SAPIENZA
UNIVERSITÀ DI ROMA

Ultrametric models for hierarchical dimensionality reduction

Scuola di Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXXIV Ciclo

Candidate

Giorgia Zaccaria

ID number 1601319

Thesis Advisor

Prof. Maurizio Vichi

Thesis submitted in October 2021

Thesis defended on February 22, 2022
in front of a Board of Examiners composed by:

Prof. Salvatore Ingrassia (chairman)

Prof.ssa Maura Mezzetti

Prof. Domenico Vistocco

Ultrametric models for hierarchical dimensionality reduction
Ph.D. thesis. Sapienza – University of Rome

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

Version: January 31, 2022

Author's email: giorgia.zaccaria@uniroma1.it

Alla mia guerriera.

Ringraziamenti

"Nel dubbio lo stolto vede un problema, il saggio una opportunità"

Antico Proverbio Cinese

Dedico questo lavoro a tutte le persone che hanno contribuito alla mia crescita personale e professionale. A chi ha saputo comprendere le difficoltà di questo percorso (soprattutto quelle organizzative!) con paziente attesa. Grazie.

Al mio supervisore, il Prof. Maurizio Vichi, che mi ha introdotto nel mondo della ricerca trasmettendomi la passione e la gioia per questo lavoro. Grazie per i consigli, i confronti metodologici ed il costante supporto professionale e umano.

Al mio collega e amico Carlo, per avermi accolta in stanza 41 quando ancora ero una studentessa poco consapevole del mio futuro, per aver costruito con me una solida collaborazione lavorativa, per tutti i pomeriggi di lavoro passati alla lavagna a fare dimostrazioni intervallati da una chiacchierata e per i preziosi suggerimenti. Ma prima di tutto questo, per essere un amico fidato.

A tutti i Professori del DSS con cui ho avuto modo di confrontarmi in questi anni. In particolare, al Prof. Marco Alfò e alla Prof.ssa Francesca Martella per i vostri consigli e le chiacchierate sul significato più profondo di questo lavoro. A tutti i dottorandi dei vecchi e dei nuovi cicli: questo anno e mezzo così difficile ci ha privati di spazi e momenti di condivisione, ma porto con me i ricordi di tutti quelli precedenti.

A Francesca, anche detta "Penta", per la sopportazione di sfoghi in note interminabili, per aver condiviso gli eventi di un percorso tanto impegnativo quanto meraviglioso, per essere un punto di riferimento.

A Lucia Colognato, che mi ha insegnato a perseguire i miei obiettivi con determinazione e sacrificio senza scoraggiarmi di fronte alle difficoltà.

Agli amici di sempre, Emilia, Alex e Simone, con i quali tante volte mi sono confrontata su questo percorso e sul futuro. A voi, che ci siete sempre stati, nonostante le distanze che ad oggi ci dividono tra Roma, Parigi e Bonn. Vi auguro di intraprendere la strada che più vi renda soddisfatti, nella speranza di convergere prima o poi su un unico punto della cartina geografica.

Alla mia famiglia, tutta, va un ringraziamento speciale.

A mia madre, che ha saputo rimanere attaccata alla vita nei momenti più difficili. Senza la tua tenacia non avrei avuto la forza di portare a compimento questa tesi.

A mio padre, per il costante supporto e confronto, per la pazienza e lo sforzo nel voler sempre comprendere le stranezze di questo lavoro, per essermi sempre vicino.

A Roberta, per l'incoraggiamento in ogni situazione, per l'affetto costantemente dimostrato e per cercare di "indicarmi la retta via anche quando parto per la tangente".

Ad Antonio, che più di tutti ha vissuto e condiviso con me le gioie e le soddisfazioni, ma anche la fatica e i turbamenti di questo percorso. Grazie per esserci sempre stato. Questa tesi e gli sforzi che ha richiesto sono anche per te, per noi.

Contents

1 Introduction **1**
1.1 Hierarchical structures on variables: an overview 3
1.2 Ultrametricity 5
1.3 Reflective and formative models 7
1.4 Chapter summaries 10

2 The ultrametric correlation matrix for modeling hierarchical latent concepts **12**
2.1 Introduction 12
2.2 Notation and basic notions 13
2.3 The model 15
2.4 Least-Squares estimation of the model and algorithm 18
2.4.1 Algorithm for detecting the LS Ultrametric Correlation Model 21
2.5 Simulation 21
2.6 Application 23
2.6.1 Bechtoldt data set 23
2.6.2 Drug consumption data set 24
2.7 Conclusions 27

3 Exploring hierarchical concepts: theoretical and application comparisons **29**
3.1 Introduction 29
3.2 Hierarchical classification of variables 30
3.3 The Ultrametric Correlation Model: a brief review 33
3.4 A comparison between the Ultrametric Correlation Model and the agglomerative clustering algorithms 34
3.5 Conclusions 39

4 Gaussian Mixture Model with an extended ultrametric covariance structure **40**
4.1 Introduction 40
4.2 Notation and theoretical background 43
4.3 Extended Ultrametric Covariance Structure 45
4.4 Gaussian Mixture Model with an Extended Ultrametric Covariance Structure 48
4.4.1 GMMEUCovS algorithm 51

4.4.2 Model selection 53
4.5 Simulation 53
4.6 Application 58
4.6.1 Well-Being Indicators data set 58
4.6.2 Coffee data set 59
4.7 Conclusions 61

5 Hierarchical Disjoint Principal Component Analysis 63
5.1 Introduction 63
5.2 Notation 66
5.3 Hierarchical Disjoint Principal Component Analysis 67
5.3.1 Least-squares estimation of HierDPCA 70
5.3.2 Coordinate descent algorithm for HierDPCA 74
5.4 Simulation study 74
5.5 Application 80
5.5.1 Big Five Personality Test 80
5.5.2 ASEM Connectivity Sustainability Index 84
5.6 Conclusions 86

6 Discussion 88
6.1 Further developments 90

A Appendix to Chapter 2
Relationship between $\mathbf{R_W}$, $\mathbf{R_B}$ and the Cronbach's α 92

B Appendix to Chapter 4
Maximum likelihood estimates of the GMMEUCovS covariance structure 94

C Appendix to Chapter 5
Proofs 97

Bibliography 99

List of Figures

1.1 Examples of an ultrametric distance matrix and an ultrametric matrix. 6
1.2 Examples of the measurement model. 8

2.1 Relationship between a $(2Q-1)$-ultrametric correlation matrix and a corresponding hierarchy of latent concepts. 17
2.2 Example of the heat maps of three (30×30) correlation matrices produced by the simulation study with different levels of error. The theoretical number of groups of variables is $Q = 7$ (Scenario 1). . . . 22
2.3 Comparison between heat maps of the observed and estimated correlation matrices. 24
2.4 Path diagram of the Bechtoldt hierarchical structure. 25
2.5 Drug consumption data set. 26
2.6 Path diagram representation of the drug consumption. 27

3.1 N-tree representation: root node (a), internal nodes (b, c, d, e, f, g), terminal nodes (h, i, j, k, l, m, n, o). 30
3.2 Heatmap of the Holzinger (14×14) correlation matrix of ability tests. 34
3.3 Dendrogram of the ultrametric distance matrix obtained by computing the Ultrametric Correlation Model on the Holzinger (14×14) correlation matrix of ability tests and applying Eq. (3.1) on the result. 36
3.4 Dedrogram of (3.4a) Single Linkage, (3.4b) Complete Linkage, (3.4c) Average Linkage, (3.4d) Ward's Method on the distance matrix obtained by transforming the Holzinger (14×14) correlation matrix of ability tests according to Eq. (3.1). 38

4.1 Examples of the extended ultrametric covariance matrices (4.1a)-(4.1c) and the corresponding path diagrams (4.1d)-(4.1f) representing different hierarchical relationships among nested concepts. 42
4.2 Relationship between EUCovS and the corresponding hierarchy of variable groups. 47
4.3 GMMEUCovS variable hierarchies for each cluster of countries. . . . 60
4.4 GMMEUCovS variable hierarchies for each cluster of coffee data, i.e., Arabica and Robusta. 61

5.1 Examples of the heat maps of two correlation matrices representing the different nature of relationships in a hierarchy over observed variables. 65

5.2 Heat maps of the correlation matrices corresponding to some generated data sets in different scenarios. 77
5.3 Path diagram of HierDPCA on the Big Five Personality Test data set. Dashed arrows connecting the five dimensions with the corresponding variables represent correlations < 0.7. 84
5.4 Path diagram of HierDPCA on the ASEM data set. Dashed arrows represent correlations between the observed variable and the corresponding pillar $< |0.7|$. Colors, purple and green, identify the pillars corresponding to *Connectivity* and *Sustainability*, respectively. 86

List of Tables

2.1 Simulation study results. 22
2.2 List of Bechtoldt variables. 24
2.3 Initial five groups identified by the Ultrametric Correlation Model. . 26

3.1 Holzinger data set: variables and latent dimensions (ability) description. 34
3.2 Variable groups of UCM with $Q = 4$. 37
3.3 Variable clusters at the 4th level ($Q = 4$) of the clustering methods hierarchy. 37
3.4 ARI between the theoretical membership matrix defined in Holzinger and Swineford (1937) and the membership matrices obtained by UCM and the traditional hierarchical clustering methods at level $Q = 4$ and their loss. 39

4.1 Mean and standard deviation of the ARI (mARI and sARI, respectively) between the generated and the estimated clusters for the three hierarchical scenarios. 55
4.2 Mean of the ARI between the generated and the estimated variable partitions for each cluster of GMMEUCovS both for the Qth level of the hierarchy (mARI) and across the hierarchical levels $Q, \ldots, 2$ (hARI) for the three hierarchical scenarios. 56
4.3 % of samples with a correct choice of G (and Q for GMMEUCovS and EPGMMs) for the three hierarchical scenarios. 57
4.4 Mean and standard deviation of the ARI (mARI and sARI, respectively) between the generated and the estimated clusters for the non-hierarchical scenario. 57
4.5 List of the OECD countries. 58
4.6 GMMEUCovS clusters of countries. 59

5.1 Scenarios of the simulation study. 76
5.2 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of HierDPCA for *Scenario 1*. 78
5.3 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of HierDPCA for *Scenario 2* and *Scenario 3*. 79
5.4 % of samples whose value of M estimated by HierDPCA corresponds to the true one, i.e., M^{th}, for each scenario in Table 5.1. 80
5.5 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of PCA + ORM for *Scenario 1*. 81

5.6 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of the hierarchical clustering algorithms for *Scenario 1.* 81
5.7 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of PCA + ORM for *Scenario 2* and *Scenario 3.* 82
5.8 Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of the hierarchical clustering algorithms for *Scenario 2* and *Scenario 3.* . . 82
5.9 Observed variables of the Big Five Personality Test data set, corresponding dimensions and their Cronbach's α. 83
5.10 ASEM Connectivity and Sustainability pillars. Source: W. Becker et al. (2018, pp. 23-25). 85

Abstract

Many relevant multidimensional phenomena, such as well-being, climate change, sustainable development, poverty and so on, are defined by nested latent concepts, which can be represented by a tree-shape structure supposing hierarchical relationships among observed variables. In literature, several methodologies have been proposed to both model the relationships among observed variables that reflect unobserved ones, and assess the existence of unobserved variables of "higher-order". Nonetheless, these methodologies are usually developed with sequential procedures that do not optimize a unique objective function, and/or a confirmatory approach, i.e., by setting the relationships between observed and unobserved variables a priori.

This dissertation discusses some new simultaneous, exploratory and parsimonious models for hierarchical dimensionality reduction, which overcome the limitations of the existing methodologies. The proposals introduced herein are based, "directly" or "indirectly", upon the definition of an ultrametric matrix, that differs from the well-known definition of an ultrametric distance matrix and is one-to-one associated with a hierarchy of latent concepts. The first proposal allows to model a nonnegative correlation matrix via an ultrametric correlation one by detecting reliable concepts, associated with disjoint groups of variables, and hierarchical relationships among them. The second work compares the first proposal with the traditional agglomerative hierarchical clustering algorithms applied on variables, after a transformation of correlations into distances, by highlighting the need for specific models to inspect the hierarchical relationships among variables. The third proposal extends the definition of an ultrametric matrix to a generic one by relaxing the non-negativity assumption and applying it to a covariance matrix. The extended ultrametric covariance matrix is then used to model the covariance structures of a Gaussian mixture model by both defining a new parsimonious parameterization of a covariance matrix and inspecting the hierarchical structure underlying multidimensional phenomena in heterogeneous populations. The fourth proposal introduces a quantification of latent concepts via a hierarchical extension of the Disjoint Principal Component Analysis. Even if not directly based on the definition of an ultrametric matrix, this proposal aims in turn at pinpointing nested partitions of variables into groups, each one associated with a component.

The proposed models are illustrated both via simulation studies and real data applications in order to study their performances and abilities.

Chapter 1

Introduction

The study of multidimensional phenomena is currently growing with the complexity of the reality, raising the need for methodologies to explore the relationships between their many facets. Multidimensional concepts often include more specific ones, highlighting the existence of an underlying tree structure to represent them. The root of the tree is a general concept usually corresponding to the multidimensional phenomenon under study; the leaves coincide with the observed variables and the internal nodes can represent specific dimensions defining the general concept.

Detecting latent dimensions with different relationship intensities is a crucial need for a correct and all-around understanding of the phenomenon under study in different fields, e.g., psychometrics, marketing, climatology, and environmental science, economics, and social sciences, and so on. In psychometric studies, many examples of multidimensional phenomena that entail the presence of a hierarchy of latent concepts are described, e.g., the cognitive abilities (the *g* factor, Spearman, 1927; Carroll, 1993), the personality traits, (the *Big Five model*, Cattell, 1947; Eysenck, 1970; Digman, 1990; Costa & McCrae, 1992; Goldberg, 1990, 1992, 2006; de Raad & Mlačić, 2015, among others), as well as allometry studies in the field of physiology (Rindskopf & Rose, 1988). For instance, according to the three-stratum theory (Carroll, 1993), the *g* factor can be conceptualized via a hierarchy made up of specific (narrow) factors - e.g., induction, quantitative reasoning - directly associated with the observed variables (first stratum); eight abilities (e.g., crystallized intelligence, broad visual perception), called broad factors (second stratum); and the general intelligence factor, which covers the total domain of cognitive abilities and accounts for all the relationships among the observed variables (third stratum). Moreover, the *Big Five model* highlights a hierarchy of latent concepts representing personality traits of human beings with different levels of abstraction, from the most specific (openness to experience, conscientiousness, extraversion, agreeableness and neuroticism) to the most abstract (intelligence). In environmental studies, climate change, that is nowadays one of the most urgent topics in public debate because of its risks for human life, is an epitome of a multidimensional phenomenon. Indeed, this is defined by different dimensions pertaining to greenhouse gas emissions, human causes of climate change, impacts on humans and natural systems, and efforts of human to avoid and adapt to the consequences, each of which is described by a set of variables directly observed (UNECE, 2017). Other examples concern the study of

well-being, poverty, sustainability and socio-economic phenomena, some of which will be analyzed throughout this thesis.

In the specialized literature, manifold models have been developed to analyze hierarchical structures that refer to a multidimensional phenomenon. Therefore, the existing methodologies are usually based upon sequential procedures for the hierarchy construction, i.e., without specifying an overall objective function, or defined in a confirmatory approach, i.e., by fixing the relationships between observed variables and latent specific dimensions a priori. In the former case, an inaccurate detection of the hierarchical relationships among observed variables may occur by leading to incorrect conclusions on the definition of the general concept; whereas in the latter case, the researcher knows the hierarchical relationships among latent concepts defining the general one and tests the hypothesized structure. However, a theoretical conceptualization of the phenomenon under study may be not available or the existing one may not be confirmed, highlighting the need for an exploratory approach.

In this thesis, we introduce new *simultaneous* and *exploratory* models for studying multidimensional phenomena. The proposals are characterized by a common feature: *ultrametricity*. This is an important notion in different fields, like mathematics (e.g., Schikhof, 1985), physics (e.g., Mézard et al., 1984; Parisi & Ricci-Tersenghi, 1999), taxonomy (e.g., Benzécri, 1973). In statistics, the ultrametric property is usually connected with distances in hierarchical clustering, where a complete hierarchy over units is associated with an ultrametric distance matrix. Nonetheless, the definition of an *ultrametric matrix*, which differs from that of an ultrametric distance matrix even if a relationship between the two exists, is certainly less known. The methodologies proposed in this dissertation are based upon the latter definition that can be applied to correlation matrices and, with some extensions proposed herein, to covariance matrices. One of the main features of an ultrametric matrix is the relation with a hierarchy of partitions and specifically, thanks to its application to correlation (and covariance) matrices, with partitions of the variable space. When multidimensional phenomena are studied, observed variables are often highly correlated in "blocks" such that they can be partitioned into groups associated with latent concepts by inspecting their correlations (covariances). In this case, the number of internal nodes of the tree used to represent multidimensional concepts is limited, thus a *parsimonious* hierarchy (tree, Gordon, 1999) can be depicted. Since the proposals define parsimonious hierarchies by firstly partitioning the variables into groups, they can be considered into the dimensionality reduction framework.

The purpose of this chapter is to provide the reader with an introduction of the problem under study, i.e., the analysis of multidimensional phenomena composed of nested latent concepts, how this problem has been addressed in the literature and which is the central idea underlying the proposals. An overview of the existing methodologies for detecting hierarchical structures of latent concepts is provided in Section 1.1, whereas ultrametricity is discussed in Section 1.2. Section 1.3 introduces the difference between reflective and formative models which define the nature of the relationships among concepts (or between observed variables and corresponding latent concepts) in two sequential levels of a hierarchical structure. The final section of this chapter (Section 1.4) gives a brief summary of the dissertation structure.

1.1 Hierarchical structures on variables: an overview

In many areas of study, real problems concern multidimensional phenomena whose complexity cannot be directly explored via observed variables. For this reason, multidimensional concepts may be hypothesized to have a hierarchical latent structure representing different levels of abstraction, from the most specific concepts to the most general one. We can observe that the design of this structure among levels can be *crossed* or *nested*. In the former, all possible combinations of concepts between two levels are considered such that a concept of lower level may affect more than one of higher level; whereas in the latter, concepts of higher level reconstruct only some of those (nested) of lower level, or observed variables, by defining a tree configuration (Gordon, 1999). The nested form is associated with a hierarchical partition of variables, that is in turn composed of disjoint groups of observed variables whose pairwise possible amalgamations give rise to broader groups according to the magnitude of the relationships between the concepts they represent. This thesis focuses on hierarchical structures composed of *nested* latent concepts.

Factor Analysis (FA, Spearman, 1904; Anderson & Rubin, 1956; Horst, 1965) is one of the most used models to reconstruct the relationships among variables, i.e., the covariance or correlation matrix, via a set of latent factors. Together with Principal Component Analysis (PCA, Pearson, 1901; Hotelling, 1933), FA aims at reducing the dimensionality of the data by computing a reduced number of unobserved variables (components or factors), but preserving as much information as possible regarding the relationships among the observed variables. Therefore, neither FA nor PCA are suitable to detect the hierarchical relationships among observed variables. In the specialized literature, different classes of models have been developed to analyze hierarchical structures that refer to a multidimensional phenomenon. First of all, two main classes of *sequential* and *exploratory* methodologies have been introduced: the Higher-Order Factor models (G. H. Thompson, 1948; Cattell, 1978b; Rindskopf & Rose, 1988; Undheim & Gustafsson, 1988) and the Bi-Factor or Hierarchical Factor models (Holzinger & Swineford, 1937; Wherry, 1959; Schmid & Leiman, 1957; Jennrich & Bentler, 2011, 2012), which mainly differ in the construction of the hierarchy. Indeed, the former aim at pinpointing higher-order of factors via sequential applications of the exploratory FA (B. Thompson, 2004) on the covariance or correlation matrix of the observed variables first, and higher-order factors then, followed each time by an oblique rotation method, until zero correlation occurs among factors or a single factor is detected (Gorsuch, 1983). Insofar each level of the hierarchy is obtained from the lower previous one, an indirect relationship between the general factor and the observed variables through the other (higher-order) latent factors is identified. Contrariwise, the Bi-Factor or Hierarchical Factor models are characterized by a single order of orthogonal hierarchical factors, usually obtained by applying the Schmid-Leiman transformation (Schmid & Leiman, 1957) to the corresponding higher-order solutions, and they identify a direct effect of the general factor on the observed variables. Several authors (Mulaik & Quartetti, 1997; Yung, Thissen, & McLeod, 1999; Gignac, 2016) have shown the equivalence of the aforementioned classes of models under certain conditions. Moreover, in both models a simple structure of the loading matrices (Thurstone, 1947) can be sought. Recently, Cavicchia and Vichi (2021) proposed a *simultaneous* and *exploratory two-*

level FA, called Second-Order Disjoint Factor Analysis, to model phenomena with an underlying hierarchical structure of latent concepts composed of two orders, in the first of which disjoint groups of observed variables are identified.

Other methodologies have been proposed to study the data with a hierarchical structure on variables. Among others, Holzinger (1944) and Jöreskog (1966, 1969, 1978) introduced the Hierarchical Confirmatory Factor Analysis by assuming that the number of factors and the relationships between factors and observed variables are known a priori. Moreover, the relationships among observed and unobserved variables, as well as among the latter ones, can be investigated via Structural Equation Modeling (SEM, Wright, 1921; Kline, 2015, for a complete overview). SEM are *simultaneous* models which combine *confirmatory* FA (B. Thompson, 2004) and regression analysis (see, for example, Seber & Lee, 2003), and can be estimated with two different approaches: the LInear Structural RELations (LISREL or SEM-ML, Jöreskog, 1970; Jöreskog & Sörbom, 1982) approach and the Partial Least Squares Path Modeling (PLS-PM or SEM-PLS, Wold, 1966, 1982, 1985; Tenenhaus et al., 2005) approach (see Jöreskog & Wold, 1982, for a comparison between the two approaches). The former, also known as the covariance-based method, is based upon the maximum likelihood estimation method, thus requiring distributional assumptions, and is related to FA. The latter, also known as the component-based method, does not assume any distributional assumption and is related to PCA. It is worth noticing that if the general factor corresponds to one of the higher-order factors, the Higher-Order Factor models can be investigated via LISREL approach (Undheim & Gustafsson, 1988). SEM models have been developed in turn in a confirmatory approach; however, they do not build a hierarchy in the broad sense over the observed variables.

In order to study data which are described by several groups of variables of different nature (qualitative and quantitative) and organized in a hierarchical structure, Le Dien and Pagès (2003) proposed a hierarchical extension of Multiple Factor Analysis (Escofier & Pagès, 1983, 1994), called Hierarchical Multiple Factor Analysis (HMFA). The latter aims at integrating different groups of variables which describe the same observations and balancing their role within each node of the hierarchy. Additionally, HMFA provides an overall, and a partial (for each node of the hierarchy), graphical display of the variable groups.

Finally, hierarchical clustering algorithms (Cliff et al., 1995; Gordon, 1999, Chapter 4; Rencher, 2002, pp. 455-481; Strauss, Bartko, and Carpenter, 1973) may be used to inspect hierarchical structures on variables. Indeed, after a transformation of correlations (measure of similarity) among variables into distances (measure of dissimilarity) by subtracting to one the absolute value or the square of the correlation coefficients (Revelle, 1979; Soffritti, 1999; Liu et al., 2012), hierarchical clustering algorithms may be implemented in order to build a hierarchy over the observed variables. It is worth highlighting that these algorithms are based upon *sequential* and *greedy* procedures carried out in an *exploratory* approach, and build a *complete hierarchy*, i.e., a tree with the maximum number of internal nodes given the number of variables. Thus, a partition of the variable space into a reduced number of groups is obtained only *a posteriori*, by cutting the complete tree. Agglomerative hierarchical clustering methods will be discussed in Chapter 3 in order to compare their performances with respect to those of the first proposal illustrated in this thesis

(Chapter 2).

In the following section, the ultrametric property is introduced and the distinction between the definition of the well-known ultrametric distance matrix and that one of the ultrametric matrix is given.

1.2 Ultrametricity

Ultrametricity is a very important notion introduced in mathematics with regard to p-adic number theory. In the last two decades, ultrametricity has gained attention in different fields, like statistical physics with application to spin glasses, taxonomy and evolutionary biology for the phylogenetic tree construction, etc., thanks to its relationship with nested partitions and tree structures.

In statistics, the ultrametric property can be found in hierarchical cluster analysis (see Gordon, 1987, for an exhaustive review), where a complete hierarchy over a set of objects[1] I is built. The graphical representation of hierarchical structures used in hierarchical cluster analysis is a particular tree, called *dendrogram* (Gordon, 1987, 1999, Chapter 4). The latter corresponds to an ultrametric distance matrix, whose definition is reported as follows.

Definition 1.1. Given a set of objects I, a matrix $\mathbf{D}$ is an ultrametric distance matrix if

(i) $d_{ij} \geq 0$ for all $i, j \in I$ (non-negativity);

(ii) $d_{ij} = d_{ji}$ for all $i, j \in I$ (symmetry);

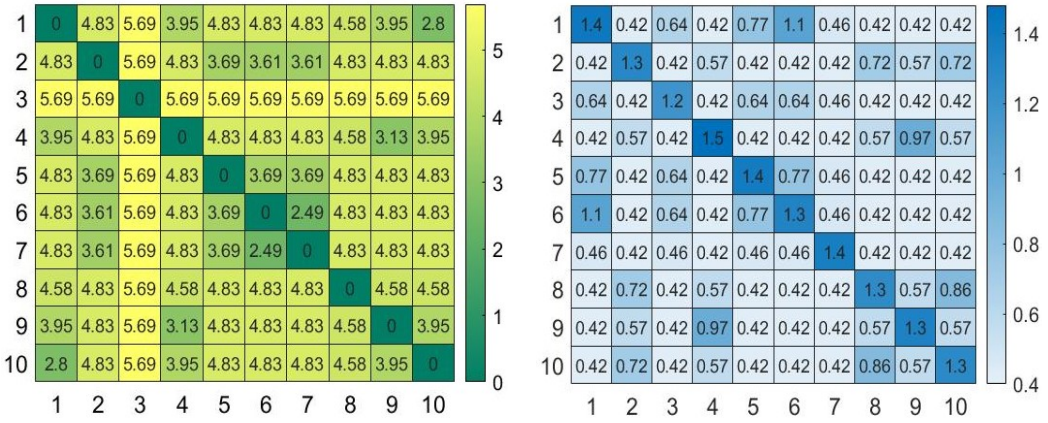
(iii) $d_{jj} = 0$ for all $j \in I$ (zeros on the main diagonal);

(iv) $d_{ij} \leq d_{il} + d_{jl}$, for all $i, j, l \in I$ (triangle inequality);

(v) $d_{ij} \leq \max\{d_{il}, d_{jl}\}$, for all $i, j, l \in I$ (ultrametric inequality).

Johnson (1967) first demonstrated that a hierarchical classification was endowed with the ultrametric property. Nonetheless, it is worth highlighting that not all hierarchical clustering algorithms induce an ultrametric metric (Milligan, 1979), like the centroid method (UPGMC) and the median method (WPGMC). Definition 1.1 states that a distance matrix - which is nonnegative, symmetric, with zeros on the main diagonal and satisfies the triangle inequality by definition - must fulfill the ultrametric inequality, also called *strong triangle inequality* (see, for instance, Contreras & Murtagh, 2015, p. 105). The latter simply implies that objects (or cluster of objects) merged earlier in the hierarchy have a distance value that is smaller than the distance value of objects (or cluster of objects) merged later in the hierarchy. Thus, no reversal occurs in the corresponding tree, but rather the distances among objects (cluster of objects) monotonically increase from the bottom of the hierarchy upwards. An example of an ultrametric distance matrix is given in Figure 1.1a.

[1] Tipically, the term *objects* refers to units in hierarchical cluster analysis. In this dissertation, we refer to objects as a synonym of both units and variables because of the application of hierarchical clustering algorithms to classify variables, as we will see in Chapter 3.



(a) *Ultrametric distance matrix* **(b)** *Ultrametric matrix*

Figure 1.1. Examples of an ultrametric distance matrix and an ultrametric matrix.

As already mentioned, the hierarchical clustering algorithms build a complete hierarchy over objects. This corresponds to a tree (dendrogram) with the maximum number of internal nodes, i.e., $|I| - 2$ (except the root) where $|I|$ represents the cardinality of the set of objects I. However, one of the most common uses of these methods is to derive a partition of the object space by cutting the dendrogram at a specific level chosen by a visual inspection of the latter, thus only *a posteriori* with respect to the hierarchy construction. Alternative procedures to obtain a parsimonious hierarchy of objects are those based on a "tandem analysis" which first applies Multidimensional Scaling (MDS, Torgerson, 1958; Gower, 1966) on a distance matrix, and then a partitioning algorithm like K-means (MacQueen, 1967; Hartigan & Wong, 1979) on the dimensions specified by MDS. Vichi (2008) introduced a class of new methodologies for hierarchical clustering in which a partition of the object space was directly estimated in a model-based approach. The parsimonious tree obtained in this way is associated with an ultrametric distance matrix that contains a reduced number of different values.

In mathematics, another definition pertaining to ultrametricity was introduced with respect to the study of M-matrices (Martínez, Michon, & San Martín, 1994): that one of an *ultrametric matrix*, which is provided as follows.

Definition 1.2. Given a set of objects I, a matrix $\mathbf{U}$ is an ultrametric matrix if

(i) $u_{ij} \geq 0$ for all $i, j \in I$ (non-negativity);

(ii) $u_{ij} = u_{ji}$ for all $i, j \in I$ (symmetry);

(iii) $u_{jj} \geq \max\{u_{kj} : k \in I\}$ for all $j \in I$ (column pointwise diagonal dominance);

(iv) $u_{ij} \geq \min\{u_{il}, u_{jl}\}$, for all $i, j, l \in I$ (ultrametric inequality).

It is worthy of remark that Definition 1.2 will be re-written in Chapter 2 in order to help the reader to follow the mathematical notation used in that chapter. An example of an ultrametric matrix is given in Figure 1.1b.

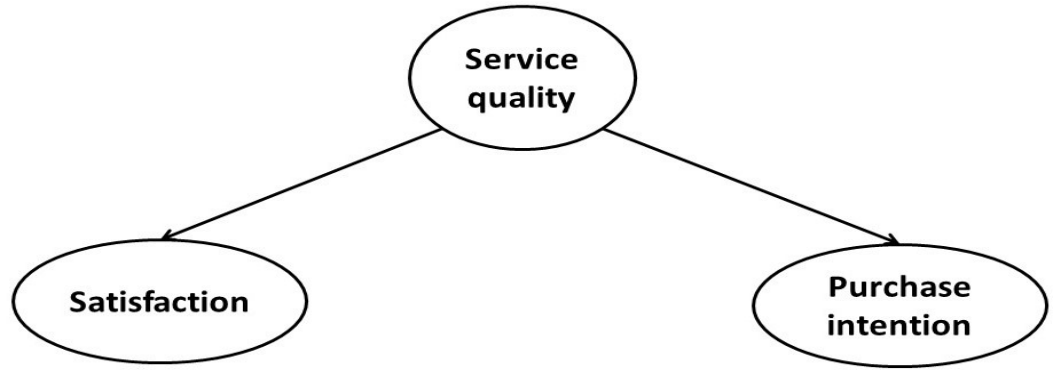
Definition 1.2 differs from an ultrametric distance matrix in Definition 1.1, because the conditions of null diagonal and triangle inequality do not hold for an ultrametric matrix. The former (condition (iii) of Definition 1.1) is replaced with the column pointwise diagonal dominance for an ultrametric matrix (condition (iii) of Definition 1.2). This implies a reverse relationship between the diagonal entries of the two matrices with respect to their off-diagonal values. Indeed, an ultrametric distance matrix has its minimum value on the main diagonal, i.e., zero for each diagonal entry, whereas in an ultrametric matrix each diagonal element corresponds to the maximum value of the corresponding column (or row thanks to condition ii), remembering that both matrices are nonnegative by definition (condition i). Moreover, the ultrametric condition in Definition 1.2 (condition iv) shows a reverse inequality with respect to the one in Definition 1.1 (condition v). The ultrametric matrix is in turn associated with a hierarchy, where objects (or cluster of objects) merged earlier have a stronger relationship than objects (or cluster of objects) merged later in the hierarchy. Thus, the interpretation of the hierarchical structure obtained from an ultrametric matrix differs from that of an ultrametric distance matrix.

In this thesis, we use the definition of an ultrametric matrix to model hierarchical relationships among variables. In fact, conditions (i), (ii), (iii) in Definition 1.2 hold for a nonnegative correlation matrix, which turns out to be ultrametric if condition (iv) in Definition 1.2 is satisfied. Moreover, as we will discuss in Chapter 2, every ultrametric matrix is positive semi-definite by allowing the application of Definition 1.2 to a nonnegative correlation matrix. Furthermore, in Chapter 4 we will propose an extension of Definition 1.2 to a generic matrix such that the positive semi-definiteness still holds even relaxing the non-negativity constraint and allowing its application to a generic covariance matrix. In Chapter 4, the definition of an extended ultrametric covariance matrix will be used to characterize a new parameterization of a covariance matrix in Gaussian mixture models (Titterington, Smith, & Makov, 1985; McLachlan & Basford, 1988; McLachlan & Peel, 2000a). It has to be highlighted that an extended ultrametric covariance matrix is in turn associated with a hierarchy, where the most concordant variables are merged earlier in the hierarchy than the less concordant ones (most discordant), and its application to Gaussian mixture models enables to study multidimensional phenomena in heterogeneous populations.

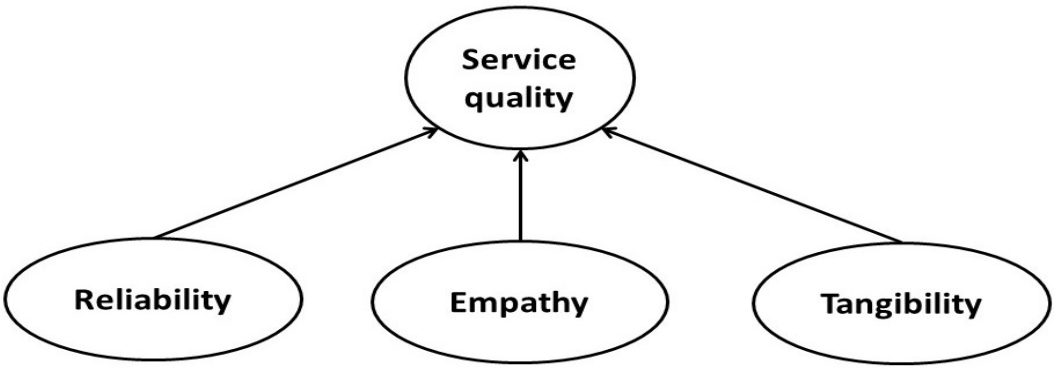
The methodologies presented in this dissertation directly pinpoint a partition of the variable space in a reduced number of groups - without cutting the tree after its construction (a posteriori), as done by the hierarchical clustering algorithms - and a parsimonious hierarchy over them by identifying specific features of the variable groups in terms of correlation and covariance, as in Chapters 2 and 4, or quantifying the latent concepts associated with them, as in Chapter 5. In Chapters 2 and 4, the parsimony of the proposals results in ultrametric and extended ultrametric matrices, respectively, which contain a reduced number of different values.

1.3 Reflective and formative models

Two main features typically characterize a hierarchical structure of nested latent concepts underlying a multidimensional phenomenon: the number of levels in the hierarchy, i.e., internal nodes of the corresponding tree, and the nature of



(a) *Reflective model*



(b) *Formative model*

Figure 1.2. Examples of the measurement model.

the relationships between (observed and unobserved) variables belonging to two sequential hierarchical levels. As already mentioned in the previous sections, a reduced number of internal nodes of the tree gives rise to a parsimonious hierarchy. The latter represents the main structure the proposals introduced in this thesis are based on. The nature of the relationships in a hierarchical structure is instead specified by their "direction", which formally describes the measurement model (Blalock, 1964; Bollen & Bauldry, 2011), and affects both the relationships between observed and unobserved variables than unobserved variables and "higher-order" ones. It is worth highlighting that we refer herein to an unobserved variable as a quantitative representation (quantification) of a latent concept.

Two different directions of the relationships among variables belonging to sequential levels of a hierarchy exist: *reflective* and *formative* (Bollen, 2001). A reflective relation occurs when a set of correlated observed (or unobserved) variables reflect a (higher-order) unobserved one, i.e., a (broader) latent concept accounts for the covariance/correlation among the observed (or unobserved of lower level) variables. Whereas, a formative relation arises when the (higher-order) unobserved variable is defined/formed by observed (or unobserved of lower level) variables, which are generally uncorrelated to each other. In this case, the observed (or lower-order unobserved) variables represent a unique part of the (higher-order) unobserved variable. Blalock (1964) referred to reflective models as *effect* models, since the

relationships among the variables of lower level depend on a common variable of higher level that explains them in a top-down approach, and to formative ones as *causal* models, since the variables of lower level determine the one of higher level in a bottom-up approach. A real example of the two kinds of relations among sequential levels of a hierarchical structure is provided in Figure 1.2. These two approaches are not necessarily referring to causality, since a relationship between two elements does not imply a causal link between the two, but rather the need for one element for the existence of the other. In order to better understand the difference between reflective and formative models, see, for example, Edwards and Bagozzi (2000) and Jarvis, MacKenzie, and Podsakoff (2003).

FA and PCA usually define reflective relationships between observed variables and factors or components, respectively. Nonetheless, they pinpoint unobserved variables which turn out to be uncorrelated and, thus, could define a higher-order one in a formative approach. Although Higher-Order Factor models and Bi-Factor or Hierarchical Factor models described in Section 1.1 are based upon FA, they were developed to build only reflective hierarchies. Indeed, these models implement an oblique rotation method after each FA application such that unobserved variables (factors) can be correlated and the higher-order level is computed if the correlation among unobserved variables of lower level occurs; otherwise, the hierarchy construction stops. Similarly, Second-Order Disjoint Factor Analysis (Cavicchia & Vichi, 2021) theorizes a reflective relationship between the observed variables and the higher-order factors, i.e., the factors of the first level, as well as between the latter and the general one, i.e., the factor of the second level. It is worth noticing that this methodology extends the Disjoint Factor Analysis (Vichi, 2017) by relaxing the orthogonal constraint on the unobserved variables and assuming that a second-order factor, i.e., the general one, exists.

When dealing with confirmatory models, the relationships among variables and their nature are set a priori. For instance, in SEM-ML the relations between observed and unobserved variables are usually modeled by FA thus with a reflective approach, whereas those among unobserved variables are modeled via multivariate regression and thus can be interpreted as formative. In SEM-PLS, J. M. Becker, Klein, and Wetzels (2012) discussed guidelines for using reflective-formative, i.e., mixed, models. It is noteworthy that an entirely *formative* model is not identified (Edwards, 2011; Bollen, 2011). To solve this problem, Hauser and Goldberger (1971) and Jöreskog and Goldberger (1975) proposed the Multiple Indicators and Multiple Causes model in which an unobserved variable represents both the effect of some observed ones and the determinant of some others.

In the real applications we will illustrate in Chapter 2, the existence of a general concept which causes some nested specific ones is theorized by identifying a reflective model. Nevertheless, in some situations researchers may not have a theoretical definition of the hierarchical relationships among variables, or maybe this may not be empirically confirmed by a test. In Chapter 5, we will propose a methodology which defines a model-based approach to choose the nature of the relationships among unobserved variables of two contiguous hierarchical levels. The latter assumes a reflective relationship between the observed variables and unobserved ones of the first bottom-up level, and then tests the correlation among the unobserved variables along the hierarchy by changing from a reflective to a formative approach at the

first level at which the correlation turns out to be not statistically significant.

1.4 Chapter summaries

In this section, an overview of the following chapters and appendices is provided.

Chapter 2 presents a novel, exploratory, parsimonious and simultaneous model, called *Ultrametric Correlation Model* (UCM), to study multidimensional phenomena. The proposed methodology reconstructs a nonnegative correlation matrix via an ultrametric correlation one that is able to pinpoint the hierarchical nature of the phenomenon under study. Indeed, UCM detects non-overlapping groups of variables, each one associated with a latent concept, and their hierarchical relationships by inspecting the internal consistency of the latent concepts and the correlation between them. A relationship between these features and a well-known measure of internal consistency of a variable group is provided. The performance of the proposed model is illustrated through a simulation study and two real applications - one on a benchmark data set regarding mental abilities and the other one on drug consumption.

The contents of Chapter 2 were developed with Prof. Maurizio Vichi and Dr. Carlo Cavicchia, and are reported in a paper which was published in *Advances in Data Analysis and Classification* in 2020, see Cavicchia, Vichi, and Zaccaria (2020b). The contents of Section 2.6.2 concerning the real data example on drug consumption were developed with Prof. Maurizio Vichi, and are reported in a volume which was published by Pearson in 2020, see Zaccaria and Vichi (2020).

Chapter 3 provides a comparison between UCM and the procedure based on well-known hierarchical clustering methods used for variable classification. The performances of the aforementioned model and methods are illustrated through an application to the Holzinger data set, which represents a real demonstration of a hierarchical structure of latent concepts.

The contents of Chapter 3 were developed with Prof. Maurizio Vichi and Dr. Carlo Cavicchia, and are reported in the chapter of a volume published by Springer in 2020, see Cavicchia, Vichi, and Zaccaria (2020a).

Chapter 4 extends the model presented in Chapter 2 to a generic covariance matrix. The definition of an *extended ultrametric covariance matrix* is stated and implemented into a Gaussian mixture model. The proposal is able to pinpoint a hierarchical structure on variables for each component of the Gaussian mixture, thus identifying a different characterization of a multidimensional phenomenon for each component (cluster, subpopulation) of the mixture. At the same time, the proposed parameterization of the covariance matrix defines a new parsimonious Gaussian mixture model since the ultrametric covariance structure reconstructs the relationships among variables with a limited number of parameters. Furthermore, a simulation study shows the performance of the proposal both in terms of cluster recovering - even in comparison with other existing methodologies - and correct identification of the variable partition and hierarchical structure over it. Two real data examples are used to illustrate the features of the proposed methodology.

The contents of Chapter 4 have been developed with Prof. Maurizio Vichi and Dr. Carlo Cavicchia, and are reported in a paper that has been accepted for publication

in *Advances in Data Analysis and Classification*, see Cavicchia, Vichi, and Zaccaria (2022).

Chapter 5 proposes a new exploratory and simultaneous model, called *Hierarchical Disjoint Principal Component Analysis* (HierDPCA), with the aim of building a parsimonious hierarchy of nested components associated with disjoint groups of observed variables. Differently from the models proposed in the previous chapters, HierDPCA introduces the quantification of the latent concepts associated with variable groups for each level of the hierarchy. Moreover, the proposed methodology allows to choose the type of the relationship among components of two sequential levels, from the lowest upwards, by testing the component correlation per level and changing from a reflective to a formative approach when this correlation turns out to be not statistically significant. The performance of the proposal is illustrated through an extensive simulation study and two real data applications.

The contents of Chapter 5 have been developed with Prof. Maurizio Vichi and Dr. Carlo Cavicchia, and are reported in a paper which has been submitted for publication and is currently under the second revision in an international journal, see Cavicchia, Vichi, and Zaccaria (2021).

All the models presented in Chapters 2-5 have been implemented in a MATLAB routine.

Appendices A-C report the supplementary materials for Chapters 2, 4 and 5.

Chapter 2

The ultrametric correlation matrix for modeling hierarchical latent concepts

2.1 Introduction

The identification of a hierarchy of nested latent concepts is a considerable aspect in the study of phenomena composed of different facets, i.e., multidimensional phenomena. These arise in many fields like psychometrics, marketing, economics and social sciences, etc., and need specific models to be studied. Manifold methodologies were proposed to deal with the problem of the construction of a general latent concept via a hierarchy of nested specific ones, as illustrated in Chapter 1 (Section 1.1). Therefore, these methodologies are based upon sequential applications of exploratory factor analysis, like the Higher-Order Factor models and Bi-Factor or Hierarchical Factor models, without specifying an overall objective function, or defined in a confirmatory approach, i.e., by fixing the relationships between observed variables and latent concepts a priori, as in SEM models.

In this chapter, we propose a *simultaneous*, *exploratory* and *parsimonious* model, named Ultrametric Correlation Model (UCM), to reconstruct a correlation matrix via an *ultrametric* correlation one in order to explore a multidimensional phenomenon (general concept) through a set of *nested* specific dimensions (concepts). UCM gives rise to a hierarchical partition of the observed variable space that can be associated with a hierarchy of latent concepts defining a general broader one. The nested design of the hierarchical structure among concepts makes them easier to interpret with respect to the crossed one. Thus, each dimension can be directly or indirectly specified by a set of observed variables. In order to detect specific concepts at each level of the hierarchy, i.e., a tree-shape structure, and different relationships between them, two main features characterize the ultrametric correlation matrix of the proposed model: the *internal consistency of a latent concept*, i.e., the concordant relations observed within a group of variables that assess the reliability of the concept, and the *correlation between concepts*, i.e., the concordant relations between two groups of observed variables. Some relationships between these two characteristics of latent concepts and the Cronbach's α (Cronbach, 1951) are here highlighted. UCM

is estimated in the Least-Squares (LS) non-parametric approach searching for the internal consistency of concepts and the correlations between them which better represent the hierarchical structure of the phenomenon under study. Furthermore, to avoid that specific concepts compensate in the definition of a latent dimension, the correlation matrix of the data is assumed to be nonnegative[1]. Even if restrictive, the latter assumption turns out to be realistic in a manifold of real applications, e.g., the g factor (Spearman, 1927; Carroll, 1993) in psychometric studies on intelligence and cognitive abilities.

The chapter is organized as follows. In Section 2.2, the notation used in the whole chapter and some basic notions on hierarchical partitions and ultrametric matrices are provided to allow the reader to follow the specification of the model herein. An in-depth description of the proposed methodology is provided in Section 2.3. Section 2.4 is dedicated to the non-parametric least-squares estimation of the model together with the description of the corresponding algorithm. Section 2.5 discusses the results of a simulation study to assess the model. Two real applications are illustrated in Section 2.6 and a final discussion completes the chapter in Section 2.7.

2.2 Notation and basic notions

For the convenience of the reader, the notation used in this chapter is listed here:

p, Q	number of variables, number of groups of the variable partition.
C	set (partition) $\{C_1, \ldots, C_Q\}$ of Q groups of variables.
$\mathbf{R} = [r_{jl}]$	$(p \times p)$ correlation matrix of the observed variables, where r_{jl} is the correlation between variables j and l $(j, l = 1, \ldots, p, l \neq j)$.
$\mathbf{R}_{\mathrm{W}} = [{}_W r_{qq}]$	$(Q \times Q)$ within-concept consistency (diagonal) matrix, where ${}_W r_{qh} = 0$ for all $q \neq h$; ${}_W r_{qq} > 0$ $(q = 1, \ldots, Q)$ represents the consistency within the qth group of variables.
$\mathbf{R}_{\mathrm{B}} = [{}_B r_{qh}]$	$(Q \times Q)$ between-concept correlation matrix, where ${}_B r_{qh}$ $(q, h = 1, \ldots, Q, h \neq q)$ denotes the correlation between latent concepts associated with the variable groups $\{C_q, C_h\} \subset C$; ${}_B r_{qq} = 1$ for all $q = 1, \ldots, Q$.
$\mathbf{V} = [v_{jq}]$	$(p \times Q)$ membership matrix, where $v_{jq} = 1$ if the jth variable belongs to the qth group C_q; $v_{jq} = 0$ otherwise. It pinpoints a partition C of variables in Q groups, $\{C_1, \ldots, C_Q\}$; thus, $\mathbf{V}$ is binary and row-stochastic, i.e., with one non-zero element per row.
$\mathbf{1}_p, \mathbf{1}_Q, \mathbf{I}_p, \mathbf{I}_Q$	unitary vector of order p and Q, identity matrix of order p and Q, respectively.
$\mathbf{E} = [e_{jl}]$	$(p \times p)$ error matrix.

The *internal consistency* of a group C_q $(q = 1, \ldots, Q)$ of observed variables is the extent to which all the variables in C_q contribute to identify the same latent concept. A measure of the internal consistency of a group C_q is the Cronbach's α

[1] A matrix $\mathbf{A} = [a_{ij}]$ is *nonnegative* if $a_{ij} \geq 0 \; \forall i, j$ (see Horn & Johnson, 2013, p. 519). It is important to stress that the definition of a nonnegative matrix is different from that of a *nonnegative definite* matrix.

(Cronbach, 1951). It ranges between 0^2 and 1, that are reached when C_q has an identity correlation matrix, i.e. $\mathbf{R}_q = \mathbf{I}_q$, and C_q has a correlation matrix made up of unitary elements, i.e. $\mathbf{R}_q = \mathbf{1}_q\mathbf{1}_q'$, respectively. In this framework, the internal consistency of the variable groups C_q, $q = 1, \ldots, Q$, is represented by the value, one per group, ${}_W r_{qq}$ computed as a function of the correlations between variables in C_q (see Section 2.4, Eq. 2.11) and arranged on the main diagonal of the matrix $\mathbf{R}_{\mathrm{W}}$, where ${}_W r_{qh} = 0$ for all $q \neq h$. It is worthy to remark that a relationship between based upo of C_q and the corresponding within-concept consistency coefficient ${}_W r_{qq}$ exists (Cronbach, 1951; Osburn, 2000; Warrens, 2015). Indeed, the standardized Cronbach's α - a generalization of the Spearman-Brown formula (Spearman, 1910; Brown, 1910) - can be written as $\alpha_q^S = \frac{J_q\, {}_W r_{qq}}{1+(J_q-1)\, {}_W r_{qq}}$, where J_q is the number of variables in C_q and $J_1 + \cdots + J_Q = p$.

Given two groups C_q and C_h $(h \neq q)$ of observed variables, the *correlation between* them measures the extent to which variables in C_q are concordant with variables in C_h and, therefore, it measures the correlation between latent concepts. For $C_1, \ldots, C_Q$, $\frac{Q(Q-1)}{2}$ correlations between pairs of variable groups - each one representing a latent concept - can be computed. These values can be arranged in a correlation matrix $\mathbf{R}_{\mathrm{B}} = [{}_B r_{qh}]$ of order Q, where ${}_B r_{qh}$ $(q, h = 1, \ldots, Q, h \neq q)$ is defined as a function of the correlations between pairs of variables, one belonging to C_q and the other one to C_h (see Section 2.4, Eq. 2.14); hence, ${}_B r_{qq} = 1$ for all q. A further relationship between α_q^S, α_h^S, ${}_B r_{qh}$ $(h \neq q)$ and the Cronbach's α of the set $C_q \cup C_h$ exists (see Appendix A).

Before going into detail of the proposal, let us provide some basic notions that turn out to be necessary for the explanation of the UCM. As done from Section 2.3 onward, let us suppose that the Q groups of variables form a partition of the observed variable space. This implies that each variable contributes to specify one and only one latent dimension. Therefore, we need to introduce the definition of a hierarchy of latent concepts and an ultrametric matrix[3], which differs from that of the well-known ultrametric distance matrix as already mentioned in Chapter 1 (Section 1.2).

Definition 2.1. A *hierarchy of latent concepts* is a set $H_Q = \{C_q, (q = 1, \ldots, Q), C_{Q+1}, \ldots, C_{2Q-1}\} = \{C, C_{Q+1}, \ldots, C_{2Q-1}\}$, composed of $2Q - 1$ groups, each one representing a latent concept, where the first $C_1, \ldots, C_Q$ stand for the subsets of an initial partition C of the observed variables with internal consistency ${}_W r_{qq}$ $(q = 1, \ldots, Q)$; the remaining groups, i.e. $C_{Q+1}, \ldots, C_{2Q-1}$, are obtained by $Q - 1$ pairwise possible amalgamations of subsets of C with correlation between groups ${}_B r_{qh}$ $(q, h = 1, \ldots, Q, h \neq q)$. Thus, $C_k, C_h \subset H_Q \Leftrightarrow C_k \cap C_h \subset \{C_k, C_h, \emptyset\}$, $k, h = 1, \ldots, 2Q - 1$.

Definition 2.2. (Dellacherie, Martínez, & San Martín, 2014, pp. 58-59) Given a set of objects I, a nonnegative matrix $\mathbf{R}$ is said to be *ultrametric* if

(i) $r_{ij} = r_{ji}$ for all $i, j \in I$ (symmetry);

[2]Negative values are not taken into account herein thanks to the assumption of nonnegative correlations that is introduced from this section on.

[3]The definition of an ultrametric matrix is rewritten herein by using the mathematical notation adopted throughout the chapter. Henceforth, we will cite Definition 2.2 instead of Definition 1.2.

(ii) $r_{jj} \geq \max\{r_{kj} : k \in I\}$ for all $j \in I$ (column pointwise diagonal dominance);

(iii) $r_{ij} \geq \min\{r_{il}, r_{jl}\}$, for all $i, j, l \in I$ (ultrametric inequality).

Property (iii) can be equivalently rewritten as follows:

(iii') for each triplet $i, j, l \in I$, there exists a reordering $\{i, j, l\}$ of the elements s.t.

$$r_{ij} \geq r_{il} = r_{jl},$$

which corresponds to say that for each triplet the smallest two elements are equal. This fact limits the number of different values in the ultrametric matrix $\mathbf{R}$.

It is straightforward to observe that each nonnegative correlation matrix $\mathbf{R}$ satisfies properties (i) and (ii), i.e., it is symmetric and column pointwise diagonally dominant.

2.3 The model

Starting from the observation of a $(p \times p)$ nonnegative correlation matrix $\mathbf{R}$, the hierarchical statistical problem we want to deal with can be formalized as follows

$$\mathbf{R} = \mathbf{R}_{\mathrm{u}} + \mathbf{E}, \tag{2.1}$$

where the $(p \times p)$ matrix $\mathbf{R}_{\mathrm{u}}$ represents the hierarchical structure of the latent concepts, i.e., the theoretical hierarchical model for the concepts, and $\mathbf{E}$ is the random error matrix, i.e., the residual matrix from the hierarchical model. The non-negativity assumption of $\mathbf{R}$ and $\mathbf{R}_{\mathrm{u}}$ is needed to pinpoint a non-compensatory hierarchical structure of the latent concepts.

The Ultrametric Correlation Model (Ultrametric Correlation Matrix model, UCM) proposes to build an ultrametric correlation matrix for modeling hierarchical latent concepts is formally specified as follows

$$\mathbf{R}_{\mathrm{u}} = \mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q)\mathbf{V}' + \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}') + \mathbf{I}_p, \tag{2.2}$$

subject to constraints

$$\mathbf{V} = [v_{jq} \in \{0, 1\} : j = 1, \ldots, p, q = 1, \ldots, Q]; \tag{2.3}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1, \ldots, p; \tag{2.4}$$

$$\mathbf{R}_{\mathrm{B}} \text{ is an ultrametric correlation matrix (Definition 2.2);} \tag{2.5}$$

$$\min\{{}_W r_{qq} : q = 1, \ldots, Q\} \geq \max\{{}_B r_{qh} : q, h = 1, \ldots, Q, h \neq q\}, \tag{2.6}$$

where $\mathrm{diag}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}')$ is a diagonal matrix with diagonal entries equal to the diagonal of the matrix $\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}'$.

It is worthy to notice that since the within-concept consistency matrix $\mathbf{R}_{\mathrm{W}}$ is diagonal, it is ultrametric by definition.

The ultrametricity constraint (2.5) implies that the following $O(Q^3)$ constraints on the triplets of $\mathbf{R}_{\mathrm{B}}$ hold:

$$\begin{cases} {}_{B}r_{qh} \geq \min({}_{B}r_{qk},\ {}_{B}r_{hk}) \\ {}_{B}r_{hk} \geq \min({}_{B}r_{qh},\ {}_{B}r_{qk}) \\ {}_{B}r_{qk} \geq \min({}_{B}r_{qh},\ {}_{B}r_{hk}) \end{cases} \qquad q = 1, \ldots, Q, h = q, \ldots, Q, k = h, \ldots, Q. \tag{2.7}$$

Before inspecting the main properties of the proposal, a basic result and definition are provided in order to state our principal findings (Lemma 2.1 and Theorem 2.1).

Proposition 2.1. *The number of different off-diagonal elements of* $\mathbf{R}_{\mathrm{u}}$ *(*$n_{\mathbf{R}_{\mathrm{u}}}$*) is*

$$\begin{cases} 1 \leq n_{\mathbf{R}_{\mathrm{W}}} \leq Q \\ 1 \leq n_{\mathbf{R}_{\mathrm{B}}} \leq Q - 1 \end{cases} \quad \Rightarrow \quad 2 \leq n_{\mathbf{R}_{\mathrm{u}}} \leq 2Q - 1, \tag{2.8}$$

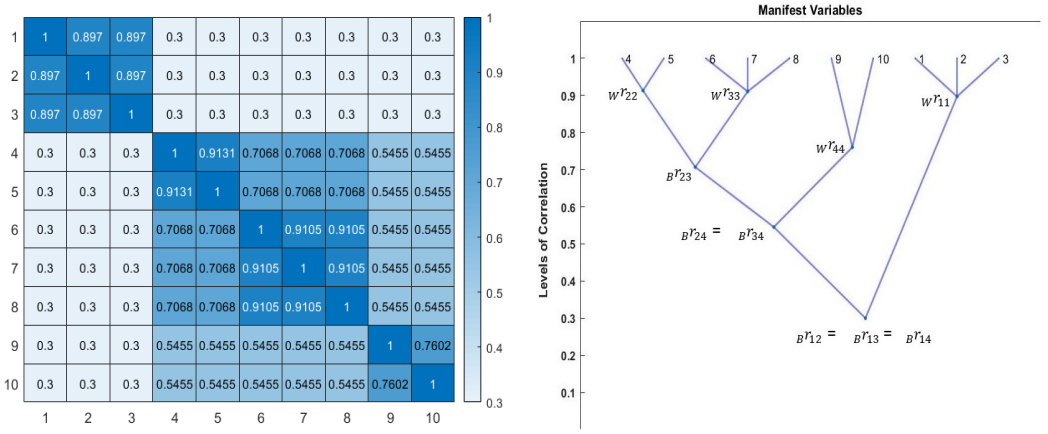
where $n_{\mathbf{R}_{\mathrm{W}}}$*,* $n_{\mathbf{R}_{\mathrm{B}}}$ *are the number of different diagonal elements of* $\mathbf{R}_{\mathrm{W}}$ *and the number of different off-diagonal elements of* $\mathbf{R}_{\mathrm{B}}$*, respectively.*

Definition 2.3. A $(2Q-1)$-ultrametric correlation matrix is a square ultrametric matrix of order p with diagonal elements equal to one and off-diagonal elements that can assume one of at most $(2Q-1)$ different values ${}_{B}r_{qh}$, ${}_{W}r_{qq}$, such that $0 \leq {}_{B}r_{qh} \leq {}_{W}r_{qq} \leq 1$.

Lemma 2.1. *A hierarchy of* Q *latent concepts, i.e.* H_Q *- starting from* p *observed variables - with within-concept consistencies, i.e., the group reliability,* ${}_{W}r_{qq}$ *(*$q = 1, \ldots, Q$*) and between-concept correlations* ${}_{B}r_{qh}$ *(*$q, h = 1, \ldots, Q$*,* $h \neq q$*), with* $0 \leq {}_{B}r_{qh} \leq {}_{W}r_{qq} \leq 1$*, is one-to-one associated with a* $(2Q-1)$*-ultrametric correlation matrix* $\mathbf{R}$*.*

Proof. Considering a hierarchy of latent concepts H_Q, it can be stated that the jth $(j = 1, \ldots, p)$ variable belongs to only one group C_q $(q = 1, \ldots, Q) \subset H_Q$. The latter statement implies that any triplet (i, j, l) of variables certainly falls into one of the following scenarios: (a) all the elements of the triplet belong to a single group C_q $(q = 1, \ldots, Q)$; (b) the elements of the triplet belong to two distinct groups C_q and $C_h \subset H_Q$ $(q, h = 1, \ldots, Q, h \neq q)$; (c) all the elements of the triplet belong to different groups $C_q, C_h, C_k \subset H_Q$ $(q, h, k = 1, \ldots, Q, k \neq h \neq q)$. According to the Definition 2.2 and 2.3 it can be gathered that (a), (b) and (c) correspond to the following correlation triplets: $({}_{W}r_{qq},\ {}_{W}r_{qq},\ {}_{W}r_{qq})$, $({}_{W}r_{qq},\ {}_{B}r_{qh},\ {}_{B}r_{qh})$ and $({}_{B}r_{qh},\ {}_{B}r_{qk},\ {}_{B}r_{hk})$, respectively. Furthermore, considering a hierarchy of latent concepts H_Q, all the triplets previously pinpointed verify the ultrametric inequality by definition (i.e., condition (iii) and (iii') of the Definition 2.2). Thus, the $(2Q-1)$-ultrametric matrix $\mathbf{R}$ has Q levels ${}_{W}r_{qq}$ $(q = 1, \ldots, Q)$ corresponding to the subsets of C, while the remaining $Q-1$ levels ${}_{B}r_{qh}$ $(q, h = 1, \ldots, Q,\ h \neq q)$ match the subsets $C_{Q+1}, \ldots, C_{2Q-1}$.

Conversely, each $(2Q-1)$-ultrametric correlation matrix $\mathbf{R}$ leads to a hierarchy of latent concepts, because for each pair of variables (j, l) they belong to the same group $C_q \subset C$ if their correlation is ${}_{W}r_{qq}$, or to different groups $C_q, C_k \subset C$ if their correlation is ${}_{B}r_{qk}$. Moreover, $C_{Q+1}, \ldots, C_{2Q-1}$ may contain only the triplets previously defined that are associated to non-overlapping groups since the ultrametricity constraint on $\mathbf{R}$. □



(a) $(2Q-1)$*-ultrametric correlation matrix* **(b)** *Hierarchy of latent concepts: path diagram representation*

Figure 2.1. Relationship between a $(2Q-1)$-ultrametric correlation matrix and a corresponding hierarchy of latent concepts.

Theorem 2.1. *The matrix* $\mathbf{R}_{\mathrm{u}}$ *defined in Eq. (2.2) is a* $(2Q-1)$*-ultrametric correlation matrix.*

Proof. The matrix $\mathbf{R}_{\mathrm{u}}$ defined in Eq. (2.2) can be rewritten as

$$\mathbf{R}_{\mathrm{u}} = \mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q + \mathbf{R}_{\mathrm{W}})\mathbf{V}' - \mathrm{diag}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}') + \mathbf{I}_p. \tag{2.9}$$

According to constraints (2.5)-(2.6) and the ultrametricity of $\mathbf{R}_{\mathrm{W}}$, it is easily to demonstrate that $\mathbf{Q} = \mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q + \mathbf{R}_{\mathrm{W}}$ turns out to be ultrametric. Indeed, $\mathbf{Q}$ is symmetric as well as $\mathbf{R}_{\mathrm{B}}$, since constraint (2.5) holds; it is column pointwise diagonally dominant since its diagonal elements are those of $\mathbf{R}_{\mathrm{W}}$ and the constraint (2.6) holds; the ultrametric inequality holds for $\mathbf{Q}$ observing that its possible triplets are contained in $\mathbf{R}_{\mathrm{B}}$ that is assumed to be ultrametric. Thus, $\mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q + \mathbf{R}_{\mathrm{W}})\mathbf{V}'$ is in turn ultrametric because it may contain the only triplets defined in Lemma 2.1, and the remaining addends of Eq. (2.9) affect only the diagonal of $\mathbf{R}_{\mathrm{u}}$, whose elements turn out to be unitary. Thus, the conditions of the Definition 2.2 and 2.3 are satisfied and the $(2Q-1)$-ultrametricity of $\mathbf{R}_{\mathrm{u}}$ is proved.

Moreover, property (ii), i.e., the column pointwise diagonal dominance, is a sufficient condition for an ultrametric matrix to be positive semi-definite (Dellacherie, Martínez, & San Martín, 2014, pp. 60-61). The matrix $\mathbf{R}_{\mathrm{u}}$ is nonnegative, with unitary diagonal elements and it is an ultrametric matrix since it is defined according to Eq. (2.2) subject to constraints (2.3)-(2.6); then, it is a correlation matrix. □

An example of a $(2Q-1)$-ultrametric correlation matrix $\mathbf{R}_{\mathrm{u}}$ and its corresponding hierarchical representation - one-to-one correspondence defined in Lemma 2.1 - is shown in Figure 2.1. In this example, four main groups which are internally strongly correlated are visible, i.e. $Q = 4$, in Figure 2.1a; indeed, the four within-concept consistency coefficients are $\{_W r_{11}, _W r_{22}, _W r_{33}, _W r_{44}\} = \{0.8970, 0.9131, 0.9105, 0.7602\}$, that correspond to the first four levels, starting from above, in Figure 2.1b, whereas

the hierarchical relationships between them are pinpointed through the matrix $\mathbf{R}_{\mathrm{B}}$ which has three different off-diagonal values, i.e., ${}_{B}r_{23} = 0.7068$, ${}_{B}r_{24}$, ${}_{B}r_{34} = 0.5455$ and ${}_{B}r_{12}$, ${}_{B}r_{13}$, ${}_{B}r_{14} = 0.3$, corresponding to the last three levels in Figure 2.1b. As shown in this example, UCM is based on a reflective approach since it assumes the existence of a general concept that causes some nested specific ones differently correlated to each other, as it happens in many applications. Nevertheless, the values of the correlation between concepts, i.e., the off-diagonal elements of $\mathbf{R}_{\mathrm{B}}$, in some situations can be very low and therefore close to zero providing the researcher with useful information about the formative nature of the general latent concept.

In the next section, the estimates of the proposed model defined in Eq. (2.2) are provided according to a non-parametric least-squares approach.

2.4 Least-Squares estimation of the model and algorithm

The least-squares estimation of the model (2.2), which provides the $(2Q-1)$-ultrametric approximation ($\mathbf{R}_{\mathrm{u}}$, see Theorem 2.1) of the nonnegative correlation matrix $\mathbf{R}$ by identifying its hierarchical structure of latent concepts, is defined as the minimization of the following constrained quadratic problem with respect to $\mathbf{R}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}$ and $\mathbf{V}$

$$\begin{aligned} F(\mathbf{R}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}, \mathbf{V}) =& \parallel \mathbf{R} - \mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q)\mathbf{V}' - \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' + \mathrm{diag}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}') - \mathbf{I}_p \parallel^2 \\ =& \sum_{q=1}^{Q}\sum_{j=1}^{p}\sum_{\substack{l=1 \\ l\neq j}}^{p}(r_{jl} - {}_{W}r_{qq})^2 v_{jq}v_{lq} \\ &+ \sum_{q=1}^{Q}\sum_{\substack{h=1 \\ h\neq q}}^{Q}\sum_{j=1}^{p}\sum_{\substack{l=1 \\ l\neq j}}^{p}(r_{jl} - {}_{B}r_{qh})^2 v_{jq}v_{lh} \quad \rightarrow \quad \min_{\mathbf{R}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}, \mathbf{V}} \end{aligned} \tag{2.10}$$

subject to constraints (2.3)-(2.6).

Therefore, before describing the algorithm to solve the aforementioned constrained quadratic optimization problem, the expression of the parameter estimators is provided.

Estimation of $\mathbf{R}_{\mathrm{W}}$

The estimators of the elements of $\mathbf{R}_{\mathrm{W}}$ are computed by differentiating Eq. (2.10) with respect to ${}_{W}r_{qq}$ $(q = 1, \ldots, Q)$ for a fixed $\widehat{\mathbf{V}}$:

$${}_{W}\hat{r}_{qq} = \frac{\sum_{j=1}^{p}\sum_{\substack{l=1 \\ l\neq j}}^{p} r_{jl}\hat{v}_{jq}\hat{v}_{lq}}{\sum_{j=1}^{p}\sum_{\substack{l=1 \\ l\neq j}}^{p} \hat{v}_{jq}\hat{v}_{lq}} \quad q = 1, \ldots, Q. \tag{2.11}$$

The above estimate of $\mathbf{R}_{\mathrm{W}}$ can be also expressed in a matrix form. Indeed, for fixed $\widehat{\mathbf{R}}_{\mathrm{B}}$ and $\widehat{\mathbf{V}}$, the loss function (2.10) can be rewritten as

$$F(\mathbf{R}_{\mathrm{W}}, \widehat{\mathbf{R}}_{\mathrm{B}}, \widehat{\mathbf{V}}) = \parallel \tilde{\mathbf{R}} - \widehat{\mathbf{V}}\mathbf{R}_{\mathrm{W}}\widehat{\mathbf{V}}' + \mathrm{diag}(\widehat{\mathbf{V}}\mathbf{R}_{\mathrm{W}}\widehat{\mathbf{V}}') \parallel^2, \tag{2.12}$$

where $\tilde{\mathbf{R}} = \mathbf{R} - \widehat{\mathbf{V}}(\widehat{\mathbf{R}}_{\mathrm{B}} - \mathbf{I}_Q)\widehat{\mathbf{V}}' - \mathbf{I}_p$ is a known residual matrix. Eq. (2.12) is minimized by

$$\widehat{\mathbf{R}}_{\mathrm{W}} = \mathrm{diag}(\widehat{\mathbf{V}}'(\mathbf{R} - \mathbf{I}_p)\widehat{\mathbf{V}})[(\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^2 - \widehat{\mathbf{V}}'\widehat{\mathbf{V}}]^+, \tag{2.13}$$

where $[\mathbf{A}]^+$ denotes the Moore-Penrose inverse of a matrix $\mathbf{A}$. The estimates in Eq. (2.11) are equivalent to the diagonal elements of $\widehat{\mathbf{R}}_{\mathrm{W}}$ in Eq. (2.13) and they must satisfy constraint (2.6).

Estimation of $\mathbf{R_B}$

The estimators of the elements of $\mathbf{R}_{\mathrm{B}}$ are computed by differentiating Eq. (2.10) with respect to ${}_Br_{qh}$ $(q, h = 1, \ldots, Q,\ h \neq q)$ for a fixed $\widehat{\mathbf{V}}$:

$${}_B\hat{r}_{qh} = \frac{\sum_{j=1}^{p}\sum_{\substack{l=1\\ l\neq j}}^{p} r_{jl}\hat{v}_{jq}\hat{v}_{lh}}{\sum_{j=1}^{p}\sum_{\substack{l=1\\ l\neq j}}^{p} \hat{v}_{jq}\hat{v}_{lh}} \quad q, h = 1, \ldots, Q, h \neq q. \tag{2.14}$$

The above estimate of $\mathbf{R}_{\mathrm{B}}$ can be also expressed in a matrix form. Indeed, for fixed $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{V}}$, the loss function (2.10) can be rewritten as

$$F(\widehat{\mathbf{R}}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}, \widehat{\mathbf{V}}) = \| \tilde{\mathbf{R}} - \widehat{\mathbf{V}}\mathbf{R}_{\mathrm{B}}\widehat{\mathbf{V}}' \|^2, \tag{2.15}$$

where $\tilde{\mathbf{R}} = \mathbf{R} - \widehat{\mathbf{V}}\widehat{\mathbf{R}}_{\mathrm{W}}\widehat{\mathbf{V}}' + \mathrm{diag}(\widehat{\mathbf{V}}\widehat{\mathbf{R}}_{\mathrm{W}}\widehat{\mathbf{V}}') - \mathbf{I}_p + \widehat{\mathbf{V}}\,\mathbf{I}_Q\widehat{\mathbf{V}}'$ is a known residual matrix. The minimization of (2.15) is a Penrose multivariate regression problem with the following matricial solution

$$\tilde{\mathbf{R}}_{\mathrm{B}} = (\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{V}}'\tilde{\mathbf{R}}\widehat{\mathbf{V}}(\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^{-1}. \tag{2.16}$$

The elements of $\tilde{\mathbf{R}}_{\mathrm{B}}$ simply define the correlations between $C_1, \ldots, C_Q$. Since constraint (2.5) must be satisfied, the estimate of $\mathbf{R}_{\mathrm{B}}$ is the closest - in the LS sense - ultrametric matrix to Eq. (2.16). This ultrametric solution, which corresponds to $\widehat{\mathbf{R}}_{\mathrm{B}}$, is computed via an average linkage UPGMA algorithm (Sokal & Michener, 1958), but for correlations, such that the Definition 2.2 is satisfied. Thus, the elements of $\widehat{\mathbf{R}}_{\mathrm{B}}$ represent the levels of correlation between the groups of H_Q (see Definition 2.1).

Estimation of V

The estimators of the elements of $\mathbf{V}$ are computed by minimizing Eq. (2.10) row by row for each $\mathbf{v}_j$ $(j = 1, \ldots, p)$ of $\mathbf{V}$, when all the remaining rows are fixed and the corresponding $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}$ have been estimated. Specifically, the jth variable is assigned to the qth group ($v_{jq} = 1$) if Eq. (2.10) reaches its minimum with respect to $\mathbf{V}$, after the corresponding estimation of $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}$. Formally, each row $\mathbf{v}_j$ of $\mathbf{V}$, $j = 1, \ldots, p$, is estimated as follows

$$\begin{cases} \hat{v}_{jq} = 1 & \text{if} \quad \underset{q=1,\ldots,Q}{\arg\min}\, F(\widehat{\mathbf{R}}_{\mathrm{W}}, \widehat{\mathbf{R}}_{\mathrm{B}}, [\hat{\mathbf{v}}_1, \ldots, \mathbf{v}_j = \mathbf{i}_q, \ldots, \hat{\mathbf{v}}_p]') \\ \hat{v}_{jq} = 0 & \text{otherwise} \end{cases} \tag{2.17}$$

where $\mathbf{i}_q$ is the qth row of the identity matrix of order Q and $\widehat{\mathbf{R}}_{\mathrm{W}}$, $\widehat{\mathbf{R}}_{\mathrm{B}}$ are the estimates of the within-concept consistency and between-concept correlation matrices, respectively, which correspond to the configuration of $\widehat{\mathbf{V}}$.

Algorithm 1: LS Estimate of the Ultrametric Correlation Model

Input: $\mathbf{R}$, Q, Random Starts

Fixed values $\epsilon \leftarrow$ *small nonnegative convergence tolerance value;*
maxiter $\leftarrow$ *maximum number of iterations;*
for $i = 1$ *to* Random Starts **do**
  Initialization $t \leftarrow 0$
  $\mathbf{V}^{(0)} \leftarrow$ *random initial partition of variables in* Q *non-empty groups s.t. constraints (2.3)-(2.4) hold;*
  $\mathbf{R}_{\mathrm{W}}^{(0)} \leftarrow$ *Eq. (2.13) subject to constraint (2.6), given* $\mathbf{V}^{(0)}$*;*
  $\mathbf{R}_{\mathrm{B}}^{(0)} \leftarrow$ *Eq. (2.16) subject to constraint (2.5), given* $\mathbf{V}^{(0)}$*;*
  if *Constraint (2.6) does not hold* **then**
    $min\{_{W}r_{qq}^{(0)} : q = 1, \ldots, Q\} \leftarrow max\{_{B}r_{qh}^{(0)} : q, h = 1, \ldots, Q, h \neq q\}^{*}$*;*
  end
  $F^{(0)} \leftarrow F(\mathbf{R}_{\mathrm{W}}^{(0)}, \mathbf{R}_{\mathrm{B}}^{(0)}, \mathbf{V}^{(0)})$ *through Eq. (2.10);*
  $F_{diff}^{(0)} \leftarrow F^{(0)}$*;*
  $F_{min} \leftarrow F^{(0)}$*;*
  while $F_{diff}^{(t)} > \epsilon$ *and* $t \leq maxiter$ **do**
    $t \leftarrow t + 1$;
    $\mathbf{V}_{temp}^{(t)} \leftarrow \mathbf{V}^{(t-1)}$;
    for $j = 1$ *to* p **do**
      $currentq \leftarrow q$th group the jth variable belongs to;
      for $q = 1$ *to* Q **do**
        $\mathbf{V}_{temp}^{(t)}(j, :) \leftarrow \mathbf{i}_q$;
        if $\mathbf{V}_{temp}^{(t)}(:, currentq)$ *is nonempty* **then**
          Step 1 Update $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$
            $\mathbf{R}_{\mathrm{W;temp}}^{(t)} \leftarrow$ Eq. (2.13) subject to constraint (2.6), given $\mathbf{V}_{temp}^{(t)}$;
            $\mathbf{R}_{\mathrm{B;temp}}^{(t)} \leftarrow$ Eq. (2.16) subject to constraint (2.5), given $\mathbf{V}_{temp}^{(t)}$;
            if *Constraint (2.6) does not hold* **then**
              $min\{_{W;temp}r_{qq}^{(t)} : q = 1, \ldots, Q\} \leftarrow max\{_{B;temp}r_{qh}^{(t)} : q, h = 1, \ldots, Q, h \neq q\}^{*}$;
            end
          endStep1
          $F_{temp}^{(t)} \leftarrow F(\mathbf{R}_{\mathrm{W;temp}}^{(t)}, \mathbf{R}_{\mathrm{B;temp}}^{(t)}, \mathbf{V}_{temp}^{(t)})$ through Eq. (2.10);
          if $F_{temp}^{(t)} < F_{min}$ **then**
            $F^{(t)} \leftarrow F_{temp}^{(t)}$;
            $F_{min} \leftarrow F^{(t)}$;
            $currentq \leftarrow q$;
            $\mathbf{R}_{\mathrm{W}}^{(t)} \leftarrow \mathbf{R}_{\mathrm{W;temp}}^{(t)}$;
            $\mathbf{R}_{\mathrm{B}}^{(t)} \leftarrow \mathbf{R}_{\mathrm{B;temp}}^{(t)}$;
          end
        end
      end
      Step 2 Update $\mathbf{V}$
        $\mathbf{V}^{(t)}(j, :) \leftarrow \mathbf{i}_{currentq}$;
      endStep2
    end
    $F_{diff}^{(t)} \leftarrow (F^{(t-1)} - F^{(t)})$;
  end
end

* It is worthy to notice that if the aforementioned replacement occurs, the relationship between the standardized Cronbach's α of C_q and the within-consistency coefficient $_{W}r_{qq}$ stressed in Section 2.2 does not hold.

Output: $\widehat{\mathbf{R}}_{\mathrm{W}}, \widehat{\mathbf{R}}_{\mathrm{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{R}}_{\mathrm{u}}, F(\widehat{\mathbf{R}}_{\mathrm{W}}, \widehat{\mathbf{R}}_{\mathrm{B}}, \widehat{\mathbf{V}})$ corresponding to the optimal solution among the whole ones obtained running *Random Starts* times the algorithm.

2.4.1 Algorithm for detecting the LS Ultrametric Correlation Model

The Algorithm 1 for the LS estimation of the UCM parameters is composed of two main steps: one to update the parameters representing the continuous part of the optimization problem in Eq. (2.10), i.e., $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$, conditionally on the configuration of $\widehat{\mathbf{V}}$ and subject to constraints (2.5) and (2.6); the other one to update the parameter representing the combinatorial part of the optimization problem in Eq. (2.10) subject to constraints (2.3) and (2.4), i.e., $\mathbf{V}$, and conditionally on the corresponding $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}$. These two main steps are iteratively repeated after the initialization step - that starts from a random initial partition of the variable space, i.e. $\mathbf{V}^{(0)}$, since the estimates of $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$ are based upon it - and at each iteration the objective function in Eq. (2.10) does not increase and generally decreases until the convergence to a stationary point. The latter is at least a local minimum; thus, to improve the chance to reach a global optimum, the algorithm is run several times starting from different random initializations of the parameters (*Random Starts* input parameter of the Algorithm 1).

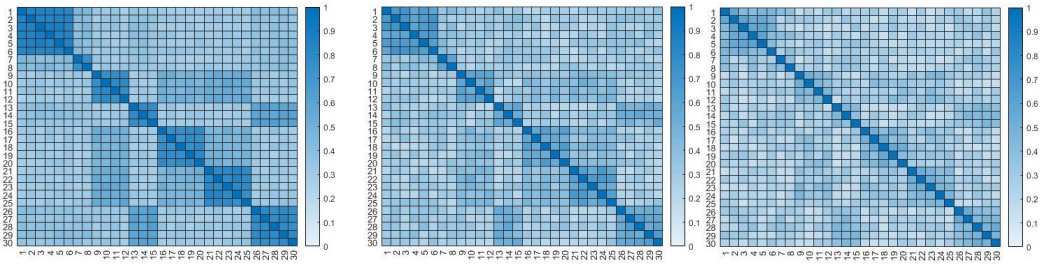
It is worthy to notice that the computational time and space are reduced relative to an algorithm on a data matrix, since at most $2Q - 1 + p$ different elements per iteration are stored.

2.5 Simulation

In order to assess the performances of the Ultrametric Correlation Model, we have implemented a simulation study by following the ultrametric correlation structure determined in Eq. (2.2). Two different scenarios have been taken into account to test the model with a small scale correlation structure, characterized by $p = 30$ and $Q = 4$ (small number of groups), 7 (large number of groups), and with a larger one, with $p = 100$ and $Q = 7$ (small number of groups), 15 (large number of groups).

The matrices of the model (2.1) have been defined as follows. The diagonal elements of $\mathbf{R}_{\mathrm{W}}$ have been generated as ${}_{W}r_{qq} = 0.85 + 0.1\,a$ where $a \sim N(0,1)$, $q = 1, \ldots, Q$, whereas the off-diagonal elements of $\mathbf{R}_{\mathrm{B}}$ have been set in the interval $[0.3, 0.7]$ such that the difference between pairs of two sequential correlation coefficients turns out to be equal and constraint (2.6) holds. In this way, the correct estimation of the hierarchical structure depends only on the noise, i.e., on the error matrix $\mathbf{E}$ which has been generated in turn starting from a uniform distribution in the interval $[0, \sigma_E]$ - the matrix has been symmetrized and its positive semi-definiteness has been verified. The membership matrix $\mathbf{V}$ has been randomly generated such that no constraint on the variable groups, e.g., the number of variables in each group, has been put. Three levels of error σ_E were fixed: $\sigma_E^S = 0.1$ (small error), $\sigma_E^M = 0.3$ (medium error) and $\sigma_E^H = 0.5$ (high error). In Figure 2.2, an example of the meaning of the error levels is illustrated. Indeed, the groups and their hierarchical structure are clearly visible when the small error is added to $\mathbf{R}_{\mathrm{u}}$ and tend to be less visible as the error grows. For each generated matrix $\mathbf{R}$ according to Eq. (2.1), we have verified if it has turned out to be a proper correlation matrix (i.e., positive semi-definite and with values between 0 and 1).

The simulated model is evaluated according to the Adjusted Rand Index (ARI, Hubert & Arabie, 1985), that compares the generated membership matrix $\mathbf{V}$ with



(a) $\sigma_E^S = 0.1$ **(b)** $\sigma_E^M = 0.3$ **(c)** $\sigma_E^H = 0.5$

Figure 2.2. Example of the heat maps of three (30×30) correlation matrices produced by the simulation study with different levels of error. The theoretical number of groups of variables is $Q = 7$ (Scenario 1).

Table 2.1. Simulation study results.

	σ_E^S		σ_E^M		σ_E^H	
	Scenario 1					
Q	4	7	4	7	4	7
MSE($\mathbf{R}_{\mathrm{W}}$)	0.0012	$8.3e^{-4}$	0.0170	0.0108	0.0342	0.0213
MSE($\mathbf{R}_{\mathrm{B}}$)	$2.4e^{-4}$	$2.4e^{-4}$	0.0061	0.0056	0.0138	0.0126
% ARI $= 1$	100.0%	100.0%	100.0%	99.5%	68.5%	68.5%
ARI Mean	1.0000	1.0000	1.0000	0.9996	0.9591	0.9711
	Scenario 2					
Q	7	15	7	15	7	15
MSE($\mathbf{R}_{\mathrm{W}}$)	0.0049	0.0034	0.0243	0.0145	0.0377	0.0211
MSE($\mathbf{R}_{\mathrm{B}}$)	0.0045	0.0062	0.0297	0.0271	0.0398	0.0401
% ARI$= 1$	100.0%	100.0%	100.0%	82.0%	92.5%	56.0%
ARI Mean	1.0000	1.0000	1.0000	0.9847	0.9987	0.9772

the estimated one $\widehat{\mathbf{V}}$. Furthermore, the Mean Squared Error (MSE) of the within-concept consistency matrix $\mathbf{R}_{\mathrm{W}}$ and the between-concept correlation matrix $\mathbf{R}_{\mathrm{B}}$ is computed. All the simulation study results are shown in Table 2.1.

We have generated 200 correlation matrices for each scenario (i.e., for each pairs (p, Q) and each level of error). The variable partition of the model with a small level of error is completely reconstructed (100% of samples with ARI equal to 1) in both scenarios and it is always correctly detected. Whereas, when the error grows it tends to mask the generated correlation structure as shown in Figure 2.2, and the detection is not always correct. Nevertheless, the misclassification concerns a reduced number of samples (at least 68.5% and 56% of samples with ARI equal to 1 in the first and the second scenario, respectively) and variables (the mean of the ARI is always greater than 0.9). The MSE for $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$ is extremely good.

In the whole scenarios the number of random starts has been set equal to 50.

The algorithm has shown impressive performances in terms of running time and it converges in few steps; thus, the chosen number of random starts turns out to be enough to find the global optimum of Eq. (2.10). Nevertheless, it is advisable to increase the initial random starts when the correlation structure is not clearly determined.

2.6 Application

In this section two real data examples are considered. The first one is a benchmark data set concerning mental abilities (Section 2.6.1), whereas the second one regards drug consumption (Section 2.6.2). On the former, UCM is able to correctly detect the theoretical variable groups associated with the six primary mental abilities (latent concepts) and allows to investigate their hierarchical relationships. In the second application, UCM is implemented to study drug consumption by identifying groups of drugs highly correlated and their hierarchical relationships. An exploratory in-depth analysis of this phenomenon through the proposed model can contribute to its better understanding and to consequently implement policies aimed at reducing it.

2.6.1 Bechtoldt data set

The Bechtoldt data set contains 17 variables, shown in Table 2.2, with 6 theoretical latent factors (Bechtoldt, 1961; Kano, 1997). Our goal is to correctly estimate the variable partition by identifying the 6 latent concepts - *Memory, Verbal, Words, Space, Number, Reasoning* - and to explore the hierarchical structure of the Bechtoldt correlation matrix in order to understand the relationships among the variable groups.

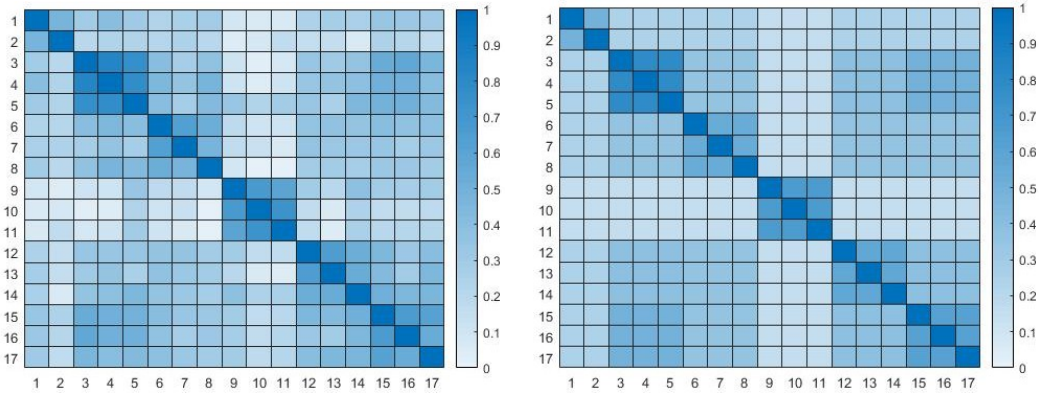
The data set is presented as a correlation matrix with all nonnegative values (Figure 2.3a). The algorithm run with 50 random starts has found the right partition in 6 main groups (Figure 2.3b) corresponding to the theoretical ones. All the groups of variables are reliable except the first, i.e. *Memory*, whose Cronbach's α is equal to 0.6413.

The hierarchical structure pinpointed by the model starts from the first aggregation of two groups, corresponding to the latent concepts *Verbal* and *Reasoning*, to which the other concepts are merged one-by-one, as shown in Figure 2.4. *Verbal* turns out to be the most reliable concept (the correlation within the group is 0.7887), whereas *Space*, whose correlation within the corresponding group is equal to 0.6687, is the concept less correlated with the others (the correlation between Space and the other concepts is on average 0.1464). It is worthy to notice that even if the path diagram in Figure 2.4 shows a constant trend in the definition of broader dimensions, we could characterize three main concepts: (i) *Intellectual Abilities* composed of the concepts *Verbal, Reasoning, Number, Words* (the correlation between them is on average 0.346), (ii) *Memory* and (iii) *Space* that are defined according to the theoretical groups of the original variables. This result can be confirmed by the analysis of the Cronbach's α at each level of the hierarchy, which increases up to the third bottom-up level (0.8991) and then decreases (0.8946). Nevertheless, by examining the path diagram in Figure 2.4 it can be observed the existence of a

Table 2.2. List of Bechtoldt variables.

Specific Concept	Variables	ID	Specific Concept	Variables	ID
Memory	First Names	1	Space	Flags	9
	Word Number	2		Figures	10
Verbal	Sentences	3		Cards	11
	Vocabulary	4	Number	Addition	12
	Completion	5		Multiplication	13
Words	First Letters	6		Three Higher	14
	Four Letter Words	7	Reasoning	Letter Series	15
	Suffixes	8		Pedigrees	16
				Letter Grouping	17

general concept, that is clearly evident and it may represent the dimension *Mental Abilities*.



(a) *Bechtoldt correlation matrix* **(b)** *Estimated correlation matrix via the UCM*

Figure 2.3. Comparison between heat maps of the observed and estimated correlation matrices.

As shown in Figure 2.3, the model (2.2) is able to reconstruct the correlation matrix of the observed variables with the main advantage of identifying its hierarchical structure through the definition of the internal consistency of the concepts - each one associated with a group of variables - and the correlations between them. Evidently, the use of the matrix $\mathbf{R}_{\mathrm{u}}$ defined in Eq. (2.2) entails biased point estimates of each element of the original matrix due to the ultrametricity assumption. Nevertheless, UCM can help the researcher to identify the hypothesized hierarchy of latent concepts, e.g., in many psychometric applications, and to define and characterize broader dimensions.

2.6.2 Drug consumption data set

Drug consumption is one of the most challenging problems in the modern societies. Indeed, it contributes to rise the risk of poor health, crimes, social harm, environ-

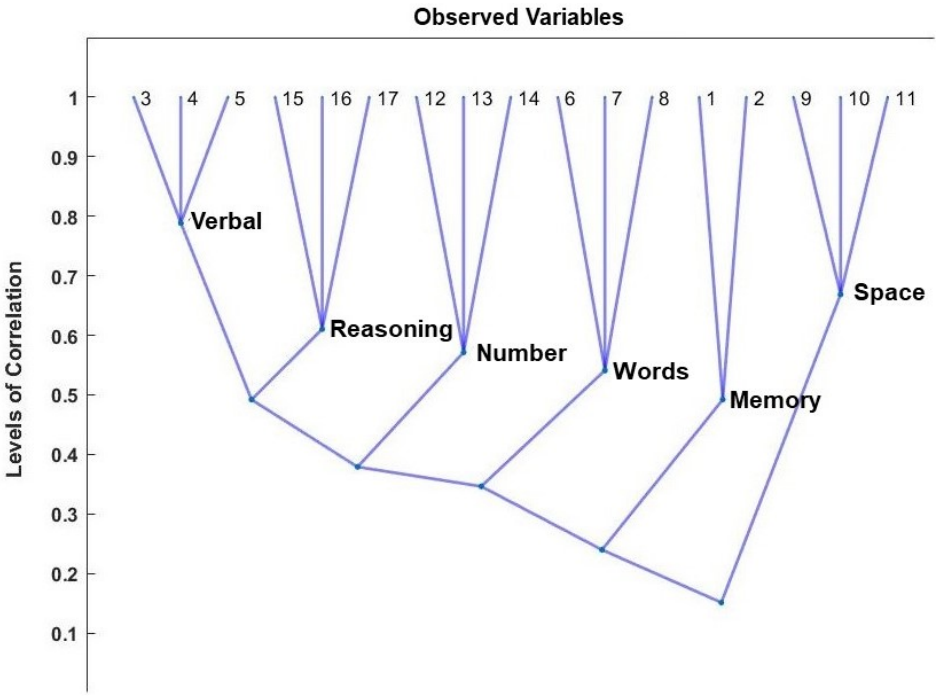


Figure 2.4. Path diagram of the Bechtoldt hierarchical structure.

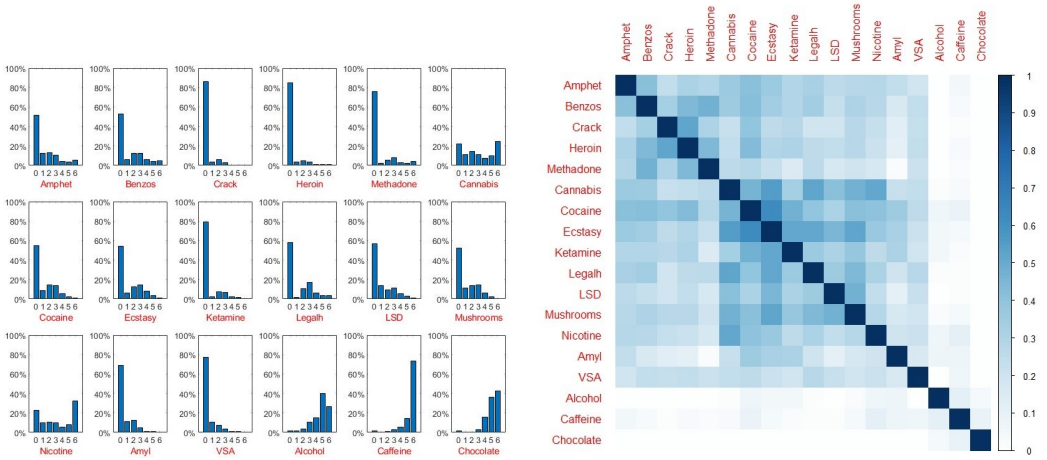
mental damage and it has become a social problem over years - especially among young people - governments have to face with. Many studies have been developed to analyze this phenomenon, its individual and community effects, e.g., McGinnis and Foege (1993).

The data set analyzed in this section[4] (Fehrman et al., 2015) contains information on 1885 respondents, mainly coming from UK (55.58%), USA (29.55%), Canada (4.62%) and Australia (2.86%) and aged from 18 years old, on their drug consumption. Specifically, the use of 18 legal (alcohol, caffeine, chocolate, nicotine) and illegal (amphetamines, amyl nitrite, benzodiazepine, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, Volatile Substance Abuse) drugs is investigated in terms of ordinal variables. The response classes are the following: *Never Used, Used over a Decade Ago, Used in Last Decade, Used in Last Year, Used in Last Month, Used in Last Week* and *Used in Last Day*.

In order to apply the methodology described in Section 2.3 to investigate the correlation structure among drugs, the ordinal variables - each one representing consumption of a specific drug - have to be quantified. This quantification is implemented via the Categorical Principal Component Analysis (CatPCA) (Gifi, 1990) and the correlation matrix of the corresponding quantitative variables is computed. Six correlation coefficients assume negative values (not lower than[5] -0.05) which turn out to be statistically nonsignificant; whereas, the variable *Chocolate* has

[4]Drug consumption (quantified) data set available at: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29.

[5]In this case, the term *not lower than* refers to small negative correlation coefficients close to zero.



(a) Bar chart of the 18 ordinal variables

(b) Correlation matrix of the quantified data via the CatPCA

Figure 2.5. Drug consumption data set.

Table 2.3. Initial five groups identified by the Ultrametric Correlation Model.

Group	Group Name	Variables
Group 1	Depressant and Artificial Drugs	Ampeth, Benzodiazepine, Crack, Heroine, Methadone
Group 2	Stimulant Drugs and Hallucinogens	Cannabis Cocaine, Ecstasy, Ketamine, Legal highs, LSD, Mushrooms, Nicotine
Group 3	Inhalant Drugs	Amyl nitrite, Volatile Substance Abuse
Group 4	Legal Drugs of Daily Use	Alcohol, Caffeine
Group 5	Chocolate	Chocolate

negative correlations with all the other drugs (Figure 2.5a) - except for *Alcohol* and *Caffeine* - which are not lower than -0.09 and considered nonsignificant in literature (Fehrman et al., 2015). For this reason, in both cases the negative correlations are set to zero such that the non-negativity condition necessary for UCM holds (Figure 2.5b). Furthermore, the number of the variable groups necessary to implement the exploratory, parsimonious model described in Section 2.3 is set according to the scree plot and it is equal to five. It is worthy of remark that hierarchical clustering methods could be implemented to study the correlation between usage of different drugs, but they would not guarantee the correct identification of the underlying hierarchical structure as we will see in Chapter 3.

The application of UCM to the aforementioned data set provides a representation of drug consumption through the identification of different groups of drugs mostly correlated (Figure 2.5b), and broader ones defined by merging the initial five groups (Figure 2.6). In this framework, a model-based approach to analyze correlations between variables can back up the experts' theories on this phenomenon. The initial five groups identified by the model are reported in Table 2.3. All of them are reliable according to the Cronbach's α, except for *Inhalant Drugs* and *Legal Drugs of Daily Use*. It is worthy of remark that the Cronbach's α of a group is affected by its number of variables.

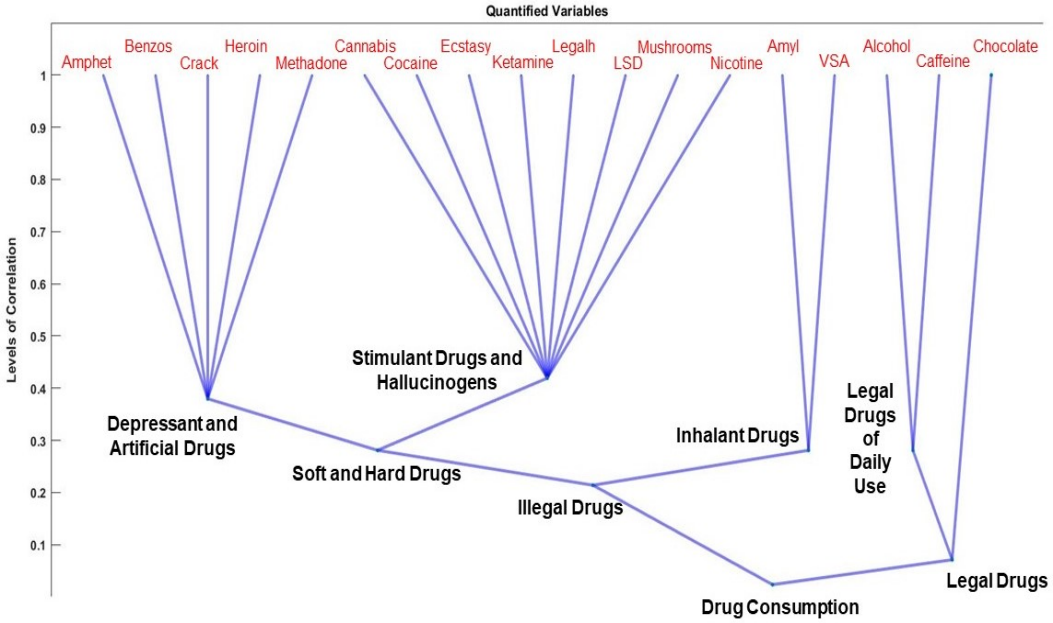


Figure 2.6. Path diagram representation of the drug consumption.

The hierarchy over the five groups gives rise to broader concepts: *Soft and Hard Drugs* obtained by lumping together Group 1 and Group 2 ($\alpha = 0.87$); *Illegal Drugs* obtained by merging the latter with Group 3 ($\alpha = 0.87$); *Legal Drugs* obtained by lumping together Group 4 and Group 5 ($\alpha < 0.7$). The existence of a general concept representing *Drug Consumption* is assessed through the Cronbach's α of the whole data set, which is equal to 0.84. These results turn out to be coherent with the specialized literature on drug consumption (e.g., Fehrman et al., 2015).

2.7 Conclusions

The model proposed herein allows to investigate the hierarchical structure of a non-negative correlation matrix of observed variables via an ultrametric correlation one in an exploratory, simultaneous and non-compensatory approach. In psychometric studies, many multidimensional phenomena underlie a hierarchy of latent concepts that defines a general concept through the identification of more specific ones. In this field, the non-negativity assumption turns out to be realistic.

The ultrametric correlation structure of the model allows to disclose a parsimonious hierarchy from the observed variables up to the most general concept, composed of all variables. The identification of this hierarchical structure of latent concepts - each one associated with a group of variables - is based upon the definition of two main features: the within-concept consistency and the between-concept correlation. These two characteristics pinpoint the reliability of the concepts and their hierarchical relationships, respectively, and they allow to define broader dimensions starting from the initial ones. Moreover, a relationship between these features and the Cronbach's α is provided. The methodology is developed in a reflective approach since it assumes the existence of a general concept that causes some nested specific

ones, differently correlated to each other, but very low or close to zero values can occur in the ultrametric matrix providing the researcher with a useful information about the formative nature of the general latent concept.

A LS estimation of UCM is proposed in order to detect consistent latent concepts and the correlation between them. Hence, the whole hierarchy of latent concepts is built by the levels of correlation. The simulation study and the two applications provided show the good performances of the model. Furthermore, the algorithm for the UCM estimation results very fast and stable.

A further development of the model presented in this chapter is its extension to general correlation (or covariance) matrices by relaxing the non-negativity constraint. In Chapter 4 the definition of an ultrametric matrix will be extended to a generic covariance matrix and then implemented into a Gaussian mixture model, in order to identify a different characterization of a multidimensional phenomenon in heterogeneous populations.

Chapter 3

Exploring hierarchical concepts: theoretical and application comparisons

3.1 Introduction

The investigation of the relationships between latent concepts defining a multidimensional phenomenon is the aim of the model proposed in Chapter 2, in which a parsimonious hierarchy of nested partitions of the variable space is formally defined via an ultrametric matrix. As already introduced in Chapter 1 (Section 1.2), an ultrametric matrix (Definition 2.2) does not correspond to an ultrametric distance matrix (Definition 1.1), even if there exists a relationship between the two. However, the researchers could think to use a procedure based on a classical (agglomerative) hierarchical clustering algorithm to inspect the hierarchical relationships among observed variables (see Chapter 1, Section 1.1). The latter can be implemented building the whole hierarchy from p observed variables up to the most general latent concept, and cutting the tree identifying Q main latent concepts by maintaining the corresponding hierarchy. Nonetheless, this sequential strategy does not guarantee an optimal solution since the classification errors made in the first steps can never be corrected; indeed, hierarchical clustering algorithms can be characterized as greedy. It is worthy of remark that the study of multidimensional phenomena needs the detection of *reliable* latent concepts together with the corresponding hierarchy.

In this chapter we compare the above described procedure based on traditional agglomerative clustering methods - in particular, single linkage (Florek et al., 1951), complete linkage (McQuitty, 1960), Ward's method (Ward, 1963) and average linkage (Sokal & Michener, 1958) - with the Ultrametric Correlation Model (UCM) proposed by Cavicchia, Vichi, and Zaccaria (2020b) and presented in Chapter 2. Their application to a benchmark data set made up of variable groups identifying latent concepts highlights the potential of UCM with respect to the classical hierarchical clustering algorithms. Furthermore, the aforementioned proposal is based upon a parsimonious representation of the relationships among variables which allows to reduce the time complexity of the bottom-up algorithms.

The chapter is organized as follows. In Section 3.2, a brief review of the four

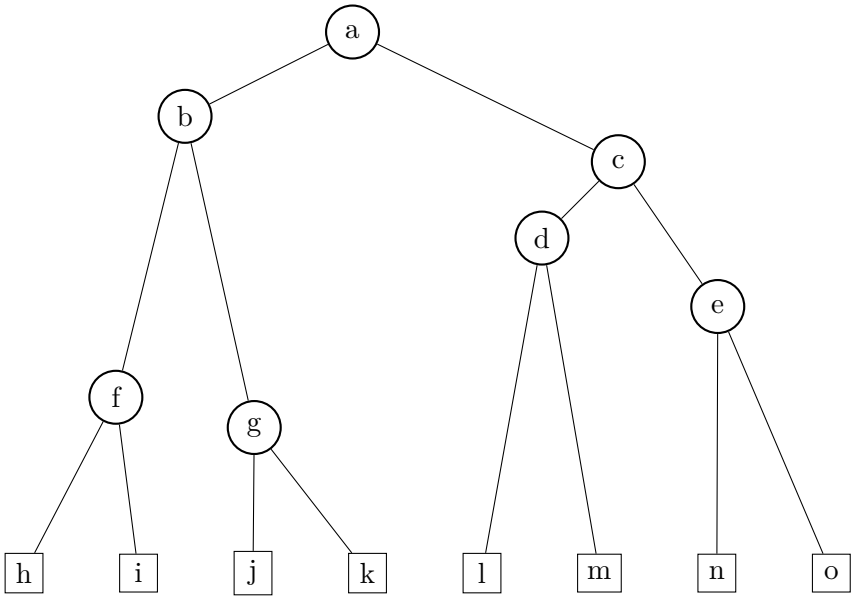


Figure 3.1. N-tree representation: root node (a), internal nodes (b, c, d, e, f, g), terminal nodes (h, i, j, k, l, m, n, o).

aforementioned agglomerative clustering methods is depicted and UCM is briefly recalled in Section 3.3. Section 3.4 provides a deep comparison among the aforementioned methods via a benchmark data set, in order to highlight their advantages and weaknesses in searching for hierarchical relationships among variables - not only with the clustering objective - associated with a latent concept structure. A final discussion completes the chapter in Section 3.5.

3.2 Hierarchical classification of variables

Hierarchical classification defines a set of methods that have been proposed to pinpoint hierarchically-nested classes of units, even variables[1], by defining a set of partitions represented by tree-shape structures. For completeness, we firstly define an *n-tree* (Bobisud & Bobisud, 1972; McMorris, Meronk, & Neumann, 1983) as follows.

Definition 3.1. An n-tree on a set of objects $O = \{1, 2, \ldots, p\}$ is a set T of subsets of O satisfying the following conditions: $O \in T$, $\emptyset \notin T$, $\{j\} \in T \ \forall j \in O$ and $A \cap B \in \{\emptyset, A, B\} \ \forall A, B \in T$.

An n-tree is composed of a root node, which represents the whole set of objects, some internal nodes, which define the nested classes of objects, and the terminal nodes (leaves), which are the observed objects. All nodes are connected by branches as represented in Figure 3.1, with at most $p - 2$ internal nodes corresponding to a binary tree.

A particular n-tree, called *dendrogram*, is defined in Definition 3.2. This is the main graphical representation used in this chapter (see Section 3.4).

[1]In this chapter we use the term *objects* as a synonym of both units and variables.

Definition 3.2. A dendrogram is a valued n-tree, where given a mapping b on $\mathbb{R}^+$ any two internal nodes A and B of T, such that $A \cap B \neq \emptyset$, then $b(A) \leq b(B) \Leftrightarrow A \subset B$.

The hierarchical classification methods usually produce a complete dendrogram. Nonetheless, especially for large data sets, a complete hierarchy of nested partitions frequently has low interest. The construction of a *parsimonious* tree, that contains a limited number of internal nodes, is preferred and turns out to be clearer albeit the loss of information related to the dimensionality reduction (Gordon, 1999).

The hierarchical clustering algorithms we take into account are the *agglomerative* ones, whose criterion for the construction of the dendrogram starts from p singleton sets of objects and recursively merges two of them - from the bottom upwards - to obtain the whole hierarchy. All these methods are computed on a distance matrix, as a measure of dissimilarity, and they differ in the way of defining distance between two groups of objects (or between a group of objects and a singleton). It is worthy of remark that the distance matrices have diagonal elements equal to zero, nonnegative off-diagonal elements and they must be symmetric.

For variables, it is often suggested to use the correlation coefficient to quantify the similarity among them (e.g., Cliff et al., 1995; Gordon, 1999; Strauss, Bartko, & Carpenter, 1973). Therefore, even if the classical hierarchical clustering methods are defined for clustering units, they can be employed for classifying variables. Indeed, it is possible to transform a measure of similarity - the correlation coefficient in this case - into a dissimilarity between objects, as follows

$$d_{jh} = 1 - r_{jh} \quad \rightarrow \quad d_{jh} \in [0,1] \text{ when } r_{jh} \text{ is assumed to be nonnegative,} \tag{3.1}$$

where d_{jh} is the distance between the object $\{j\}$ and the object $\{h\}$ of O, and r_{jh} is their correlation coefficient. Moreover, if a similarity matrix is positive semi-definite, as the correlation matrix is, then the distance matrix defined by

$$d_{jh} = \sqrt{1 - r_{jh}} \tag{3.2}$$

is Euclidean (Gower, 1966).

The four hierarchical clustering methods we consider herein - single linkage, complete linkage, average linkage, Ward's method - can be obtained as special cases of the following equation proposed by Lance and Williams (1966, 1967), and generalized by Jambu (1978),

$$\begin{aligned} d(C_i \cup C_h, C_k) = {} & \alpha_i \, d(C_i, C_k) + \alpha_h \, d(C_h, C_k) + \beta \, d(C_i, C_h) \\ & + \gamma \, |d(C_i, C_k) - d(C_h, C_k)|, \end{aligned} \tag{3.3}$$

where C_i, C_h, C_k are clusters of objects of O with $1 \leq |C_i| \leq p-2$, $\forall \, C_i \in O$. The parameters $\alpha_i, \alpha_h, \beta, \gamma$ in Eq. (3.3) define different clustering techniques, as shown in Lance and Williams (1967) and Everitt et al. (2011).

All these methods are agglomerative techniques which do not produce *reversals* in the dendrogram representation, i.e., the following conditions for Lance and William's Eq. (3.3) hold:

$$\gamma \geq -\min\{\alpha_i, \alpha_h\},$$

$$\alpha_i + \alpha_h \geq 0,$$
$$\alpha_i + \alpha_h + \beta \geq 1.$$

Moreover, these methods - as Definition 3.2 in turn - satisfy a fundamental condition: the ultrametric property (e.g., Hartigan, 1967). This property may be expressed in two different ways, i.e., with respect to distances and the components of a dendrogram, respectively as follows:

$$d(C_i, C_h) \leq \max\{d(C_i, C_k), d(C_h, C_k)\} \quad C_i, C_h, C_k \in O, \tag{3.4}$$
$$b(A, B) \leq \max\{b(A, C), b(B, C)\} \quad A, B, C \in T. \tag{3.5}$$

Starting from a distance matrix, the hierarchical clustering algorithms produce a complete dendrogram. In this framework, the optimal number of clusters is chosen by cutting the n-tree at a specific level. For a deeper review of the hierarchical classification algorithms see Gordon (1987).

The procedure based on the above described hierarchical clustering algorithms for the classification of variables works as follows:

Step 1 (*Transformation of correlations into distances*) Given a nonnegative correlation matrix $\mathbf{R}$, the corresponding distance matrix is obtained by applying Eq. (3.1) w.r.t. the elements of $\mathbf{R}$.

Step 2 (*Hierarchical clustering algorithm*) According to Eq. (3.3), a hierarchical clustering algorithm is chosen and computed on the distance matrix defined in Step 1. A complete dendrogram and the corresponding estimated ultrametric distance matrix are obtained.

Step 3 (*Parsimonious hierarchy*) To define a parsimonious hierarchy in Q groups of variables, for a given Q, that may correspond to Q latent concepts, the dendrogram obtained in Step 2 is cut at the Qth level. The bottom-up aggregations from the aforementioned level upwards identify the parsimonious hierarchy.

Step 4 (*Model fit*) To evaluate the solution obtained in Step 3, the estimated ultrametric distance matrix has to be transformed into an ultrametric correlation matrix through the inverse relationship to that of Eq. (3.1). The least-squares difference between the nonnegative correlation matrix $\mathbf{R}$ and the estimated - according to the hierarchical clustering method chosen in Step 2 - correlation matrix is computed with respect to the total correlation of the data[2].

It is worth noticing that herein we consider a nonnegative correlation matrix $\mathbf{R}$ in order to compare the hierarchical clustering algorithms with the model presented in Chapter 2. If the non-negativity assumption does not hold, Eq. (3.1) can be used to transform similarities into dissimilarities by taking the absolute value or the square of the correlation coefficients (Revelle, 1979; Soffritti, 1999; Liu et al., 2012). However, the latter case is out of the scope of this chapter.

[2]We use the term *loss* to refer to it.

3.3 The Ultrametric Correlation Model: a brief review

In this section we briefly recall the main features of the model presented in Chapter 2.

Considering a *nonnegative* correlation matrix $\mathbf{R}$ of order p, the Ultrametric Correlation Model is defined by the following equation

$$\mathbf{R} = \mathbf{R}_{\mathrm{u}} + \mathbf{E}, \qquad \text{(as Eq. 2.1)}$$

where $\mathbf{R}_{\mathrm{u}}$ is the $(p \times p)$ matrix representing the hierarchical structure of latent concepts and $\mathbf{E}$ is the $(p \times p)$ random error matrix, i.e., the residual matrix. The nonnegativity assumption on $\mathbf{R}$, and consequently $\mathbf{R}_{\mathrm{u}}$, lets avoid a compensatory effect into the hierarchy. Therefore, this assumption allows to compare the hierarchical clustering methods recalled in Section 3.2 with UCM, since both distances and correlations turn out to be nonnegative. Moreover, they belong to the interval $[0, 1]$ according to Eq. (3.1).

$\mathbf{R}_{\mathrm{u}}$ is an ultrametric correlation matrix, where the ultrametric property formalizes the mathematical counterpart of the latent concept hierarchy. It is formally specified as follows

$$\mathbf{R}_{\mathrm{u}} = \mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q)\mathbf{V}' + \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}') + \mathbf{I}_p, \qquad \text{(as Eq. 2.2)}$$

subject to constraints

$$\mathbf{V} = [v_{jq} \in \{0, 1\} : j = 1, \ldots, p, q = 1, \ldots, Q]; \qquad \text{(as constr. 2.3)}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1, \ldots, p; \qquad \text{(as constr. 2.4)}$$

$$\mathbf{R}_{\mathrm{B}} \text{ is an ultrametric correlation matrix (Definition 2.2);} \qquad \text{(as constr. 2.5)}$$

$$\min\{{}_W r_{qq} : q = 1, \ldots, Q\} \geq \max\{{}_B r_{qh} : q, h = 1, \ldots, Q, h \neq q\}, \qquad \text{(as constr. 2.6)}$$

where $\mathbf{V}$ is the $(p \times Q)$ membership matrix that defines a partition of the variable space, i.e., it identifies Q non-overlapping groups of variables $(C_1, \ldots, C_Q)$; $\mathbf{R}_{\mathrm{B}}$ is the $(Q \times Q)$ between-concept correlation matrix, whose off-diagonal elements ${}_B r_{qh}$ $(q, h = 1, \ldots, Q, h \neq q)$ denote the correlation between two latent concepts, each one associated with a variable group (C_q and C_h), and $\mathbf{R}_{\mathrm{W}}$ is the $(Q \times Q)$ diagonal within-concept consistency matrix, whose diagonal elements ${}_W r_{qq}$ $(q = 1, \ldots, Q)$ represent the consistency within each group of variables. The two latter matrices, $\mathbf{R}_{\mathrm{B}}$ and $\mathbf{R}_{\mathrm{W}}$, embody two different features related to the variable groups: the correlation between concepts and the internal consistency of a concept, respectively (Cavicchia, Vichi, & Zaccaria, 2019).

UCM is estimated in a least-squares framework, minimizing the squared norm of the difference between the observed correlation matrix $\mathbf{R}$ and the reconstructed ultrametric correlation matrix $\mathbf{R}_{\mathrm{u}}$ (see Section 2.4).

In the next section, a comparison between the models described herein is carried out. This stresses the strong potential of UCM in investigating the hierarchical relationships between latent concepts, whenever they exist and even if not known a priori, with respect to the traditional agglomerative clustering algorithms.

3.4 A comparison between the Ultrametric Correlation Model and the agglomerative clustering algorithms

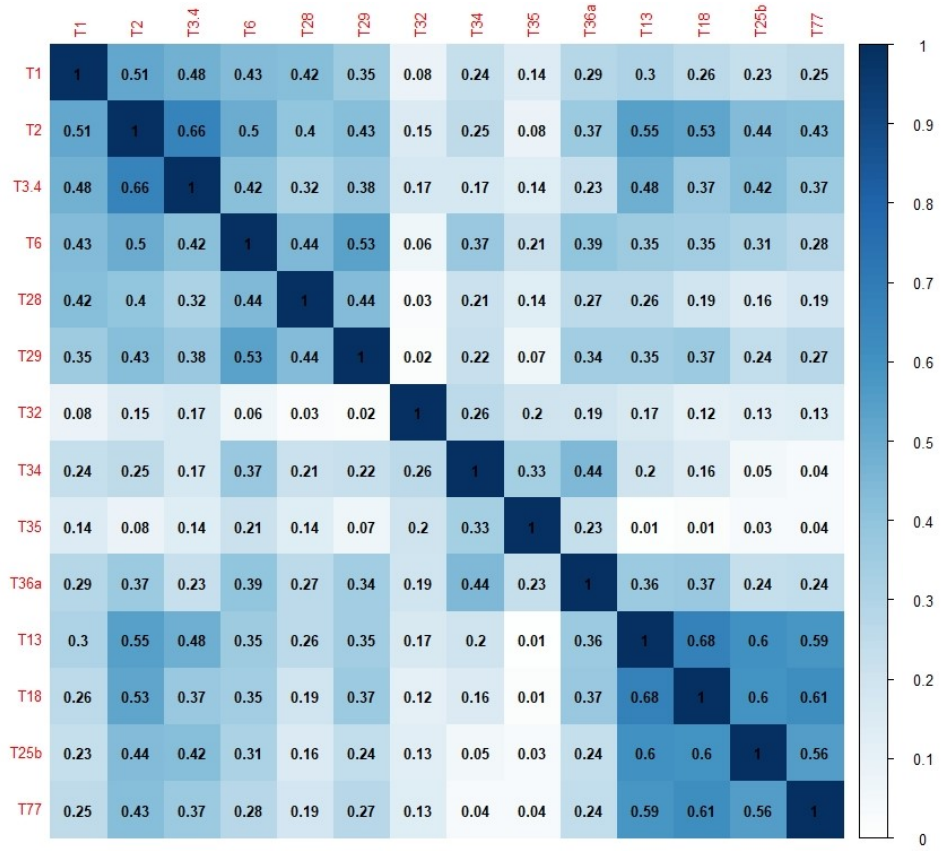


Figure 3.2. Heatmap of the Holzinger (14×14) correlation matrix of ability tests.

Table 3.1. Holzinger data set: variables and latent dimensions (ability) description.

C_q^{th}	Latent concept (ability)	Variables
C_1^{th}	Spatial Tests	T1, T2, T3.4
C_2^{th}	Mental Speed Tests	T6, T28, T29
C_3^{th}	Motor Speed Tests	T32, T34, T35, T36a
C_4^{th}	Verbal Tests	T13, T18, T25b, T77

The Holzinger data set[3] (Holzinger & Swineford, 1937) is a benchmark example very useful to inspect the hierarchical factorial structure of a multidimensional phenomenon. In this case, the latter is represented by the general ability of an individual that is composed of different latent dimensions (concepts). The data set is

[3]Available on `psych` package in R.

defined as a (14×14) correlation matrix ($p = 14$), with $Q = 4$ latent concepts (*Spatial*, *Mental Speed*, *Motor Speed* and *Verbal*) corresponding to the abilities tested for 355 individuals and described in Table 3.1. The optimal number of latent concepts may be assessed by means of the classical criteria to choose the number of factors or principal components (e.g., the Kaiser's method, Kaiser, 1960).

We aim at defining reliable concepts and, additionally, identifying the hierarchical structure over them. To achieve this goal we assess the potential of the model described in Section 3.3 with respect to the traditional hierarchical clustering methods cited in Section 3.2, when there exists a particular hierarchical latent structure underlying the data.

Firstly, we implemented UCM on the Holzinger correlation matrix to obtain the parsimonious bottom-up structure of the four latent concepts. It is worth noticing that in order to apply UCM the non-negativity condition on the correlation matrix must hold. The original one has three negative correlation coefficients very close to zero which turn out to be statistically nonsignificant (Holzinger & Swineford, 1937), as well as the other values in the Holzinger correlation matrix whose magnitude is lower than 0.1; we can thus take the absolute value of these negative terms. The resulting nonnegative correlation matrix shown in Figure 3.2 points out the existence of the four theoretical groups - corresponding to the latent concepts of the *Spatial*, *Mental Speed*, *Motor Speed* and *Verbal* abilities: three out of four are internally highly correlated, whereas the variables representing the *Motor Speed* ability have lower correlations within the group. As a result, this weak relationship between the objects in C_3^{th} could entail their misclassification, as we will see thereafter. Moreover, the variables belonging to the first two groups, which correspond to the *Spatial* and *Mental Speed* abilities, are highly correlated between them.

Unlike the traditional hierarchical clustering algorithms that produce complete dendrograms, i.e., they define a complete hierarchy over p variables, UCM starts from the classification of variables in $Q < p$ groups before searching for their optimal bottom-up aggregations. As shown in Table 3.2, UCM groups together the *Spatial* and *Mental Speed* abilities into C_1^{UCM}, and let the variable T36a define a singleton, i.e., C_2^{UCM}, rather than be merged with the variables belonging to C_4^{UCM}. C_3^{UCM} is instead well defined and associated with the *Verbal* ability as in C_4^{th}.

The UCM estimates of the between-concept correlation matrix and the within-concept consistency matrix are the following

$$\mathbf{R}_{\text{B}} = \begin{bmatrix} 1.000 & 0.309 & 0.332 & 0.143 \\ 0.309 & 1.000 & 0.309 & 0.143 \\ 0.332 & 0.309 & 1.000 & 0.143 \\ 0.143 & 0.143 & 0.143 & 1.000 \end{bmatrix} \quad \mathbf{R}_{\text{W}} = \begin{bmatrix} 0.447 & 0 & 0 & 0 \\ 0 & 1.000 & 0 & 0 \\ 0 & 0 & 0.606 & 0 \\ 0 & 0 & 0 & 0.332 \end{bmatrix}.$$

It has to be highlighted that the UCM algorithm involves an UPGMA step, adapted for correlations, in order to obtain the ultrametric correlation matrix $\mathbf{R}_{\text{B}}$ (see Section 2.4). The results of the application of UCM on the Holzinger correlation matrix are shown in Figure 3.3. The dendrogram is obtained by applying Eq. (3.1) to the ultrametric correlation matrix estimated by UCM. Since the observed correlation coefficients are nonnegative by hypothesis, the distances belong to the interval $[0, 1]$ and vice versa. Looking at Figure 3.3, the hierarchy over the four variable groups and the corresponding latent concepts is built by merging the *Spatial* and *Mental*

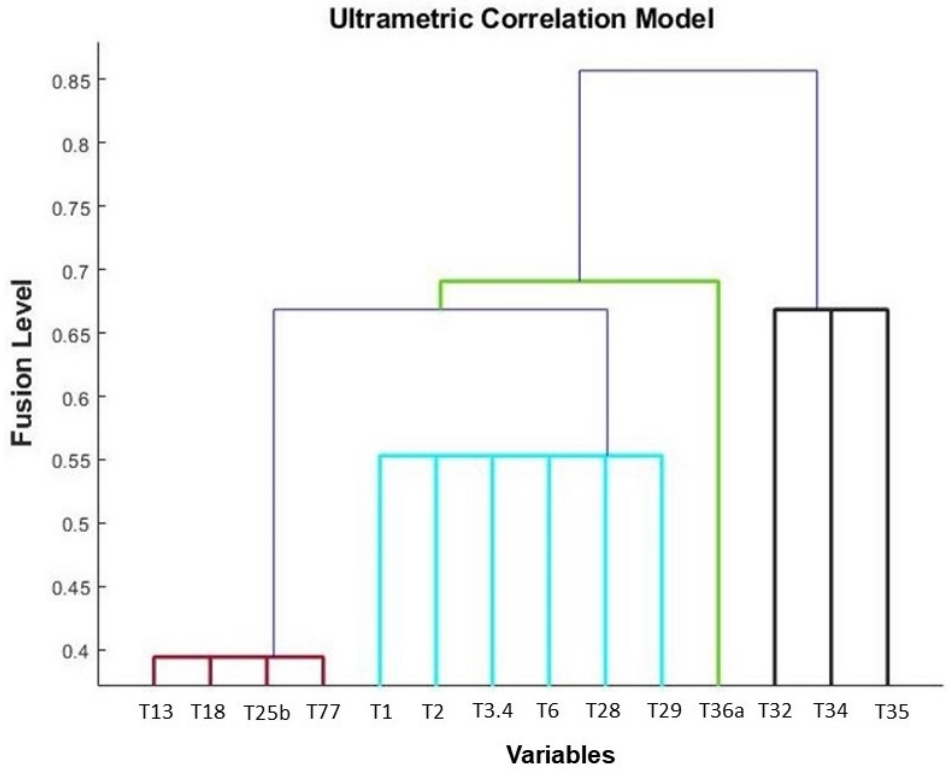


Figure 3.3. Dendrogram of the ultrametric distance matrix obtained by computing the Ultrametric Correlation Model on the Holzinger (14×14) correlation matrix of ability tests and applying Eq. (3.1) on the result.

Speed abilities with the *Verbal* ability first, that are all related to the brain; then, this broad group with the variable T36a. It has to be noticed that this variable has a lower correlation (on average) with those in C_3^{th} than with those in the other groups (Figure 3.2). The last aggregation lumps together the broad group composed of $\{C_1^{\text{UCM}}, C_2^{\text{UCM}}, C_3^{\text{UCM}}\}$ and C_4^{UCM}, whose corresponding latent concept is related to the movement ability. It is worthy of remark that the Holzinger data set has been analyzed by many authors; in particular, Loehlin and Beaujean (2017, pp. 235-239) conducted a higher-order exploratory analysis (Schmid & Leiman, 1957) on the Holzinger correlation matrix by pinpointing a strong correlation between the *Spatial* and *Verbal* abilities, and a low correlation between the *Motor Speed* and the other abilities. The latter bears UCM out since C_4^{UCM} is lumped together with the other groups in the last aggregation, as an additional ability. Therefore, the examination of the Cronbach's α (Cronbach, 1951) was carried out, revealing the existence of a general latent concept associated with all variables.

In order to make a comparison between UCM and the hierarchical classification methods, we applied the Single, Complete, Average Linkage and Ward's Method to the distance matrix obtained by transforming the Holzinger correlation matrix by means of Eq. (3.1), and complying the procedure described in Section 3.2. The results are shown in Figure 3.4, where the groups corresponding to the 4th level of the hierarchy ($Q = 4$) are colored. It is worthy of remark that in order to identify the

Table 3.2. Variable groups of UCM with $Q = 4$.

C_q^{UCM}	Variables
C_1^{UCM}	T1, T2, T3.4, T6, T28, T29
C_2^{UCM}	T36a
C_3^{UCM}	T13, T18, T25b, T77
C_4^{UCM}	T32, T34, T35

Table 3.3. Variable clusters at the 4th level ($Q = 4$) of the clustering methods hierarchy.

C_q^{m}	Single Link and Average Link	Complete Link and Ward's Method
C_1^{m}	T1, T2, T3.4, T6, T28, T29, T13, T18, T25b, T77	T1, T2, T3.4, T6, T28, T29
C_2^{m}	T34, T36a	T32
C_3^{m}	T32	T13, T18, T25b, T77
C_4^{m}	T35	T34, T35, T36a

aforementioned level of the hierarchy, and the corresponding partition of variables, a complete dendrogram must be computed. Indeed, the hierarchical clustering algorithms taken into account herein do not allow to choose the optimal number of clusters a priori, as UCM does. The four groups identified at the 4th top-down level of the hierarchy by each hierarchical clustering method are illustrated in Table 3.3. The different composition of these groups with respect to the theoretical and the UCM ones stands out. On one hand, the variable T36a is merged with at least one of the other variables of C_3^{th}, conversely to UCM. On the other hand, it is evident that the partitions in 4 groups obtained by the Complete Linkage and the Ward's Method are more similar to that of UCM than those of the Single and Average Linkage. For all methods, the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) is computed at the Qth level of the hierarchy in order to compare the theoretical variable partition in 4 groups with the estimated ones. Looking at Table 3.4, it can be noticed that the Complete Linkage, Ward's Method and UCM have the same ARI, even if the singleton of UCM is composed of the variable T36a instead of T32.

To have a deeper comparison between the hierarchical clustering methods and UCM in terms of the hierarchical relationship detection, we computed their loss as the least-squares difference between the observed correlation matrix and the estimated one over the total correlation of the data. The results are shown in Table 3.4. The loss of UCM turns out to be lower than that of the other methods by revealing that UCM is better able to reconstruct the observed data matrix than the competing methods. Indeed, even if the ARI of UCM is equal to that of the Complete Linkage and Ward's Method, the UCM hierarchy over the 4 variable groups better reconstructs the hierarchical relationships among the fourteen variables. Only the Average Linkage has a similar loss to UCM, whereas the Single, Complete Linkage and Ward's method have a three times higher loss. Therefore, the ARI of the Average Linkage is extremely lower than that of UCM. Thus, we can state that UCM is able to balance a good performance on the variable partition recovery in Q groups and

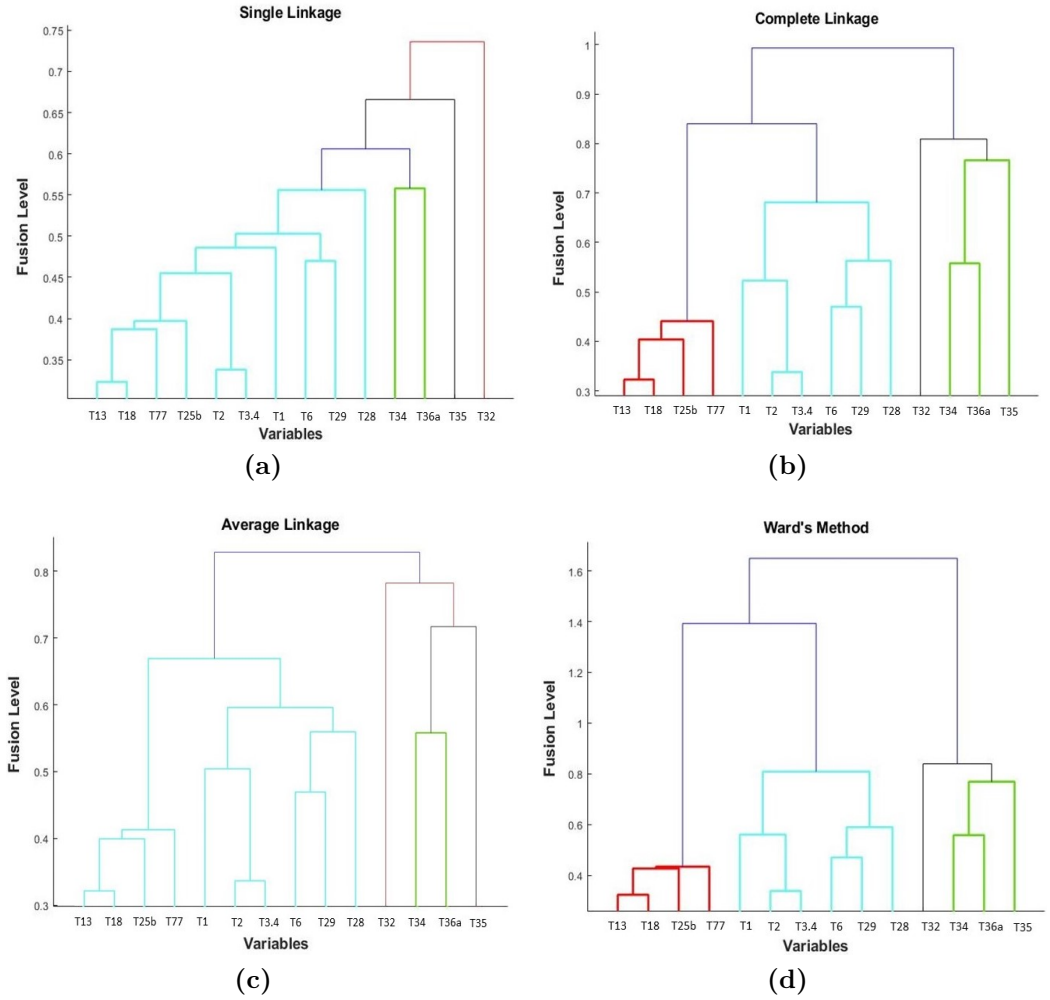


Figure 3.4. Dedrogram of (3.4a) Single Linkage, (3.4b) Complete Linkage, (3.4c) Average Linkage, (3.4d) Ward's Method on the distance matrix obtained by transforming the Holzinger (14×14) correlation matrix of ability tests according to Eq. (3.1).

on detecting the hierarchical relationships among them.

The results shown in this section illustrate the difference between the procedure based on the traditional hierarchical classification methods (see Section 3.2) and the Ultrametric Correlaton Model described in Section 3.3. Indeed, starting from p observed variables and building a complete hierarchy over them turns out to be not sufficient to properly identify a hierarchy of latent concepts, when there exists a hierarchical latent structure in the data. Moreover, if the number of the original variables is too large a dimensionality reduction of the problem, means a parsimonious representation, is needed. UCM provides both a parsimonious representation of the relationships among variables and a model-based approach to build a hierarchy starting from Q variable groups, each one associated with a latent concept. It is worth highlighting once again that the hierarchical clustering methods are not able to repair the errors done in the initial levels of the complete hierarchy,

Table 3.4. ARI between the theoretical membership matrix defined in Holzinger and Swineford (1937) and the membership matrices obtained by UCM and the traditional hierarchical clustering methods at level $Q = 4$ and their loss.

Method/Model	ARI	Loss
UCM	0.6308	0.0423
Single Link	0.1703	0.1440
Complete Link	0.6308	0.1536
Average Link	0.1703	0.0449
Ward's Method	0.6308	0.1354

whenever they occur. Conversely, thanks to its features, UCM does not suffer from the errors underneath the Qth level of the hierarchy, because Q variable groups are directly pinpointed in a dimensionality reduction approach, without going through binary aggregations of variables from p up to Q. Moreover, since UCM works on a correlation matrix, the latent concepts of a phenomenon might also be quantified.

3.5 Conclusions

In this chapter a comparison between a procedure based on the well-known hierarchical clustering methods applied on variables and the novelty model proposed in Chapter 2 is provided. The latter allows to pinpoint a parsimonious representation of multidimensional phenomena through the partition of the observed variables into a reduced number of groups, each one associated with a latent concept, and to study the relationships among them. The difference between the procedure based on the traditional clustering algorithms and the Ultrametric Correlation Model is appreciated thanks to their application to a benchmark data set with a hierarchical "factorial" structure.

UCM entails a dimensionality reduction of the problem under study, starting from a parsimonious representation of the variables into groups, and the construction of a hierarchy of latent concepts. The number of groups is chosen a priori by means of the traditional criteria for selecting the optimal number of factors/components instead of cutting the n-tree, as usually done for the hierarchical clustering methods. Differently from the latter algorithms, UCM does not suffer from the errors that could turn up in the lower levels of the dendrogram, since it starts from a partition of variables into groups.

Chapter 4

Gaussian Mixture Model with an extended ultrametric covariance structure

4.1 Introduction

Finite mixture models are one of the most widespread methodologies to model the density of a heterogeneous population. They assume that the observed data are collected from a population composed of a finite set of G homogeneous subpopulations with a given distribution (Titterington, Smith, & Makov, 1985; McLachlan & Basford, 1988; McLachlan & Peel, 2000a). When each distribution has a multivariate Gaussian form, the model is called Gaussian Mixture Model (GMM). In general, the GMM density assumes the form

$$f(\mathbf{x}|\boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \, \phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{4.1}$$

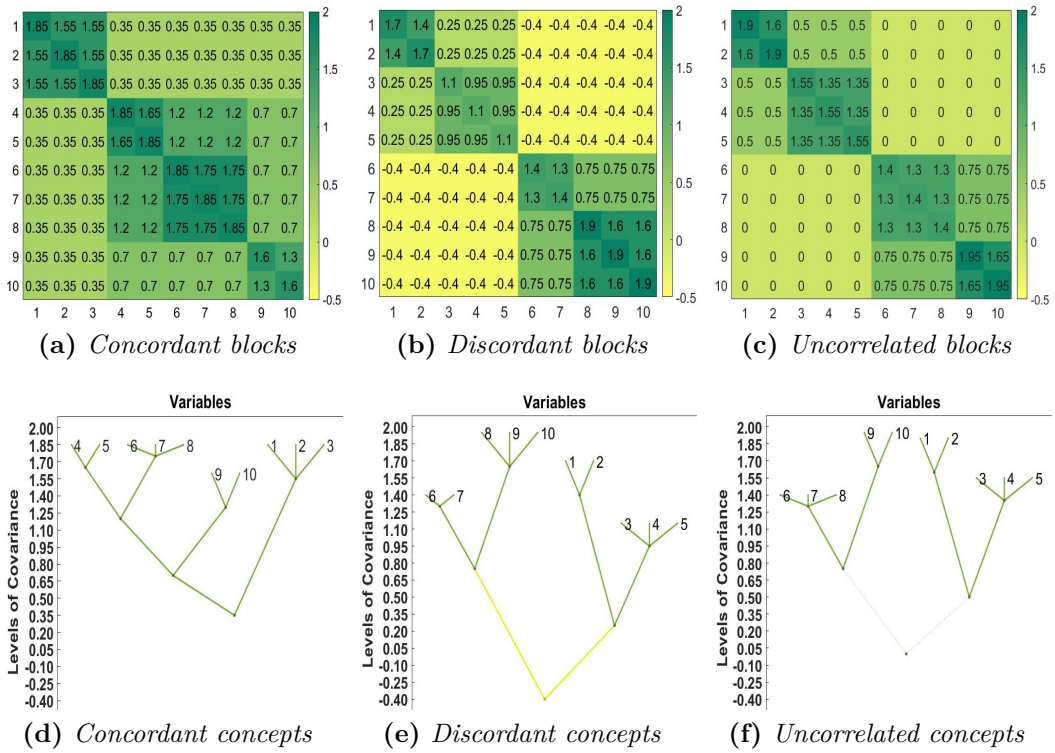
where each component of the mixture has the density of a multivariate Gaussian, denoted by $\phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, with a p-dimensional mean vector $\boldsymbol{\mu}_g$ and a covariance matrix $\boldsymbol{\Sigma}_g$. The quantities $\pi_1, \ldots, \pi_G$ are the mixing proportions (prior probabilities) such that $\pi_g \geq 0$ and $\sum_{g=1}^{G} \pi_g = 1$, and $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}$ is the overall parameter vector. Relevant specialized literature and complete reviews of the GMM are found in McLachlan and Basford (1988), McLachlan and Peel (2000a), Fraley and Raftery (2002), and Bouveyron et al. (2019).

Finite mixture models are used for model-based clustering (McNicholas, 2016; Bouveyron et al., 2019), as well as discriminant analysis and multivariate density estimation. Considering the GMM, in the model-based approach to clustering each component of the mixture is associated with an ellipsoidal cluster, centered at the mean vector $\boldsymbol{\mu}_g$ and with volume, shape and orientation derived by the covariance matrix $\boldsymbol{\Sigma}_g$.

Although GMMs represent a conceptually and mathematically elegant class of models, they suffer from the *curse of dimensionality* (Bellman, 1957) due to the fact that their application for clustering with high-dimensional data is often

computationally demanding; indeed, this requires a large amount of data for the estimation of the large number of parameters, that is $G-1+Gp+Gp(p+1)/2$, (i.e., $G-1$ for mixing proportions; Gp for mean vectors; $Gp(p+1)/2$ for covariance matrices). Since most parameters are produced by the covariance matrices, parsimonious parameterizations of the latter were proposed in literature. One of the most used is the eigen-decomposition (Banfield & Raftery, 1993) of the form $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, where λ_g is a scalar determining the cluster volume, $\mathbf{A}_g$ is a diagonal matrix controlling the cluster shape, and $\mathbf{D}_g$ is an orthogonal matrix which specifies the cluster orientation. The eigen-decomposition allows defining different parsimonious GMMs, called Gaussian Parsimonious Clustering Models (GPCMs), by imposing specific geometric features to the cluster covariance structure and/or by constraining the covariance components to be equal or unequal across clusters (Celeux & Govaert, 1995; Fraley & Raftery, 1998, 2002). The fourteen different models based on the eigen-decomposition are implemented into the R packages `mixture` (Langrognet et al., 2020), `mclust` (Fraley & Raftery, 1999; Scrucca et al., 2016), `Rmixmod` (Biernacki et al., 2006). McNicholas and Murphy (2008) proposed a class of eight Parsimonious GMMs (PGMMs), then increased to twelve (Expanded PGMMs, EPGMMs, McNicholas & Murphy, 2010), based on Factor Analysis (FA, Spearman, 1904; Anderson & Rubin, 1956; Horst, 1965) by extending both the mixtures of factor analyzers (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000b; McLachlan, Peel, & Bean, 2003) and the mixtures of probabilistic principal component analyzers (Tipping & Bishop, 1999b, 1999a). The mixture of factor analyzers model assumes a cluster covariance structure of the form $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$, where $\boldsymbol{\Lambda}_g$ is the $(p \times Q)$ factor loading matrix and $\boldsymbol{\Psi}_g$ is the diagonal covariance matrix of the error of order p. Also in this case, the twelve models are obtained by considering equal or unequal covariance components across clusters. EPGMMs are implemented into the R package `pgmm` (McNicholas et al., 2019). In order to further reduce the number of parameters of the cluster covariance matrices, Baek, McLachlan, and Flack (2010) provided an extension of the mixture of factor analyzers with common component-factor loadings. The latter results effective when the number of dimensions p is large relative to the sample size n and/or the number of clusters G is not small. In the high-dimensional context, Bouveyron, Girard, and Schmid (2007) proposed the High-Dimensional Data Clustering (HDDC) model which is a GMM based on the eigen-decomposition with a reduced number of different eigenvalues for each cluster covariance matrix. HDDC therefore parameterizes the diagonal matrix of the eigenvalues as $\mathbf{A}_g = \text{diag}([a_{g1}, \ldots, a_{gd_g}, b_g, \ldots, b_g]')$, where $\text{diag}(\mathbf{a})$ is a diagonal matrix with diagonal entries equal to the vector $\mathbf{a}$, $d_g \in \{1, \ldots, p-1\}$ is the intrinsic dimension of each mixture component, $a_{gj}, j = 1, \ldots, d_g$, are the first d_g greatest eingenvalues of $\boldsymbol{\Sigma}_g$ modeling the variance in the cluster-specific subspace, and b_g represents the variance of the noise. As well as GPCMs and EPGMMs, HDDC enables to define a family of parsimonious models by fixing some parameters to be common between and/or within clusters. A subset of the twenty-eight models resulting from the HDDC parameterization is implemented in the R package `HDClassif` (Bergé, Bouveyron, & Girard, 2012, 2019). For a complete review of the existing methodologies for model-based clustering in high-dimensional spaces see Bouveyron and Brunet-Saumard (2014) and Fop and Murphy (2018).

In this chapter, we introduce a new GMM with a parameterization of the



(a) *Concordant blocks* **(b)** *Discordant blocks* **(c)** *Uncorrelated blocks*

(d) *Concordant concepts* **(e)** *Discordant concepts* **(f)** *Uncorrelated concepts*

Figure 4.1. Examples of the extended ultrametric covariance matrices (4.1a)-(4.1c) and the corresponding path diagrams (4.1d)-(4.1f) representing different hierarchical relationships among nested concepts.

covariance matrix by assuming an *extended ultrametric covariance matrix* for each cluster. The latter extends the definition of an ultrametric matrix provided in Chapter 2 (Definition 2.2) to a generic covariance matrix, i.e., by relaxing the non-negativity constraint. The extended ultrametric covariance matrix defined in this chapter is thus modeled with a similar structure to UCM (Chapter 2) - but for covariances - which is one-to-one associated in turn with a tree describing a hierarchy of relations among groups of variables. The hierarchy defines frequently used hierarchical relations: (i) a unique, consistent and reliable general latent concept identified by even more reliable, nested and specific concepts. In this case, the extended ultrametric covariance matrix has all nonnegative values and it is formed by nested blocks of more positive covariance sub-matrices (Fig. 4.1a). Thus, there is a relevant internal consistency (concordance, agreement) among all observed variables that can allow one to identify the general latent concept at the root of the tree (Fig. 4.1d). From this matrix a composite indicator corresponding to the general concept can be estimated (OECD, 2008); (ii) an inconsistent and unreliable general latent concept, formed by two or more discordant ones which are characterized by specific and internally consistent concepts. Thus, there exists a general concept, but it is formed by discordant specific ones. The extended ultrametric covariance matrix has positive values that define the specific initial concepts in the hierarchy, with nested blocks of more positive covariance sub-matrices. However, between the last two (or more) blocks of nested sub-matrices the covariances are on average

negative (Fig. 4.1b). For several levels, the hierarchy finds at least two (or more) internally consistent sub-hierarchies, which identify two (or more) discordant (on average) groups of variables (Fig. 4.1e); (iii) no general concept, since the last two (or more), which should form the general one, are substantially uncorrelated. Moreover, it should be noted that, in this case, the hierarchy does not form a unique tree (Fig. 4.1f). The extended ultrametric covariance matrix has positive values, with nested blocks of sub-matrices, to define the specific concepts, but the covariance is (on average) null between the last two (or more) (Fig. 4.1c). This situation is frequently observed in higher-order factor models (G. H. Thompson, 1948; Cattell, 1978a; Gorsuch, 1983).

The extended ultrametric covariance matrix, which has this important flexibility in modeling hierarchical relationships among variables, is finally implemented into a GMM. On one hand, our approach allows modeling multidimensional phenomena which present a nested hierarchical structure on variables by considering a limited number of parameters, and, on the other hand, it allows defining a new parsimonious GMM. The parsimony of the ultrametric structure motivates its use to model complex multidimensional phenomena.

The chapter is organized as follows. In Section 4.2, the notation used throughout the chapter and a background about ultrametric matrices are given to allow the reader to follow the specification of the model herein. Section 4.3 introduces the extended ultrametric covariance structure with its features. The Gaussian Mixture Model with an Extended Ultrametric Covariance Structure is provided in Section 4.4, along with computational aspects and the model selection criterion. Section 4.5 shows the performance of the proposal on synthetic data and Section 4.6 on real data. A final discussion completes the chapter in Section 4.7.

4.2 Notation and theoretical background

For the convenience of the reader, the notation used in this chapter is listed here.

n, p, G, Q	Number of observations, variables, clusters, groups of variables, respectively.
$\mathbf{\Sigma} = [\sigma_{jl}]$	Covariance matrix of order p.
$\mathbf{V} = [v_{jq}]$	$(p \times Q)$ membership matrix, where $v_{jq} = 1$ if the jth variable belongs to the qth group; $v_{jq} = 0$ otherwise. It is binary and row-stochastic, i.e., with one non-zero element per row, identifying a partition of variables in Q groups.
$\mathbf{\Sigma}_{\mathrm{V}} = [{}_{V}\sigma_{qq}]$	Diagonal matrix of order Q with diagonal entries representing variances of the groups of variables.
$\mathbf{\Sigma}_{\mathrm{W}} = [{}_{W}\sigma_{qq}]$	Diagonal matrix of order Q with diagonal entries representing covariances within groups of variables.
$\mathbf{\Sigma}_{\mathrm{B}} = [{}_{B}\sigma_{qh}]$	Matrix of order Q with off-diagonal entries representing covariances between groups of variables, and diagonal ones equal to zero.

In Chapter 2, we introduce the Ultrametric Correlation Model (UCM, Cavicchia, Vichi, & Zaccaria, 2020b) to reconstruct a nonnegative correlation matrix by using the definition of an ultrametric matrix (Dellacherie, Martínez, & San Martín, 2014, pp. 58-59). The latter differs from an ultrametric distance matrix, but preserves

the non-negativity property, i.e., the entries of the matrix are nonnegative (see Chapter 1, Section 1.2). The ultrametric correlation matrix $\mathbf{R}_{\text{u}} = [{}_{\text{u}}r_{jl}]$ is related to an ultrametric distance matrix $\mathbf{D}_{\text{u}} = [{}_{\text{u}}d_{jl}]$ by the relationship ${}_{\text{u}}d_{jl} = 1 - {}_{\text{u}}r_{jl}$, $j, l = 1, \ldots, p$.

In this chapter, we extend the definition of a nonnegative ultrametric correlation matrix to a generic covariance matrix by relaxing the non-negativity constraint. Let us recall that a covariance matrix $\mathbf{\Sigma}$ of order p, with elements $\sigma_{jl} \in \mathbb{R}$, $j, l = 1, \ldots, p$, satisfies the following properties:

(i) *symmetry*: $\sigma_{jl} = \sigma_{lj}$ for $j, l = 1, \ldots, p$;

(ii) *non-negativity of the diagonal*: $\sigma_{jj} \geq 0$ for all $j = 1, \ldots, p$;

(iii) *positive semi-definiteness*: $\mathbf{x}'\mathbf{\Sigma}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^p$.

The ultrametric definition of a nonnegative matrix (Dellacherie, Martínez, & San Martín, 2014, pp. 58-59) requires to satisfy property (i) and the following two additional properties (see Definition 2.2), which can be extended to a matrix $\mathbf{\Sigma}$ with real values:

(iv) *ultrametric inequality*: $\sigma_{jl} \geq \min\{\sigma_{jh}, \sigma_{lh}\}$, for $j, l, h = 1, \ldots, p$;

(v) *column pointwise diagonal dominance*: $\sigma_{jj} \geq \max\{|\sigma_{lj}|, l = 1, \ldots, p\}$ for $j = 1, \ldots, p$.

Condition (iv) can be equivalently rewritten as follows

(iv') for each triplet $j, l, h = 1, \ldots, p$, there exists a reordering $\{j, l, h\}$ of the elements s.t. $\sigma_{jl} \geq \sigma_{jh} = \sigma_{lh}$. It corresponds to state that for each triplet the smallest two elements are equal;

and, together with (i), it implies condition (v) when a matrix is nonnegative.

Definition 4.1 (Weak Extended Ultrametric Matrix)**.** A matrix $\mathbf{\Sigma}$ is a weak extended ultrametric matrix if all its elements $\sigma_{jl} \in \mathbb{R}$, for $j, l = 1, \ldots, p$, and conditions (i), (ii), (iv), (v) hold.

Remark 4.1. Condition (v) is sufficient for a nonnegative ultrametric matrix to be positive semi-definite (Dellacherie, Martínez, & San Martín, 2014, pp. 60-61). However, if a matrix meets conditions (i), (ii) and (iv), without the non-negativity constraint on the off-diagonal elements, condition (v) is not sufficient to guarantee its positive semi-definiteness.

According to Remark 4.1, under condition (v) $\mathbf{\Sigma}$ is a weak extended ultrametric matrix, but not a covariance matrix, since condition (iii) cannot be necessarily satisfied. A stronger condition is thus needed to guarantee $\mathbf{\Sigma}$ to be weak extended ultrametric and positive semi-definite. This is achieved with the following property:

(v') *diagonal dominance*: $\sigma_{jj} \geq \sum_{\substack{l=1 \\ l \neq j}}^{p} |\sigma_{jl}|$ for $j = 1, \ldots, p$.

In fact, condition (v'), together with conditions (i) and (ii), is sufficient for the positive semi-definiteness of a matrix as shown in Brouwer and Haemers (2012, pp. 30-31). Therefore, we can extend the definition of a nonnegative ultrametric correlation matrix to a covariance matrix with real values.

Definition 4.2 (Weak Extended Ultrametric Covariance Matrix)**.** A matrix $\mathbf{\Sigma}$ is said to be a weak extended ultrametric covariance matrix if all its elements $\sigma_{jl} \in \mathbb{R}$, for $j, l = 1, \ldots, p$, and conditions (i), (ii), (iv), (v') hold.

Remark 4.2. If properties (ii) and (v') are strict in Definition 4.2, then we say that $\mathbf{\Sigma}$ is a (*strict*) *extended ultrametric covariance matrix*. The latter is positive definite (Gerschgorin, 1931; Horn & Johnson, 2013, Corollary 7.2.3).

Remark 4.3. It is worth highlighting that the diagonal dominance and the column pointwise diagonal dominance are defined for a generic matrix $\mathbf{\Sigma}$ (i.e., not necessarily with nonnegative diagonal entries) as $|\sigma_{jj}| \geq \sum_{l=1, l\neq j}^{p} |\sigma_{jl}|$ and $|\sigma_{jj}| \geq \max\{|\sigma_{lj}|, l = 1, \ldots, p\}$, $j = 1, \ldots, p$, respectively. However, they can be respectively written as (v') and (v) under condition (ii).

Remark 4.4. If $\mathbf{\Sigma}$ is a weak extended ultrametric covariance matrix, it can be transformed into an ultrametric distance matrix $\mathbf{D}_{\mathrm{u}} = [{}_{\mathrm{u}}d_{jl}]$ by the relationship ${}_{\mathrm{u}}d_{jl} = \frac{\sqrt{\sigma_{jj}\sigma_{ll}} - \sigma_{jl}}{2\sqrt{\sigma_{jj}\sigma_{ll}}}$, for $j, l = 1, \ldots, p$, where σ_{jj}, σ_{ll} are the variances of the jth and lth variables, respectively, and σ_{jl} is their covariance.

In the next section, we formalize the extension of UCM to an extended ultrametric covariance matrix.

4.3 Extended Ultrametric Covariance Structure

We introduce a parameterization of an extended ultrametric covariance matrix, called Extended Ultrametric Covariance Structure (EUCovS), which is formally defined as follows

$$\mathbf{\Sigma}_{\mathrm{u}} = \mathbf{V}(\mathbf{\Sigma}_{\mathrm{W}} + \mathbf{\Sigma}_{\mathrm{B}})\mathbf{V}' - \mathrm{diag}(\mathbf{V}\mathbf{\Sigma}_{\mathrm{W}}\mathbf{V}') + \mathrm{diag}(\mathbf{V}\mathbf{\Sigma}_{\mathrm{V}}\mathbf{V}') \tag{4.2}$$

subject to constraints

$$\mathbf{V} = [v_{jq} \in \{0, 1\} : j = 1, \ldots, p, q = 1, \ldots, Q]; \tag{4.3}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1, \ldots, p; \tag{4.4}$$

$$\mathbf{\Sigma}_{\mathrm{B}} = \mathbf{\Sigma}'_{\mathrm{B}}, \mathrm{diag}(\mathbf{\Sigma}_{\mathrm{B}}) = \mathbf{0}, {}_{B}\sigma_{qh} \geq \min\{{}_{B}\sigma_{qs}, {}_{B}\sigma_{hs}\}\ q, h, s = 1, \ldots, Q,$$
$$s \neq h \neq q; \tag{4.5}$$

$$\min\{{}_{W}\sigma_{qq} : q = 1, \ldots, Q\} \geq \max\{{}_{B}\sigma_{qh} : q, h = 1, \ldots, Q, h \neq q\}; \tag{4.6}$$

$${}_{V}\sigma_{qq} > |{}_{W}\sigma_{qq}|\Big(\sum_{l=1}^{p} v_{lq} - 1\Big) + \sum_{\substack{h=1 \\ h\neq q}}^{Q} |{}_{B}\sigma_{qh}| \sum_{l=1}^{p} v_{lh} \quad q = 1, \ldots, Q, \tag{4.7}$$

where $\mathrm{diag}(\mathbf{A})$ is a diagonal matrix with diagonal entries equal to the diagonal of the matrix $\mathbf{A}$, $\mathbf{1}_p$ and $\mathbf{1}_Q$ are the unitary vectors of order p and Q respectively.

$\boldsymbol{\Sigma}_{\mathrm{u}}$ is an extended ultrametric covariance matrix. In fact, it is symmetric since (4.5) holds; it has positive entries on the main diagonal (4.7); it satisfies the ultrametric inequalities due to (4.5), (4.6) and (4.7); and it is strictly diagonally dominant given (4.7). Therefore, $\boldsymbol{\Sigma}_{\mathrm{u}}$ represents a covariance matrix where the variances on the diagonal are expressed by the diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{V}}$, while the covariances (i.e., the off-diagonal elements) are expressed by the diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{W}}$ and the off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}}$. The strict inequality in (4.7) guarantees to obtain $\boldsymbol{\Sigma}_{\mathrm{u}}$ as an extended ultrametric covariance matrix; however, if the equality is included, $\boldsymbol{\Sigma}_{\mathrm{u}}$ results in a weak extended ultrametric covariance matrix. We define $\boldsymbol{\Sigma}_{\mathrm{u}}$ as an extended ultrametric covariance matrix in order to allow the use of the EUCovS in the GMM, as we will show in the following section.

Remark 4.5. The strictly diagonal dominance in (4.7) is a strong condition which may lead to an overestimation of the parameter $\boldsymbol{\Sigma}_{\mathrm{V}}$.

A solution to the overestimation problem is given by replacing (4.7) with the following constraints

$$ {}_{V}\sigma_{qq} \geq \max\{|{}_{W}\sigma_{qq}|, |{}_{B}\sigma_{qh}|, h = 1, \ldots, Q, h \neq q\} \quad q = 1, \ldots, Q, \tag{4.8} $$

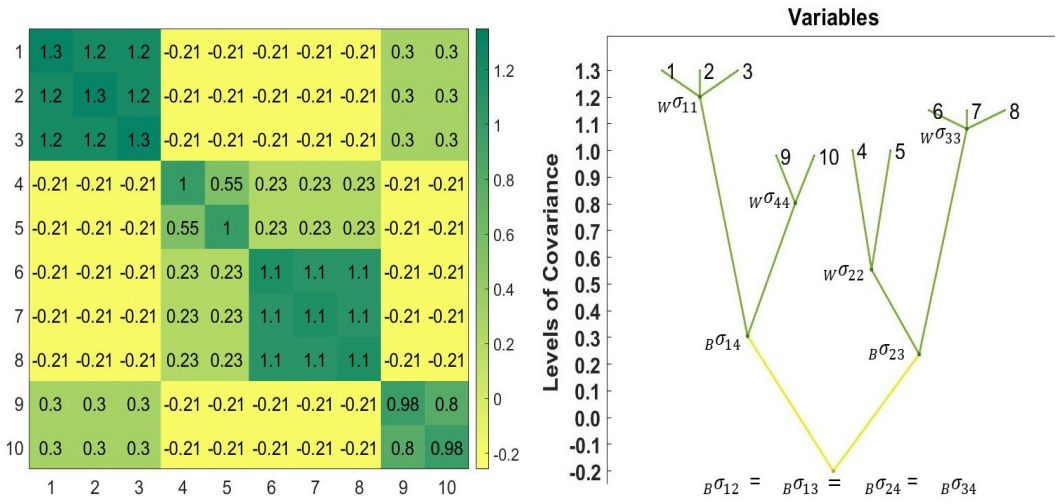
$$ \boldsymbol{\Sigma}_{\mathrm{u}} = \boldsymbol{\Sigma}_{\mathrm{u}} + a\mathbf{I}_p, \text{ with } a > 0, \text{ and such that } \boldsymbol{\Sigma}_{\mathrm{u}} \text{ is positive definite,} \tag{4.9} $$

where $\mathbf{I}_p$ is the identity matrix of order p. The factor a is the absolute value of the smallest eigenvalue of $\boldsymbol{\Sigma}_{\mathrm{u}}$ (Cailliez, 1983) plus an arbitrary small positive constant (e.g., in our algorithm it is equal to 0.1^6). Under constraints (4.3)-(4.6), (4.8) and (4.9), $\boldsymbol{\Sigma}_{\mathrm{u}}$ is still an extended ultrametric covariance matrix. It is worth noticing that (4.9) guarantees the positive definiteness of $\boldsymbol{\Sigma}_{\mathrm{u}}$ by changing as few elements as possible, i.e., only the elements of $\boldsymbol{\Sigma}_{\mathrm{V}}$ and not those of $\boldsymbol{\Sigma}_{\mathrm{W}}$ and/or $\boldsymbol{\Sigma}_{\mathrm{B}}$.

One of the main properties of the EUCovS in (4.2) is to be parsimonious in terms of the number of parameters involved. Indeed, the number of different diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{V}}$ and $\boldsymbol{\Sigma}_{\mathrm{W}}$ varies between 1 and Q, $Q \in \{1, \ldots, p\}$; whereas, the number of different off-diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{B}}$ ranges between 0 (all the variables are in the same group) and $Q - 1$, $Q \in \{1, \ldots, p\}$. Thus, $\boldsymbol{\Sigma}_{\mathrm{u}}$ can have as few as 2 and as many as $3Q - 1$ different elements.

Remark 4.6. Given the EUCovS in (4.2), then $3Q - 1$ is an upper bound for the number of different elements of $\boldsymbol{\Sigma}_{\mathrm{u}}$. Indeed, if $Q \geq p/2$ and p is an even number or $Q \geq \lfloor p/2 \rfloor + 1$ and p is an odd number, then the maximum number of different elements of $\boldsymbol{\Sigma}_{\mathrm{u}}$ might be lower than $3Q - 1$ since some groups of variables can be singletons, and thus the corresponding diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{V}}$ can be equal to the corresponding diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{W}}$.

Corollary 4.1. $\boldsymbol{\Sigma}_{\mathrm{u}}$ is one-to-one associated with a hierarchy of Q variable groups - each one representing a specific concept (dimension) of a multidimensional phenomenon characterized by at most $3Q - 1$ hierarchical levels. Values ${}_{V}\sigma_{qq}$, $q = 1, \ldots, Q$, define the initial level of the hierarchy for each group, ${}_{W}\sigma_{qq}$, $q = 1, \ldots, Q$, are associated with the first aggregation levels of the hierarchy and represent the covariance within the first Q groups. While, values ${}_{B}\sigma_{qh}$, $q, h = 1, \ldots, Q, h \neq Q$, identify the remaining $Q - 1$ levels and represent the covariance between groups of variables. The hierarchy therefore depicts the relationships within and between groups of variables, from the most concordant to the most discordant.



(a) *EUCovS*

(b) *Hierarchy over Q variable groups: path diagram representation*

Figure 4.2. Relationship between EUCovS and the corresponding hierarchy of variable groups.

Corollary 4.1 is based upon Lemma 1 in Cavicchia, Vichi, and Zaccaria (2020b) for a $(2Q-1)$-ultrametric correlation matrix and it can be demonstrated in the same way. This means that each $(3Q-1)$-EUCovS $\boldsymbol{\Sigma}_{\mathrm{u}}$ defines a hierarchy of variables, because each pair of variables might belong either to the same variable group (i.e., latent concept) if their covariance is ${}_W\sigma_{qq}$, or to distinct groups if their covariance is ${}_B\sigma_{qh}$. Furthermore, it is worth underlying that the variable groups are disjoint (as defined by the membership matrix $\mathbf{V}$) and nested due to the ultrametricity of $\boldsymbol{\Sigma}_{\mathrm{u}}$. Note that $\boldsymbol{\Sigma}_{\mathrm{W}}$ and $\boldsymbol{\Sigma}_{\mathrm{B}}$ only determine the $2Q-1$ aggregation levels of the latent concepts, whereas $\boldsymbol{\Sigma}_{\mathrm{V}}$ defines the starting position of the variables in the path diagram (Figure 4.2). EUCovS therefore models the relationships among the dimensions defining a multidimensional phenomenon by means of a set of hierarchically nested latent concepts, each one associated with a group of variables, from the most concordant to the most discordant. The specific concepts are located at the beginning of the hierarchy and are the most internally highly consistent and reliable; the groups are therefore visible along the diagonal blocks of EUCovS, after a row permutation of $\mathbf{V}$ such that the variables belonging to the same group are contiguous. The covariance tends to decrease in the hierarchy when groups of variables aggregate and the associated latent concepts become less consistent (concordant), thus the aggregation levels are discernible as the off-diagonal blocks of EUCovS. The last aggregations in the hierarchy may occur between: (i) concordant concepts defining a general one; (ii) discordant concepts with negative between-group covariance; (iii) uncorrelated concepts. These three scenarios are graphically represented in Figure 4.1.

Finally, the point where each of the Q groups enters the hierarchy assesses the group's initial average level of consistency. The distance between the initial consistency of each group and the corresponding internal node of the tree measures

the non-uniqueness of the concept associated with the group. If the distance is zero, the variance of the group is equal to the covariance within the group and, thus, a unique eigenvalue explains the overall variance of the group.

Our proposal differs from methods which aim at clustering variables by considering - after the transformation of covariances into correlations - the magnitude of the correlation coefficient as a measure of similarity (e.g., taking into account the absolute value or the square of the correlation coefficient, Revelle, 1979; Soffritti, 1999; Liu et al., 2012). The correlation coefficient is thereby transformed into distance in order to apply a hierarchical clustering algorithm (Cliff et al., 1995; Gordon, 1999; Strauss, Bartko, & Carpenter, 1973). Cavicchia, Vichi, and Zaccaria (2020a) showed that this approach has several limitations (see Chapter 3).

The elements of the matrices $\boldsymbol{\Sigma}_{\mathrm{W}}$ and $\boldsymbol{\Sigma}_{\mathrm{B}}$ represent a concordance measure among variables. Equation (4.6) ensures that variables belonging to the same group are more concordant than those in two different groups, i.e., the covariances within groups are greater than the covariances between groups. Thus, the higher the covariances measured on $\mathbb{R}$ are, the stronger the concordance between the two corresponding variables is; the lower the covariances measured on $\mathbb{R}$ are, the higher the discordance between the two corresponding variables is.

In the next section, we introduce a new Gaussian Mixture Model with the proposed covariance structure by inspecting the advantages of assuming this parsimonious parameterization for a covariance matrix.

4.4 Gaussian Mixture Model with an Extended Ultrametric Covariance Structure

The proposal consists of a new Gaussian Mixture Model with the assumption of the Extended Ultrametric Covariance Structure (GMMEUCovS) defined in Section 4.3 for each component of the mixture. On one hand, GMMEUCovS aims at clustering observations by assuming a parsimonious covariance structure for each component of the mixture; on the other hand, when the phenomenon under study is characterized by different dimensions, GMMEUCovS is able to pinpoint the hierarchical relationships among variables within each cluster. We can now formalize the proposed model.

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a random sample - where $\mathbf{x}_i$ is a p-dimensional random vector - which is drawn from a population composed of G subpopulations. Suppose that, conditional on the membership to the subpopulation, the density of $\mathbf{x}_i$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_{\mathrm{u}_g} = \mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')$, where $\boldsymbol{\Sigma}_{\mathrm{u}_g}$ is the EUCovS defined in (4.2) and subject to constraints (4.3)-(4.7) - or, under a less strong condition, constraints (4.3)-(4.6), (4.8) and (4.9). The GMMEUCovS density is

$$f(\mathbf{x}_i|\boldsymbol{\Psi}) = \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}|\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')|^{1/2}}$$
$$\times \exp\Big\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)'[\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')]^{-1}$$
$$(\mathbf{x}_i - \boldsymbol{\mu}_g)\Big\}, \tag{4.10}$$

where $\pi_g, g = 1, \ldots, G$, are the mixing proportions (prior probabilities) and $\boldsymbol{\Psi} = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g : g = 1, \ldots, G\}$ is the overall parameter vector.

The log-likelihood of GMMEUCovS in (4.10) is

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \Big(\sum_{g=1}^{G} \pi_g \, \phi(\mathbf{x}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g) \Big). \tag{4.11}$$

As shown by Hathaway (1986), maximizing (4.11) is equivalent to maximize

$$\begin{aligned} \ell_{\mathrm{H}}(\mathbf{W}, \boldsymbol{\Psi}) = & \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} \log \Big(\pi_g \, \phi(\mathbf{x}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g) \Big) \\ & - \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} \log(w_{ig}), \end{aligned} \tag{4.12}$$

w.r.t. $\mathbf{W} = [w_{ig}] \in M = \{\mathbf{W} \in \mathbb{R}^{nG} : 0 \leq w_{ig} \leq 1, \sum_{g=1}^{G} w_{ig} = 1, 1 < \sum_{i=1}^{n} w_{ig} < n, i = 1, \ldots, n, g = 1, \ldots, G\}$ and $\boldsymbol{\Psi}$.

GMMEUCovS is estimated via a grouped coordinate ascent algorithm (Zangwill, 1969; Bezdek et al., 1987) by maximizing (4.12) w.r.t. the parameters $\mathbf{W}$ and $\boldsymbol{\Psi}$. As demonstrated by Hathaway (1986), the EM algorithm (Dempster, Laird, & Rubin, 1977; Redner & Walker, 1984) usually used to estimate the parameters of a GMM can be interpreted as a method of coordinate ascent on a particular objective function, i.e., (4.12) with $\boldsymbol{\Sigma}_g$ as a generic covariance matrix.

The fundamental steps of the algorithm for the estimation of GMMEUCovS are described as follows.

(a) Estimation of $\mathbf{W} = [w_{ig}]$: it can be easily demonstrated that the estimates of w_{ig} are obtained by maximizing (4.12) over M. Thus,

$$\widehat{w}_{ig} = \frac{\widehat{\pi}_g \, \phi(\mathbf{x}_i | \widehat{\boldsymbol{\mu}}_g, \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}, \widehat{\mathbf{V}}_g)}{\sum_{h=1}^{G} \widehat{\pi}_h \, \phi(\mathbf{x}_i | \widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_h}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_h}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_h}, \widehat{\mathbf{V}}_h)}, \tag{4.13}$$

where $\widehat{w}_{ig}$ is the posterior probability that the ith observation belongs to the gth component ($i = 1, \ldots, n, g = 1, \ldots G$).

(b) Estimation of $\boldsymbol{\pi} = [\pi_g]$: for the estimates of the mixing proportions we can note that (4.12) can be written as

$$\ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \widehat{w}_{ig} \log(\pi_g) + C, \tag{4.14}$$

where C is a constant function w.r.t. $\pi_1, \ldots, \pi_G$. (4.14) is maximized when

$$\widehat{\pi}_g = \frac{n_g}{n} \qquad g = 1, \ldots, G, \tag{4.15}$$

where $n_g = \sum_{i=1}^{n} \widehat{w}_{ig}$.

(c) Estimation of $\boldsymbol{\mu} = [\boldsymbol{\mu}_g]$: for the estimates of the mean vectors we can note that (4.12) can be written as

$$\ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{w}_{ig}\Big[(\mathbf{x}_i - \boldsymbol{\mu}_g)'[\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')]^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\Big] + C, \tag{4.16}$$

where C is a constant function w.r.t. $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$. Equation (4.16) is maximized when

$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n}\widehat{w}_{ig}\,\mathbf{x}_i}{n_g} \quad g = 1, \ldots, G. \tag{4.17}$$

(d) Estimation of the parameters of $\boldsymbol{\Sigma}_{\mathrm{u}} = [\boldsymbol{\Sigma}_{\mathrm{u}_g}]$: for the estimates of $\boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}$, $\boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g$ we can note that (4.12) can be written as

$$\begin{aligned}\ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi}) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{w}_{ig}\Big[\log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|) + (\mathbf{x}_i - \boldsymbol{\mu}_g)'[\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' \\ &\quad - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')]^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\Big] + C \\ &= -\frac{1}{2}\sum_{g=1}^{G} n_g\Big[\log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|) + \mathrm{tr}\Big([\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g \\ &\quad \boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')]^{-1}\mathbf{S}_g\Big)\Big] + C,\end{aligned} \tag{4.18}$$

where C is a constant function w.r.t. $\boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \mathbf{V}_g$, and $\mathbf{S}_g = (1/n_g)$ $\sum_{i=1}^{n}\widehat{w}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)'$ $(g = 1, \ldots, G)$.

(d1) Estimation of $\boldsymbol{\Sigma}_{\mathrm{V}} = [\boldsymbol{\Sigma}_{\mathrm{V}_g}]$: given $\widehat{\mathbf{V}}_g, g = 1, \ldots, G$, we have

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\mathrm{diag}(\mathbf{S}_g)\widehat{\mathbf{V}}_g \qquad g = 1, \ldots, G, \tag{4.19}$$

subject to constraint (4.7).

(d2) Estimation of $\boldsymbol{\Sigma}_{\mathrm{W}} = [\boldsymbol{\Sigma}_{\mathrm{W}_g}]$: given $\widehat{\mathbf{V}}_g, \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}, g = 1, \ldots, G$, we have

$$\begin{aligned}&\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g} = [(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^2 - \widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g]^{-1}\mathrm{diag}\Big[\widehat{\mathbf{V}}_g'\Big(\mathbf{S}_g - \mathrm{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}\widehat{\mathbf{V}}_g')\Big)\widehat{\mathbf{V}}_g\Big] \\ &g = 1, \ldots, G,\end{aligned} \tag{4.20}$$

subject to constraint (4.6).

(d3) Estimation of $\boldsymbol{\Sigma}_{\mathrm{B}} = [\boldsymbol{\Sigma}_{\mathrm{B}_g}]$: given $\widehat{\mathbf{V}}_g, g = 1, \ldots, G$, we have

$$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{B}_g} = \widehat{\mathbf{V}}_g^{+}\mathbf{S}_g(\widehat{\mathbf{V}}_g')^{+} \qquad g = 1, \ldots, G, \tag{4.21}$$

where $\widehat{\mathbf{V}}_g^{+}$ represents the Moore-Penrose inverse of $\widehat{\mathbf{V}}_g$. It is worth noticing that the off-diagonal elements of $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{B}_g}$ simply denote the covariances between Q groups of variables and they do not fulfill the ultrametric condition. To fully satisfy constraint (4.5), $\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}$ is computed such that property (iv) holds.

(d4) Estimation of $\mathbf{V} = [\mathbf{V}_g]$: $\mathbf{V}_g, g = 1, \ldots, G$, is estimated row by row, i.e., for each $\mathbf{v}_{g_j}$, $j = 1, \ldots, p$, $g = 1, \ldots, G$, when all the remaining rows are fixed and the corresponding $\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}, g = 1, \ldots, G$, are computed. This means that the jth variable is assigned to the group q that most increases (4.18). Formally, each row $\mathbf{v}_{g_j}, j = 1, \ldots, p$, of $\mathbf{V}_g$ is estimated by

$$\begin{cases} \widehat{v}_{g_{jq}} = 1 & \text{if} \quad \underset{q=1,\ldots,Q}{\arg\max}\, \ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \widehat{\boldsymbol{\Psi}}_{-g}, [\widehat{\mathbf{v}}_{g_1}, \ldots, \mathbf{v}_{g_j} = \mathbf{i}_q, \ldots, \widehat{\mathbf{v}}_{g_p}]') \text{ in (4.18)}, \\ \widehat{v}_{g_{jq}} = 0 & \text{otherwise} \end{cases} \tag{4.22}$$

where $\widehat{\boldsymbol{\Psi}}_{-g} = \{\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}}, \widehat{\mathbf{V}}_h, h = 1, \ldots, G, h \neq g\}$, $g = 1, \ldots, G$, and $\mathbf{i}_q$ is the qth row of the identity matrix of order Q.

The details of the estimates in steps (d1)-(d4) are provided in Appendix B.

Remark 4.7. The solution $\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}$ is obtained by optimizing (4.18) w.r.t. $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ only, when all the other parameters are fixed, by requiring that constraints in (4.5) are verified. The multivariate constraint problem is solved with an "interior-point" algorithm that satisfies bounds at all iterations. Alternatively to the interior-point algorithm, an adapted average linkage (UPGMA) algorithm for covariance matrices can be used in order to satisfy property (iv). It was used with success in the proposed algorithm.

4.4.1 GMMEUCovS algorithm

The steps described in the previous section are iteratively alternated until convergence. We can briefly show the steps of the algorithm.

Step 0: Initial values for $\widehat{\mathbf{W}} = [\widehat{w}_{ig}]$ and $\widehat{\mathbf{V}} = [\widehat{\mathbf{V}}_g]$ are chosen. Then, initial values for $\widehat{\boldsymbol{\pi}} = [\widehat{\pi}_g], \widehat{\boldsymbol{\mu}} = [\widehat{\boldsymbol{\mu}}_g], \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}} = [\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}], \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}} = [\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}], \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}} = [\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}]$ are computed according to (4.15), (4.17), (4.19), (4.20), (4.21), respectively, subject to the corresponding constraints. $\widehat{\mathbf{W}}^{(1)} = [\widehat{w}_{ig}^{(1)}]$ is computed according to (4.13) given the initial values of the other parameters.

For iteration $t = 1, \ldots, T$:

Step 1: each $\widehat{\pi}_g^{(t)}$ is updated by (4.15), given $\widehat{w}_{ig}^{(t)}, i = 1, \ldots, n, g = 1, \ldots, G$;

Step 2: each $\widehat{\boldsymbol{\mu}}_g^{(t)}$ is updated by (4.17), given $\widehat{w}_{ig}^{(t)}, i = 1, \ldots, n, g = 1, \ldots, G$;

Step 3: each $\widehat{\boldsymbol{\Sigma}}_{\mathrm{u}_g}^{(t)}$ is computed by updating $\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}^{(t)}$ according to (4.19), (4.20), the closest - in the Frobenius norm - matrix to (4.21) which satisfies (4.5), respectively, and subject to the corresponding constraints, which correspond to the configuration of $\widehat{\mathbf{V}}_g^{(t)}$ in (4.22), given $\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_h}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_h}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_h}^{(t)}$, for $h < g$, and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_h}^{(t-1)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_h}^{(t-1)}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_h}^{(t-1)}$, for $h > g$, $g, h = 1, \ldots, G, h \neq g$;

Step 4: each $\widehat{w}_{ig}^{(t+1)}$ is updated according to (4.13), given $\widehat{\mathbf{\Psi}}^{(t)}$.

Stopping rule: Compute $\ell_{\mathrm{H}}(\widehat{\mathbf{W}}^{(t+1)}, \widehat{\mathbf{\Psi}}^{(t)})$ in (4.12). **Steps** from **1** to **4** are repeated if

$$\frac{\ell_{\mathrm{H}}(\widehat{\mathbf{W}}^{(t+1)}, \widehat{\mathbf{\Psi}}^{(t)}) - \ell_{\mathrm{H}}(\widehat{\mathbf{W}}^{(t)}, \widehat{\mathbf{\Psi}}^{(t-1)})}{|\ell_{\mathrm{H}}(\widehat{\mathbf{W}}^{(t)}, \widehat{\mathbf{\Psi}}^{(t-1)})|} > \epsilon,$$

where ϵ is an arbitrary small positive constant, and $t < T$.

Some remarks on the algorithm are necessary. Firstly, the posterior probabilities $\widehat{w}_{ig}$ at convergence are computed and used to determine the cluster membership of observations according to the Maximum A Posteriori (MAP) approach, when a hard partition is required. Secondly, the stopping rule is based upon the sequence of likelihood values as reported in McLachlan and Krishnan (2008). The log-likelihood function generally increases, or does not decrease, at each iteration fulfilling the coordinate ascent (and EM) algorithm properties. It can be noted that $\ell(\widehat{\mathbf{\Psi}}^{(t)})$ is equal to $\ell_{\mathrm{H}}(\widehat{\mathbf{W}}^{(t+1)}, \widehat{\mathbf{\Psi}}^{(t)})$, for all t (Hathaway, 1986). The arbitrary constant ϵ was set to 0.1^{10}, which was considered small enough to be neglected, whereas the maximum number of iterations T was set to 500. It is worth underscoring that in our experiments the algorithm always stopped after a limited number of steps very far from the maximum number of iterations showing that it converges in a finite number of iterations.

One product of GMMEUCovS is also the classification of variables for each component; however, since the problem of optimally partitioning a set of multivariate objects is known to be an NP-hard problem (Krivánek & Morávek, 1986), the global optimal solution cannot be guaranteed. The solution found at convergence is thus at least a local optimum, and to increase the chance to reach the global optimum the algorithm is run several times starting from different initial values. In our experiments, the number of running times was set to 20 and this was sufficient to obtain the optimal solution. As shown in **Step 0**, in order to start the algorithm the posterior probabilities $w_{ig}, i = 1, \ldots, n, g = 1, \ldots, G$, and the membership matrices $\mathbf{V}_g, g = 1, \ldots, G$, are needed. They can be initialized randomly (i.e., random values for $\mathbf{W} \in M$ and random partitions for $\mathbf{V}_g, g = 1, \ldots, G$, with nonempty groups of variables). However, we suggest to start the algorithm from the solution of k-means with $k = G$ in order to find the starting values of $\mathbf{W} = [w_{ig}]$, and from the solution of an adapted UCM algorithm to covariance matrices applied to $\mathbf{S}_g, g = 1, \ldots, G$, to find the partitions of variables in $\mathbf{V}_g, g = 1, \ldots, G$. The initial values of $\mathbf{\Sigma}_{\mathrm{V}_g}$, $\mathbf{\Sigma}_{\mathrm{W}_g}$ and $\mathbf{\Sigma}_{\mathrm{B}_g}$ are accordingly obtained as reported in **Step 0**.

Remark 4.5 stated that the diagonal dominance in (4.7) results in a strong condition which generally leads to an overestimation of the parameters on the diagonal of $\mathbf{\Sigma}_{\mathrm{u}}$. The proposed algorithm presented above and used in Section 4.5 and 4.6 replaces this condition with (4.8) and (4.9), searching for the solution of $\mathbf{\Sigma}_{\mathrm{u}}$ in the positive definite and column pointwise diagonally dominant matrix space.

It is worth highlighting that GMMEUCovS is a parsimonious model since the EUCovS assumed for each component of the GMM allows modeling a generic covariance structure via a limited number of parameters. Specifically, $3Q - 1$

parameters are needed to build a consistent hierarchy of variable groups, each one associated with a concept/dimension (see Corollary 4.1). EUCovS therefore reconstructs a covariance structure in terms of $2Q + p - 1$ unknown free parameters in $\mathbf{\Sigma}_{\mathrm{V}}$, $\mathbf{\Sigma}_{\mathrm{W}}$, $\mathbf{\Sigma}_{\mathrm{B}}$ and $\mathbf{V}$. In detail, Q parameters in $\mathbf{\Sigma}_{\mathrm{V}}$, Q parameters in $\mathbf{\Sigma}_{\mathrm{W}}$, $Q-1$ parameters in $\mathbf{\Sigma}_{\mathrm{B}}$ and $p-Q$ parameters in $\mathbf{V}$ must be considered. The parameters of GMMEUCovS, hence, are $\nu = G(2p+2Q)-1$, by including also the $G-1$ parameters from the mixing proportions and the Gp parameters from the mean vectors.

4.4.2 Model selection

One popular model selection criterion is the Bayesian Information Criterion (BIC, Schwarz, 1978). For a model with parameter vector $\boldsymbol{\theta}$, the BIC is given by

$$\mathrm{BIC} = 2\ell(\widehat{\boldsymbol{\theta}}) - \nu \log n, \tag{4.23}$$

where $\ell(\widehat{\boldsymbol{\theta}})$ is the maximized log-likelihood, $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and ν is the number of free parameters in the model. Given a set of candidate models, specified for different values of G and Q, the preferred model is the one which maximizes BIC. The BIC has several good properties. It is usually considered as standard in clustering context due to the consistent estimates of the number of components of a mixture model (Keribiin, 1998, 2000), and the results provided by Fraley and Raftery (1998, 2002) show that BIC performs well as model selection criterion for mixture models. Moreover, BIC is selected as default model selection criterion in Scrucca et al. (2016) and McNicholas and Murphy (2008), among others.

There are several alternatives to the BIC for GMMs. One of the most popular is the Integrated Completed Likelihood (ICL, Biernacki, Celeux, & Govaert, 2000), which is based on the BIC and penalizes the latter by subtracting an entropy term measuring the clusters' overlapping. It means that ICL prefers solutions with well-separated clusters. Although Scrucca et al. (2016) showed that the performance of BIC and ICL is often comparable, we propose using BIC to choose the number of mixture components G and groups of variables Q in the proposed model.

The performance of GMMEUCovS is evaluated in Section 4.5 through the cross-table of the MAP classification of the observations and the true (generated) cluster membership, which may be quantified using the Adjusted Rand Index (ARI, Hubert & Arabie, 1985). The variable partition in Q groups and the hierarchy over them are evaluated according to the ARI as well.

4.5 Simulation

A simulation study was implemented in order to evaluate the clustering performance of GMMEUCovS. We assessed the classification performance of the proposed model in terms of recovering both the generated cluster structure, and the true classification of variables in Q groups together with their hierarchical relationships. In detail, three hierarchical scenarios were considered (Scenario 1, Scenario 2, Scenario 3) and one non-hierarchical scenario (Scenario 4) completed the simulation study in order to evaluate the performance of GMMEUCovS when the component covariance structures were non-hierarchical.

For all scenarios, each simulated random sample was generated from the density (4.1), with the component label of each unit generated from a multinomial distribution with fixed probabilities $\pi_1, \ldots, \pi_G$, determining the cluster sizes, vector means $\boldsymbol{\mu}_g, g, = 1, \ldots, G$, generated from a uniform distribution in $[0, 10]$ and covariance matrices generated according to the EUCovS in (4.2) for the hierarchical scenarios and non-spherical and heteroscedastic structure for the non-hierarchical scenario. Moreover, different levels of overlapping were set for all scenarios. The overlapping measure chosen for the simulation study is the one proposed by Maitra and Melnykov (2010) and implemented in the R package `MixSim`, used for the simulation of the random samples in the non-hierarchical scenario, and the MATLAB package `FSDA` (Riani, Perrotta, & Torti, 2012; Riani et al., 2015), whose function `MixSim()` was modified to impose the EUCovS on the cluster covariance matrices in the hierarchical scenarios. Three levels of maximum overlapping ($\omega_{\max}$) were set in the simulation study: $0.01, 0.1, 0.2$. Thus, 2400 random samples were generated in the whole simulation study, i.e., 200 samples for each scenario and overlapping level.

GMMEUCovS was also compared to GPCMs (R package `mixture`), EPGMMs (R package `pgmm`) and HDDC (R package `HDClassif`) in all scenarios. For comparability reasons, the R package `mixture` was preferred to the other ones cited in Section 4.1 for GPCMs since its `gpcm()` function enables to use a k-means start. The R packages `pgmm` and `HDClassif` provide an initialization via k-means as well. GPCMs, EPGMMs and HDDC were let free to choose their best model, i.e., the constrained or fully unconstrained covariance structures, according to the BIC. It is worth noting that, as for the competitors, the covariance structure of the proposed model could be constrained to be equal across and within clusters. This extension of GMMEUCovS will be introduced in a future work together with an extensive comparison with the aforementioned models.

In the hierarchical scenarios, the covariance matrices were generated according to the EUCovS in (4.2). The parameters of $\boldsymbol{\Sigma}_{\mathrm{u}_g}$ were generated as follows. For each component of the mixture, $\mathbf{v}_{g_j} \sim \text{Multinomial}(Q : \text{prob}(q) = \frac{1}{Q},\ q = 1, \ldots, Q),\ j = 1, \ldots, p$; each diagonal element of $\boldsymbol{\Sigma}_{\mathrm{V}_g}$, $\boldsymbol{\Sigma}_{\mathrm{W}_g}$, and off-diagonal element of $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ was uniformly distributed according to different ranges specifically chosen for each hierarchical scenario and such that constraints (4.3)-(4.6), (4.8) and (4.9) held. Positive values were generated for the diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{W}_g}$, $g = 1, \ldots, G$. In detail, the three hierarchical scenarios are defined as follows.

- *Scenario 1*: $n = 200$, $p = 10$, $G^* = 3$, $Q^* = 3$, $\boldsymbol{\pi} = [0.25, 0.25, 0.5]$, negative, near-zero and positive values for the off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_1}$, $\boldsymbol{\Sigma}_{\mathrm{B}_2}$ and $\boldsymbol{\Sigma}_{\mathrm{B}_3}$, respectively.

- *Scenario 2*: $n = 300$, $p = 25$, $G^* = 4$, $Q^* = 3$, $\boldsymbol{\pi} = [0.15, 0.15, 0.3, 0.4]$, negative values for the off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_1}$, $\boldsymbol{\Sigma}_{\mathrm{B}_4}$ and positive values for the off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_2}$, $\boldsymbol{\Sigma}_{\mathrm{B}_3}$.

- *Scenario 3*: $n = 400$, $p = 50$, $G^* = 5$, $Q^* = 4$, $\boldsymbol{\pi} = [0.1, 0.15, 0.15, 0.5, 0.1]$, one positive and two negative values for the different off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_1}$, one positive and two near-zero values for the different off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_2}$, three positive values for the different off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_3}$ and

Table 4.1. Mean and standard deviation of the ARI (mARI and sARI, respectively) between the generated and the estimated clusters for the three hierarchical scenarios.

	Scenario 1							
	GMMEUCovS		**GPCMs**		**EPGMMs**		**HDDC**	
	mARI	sARI	mARI	sARI	mARI	sARI	mARI	sARI
$\omega_{\max} = 0.01$	0.982	0.050	0.902	0.154	0.958	0.098	0.954	0.096
$\omega_{\max} = 0.10$	0.800	0.081	0.444	0.198	0.681	0.217	0.581	0.239
$\omega_{\max} = 0.20$	0.572	0.110	0.215	0.121	0.232	0.228	0.280	0.195
	Scenario 2							
	GMMEUCovS		**GPCMs**		**EPGMMs**		**HDDC**	
	mARI	sARI	mARI	sARI	mARI	sARI	mARI	sARI
$\omega_{\max} = 0.01$	0.993	0.008	0.474	0.128	0.896	0.149	0.873	0.137
$\omega_{\max} = 0.10$	0.919	0.042	0.268	0.059	0.608	0.206	0.623	0.169
$\omega_{\max} = 0.20$	0.803	0.082	0.242	0.046	0.435	0.153	0.487	0.143
	Scenario 3							
	GMMEUCovS		**GPCMs**		**EPGMMs**		**HDDC**	
	mARI	sARI	mARI	sARI	mARI	sARI	mARI	sARI
$\omega_{\max} = 0.01$	0.947	0.072	0.040	0.049	0.363	0.116	0.353	0.103
$\omega_{\max} = 0.10$	0.812	0.053	-0.018	0.014	0.111	0.070	0.046	0.040
$\omega_{\max} = 0.20$	0.747	0.104	0.515	0.222	0.574	0.225	0.587	0.256

$\boldsymbol{\Sigma}_{\mathrm{B}_4}$, and two positive and one negative values for the different off-diagonal entries of $\boldsymbol{\Sigma}_{\mathrm{B}_5}$.

It is worth noting that in order to correctly compare the generated and the estimated partitions of variables in Q groups and their hierarchical relationships, i.e., to compare the generated and estimated variable classification w.r.t. the same cluster of units, the label switching problem was solved by implementing the *Complete likelihood-based labelling* (COMPLH) method proposed by Yao (2015). For each random sample and cluster, the variable partition recovery was assessed by means of the ARI and the hierarchy evaluation was performed by computing the mean of the ARI across the hierarchical levels, i.e., from Q to 2.

The four models' performances were compared on the correct identification of the cluster structure by fixing $G = G^*$ (and $Q = Q^*$ for GMMEUCovS and EPGMMs) and letting GPCMs, EPGMMs and HDDC free to choose their best model, i.e., the covariance structure, according to the BIC. Additionally, on the same generated random samples for each hierarchical scenario we fitted the four models with G in $\{G^* - 1, G^*, G^* + 1\}$ (and Q in $\{Q^* - 1, Q^*, Q^* + 1\}$, $Q \leq G$, for GMMEUCovS and EPGMMs) in order to evaluate the performance of the models in correctly identifying G (and Q for GMMEUCovS and EPGMMs). We used the BIC to choose G and Q for GMMEUCovS, as described in Section 4.4.2, and to choose both G (and Q for EPGMMs) and the covariance structure for the competitors.

Table 4.1 provides the results of the simulation study evaluated in terms of the mean and standard deviation of the ARI between the generated and the estimated

Table 4.2. Mean of the ARI between the generated and the estimated variable partitions for each cluster of GMMEUCovS both for the Qth level of the hierarchy (mARI) and across the hierarchical levels $Q, \ldots, 2$ (hARI) for the three hierarchical scenarios.

			$\omega_{\max} = 0.01$	$\omega_{\max} = 0.10$	$\omega_{\max} = 0.20$
Scenario 1	Cluster 1	mARI	0.988	0.982	0.981
		hARI	0.872	0.868	0.868
	Cluster 2	mARI	0.965	0.936	0.786
		hARI	0.898	0.860	0.742
	Cluster 3	mARI	0.998	0.991	0.992
		hARI	0.966	0.962	0.968
Scenario 2	Cluster 1	mARI	0.993	0.967	0.793
		hARI	0.838	0.765	0.661
	Cluster 2	mARI	0.999	0.981	0.942
		hARI	0.930	0.902	0.898
	Cluster 3	mARI	1.000	0.987	0.977
		hARI	0.984	0.949	0.966
	Cluster 4	mARI	1.000	0.995	0.998
		hARI	0.905	0.885	0.901
Scenario 3	Cluster 1	mARI	0.895	0.964	0.447
		hARI	0.839	0.885	0.429
	Cluster 2	mARI	0.946	0.995	0.692
		hARI	0.829	0.864	0.606
	Cluster 3	mARI	0.916	0.927	0.599
		hARI	0.807	0.796	0.523
	Cluster 4	mARI	0.999	1.000	0.972
		hARI	0.879	0.884	0.833
	Cluster 5	mARI	0.904	0.980	0.414
		hARI	0.853	0.924	0.391

clusters, for each hierarchical scenario and overlapping level. The results show that GMMEUCovS outperforms GPCMs, EPGMMs and HDDC in identifying the true clusters. As the maximum overlapping grows, the performance of the four models deteriorates; however, the proposed model always performs better than the competitors even when the maximum overlapping is high. The simulation study underlines the need to use the specific GMMEUCovS methodology when data have a hierarchical covariance structure, since this last strongly influences the clustering results of the other GMM considered competitors.

The GMMEUCovS performance was also evaluated in terms of the correct classification of variables in Q groups and the overall hierarchy in the hierarchical scenarios (Table 4.2). The proposed model shows good results in identifying the variable partitions, especially when $\omega_{\max} = 0.01, 0.1$. Finally, the four models were compared in the correct identification of G^* for GPCMs and HDDC, and G^* and

Table 4.3. % of samples with a correct choice of G (and Q for GMMEUCovS and EPGMMs) for the three hierarchical scenarios.

	Scenario 1			
	GMMEUCovS	**GPCMs**	**EPGMMs**	**HDDC**
$\omega_{\max} = 0.01$	98.5%	85.5%	84.5%	91.5%
$\omega_{\max} = 0.10$	97.0%	41.0%	61.5%	53.0%
$\omega_{\max} = 0.20$	99.5%	15.5%	33.5%	22.0%
	Scenario 2			
	GMMEUCovS	**GPCMs**	**EPGMMs**	**HDDC**
$\omega_{\max} = 0.01$	92.0%	1.5%	62.0%	53.5%
$\omega_{\max} = 0.10$	80.5%	5.5%	37.0%	31.0%
$\omega_{\max} = 0.20$	84.0%	5.5%	29.0%	16.0%
	Scenario 3			
	GMMEUCovS	**GPCMs**	**EPGMMs**	**HDDC**
$\omega_{\max} = 0.01$	94.5%	23.5%	6.0%	25.0%
$\omega_{\max} = 0.10$	91.0%	44.5%	4.0%	33.0%
$\omega_{\max} = 0.20$	21.0%	7.5%	13.0%	12.0%

Table 4.4. Mean and standard deviation of the ARI (mARI and sARI, respectively) between the generated and the estimated clusters for the non-hierarchical scenario.

	Scenario 4							
	GMMEUCovS		**GPCMs**		**EPGMMs**		**HDDC**	
	mARI	sARI	mARI	sARI	mARI	sARI	mARI	sARI
$\omega_{\max} = 0.01$	0.953	0.009	0.901	0.019	0.893	0.022	0.894	0.022
$\omega_{\max} = 0.10$	0.730	0.004	0.733	0.041	0.767	0.017	0.636	0.050
$\omega_{\max} = 0.20$	0.255	0.091	0.205	0.068	0.214	0.058	0.192	0.092

Q^* for GMMEUCovS and EPGMMs, as shown in Table 4.3.

In the non-hierarchical scenario, the covariance matrices were generated such that the mixture components turned out to be non-spherical and heteroscedastic. In detail,

- *Scenario 4*: $n = 200$, $p = 10$, $G^* = 3$, $\pi_g^{\min} = 0.25$, non-spherical and heteroscedastic components.

As shown in Table 4.4, the results of the proposal are comparable with those of the competitors with fixed $G = G^*$ (and $Q = Q^*$ for GMMEUCovS and EPGMMs). Indeed, GMMEUCovS often outperforms GPCMs, EPGMMs and HDDC even if the differences between the ARI of the four models are much weaker than those in the hierarchical scenarios. This last scenario verifies that GMMEUCovS still performs

Table 4.5. List of the OECD countries.

Country	Code	Country	Code	Country	Code	Country	Code
Australia	AUS	France	FRA	Korea	KOR	Portugal	PRT
Austria	AUT	Germany	DEU	Latvia	LVA	Slovak Republic	SVK
Belgium	BEL	Greece	GRC	Lithuania	LTU	Slovenia	SVN
Canada	CAN	Hungary	HUN	Luxembourg	LUX	Spain	ESP
Chile	CHL	Iceland	ISL	Mexico	MEX	Sweden	SWE
Czech Republic	CZE	Ireland	IRL	Netherlands	NLD	Switzerland	CHE
Denmark	DNK	Israel	ISR	New Zealand	NZL	Turkey	TUR
Estonia	EST	Italy	ITA	Norway	NOR	United Kingdom	GBR
Finland	FIN	Japan	JPN	Poland	POL	United States	USA

well also when a general (non-hierarchical) covariance structure is observed for the data. Also in this case there is not risk to strongly fail in the correct classification when the data are heterogeneous and well-structured.

4.6 Application

In this section we consider two real data examples: the first one concerning well-being (Section 4.6.1) and the second one on the chemical properties of coffee (Section 4.6.2). The application of GMMEUCovS on the Well-Being Indicators data set pinpoints positive (concordant) and near-zero hierarchical relationships among variable groups. In the second example, a benchmark data set is considered (see, for example, McNicholas & Murphy, 2008) in order to evaluate GMMEUCovS on well-known data. The latter example also shows the capability of the proposal in identifying discordant latent concepts, since it assesses negative hierarchical relationships among the variable groups.

4.6.1 Well-Being Indicators data set

GMMEUCovS was applied on the Well-Being Indicators data set[1] provided by the Organization for Economic Co-operation and Development (OECD) and collected on the 36 OECD countries at 2018 (Table 4.5).

The data set is comprised of eleven dimensions of well-being: *Education* (1), *Jobs* (2), *Income* (3), *Safety* (4), *Health* (5), *Environment* (6), *Civic Engagement* (7), *Accessibility to Services* (8), *Housing* (9), *Community* (10) and *Life Satisfaction* (11). They are related to material living conditions concerning economic aspects $(2, 3, 9)$, and to quality of life pertaining to individual aspects $(1, 4, 5, 6, 8, 11)$ and relational ones $(7, 10)$. Notwithstanding the two aforementioned broader concepts, i.e, material living conditions and quality of life, countries can differ for the importance of the eleven dimensions in the definition of well-being. In order to investigate these differences, GMMEUCovS was fitted to the Well-Being Indicators data set for $G = 1, 2, \ldots, 5$ and $Q = 1, 2, \ldots, 5$. The optimal model was selected according to the highest BIC, as described in Section 4.4.2, and corresponded to $G = 2$ and $Q = 4$. The resulting clusters (Table 4.6) separate the 36 OECD countries into more

[1]Source: https://www.oecdregionalwellbeing.org/.

Table 4.6. GMMEUCovS clusters of countries.

Cluster	Countries
Cluster 1	AUS, AUT, BEL, CAN, DNK, FIN, FRA, DEU, ISL, IRL, ITA, JPN, LUX, NLD, NZL, NOR, ESP, SWE, CHE, GBR, USA
Cluster 2	CHL, CZE, EST, GRC, HUN, ISR, KOR, LVA, LTU, MEX, POL, PRT, SVK, SVN, TUR

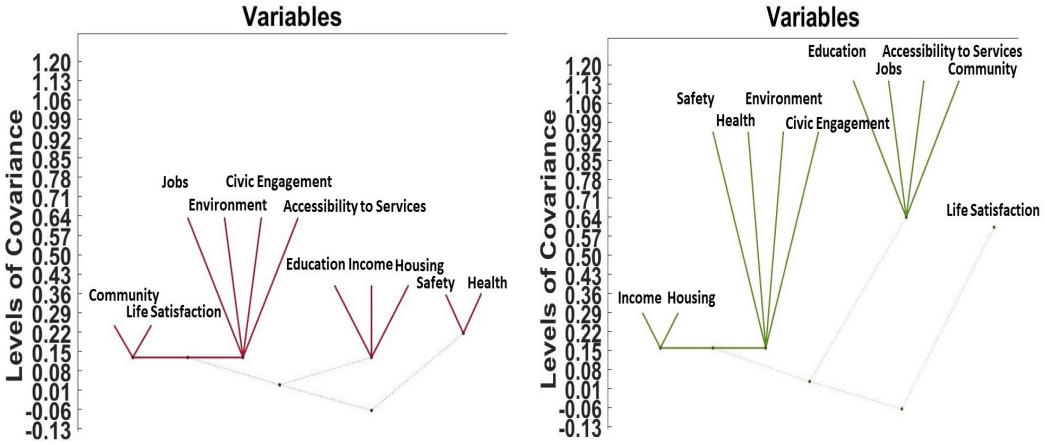
developed and less developed economies. These two clusters differ in the partition of the eleven dimensions into groups and their hierarchical structures defining well-being, as shown in Figure 4.3, even if some similarities between the two hierarchies can be highlighted. Indeed, *Income* and *Housing* are in the same group of variables both for cluster 1 and 2 since they are strictly related to monetary dimensions, but associated with *Education* for more developed economies. Thus, in those countries the higher the level of education, the higher the income. In both cases, the group associated with those dimensions seems to identify a unique broader concept, specifically in Figure 4.3b where the distance between the initial consistency of the group and the corresponding internal node is very small. *Safety* and *Health* are merged in both clusters and grouped together with *Environment* and *Civic Engagement* in less developed economies. Since the latter are related to air and water quality, and voter turnout and consultation on rule making, respectively, we can state that, in cluster 2, less polluted countries are those where the health conditions are better and countries where the population is more involved in political life are those where citizens feel safer. Furthermore, *Life Satisfaction* is associated with *Community* only for more developed economies by detecting that citizens with high social connections are more satisfied in their lives. The last two groups of dimensions described are strongly associated with a unique concept in Figure 4.3a (trivial for *Life Satisfaction* in Figure 4.3b). Finally, *Jobs* is lumped together with *Accessibility to Services* in both clusters of countries, but associated with *Environment* and *Civic Engagement* in more developed economies and with *Education* and *Community* in less developed ones.

The hierarchical structures built over the aforementioned groups of variables pinpoint a broader group which turns out to be uncorrelated with the others. Indeed, after a first aggregation which lumps together *Jobs*, *Environment*, *Civic Engagement*, *Accessibility to Services*, *Community*, *Life Satisfaction* for more developed economies, and *Income*, *Housing*, *Safety*, *Health*, *Environment*, *Civic Engagement* for less developed economies, these broader groups are faintly associated with the others. Thus, GMMEUCovS identified two different structures of the eleven dimensions, which differently characterize well-being in the two clusters of countries, but detect uncorrelated concepts associated with groups of dimensions.

4.6.2 Coffee data set

We applied GMMEUCovS on coffee data[2] (Streuli, 1973) in order to assess its performance also on a benchmark data set. The latter is composed of 43 coffee

[2]Available within the R package `pgmm`.



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

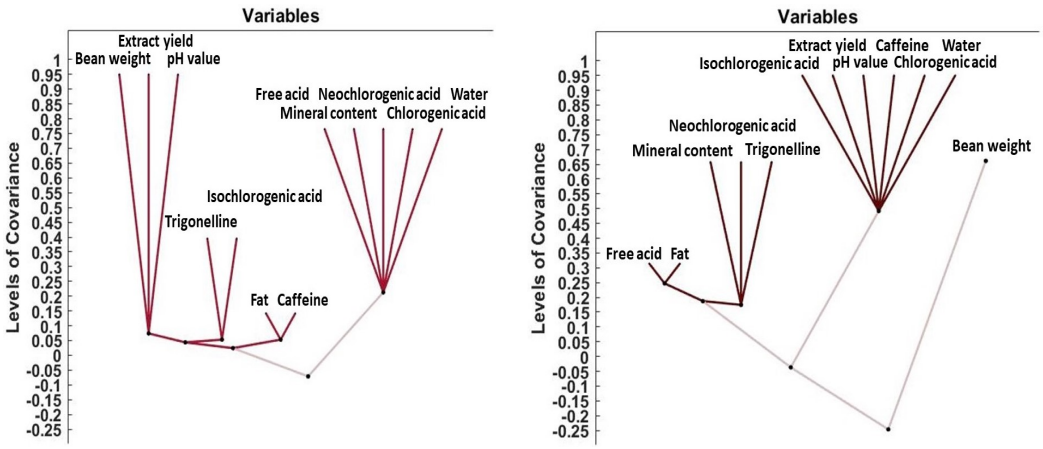
Figure 4.3. GMMEUCovS variable hierarchies for each cluster of countries.

samples pertaining two varieties of beans, namely Arabica and Robusta, which 13 chemical properties were measured on. We compared the results of GMMEUCovS with the ones reported in McNicholas and Murphy (2008), other than GPCMs and HDDC. For comparability reasons, we did not consider the variable *total chlorogenic acid* as done by McNicholas and Murphy (2008), since it is the sum of other three chemical constituents. Twelve variables were therefore taken into account.

GMMEUCovS was fitted to the coffee data with $G = 1, \ldots, 5$ and $Q = 1, \ldots, 5$. According to the BIC, the best model is that one with $G = 2$ and $Q = 4$, as also found by PGMMs in McNicholas and Murphy (2008). The two theoretical clusters corresponding to the coffee species, composed of 36 samples of Arabica and 7 of Robusta, are perfectly recovered (ARI = 1) by GMMEUCovS. Compared to the results in McNicholas and Murphy (2008), the proposal performs as well as PGMMs, whereas the ARI for `Mclust()` in the R package `mclust` is 0.38 with $G = 3$. In order to complete the comparison w.r.t. the other models used in Section 4.5, GPCMs and HDDC were fitted to the data with $G = 1, \ldots, 5$. The former perfectly recovers the theoretical clusters ($G = 2$, ARI = 1), whereas the latter identifies 3 clusters with ARI = 0.97. It is worth noting that the coffee data set is not high-dimensional, so that the HDDC performance is not as good as the other competitors and numerical instability occurs in the algorithm.

The four groups of variables pinpointed by GMMEUCovS are the following (Figure 4.4).

- *Arabica*
  1. Fat, Caffeine;
  2. Bean weight, Extract yield, pH value;
  3. Trigonelline, Isochlorogenic acid;
  4. Water, Free acid, Mineral content, Chlorogenic acid, Neochlorogenic acid.
- *Robusta*



(a) *Arabica* **(b)** *Robusta*

Figure 4.4. GMMEUCovS variable hierarchies for each cluster of coffee data, i.e., Arabica and Robusta.

1. Water, Extract yield, pH value, Caffeine, Chlorogenic acid, Isochlorogenic acid;
2. Free acid, Fat;
3. Bean Weight;
4. Mineral content, Trigonelline, Neochlorogenic acid.

As shown in Figure 4.4a, for the first cluster (Arabica) GMMEUCovS merges group 2 and group 3, then this broader group with group 1, and finally group 4. Only the last aggregation corresponds to a negative covariance between the corresponding variable groups by showing that group 4 is discordant with the others. Looking at Figure 4.4b, for the second cluster (Robusta) GMMEUCovS lumps together group 2 and group 4 with a positive covariance between them, whereas the other aggregations correspond to negative covariances between (discordant) groups - group 1 is firstly merged with the broader group composed of groups 2 and 4, and then with group 3.

4.7 Conclusions

We propose a parsimonious Gaussian Mixture Model with a new parameterization of the component covariance structure. After having extended the definition of a nonnegative ultrametric correlation matrix given in Chapter 2 (Definition 2.2) to a generic covariance one by illustrating its properties, we assume that each component of the GMM has an extended ultrametric covariance structure. The proposal, called Gaussian Mixture Model with the Extended Ultrametric Covariance Structure (GMMEUCovS), aims at modeling multidimensional phenomena which are usually defined by hierarchically nested latent concepts, by inspecting the different characterizations that these phenomena can have in subpopulations. The model is able to identify common hierarchical scenarios: firstly, the presence of a unique, consistent and reliable general concept which is defined by even more

reliable, nested and specific concepts. This situation is characterized by a general concordance (positive covariances) among all observed variables. Secondly, two (or more) discordant concepts which form an unreliable general concept. These concepts are internally consistent yet discordant to each other corresponding to negative covariances among some of the variables. Thirdly, the absence of a general concept. In detail, the latter means that the last two (or more) concepts, which should form the general one, are substantially uncorrelated and define two (or more) separated hierarchies. The last scenario corresponds to sparse covariance matrices. In literature, some methodologies were proposed to deal with model-based clustering with sparse covariance matrices. Among others, Galimberti and Soffritti (2013) introduced a parsimonious Gaussian Mixture Model with block diagonal covariance matrices derived from a partition of variables into groups, which were conditionally independent within clusters. Moreover, Fop, Murphy, and Scrucca (2019) proposed a mixture of Gaussian covariance graph models in which the component covariance matrices were sparse, without necessarily imposing a block structure on them. These two methodologies differ from GMMEUCovS since the latter also allows inspecting the hierarchical relationships among variables. Additionally, GMMEUCovS can be interpreted as a bi-clustering method with the special feature that it produces a partition or hierarchy of variables per cluster of units, whereas a bi-clustering method generally produces a unique variable partition (see Rocci & Vichi, 2008, for some extensions).

In order to estimate the GMMEUCovS parameters, we propose a coordinate ascent algorithm which is strictly related to the EM algorithm. Its application on synthetic data shows good performance in terms of cluster recovering, even by comparison with the Gaussian Parsimonious Clustering Models, Parsimonious Gaussian Mixture Models and High-Dimensional Data Clustering models, and correct identification of the variable groups and their hierarchical relationships. The proposed model works particularly well in situations where groups of highly concordant variables exist and a hierarchy over them can be identified. Nonetheless, the proposal shows good performance also when a general (non-hierarchical) covariance structure is assumed for the data. The application of the proposal to real data concerning well-being and a benchmark data set illustrates its potentials to explore multidimensional phenomena in a heterogeneous population.

The number of the GMMEUCovS parameters grows linearly with both the data dimension and the number of the variable groups; this sheds light on the parsimony of the proposed model. Our goal for future studies is to extend our proposal by including constrained covariance structures across and within clusters which further reduce the number of parameters of the model.

Chapter 5

Hierarchical Disjoint Principal Component Analysis

5.1 Introduction

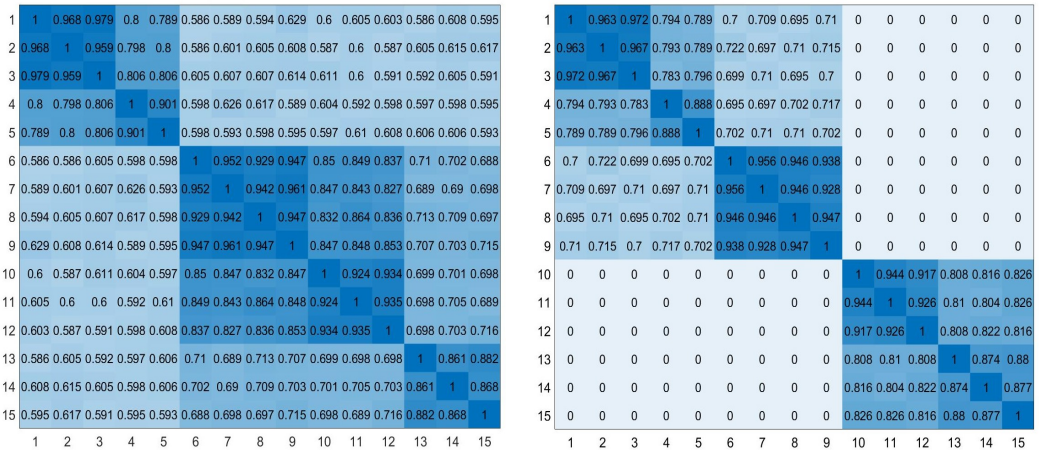
The methodologies proposed in Chapter 2 and 4 aim at modeling a hierarchical structure of nested latent concepts underlying a multidimensional phenomenon via an ultrametric correlation matrix and its extension to a generic covariance one (i.e., including also negative values) implemented into a Gaussian mixture model, respectively. The purpose of this chapter is to introduce a new hierarchical model which inspects the relationships among latent dimensions, as well as the aforementioned methodologies, but quantifying the latent concepts throughout the hierarchy. Differently from the proposals illustrated in Chapter 2 and 4, the model proposed herein is not directly based upon an ultrametric (covariance or correlation) matrix; nonetheless, it works toward the construction of a hierarchy of nested latent dimensions whose quantification is computed via an extension of Principal Component Analysis (Pearson, 1901; Hotelling, 1933). It is worth underlying that the Ultrametric Correlation Model presented in Chapter 2 is implemented herein in order to generate data with a hierarchical structure.

One of the main ideas motivating Principal Component Analysis (PCA) and Factor Analysis (FA, Spearman, 1904) is to reduce the dimensionality of the data by computing a reduced number of components or factors, respectively, but preserving as much information as possible regarding the relationships among the observed variables. Despite the difference between the two methodologies (e.g., Jolliffe, 2002), both suffer from the interpretation problem of the components or factors. In the case of PCA, uncorrelated components are identified as linear combinations of *all* observed variables by maximizing the explained total variance of the data. The weights of the observed variables in each linear combination are different in magnitude, and sometimes irrelevant insofar they are artificially set to zero. Cadima and Jolliffe (1995) demonstrated that this procedure may be misleading. Several methodologies were proposed in order to improve the component interpretation by identifying consistent subsets of observed variables (Cadima & Jolliffe, 1995; Vines, 2000; Jolliffe, Trendafilov, & Uddin, 2003; Zou, Hastie, & Tibshirani, 2006; d'Aspremont et al., 2007; Ferrara, Martella, & Vichi, 2016, among others). These methodologies

result in a sparse structure of the component loading matrix which might be a simple structure when the subsets of the observed variables are not necessarily disjoint (Thurstone, 1947), or the sparsest one when they are disjoint. In those cases, as well as after an oblique rotation, components can be significantly correlated and a dimension reduction technique as PCA can be further applied on the component score vectors thereby obtaining a hierarchy of higher-order components. Several methods were introduced in an attempt to investigate hierarchical relationships among observed variables, such as Higher-Order Factor Models (HOFMs, Cattell, 1978b; Rindskopf & Rose, 1988; Undheim & Gustafsson, 1988; Le Dien & Pagès, 2003) and Bi-Factor or Hierarchical Factor models (Holzinger & Swineford, 1937; Wherry, 1959), illustrated in Chapter 1 (Section 1.1). HOFMs are based upon *sequential* applications of FA on the covariance or correlation matrix of the observed variables first, and on that of higher-order factors then, followed each time by an Oblique Rotation Method (ORM) until zero correlation occurs among factors or a single factor is detected (Gorsuch, 1983). Hierarchical models are instead based on the solution of HOFMs followed by the Schmid-Leiman transformation (Schmid & Leiman, 1957) in order to pinpoint a single general factor and a set of orthogonal specific ones, all directly associated with the observed variables. Both methodologies can be applied in an exploratory approach, i.e., without fixing the relationships between observed and unobserved variables a priori. Contrariwise, simultaneous methodologies like Structural Equation Modeling (SEM, Kline, 2015) were introduced in order to evaluate the relationships among observed and unobserved variables, as well as among the latter ones, but in a confirmatory approach and without a hierarchy in the broad sense over them (see Chapter 1, Section 1.1).

By considering the limitations of PCA as a technique of dimension reduction when complex relations among variables exist, and especially when these relations have a hierarchical structure, in this chapter we propose a new methodology to identify a reduced (parsimonious) hierarchy of disjoint principal components with the highest variance. In order to define a hierarchical structure, the nature of the relationships among variables belonging to two sequential levels of the hierarchy needs to be established. Their "direction" formally describes the measurement model (Blalock, 1964; Bollen & Bauldry, 2011) and affects the relationships between observed and unobserved variables as well as unobserved variables and "higher-order" ones. For simplicity and clarity reasons, we focus on the relationships between observed and unobserved variables. Nonetheless, the following definition can be easily extended to the relationships between unobserved variables of two sequential hierarchical levels.

Two different directions exist, *reflective* and *formative*, as described in Chapter 1, Section 1.3. We recall that the former occurs when a set of correlated observed variables reflects the unobserved one; whereas the latter arises when the unobserved variable is defined/formed by a set of observed variables - each one representing a unique part of it - that are generally uncorrelated to each other or little correlated. Reflective relationships are usually modeled via FA, whereas formative relationships are formalized via multiple and multivariate linear regression (Blalock, 1964; Bollen & Lennox, 1991, among others). PCA is often considered as a formative measurement model (Edwards & Bagozzi, 2000); nonetheless, Mazziotta and Pareto (2019) demonstrated that PCA is more suited to model reflective relationships than



(a) *Fully reflective* **(b)** *Reflective/formative*

Figure 5.1. Examples of the heat maps of two correlation matrices representing the different nature of relationships in a hierarchy over observed variables.

formative ones (see also Götz, Liehr-Gobbers, & Krafft, 2010). With respect to the aforementioned models, HOFMs and Hierarchical Factor models were developed to build only a reflective hierarchy. In SEM, the measurement model underlying a multidimensional phenomenon is usually chosen by the researcher. Nevertheless, in some situations researchers may not have a theoretical definition of the hierarchical relationships among variables, or maybe this may not be empirically confirmed by a test.

In order to understand the hierarchical relationships among observed and unobserved variables we show two examples in which a hierarchy over observed variables measuring a multidimensional phenomenon is associated with a correlation matrix. Let 15 observed variables be arranged into 5 thoroughly internally correlated groups, which correspond to the blocks of higher correlations in Figure 5.1. The relationships among these groups are depicted by nested blocks of correlation sub-matrices. In Figure 5.1a, statistically significant correlations among the 5 groups exist and can be sorted in a decreasing order according to their magnitude by defining a hierarchical structure of relationships. Indeed, the highest correlation (0.84 on average) occurs between the variables belonging to the third and the fourth groups, then between the ones in the first and second groups (0.80 on average), and so on, up to the last one (0.60 on average) between the two nested groups composed of the first two and the last three groups out of the five original ones. Thus, we can assume that the hierarchy associated with this block correlation matrix has a reflective nature. Contrariwise, in Figure 5.1b two uncorrelated nested groups of observed variables are identified, which, in turn, are composed of 3 and 2 groups. Hence, if each of the five original groups and the broader ones are associated with an unobserved variable, we can state that the hierarchy ends with the definition of two uncorrelated unobserved variables which do not reflect the general one, but rather form it.

In this chapter, we propose a hierarchical extension of the Disjoint Principal Component Analysis (Ferrara, Martella, & Vichi, 2016) which aims at detecting a parsimonious hierarchy of disjoint principal components of maximum variance, and

defining a model-based approach to choose the "direction" (reflective or formative) of relationships among disjoint principal components in two contiguous levels of the hierarchy. The new methodology, called Hierarchical Disjoint Principal Component Analysis (HierDPCA), is exploratory, but the researcher can fix some (or all) relations between observed variables and disjoint principal components of the lowest level of the hierarchy so that the methodology becomes partially (or fully) confirmatory. Moreover, HierDPCA overcomes the duality between HOFMs and Hierarchical Factor models (Yung, Thissen, & McLeod, 1999) by quantifying direct relationships between observed variables and disjoint principal components at each hierarchical level, but building a hierarchy over them thanks to the nestedness assumption among variable partitions in two contiguous hierarchical levels. The proposed methodology can also be considered for preliminary analyses in other dimension reduction techniques in order to choose the optimal number of components to retain. Indeed, HierDPCA builds a hierarchy of disjoint principal components which is associated with a tree structure (Gordon, 1999), from the leaves to the root which represent the specific (disjoint principal) components and the general one, respectively. According to this representation, the number of components to be used in a further dimensionality reduction analysis can be chosen by cutting the tree visually evaluating the difference among levels or, for instance, where the hierarchy turns from reflective to formative.

The overview of the chapter is defined as follows. In Section 5.2, the notation used herein is listed to help the reader following the mathematical parts of the chapter. Section 5.3 is dedicated to the in-depth explanation of the Hierarchical Disjoint Principal Component Analysis model: the least-squares estimation of the proposal is given (Section 5.3.1) and a coordinate descent algorithm is provided (Section 5.3.2). Section 5.4 illustrates a simulation study in order to assess the performances of the model, and two applications on real data are implemented in Section 5.5. A final discussion completes the chapter in Section 5.6.

5.2 Notation

For the convenience of the reader, the notation and terminology used in all sections of this chapter are set out below.

n, p	Number of units and observed variables, respectively.
Q	Number of disjoint principal components that firstly reduce the dimensionality of the data.
M	First bottom-up level of the hierarchy at which the correlation among disjoint principal components turns out to be not statistically significant. If $M = 1$, the hierarchy is reflective; if $M > 1$ (up to Q), the hierarchy is reflective/formative.
$\mathbf{X} = [x_{ij} : i = 1, \ldots, n, j = 1, \ldots, p]$	$(n \times p)$ data matrix, where x_{ij} is the observed value on the ith unit for the jth variable.

$\mathcal{P}_H = \{\mathcal{P}_{MH}, \ldots, \mathcal{P}_{QH}\}$	Hierarchical partition set of the variable space in $M, \ldots, Q$ non-overlapping groups, i.e., set of partitions for each hierarchical level $q = M, \ldots, Q, M \in \{1, \ldots, Q\}$.
$J_{h(q)}$	Number of observed variables in the hth group of the qth partition $\mathcal{P}_{qH}$. For any $q = M, \ldots, Q$, $J_{1(q)} + J_{2(q)} + \cdots + J_{q(q)} = p$.
$\text{diag}(\mathbf{a}), \text{diag}(\mathbf{A})$	Diagonal matrix with diagonal entries equal to the vector $\mathbf{a}$ and diagonal matrix with diagonal entries equal to the diagonal of the matrix $\mathbf{A}$, respectively.
$\text{blkdiag}([\mathbf{A}_1, \ldots, \mathbf{A}_q])$	Block diagonal matrix with diagonal entries equal to the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_q$.
$\mathbf{1}_Q, \mathbf{1}_p$	$(Q \times 1)$ and $(p \times 1)$ unitary vectors, respectively.
$\mathbf{V}_q = [v_{jh(q)}] = [\mathbf{v}_{h(q)} : h = 1, \ldots, q]$	$(p \times q)$ membership matrix defining a partition of the p observed variables into q groups, each one represented by a (disjoint principal) component. $v_{jh(q)} = 1$ if the jth observed variable belongs to the hth group of the qth partition $\mathcal{P}_{qH}$; $v_{jh(q)} = 0$ otherwise.
$\mathbf{B}_q = \text{diag}(\mathbf{b}_{(q)}) = \text{diag}([b_{1(q)}, \ldots, b_{p(q)}])$	Diagonal matrix of order p, whose diagonal elements $b_{j(q)}$, $j = 1, \ldots, p$, represent the unique loading of each observed variable on the corresponding disjoint principal component of $\mathcal{P}_{qH}$, i.e., according to $\mathbf{V}_q$.
$\mathbf{Y}_q = [y_{ih(q)}] = [\mathbf{y}_{h(q)} : h = 1, \ldots, q]$	$(n \times q)$ component score matrix, where $y_{ih(q)}$ is the value on the ith unit for the hth disjoint principal component of $\mathcal{P}_{qH}$.
$\mathbf{g}$	$(n \times 1)$ vector whose elements represent a measure of synthesis for the n units.
$\mathbf{E}_q = [e_{ij(q)}] = [\mathbf{e}_{j(q)} : j = 1, \ldots, p]$	$(n \times p)$ error matrix associated with the partition $\mathcal{P}_{qH}$.

5.3 Hierarchical Disjoint Principal Component Analysis

Let $\mathbf{X}$ be a $(n \times p)$ centered data matrix which consists of n units and p quantitative observed variables. Let us recall that PCA can be written in a model form, i.e., $\mathbf{X} = \mathbf{YA}' + \mathbf{E}$, where $\mathbf{Y} = \mathbf{XA}$ is the $(n \times Q)$ component score matrix, $\mathbf{A}$ is the $(p \times Q)$ component loading matrix s.t. $\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$, and $\mathbf{E}$ is the $(n \times p)$ error matrix. The LS estimates of the PCA model parameters are the matrix $\widehat{\mathbf{A}}$ of orthonormal eigenvectors associated with the Q largest eigenvalues of $\mathbf{X}'\mathbf{X}$, and the matrix $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{A}}$. It is worth noting that $\mathbf{YA}'$ is the best approximation of rank $Q \leq p$ to $\mathbf{X}$ in the least-squares sense (Magnus & Neudecker, 2007, Th. 7, p. 402). Based on the model formalization of PCA, Ferrara, Martella, and Vichi (2016) proposed the Disjoint Principal Component Analysis (DPCA) in which the component loading matrix $\mathbf{A}$ had a single non-null value per row, and thus was rewritten into the product of a weighting matrix $\mathbf{B}$ of order p and a $(p \times Q)$ membership matrix $\mathbf{V}$. DPCA for centered data was formally specified as $\mathbf{X} = \mathbf{YV}'\mathbf{B} + \mathbf{E}$, subject to constraints that (i) $\mathbf{V}$ was binary and row stochastic, (ii) $\mathbf{B}$ was diagonal and (iii) $\mathbf{A} = \mathbf{BV}$

was semi-orthogonal. It is worth highlighting that constraint (i) corresponds to a partition of the variable space, and, consequently, $\mathbf{Y} = \mathbf{XBV}$ contains Q disjoint principal components which can be correlated unlike standard principal components. The estimates of DPCA parameters are obtained in a LS framework by searching for the disjoint principal components of maximum variance (see Ferrara, Martella, & Vichi, 2016, 2019, for further details).

The *Hierarchical Disjoint Principal Component Analysis* (HierDPCA) can be formally specified by the following system of $Q - M + 1$ simultaneous equations,

$$\begin{cases} \mathbf{X} = \mathbf{Y}_M \mathbf{V}'_M \mathbf{B}_M + \mathbf{E}_M \\ \quad \dots \quad \dots \quad \dots \\ \mathbf{X} = \mathbf{Y}_{Q-1} \mathbf{V}'_{Q-1} \mathbf{B}_{Q-1} + \mathbf{E}_{Q-1} \\ \mathbf{X} = \mathbf{Y}_Q \mathbf{V}'_Q \mathbf{B}_Q + \mathbf{E}_Q \end{cases} \tag{5.1}$$

subject to constraints (for $q = M, \dots, Q$)

$$\mathbf{V}_q = [v_{jh(q)} \in \{0, 1\} : j = 1, \dots, p, h = 1, \dots, q] \quad \text{(binary);} \tag{5.2}$$

$$\mathbf{V}_q \mathbf{1}_q = \mathbf{1}_p \quad \text{(row stochastic);} \tag{5.3}$$

$$\mathbf{V}_{q-1} = [\mathbf{V}_q \setminus \{\mathbf{v}_{q-1(q)}, \mathbf{v}_{q(q)}\}, \mathbf{v}_{q-1(q-1)}] \quad \text{(nested partitions);} \tag{5.4}$$

$$\mathbf{v}_{q-1(q-1)} = \mathbf{v}_{q-1(q)} + \mathbf{v}_{q(q)}$$

$$\mathbf{B}_q = \text{diag}([b_{1(q)}, \dots, b_{p(q)}]) \quad \text{(diagonal);} \tag{5.5}$$

$$\mathbf{V}'_q \mathbf{B}_q \mathbf{B}_q \mathbf{V}_q = \mathbf{I}_q \quad \text{(semi-orthogonal).} \tag{5.6}$$

It has to be pointed out that constraint (5.4) does not hold for $q = M$.

Each equation of model (5.1) is expressed by means of a DPCA, where the component loading matrix $\mathbf{A}_q$ is rewritten into the product of $\mathbf{B}_q$ and $\mathbf{V}_q$, $q = M, \dots, Q$. The former is a diagonal matrix of order p (constraint 5.5), whose diagonal elements represent the unique loading of each observed variable on the corresponding disjoint principal component, whilst the latter identifies the variable partition $\mathcal{P}_{qH}$ in q groups. Without loss of generality, we define the last column of $\mathbf{V}_{q-1}$, i.e. $\mathbf{v}_{q-1(q-1)}$, by the sum of the last two columns of $\mathbf{V}_q$, i.e., $\mathbf{v}_{q-1(q)}$ and $\mathbf{v}_{q(q)}$, for $q = M + 1, \dots, Q$. On the whole, the equations of model (5.1) aim at reconstructing the data matrix by means of a decreasing number of disjoint principal components

$$\mathbf{Y}_q = \mathbf{XB}_q \mathbf{V}_q \qquad q = M, \dots, Q, \tag{5.7}$$

linked by constraint (5.4). HierDPCA therefore gives rise to a parsimonious tree of nested partitions, starting from Q up to M groups of observed variables.

Model (5.1) represents the reflective part of the hierarchy. The choice of M is crucial to determine the remaining hierarchical levels of the model. Indeed, $M \in \{1, \dots, Q\}$ is defined as the first bottom-up level at which no statistically significant correlation occurs among the disjoint principal components. By testing for this correlation, HierDPCA provides a model-based approach to select the measurement model from the Mth level upwards. If the correlation among the M disjoint principal components results to be not statistically significant, the remaining part of the hierarchy is modeled via a multiple linear regression, as we will depict

hereinafter. It is worth noticing that the relationship between the p observed variables and the Q disjoint principal components is always supposed reflective for $Q < p$ by obtaining a dimensionality reduction of the data. Under the same assumptions of the Probabilistic Disjoint Principal Component Analysis (PDPCA, Ferrara, Martella, & Vichi, 2019), M can be determined according to the Student's T statistic (Cramer, 1946, p. 400) by testing for the absence of correlation among any pair of disjoint principal components. Therefore, this parametric approach needs assumptions which can turn out to be unrealistic in some applications, and does not implement a multivariate test for the component correlation matrix. For these reasons, HierDPCA carries out a test for identity of covariance (or correlation) matrices proposed by Chen, Zhang, and Zhong (2010) in a non-parametric framework, i.e., without assuming a specific distribution for the data. The following hypothesis system has to be tested

$$\begin{cases} \mathrm{H}_0 : \mathbf{R}_{\mathbf{Y}_q} = \mathbf{I}_q \\ \mathrm{H}_1 : \mathbf{R}_{\mathbf{Y}_q} \neq \mathbf{I}_q \end{cases} \tag{5.8}$$

for each $q = Q, \ldots, 2$, where $\mathbf{R}_{\mathbf{Y}_q}$ is the correlation matrix of the q disjoint principal components. M is thus defined as the *first* bottom-up level $q \in \{2, \ldots, Q\}$ such that $\mathbb{P}\left((1/2)nT^{\mathrm{CZZ}_n} \geq z_\alpha\right) > \alpha$, where T^{CZZ} is the Chen-Zhang-Zhong statistic with an asymptotic normal distribution under the null hypothesis H_0 in (5.8), and z_α is the α-upper quantile of the standard normal distribution (for further details see Chen, Zhang, & Zhong, 2010).

If the null hypothesis in (5.8) is rejected for all hierarchical levels, M is settled on one and the equations in (5.1) model the whole hierarchy. It is worth highlighting that for $q = 1$ constraints (5.2)-(5.4) are trivial since $\mathbf{V}_1 = \mathbf{1}_p$, thus the definition of the general concept $\mathbf{g}$[1] is independent of the whole hierarchy. In order to determine the last equation of model (5.1) such that it depends on the previous hierarchical level, we can replace it with $\mathbf{Y}_2 = \mathbf{g}\mathbf{V}_1'\mathbf{B}_1 + \mathbf{E}_1$, where $\mathbf{g}$ is the column vector corresponding to the synthesis of the two disjoint principal components in $\mathbf{Y}_2$. Whereas if the null hypothesis in (5.8) is not rejected at the Mth level of the hierarchy, the correlation among the M disjoint principal components is not statistically significant and they are supposed to form a general concept $\mathbf{g}$, rather than reflect disjoint principal components of "higher-order". The model that formalizes the relationship between the M uncorrelated (or little correlated) disjoint principal components and the general concept $\mathbf{g}$ is the following multiple linear regression

$$\mathbf{g} = \mathbf{Y}_M\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.9}$$

where $\boldsymbol{\beta}$ is the $(M \times 1)$ regression coefficient vector and $\boldsymbol{\epsilon}$ is the $(n \times 1)$ error vector. It is worth noticing that in order to apply a multiple linear regression model, the response variable, that is represented by the general concept $\mathbf{g}$, needs to be quantified. When $M > 1$, $\mathbf{g}$ is assumed to be the first (standard) principal component of the p observed variables ($\mathbf{g}^{\mathrm{PC}}$). Thus, the relationship between the general concept and the M disjoint principal components is inspected by estimating the regression coefficients.

The hierarchy ensued from HierDPCA strictly depends on the initial choice of the number of disjoint principal components Q, that might be determined by several

[1] $\mathbf{g}$ stands for $\mathbf{Y}_1$, i.e., the general concept identified at the last level of the hierarchy.

methods, such as the Kaiser's one (Kaiser, 1960). Moreover, Q can be chosen as the minimum number of disjoint groups of observed variables such that each one is represented by a unidimensional component. The unidimensionality is evaluated according to the second largest eigenvalue of the covariance sub-matrix related to each variable group: if it is lower than the mean of the eigenvalues, the disjoint principal component associated with the variable group is unidimensional. The optimal number of disjoint principal components for the first bottom-up level of the hierarchy is therefore chosen by starting from 1 up to the value which corresponds to the first Q with unidimensional disjoint principal components. If Q is set greater than the optimal one, the identified disjoint principal components result highly correlated and there is no real interest that this duplication of components is preserved. In addition, the increase of the explained variance is negligible with respect to the ones obtained up to the optimal Q.

5.3.1 Least-squares estimation of HierDPCA

The least-squares estimations of model (5.1) are obtained by minimizing the following loss function representing a quadratic mixed continuous and combinatorial problem

$$F(\mathbf{B}_M, \ldots, \mathbf{B}_Q, \mathbf{V}_M, \ldots, \mathbf{V}_Q, \mathbf{Y}_M, \ldots, \mathbf{Y}_Q) = \sum_{q=M}^{Q} \|\mathbf{X} - \mathbf{Y}_q \mathbf{V}_q' \mathbf{B}_q\|^2 \qquad (5.10)$$

subject to constraints (5.2)-(5.6), where $\|\cdot\|$ denotes the Euclidean norm. If $M = 1$, the sum in (5.10) ends with $q = 2$ and $\|\mathbf{Y}_2 - \mathbf{g}\mathbf{V}_1'\mathbf{B}_1\|^2$ is added to it. If $M > 1$, $\|\mathbf{g} - \mathbf{Y}_M\boldsymbol{\beta}\|^2$, where $\mathbf{g} = \mathbf{g}^{\mathrm{PC}}$ (see Section 5.3), is added to (5.10). From now on, to simplify we will refer to $F(\mathbf{B}_M, \ldots, \mathbf{B}_Q, \mathbf{V}_M, \ldots, \mathbf{V}_Q, \mathbf{Y}_M, \ldots, \mathbf{Y}_Q)$ as F.

Before illustrating the least-squares estimators, some properties of HierDPCA are provided.

Property 0. According to Eq. (5.7), each equation of model (5.1) can be re-written as $\mathbf{X} = \mathbf{X}\mathbf{B}_q\mathbf{V}_q\mathbf{V}_q'\mathbf{B}_q + \mathbf{E}_q$, $q = M, \ldots, Q$, where, given constraint (5.6), the first term of the r.h.s. is the orthogonal projection of the rows of $\mathbf{X}$ onto the q-dimensional subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{B}_q\mathbf{V}_q$.

Property 1. The following decomposition of the total deviance, multiplied by a constant, holds

$$(Q - M + 1)\,\|\mathbf{X}\|^2 = \sum_{q=M}^{Q} \|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + \sum_{q=M}^{Q} \|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2. \qquad (5.11)$$

The proof of (5.11) is provided in Appendix C.

If $M = 1$ and the last equation of model (5.1) is replaced such that it depends on the whole hierarchy, the decomposition of the total deviance in Property 1 results into

$$(Q-1)\,\|\mathbf{X}\|^2 = \sum_{q=3}^{Q}\|\mathbf{X}-\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + \sum_{q=3}^{Q}\|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + \|\mathbf{X}-\mathbf{g}\mathbf{V}_1'\mathbf{B}_1\mathbf{V}_2'\mathbf{B}_2\|^2 + \|\mathbf{g}\mathbf{V}_1'\mathbf{B}_1\mathbf{V}_2'\mathbf{B}_2\|^2.$$

This can be easily proved by extending the proof of (5.11) and substituting the last equation of model (5.1).

Remark 5.1. Since the l.h.s. of (5.11) is known and fixed given $\mathbf{X}$ and Q, minimizing the first term of the r.h.s., i.e., the total *residual* deviance of model (5.1), corresponds to maximize the second term of the r.h.s., i.e., the total deviance *reconstructed* by model (5.1).

Property 2. HierDPCA defines a parsimonious hierarchy formed by $2Q - M$ components, corresponding to disjoint or nested groups of observed variables, that maximize the total explained variance.

Property 2 therefore shows that maximizing the total reconstructed deviance corresponds to maximizing the explained variance of the disjoint principal components for each bottom-up level $q = Q, \ldots, M$, that matches in turn the selection of the first principal component for each non-overlapping variable group of $\mathcal{P}_{qH}$:

$$\sum_{q=M}^{Q}\|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 = \sum_{q=M}^{Q}\mathrm{tr}(\mathbf{B}_q\mathbf{V}_q\mathbf{Y}_q'\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q) = \sum_{q=M}^{Q}\mathrm{tr}(\mathbf{Y}_q'\mathbf{Y}_q) = n\sum_{q=M}^{Q}\mathrm{tr}(\Sigma_{\mathbf{Y}_q}),$$

where $\Sigma_{\mathbf{Y}_q}$ is the covariance matrix of the q disjoint principal components. Indeed, $\mathrm{tr}(\Sigma_{\mathbf{Y}_q}) = \mathrm{tr}(\frac{1}{n}\mathbf{V}_q'\mathbf{B}_q\mathbf{X}'\mathbf{X}\mathbf{B}_q\mathbf{V}_q) = \mathrm{tr}(\mathbf{V}_q'\mathbf{B}_q\Sigma_{\mathbf{X}}\mathbf{B}_q\mathbf{V}_q) = \sum_{h=1}^{q}\mathbf{v}_{h(q)}'\mathbf{B}_q\,\Sigma_{\mathbf{X}}\mathbf{B}_q\mathbf{v}_{h(q)}$, where $\mathbf{v}_{h(q)}'\mathbf{B}_q\Sigma_{\mathbf{X}}\mathbf{B}_q\mathbf{v}_{h(q)}$ represents the variance of the first principal component associated with the hth variable group, $h = 1, \ldots, q$, of $\mathcal{P}_{qH}$.

Property 3. HierDPCA minimizes the decrease of the explained deviance that occurs after having merged two groups of $\mathcal{P}_{qH}$ to obtain the partition $\mathcal{P}_{q-1H}$, which is formally defined as

$$I_d(\mathbf{V}_q, \mathbf{V}_{q-1}) = \|\mathbf{X}-\mathbf{Y}_{q-1}\mathbf{V}_{q-1}'\mathbf{B}_{q-1}\|^2 - \|\mathbf{X}-\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2$$
$$= \|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 - \|\mathbf{Y}_{q-1}\mathbf{V}_{q-1}'\mathbf{B}_{q-1}\|^2 \overset{\text{Constr. (5.6)}}{=} \|\mathbf{Y}_q\|^2 - \|\mathbf{Y}_{q-1}\|^2 \qquad (5.12)$$

and is always nonnegative.

$I_d(\mathbf{V}_q, \mathbf{V}_{q-1})$ actually depends only on $\mathbf{v}_{q-1(q)}$, $\mathbf{v}_{q(q)}$ and $\mathbf{v}_{q-1(q-1)}$. Indeed, inasmuch as the membership matrix $\mathbf{V}_{q-1}$ has the first $q-2$ columns equal to $\mathbf{V}_q$, then $I_d(\mathbf{V}_q, \mathbf{V}_{q-1}) = \|\mathbf{y}_{q-1(q)}\|^2 + \|\mathbf{y}_{q(q)}\|^2 - \|\mathbf{y}_{q-1(q-1)}\|^2$. Moreover, it holds (Vigneau & Qannari, 2003) that

$$\|\mathbf{y}_{q-1(q-1)}\|^2 \leq \|\mathbf{y}_{q-1(q)}\|^2 + \|\mathbf{y}_{q(q)}\|^2,$$

where the equality occurs when the covariance between $\mathbf{y}_{q-1(q)}$ and $\mathbf{y}_{q(q)}$ reaches its maximum. $I_d(\mathbf{V}_q, \mathbf{V}_{q-1})$ is thus nonnegative and the variance explained by the disjoint principal components declines at each level of the hierarchy with the component number reduction.

Remark 5.2. According to (5.12), when $M = 1$ the loss function (5.10) can be rewritten as follows

$$F = Q\|\mathbf{X} - \mathbf{Y}_Q\mathbf{V}_Q'\mathbf{B}_Q\|^2 + \sum_{q=2}^{Q}(q-1)\, I_d(\mathbf{V}_q, \mathbf{V}_{q-1}). \tag{5.13}$$

The proof of (5.13) is provided in Appendix C.

In order to present the least-squares estimators of the HierDPCA parameters, let us simplify the structure of the weighting and membership matrices. Given a partition $\mathcal{P}_{qH}$, $q = M, \ldots, Q$, we consider an appropriate row permutation of the membership matrix $\mathbf{V}_q$ such that the observed variables associated with the same disjoint principal component are contiguous. It follows that the weighting matrix $\mathbf{B}_q = \text{blkdiag}([\mathbf{B}_{1(q)}, \ldots, \mathbf{B}_{q(q)}])$ is a block diagonal matrix, where each $\mathbf{B}_{h(q)}$, $h = 1, \ldots, q$, is the diagonal sub-matrix of order $J_{h(q)}$ whose diagonal elements are the loadings of the contiguous observed variables associated with the hth disjoint principal component. The columns of the data matrix $\mathbf{X}$ can be ordered as well.

Given Q, the HierDPCA parameter estimates are the following.

(0) **Estimation of $\mathbf{V}_Q$, $\mathbf{B}_Q$ and $\mathbf{Y}_Q$:** the estimates of $\mathbf{V}_Q$, $\mathbf{B}_Q$ and $\mathbf{Y}_Q$ are obtained according to the DPCA estimation (Ferrara, Martella, & Vichi, 2016). See also Vichi and Saporta (2009) for the details on the DPCA estimates, where $K = n$, i.e., the clustering structure for the observations is not taken into account.

The hierarchy is estimated as follows.

(a) **Test on the component correlation matrix**: the Chen, Zhang, and Zhong (2010) nonparametric test is applied to the correlation matrix of $\widehat{\mathbf{Y}}_q$. If the null hypothesis in (5.8) is rejected, the number of disjoint principal components is reduced by one and the estimates of $\mathbf{V}_{q-1}$, $\mathbf{B}_{q-1}$ and $\mathbf{Y}_{q-1}$ are computed as in (b), (c) and (d), respectively; otherwise, $M = q$ and the estimates of the regression coefficients, which link the general concept to the M disjoint principal components, are computed as in (e).

(b) **Estimation of $\mathbf{V}_{q-1}$**: the membership matrix $\mathbf{V}_{q-1}$ is estimated by merging two columns of $\widehat{\mathbf{V}}_q$, say $s, h \in \{1, \ldots, q\}$, such that $\|\mathbf{X} - \widehat{\mathbf{Y}}_{q-1}\widehat{\mathbf{V}}_{q-1}'\widehat{\mathbf{B}}_{q-1}\|^2$, or equivalently $I_d(\widehat{\mathbf{V}}_q, \widehat{\mathbf{V}}_{q-1})$, is minimum.

Formally, recalling constraint (5.4), $\widehat{\mathbf{V}}_{q-1} = [\{(\hat{\mathbf{v}}_{1(q)}, \ldots, \hat{\mathbf{v}}_{q(q)}) \setminus \hat{\mathbf{v}}_{s(q)}, \hat{\mathbf{v}}_{h(q)}\}, \hat{\mathbf{v}}_{q-1(q-1)}]$, where $\hat{\mathbf{v}}_{q-1(q-1)} = \hat{\mathbf{v}}_{s(q)} + \hat{\mathbf{v}}_{h(q)}$. It is worth noting that, only for simplicity reasons, constraint (5.4) is written with respect to $\mathbf{v}_{q-1(q)}$ and $\mathbf{v}_{q(q)}$ by assuming a reordering of the columns of $\mathbf{V}_q$ such that the ones to be merged are the last two.

(c) **Estimation of $\mathbf{B}_{q-1}$**: given $\widehat{\mathbf{V}}_{q-1}$, let us consider the spectral decomposition of the covariance matrix associated with the $(s \cup h)$th subset of $\mathcal{P}_{q-1H}$, $\Sigma_{\mathbf{X}_{s\cup h(q-1)}} = \mathbf{D}_{s\cup h(q-1)}\mathbf{\Lambda}_{s\cup h(q-1)}\mathbf{D}'_{s\cup h(q-1)}$, where $\mathbf{D}_{s\cup h(q-1)}$ and $\mathbf{\Lambda}_{s\cup h(q-1)}$ are the orthonormal matrix of the eigenvectors and the diagonal matrix of the eigenvalues of $\Sigma_{\mathbf{X}_{s\cup h(q-1)}}$, respectively. $\widehat{\mathbf{B}}_{q-1(q-1)} = \text{diag}(\mathbf{D}^{(1)}_{s\cup h(q-1)})$, where $\mathbf{D}^{(1)}_{s\cup h(q-1)}$ is therefore the eigenvector corresponding to the largest eigenvalue $\lambda^{(1)}_{s\cup h(q-1)}$ of the matrix $\Sigma_{\mathbf{X}_{s\cup h(q-1)}}$. Inasmuch constraint (5.4) holds, the weighting matrix estimator is computed by substituting the loadings of the observed variables which belong to the sth and hth group with the one obtained by merging them, i.e., $\widehat{\mathbf{B}}_{q-1} = \text{blkdiag}([\{[\widehat{\mathbf{B}}_{1(q)}, \ldots, \widehat{\mathbf{B}}_{q(q)}] \setminus \widehat{\mathbf{B}}_{s(q)}, \widehat{\mathbf{B}}_{h(q)}\}, \widehat{\mathbf{B}}_{q-1(q-1)}])$.

(d) **Estimation of $\mathbf{Y}_{q-1}$**: given $\widehat{\mathbf{V}}_{q-1}$ and $\widehat{\mathbf{B}}_{q-1}$, the component score matrix is computed as $\widehat{\mathbf{Y}}_{q-1} = \mathbf{X}\widehat{\mathbf{B}}_{q-1}\widehat{\mathbf{V}}_{q-1}$, which corresponds to the Bartlett's weighted least-squares scores (Ferrara, Martella, & Vichi, 2019).

(e) **Estimation of $\boldsymbol{\beta}$ and $\mathbf{g}$:** given $\widehat{\mathbf{Y}}_M$, the regression coefficient vector is estimated as in a multiple linear regression analysis, i.e., $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{Y}}_M$, and $\widehat{\mathbf{g}} = \widehat{\mathbf{Y}}_M\widehat{\boldsymbol{\beta}}$.

If $M = 1$ and the last equation of model (5.1) is replaced such that it depends on the whole hierarchy, the estimates of $\mathbf{B}_1$ and $\mathbf{Y}_1$ are obtained by replacing $\mathbf{X}$ with $\widehat{\mathbf{Y}}_2$.

Property 4. HierDPCA identifies the sparsest component loading matrices which correspond to disjoint principal components of maximum variance and cannot be further simplified. Additionally, the solution is unique.

In fact, since model (5.1) identifies a partition of the variable space for each hierarchical level, the component loading matrix turns out to be the sparsest one which corresponds to disjoint principal components of maximum variance and no orthogonal transformation can improve its simplification because, given an orthogonal matrix $\mathbf{Q}$ of order q and the component loading matrix $\mathbf{A}_q = \mathbf{B}_q\mathbf{V}_q$, the transformation $\mathbf{A}_q\mathbf{Q} = \mathbf{B}_q\mathbf{V}_q\mathbf{Q}$ $(q = Q, \ldots, M)$ does not return a matrix with only one non-null loading per row. Property 4 is formalized in the following theorem, whose proof is similar to that provided by Vichi (2017, pp. 573-574).

Theorem 5.1. *[Uniqueness] For each level q $(q = M, \ldots, Q)$ of HierDPCA, the component loading matrix is unique and no Orthomax rotation $\gamma\mathbf{Q}$ satisfies the parameterization constraint $\mathbf{A}_q = \mathbf{B}_q\mathbf{V}_q$, i.e. $\mathbf{A}_q\gamma\mathbf{Q} = \mathbf{B}_q\mathbf{V}_q\gamma\mathbf{Q}$, other than the identity matrix $\mathbf{I}_q$, for all $q = M, \ldots, Q$.*

The solution of the minimization of (5.10) is thus unique, i.e., the whole hierarchy obtained by the least-squares estimation of the model parameters is unique both in the reflective and reflective/formative cases. HierDPCA is therefore identifiable thanks to the uniqueness of the weighting matrices $\mathbf{B}_q$ and the membership matrices $\mathbf{V}_q$ $(q = M, \ldots, Q)$, as well as $\boldsymbol{\beta}$ in (5.9). Nevertheless, since the tight relation with PCA, model (5.1) is *not* scale-invariant. For this reason, the analysis has to be performed by standardizing the observed variables when data have different measurement scales.

5.3.2 Coordinate descent algorithm for HierDPCA

Given Q, the least-squares estimates of the HierDPCA parameters are computed by using a coordinate descent algorithm implemented in MATLAB, which is described in Algorithm 2. The estimates recalled in Algorithm 2 have been presented in Section 5.3.1.

The HierDPCA algorithm converges to a solution which is at least a local minimum. To increase the chance to find a global optimum, Algorithm 2 should be run several times starting from different random initial partitions $\mathbf{V}_Q^{(0)}$.

Albeit the HierDPCA algorithm is NP-hard (Krivánek & Morávek, 1986), it is computationally efficient. The parsimony and the sparsity properties of model (5.1) guarantee a low complexity both in terms of the execution time and the storage space needed for computation. To improve the computational efficiency of the HierDPCA algorithm, **Step 1**, which represents a step of the DPCA algorithm, can be led to convergence and run several times before building the whole hierarchy. Recalling that the latter is in turn a coordinate descent algorithm and according to our experience based on the simulation study presented in Section 5.4, **Step 1** could be run at least 30 times so that the starting point of the hierarchy completion is the optimal solution of DPCA.

Furthermore, the HierDPCA algorithm allows some options. The first one is the possibility to constrain some (or all) observed variables to load, necessarily, on a disjoint principal component at the Qth level of the hierarchy if some (or all) relationships between the observed variables and the (disjoint principal) components are known a priori. Moreover, the researcher may impose the non-negativity of the loadings s.t. positive and negative relationships between observed variables and ("higher-order") disjoint principal components do not balance out in the estimation of the latter.

5.4 Simulation study

The performances of HierDPCA were evaluated through a large-scale simulation study with different scenarios. Before analyzing its results in details, let us illustrate the data generation process.

Each simulated random sample of $n > p$ multivariate observations $\mathbf{x}_i$ $(i = 1, \ldots, n)$ is generated according to $\mathbf{X} = \mathbf{X}_t + \mathbf{E}$, where $\mathbf{X}_t \sim MVN_p(\mathbf{0}, \mathbf{R_u})$ and $\mathbf{E} \sim \sigma_{\mathbf{E}} \cdot MVN_p(\mathbf{0}, \mathbf{I})$, with $\sigma_{\mathbf{E}}$ which represents the error level. The correlation structure $\mathbf{R}_\text{u}$ is modeled as presented by Cavicchia, Vichi, and Zaccaria (2020b, see Chapter 2)

$$\mathbf{R}_\text{u} = \mathbf{V}_Q \left(\mathbf{R}_\text{B} - \mathbf{I}_Q\right) \mathbf{V}_Q' + \mathbf{V}_Q \mathbf{R}_\text{W} \mathbf{V}_Q' - \text{diag}(\mathbf{V}_Q \mathbf{R}_\text{W} \mathbf{V}_Q') + \mathbf{I}_p. \tag{5.14}$$

In detail, $\mathbf{V}_Q$ is a $(p \times Q)$ membership matrix which partitions p variables into Q groups, $\mathbf{R}_\text{W}$ a $(Q \times Q)$ diagonal matrix describing correlations within the groups and $\mathbf{R}_\text{B}$ a $(Q \times Q)$ ultrametric correlation matrix representing the hierarchical relationships among the Q groups. The parameters of $\mathbf{R}_\text{u}$ were set as follows. The membership matrix $\mathbf{V}_Q$ was randomly generated such that constraints (5.2) and (5.3) held and no empty group occurred. Moreover, the cardinality of the variable groups

Algorithm 2: HierDPCA

Input: $\mathbf{X}$, Q

Fixed values $\epsilon \leftarrow$ *small nonnegative converge tolerance value;*
maxiter $\leftarrow$ *maximum number of iterations;*
Initialization $t \leftarrow 0$*;*
$\mathbf{V}_Q^{(0)} \leftarrow$ *random variable partition s.t. (5.2) and (5.3) are satisfied and the* Q *variable groups are non-empty;*
$\widehat{\mathbf{B}}_Q^{(0)} \leftarrow$ *see (0) in Section 5.3.1;*
$\widehat{\mathbf{Y}}_Q^{(0)} \leftarrow$ *see (0) in Section 5.3.1;*
Hierarchy: *build the hierarchy by testing the correlation among the disjoint principal components at each hierarchical level. If the null hypothesis is rejected, reduce the number of disjoint principal components by one, compute (b), (c), (d) and (a); otherwise the hierarchy ends with (e);*
$F^{(0)} \leftarrow$ *objective function in (5.10) plus* $\|\widehat{\mathbf{Y}}_2^{(0)} - \widehat{\mathbf{g}}^{(0)}\widehat{\mathbf{V}}_1^{(0)\prime}\widehat{\mathbf{B}}_1^{(0)}\|^2$ *if* $M = 1$*, or* $\|\mathbf{g}^{PC(0)} - \widehat{\mathbf{Y}}_M^{(0)}\widehat{\boldsymbol{\beta}}^{(0)}\|^2$ *if* $M > 1$*;*
$F_{diff}^{(0)} \leftarrow F^{(0)}$*;*
while $F_{diff}^{(t)} > \epsilon$ *and* $t \leq maxiter$ **do**
  $t \leftarrow t + 1$;
  Step 1. Given $\widehat{\mathbf{V}}_Q^{(t-1)}$, compute $\widehat{\mathbf{V}}_Q^{(t)}$, $\widehat{\mathbf{B}}_Q^{(t)}$ and $\widehat{\mathbf{Y}}_Q^{(t)}$ s.t. $\|\mathbf{X} - \widehat{\mathbf{Y}}_Q^{(t)}\widehat{\mathbf{V}}_Q^{(t)\prime}\widehat{\mathbf{B}}_Q^{(t)}\|^2 \leq \|\mathbf{X} - \widehat{\mathbf{Y}}_Q^{(t-1)}\widehat{\mathbf{V}}_Q^{(t-1)\prime}\widehat{\mathbf{B}}_Q^{(t-1)}\|^2$, i.e., starting from $\widehat{\mathbf{V}}_Q^{(t-1)}$ and searching for a new variable partition s.t. the DPCA objective function is improved;
  for *q = Q-1, . . . , 1* **do**
    Step 2. Test for the correlation matrix $\mathbf{R}_{\widehat{\mathbf{Y}}_{q+1}^{(t)}}$ as in (a);
    if *the null hypothesis in (5.8) is rejected* **then**
      Step 3a. Compute $\widehat{\mathbf{V}}_q^{(t)}$, $\widehat{\mathbf{B}}_q^{(t)}$ and $\widehat{\mathbf{Y}}_q^{(t)}$ according to (b), (c) and (d), respectively;
      if *q == 1* **then**
        $M \leftarrow 1$;
      end
    end
    else
      Step 3b. $M \leftarrow q$;
      Given $\widehat{\mathbf{Y}}_M^{(t)}$, compute $\widehat{\boldsymbol{\beta}}^{(t)}$ and $\widehat{\mathbf{g}}^{(t)}$ as in (e);
      break
    end
  end
  $F^{(t)} \leftarrow$ objective function in (5.10) plus $\|\widehat{\mathbf{Y}}_2^{(t)} - \widehat{\mathbf{g}}^{(t)}\widehat{\mathbf{V}}_1^{(t)\prime}\widehat{\mathbf{B}}_1^{(t)}\|^2$ if $M = 1$, or $\|\mathbf{g}^{\mathrm{PC}(t)} - \widehat{\mathbf{Y}}_M^{(t)}\widehat{\boldsymbol{\beta}}^{(t)}\|^2$ if $M > 1$;
  $F_{\mathrm{diff}}^{(t)} \leftarrow F^{(t-1)} - F^{(t)}$.
end

Table 5.1. Scenarios of the simulation study.

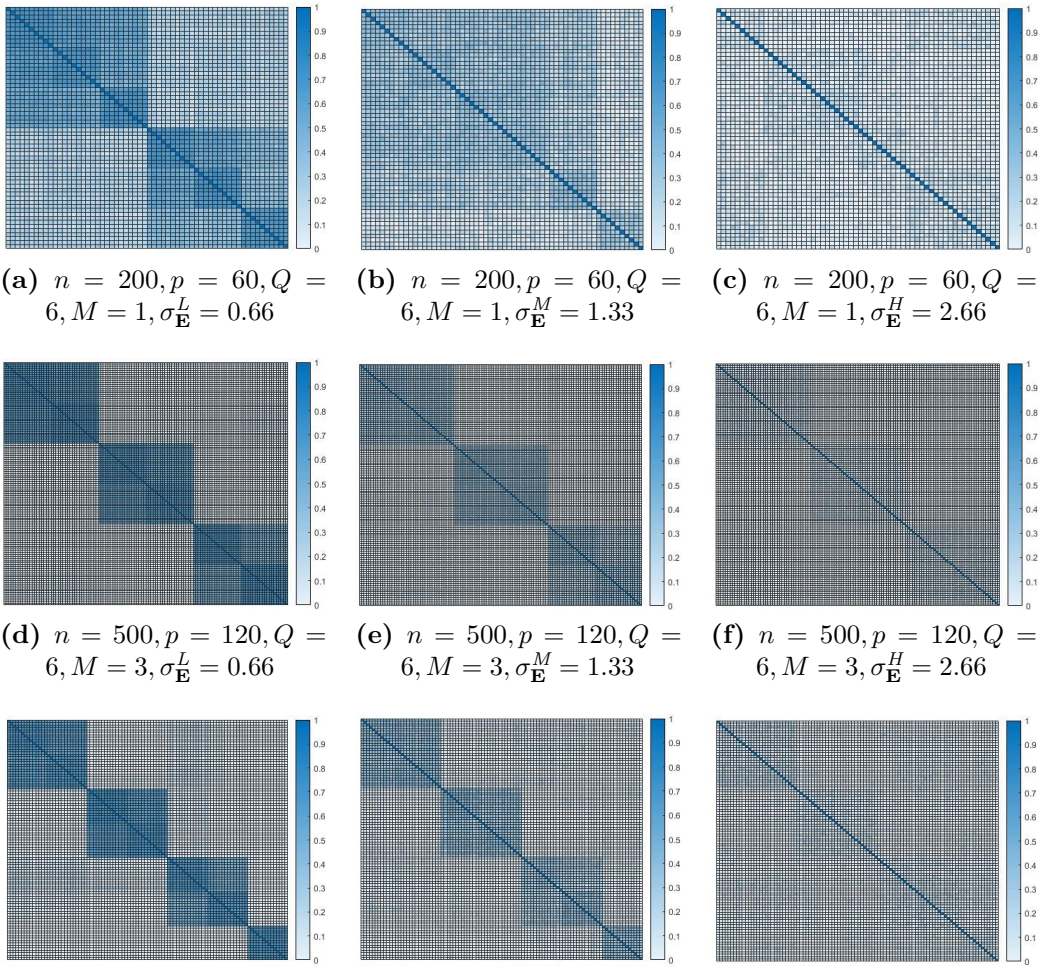
	Scenario 1	*Scenario* 2	*Scenario* 3
N. units n	$200, 500$	$200, 500$	$200, 500$
N. variables p	$60, 120$	$60, 120$	98
N. groups Q	6	6	7
M^{th}	1	3	4
$\mathbf{R}_{\text{B}}$ off-diagonal values	{0.9 0.85 0.78 0.65 0.37}	{0.9 0.85 0.7 0 0}	{0.9 0.85 0.7 0 0}
Error levels $\sigma_{\mathbf{E}}$	$0.66, 1.33, 2.66$	$0.66, 1.33, 2.66$	$0.66, 1.33, 2.66$

was fixed to p/Q in order to evaluate their bottom-up aggregations net of their size. For simplicity, $\mathbf{R}_{\text{W}}$ was assumed to be an identity matrix, i.e., the highest linear relationship within groups existed. The $Q-1$ different off-diagonal elements of $\mathbf{R}_{\text{B}}$ were set such that the relationships among the Q groups were hierarchically ordered. These values defined the membership matrices $\mathbf{V}_q$, $q = M^{\text{th}}, \ldots, Q-1$, where M^{th} is a parameter as well. The simulation study was implemented by generating 200 random samples according to the aforementioned structure of the data matrix for each scenario reported in Table 5.1. In each one, both a small and large scale data generation were provided in terms of units and variables (except for *Scenario 3*) and both reflective (*Scenario 1*) and reflective/formative hierarchies (*Scenario 2* and *Scenario 3*) were considered in the simulation study. Overall, 6000 random samples were generated.

It can be highlighted that the data matrix $\mathbf{X}$ was not directly generated from HierDPCA by avoiding a privileged position for the proposal, especially in comparisons with other methodologies, and its correlation matrix was not exactly equal to $\mathbf{R}_{\text{u}}$ since an error perturbed this structure. In Figure 5.2, a representation of the error level effect is provided. Some correlation matrices of the generated samples are represented for each scenario. Increasing $\sigma_{\mathbf{E}}$ corresponds to mask the ultrametric correlation structure of the generated data by making the distinct groups of variables less clearly distinguishable, as well as the difference of their hierarchical relationships.

HierDPCA was also compared to other two existing methods: PCA + Oblique Rotation Method (oblimin, quartimin, geomin) and the hierarchical clustering algorithms (single, complete, average linkage and Ward's method). The former is a sequential procedure similar to HOFMs, where PCA was used instead of FA for comparability reasons. For each hierarchical level $q = Q, \ldots, M^{\text{th}}$, the variable partition of PCA + ORM was obtained by assigning each variable (or component) to the higher-order component it loads more on in absolute term after an oblique rotation, whereas the one resulting from the hierarchical clustering algorithms was derived by cutting the hierarchy at the qth level. It is worth noticing that PCA + ORM was firstly applied on the variable correlation matrix to obtain Q components, and then on the component correlation matrix by reducing the number of components by one. Both PCA + ORM and hierarchical clustering algorithms were implemented by fixing Q and $M = M^{\text{th}}$.

The models' performances were evaluated in terms of similarity between the true (generated) and the estimated (through HierDPCA, PCA + ORM, hierarchical clustering algorithms) partitions of variables for each hierarchical level via the



(a) $n = 200, p = 60, Q = 6, M = 1, \sigma^L_{\mathbf{E}} = 0.66$ **(b)** $n = 200, p = 60, Q = 6, M = 1, \sigma^M_{\mathbf{E}} = 1.33$ **(c)** $n = 200, p = 60, Q = 6, M = 1, \sigma^H_{\mathbf{E}} = 2.66$

(d) $n = 500, p = 120, Q = 6, M = 3, \sigma^L_{\mathbf{E}} = 0.66$ **(e)** $n = 500, p = 120, Q = 6, M = 3, \sigma^M_{\mathbf{E}} = 1.33$ **(f)** $n = 500, p = 120, Q = 6, M = 3, \sigma^H_{\mathbf{E}} = 2.66$

(g) $n = 200, p = 98, Q = 7, M = 4, \sigma^L_{\mathbf{E}} = 0.66$ **(h)** $n = 200, p = 98, Q = 7, M = 4, \sigma^M_{\mathbf{E}} = 1.33$ **(i)** $n = 200, p = 98, Q = 7, M = 4, \sigma^H_{\mathbf{E}} = 2.66$

Figure 5.2. Heat maps of the correlation matrices corresponding to some generated data sets in different scenarios.

Adjusted Rand Index (ARI, Hubert & Arabie, 1985). It ranges between $-\infty$ and 1, assuming the latter in case of perfect agreement between the true and the estimated membership matrices. Moreover, since the positiveness of the off-diagonal values of $\mathbf{R}_{\mathrm{B}}$, the Cronbach's α (Cronbach, 1951) was computed on the estimated variable groups in order to evaluate their internal consistency for each level of the hierarchy. Specifically, for all samples, the Cronbach's α of each group of $\mathcal{P}_{qH}$, $q = M^{\mathrm{th}}, \ldots, Q$, was calculated and its mean was computed for $q = M^{\mathrm{th}}, \ldots, Q$, i.e. $\bar{\alpha}_q = \frac{1}{200} \sum_{s=1}^{200} \frac{1}{q} \sum_{h=1}^{q} \alpha_{hs(q)}$.

Table 5.2 and Table 5.3 report the performances of HierDPCA in terms of the ARI and $\bar{\alpha}$ for *Scenario 1* and *Scenario 2* and *3*, respectively, by imposing $M = M^{\mathrm{th}}$. The results for the last aggregation of *Scenario 1* were not reported inasmuch as the membership matrix turned out to be a unitary vector whatever the variable partition in two groups was. In Table 5.3, the only part of the hierarchy which

Table 5.2. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of HierDPCA for *Scenario 1*.

n	p	q	$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$	
			ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	2	1.000 (100.0)	0.977	0.997 (99.5)	0.900	0.802 (8.0)	0.776
		3	1.000 (100.0)	0.970	0.978 (93.0)	0.882	0.623 (0.5)	0.709
		4	1.000 (100.0)	0.966	0.934 (64.0)	0.871	0.474 (0.0)	0.671
		5	1.000 (100.0)	0.962	0.883 (25.0)	0.861	0.347 (0.0)	0.641
		6	1.000 (100.0)	0.958	0.855 (7.0)	0.848	0.287 (0.0)	0.618
	120	2	1.000 (100.0)	0.988	1.000 (100.0)	0.963	0.926 (15.5)	0.862
		3	1.000 (100.0)	0.985	1.000 (100.0)	0.944	0.646 (1.0)	0.816
		4	1.000 (100.0)	0.983	0.973 (84.0)	0.930	0.474 (0.0)	0.790
		5	1.000 (100.0)	0.980	0.949 (33.5)	0.925	0.383 (0.0)	0.771
		6	1.000 (100.0)	0.979	0.932 (5.0)	0.917	0.378 (0.0)	0.754
500	60	2	1.000 (100.0)	0.977	1.000 (100.0)	0.929	0.991 (88.0)	0.775
		3	1.000 (100.0)	0.970	1.000 (100.0)	0.896	0.926 (39.5)	0.697
		4	1.000 (100.0)	0.966	0.982 (95.5)	0.872	0.769 (2.5)	0.655
		5	1.000 (100.0)	0.962	0.990 (92.5)	0.863	0.618 (0.0)	0.618
		6	1.000 (100.0)	0.958	0.981 (66.0)	0.849	0.530 (0.0)	0.592
	120	2	1.000 (100.0)	0.988	1.000 (100.0)	0.947	0.999 (97.5)	0.867
		3	1.000 (100.0)	0.986	1.000 (100.0)	0.945	0.970 (37.0)	0.835
		4	1.000 (100.0)	0.983	1.000 (100.0)	0.936	0.928 (8.5)	0.788
		5	1.000 (100.0)	0.980	0.999 (98.5)	0.926	0.824 (0.0)	0.758
		6	1.000 (100.0)	0.979	0.996 (84.5)	0.919	0.651 (0.0)	0.738

was considered is that one pertaining the system in (5.1). It can be observed that the hierarchy is always covered by the model for a low level of error (σ^L_E), both in Table 5.2 and Table 5.3. For medium (σ^M_E) and high error (σ^H_E), the mean of the ARI increases from the Qth level of the hierarchy upwards as expected, since the possibility of misclassification is reduced as the number of groups decreases. $\bar{\alpha}$ has an increasing - from Q upwards - behavior in turn; it is greater than 0.95, 0.80 and, almost always, 0.60 in the whole hierarchy for the low, medium and high levels of error, respectively. It has to be noticed that the Cronbach's α is affected by the number of variables in a group, which is hypothesized to be the same at the Qth level of the hierarchy, but changes when the number of groups decreases. The strong internal consistency of the variable groups stresses that HierDPCA pinpoints groups of highly correlated variables, even if the ARI differs from 1. The theorized reflective structure of the three scenarios therefore turns out to be correctly estimated.

The simulations reported in Table 5.2 and Table 5.3 were computed by fixing M to the theoretical one, i.e., M^{th}. To fully evaluate the performances of HierDPCA, all scenarios reported in Table 5.1 were assessed according to the correct identification of M on the samples previously generated. As shown in Table 5.4, M was almost always correctly estimated in *Scenario 1*. For the two samples in which M differed from one, it was estimated to be two. In *Scenario 2* and *Scenario 3*, the magnitude

Table 5.3. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of HierDPCA for *Scenario 2* and *Scenario 3*.

n	p	q	$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$	
			ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.905	0.976 (63.5)	0.715
		4	1.000 (100.0)	0.966	0.998 (98.5)	0.879	0.826 (3.0)	0.662
		5	1.000 (100.0)	0.961	0.931 (47.0)	0.860	0.649 (0.0)	0.628
		6	1.000 (100.0)	0.958	0.882 (9.0)	0.847	0.489 (0.0)	0.605
	120	3	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.993 (75.0)	0.832
		4	1.000 (100.0)	0.983	1.000 (100.0)	0.936	0.915 (1.0)	0.788
		5	1.000 (100.0)	0.980	0.959 (49.5)	0.925	0.715 (0.0)	0.762
		6	1.000 (100.0)	0.978	0.942 (7.5)	0.918	0.548 (0.0)	0.745
500	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.906	1.000 (100.0)	0.715
		4	1.000 (100.0)	0.966	1.000 (100.0)	0.880	0.952 (37.5)	0.654
		5	1.000 (100.0)	0.962	0.986 (93.5)	0.862	0.767 (0.0)	0.615
		6	1.000 (100.0)	0.958	0.986 (73.0)	0.849	0.607 (0.0)	0.590
	120	3	1.000 (100.0)	0.986	1.000 (100.0)	0.951	1.000 (100.0)	0.833
		4	1.000 (100.0)	0.983	1.000 (100.0)	0.936	1.000 (100.0)	0.787
		5	1.000 (100.0)	0.980	0.996 (97.5)	0.925	0.841 (0.0)	0.757
		6	1.000 (100.0)	0.979	0.997 (88.0)	0.918	0.692 (0.0)	0.738
200	98	4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.979 (49.0)	0.747
		5	1.000 (100.0)	0.974	1.000 (100.0)	0.906	0.883 (1.0)	0.712
		6	1.000 (100.0)	0.972	0.955 (50.5)	0.895	0.721 (0.0)	0.689
		7	1.000 (100.0)	0.970	0.931 (10.0)	0.886	0.590 (0.0)	0.672
500		4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	1.000 (100.0)	0.749
		5	1.000 (100.0)	0.974	1.000 (100.0)	0.906	0.986 (45.5)	0.712
		6	1.000 (100.0)	0.972	0.997 (98.0)	0.895	0.842 (0.0)	0.685
		7	1.000 (100.0)	0.970	0.992 (75.5)	0.887	0.716 (0.0)	0.666

of the error affects the estimation of M, although not equivalently in all settings. Indeed, in some cases (e.g., $n = 200, p = 60, Q = 6, M^{\mathrm{th}} = 3$) the percentage of samples in which M is correctly estimated decreases when the error level increases, whereas in some other cases (e.g., $n = 500, p = 60, Q = 6, M^{\mathrm{th}} = 3$) the opposite occurs.

Tables 5.5 and 5.7 provide the results of PCA + ORM, whereas Tables 5.6 and 5.8 those of the hierarchical clustering algorithms in the three scenarios with fixed Q and $M = M^{\mathrm{th}}$. It can be noticed that in all scenarios HierDPCA performs better than or equal to (e.g., to Ward's method with the low level of error) the competing methods. For PCA + ORM, the mean of the ARI decreases from the Qth level of the hierarchy upwards by stressing the incorrectness of component aggregations, although, for instance, the variable partition at the Qth level is always perfectly recovered for the low level of error. Contrariwise, the mean of the ARI for the hierarchical clustering algorithms increases from the Qth level of the hierarchy

Table 5.4. % of samples whose value of M estimated by HierDPCA corresponds to the true one, i.e., M^{th}, for each scenario in Table 5.1.

N. units	N. variables	M^{th}	$\sigma^{\text{L}}_{\mathbf{E}} = 0.66$	$\sigma^{\text{M}}_{\mathbf{E}} = 1.33$	$\sigma^{\text{H}}_{\mathbf{E}} = 2.66$
$n = 200$	$p = 60$	1	100.0	99.5	99.5
		3	96.5	95.5	93.0
	$p = 120$	1	100.0	100.0	100.0
		3	95.0	95.5	92.0
$n = 500$	$p = 60$	1	100.0	100.0	100.0
		3	94.5	96.0	98.0
	$p = 120$	1	100.0	100.0	100.0
		3	99.5	95.0	98.5
$n = 200$	$p = 98$	4	89.5	89.0	81.0
$n = 500$			88.5	90.0	89.0

upwards, as well as HierDPCA, by recovering the errors which occur at the lower levels of the hierarchy as the number of groups decreases, even if the mean of the ARI is always lower than that of HierDPCA.

5.5 Application

The proposed methodology described in Section 5.3 was applied to two real data sets. The first one is the Big Five Personality Test data set (Adachi & Trendafilov, 2018) on personality traits (Section 5.5.1) and the second one is the ASia-Europe Meeting (ASEM) Connectivity Sustainability Index data set (W. Becker et al., 2018) on relationships among countries, people and societies of the two regions (Section 5.5.2). The data sets are analyzed in an exploratory and mixed exploratory and confirmatory approach, respectively.

5.5.1 Big Five Personality Test

The Big Five Personality Test data set[2] detects personality traits by means of self-ratings reported on 25 items by 190 university students. The Big Five model (Digman, 1990; Goldberg, 1990, 1992; Costa & McCrae, 1992) defines the personality trait organization in five dimensions shown in Table 5.9. As observed by Digman (1997), in many studies two higher-order dimensions arise: *Alpha* and *Beta*, also called *Stability* and *Plasticity*, respectively, by DeYoung, Peterson, and Higgins (2002). The former is a combination of *Neuroticism*, *Agreeableness* and *Conscientiousness*, whereas the latter of *Extraversion* and *Openness* (see Digman, 1990, Table 1 for a complete review of the five dimensions' names). Some authors as Musek (2007) theorized the existence of a general single superordinate dimension, i.e., the Big One.

HierDPCA was applied on this data set in an exploratory approach in order to test its performances on recovering the theorized Big Five structure. Before

[2] Available at http://bstat.jp/en_material/.

Table 5.5. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of PCA + ORM for *Scenario 1*.

			Oblimin						Quartimin						Geomin					
			$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$		$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$		$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	2	0.131 (0.0)	0.967	0.632 (55.0)	0.897	0.053 (0.0)	0.679	0.131 (0.0)	0.967	0.632 (55.0)	0.897	0.053 (0.0)	0.679	0.227 (6.5)	0.967	0.183 (15.0)	0.902	0.061 (0.0)	0.693
		3	0.508 (0.5)	0.963	0.735 (37.0)	0.878	0.125 (0.0)	0.624	0.508 (0.5)	0.963	0.735 (37.0)	0.878	0.125 (0.0)	0.624	0.456 (3.0)	0.963	0.390 (18.5)	0.879	0.078 (0.0)	0.620
		4	0.631 (3.0)	0.962	0.643 (5.0)	0.869	0.154 (0.0)	0.596	0.631 (3.0)	0.962	0.643 (5.0)	0.869	0.154 (0.0)	0.596	0.604 (2.5)	0.962	0.528 (7.0)	0.867	0.109 (0.0)	0.589
		5	0.976 (93.0)	0.961	0.719 (4.0)	0.851	0.156 (0.0)	0.566	0.976 (93.0)	0.961	0.719 (4.0)	0.851	0.156 (0.0)	0.566	0.938 (82.0)	0.961	0.662 (0.0)	0.857	0.139 (0.0)	0.565
		6	1.000 (100.0)	0.958	0.725 (0.5)	0.840	0.157 (0.0)	0.546	1.000 (100.0)	0.958	0.725 (0.5)	0.840	0.157 (0.0)	0.546	1.000 (100.0)	0.958	0.716 (0.5)	0.838	0.150 (0.0)	0.536
	120	2	0.329 (6.0)	0.984	0.153 (1.0)	0.943	0.199 (0.0)	0.797	0.329 (6.0)	0.984	0.153 (1.0)	0.943	0.199 (0.0)	0.797	0.294 (14.5)	0.984	0.270 (5.0)	0.948	0.071 (0.0)	0.820
		3	0.498 (2.0)	0.982	0.524 (1.0)	0.933	0.373 (0.0)	0.762	0.498 (2.0)	0.982	0.524 (1.0)	0.933	0.373 (0.0)	0.762	0.513 (1.0)	0.981	0.506 (3.5)	0.934	0.082 (0.0)	0.766
		4	0.675 (5.0)	0.981	0.853 (30.0)	0.928	0.261 (0.0)	0.742	0.675 (5.0)	0.981	0.853 (30.0)	0.928	0.261 (0.0)	0.742	0.614 (2.5)	0.981	0.777 (30.0)	0.928	0.131 (0.0)	0.744
		5	0.978 (93.5)	0.980	0.877 (8.5)	0.922	0.220 (0.0)	0.713	0.978 (93.5)	0.980	0.877 (8.5)	0.922	0.220 (0.0)	0.713	0.952 (85.5)	0.980	0.830 (5.0)	0.923	0.181 (0.0)	0.725
		6	1.000 (100.0)	0.978	0.849 (0.5)	0.914	0.209 (0.0)	0.705	1.000 (100.0)	0.978	0.849 (0.5)	0.914	0.209 (0.0)	0.705	1.000 (100.0)	0.978	0.843 (0.5)	0.911	0.206 (0.0)	0.701
500	60	2	0.131 (0.0)	0.967	0.133 (2.0)	0.895	0.109 (0.0)	0.688	0.131 (0.0)	0.967	0.133 (2.0)	0.895	0.109 (0.0)	0.688	0.259 (5.5)	0.967	0.270 (6.5)	0.903	0.116 (0.0)	0.693
		3	0.526 (0.5)	0.964	0.547 (1.0)	0.878	0.310 (0.0)	0.640	0.526 (0.5)	0.964	0.547 (1.0)	0.878	0.310 (0.0)	0.640	0.500 (2.5)	0.963	0.532 (5.5)	0.879	0.204 (0.0)	0.625
		4	0.625 (2.0)	0.963	0.926 (77.0)	0.870	0.323 (0.0)	0.603	0.625 (2.0)	0.963	0.926 (77.0)	0.870	0.323 (0.0)	0.603	0.600 (1.0)	0.962	0.792 (60.0)	0.869	0.276 (0.0)	0.592
		5	0.991 (97.5)	0.962	0.940 (67.0)	0.861	0.347 (0.0)	0.566	0.991 (97.5)	0.962	0.940 (67.0)	0.861	0.347 (0.0)	0.566	0.956 (87.0)	0.961	0.886 (54.5)	0.861	0.315 (0.0)	0.564
		6	1.000 (100.0)	0.958	0.946 (40.5)	0.848	0.330 (0.0)	0.543	1.000 (100.0)	0.958	0.946 (40.5)	0.848	0.330 (0.0)	0.543	1.000 (100.0)	0.958	0.942 (41.0)	0.846	0.333 (0.0)	0.533
	120	2	0.015 (5.0)	0.983	0.864 (86.5)	0.948	0.342 (0.0)	0.828	0.015 (5.0)	0.983	0.864 (86.5)	0.948	0.342 (0.0)	0.828	0.176 (8.5)	0.984	0.563 (45.5)	0.948	0.218 (0.0)	0.815
		3	0.352 (0.0)	0.981	0.530 (0.5)	0.940	0.383 (0.0)	0.786	0.352 (0.0)	0.981	0.530 (0.5)	0.940	0.383 (0.0)	0.786	0.450 (0.0)	0.982	0.491 (3.5)	0.936	0.320 (0.0)	0.767
		4	0.660 (12.5)	0.980	0.699 (9.5)	0.933	0.477 (0.0)	0.759	0.660 (12.5)	0.980	0.699 (9.5)	0.933	0.477 (0.0)	0.759	0.695 (14.0)	0.981	0.612 (2.5)	0.930	0.439 (0.0)	0.743
		5	0.988 (96.5)	0.980	0.988 (94.0)	0.925	0.482 (0.0)	0.727	0.988 (96.5)	0.980	0.988 (94.0)	0.925	0.482 (0.0)	0.727	0.983 (95.0)	0.980	0.963 (86.0)	0.925	0.497 (0.0)	0.717
		6	1.000 (100.0)	0.979	0.993 (70.0)	0.918	0.440 (0.0)	0.704	1.000 (100.0)	0.979	0.993 (70.0)	0.918	0.440 (0.0)	0.704	1.000 (100.0)	0.979	0.993 (71.0)	0.918	0.439 (0.0)	0.687

Table 5.6. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of the hierarchical clustering algorithms for *Scenario 1*.

			Single Linkage						Complete Linkage					
			$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$		$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	2	1.000 (100.0)	0.977	0.878 (87.5)	0.907	0.000 (0.0)	0.814	1.000 (100.0)	0.977	0.954 (92.5)	0.900	0.362 (0.0)	0.747
		3	0.978 (95.5)	0.970	0.529 (3.0)	0.895	0.009 (0.0)	0.804	0.982 (96.0)	0.970	0.802 (50.0)	0.883	0.340 (0.0)	0.682
		4	0.923 (79.5)	0.966	0.326 (0.0)	0.890	0.014 (0.0)	0.790	0.963 (91.0)	0.966	0.643 (5.0)	0.866	0.293 (0.0)	0.634
		5	0.884 (47.0)	0.963	0.212 (0.0)	0.878	0.006 (0.0)	0.777	0.937 (80.5)	0.961	0.609 (0.5)	0.847	0.228 (0.0)	0.600
		6	0.753 (6.0)	0.961	0.150 (0.0)	0.869	0.003 (0.0)	0.757	0.983 (90.0)	0.957	0.595 (0.0)	0.825	0.184 (0.0)	0.571
	120	2	1.000 (100.0)	0.988	0.570 (57.0)	0.965	0.008 (0.0)	0.904	1.000 (100.0)	0.988	0.997 (98.5)	0.963	0.395 (0.0)	0.844
		3	0.998 (99.5)	0.985	0.561 (0.5)	0.961	0.020 (0.0)	0.894	0.998 (99.5)	0.985	0.927 (58.5)	0.943	0.375 (0.0)	0.793
		4	0.919 (77.0)	0.982	0.526 (0.0)	0.959	0.013 (0.0)	0.884	0.959 (89.5)	0.982	0.792 (12.5)	0.927	0.255 (0.0)	0.754
		5	0.874 (33.5)	0.981	0.367 (0.0)	0.952	0.006 (0.0)	0.862	0.929 (77.5)	0.980	0.684 (0.0)	0.913	0.195 (0.0)	0.721
		6	0.731 (1.0)	0.980	0.286 (0.0)	0.951	0.001 (0.0)	0.842	0.984 (85.5)	0.978	0.633 (0.0)	0.896	0.166 (0.0)	0.695
500	60	2	1.000 (100.0)	0.977	0.996 (99.5)	0.929	0.000 (0.0)	0.817	1.000 (100.0)	0.977	1.000 (100.0)	0.929	0.755 (25.5)	0.761
		3	1.000 (100.0)	0.970	0.927 (69.5)	0.903	0.009 (0.0)	0.799	1.000 (100.0)	0.970	0.973 (89.5)	0.896	0.642 (1.5)	0.678
		4	0.976 (94.0)	0.966	0.850 (13.5)	0.893	0.011 (0.0)	0.781	0.990 (97.5)	0.966	0.856 (56.5)	0.871	0.473 (0.0)	0.624
		5	0.958 (87.5)	0.961	0.604 (0.0)	0.885	0.024 (0.0)	0.751	0.956 (87.0)	0.961	0.797 (17.5)	0.855	0.399 (0.0)	0.586
		6	0.937 (70.0)	0.959	0.464 (0.0)	0.874	0.018 (0.0)	0.719	1.000 (100.0)	0.958	0.810 (5.5)	0.836	0.337 (0.0)	0.553
	120	2	1.000 (100.0)	0.988	1.000 (100.0)	0.947	0.001 (0.0)	0.900	1.000 (100.0)	0.988	1.000 (100.0)	0.947	0.823 (18.0)	0.861
		3	1.000 (100.0)	0.986	0.565 (29.0)	0.945	0.000 (0.0)	0.890	1.000 (100.0)	0.986	0.985 (95.5)	0.945	0.566 (0.0)	0.811
		4	0.997 (99.0)	0.983	0.505 (0.5)	0.943	0.002 (0.0)	0.889	0.994 (98.0)	0.983	0.940 (71.0)	0.934	0.531 (0.0)	0.761
		5	0.982 (94.0)	0.980	0.455 (0.0)	0.940	0.002 (0.0)	0.878	0.967 (90.0)	0.980	0.892 (38.0)	0.922	0.478 (0.0)	0.721
		6	0.922 (59.5)	0.979	0.376 (0.0)	0.935	0.003 (0.0)	0.856	1.000 (100.0)	0.979	0.853 (3.5)	0.908	0.354 (0.0)	0.688

			Average Linkage						Ward's Method					
			$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$		$\sigma^{\mathrm{L}}_{\mathbf{E}} = 0.66$		$\sigma^{\mathrm{M}}_{\mathbf{E}} = 1.33$		$\sigma^{\mathrm{H}}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	2	1.000 (100.0)	0.977	1.000 (99.5)	0.901	0.492 (0.5)	0.748	1.000 (100.0)	0.977	1.000 (99.5)	0.901	0.637 (1.5)	0.770
		3	0.993 (98.5)	0.970	0.934 (72.5)	0.879	0.489 (0.0)	0.659	0.995 (99.0)	0.970	0.959 (72.0)	0.881	0.492 (0.0)	0.699
		4	0.963 (91.0)	0.966	0.721 (12.0)	0.860	0.470 (0.0)	0.605	1.000 (100.0)	0.966	0.869 (18.0)	0.870	0.390 (0.0)	0.656
		5	0.951 (85.0)	0.961	0.574 (1.0)	0.839	0.325 (0.0)	0.571	0.984 (94.0)	0.962	0.763 (1.5)	0.859	0.277 (0.0)	0.623
		6	0.945 (73.5)	0.958	0.494 (0.0)	0.815	0.237 (0.0)	0.545	0.999 (97.0)	0.958	0.720 (0.0)	0.840	0.223 (0.0)	0.596
	120	2	1.000 (100.0)	0.988	1.000 (99.5)	0.963	0.513 (1.0)	0.821	1.000 (100.0)	0.988	1.000 (99.0)	0.963	0.717 (0.0)	0.855
		3	1.000 (100.0)	0.985	0.956 (68.5)	0.944	0.647 (0.0)	0.732	0.998 (99.5)	0.985	0.995 (76.5)	0.944	0.522 (0.0)	0.803
		4	0.966 (91.5)	0.982	0.882 (13.0)	0.920	0.433 (0.0)	0.665	1.000 (100.0)	0.983	0.956 (28.5)	0.930	0.339 (0.0)	0.772
		5	0.934 (79.0)	0.980	0.664 (0.0)	0.895	0.282 (0.0)	0.624	0.988 (96.0)	0.980	0.841 (0.0)	0.924	0.255 (0.0)	0.748
		6	0.943 (66.0)	0.978	0.546 (0.0)	0.877	0.201 (0.0)	0.600	0.998 (91.5)	0.978	0.781 (0.0)	0.914	0.230 (0.0)	0.724
500	60	2	1.000 (100.0)	0.977	1.000 (100.0)	0.929	0.923 (35.5)	0.774	1.000 (100.0)	0.977	1.000 (100.0)	0.929	0.916 (31.0)	0.773
		3	1.000 (100.0)	0.970	1.000 (100.0)	0.896	0.759 (2.0)	0.672	1.000 (100.0)	0.970	1.000 (100.0)	0.896	0.805 (3.5)	0.690
		4	0.988 (97.0)	0.966	0.939 (73.5)	0.874	0.586 (0.0)	0.610	1.000 (100.0)	0.966	0.976 (89.0)	0.872	0.609 (0.0)	0.645
		5	0.962 (89.0)	0.961	0.816 (28.0)	0.859	0.494 (0.0)	0.566	0.997 (99.0)	0.962	0.952 (45.5)	0.862	0.483 (0.0)	0.605
		6	1.000 (100.0)	0.958	0.749 (9.0)	0.842	0.381 (0.0)	0.533	1.000 (99.5)	0.958	0.917 (19.5)	0.848	0.408 (0.0)	0.573
	120	2	1.000 (100.0)	0.988	1.000 (100.0)	0.947	0.950 (26.5)	0.866	1.000 (100.0)	0.988	1.000 (100.0)	0.947	0.946 (21.5)	0.865
		3	1.000 (100.0)	0.986	0.999 (98.0)	0.945	0.611 (0.0)	0.766	1.000 (100.0)	0.986	1.000 (99.0)	0.945	0.791 (0.0)	0.830
		4	0.997 (99.0)	0.983	0.958 (78.0)	0.936	0.591 (0.0)	0.693	1.000 (100.0)	0.983	0.997 (86.5)	0.936	0.731 (0.0)	0.777
		5	0.982 (94.5)	0.980	0.890 (29.5)	0.926	0.567 (0.0)	0.658	1.000 (100.0)	0.980	0.985 (53.5)	0.925	0.638 (0.0)	0.741
		6	0.994 (97.0)	0.979	0.788 (1.5)	0.913	0.451 (0.0)	0.624	1.000 (100.0)	0.979	0.951 (7.0)	0.918	0.458 (0.0)	0.710

Table 5.7. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of PCA + ORM for *Scenario 2* and *Scenario 3*.

			Oblimin						Quartimin						Geomin					
			$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$		$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$		$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	3	0.308 (0.0)	0.951	0.363 (0.0)	0.854	0.245 (0.0)	0.560	0.308 (0.0)	0.951	0.363 (0.0)	0.854	0.245 (0.0)	0.560	0.386 (0.0)	0.951	0.379 (0.0)	0.849	0.225 (0.0)	0.543
		4	0.610 (4.0)	0.955	0.663 (5.0)	0.853	0.331 (0.0)	0.538	0.610 (4.0)	0.955	0.663 (5.0)	0.853	0.331 (0.0)	0.538	0.638 (5.5)	0.954	0.601 (0.5)	0.843	0.297 (0.0)	0.516
		5	0.971 (91.5)	0.961	0.835 (10.5)	0.848	0.348 (0.0)	0.513	0.971 (91.5)	0.961	0.835 (10.5)	0.848	0.348 (0.0)	0.513	0.954 (86.5)	0.960	0.744 (2.5)	0.830	0.327 (0.0)	0.491
		6	1.000 (99.5)	0.958	0.775 (1.0)	0.831	0.303 (0.0)	0.485	1.000 (99.5)	0.958	0.775 (1.0)	0.831	0.303 (0.0)	0.485	1.000 (99.5)	0.958	0.726 (0.5)	0.793	0.304 (0.0)	0.458
	120	3	0.308 (0.0)	0.975	0.329 (0.0)	0.922	0.320 (0.0)	0.723	0.308 (0.0)	0.975	0.329 (0.0)	0.922	0.320 (0.0)	0.723	0.480 (1.5)	0.976	0.394 (0.0)	0.922	0.296 (0.0)	0.695
		4	0.610 (2.5)	0.977	0.625 (6.0)	0.922	0.437 (0.0)	0.705	0.610 (2.5)	0.977	0.625 (6.0)	0.922	0.437 (0.0)	0.705	0.712 (15.5)	0.977	0.639 (1.5)	0.920	0.385 (0.0)	0.671
		5	0.987 (96.0)	0.980	0.921 (22.0)	0.924	0.438 (0.0)	0.674	0.987 (96.0)	0.980	0.921 (22.0)	0.924	0.438 (0.0)	0.674	0.978 (93.5)	0.980	0.867 (14.5)	0.918	0.416 (0.0)	0.644
		6	1.000 (100.0)	0.978	0.875 (0.0)	0.915	0.359 (0.0)	0.649	1.000 (100.0)	0.978	0.875 (0.0)	0.915	0.359 (0.0)	0.649	1.000 (100.0)	0.978	0.860 (0.0)	0.905	0.366 (0.0)	0.615
500	60	3	0.315 (0.0)	0.951	0.321 (0.0)	0.856	0.357 (0.0)	0.591	0.315 (0.0)	0.951	0.321 (0.0)	0.856	0.357 (0.0)	0.591	0.410 (0.5)	0.952	0.386 (0.0)	0.858	0.355 (0.0)	0.574
		4	0.607 (3.5)	0.955	0.613 (4.0)	0.856	0.534 (0.0)	0.571	0.607 (3.5)	0.955	0.613 (4.0)	0.856	0.534 (0.0)	0.571	0.658 (7.0)	0.954	0.620 (4.5)	0.856	0.456 (0.0)	0.540
		5	0.981 (94.5)	0.961	0.965 (81.5)	0.861	0.559 (0.0)	0.546	0.981 (94.5)	0.961	0.965 (81.5)	0.861	0.559 (0.0)	0.546	0.961 (88.5)	0.961	0.907 (61.5)	0.858	0.498 (0.0)	0.511
		6	1.000 (100.0)	0.958	0.957 (47.5)	0.847	0.456 (0.0)	0.515	1.000 (100.0)	0.958	0.957 (47.5)	0.847	0.456 (0.0)	0.515	1.000 (100.0)	0.958	0.944 (44.0)	0.843	0.441 (0.0)	0.474
	120	3	0.313 (0.0)	0.975	0.304 (0.0)	0.922	0.404 (0.0)	0.756	0.313 (0.0)	0.975	0.304 (0.0)	0.922	0.404 (0.0)	0.756	0.522 (3.5)	0.977	0.460 (1.0)	0.927	0.403 (0.0)	0.730
		4	0.615 (3.5)	0.977	0.609 (2.0)	0.922	0.652 (0.0)	0.744	0.615 (3.5)	0.977	0.609 (2.0)	0.922	0.652 (0.0)	0.744	0.741 (19.5)	0.977	0.691 (10.5)	0.924	0.558 (0.0)	0.707
		5	0.993 (98.0)	0.980	0.993 (94.5)	0.925	0.669 (0.0)	0.714	0.993 (98.0)	0.980	0.993 (94.5)	0.925	0.669 (0.0)	0.714	0.985 (95.5)	0.980	0.983 (92.0)	0.925	0.601 (0.0)	0.676
		6	1.000 (100.0)	0.979	0.991 (68.0)	0.918	0.545 (0.0)	0.685	1.000 (100.0)	0.979	0.991 (68.0)	0.918	0.545 (0.0)	0.685	1.000 (100.0)	0.979	0.990 (68.5)	0.918	0.535 (0.0)	0.639
200	98	4	0.420 (0.0)	0.966	0.465 (0.0)	0.889	0.375 (0.0)	0.620	0.420 (0.0)	0.966	0.465 (0.0)	0.889	0.375 (0.0)	0.620	0.511 (0.0)	0.966	0.494 (0.0)	0.885	0.350 (0.0)	0.603
		5	0.682 (7.0)	0.968	0.712 (3.5)	0.889	0.459 (0.0)	0.605	0.682 (7.0)	0.968	0.712 (3.5)	0.889	0.459 (0.0)	0.605	0.713 (6.5)	0.968	0.708 (1.0)	0.882	0.413 (0.0)	0.579
		6	0.981 (93.5)	0.972	0.904 (18.0)	0.886	0.464 (0.0)	0.581	0.981 (93.5)	0.972	0.904 (18.0)	0.886	0.464 (0.0)	0.581	0.970 (89.5)	0.971	0.840 (9.0)	0.871	0.445 (0.0)	0.560
		7	1.000 (100.0)	0.970	0.844 (1.0)	0.874	0.401 (0.0)	0.558	1.000 (100.0)	0.970	0.844 (1.0)	0.874	0.401 (0.0)	0.558	1.000 (100.0)	0.970	0.812 (1.0)	0.846	0.409 (0.0)	0.525
500		4	0.424 (0.0)	0.966	0.412 (0.0)	0.890	0.508 (0.0)	0.677	0.424 (0.0)	0.966	0.412 (0.0)	0.890	0.508 (0.0)	0.677	0.554 (1.0)	0.966	0.513 (1.0)	0.893	0.465 (0.0)	0.635
		5	0.669 (3.0)	0.968	0.669 (3.0)	0.892	0.676 (0.0)	0.663	0.669 (3.0)	0.968	0.669 (3.0)	0.892	0.676 (0.0)	0.663	0.751 (11.0)	0.967	0.712 (5.5)	0.892	0.608 (0.0)	0.621
		6	0.993 (97.5)	0.972	0.994 (94.0)	0.895	0.678 (0.0)	0.635	0.993 (97.5)	0.972	0.994 (94.0)	0.895	0.678 (0.0)	0.635	0.989 (96.0)	0.972	0.984 (89.5)	0.895	0.633 (0.0)	0.597
		7	1.000 (100.0)	0.970	0.977 (48.5)	0.887	0.583 (0.0)	0.609	1.000 (100.0)	0.970	0.977 (48.5)	0.887	0.583 (0.0)	0.609	1.000 (100.0)	0.970	0.974 (47.5)	0.885	0.572 (0.0)	0.561

Table 5.8. Mean of the ARI, % of samples with ARI equal to one (in brackets) and $\bar{\alpha}$ of the hierarchical clustering algorithms for *Scenario 2* and *Scenario 3*.

			Single Linkage						Complete Linkage					
			$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$		$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.905	0.026 (0.0)	0.679	1.000 (100.0)	0.971	1.000 (100.0)	0.905	0.743 (4.5)	0.687
		4	0.992 (96.5)	0.966	0.856 (5.5)	0.899	0.051 (0.0)	0.663	0.990 (96.5)	0.966	0.934 (60.5)	0.875	0.678 (0.0)	0.628
		5	0.909 (54.0)	0.962	0.714 (0.0)	0.889	0.064 (0.0)	0.653	0.908 (73.0)	0.960	0.810 (7.0)	0.848	0.548 (0.0)	0.588
		6	0.768 (6.0)	0.959	0.553 (0.0)	0.882	0.058 (0.0)	0.636	0.975 (85.5)	0.956	0.705 (0.0)	0.824	0.419 (0.0)	0.558
	120	3	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.000 (0.0)	0.800	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.824 (4.5)	0.817
		4	0.994 (97.5)	0.983	0.863 (3.0)	0.949	0.006 (0.0)	0.793	0.993 (97.5)	0.983	0.951 (59.0)	0.933	0.698 (0.0)	0.753
		5	0.905 (44.0)	0.981	0.717 (0.0)	0.948	0.006 (0.0)	0.784	0.929 (78.5)	0.980	0.827 (2.5)	0.913	0.581 (0.0)	0.712
		6	0.751 (1.5)	0.980	0.559 (0.0)	0.944	0.009 (0.0)	0.770	0.981 (87.0)	0.977	0.705 (0.0)	0.895	0.439 (0.0)	0.682
500	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.906	0.610 (27.5)	0.700	1.000 (100.0)	0.971	1.000 (100.0)	0.906	0.986 (81.0)	0.713
		4	1.000 (100.0)	0.966	0.940 (61.5)	0.889	0.602 (0.0)	0.694	0.999 (99.5)	0.966	0.974 (90.5)	0.879	0.825 (4.5)	0.636
		5	0.967 (90.0)	0.961	0.793 (0.0)	0.881	0.533 (0.0)	0.687	0.952 (86.0)	0.961	0.875 (38.0)	0.856	0.666 (0.0)	0.590
		6	0.930 (66.5)	0.959	0.626 (0.0)	0.870	0.433 (0.0)	0.678	0.997 (98.0)	0.958	0.835 (7.0)	0.835	0.512 (0.0)	0.556
	120	3	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.465 (13.0)	0.826	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.994 (83.0)	0.832
		4	1.000 (100.0)	0.983	0.909 (37.0)	0.944	0.481 (0.0)	0.821	1.000 (100.0)	0.983	0.986 (94.0)	0.935	0.846 (0.5)	0.765
		5	0.984 (94.0)	0.980	0.781 (0.0)	0.940	0.471 (0.0)	0.816	0.974 (92.0)	0.980	0.901 (44.5)	0.923	0.691 (0.0)	0.724
		6	0.905 (51.5)	0.979	0.620 (0.0)	0.936	0.385 (0.0)	0.813	0.999 (99.5)	0.979	0.861 (6.0)	0.908	0.523 (0.0)	0.695
200	98	4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.006 (0.0)	0.720	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.704 (0.0)	0.708
		5	0.998 (99.0)	0.974	0.879 (1.5)	0.918	0.009 (0.0)	0.700	0.991 (96.5)	0.974	0.946 (55.5)	0.903	0.656 (0.0)	0.674
		6	0.914 (43.5)	0.972	0.750 (0.0)	0.915	0.015 (0.0)	0.678	0.942 (79.5)	0.971	0.830 (2.0)	0.883	0.575 (0.0)	0.641
		7	0.785 (1.5)	0.971	0.607 (0.0)	0.912	0.015 (0.0)	0.663	0.978 (83.5)	0.969	0.747 (0.0)	0.865	0.472 (0.0)	0.616
500		4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.501 (12.0)	0.735	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.984 (70.5)	0.747
		5	1.000 (100.0)	0.974	0.937 (50.0)	0.913	0.493 (0.0)	0.730	1.000 (100.0)	0.974	0.991 (95.0)	0.905	0.864 (1.0)	0.692
		6	0.979 (92.0)	0.972	0.816 (0.0)	0.909	0.472 (0.0)	0.726	0.963 (87.0)	0.972	0.908 (38.5)	0.892	0.724 (0.0)	0.653
		7	0.929 (59.0)	0.970	0.673 (0.0)	0.905	0.405 (0.0)	0.721	1.000 (100.0)	0.970	0.865 (5.5)	0.875	0.591 (0.0)	0.625

			Average Linkage						Ward's Method					
			$\sigma^{L}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$		$\sigma^{M}_{\mathbf{E}} = 0.66$		$\sigma^{M}_{\mathbf{E}} = 1.33$		$\sigma^{H}_{\mathbf{E}} = 2.66$	
n	p	q	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$	ARI	$\bar{\alpha}$
200	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.905	0.886 (12.0)	0.706	1.000 (100.0)	0.971	1.000 (100.0)	0.905	0.871 (9.0)	0.706
		4	0.994 (98.0)	0.966	0.951 (65.0)	0.880	0.773 (0.0)	0.639	1.000 (100.0)	0.966	0.993 (80.0)	0.878	0.737 (0.0)	0.645
		5	0.938 (81.5)	0.961	0.817 (5.0)	0.854	0.632 (0.0)	0.589	0.981 (94.0)	0.961	0.888 (15.5)	0.858	0.588 (0.0)	0.607
		6	0.923 (62.5)	0.956	0.669 (0.0)	0.834	0.481 (0.0)	0.556	0.997 (93.5)	0.958	0.791 (0.5)	0.840	0.451 (0.0)	0.579
	120	3	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.913 (8.5)	0.826	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.903 (4.0)	0.827
		4	0.997 (99.0)	0.983	0.956 (60.5)	0.935	0.796 (0.0)	0.736	1.000 (100.0)	0.983	0.996 (75.0)	0.935	0.795 (0.0)	0.773
		5	0.940 (81.0)	0.980	0.817 (0.0)	0.908	0.663 (0.0)	0.669	0.990 (97.0)	0.980	0.922 (5.0)	0.924	0.637 (0.0)	0.737
		6	0.939 (67.5)	0.977	0.667 (0.0)	0.886	0.513 (0.0)	0.629	0.999 (95.0)	0.978	0.832 (0.0)	0.916	0.475 (0.0)	0.710
500	60	3	1.000 (100.0)	0.971	1.000 (100.0)	0.906	0.994 (89.5)	0.714	1.000 (100.0)	0.971	1.000 (100.0)	0.906	0.989 (80.0)	0.714
		4	1.000 (100.0)	0.966	0.993 (96.5)	0.880	0.865 (5.5)	0.658	1.000 (100.0)	0.966	1.000 (99.0)	0.880	0.904 (8.0)	0.649
		5	0.971 (91.5)	0.961	0.879 (35.5)	0.864	0.723 (0.0)	0.610	0.998 (99.5)	0.962	0.975 (70.0)	0.861	0.728 (0.0)	0.606
		6	0.995 (97.0)	0.958	0.784 (5.0)	0.849	0.556 (0.0)	0.571	1.000 (99.5)	0.958	0.937 (22.5)	0.847	0.554 (0.0)	0.572
	120	3	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.996 (85.0)	0.832	1.000 (100.0)	0.985	1.000 (100.0)	0.951	0.995 (82.5)	0.832
		4	1.000 (100.0)	0.983	0.996 (98.0)	0.936	0.873 (1.0)	0.774	1.000 (100.0)	0.983	1.000 (99.5)	0.936	0.944 (2.0)	0.783
		5	0.987 (96.0)	0.980	0.907 (39.5)	0.922	0.742 (0.0)	0.727	1.000 (100.0)	0.980	0.988 (65.0)	0.925	0.763 (0.0)	0.747
		6	0.995 (97.5)	0.979	0.785 (3.0)	0.909	0.584 (0.0)	0.686	1.000 (100.0)	0.979	0.953 (10.5)	0.918	0.588 (0.0)	0.720
200	98	4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.881 (1.0)	0.736	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.862 (2.0)	0.736
		5	0.998 (99.0)	0.974	0.947 (52.5)	0.904	0.785 (0.0)	0.673	1.000 (100.0)	0.974	0.994 (73.5)	0.905	0.763 (0.0)	0.693
		6	0.940 (78.0)	0.971	0.847 (2.0)	0.887	0.675 (0.0)	0.629	0.983 (94.0)	0.972	0.928 (9.0)	0.894	0.636 (0.0)	0.664
		7	0.934 (60.5)	0.968	0.714 (0.0)	0.867	0.544 (0.0)	0.596	0.999 (95.5)	0.970	0.845 (0.0)	0.882	0.512 (0.0)	0.640
500		4	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.994 (79.0)	0.749	1.000 (100.0)	0.977	1.000 (100.0)	0.920	0.990 (71.0)	0.748
		5	1.000 (100.0)	0.974	0.997 (98.0)	0.906	0.890 (1.5)	0.700	1.000 (100.0)	0.974	1.000 (99.0)	0.906	0.941 (3.5)	0.708
		6	0.979 (92.5)	0.972	0.909 (31.5)	0.895	0.768 (0.0)	0.662	1.000 (100.0)	0.972	0.985 (62.0)	0.895	0.780 (0.0)	0.676
		7	0.996 (97.5)	0.970	0.807 (3.0)	0.880	0.628 (0.0)	0.636	1.000 (100.0)	0.970	0.944 (13.0)	0.886	0.640 (0.0)	0.648

Table 5.9. Observed variables of the Big Five Personality Test data set, corresponding dimensions and their Cronbach's α.

Dimension	Variable ID	Variable	α	Dimension	Variable ID	Variable	α
Neuroticism	1	Worry	0.789	Agreeableness	16	Mild	0.787
	2	Sensitive			17	Tenderhearted	
	3	Pessimistic			18	Altruistic	
	4	Unrest			19	Cooperative	
	5	Careful			20	Sympathetic	
Extraversion	6	Sociable	0.874	Conscientiousness	21	Deliberate	0.807
	7	Talkative			22	Reliable	
	8	Voluntary			23	Diligent	
	9	Cheerful			24	Systematic	
	10	Showy			25	Methodical	
Openness	11	Creative	0.754				
	12	Adventurous					
	13	Progressive					
	14	Flexible					
	15	Imaginative					

analyzing its hierarchical structure, the optimal Q was chosen. The implemented criterion based on the unidimensionality of each variable group converges to retain Q equal to five. Figure 5.3 displays the results of the HierDPCA application on the Big Five Personality Test data set. The path diagram is built such that the aggregation level of each node representing a disjoint principal component is the sum of the disjoint principal component deviance and the maximum disjoint principal component deviance of the lower hierarchical level over the total deviance of the data matrix. If $M > 1$, the aggregation level of the last node is the sum of the deviance reconstructed by model (5.9) and the maximum disjoint principal component deviance of the lower hierarchical level over the total deviance of the data matrix. To help the reader to follow the path diagram representation, dashed horizontal lines are included, that correspond to the highest aggregation level for each $q = Q, \ldots, M$; the difference between the node and the lower dashed horizontal line represents the ratio between the disjoint principal component deviance (or the reconstructed one for the formative index) and the total deviance of the data. The size of each node of the reflective part of the hierarchy is set according to the Cronbach's α of the corresponding variable group: the thresholds are set according to the rule of thumb given by George and Mallery (2003), by considering 0.7 as the acceptable level of α (Nunnally, 1978). As shown in Figure 5.3, HierDPCA perfectly identifies the five dimensions as well as the "higher-order" two, i.e., *Alpha/Stability* and *Beta/Plasticity*. The five dimensions are reliable, unidimensional and positively correlated with the corresponding observed variables; the two "higher-order" dimensions are reliable in turn. The two broader dimensions turn out to be uncorrelated such that $M = 2$ and the general personality index is built in a formative approach. *Alpha/Stability* and *Beta/Plasticity* contribute to the definition of the general index with coefficients equal to 0.773 and 0.635, respectively. HierDPCA thus seems to be suitable to detect the Big Five structure of personality traits since it is able to pinpoint the five variable groups and the whole hierarchy over them.

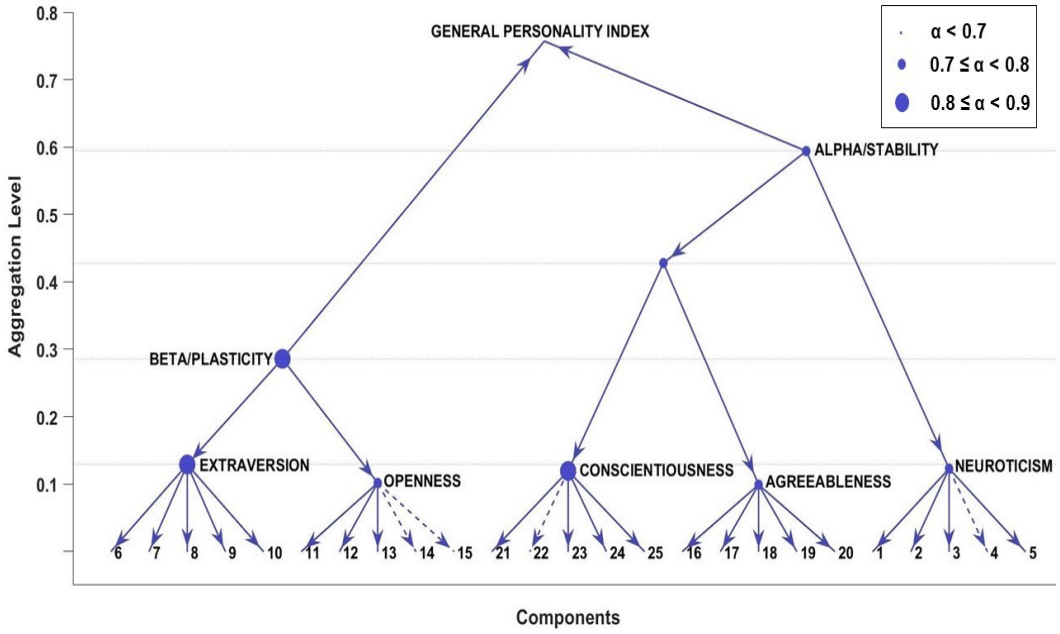


Figure 5.3. Path diagram of HierDPCA on the Big Five Personality Test data set. Dashed arrows connecting the five dimensions with the corresponding variables represent correlations < 0.7.

5.5.2 ASEM Connectivity Sustainability Index

The ASia-Europe Meeting (ASEM) Connectivity Sustainability Index aims at quantifying the relationships among countries, people and societies of the two regions in a social and economic sense. The ASEM data set[3] comprises 51 countries - 31 European and 20 Asian - and 49 observed variables grouped in two main dimensions: *Connectivity* and *Sustainability*. These are in turn composed of five and three pillars, respectively (Table 5.10). The *Connectivity* and *Sustainability* indices were calculated by firstly averaging the normalized indicators at pillar level, and then the pillars to obtain the two composite indicators. This approach was based upon experts' consultations (see W. Becker et al., 2018, for further details).

HierDPCA was implemented on this data set in a mixed confirmatory and exploratory approach, i.e., by imposing the membership of each observed variable to the corresponding pillar at the Qth level of the hierarchy and letting the pillar aggregations be chosen by the model. The application of HierDPCA presented herein aims at investigating the conceptual framework of the ASEM sustainable connectivity, which is based on experts' review of literature on globalization indicators, and inspecting the relationships among pillars with a model-based approach. Data were normalized via the min-max normalization, s.t. each observed variable ranges between 0 and 100, as done by W. Becker et al. (2018), and centered.

The results of the HierDPCA application on the ASEM data set are shown in Figure 5.4. The proposed methodology detects some negative correlations between

[3]Available at https://composite-indicators.jrc.ec.europa.eu/asem-sustainable-connectivity/ repository.

Table 5.10. ASEM Connectivity and Sustainability pillars. Source: W. Becker et al. (2018, pp. 23-25).

		Connectivity
ID vars.	Pillar	Description
1 : 8	Physical	Physical infrastructure in terms of transport, energy and ICT between countries.
9 : 13	Economic/Financial	Trade of goods and services and financial flows.
14 : 16	Political	Political relations with other countries.
17 : 22	Institutional	Regulatory environment to facilitate trade, investment, mobility of people.
23 : 30	People-to People	Mobility of people in education, tourism and migration, exchange of culture and communication.
		Sustainability
ID vars.	Pillar	Description
31 : 35	Environmental	Countries' CO_2 emissions, domestic material consumption, forest loss, intensity of renewable energies.
36 : 44	Social	Poverty, inequality, education, gender balance and inclusive and open societies.
45 : 49	Economic/Financial	Financial sustainability, economic growth, research expenditure and youth unemployment.

variables and pillars, specifically for variables $19, 35, 48, 49$, and/or low correlations, i.e., $< |0.7|$ (dashed arrows in Figure 5.4). All pillars are not reliable, except for the *People-to-People* and *Social* ones. HierDPCA builds a hierarchy over the eight pillars which turns from reflective to formative at level $M = 2$. Before analyzing the two broader variable (and pillars) groups, it is noteworthy to highlight some lower level aggregations. Indeed, HierDPCA merges: *Economic/Financial (Connectivity)* with *People-to-People* first, whose indicators are connected with trades in different areas (e.g., goods, services, culture and research), and then these pillars with *Physical*; *Political* and *Institutional*, whose indicators are connected with international networks and agreements; and the three pillars which define the theoretical *Sustainability* index. The two broader variable groups are composed of three out of the five pillars of *Connectivity* and the *Sustainability* pillars merged with the *Political* and *Institutional* ones. This result highlights that the *Sustainability* pillars are significantly related to the ones affecting political relations and international agreements on flows (of goods, people and investments) among countries.

In order to compare the theoretical structure in two dimensions (*Connectivity* and *Sustainability*) with that one obtained by HierDPCA at $q = 2$, we can compute a confirmatory DPCA on the ASEM data set by fixing the membership of the variables to the corresponding dimensions. The percentage of the variance explained by the components corresponding to *Connectivity* and *Sustainability* is 40.29%, whereas the two components identified by HierDPCA at $q = 2$ explain 45.17% of the variance of the data.

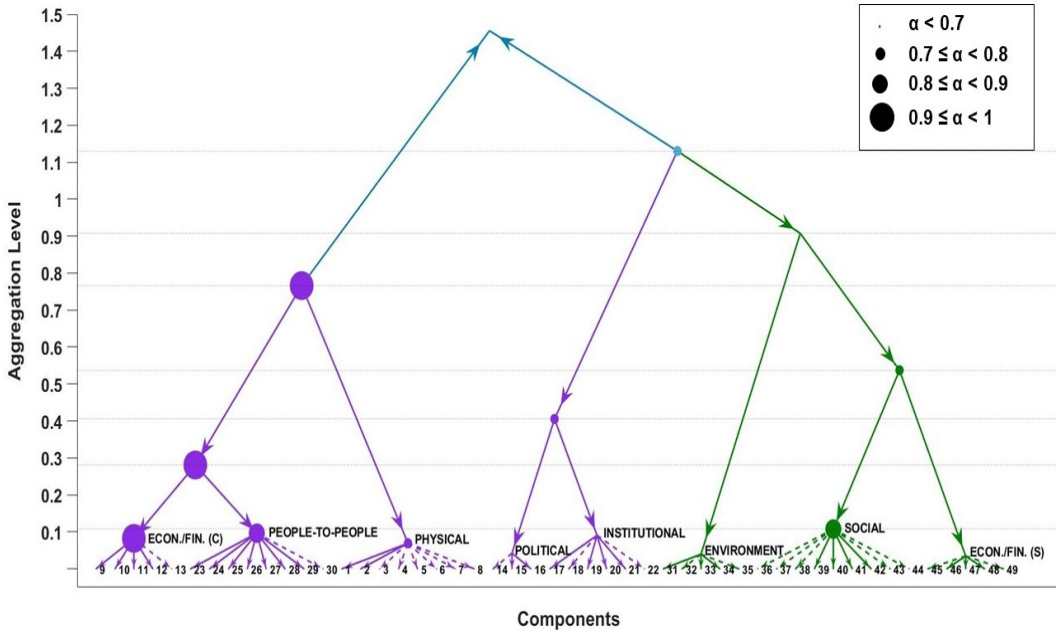


Figure 5.4. Path diagram of HierDPCA on the ASEM data set. Dashed arrows represent correlations between the observed variable and the corresponding pillar $< |0.7|$. Colors, purple and green, identify the pillars corresponding to *Connectivity* and *Sustainability*, respectively.

5.6 Conclusions

In this chapter a hierarchical extension of DPCA, called Hierarchical Disjoint Principal Component Analysis, is proposed. The proposal aims at: 1) building a hierarchy of disjoint principal components with the largest explained variance by starting from a dimensionality reduction of the observed variables into Q disjoint correlated groups, each one associated with a component, and, eventually, merging them in pairs up to the identification of a general component; 2) estimating the variable partition for each hierarchical level, that is connected with the one of lower level of the hierarchy by the nestedness assumption, is identified; 3) defining a model-based approach to choose the "direction" of relationships (reflective or formative) between the disjoint principal components of two contiguous hierarchical levels. The latter is usually based upon a theoretical conceptualization of the phenomenon under study and defined a priori by the researcher, without any empirical confirmation. Contrariwise, HierDPCA tests for correlation among disjoint principal components at each bottom-up level of the hierarchy such that the model turns from reflective to formative if this correlation is not statistically significant. The analysis can be conducted either through a confirmatory approach, in case the researcher has a theory to confirm, or through an exploratory approach guided by the observed data.

Differently from the existing methodologies proposed to investigate hierarchical relationships among observed variables, which are based upon sequential application of Factor Analysis followed by an Oblique Rotation Method, HierDPCA is a simultaneous model. The proposal allows to overcome the duality between

Higher-Order Factor and Bi-Factor or Hierarchical Factor models. Indeed, HierDPCA both estimates direct relationships between observed variables and disjoint principal components for each hierarchical level, and defines a hierarchy, since the disjoint principal components of two sequential hierarchical levels are linked by the nestedness assumption among the corresponding variable partitions.

The parameters of HierDPCA are estimated in a least-squares framework according to a simultaneous approach in which a constrained minimization problem is solved for the whole hierarchy. A coordinate descent algorithm is proposed for the parameter estimation. Some properties of the proposed methodology are illustrated; these highlight the similarities (e.g., the maximization of the total variance for each level of the hierarchy) and differences (e.g., the uniqueness of the component loading matrices for each level of the hierarchy) with the modeling approach to PCA. Furthermore, it is worth noticing that HierDPCA can be applied also in the context of wide data (more variables than observations), and the fact that variables are partitioned into groups allows computing the eigen-decomposition also for a large number of variables. A large-scale simulation study demonstrates the good performances of HierDPCA in identifying the variable partition for each hierarchical level, also w.r.t. competing methods as PCA + ORM and hierarchical clustering algorithms, and correctly choosing the level at which the model turns from reflective to formative, i.e., the "direction" of the relationship among disjoint principal components in two contiguous levels. Two real data set are analyzed according to the proposal by highlighting its potential in detecting a hierarchy of disjoint principal components corresponding to different dimensions of a multidimensional phenomenon.

Chapter 6

Discussion

In the last decades, the complexity of real phenomena has grown. Consequently, the need for methodologies to study multidimensional phenomena in different fields like psychometrics, economics, social sciences, environment, etc., arises. This thesis intends to contribute to the current research on modeling multidimensional phenomena via a hierarchical dimensionality reduction approach by introducing a new class of simultaneous, exploratory and parsimonious models. The dissertation is composed of an introduction (Chapter 1), and four main chapters (Chapters 2-5) in which the proposals are illustrated. In particular, Chapter 1 introduces the reader into the problem under study by illustrating the existing methodologies proposed to deal with the construction of hierarchical structures of nested latent concepts, and the key notion of the thesis: ultrametricity. The latter arises in different fields, as mathematics, physics and taxonomy, thanks to its relation with nested partitions and tree-shape structures, and usually known in statistics in connection with distances in hierarchical cluster analysis. In Chapter 1, we introduce the little-known definition of an ultrametric matrix, which is related to a hierarchy of latent concepts and underlies the models proposed in Chapter 2 and 4. The nature of the relationships among levels in a hierarchy is discussed by presenting the difference between reflective and formative models.

Chapter 2 presents a new model, called Ultrametric Correlation Model (UCM), with the aim of detecting consistent latent concepts and their hierarchical relationships. UCM is an exploratory, parsimonious and simultaneous model that reconstructs a nonnegative correlation matrix via an ultrametric correlation one, and supplies a parsimonious representation of multidimensional phenomena through a partition of the observed variables into a reduced number of groups, each one associated with a latent concept. Two main features related to concepts are highlighted: the within-concept consistency and the between-concept correlation. A relationship between these features and the Cronbach's α (Cronbach, 1951), which is a well-known measure of internal consistency, is provided in the chapter. It has to be highlighted that a measure of internal consistency does not correspond to a measure of unidimensionality (homogeneity), even if the former is necessary for the latter (see Schmitt, 1996, for further details). UCM is developed in a reflective approach by assuming the existence of a general concept which accounts for the relationships among the more specific ones, and illustrated through two real data applications.

Chapter 3 discusses a comparison between the traditional, agglomerative hierarchical clustering algorithms and the model presented in Chapter 2. The former are usually implemented to build a hierarchy of units and associated with an ultrametric distance matrix. Therefore, one could assume to transform a correlation matrix into a distance one (Gordon, 1999), and then apply a hierarchical clustering algorithm to the latter. We compare UCM and four (agglomerative) hierarchical clustering methods through their application on the Holzinger data set, that represents a benchmark epitome for the study of a hierarchical structure underlying a multidimensional phenomenon. The data set is presented as a nonnegative correlation matrix of fourteen observed variables associated with four latent concepts. UCM turns out to be more suitable (efficient) in terms of hierarchy construction and goodness of fit than the hierarchical clustering algorithms, when dealing with variables. Moreover, UCM directly pinpoints a partition of the variable space in a reduced number of groups without suffering from the aggregation errors in the first levels of a complete hierarchy.

Chapter 4 presents an extension of the model developed in Chapter 2 to a generic covariance matrix with its implementation into a Gaussian Mixture Model (GMM). Specifically, we first extend the definition of an ultrametric matrix to an extended ultrametric matrix, which allows to include negative values by preserving the semi-definiteness of the matrix; then, we use this definition to model the covariance matrix of each component of a GMM. The proposal, called Gaussian Mixture Model with an Extended Ultrametric Covariance Structure (GMMEUCovS), is able to pinpoint a hierarchical structure on variables for each component of the GMM, thus identifying a different characterization of a multidimensional phenomenon for each component (cluster, subpopulation) of the mixture. It is worth noticing that the large number of parameters of a GMM is produced by the covariance matrices. In order to reduce this number, parsimonious parameterizations of the latter were proposed in literature, e.g., the eigen-decomposition (Banfield & Raftery, 1993) and the parsimonious GMMs based on mixtures of probabilistic principal component analyzers (Tipping & Bishop, 1999b, 1999a) and mixtures of factor analyzers (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000b; McLachlan, Peel, & Bean, 2003). GMMEUCovS defines a new parsimonious GMM since the ultrametric covariance structure reconstructs the relationships among variables with a limited number of parameters. The model proposed in Chapter 4 is illustrated via two real data examples. The first one concerns the study of well-being in the OECD countries. The application pinpoints a different characterization of this phenomenon in more developed and less developed economies, even if some similarities between the hierarchies over the eleven variables can be detected. The second example is the coffee data set on which GMMEUCovS is applied in order to both assess the performance of the proposal in recovering the true clustering structure, represented by the two varieties of beans (Arabica and Robusta), in comparison with Gaussian Parsimonious Clustering Models (GPCMs, Celeux & Govaert, 1995; Fraley & Raftery, 1998, 2002), Parsimonious Gaussian Mixture Models (PGMMs, McNicholas & Murphy, 2008, 2010) and High-Dimensional Data Clustering (HDDC, Bouveyron, Girard, & Schmid, 2007), and to identify a hierarchy composed also of discordant concepts, i.e., negative covariances among groups.

Chapter 5 illustrates a new simultaneous, exploratory and parsimonious model, called Hierarchical Disjoint Principal Component Analysis (HierDPCA), for hier-

archical dimensionality reduction. Principal Component Analysis (PCA) is often employed to obtain a dimensionality reduction of the variable space via a reduced set of components, but preserving the largest possible part of the total variance of the data. Nonetheless, PCA is not suitable to detect hierarchical relationships among variables. Contrariwise, HierDPCA aims at building a hierarchy of nested components associated with disjoint groups of observed variables, by improving their interpretation. Moreover, HierDPCA allows to choose the type of the relationship among components of two sequential levels, from the lowest upwards, by testing the component correlation per level and changing from a reflective to a formative approach when this correlation turns out to be not statistically significant. The goal of this proposal is in turn to build a hierarchical structure of nested latent concepts, even if this is not directly associated with an (extended) ultrametric matrix as the models illustrated in Chapter 2 and 4. Additionally, HierDPCA introduces the quantification of the latent concepts for each level of the hierarchy. The proposal is implemented on two real data examples, in an exploratory and mixed exploratory and confirmatory approach.

Appendices to Chapters 2, 4 and 5 provide supplementary materials relating to further properties of the proposed methodology, details on the estimates of the model parameters and proofs of some equations, respectively.

6.1 Further developments

The models introduced herein are all developed in the context of hierarchical dimensionality reduction and associated with *binary* tree (Gordon, 1987, 1999, among others). Indeed, they aim at building a hierarchy of nested variable partitions, whose groups are characterized by specific features (Chapter 2) and distinguished in heterogeneous populations (Chapter 4) or associated with unobserved variables (Chapter 5). Some further developments for each proposal of this thesis may be outlined.

Firstly, latent concepts identified by UCM may be quantified. For instance, components associated with variable groups may be computed by maximizing the explained variance for each level of the hierarchy, thus optimizing a common objective function between the models presented in Chapter 2 and 5, in a LS framework.

One of the main goals for future studies on GMMEUCovS presented in Chapter 4 is to constrain the parameters of EUCovS, i.e., $\mathbf{V}, \boldsymbol{\Sigma}_{\mathrm{V}}, \boldsymbol{\Sigma}_{\mathrm{W}}, \boldsymbol{\Sigma}_{\mathrm{B}}$, to be equal across and within clusters. This extension gives rise to a new class of parsimonious models that further reduces the number of parameters of GMMEUCovS, and allows an extensive comparison with GPCMs, PGMMs, HDDC also in terms of the model parsimony.

By considering the model presented in Chapter 5, it may be of interest to extend HierDPCA in a cluster analysis framework, i.e., by getting a simultaneous hierarchical parsimonious clustering of units, aggregated around centroids, and dimensionality reduction of variables, aggregated around components. This simultaneous model might also be seen as a hierarchical extension of the Clustering and Disjoint Principal Component Analysis proposed by Vichi and Saporta (2009), where the membership matrices of units and variables in two sequential levels of the corresponding hierarchy

are linked by the nestedness assumption.

The ultrametricity notion illustrated in this dissertation has a manifold of applications. One of the most important development of the ultrametric models presented herein is the definition of the Ultrametric Factor Analysis (UFA). As already mentioned in Chapter 1 (Section 1.1), Factor Analysis is one of the most used models to reconstruct relationships among variables via a reduced number of factors, but is not apt to pinpoint hierarchical structures over them. The idea we will work on is to define a new structure of the loading matrix $\mathbf{A}_{\mathrm{u}}$ such that $\boldsymbol{\Sigma}_{\mathrm{u}} = \mathbf{A}_{\mathrm{u}}\mathbf{A}_{\mathrm{u}}' + \boldsymbol{\Psi}_{\mathrm{u}}$ is an extended ultrametric covariance matrix. UFA will overcome the limitations of UCM and EUCovS, i.e., the lack of the latent concept quantification, by identifying a set of hierarchically nested partitions of variables into groups, each one associated with a factor, and detecting an ultrametric structure. UFA could also be implemented into a GMM in order to define an ultrametric extension of the mixture of factor analyzers with the edge of studying different characterizations of a multidimensional phenomenon in heterogeneous populations and, at the same time, quantifying its latent dimensions. Another interesting application of the ultrametric models presented in this thesis, that we would inspect in the future, is their use into (Gaussian) graphical models (Whittaker, 1990). The latter are very useful to model the relationships among random variables, especially when their number is very high, and the introduction of a (strict) extended ultrametric covariance matrix into these models can tackle the problem of the curse of dimensionality (Bellman, 1957) and the precision matrix, i.e., the inverse of the covariance matrix, estimation.

Finally, all the methodologies presented in this thesis aim at being applicable by other researchers, hence it is essential to let them available for free. For this reason, we will develop an R and/or MATLAB package containing all the routines used to implement the models proposed herein.

Appendix A

Appendix to Chapter 2 Relationship between $\mathbf{R_W}$, $\mathbf{R_B}$ and the Cronbach's $\boldsymbol{\alpha}$

As shown in Section 2.2, the standardized Cronbach's α of a variable group C_q $(q = 1, \ldots, Q)$, i.e. α_q^S, can be rewritten in terms of the within-concept consistency of C_q, i.e. ${}_W r_{qq}$. Furthermore, if we consider two groups of variables C_q and C_h $(h \neq q)$ merged together, the *total* within-concept consistency coefficient ${}_W r_{\text{tot}}$ of $C_q \cup C_h$ $(h \neq q)$ can be written as a function of the within-concept consistency of the two groups C_q and C_h, i.e., ${}_W r_{qq}$ and ${}_W r_{hh}$ respectively, and the correlation between them, i.e. ${}_B r_{qh}$. As a consequence, the standardized Cronbach's α of the merged group $C_q \cup C_h$ (α_{tot}^S) can in turn be rewritten as a function of the standardized Cronbach's α of the two groups C_q and C_h, i.e., α_q^S and α_h^S respectively, and the correlation between them. These two relationships are stated and proved as follows.

Let $\mathbf{R}$ be a nonnegative correlation matrix of order p. For simplicity, let us assumes that UCM estimates only two groups, i.e., $\widehat{\mathbf{V}}$ is a $(p \times 2)$ membership matrix where two groups of variables C_1 and C_2 are pinpointed. Let us suppose that the first J_1 variables belong to C_1 and the remaining $p - J_1$ to C_2, so that the variables which belong to the same group are contiguous in $\mathbf{R}$. Since the hierarchical nature of UCM, the researcher could be interested in evaluating the consistency of the broader dimension obtained by merging the two groups C_1 and C_2, i.e. $C_1 \cup C_2$.

Firstly, we can rewritten the diagonal elements of $\widehat{\mathbf{R}}_{\text{W}}$ estimated by Eq. (2.11) and the off-diagonal values of $\widehat{\mathbf{R}}_{\text{B}}$ estimated by Eq. (2.14) as follows:

$$ {}_W\hat{r}_{11} = \frac{\sum_{j=1}^{J_1} \sum_{\substack{l=1 \\ l \neq j}}^{J_1} r_{jl}}{J_1(J_1 - 1)}; \quad {}_W\hat{r}_{22} = \frac{\sum_{j=J_1+1}^{p} \sum_{\substack{l=J_1+1 \\ l \neq j}}^{p} r_{jl}}{(p - J_1)(p - J_1 - 1)}; \quad {}_B\hat{r}_{12} = \frac{\sum_{j=1}^{J_1} \sum_{\substack{l=J_1+1 \\ l \neq j}}^{p} r_{jl}}{J_1(p - J_1)}. $$

Thus, the total within-concept consistency coefficient of $C_1 \cup C_2$ turns out to be

$$ {}_W\hat{r}_{\text{tot}} = \frac{\sum_{j=1}^{p} \sum_{\substack{l=1 \\ l \neq j}}^{p} r_{jl}}{p(p-1)} = \frac{\sum_{j=1}^{p} \left[\sum_{l=1}^{p} r_{jl} - r_{jj} \right]}{p(p-1)} = \frac{\sum_{j=1}^{p} \sum_{l=1}^{p} r_{jl} - p}{p(p-1)} $$

$$= \frac{\sum_{j=1}^{J_1}\sum_{l=1}^{J_1} r_{jl} + \sum_{j=J_1+1}^{p}\sum_{l=J_1+1}^{p} r_{jl} + 2\sum_{j=1}^{J_1}\sum_{l=J_1+1}^{p} r_{jl} - p}{p(p-1)}$$

$$= \frac{\Big[\sum_{j=1}^{J_1}\sum_{l=1}^{J_1} r_{jl} - J_1\Big] + \Big[\sum_{j=J_1+1}^{p}\sum_{l=J_1+1}^{p} r_{jl} - (p-J_1)\Big] + 2\sum_{j=1}^{J_1}\sum_{l=J_1+1}^{p} r_{jl}}{p(p-1)}$$

$$= \frac{\sum_{j=1}^{J_1}\sum_{\substack{l=1\\ l\neq j}}^{J_1} r_{jl} + \sum_{j=J_1+1}^{p}\sum_{\substack{l=J_1+1\\ l\neq j}}^{p} r_{jl} + 2\sum_{j=1}^{J_1}\sum_{l=J_1+1}^{p} r_{jl}}{p(p-1)}$$

$$= \frac{J_1(J_1-1)_W\hat{r}_{11} + (p-J_1)(p-J_1-1)_W\hat{r}_{22} + 2\,J_1(p-J_1)_B\hat{r}_{12}}{p(p-1)}.$$

Considering the relationship with the within-concept consistency coefficient defined in Section 2.2, the standardized Cronbach's α of $C_1 \cup C_2$ can be rewritten as a function of $\hat{\alpha}_1^S$ and $\hat{\alpha}_2^S$ and ${}_B\hat{r}_{12}$ as follows

$$\begin{aligned}\hat{\alpha}_{\text{tot}}^S &= \frac{p\;{}_W\hat{r}_{\text{tot}}}{1+(p-1)_W\hat{r}_{\text{tot}}} = \frac{p}{1+(p-1)_W\hat{r}_{\text{tot}}}\,\frac{1}{p(p-1)}\Big[J_1(J_1-1)_W\hat{r}_{11}\\ &+ (p-J_1)(p-J_1-1)_W\hat{r}_{22} + 2\,J_1(p-J_1)_B\hat{r}_{12}\Big]\\ &= \frac{1}{(p-1)[1+(p-1)\,{}_W\hat{r}_{\text{tot}}]}\Big[(J_1-1)[1+(J_1-1)\,{}_W\hat{r}_{11}]\;\hat{\alpha}_1^S\\ &+ (p-J_1-1)[1+(p-J_1-1)\,{}_W\hat{r}_{22}]\;\hat{\alpha}_2^S + 2\,J_1(p-J_1)_B\hat{r}_{12}\Big].\end{aligned}$$

The above decomposition of ${}_W\hat{r}_{\text{tot}}$ and $\hat{\alpha}_{\text{tot}}^S$ can be easily generalized to all pairs of groups belonging to the hierarchical partition H_Q, which is obtained by applying UCM on $\mathbf{R}$ and then computing a row-column permutation such that the variables belonging to the same group turn out to be contiguous - with p replaced by $J_1 + J_2$.

Appendix B

Appendix to Chapter 4 Maximum likelihood estimates of the GMMEUCovS covariance structure

We provide here the detail of the GMMEUCovS covariance structure estimation. For the compactness of the equations, we substitute $\boldsymbol{\Sigma}_{\mathrm{u}_g}$ to its definition $\mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g')$. The following results are based on Lütkepohl (1996, Chapter 9) and Magnus and Neudecker (2007, Chapters 8 and 9).

The maximum likelihood estimate of $\boldsymbol{\Sigma}_{\mathrm{V}_g}, g = 1, \ldots, G$, is obtained by differentiating Eq. (4.18) with respect to $\boldsymbol{\Sigma}_{\mathrm{V}_g}$

$$\frac{\partial \ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = -\frac{n_g}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}}\Big[\log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|) + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\mathbf{S}_g)\Big] = -\frac{n_g}{2}\left[\frac{\partial \log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} + \frac{\partial \,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}}\right].$$

(A) $\dfrac{\partial \log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = \boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\dfrac{\partial \boldsymbol{\Sigma}_{\mathrm{u}_g}}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = \mathbf{V}_g'\Big[\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{V}_g$, remembering that $\mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g') = \mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\mathbf{V}_g' \odot \mathbf{I}_p$, where $\odot$ is the Hadamard product.

(B) $\dfrac{\partial \,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = -\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\dfrac{\partial \boldsymbol{\Sigma}_{\mathrm{u}_g}}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}}\mathbf{S}_g\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} = -\mathbf{V}_g'\Big[\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{S}_g\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\mathbf{V}_g$.

Given the other parameters of $\boldsymbol{\Sigma}_{\mathrm{u}_g}$, we equal to zero the partial derivative of $\ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})$ w.r.t. $\boldsymbol{\Sigma}_{\mathrm{V}_g}$ as follows

$$\begin{aligned}
&\frac{\partial \ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = 0 \Rightarrow \frac{\partial \log(|\boldsymbol{\Sigma}_{\mathrm{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} + \frac{\partial \,\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\mathrm{V}_g}} = 0 \\
&\Rightarrow \widehat{\mathbf{V}}_g'\Big[\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} \odot \mathbf{I}_p\Big]\widehat{\mathbf{V}}_g - \widehat{\mathbf{V}}_g'\Big[\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{S}_g\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1}\widehat{\mathbf{V}}_g = 0
\end{aligned}$$

which holds if and only if $\mathbf{S}_g\boldsymbol{\Sigma}_{\mathrm{u}_g}^{-1} = \mathbf{I}_p$. Thus,

$$\mathbf{S}_g = \boldsymbol{\Sigma}_{\mathrm{u}_g} \Rightarrow \mathbf{S}_g = \widehat{\mathbf{V}}_g(\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g} + \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g})\widehat{\mathbf{V}}_g' - \mathrm{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}\widehat{\mathbf{V}}_g') + \mathrm{diag}(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\widehat{\mathbf{V}}_g') \Rightarrow$$

$$\text{diag}(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{V}_g}\widehat{\mathbf{V}}_g') = \text{diag}(\mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g' + \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g') - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g') \Rightarrow$$
$$\text{diag}(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{V}_g}\widehat{\mathbf{V}}_g') = \text{diag}(\mathbf{S}_g) \Rightarrow$$
$$(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{V}_g}\widehat{\mathbf{V}}_g' \odot \mathbf{I}_p)\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\text{diag}(\mathbf{S}_g)\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} \Rightarrow$$
$$(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{V}_g})(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\text{diag}(\mathbf{S}_g)\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} \Rightarrow$$
$$\boldsymbol{\Sigma}_{\text{V}_g}(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\text{diag}(\mathbf{S}_g)\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} \Rightarrow$$
$$\widehat{\boldsymbol{\Sigma}}_{\text{V}_g} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\text{diag}(\mathbf{S}_g)\widehat{\mathbf{V}}_g,$$

where $\text{diag}(\mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g' + \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g') - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g') = \text{diag}(\mathbf{S}_g)$ since the diagonal of $\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g'$ is equal to the diagonal of $\text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{W}_g}\widehat{\mathbf{V}}_g')$, and the diagonal of $\text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g')$ is equal to zero.

The maximum likelihood estimate of $\boldsymbol{\Sigma}_{\text{W}_g}, g = 1, \ldots, G$, is obtained by differentiating Eq. (4.18) with respect to $\boldsymbol{\Sigma}_{\text{W}_g}$

$$\frac{\partial \ell_{\text{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = -\frac{n_g}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_{\text{W}_g}}\Big[\log(|\boldsymbol{\Sigma}_{\text{u}_g}|) + \text{tr}(\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g)\Big] = -\frac{n_g}{2}\left[\frac{\partial \log(|\boldsymbol{\Sigma}_{\text{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} + \frac{\partial \,\text{tr}(\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}}\right].$$

(A) $\dfrac{\partial \log(|\boldsymbol{\Sigma}_{\text{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = \boldsymbol{\Sigma}_{\text{u}_g}^{-1}\dfrac{\partial \boldsymbol{\Sigma}_{\text{u}_g}}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = \mathbf{V}_g'\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{V}_g - \mathbf{V}_g'\Big[\boldsymbol{\Sigma}_{\text{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{V}_g$, recalling that $\text{diag}(\mathbf{V}_g\,\boldsymbol{\Sigma}_{\text{W}_g}\mathbf{V}_g') = \mathbf{V}_g\boldsymbol{\Sigma}_{\text{W}_g}\mathbf{V}_g' \odot \mathbf{I}_p$.

(B) $\dfrac{\partial \,\text{tr}(\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = -\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\dfrac{\partial \boldsymbol{\Sigma}_{\text{u}_g}}{\partial \boldsymbol{\Sigma}_{\text{W}_g}}\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1} = -\mathbf{V}_g'\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{V}_g + \mathbf{V}_g'\Big[\boldsymbol{\Sigma}_{\text{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{V}_g$.

Given the other parameters of $\boldsymbol{\Sigma}_{\text{u}_g}$, we equal to zero the partial derivative of $\ell_{\text{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})$ w.r.t. $\boldsymbol{\Sigma}_{\text{W}_g}$ as follows

$$\frac{\partial \ell_{\text{H}}(\widehat{\mathbf{W}}, \boldsymbol{\Psi})}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = 0 \Rightarrow \frac{\partial \log(|\boldsymbol{\Sigma}_{\text{u}_g}|)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} + \frac{\partial \,\text{tr}(\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g)}{\partial \boldsymbol{\Sigma}_{\text{W}_g}} = 0$$
$$\Rightarrow \widehat{\mathbf{V}}_g'\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\widehat{\mathbf{V}}_g - \widehat{\mathbf{V}}_g'\Big[\boldsymbol{\Sigma}_{\text{u}_g}^{-1} \odot \mathbf{I}_p\Big]\widehat{\mathbf{V}}_g - \widehat{\mathbf{V}}_g'\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\widehat{\mathbf{V}}_g + \widehat{\mathbf{V}}_g'\Big[\boldsymbol{\Sigma}_{\text{u}_g}^{-1} \odot \mathbf{I}_p\Big]\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1}\widehat{\mathbf{V}}_g = 0$$

which holds if and only if $\mathbf{S}_g\boldsymbol{\Sigma}_{\text{u}_g}^{-1} = \mathbf{I}_p$. Thus,

$$\mathbf{S}_g = \boldsymbol{\Sigma}_{\text{u}_g} \Rightarrow \mathbf{S}_g = \widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g' + \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g' - \text{diag}(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g') + \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g') \Rightarrow$$
$$\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g' - \widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g' \odot \mathbf{I}_p = \mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g' - \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g') \Rightarrow$$
$$\boldsymbol{\Sigma}_{\text{W}_g} - (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'(\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g' \odot \mathbf{I}_p)\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} = (\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1}\widehat{\mathbf{V}}_g'\Big[\mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g'$$
$$- \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g')\Big]\widehat{\mathbf{V}}_g(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^{-1} \Rightarrow$$
$$\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g}\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g - \widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g} = \widehat{\mathbf{V}}_g'\Big[\mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g' - \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g')\Big]\widehat{\mathbf{V}}_g \Rightarrow$$
$$\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g} - \widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g\boldsymbol{\Sigma}_{\text{W}_g} = \widehat{\mathbf{V}}_g'\Big[\mathbf{S}_g - \widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{B}_g}\widehat{\mathbf{V}}_g' - \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g')\Big]\widehat{\mathbf{V}}_g \Rightarrow$$
$$\widehat{\boldsymbol{\Sigma}}_{\text{W}_g} = [(\widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g)^2 - \widehat{\mathbf{V}}_g'\widehat{\mathbf{V}}_g]^{-1}\text{diag}\Big[\widehat{\mathbf{V}}_g'\Big(\mathbf{S}_g - \text{diag}(\widehat{\mathbf{V}}_g\widehat{\boldsymbol{\Sigma}}_{\text{V}_g}\widehat{\mathbf{V}}_g')\Big)\widehat{\mathbf{V}}_g\Big].$$

The maximum likelihood estimate of $\mathbf{\Sigma}_{\mathrm{B}_g}, g = 1, \ldots, G$, is obtained by differentiating Eq. (4.18) with respect to $\mathbf{\Sigma}_{\mathrm{B}_g}$

$$\frac{\partial \ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \mathbf{\Psi})}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = -\frac{n_g}{2} \frac{\partial}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} \Big[\log(|\mathbf{\Sigma}_{\mathrm{u}_g}|) + \mathrm{tr}(\mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g) \Big] = -\frac{n_g}{2} \left[\frac{\partial \log(|\mathbf{\Sigma}_{\mathrm{u}_g}|)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} + \frac{\partial \, \mathrm{tr}(\mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} \right].$$

$$\text{(A)} \quad \frac{\partial \log(|\mathbf{\Sigma}_{\mathrm{u}_g}|)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \frac{\partial \mathbf{\Sigma}_{\mathrm{u}_g}}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = \mathbf{V}_g' \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{V}_g.$$

$$\text{(B)} \quad \frac{\partial \, \mathrm{tr}(\mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = -\mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \frac{\partial \mathbf{\Sigma}_{\mathrm{u}_g}}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} \mathbf{S}_g \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} = -\mathbf{V}_g' \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{V}_g.$$

Given the other parameters of $\mathbf{\Sigma}_{\mathrm{u}_g}$, we equal to zero the partial derivative of $\ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \mathbf{\Psi})$ w.r.t. $\mathbf{\Sigma}_{\mathrm{B}_g}$ as follows

$$\begin{aligned} &\frac{\partial \ell_{\mathrm{H}}(\widehat{\mathbf{W}}, \mathbf{\Psi})}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = 0 \Rightarrow \frac{\partial \log(|\mathbf{\Sigma}_{\mathrm{u}_g}|)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} + \frac{\partial \, \mathrm{tr}(\mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g)}{\partial \mathbf{\Sigma}_{\mathrm{B}_g}} = 0 \\ &\Rightarrow \mathbf{V}_g' \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{V}_g - \mathbf{V}_g' \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{S}_g \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} \mathbf{V}_g = 0 \end{aligned}$$

which holds if and only if $\mathbf{S}_g \mathbf{\Sigma}_{\mathrm{u}_g}^{-1} = \mathbf{I}_p$. Thus,

$$\begin{aligned} &\mathbf{S}_g = \mathbf{\Sigma}_{\mathrm{u}_g} \Rightarrow \mathbf{S}_g = \widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g' + \widehat{\mathbf{V}}_g \mathbf{\Sigma}_{\mathrm{B}} \widehat{\mathbf{V}}_g' - \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g') + \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{V}_g} \widehat{\mathbf{V}}_g') \Rightarrow \\ &\widehat{\mathbf{V}}_g \mathbf{\Sigma}_{\mathrm{B}} \widehat{\mathbf{V}}_g' = \mathbf{S}_g - \widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g' + \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g') - \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{V}_g} \widehat{\mathbf{V}}_g') \Rightarrow \\ &\widetilde{\mathbf{\Sigma}}_{\mathrm{B}_g} = \widehat{\mathbf{V}}_g^{+} \Big[\mathbf{S}_g - \widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g' + \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{W}_g} \widehat{\mathbf{V}}_g') - \mathrm{diag}(\widehat{\mathbf{V}}_g \widehat{\mathbf{\Sigma}}_{\mathrm{V}_g} \widehat{\mathbf{V}}_g') \Big] (\widehat{\mathbf{V}}_g')^{+} \Rightarrow \\ &\widetilde{\mathbf{\Sigma}}_{\mathrm{B}_g} = \widehat{\mathbf{V}}_g^{+} \mathbf{S}_g (\widehat{\mathbf{V}}_g')^{+}. \end{aligned}$$

Appendix C

Appendix to Chapter 5 Proofs

In this Appendix the proofs of (5.11) and (5.13) defined in Section 5.3.1 are provided.

Eq. (5.11) can be proved by recalling $\mathbf{Y}_q = \mathbf{X}\mathbf{B}_q\mathbf{V}_q$ for $q = M, \ldots, Q$, the trace additive and invariance under scale permutation properties and constraint (5.6) of HierDPCA.

Proof.

$$\sum_{q=M}^{Q} \|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + \sum_{q=M}^{Q} \|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 = \sum_{q=M}^{Q} (\|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + \|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2)$$

$$= \sum_{q=M}^{Q} \left\{ \mathrm{tr}[(\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q)'(\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q)] + \mathrm{tr}[(\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q)'(\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q)] \right\}$$

$$= \sum_{q=M}^{Q} \left[\mathrm{tr}(\mathbf{X}'\mathbf{X}) - \mathrm{tr}(\mathbf{X}'\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q) - \mathrm{tr}(\mathbf{B}_q\mathbf{V}_q\mathbf{Y}_q'\mathbf{X}) + 2\,\mathrm{tr}(\mathbf{B}_q\mathbf{V}_q\mathbf{Y}_q'\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q) \right]$$

$$= \sum_{q=M}^{Q} [\mathrm{tr}(\mathbf{X}'\mathbf{X}) - \mathrm{tr}(\mathbf{X}'\mathbf{X}\mathbf{B}_q\mathbf{V}_q\mathbf{V}_q'\mathbf{B}_q) - \mathrm{tr}(\mathbf{B}_q\mathbf{V}_q\mathbf{V}_q'\mathbf{B}_q\mathbf{X}'\mathbf{X})$$
$$+ 2\,\mathrm{tr}(\mathbf{V}_q'\mathbf{B}_q\mathbf{X}'\mathbf{X}\mathbf{B}_q\mathbf{V}_q\mathbf{V}_q'\mathbf{B}_q\mathbf{B}_q\mathbf{V}_q)]$$

$$= \sum_{q=M}^{Q} \mathrm{tr}(\mathbf{X}'\mathbf{X}) = \sum_{q=M}^{Q} \|\mathbf{X}\|^2 = (Q - M + 1)\|\mathbf{X}\|^2.$$

□

The proof of (5.13) is provided as follows by remembering that the difference between two nested partitions $\mathbf{V}_q$ and $\mathbf{V}_{q-1}$ is written in (5.12).

Proof.

$$\sum_{q=1}^{Q} \|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 = \sum_{q=1}^{Q} \|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + Q\,\|\mathbf{X}\|^2 - Q\,\|\mathbf{X}\|^2$$

$$= -\left[\sum_{q=1}^{Q} (\|\mathbf{X}\|^2 - \|\mathbf{X} - \mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2) \right] + Q\,\|\mathbf{X}\|^2 \overset{\text{Eq. (5.11)}}{=} -\sum_{q=1}^{Q} \|\mathbf{Y}_q\mathbf{V}_q'\mathbf{B}_q\|^2 + Q\|\mathbf{X}\|^2$$

$$
\begin{aligned}
&= -\sum_{q=1}^{Q-1} \|\mathbf{Y}_q \mathbf{V}_q' \mathbf{B}_q\|^2 - \|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 + Q \|\mathbf{X}\|^2 \\
&= -\sum_{q=1}^{Q-1} \|\mathbf{Y}_q \mathbf{V}_q' \mathbf{B}_q\|^2 - \|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 + Q \left(\|\mathbf{X} - \mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 + \|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2\right) \\
&= Q \|\mathbf{X} - \mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 - \sum_{q=1}^{Q-1} \|\mathbf{Y}_q \mathbf{V}_q' \mathbf{B}_q\|^2 + (Q-1) \|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 \\
&= Q \|\mathbf{X} - \mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 - \|\mathbf{Y}_1 \mathbf{V}_1' \mathbf{B}_1\|^2 - \|\mathbf{Y}_2 \mathbf{V}_2' \mathbf{B}_2\|^2 - ... - \|\mathbf{Y}_{Q-1} \mathbf{V}_{Q-1}' \mathbf{B}_{Q-1}\|^2 \\
&\quad + (Q-1) \|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 \\
&= Q \|\mathbf{X} - \mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 + \|\mathbf{Y}_2 \mathbf{V}_2' \mathbf{B}_2\|^2 - \|\mathbf{Y}_1 \mathbf{V}_1' \mathbf{B}_1\|^2 + 2 \left(\|\mathbf{Y}_3 \mathbf{V}_3' \mathbf{B}_3\|^2 - \|\mathbf{Y}_2 \mathbf{V}_2' \mathbf{B}_2\|^2\right) \\
&\quad + ... + (Q-1) \left(\|\mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 - \|\mathbf{Y}_{Q-1} \mathbf{V}_{Q-1}' \mathbf{B}_{Q-1}\|^2\right) \\
&= Q \|\mathbf{X} - \mathbf{Y}_Q \mathbf{V}_Q' \mathbf{B}_Q\|^2 + \sum_{q=2}^{Q} (q-1)\, I_d(\mathbf{V}_q, \mathbf{V}_{q-1}).
\end{aligned}
$$

□

Bibliography

Adachi, K., & Trendafilov, N. T. (2018). Sparsest factor analysis for clustering variables: A matrix decomposition approach. *Advances in Data Analysis and Classification*, *12*(3), 559–585.

Anderson, T. W., & Rubin, H. (1956). Statistical inferences in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (5th ed., pp. 111–150). University of California Press, Berkeley.

Baek, J., McLachlan, G. J., & Flack, L. K. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1298–1309.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Bechtoldt, H. (1961). An empirical study of the factor analysis stability hypothesis. *Psychometrika*, *26*(4), 405–432.

Becker, J. M., Klein, K., & Wetzels, M. (2012). Hierarchical latent variable models in PLS-SEM: Guidelines for using reflective-formative type models. *Long Range Planning*, *45*(5), 359–394.

Becker, W., Dominguez-Torreiro, M., Neves, A. R., Tacao Moura, C. J., & Saisana, M. (2018). *Exploring ASEM Sustainable Connectivity – What brings Asia and Europe together?* (PUBSY JRC112998). Luxembourg: Publications Office of the European Union. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC112998/asem-report_online.pdf

Bellman, R. (1957). *Dynamic programming*. Princeton University Press, Princeton.

Benzécri, J. P. (1973). *L'analyse des données, tome I: La taxonomie*. Dunod, Paris.

Bergé, L., Bouveyron, C., & Girard, S. (2012). HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, *46*(6), 1–29.

— (2019). *HDclassif: High dimensional supervised classification and clustering* [R package version 2.2.0]. https://cran.r-project.org/web/packages/HDclassif

Bezdek, J. C., Hathaway, R. J., Howard, R. E., Wilson, C. A., & Windham, M. P. (1987). Local convergence analysis of a grouped variable version of coordinate descent. *Journal of Optimization Theory and Applications*, *54*(3), 471–477.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Biernacki, C., Celeux, G., Govaert, G., & Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, *51*(2), 587–600.

Blalock, H. M. (1964). *Causal inferences in nonexperimental research*. Norton, New York.

Bobisud, H. M., & Bobisud, L. E. (1972). A metric for classifications. *Taxon*, *21*(5/6), 607–613.

Bollen, K. A. (2001). Indicator: Methodology. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 7282–7287). Elsevier Science, Oxford.

— (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, *35*(2), 359–372.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological methods*, *16*(3), 265–284.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*(100), 52–78.

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

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis*, *52*(1), 502–519.

Brouwer, A. E., & Haemers, W. H. (2012). *Spectra of graphs*. Springer, New York.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322.

Cadima, J., & Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, *22*(2), 203–214.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, *48*(2), 305–308.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, Cambridge.

Cattell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika*, *12*(3), 197–220.

— (1978a). Higher-order factors: Models and formulas. *The scientific use of factor analysis in behavioral and life sciences* (pp. 192–228). Springer, Boston.

— (1978b). *The scientific use of factor analysis in behavioral and life sciences*. Plenum, New York.

Cavicchia, C., & Vichi, M. (2021). Second-order disjoint factor analysis. *Psychometrika*. https://doi.org/10.1007/s11336-021-09799-6

Cavicchia, C., Vichi, M., & Zaccaria, G. (2019). Dimensionality reduction via hierarchical factorial structure. In G. C. Porzio, F. Greselin, & S. Balzano (Eds.), *CLADAG 2019, 11-13 September 2019, Cassino: Book of Short Papers* (pp. 116–119). Centro Editoriale di Ateneo Università di Cassino e del Lazio Meridionale, Cassino.

Cavicchia, C., Vichi, M., & Zaccaria, G. (2020a). Exploring hierarchical concepts: Theoretical and application comparison. In T. Imaizumi, A. Nakayama, & S. Yokoyama (Eds.), *Advanced Studies in Behaviormetrics and Data Science. Behaviormetrics: Quantitative Approaches to Human Behavior* (pp. 315–328). Springer, Singapore.

— (2020b). The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification*, *14*(4), 837–853.

— (2021). Hierarchical disjoint principal component analysis. *Manuscript submitted for publication.*

— (2022). Gaussian mixture model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification. Accepted.*

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793.

Chen, S. X., Zhang, L., & Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, *105*(490), 810–819.

Cliff, A. D., Haggett, P., Smallman-Raynor, M. R., Stroup, D. F., & Williamson, G. D. (1995). The application of multidimensional scaling methods to epidemiological data. *Statistical Methods in Medical Research*, *4*(2), 102–123.

Contreras, P., & Murtagh, F. (2015). Hierarchical clustering. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of Cluster Analysis* (1st ed., pp. 103–123). Chapman & Hall/CRC, New York.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual: Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI).* Psychological Assessment Resources, Odessa.

Cramer, H. (1946). *Mathematical methods of statistics.* Princeton University Press, Princeton.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., & Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, *49*(3), 434–446.

de Raad, B., & Mlačić, B. (2015). Big five factor model, theory and structure. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 559–566). Elsevier, Oxford.

Dellacherie, C., Martínez, S., & San Martín, J. (2014). *Inverse M-matrices and ultrametric matrices.* Lecture Notes in Mathematics, Springer International Publishing.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *39*(1), 1–38.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the big five predict conformity: Are there neuroses of health? *Personality and Individual Differences*, *33*(4), 533–552.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440.

— (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246–1256.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370–388.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of the relationship between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Escofier, B., & Pagès, J. (1983). *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation* (1st ed.). Dunod, Paris.

— (1994). Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis*, *18*(1), 121–140.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley, New York.

Eysenck, H. J. (1970). *The structure of human personality* (3rd ed.). Methuen, London.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2015). The Five Factor Model of personality and evaluation of drug consumption risk. *arXiv*. https://arxiv.org/abs/1506.06297

Ferrara, C., Martella, F., & Vichi, M. (2016). Dimensions of well-being and their statistical measurements. In G. Alleva & A. Giommi (Eds.), *Topics in Theoretical and Applied Statistics* (pp. 85–99). Springer International Publishing.

— (2019). Probabilistic disjoint principal component analysis. *Multivariate Behavioral Research*, *54*(1), 47–61.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, *2*(3/4), 282–285.

Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, *12*, 18–65.

Fop, M., Murphy, T. B., & Scrucca, L. (2019). Model-based clustering with sparse covariance matrices. *Statistics and Computing*, *29*(4), 791–819.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, and density estimation. *The Computer Journal*, *41*(8), 578–588.

— (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, *16*(2), 297–306.

— (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Galimberti, G., & Soffritti, G. (2013). Using conditional independence for parsimonious model-based Gaussian clustering. *Statistics and Computing*, *23*(5), 625–638.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Allyn & Bacon, Boston.

Gerschgorin, S. (1931). Uber die abgrenzung der eigenwerte einer matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika*, *7*, 749–754.

Ghahramani, Z., & Hinton, G. H. (1997). The EM algorithm for factor analyzers. *Technical report CRG-TR-96-1, University of Toronto, Toronto.*

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley, New York.

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57–68.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229.

— (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*(1), 26–42.

— (2006). Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40*(4), 347–358.

Gordon, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A-G*, *150*(2), 119–137.

— (1999). *Classification* (2nd ed.). Monographs on Statistics & Applied Probability, Chapman & Hall/CRC, Boca Raton.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Erlbaum, Hillsdale.

Götz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (pp. 691–711). Springer.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3/4), 325–338.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, *62*(320), 1140–1158.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108.

Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, *4*(2), 53–56.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *3*, 81–117.

Holzinger, K. J. (1944). A simple method of factor analysis. *Psychometrika*, *9*(4), 257–262.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Horn, R. A., & Johnson, C. R. (2013). *Matrix analysis* (2nd ed.). Cambridge University Press, Cambridge.

Horst, P. (1965). *Factor analysis of data matrices*. Holt Rinehart; Winston, New York.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *ournal of Educational Psychology*, *24*(6), 417–441, 498–520.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Jambu, M. (1978). *Classification authomatique pour l'analyse des données, tome 1*. Dunod, Paris.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*(4), 537–549.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*(3), 442–454.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer-Verlag, New York.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, *12*(3), 531–547.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, *31*(2), 165–178.

— (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

— (1970). A general method for analysis of covariance structure. *Biometrika*, *57*(2), 239–251.

— (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*(4), 443–477.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639.

Jöreskog, K. G., & Sörbom, D. (1982). Recent developments in Structural Equation Modeling. *Journal of Marketing Research*, *19*(4), 404–416.

Jöreskog, K. G., & Wold, H. (1982). The ML and PLS technique for modeling with latent variables: Historical and comparative aspects. In K. G. Jöreskog & H. Wold (Eds.), *Systems Under Indirect Observation, Part I* (pp. 263–270). North-Holland, Amsterdam.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kano, Y. (1997). Exploratory factor analysis with a common factor with two indicators. *Behaviormetrika*, *24*(2), 129–145.

Keribiin, C. (1998). Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, *326*(2), 243–248.

— (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, *62*(1), 49–66.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press, New York.

Kriv́anek, M., & Morávek, J. (1986). NP-hard problems in hierarchical-tree clustering. *Acta Informatica*, *23*(3), 311–323.

Lance, G. N., & Williams, W. T. (1966). A generalized sorting strategy for computer classifications. *Nature*, *212*, 218.

— (1967). A general theory of classificatory sorting strategy: 1. Hierarchical systems. *The Computer Journal*, *9*(4), 373–380.

Langrognet, F., Lebret, R., Poli, C., Iovleff, S., Auder, B., Bhatia, P., Echenim, A., Biernacki, C., Celeux, G., Govaert, G., & Grimonprez, Q. (2020). *Rmixmod: Classification with mixture modelling* [R package version 2.1.5]. https://cran.r-project.org/web/packages/Rmixmod

Le Dien, S., & Pagès, J. (2003). Analyse factorielle multiple hiérarchique. *Revue de statistique appliquée*, *51*(2), 47–73.

Liu, X., Zhu, X. H., Qiu, P., & Chen, W. (2012). A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. *Journal of Neuroscience Methods*, *211*(1), 94–102.

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5th ed.). Routledge, New York.

Lütkepohl, H. (1996). *Handbook of matrices*. Wiley, Chichester.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Statistics* (pp. 281–297). University of California Press, Berkeley.

Magnus, R. J., & Neudecker, H. (2007). *Matrix differential calculus with application in statistics and econometrics* (3rd ed.). Wiley, Chichester.

Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, *19*(2), 354–376.

Martínez, S., Michon, G., & San Martín, J. (1994). Inverses of ultrametric matrices are of Stieltjes types. *SIAM Journal on Matrix Analysis and Applications*, *15*(1), 98–106.

Mazziotta, M., & Pareto, A. (2019). Use and misuse of PCA for measuring well-being. *Social Indicators Research*, *142*(2), 451–476.

McGinnis, J. M., & Foege, W. (1993). Actual causes of death in the united states. *JAMA*, *270*(18), 2207–2212.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York.

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). Wiley, Hoboken.

McLachlan, G. J., & Peel, D. (2000a). *Finite mixture models*. Wiley, New York.

— (2000b). Mixtures of factor analyzers. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning* (pp. 599–606). Morgan Kaufmann, San Francisco.

McLachlan, G. J., Peel, D., & Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics Data Analysis*, *41*(3), 379–388.

McMorris, F. R., Meronk, D. B., & Neumann, D. A. (1983). A view of some consensus methods for tree. In J. Felsenstein (Ed.), *Numerical Taxonomy* (pp. 122–126). Springer-Verlag, Berlin.

McNicholas, P. D. (2016). *Mixture model-based classification*. Chapman & Hall/CRC, Boca Raton.

McNicholas, P. D., ElSherbiny, A., Jampani, K. R., McDaid, A. F., Murphy, T. B., & Banks, L. (2019). *pgmm: Parsimonious Gaussian mixture models* [R package version 1.2.4]. https://cran.r-project.org/web/packages/pgmm/

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, *18*(3), 285–296.

McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, *26*(21), 2705–2712.

McQuitty, L. L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, *20*(1), 55–67.

Mézard, M., Parisi, G., Sourlas, N., Toulouse, G., & Virasoro, M. (1984). Nature of the spin-glass phase. *Physical Review Letter*, *52*, 1156–1159.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika*, *44*(3), 343–346.

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(3), 193–211.

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, *41*(6), 1213–1233.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill, New York.

OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing, Paris.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343–355.

Parisi, G., & Ricci-Tersenghi, F. (1999). On the origin of ultrametricity. *33*(1), 113–129.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195–239.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). Wiley, New York.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74.

Riani, M., Cerioli, A., Perrotta, D., & Torti, F. (2015). Simulating mixtures of multivariate data with fixed cluster overlap in FSDA library. *Advances in Data Analysis and Classification*, *9*(4), 461–481.

Riani, M., Perrotta, D., & Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, *116*, 17–32.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51–67.

Rocci, R., & Vichi, M. (2008). Two-mode multi-partitioning. *Computational Statistics & Data Analysis*, *52*(4), 1984–2003.

Schikhof, W. H. (1985). *Ultrametric calculus: An introduction to p-Adic analysis*. Cambridge University Press, Cambridge.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factorial solutions. *Psychometrika*, *22*(1), 53–61.

Schmitt, N. (1996). Uses and abuses of coefficient Alpha. *Psychological Assessment*, *8*(4), 350–353.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). Wiley, New York.

Soffritti, G. (1999). Hierarchical clustering of variables: A comparison among strategies of analysis. *Communications in Statistics - Simulation and Computation*, *28*(4), 977–999.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Spearman, C. E. (1904). "General intelligence,' objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293.

— (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.

— (1927). *The abilities of man: Their nature and measurement*. Macmillan, New York.

Strauss, J. S., Bartko, J. J., & Carpenter, W. T. (1973). The use of clustering techniques for the classification of psychiatric patients. *The British Journal of Psychiatry*, *122*(570), 531–540.

Streuli, H. (1973). Der heutige stand der kaffeechemie. *Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemisrty, Bogata*, 61–72.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, *48*(1), 159–205.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

Thompson, G. H. (1948). *The factorial analysis of human ability*. Houghton Mifflin, New York.

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press, Chicago.

Tipping, M. E., & Bishop, C. M. (1999a). Mixtures of probabilistic principal component analysers. *Neural Computation*, *11*(2), 443–482.

— (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *61*(3), 611–622.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture models*. Wiley, Chichester.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley, New York.

Undheim, J. O., & Gustafsson, J. E. (1988). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of Linear Structural Relations (LISREL). *Multivariate Behavioral Research*, *22*(2), 149–171.

UNECE. (2017). *A set of key climate change-related statistics using the system of environmental-economic accounting* (tech. rep.). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2016/mtg/19-Report_on_climate_indicators_final.pdf

Vichi, M. (2008). Fitting semiparametric clustering models to dissimilarity data. *Advances in Data Analysis and Classification*, *2*(2), 121–161.

Vichi, M. (2017). Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification*, *11*(3), 563–591.

Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, *53*(8), 3194–3208.

Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics - Simulation and Computation*, *32*(4), 1131–1150.

Vines, S. K. (2000). Simple principal components. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *49*(4), 441–451.

Ward, J. H. (1963). Hierarchical grouping to optimize and objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Warrens, M. J. (2015). Some relationships between Cronbach's alpha and the Spearman-Brown formula. *Journal of Classification*, *32*(1), 127–137.

Wherry, R. J. (1959). Hierarchical factorial solutions without rotation. *Psychometrika*, *24*(1), 45–51.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, London.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah (Ed.), *Multivariate Analysis* (pp. 391–420). Academic Press, New York.

— (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under Indirect Observation: Part II* (pp. 1–54). North-Holland, Amsterdam.

— (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, Vol. 6* (pp. 581–591). Wiley, New York.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *23*(3), 557–585.

Yao, W. (2015). Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation*, *85*(5), 1000–1012.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128.

Zaccaria, G., & Vichi, M. (2020). Exploring drug consumption via an ultrametric correlation matrix. In A. Pollice, N. Salvati, & F. Schirripa Spagnolo (Eds.), *Book of Short Papers SIS 2020* (pp. 372–377). Pearson.

Zangwill, W. I. (1969). *Nonlinear programming: A unified approach*. Prentice-Hall, Englewood Cliffs.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.