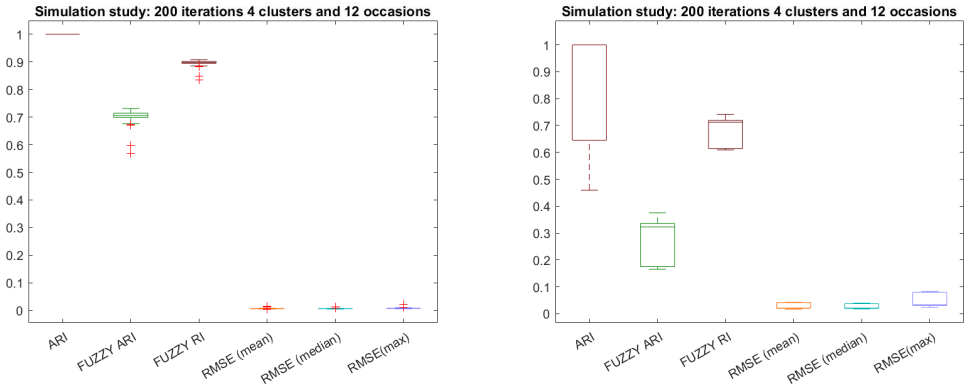
Table 3.1. Local minima occurrences (%) under Scenario 1 with high error.

RndStarts	1	2	3	5	10	20	30	40	50
%	37.5	28.5	26.5	17.5	9.5	5.0	2.5	0.0	0.0

From Table 4 it can be observed that with 1 random start of the algorithm and high error, the new methodology performs well, identifying the optimal solution in a large majority of cases (about 63%). It is worth noticing that with 20 random starts, the algorithm identifies the global optimal solution in 95% of cases, and with 40 random starts the algorithm was never trapped in local minima. Thus, in the following scenarios of the simulation study, the number of random starts was fixed to 40.

3.4.7 Results of the simulation study under Scenario 1

Figures 3.3 (a) and (b) show the results of the performances of the new methodology, for Scenario 1 (H=12, K=4) when low and high error levels are used; Table 3.2 lists the corresponding summary statistics of the performance indicators described in Section 3.4.5. Particularly, both the mean and the median of the distributions of the indices are reported. Moreover, it is important to clarify how mean RMSE, median RMSE, and max RMSE are computed. At each iteration, the mean, the median, and the maximum values of the 4 RMSE values are computed, where the 4 RMSE values compare the ultrametric matrices found by the algorithm and the true original ultrametric matrices. This is done in each iteration, to obtain the whole distribution.



(a) Low level of error **(b)** High level of error

Figure 3.3. Performance of the new methodology. Scenario 1 with low and high errors.

Table 3.2. Summary statistics. Scenario 1 with low and high errors.

Level of error	local minimum (%)	ARI (median, mean)	Fuzzy ARI (median, mean)	Fuzzy RI (median, mean)	mean RMSE (median, mean)	median RMSE (median, mean)	max RMSE (median, mean)
Low	0.000	1.000, 1.000	0.706, 0.704	0.898, 0.897	0.006, 0.006	0.006, 0.006	0.007, 0.008
High	0.000	1.000, 0.870	0.712, 0.673	0.323, 0.265	0.021, 0.029	0.021, 0.027	0.033, 0.052

When using low error, optimal performances of the new methodology are observed, with very high (close to 1) values for ARI, fuzzy ARI, and fuzzy RI, and low (close to 0) RMSE. Therefore, in all generated samples, on the one hand, the new methodology perfectly detects the true partition, assigning dissimilarity matrices to the correct clusters and, on the other, identifies as the consensus ultrametric matrix for each cluster of dissimilarity matrices the one that is very close to the true ultrametric matrix which has generated the data.

Note that the new methodology recognizes the generated hard partition and gives a membership value close to 1 to each dissimilarity matrix of the cluster. The mean membership matrix obtained averaging all the obtained membership matrices is reported in Table 3.3 (a) (results are approximated to 3 digits). Results confirm that the membership is always larger than 0.7 as expected.

When using high error, from Figure 3.3 (b) it can be observed that, even if in a large majority of times the new methodology recognizes the true partition, in more than the remaining 30% of cases, the ARI is quite different from 1, reaching a minimum at level of 0.45. Moreover, the fuzzy version of the ARI is quite low reaching a minimum at level of 0.2. This is due to the resulting membership matrix, which is no longer nearly hard, but closer to a fuzzier one. Indeed, looking at the memberships mean matrix in Table 3.3 (b), it is observed that the resulting partition has a fuzzier form: for example, the membership degrees of Cluster 2 and Cluster 4 for the second dissimilarity matrix are very close to each other. The fuzzy structure of Table 3.3 (b) represents what was expected in the setting of Scenario 1 (Section 3.4.2). Indeed, fuzzification of the resulting partition occurs because the error in the generating model is increased. This makes the algorithm misidentify the generating model and makes it almost fail to reconstruct the original consensus ultrametric matrices (see max RMSE in Table 3.2). However, note that to evaluate the partition with the ARI index, the MAP is applied to the resulting fuzzy membership matrices, and therefore this explains why most of the ARI turns out to be 1.

3.4.8 Results of the simulation study under Scenario 2

The performance of the methodology in recognizing, on the one hand, the hard nature of the first 6 generated dissimilarity matrices, and, on the other hand, the fuzzy nature of the last 3 generated dissimilarity matrices is now evaluated.

Here, in particular, it is worth focusing on the ability of the methodology to recognize the membership degrees that the new methodology assigns to the 'hybrid' generated dissimilarity matrices. Indeed, the new fuzzy methodology must also take into account the 'grey-scale' nature of some dissimilarity matrices: a matrix can indeed belong to one cluster, to one another, and even to both of them, with a membership degree that indicates the strength of that membership. Results for low and high errors are provided in Figure 3.4 (a) and (b), respectively and Table 3.4 lists the summary statistics. Particularly, both the mean and the median of the distributions of the indices are reported. Moreover, it is important to clarify what mean RMSE, median RMSE, and max RMSE mean: at each iteration, the mean, the median, and the maximum values of the 2 RMSE values are computed, where the 2 RMSE values compare the resulting and the true original ultrametric matrices. This is done in each iteration, to obtain the whole distribution.
By analyzing the low error results, in Figure 3.4 (a) and Table 3.4, the high ARI, fuzzy ARI, and fuzzy RI show that the methodology is able to recognize the underlying true partition.

Moreover, low values of RMSE mean that the obtained consensus ultrametric matrices are very close to the true ones. Finally, from the last column of Table 3.4,

Table 3.3. Mean membership matrices. Scenario 1 with low and high errors.

(a) Low level of error

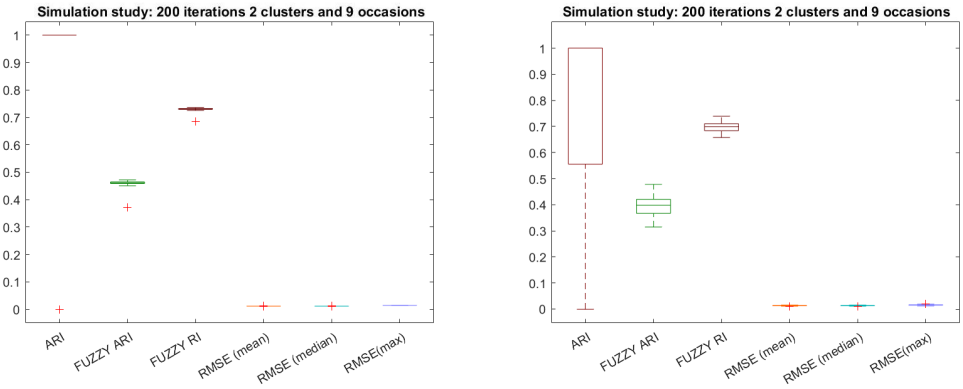
Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.903	0.037	0.024	0.036
0.030	**0.717**	0.068	0.185
0.022	0.077	**0.839**	0.062
0.029	0.187	0.054	**0.730**
0.903	0.037	0.024	0.036
0.030	**0.717**	0.068	0.185
0.022	0.078	**0.837**	0.063
0.029	0.188	0.054	**0.729**
0.901	0.037	0.025	0.037
0.030	**0.717**	0.068	0.185
0.022	0.077	**0.839**	0.062
0.029	0.188	0.055	**0.728**

(b) High level of error

Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.752	0.088	0.072	0.088
0.056	**0.369**	0.212	0.363
0.063	0.223	**0.494**	0.220
0.062	0.364	0.200	**0.374**
0.752	0.088	0.072	0.088
0.056	**0.368**	0.213	0.363
0.063	0.224	**0.492**	0.221
0.062	0.364	0.200	**0.374**
0.750	0.089	0.072	0.089
0.056	**0.369**	0.213	0.362
0.063	0.223	**0.493**	0.221
0.062	0.364	0.200	**0.374**

it is seen that the methodology always recognizes the fuzzy assignment of the last three dissimilarity matrices to the two clusters. In particular, Table 3.5 (a) (results are approximated to 3 digits) shows the membership degree matrix obtained by averaging over the whole simulation experiment.
From Table 3.5 (a), the behavior of the methodology is assessed. Indeed, for the first 6 dissimilarity matrices which are generated from one original ultrametric matrix, the corresponding membership values are close to 1, meaning that they are fully assigned to their cluster, in a hard sense. For the last 3 dissimilarity matrices, instead, the membership degrees fully respect their fuzzy nature: indeed, values corresponding to the first two clusters are very close to one another, meaning that the dissimilarity matrix belongs to both clusters, or in other words, is almost exactly in the middle of the two. This result confirms what was expected and highlights the flexibility of the new methodology when used in a fuzzy context.

Turning to the high level of error, good results are again observed by comparing the true and the obtained partitions, as the ARI, fuzzy ARI, fuzzy RI in Figure 3.4 (b) and Table 3.4 and the membership matrix in Table 3.5 (b) show. Moreover, Table 3.5 (b) shows that the methodology is able to recognize which dissimilarity



(a) Low level of error **(b)** High level of error

Figure 3.4. Performance of the new methodology. Scenario 2 with low and high errors.

Table 3.4. Summary statistics. Scenario 2 with low and high errors.

Level of error	local minimum (%)	ARI (median, mean)	Fuzzy ARI (median, mean)	Fuzzy RI (median, mean)	mean RMSE (median, mean)	median RMSE (median, mean)	max RMSE (median, mean)	Detection of fuzziness (%)
Low	0.000	1.000, 0.995	0.462, 0.461	0.731, 0.731	0.012, 0.012	0.012, 0.012	0.014, 0.014	100
High	0.000	1.000, 0.698	0.398, 0.394	0.699, 0.697	0.014, 0.014	0.014, 0.014	0.016, 0.016	100

Table 3.5. Mean membership matrices. Scenario 2 with low and high errors.

(a) Low level of error

Cluster 1	Cluster 2
0.995	0.005
0.011	**0.989**
0.995	0.005
0.011	**0.989**
0.995	0.005
0.011	**0.989**
0.454	**0.546**
0.454	**0.546**
0.454	**0.546**

(b) High level of error

Cluster 1	Cluster 2
0.940	0.060
0.061	**0.939**
0.940	0.060
0.062	**0.938**
0.940	0.060
0.061	**0.939**
0.462	**0.538**
0.460	**0.540**
0.460	**0.540**

matrices are fully (in a hard way) assigned to clusters and which are fuzzily assigned. Also in this case, it is able to reconstruct the true ultrametric matrices, as the low values of the RMSE show.

3.4.9 Results of the simulation study under Scenario 3

The optimal number of clusters has been evaluated using the Fuzzy Silhouette (Fuzzy Sil.) index (Campello and Hruschka, 2006a), the pseudo F index (Caliński and Harabasz, 1974), and the Xie-Beni (XB) index (Xie and Beni, 1991). Values of K that optimize such indices under Scenario 1 (Section 3.4.2) are reported in Table 3.6 (a) and (b), with low and high levels of error, respectively. Values of K that optimize such indices under Scenario 2 (Section 3.4.3) are reported in Table 3.7 (a) and (b), with low and high levels of error, respectively. Specifically, the Fuzzy Silhouette and the pseudo F indices are maximized, while the Xie-Beni index is minimized.

Table 3.6. Optimal number of clusters according to the Fuzzy Silhouette, Pseudo F, and Xie-Beni indices: occurrences (%) under Scenario 1 with low and high errors. Expected value $K^* = 4$.

(a) Low level of error

K*	Fuzzy Sil.	Pseudo F	XB index
2	0.0	0.0	0.0
3	0.0	0.0	0.0
4	100.0	100.0	100.0
5	0.0	0.0	0.0
6	0.0	0.0	0.0

(b) High level of error

K*	Fuzzy Sil.	Pseudo F	XB index
2	34.0	55.0	67.5
3	45.0	37.0	24.5
4	21.0	8.0	8.0
5	0.0	0.0	0.0
6	0.0	0.0	0.0

Table 3.7. Optimal number of clusters according to the Fuzzy Silhouette, Pseudo F, and Xie-Beni indices: occurrences (%) under Scenario 2 with low and high errors. Expected value $K^* = 2$.

(a) Low level of error

K*	Fuzzy Sil.	Pseudo F	XB index
2	100.0	100.0	100.0
3	0.0	0.0	0.0
4	0.0	0.0	0.0

(b) High level of error

K*	Fuzzy Sil.	Pseudo F	XB index
2	75	78	59.5
3	19	20.5	36.5
4	6	1.5	4

Table 3.6 (a) shows that in all the 200 iterations, when generating the dissimilarity matrices under Scenario 1 with low error, the methodology detects the true number of clusters ($K^* = 4$), according to all the indices. The same happens when the dissimilarity matrices are generated under Scenario 2 with low error (see Table 3.7 (a)). When using high error under the conditions of Scenario 2, Table 3.7 (b) shows that in the majority of the simulations the algorithm detects again the true number of clusters ($K^* = 2$). The high error causes the indices to be more imprecise. However, all the indices detect simultaneously $K^* = 2$ as optimum in more than half of the simulations (53.5%). On the contrary, Table 3.6 (b) shows that under the conditions of Scenario 1 according to both XB and Pseudo F indices, the methodology chooses $K^* = 2$ as the optimal number of clusters, as this is the mode of the distributions for these two indices. Instead, Fuzzy Sil. leads the methodology to choosing $K^* = 3$. It is worth mentioning that the methodology in some occasions leads to choose

$K^* = 4$: in particular, Fuzzy Sil. selects $K^* = 4$ in 21% of the iterations; instead, XB and Pseudo F return $K^* = 4$ in only 8% of the iterations, the second of these being a conservative index. In more details, XB, and Pseudo F simultaneously select $K^* = 4$ as optimum, while all the three indices simultaneously return $K^* = 4$ in only 6% of the iterations. Therefore, the optimal $K^* = 4$ is chosen quite rarely, probably because the error too strongly masks the four ultrametric matrices. Indeed, when using high error the membership degree matrices tend to be very fuzzy, i.e. each dissimilarity matrix tends to have membership degrees very similar to one another: consequently, the level of error makes the indices suffer in recognizing that the true partition is given by four ultrametrics. For this reason, additional methods for choosing the number of clusters should be investigated and a specific simulation study fully dedicated to the choice of K should be carried out. This is beyond the scope of this work, but it is emphasized that additional investigation is needed. However, it is recommended to base the decision regarding the selection of the number of clusters not only on the use of indices appropriately designed to select the optimal K, but also on the interpretation of the resulting partitions.

3.5 Real dataset application

To fully motivate the utility of the methodology that has been proposed, and clearly illustrate the concepts of single secondary consensus dendrogram, and multiple secondary consensus dendrograms with a fuzzy partition of primary dendrograms, a real data-set showing the long-term scenario 2005-2020 of macroeconomic performances, as measured by 6 major economic indicators, of the national economies of the G7 most-industrialized countries is considered. These panel data form a three-way data array ($7 \times 6 \times 16$), where the first dimension represents the G7 countries: Canada (CAN,1), Germany (DEU,2), France (FRA,3), Great Britain (GBR,4), Italy (ITA,5), Japan (JPN,6), United States of America (USA,7); while the second dimension corresponds to the six economic indicators: Gross Domestic Product (US dollars/capita) measures the value added created through the production of goods and services in a country during a year (GDP); Composite Leading Indicator shows short-term economic movements in qualitative rather than the quantitative term (Long-term average =100) (CLI); Long-Term Unemployment Rate for people who have been unemployed for 12 months or more shows the proportion of these long-term unemployed among all unemployed (LUR); Short-term Interest rate measures the rates at which short-term borrowings are effected between financial institutions (SIR); Current Account Balance of payments measures the country's international transactions with the rest of the world as a percentage of the gross domestic product (CAB); and Saving Rate (SR) measures the difference between disposable income and final consumption expenditure divided by gross domestic product (SR). Finally, the third dimension of the three-way data array regards the time period from 2005 to 2020. These data have been downloaded from OECD.Stat which includes data and metadata for OECD countries and is frequently used by OECD to depict the macroeconomic outlook.

Each data matrix (units-by-variables (7×6)) $\mathbf{X}_h$, for $h = 1, \ldots, 16$, (period $2005, \ldots, 2020$) has been normalized by using the min-max normalization, where the min and the max of variables are over the entire period 2005-2020. For each matrix $\mathbf{X}_h$ the Euclidean (7×7) distance matrix $\mathbf{D}_h$ between units has been computed. In Figure 3.5 the dendrograms $\Delta = \{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_{16}\}$ of Ward's method of hierarchical clustering, computed on matrices $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{16}$ are shown for the years from 2005 to 2020. The hierarchical clustering of the G7 countries in different years exposes

dissimilarities and convergences in the economic characteristics of G7 countries.

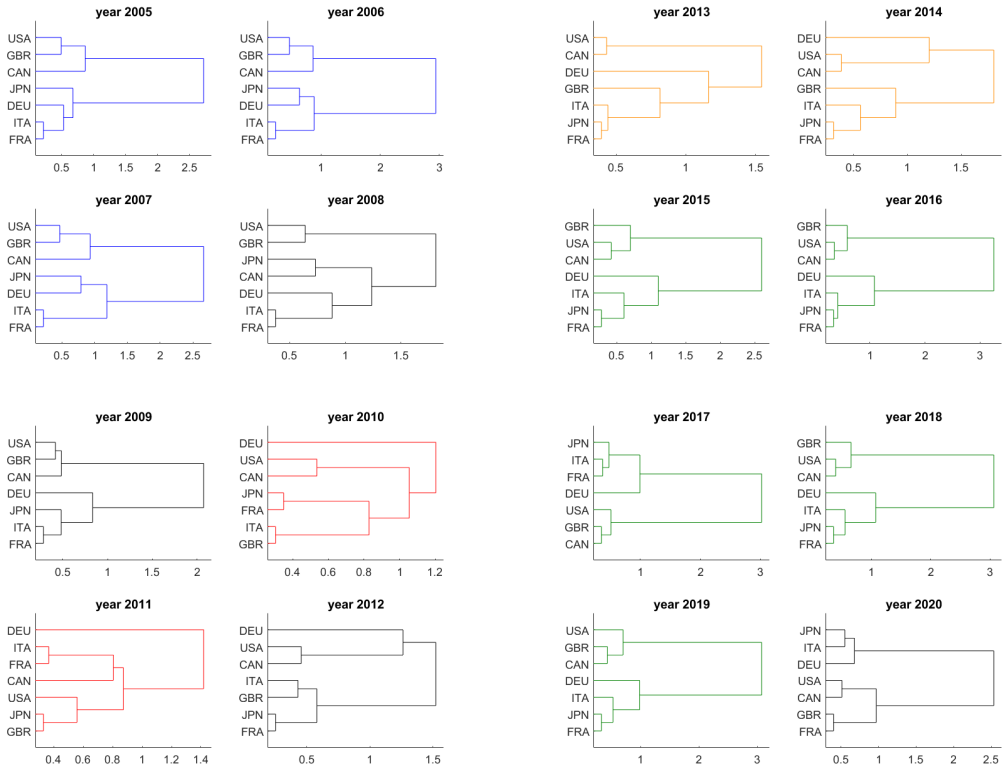


Figure 3.5. Hierarchical clustering of G7 countries by 6 economic variables from 2005 to 2020

The hierarchical clustering of the G7 countries is obtained by one single or multiple consensuses of primary dendrograms of the different years. A single consensus would be sufficient if all the dendrograms from 2005 to 2020 were similar or nearly so. To assess the similarity between dendrograms, we examine the associated N-trees (as defined in Section 2). Concerning the real dataset, the N-trees of the dendrograms from 2005 to 2020, which include only internal nodes of the tree, and thus exclude singleton clusters $\{i\}$, $i = 1, \ldots, 7$, are shown in Table 3.8. Two dendrograms $\boldsymbol{\delta}_h$ and $\boldsymbol{\delta}_m$ are similar if they have the same N-trees $\mathbf{T}_h = \mathbf{T}_m$. When dendrograms are similar, they have a perfect strict consensus tree (Sokal and Rohlf, 1981) since each class of the consensus belongs to each original similar tree. It is 'perfect' because 100% of classes coincide between dendrograms. For example, the dendrogram of 2006 is similar to the one of 2007. They differ only according to the levels of aggregation and therefore they may be represented by a perfect strict consensus tree. However, during 2008-2009 -probably due to the deep economic crisis- countries' economies start to change and different clusters of countries appear. As a consequence, the dendrograms change more radically from the previous ones. Another three similar dendrograms are indeed found only several years later in 2015, 2016, and 2018. However, these last ones show different initial aggregations of countries from those of 2006 and 2007. This suggests that multiple secondary consensuses are needed to show the clustering changes for the entire period. On the other hand, dendrograms in different years have some classes of countries in common. For example {FRA, ITA} and {GBR, USA} or {CAN, USA} are frequently found in all dendrograms. This

suggests that all dendrograms may contribute to defining the consensuses. Thus, great flexibility in the new clustering methodology is needed to understand and assess how much these common classes contribute to the definition of the consensuses. Vichi (1998), with Principal Classification Analysis (PRINCLA), considers as consensuses the orthogonal linear combinations of the original dendrograms. The orthogonality produces a hard *pseudo*-partition of dendrograms (it is not exactly a partition, as it excludes dendrograms that alone contribute to one principal component), so only disjoint subsets of original dendrograms may contribute to defining the consensuses. Vichi (1999), with Partition and Least-Squares Consensus Classification Analysis (PARSCLA), defines the partition of dendrograms in K homogeneous classes and the LS consensus for each class. Therefore, PRINCLA and PARLSCLA identify consensuses related to disjoint subsets of dendrograms. In other words, an original dendrogram can contribute to the definition of a unique consensus only. In this work, to allow the required flexibility to define consensuses with the contribution of all original primary dendrograms, the Least-Square fuzzy secondary partition of the dendrograms with a consensus for each class is proposed.

Table 3.8. The N-trees of the dendrograms from 2005 to 2020; these exclude singleton clusters {i}, i=1,..,7.

T_{2005}={{FRA,ITA},{GBR,USA},{DEU,FRA,ITA},{DEU,FRA,ITA,JPN},{CAN,GBR,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2006}={{FRA,ITA},{GBR,USA},{DEU,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2007}={{FRA,ITA},{GBR,USA},{DEU,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2008}={{FRA,ITA},{GBR,USA},{CAN,JPN},{DEU,FRA,ITA},{CAN,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2009}={{FRA,ITA},{GBR,USA},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2010}={{GBR,ITA},{FRA,JPN},{CAN,USA},{CAN,GBR,USA},{CAN,FRA,GBR,ITA,JPN,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2011}={{GBR,JPN},{FRA,ITA},{GBR,JPN,USA},{CAN,FRA,ITA},{CAN,FRA,GBR,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2012}={{FRA,JPN},{GBR,ITA},{CAN,USA},{FRA,GBR,ITA,JPN},{DEU,GBR,JPN,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2013}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{FRA,GBR,ITA,JPN},{DEU,FRA,GBR,JPN,ITA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2014}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{FRA,GBR,ITA,JPN},{CAN,DEU,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2015}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2016}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2017}={{CAN,GBR},{FRA,ITA},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN,},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2018}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{2019}={{FRA,JPN},{CAN,GBR},{FRA,ITA,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

T_{2020}={{FRA,GBR},{CAN,USA},{ITA,JPN},{CAN,GBR,USA},{DEU,ITA,JPN,},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

Indeed, the methodology proposed here, when applied to the OECD panel data, gives very reliable results with a fuzzy partition of the dendrograms and a consensus dendrogram for each class of the partition. The 16 original dendrograms, shown in Figure 3.5, have been split into 4 clusters (the number K=4 was chosen by using the Fuzzy Silhouette index, see Section 3.4.4 and 3.4.9) with 4 consensus dendrograms, and the corresponding N-trees reported in Figure 3.6 and Table 3.9, respectively.

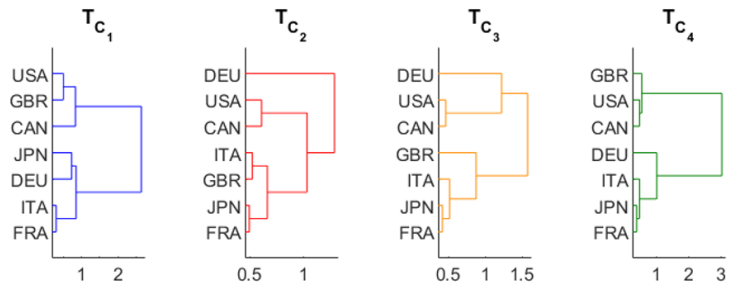


Figure 3.6. Resulting consensus dendrograms, representing hierarchical clustering of G7 countries by 6 economic variables

Focusing on the memberships of the original dendrograms to each cluster of the secondary partition, it can be observed that some dendrograms have nearly hardly assignments to one cluster, but also that some others are softly assigned to more than one cluster. The highest memberships of the 16 dendrograms to the 4 clusters of dendrograms of the new proposed methodology are reported in Table 3.10. It can be observed that dendrograms corresponding to the years 2015, 2016, 2017, 2018, and 2019 have a nearly hard membership to the fourth cluster with consensus T_{C_4} and indeed, these dendrograms share with the corresponding consensus dendrogram almost all clusters.

Dendrograms which are more flexibly associated with the resulting consensus ultrametrics are those corresponding to the years 2008, 2012, and 2020 which identify three distinct periods. The year 2008 may be considered the initial year of the great recession observed globally in national economies, while the year 2012 can be seen as the first year in which there was partial recovery from the recession (source: Investopedia). Finally, the year 2020 relates to the COVID-19 recession. In 2008, but also 2009, dendrograms have clusters between those of the period 2005-2007 (cluster 1) and those of the period 2010-2011 (cluster 2). The dendrogram for 2008 is characterized, among the others, by classes {FRA, ITA}, and {GBR, USA} that are typical of the period 2005-2007. Germany in 2008 started to show a larger distance with respect to the other countries and this is typical of the dendrograms for 2010, 2011, and 2012. The dendrogram related to the year 2012 has characteristics of the dendrograms of the period 2010-2011 and the period 2013-2014, sharing with them the classes {CAN, USA}, {FRA, JPN}. Finally, the dendrogram related to the year 2020 has a class {CAN, USA} that makes it belong to both Clusters 3 and 4 and a class {CAN, GRB, USA} that makes it also belong to Clusters 1 and 2 and increases the membership to Cluster 3. Thus, it can be observed that dendrograms for the years 2008, 2009, 2012, and 2020 reflect the changes in the heterogeneity of the G7 countries, as their soft fuzzy membership confirms. To be more precise, the changes in the hierarchical clustering of the G7 countries are due to the different economic changes in the period 2005 - 2020.

The reader can notice that clusters substantially keep within them the chronological order of the dendrograms.

Table 3.9. The N-trees of the resulting consensus dendrograms representing hierarchical clustering of G7 countries by 6 economic variables; these exclude singleton clusters {i}, i=1,..,7.

T_{C_1}={{ITA,FRA},{GBR,USA},{DEU,JPN},{CAN,GBR,USA},{DEU,FRA,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{C_2}={{FRA,JPN},{GBR,ITA},{CAN,USA},{FRA,GBR,ITA,JPN},{CAN,FRA,GBR,ITA,JPN,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{C_3}={{FRA,JPN},{CAN,USA},{FRA,ITA,JPN},{FRA,GBR,ITA,JPN},{CAN,DEU,USA},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}
T_{C_4}={{FRA,JPN},{FRA,ITA,JPN},{CAN,USA},{CAN,GBR,USA},{FRA,DEU,ITA,JPN},{CAN,DEU,FRA,GBR,ITA,JPN,USA}}

Table 3.10. Cluster assignment of the original dendrograms to 4 clusters, with the highest membership degree.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Cluster	1	1	1	1, 2	1, 3	2	2	2, 3	3	3	4	4	4	4	4	1,2,3,4
Membership degree	0.79	0.77	0.77	0.28,0.25	0.34, 0.30	0.89	0.78	0.46, 0.48	0.93	0.88	0.87	0.87	0.91	0.87	0.93	0.28, 0.23, 0.28, 0.21

3.6 Conclusion

A set of primary hierarchies (dendrograms) of the same set of units is often available to the researcher. For example, in economic applications, primary hier-

archies can be found by hierarchically clustering countries according to variables for each different year or applying, on the same dissimilarity matrix, different agglomerative algorithms. In psychometric studies, the hierarchies are obtained from data cards, a well-known data-gathering technique, while in marketing applications, hierarchies are given by the customers who rate and classify some products and then continue to classify classes to obtain a hierarchy. A relevant research problem discussed in this chapter is to find single or multiple consensuses of the set of primary dendrograms. When all primary hierarchies are similar to one another, a single secondary hierarchical consensus is sufficient. This is generally achieved by multi-view clustering methods that are highly appreciated because they improve single-view clustering. However, when primary hierarchies change quite drastically in the given set, a unique consensus of the whole set of primary hierarchies would be a too unrealistic and narrow synthesis of the dendrograms. In addition, it is also possible to hypothesize situations where groups of dendrograms may share some elements, i.e. the corresponding N-trees may have similar classes. In such situations, not only must the methodology be able to identify multiple consensuses, but it must also allow each dendrogram to contribute with a different weight in defining each consensus. Therefore, in this work, it is proposed to develop a clustering methodology, named PARoDENo3WD, that allows identification of classes of primary dendrograms perceived as similar (secondary partition), synthesis of each class of the primary dendrograms by a consensus dendrogram, and the use of a fuzzy approach to associate to each primary dendrogram a membership degree for each class of the secondary partition.

In the application to a real dataset described in Section 3.5, this methodology was extremely helpful to classify G7 countries observed in a period from 2005 to 2020 in order to show dissimilarities and convergences in the economic characteristics of G7 countries.

The proposed methodology PARoDENo3WD has also been extensively tested in a simulation study, by generating 1800 three-way datasets. In the design of the study, the parametric uncertainty has been defined by using a multivariate normal distribution. This was useful to compare the performance of the new methodology in different scenarios with different levels of error in the data. However, it is important to observe that PARoDENo3WD is a non-parametric methodology, which does not require knowing the data distribution and can also be applied for large three-way data arrays. A fuzzy clustering approach has been adopted to handle the clustering uncertainty, i.e., the uncertainty in the assignment of the units to clusters. In the simulation study, this uncertainty has been fully analyzed. The results show that PARoDENo3WD is able to identify both the hard and the fuzzy partitions of the set of dendrograms and also to summarize the starting dissimilarities using consensus dendrograms which are very close to the original generated ones. Moreover, the simulation was used to study the performance of several methodologies to evaluate how to choose the number of classes of the secondary fuzzy partition. The Fuzzy Silhouette, the Pseudo F, and the Xie-Beni indices were used to detect the number of classes. Specifically, they always select the true number of classes when a low level of error is used and naturally tend to be more imprecise when a high level of error is used. In addition, when the underlying partition is hard, the indices tend to underestimate the number of classes when a high level of error is used. PARoDENo3WD does not guarantee the global optimal solution because of the clustering problem that is well-known to be NP-hard. Hence, the simulation study analyzes the problem of local minimum solutions. It is observed that 40 random starts will drastically reduce this problem. The algorithm is fast, even facing some issues with data storage: indeed, dissimilarity matrices are of order N^2.

This work contributes to the introduction of a new methodology in the three-way clustering literature and the multi-view clustering and opens up the possibilities for new applications in such a framework. In addition, in order to face the complexity issue due to data storage, a future project could be the development of a new methodology based on a parsimonious approach, in such a way that the first aggregations of the unit in the dendrogram are ignored and only the ones starting from a specific level are considered by the algorithm. In this way, the secondary consensus dendrograms will have a parsimonious structure and the whole complexity will be reduced.
In addition, as highlighted during the illustration of the methodology, an important aspect that requires further consideration and study relates to the evaluation of how and to what extent the choice of the number of clusters affects the choice of fuzzyness parameters and vice versa. Moreover, it might be interesting to study in depth how sensitive the algorithm is to the choice of linkage method used in hierarchical clustering.

Chapter 4

Parsimonious consensus hierarchies, partitions and fuzzy partitioning of a set of hierarchies

4.1 Introduction

This chapter addresses the problem of obtaining partitions of the set of hierarchical classification of objects. These will be referred to as primary hierarchies with associated primary ultrametric matrices, knowing that there is a bijection between hierarchies and ultrametric matrices (Johnson, 1967). A fuzzy partition of a set of primary hierarchies will be referred to as a secondary fuzzy partition. The aim of the methodology described in this chapter is to obtain a secondary fuzzy partition of the set of primary hierarchies into classes with the property that primary hierarchies with a relevant membership degree for the same class are perceived as similar to one another. Each of the classes will have an associated parsimonious consensus hierarchy, which serves as a summary of the set of primary hierarchies belonging to the class. The secondary partition is *fuzzy* because it can describe some "uncertainties" that occur in the observed set of primary hierarchies and it provides further information: the membership degrees can show for each class which primary hierarchies are more strongly associated with it and which hierarchies have only a loose association. Therefore, each hierarchy contributes to the definition of all classes according to different membership degrees. The consensus hierarchy (tree) is *parsimonious*, because it limits its internal nodes to a reduced number G (where G is much smaller than the number N of objects). Thus, the parsimonious trees has the property that clusters appearing in excess of K are viewed as very close to each other and perceived as almost indistinguishable and irrelevant in the hierarchy. In addition, the consensus includes the optimal partition into K clusters. Frequently investigators wish to identify this optimal partition in the hierarchy to detect the most relevant classification of its nested partitions. A flowchart describing the methodology and clarifying the process is displayed in Figure 4.1. The flowchart is also useful to synthesize graphically the data structure.

The remainder of this chapter is organized as follows. Section 4.2 is fully dedicated to recall the terminology and the review of the literature; Section 4.3 describes the proposed methodology and its estimation. The performance of the new methodology is tested in an extended simulation study presented and discussed in Section 4.4

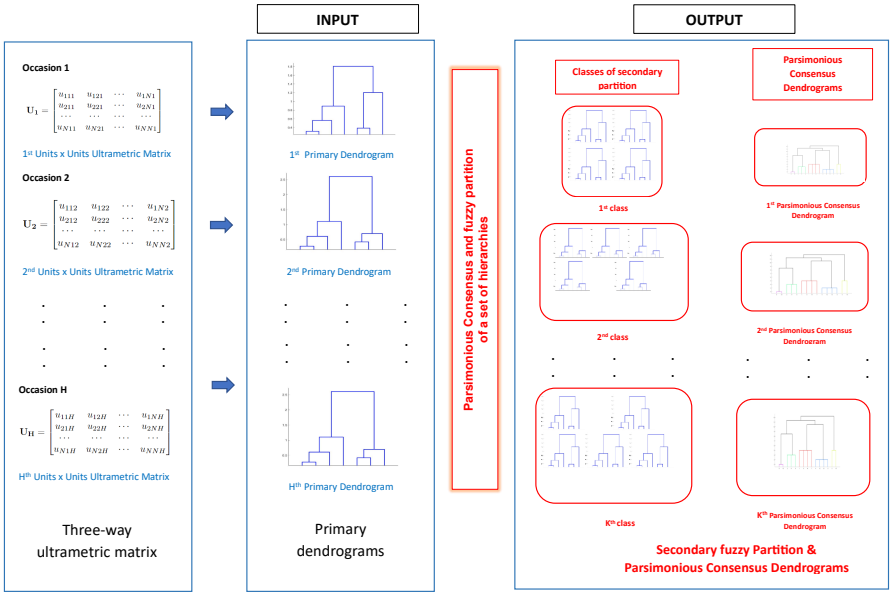


Figure 4.1. Flowchart describing the proposed methodology

and Section 4.5 includes the applications to real datasets. Finally, Section 4.6 gives remarks and considerations on future developments.

4.2 Notation and theoretical background

The notation used in this framework has been already formalized in Chapter 2; relatively to this chapter, the three-way consensus matrix $\mathbf{U}^* = [u^*_{ilk}]$ $(N \times N \times K)$ are formed by $(2G-1)$-ultrametric matrices, where G is the number of clusters forming a partition of the N units. Formally, $\mathbf{U}^* = [\mathbf{U}^*_1, \ldots, \mathbf{U}^*_K]$, where $\mathbf{U}^*_k$ is a $N \times N$ $(2G-1)$-ultrametric matrix, for $k = 1, \ldots, K$. The H primary hierarchies are supposed observed with associated ultrametric matrices $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_H$. When they are not observed, then they are built by applying a fixed hieararchical clustering algorithm (Gordon, 1999) to each dissimilarity matrix $\mathbf{D}_h$ related to the data matrix $\mathbf{X}_h$, $h = 1, \ldots, H$ which make up the three-way three-mode data array $\mathbf{X} = [x_{ijh} : i \in \mathcal{I}, j \in \mathcal{J}, h \in \mathcal{H}]$.

From the H given ultrametric matrices, K parsimonious consensus dendrograms, i.e. K $(2G-1)$-ultrametric matrices, summarizing the original H hierarchies will be identified. Each $(2G-1)$-ultrametric matrix, is a square N dimensional matrix with elements satisfying ultrametric inequalities and with off-diagonal elements that can assume one of at most $(2G-1)$ positive different values.

It is now necessary to introduce the model used to obtain a parsimonious tree, associated to a $(2G-1)$-ultrametric matrix.
Therefore the theoretical background on parsimonious hierarchy necessary for the reader to follow the new methodology is reported in the following. The theoretical data structure for multivariate objects and dissimilarities examined on different occasions used in this chapter is the three-way array.

4.2.1 Well-Structured Partition (WSP)

A partition of objects into G clusters has two main characteristics: the *isolation between clusters* and the *heterogeneity within clusters*. The partition is usually represented by a classification matrix, which is a clustering model for dissimilarities in the well-known model-based clustering framework. Rubin (1967) proposed to model the classification matrix by three matrices: the diagonal matrix ${}_W\mathbf{D}_k = [{}_Wd^k_{gg} > 0 : {}_Wd^k_{gt} = 0, g, t = 1, \ldots, G, \ (g \neq t)]$, the squared matrix ${}_B\mathbf{D}_k = [{}_Bd^k_{gt} > 0 : {}_Bd^k_{gg} = 0, t, g = 1, \ldots, G, \ (t \neq g)]$ and the membership matrix $\mathbf{M}_k = [m^k_{ig} : m^k_{ig} \in \{0, 1\}$ for $i = 1, \ldots, N, g = 1, \ldots, G,$ and $\sum_{g=1}^{G} m_{ig} = 1 \ \forall \ i = 1, \ldots, N]$, modelling heterogeneity within clusters, isolation between clusters and the partition into G classes, respectively. Thus, the classification matrix identifying a partition is modelled as follows

$$\mathbf{U}^*_k = \mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}'_k + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}'_k - \text{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}'_k), \tag{4.1}$$

In order to obtain a Well-Structured Partion (Rubin, 1967), Equation (4.1) is subject to the constraint

$$\max\{{}_Wd^k_{gg} : g = 1, \ldots, G\} \leq \min\{{}_Bd^k_{gt} : t, g = 1, \ldots, G, \ (g \neq t)\} \tag{4.2}$$

In other words, dissimilarities within clusters must be smaller than the dissimilarities between clusters. For the sake of brevity, the matrix form of constraint (4.2) will be used in the rest of the chapter, i.e.

$${}_B\mathbf{D}_k > {}_W\mathbf{D}_k. \tag{4.3}$$

4.2.2 Parsimonious Hierarchies

When matrix ${}_B\mathbf{D}_k$ satisfying constraint 4.2 is also an ultrametric matrix of order G, then $\mathbf{U}^*_k$ is a square $(2G-1)$-ultrametric matrix of order N, with off-diagonal elements that can assume one of at most $(2G-1)$ different values: $0 < {}_Wd^k_{gg} \leq {}_Bd^k_{gt} (g, t = 1, \ldots, G; \ g \neq t)$ (Vichi, 2008).

More formally: $\mathbf{U}^*_k = [u^{k*}_{il}], u^{k*}_{ii} = 0, u^{k*}_{il} \geq 0, u^{k*}_{il} = u^{k*}_{li}, u^{k*}_{il} \leq \max(u^{k*}_{ir}, u^{k*}_{lr}) \ \forall \ (i, l, r)$; furthermore $u^{k*}_{il} \in \{0, \ {}_Wd^k_{gg}, \ {}_Bd^k_{gt}\}$, with $0 < {}_Wd^k_{gg} \leq {}_Bd^k_{gt} \ \forall (g, t : g \neq t)$.

There exists a bijection between ultrametric matrices $\mathbf{U}_h$ and dendrograms (hierarchies), which has been proved by Johnson (1967). Thus, H ultrametric matrices are associated with a set of H dendrograms representing the primary hierarchies $\Delta = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_H]$.

To clearly show what is meant by parsimonious hierarchy, in the following we consider a $(2G-1)$ dendrogram when $G = 5$ and also its corresponding ${}_B\mathbf{D}$ and ${}_W\mathbf{D}$ matrices. More precisely, on the one hand, the entries on the main diagonal of the matrix ${}_W\mathbf{D}$ are the values of heterogeneity within the $G = 5$ clusters and are displayed on the y-axis of Figure 4.2, representing the levels of aggregation between the units of each cluster; on the other hand, the off-diagonal entries of the matrix ${}_B\mathbf{D}$ are the values of isolation between the clusters and are displayed on the y-axis of Figure 4.2, representing the levels of aggregation between each pair of the considered G clusters of the partition.

Table 4.1. Matrices of isolation between clusters ($_B\mathbf{D}$) and heterogeneity with clusters ($_W\mathbf{D}$)

$$
{}_B\mathbf{D} = \begin{bmatrix} 0 & {}_Bd_{12} & {}_Bd_{13} & {}_Bd_{14} & {}_Bd_{15} \\ {}_Bd_{21} & 0 & {}_Bd_{23} & {}_Bd_{24} & {}_Bd_{25} \\ {}_Bd_{31} & {}_Bd_{32} & 0 & {}_Bd_{34} & {}_Bd_{35} \\ {}_Bd_{41} & {}_Bd_{42} & {}_Bd_{43} & 0 & {}_Bd_{45} \\ {}_Bd_{51} & {}_Bd_{52} & {}_Bd_{53} & {}_Bd_{54} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 10.5 & 11.375 & 11.375 & 12.25 \\ 10.5 & 0 & 11.375 & 11.375 & 12.25 \\ 11.375 & 11.375 & 0 & 8.5 & 12.25 \\ 11.375 & 11.375 & 8.5 & 0 & 12.25 \\ 12.25 & 12.25 & 12.25 & 12.25 & 0 \end{bmatrix}
$$

$$
{}_W\mathbf{D} = \begin{bmatrix} {}_Wd_{11} & 0 & 0 & 0 & 0 \\ 0 & {}_Wd_{22} & 0 & 0 & 0 \\ 0 & 0 & {}_Wd_{33} & 0 & 0 \\ 0 & 0 & 0 & {}_Wd_{44} & 0 \\ 0 & 0 & 0 & 0 & {}_Wd_{55} \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}
$$

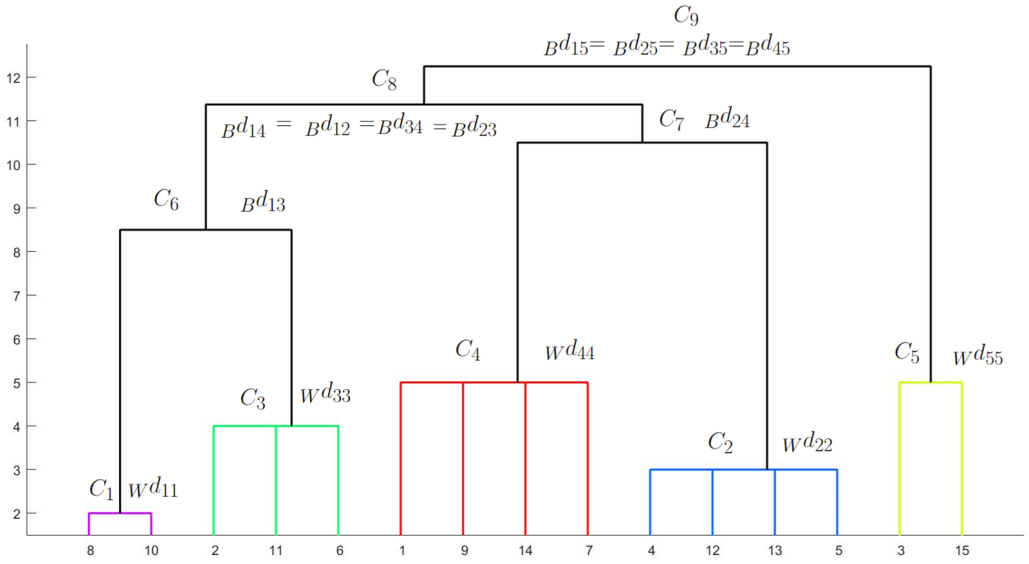


Figure 4.2. Representation of a $(2G-1)$-dendrogram when $G=5$. A 9-dendrogram is shown; the first five clusters $(C_1,\ldots,C_5)$ form a partition; clusters $C_6=\{C_1,C_3\}, C_7=\{C_2,C_4\}, C_8=\{C_6,C_7\}, C_9=\{C_5,C_8\}$, specify the hierarchical structure of the partition

Clearly, the parsimonious dendrogram (PD) displayed in Figure 4.2 is associated with a parsimonious hierarchy. Moreover, it is worth noting that the associated isolation and heterogeneity matrices displayed in Table 4.1 have the following characteristics: matrix $_W\mathbf{D}$ is a diagonal matrix with positive entries on the main diagonal, the matrix $_B\mathbf{D}$ is an ultrametric matrix, and the WSP constraint (4.2 or 4.3) holds, with the maximum value of $_W\mathbf{D}$, i.e. 5 being smaller than the minimum value of matrix $_B\mathbf{D}$, i.e. 8.5.

The choice of G (in general, of the number of classes in a partition problem) is an open question and requires further consideration. In our applications (Section 4.5), the choice of G is either suggested in the literature or known *a priori*. Nevertheless, it is important to discuss and make a few remarks about it. It is crucial to remember

that the choice of G is related to the WSP model subject to the constraint of matrix ${}_B\mathbf{D}$ being an ultrametric matrix of order G. In fact, in this way the obtained consensus matrix $\mathbf{U}_k^*$ corresponds to a parsimonious hierarchy of order $(2G-1)$. As a first consideration, it should be noted that the choice of G influences the number of parameters to estimate the inter-cluster isolation (${}_B\mathbf{D}$) and within-cluster heterogeneity (${}_W\mathbf{D}$) matrices; more precisely, the larger G is, the larger the number of parameters to be estimated and the smaller the fit of the objective function. Therefore, by letting G vary in a reasonable range and storing the solutions of the objective function, it is suggested to keep the smallest G that produces the strongest variation in the solutions of the objective function, as Cattell (1966) proposed in a factor analysis framework. In this way, a reasonably interpretable final partition can be obtained (Vichi, 2008).

4.3 Fuzzy partition of hierarchies and their parsimonious consensus dendrograms

The methodology proposed in this chapter aims to find a fuzzy partition in K classes of the primary hierarchies with $(2G-1)$-ultrametric consensuses (parsimonious trees) for each class of the partition.
In order to achieve this goal the following optimization problem has to be solved w.r.t. $\mathbf{M}_k$, ${}_B\mathbf{D}_k$, ${}_W\mathbf{D}_k$, and μ_{hk}^m,

$$
\left\{
\begin{array}{lll}
\text{minimize} \displaystyle\sum_{k=1}^{K}\sum_{h=1}^{H} \|\mathbf{U}_h - \mathbf{U}_k^*\|^2 \mu_{hk}^m = & & \\
\text{minimize} \displaystyle\sum_{k=1}^{K}\sum_{h=1}^{H} \|\mathbf{U}_h - (\mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}_k^{'} + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'} + & & \\
\qquad - \operatorname{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'}))\|^2 \mu_{hk}^m & & (4.\text{P1}) \\
\text{s.t.} & & \\
\displaystyle\sum_{k=1}^{K} \mu_{hk} = 1 & \text{for } h = 1,\ldots,H & (4.C_1) \\
\mu_{hk} \in [0,1] & \text{for } h = 1,\ldots,H,\ k = 1,\ldots,K & (4.C_2) \\
m_{ig}^k \in \{0,1\} & \text{for } i = 1,\ldots,N, g = 1,\ldots,G & (4.C_3) \\
\displaystyle\sum_{g=1}^{G} m_{ig} = 1 & \text{for } i = 1,\ldots,N & (4.C_4) \\
{}_B\mathbf{D}_k > {}_W\mathbf{D}_k & & (4.C_5) \\
{}_Bd_{il}^k \le \max\{{}_Bd_{ip}^k,\ {}_Bd_{lp}^k\} & & (4.C_6) \\
{}_Bd_{ip}^k \le \max\{{}_Bd_{il}^k,\ {}_Bd_{lp}^k\} & & (4.C_7) \\
{}_Bd_{lp}^k \le \max\{{}_Bd_{ip}^k,\ {}_Bd_{il}^k\} & \text{for } i = 1,\ldots,G-2,\ l = i+1,\ldots,G-1,\ p = l+1,\ldots,G. & (4.C_8)
\end{array}
\right.
$$

consensuses fitting WSP and fuzzy partition of hierarchies

Constraints $4.C_1$ and $4.C_2$ guarantee that the set of ultrametric matrices $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_H$ is partitioned in a fuzzy way, i.e., into K classes: each ultrametric matrix belongs to the k-th class with the h-th membership μ_{hk}. Constraints $4.C_3$, $4.C_4$ and $4.C_5$ are needed to guarantee that the partition is well-structured. Finally, the last triplet of constraints, i.e. constraints $4.C_6$, $4.C_7$ and $4.C_8$, guarantees that the matrix ${}_B\mathbf{D}$ is ultrametric. The whole set of constraints in 4.P1 allows us to obtain a fuzzy partition of the primary hierarchies into K classes, by identifying K

parsimonious ultrametric matrices $\mathbf{U}_k^*$: in this way, each consensus is a $(2G-1)$-ultrametric matrix, and therefore has a parsimonious tree associated with it. The reader can see that if the last triplet of constraints, i.e. constraints 4.C_6, 4.C_7 and 4.C_8, is ignored, then the K consensus matrices are matrices that identify just a well-structured partition and not a parsimonious hierarchy. Finally, the fuzziness of the partition is controlled by the parameter m, named *fuzzifier*. In particular, when $m \to 1$ the partition tends to become hard, i.e. the membership degrees tend to be either 0 or 1; for $m \to \infty$ membership tend to be constant and equal to $1/K$.

Therefore, problem (4.P1) can be used in order to solve the following sub-problems:

(4.P1.a) Given H primary hierarchies, obtain a fuzzy secondary partition of the primary hierarchies, and for each class of the secondary partition identify a consensus well-structured partition. This problem consists of solving 4.P1 subject to constraints 4.C_1-4.C_5.

(4.P1.b) Given H primary hierarchies, obtain a fuzzy secondary partition of the primary hierarchies and for each class of the secondary partition identify a consensus hierarchy with a parsimonious structure. This problem consists of solving 4.P1 subject to constraints 4.C_1-4.C_8.

(4.P1.c) Given a single hierarchy (dendrogram), find the closest well-structured partition. If the hierarchy is not initially given, i.e. if a dissimilarity matrix is given, then its corresponding hierarchy or ultrametric matrix can be obtained by applying UPGMA, or any other hierarchical clustering algorithm, to the dissimilarity matrix. This problem consists of solving 4.P1 subject to constraints 4.C_1-4.C_5 when $H = 1$ and $K = 1$.

(4.P1.d) Given a single hierarchy (dendrogram), find the closest parsimonious dendrogram. If the hierarchy is not initially given, i.e. if a dissimilarity matrix is given, then its corresponding hierarchy or ultrametric matrix can be obtained by applying UPGMA, or any other hierarchical clustering algorithm, to the dissimilarity matrix. This problem consists of solving 4.P1 subject to constraints 4.C_1-4.C_8 when $H = 1$ and $K = 1$.

4.3.1 Least-Squares Estimation

In order to implement (4.P1), it is worth noting that it can be decomposed into two alternating minimization sub-problems:

(A) the partial minimization of the objective function of (4.P1) with respect to centroid matrices when these are the parsimonious hierarchies (1). i.e., $\mathbf{U}_k^*$, and $\hat{\mu}_{hk}$ is given.

$$
\begin{cases}
\text{minimize } \sum_{k=1}^{K}\sum_{h=1}^{H} \|\mathbf{U}_h - (\mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}_k^{'} + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'} - \text{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'}))\|^2 \hat{\mu}_{hk}^m & \text{(4.P2)} \\
\text{s.t.} & \\
m_{ig}^k \in \{0,1\} \qquad \text{for } i=1,\ldots,N, g=1,\ldots,G & (4.C_3) \\
\sum_{g=1}^{G} m_{ig} = 1 \qquad \text{for } i=1,\ldots,N & (4.C_4) \\
{}_B\mathbf{D}_k > {}_W\mathbf{D}_k & (4.C_5) \\
{}_Bd_{il}^k \leq \max\{{}_Bd_{ip}^k,\ {}_Bd_{lp}^k\} & (4.C_6) \\
{}_Bd_{ip}^k \leq \max\{{}_Bd_{il}^k,\ {}_Bd_{lp}^k\} & (4.C_7) \\
{}_Bd_{lp}^k \leq \max\{{}_Bd_{ip}^k,\ {}_Bd_{il}^k\} \qquad \text{for } i=1,\ldots,G-2,\ l=i+1,\ldots,G-1,\ p=l+1,\ldots,G & (4.C_8)
\end{cases}
$$

The solution of this sub-probem (A) can be found by using the Sequential Quadratic Programming (SQP) algorithm (Powell, 1983).

It is worth noting that the unconstrained least square solution of (4.P2) is given by $\bar{\mathbf{U}}_k$, for $k=1,\ldots,K$, where

$$\bar{\mathbf{U}}_k = \frac{1}{\sum_{h=1}^{H}\hat{\mu}_{hk}^m}\sum_{h=1}^{H}\hat{\mu}_{hk}^m\mathbf{U}_h \tag{4.4}$$

is the weighted arithmetic mean matrix of $\mathbf{U}_h$, for $h=1,\ldots,H$, weighted by $\hat{\mu}_{hk}^m$.

Typically, matrices $\bar{\mathbf{U}}_k$ are not (2G-1)-ultrametrics. However, only a few iterations are needed for the SQP algorithm to run, if the problem (4.P2) takes as initial values the matrices $\bar{\mathbf{U}}_k$. For this reason, the following problem is minimized with respect to $\mathbf{U}_k^*$:

$$
\begin{cases}
\text{minimize } \sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - (\mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}_k^{'} + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'} - \text{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k^{'}))\|^2 \sum_{h=1}^{H}\hat{\mu}_{hk}^m & \text{(4.P3)} \\
\text{s.t.} & \\
m_{ig}^k \in \{0,1\} \qquad \text{for } i=1,\ldots,N, g=1,\ldots,G & (4.C_3) \\
\sum_{g=1}^{G} m_{ig} = 1 \qquad \text{for } i=1,\ldots,N & (4.C_4) \\
{}_B\mathbf{D}_k > {}_W\mathbf{D}_k & (4.C_5) \\
{}_Bd_{il}^k \leq \max\{{}_Bd_{ip}^k,\ {}_Bd_{lp}^k\} & (4.C_6) \\
{}_Bd_{ip}^k \leq \max\{{}_Bd_{il}^k,\ {}_Bd_{lp}^k\} & (4.C_7) \\
{}_Bd_{lp}^k \leq \max\{{}_Bd_{ip}^k,\ {}_Bd_{il}^k\} \qquad \text{for } i=1,\ldots,G-2,\ l=i+1,\ldots,G-1,\ p=l+1,\ldots,G & (4.C_8)
\end{cases}
$$

by using SQP. An alternative way to optimize (4.P3) is to solve problem (4.P3), by using a coordinate descent algorithm where in the step of computing ${}_BD_k$ the UPGMA algorithm is applied on the matrix ${}_BD_k$, since UPGMA is known to find an optimal LS ultrametric transformation of ${}_BD_k$. In this way, the WSP model (model 4.1) is solved subject to the ultrametricity constraint of the matrix ${}_B\mathbf{D}$ (i.e. constraints 4.C_6, 4.C_7, 4.C_8) on matrices $\bar{\mathbf{U}}_k$, for $k=1,\ldots,K$ to obtain the corresponding parsimonious ultrametric matrix. In practice, (4.P3) transforms the dissimilarity matrix $\bar{\mathbf{U}}_k$ into the closest $(2G-1)$-ultrametric matrix.

(B) the partial minimization of the objective function of (4.P2) with respect to the fuzzy partition $[\mu_{hk}]$ when $\hat{\mathbf{U}}_k^*$ is given

$$\begin{cases} \text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{U}_h - \hat{\mathbf{U}}_k^*\|^2 \mu_{hk}^m & \text{(4.P4)} \\ \text{s.t.} \\ \sum_{k=1}^{K} \mu_{hk} = 1 \qquad \text{for } h = 1,\dots,H & (4.C_1) \\ \mu_{hk} \in [0,1] \qquad \text{for } h = 1,\dots,H,\ k = 1,\dots,K. & (4.C_2) \end{cases}$$

This sub-problem (B) is obtained by solving it by means of the first-order conditions for stationarity. In fact, the stationary point can be found by considering the Lagrangian function

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \left\|\mathbf{U}_h - \hat{\mathbf{U}}_k^*\right\|^2 \mu_{hk}^m - \sum_{h=1}^{H} \lambda_h \left(\sum_{k=1}^{K} \mu_{hk} - 1\right), \tag{4.5}$$

where the solution with respect to μ_{hk} is

$$\mu_{hk} = \frac{1}{\sum_{j=1}^{K} \left(c_{hk}/c_{hj}\right)^{\frac{2}{m-1}}}, \qquad \text{for}, h = 1,\dots,H, k = 1,\dots,K. \tag{4.6}$$

where $c_{lp} = tr[(\mathbf{U}_l - \hat{\mathbf{U}}_p^*)'(\mathbf{U}_l - \hat{\mathbf{U}}_p^*)]$.

After the solution of the two sub-problems (A) and (B) the objective function generally reduces w.r.t. the previous iteration, or at least does not increase. Since it is bounded below by zero, after some iterations the algorithm stops to a stationary point that is not guaranteed to be the global minimum of the problem. For this reason, the algorithm is recommended to be run from several initial starting points to improve the chance of identifying the global optimal solution. The steps of the algorithm can now be formally presented.

ALGORITHM for (4.P1):

0. Initialization
   Set $t = 0$; $\epsilon > 0$ convergence constant; and randomly generate the membership degree matrix $[\mu_{hk}]$, with $k = 1,\dots,K$, $h = 1,\dots H$ from a uniform distribution and make it row-stochastic.

1. Do $t = t + 1$

2. Given $[\hat{\mu}_{hk}]$, solve sub-problem (A) with SQP algorithm or considering the following steps:

   (a) Compute $\bar{\mathbf{U}}_k$, for $k = 1,\dots,K$ as follows:

$$\bar{\mathbf{U}}_k = \frac{1}{\sum_{h=1}^{H} \hat{\mu}_{hk}^*} \sum_{h=1}^{H} \hat{\mu}_{hk}^* \mathbf{U}_h \tag{4.7}$$

(b) Solve problem (4.P3) as follows. For sake of simplicity, we let F be the objective function of problem (4.P3), namely:

$$F({}_B\mathbf{D}_k,{}_W\mathbf{D}_k,\mathbf{M}) = \sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - (\mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}_k' + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k' + - \operatorname{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k'))\|^2 \sum_{h=1}^{H} \hat{\mu}_{hk}^m \quad (4.8)$$

It is worth noting that when minimizing (4.8) w.r.t. ${}_B\mathbf{D}_k$, ${}_W\mathbf{D}_k$ and $\mathbf{M}_k$, $\hat{\mu}_{hk}$ is fixed (constant) and therefore only

$$F({}_B\mathbf{D}_k,{}_W\mathbf{D}_k,\mathbf{M}) = \sum_{k=1}^{K} \|\bar{\mathbf{U}}_k - (\mathbf{M}_k({}_B\mathbf{D}_k)\mathbf{M}_k' + \mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k' + - \operatorname{diag}(\mathbf{M}_k({}_W\mathbf{D}_k)\mathbf{M}_k'))\|^2 \quad (4.9)$$

will be minimized w.r.t. ${}_B\mathbf{D}_k$, ${}_W\mathbf{D}_k$ and $\mathbf{M}_k$.

i. Fixing $\hat{\mathbf{M}}_k$, differentiate the objective function of (4.P3) (Equation 4.9) w.r.t. ${}_W\mathbf{D}_k$ and equate to zero. The solution ${}_W\hat{\mathbf{D}}_k$ will have as generic element on the main diagonal:

$${}_W\hat{d}_{gg}^k = \frac{\sum_{i=1}^{N}\sum_{l=1,i\neq l}^{N} \bar{\mathbf{U}}_{il}^k \hat{m}_{ig}^k \hat{m}_{lg}^k}{\sum_{i=1}^{N}\sum_{l=1,i\neq l}^{N} \hat{m}_{ig}^k \hat{m}_{lg}^k} \qquad (g = 1,\ldots,G); \quad (4.10)$$

ii. Fixing $\hat{\mathbf{M}}_k$, differentiate the objective function of (4.P3) (Equation 4.9) w.r.t. ${}_B\mathbf{D}_k$ and equate to zero. The solution ${}_B\hat{\mathbf{D}}_k$ will have as generic element:

$${}_B\hat{d}_{gf}^k = \frac{\sum_{i=1}^{N}\sum_{l=1,i\neq l}^{N} \bar{\mathbf{U}}_{il}^k \hat{m}_{ig}^k \hat{m}_{lf}^k}{\sum_{i=1}^{N}\sum_{l=1}^{N} \hat{m}_{ig}^k \hat{m}_{lf}^k} \qquad (g,f = 1,\ldots,G); \quad (4.11)$$

iii. Fixing ${}_W\hat{\mathbf{D}}_k$ and ${}_B\hat{\mathbf{D}}_k$, minimize the objective function of (4.P3) (Equation 4.9) w.r.t. $\mathbf{M}_k$. The minimization is done row by row, namely minimizing the objective function w.r.t. row i of $\mathbf{M}_k$ ($\mathbf{m}_i^k$), fixing the other rows of $\mathbf{M}_k$; formally the minimization will be done considering $\mathbf{M}_k = [\hat{\mathbf{m}}_1^k, \hat{\mathbf{m}}_2^k, \ldots, \mathbf{m}_i^k, \ldots, \hat{\mathbf{m}}_N^k]'$. Therefore, unit i belongs to the gth class, $m_{ig}^k = 1$, if the objective function of (4.P3) reaches its minimum compared to the situations where unit i is assigned to any other class $v = 1,\ldots,G$, $v \neq g$. Otherwise, unit i does not belong to class g, i.e. $m_{ig}^k = 0$. Formally, for each $i = 1,\ldots,N$:

$$\hat{m}_{ig}^k = 1, \text{ if } F({}_B\mathbf{D}_k,{}_W\mathbf{D}_k,[\hat{\mathbf{m}}_1^k,\hat{\mathbf{m}}_2^k,\ldots,\mathbf{m}_i^k = \mathbf{i}_g,\ldots,\hat{\mathbf{m}}_N^k]') = \\ = \min\{F({}_B\mathbf{D}_k,{}_W\mathbf{D}_k,[\hat{\mathbf{m}}_1^k,\hat{\mathbf{m}}_2^k,\ldots,\mathbf{m}_i^k = \mathbf{i}_f,\ldots,\hat{\mathbf{m}}_N^k]') : \\ : f = 1,\ldots,G\ (f \neq g)\},$$

$$\hat{m}_{ig}^k = 0, \text{ otherwise,}$$

where $\mathbf{i}_f$ is the fth row of the identity matrix of order G.

The proofs of the aforementioned estimates are given by Vichi (2008).

3. Given $\hat{\mathbf{U}}$, solve sub-problem (B)
The solution of (4.P4) is given by:

$$\mu_{hk} = \frac{1}{\sum_{j=1}^{K} \left(c_{hk}/c_{hj}\right)^{\frac{2}{m-1}}}, \qquad \text{for}, h = 1, \ldots, H, k = 1, \ldots, K. \tag{4.12}$$

where $c_{lp} = tr[(\mathbf{U}_l - \hat{\mathbf{U}}_p^{(t)})'(\mathbf{U}_l - \hat{\mathbf{U}}_p^{(t)})]$.

4. Stopping Rule
Repeat steps 1-3 until the difference between the objective function at iteration t and the objective function at iteration $t-1$ is greater than ϵ.

4.4 Simulation study

To assess the performance of the proposed methodology, an extended simulation study has been developed. It consists mainly of two experiments. The former aims to assess whether the proposed methodology is able to recognize the underlying generated hard partition; the latter studies the performance of the proposed methodology in recognizing the underlying generated fuzzy partition. The simulation is organized in the two above briefly described experiments by considering two levels of errors in the data generation process. The errors have been generated from a normal distribution and then symmetrized. For each experiment and error level 200 three-way ultrametric matrices have been generated for a total of 800 samples. Details are provided in the corresponding Sections 4.4.1 and 4.4.2. In addition, 200 three-way ultrametric matrices have been generated to study how to avoid local minima in the final solution of the algorithm.

Clearly, since the partitioning problem of a set of multivariate objects is an NP-hard problem (Křivánek and Morávek, 1986), there is no guarantee that the new methodology finds a global optimum; indeed, it is possible that the obtained minimum is just a local one. For this reason, the algorithm for each data set is run by using several randomly generated partitions (briefly, "random starts") and the best solution is retained in order to increase the chance of identifying the global minimum solution. More specifically, the correct choice of the number of random starts has been decided by running an experiment, using a high level of error in the generated ultrametric matrices (see Section 4.4.1). The new algorithm was run by letting the number of random starts be 1, 5, 10, 20, 30, and 40. Then, the percentage of the final solutions ending in a local minimum has been computed. Table 4.2 reports the local minima occurrence (percentage), as the number of random starts increases. It has to be noted that when the number of random starts is set equal to 10, local minima do not occur. Thus, the number of random starts for the whole simulation study was set $RndStarts = 10$. From Table 4.2 it can be observed that

Table 4.2. Local minima occurrences (%)

RndStarts	1	5	10	20	30	40
%	20	5	0	0	0	0

even with only 1 random start the performance of the algorithm is good, with only 20% local minima occurrences. When 5 random starts are used, the percentage of

local minima strongly decreases (5%), thus identifying the global minimum in 95% of cases.

The difference in the number of random starts used to avoid local minima in the simulation studies discussed in Chapter 3 and in this Chapter is mainly due to the following reasons. Firstly, because of the process of generating the input matrices: in fact, in Chapter 3, Section 3.4, 12 dissimilarity matrices were generated from 4 ultrametric matrices, while in this simulation study, 12 ultrametric matrices were generated from 4 parsimonious ultrametric matrices. Secondly, because of the optimization problem: when the number of constraints of an optimization problem (here 3.P1) is smaller than the number of constraints of another optimization problem (here 4.P1), the dimension of the search space of the former is higher than the dimension of the latter. Consequently, the probability of local minima occurring is lower in the latter problem than that in the former. In fact, when the search space is large or high-dimensional, the task of finding the global solution is arduous due to the presence of a large search space and a large number of local minima. Moreover, the complexity of the problem increases with the size of the search space; in other words, the proportion of the region attracting the global optimum to the entire search space (region) decreases with dimensionality and the depth of local minima increases (D'Angelo and Palmieri, 2021). Also, it is important to remember that in the two simulation studies, while the number of matrices generated is the same, the size of the search space is different. Thus, to reduce the risk of getting trapped in local minima, in addition to the different number of random starts to be used, it would be necessary to increase the number of input matrices needed to cover the entire search space in the first problem (here 3.P1) (Sharma and Jabeen, 2023).

The results of the simulation studies are analyzed by considering several external validity indices to compare the obtained partition with the true one. Adjusted Rand Index (ARI, by Hubert and Arabie (1985)), fuzzy Adjusted Rand Index (Fuzzy ARI, by Campello (2007) and fuzzy Rand Index (Fuzzy RI, by Campello (2007)) have been used. In addition, the Normalized Root Mean Square Error (NRMSE) has been used to compare the obtained consensus matrices with the true ones. Finally, the Mean Membership Matrices are computed to assess whether the methodology is able to recognize the fuzzy or hard assignment: these matrices are obtained by averaging all the membership matrices resulting from each run of the algorithm after optimally permuting their columns in order to avoid the label switching problem.

4.4.1 First simulation: hard assignment experiment

The first simulation has been developed by considering four $(2G-1)$-ultrametric matrices, with $G = 4$. Each of these matrices is associated with a parsimonious dendrogram, as shown in Figure 4.3, where the 4 clusters $(C_1 - C_4)$ are clearly visible.

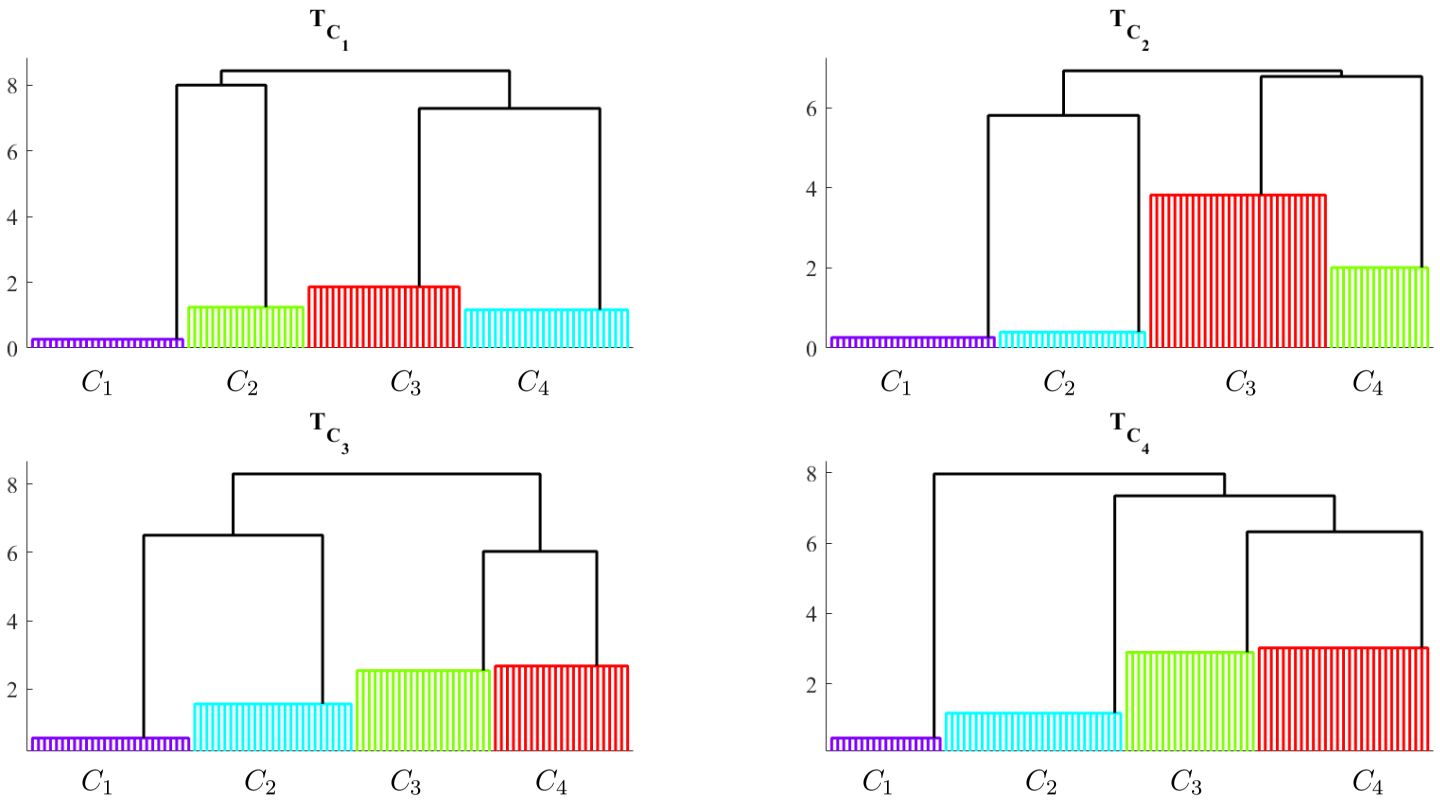


Figure 4.3. Consensus parsimonious dendrograms (hard assignment experiment)

Those four $(2G-1)$-ultrametric matrices ($\mathbf{U}_k^*,\ k=1\ldots,4$) are used to generate the $H=12$ starting ultrametric matrices (primary hierarchies) ($\mathbf{U}_h,\ h=1,\ldots,12$). In fact, from each $\mathbf{U}_k^*,\ k=1,\ldots,4$, three different ultrametric matrices are generated by adding a symmetric error matrix to $\mathbf{U}_k^*$ and forcing the resulting dissimilarity matrix to be ultrametric, by using an average linkage method (UPGMA). Thus, a total of $H=12$ ultrametric matrices are obtained and given as input to the algorithm to recognize the hard assignment, since each of the H ultrametric matrices is associated with the single consensus matrix. The algorithm returns as output not only the obtained secondary partition, but also the parsimonious hierarchy associated with each class of the partition.

It has to be noted that two levels of errors are considered. A low error should guarantee that the algorithm works in optimal conditions and it should always be able to find the global optimum solution with an ARI always equal to 1. In other words, the algorithm always detects the true (secondary) partition. The high error should identify a strongly biased situation, where the algorithm is able to recognize the true (secondary) partition in the majority of cases.

Table 4.3 reports the corresponding summary statistics of the performance aforementioned indicators. Particularly, both the mean and the median of the indices regarding 200 iterations are shown. The NRMSE is reported with three different statistics: indeed, in each iteration K NRMSE are computed, each of those measuring the difference between the k-th resulting ultrametric and the k-th original true one; then, the mean, the median and the maximum values are computed.

Table 4.3. Summary statistics. Experiment under a hard assignment with low and high errors.

Level of error	Statistics	local minimum (%)	ARI	Fuzzy ARI	Fuzzy RI	mean NRMSE	median NRMSE	max NRMSE
Low	median	0.000	1.000	0.835	0.947	0.006	0.006	0.007
Low	mean	0.000	1.000	0.835	0.947	0.005	0.005	0.007
High	median	0.000	1.000	0.590	0.504	0.427	0.444	0.469
High	mean	0.000	0.869	0.610	0.505	0.415	0.423	0.468

When using low error, the algorithm performed very well. Indeed, values of ARI, fuzzy ARI and fuzzy RI are close to 1, while values of NRMSE are quite low (Table 4.3). When using high error, the methodology detects only few times the true partition and low values of ARI, Fuzzy ARI and Fuzzy RI are shown in Table 4.3. The percentage of ARI equal to one is 86.9%, as hypothesised from a high level of error. In addition, the values of the NRMSE are significantly larger than zero, meaning that the true parsimonious consensus dendrograms are not perfectly detected.

Moreover, the methodology is able to recognize the underlying hard secondary partition of the ultrametric matrices (primary hiearchies). Indeed, the membership value of each ultrametric matrix to the corresponding cluster is frequently close to 1. The mean membership matrix obtained averaging the 200 obtained matrices is reported in Table 4.4. Results confirm that when using low error the membership is always larger than 0.8 as expected (Table 4.4 (a)). When using high error, the true partition is still detected, but the highest value (indicating the strongest membership) is about 0.5 (Table 4.4 (b)).

4.4.2 Second simulation: fuzzy assignment experiment

The second simulation has been developed by considering two $(2G-1)$-ultrametric matrices, with $G=5$. The associated parsimonious dendrograms are shown in Figure 4.4.

Table 4.4. Mean membership matrices. Experiment under a hard assignment with low and high error.

(a) low error

Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.875	0.042	0.040	0.043
0.874	0.042	0.040	0.043
0.875	0.042	0.040	0.043
0.041	**0.857**	0.049	0.053
0.041	**0.857**	0.049	0.053
0.041	**0.856**	0.049	0.053
0.040	0.049	**0.863**	0.048
0.040	0.049	**0.863**	0.048
0.040	0.049	**0.863**	0.048
0.042	0.053	0.048	**0.857**
0.042	0.053	0.048	**0.857**
0.042	0.053	0.047	**0.857**

(b) high error

Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.540	0.155	0.149	0.156
0.537	0.156	0.150	0.157
0.540	0.155	0.149	0.156
0.147	**0.469**	0.188	0.196
0.147	**0.470**	0.187	0.196
0.146	**0.473**	0.186	0.195
0.144	0.189	**0.483**	0.184
0.144	0.189	**0.484**	0.183
0.144	0.189	**0.484**	0.183
0.149	0.195	0.182	**0.474**
0.149	0.195	0.181	**0.475**
0.149	0.195	0.181	**0.475**

Those $K = 2$ $(2G - 1)$-ultrametric matrices ($\mathbf{U}_k^*,\ k = 1 \ldots, 2$) are used to generate the $H = 9$ starting ultrametric matrices (primary hierarchies) ($\mathbf{U}_h,\ h = 1, \ldots, 9$) under a fuzzy assignment scenario. Specifically, from each $\mathbf{U}_k^*,\ k = 1, \ldots, K$, three different ultrametric matrices are generated by adding a symmetric error matrix to $\mathbf{U}_k^*$ and forcing the resulting dissimilarity matrix to be ultrametric, by using an averaging linkage method (UPGMA). Thus, 6 ultrametric matrices are generated and expected to be hardly associated to the corresponding cluster, being themselves generated by one consensus matrix. Moreover, an additional 3 ultrametric matrices are generated by averaging the two consensus matrices, and then adding a symmetric error term, and forcing the resulting matrix to be ultrametric by UPGMA. In this way, the last three ultrametric matrices are expected to be softly associated with both clusters, being themselves generated by a linear combination of the two consensus parsimonious ultrametric matrices.

Therefore, a total of $H = 9$ ultrametric matrices are obtained and given as input to the algorithm in order to recognize the fuzzy assignment. The algorithm returns as output not only the obtained secondary partition but also the parsimonious

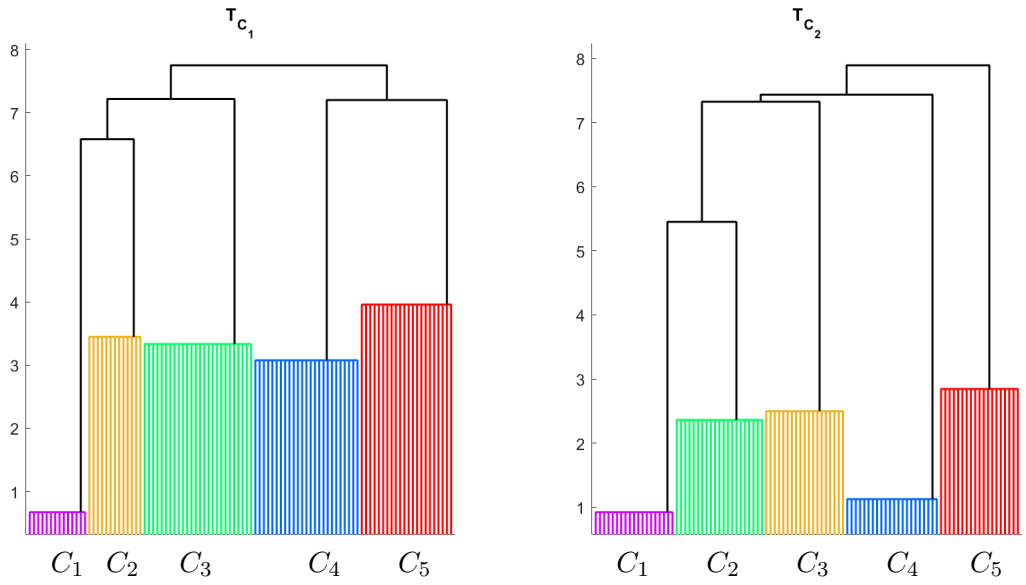


Figure 4.4. Consensus parsimonious dendrograms (fuzzy assignment experiment)

hierarchy associated with each class of the partition.

It has to be noted that also in this case two levels of errors are considered. A low error guarantees that the partition is always detected and therefore all the ARI are equal to 1, while a high error masks the true partition, but still the algorithm detects the partition in the majority of cases.

For the results, we expect that the algorithm almost hardly assigns the first six ultrametric matrices (primary hierarchies) to the corresponding cluster and softly assigns the last three ultrametric matrices (primary hierarchies) to both the clusters. We ran the experiment with both low and high error levels. The results are shown in Table 4.5, which reports the main statistics of interest and also the percentage of fuzziness detection, i.e. the proportion of occurrences in which the methodology is able to recognize that the last three ultrametric matrices are generated by both the consensuses. When using low error, the percentage of ARI equal to 1 is 100%. When the level of error is high, the percentage of ARI exactly equal to 1 is 62%.

Table 4.5. Summary statistics. Experiment under a fuzzy assignment with low and high errors.

Level of error	Statistics	local minimum (%)	ARI	Fuzzy ARI	Fuzzy RI	mean NRMSE	median NRMSE	max NRMSE	Fuzzyness detection
Low	median	0.000	1.000	0.490	0.760	0.229	0.229	0.326	1.000
Low	mean	0.000	1.000	0.450	0.720	0.219	0.219	0.316	1.000
High	median	0.000	1.000	0.480	0.740	0.252	0.252	0.286	1.000
High	mean	0.000	0.620	0.429	0.715	0.252	0.252	0.286	1.000

From Table 4.5, we notice that the methodology is able to recognize the underlying partition. For both errors, the mean values of Fuzzy ARI and Fuzzy RI are about 0.5. Clearly, when low error is used, the performance is slightly better. Moreover, the NRMSE are significantly larger than zero, showing differences between generated and obtained consensus parsimonious matrices. In addition, we notice that the proportion of occurrences in which the methodology is able to recognize the fuzzy nature of the last three ultrametric matrices is 1, meaning that the methodology always softly assigns those matrices to both the clusters, regardless the level of error used.

It is worth observing that the new methodology is able to recognize the fuzzy nature of the last three ultrametric matrices (primary hierarchies) and also that the first six are generated by just one consensus matrix. Table 4.6 shows the mean membership matrix, highlighting that for the first 6 ultrametric matrices the highest membership value is close to 0.9; instead, for the last three ultrametric matrices, both memberships are approximately close to 0.5, meaning that those matrices are softly assigned to both clusters, as expected.

Table 4.6. Mean membership matrices: experiment under a fuzzy assignment with low and high errors.

(a) low error

Cluster 1	Cluster 2
0.987	0.013
0.987	0.013
0.987	0.013
0.010	**0.990**
0.010	**0.990**
0.010	**0.990**
0.538	**0.462**
0.538	**0.462**
0.538	**0.462**

(b) high error

Cluster 1	Cluster 2
0.985	0.015
0.985	0.015
0.985	0.015
0.016	**0.984**
0.016	**0.984**
0.016	**0.984**
0.495	**0.505**
0.495	**0.505**
0.494	**0.506**

4.5 Real applications

In the following, two applications to real data are analyzed. The former consists of applying the methodology to the *zoo dataset* (UCI repository) and refers to problem (4.P1.d): given a dendrogram, find the closest Least-Square parsimonious dendrogram. The latter consists of applying the methodology to the *girls' growth curves dataset* (Sempé and Médico-Sociale, 1987) and refers to problem (4.P1.b): given a set of primary hierarchies, find a fuzzy secondary partition of them, and within each class of the secondary partition, identify a consensus parsimonious dendrogram. Details on the dataset descriptions and on the results of the analyses are provided below.

4.5.1 Zoo data

For the zoological dataset (dowloaded from the UCI Machine Learning Repository and donated by Richard Forsyth's) the problem will be reduced in finding the closest parsimonious dendrogram to a given one.

The dataset consists of 101 observations (animals) and 18 variables; more in detail, 15 variables are binary, highlighting in each animal the presence/absence of hair, feathers, eggs, milk, airbone, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize; one variable is categorical and refers to the number of legs of each animal; one variable refers to the animal name; finally, the last variable is a class attribute, providing the animals' taxonomy in seven classes: mammals, birds, reptiles, fishes, amphibians, insects, and invertebrates. The whole dataset does not contain any missing value.

For the application, we used the 15 binary variables only. From the units-by-variables data matrix, the dissimilarity matrix $\mathbf{D}_1$ of dimension 101×101 was

obtained by computing the squared Euclidean distance between each pair of units. Then, given $\mathbf{D}_1$, its closest ultrametric matrix $\mathbf{U}_1$ was found by applying the UPGMA algorithm on $\mathbf{D}_1$. Finally, the proposed algorithm, applied on $\mathbf{U}_1$ by setting $G = 7$ and using 100 random starts, found a unique ($K = 1$) consensus parsimonious dendrogram. In Figure 4.5 and 4.6, the starting ultrametric matrix and the closest parsimonious dendrogram are shown, respectively. As it is shown in Figure 4.5, the partition of the animals in $G = 7$ clusters with cutoff level 1.94 is not clearly identifiable, because by moving the cutoff level slightly up (level 1.95) or down (level 1.91) the number of clusters of the partition varies from 6 to 8. Thus, there is an uncertainty in the identification of the cutting level. In practice, the visual inspection of the dendrogram does not show a clear distinction between the partitions on 6, 7, or 8 clusters. This situation does not occur in Figure 4.6, where the $G = 7$ classes are clearly visible and identifiable by the investigator. In this case, the taxonomy of animals (mammals, birds, reptiles, fishes, amphibians, insects, and invertebrates) is clearly identified and their clustering aggregations (such as oviparous vs mammals, non-toothed vs toothed and non-aquatic vs aquatic) can be appreciated.

In order to understand whether the classification taxonomy is recognized, we compared it with the partition of the animals in $G = 7$ classes derived from the complete (UPGMA) dendrogram in Figure 4.5 and with the partition corresponding to the consensus parsimonious dendrogram in Figure 4.6. The ARI values are equal to 0.796 and 0.853, respectively. Thus, the taxonomy in 7 classes is better recovered by the PD. Therefore, in terms of classification tasks, our proposal performs better than the standard methodology. The confusion matrix between true partition in 7 classes and the one of PD is displayed in Table 4.7. We observe that most of the animals are correctly classified (bold on the diagonal) and only 13% of animals are misclassified (13 animals out of $n = 101$ animals) and are highlighted in red in the Table 4.7.

Table 4.7. Confusion matrix: true partition compared to obtained partition of animals of the PD (zoo dataset). In bold the correctly classified animals, in red the misclassified animals.

36	0	0	0	5	0	0
0	**20**	0	0	0	0	0
0	1	**4**	0	0	0	0
0	0	0	**13**	0	0	0
0	0	4	0	**0**	0	0
0	0	0	0	0	**8**	0
0	0	0	0	0	3	**7**

It is worthy to observe that the hierarchical aggregations in PD (in Figure 4.6) have a very clear meaning. For $G = 7$, we have: mollusks (aquatic animals of the class 'invertebrates') (C_1), bugs and worm, slug, scorpion (terrestrial animals of the class 'invertebrates') (C_2), birds and tortoile (one animal of class 'reptiles') (C_3), fishes (C_4), amphibians and reptiles (all but tortoile) (C_5), terrestrial mammals (C_6) and finally aquatic mammals (e.g. "dolphin", "platypus", "sealion", "porpoise" and "seal") (C_7). Moreover, the parsimonious dendrogram allows the study of all the aggregations of those clusters into wider ones: the first aggregations into wider clusters occur by grouping terrestrial mammals and acquatic mammals (C_6 and C_7) in the 'mammals' cluster (C_6+C_7) thanks to the variable 'acquatic' and by

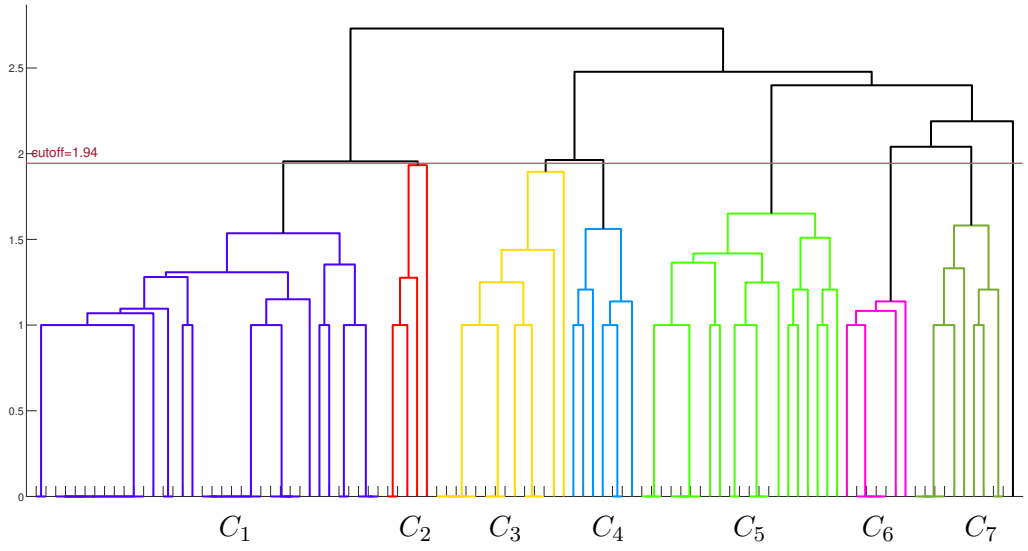


Figure 4.5. Original dendrogram (zoo dataset)

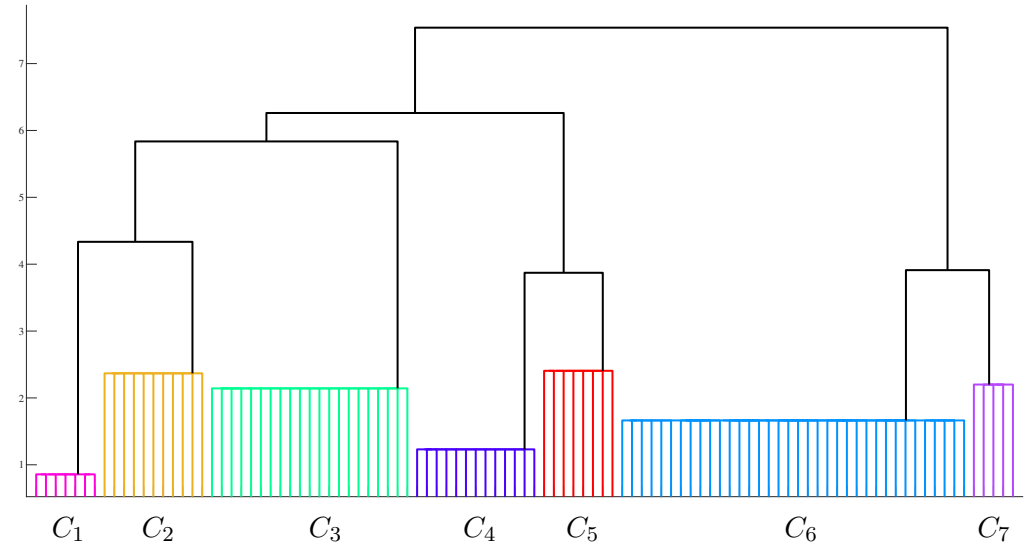


Figure 4.6. Obtained parsimonious dendrogram (zoo dataset)

grouping fishes and amphibians+reptiles (C_4 and C_5) thanks to the variables related to the presence/absence of 'breath' and 'fins'; moreover, the partition into 4 clusters is obtained by grouping mollusks and insects (C_1 and C_2); then, birds (C_3) join the cluster with mollusks and bugs (C_1+C_2) thanks to the variable related to the presence/absence of 'feathers', 'tails' and 'backbone' and thus creating a partition with $G = 3$ clusters. Finally, in order to obtain a partition with $G = 2$ clusters, thanks to the variable 'toothed', this new cluster (C_1+C_2+C_3), characterized by all non-toothed and mostly terrestrial animals and with no fins and no hair, is aggregated with the cluster including fishes, amphibians and reptiles (C_4+C_5), characterized by all toothed animals with no feathers, no hair, mostly aquatic and vertebrates. The obtained cluster ($C_1 + C_2 + C_3 + C_4 + C_5$) referring to 'oviperous' animals and the cluster (C_7+C_8) referring to the 'mammals' make up the partition in only two clusters, where the discriminant variable is the one referring to presence/absence of 'milk'.

4.5.2 Girls' growth curves

For the second application we use the girls' growth curves dataset (Sempé and Médico-Sociale, 1987), downloaded from the webpage of Prof. P.M Kroonenberg and donated by Prof. M. Sempé. The dataset includes 8 physical measurements of 30 girls collected from 1953 until 1975 during a French auxiological study: particularly, the biometric variables related to physical growth (weight, length, crown-rum length, head circumference, chest circumference, arm, calf, pelvis) are measured yearly in the selected girls, who started the experiment at age 4 and ended the experiment at age 15. The data set is therefore a three-way data array with three modes: the first refers to 30 girls, the second to 8 variables, and the third to 12 years.

The objective of the analysis is to compute 12 dendrograms (primary hierarchies) and apply our methodology to identify a fuzzy secondary partition of them and within each class of the secondary partition, identify a consensus parsimonious dendrogram. Before applying the new methodology, a preliminary data manipulation is needed, by normalizing the overall dataset with min-max normalization, where the min and the max of variables are over the entire period (4-15 years old). Then, the overall average trends of the 8 observed variables among the 30 girls are shown in Figure 4.7. The trends are clearly increasing and it is possible to observe a change in the slope of the growth around age 9-10.

The starting $H = 12$ dendrograms (H ultrametric matrices or primary hierarchies) are obtained by considering the H matrices $\mathbf{X}_h^N$, $h = 1, \ldots, 12$, where $\mathbf{X}_h^N$ is the 30×8 normalized data matrix referring to the physical measurements at age h. Then, we obtained the H dissimilarity matrices $\mathbf{D}_h$, $h = 1, \ldots, 12$ by computing the Euclidean distance between each pair of units. Finally, in Figure 4.8 the dendrograms $\Delta = \{\delta_1, \ldots, \delta_{12}\}$ of Ward's method of hierarchical clustering, computed on matrices $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{12}$ are from age 4 to age 15 years old.

Given the $H = 12$ primary hierarchies, the algorithm is applied on the corresponding ultrametric matrices $\mathbf{U}_1, \ldots, \mathbf{U}_{12}$, by using 100 random starts and setting $G = 3$ and $K = 2$, as suggested by Kroonenberg et al. (1987). The algorithm finds a fuzzy partition of the primary hierarchies into $K = 2$ clusters and within each class of the secondary partition identifies a consensus parsimonious dendrogram, i.e. a $(2G - 1)$-ultrametric matrix, where $G = 3$ identifies the number of classes of the girls.

The obtained fuzzy partition is illustrated in Table 4.8, where for each age of the girls the corresponding cluster and the related membership degree are reported. In particular, we observe that the chronological order is retained, as ages 4-8 belong

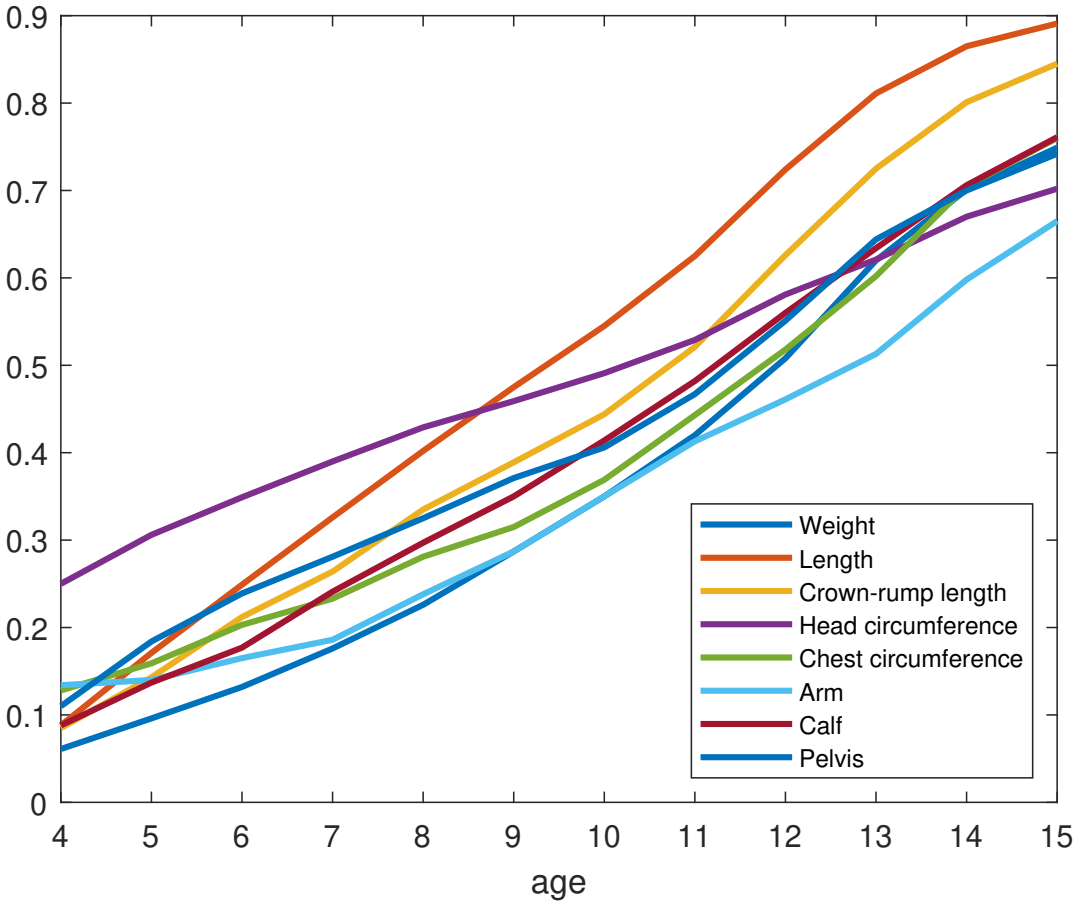


Figure 4.7. Average trends of the variables of interests from age 4 until age 15 (girls' growth curves dataset).

almost hardly to the first cluster, and ages 11-15 belong almost hardly to the second cluster. In addition, ages 9 and 10 are more softly associated to both clusters, having membership degrees quite fuzzier and closer to one another. This result is interesting and meaningful: indeed, at ages 9 and 10 we observed in Figure 4.7 that several curves change their slopes. More generally, it has been shown by many research studies (Breehl and Caban, 2021; Farello et al., 2019) that the puberty period for girls starts around age 8 and therefore ages 9 and 10 are exactly when the puberty period is in progress. For this reason, we can conclude that the proposed approach allows us to detect the ages which can be considered as a transitional period in these data .

Table 4.8. Cluster assignment of the original dendrograms to 2 clusters, with the highest membership degree.

Age	4	5	6	7	8	9	10	11	12	13	14	15
Cluster	1	1	1	1	1	1,2	1,2	2	2	2	2	2
Membership degree	0.91	0.95	0.96	0.96	0.88	0.67,0.33	0.44, 0.56	0.82	0.90	0.91	0.87	0.78

In addition, the resulting parsimonious consensus dendrograms are shown in

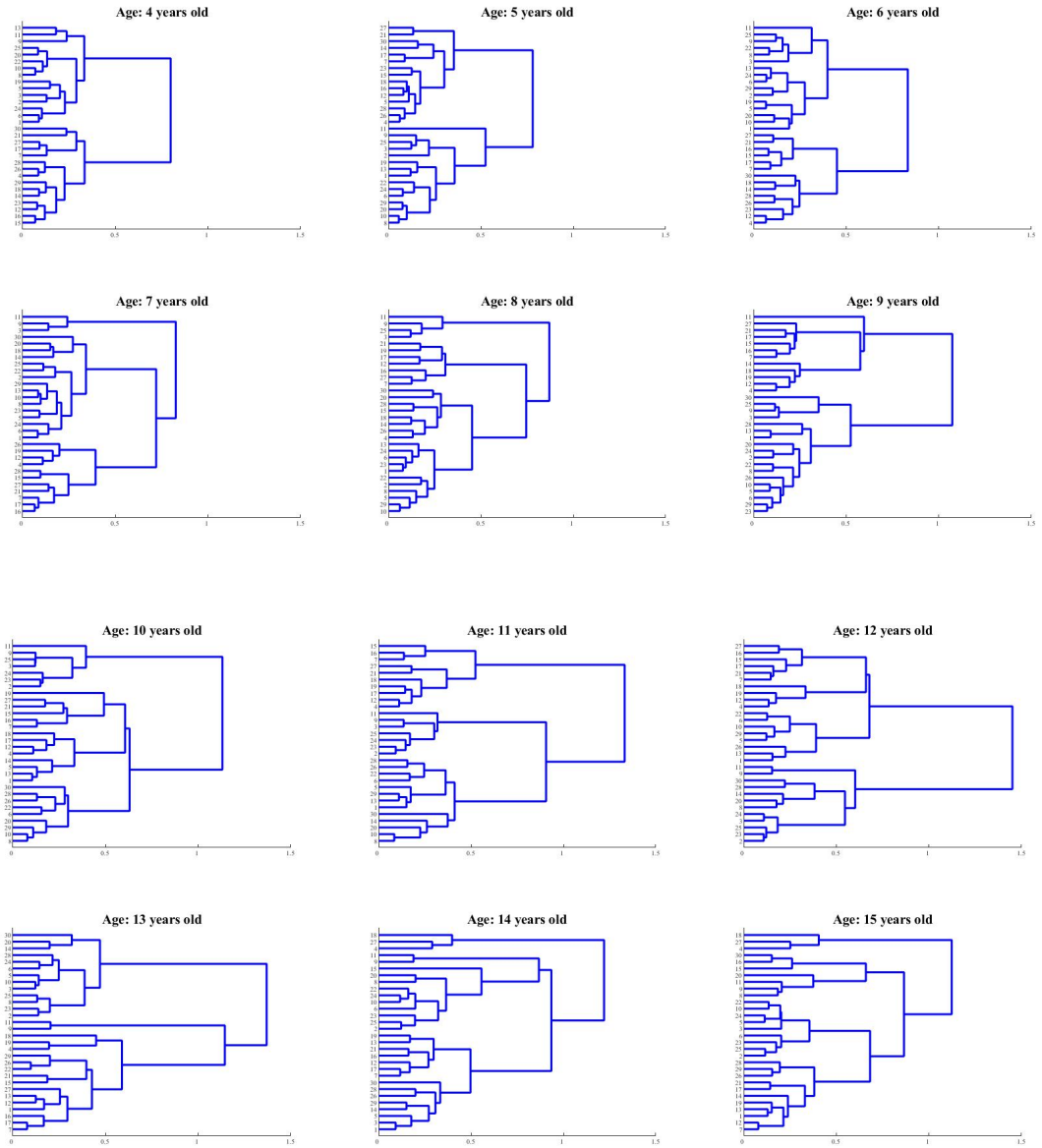


Figure 4.8. Hierarchical clustering of the girls by 8 biometric variables from age 4 until age 15 (girls' growth curves dataset).

Figure 4.9, where we can clearly see the aggregations of the girls into $G = 3$ clusters and the distinct agglomerations of these clusters. It is worthy to notice that clusters

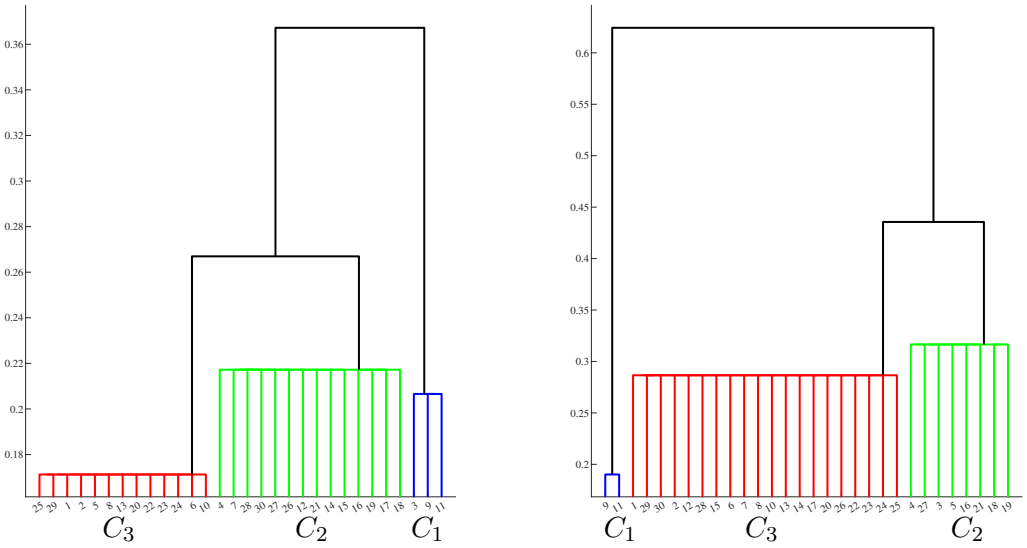


Figure 4.9. Resulting consensus dendrograms, representing hierarchical clustering of girls by 8 biometric variables (girls' growth curves dataset).

C_1, C_2 and C_3 of girls, identified in the two parsimonious dendrograms, have some common individuals, but also show some differences due to the fact that some girls have changed the pattern of growth from the first period to the second, thus, moving from one cluster to another. In order to better visualize and interpret the results, the hard partition of ages was considered by applying MAP (maximum a posteriori) to the membership degree matrix. This allowed us to have separate plots of the trends of the variables for the two clusters of ages, and the three clusters of girls. For the visualization task and to reduce the amount of trends to be plotted it was decided to plot only the three dimensions identified by Kroonenberg et al. (1987) and named *Skeletal Length*, *Skeletal Width* and *Stoutness*: *Skeletal Length* is referred to variables length and crown-rump length, *Skeletal Width* to variables head and pelvis, *Stoutness* to variables weight, chest, arm and calf. The trends of these dimensions are shown in Figure 4.10 (solid lines) separately by cluster of ages and cluster of girls, as well as the average trend of each dimension in that specific period (i.e. cluster of ages) (dotted lines).

From Figure 4.10, it is possible to comment on the clusters of girls obtained separately for each cluster of ages. We observe that the trends of the girls who belong to the first cluster between ages 4 and 9 (C_1 in the left dendrogram in Figure 4.9) are quite far below the average level, meaning that those girls are *below average stature*, characterized by a less rapid growth and low levels of biometric variables. Girls belonging to the second cluster when they are between 4 and 9 years old (C_2 in the left dendrogram in Figure 4.9) grow on average: trends are very close to the corresponding average; they can be considered a cluster of *average stature girls*. Finally, those who belong to the third cluster of girls between ages 4 and 9 (C_3 in the left dendrogram in Figure 4.9) have trends far above the average ones. This means that those girls are the most robust and tallest ones (*above average stature girls*). In addition, focusing on the ages 10-15, girls who fall into the first cluster are the

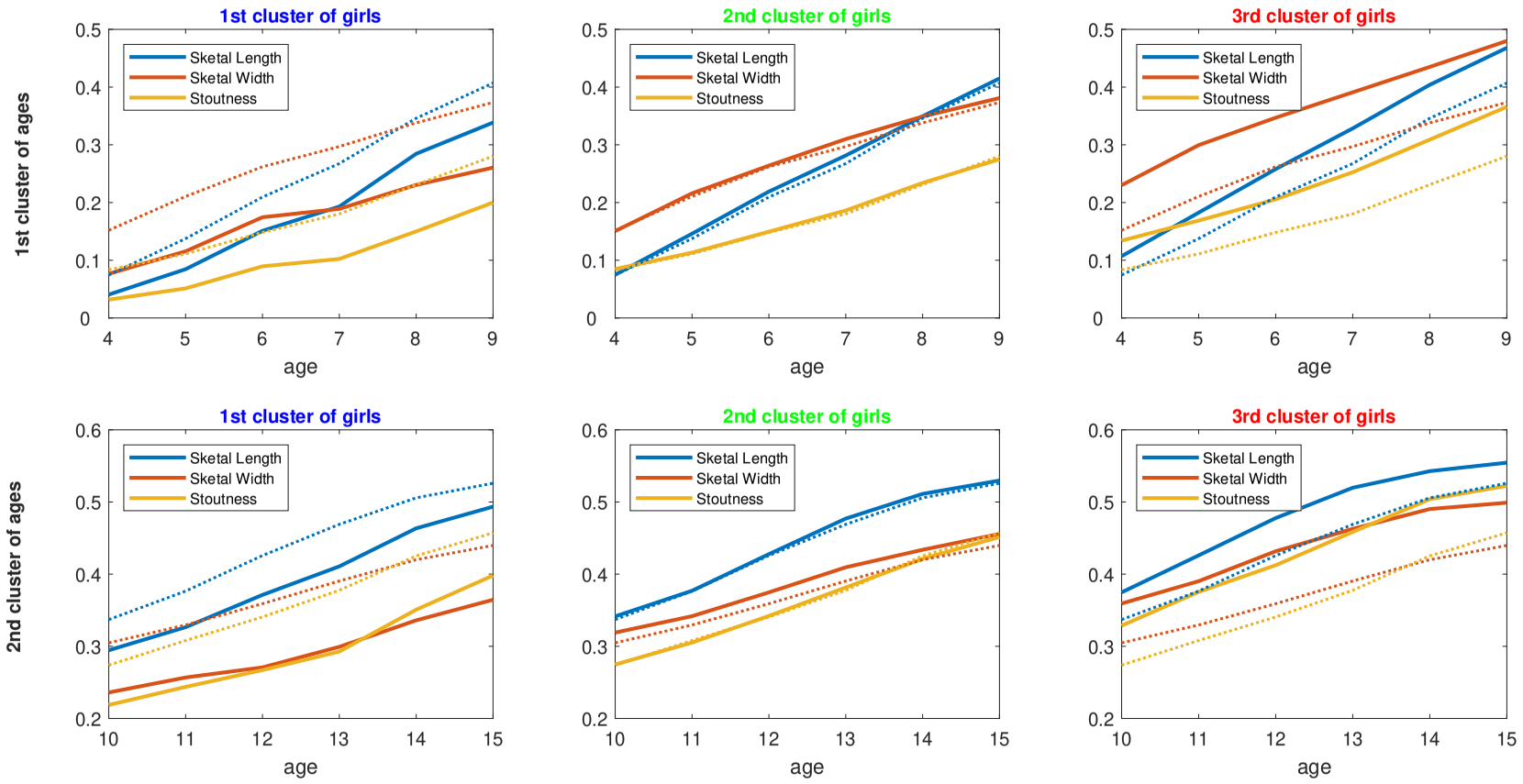


Figure 4.10. Solid lines: trends of the dimensions taken from the variables of interest separately per cluster of ages and class of girls. Dotted lines: average trends of the dimensions in the entire period. Title of the subplots are colored with the color of the class of girls in Figure 4.9 (girls' growth curves dataset).

ones in the first cluster between ages 4-9 except for unit 3 (as depicted in Figure 4.9, cluster C_1): their trends have similar behaviour as in the earlier ages, being far below the average; therefore this cluster identifies the *below average stature girls*. The second cluster of girls being between 10 and 15 years old (C_2 in the right dendrogram in Figure 4.9) have trends following the average, as happens in the earlier ages, except for Skeletal Length, which is slightly above the average. Therefore, the cluster groups together the *average stature girls*; it is worth mentioning that unit 3, who falls in the *below average stature girls* cluster between ages 4 and 9, joins the *average stature girls* cluster in the next ages' period, meaning that her biometric variables' trends returned to the average level. Finally, the third cluster of girls between ages 10-15 (C_3 in the right dendrogram in Figure 4.9) have similar trends as the earlier ages, thus identifying the *above average stature girls*.

In conclusion, the analysis allowed us to identify two distinct clusters of ages (one for ages 4-8, one for ages 11-15), except for ages 9-10 which are in the middle of the two, characterizing a transitional period in the girls' physical growth. In addition, for each cluster of ages a consensus parsimononious dendrogram has been identified. Both of the consensus dendrograms identify three distinct clusters of girls. By analyzing these separately per cluster of ages, we noticed that they correspond to *below average stature girls* (C_1 in Figure 4.9), *average stature girls* (C_2 in Figure 4.9) and *above average stature girls* (C_3 in Figure 4.9); more specifically, considering the entire period, very few girls are always under the average (see clusters C_1 of Figure 4.9), some girls who were on average during ages 4-9 became above the average in the following ages (see for example unit 7 and 28 move from C_2 to C_3 in Figure 4.9), and some girls who were above the average during ages 4-9 became on average during the next years (see for example unit 5 moves from C_3 to C_2 in Figure 4.9).

4.6 Conclusion

The new methodology proposed in this chapter makes it possible to solve several problems:

(i) Given H primary hierarchies, obtain a fuzzy secondary partition of the primary hierarchies, and for each class of the secondary partition identify a consensus well-structured partition (where within-cluster distances are all smaller than between-cluster distances). This problem consists of solving simultaneously a fuzzy partitioning problem to identify the secondary partition and K least-squares optimal differences between ultrametric matrices of a cluster of the secondary partition and a consensus well-structured partition that should identify the partition closest to the hierarchies (see the problem (4.P1.a) in Section 4.3);

(ii) Given H primary hierarchies, obtain a fuzzy secondary partition of the primary hierarchies, and for each class of the secondary partition identify a consensus parsimonious dendrogram. This problem consists of solving simultaneously a fuzzy partitioning problem to identify the secondary partition and K least-squares optimal differences between a subset of ultrametric matrices and a consensus parsimonious dendrogram (see the problem (4.P1.b) in Section 4.3);

(iii) Given a single hierarchy (dendrogram), find the closest well-structured partition. This is a problem frequently considered in hierarchical clustering, where the investigator has to find an optimal partition by the visual inspection of the dendrogram or by means of a specific methodology. This problem consists

of solving the problem (i) above when a single dendrogram is observed or computed, and it is necessary to find a single well-structured partition (see the problem (4.P1.a) with $H = 1$ and $K = 1$, in Section 4.3).

(iv) Given a single hierarchy (dendrogram), find the closest parsimonious dendrogram. This is an evolution of the previous problem (iii) where the investigator wishes to find an optimal partition in G classes in the ultrametric matrix (dendrogram) and the corresponding optimal aggregations from G to 1. This problem consists of solving the problem (ii), when a single dendrogram is observed or computed and it is necessary to find a single consensus parsimonious dendrogram (see the problem (4.P1.b) with $H = 1$ and $K = 1$, in Section 4.3).

For problems (iii) and (iv) if the hierarchy is not initially given, i.e. if a dissimilarity matrix is given, then its corresponding hierarchy or ultrametric matrix can be obtained by applying UPGMA, or any other hierarchical clustering algorithm, to the dissimilarity matrix.
For problems (i) and (ii) a secondary fuzzy partition that allows each dendrogram of the primary partition to belong to all clusters of the secondary partition according to different membership degrees is required. This guarantees great flexibility in the results and their interpretation.
For each class of the fuzzy partition, a consensus hierarchy (dendrogram) is identified. However, several authors have noted that the complete sets of partitions and clusters of the dendrogram are not all used by investigators, even hindering interpretation (Gordon, 1999). One approach for resolving this difficulty has involved the construction of a parsimonious dendrogram that contains a limited number of internal nodes. Some information is lost here, but the main features of the data are represented more clearly (Gordon, 1999). For this reason, the consensus hierarchy in this chapter has a parsimonious structure.
It is important to recall that, using a fuzzy approach to clustering, all primary hierarchies contribute to the definition of each consensus hierarchy according to their membership degree. Therefore, each consensus hierarchy is mainly determined by the primary hierarchy whose membership degree is sufficiently high (e.g. >0.7), while the contribution of primary hierarchies whose membership degree is low is less relevant. Therefore, it is important to emphasize that the contribution of each primary hierarchy is taken into account in the definition of consensus hierarchies by fuzzy assignment to clusters. However, it is important to mention that, especially to better visualize and interpret the results, very often MAP is applied to the membership degree matrix to hardly divide the occasions into clusters.

The proposed methodology has been tested in an extended simulation study, where 1000 three-way arrays of ultrametric matrices have been generated. Two scenarios of hard assignment and fuzzy assignment of the primary hierarchies to the consensus hierarchies have been considered. The study showed good results, not only in recovering the underlying true secondary partition but also in identifying consensus parsimonious dendrograms very similar to the original ones.

The methodology has also been applied to real datasets; the results of the analyses show that the proposed methodology is helpful in partitioning the primary hierarchies in a fuzzy manner, by identifying correctly the hierarchies which share characteristics with more than one cluster of the secondary partition: for example, in the application to girls' growth curves dataset, two periods of contiguous ages are identified and the hierarchies corresponding to two transitional years from one period to the following are reasonably softly assigned to both periods. In addition, for each class (period) of the fuzzy partition, the methodology identifies a consensus parsimonious dendrogram, which really facilitates the interpretation of the aggregation of the girls.

This research work introduces a new methodology in multidimensional data analysis and opens up the possibility to new applications and further developments. Among the further developments, it might be interesting to study in depth how sensitive the algorithm is to the choice of linkage method used in hierarchical clustering.

Chapter 5

Fuzzy clustering and dimensionality reduction of a three-way data matrix

5.1 Introduction

This chapter addresses the problem of obtaining a fuzzy partition of the set of units-by-variables matrices. A fuzzy clustering technique is proposed to identify a set of K clusters, their associated K consensus matrices, and the membership degrees of each original data matrix to the detected clusters. It has to be observed that the clusters of years and the corresponding consensus matrices are identified supposing that the dissimilarity relation between units and covariance structure between variables do not change much so that data matrices in the cluster are perceived to be one similar to the other. Therefore, the investigator is also interested to synthesize the covariance structure of each cluster by means of a dimensional reduction model. Since variables are supposed correlated, second-order disjoint factor analysis (Cavicchia and Vichi, 2022) is supposed to identify a latent hierarchical structure of the variables.

Thus to summarize, given a three-way three-mode data matrix $\mathbf{X}$, in this chapter, a simultaneous fuzzy partitioning of the occasions is found that best identifies a reduced set of K consensus matrices, each one with covariance structure summarized by a different second-order disjoint factor analysis. For each consensus matrix is identified: *a*) a set of first-order factors with the corresponding loading matrix and *b*) a second-order factor, namely a general composite indicator. Formally, given K consensus matrices, for each $k = 1, \ldots, K$, a set of factors $\mathbf{Y}_k$ and their corresponding loading matrix A_k and a general composite indicator g_k will be identified, reporting the scores of each of the N units in the latent dimension. The ranking of the values of the latent dimension (composite indicator) allows for assessing differences between units and defining their total order for each class of occasions. Actually, the final clustering of the units according to the latent dimension will allow the identification of equivalence classes of units, where it is not possible to appreciate their differences, while differences will be assessed between classes of units. This means that the ranking will be considered only between clusters of units.
A flowchart describing the methodology and clarifying the process is displayed in Figure 5.1.

It has to be clearly observed that the fuzzy clustering for the occasions and the second-order factor analysis for the variables are estimated simultaneously in the

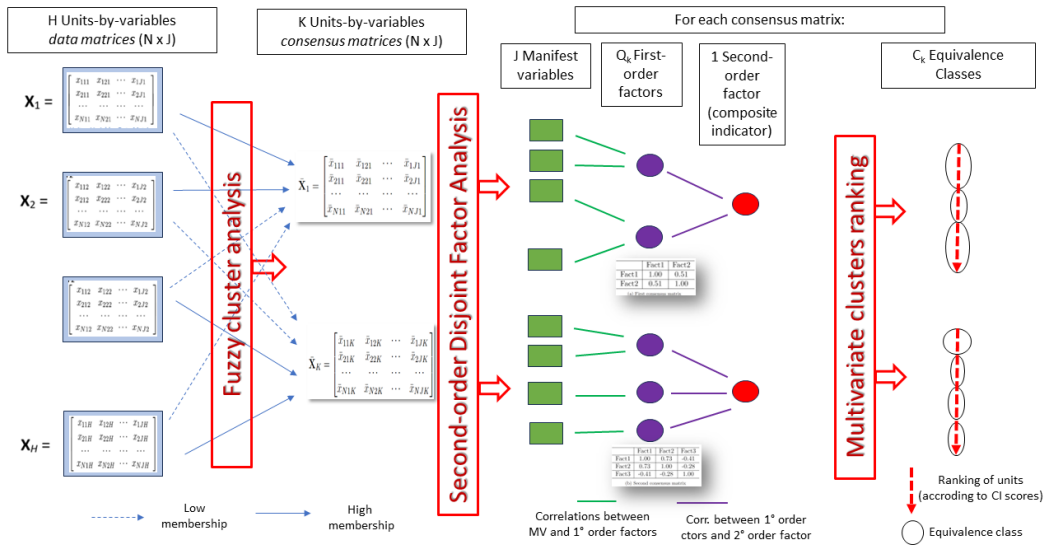


Figure 5.1. Flowchart describing the proposed methodology

three-way data matrix to avoid the sequential estimation of the clustering and factor analysis that may produce masking effects of the global optimal solution.

The new methodology is estimated according to a least-square coordinate descent method by using an efficient algorithm. As usual in many multivariate methodologies, the algorithm does not guarantee that the global optimal solution of the clustering and the simultaneous dimensional reduction of the three-way array is achieved. For this reason, it is advised to run the algorithm from different starting points to increase the chance to detect the global optimal solution.

The remainder of this chapter is organized as follows. Section 5.1.1 is fully dedicated to the review of the literature in this framework; Section 5.2 is useful to recall the basic notions and models used in the methodological proposal; Section 5.3 describes the proposed methodology and its estimation. In Section 5.4, the new methodology is applied to a real-case study, namely to the well-being dataset How's Life. Finally, Section 5.5 gives remarks and considerations on future developments.

5.1.1 Related Literature

The proposed methodology is included in a general framework of the clustering techniques for three-way data, or, from an Information Technology (IT) point-of-view, in a 'multi-view data' background. Most of the proposed approaches focus on clustering with a hard assignment of units to clusters. For example, a constrained algorithm to hardly partition the occasions according to units and variables has been proposed by Cariou, Alexandre-Gouabau, and Wilderjans (2021) and an algorithm in a maximum likelihood framework performing clustering to obtain a hard partition of the occasions has been proposed by Cappozzo, Alessandro, Michael, et al. (2021). A different problem is addressed by Bocci and Vicari (2019) who proposed an algorithm to cluster a three-way two-mode data. Another approach of clustering in a three-way three-mode data framework is proposed by Schoonees, Groenen, and van de Velden (2021) who simultaneously hardly partition the three modes.

A completely different approach is provided by Durieux and Wilderjans (2019),

who provide a two-way symmetric dissimilarity matrix synthesizing the three-way data. Then, they apply standard hard clustering algorithms, such as Partitioning around Medoids, to the matrix.

In a framework of multi-view data, collected from multiple sources or containing multiple features, many research works have been published. For example, Khan et al. (2022) have proposed a multi-view data clustering via nonnegative matrix factorization with manifold regularization. The non-negative matrix factorization is applied before clustering to obtain a meaningful clustering solution. However, the factorization frequently does not retain some characteristics of the data structure, and a manifold regularization is needed to retain the geometrical feature of the data. In the same framework of multi-view clustering with matrix factorization, also the contribution of Yang et al. (2020) is interesting. They suppose that the clustering performance is characterized by the data distribution, and they have decided to propose a tri-factorization based on the Non-Negative Matrix Factorization model with an embedding matrix. In this way, they claim that the obtained consensus matrices better represent the multi-view data in the subspace. With a similar aim to improve clustering performance, Zhao et al. (2023) have proposed to combine multi-view clustering with binary code learning including several new useful features. Indeed, they have proposed an orthogonal mapping binary graph method (OMBG), which makes every view of the multi-view data orthogonal and embeds a binary graph structure into the unified binary multi-view clustering framework. The orthogonalization allows to eliminate redundant information and the binary graph structure permits to achieve an optimal clustering result.

5.2 Theoretical Framework

In order to allow the reader to clearly follow the proposed new methodology, it is worth mentioning the necessary theory, which is based on the deep knowledge of the three-way data analysis, together with simultaneous fuzzy clustering and factorial methods of dimensional reduction.

In a framework of three-way three-mode data, our interest is to consider H unit-by-variable matrices and to apply a fuzzy clustering technique to softly group them into K clusters, with $K << H$. So this is a generalization of the usual partitioning problem because the elements to be classified are not units but matrices of data unit-by-variable. In this case, several occasions (layers), such as years, can belong to more than one cluster of matrices and therefore can contribute to the definition of more than one consensus matrix. The membership degree matrix $[\mu_{hk}]_{h=1,\ldots,H,k=1,\ldots,K}$ considered in this chapter has dimension $H \times K$ and its generic element μ_{hk} belongs to the interval $[0, 1]$, $\forall\ h = 1, \ldots, H,\ \forall\ k = 1, \ldots, K$.

Factor analysis techniques are used to reduce the space of variables by using latent variables (LV) factors that produce a dimensionality reduction of the variables. These methodologies are often applied when the given set of J manifest variables (MV) are correlated and represent one or more latent concepts that are not directly measurable. The latent concepts reconstruct the MVs with different levels of abstraction, from the most specific, closer to the observed variables, to the most general ones synthesizing the common relationships of the MVs. Usually, a reduced set of LVs, say Q, (factors, latent dimensions, or components are terms used interchangeably) are able to explain and reconstruct most of the information contained in the original data-set. Exploratory Factor Analysis (EFA), Principal Components Analysis (PCA), and many variations and extensions are generally used to solve this task.

In general, LVs, and MVs (observed indicators) are statistically related and the

Factor Analysis (EFA) model describes their relationships by measuring with the LVs the unobservable latent concepts (constructs).

In general, two types of relationships can exist between LVs and MVs: they can be either reflective or formative. When a "reflective" relationship exists it means that LVs cause the MVs, or in other words, MVs reflect the influence of LVs, or, are affected by the LVs. On the contrary, a "formative" relationship between MVs and LVs exists when MVs cause (define, explain) the LVs. The reflective relation assumes that the MVs are correlated and the LVs explain this correlation, while the formative approach supposes that MVs are uncorrelated and that they are used to form the LVs.

Based on the type of relationship that is supposed to be observed, two approaches are possible in FA. If the researcher assumes a model behind the data, because a theory supports the existence of fixed relationships between MVs and LVs, the analysis is used to confirm or reject the model, that is, the presence and the level of these relations. In this case, the analysis follows a confirmative approach, and a Confirmatory Factor Analysis is used (CFA). On the contrary, if the researcher has not a theory that defines the existence and the level of relationships, which are unknown, the analysis follows an explorative approach, or, a mixed confirmative/explorative approach. In this last case, the known relations between MVs and LVs are fixed and all the other relationships between LVs and MVs are explored by the analysis.

The aforementioned methodologies aim to define uncorrelated LVs, but very often these last are correlated because there are cross-loadings, i.e., MVs that are correlated with two or more LVs and that induce a correlation between LVs. Thus, researchers wish to explain this remaining correlation between LVs by means of a general LV of the "second-order" with respect to those obtained by the "first-order" FA. For example, Vichi (2017) has proposed the Disjoint Factor Analysis (DFA) that can identify a reduced set of first-order factors with a sparse loading matrix and define two or more factors that could be correlated, although not highly correlated because otherwise these factors would have been further combined into a smaller number. In this context, Cavicchia and Vichi (2022), have introduced the second-order DFA that identifies a two-level hierarchy of factors. In practice, they first obtain a first-order factor analysis identifying a reduced set of multidimensional concepts, then the second order allows them to obtain a single general factor. The first-order and second-order LVs are estimated simultaneously. It is useful to explain this methodology with additional details and clear formalization because it will be included in the new proposed methodology for three-way data.

5.2.1 Second-order disjoint factor analysis

For the moment let the three-way data matrix degenerate into an $N \times J$ data matrix $\mathbf{X}$, that is, H=1. The Second-Order Factor Analysis (2O-FA) model can be obtained by merging two different nested models involving two typologies of latent factors: Q ($Q << J$) first-order factors and a single (nested) general factor. In particular, letting $\mathbf{x}$ be the $(J \times 1)$ multivariate observation, then:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{w}, \tag{5.1}$$

$$\mathbf{y} = \mathbf{c}g + \mathbf{u}, \tag{5.2}$$

where $\mathbf{A}$ is the $(J \times Q)$ matrix of first-order factors loadings, $\mathbf{y}$ is the $(Q \times 1)$ vector denoting the first-order factor scores, and $\mathbf{w}$ is a $(J \times 1)$ random vector of errors; in addition, g is the general factor and $\mathbf{c}$ is the $(Q \times 1)$ vector of general factor loadings. Finally, $\mathbf{u}$ is a $(Q \times 1)$ random vector of errors.

Model (5.1) finds Q specific theoretical constructs by means of a common factor model that identifies common information with Q factors related to the Manifest Variables, while model (5.2) detects the general latent construct by means of a one-factor model.

Therefore, given an $N \times J$ data matrix $\mathbf{X}$, by merging models (5.1) and (5.2), the 2O-FA model can be written in a matrix form as follows:

$$\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{A}' + \mathbf{E}, \tag{5.3}$$

where $\mathbf{g} = [g_1, \ldots, g_N]'$ is the $N \times 1$ vector denoting the second-order (general) factor scores, and $\mathbf{E}$ is the $N \times J$ matrix of errors.

Since the *Disjoint* FA (DFA) is used as a general framework to simplify factors interpretation, it is worth recalling the additional property of the loading matrix $\mathbf{A}$ and the advantages of the model. According to disjoint models, the loading matrix $\mathbf{A}$ is of the following form:

$$\mathbf{A} = \mathbf{B}\mathbf{V}, \tag{5.4}$$

where $\mathbf{V} = [v_{jq}]$ is a $(J \times Q)$ binary and row stochastic matrix identifying a partition of MVs into Q subsets corresponding to Q factors. If the jth MV belongs to the qth subset then $v_{jq} = 1$, otherwise, $v_{jq} = 0$; whereas, $\mathbf{B} = \text{diag}(b_1, \ldots, b_J)$ is a $(J \times J)$ scaling diagonal matrix weighting variables and such that $b_j^2 > 0$, and the operator $\text{diag}(\cdot)$ of a vector produces a diagonal matrix of that vector.
The main advantage of this *Disjoint* model is the simplification of the loading matrix $\mathbf{A}$ that captures the simplest structure, i.e. the sparsest, which means that each MV is reconstructed only by a single factor. In other words, a disjoint class of variables reflects the influence of a single factor, and all classes form a partition of the MVs.

It is worth specifying that even if DFA assumes orthogonal factors, that is, $\mathbf{\Sigma_y} = \mathcal{I}_Q$, in the 2O-DFA this condition is relaxed in order to allow a hierarchical structure of the data.

5.3 The new methodological proposal for three-way data

The methodology proposed in this chapter aims to find a fuzzy partition in K clusters of the H unit-by-variable matrices and identify for each class a consensus matrix. Simultaneously, it aims to reduce the dimension of the obtained consensus matrices, by applying a 2O-DFA to each of them. Therefore, given a set of $H > 1$ matrices, the goal is to softy cluster them into $K > 1$ groups, each of them summarized by a set of Q_k first-order latent variables and one second-order latent variable (or general composite indicator). The number Q_k of first-order latent variables can be different in each consensus matrix.
In order to achieve this goal the following optimization problem has to be solved

w.r.t. continuous variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k, \mu_{hk}^m$ and discrete variables $\mathbf{V}_k$:

$$
\begin{cases}
\text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\mu_{hk}^m & \text{(5.P1)} \\
\text{s.t.} & \\
\sum_{k=1}^{K} \mu_{hk} = 1 & \text{for } h = 1,\dots,H \quad \text{(5.}C_1\text{)} \\
\mu_{hk} \in [0,1] & \text{for } h = 1,\dots,H,\ k = 1,\dots,K \quad \text{(5.}C_2\text{)} \\
v_{jqk} \in \{0,1\} & \text{for } j = 1,\dots,J,\ q = 1,\dots,Q_k \quad \text{(5.}C_3\text{)} \\
 & \quad k = 1,\dots,K \\
\sum_{q=1}^{Q_k} v_{jqk} = 1 & \text{for } j = 1,\dots,J,\ k = 1,\dots,K \quad \text{(5.}C_4\text{)} \\
\mathbf{B}_k = \text{diag}(b_{1k},\dots,b_{Jk}) & \text{for } k = 1,\dots,K \quad \text{(5.}C_5\text{)} \\
b_{jk}^2 > 0 & \text{for } j = 1,\dots,J,\ k = 1,\dots,K \quad \text{(5.}C_6\text{)}
\end{cases}
$$

Constraints (5.C_1) and (5.C_2) guarantee that the set of unit-by-variable matrices $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_H$ is partitioned in a fuzzy way, i.e., into K clusters: each matrix belongs to the k-th cluster with the h-th membership degree μ_{hk}. In addition, constraints (5.C_3), (5.C_4), (5.C_5), (5.C_6) allow the dimensionality reduction model to be disjoint. More in details, the binary and row-stochasticity properties of $\mathbf{V}_k, k = 1,\dots,K$ (constraints (5.C_3) and (5.C_4)) allow each MV to contribute to the definition of one factor only, while the constraints on $\mathbf{B}_k, k = 1 \dots, K$ (constraints (5.C_5) and (5.C_6)) allow the matrix to give each variable the weight in defining each component.
Finally, the fuzziness of the partition is controlled by the parameter m, named *fuzzifier*. In particular, when $m \to 1$ the partition tends to become hard, i.e. the membership degrees tend to be either 0 or 1; for $m \to \infty$ membership tend to be constant and equal to $1/K$.

5.3.1 Least-Squares Estimation

In order to implement (5.P1), it is worth noting that it can be decomposed into two alternating minimization sub-problems:

(A) the partial minimization of the objective function of (5.P1) w.r.t. continuous

variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k$, and discrete variables $\mathbf{V}_k$, when $\hat{\mu}_{hk}$ is given.

$$\begin{cases} \text{minimize} \displaystyle\sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\hat{\mu}_{hk}^m & (5.\text{P2}) \\ \text{s.t.} \\ v_{jqk} \in \{0,1\} & \text{for } j=1,\ldots,J,\ q=1,\ldots,Q_k,\ k=1,\ldots,K \quad (5.C_3) \\ \displaystyle\sum_{q=1}^{Q_k} v_{jqk} = 1 & \text{for } j=1,\ldots,J,\ k=1,\ldots,K \quad (5.C_4) \\ \mathbf{B}_k = \text{diag}(b_{1k},\ldots,b_{Jk}) & \text{for } k=1,\ldots,K \quad (5.C_5) \\ b_{jk}^2 > 0 & \text{for } j=1,\ldots,J,\ k=1,\ldots,K \quad (5.C_6) \end{cases}$$

The solution of this sub-problem (A) can be found by using the Sequential Quadratic Programming (SQP) algorithm (Powell, 1983).

It is worth noting that objective function of (5.P2) can be rewritten as follows:

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 = \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \bar{\mathbf{X}}_k\|^2 + \sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \quad (5.5)$$

where,

$$\bar{\mathbf{X}}_k = \frac{1}{\sum_{h=1}^{H}\hat{\mu}_{hk}^m}\sum_{h=1}^{H}\hat{\mu}_{hk}^m\mathbf{X}_h, \quad (5.6)$$

is the weighted arithmetic mean matrix of $\mathbf{X}_h$, for $h=1,\ldots,H$, weighted by $\hat{\mu}_{hk}^m$, which represents the closest LS solution of the unconstrained (5.P2) problem (first term of the right hand side of (5.5)). Thus, it remains to minimize the second term of the right hand side of (5.5) under the constraints ($5.C_3$)-($5.C_6$), that is, the following optimization problem:

$$\begin{cases} \text{minimize} \displaystyle\sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\hat{\mu}_{hk}^m \\ \text{minimize} \displaystyle\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \sum_{h=1}^{H}\hat{\mu}_{hk}^m & (5.\text{P3}) \\ \text{s.t.} \\ v_{jqk} \in \{0,1\} & \text{for } j=1,\ldots,J,\ q=1,\ldots,Q,\ k=1,\ldots,K \quad (5.C_3) \\ \displaystyle\sum_{q=1}^{Q} v_{jqk} = 1 & \text{for } j=1,\ldots,J,\ k=1,\ldots,K \quad (5.C_4) \\ \mathbf{B}_k = \text{diag}(b_{1k},\ldots,b_{Jk}) & \text{for } k=1,\ldots,K \quad (5.C_5) \\ b_{jk}^2 > 0 & \text{for } j=1,\ldots,J,\ k=1,\ldots,K \quad (5.C_6) \end{cases}$$

To prove the equality between problems (5.P2) and (5.P3), we prove that the minimization of

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\|\mathbf{X}_h-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\hat{\mu}_{hk}^m \tag{5.7}$$

w.r.t. continuous variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k$, and discrete variables $\mathbf{V}_k$, under the constraints ($5.C_3$)-($5.C_6$), is equivalent to the minimization of

$$\sum_{k=1}^{K}\|\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\sum_{h=1}^{H}\hat{\mu}_{hk}^m \tag{5.8}$$

w.r.t. continuous variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k$, and discrete variables $\mathbf{V}_k$, under the constraints ($5.C_3$)-($5.C_6$). It has to be proved that the following decomposition holds:

$$\|\mathbf{X}_h-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2=\|\mathbf{X}_h-\bar{\mathbf{X}}_k\|^2+\|\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \tag{5.9}$$

Proof.

$$\begin{aligned}
&\|\mathbf{X}_h-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2=\\
&\|\mathbf{X}_h-\bar{\mathbf{X}}_k+\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2=\\
&=tr[((\mathbf{X}_h-\bar{\mathbf{X}}_k)+(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'))'((\mathbf{X}_h-\bar{\mathbf{X}}_k)+\\
&+(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'))]=\\
&=tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\mathbf{X}_h-\bar{\mathbf{X}}_k)]+\\
&+tr[(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]+\\
&+tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]+\\
&+tr[(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')'(\mathbf{X}_h-\bar{\mathbf{X}}_k)]=\\
&=tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\mathbf{X}_h-\bar{\mathbf{X}}_k)]+\\
&+tr[(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]+\\
&+2tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]=\\
&=\|\mathbf{X}_h-\bar{\mathbf{X}}_k\|^2+\|\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2+\\
&+2tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]
\end{aligned} \tag{5.10}$$

The second-last equality is given by the fact that $tr[\mathbf{A}'\mathbf{B}]=tr[\mathbf{B}'\mathbf{A}]$. Note that $tr[(\mathbf{X}_h-\bar{\mathbf{X}}_k)'(\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k')]$ is 0, as $\bar{\mathbf{X}}_\mathbf{k}$ is the weighted arithmetic mean matrix of matrices $\mathbf{X}_h$. □

Therefore, we have that

$$\text{minimize}\sum_{h=1}^{H}\sum_{k=1}^{K}\|\mathbf{X}_h-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\hat{\mu}_{hk}^m, \tag{5.11}$$

is equivalent to

$$\text{minimize}\Big(\sum_{h=1}^{H}\sum_{k=1}^{K}\|\mathbf{X}_h-\bar{\mathbf{X}}_k\|^2\hat{\mu}_{hk}^m+\sum_{h=1}^{H}\sum_{k=1}^{K}\|\bar{\mathbf{X}}_k-\mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2\hat{\mu}_{hk}^m\Big), \tag{5.12}$$

which is equivalent to

$$\text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \bar{\mathbf{X}}_k\|^2 \hat{\mu}_{hk}^m + \text{minimize} \sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \hat{\mu}_{hk}^m \quad (5.13)$$

because, given $\hat{\mu}_{hk}$, the two minimization problems are independent.
Moreover, we observe that $\sum_{h=1}^{H}\sum_{k=1}^{K} \|\mathbf{X}_h - \bar{\mathbf{X}}_k\|^2 \hat{\mu}_{hk}^m$ is already minimized being $\bar{\mathbf{X}}_k$ the minimum for the unconstrained version of Problem (5.P2).
Therefore, the solution of Equation (5.11) is equivalent to the solution of the minimization of

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \hat{\mu}_{hk}^m \quad (5.14)$$

which can be also written as

$$\sum_{k=1}^{K} \|\bar{\mathbf{X}}_k - \mathbf{g}_k\mathbf{c}_k'\mathbf{B}_k\mathbf{V}_k'\|^2 \sum_{h=1}^{H} \hat{\mu}_{hk}^m \quad (5.15)$$

Given $\hat{\mu}_{hk}$, then, the minimization of (5.15) can be obtained by solving K separate independent minimization problems with respect to continuous variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k$, and discrete variables $\mathbf{V}_k$, under the constraints (5.C_3)-(5.C_6). This holds because for each k, $\sum_{h=1}^{H} \hat{\mu}_{hk}^m$ is a constant in the objective function.

The solution of problem (5.P3) can be found by using SQP. An alternative way to optimize (5.P3) is to solve it by using a coordinate descent algorithm where at each step the 2O-DFA is applied to each of the consensus matrices $\bar{\mathbf{X}}_k$ subject to the constraints (5.C_3)-(5.C_6).

(B) the partial minimization of the objective function of (5.P1) with respect to the fuzzy partition $[\mu_{hk}]$ when $\hat{\mathbf{g}}_k$, $\hat{\mathbf{c}}_k$, $\hat{\mathbf{B}}_k$, $\hat{\mathbf{V}}_k$ are given:

$$\begin{cases} \text{minimize} \sum_{k=1}^{K}\sum_{h=1}^{H} \|\mathbf{X}_h - \hat{\mathbf{g}}_k\hat{\mathbf{c}}_k'\hat{\mathbf{B}}_k\hat{\mathbf{V}}_k'\|^2 \mu_{hk}^m & (5.\text{P4}) \\ \text{s.t.} & \\ \sum_{k=1}^{K} \mu_{hk} = 1 \qquad \text{for } h = 1,\ldots,H & (5.C_1) \\ \mu_{hk} \in [0,1] \qquad \text{for } h = 1,\ldots,H,\ k = 1,\ldots,K & (5.C_2) \end{cases}$$

This sub-problem (B) can be minimized by considering the Lagrangian function

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \left\|\mathbf{X}_h - \hat{\mathbf{g}}_k\hat{\mathbf{c}}_k'\hat{\mathbf{B}}_k\hat{\mathbf{V}}_k'\right\|^2 \mu_{hk}^m - \sum_{h=1}^{H} \lambda_h \left(\sum_{k=1}^{K} \mu_{hk} - 1\right), \quad (5.16)$$

where the solution with respect to μ_{hk} is

$$\mu_{hk} = \frac{1}{\sum_{j=1}^{K} \left(d_{hk}/d_{hj}\right)^{\frac{2}{m-1}}}, \qquad \text{for}, h = 1,\ldots,H, k = 1,\ldots,K. \quad (5.17)$$

where $d_{lp} = tr[(\mathbf{X}_l - \hat{\mathbf{g}}_p\hat{\mathbf{c}}_p'\hat{\mathbf{B}}_p\hat{\mathbf{V}}_p')'(\mathbf{X}_l - \hat{\mathbf{g}}_p\hat{\mathbf{c}}_p'\hat{\mathbf{B}}_p\hat{\mathbf{V}}_p')]$.

Proof. To prove the solution stated in Equation 5.17, the Lagrangian function in Equation 5.16 has to be derived with respect to μ_{hk} and with respect to λ_h. Then, set the two derivatives equal to zero. In the following, for the sake of brevity and simplicity, we let d_{hk} denote $\left\|\mathbf{X}_h - \hat{\mathbf{g}}_k\hat{\mathbf{c}}_k'\hat{\mathbf{B}}_k\hat{\mathbf{V}}_k'\right\| = tr[(\mathbf{X}_h - \hat{\mathbf{g}}_k\hat{\mathbf{c}}_k'\hat{\mathbf{B}}_k\hat{\mathbf{V}}_k')'(\mathbf{X}_h - \hat{\mathbf{g}}_k\hat{\mathbf{c}}_k'\hat{\mathbf{B}}_k\hat{\mathbf{V}}_k')]$.
The first derivative with respect to μ_{hk} is

$$\frac{d}{d\mu_{hk}}\mathcal{L} = d_{hk}^2 m\mu_{hk}^{m-1} - \lambda_h \tag{5.18}$$

The first derivative with respect to λ_h is

$$\frac{d}{d\lambda_h}\mathcal{L} = \sum_{k=1}^{K}\mu_{hk} - 1 \tag{5.19}$$

By setting Equation 5.18 equal to zero, we obtain:

$$\begin{aligned} &\frac{d}{d\mu_{hk}}\mathcal{L} = d_{hk}^2 m\mu_{hk}^{m-1} - \lambda_h = 0 \iff \\ \iff \mu_{hk}^{m-1} = \frac{\lambda_h}{md_{hk}^2} &\iff \mu_{hk} = \left(\frac{\lambda_h}{md_{hk}^2}\right)^{\frac{1}{m-1}} \end{aligned} \tag{5.20}$$

By setting Equation 5.19 equal to zero, we obtain:

$$\frac{d}{d\lambda_h}\mathcal{L} = \sum_{k=1}^{K}\mu_{hk} - 1 = 0 \iff \sum_{k=1}^{K}\mu_{hk} = 1 \tag{5.21}$$

By inserting Equation 5.20 into Equation 5.21:

$$\begin{aligned} &\sum_{k'=1}^{K}\mu_{hk'} = \sum_{k'=1}^{K}\left(\frac{\lambda_h}{md_{hk'}^2}\right)^{\frac{1}{m-1}} = 1 \iff \\ \iff &\left(\frac{\lambda_h}{m}\right)^{\frac{1}{m-1}}\sum_{k'=1}^{K}\left(\frac{1}{d_{hk'}^2}\right)^{\frac{1}{m-1}} = 1 \iff \end{aligned} \tag{5.22}$$

$$\iff \left(\frac{\lambda_h}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{k'=1}^{K}\left(\frac{1}{d_{hk'}^2}\right)^{\frac{1}{m-1}}} \tag{5.23}$$

By inserting Equation 5.23 into Equation 5.20,

$$\mu_{hk} = \left(\frac{\lambda_h}{md_{hk}^2}\right)^{\frac{1}{m-1}} \iff \mu_{hk} = \frac{1}{\sum_{k'=1}^{K}\left(\frac{d_{hk}^2}{d_{hk'}^2}\right)^{\frac{1}{m-1}}} \iff \tag{5.24}$$

$$\iff \mu_{hk} = \frac{1}{\sum_{k'=1}^{K}\left(\frac{d_{hk}}{d_{hk'}}\right)^{\frac{2}{m-1}}} \tag{5.25}$$

Equation 5.24 proves the solution stated in Equation 5.17.

□

After the solution of the two sub-problems (A) and (B) the objective function generally reduces w.r.t. the previous iteration, or at least does not increase. Then, the algorithm stops to a stationary point which is not guaranteed to be the global minimum. This is due to the fact that a partitioning problem is included in (5.P1), which corresponds to an np-hard problem. For this reason, the algorithm is recommended to be run from several initial starting points to improve the chance of identifying the global optimal solution. The steps of the algorithm can now be formally presented.

ALGORITHM for (5.P1):

0. Initialization
   Set $t = 0$; $\epsilon > 0$ convergence constant; and randomly generate the membership degree matrix $[\mu_{hk}]$, with $k = 1, \ldots, K$, $h = 1, \ldots H$ from a uniform distribution and make it row-stochastic.

1. Do $t = t + 1$

2. Given $[\hat{\mu}_{hk}]$, solve sub-problem (A) with SQP algorithm or considering the following steps:

   (a) Compute $\bar{\mathbf{X}}_k$, for $k = 1, \ldots, K$ as follows:

$$\bar{\mathbf{X}}_k = \frac{1}{\sum_{h=1}^{H} \hat{\mu}_{hk}^m} \sum_{h=1}^{H} \hat{\mu}_{hk}^m \mathbf{X}_h \tag{5.26}$$

   (b) Solve problem (5.P3).

3. Given $\hat{\mathbf{g}}_k$, $\hat{\mathbf{c}}_k$, $\hat{\mathbf{B}}_k$, $\hat{\mathbf{V}}_k$, solve sub-problem (B)
   The solution of (5.P4) is given by:

$$\mu_{hk} = \frac{1}{\sum_{j=1}^{K} \left(d_{hk}/d_{hj}\right)^{\frac{2}{m-1}}}, \qquad \text{for}, h = 1, \ldots, H, k = 1, \ldots, K. \tag{5.27}$$

   where $d_{lp} = tr[(\mathbf{X}_l - \hat{\mathbf{g}}_p \hat{\mathbf{c}}_p' \hat{\mathbf{B}}_p \hat{\mathbf{V}}_p')'(\mathbf{X}_l - \hat{\mathbf{g}}_p \hat{\mathbf{c}}_p' \hat{\mathbf{B}}_p \hat{\mathbf{V}}_p')]$.

4. Stopping Rule
   Repeat steps 1-3 until the difference between the objective function at iteration t and the objective function at iteration $t - 1$ is greater than ϵ.

5.3.2 Remarks

It is worth mentioning that in order to allow the researcher to use either a confirmative approach or an explorative approach, explained in Section 5.2, in the 2O-DFA, it is possible to specify a constraints option, i.e. a vector of length equal to the number of manifest variables in the data set (J). When a confirmative approach is used, then the vector is a vector specifying which latent concepts define the corresponding manifest variable, while when an explorative approach is used, then the vector is left empty. Obviously, in a confirmative approach, the vector of constraints directly influence the structure of the matrix $\mathbf{V}_k$.

The main outputs of the algorithm implementing our proposed methodology are the membership degree matrix $[\mu_{hk}]$, the continuous variables $\mathbf{g}_k$, $\mathbf{c}_k, \mathbf{B}_k$, and discrete variables $\mathbf{V}_k$, which satisfy the constraints of model (5.3).

In particular, once the general composite indicators are obtained, then it is possible to rank units according to their scores and to discuss and make considerations on the results. However, frequently very small differences between scores of several units are observed, These differences cannot be considered statistically significant to decide if one unit comes before another. Thus, these units actually form equivalence classes and for these, the individual ranking cannot be realistically assessed. For this reason, instead of considering the complete ranking of all units, ignoring the statistical uncertainty, we find the ranking of the classes, i.e., the multivariate partial ranking defined in the "poset" literature Linear Ordered Partition (Stanley, 1997). In this way, we find C classes of units, by considering the small differences between scores as negligible. Given that we are seeking a hard partition of N units into C groups according to one variable and assuming that we are in a framework in which there exists a latent variable that defines the ranking of the units, then we can apply the Kmeans (Lloyd, 1982a; MacQueen, 1967a) algorithm on the general composite indicator in order to find the optimal partition. By letting C vary in an appropriate and reasonable interval, we apply the Kmeans algorithm and we compute the Fmax statistic (Caliński and Harabasz, 1974), which is proved to be a good measure for identifying the optimal C. Since in this case we are interested in having a large number of clusters to rank, we choose the highest C provided that the corresponding index value does not decrease more than 5% from the maximum value that the index takes in the range $\{1, \ldots, C\}$. In other words, we choose the number of clusters c at which the value of Fmax is between $[mFmax - 0.05 \cdot mFmax, mFmax + 0.05 \cdot mFmax]$, where $mFmax := \max_c Fmax(c), \forall\ c \in \{1, \ldots, C\}$.

Once the optimal partition of units into C cluster is obtained, then it is possible to represent clusters and their centroids as the values which characterize the clusters, as is usually done in the clustering problems. Then the final ranking between clusters of units can be provided. In this way, units belonging to the same cluster are not comparable in terms of their scores, but they are statistically different from units in the other clusters.

By summing up our methodology, we can state that given a three-way three-mode data array, the proposed methodology allows us to find a fuzzy partition of occasions into K classes, and within each class, identify a consensus matrix of dimensions $N \times J$ which serves as "centroid". In addition, simultaneously, from each of the K consensus matrices, it is possible to apply a hierarchical factor analysis model in order to find several factors (specific composite indicators) and a single general composite indicator. It is worthy recalling that the number of specific composite indicators can vary among consensus matrices. The proposed methodology, therefore, allows the reduction of the number of occasions H by identifying K groups of occasions and simultaneously the reduction of the number of variables J of each consensus matrix by, firstly finding Q_k factors and finally identifying 1 single general composite indicator.

Moreover, in order to better represent the results, an additional dimensionality reduction is considered. Indeed, once K general composite indicators (of dimension $N \times 1$) are obtained, then a linear ordered partition of the units is applied in order to find C groups of units and therefore obtaining K synthesized general composite indicators (of dimension $C \times 1$), given by the "centroids" of the groups.

5.4 Application to well being dataset

The proposed methodology is extremely useful when a set of multivariate observations of the same statistical units is provided for a given period of time. In this case, an interesting objective is to identify groups of years and simultaneously for each group detect a general composite indicator that is able to capture the differences between units in that specific period of time, by considering their ranking.

This is the case for example of the well-being dataset. In our analysis, How's Life- Well Being dataset (downloadable here), was considered following the OECD website that warns on the use of the classical Well Being Dataset (downloadable here).[1] The How's Life-Well Being (HLWB) dataset measures similar dimensions and variables to the ones of the classical Well-Being Dataset. It reports data for different years and for 38 OECD countries, which are members of the Organisation for Economic Cooperation and Development, (OECD), which includes most of the world's developed economies and several emerging economies. We consider the whole time period from 2005 to 2021.

However, several data processing steps were needed before implementing our methodology. First of all, there was the need to impute missing data and this has been done by using the k-Nearest Neighbors imputation (Troyanskaya et al., 2001) Secondly, data have been normalized by using the min-max normalization, where the min and the max are taken over the entire period (2005-2021). In addition, the variable Road Deaths was further manipulated so that the observation is obtained by subtracting the normalized value from the observed maximum. In this way, the variable Road Deaths was polarized correctly. Of course, the interpretation is reversed and the variable could be renamed "Difference about max Road Deaths". Finally, since our methodology requires a three-way array formed by the same units and variables over different years, some variables not present in all years have not been included. Therefore, the three-way array is formed by H =17 unit-by-variable matrices, with N =38 countries of OECD and referring to J =10 variables. The data were normalized and imputed as described above. Each variable is associated with a known dimension of Well-being as described by the OECD. More specifically, the following variables were considered: *Household income* (dimension: *Income and Wealth*), *Employment rate* (dimension: *Work and Job Quality*), *Gender wage gap* (dimension: *Work and Job Quality*), *Earnings* (dimension: *Work and Job Quality*), *Housing affordability* (dimension: *Housing*), *Households with internet access at home* (dimension: *Housing*), *Life expectancy at birth* (dimension: *Health*), *Perceived health* (dimension: *Health*), *Voter turnout* (dimension: *Civic Engagement*), *Road deaths* (dimension: *Safety*). In addition, OECD in its report clearly identifies two main larger dimensions: indeed, well-being can be considered in terms of *material living conditions* (housing, income, jobs) and *quality of life*(community, education, environment, governance, health, life satisfaction, safety, and work-life balance). In the remainder of the paper, the two dimensions Qualify of Life and Material Living Conditions will be referred to as QL and MLC, respectively. Merging all the information stated above, we can summarize the characteristics of our dataset in Tables 5.1 and 5.2.

[1] In fact, it is written 'Data cannot be compared between editions of the Better Life Index. For time series, please refer to the How's Life – Well-being database'. This means that the Better Life Index (classical Well-Being Dataset) can be only used to measure the *current* well-being condition; instead, the How's Life Well Being can be used to make comparisons *over time*.

Table 5.1. Information on three-way three-modes data: well-being data

Modes	Modes' Representation	Cardinality
Units	OECD Countries	N=38
Variables	Indicators of WB	J=10
Layers	Years	H=17

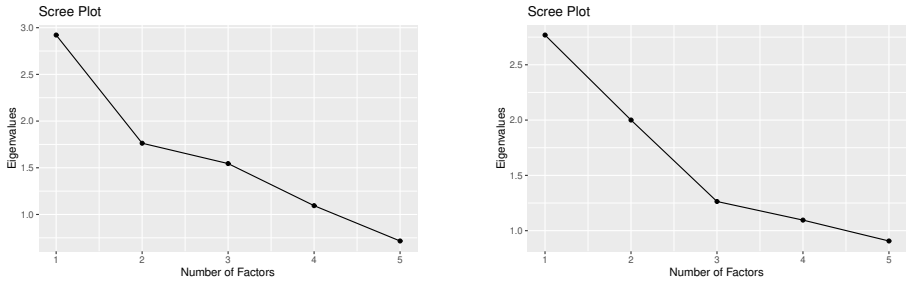
Table 5.2. Information on variables of the three-way three-modes data: well-being data

Variable	Type of indicator	Dimension	Dimension OECD
Household income	Average	Income and Wealth	Material living conditions
Employment rate	Average	Work and Job Quality	Material living conditions
Gender wage gap	Average	Work and Job Quality	Material living conditions
Earnings	Average	Work and Job Quality	Material living conditions
Housing affordability	Average	Housing	Material living conditions
Households with internet access at home	Average	Housing	Material living conditions
Life expectancy at birth	Average	Health	Quality of life
Perceived health	Average	Health	Quality of life
Voter turnout	Average	Civic Engagement	Quality of life
Difference about max Road Deaths	Average	Safety	Quality of life

5.4.1 Data Analysis

We applied our methodology to the aforementioned described dataset. The parameters K (number of clusters) and $Q_1, \ldots, Q_K$ (number of factors for each consensus matrix $\bar{\mathbf{X}}_k$, $\forall\ k = 1, \ldots, K$) are appropriately chosen.

The parameter K was a priori selected by considering internal validity indices, Fuzzy Silhouette (FS), Xie-Beni (XB), and Calinski-Harabasz (CH) indices (Campello and Hruschka, 2006a; Xie and Beni, 1991; Caliński and Harabasz, 1974). The algorithm has been run from 200 random starts to avoid local minima, by letting K vary in $[2, 6]$. For each iteration, and for each value of K: i) we computed the aforementioned indices, ii) we selected the number of clusters (K) that optimizes the indices; at the end of 200 iterations, we chose the value of K which was selected the most. According to FS and CH indices, the best K was set equal to 2. We intuitively consider this choice as sense-full also because, the period 2005-2021 can be divided at least into two parts. In the period 2005 - 2009, the year 2008 may be considered the initial year of the great recession that in 2009 continued to be observed globally in national economies. The effects of the crisis are evident in 2010 and 2011 when a period of instability is still observed: indeed, according to the OECD statement, 2010 was a year when the impacts of the financial crisis continued to be deeply felt in many OECD countries (source: OECD website). The year 2012 can be seen as the first year of the beginning of the recovery from the recession (source: Investopedia website) and therefore from 2012 a period of slow recovery with some stability is observed. In 2020 the infectious disease Covid-19 has spread all over the world in a few short months, but its effects on the national economies are not visible until 2021: indeed, as the World Bank website shows, the Covid-19 crisis had a major impact in 2021, a year that thus demonstrated that the pandemic has a far-reaching impact and touched every possible area of development. In addition, the OECD website reports that as the pandemic progressed, many people began to feel exhausted, especially in 2021. So this last crisis is not visible yet in the data that stop in 2021.
We will explore if the economic situation now described is recognized by the fuzzy clustering with some years hardly assigned to one of the two stability periods, and with some others more softly assigned to unstable periods.



(a) screeplot for $\bar{\mathbf{X}}_1$ **(b)** screeplot for $\bar{\mathbf{X}}_2$

Figure 5.2. Screeplot to choose the number of factors

Clearly, the choice of K influences the choice of different number of factors $Q_1, \ldots, Q_K$. Regarding the number of factors for each consensus matrix, we selected the best one according to the elbow method. In particular, for the first consensus matrix, Q_1 was set as equal to 2, while for the second the elbow was observed when Q_2 was set as equal to 3. The elbow plots can be analyzed in Figure 5.2.

However, it is important to mention that the 2O-DFA can be also run by imposing several constraints on the loading matrix. For this reason, we mainly implemented the methodology on the dataset by using two scenarios:

- By using the unconstrained version of the algorithm, we let Q_1 and Q_2 be data-driven chosen (explorative approach). Therefore, $Q_1 = 2, Q_2 = 3$, as discussed above.
- By using the constrained option of the algorithm, we set $Q_1 = Q_2 = 2$ and impose constraints on the variables' contributions to factors as identified by OECD (see last column in Table 5.2) (confirmative approach). In particular, for each consensus matrix $\bar{\mathbf{X}}_k$, the constraint vector $constr^k$ of dimension $(J \times 1)$ was used to indicate for each variable if the variable is constrained to be in a fixed class; the generic element of the vector is the number identifying the factor the variable contributes to defining. Formally, $constr^k_j = q,\ \forall\ j = 1, \ldots, J,\ \forall\ q = 1, \ldots, Q,\ \forall\ k = 1, \ldots, K$. When constraints are fixed, the 2O-DFA can be seen as a confirmatory second-order disjoint factor analysis.

Once K and $Q_k,\ \forall\ k = 1, \ldots, K$ are set, the algorithm is able to find the fuzzy partition of the H original data matrices, identify K consensus matrices, identify K matrices of correlations between variables and factors of dimension $J \times Q_k$ and the general composite indicator g_k, with the corresponding vector of coefficient $c_k,\ \forall\ k = 1, \ldots, K$.

The best solution (in terms of objective function minimization) over 200 runs was retained.

5.4.1.1 Membership Degree Matrix

The first output we analyze refers to the fuzzy membership degree matrix which is shown in Table 5.3.

We observe that our proposed approach is very appropriate for this application. As can be noticed, a fuzzy approach for clustering matrices allows these (in this application data matrices represent years) to belong either hardly or softly to clusters. In particular, we notice that the chronological order is retained and the whole period

Table 5.3. Membership degree matrix: well-being data

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Cluster 1	**0.858**	**0.870**	**0.870**	**0.888**	**0.769**	**0.561**	**0.388**	0.257	0.221	0.162	0.108	0.102	0.080	0.102	0.133	0.138	0.226
Cluster 2	0.142	0.130	0.130	0.112	0.231	**0.439**	**0.612**	**0.743**	**0.779**	**0.838**	**0.892**	**0.898**	**0.920**	**0.898**	**0.867**	**0.862**	**0.774**

Figure 5.3. Mosaic plot displaying on the x-axis the % membership degree to each cluster

is mainly split into two sub-periods (years: 2005-2009 and 2012-2021); the years 2005-2009 are almost hardly assigned to the first cluster, the years 2012-2021 are almost hardly assigned to the second cluster, while the two years exactly in the middle of the two clusters (2010 and 2011) are instead softly assigned to both of them. Indeed, we realized that the economic and social shock occurred and the transition from a sub-period to another one started in the year 2010 – a year when the impacts of the financial crisis continued to be deeply felt in many OECD countries (see OECD website).
The membership degree can be better visualized when a mosaic plot is used. In Figure 5.3, we observe for a given year a stacked barplot characterized by two colors, one for each cluster and where the length of each bar is given by the membership degree (in %) of the corresponding cluster. As can be seen, from 2005 to 2009 the bar is almost completely green, meaning high membership for cluster one, while from years 2012 to 2021 almost completely yellow, meaning high membership for cluster two. Finally, the years 2010-2011 which correspond to a transitional period, have yellow and green bars of similar length, meaning similar membership for both clusters.

Hereafter, to better interpret and visualize the results of the applications of 2O-DFA and Multivariate Clusters Ranking, the whole period 2005 – 2021 was divided into two sub-periods, the first from 2005 to 2010, the second from 2011 to 2021. The subdivision was implemented using the maximum a posteriori (MAP) technique applied to the membership degree matrix in Table 5.3. The MAP is a tool to obtain the hard counterpart of a fuzzy partition by hardly assigning each unit to the cluster whose membership degree is highest.
Therefore, in the following, results are presented and commented on separately by sub-period. It is worth remembering that the consensus matrices are obtained as Least-Squares approximation of the original data matrices. This corresponds to computing the weighted average of matrices, where weights are given by the degrees

to which each original data matrix belongs to the cluster.
For this reason, the years 2005-2010 (hardly belonging to the first cluster) are the main contributors to the first consensus matrix, while the years 2011-2021 (hardly belonging to the second cluster) are the main contributors to the second consensus matrix as can be seen from Table 3.

5.4.1.2 First order DFA

The results of applying 2O-DFA consist of Q_k first-order factors and a single second-order factor, or general composite indicator, for each $k = 1, \ldots, K$. In this section, we comment on the first-order DFA.

It is worthwhile to analyze and discuss the matrices of correlation values between each manifest variable and each indicator (first-order factor) obtained by applying the proposed algorithm to the data set in both a constrained and unconstrained scenario. The matrix of correlations between variables and factors, considering the two consensus matrices separately, is shown in Table 5.4. Note that the resulting matrices are disjoint: in fact, we applied 2O-DFA, which causes each variable to be associated with only one factor. Also, in Table 5.4, the factors are sorted from left to right according to the descending order of their explained variance; the name of the columns and their colors help the reader to understand, compare and grasp the similarities and dissimilarities between the dimensions identified in the two scenarios.

First, we note that once the constraints on the $\mathbf{V}$ matrix are given, the algorithm calculates the covariances (and thus the correlations) between the variables and the factors. We observe that for both consensus matrices, the variables are highly and positively correlated with the corresponding factor, with the only exceptions being *Gender wage gap* and *Housing affordability*. This is an indication that the current location of these variables may not seem optimal and/or that these two variables contribute little to the definition of material living conditions as defined by the OECD. This result is confirmed passing from the period 2005-2010 to 2011-2021 where the correlation tends to nullify for these two variables. Thus, excluding *Gender wage gap* and *Housing affordability*, correlations confirm the classification provided by the OECD and shown in table 5.2.

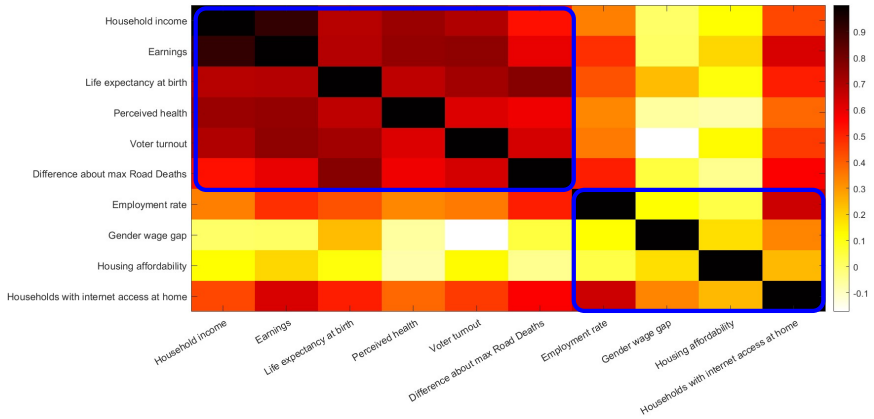
In the unconstrained scenario, in which we let the algorithm choose the number of factors in a data-driven manner, two factors are considered for the first consensus matrix and three for the second. The number of factors is chosen by the elbow method (Figure 5.2), and from the heatmaps showing the correlations between the variables in each consensus matrix (see Figure 5.4), these two parameters seem to be a good choice, consistent with the structure of the data: in fact, two and three blocks of variables are clearly visible.

Focusing on the first consensus matrix (years 2005-2010), the factors do not completely correspond to those provided by the OECD (constrained version), although some similarities can be noted: in fact, the QL dimension is the same as in the constrained scenario, but two additional variables, namely *Household income* and *Earnings*, also contribute to its definition. Moreover, the two variables whose correlations with the MLC dimension in the constrained scenario were very low (close to 0.3), namely *Gender wage gap* and *Housing affordability*, in the unconstrained scenario are highly and positively correlated with the MLC dimension. Thus, the empirical evidence does not confirm that these two variables are coherent with the QL. Shifting *Household income* and *Earnings* from MLC to QL produces from the empirical evidence a more consistent measure of QL with respect to the one defined theoretically by OECD with an increase of Cronbach's alpha from 0.018 (near-zero consistency) to 0.694, at the expense of a slight reduction of the Cronbach's alpha

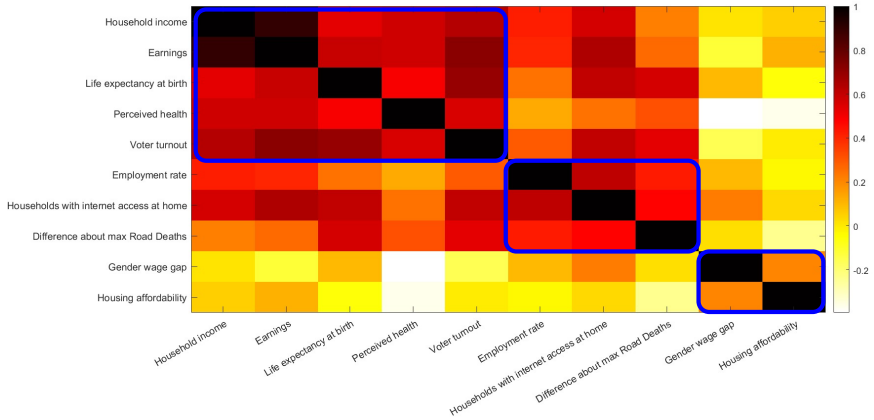
Table 5.4. Disjointed matrices of correlations between the variables and the factors under two scenarios: constrained and unconstrained (well-being dataset).

	first consensus: 2005-2010			
	unconstrained		constrained	
Variable	QL	MLC	QL	MLC
Household income	**0.889**	0.000	0.000	**0.810**
Employment rate	0.000	**0.773**	0.000	**0.709**
Gender wage gap	0.000	**0.542**	0.000	**0.234**
Earnings	**0.915**	0.000	0.000	**0.909**
Housing affordability	0.000	**0.401**	0.000	**0.298**
Households with internet access at home	0.000	**0.902**	0.000	**0.839**
Life expectancy at birth	**0.877**	0.000	**0.917**	0.000
Perceived health	**0.844**	0.000	**0.826**	0.000
Voter turnout	**0.855**	0.000	**0.858**	0.000
Difference about max Road Deaths	**0.790**	0.000	**0.864**	0.000
Cronbach's alpha	0.694	0.586	0.018	0.732
Explained variance (%)	44.646	18.658	30.066	28.329
Total explained variance (%)	63.304		58.395	

	second consensus: 2011-2021				
	unconstrained			constrained	
Variable	QL	MLC1	MLC2	MLC	QL
Household income	**0.878**	0.000	0.000	**0.880**	0.000
Employment rate	0.000	**0.839**	0.000	**0.69**8	0.000
Gender wage gap	0.000	0.000	**0.777**	**0.100**	0.000
Earnings	**0.914**	0.000	0.000	**0.897**	0.000
Housing affordability	0.000	0.000	**0.777**	**0.096**	0.000
Households with internet access at home	0.000	**0.859**	0.000	**0.848**	0.000
Life expectancy at birth	**0.784**	0.000	0.000	0.000	**0.868**
Perceived health	**0.751**	0.000	0.000	0.000	**0.720**
Voter turnout	**0.861**	0.000	0.000	0.000	**0.878**
Difference about max Road Deaths	0.000	**0.755**	0.000	0.000	**0.744**
Cronbach's alpha	0.894	0.357	0.342	0.695	0.187
Explained variance (%)	35.252	20.118	12.063	28.048	25.96
Total explained variance (%)		67.433		54.008	



(a) First consensus matrix



(b) second consensus matrix

Figure 5.4. Correlation matrices of the consensus matrices

from 0.732 (MLC, according to OECD definition) to 0.586.

Focusing on the second consensus matrix (years 2011-2021), we note that *Household income* and *Earnings* still remain, together with the other variables, the main contributors to the QL dimension. Also in this sub-period, this variables' aggregation produces from the empirical evidence a more consistent measure of QL with respect to the one defined theoretically by OECD with an increase of Cronbach's alpha from 0.187 (low consistency) to 0.894 (high consistency). Thus the perception of QL remains the same in the two periods with the only exception of the variable *Difference about max Road Deaths*. In addition, we notice that in the second sub-period the variables related to the MLC dimension according to the OECD classification, in the unconstrained scenario, are divided into two subsets. So in this period after the crisis, the shock has produced a different perception of the MLC. Variables *Employment rate*, *Households with internet access at home* and *Difference about max Road Deaths* are the first specific dimension of the MLC (referred to as MLC1). Variables *Gender wage gap* and *Housing affordability* form the second specific dimension of MLC (named MLC2). After the crisis, there was major attention to Employment and Households with the internet from one side and the Gender wage gap together with Housing affordability on the other side. Furthermore, we note that in the constrained

scenario *Gender wage gap* and *Housing affordability* are very poorly correlated, with a correlation coefficient close to zero, to the MLC dimension: therefore, in the second sub-period (years 2011-2021) the OECD classification of dimensions leads to deleting the above variables in terms of their contribution to the MLC dimension. In contrast, in the unconstrained scenario, this no longer the case: in fact, the two variables are highly correlated with the associated factor and alone define the latent dimension (MLC2). Therefore, in this application the unconstrained scenario allows for a better explanation and identification of the contributions of the variables to the factors.

Focusing on the objective function, the 2O-DFA uses the explained variance as a measure of the goodness of fit of the model. Analysis of the model fit statistics shows that the fit is greater in the unconstrained scenario than in the constrained one. In this sense, we can conclude that using the factors identified by a data-driven approach resulted in a solution that suggests the strengths and weaknesses of the OECD characterization of Well-being and allows the researchers to understand how to modify the definitions that do not seem sustained in the years by empirical pieces of evidence. The strength of this methodology is to define factors that can be seen as the consensus of the factors in sub-periods of years.

Finally, it is worth commenting on the correlations between the factors obtained in this first-order DFA, shown in the Table 5.5.

(a) First consensus matrix

	Fact1	Fact2
Fact1	1.00	0.51
Fact2	0.51	1.00

(b) Second consensus matrix

	Fact1	Fact2	Fact3
Fact1	1.00	0.73	-0.41
Fact2	0.73	1.00	-0.28
Fact3	-0.41	-0.28	1.00

Table 5.5. Correlations between factors resulted from first-order DFA applied on the two consensus matrices

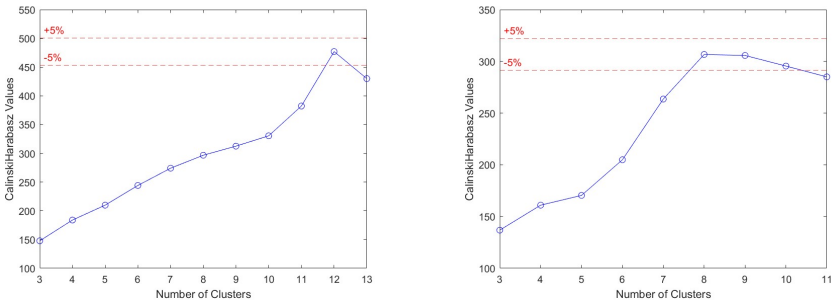
Looking at Table 5.5, we note that the factors identified by applying first-order DFA to both consensus matrices are highly correlated. In the context of higher-order factor analysis, this result clearly demonstrates that a second-order DFA is needed so that a single factor (general composite indicator) explains this residual part of correlation and represents an overall synthesis of the original J variables. The $K = 2$ composite indicators, one for each consensus matrix, are analyzed and discussed in the next section. The product of the correlation coefficients of the first and second order defines the contribution of each original variable to the general composite indicator.

5.4.1.3 General Composite Indicators

Once the first-order dimensions have been found, the second-order DFA makes it possible to identify $K = 2$ general composite indicators, one for each consensus matrix $\bar{\mathbf{X}}_k$, for $k = 1, 2$. Since better results are obtained in the unconstrained scenario, the general composite indicators we analyze here are those obtained in this scenario.

For each consensus, a ranking of the $N = 38$ OECD countries can be obtained. In fact, the general composite indicator provides the score for each unit and can be considered as the *well-being indicator*, measuring a nation's progress in terms of health, wealth, and personal well-being.

However, it is necessary to remember that the ranking of countries based on the scores of the overall composite indicator cannot be used directly because very small differences are observed between the scores of different units, so a meaningful and unambiguous interpretation cannot be reached. Using the complete ranking of all units corresponds to ignoring the statistical uncertainty around the units. For example, the scores of Finland and Korea are 0.239 and 0.235 for the first period 2005-2010, so they are basically incomparable, meaning that Finland cannot be said to come before Korea. Similar situations are observed for the Slovak Republic and Hungary (-0.602 and -0.606), Australia and Denmark (0.316 and 0.301), the United States, and Luxembourg (0.382 and 0.377), which have almost the same score in the first period. This situation is also repeated for the ranking of the second period 2011-2021, where, for example, Portugal and the Slovak Republic have exactly the same score (-0.338), France and Finland have very similar scores (0.144 and 0.142) so as Sweden and Australia (0.316 and 0.315, respectively). Therefore, we apply the specialized K-Means for unidimensional variables to the scores of the general composite indicator to find C homogeneous classes of countries. In this way, countries in the same cluster are considered incomparable and define equivalence classes in terms of ranking, while counties in one cluster differ from those in other clusters. The choice of C should not be parsimonious because we want to have a granular ranking of clusters. This is explained in detail in Section 5.3.2. We let C be chosen from the Fmax statistics with additional flexibility shown in Figure 5.5 and $C = 12$ and $C = 10$ were considered, respectively.



(a) First consensus matrix **(b)** Second consensus matrix

Figure 5.5. Fmax values of resulting partition by applying K-Means on the general composite indicator scores for both the consensus matrices as c varies (well-being dataset)

The final solution, obtained when $C = 12$ in the first case and $C = 10$ in the second, leads to the partition of countries shown in Table 5.6, where the values of the cluster centroids are also given. The analysis was conducted using the R package `Ckmeans.1d.dp` (Song, Wang, and Song, 2022), and more specifically the `Ckmeans.1d.dp` function, which implements the Ckmeans.1d.dp algorithm to cluster univariate data given by a numeric vector into C groups by dynamic programming (Wang and Song, 2011; Song and Zhong, 2020) and guarantees the optimality of clustering, as the total of within-cluster sums of squares is always the minimum given the number of clusters C.

Looking at the obtained partitions, it can be seen that several countries have improved their well-being indicator over time: this is the case, for example, of Luxembourg, Canada, Germany, New Zealand, Spain, Colombia, and Turkey. However, comparing the two clustering partitions, corresponding to the two sub-periods, makes

Table 5.6. Partition of countries in $C = 12$ and $C = 10$ clusters and related centroids, when K-Means is applied to the composite indicator scores associated to the first and to the second consensus, respectively (well-being dataset)

(a) First consensus matrix

cluster	centroid	countries
1	0.609	Iceland, Switzerland
2	0.408	Netherlands, Norway, United States, Luxembourg
3	0.323	Canada, Sweden, Australia, Denmark
4	0.242	Austria, Japan, United Kingdom, Finland, Korea, Belgium
5	0.148	Germany, Israel, France, Ireland
6	0.036	New Zealand, Spain
7	-0.107	Italy, Chile, Slovenia
8	-0.201	Costa Rica, Portugal, Greece
9	-0.335	Czech Republic, Estonia
10	-0.492	Colombia, Poland, Lithuania
11	-0.613	Latvia, Slovak Republic, Hungary, Mexico
12	-0.744	Türkiye

(b) Second consensus matrix

cluster	centroid	countries
1	0.505	Switzerland, Luxembourg, Iceland
2	0.327	Norway, Netherlands, Denmark, Sweden, Australia, Canada, United States, New Zealand
3	0.219	Belgium, Austria, Germany, United Kingdom, Ireland
4	0.141	France, Finland, Israel
5	0.043	Japan, Spain, Italy
6	-0.072	Chile, Slovenia
7	-0.234	Colombia, Czech Republic, Costa Rica, Greece
8	-0.371	Portugal, Slovak Republic, Korea, Poland
9	-0.470	Estonia, Türkiye, Hungary, Lithuania
10	-0.667	Latvia, Mexico

it possible to show that some other countries have worsened their well-being index: this is the case, for example, of Mexico, Latvia, Estonia, Portugal, Japan, and Korea. For the sake of clarity and interpretation of the results, among the aforementioned countries, we analyze New Zealand and Korea in particular: the former is characterized by a well-being index close to zero, signifying an average behaviour in terms of well-being policies, in the first sub-period; in the second sub-period, on the other hand, it scores similar to the countries in the second position, meaning that over time it has implemented good policies, not only social but also economic, in order to increase the well-being of its citizenry. On the contrary, Korea scored positive and relatively high in the first sub-period, while in the second sub-period, it ranked substantially worse, below the mean of the countries, meaning that its policies have been able to reduce the well-being index.

As additional note, it has to be considered that Iceland, Switzerland, the Netherlands, Norway, Luxembourg and the United States are always in the top two clusters, which means that their scores have always been at the top in the two sub-periods; in other words, in terms of comparison with other countries over the period under review, they have not changed their policies either to improve or to worsen the state of wealth, health, personal and social well-being of their citizens. In particular, Switzerland and Iceland are always in the top group, which means that over the entire period, they were able to provide good services. Indeed, by examining the original variables related to health, wealth, and personal and social well-being, we can see that the values of these variables for the countries mentioned above are consistently at the top of their rankings throughout the period. In fact, looking at Table 5.7, we can see that ranking Iceland and Switzerland according to these variables separately for each year results in ranking positions that are often among the top.

It has to be noted that the partition into $C = 12$ clusters of countries in the first consensus matrix leads to one singleton clusters, namely Turkey (cluster 12); the identification of singleton cluster is strictly related to the country behavior in terms of the well-being index: in fact, this country is standing alone and behaving differently from other countries during the period under consideration, having a score that is significantly different from that of other countries, a score that deviates

Table 5.7. Ranking position of Iceland and Switzerland based on the normalized values of Earning and Employment Rate reported at years 2005, 2006, 2007, 2017, 2018, 2019

(a) Earnings

country	2005	2006	2007	2017	2018	2019
Iceland	3	2	2	2	1	2
Switzerland	2	3	3	4	4	4

(b) Employment Rate

country	2005	2006	2007	2017	2018	2019
Iceland	1	1	1	1	1	2
Switzerland	2	2	4	3	3	3

significantly from the nearest centroid and is itself a centroid.

To summarize the results, countries whose centroid values of the overall composite indicator are high and positive correspond to countries with a high well-being index; countries whose centroid values of the overall composite indicator are medium-high and positive are those countries characterized by a better-than-average well-being index; finally, countries whose centroid values are negative and low correspond to countries that are always below the average level of the well-being index.

It is possible to appreciate the differences in well-being indicator scores by looking at the maps shown in Figure 5.6.

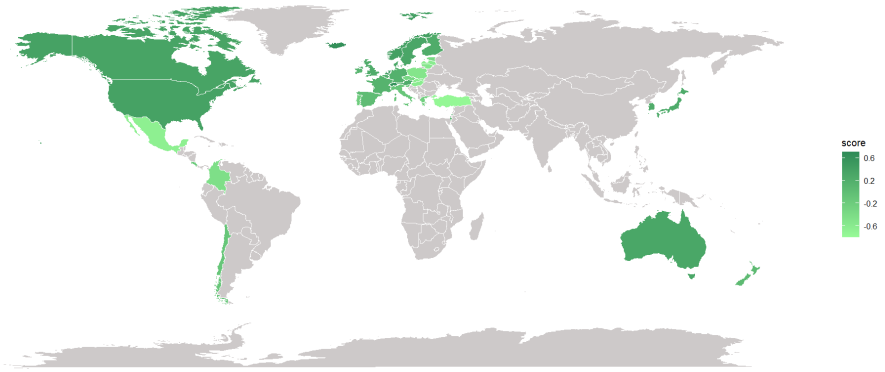
The analysis of the rankings and, more specifically, analysis of the identified groups of countries allowed us to compare countries' policies and citizens' perceptions of personal and social comfort, health, wealth, and well-being.

In conclusion, our methodology allowed us to start from a complex data structure, i.e., a set of H data matrices $(N \times J)$ and obtain, at the end of the process, K composite indicators of N units, where $K = 2$ are the clusters of years; most years are hardly assigned to the corresponding cluster, while two years exactly in the middle of the two periods (clusters), corresponding to transition years, were correctly identified by the fuzzy membership. In addition, a partitioning algorithm was applied to the two composite indicators to obtain C classes of countries in both sub-periods, where the ranking can be considered more realistically and hence the final groups of countries are more easily interpreted.

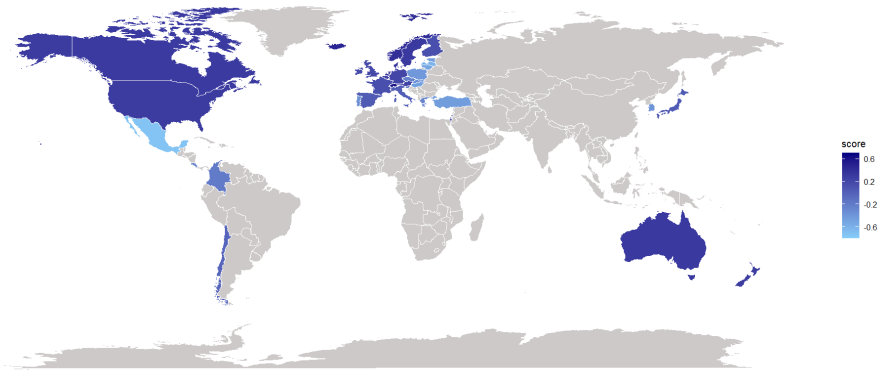
5.5 Conclusion

Given a three-way data matrix characterized by three modes, N units, J variables, and H occasions, a new methodology is introduced that allows the fuzzy partitioning of occasions into K clusters and identification of K consensus data matrices. At the same time, the covariance structures of each of the K consensuses are summarized by the corresponding disjoint second-order factor analysis. In more detail, for each consensus data matrix, we obtain a) a set of first-order factors with the corresponding loadings matrix and b) a second-order factor analysis, i.e., a general composite indicator, and Q specific composite indicators.
In addition, the scores of each general composite indicator can be ranked, and such ranking allows for the evaluation of differences between units. However, it is suggested to apply a partitioning algorithm to obtain C equivalence classes of units, each of which groups countries whose difference between scores cannot be appreciated. In this way, it is possible to appreciate the differences between classes and produce their ranking.

It should be emphasized that the fuzzy clustering of occasions and the second-order factor analysis are computed simultaneously in the three-way data matrix to



(a) First consensus matrix



(b) Second consensus matrix

Figure 5.6. Composite indicator scores associated to the first and to the second consensus, respectively (well-being dataset)

avoid the sequential estimation of clustering and factorial methods that, as it is well-known, could produce masking effects of the overall optimal solution.

The new methodology is estimated by a least-squares coordinate descent method by using an efficient algorithm.

The proposed methodology is extremely useful when one has a set of multivariate observations of the same statistical units for a specific time period. In this case, the goal is to identify year groups and, simultaneously, for each group, detect an overall composite indicator that is able to capture the differences between the units over those specific years, considering their compote indicators and corresponding ranking.

Within this framework, the proposed methodology was applied to the well-being dataset. The How's Life-Well Being (HLWB) dataset measures several variables and reports data for several years and 38 OECD countries. We consider the entire time period from 2005 to 2021. Therefore, the three-way matrix consists of $H = 17$ unit-by-variable matrices, with $N = 38$ OECD countries monitored on $J = 10$ variables considered proxies of the well-being.

The fuzzy approach to clustering is very appropriate for this application. In fact, the fuzzy partitioning of years can be easily interpreted and the results are realistic and meaningful: first, the chronological order is maintained; second, the whole period is mainly divided into two sub-periods: most years are hardly assigned to clusters, while the two years (2010 and 2011) exactly in the middle of the two clusters are instead softly assigned to both. In fact, the transition from the first sub-period to the second began in 2010, a year in which the impacts of the financial crisis were deeply felt in many OECD countries. Therefore, the soft allocation of these two years is significant and fully consistent with their transitional characterization.

In addition, with the second-order DFA it was possible to identify $K = 2$ overall composite indicators and to obtain a ranking of $N = 38$ OECD countries. For the reasons mentioned above, we find a multivariate partial ranking by applying a K-Means to the scores of general composite indicators to find C groups of countries. The analysis of the identified countries' groups allowed for a reasonable comparison of countries' policies and people's perceptions of health, wealth, and personal so as social personal comfort.

The proposed methodology is a powerful tool for multivariate data analysis because it includes several tasks, simultaneously solved, of clustering, dimensional reduction, and ranking, typically sequentially computed in many applications with an understandable reduction of optimality of the final solution. Specifically, given a three-way data matrix, the new methodology allows for the simultaneous fuzzy partitioning of occasions into K clusters and ranking of units within each cluster according to general composite indicator scores obtained by disjoint second-order factor analysis applied to each consensus matrix. There are many possible applications of the methodology in different areas: for example, in economic studies, it can be used to obtain a fuzzy partition of years and simultaneously rank countries in each sub-periods according to their economic behavior.

The methodology opens up the possibilities of further developments, according to different data analysis aims.

Chapter 6

Representing ensembles of networks for fuzzy cluster analysis

6.1 Introduction

Networks represent a powerful model for problems in different scientific and technological fields, such as neuroscience (Simpson, Hayasaka, and Laurienti, 2011; Obando and de Vico Fallani, 2017), molecular biology (Grazioli, Martin, and Butts, 2019), biomedicine (Yang et al., 2014; Granata et al., 2020a), sociology (Heckerman, 1997; Jiang et al., 2014; Slaughter and Koehly, 2016), social network analysis (Tang and Liu, 2011; Tagarelli, Amelio, and Gullo, 2017) and political science (Moody and Mucha, 2013). As the number of network applications increases, so does the need for novel data analysis techniques, particularly for fuzzy cluster analysis.

A well-known approach to the clustering problem on a network is the detection of clusters of nodes (or *communities*). This task is widely explored in literature, and many research works focus on applying both hard and fuzzy clustering algorithms to a network to detect the underlying community structure. Among the studies utilising hard clustering algorithms to detect node clusters, we recall Asur, Ucar, and Parthasarathy (2007) who proposed a methodology to be applied on protein network where they applied standard conventional graph clustering techniques to find communities of nodes inside a weighted network and combined the clustering results to get a consensus clustering. A systematic investigation of consensus clustering for detection of community structures in complex networks was performed by Lancichinetti and Fortunato (2012). A general overview of the existing methodologies used to summarize clustering results is given by Ghosh and Acharya (2011), who provided a detailed review of cluster ensembles. Ou-Yang, Yan, and Zhang (2017) proposed a methodology to obtain communities of nodes or protein complexes by using information taken from multiple heterogenous protein networks.
Instead, among the research works that have focused on applying fuzzy clustering algorithms to detect clusters of nodes, one example is the study conducted by Havens et al. (2013), in which they employed the Fuzzy k-Means (FkM) algorithm to identify fuzzy communities within social networks. A different proposal is addressed by Runkler and Ravindra (2015) who apply the Non-Euclidean Relational Fuzzy k-Means (NERFkM) algorithm to the dissimilarity matrices obtained using three algorithms for crisp graph clustering: the Newman-Girvan, the Small World, and the Signal algorithms. Bhatia and Rani (2017) use a Parallel Fuzzy Clustering Algorithm

for handling scalable graph data and Zaidi (2012) focuses on a network structure to find clusters of web pages according to related keywords. In particular, the author uses the Fuzzy Agglomerative Hierarchical clustering algorithm and Hierarchical Hyperspherical Divisive Fuzzy k-Means.

All the articles mentioned above focus on considering a single node as the statistical unit. On the contrary, we consider a single network as the unit of interest. Thanks to its structure, a network provides a more detailed representation of the problem yet introduces new complexity. In this context, with the fast-growing availability of data and ensembles of networks, several research areas can focus on clustering networks. In biology, networks can represent, for example, tumour metabolism: the nodes are the metabolites, and edges connect pairs of nodes involved in the same reaction. In this case, the interest is to characterise groups of patients based on the tumor type (Manipur et al., 2020a). In air passenger transport, each graph represents an airlines company, nodes are airports and edges flight routes. The aim is to study how the airline companies can be grouped (Carpi et al., 2019; Tantardini et al., 2019), accordingly to the structure of their routes. In the field of commerce and trade, products can be represented by networks, whose nodes are countries and edges are the export/import trades, and the interest is to group products with similar trade behaviour (Tantardini et al., 2019). In political science, Yin, Shen, and Butts (2022) propose a model-based clustering to identify groups of networks: in their case study, networks represent co-voting patterns, nodes represent Senators, and edges link Senators that vote concurrently. In a sociological study, Brandes, Lerner, and Nagel (2011) apply a clustering technique to detect clusters of networks of migrants in Spain and USA and for each cluster of networks determine its role structure: in particular, they extract a feature vector from each network and apply standard clustering methods (e.g. k-Means). In medicine, Duroux and Van Steen (2023) identify groups of networks by computing similarity between networks via appropriate distance measures; they apply the methodology to nitroaromatic compound networks and brain networks. Finally, several applications in different areas are provided by Ni et al. (2017), whose aim is to find groups of networks while detecting common clusters within each network group.

To the best of our knowledge, this is the first work focusing on identifying how to represent ensembles of networks for fuzzy cluster analysis. Indeed, related works, such as the aforementioned ones, use a single hard approach to clustering: this means that each network can belong to one cluster only. However, depending on their representation, networks may have characteristics in common to more than one cluster, and therefore in such situations, a more flexible approach is more adequate. In this sense, the fuzzy approach guarantees major flexibility than the hard approach, by allowing each network to belong to all clusters according to different membership degrees.

It is important to emphasise the distinctive nature of our work in comparison to previous proposals regarding ensemble network clustering. While prior approaches either concentrate on applying fuzzy and hard node clustering methods to detect clusters of nodes or focus solely on using hard clustering techniques to ensemble of networks to detect clusters of networks, our proposal distinguishes itself by employing a fuzzy clustering approach to identify clusters of networks.

To cluster networks, we need to find an adequate representation. In the early proposals on this topic, networks have been represented using some topological characteristics, such as density, the average number of nodes per edge, centrality indexes, to name a few. The pros of such a choice are that even if the networks do not share a common set of nodes, it is still possible to obtain a representation in some vector space and apply standard clustering techniques. Moreover, networks of

different sizes find a representation in the same vector space. The cons rely mainly on the fact that very different networks might be represented by the same values of the chosen features, making the data analysis difficult or impossible.

Another option would be to represent the network using its *adjacency matrix*, a $0, 1$ matrix, whose dimension is the number of nodes, encoding the presence of a connection between nodes i and i' with a 1 in position i, i'. This matrix representation would make it possible to use matrix norms to induce a distance. The limit of this solution is that a matrix norm cannot account for differences in specific parts of the network and therefore ignores its topological characteristics.

To overcome these limits, we study two types of network representations: a probabilistic representation of graphs where the Jensen-Shannon Divergence is used to compute pairwise distances and a whole-graph embedding representation. The embedding techniques provide a vector space representation of the networks to identify a space that is optimal with respect to some characteristics.

Once we have chosen how to adequately represent the networks, it is possible to apply fuzzy clustering algorithms. More in detail, we applied Non-Euclidean Fuzzy Relational Clustering, introduced by Davé and Sen (2002), and the Fuzzy Analysis clustering, introduced by Rousseeuw and Kaufman (1990), when the networks are represented by a matrix of distances; instead, we applied the Fuzzy k-Means (Bezdek, 1981), the Fuzzy k-Means with polynomial fuzzifier (Klawonn and Höppner, 2003), the Fuzzy k-Means based on L1 metric (Jajuga, 1991) and the Fuzzy k-Medoids (Krishnapuram et al., 2001), when they are in form of a feature matrix. We empirically compare the strategies, highlighting their possible uses in different scenarios. Finally, we analyze the performances of the algorithms applied to the proposed networks representations of two ensemble of networks: the former is the European Air Transportation Network (Cardillo et al., 2013), the latter is obtained from the FAOSTAT, the Food and Agriculture Organization of United Nations database.

The chapter is organised as follows: Section 6.2 gives a brief introduction to graphs (in this context synonymous of networks); Section 6.3 explains the methodology and details the algorithms used in our analysis; Section 6.4 focuses on the applications on simulated and real datasets: both the networks descriptions and the main clustering results are provided. Finally, Section 6.5 contains concluding remarks and future work.

6.2 Definitions

6.2.1 Graphs

A graph or a network is a mathematical entity representing connections or relationships between pairs of objects, or more in general, between several objects. A graph $\mathcal{G} = (V, E)$ is an ordered tuple of 2 sets: $V = \{v_1, \ldots, v_N\}$ is the set of its N unique nodes, and $E = \{e_1, \ldots, e_M\}$ is the set of its M edges. The set E is a subset of $V \times V$, $E \subseteq V \times V$, the set of all possible edges. In the case of *undirected* graphs, each edge is an (un)ordered node pair, $e_l = (v_i, v_{i'})\ \forall\ l = 1, \ldots, M$, and $\forall\ i, i' = 1, \ldots, N$, not necessarily $i \neq i'$, as it is possible to have self loops. If the edges have an orientation, we call the graph oriented (*directed*). In non-oriented graphs, the relation represented by the edge between a pair of objects is symmetric. In contrast, in the case of directed graphs, the orientation of the edges (that are represented by arrows) points out a two-level relation. Aside from orientation, edges might be *weighted* or *unweighted*; the weights, representing the strength of the connection between two nodes, are usually shown as numbers just above the

corresponding edges. Focusing on nodes, they usually identify units inside the network. Therefore, they are unique, and their labels are the key identifiers. The label can either be a number, a letter, or a word.
Finally, in a graph $\mathcal{G} = (V, E)$, we define a path (walk) from vertex v_i to vertex $v_{i'}$ as a sequence of edges $\{e_0, e_1, \ldots, e_z\}$ joining a sequence of vertices, where v_i is the starting node, and $v_{i'}$ is the final node. If there exists a path from node v_i to node $v_{i'}$ we will say that $v_{i'}$ is *reachable* from v_i. In an unweighted network, the length of a path is equal to the number of its edges, In a weighted one, the length of a path is given by the sum of weights of its edges. We refer to the *shortest* (geodesic) *path* as the path connecting two nodes with the shortest length. The length of the longest shortest path is called the *diameter* of a network.

6.2.2 Network representations

We adapted diverse network representations to assess the fuzzy clustering algorithms. These representations are based on probability distributions of topological network properties and whole-graph embeddings.

Networks are often represented as probability distributions of their topological features. In this study, we use two such distributions, (Carpi et al., 2019; Granata et al., 2018; Granata et al., 2020b), describing global and local network properties, the Node Distance Distribution and the Transition Matrix, which will be explained in the following subsections.

As additional networks' representation, it is also possible to use the whole-graph embedding representation. The embedding techniques provide a vector space representation of the networks to identify a space that is optimal with respect to some characteristics. Particularly, the Joint Emdedding Technique (Wang et al., 2021) and the Denoising Autoencoder (Gutiérrez-Gómez and Delvenne, 2019) are described in the following subsections.

6.2.2.1 Node Distance Distribution

The Node Distance Distribution (NDD) summarizes the graph by using a row-stochastic matrix of dimension $N \times d$, where N is the cardinality of the set V and d is the diameter of the graph, i.e. the length of the longest shortest path existing in the graph.

More in details, the NDD of node i, denoted as $N_i(l)$, is defined as the fraction of nodes reachable with a shortest path of length l from node i. The matrix describing the NDD of a graph $\mathcal{G}$ is obtained by computing $\mathcal{N}_i(l)$ for each node $i = 1, \ldots, N$ and for each length $l \in L$, where L is the set of lengths of all the shortest paths inside the graph $\mathcal{G}$. More formally, the NDD of graph $\mathcal{G}$ is defined as $N_1, \ldots, N_N$, where $N_i = [N_i(1), \ldots, N_i(d)]$, $\forall\ i = 1, \ldots, N$ and contains information about the global topology of graph $\mathcal{G}$.
Operationally, in order to represent a graph $\mathcal{G}$ by its NDD the following steps are needed:

1. Compute a matrix $\mathbf{M_1}$ of dimension $N \times N$, filled by the lengths of the shortest paths joining each pair of nodes in the graph.
2. Build a new matrix $\mathbf{M_2}$ of dimension $N \times d$, where d is the diameter of the network: each entry is the number of shortest paths starting from the node in the row-position and having length equal to the one in the column-position.
3. Divide each entry by $N - 1$: this is the final matrix $\mathbf{M_3}$, i.e. the NDD representation of the graph.

As an example, Table 6.1 shows the aforementioned steps used to obtain the NDD the unweighted, undirected graph represented in Figure 6.1.

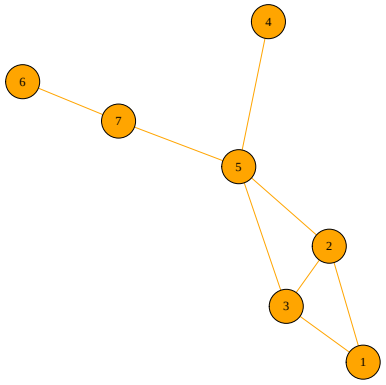


Figure 6.1. An example of unweighted, undirected network.

The matrix in 6.1 (c) is the final representation of the graph through its Node Distance Distribution. It is a row-stochastic matrix, having as generic element in position (i, l) the fraction of nodes reachable from node i with a shortest path of length l. For example, half of the nodes are reached from node labeled with 7 with a shortest path of length 2.

It is worthy mentioning that the NDD of a graph provides information about the global topology of the graph.

The NDD is a very important and useful tool in order to obtain a synthetic representation of the network. Indeed, each graph is represented by a matrix of dimension $N \times d$, and even in a network with very large vertex cardinality, the diameter d is often very short; when this is the case, the network can be referred to as a *small-world network*. In other words, a network satisfies the *small-world property* if the average geodesic distance between pairs of nodes is small relative to the total number of nodes in the network.

The most popular example related to the small-world network refers to Milgram's studies. In the 1960s, sociologist Stanley Milgram conducted several studies to verify the existence of short social connection paths between people and to quantify the average distance between entities in a social network; in other words, his goal was to measure the length of paths between any two nodes in a social network by counting the number of friendship ties between any two people. His studies showed that human society can be regarded as a small-world network, and his discovery is often associated to the sociological theory *Six degrees of separation*: according to the theory, any two people in the world are likely to be connected by a path with no more than 6 edges. In particular, the researcher carried out an experiment in the United States whose purpose was to count how many nodes (people) were needed to send a packet from one randomly selected person to another; the results showed that about an average of 6 friendship ties united those two people. Similar experiments

Table 6.1. Operational procedure to obtain the NDD of the graph $\mathcal{G}$ displayed in Figure 6.1.

(a) step 1

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	1	3	2	4	3
2	1	0	1	2	1	3	2
3	1	1	0	2	1	3	2
4	3	2	2	0	1	3	2
5	2	1	1	1	0	2	1
6	4	3	3	3	2	0	1
7	3	2	2	2	1	1	0

(b) step 2

	1	**2**	**3**	**4**
1	2	1	2	1
2	3	2	1	0
3	3	2	1	0
4	1	3	2	0
5	4	2	0	0
6	1	1	3	1
7	2	3	1	0

(c) step 3

	1	**2**	**3**	**4**
1	0.33	0.17	0.33	0.67
2	0.50	0.33	0.17	0.00
3	0.50	0.33	0.17	0.00
4	0.17	0.50	0.33	0.00
5	0.67	0.33	0	0.00
6	0.17	0.17	0.50	0.17
7	0.33	0.50	0.17	0.00

have been conducted to investigate this interesting topic more thoroughly; among the others, a fairly recent experiment by Dodds, Muhamad, and Watts (2003), characterized by randomly selected sample and by the use of email addresses instead of physical addresses as in Milgram (1967), shows that there is a median of five to seven steps between each pair of people in the world. As a final note, it is interesting to know that in his article Milgram never mentioned the notion of *degree of separation*; it was later coined by John Guare, an American playwrite who titled one of his plays *six degree of separation* (Guare, Sandrich, and Loewenberg, 2000) and his brilliant idea popularized the phrase and related social theory.

6.2.2.2 Transition Matrix

A Transition Matrix (TM) of order s represents a graph using a row-stochastic $N \times N$ matrix (TM of order s), having element $T^s_{ii'}$ equal to the probability of node i' to be reached in s steps by a random walker located in position i, $\forall\, i, i' \in \{1, \ldots, N\}$. Operationally, in order to represent a graph $\mathcal{G}$ by its Transition Matrix of order s, the following steps are needed:

1. Build a matrix $\mathbf{M}$ of dimension $N \times N$, a 0/1 matrix filled by 1, if starting from the row-node it is possible to reach the column-node in s steps.
2. Compute the row sums of $\mathbf{M}$.
3. Divide each entry of $\mathbf{M}$ by the n-column vector of row sums.

In practice, as an example, Table 6.2 shows how to obtain the TM of second order ($s = 2$) representing the network in Figure 6.1.

Table 6.2. Operational procedure to obtain the TM of order 2 of the graph $\mathcal{G}$ displayed in Figure 6.1.

(a) step 1

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	0	0	0	1	0	0
2	0	0	0	1	0	0	1
3	0	0	0	1	0	0	1
4	0	1	1	0	0	0	1
5	1	0	0	0	0	1	0
6	0	0	0	0	1	0	0
7	0	1	1	1	0	0	0

(b) step 2

	1	**2**	**3**	**4**	**5**	**6**	**7**	**sum**
1	0	0	0	0	1	0	0	1
2	0	0	0	1	0	0	1	2
3	0	0	0	1	0	0	1	2
4	0	1	1	0	0	0	1	3
5	1	0	0	0	0	1	0	2
6	0	0	0	0	1	0	0	1
7	0	1	1	1	0	0	0	3

(c) step 3

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0.00	0.00	0.00	0.00	1	0.00	0.00
2	0.00	0.00	0.00	0.50	0.00	0.00	0.50
3	0.00	0.00	0.00	0.50	0.00	0.00	0.50
4	0.00	0.33	0.33	0.00	0.00	0.00	0.33
5	0.5	0.00	0.00	0.00	0.00	0.50	0.00
6	0.00	0.00	0.00	0.00	1	0.00	0.00
7	0.00	0.33	0.33	0.33	0.00	0.00	0.00

The matrix in Table 6.2 (c) is the final row-stochastic matrix of dimension $N \times N$, representing the graph through its Transition Matrix of order 2.

It is worthy mentioning that the transition matrix of order 1 of the graph $\mathcal{G}$ is just the Adjacency matrix of $\mathcal{G}$, rescaled by the degree of each node (number of edges adjacent to the node). In addition, the TM of a generic order s can be obtained as $\mathrm{TM}^s = (\mathrm{TM}^1)^s$.

Finally, the TM of a graph provides information about the connectivity of the graph.

6.2.2.3 Distribution distances

Networks are often represented as probability distributions of their topological features, such as the Node Distance Distribution and the Transition Matrix, which can be summarized as follows:

- *Node Distance Distribution* (NDD): $\mathcal{N}_i^r(l)$, the NDD of node i in graph $\mathcal{G}_r$, is the fraction of nodes in $\mathcal{G}_r$ reachable with the shortest path of length l from node i.
- $\mathcal{T}^r(s)$, the *Transition Matrix* (TM) of a graph $\mathcal{G}_r$ of order s: $\mathcal{T}_{ii'}^r(s)$ is the probability of a node i to be reached in s steps by a random walker located at node i' in the graph $\mathcal{G}_r$.

Once the graph are represented as probability distributions, then a dissimilarity metric between probability distribution can be used to obtain the dissimilarities between graphs. Formally, let $\mathcal{G}_p$ and $\mathcal{G}_q$ be two graphs, with NDDs for node i

$\mathcal{N}_i^p$ and $\mathcal{N}_i^q$ respectively, and TMs of order s=2: $\mathcal{T}^p(2)$ and $\mathcal{T}^q(2)$ respectively. By averaging over all N nodes, the $\mathcal{M}^N$ and $\mathcal{M}^T$ graph distances are defined as:

$$\mathcal{M}^N(\mathcal{G}_p, \mathcal{G}_q) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathcal{J}(\mathcal{N}_i^p, \mathcal{N}_i^q)}, \tag{6.1}$$

$$\mathcal{M}^T(\mathcal{G}_p, \mathcal{G}_q) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathcal{J}(\mathcal{T}_i^p(2), \mathcal{T}_i^q(2))}. \tag{6.2}$$

where $\mathcal{J}$ is the *Jensen-Shannon* distance of distributions (Fuglede and Topsoe, 2004a), which is recalled for the reader in Section 6.2.2.4. Thus each row of $\mathcal{M}^N$ and $\mathcal{M}^T$ contain the networks' dissimilarities. More precisely, the resulting matrix, usually called *Gram matrix* or *Distance matrix* (Manipur et al. (2020a)), is a squared symmetric matrix, having null diagonal elements and non-negative off-diagonal elements. According to Schieber et al. (2017a), such network dissimilarity measure matrix is very precise. Indeed it compares, through the Jensen–Shannon divergence, topological differences between networks. In particular, the generic element of the dissimilarity matrix is equal to 0 if the corresponding graphs are isomorphic and is positive if the pair of graphs are not isomorphic: in this latter case, we can also say that the dissimilarity value quantifies the topological differences between the two graphs.

6.2.2.4 Jensen-Shannon Divergence

According to Fuglede and Topsoe (2004b), the Jensen-Shannon Divergence is a smoothed and symmetrized version of the Kullback-Leibler divergence, one of the most important divergence measure of information theory (Kullback and Leibler, 1951a). The Jensen-Shannon Divergence (JSD) measures how different is the probability distribution P compared to the reference probability distribution Q and it is defined as follows:

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{6.3}$$

where

$$M = \frac{P+Q}{2}$$

and

$$D(A||B) = \sum_{x\in\mathcal{X}} A(x)\log\left(\frac{A(x)}{B(x)}\right) \tag{6.4}$$

is the Kullback-Leibler divergence between two generic discrete distributions A and B with support over $\mathcal{X}$, or, when the two probability distributions are continuous,

$$D(A||B) = \int_{\mathcal{X}} A(x)\log\left(\frac{A(x)}{B(x)}\right)dx \tag{6.5}$$

In other words, the Kullback-Leibler divergence, also known as directed divergence in Kullback and Leibler (1951a) or discrimination information in Kullback (1997a), is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probability P.
It is important to notice that the logarithm in (6.4) and in (6.5) can be the natural logarithm, i.e. $\ln(\cdot)$, or can be at base 2.

6.2.2.5 Denoising Autoencoder

Neural network approaches to generate graph embeddings have been used for several clustering and classification applications. The main task of graph embedding is to project the graphs into a vector space while preserving their properties. Here, we consider the Denoising Autoencoder (DAE) approach by Gutiérrez-Gómez and Delvenne (2019) for generating unsupervised embeddings of graphs sharing the same set of nodes.

A corrupted version of a network is fed to the autoencoder, following which it is trained to reconstruct a clean version of the original input graph. The powers of the adjacency matrices are vectorised and used as input to the DAE. Finally, euclidean distances are computed to construct a distance matrix between the embeddings. In our experiments, we evaluated the embedding dimensions: 800 and 1600; we fixed the power of the adjacency matrices at 3.

6.2.2.6 Joint Embedding

Joint Embedding (JE) is a matrix factorization-based embedding method proposed by Wang et al. (2021). Given a set of undirected graphs, the method first identifies a linear subspace spanned by rank one symmetric matrices; and then projects the adjacency matrices of the graphs into this subspace. The projection coefficients give the features of each graph.

Given R graphs, $\{Gr_i\}_{i=1}^{R}$, with $\mathbf{A}_i$ being the corresponding adjacency matrix, the d-dimensional Joint Embedding of graphs $\{Gr_i\}_{i=1}^{R}$ is given by

$$(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_R, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_d) = \underset{\lambda_i, \|\alpha_k\|=1}{\operatorname{argmin}} \sum_{i=1}^{R} \|\mathbf{A}_i - \sum_{k=1}^{d} \lambda_i[k]\alpha_k * \alpha_k^T\|^2 \tag{6.6}$$

where d is the embedding dimension, $\|\cdot\|$ is the Frobenius norm, λ_i is the loading for graph i, $\lambda_i[k]$ is the kth entry of the vector λ_i and α represents the vector with latent positions of the embedding.

The embedding dimensions d of 2 and 10 are evaluated in the clustering experiments.

6.3 Fuzzy networks clustering

Cluster analysis in a network framework may be applied with a dual purpose: on the one hand, to detect clusters of networks, on the other hand, to recognize clusters of nodes (communities) in one network. We are interested in considering each network as a statistical unit and investigating how the graphs can be grouped and the algorithm's performance through some validity indices. Generally, clustering algorithms apply either to a relational data matrix or a feature (or object) matrix. In the former case, the matrix represents relationships (e.g., dissimilarities) between the units; in the latter case, the matrix has row vectors (one for each unit) representing some features of the corresponding unit. Therefore, we need to obtain a well-defined matrix before applying clustering techniques. The clustering algorithms differ from one another for the input matrix they require. In particular, we focus on algorithms that use a fuzzy approach. Unlike the classical/hard approach, the fuzzy one assigns each unit to a cluster with a membership degree, taking values in the interval $[0, 1]$. The unit interval limits indicate complete non-membership and complete membership, respectively. In this work, on the one hand, we use the so-called

Non-Euclidean Relational Fuzzy Clustering (Davé and Sen, 2002) and the so-called Fuzzy Analysis Clustering or FANNY (Rousseeuw and Kaufman, 1990), which take as input relational data (such as a distance matrix); on the other, we use the Fuzzy k-Means (Bezdek, 1981), the Fuzzy k-Means with polynomial fuzzifier (Klawonn and Höppner, 2003), the Fuzzy k-Means based on L1 metric (Jajuga, 1991) and the Fuzzy k-Medoids (Krishnapuram et al., 2001), that take as input a feature matrix.

6.3.1 Fuzzy clustering algorithms for feature matrix

This Section is fully devoted to recall several clustering algorithms which take as input a feature matrix: in particular, the Fuzzy k-Means algorithms, as well as some of its variations. Clearly, as the algorithms take as input a feature matrix, the technique described in Section 6.2.2.6 (Joint Embedding) is used to represent networks by rows.

A *de-facto* standard fuzzy clustering method is the Fuzzy k-Means (FkM) (Bezdek, 1981), the fuzzy generalisation of the k-means algorithm (MacQueen, 1967b). Given a data matrix $\mathbf{X}$ of order $N \times J$, where N and J are the number of units and features/variables, respectively, and letting $d(\mathbf{x}_i, \mathbf{x}_{i'})$ denote the dissimilarity between objects $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, the FkM objective function to be minimised to find K clusters is:

$$J_{\text{F}k\text{M}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik}^m d^2(\mathbf{x}_i, \mathbf{h}_k) \tag{6.7}$$

s.t.

$$\mu_{ik} \in [0,1] \qquad \forall\, i = 1, \ldots, N, \qquad \forall\, k = 1 \ldots, K \tag{6.8}$$

$$\sum_{k=1}^{K} \mu_{ik} = 1 \qquad \forall\, i = 1, \ldots, N, \tag{6.9}$$

The matrix $[\mu_{ik}]_{i=1,\ldots,N,k=1,\ldots,K}$ is of order $N \times K$ and contains the membership degrees, while $\mathbf{H}$ is the $K \times J$ prototype matrix. The parameter m (> 1) indicates the fuzziness of the partition. A common choice for m is in the interval $[1.5, 2]$. The clustering is not meaningful for $m \to \infty$, leading to the same constant membership degree for each unit. In contrast, the fuzzy approach will become the classical hard one when $m \to 1$, the membership degree tends to be either 0 or 1.

A variation of FkM algorithm is introduced by Jajuga (1991), who proposed the Fuzzy k-Means based on L1 metric (FkM.L1), i.e. a variation of the standard Fuzzy k-Means, which generally use the Euclidean metric to measure the distance between the units. In this variation, instead, the L1 metric is used.

A particular version of Fuzzy k-Means, i.e. the Fuzzy k-Medoids (FkMed), introduced by Krishnapuram et al. (2001), consists in considering the medoids as centroids. In particular, letting $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ $\mathbf{c}_i \in \mathbf{X}$ represent a subset of $\mathbf{X}$ with cardinality K, $\mathbf{X}^k$ represent the set of all subsets of $\mathbf{X}$ with cardinality K, the objective function is the following:

$$J_{\text{F}k\text{Med}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik}^m d^2(\mathbf{x}_i, \mathbf{c}_k) \tag{6.10}$$

s.t.

$$\mu_{ik} \in [0,1] \qquad \forall\, i = 1, \ldots, N, \qquad \forall\, k = 1 \ldots, K \tag{6.11}$$

$$\sum_{k=1}^{K} \mu_{ik} = 1 \qquad \forall\, i = 1, \ldots, N, \tag{6.12}$$

The matrix $\mathbf{C}$ is a $K \times J$ matrix having K medoids, i.e. representative objects of the data set, by rows. Each cluster is therefore represented by one medoid.

A later proposal is provided by Klawonn and Höppner (2003), who introduced the Fuzzy k-Means with polynomial fuzzifier (FkM.pf), which is an extension of the Fuzzy k-Means algorithm. More in details, the authors proposed to use an alternative fuzzifier function, i.e. the polynomial fuzzifier function. Formally, a fuzzifier function is a continuous, strictly increasing function $f = [0,1] \rightarrow [0,1]$, with $f(0) = 0$ and $f(1) = 1$. Klawonn and Höppner (2003) proposed to use the following fuzzifier function: $\frac{1-\beta}{1+\beta}\mu_{ik}^2 + \frac{2\beta}{1+\beta}\mu_{ik},\ \beta \in [0,1)$, which replaces the fuzzifier function of FkM in Equation 6.7, i.e. u_{ik}^m. The parameter β describes the ratio of distances at which the clustering result becomes crisp. The polynomial fuzzifier function is in some way a linear combination between hard clustering and FkM with fuzzifier $m = 2$.

6.3.2 Fuzzy clustering algorithms for relational data matrix

Some fuzzy clustering algorithms which take as input a relational data matrix are now recalled. Clearly, to apply the following clustering algorithms, it is needed to obtain a matrix of distances between networks and for this reason the techniques explained in Sections 6.2.2.3 and 6.2.2.5 are used.

The Fuzzy Analysis clustering (FANNY) algorithm (Rousseeuw and Kaufman, 1990) is a fuzzy clustering method for relational data, such as distances/dissimilarities. FANNY consists in minimizing the following optimization problem:

$$J_{\text{FANNY}} = \sum_{k=1}^{K} \frac{\sum_{i'=1}^{N}\sum_{i=1}^{N} \mu_{ik}^2 \mu_{i'k}^2 d(\mathbf{x}_i, \mathbf{x}_{i'})}{2\sum_{s=1}^{N} \mu_{sk}^2} \tag{6.13}$$

s.t.

$$\mu_{ik} \in [0,1] \qquad \forall\, i = 1, \ldots, N, \qquad \forall\, k = 1 \ldots, K \tag{6.14}$$

$$\sum_{k=1}^{K} \mu_{ik} = 1 \qquad \forall\, i = 1, \ldots, N, \tag{6.15}$$

In Equation 6.13, $d(\cdot,\cdot)$ usually is the L1-distance.

A generalization of the FANNY algorithm is the so-called Non-Euclidean Fuzzy Relational Clustering (NEFRC) algorithm (Davé and Sen, 2002). The authors, indeed, allow for a general fuzzifier m: in Equation 6.13, they replaced the exponent 2 of μ_{ik} with a general m. In addition, they allow any relational data matrix coming from a general distance.

6.4 Empirical analysis

Before describing the datasets and showing the application of clustering algorithms on simulated and real networks and the main results, a brief *excursus* on the graphical representation of the results and on the clustering evaluation metrics is needed.

6.4.1 Visual exploratory analysis and evaluation metrics

Two-dimensional space coordinates for graphs are obtained by reducing the mutual distance matrix D through a dimensionality reduction method. In particular, we use the t-distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique allowing the embedding of high-dimensional data for visualization in a low-dimensional space. Nearby points model similar observations and dissimilar observations are modelled by distant points (Van der Maaten and Hinton, 2008).
Here, we use the Barnes-Hut implementation of t-SNE introduced in Van Der Maaten (2014), which is available in the R package, `Rtsne` (Krijthe, 2015). In the following sections, we present the results through the t-SNE and analytically, by using some external validity indices, such as the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and the Adjusted Mutual Information (AMI) (Vinh, Epps, and Bailey, 2010) when the true partition is available. When the true partition is not provided, we use the Silhouette index (Sil) (Rousseeuw and Kaufman, 1990) and its fuzzy version (Fuzzy Sil) (Campello and Hruschka, 2006b) as internal validity indices. Clustering partition is obtained by "discretizing" the membership degree matrix: each unit is assigned to the cluster whose membership degree is the highest.

6.4.2 Simulated data

In this Section the application to two different sets of simulated data is shown, in particular, the application of NEFRC and FANNY algorithms to the distance matrices while FkM, FkMed, FkM.pf and FkM.L1 to the feature matrix. We let the fuzzifier m range in a sequence from 1.1 to 2 by 0.1 to have as many clustering validity indices as the m's. We show the median result and deviation measures such as the interquartile range (IQR) and the standard deviation (SD).

The first set of simulated networks is generated using the Multiple Random Eigen Graphs (MREG) model, defined in Wang et al. (2021) as

$$(\lambda_i, A_i)_{i=1}^{R} = MREG(F, h_1, \ldots, h_d).$$

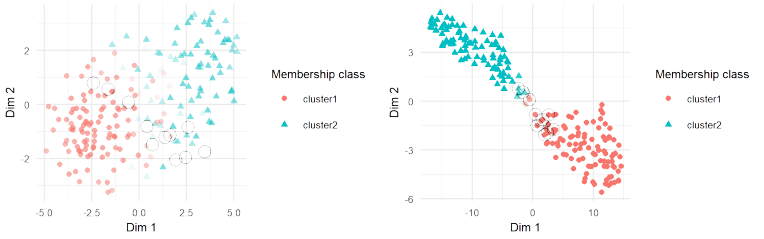
Given R graphs, $A_1, \ldots A_R$ are their random adjacency matrices, generated with the d-dimensional MREG model. $\{\lambda_i\}_{i=1}^{R}$ are random variables, and F denotes their distribution on χ, where $\chi \subseteq \mathbb{R}^d$ such that $x^T y \in [0, 1]$, for all $x, y \in \chi$. $\{h_k\}_{k=1}^{d}$ are vectors which satisfy $\sum_{k=1}^{d} \lambda_i[k] h_k h_k^T \in [0, 1]^{n \times n}$, for all $\lambda \in \chi$.

A d=2 dimensional MREG dataset with 200 graphs having 100 nodes each was generated using this model. The graphs belong to 2 classes, with 100 graphs in each class. We set λ=[24.5, 4.75] for class 1 and λ=[20.75, 2.25] for class 2. The entries of h_1 are all set to 0.1; we set the first half entries of h_2 to -0.1, and the remaining to 0.1 (Wang et al., 2021). The clustering task consists of grouping networks with a similar distribution of edges.

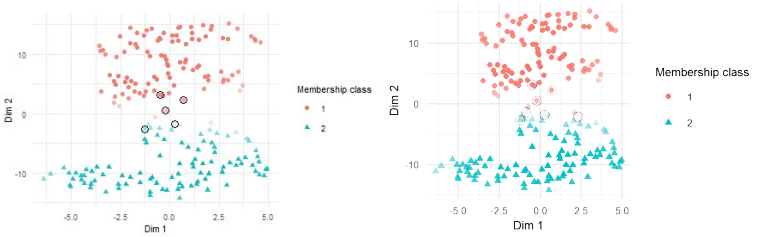
The Lancichinetti–Fortunato–Radicchi (LFR) benchmark generator (Lancichinetti, Fortunato, and Radicchi, 2008) is used to construct the second simulated dataset of undirected and unweighted networks. The parameter μ controls the strength of the communities in the dataset: small values of μ result in well-defined communities. Therefore, the clustering task consists in detecting networks with similar communities structures. The LFR dataset contains three classes with 100 graphs each, generated using three different values: $\mu = 0.1$, 0.5, and 0.8, with all graphs containing 100 nodes. In the simulation study we represent each ensemble of networks using distance matrices, i.e. $\mathcal{M}^N$, $\mathcal{M}^T$, DAE, and the feature matrix JE, and we apply clustering

on them. For simulated datasets, the number of clusters is a priori known and therefore the clustering algorithms were run by asking for the true number of clusters. The obtained partitions were then compared to the true ones by means of external cluster validity indices, such as ARI and AMI indices.
For the MREG networks best results are obtained by using $\mathcal{M}^N$, $\text{DAE}_{d=1600}$ and $\text{JE}_{d=2}$. Table 6.3, Table 6.4 and Figure 6.2 show results of the application of NEFRC and FANNY to $\mathcal{M}^N$ and DAE and of FkM, FkMed, FkM.pf and FkM.L1 to JE.

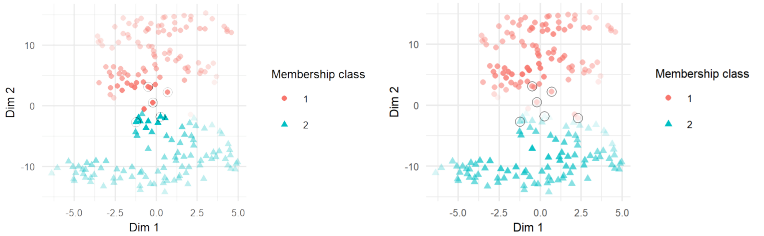
(a) NEFRC applied on $\mathcal{M}^N$ **(b)** NEFRC applied on $\text{DAE}_{d=1600}$



(c) FANNY applied on $\mathcal{M}^N$ **(d)** FANNY applied on $\text{DAE}_{d=1600}$



(e) FkM applied on $\text{JE}_{d=2}$ **(f)** FkMed applied on $\text{JE}_{d=2}$



(g) FkM.pf applied on $\text{JE}_{d=2}$ **(h)** FkM.L1 applied on $\text{JE}_{d=2}$

Figure 6.2. t-SNE representation of clustering results of NEFRC, FANNY, FkM, FkMed, FkM.pf and FkM.L1 (MREG networks). Misclassified units are circled in black. The intensity of the colors is given by the membership degree of each network to the corresponding assigned cluster.

Tables 6.3 and 6.4 show the algorithm's performance using the clustering validity indices. When NEFRC is applied to $\mathcal{M}^N$, and FkM is applied to JE, high ARI and AMI indices are obtained, which show that most of the networks are correctly assigned to their original clusters. By analyzing the results of the application of FANNY to the distance matrices, we observe that the validity indices are slightly lower than the ones by using the NEFRC algorithm. The same occurs when comparing the results of the application of FkMed, FkM.pf and FkM.L1 with the results of the

Table 6.3. Main results of the application of NEFRC and FANNY to Distance Matrices (MREG networks)

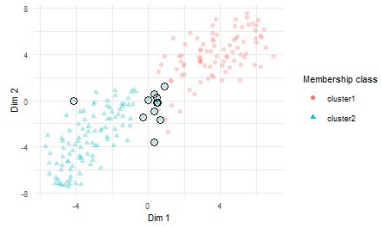
	NEFRC				FANNY			
	$\mathcal{M}^N$		$\text{DAE}_{d=1600}$		$\mathcal{M}^N$		$\text{DAE}_{d=1600}$	
	ARI	AMI	ARI	AMI	ARI	AMI	ARI	AMI
Median	**0.81**	0.72	0.76	0.67	0.64	0.64	0.76	0.67
IQR	0	0	0.04	0.03	0.1	0.09	0.01	0.01
SD	0.01	0.02	0.02	0.02	0.06	0.05	0.02	0.01

Table 6.4. Main results of the application of FkM, FkMed, FkM.pf and FkM.L1 to Joint Embeddings Matrix (MREG networks)

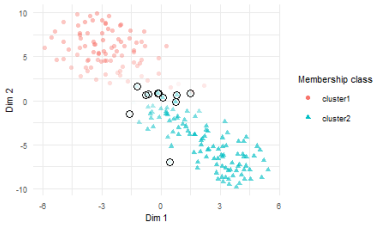
	FkM		FkMed		FkM.pf		FkM.L1	
	$\text{JE}_{d=2}$		$\text{JE}_{d=2}$		$\text{JE}_{d=2}$		$\text{JE}_{d=2}$	
	ARI	AMI	ARI	AMI	ARI	AMI	ARI	AMI
Median	**0.9**	0.83	0.86	0.78	0.87	0.79	0.88	0.80
IQR	0	0	0	0	0.06	0.08	0	0
SD	0.01	0.01	0	0	0.05	0.05	0	0

application of the classical FkM.
Figure 6.2 shows that the two clusters are well separated; the circled points highlight networks that the algorithms have misclassified. We can indeed study the misclassified units in-depth using the fuzzy membership degree matrix. By applying NEFRC to distance matrices, we notice that, on average, around 40% of misclassified networks are in the middle of the two cluster prototypes. Therefore, the unit can be assigned to any cluster, with both membership degrees close to 0.5. Regarding the application of FkM to $\text{JE}_{d=2}$, we notice that out of 5 misclassifications, one is approximately in the middle of the two clusters prototypes whose membership degrees are 0.58 and 0.42. Therefore, membership degrees allow us to consider the uncertainty of an assignment of a unit to a cluster and then eventually add information on clustering interpretation: this represents one of the main advantages of a fuzzy approach.

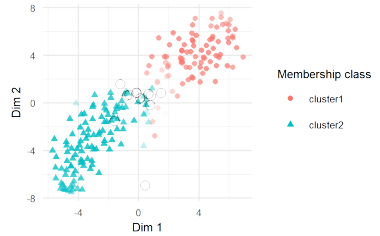
For the second set of simulated networks, i.e., LFR, we performed pairwise comparisons. More specifically, we applied clustering on the restricted set of networks having $\mu = 0.1$ and $\mu = 0.5$; the same is done on the restricted set of networks having $\mu = 0.1$ and $\mu = 0.8$ and on the one with networks characterised by $\mu = 0.5$ and $\mu = 0.8$. Finally, a triple-wise comparison also has been made.
By applying the NEFRC algorithm to $\mathcal{M}^N$, the true partitions are well detected when analyzing pairwise comparisons between the first class of networks (i.e., those generated using $\mu = 0.1$) and the second one ($\mu = 0.5$) and between the first one and the third one ($\mu = 0.8$). The same analysis has been carried out by using the FANNY algorithm, but the results are slightly worse in terms of validity indices, being them lower than the ones obtained by applying the NEFRC algorithm to the same matrices. A summary of the results is provided in Table 6.5 and Figure 6.3.
From the summary table (Table 6.5), high clustering indices show that most units are correctly assigned to their original cluster; from the visual representation (Figure 6.3), the clusters are well separated. When we investigate the membership degrees, we observe that for the first pairwise comparison ($\mu = 0.1$ and $\mu = 0.5$), all the



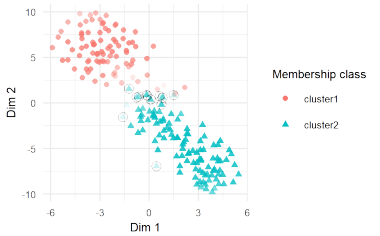
(a) NEFRC applied on $\mathcal{M}^N$ ($\mu = 0.1 - \mu = 0.5$)



(b) NEFRC applied on $\mathcal{M}^N$ ($\mu = 0.1 - \mu = 0.8$)



(c) FANNY applied on $\mathcal{M}^N$ ($\mu = 0.1 - \mu = 0.5$)



(d) FANNY applied on $\mathcal{M}^N$ ($\mu = 0.1 - \mu = 0.8$)

Figure 6.3. t-SNE representation of clustering results of NEFRC and FANNY (LFR networks). Misclassified units are circled in black. The intensity of the colors is given by the membership degree of each network to the corresponding assigned cluster.

Table 6.5. Main results of the application of NEFRC and FANNY to Distance Matrices (LFR networks).

	NEFRC				FANNY			
	$\mathcal{M}^N$($\mu = 0.1$ and $\mu = 0.5$)		$\mathcal{M}^N$($\mu = 0.1$ and $\mu = 0.8$)		$\mathcal{M}^N$ ($\mu = 0.1$ and $\mu = 0.5$)		$\mathcal{M}^N$ ($\mu = 0.1$ and $\mu = 0.8$)	
	ARI	AMI	ARI	AMI	ARI	AMI	ARI	AMI
Median	**0.77**	0.72	**0.79**	0.72	0.75	0.71	0.77	0.72
IQR	0.02	0.02	0.02	0.02	0	0	0	0
SD	0.01	0.01	0.01	0.01	0.01	0.01	0	0

misclassified units by NEFRC are in the middle of the two cluster prototypes with both membership degrees close to 0.5. For the second pairwise comparison ($\mu = 0.1$ and $\mu = 0.8$), we can state similarly: in this case, among the 11 misclassifications by NEFRC, 45% is close to both cluster prototypes.

The results show that fuzzy algorithms essentially recognize the true partition when representing the networks with the $\mathcal{M}^N$ distance matrix. Moreover, the fuzzy approach gives the membership degree matrix, which is helpful in terms of clustering interpretation, as stated previously. Indeed, we can understand the results deeply and quantify the uncertainty related to the assignment of a given network to a given cluster.

6.4.2.1 European Air Transportation Network

This section is devoted to the analysis of European Air Transportation Network. The European Air Transportation Network (ETN) (Cardillo et al. (2013)) is a multiple network made of 37 networks: each network represents a different European

airline company. All network share the same set of nodes, i.e. 450 nodes, representing European airports and they have different edges, representing flights routes. These network layers are undirected and unweighted and in our analysis we consider each layer as an individual network to which the $\mathcal{M}^N$ and DAE are applied, followed by network clustering.

The aim of the analysis is to obtain groups of European airline companies (networks) characterized by similar structural features. In this real case study, the true partition is not known, i.e. the ground truth is not available, and therefore we used the internal validity indices to evaluate the obtained partitions. In particular, the Silhouette index (Sil) and its fuzzy version (Fuzzy Sil) were computed to choose the optimal number of clusters K. Furthermore, we ran all the algorithms by letting the number of clusters K range in $[2, 6]$ and we kept the solution with the value of K that maximizes the values of the validity indices.

In this work, we applied NEFRC to $\mathcal{M}^N$ and DAE and FkM to $\text{JE}_{d=2}$. The results highlight that the FkM algorithm applied to $\text{JE}_{d=2}$ with $K = 2$ clusters leads to a Sil value equal to 0.88 and to a Fuzzy Sil value equal to 0.95. NEFRC algorithm applied on $\text{DAE}_{d=800}$ network representation with $K = 2$ clusters leads to a Sil index of 0.69 and to a Fuzzy Sil index of 0.76. Table 6.6 reports the values of the cluster validity indices and the values of K corresponding to the best partition ($K = 2$ in all the cases).

Table 6.6. Main results of the application of NEFRC to Distance Matrices and of FkM to Joint Embeddings Matrix (European Air Transportation Networks).

	NEFRC		FkM
	$\mathcal{M}^N$	$\text{DAE}_{d=800}$	$\text{JE}_{d=2}$
Sil	0.48	**0.69**	**0.88**
Fuzzy Sil	0.48	**0.76**	**0.95**
K^*	2	2	2

The FkM algorithm applied to JE actually detects two clusters: it groups together in one cluster all but one networks; the other cluster instead is made by the network representing Ryanair Airline company. We observe that *Ryanair* has one peculiar characteristic: it is the network that has the lowest number of isolated nodes, i.e. it is the company that, among all the other companies considered in this study, covers the most number of airports (nodes).

The NEFRC algorithm applied to DAE tends to put together companies on regional basis: it pairs *Lufthansa* and *Air Berlin* (Germany), *Scandinavian Airlines* and *Norwegian* (Scandinavia), *Aegean Airlines and Olympic Air* (Greece), *Ryanair* and *Air Lingus* (Ireland), *Germanwings* and *SunExpress* (Germany). As highlighted also in Tantardini et al. (2019), in many instances the two paired airlines are the leading national company and a low-cost company, offering different journey and price conditions on the same routes: this is the case for example for *Ryanair* (Low Cost) and *Aer Lingus* (Leading National Company) in Ireland or for *Air Berlin* (Low cost) and *Lufthansa* (Leading National Company) in Germany. Indeed, airlines based in the same region share the same airports and therefore the nodes of two airlines based in the same nation are typically similar. This reduces the distance

between the corresponding networks and explains why the two companies (i.e. the two networks) are placed in the same cluster.

By focusing on networks' structure, as in Tantardini et al. (2019), we observe that the first cluster groups together non-star networks, while the second cluster groups together all pure-star networks and all networks close to star topology. A star topology is characterized by nodes that are not directly connected to each other, but they are connected to a central node, normally called a *hub*; as a consequence all nodes are indirectly connected one another through the central one; so we can think about the hub as the center of the star and the other nodes different from the hub as the points of the star.

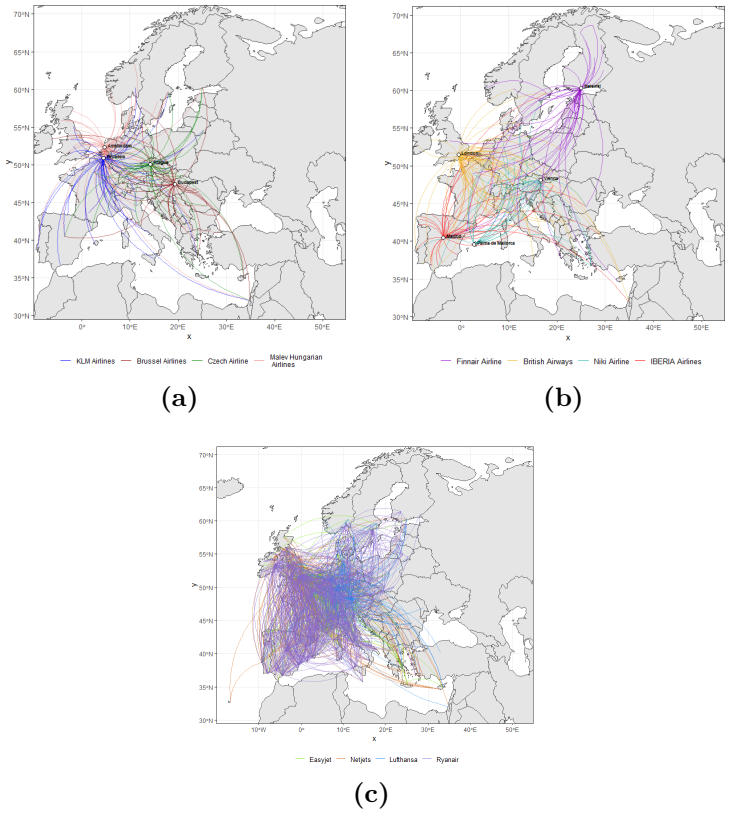


Figure 6.4. European Air Transportation Networks: Pure-star networks (a) and networks close to a star topology (b) belonging to the second cluster; some of non-star networks (c) belonging to the first cluster (according to NEFRC results applied on DAE).

Indeed, in our clustering results, the second cluster groups together all pure-star networks (see Figure 6.4 (a)): *KLM*, whose hub is Amsterdam Airport Schipol, having degree 62; *Brussel Airlines* whose hub is Brussels Airport, having degree 43; *Czech Airlines*, having as hub Václav Havel Airport Prague with degree 26; the last pure-star network is the one representing *Malev Hungarian Airlines*, whose hub is Budapest Liszt Ferenc International Airport, having degree 34. All these networks are therefore represented in the same way, as a star, with one node (airport) at the center of the star and all the others nodes as the points of the star. The difference among those networks is the number of nodes. The second cluster includes also networks close to a star topology (see Figure 6.4 (b)). In particular: Iberia, British Airways, Niki and Finnair. On the other hand, the first cluster groups networks that

have no hub and whose nodes have similar nodes degrees: graphically (see Figure 6.4 (c)) it is not possible to identify a main central hub and most of the nodes are connected one another.

Finally, by following the idea of Carpi et al. (2019), we notice that among the 37 airlines companies, 17 of them joined three airlines alliances: Star Alliance, One World, Skyteam. We then investigate whether our clustering results can match this classification; so, by considering only those 17 out of 37 networks, we find out that our algorithm (NEFRC applied to DAE) detects only two clusters. This is not surprising: indeed, we realize that airlines' alliances include companies that are not similar each other in terms of covered regions; indeed, airlines belonging to the same alliance fly from and to different airports and then the corresponding networks have dissimilar nodes. Instead, our algorithm recognizes in the same cluster networks that have similar nodes, i.e. companies that share the most of the flight routes and airports. For example, *Air Berlin* that joined One World Alliance is placed in the same cluster of *Lufthansa*, that instead is part of Star Alliance. This result demonstrates the importance of properly representing networks: indeed, the DAE representation together with the application of NEFRC have been useful to recognize clusters characterized by similar flight routes, without being influenced by the existing theoretically underling grouping structure.

This real application allowed us to study the performance of the proposed clustering algorithms obtaining good results in terms of validity indices, and also meaningful considerations regarding clustering interpretation.

6.4.3 FAO correlations networks

This section is devoted to the analysis of FAO correlations networks. The dataset consists of 140 networks where each network represents a country and share the same set of nodes. The nodes represent 21 variables related to sustainability, climate change, economic and production indicators (downloaded from the FAOSTAT database `https://www.fao.org/faostat/en/#data`). For each country, we generated a network as follows: we consider 21 time series of the indicators of interest from 1990 to 2019. In order to avoid autocorrelation, we detrended the data using the function `detrend` of the R package `pracma` (Borchers, 2021). The function removes the linear trend from the data by computing the least-squares fit of a line and subtracting the resulting function from the data. After detrending, we computed the correlation matrix for each country and transformed each correlation matrix into an adjacency matrix by adding an edge between any pair of variables whenever the correlation between them, in its absolute value, is higher than a threshold (0.7). Therefore, the generic element of the adjacency matrix for country c, A^c_{ij}, is 1 if variables i and j are highly correlated, 0 otherwise. Table 6.7 provides a detailed overview of the variables (nodes).

In this real case study, the true partition is not known, i.e. the ground truth is not available, and therefore we used the internal validity indices to evaluate the obtained partitions. In particular, the Silhouette index (Sil) and its fuzzy version (Fuzzy Sil) were computed to choose the optimal number of clusters K. Furthermore, we ran all the algorithms by letting the number of clusters K range in $[2, 6]$ and we kept the solution with the value of K that maximizes the values of the validity indices.

By applying the clustering algorithms to the matrices representation of the networks, we noticed that $\mathcal{M}^T$, DAE and JE have good results. Table 6.8 reports the values of the cluster validity indices and the values of K corresponding to the best partition, i.e. the number of clusters that maximizes the validity indices. Figure

Table 6.7. Node labels and description (source: FAOSTAT).

Node label	Node name	Node description
v 1	CH4 LULUCF	total methane (CH4) emissions, measured in kilotonnes, from land-use, land-use change and Forestry.
v 2	CH4 AFOLU	total methane (CH4) emissions, measured in kilotonnes, from Agriculture, Forestry, and Other Land Use.
v 3	CH4 Emissions on agricultural land	total methane (CH4) emissions, measured in kilotonnes, from Agriculture.
v 4	CH4 Farm-gate emissions	Farm-gate total methane (CH4) emissions, measured in kilotonnes.
v 5	CH4 land-use change	total methane (CH4) emissions, measured in kilotonnes, from land-use change.
v 6	CO2 LULUCF	total emissions of carbon dioxide (CO2), measured in kilotonnes, from land-use, land-use change and Forestry
v 7	CO2 AFOLU	total carbon dioxide (CO2) emissions, measured in kilotonnes, from Agriculture, Forestry, and other land-use.
v 8	CO2 Emissions on agricultural land	total carbon dioxide (CO2) emissions, measured in kilotonnes, from Agriculture.
v 9	CO2 Farm-gate emissions	Farm-gate total carbon dioxide (CO2) emissions, measured in kilotonnes.
v 10	CO2 land-use change	total carbon dioxide (CO2) emissions, measured in kilotonnes, from land-use change.
v 11	N2O LULUCF	total nitrous oxide (N2O) emissions, measured in kilotonnes, from land-use, land-use change and Forestry.
v 12	N2O AFOLU	total nitrous oxide (N2O) emissions, measured in kilotonnes, from Agriculture, Forestry, and Other Land Use.
v 13	N2O Emissions on agricultural land	total nitrous oxide (N2O) emissions, measured in kilotonnes, from Agriculture.
v 14	N2O Farm-gate emissions	Farm-gate total nitrous oxide (N2O) emissions, measured in kilotonnes.
v 15	N2O land-use change	total nitrous oxide (N2O) emissions, measured in kilotonnes, from land-use change.
v 16	Agriculture	Gross per capita Production Index Number (2014-2016 = 100) for agricultural products.
v 17	Livestock	Gross per capita Production Index Number (2014-2016 = 100) for livestock.
v 18	Vegetables and Fruit Primary	Gross per capita Production Index Number (2014-2016 = 100) for vegetables and fruit.
v 19	Gross Domestic Product	gross domestic product (in value US$ per capita, 2015 prices).
v 20	Agricultural land	share of the agricultural land in land area (%).
v 21	Temperature change	increment of temperature (measured in °C) in the meteorological year w.r.t the previous year.

6.5 shows the resulting clusters.

Table 6.8. Values of the cluster validity indices and of the optimal numbers of clusters (K^*) related to the application of NEFRC and FANNY to Distance Matrices and of FkM, FkMed, FkM.pf, FkM.L1 to Joint Embeddings Matrix (FAO correlation networks).

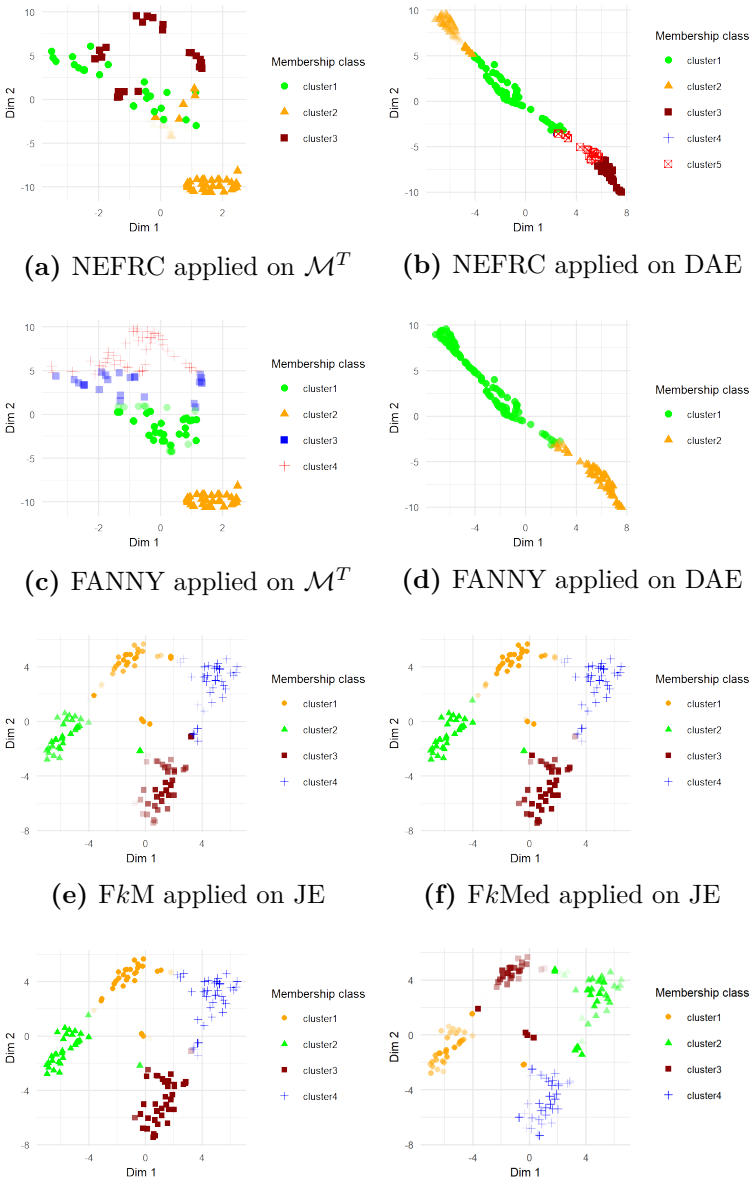
	NEFRC		FANNY		FkM	FkMed	FkM.pf	FkM.L1
	$\mathcal{M}^T$	DAE	$\mathcal{M}^T$	DAE	JE	JE	JE	JE
Sil	0.1	0.33	-0.1	0.53	0.51	0.51	0.51	0.50
Fuzzy Sil	**0.77**	0.66	0.72	0.6	**0.78**	0.75	0.73	**0.78**
K^*	3	5	3	2	4	4	4	4

FkM algorithm applied to JE leads to a Sil value equal to 0.51 and a Fuzzy Sil value equal to 0.78 and its performance is slightly better than the ones of FkMed, FkM.pf and FkM.L1. Moreover, NEFRC algorithm applied on $\mathcal{M}^T$ and DAE network representations leads to a Fuzzy Sil index of 0.77 and 0.66, respectively. Also in this case, NEFRC performs better than FANNY, as the indices show. Moreover, NEFRC applied to DAE representation of the networks identifies $K = 5$ clusters, but one of them is only made up with one country, Chile.

From Figure 6.5 we can see that clusters are well separated when using JE and DAE representation. Since FkM applied on the JE representation of networks leads to the highest cluster validity indices, we will discuss its resulting partition. Figure 6.6 depicts the obtained clustering.

The map shows some European countries grouped with some Asian countries and Canada. Moreover, Latin American countries are in the same cluster as some African countries. The USA and Central America share some characteristics with Central African countries and some Asian countries and islands; finally, Australia is in the same cluster as Spain and some Asian and African countries.

We can explore cluster characteristics and structure by looking at Figure 6.7 with graphs that summarize the four clusters. In Figure 6.7, edges are colored differently



(a) NEFRC applied on $\mathcal{M}^T$ **(b)** NEFRC applied on DAE

(c) FANNY applied on $\mathcal{M}^T$ **(d)** FANNY applied on DAE

(e) FkM applied on JE **(f)** FkMed applied on JE

(g) FkM.pf applied on JE **(h)** FkM.L1 applied on JE

Figure 6.5. t-SNE representation of clustering results of NEFRC, FANNY, FkM, FkMed, FkM.pf and FkM.L1 (FAO correlation networks). The intensity of the colours is given by the membership degree of each network to the corresponding assigned cluster.

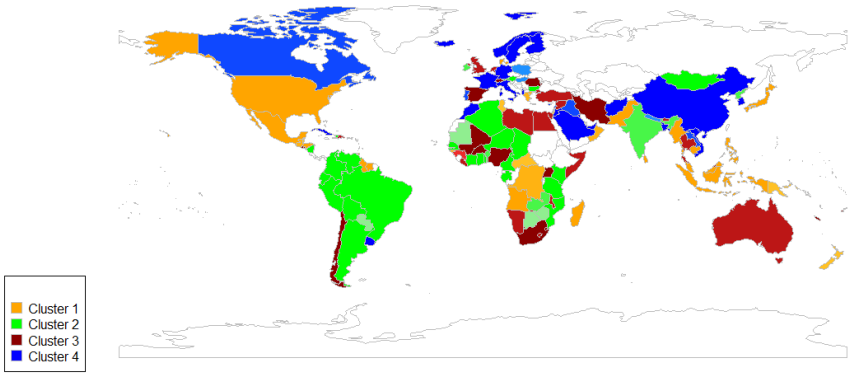
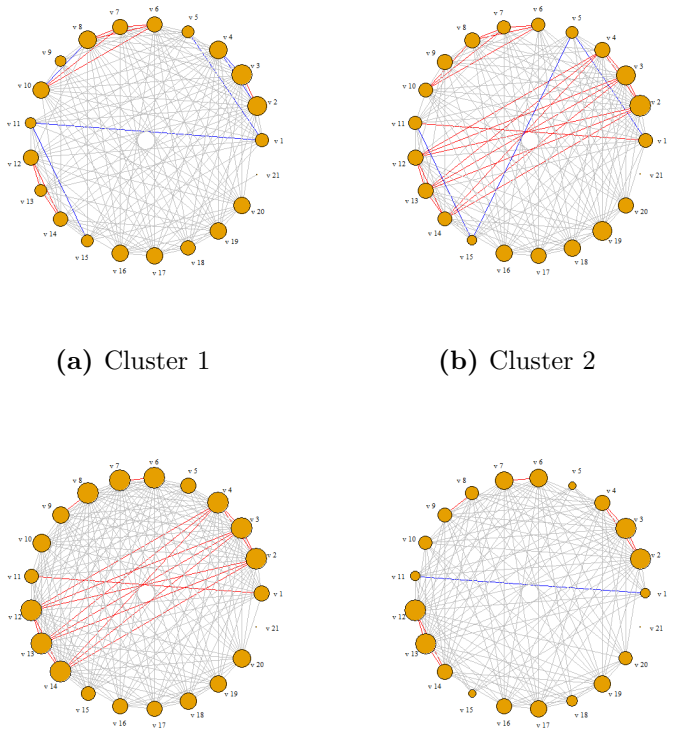


Figure 6.6. Clusters obtained by applying FkM to the JE representation of FAO correlations networks. The intensity of the colors is given by the membership degree of each network to the corresponding assigned cluster.

according to the frequency they appear within the networks belonging to the same cluster. In particular, grey edges represent those shared by less than 50% of graphs belonging to the same cluster; blue edges represent those in common between 50% and 80% of networks belonging to the same cluster; finally, red edges are those edges that appear in most of the networks (i.e. more than 80%) in the same cluster.

We can characterise the first cluster as the one where most networks show a high correlation between N2O AFOLU emissions and N2O emissions from agricultural land and farm-gate. CO2 land-use change emissions are highly correlated with CO2 LULUCF and CO2 AFOLU emissions. In most countries, there is a high correlation between CO2 emissions on agricultural land and CO2 AFOLU emissions and CO2 emissions on agricultural land and CO2 LULUCF emissions. Moreover, CO2 LULUCF and CO2 AFOLU emissions are highly correlated, as CH4 emissions on agricultural land and CH4 AFOLU emissions. More than half of countries belonging to the first cluster have a high correlation between CO2 emission for land-use change and CO2 emissions on agricultural land, CH4 LULUCF and CH4 land-use change, CH4 LULUCF and N2O LULUCF, N2O LULUCF and N2O land-use change. We conclude that these countries have high correlations between variables related to CO2 emissions and between variables related to N2O emissions.
Following the same rationale, countries in the second cluster have a high correlation between variables related to emissions of the same gas (CH4, N20 and CO2) and between variables related to emissions of different gases (CH4 and N2O). That means in almost all the countries belonging to the second cluster, emissions of the same gas for a different purpose are highly correlated, and emissions of different gas for a different purpose. Also, we observe that emissions of CO2 highly correlate with each other.
Cluster 3 is again characterised by a high correlation between CH4 and N2O emissions. However, differently from before, we do not observe neither a high correlation between the variables related to CO2 emissions, nor a high correlation between land-use change emissions.
Finally, the fourth cluster groups countries sharing high pairwise correlations between



(a) Cluster 1 **(b)** Cluster 2

(c) Cluster 3 **(d)** Cluster 4

Figure 6.7. FAO correlation networks: results from the application of FkM to JE representation.

N2O emissions stemming from agricultural activities, those from farm-gate and those from AFOLU. Also, high pairwise correlations appear between CH4 emissions stemming from agricultural, farm-gate and AFOLU.

By summing up, we notice that the clusters are different from one another according to the different edges they contain. In other words, clusters can be described and interpreted using the most common edges between the networks in the same cluster. We observe that the main difference is mainly driven by the relationships (i.e. high correlations) between variables related to emissions.

By analyzing our clustering results and comparing them with the reports by the UN, we can make the following considerations. In the first cluster one of the top four emitters, i.e. USA, (see United Nations Environment Programme (2019)) is grouped together with Sub-Saharan Countries, where there are energy and land-based emissions (see United Nations Climate Change (2022)), and Pakistan, that has as main contributors to the total emissions energy, agriculture and forestry sectors (see Ministry of Climite Change Government of Pakistan (2022)). In the second cluster, one of the top four emitters, i.e. India, (see United Nations Environment Programme (2019)) is grouped together mainly with Latin American Countries: in India, the main driver to emission has been electricity production (see Ministry of Environment Forest and Climate Change Government of India (2021)), in Latin American Countries, for many years the share of national emissions from the Land Use, Land Use Change

and Forestry sectors was huge; due to the reduction of deforestation, the share from Land Use, Land Use Change and Forestry is decreasing, and consequently the share of the Energy and Agriculture sectors is becoming larger (see for example Ministry of Foreign Affairs, Ministry of Science, Technology and Innovations (2020), World Bank Group (2022b), and World Bank Group (2022c)). The third cluster groups countries, such as South Africa, Australia, Chile, and Egypt, whose main contributor to the emissions is the energy sector (see for example Department of Forestry, Fisheries and the Environment (2021), Australian Government, Department of Environment and Energy (2019), Minister of Environment of Chile (2018), and Ministry of Environment, Egyptian Environmental Affairs Agency (2018)). Finally, the last cluster groups two of the top four emitters, i.e. Europe and Canada (United Nations Environment Programme, 2019), and mainly Saudi Arabic and China: we observe that in this case, those countries do not share the main emission driver, being oil and gas sector for Canada (see Environment and Climate Change Canada (2020)), energy sector for Europe and Saudi Arabic (see European Commission (2020) and Ministry of Energy, Industry and Mineral Resources, Kingdom of Saudi Arabia (2018), respectively) and power sector for China (see World Bank Group (2022a)).

This application highlights how to study the performance of clustering algorithms on network ensembles. As a final remark, we note that a simple application of a clustering technique to the matrix having units on the rows and covariates based on correlation on the columns, could possibly work in case of correlation networks, but would not scale up with an increasing number of variables. Indeed, let p be the number of indicators we are considering. The number of columns of such a matrix will be the cardinality of the set of all the possible correlations between the p indicators, i.e. the cardinality of the set of all the possible combinations of 2 indicators among p, that is given by the binomial coefficient $\binom{p}{2} = \frac{p!}{2!(p-2)!} = \frac{p\cdot(p-1)\cdot(p-2)!}{2!(p-2)!} = \frac{p\cdot(p-1)}{2} \simeq p^2$. Therefore, describing our observations using covariates based on correlation, rather than networks, would make the complexity of our clustering problem increase quadratically with the number of indicators, while it is only linearly using a network representation.

6.5 Final remarks

This study explores clustering analysis when the statistical units are networks. To this extent, we focus on different methodologies that can provide a suitable representation of the sample of the networks for subsequent data analysis. Our exploration moves along two different directions. In the first case, we represent networks in terms of their topological characteristics in the node distance distribution. The distance among these representations can be evaluated using probability distribution distances, resulting in a matrix of pairwise distances between networks. In the second case, we use a whole network embedding approach, transforming the networks into a subspace of a fixed dimension. We applied fuzzy clustering algorithms in both cases, using standard metrics to evaluate their performance on synthetic and real datasets. Our analysis provides valuable hints for cluster analysis and highlights the pros and cons of the different obtained combinations. The present chapter is a first step in exploring these methodologies, which we believe will provide a path for further exploration and development of novel methodologies. For example, among the further developments, it may be of interest to apply a similar methodology to weighted and oriented networks.

Chapter 7

Final discussion

Nowadays, as the complexity of real-world phenomena increases, so does the need for new data structures and methodologies to handle that complexity. In particular, cluster analysis is one of the most popular and successful techniques for data exploration and characterization. In this thesis, it is proposed to contribute to current research on modeling multidimensional phenomena by introducing the application of fuzzy clustering techniques to complex data structures such as three-dimensional data and networks. In particular, new methodologies are presented for soft partitioning a set of units-by-variables matrices, a set of hierarchies, and a set of networks.

Chapter 1 provides a general introduction to data structures and the position of the problem. Theoretical motivations and real-world examples useful for giving concrete meaning to the problem are presented. Basic theoretical concepts and general notation are introduced and formalized in Chapter 2. Chapters 3, 4, 5, 6 are the main chapters that include the methodological proposals.

Chapter 3 focuses on finding a fuzzy partition of a set of hierarchies and for each class in the partition identifies a consensus hierarchy. Indeed, in cluster analysis, a problem often faced is to find a consensus on a set of hierarchical classifications of the same set of objects, called primary hierarchies (dendrograms). A unique consensus of the primary hierarchies, called the secondary hierarchy, is sufficient to synthesize relevant clustering information only when the primary dendrograms are similar to each other. In contrast, when the primary hierarchies change dramatically, a unique consensus of the entire set of primary hierarchies would be an overly unrealistic and narrow synthesis of the dendrograms. In these situations, when several differences between the original dendrograms are observed, more than one secondary consensus hierarchy is needed to clearly synthesize the different primary hierarchies. Furthermore, it can happen that different groups of dendrograms share some characteristics, and thus in such situations each original dendrogram must contribute to the definition of all consensus dendrograms. The required flexibility is provided by the use of a fuzzy approach to clustering. Our methodology, PARoDENo3WD (PARtition of DENdrograms of a 3-Way Data array), aims to obtain a secondary fuzzy partition of primary hierarchies, where hierarchies belonging to the same class are perceived as similar. Each class is associated with a consensus hierarchy. The fuzzy approach allows each primary hierarchy to contribute to the definition of all classes in the secondary partition, according to different degrees of membership. In this way, "clustering uncertainty" is taken into account. The performance of PARoDENo3WD was evaluated with an extended simulation study, generating 1800 three-way datasets. The results show that the methodology can identify both hard and fuzzy partitions of the dendrogram set

and also identify consensus dendrograms close to those originally generated. An application of PARoDENo3WD on a real data set is provided. In particular, the application to OECD data reporting economic indicators for G7 countries from 2005 to 2020 has proved extremely useful in obtaining consensus hierarchies of G7 countries and in identifying groups of years characterized by stability and, among these groups, years that correspond to periods of transition. The fuzzy approach in this application is extremely useful and its advantages are clearly highlighted. Indeed, the real application to the OECD countries confirms the ability of the proposed methodology to obtain a secondary fuzzy partition of the set of primary hierarchies capable of describing the uncertainty of the hierarchical relations of countries during years in which an economic shock was observed.

An extension of the proposal presented in Chapter 3 is provided in Chapter 4, which focuses on identifying a fuzzy partition of primary hieararchies and identifying, withing each class of the partition, a consensus hierarchy, characterized by the peculiar property of being parsimonious. More in details, a methodology for fitting a fuzzy partition and a parsimonious consensus hierarchy (ultrametric matrix) to a set of hierarchies of the same set of objects is described. The reason behind the use of parsimonious hierarchy lies in the fact that the complete sets of partitions and clusters of the dendrograms are not all used by investigators, even hindering interpretation, as noted by several authors (Gordon, 1999). One approach for resolving this difficulty has involved the construction of parsimonious trees that contain a limited number of internal nodes. In this way, some information is lost, but the main features of the data are more clearly represented. Therefore, in order to better study the partitioning of the set of dendrograms and identify the most relevant consensus parsimonious dendrograms we present a new methodology that identifies a secondary fuzzy partition of the original primary dendrograms and a parsimonious consensus dendrogram for each class of the secondary fuzzy partition. Each consensus includes an optimal consensus hard partition of objects and all agglomerative hierarchical aggregations among the clusters of the consensus partition. The performance of the methodology is illustrated by an extended simulation study and applications to real data. In more detail, the proposed methodology was tested in an extended simulation study, in which 1000 three-way ultrametric matrices were generated in two scenarios of hard and fuzzy assignment of primary hierarchies to consensus hierarchies. The study showed good results, not only in recovering the true underlying secondary partition, but also in identifying consensus parsimonious dendrograms very similar to the original ones. In addition, the results of applying the methodology to a real case show that the proposed methodology is useful for partitioning primary hierarchies in a fuzzy manner, correctly identifying hierarchies that share features with more than one cluster of the secondary partition: for example, in the application to the girls' growth curve dataset, two contiguous age periods were identified, and the hierarchies corresponding to two years of transition from one period to the next are assigned reasonably softly to both periods. Furthermore, for each class (period) of the fuzzy partition, the methodology identifies a consensus parsimonious dendrogram, which really facilitates the interpretation of the aggregation of girls.

When researchers are interested in comprehensively and statistically analyzing a collective phenomenon, a three-mode data matrix $\mathbf{X}$, where the modes are the units, the variables and the occasions, is the data structure to use properly. The occasions are often different times, so that the units can be observed in their natural complexity across a large set of variables, and also the history of the units can be followed and analyzed over time in the same descriptive statistical analysis. When $\mathbf{X}$ has a large dimension, it is important to synthesize information by identifying classes of similar occasions in which units are described by a reduced set of latent variables.

Within this framework, a simultaneous reduction of occasions and variables of $\mathbf{X}$ is proposed, and the methodological formulation and its application are provided in Chapter 5. It is assumed that the phenomenon being analyzed has a reduced set of views, each consisting of a set of occasions in which the units do not change the cross-sectional structure of the variables very much and thus the corresponding data matrices are perceived to be similar to each other. It is further assumed that the variables are correlated in these classes of occasions and a reduced set of correlated latent variables is identified. A fuzzy clustering of occasions allows the identification of K clusters of multivariate data matrices that are similar within the cluster. For each cluster, the closest data matrix that represents a consensus matrix of those in the cluster is identified. The variables in the cluster are correlated and retain their covariance structure, which can be summarized for each consensus matrix by applying a disjoint second-order factorial analysis. Then, the proposal allows for soft clustering of occasions into K clusters and for each consensus matrix to first identify a set of Q first-order factors and the corresponding loadings matrix and second identify a single general factor, which can be considered as the most synthetic indicator summarizing the original J variables. The proposed methodology is extremely useful when you have a set of multivariate observations of the same statistical units for a specific time period. In this case, the goal is to identify groups of years and, simultaneously, for each group, identify an overall composite indicator that is able to capture the differences between the units in those specific years by considering their composite indicators and corresponding ranking. Within this framework, the proposed methodology was applied to the well-being dataset, where its strength and usefulness were revealed. The How's Life-Well Being (HLWB) dataset measures several variables and reports data for several years and OECD countries. The fuzzy approach to clustering is very appropriate for this application: in fact, it allows the identification of groups of years of stability and years that correspond to periods of transition. In addition, with second-order DFA it was possible to identify K overall composite indicators and obtain a ranking of OECD countries. A multivariate partial ranking was obtained by applying a K-Means to the scores of the overall composite indicators to identify C groups of countries. The analysis of the identified country groups allowed a reasonable comparison of countries' policies and people's perceptions of health, wealth, and personal and social comfort.

In the end, as statistical network analysis finds application in an increasing number of disciplines, new methodologies are needed to handle such complexity. In this framework, Chapter 6 focuses on how to represent sets of networks for fuzzy clustering. In detail, computational procedures are provided to identify clusters of networks, where each network represents an object, based on fuzzy clustering algorithms, particularly Non-Euclidean Fuzzy Relational Clustering (NEFRC), Fuzzy Clustering Analysis (FANNY) and Fuzzy k-Means (FkM) algorithms with their variants (such as Fuzzy k-Means with polynomial fuzzifier, Fuzzy k-Means based on L1 metric, Fuzzy k-Medoids). Clustering algorithms are applied to networks represented through probability distributions of their topological properties or through vector representations derived from integer graph embedding methods. In particular, we explore three different network representations based on probability distributions, autoencoder and joint embedding. We verify the suitability and discuss the characteristics of the algorithms through simulations and real case studies. In terms of real applications, our methodology has proven extremely useful in finding groups of airlines with similar flight route structures and groups of countries characterized by similar correlations between environmental variables. Our analysis provides valuable insights for cluster analysis and highlights the pros and cons of

the different combinations obtained.

To conclude, what is proposed in this dissertation aims to contribute to the research of multidimensional phenomena, which are evolving not only in dimension but also in complexity. The growing dimensions and complexity of real-world problems makes the use of complex data structures necessary and the need of novel methodology to handle such kind of data urgent. The proposed methodologies have been tested on several real-world cases and the results are promising. The results of the application to real data sets emphasize and reinforce the importance of using the fuzzy approach to clustering, as it is extremely useful in obtaining meaningful considerations of clustering interpretation. As the history of scientific, economic and sociological events teaches us, multidimensional and multiyear phenomena develop in a period composed of stable and transitional sub-periods, and the proposed methodology allows us also to identify this peculiarity.

Bibliography

Abu Arqub, O., Singh, J., and Alhodaly, M. (2021). "Adaptation of kernel functions-based approach with Atangana–Baleanu–Caputo distributed order derivative for solutions of fuzzy fractional Volterra and Fredholm integrodifferential equations". *Mathematical Methods in the Applied Sciences*.

Abu Arqub, O., Singh, J., Maayah, B., and Alhodaly, M. (2021). "Reproducing kernel approach for numerical solutions of fuzzy fractional initial value problems under the Mittag–Leffler kernel differential operator". *Mathematical Methods in the Applied Sciences*.

Asur, S., Ucar, D., and Parthasarathy, S. (2007). "An ensemble framework for clustering protein–protein interaction networks". *Bioinformatics* 23.13, i29–i40.

Australian Government, Department of Environment and Energy (2019). *Australia's Fourth Biennial Report*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/Australia%20Fourth%20Biennial%20Report.pdf.

Bavelas, A. (1950). "Communication patterns in task-oriented groups". *The journal of the acoustical society of America* 22.6, 725–730.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithm*. Plenum Press, New York.

— (1987). "Some non-standard clustering algorithms". In: *Develoments in Numerical Ecology*. Springer, pp. 225–287.

Bhatia, V. and Rani, R. (2017). "A parallel fuzzy clustering algorithm for large graphs using Pregel". *Expert Systems with Applications* 78, 135–144.

Bickel, S. and Scheffer, T. (2004). "Multi-view clustering." In: *International Conference on Data Mining*. Vol. 4. 2004. Citeseer, pp. 19–26.

Bocci, L. and Vicari, D. (2019). "Rootclus: Searching for "root clusters" in three-way proximity data". *psychometrika* 84.4, 941–985.

Bombelli, I., Ferraro, M. B., and Vichi, M. (2023). "Consensus and fuzzy partition of dendrograms from a three-way dissimilarity array". *Information Sciences* 637, 118948. DOI: https://doi.org/10.1016/j.ins.2023.118948.

Bombelli, I., Manipur, I., Guarracino, M. R., and Ferraro, M. B. (2023). "Representing ensembles of networks for fuzzy cluster analysis: a case study". *Data Mining and Knowledge Discovery*. DOI: https://doi.org/10.1007/s10618-023-00977-x.

Bonacich, P. (1987). "Power and centrality: A family of measures". *American journal of sociology* 92.5, 1170–1182.

Borchers, H. W. (2021). *pracma: Practical Numerical Math Functions*. R package version 2.3.3. URL: https://CRAN.R-project.org/package=pracma.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*. Vol. 50. Cambridge University Press.

Brandes, U. (2001). "A faster algorithm for betweenness centrality". *The Journal of Mathematical Sociology* 25.2, 163–177.

Brandes, U., Lerner, J., and Nagel, U. (2011). "Network ensemble clustering using latent roles". *Advances in Data Analysis and Classification* 5, 81–94.

Breehl, L. and Caban, O. (2021). "Physiology, puberty". In: *StatPearls [Internet]*. StatPearls Publishing.

Caliński, T. and Harabasz, J. (1974). "A dendrite method for cluster analysis". *Communications in Statistics-theory and Methods* 3.1, 1–27.

Campello, R. J. (2007). "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment". *Pattern Recognition Letters* 28.7, 833–841.

Campello, R. J. and Hruschka, E. R. (2006a). "A fuzzy extension of the silhouette width criterion for cluster analysis". *Fuzzy Sets and Systems* 157.21, 2858–2875.

— (2006b). "A fuzzy extension of the silhouette width criterion for cluster analysis". *Fuzzy Sets and Systems* 157.21, 2858–2875.

Cappozzo, A., Alessandro, C., Michael, F., et al. (2021). "Penalized model-based clustering for three-way data structures". In: *SIS 2021: Book of Short Papers*. Pearson, pp. 758–763.

Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F., and Boccaletti, S. (2013). "Emergence of network features from multiplexity". *Scientific reports* 3.1, 1–6.

Cariou, V., Alexandre-Gouabau, M.-C., and Wilderjans, T. F. (2021). "Three-way clustering around latent variables approach with constraints on the configurations to facilitate interpretation". *Journal of Chemometrics* 35.2, e3269.

Carpi, L. C., Schieber, T. A., Pardalos, P. M., Marfany, G., Masoller, C., Díaz-Guilera, A., and Ravetti, M. G. (2019). "Assessing diversity in multiplex networks". *Scientific reports* 9.1, 1–12.

Cattell, R. B. (1966). "The scree test for the number of factors". *Multivariate behavioral research* 1.2, 245–276.

Cavicchia, C. and Vichi, M. (2022). "Second-order disjoint factor analysis". *psychometrika* 87.1, 289–309.

Chao, G., Sun, J., Lu, J., Wang, A.-L., Langleben, D. D., Li, C.-S., and Bi, J. (2019). "Multi-view cluster analysis with incomplete data to understand treatment effects". *Information sciences* 494, 278–293.

Coppi, R. and Bolasco, S. (1989). *Multiway data analysis*. North-Holland Publishing Co.

Csardi, G. and Nepusz, T. (2006). "The igraph software package for complex network research". *InterJournal* Complex Systems, 1695.

Davé, R. N. and Sen, S. (2002). "Robust fuzzy clustering of relational data". *IEEE Transactions on Fuzzy Systems* 10.6, 713–727.

De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (Apr. 2015). "Structural reducibility of multilayer networks". *Nature Communications* 6, 6864. DOI: 10.1038/ncomms7864.

Department of Forestry, Fisheries and the Environment (2021). *SOUTH AFRICA'S 4TH BIENNIAL UPDATE REPORT TO THE UNITED NATIONS FRAMEWORK CONVENTION ON CLIMATE CHANGE*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/South%20Africa%20BUR4%20to%20the%20UNFCCC.pdf.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford university press.

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). "An experimental study of search in global social networks". *science* 301.5634, 827–829.

Donaldson, J. (2016). *tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE)*. R package version 0.1-3.

Dunn, J. C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". *Journal of Cybernetics* 3.3, 32–57.
— (1974). "Well-separated clusters and optimal fuzzy partitions". *Journal of cybernetics* 4.1, 95–104.
Durieux, J. and Wilderjans, T. F. (2019). "Partitioning subjects based on high-dimensional fMRI data: comparison of several clustering methods and studying the influence of ICA data reduction in big data". *Behaviormetrika* 46.2, 271–311.
Duroux, D. and Van Steen, K. (2023). "netANOVA: novel graph clustering technique with significance assessment via hierarchical ANOVA". *Briefings in Bioinformatics* 24.2, bbad029.
D'Angelo, G. and Palmieri, F. (2021). "GGA: A modified genetic algorithm with gradient-based local search for solving constrained optimization problems". *Information Sciences* 547, 136–162.
El Hajjar, S., Dornaika, F., and Abdallah, F. (2022). "One-step multi-view spectral clustering with cluster label correlation graph". *Information Sciences* 592, 97–111.
Environment and Climate Change Canada (2020). *CANADA'S FOURTH BIENNIAL REPORT ON CLIMATE CHANGE*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/br4_final_en.pdf.
European Commission (2020). *Second Biennial Report of the European Union under the UN Framework Convention on Climate Change*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/European%20Union_second_biennial_report_under_the_unfccc_%282%29.pdf.
Farello, G., Altieri, C., Cutini, M., Pozzobon, G., and Verrotti, A. (2019). "Review of the literature on current changes in the timing of pubertal development and the incomplete forms of early puberty". *Frontiers in pediatrics* 7, 147.
Ferraro, M., Giordani, P., and Serafini, A. (2019). "fclust: An R Package for Fuzzy Clustering". *The R Journal* 11.
Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). "Sur la liaison et la division des points d'un ensemble fini". In: *Colloquium mathematicum*. Vol. 2. 3-4, pp. 282–285.
Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., et al. (2021). "Is there a role for statistics in artificial intelligence?" *Advances in Data Analysis and Classification*, 1–24.
Fu, L., Lin, P., Vasilakos, A. V., and Wang, S. (2020). "An overview of recent multi-view clustering". *Neurocomputing* 402, 148–161.
Fuglede, B. and Topsoe, F. (2004a). "Jensen-Shannon divergence and Hilbert space embedding". In: *Proceedings of International Symposium on Information Theory, ISIT 2004*, pp. 31–.
— (2004b). "Jensen-Shannon divergence and Hilbert space embedding". In: *Proceedings of International Symposium on Information Theory, ISIT 2004*, pp. 31–.
— (2004c). "Jensen-Shannon divergence and Hilbert space embedding". In: *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.* IEEE, p. 31.
Ghosh, J. and Acharya, A. (2011). "Cluster ensembles". *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 1.4, 305–315.
Gordon, A. (1999). *Classification, 2nd Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press. ISBN: 9781584888536. URL: https://books.google.it/books?id=_w5AJtbfEz4C.
Granata, I., Guarracino, M. R., Kalyagin, V. A., Maddalena, L., Manipur, I., and Pardalos, P. M. (2018). "Supervised classification of metabolic networks". In:

2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 2688–2693.

Granata, I., Guarracino, M. R., Kalyagin, V. A., Maddalena, L., Manipur, I., and Pardalos, P. M. (2020a). "Model simplification for supervised classification of metabolic networks". *Annals of Mathematics and Artificial Intelligence* 88.1, 91–104.

Granata, I., Guarracino, M. R., Maddalena, L., and Manipur, I. (2020b). "Network Distances for Weighted Digraphs". In: *International Conference on Mathematical Optimization Theory and Operations Research*. Springer, pp. 389–408.

Grazioli, G., Martin, R. W., and Butts, C. T. (2019). "Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods". *Frontiers in molecular biosciences* 6, 42.

Guare, J., Sandrich, J., and Loewenberg, S. A. (2000). *Six degrees of separation*. LA Theatre Works.

Gutiérrez-Gómez, L. and Delvenne, J.-C. (2019). "Unsupervised network embeddings with node identity awareness". *Applied Network Science* 4.1, 82.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hathaway, R. J. and Bezdek, J. C. (1994). "NERF c-means: Non-Euclidean relational fuzzy clustering". *Pattern recognition* 27.3, 429–437.

Havens, T. C., Bezdek, J. C., Leckie, C., Chan, J., Liu, W., Bailey, J., Ramamohanarao, K., and Palaniswami, M. (2013). "Clustering and visualization of fuzzy communities in social networks". In: *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, pp. 1–7.

Heckerman, D. (1997). "Bayesian networks for data mining". *Data mining and knowledge discovery* 1.1, 79–119.

Hubert, L. and Arabie, P. (1985). "Comparing partitions". *Journal of classification* 2.1, 193–218.

Hussain, S. F., Mushtaq, M., and Halim, Z. (2014). "Multi-view document clustering via ensemble method". *Journal of Intelligent Information Systems* 43.1, 81–99.

Institution, B. S. (1994). *Accuracy (trueness and precision) of measurment methods and results: general principals and definitions*.

Jajuga, K. (1991). "L1-norm based fuzzy clustering". *Fuzzy Sets and Systems* 39.1, 43–50.

Jiang, M., Cui, P., Beutel, A., Faloutsos, C., and Yang, S. (2014). "Inferring strange behavior from connectivity pattern in social networks". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 126–138.

Johnson, S. C. (1967). "Hierarchical clustering schemes". *Psychometrika* 32.3, 241–254.

Khan, G. A., Hu, J., Li, T., Diallo, B., and Wang, H. (2022). "Multi-view data clustering via non-negative matrix factorization with manifold regularization". *International Journal of Machine Learning and Cybernetics*, 1–13.

Klawonn, F. and Höppner, F. (2003). "What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier". In: *International symposium on intelligent data analysis*. Springer, pp. 254–264.

Klawonn, F., Kruse, R., and Winkler, R. (2015). "Fuzzy clustering: More than just fuzzification". *Fuzzy sets and systems* 281, 272–279.

Krijthe, J. H. (2015). "Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation". *R package version 0.13, URL https://github.com/jkrijthe/Rtsne*.

Krishnapuram, R., Joshi, A., Nasraoui, O., and Yi, L. (2001). "Low-complexity fuzzy relational clustering algorithms for web mining". *IEEE transactions on Fuzzy Systems* 9.4, 595–607.

Křivánek, M. and Morávek, J. (1986). "NP-hard problems in hierarchical-tree clustering". *Acta informatica* 23.3, 311–323.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Vol. 702. John Wiley & Sons.

Kroonenberg, P. M., Janssen, J, Marcotorchino, F, and Proth, J. (1987). "Multivariate and longitudinal data on growing children. Solutions using a three-mode principal component analysis and some comparison results with other approaches". *Data analysis. The ins and outs of solving real problems*, 89–112.

Kuhn, M. (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Kullback, S. (1997a). *Information theory and statistics*. Courier Corporation.

— (1997b). *Information theory and statistics*. Courier Corporation.

Kullback, S. and Leibler, R. A. (1951a). "On information and sufficiency". *The annals of mathematical statistics* 22.1, 79–86.

— (1951b). "On information and sufficiency". *The annals of mathematical statistics* 22.1, 79–86.

Lancichinetti, A. and Fortunato, S. (2012). "Consensus clustering in complex networks". *Scientific reports* 2.1, 1–7.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). "Benchmark graphs for testing community detection algorithms". *Physical review E* 78.4, 046110.

Langari, R. K., Sardar, S., Mousavi, S. A. A., and Radfar, R. (2020). "Combined fuzzy clustering and firefly algorithm for privacy preserving in social networks". *Expert Systems with Applications* 141, 112968.

Lapointe, F.-J. and Cucumel, G. (1997). "The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa". *Systematic Biology* 46.2, 306–312.

Lloyd, S. (1982a). "Least squares quantization in PCM". *IEEE transactions on information theory* 28.2, 129–137.

— (1982b). "Least squares quantization in PCM". *IEEE transactions on information theory* 28.2, 129–137.

Lu, H., Liu, S., Wei, H., and Tu, J. (2020). "Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network". *Expert Systems with Applications* 159, 113513.

MacQueen, J (1967a). "Classification and analysis of multivariate observations". In: *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, pp. 281–297.

MacQueen, J. (1967b). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 281–297.

Madaan, H. and Gosain, A. (2022). "Prioritized dynamic cube selection in data warehouse". *Multimedia Tools and Applications*, 1–23.

Manipur, I., Granata, I., Maddalena, L., and Guarracino, M. R. (2020a). "Clustering analysis of tumor metabolic networks". *BMC bioinformatics* 21.10, 1–14.

Manipur, I., Granata, I., Maddalena, L., and Guarracino, M. R. (2020b). "Clustering analysis of tumor metabolic networks". *BMC Bioinformatics* 21.10, 349.

MATLAB (2021). *MATLAB version 9.11.0.1769968 (R2021b)*. The Mathworks, Inc. Natick, Massachusetts.

McQuitty, L. L. (1960). "Hierarchical linkage analysis for the isolation of types". *Educational and Psychological Measurement* 20.1, 55–67.

Milgram, S. (1967). "The small world problem". *Psychology today* 2.1, 60–67.

Minister of Environment of Chile (2018). *CHILE´S THIRD BIENNIAL UPDATE REPORT To the United Nations Framework Convention on Climate Change*.

Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/5769410_Chile-BUR3-1-Chile_3BUR_English.pdf.

Ministry of Climite Change Government of Pakistan (2022). *PAKISTAN'S FIRST BIENNIAL UPDATE REPORT (BUR-1) TO THE UNITED NATIONS FRAMEWORK CONVENTION ON CLIMATE CHANGE (UNFCCC)*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/Pakistan%E2%80%99s%20First%20Biennial%20Update%20Report%20%28BUR-1%29%20-%202022.pdf.

Ministry of Energy, Industry and Mineral Resources, Kingdom of Saudi Arabia (2018). *The First Biennial Update Report*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/18734625_Saudi%20Arabia-BUR1-1-BUR1-Kingdom%20of%20Saudi%20Arabia.pdf.

Ministry of Environment, Egyptian Environmental Affairs Agency (2018). *EGYPT'S FIRST BIENNIAL UPDATE REPORT to the UNITED NATIONS FRAMEWORK CONVENTION ON CLIMATE CHANGE*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/BUR%20Egypt%20EN.pdf.

Ministry of Environment Forest and Climate Change Government of India (2021). *India, Third Biennial Update Report to The United Nations Framework Convention on Climate Change*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/INDIA_%20BUR-3_20.02.2021_High.pdf.

Ministry of Foreign Affairs, Ministry of Science, Technology and Innovations (2020). *FOURTH BIENNIAL UPDATE REPORT OF BRAZIL TO THE UNITED NATIONS FRAMEWORK CONVENTION ON CLIMATE CHANGE*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/BUR4.Brazil.pdf.

Moody, J. and Mucha, P. J. (2013). "Portrait of political party polarization1". *Network Science* 1.1, 119–121.

Muller, E., Gunnemann, S., Farber, I., and Seidl, T. (2012). "Discovering multiple clustering solutions: Grouping objects in different views of the data". In: *2012 IEEE 28th international conference on data engineering*. IEEE, pp. 1207–1210.

Nawaz, M. and Yan, H. (2020). "Saliency detection via multiple-morphological and superpixel based fast fuzzy C-mean clustering network". *Expert Systems with Applications* 161, 113654.

Ni, J., Cheng, W., Fan, W., and Zhang, X. (2017). "ComClus: A self-grouping framework for multi-network clustering". *IEEE transactions on knowledge and data engineering* 30.3, 435–448.

Obando, C. and de Vico Fallani, F. (2017). "A statistical model for brain networks inferred from large-scale electrophysiological signals". *Journal of The Royal Society Interface* 14.128, 20160940.

Ou-Yang, L., Yan, H., and Zhang, X.-F. (2017). "A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks". *BMC bioinformatics* 18.13, 23–34.

Petersen, K. B., Pedersen, M. S., et al. (2008). "The matrix cookbook". *Technical University of Denmark* 7.15, 510.

Powell, M. J. (1983). "Variable metric methods for constrained optimization". In: *Mathematical programming the state of the art*. Springer, pp. 288–311.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). "Defining and identifying communities in networks". *Proceedings of the national academy of sciences* 101.9, 2658–2663.

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical association* 66.336, 846–850.

Rosenberg, S. and Kim, M. P. (1975). "The method of sorting as a data-gathering procedure in multivariate research". *Multivariate behavioral research* 10.4, 489–502.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of computational and applied mathematics* 20, 53–65.

Rousseeuw, P. J. and Kaufman, L (1990). "Finding groups in data". *Hoboken: Wiley Online Library* 1.

Rubin, J. (1967). "Optimal classification into groups: an approach for solving the taxonomy problem". *Journal of Theoretical Biology* 15.1, 103–144.

Runkler, T. A. and Bezdek, J. C. (2013). "Fuzzy relational approaches to graph clustering and visualization". In: *Proc. GMA/GI Workshop Computational Intelligence, Dortmund*, pp. 39–56.

Runkler, T. A. and Ravindra, V. (2015). "Fuzzy Graph Clustering based on Non-Euclidean Relational Fuzzy c-Means". In: *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*. Atlantis Press.

Ruspini, E. H. (1969a). "A new approach to clustering". *Information and control* 15.1, 22–32.

— (1969b). "A new approach to clustering". *Information and control* 15.1, 22–32.

Sato, M. and Sato, Y. (1994). "On a multicriteria fuzzy clustering method for 3-way data". *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 2.02, 127–142.

Schieber, T. A., Carpi, L., Díaz-Guilera, A., Pardalos, P. M., Masoller, C., and Ravetti, M. G. (2017a). "Quantification of network structural dissimilarities". *Nature communications* 8.1, 1–10.

— (2017b). "Quantification of network structural dissimilarities". *Nature communications* 8.1, 13928.

Schoonees, P. C., Groenen, P. J., and van de Velden, M. (2021). "Least-squares bilinear clustering of three-way data". *Advances in Data Analysis and Classification*, 1–37.

Sempé, M. and Médico-Sociale, G. d. (1987). "Presentation of the French Auxological Survey". In: *Data Analysis*. Springer, pp. 3–6.

Sharma, D. and Jabeen, S. D. (2023). "Hybridizing interval method with a heuristic for solving real-world constrained engineering optimization problems". In: *Structures*. Vol. 56. Elsevier, p. 104993.

Simpson, S. L., Hayasaka, S., and Laurienti, P. J. (2011). "Exponential random graph modeling for complex brain networks". *PloS one* 6.5, e20039.

Slaughter, A. J. and Koehly, L. M. (2016). "Multilevel models for social networks: Hierarchical Bayesian approaches to exponential random graph modeling". *Social networks* 44, 334–345.

Sokal, R. R. (1958). "A statistical method for evaluating systematic relationships." *Univ. Kansas, Sci. Bull.* 38, 1409–1438.

Sokal, R. R. and Rohlf, F. J. (1981). "Taxonomic congruence in the Leptopodomorpha re-examined". *Systematic Zoology* 30.3, 309–325.

Song, J., Wang, H., and Song, M. J. (2022). "Package 'Ckmeans. 1d. dp'".

Song, M. and Zhong, H. (2020). "Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers". *Bioinformatics* 36.20, 5027–5036.
Stanley, R. P. (1997). "Enumerative combinatorics. Vol 2". *Cambridge Studies in Advanced Mathematics* 62, 297.
Tagarelli, A., Amelio, A., and Gullo, F. (2017). "Ensemble-based community detection in multilayer networks". *Data Mining and Knowledge Discovery* 31.5, 1506–1543.
Tang, L. and Liu, H. (2011). "Leveraging social media networks for classification". *Data Mining and Knowledge Discovery* 23.3, 447–478.
Tantardini, M., Ieva, F., Tajoli, L., and Piccardi, C. (2019). "Comparing methods for comparing networks". *Scientific reports* 9.1, 1–19.
TGA (n.d.). *The Cancer Genome Atlas*. URL: https://tcga-data.nci.nih.gov.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). "Missing value estimation methods for DNA microarrays". *Bioinformatics* 17.6, 520–525.
United Nations Climate Change (2022). *Output Report of Africa Climate Week 2022*. Accessed = 2022-11-30. URL: https://unfccc.int/sites/default/files/resource/ACW2022_OutputReport_10102022.pdf.
United Nations Environment Programme (2019). *Emissions gap report 2019*. Accessed = 2022-11-30. URL: https://wedocs.unep.org/bitstream/handle/20.500.11822/30797/EGR2019.pdf?sequence=1&isAllowed=y.
United Nations International Trade Statistics (2017). *Harmonized Commodity Description and Coding Systems (HS)*. https://unstats.un.org/unsd/tradekb/Knowledgebase/50018/Harmonized-Commodity-Description-and-Coding-Systems-HS. Online; accessed 18 April 2021.
Van der Maaten, L. and Hinton, G. (2008). "Visualizing data using t-SNE." *Journal of machine learning research* 9.11.
Van Der Maaten, L. (2013). "Barnes-hut-sne". *arXiv preprint arXiv:1301.3342*.
— (2014). "Accelerating t-SNE using tree-based algorithms". *The journal of machine learning research* 15.1, 3221–3245.
Vathy-Fogarassy, Á. and Abonyi, J. (2013). *Graph-based clustering and data visualization algorithms*. Springer.
Vicari, D. and Vichi, M. (2009). "Structural classification analysis of three-way dissimilarity data". *Journal of classification* 26.2, 121–154.
Vichi, M. (1993). "Un algoritmo dei minimi quadrati per interpolare un insieme di classificazioni gerarchiche con una classificazione consenso". *Metron* 51.3–4, 139–163.
— (1998). "Principal classifications analysis: a method for generating consensus dendrograms and its application to three-way data". *Computational statistics & data analysis* 27.3, 311–331.
— (1999). "One-mode classification of a three-way data matrix". *Journal of Classification* 16.1, 27–44.
— (2008). "Fitting semiparametric clustering models to dissimilarity data". *Advances in Data Analysis and Classification* 2.2, 121–161.
— (2017). "Disjoint factor analysis with cross-loadings". *Advances in Data Analysis and Classification* 11.3, 563–591.
Vichi, M., Cavicchia, C., and Groenen, P. J. (2022). "Hierarchical Means Clustering". *Journal of Classification* 39.3, 553–577.
Vinh, N., Epps, J, and Bailey, J (2010). "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". *Journal of Machine Learning Research* 11, 2837–2854.

Wagner, S. and Wagner, D. (2007). *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Wang, H. and Song, M. (2011). "Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming". *The R journal* 3.2, 29.

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2021). "Joint embedding of graphs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1324–1336.

Wang, Y., Ma, X., Lao, Y., and Wang, Y. (2014). "A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization". *Expert systems with applications* 41.2, 521–534.

Ward Jr, J. H. (1963). "Hierarchical grouping to optimize an objective function". *Journal of the American statistical association* 58.301, 236–244.

Wen, J., Zhang, Z., Fei, L., Zhang, B., Xu, Y., Zhang, Z., and Li, J. (2022). "A survey on incomplete multiview clustering". *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Whaley, A. L. and Longoria, R. A. (2009). "Preparing card sort data for multidimensional scaling analysis in social psychological research: a methodological approach". *The Journal of social psychology* 149.1, 105–115.

World Bank Group (2022a). *China, COUNTRY CLIMATE AND DEVELOPMENT REPORT*. Accessed = 2022-11-30. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/38136/FullReport.pdf.

— (2022b). *COUNTRY CLIMATE AND DEVELOPMENT REPORT: ARGENTINA*. Accessed = 2022-11-30. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/38252/ARG_CCDR_FullReport.pdf?sequence=6&isAllowed=y.

— (2022c). *COUNTRY CLIMATE AND DEVELOPMENT REPORT: Perù*. Accessed = 2022-11-30. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/38251/EnglishReport.pdf?sequence=2&isAllowed=y.

Wright, S. J. and Nocedal, J. (2006). *Numerical optimization*. Springer New York, NY.

Xie, X. L. and Beni, G. (1991). "A validity measure for fuzzy clustering". *IEEE Transactions on pattern analysis and machine intelligence* 13.8, 841–847.

Yağ, I. and Altan, A. (2022). "Artificial Intelligence-Based Robust Hybrid Algorithm Design and Implementation for Real-Time Detection of Plant Diseases in Agricultural Environments". *Biology* 11.12, 1732.

Yang, M.-S. and Hussain, I. (2023). "Unsupervised multi-view K-means clustering algorithm". *IEEE Access* 11, 13574–13593.

Yang, M., Li, Y., Hu, P., Bai, J., Lv, J., and Peng, X. (2022). "Robust multi-view clustering with incomplete information". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1, 1055–1069.

Yang, X., Liu, J., Cheung, W. K. W., and Zhou, X.-N. (2014). "Inferring metapopulation based disease transmission networks". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 385–399.

Yang, Z., Liang, N., Yan, W., Li, Z., and Xie, S. (2020). "Uniform distribution non-negative matrix factorization for multiview clustering". *IEEE transactions on cybernetics* 51.6, 3249–3262.

Yin, F., Shen, W., and Butts, C. T. (2022). "Finite Mixtures of ERGMs for Modeling Ensembles of Networks". *Bayesian Analysis* 1.1, 1–39.

Yin, Q., Wu, S., He, R., and Wang, L. (2015). "Multi-view clustering via pairwise sparse subspace representation". *Neurocomputing* 156, 12–21.

Yu, H., Wang, X., Wang, G., and Zeng, X. (2020). "An active three-way clustering method via low-rank matrices for multi-view data". *Information Sciences* 507, 823–839.

Zaidi, F. (2012). "Fuzzy Clustering and Visualization of Information for Web Search Results". *Journal of Internet Technology* 13, 939–952.

Zhao, J., Kang, F., Zou, Q., and Wang, X. (2023). "Multi-view clustering with orthogonal mapping and binary graph". *Expert Systems with Applications* 213, 118911.

Zhao, L., Ma, Y., Chen, S., and Zhou, J. (2022). "Multi-view co-clustering with multi-similarity". *Applied Intelligence*, 1–12.