SAPIENZA
UNIVERSITÀ DI ROMA

Hierarchical latent variable models for dimensionality reduction: an application on composite indicators

Scuola di Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXXII Ciclo

Candidate

Carlo Cavicchia

ID number 1348873

Thesis Advisor

Prof. Maurizio Vichi

Thesis submitted in 2019

Thesis defended on February 17, 2020
in front of a Board of Examiners composed by:

Prof. Roberto Rocci
Prof. Bruno Toaldo
Prof. Simone Vantini

Hierarchical latent variable models for dimensionality reduction: an application on composite indicators
Ph.D. thesis. Sapienza – University of Rome

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

Version: February 17, 2020

Author's email: carlo.cavicchia@uniroma1.it

A quel bottone di rosa.

Acknowledgments

I never expected to start a Ph.D. but the last three years have been the most intense and powerful experience of my life. Most importantly, I have never been alone. I am grateful to all my colleagues and the professors of the Department of Statistical Sciences who have been extraordinarily important for me. Let me just cite two wonderful girls who made these last years special, Tullia and Silvia. Since you have introduced me in Stanza 41, you have been more than simple friends. The list of all the people who contributed to make my Ph.D. experience so great is too long, so I can simply thank each and all of you.

Since my Bachelor's degree, I was guided and supported by Prof. Maurizio Vichi. I would like to express all my gratitude to you for everything you did for me, both as an academic advisor and a man: you fully contributed to my personal growth.

This thesis is a collection of different works developed throughout these years. I therefore want to wholeheartedly thank who collaborated with me, namely Prof. Pasquale Sarnacchiaro for all his support and help, and my colleague and friend Giorgia, for always putting up with me.

Amongst my lifetime friends, I pick just two special guys who supported me day by day, Luca and Alessio. Thank you for all the moments together and for pushing me into the craziest thing of my life: ASD Valcanneto Futsal.

I would like to dedicate not only this thesis but also all my Ph.D. to my parents, Susanna and Antonio, for giving me your best every single day. Even if what I do is not clear yet, you always work in order to guarantee my well being. Matteo, this work is also for you, for having inspired my life from the very start.

I should thank too many people who contributed to make me who I am now, especially in these last three years. So then again, simply a heartfelt thanks to all of you.

Last, all my work is for you Silvia, you know. Lysm.

Contents

List of Figures

List of Tables

Abstract

This thesis is devoted to the development of new hierarchical latent variable models for *Dimensionality Reduction* with a specific focus on the construction of Composite Indicators (CIs). Since our society is producing a huge quantity of data, the construction of model-based CIs represents an interesting and still open methodological challenge. This dissertation is motivated by the necessity to provide model-based CIs which are built according to a *statistical* approach avoiding subjective choices (e.g., normative weights), therefore, the new insights and proposals hope to represent a contribution to the current literature. This thesis provides an introduction to CIs, a brief review of the most used methods in Multidimensional Data Analysis framework, a discussion about *measurement models* and methodological proposals to model latent concepts. Factor Analysis and its hierarchical extensions have been introduced in order to set the starting point of the analysis. A first proposal represents a new latent factor model that could be used for building CIs, it aims to investigate the hierarchical structure of the data in order to define two levels of CIs. The model, named Hierarchical Disjoint Non-Negative Factor Analysis is composed of two novelties: a model which is the two level hierarchical extension of FA and its disjoint extension with non-negative loadings. The latter model is enriched by considerations about the CIs used for tracking coherent policy conclusions. A set of features, properties and rules useful to build "good" CIs have been presented and explained. The last proposal in the thesis represents a new model for positive data correlation matrices which aims to detect reliable concepts and to build the hierarchy from them to the most general one. The proposed models are illustrated both via simulation studies and real data applications, to analyze their performances and abilities. In particular, the main application in this thesis regards the construction of a hierarchically aggregated index for the multidimensional phenomenon *Waste Management* in European Union. *Waste Management* is becoming even more important for its impact on human-being's lives, and many data have been produced about it, therefore the construction of a CI able to reduce its dimensionality and to highlight the main dimensions of it has a extraordinary usefulness in order to provide support to EU countries' action and policies.

Acknowledgments
A special thank goes to the reviewers, Marco Fattore and Andrea Cerioli, for their precious comments and feedback.

Chapter 1

Introduction

1.1 Introduction

In the last years, with the data revolution and the use of new technologies, phenomena are frequently described by a huge quantity of information useful for making strategic decisions. A priority for policymakers is having simple statistical tools useful to synthesize data. Such tools are represented by Composite Indicators (CIs). In policy-making, it is important underline that CIs should never be meant as a goal per se, they are starting points for public interest discussion and they detect the trend of the phenomenon under study.

According to OECD Glossary of Statistical terms, a Composite Indicator (CI) is formed when individual indicators (i.e., manifest or observed variables) are compiled into a single non-observable index, on the basis of an underlying model for the multidimensional concept that is being measured. Multidimensional concepts are also defined as latent constructs, like well-being (Stiglitz, Sen, and Fitoussi 2009), development, poverty, environmental conditions, etc., which cannot be adequately captured by a single observed indicator, and a set of individual indicators is needed. A CI, for the JRC - Composite Indicator Research Group (JRC-COIN)[1], is a mathematical combination, or aggregation (weighting), of the manifest indicators (MIs) that generally have different unit of measurement and can be differently combined (Nardo, Saisana, Saltelli, Tarantola, et al. 2005; OECD-JRC 2008). CIs are easier to interpret than trying to find a common trend in many separated indicators. Therefore, CIs are increasingly used for bench-marking countries' performances and the methodological challenges raise a series of technical issues that, if not adequately addressed, can lead to CIs being misinterpreted or manipulated. Yet doubts are often raised about the robustness of the resulting countries' rankings and about the significance of the associated policy message. According to Nardo, Saisana, Saltelli, and Tarantola (2005), arguments in favour of CIs are that they: summarise complex multi-dimensional phenomena in view of supporting decision-makers; provide an overall picture of the separate indicators thus simplifying the interpretation; facilitate the task of ranking countries on complex issues; attract public interest by providing a summary figure to use for assessing performances; reduce the size of a list of indicators. Whereas, contrary arguments are that CIs: may send misleading

[1]The Joint Research Centre (JRC) is the European Commission's science and knowledge service which employs scientists to carry out research in order to provide independent scientific advice and support to EU policy. The JRC is a Directorate-General of the European Commission under the responsibility of Tibor Navracsics, Commissioner for Education, Culture, Youth Sport.

and/or non-robust policy messages if they are poorly constructed or misinterpreted, and thus, they may invite politicians to draw simplistic policy conclusions; may involve subjective decisions in the CI construction phase that can influence the final indicator. Even if the issue of the construction of composite indicators has been discussed many times by many authors with different approaches, CIs do not have a good reputation because the statistical methods used are not always mathematically rigorous and often they are based on theories which do not seem to have a solid foundation (Mazziotta and Pareto 2013). Often CIs are computed as the weighted mean of the MIs (e.g., Multidimensional Poverty Index, Alkire and Foster (2011), Alkire, Foster, et al. (2015), and UNDP (2018)), where the weights represent the importance of each MI and they are chosen by the researcher subjectively or according to a known theory (i.e., normative weights). In particular, many authors do not appreciate CIs determined by subjective weights on the MIs because it can lead to misinterpretation of the results (Nardo, Saisana, Saltelli, and Tarantola 2005; Saltelli et al. 2004).

The definition of CI includes the notions of *concept* and *model* that need additional clarifications. The *concept* in the CI can be measured only indirectly and it relates to a fact, in general to a phenomenon of interest that, for its nature, cannot be sufficiently described by a single MI[2]. This is a typical situation in multivariate statistics, where the simultaneous observations and analysis of more than one indicator is needed to understand the phenomenon being studied. Concepts such as for example poverty, human development, gender equality, well-being cannot be satisfactorily represented by individual indicators and therefore need to be described by several indicators. The *model* is the second notion considered to characterize a CI. It is required to simplify and synthesize the complexity of the reality by means of a mathematically-formalized reconstruction of the observed data and their main relations. Models are needed for investigating phenomena in order to achieve awareness driven by the empirical evidence. In the development process used to specify the appropriate CI model for the studied phenomenon, three integral parts are needed: the variable selection, to properly characterize the phenomenon under study, the model selection from a set of candidate models and the model evaluation to assess the performances of the CI.

In order to evaluate and/or confirm the quality of building approach, even when CIs are built as average of MIs as it is the case of Human Development Index (HDI, (Alkire 2010)) and MPI, the researcher needs a set of properties able to test the CI's capacity of reconstructing the manifest data or its ability of forming the studied concept. It is important to notice how the choices of the researcher might affect the result of the analysis. The use of a modeling approach, including selection and assessment, guarantees the researcher who wishes to construct the CI based on a scientific method, to build it on a body of techniques with desirable properties to guarantee that the CI will work appropriately in different conditions and with past and future data.

Our society is producing a huge quantity of data in every moment, a model-based approach able to reduce the dimensionality of data has a crucial applicability and an extraordinary usefulness. The methodological challenges of constructing well-done CIs in order to acquire new knowledge based on empirical evidence is even more important. Models are crucial in the development process used to specify the appropriate CI for the studied

[2] A MI is a variable that indicates something, thus it should provide information about the phenomenon, otherwise it might be meant as a simple statistic. If a concept has only one dimension with one indicator, the concept is practically equivalent to a variable.

phenomenon.

The use of a model-based CI classifies this approach as *statistical* because hypotheses formulated in the specification of the CI are empirically tested. The methodology wishes to be objective and transparent, avoiding - as far as possible - the use of arbitrary choices that cannot be tested.

In this thesis, we suppose that the model for CIs is a *Dimensionality Reduction* (DR) model with a hierarchical structure that goes from the original MIs to the final General Composite Indicator (GCI), passing through a reduced set of Specific Composite Indicators (SCIs), i.e., dimensions, which measure specific concepts describing the main components of the phenomenon under study. CIs are usually computed starting from a reduced number of MIs (e.g., MPI and HDI) whereas it is more and more necessary to have tools able to handle a large number of MIs. The proposal should be considered into the *Dimensionality Reduction* framework for its capacity of compiling big quantity of information into a single CI, by way of many steps of aggregation.

The process of *Dimensionality Reduction* produce, for sure, a loss of information and the presentation of the single CI without additional information may denote a too simple picture of the complex phenomenon that it synthesizes. Indeed, the goal of the analysis is to show all the process of reduction, that is, the MIs (e.g., scoreboard, portfolio of variables) used to describe the phenomenon, the SCIs illustrating their synthesized dimensions and representing the most relevant components of the phenomenon, and the final GCI. Together with these information, it is useful to show the entire hierarchical process that goes from the MIs to the GCI, that is, the relationships, internal to the indicators (manifest or composite) used to describe the phenomenon. In addition, it would be necessary to add at each step of the reduction a model fit or measures of reliability and unidimensionality that allow the researcher to assess if the model correctly describes and synthesizes the data. In this way, the researchers have all the information at hand and can decide the most appropriate level of description to use for their purposes. For example, to measure the dimensions of well-being is appropriate to stop the analysis at the SCIs' level such as *Material Living Conditions* and *Quality of Life* (Stiglitz, Sen, and Fitoussi 2009).

Thus, in this thesis we will study methodologies of *Dimensionality Reduction* to build model-based CIs, underlying a body of properties and features that a good CI should respect in order to avoid poorly constructed indexes that induce simplistic and misleading policy conclusions. This thesis is framed in *Latent variable models* context because it refers to indicators (i.e., latent) that are not directly observable but they affect the manifest ones.

1.2 Hierarchical structure

CIs usually have a hierarchical structure, in fact, they identify a multidimensional concept that summarizes other multidimensional concepts, each one characterized by other nested multidimensional concepts or by subsets of MIs. For example, the CI of well-being defined in the Better Life Initiative of OECD (2001) is seen as a mathematical combination of *Material Living Conditions* (MLC) and *Quality of Life* (QL). The MLC is certainly a multidimensional concept characterized by *housing*, *jobs* and *income*. QL is the second multidimensional concept described by *community*, *education*, *safety*, *health*, *environment*, *civic engagements*, *life satisfaction* and *work-life balance*. Each one of these last dimensions of the well-being are also multidimensional concepts described by several indicators or by

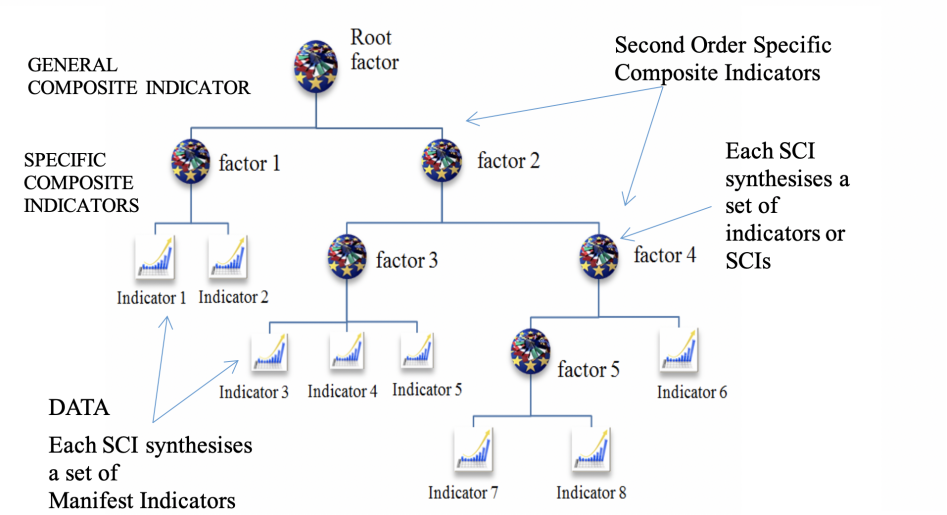


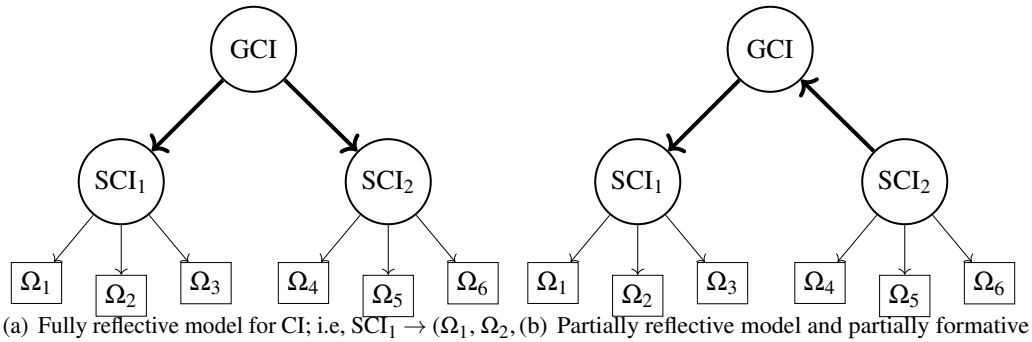
Figure 1.1. Hierarchical CI. This is the graphical representation considered in the Structural Equation Modeling

other multidimensional concepts. For example, *heath* can be described by *child health* and *reproductive health*. A field of study where the hierarchical approach is used to model and build CIs is psychometrics. For instance, hierarchical models have been used frequently in the cognitive domain (J. Carroll 1993; Horn and Cattell 1966; Horn and Cattell 1982); in the personality domain (Cattell 1957; Cattell 1966; Eysenck and Himmelweit 1947; Eysenck 1967; DeYoung, Peterson, and Higgins 2002; Digman 1997), and for modeling anxiety (R.E. Zinbarg and Barlow 1996; R.E. Zinbarg, Barlow, and Brown 1997).

It is worth to observe that these examples refer to the definition of *Latent variable models*, the aim is representing individual characteristics that cannot be directly observed (e.g., well-being, quality of life, personality, intelligence, etc.).

In general, the initial set of MIs is aggregated into classes (i.e., sets, clusters), each one representing a multidimensional concept summarised by a theoretical indicator, also named latent variable or factor. MIs within classes are supposed highly correlated so that the common information can be summarised by factors. Classes of measured indicators are continuously aggregated and represented by new factors until when all initial MIs are together into a single class and represented by a general factor that corresponds to the CI. In other terms, data useful to accurately describe the studied phenomenon are “best” reconstructed at different levels of syntheses by means of DR, considering some Specific Composite Indicators as representative of subsets of MIs. The process of reduction has a hierarchically nested form, as shown in Figure 1.1, with a graphical configuration of a tree. Leaves represent the MIs, while internal nodes denote SCIs (i.e., latent variables, factors, linear or non-linear combinations of MIs) that synthesize common information in the data. In order to represent CIs and their hierarchies, the path diagram (i.e. tree graph) is usually used: the GCI is at the top of the graph, as root of the tree and the leaves are the MIs. MIs are described by rectangles, while SCIs or GCI are represented by circles, which are seen as latent factors corresponding to theoretical concepts.

The aim of the analysis is the definition of nested latent concepts from the MIs to the GCI in order to reconstruct the hierarchical structure of the data correlation matrix. In the



(a) Fully reflective model for CI; i.e, $SCI_1 \rightarrow (\Omega_1, \Omega_2, \Omega_3$, $SCI_2 \rightarrow (\Omega_4, \Omega_5, \Omega_6)$, GCI $\rightarrow (SCI_1, SCI_2)$.

(b) Partially reflective model and partially formative model for CI; i.e, $SCI_1 \rightarrow (\Omega_1, \Omega_2, \Omega_3$, $SCI_2 \rightarrow (\Omega_4, \Omega_5, \Omega_6)$, GCI $\rightarrow SCI_1$, $SCI_2 \rightarrow$ GCI.

Figure 1.2. Reflective and Formative Models

simplest case, a two-levels structure is hypothesised taking into account only the levels of SCIs and the GCI, however when we are interested to investigate about all the possible aggregations from SCIs level to the root of the hierarchy it is possible to detect all the intermediate concepts via an ultrametric correlation matrix.

1.3 Measurement Models and Multidimensional Data Analysis

Measurement models refer to the implicit or explicit models that relate the latent variable to its MIs (Bollen 2001). Two different approaches might be distinguished with respect to the nature of the relationships between MIs and latent constructs (i.e., SCIs and GCI) which formally describe the *measurement model*, and thus, define the direction of the these relationships (Blalock 1964; Bollen and Bauldry 2011).

An arrow starting from a CI and pointing to a MI describes a casual relationship from CI to MI (CI $\rightarrow$ MI). In other terms, the CI reconstructs (predicts, explains) the MIs. The relation of the type CI $\rightarrow$ MI is called *reflective*, i.e., the CI explains the correlation among MIs contributing to specify the *concept* associated with the CI. Blalock (1964) referred to this kind of indicators as *effect* indicators. This is the typical relation in Factor Analysis (FA, (T. Anderson and Rubin 1956; Horst 1965)) where a set of latent variables is used to reconstruct the MIs. So, latent constructs are determinant of the MIs, it is a *top-down* approach. In Figure 1.2(a) a fully reflective GCI is specified with two SCIs. If a SCI is pointing to a set of MIs it means that the *concept* corresponding to the SCI reconstructs the common variance of the MIs. On the arrow, a loading (i.e., weight) is generally reported which generally represents the correlation between SCI and MI. The example in Figure 1.3 can help the reader to understand the reflective relation.

The relation of the type MI $\rightarrow$ CI is *formative* and describes a causal relation from MI to CI. Blalock (1964) explained this kind of relation saying that the MIs influence the latent variable rather than the reverse, and some refer to these as *formative* indicators or *casual* indicators. A formative relation can also have the form SCI $\rightarrow$ GCI from SCI to GCI. Indicators (MIs or SCIs) that form a concept with a formative relation generally are not correlated to each other. This is the typical relation in multiple and multivariate regression. In this case the links between MIs and contructs are measured through a *bottom-up* approach. In Figure 1.2(b), a GCI in part reflective and in part formative is shown. It is important

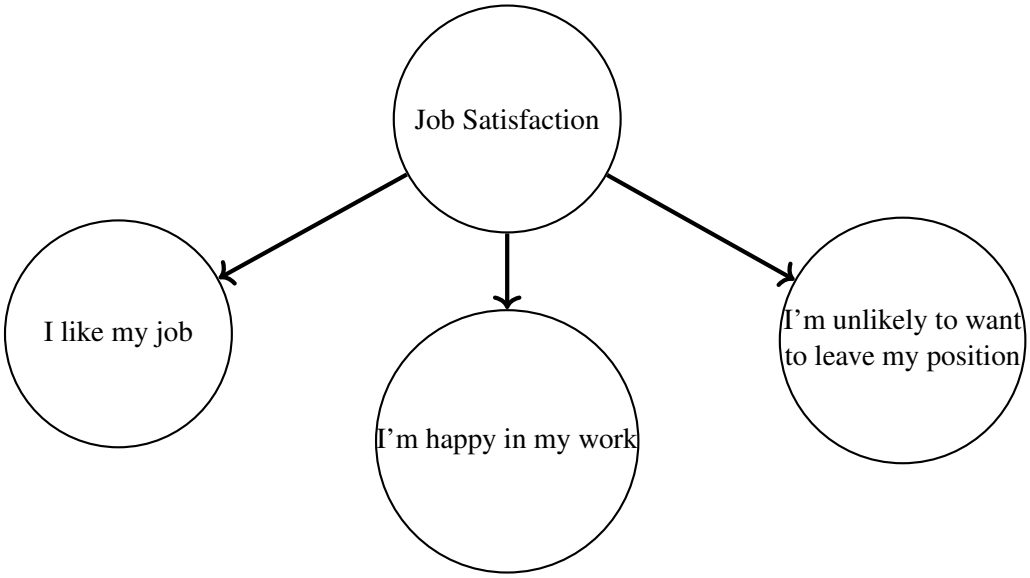


Figure 1.3. Example of Reflective construct

to highlight that an entirely *formative* model is not identified (Edwards 2011), such that without additional information it is impossible to find unique values for the parameters in the model and hence cannot assess indicator validity (Bollen 2011). In order to solve this problem, the alternative model named *MIMIC* has been considered (Hauser and Goldberger 1971; Jöreskog and Goldberger 1975) by taking into account the latent variable as effect of some MIs and, at the same time, cause of some others. Edwards (2011) proposed a critique about *formative* models by arguing that *formative* measures are often based on expressed beliefs about constructs and MIs that do not respect condition of unidimensionality and reliability, and that are difficult to defend.

The CI may reconstruct other CIs in the hierarchy ($SCI_1 \rightarrow SCI_2$). In this case, a second order or in general a higher order SCI is used. Also, in this case the weight on the arrow representing the formative relation corresponds to the correlation between the involved indicators. An example can help the reader to understand the formative relation (see Figure 1.4).

So, in a reflective relation, MIs depend on the latent indicators (i.e. SCIs); whereas in a formative relation, MIs affect the latent indicators (i.e. SCIs).

Within the methodological literature, the discussion between reflective and formative is crucial and far from a possible solution. *Reflective* indicators are appropriate for many situations in psychological measurement, but they are not appropriate for all situations. Several authors (Blalock 1964; Land 1970; Bollen 1984; Bollen 1989; Bollen and Lennox 1991; Hayduk 1987; MacCallum and Browne 1993) have underlined that some MIs should be, for their nature, more appropriately treated as casual rather than effects of the latent construct. It is important to remark that the choice between the two kinds of model should not depend directly on the researcher, but it should depend exclusively on the nature of the phenomenon and its definition (Diamantopoulos and Siguaw 2006; Maggino 2017). Moreover, it is not possible to test a priori the nature of the phenomenon in order to establish if it is *reflective* or *formative* (Maggino and Zumbo 2012). Often researchers do not consider that a latent indicator may be *causal* and treat all MIs as *effect* indicators. This is not an

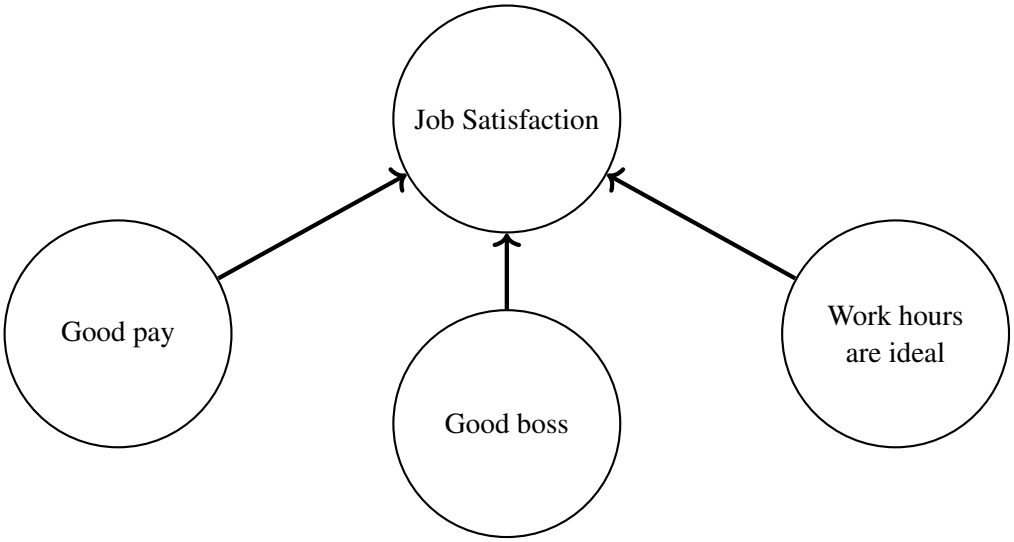


Figure 1.4. Example of Formative construct

inconsequential decision since *causal* and *effect* indicators have different properties and their incorrect classification can bias estimates from a model (Bollen and Lennox 1991). In order to better understand the difference between the two aforementioned approaches of the measurement model, see Jarvis, MacKenzie, and Podsakoff (2003) and Edwards and R.P. Bagozzi (2000).

In other cases, the indicators of a psychological construct might be a mixture of effect and causal indicators. For instance, as reported by Bollen and Ting (2000), Bollen and Lennox (1991) had considered that the Center for Epidemiological Studies Depression Scale (CES-D; Radloff (1977)) has some effect indicators (e.g., "I felt depressed" and "I felt sad") and some causal indicators (e.g., "I felt lonely").

Another situation is considered by Cavicchia, Vichi, and Zaccaria (2020a), the hierarchical aggregations of MIs start as *reflective* and become *formative* at a certain level of the hierarchy itself. Thus, the type of the latent constructs definition changes at first level where no correlation exists among constructs (the correlations are tested according to Chen, Zhang, and Zhong (2010)). This consideration can be explained by saying that until correlations between latent constructs are statistically significant, the latent constructs reflect the same upper latent concepts, when the correlations stop being statistically significant, the latent constructs form the upper latent concepts.

The hierarchical path model associated with the process to define the CI highlights the different levels of description of the phenomenon associated with the CI can be derived, starting from MIs (scoreboard) and considering one or more sets of SCIs to arrive to the final GCI. All the hierarchy is given; thus, the researcher will have the possibility to use only the scoreboards reorganized according to the given model, or additionally will consider the different levels of reduction given by the SCIs and the GCI.

Formally, the CI is a model-based indicator if the CI assumes a statistical model of the type

$$\mathbf{X} = \mathbf{CI}_M + \mathbf{E} \tag{1.1}$$

where $\mathbf{X}$ is the matrix of the MIs, $\mathbf{CI}_M$ is the matrix associated with the CI model, i.e.,

the matrix of the MIs reconstructed by the CI model, and **E** is the residual, that is, the difference between **X** and $\mathbf{CI}_M$. **E** corresponds to the part of data not explained by the CI model and generally, it includes the measurement error (i.e., a random error, sampling bias) always present in the measurement of the MIs. When the model-based CI is used, it can be estimated by considering an estimation method which minimizes a badness-of-fit or discrepancy function, i.e., a function of the difference between the MIs and the fitted CI (i.e., predicted MIs).

Multidimensional Data Analysis (MDA) approaches, like FA or Principal Component Analysis (PCA) (Pearson 1901; Hotelling 1933), are considered valid in order to build CIs for multidimensional phenomena. In PCA and FA, the weights are computed by taking into account the statistical relations among MIs. They represent reflective models thus they can be used whether all the indicators refer to a general latent concept. Another widely used methods for the construction of CIs are Structural Equation Modeling (SEM) (Jöreskog 1970; Bollen 1989; Kaplan 2000), they are used in order to build a flexible system of composite indicators able to model causal relations among them. Trinchera and Russolillo (2010) provided a useful review on the use of SEM to build CIs.

In OECD (2004) and OECD-JRC (2008), the FA is considered as a weighting method in order to combine MIs. FA has the advantage to define loadings that best reconstruct the MIs according to the estimation method chosen avoiding the subjective choice of a system of weights by the researcher. However, some choices are needed, for instance the choice of type of rotation (varimax, equimax, orthomax, etc.) (Browne 2001) and the size of the soft threshold[3]. The concept of oblique (i.e., non-orthogonal) factors in order to get more interpretable simple structure has been subject of study of many authors (Cattell and Khanna 1977; Harman 1976; Jennrich and Sampson 1966; Clarkson and Jennrich 1988). Specifically, strategies have been proposed to rotate factors so as to best represent "clusters" of MIs, without the constraint of orthogonality of factors. However, the oblique factors produced by such rotations are often not easily interpreted.

When the model presents a hierarchical structure, FA is not an appropriate method because it is neither able to get a partition of the MIs nor to get the hierarchical structure of the concepts, therefore a valid alternative method to construct a CI into the factor analysis framework is the Hierarchical Confirmatory Factor Analysis (HCFA, Holzinger (1944), Jöreskog (1966), Jöreskog (1969), Jöreskog (1978), and Jöreskog (1979)), which can be used by first theorizing the correlation structure between factors and the related MIs. However, the prior knowledge on both the number factors and the most relevant relations between MIs and factors represents a limitation in the construction of the CI in the situation, quite realistic, when the researcher has not an exact theory in mind or when the theory demonstrates to be erroneous in some parts or at least in its empirical application. Hierarchical Factor Analysis (HFA) strategy, proposed by Thompson (1951) and Schmid and Leiman (1957), and detailed by Wherry (1959), Wherry (1975), and Wherry (1984), first identifies clusters of MIs and rotates axes through those clusters; next it computes the correlation matrix of the SCIs (oblique factors). The correlation matrix of oblique factors is further factor-analyzed to yield a set of orthogonal factors that divide the variability in the MIs into that due to shared or common variance (secondary factors), and unique variance due to the clusters of similar MIs

[3]Soft thresholding is a raw solution to improve interpretability. It consists of ignoring MIs with small-magnitude loadings, and therefore, each factor is approximated by a linear combination of only a subset of MIs, namely those with not small-magnitude loadings. This is also named "Truncated Factor Analysis". The size of the threshold is subjective.

in the analysis (primary factors). Also Le Dien and Pages (2003) proposed a hierarchical extension of multiple factor analysis (Escofier and Pages 1998), called *Hierarchical Multiple Factor Analysis*, which aims to provide graphical displays of the MIs groups for each level of the hierarchy or superimposed representations of partial individuals.

The distinction between confirmatory and exploratory approach is explained in the context of FA to highlight the difference between Confirmatory FA and Exploratory FA by Ullman (2006).

In this situation, the researcher can assume that each manifest indicator is mainly related with a latent construct only, although this relation is not assumed known a priori. In this case an unknown Simple Structure Model (SSM) is hypothesised for the data.

In recent years several methods have been proposed in order to provide a clearer interpretation of the factors (or components) by means of a sparse structure of the loading matrix, i.e., assuming that some loadings are equal to zero (d'Aspremont et al. 2007; Grbovic, Dance, and Vucetic 2012; Jolliffe, Trendafilov, and Uddin 2003; Zou, Hastie, and Tibshirani 2006; Ferrara, Martella, and Vichi 2016; Ferrara, Martella, and Vichi 2018). Ferrara, Martella, and Vichi (2016) proposed a constrained PCA called *Disjoint Principal Component Analysis* (DPCA), as particular case of *Clustering Disjoint Principal Component Analysis* (Vichi and Saporta 2009), which allows to identify the sparsest classification of variables via principal components with maximum variance, summarising information held by the corresponding subset of variables. Whereas, Disjoint Factor Analysis (DFA) (Vichi 2017) has been proposed to identify the best SSM for the data in FA framework. Maximum Likelihood Estimation (MLE) of DFA allows to make inference on the number of factors, on the relations between MIs and factors (loadings), and to assess the validity of the SSM for the observed data. In addition, DFA allows to identify the eventual presence of relevant cross-loadings (Vichi 2017) and hence, identify multi-factor indicators, that tend to complicate the naming and interpretation of factors. Finally, DFA permits to fix one or more indicators that must specify the same factor. This allows the researcher to hypothesise even a part of a theory that he/she has in mind and believes is important to be included in the model in order to be successively confirmed.

However, DFA is not appropriate to identify the hierarchical structure of factors. This is so because DFA assumes that factors are orthogonal and therefore it is not supposed observable a hierarchical structure of factors and it is not hypothesised that factors are mutually related and share some common information that can be summarised by the general indicator through reflective relationships. The aggregation of uncorrelated factors is own of formative models.

In order to solve this problem, it is assumed that correlation structure in the data has a hierarchical form, albeit unknown, with at least two levels; the first denoted root or general level, representing the CI and at least a single specific level indicating a reduced set of multidimensional concepts described by disjoint classes of MIs. In this thesis we will concentrate on the two level hierarchical structure which is generally the most proposed in CFA. Thus, it is supposed that the model for CIs is a *Dimensionality Reduction* model with a hierarchical structure that goes from the original MIs to the final General Composite Indicator (GCI), passing through a reduced set of Specific Composite Indicators (SCIs), which are dimensions that measure specific concepts describing the main components of the phenomenon taken into study.

1.4 Non-aggregative approach

Although this dissertation hinges on a specific reflective hierarchical approach to build CIs which characterizes the general CI as a weighted sum of factor scores (i.e. aggregative approach), in the interests of providing a complete introduction of CIs, a brief description of the non-aggregative approach is presented in this Section.

Rankings, which are very common, cannot be aggregated through linear combinations, averages or other functionals, designed for numerical variables. Therefore, in order to treat ordinal attributes as numerical, before aggregation they are often transformed into numerical scores, through more or less sophisticated scaling tools (Fattore 2017). However Madden (2010) showed evidence that such procedures may lead to controversial results and questioned whether it is correct to force concepts which are naturally conceived in ordinal terms into numerical settings. In the non-aggregative approach, thus, the synthesis should not be conducted through a weighted sum of MIs and, in general, it should not involve any explicit function of them. Consequently, the need for properly managing rankings has led to the development of modern tools and methodologies into the framework of partial order theory to prioritize observations by many different ordinal characteristics (Brüggemann and Patil 2011).

Partial order theory has been considered a fundamental branch of mathematics, only of theoretical interest, for a long time. However, in the last years, its effectiveness as a tool for data analysis is being increasingly realized and many applications of partially ordered sets to real problems in statistics and applied sciences have appeared, hence demonstrating their insightfulness and usefulness. Main examples pertain to the analysis of complex and multidimensional systems of ordinal data and to problems of multi-criteria decision making, both extremely relevant in socio-economic and environmental sciences (Brüggemann and Patil 2011; Fattore and Maggino 2014).

Following the definition in (Davey and Priestley 2002), a partially ordered set (or a POSET) $P = (X, <)$ is a set X (called the ground set) equipped with a partial order relation $<$, that is, a binary relation satisfying the properties of reflexivity, antisymmetry and transitivity. Let $v_1, \ldots, v_k$ be k ordinal evaluation variables, profiles (each possible sequence of scores on $v_1, \ldots, v_k$) can be (partially) ordered in a natural way by a given dominance criterion. Poset tools, thus, allow to describe and to exploit the relational structure of the data, so as to compute evaluation scores in purely ordinal terms and avoiding any aggregation of variables. Since ordinal phenomena cannot be measured against an absolute scale, the evaluation scores to be assigned to the profiles are computed comparing them against some reference profiles selected as benchmarks.

Eventually, the evaluation is addressed in terms of multidimensional comparisons among profiles rather than through attribute score aggregations. In other words, the central idea in partial order is comparison, and the most general outcome is a net of relations between observations according to their indicator values, respecting the ranking aim (Carlsen and Brüggemann 2013). This makes it not needed to scale ordinal attributes into numerical variables, overcoming the issues of aggregative procedures and counting approaches. The evaluation procedure is fuzzy by nature, and it accounts for both vagueness and intensity of deprivation, thus producing a comprehensive set of synthetic indicators for policy-makers (Fattore 2016).

From this moment onwards, we will focus on hierarchical latent variable models for *Dimensionality Reduction*, as this thesis presents hierarchical factorial models to build CIs.

1.5 Chapter Summaries

Chapter 2 presents a new latent factor model that could be used for modeling CIs named *Hierarchical Disjoint Non-Negative Factor Analysis* (HDNFA). The proposal is grounded on the common definition of latent variable model and it is an exploratory methodology to model the hierarchical structure of factors, which is supposed in part or totally unknown, identifying a reduced set of factors (i.e., SCIs) each one related to a disjoint subset of MIs. Each subset of MIs must be internally consistent and reliable, that is, MIs are concordant with the related SCI and loadings must be positive. Properties are discussed for HDNFA. Furthermore, an application to optimally identify the dimensions of well-being is used to illustrate the characteristics of the model. The proposed model has been implemented in a MATLAB routine.

The contents of Chapter 2 have been developed with Prof. Maurizio Vichi, and are reported in a paper which has been submitted for publication and it is currently under first revision in a international journal, see Cavicchia and Vichi (2020a).

Chapter 3 introduces a model-based approach for construction of CIs as an alternative method of estimation for the model HDNFA, and a body of properties and features that a good CI shoud have. In order to assess some well-know CI such as HDI and MPI, a comparison with the proposed model-based CI based on the same properties has been presented. Chapter 3 provides useful tools for studying CIs which are of interest in many applicative fields.

The contents of Chapter 3 have been developed with Prof. Maurizio Vichi, and are reported in a paper which has been submitted for publication and it is currently under second revision in a international journal, see Cavicchia and Vichi (2020b).

Chapter 4 presents a hierarchical CI for *Waste Management* in Europe by using the HDFNA model. The model has detected three main reliable aspects of the phenomenon of interest. The analysis, from the MIs selection to the definition of the CI, represents an application of the model presented in Chapter 2 and it has a crucial importance for supporting policy-making. *Waste Management* is increasing its importance through the years and the number of statistics dedicated to it is getting larger and larger. The construction of a CI which highlights the main aspects of it has been a interesting challenge with the aim to provide support to EU countries' action and policies.

The contents of Chapter 4 have been developed with Prof. Pasquale Sarnacchiaro and Prof. Maurizio Vichi, and are reported in a paper which has been submitted for publication and it is currently under first revision in a international journal, see Cavicchia, Sarnacchiaro, and Vichi (2020).

Chapter 5 introduces a novelty model that aims to reconstruct the data correlation matrix through an ultrametric correlation matrix which is able to pinpoint the hierarchical nature of the phenomenon under study. The model is able to detect consistent latent concepts and the relationship among them starting their correlation structure, as well as it provides a measure for the internal consistency of concepts and a measure for the correlation among concepts. An application on Bechtoldt dataset is proposed to show the performance of the model. The proposed model has been implemented in a MATLAB routine.

The contents of Chapter 5 have been developed with Prof. Maurizio Vichi and Dr. Giorgia Zaccaria, and are reported in a paper which has been submitted for publication and it is currently under second revision in a international journal, see Cavicchia, Vichi, and Zaccaria (2020b).

Chapter 2

Hierarchical Disjoint Non-Negative Factor Analysis

2.1 Introduction

Hierarchical Disjoint Non-Negative Factor Analysis (HDNFA) is a new latent factor model that could be used for modeling Composite Indicators (CIs). CIs, in general, are multidimensional concepts described by at least a theoretical construct (i.e., factor) which is related to a cluster of MIs. Frequently CIs have a hierarchical structure, i.e., they are characterized by a set of specific factors (i.e., SCIs, dimensions) each one corresponding to disjoint, or nested clusters of MIs. Hierarchical Confirmatory Non-Negative Factor Analysis can be used to assess the hierarchical structure of the CIs. Hierarchical Disjoint Factor Analysis is grounded on the common definition of a latent variable model and it is presented as an exploratory methodology to model the hierarchical structure of factors, which is supposed in part or totally unknown, identifying a reduced set of factors each one related to a disjoint cluster of MIs.

Furthermore, in a definition of a CI, each cluster of MIs (i.e., measured variables, observed variables) must be internally consistent and reliable, that is, MIs related to the factor measure "consistently" a unique theoretical construct. This implies that MIs are concordant with the related factor and loadings must be positive. This last requirement is included as a constraint in the new methodology that for this reason is named Hierarchical Disjoint Non-Negative Factor Analysis.

Properties are discussed for HDNFA. The model and its algorithm has also the option to constraint a MI to load on a pre-specified factor in order to hypothesize, a priori, some relations between MIs and factors that the researcher wishes to fix. An application to optimally identify the dimensions of well-being is used to illustrate the characteristics of the new methodology. A final discussion completes the paper.

The paper is organised as follows. two level Hierarchical Factor Analysis (HFA) model is proposed in Section 2.2. Section 2.3 includes an overview of disjoint models. Section 2.4 introduces the non-negative constraints of the factors necessary to specify consistent latent indicators. A simulation study is considered in Section 2.5. Section 2.6 shows an application of the HNDFA. A final discussion completes the paper in Section 2.7.

It is important to underline that, in this paper, latent constructs are computed as factors, thus the term "factors" and "composite indicators" (i.e., SCIs and GCI) are used with the

same meaning.

2.2 Two level Hierarchical Factor Analysis

Let $\mathbf{x}$ be the $(J \times 1)$ random vector, representing the generic multivariate observation, with mean vector $\mu_{\mathbf{x}} = [\mu_1, \ldots, \mu_J]'$, and J-dimensional variance-covariance matrix $\Sigma_{\mathbf{x}}$. The two level Hierarchical Factor Analysis is a new factorial model that considers two typologies of latent unknown constructs: H specific factors and a single (nested) general factor identified by the two simultaneous models (equations)

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{e}_{\mathbf{x}} \tag{2.1}$$

$$\mathbf{y} = \mathbf{c}\mathbf{g} + \mathbf{e}_{\mathbf{y}} \tag{2.2}$$

where $\mathbf{A}$ is the $(J \times H)$ matrix of unknown specific factors loadings and $\mathbf{e}_{\mathbf{x}}$ is a $(J \times 1)$ random vector of errors.

It is assumed that $\mathbf{y} \sim N_H(\mathbf{0}, \Sigma_{\mathbf{y}})$ where $\Sigma_{\mathbf{y}}$ is the correlation matrix of the specific factors, and $\mathbf{e}_{\mathbf{x}} \sim N_J(\mathbf{0}, \Psi_{\mathbf{x}})$, where $\text{cov}(\mathbf{e}_{\mathbf{x}}) = \Psi_{\mathbf{x}}$ is the J dimensional diagonal positive definite variance-covariance matrix of the error of model 2.1 and $\text{cov}(\mathbf{e}_{\mathbf{x}}, \mathbf{y}) = \Sigma_{\mathbf{y}\mathbf{e}_{\mathbf{x}}} = 0$.

Furthermore, $\mathbf{g}$ is the random general factor with mean 0 and variance $\sigma^2_{\mathbf{g}} = 1$, denoting the composite indicator related to a reduced set of specific factors and $\mathbf{c}$ is the $(H \times 1)$ vector of unknown general factor loadings. In addition, $\mathbf{e}_{\mathbf{y}}$ is a non-observable $(H \times 1)$ random vector of errors. It is assumed that $\mathbf{g} \sim N(0,1)$ and $\mathbf{e}_{\mathbf{y}} \sim N_H(0, \Psi_{\mathbf{y}})$, where $\text{cov}(\mathbf{e}_{\mathbf{y}}) = \Psi_{\mathbf{y}}$ is the H-dimensional diagonal positive definite variance-covariance matrix of the error of model 2.2. In addition it is assumed that errors in the two models are uncorrelated $\text{cov}(\mathbf{e}_{\mathbf{x}}, \mathbf{e}_{\mathbf{y}}) = \Sigma_{\mathbf{e}_{\mathbf{y}}\mathbf{e}_{\mathbf{x}}} = 0$; and errors and factors are uncorrelated, i.e., $\text{cov}(\mathbf{e}_{\mathbf{y}}, \mathbf{g}) = \Sigma_{\mathbf{e}_{\mathbf{y}}\mathbf{g}} = 0$.

Model 2.1 identifies H specific theoretical constructs by means of a common factor model that identifies common information with H factors related to the MIs; while model 2.2 detects the general theoretic construct (the composite indicator) by means of a one-factor model that identifies common information with one general factor related to the H specific factors. Given these assumptions and including model 2.2 into model 2.1 the two level Hierarchical Factor Analysis model is defined

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{A}(\mathbf{c}\mathbf{g} + \mathbf{e}_{\mathbf{y}}) + \mathbf{e}_{\mathbf{x}} = \mathbf{A}\mathbf{c}\mathbf{g} + \mathbf{A}\mathbf{e}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}} \tag{2.3}$$

from which it can be derived that $\mathbf{x} \sim N_J(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, where the variance-covariance matrix $\Sigma_{\mathbf{x}}$ is

$$\begin{aligned}
\Sigma_{\mathbf{x}} &= \text{cov}(\mathbf{A}\mathbf{c}\mathbf{g} + \mathbf{A}\mathbf{e}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}}) \\
&= \mathbf{A}\mathbf{c}\text{cov}(\mathbf{g})\mathbf{c}'\mathbf{A}' + \mathbf{A}\text{cov}(\mathbf{e}_{\mathbf{y}})\mathbf{A}' + \text{cov}(\mathbf{e}_{\mathbf{x}}) \\
&= \mathbf{A}\mathbf{c}\mathbf{c}'\mathbf{A}' + \mathbf{A}\Psi_{\mathbf{y}}\mathbf{A}' + \Psi_{\mathbf{x}} \\
&= \mathbf{A}\Sigma_{\mathbf{y}}\mathbf{A}' + \Psi_{\mathbf{x}}
\end{aligned} \tag{2.4}$$

with

$$\Sigma_{\mathbf{y}} = \mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}} \tag{2.5}$$

Matrix $\Sigma_{\mathbf{y}}$ is a correlation matrix since specific factors are standardized. Similarly to Exploratory Factor Analysis, the HFA model (2.3) does not imply the a priori knowledge of

relations between MIs and factors. Model 2.4 is a restricted Confirmatory Factor Analysis model, that is an oblique common factor model, where $\Sigma_{\mathbf{y}}$ has the form $\Sigma_{\mathbf{y}} = \mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}}$.

Many hierarchical extensions of FA have been proposed by many authors through the years (e.g., Thompson (1951), Schmid and Leiman (1957), Wherry (1959), Wherry (1975), Wherry (1984), and Le Dien and Pages (2003)); the novelty of the proposal consists of the covariance structure given by the model (2.4 and 2.5). Although the proposal has developed as two nested models: first order factors model (2.1) and second order factor model (2.2), all parameters of the model (2.3) are estimated simultaneously via a descent coordinate algorithm (Section 2.2.2).

Property 1. *A necessary but not sufficient condition for the HFA model identification is that the number of unknown parameters must be smaller or equal to the number of equations*

$$JH + 2H + J \leq \frac{J(J+1)}{2} \tag{2.6}$$

where JH, H, H, J, parameters are in $\mathbf{A}$, $\mathbf{c}$, $\Psi_{\mathbf{y}}$ *and* $\Psi_{\mathbf{x}}$, *respectively. In fact, for the estimation of model parameters in HFA the number of equations must be larger than or equal to the number of parameters. The equations of the HFA model are* $\frac{J(J+1)}{2}$, *while* $JH + 2H + J$ *are the number of parameters.*

Property 2. *Given H and* $\Psi_{\mathbf{x}}$ *the parameters* $\mathbf{A}$, $\mathbf{c}$, $\Psi_{\mathbf{y}}$ *are not unique.*

To see this, let matrix $\mathbf{Z}$ *be any non-singular matrix of order H such that*

$$\mathbf{A}^* = \mathbf{A}\mathbf{Z} \tag{2.7}$$

$$\Sigma_{\mathbf{y}}^* = \mathbf{Z}^{-1}\Sigma_{\mathbf{y}}\mathbf{Z}'^{-1} = \mathbf{c}^*\mathbf{c}'^* + \Psi_{\mathbf{y}}^* \tag{2.8}$$

with

$$\mathbf{c}^* = \mathbf{Z}^{-1}\mathbf{c} \tag{2.9}$$

$$\mathbf{e}_{\mathbf{y}}^* = \mathbf{Z}^{-1}\mathbf{e}_{\mathbf{y}} \tag{2.10}$$

$$cov(\mathbf{e}_{\mathbf{y}}^*) = \Psi_{\mathbf{y}}^* = \mathbf{Z}^{-1}cov(\mathbf{e}_{\mathbf{y}})\mathbf{Z}'^{-1} = \mathbf{Z}^{-1}\Psi_{\mathbf{y}}\mathbf{Z}'^{-1} \tag{2.11}$$

thus

$$\mathbf{A}^*\Sigma_{\mathbf{y}}^*\mathbf{A}'^* + \Psi_{\mathbf{x}} = \mathbf{A}\mathbf{Z}\mathbf{Z}^{-1}\Sigma_{\mathbf{y}}\mathbf{Z}'^{-1}\mathbf{Z}'\mathbf{A}' + \Psi_{\mathbf{x}} = \mathbf{A}\Sigma_{\mathbf{y}}\mathbf{A}' + \Psi_{\mathbf{x}} \tag{2.12}$$

This illustrated a fundamental indeterminacy in the Hierarchical Factor Analysis model similar to EFA. In fact,

$$\begin{aligned}\mathbf{x} - \mu_{\mathbf{x}} &= \mathbf{A}^*\mathbf{c}^*\mathbf{g} + \mathbf{A}^*\mathbf{e}_{\mathbf{y}}^* + \mathbf{e}_{\mathbf{x}} \\ &= \mathbf{A}\mathbf{Z}^{-1}\mathbf{Z}\mathbf{c}\mathbf{g} + \mathbf{A}\mathbf{Z}^{-1}\mathbf{Z}\mathbf{e}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}} \\ &= \mathbf{A}\mathbf{c}\mathbf{g} + \mathbf{A}\mathbf{e}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}}\end{aligned} \tag{2.13}$$

Therefore, without further restrictions the model parameters $\mathbf{A}$, $\mathbf{c}$, $\Psi_{\mathbf{y}}$ *are not uniquely identified. This indeterminacy will be used to require that the total variance of the factors in model (2.13) is maximized.*

2.2.1 Maximum Likelihood estimation of the two level Hierarchical Factor Analysis model

Suppose that a random sample of $n > J$ multivariate observations $\mathbf{x}_i = [x_{i1}, \ldots, x_{iJ}]'$, $i = 1, \ldots, n$ of $\mathbf{x}$ is observed; thus, the log-likelihood is

$$L(\mathbf{x}_i, \mu_{\mathbf{x}}, \mathbf{A}, \mathbf{c}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = \\ = -\frac{nJ}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \mu_{\mathbf{x}}) \quad (2.14)$$

Maximizing L with respect to $\mu_{\mathbf{x}}$ gives the sample mean; thus rewriting $\sum_{i=1}^{n}(\mathbf{x}_i - \widehat{\mu}_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \widehat{\mu}_{\mathbf{x}}) = n\mathbf{S}$ yields to the reduced log-likelihood

$$L(\mathbf{x}_i, \mathbf{A}, \mathbf{c}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = \\ = -\frac{nJ}{2}\log(2\pi) - \frac{n}{2}\{\log|\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| + \mathrm{tr}[(\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}})^{-1}\mathbf{S}]\} \quad (2.15)$$

It is worth to note that the maximization of $L(\mathbf{x}_i, \mathbf{A}, \mathbf{c}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}})$ is equivalent to the minimization of the discrepancy function

$$D(\mathbf{x}_i, \mathbf{A}, \mathbf{c}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = \\ = \log|\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| - \log|\mathbf{S}| - J + \mathrm{tr}\{[\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\} \to \min_{\mathbf{A}, \mathbf{c}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}} \quad (2.16)$$

Let $\mathbf{S}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}$ be positive definite matrices, $\Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}$ diagonal and $\mathbf{c}$ real $(H \times 1)$ vector. The minimization of the discrepancy function (2.16) is equivalent to the minimization of the function

$$f = \log|\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}| + \mathrm{tr}\{[\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\} \quad (2.17)$$

Define $\Sigma_{\mathbf{x}} = \mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}$ and $\mathbf{C} = \Sigma_{\mathbf{x}}^{-1} - \Sigma_{\mathbf{x}}^{-1}\mathbf{S}\Sigma_{\mathbf{x}}^{-1}$, by differentiating function f with respect to $\mathbf{A}$ gives $\partial f/\partial \mathbf{A} = 2\mathrm{tr}(\mathbf{A}'\mathbf{C})\partial \mathbf{A}$.

Thus, the first-order condition is $\mathbf{C}\mathbf{A} = \mathbf{0}$, or equivalently $\mathbf{A} = \mathbf{S}\Sigma_{\mathbf{x}}^{-1}\mathbf{A}$, which corresponds to require $\Sigma_{\mathbf{x}} = \mathbf{S}$.

Therefore,

$$\begin{aligned} \mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' &= \mathbf{S} - \Psi_{\mathbf{x}} \\ &= \Psi_{\mathbf{x}}^{\frac{1}{2}}(\Psi_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{S}\Psi_{\mathbf{x}}^{-\frac{1}{2}} - \mathbf{I})\Psi_{\mathbf{x}}^{\frac{1}{2}} \\ &= \Psi_{\mathbf{x}}^{\frac{1}{2}}(\mathbf{T}\Lambda\mathbf{T}' - \mathbf{I})\Psi_{\mathbf{x}}^{\frac{1}{2}} \\ &= \Psi_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}(\Lambda - \mathbf{I})\mathbf{T}'\Psi_{\mathbf{x}}^{\frac{1}{2}} \end{aligned} \quad (2.18)$$

where $\Psi_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{S}\Psi_{\mathbf{x}}^{-\frac{1}{2}} = \mathbf{T}\Lambda\mathbf{T}'$ is a spectral decomposition, with $\mathbf{T}$ orthonormal eigenvectors and Λ the diagonal matrix of eigenvalues.

Thus,

$$\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})^{\frac{1}{2}} = \Psi_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}(\Lambda - \mathbf{I})^{\frac{1}{2}} \quad (2.19)$$

and therefore

$$\mathbf{A} = \Psi_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}(\Lambda - \mathbf{I})^{\frac{1}{2}}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})^{-\frac{1}{2}} \tag{2.20}$$

Differentiating f with respect to $\Psi_{\mathbf{x}}$ gives $\partial f/\partial\Psi_{\mathbf{x}} = \mathrm{tr}\mathbf{C}\partial\Psi_{\mathbf{x}}$. Since $\Psi_{\mathbf{x}}$ is diagonal and since the first-order condition for $\mathbf{A}$ is $\mathbf{S} = \Sigma_{\mathbf{x}}$, thus the first-order condition for $\Psi_{\mathbf{x}}$ is $\mathrm{diag}(\mathbf{S}) = \mathrm{diag}(\Sigma_{\mathbf{x}})$, hence

$$\mathrm{diag}(\mathbf{S}) = \mathrm{diag}(\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}' + \Psi_{\mathbf{x}}) \tag{2.21}$$

Therefore,

$$\Psi_{\mathbf{x}} = \mathrm{diag}(\mathbf{S} - (\mathbf{A}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}})\mathbf{A}') \tag{2.22}$$

To compute $\mathbf{c}$ also the first-order conditions for $\mathbf{A}$ have to be satisfied

$$\begin{aligned}
\mathbf{c}\mathbf{c}' &= \mathbf{A}^{+}(\mathbf{S} - \Psi_{\mathbf{x}})\mathbf{A}'^{+} - \Psi_{\mathbf{y}} \\
&= \Psi_{\mathbf{y}}^{\frac{1}{2}}(\Psi_{\mathbf{y}}^{-\frac{1}{2}}(\mathbf{A}^{+}(\mathbf{S} - \Psi_{\mathbf{x}})\mathbf{A}'^{+})\Psi_{\mathbf{y}}^{-\frac{1}{2}} - \mathbf{I})\Psi_{\mathbf{y}}^{\frac{1}{2}} \\
&= \Psi_{\mathbf{y}}^{\frac{1}{2}}(\mathbf{U}\mathbf{L}\mathbf{U}' - \mathbf{I})\Psi_{\mathbf{y}}^{\frac{1}{2}} \\
&= \Psi_{\mathbf{y}}^{\frac{1}{2}}\mathbf{U}(\mathbf{L} - \mathbf{I})\mathbf{U}'\Psi_{\mathbf{y}}^{\frac{1}{2}}
\end{aligned} \tag{2.23}$$

where $\Psi_{\mathbf{y}}^{-\frac{1}{2}}(\mathbf{A}^{+}(\mathbf{S} - \Psi_{\mathbf{x}})\mathbf{A}'^{+})\Psi_{\mathbf{y}}^{-\frac{1}{2}} = \mathbf{U}\mathbf{L}\mathbf{U}'$ is a spectral decomposition with $\mathbf{U}$ orthonormal eigenvectors, $\mathbf{L}$ the diagonal matrix of eigenvalues and $\mathbf{A}^{+}$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$ (i.e., $\mathbf{A}^{+} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$). Note that matrix $\mathbf{A}'\mathbf{A}$ is diagonal, so its inverse is easy to compute.

Thus,

$$\mathbf{c} = \Psi_{\mathbf{y}}^{\frac{1}{2}}\mathbf{U}(\mathbf{L} - \mathbf{I})^{\frac{1}{2}} \tag{2.24}$$

To compute $\Psi_{\mathbf{y}}$ the first order conditions for $\Psi_{\mathbf{x}}$ have to be satisfied

$$\mathrm{diag}(\mathbf{A}^{+}(\mathbf{S} - \Psi_{\mathbf{x}})\mathbf{A}'^{+}) = \mathrm{diag}(\mathbf{c}\mathbf{c}' + \Psi_{\mathbf{y}}) \tag{2.25}$$

Therefore,

$$\Psi_{\mathbf{y}} = \mathrm{diag}(\mathbf{A}^{+}(\mathbf{S} - \Psi_{\mathbf{x}})\mathbf{A}'^{+} - \mathbf{c}\mathbf{c}') \tag{2.26}$$

In the next section, the five steps of the algorithm of HFA are summarized.

2.2.2 Algorithm of HFA

Given H, a coordinate descent algorithm for the estimation of the model (2.3) can be described by five steps which are sequentially repeated until a stopping rule is satisfied.

Step 0 [*Initialization*] Fix the matrix $\widehat{\mathbf{\Psi}}_{\mathbf{x}} = \mathrm{diag}(\mathbf{S})$, then the algorithm starts by computing $\widehat{\mathbf{A}} = \widehat{\mathbf{\Psi}}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}(\Lambda - \mathbf{I})^{\frac{1}{2}}$, where $\mathbf{T}$ and Λ are eigenvectors and eigenvalues of $\widehat{\mathbf{\Psi}}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{S}\widehat{\mathbf{\Psi}}_{\mathbf{x}}^{-\frac{1}{2}}$; then compute $\widehat{\mathbf{\Psi}}_{\mathbf{y}} = \mathrm{diag}(\widehat{\mathbf{A}}^{+}(\mathbf{S} - \mathbf{\Psi}_{\mathbf{x}})\widehat{\mathbf{A}}'^{+})$.

Step 1 Given $\widehat{\mathbf{\Psi}}_{\mathbf{x}}$,$\widehat{\mathbf{A}}$ and $\widehat{\mathbf{\Psi}}_{\mathbf{y}}$ the minimization of the discrepancy function (2.16) with respect to $\mathbf{c}$ is given by (2.24).

Step 2 Given $\widehat{\mathbf{\Psi}}_{\mathbf{x}}$,$\widehat{\mathbf{A}}$ and $\widehat{\mathbf{c}}$ the minimization of the discrepancy function (2.16) with respect to $\mathbf{\Psi}_{\mathbf{y}}$ is given by (2.26).

Step 3 Compute $\boldsymbol{\Sigma}_{\mathbf{y}} = \widehat{\mathbf{c}}\widehat{\mathbf{c}}' + \widehat{\boldsymbol{\Psi}}_{\mathbf{y}}$, then transform covariances in correlations.

Step 4 Given $\widehat{\boldsymbol{\Psi}}_{\mathbf{x}}$,$\widehat{\boldsymbol{\Psi}}_{\mathbf{y}}$ and $\widehat{\mathbf{c}}$ the minimization of the discrepancy function (2.16) with respect to $\mathbf{A}$ is given by (2.20).

Step 5 Given $\widehat{\mathbf{A}}$,$\widehat{\boldsymbol{\Psi}}_{\mathbf{y}}$ and $\widehat{\mathbf{c}}$ the minimization of the discrepancy function (2.16) with respect to $\boldsymbol{\Psi}_{\mathbf{x}}$ is given by (2.22).

The five **Steps 1**, **2**, **3**, **4** and **5** are alternated repeatedly. At each step the discrepancy function decreases or at least does not increase. If the reduction of the discrepancy is larger than an arbitrary small positive constant the algorithm continues to iterate, otherwise the algorithm stops and is considered to have converged to a solution that is not guaranteed to be the global minimum. However, repeating the analysis from different random starts, it is observed that in our experiments the solutions are always the same. As shown in Property 2, the solution of HFA is not unique, so it is necessary to include some constraints.

2.2.3 Maximization of the explained variance

Given H, let $\widehat{\Psi}_{\mathbf{x}}$, $\widehat{\mathbf{A}}$, $\widehat{\Psi}_{\mathbf{y}}$ and $= \widehat{\mathbf{c}}$ the maximum estimates of the model (2.3); thus, the estimated variance-covariance matrix is:

$$\widehat{\Sigma}_{\mathbf{x}} = \widehat{\mathbf{A}}(\widehat{\mathbf{c}}\widehat{\mathbf{c}}' + \widehat{\Psi}_{\mathbf{y}})\widehat{\mathbf{A}}' + \widehat{\Psi}_{\mathbf{x}} \tag{2.27}$$

As shown by Property (2) the MLE solution (2.27) is not unique, thus a non-singular matrix $\mathbf{Z}$ can be defined

$$\mathbf{A}^* = \widehat{\mathbf{A}}\mathbf{Z} \tag{2.28}$$

$$\Psi^*_{\mathbf{y}} = \mathbf{Z}^{-1}\widehat{\Psi}_{\mathbf{y}}\mathbf{Z}'^{-1} \tag{2.29}$$

$$\mathbf{c}^* = \mathbf{Z}^{-1}\widehat{\mathbf{c}} \tag{2.30}$$

such that minimizes

$$||_{(H)}\widehat{\Sigma}_{\mathbf{x}} - \widehat{\mathbf{A}}\mathbf{Z}\mathbf{Z}'\widehat{\mathbf{A}}'||^2 = ||_{(H)}\widehat{\Sigma}^{\frac{1}{2}}_{\mathbf{x}} - \widehat{\mathbf{A}}\mathbf{Z}||^2 \tag{2.31}$$

where $_{(H)}\widehat{\Sigma}_{\mathbf{x}} = \mathbf{P}_{(H)}\mathbf{L}_{(H)}\mathbf{P}'_{(H)}$ is the rank-H approximation from the spectral decomposition of the matrix $\widehat{\Sigma}_{\mathbf{x}}$ and the columns of $\mathbf{P}_{(H)}$ are H eigenvectors, whereas the values on the diagonal of $\mathbf{L}_{(H)}$ the corresponding H eigenvalues of the matrix $\widehat{\Sigma}_{\mathbf{x}}$.

From the solution $\mathbf{Z}\mathbf{Z}' = \widehat{\mathbf{A}}^{+}_{(H)}\widehat{\Sigma}_{\mathbf{x}}\widehat{\mathbf{A}}'^{+}$, matrix $\mathbf{Z}$ is derived

$$\mathbf{Z} = \widehat{\mathbf{A}}^{+}\mathbf{P}_{(H)}\mathbf{L}^{\frac{1}{2}}_{(H)} \tag{2.32}$$

Now, $\text{tr}(_{(H)}\widehat{\Sigma}_{\mathbf{x}})$ is the total variance that can be explained by a rank-H approximation of $\widehat{\Sigma}_{\mathbf{x}}$. Since $\text{tr}(_{(H)}\widehat{\Sigma}_{\mathbf{x}}) = ||_{(H)}\widehat{\Sigma}^{\frac{1}{2}}_{\mathbf{x}} - \widehat{\mathbf{A}}\mathbf{Z}||^2 + ||\widehat{\mathbf{A}}\mathbf{Z}||^2$, by minimizing (2.31) is equivalent to maximize the variance $||\mathbf{A}^*||^2$ explained by the H-dimensional solution of HFA.

Therefore, the indeterminacy can be positively used to find among the infinite solutions the one that most explains the variance reconstructed by the H-dimensional solution of

the HFA. However, this solution can still be rotated if together with the transformations (2.28,2.29 and 2.30) also a rotation is applied, that is

$$\mathbf{A}^* = \widehat{\mathbf{A}}\mathbf{Z}\mathbf{M} \tag{2.33}$$

$$\Psi_{\mathbf{y}}^* = \mathbf{M}'\mathbf{Z}^{-1}\widehat{\Psi}_{\mathbf{y}}\mathbf{Z}'^{-1}\mathbf{M} \tag{2.34}$$

$$\mathbf{c}^* = \mathbf{M}'\mathbf{Z}^{-1}\widehat{\mathbf{c}} \tag{2.35}$$

where $\mathbf{M}\mathbf{M}' = \mathbf{I}$. Varimax rotation is suggested to improve interpretation of results (Kaiser 1958; Sherin 1966; J.M.F. ten Berge 1984; J.M.F ten Berge, Knol, and Kiers 1988; Thurstone 1947).

2.3 Disjoint models

2.3.1 Disjoint Non-Orthogonal Factor Analysis

The Disjoint Orthogonal Factor Analysis (DFA) (Vichi 2017) assumes that observations can be reconstructed by a non-observable $(H \times 1)$ random vector $\mathbf{y}$ denoting a reduced set of $(H \leq J)$ common factors by considering the following model

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{e}_{\mathbf{x}}\mathbf{X} = \mathbf{Y}\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{X}} \tag{2.36}$$

with

$$\Sigma_{\mathbf{x}} = \mathbf{A}\Sigma_{\mathbf{y}}\mathbf{A}' + \Psi_{\mathbf{x}} \tag{2.37}$$

where, here, factors are correlated, and the loading matrix $\mathbf{A}$ is restricted to the product

$$\mathbf{A} = \mathbf{B}\mathbf{V} \tag{2.38}$$

where $\mathbf{V} = [v_{jh}]$ is a $(J \times H)$ binary and row stochastic matrix identifying a partition of MIs into H clusters corresponding to H factors, with $v_{jh} = 1$, if the MI j-th belongs to the h-th cluster; $v_{jh} = 0$ otherwise; whereas, $\mathbf{B} = \text{diag}(b_1, \ldots, b_J)$ is a $(J \times J)$ diagonal matrix weighting MI j-th such that

$$b_j^2 > 0 \tag{2.39}$$

If it is allowed $b_j^2 \geq 0$, when $b_j^2 = 0$ the DFA admits a model selection feature, that is, an MI is assigned to a cluster with a loading equal to zero. In this case MI j is essentially discarded from the model. DFA assumes orthogonal factors, that is, $\Sigma_{\mathbf{y}} = \mathbf{I}_H$, however in this paper this condition is relaxed in order to allow a hierarchical structure of the data.

Thus, the DFA model can be rewritten as follows

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{B}\mathbf{V}\mathbf{y} + \mathbf{e}_{\mathbf{x}} \tag{2.40}$$

with

$$\Sigma_{\mathbf{x}} = \mathbf{B}\mathbf{V}\Sigma_{\mathbf{y}}\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{x}} \tag{2.41}$$

subject to constraints

$$\mathbf{V} = [v_{jh} \in \{0,1\} : j = 1,...,J, h = 1,...,H] \tag{2.42}$$

$$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J \quad \text{i.e.} \quad \sum_{h=1}^{H} v_{jq} = 1 \quad j = 1,...,J \tag{2.43}$$

$$\mathbf{B} = \text{diag}(b_1, \ldots, b_J \tag{2.44}$$

$$\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = \text{diag}(a_{\cdot 1}^2, \ldots, a_{\cdot H}^2) \tag{2.45}$$

If $\Sigma_{\mathbf{y}} = \mathbf{I}_H$ (Disjoint Orthogonal Factor Analysis) the variance-covariance $\Sigma_{\mathbf{x}}$ is block diagonal

$$\Sigma_{\mathbf{x}} = \text{blkdiag}(\Sigma_{11},\ldots,\Sigma_{hh},\ldots,\Sigma_{HH}) = \begin{bmatrix} \Sigma_{11} & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \Sigma_{22} & 0 & \ldots & \ldots & 0 \\ \vdots & 0 & \ddots & 0 & \ldots & 0 \\ \vdots & \vdots & 0 & \Sigma_{hh} & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \Sigma_{HH} \end{bmatrix} \tag{2.46}$$

where each block is the variance-covariance matrix Σ_{hh} of the MIs relating to the factor h,

$$\Sigma_{hh} = \mathbf{B}_h(\mathbf{1}_{n_h}\mathbf{1}'_{n_h})\mathbf{B}_h + \Psi_h = \mathbf{b}_h\mathbf{b}'_h + \Psi_h \tag{2.47}$$

and $\mathbf{B}_h = \text{diag}(\mathbf{b}_h), \mathbf{b}_h = [b_{1h},\ldots,b_{n_hh}]'$ and $\Psi_h = \text{diag}(\psi_h), \psi_h = [\psi_{1h},\ldots,\psi_{n_hh}]'$; where n_h is the number of MIs related to the latent factor h.

Therefore, the DFA model assumes that a relevant correlation between MIs related to the same latent factor is observed and a zero, or at least a negligible, correlation between MIs each one related to a different latent factor is detected.

If $\Sigma_{\mathbf{y}}$ has non-diagonal elements different from zero, (Disjoint Non-Orthogonal Factor Analysis) the block diagonal variance-covariance disappears and $\Sigma_{\mathbf{x}}$ has the form

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1H} \\ \Sigma_{21} & \Sigma_{22} & \ldots & \Sigma_{2H} \\ \vdots & \vdots & \Sigma_{hh} & \vdots \\ \Sigma_{H1} & \Sigma_{H2} & \ldots & \Sigma_{HH} \end{bmatrix} \tag{2.48}$$

where the generic correlation between two clusters of MIs are constrained in the

$$\Sigma_{hk} = \mathbf{B}_h(\mathbf{1}_{n_h}\mathbf{1}'_{n_k})\mathbf{B}_k = r_{hk}\mathbf{b}_h\mathbf{b}'_k \tag{2.49}$$

and r_{hk} is the correlation between factor h and factor k.

It is important to observe that in order to identify two distinct factors h and k with associated two disjoint clusters of MIs defining matrices Σ_{hh} and Σ_{kk}, high correlation within these matrices and lower correlation within Σ_{hk} needs to be observed. In fact, if also highly correlation is observed in Σ_{hk} thus a single factor is present in the data since the two clusters of MIs are not distinct and actually form a single cluster in which MIs are all highly correlated. Thus, (2.49) guarantees that correlations in Σ_{hk} are lower than correlations within Σ_{hh} and Σ_{kk}.

2.3.2 Hierarchical Disjoint Factor Analysis

If in the HFA some a priori substantive knowledge is incorporated in the form of restrictions on the loading matrix, this usually improves the description of the latent factors and leads to a parsimonious model with a simple loading matrix structure. Therefore, if in HFA a Simple Structure Model (SSM) is assumed observed for the data, this means that the factor loading matrix has the form $\mathbf{A} = \mathbf{BV}$ and the HFA model becomes the *Hierarchical Disjoint Factor Analysis* model

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{BV}(\mathbf{cg} + \mathbf{e}_{\mathbf{y}}) + \mathbf{e}_{\mathbf{x}} = \mathbf{BVcg} + \mathbf{BVe}_{\mathbf{y}} + \mathbf{e}_{\mathbf{x}} \tag{2.50}$$

It can be derived that $\mathbf{x} \sim N_J(\mu_\mathbf{x}, \Sigma_\mathbf{x})$ with

$$\Sigma_\mathbf{x} = \mathbf{B}\mathbf{V}\Sigma_\mathbf{y}\mathbf{V}'\mathbf{B} + \Psi_\mathbf{x} \tag{2.51}$$

where

$$\Sigma_\mathbf{y} = \mathbf{c}\mathbf{c}' + \Psi_\mathbf{y} \tag{2.52}$$

The discrepancy function (2.16) can be rewritten

$$\begin{aligned} &D(\mathbf{x}_i, \mathbf{B}, \mathbf{V}, \mathbf{c}, \Psi_\mathbf{x}, \Psi_\mathbf{y}) = \\ &= \log|\mathbf{B}\mathbf{V}(\mathbf{c}\mathbf{c}' + \Psi_\mathbf{y})\mathbf{V}'\mathbf{B} + \Psi_\mathbf{x}| - \log|\mathbf{S}| - J + \mathrm{tr}\{[\mathbf{B}\mathbf{V}(\mathbf{c}\mathbf{c}' + \Psi_\mathbf{y})\mathbf{V}'\mathbf{B} + \Psi_\mathbf{x}]^{-1}\mathbf{S}\} \to \min_{\mathbf{B},\mathbf{V},\mathbf{c},\Psi_\mathbf{x},\Psi_\mathbf{y}} \end{aligned} \tag{2.53}$$

2.3.3 HDFA Algorithm

Given H, a coordinate descent algorithm for the estimation of the model (2.50) can be described by five steps which are sequentially repeated until a stopping rule is satisfied.

Step 0 [*Initialization*] A random partition $\widehat{\mathbf{V}}$ is generated from a multinomial distribution in H categories each one with equal probability, where categories are not empty. Matrix $\widehat{\mathbf{\Psi}}_\mathbf{x} = \mathrm{diag}(\mathbf{S})$.

Step 1 Given $\widehat{\mathbf{V}} = [\mathbf{v}_{\cdot 1}, \ldots, \mathbf{v}_{\cdot H}]$ for each column $h = 1, \ldots, H$ compute the rank-1 approximation of the positive semi-definite matrix $\widehat{\mathbf{\Psi}}_{\mathbf{x}h}^{-\frac{1}{2}}\mathbf{S}_h\widehat{\mathbf{\Psi}}_{\mathbf{x}h}^{-\frac{1}{2}} \cong \lambda_{1h}\mathbf{u}_{1h}\mathbf{u}_{1h}'$, where λ_{1h} is the largest eigenvalue and $\mathbf{u}_{1h}$ is the associated eigenvector of the variance-covariance matrix $\widehat{\mathbf{\Psi}}_{\mathbf{x}h}^{-\frac{1}{2}}\mathbf{S}_h\widehat{\mathbf{\Psi}}_{\mathbf{x}h}^{-\frac{1}{2}}$ corresponding to MIs identified by $\mathbf{v}_{\cdot h}$, the h-th column of $\widehat{\mathbf{V}}$. Thus, the discrepancy function $D(\mathbf{B}, \mathbf{V}, \mathbf{c}, \mathbf{\Psi}_\mathbf{x}, \mathbf{\Psi}_\mathbf{y})$ (2.53) is minimized with respect to $\mathbf{B}_h = \mathrm{diag}(\mathbf{b}_h)$ by

$$\widehat{\mathbf{b}}_h = \widehat{\mathbf{\Psi}}_{\mathbf{x}h}^{-\frac{1}{2}}\mathbf{u}_{1h}(\lambda_{1h} - 1)^{\frac{1}{2}} \quad h = 1, \ldots, H \tag{2.54}$$

Step 2 Given $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}$, $\widehat{\mathbf{c}}$ and $\widehat{\mathbf{\Psi}}_\mathbf{y}$ the minimization of the discrepancy function (2.53) with respect to $\mathbf{\Psi}_\mathbf{x}$ is given by

$$\widehat{\mathbf{\Psi}}_\mathbf{x} = \mathrm{diag}(\mathbf{S} - \widehat{\mathbf{A}}(\widehat{\mathbf{c}}\widehat{\mathbf{c}}' + \widehat{\mathbf{\Psi}}_\mathbf{y})\widehat{\mathbf{A}}') \tag{2.55}$$

Step 3 Given $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}$, $\widehat{\mathbf{\Psi}}_\mathbf{x}$ and $\widehat{\mathbf{\Psi}}_\mathbf{y}$ the minimization of the discrepancy function (2.53) with respect to $\mathbf{c}$ is given by

$$\widehat{\mathbf{c}} = \widehat{\mathbf{\Psi}}_y^{\frac{1}{2}}\mathbf{U}(\mathbf{L} - \mathbf{I})^{\frac{1}{2}} \tag{2.56}$$

as already seen in (2.24), where $\widehat{\mathbf{\Psi}}_\mathbf{y}^{-\frac{1}{2}}(\widehat{\mathbf{A}}^+(\mathbf{S} - \widehat{\mathbf{\Psi}}_\mathbf{x})\widehat{\mathbf{A}}'^+)\widehat{\mathbf{\Psi}}_\mathbf{y}^{-\frac{1}{2}} = \mathbf{U}\mathbf{L}\mathbf{U}'$ is the spectral decomposition with $\mathbf{U}$ orthonormal eigenvectors and $\mathbf{L}$ is the diagonal matrix of eigenvalues.

Step 4 Given $\widehat{\mathbf{A}} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}$, $\widehat{\mathbf{c}}$ and $\widehat{\mathbf{\Psi}}_\mathbf{x}$ the minimization of the discrepancy function (2.53) with respect to $\mathbf{\Psi}_\mathbf{y}$ is given by

$$\widehat{\mathbf{\Psi}}_\mathbf{y} = \mathrm{diag}(\widehat{\mathbf{A}}^+(\mathbf{S} - \widehat{\mathbf{\Psi}}_\mathbf{x})\widehat{\mathbf{A}}'^+ - \widehat{\mathbf{c}}\widehat{\mathbf{c}}') \tag{2.57}$$

Step 5 Partition $\widehat{\mathbf{V}} = [\widehat{\mathbf{v}}_{\cdot 1}, \ldots, \widehat{\mathbf{v}}_{\cdot H}]$ is obtained row by row by assigning each MI j-th to the cluster h-th that most decreases the discrepancy function (2.53). Thus, formally

$$\begin{cases} \widehat{v}_{jh} = 1 & \text{if} \quad \underset{h=1,\ldots,H}{\arg\min} D(\widehat{\mathbf{B}}, [\widehat{\mathbf{v}}_{1\cdot}, ..., \widehat{\mathbf{v}}_{j\cdot} = \mathbf{i}_{h\cdot}, ..., \widehat{\mathbf{v}}_{J\cdot}]', \widehat{\mathbf{c}}, \widehat{\mathbf{\Psi}}_{\mathbf{x}}, \widehat{\mathbf{\Psi}}_{\mathbf{y}}) \\ \widehat{v}_{jh} = 0 & \text{otherwise} \end{cases} \tag{2.58}$$

where $\mathbf{i}_{h\cdot}$ is the h-th column of the identity matrix of order H. Note that the update of each row of $\mathbf{V}$ induces the update of the loadings of the two clusters of MIs that are eventually changed.

The five **Steps 1**, **2**, **3**, **4** and **5** are alternated repeatedly. At each step the discrepancy function decreases or at least does not increase. If the reduction of the discrepancy is larger than an arbitrary small positive constant the algorithm continues to iterate, otherwise the algorithm stops and is considered to have converged to a solution that is not guaranteed to be global minimum. To avoid the well-known sensitivity of the coordinate descent algorithms to the starting values and to increase the chance of finding the global minimum, the algorithm should be run several times starting from different initial estimates of $\mathbf{V}$ and retaining the best solution. The algorithm generally stops after a few iterations (in our experiments and simulation studies less than 15).

Until this moment the problem of *indeterminacy* in the HFA has been presented by taking into account only the non-unique identification of the parameters. It has been shown in Property 2 and it has been positively used in Section 2.2.3. Therefore, showing that different estimations of the parameters can produce the same covariance structure of the model.

It is important to remark that, for fixed $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{\Psi}}_{\mathbf{x}}$, factor scores can be estimated by considering different approaches. The same estimated covariance structure, thus, may correspond to infinite distinct potential factor scores (Bollen 1989; Schönemann 1973; Schönemann and Steiger 1978). This problem, referred to as *factor scores indeterminacy*, has been a source of considerable controversy ever since.

Many methods to estimate the factor scores have been proposed throughout the years. For instance, we consider the weighted least square estimation of $\mathbb{E}(\mathbf{X}|\mathbf{Y})$

$$\widehat{\mathbf{Y}} = \mathbf{X}(\widehat{\mathbf{\Psi}}_{\mathbf{x}}^{-1}\widehat{\mathbf{A}}(\widehat{\mathbf{A}}'(\widehat{\mathbf{\Psi}}_{\mathbf{x}}^{-1}\widehat{\mathbf{A}})^{-1}) \tag{2.59}$$

as proposed by Bartlett (1937), or simply the Thompson (1934) regression estimator

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{S}^{-1}\widehat{\mathbf{A}}) \tag{2.60}$$

The problem of factor scores indeterminacy can be resumed saying that an individual in the analysis with a high ranking according to one set of factor scores, could receive a low ranking according to another set of factor scores, when both sets are consistent with the pattern of estimated coefficients. Therefore, it is worth observing that the debates surrounding factor scores indeterminacy have been harsh and with diverging opinions (Bartholomew 1981; Guttman 1955; Maraun 1996; McDonald 1974; S. Mulaik 2009; S. Mulaik and McDonald 1977; Rozeboom 1996; Wilson 1996); however, these will not be source for further discussion here.

2.4 Hierarchical Disjoint Non-Negative Factor Analysis

Let us recall that the discrepancy function (2.53) is minimized with respect to $\mathbf{B}_h = \text{diag}(\mathbf{b}_h)$ by

$$\widehat{\mathbf{b}}_h = \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{u}_{1h} (\lambda_{1h} - 1)^{\frac{1}{2}} \quad h = 1, \ldots, H \tag{2.61}$$

where λ_{1h} is the largest eigenvalue and $\mathbf{u}_{1h}$ is the associated eigenvector of the variance-covariance matrix $\widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}}$ corresponding to MIs identified by $\mathbf{v}_{\cdot h}$, the h-th column of $\widehat{\mathbf{V}}$. Values λ_{1h} and $\mathbf{u}_{1h}$ minimize $||\widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \lambda_{1h} \mathbf{u}_{1h} \mathbf{u}'_{1h}||^2$, or equivalently

$$||\mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}'_{1h}||^2 \tag{2.62}$$

where $\mathbf{X}_h$ is the centred data matrix formed by MIs identified by $\mathbf{v}_{\cdot h}$ and $\mathbf{y}_h$ is the factor score vector. The problem in (2.62) can be solved by an alternating least squares algorithm that alternates the solution of two regression problems. Given $\widehat{\mathbf{u}}_{1h}$ compute $\widehat{\mathbf{y}}_h$ by

$$\widehat{\mathbf{y}}_h = \mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{u}}_{1h} (\widehat{\mathbf{u}}'_{1h} \widehat{\mathbf{u}}_{1h})^{-1} \tag{2.63}$$

Given $\widehat{\mathbf{y}}_h$ compute $\widehat{\mathbf{u}}_{1h}$ by

$$\widehat{\mathbf{u}}_{1h} = \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{X}'_h \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}'_h \widehat{\mathbf{y}}_h)^{-1} \tag{2.64}$$

At each iteration of **Steps 1**, **2**, **3**, **4** and **5** (corresponding to 2.54,2.55,2.56,2.57 and 2.58), the discrepancy function (2.53) decreases or at least does not increase. The algorithm stops when the discrepancy function (2.53) decreases less than a positive arbitrary constant. Now it is required that the vector $\mathbf{u}_{1h}$ is non-negative. Therefore, the algorithm based on (2.53, 2.54,2.55,2.56,2.57 and 2.58) has to be modified to include a non-negative constraint on $\mathbf{u}_{1h}$. The solution can be found by the Non-Negative Least Squares Algorithm (Lawson and Hanson 1974). This is an active set algorithm, where the H inequality constraints are active if the regression coefficient $\mathbf{u}'_{1h}$ in the loss function $||\mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}'_{1h}||^2$ (2.62) will be negative (or zero) when estimated unconstrained, otherwise constrains are passive. The non-negative solution of (2.53) with respect to $\mathbf{u}_{1h}$ will simply be the unconstrained least squares solution using only the MIs corresponding to the passive set, setting the regression coefficients of the active set to zero. Therefore

$$\widehat{\mathbf{u}}_{1h} = \begin{cases} \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{X}'_{h+} \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}'_h \widehat{\mathbf{y}}_h)^{-1} \\ 0 \quad \text{otherwise} \end{cases} \tag{2.65}$$

where $\mathbf{X}_{h+}$ is the set of passive MIs.

A similar step is used to constrain $\widehat{\mathbf{c}}$ to be non-negative.

2.5 Simulation study

In this section, a simulation study has been implemented in order to assess the classification of MIs and evaluate the reliability of the specific factors. Each simulated random sample of $n > J$ multivariate observation $\mathbf{x}_i$ $(i = 1, \ldots, n)$ is generated by using the underlying HDFA model (2.50) with $\mathbf{e}_{\mathbf{x}_i} \sim N_J(\mathbf{0}, \alpha \mathbf{I}_J)$, where the constant α sets the different levels of error.

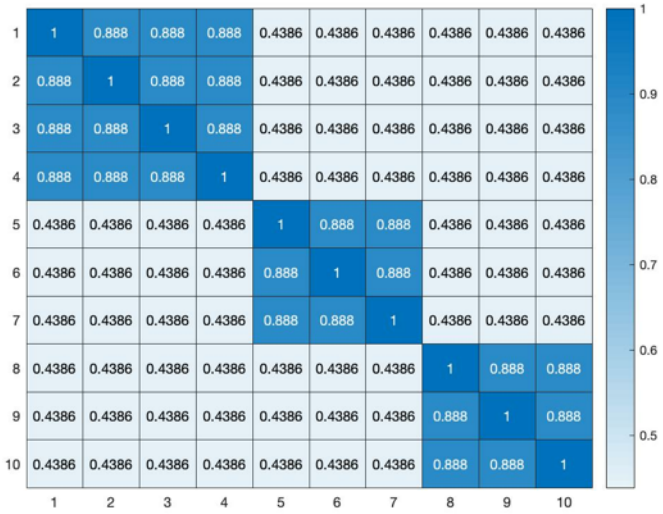


Figure 2.1. An example of heatmap of $\mathbf{R_x}$

By considering the matrix $\mu_{\mathbf{x}} = 0$ in (2.50)

$$\mathbf{x}_i = \mathbf{BVy}_i + \mathbf{e}_{\mathbf{x}_i} = \mathbf{BV}(\mathbf{cg}_i + \mathbf{e}_{\mathbf{y}}) + \mathbf{e}_{\mathbf{x}_i} = \mathbf{BVCg}_i + \mathbf{BVe}_{\mathbf{y}_i} + \mathbf{e}_{\mathbf{x}_i} \tag{2.66}$$

$$\Sigma_{\mathbf{x}} = \mathbf{BV}\Sigma_{\mathbf{y}}\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{x}} = \mathbf{BVcc}'\mathbf{V}'\mathbf{B} + \mathbf{BV}\Psi_{\mathbf{y}}\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{x}} \tag{2.67}$$

We can assume a correlation matrix for $\mathbf{X}$ as follows

$$\mathbf{R_x} = \beta(\mathbf{VV}' - \mathbf{I}_J) + \gamma(\mathbf{1}_J\mathbf{1}_J' - \mathbf{VV}') + \mathbf{I}_J \tag{2.68}$$

where $\beta = 0.8 + 0.05\delta$, $\gamma = 0.35 + 0.05\delta$, with $\delta \sim N(0,1)$ and $\beta > \gamma$.

Let consider, $\mathbf{X}_t \sim N_J(\mathbf{0}, \mathbf{R_x})$ and $\mathbf{e}_{\mathbf{x}_i} \sim N_J(\mathbf{0}, \alpha\mathbf{I}_J)$ in order to have $\mathbf{X} = \mathbf{X}_t + \mathbf{E}$. In particular, the following scenarios have been considered in the simulation study:

- number of observed MIs, $J = 20, 50$;
- number of the specific factors, $H = 5, 10$;
- different error levels, $\alpha = 0.3, 0.5, 0.8$.

For all settings, the membership matrix $\mathbf{V}$ is a random partition of J MIs in H nonempty clusters. Thus, $\mathbf{V} = [\mathbf{v}_{\cdot 1}, \ldots, \mathbf{v}_{\cdot H}]$, with $v_j \sim Multinomial(H : p_h = \frac{1}{H},\ h = 1, \ldots, H)$, $\forall j = 1, \ldots, H$.

The generated data are standardized, and the sample size has been set equal to 200. $\mathbf{R_x}$ is a correlation matrix with a block diagonal structure where correlations within blocks are around 0.8 and the between blocks correlations are around 0.35 (Figure 2.1).

For each setting, 500 datasets have been generated, and for each dataset the algorithm has been run from 30 random starting points to ensure to find the optimal solution although it has been observed that a few random starting points are generally enough. In this simulation study, we wish to show the performance of *Hierarchical Disjoint Non-Negative Factor Analysis* (HDNFA) in identifying the true correlation structure generated for the data. The

model performance has been analyzed by using the Adjusted Rand Index (ARI) (Hubert and Arabie 1985), comparing the simple structure of the MIs generated by the true matrix $\mathbf{V}$ with the one $\widehat{\mathbf{V}}$ estimated by the HDNFA. In fact, ARI ranges between 0 and 1, where 1 implies that the estimated partition $\widehat{\mathbf{V}}$ is identical to the true one $\mathbf{V}$.

We therefore count the percentage of times that ARI is equal to 1 and the average of ARI in the 500 experiments. Furthermore, the model has been evaluated with goodness of fit, adjusted goodness of fit, BIC, AIC and RMSR (root mean square of residuals for the correlation matrix). The goodness of fit is measured with the index proposed by Jöreskog and Sörbom (1984) for Structural Equation Model:

$$GFI = 1 - \frac{\text{tr}[(\Sigma_{\mathbf{x}}^{-1}\mathbf{S} - \mathbf{I})^2]}{\text{tr}[(\Sigma_{\mathbf{x}}^{-1}\mathbf{S})^2]} \tag{2.69}$$

where $\mathbf{S}$ is the observed variance-covariance matrix. It is upper bounded to 1, where 1 implies the perfect fit of the model; negative values can be observed. The adjusted goodness of fit sets the GFI for degrees-of-freedom in the model:

$$AGFI = 1 - \left[\frac{J(J+1)}{2df}\right](1 - GFI) \tag{2.70}$$

where $df = \frac{J(J+1)}{2} - 2J - H$ is the number of degree of freedom for HDNFA.

AIC and BIC are the two popular model selection criteria; Akaike's information criterion (Akaike 1973) and Bayesian information criterion (Schwarz 1978).

$$AIC = -2L(\mathbf{x}_i, \mu_{\mathbf{x}}, \mathbf{B}, \mathbf{V}, \Psi_{\mathbf{x}}, \mathbf{c}, \Psi_{\mathbf{y}}) + 2p \tag{2.71}$$

$$BIC = -2L(\mathbf{x}_i, \mu_{\mathbf{x}}, \mathbf{B}, \mathbf{V}, \Psi_{\mathbf{x}}, \mathbf{c}, \Psi_{\mathbf{y}}) + \log(n)p \tag{2.72}$$

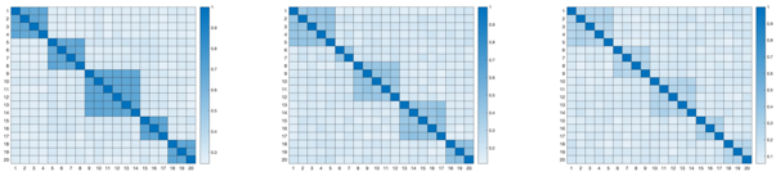
where $p = 2J + H$ is the number of parameters for HDNFA. In this formulation, the best model is the model associated with the minimum value of BIC or AIC. The RMSR (root mean square of residuals for the correlation matrix) measures the square root of the discrepancy between the sample covariance matrix and the model covariance matrix (Hooper, Coughlan, and Mullen 2008):

$$RMSR = \frac{\text{tr}[(\mathbf{S} - \Sigma_{\mathbf{x}})'(\mathbf{S} - \Sigma_{\mathbf{x}})]}{J^2} \tag{2.73}$$

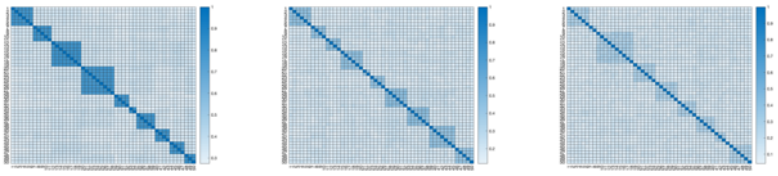
Values for RMSR range from 0 to 1[1]. In Figure 2.2 examples of generated correlation matrices with different level of errors are shown in order to allow the reader to visually appreciate the meaning of low, medium and high errors. In Figure 2.2(a) and Figure 2.2(d), correlation matrices are generated with low error: the block diagonal structure is clearly visible, whereas in Figure 2.2(b) and Figure 2.2(e) and Figure 2.2(c) and Figure 2.2(f) the correlation matrices are generated with medium and high error, respectively, and the block diagonal structure is more masked and starts to be less clear, but still visible, since otherwise the model (2.50) is not anymore true.

Table 2.1 and Table 2.2 reports the performances of the model as value of following indices: average of ARI, GFI, AGFI, BIC, AIC and RMSR on the 500 generated datasets and the percentage of ARI equal to 1. In Table 1, ARI decreases on average from 0.98 to 0.92 when the error increases and the model is reconstructed at least in 88% of cases for all

[1] Byrne (1998) considers "well-fitting" models the ones which have RMSR less than 0.05.



(a) $n = 200$, $J = 20$, $H = 5$, $\alpha = 0.3$ (b) $n = 200$, $J = 20$, $H = 5$, $\alpha = 0.5$ (c) $n = 200$, $J = 20$, $H = 5$, $\alpha = 0.8$



(d) $n = 200$, $J = 50$, $H = 10$, $\alpha = 0.3$ (e) $n = 200$, $J = 50$, $H = 10$, $\alpha = 0.5$ (f) $n = 200$, $J = 50$, $H = 10$, $\alpha = 0.8$

Figure 2.2. Heatmaps of examples of correlation matrix produced by the simulation study with different levels of error.

Table 2.1. Simulated datasets with $n = 200$, $J = 20$, $H = 5$ and different levels of error.

Criterion	Low Error	Medium Error	High Error
ARI	0.98	0.94	0.92
% ARI	95	90	88
GFI	0.88	0.84	0.82
AGFI	0.85	0.80	0.78
BIC	8339	12073	16432
AIC	8119	11852	16212
RMSR	0.04	0.06	0.08

level of error. So, the performances of the model in scenario $n = 200$, $J = 20$, $H = 5$ are appreciable even when the level of error is high, and the block diagonal structure tends to be less visible. In Table 22.2, ARI mean values are 0.99 and 0.94 for low and medium error, respectively. ARI in on average equal to 0.82 when the level of error is high.

The performances of the model in scenario $n = 200$, $J = 50$, $H = 10$ are noticeable when the level of error is low and medium, while HDNFA identifies the true partition only the 65% of all cases when the error is high. RMSR assumes value equal to 0.04 in both scenarios in settings of low error and never a value higher than 0.09 in the other settings.

The simulation study additionally examined the ability of the model to build reliable Specific Composite Indicators via Cronbach's α (Cronbach 1951) which can be seen as the expected correlation between all pairs of MIs used to specify the concept compiled by the CI, (Fornell and Larcker 1981; Kline 2000; Chaouachi and Rached 2012)[2]. Cronbach's α ranges between negative infinity and 1. Table 2.3 and Table 2.4 show evaluations of internal consistency of the specific factors estimated by HDNFA.

Looking at Table 2.4, we can state that the Cronbach's α, computed for each cluster

[2]Cronbach's α measures an internal consistency: $\alpha \geq 0.9$, excellent; $0.7 \leq \alpha < 0.9$, good; $0.6 \leq \alpha < 0.7$, acceptable; $0.5 \leq \alpha < 0.6$ poor; $\alpha < 0.5$, unacceptable

Table 2.2. Simulated datasets with $n = 200$, $J = 50$, $H = 10$ and different levels of error.

Criterion	Low Error	Medium Error	High Error
ARI	0.99	0.94	0.82
% ARI	97	91	65
GFI	0.78	0.74	0.60
AGFI	0.77	0.71	0.60
BIC	17296	24483	36044
AIC	16756	23943	35471
RMSR	0.04	0.07	0.09

Table 2.3. Percentage of times Cronbach's alpha (computed for each cluster of MIs) > 0.9 with different scenario and different levels of error.

Scenario	Low Error	Medium Error	High Error
$n = 200, J = 20, H = 5$	90.5	70.8	14.4
$n = 200, J = 50, H = 10$	86.6	84	33

of MIs, is always higher than 0.70 in settings with low and medium error for the scenario $n = 200, J = 20, H = 5$ and almost always (99.6%) for the scenario $n = 200, J = 50, H = 10$. In settings with low and medium error, we can see that the specific factors are quite often reliable (Table 2.3).

2.6 Application

The HDNFA has been applied on the OECD data on *Better Life Index 2015*, an initiative started in 2011 and focusing on statistics that capture important aspects of life for people and that characterize the quality of their lives. The dataset is formed by 34 countries on which 24 indicators have been considered reflecting two pillars that OECD has identified as essential to well-being:

Material Living Conditions (MLC) specified by three dimensions: *Housing*, *Income* and *Jobs* so characterised:

1 *Housing*:
  1 *Dwellings without basic facilities*;
  2 *Housing expenditure*;
  3 *Rooms per person*;

2 *Income*:
  4 *Household net adjusted disposable income*;
  5 *Household net financial wealth*;

Table 2.4. Percentage of times Cronbach's alpha (computed for each cluster of MIs) < 0.7 with different scenario and different levels of error.

Scenario	Low Error	Medium Error	High Error
$n = 200, J = 20, H = 5$	0	0	9.2
$n = 200, J = 50, H = 10$	0.4	0.4	6.2

3 *Jobs*:
  6 *Employment rate*;
  7 *Job security*;
  8 *Long-term unemployment rate*;
  9 *Personal earnings*;

Quality of Life (QL) identified by eight dimensions: *Community*, *Education*, *Environment*, *Civic Engagement*, *Health*, *Life Satisfaction*, *Safety* and *Work-Life Balance*, described by

4 *Community*:
  10 *Quality of social support network*;

5 *Education*:
  11 *Educational attainment*;
  12 *Student skills*;
  13 *Years in education*;

6 *Environment*:
  14 *Air pollution*;
  15 *Water quality*;

7 *Civic Engagement*:
  16 *Consultation on rule-making*;
  17 *Voter turnout*;

8 *Health*:
  18 *Life expectancy*;
  19 *Self-reported health*;

9 *Life Satisfaction*:
  20 *Life satisfaction*;

10 *Safety*:
  21 *Assault rate*;
  22 *Homicide rate*;

11 *Work-Life Balance*:
  23 *Employees working very long hours*;
  24 *Time devoted to leisure and personal care*.

MIs have been standardized. Since HDNFA is scale invariant, the min-max normalization ${}_n\mathbf{x}_{ij} = (\mathbf{x}_{ij} - \min(\mathbf{x}_{ij}))/(\max(\mathbf{x}_{ij}) - \min(\mathbf{x}_{ij}))$ can also be used, obtaining exactly the same results. This scale invariance property produces an important simplification in the composite indicators construction because the impact of the normalization in the results[3] is avoided (up to a linear transformation of MIs) and the sensitivity analysis on normalization is not any more necessary.

Since HDNFA identifies a system of non-negative loadings the first application is used to identify possible indicators that measure a negative component of the general latent construct, in our case the well-being. The number of specific constructs H is fixed equal

[3] In the Handbook on Constructing Composite Indicators OECD (2004) includes the normalization of the indicators (*Step 5*) to make the MIs comparable. Normalisation is considered a source of uncertainty that can modify the final composite indicator and therefore, OECD suggests to verify with a sensitivity analysis its impact.

to the minimum 2. If indicators have positive loadings this means that they are positive dimensions of the general latent construct. Please, note that if the positive loading is "small", that is, it is not statistically significantly different from zero, this means that a larger number of clusters H is required.

However, if H is optimally chosen and still there is an indicator with small not significant loading, this suggests discarding the indicator from the analysis because it is irrelevant into the model, and actually can be confounding for the analysis and could be removed without incurring in a significant loss of information. Finally, if some indicators have exactly zero loadings on the two factors of the HDNFA, this means that they measure a negative component of the general latent construct and therefore they must be reversed[4].

In this analysis indicators (1), (2), (7), (8), (14), (21), (22) and (23) load zero on the two factors (i.e., a row with zero loadings, for each of the above indicators was observed). In fact, these indicators measure a negative component of well-being and therefore have no positive loadings to show; thus, they have been reversed by changing the sign (i.e., by multiplying each indicator by -1). Reapplying HDNFA with $H = 2$, this time, all indicators loaded on factors with a significant positive weight unless for indicators (2) (*Housing expenditure*, with loading 0.1261, $p-value = 0.48$) and (16) (*Consultation on rule-making*, with loading 0.3014, $p-value = 0.08$). Now, in this analysis of well-being, $H = 2$ is specified by OECD by hypothesis. In fact, OECD's Better Life Initiative, presented in 2011, identifies two pillars for understanding and measuring people's well-being: the MLC and the QL; these are specific latent constructs for the HDNFA. Thus, under the hypothesis $H = 2$, the indicators (2) and (16) are irrelevant in the HDNFA and consequently have been removed. The successive analyzes have been repeated without these two indicators.

In the first application of the HDNFA, the 22 MIs have been constrained to define the MLC and QL factors as indicated by the OECD (i.e., MLC described by indicators 1,3-9; QL defined by indicators 10-15,17,24). The results of the analysis are reported in Table 2.5 columns 1 - 3. The HDNFA methodology is now a Hierarchical Confirmative "non-negative" Factor Analysis. It is a Confirmatory Factor Analysis because all indicators are forced a priori to load on the specified MLC and QL as indicated by OECD; while the analysis is non-negative since loadings are forced to be non-negative. BIC and AIC resulted 6725.06 and 6618.21, respectively. The two identified factors are not unidimensional constructs, since the eigenvalue of the second component of each of the two clusters of indicators is clearly larger than 1 and equal to 1.88 and 2.16. However, the reliability or internal consistency of each factor measured by the Cronbach's alpha is good (Kline 2000) and equal to 0.87 for both clusters. The internal consistency is the extent to which indicators, associated with a factor, measure a unique theoretical construct. However, *Job security* and *Long-term unemployment rate* do not load significantly on the MLC construct, whereas *Voter turnout* and *Self-reported health* do not load significantly on the QL. So at least these four MIs do not seem to empirically sustain the OECD definition of *Material Living Conditions* and *Quality of Life*. This observed evidence suggests verifying if the selected model is statistically confirmed by the data of 2015. A common approach for model selection used also in HDNFA is to use theory to specify an initial model, in our case the MLC and QL dimensions defined by OECD and then to use the likelihood ratio χ^2_v test to decide whether

[4]If the indicator is standardized the reversed indicator is obtained by multiplying the standardized values of the indicator by -1. If the indicator is min-max normalized the reverse indicator is obtained by subtracting from 1 the min-max normalised values.

the model is confirmed or should be simplified or expanded (Bollen 1989). However, this test is problematic in practice because is sensitive to the sample size and tends to reject good models[5]. This situation is well-known in the literature and for this reason Marsh, Balla, and McDonald (1988) and several other authors have proposed more than 30 alternative measures. Among these, we have considered two information criteria, BIC and AIC, that have proved to work well on SEM and Confirmatory Factor Analysis problems (Raftery 1995).

To examine how much the latent constructs MLC and QL change by considering a less constrained model, the HDNFA analysis was repeated by forcing this time only a single indicator for each of the 11 dimensions to load on MLC or QL as indicated by OECD. In practice the indicator, that in the first analysis had the maximal within dimension loading, was forced to load as indicated by OECD on MLC or QL [6]. The 10 constrained indicators are identified by the coloured background in columns 4 and 5 of Table 2.5. Thus, the construct MLC has: *Housing* represented by *Rooms per person* (3); *Income by Household net adjusted disposable income* (4); and *Jobs* by *Personal earnings* (9). The construct QL has: *Community* represented by *Quality of support network* (10); *Education* by *Student skills* (12); *Environment* by *Water quality* (15); *Health* by *Life expectancy* (18); *Life Satisfaction* by *Life Satisfaction* (20); *Safety* by *Homicide rate* (22); *Work-Life Balance* by *Employees working very long hours* (23). The rest of indicators was left free to load on one of the two latent constructs MLC or QL.

Thus, after that 10 of the 22 indicators have been constrained to define MLC and QL as indicated by OECD, the new HDNFA analysis should confirm that also the rest of indicators will follow the two hypothesised latent constructs. This is not true for 7 of the remaining 11 indicators. In fact, in Table 2.5, columns 4, 5 and 6 the best solution over 100 reiterations of the algorithm, starting from different random solutions, is reported. In this second analysis, the values of BIC and AIC are considerably lower and equal to 4856.39 and 4793.80, respectively. According to the study of Raftery (1995), there is a "very strong" evidence that this model is better than the previous of the OECD, since there is a difference in BIC between models larger than 10[7] ($\text{BIC}_{\text{OECD}} - \text{BIC}_{\text{HDNFA,Con}} = 1766.41$).

Furthermore, now, all indicators load significantly on the two factors. In this new application of HDNFA indicators *Employment rate*, *Job security* (reversed), *Long-term unemployment rate* (reversed) show to be better related to the *Quality of Life*, while Voter turnout, *Self-reported health* and *Assault rate* (reversed) are better related to *Material Living Conditions*. Therefore, the *Quality of Life* defined by this constrained HDNFA model includes also the need of a secure job and for a long term, while the *Material Living Conditions* include the need of a good personal health condition, the personal safety and the interest to become involved in the political process.

Now, it remains to understand if these two latent constructs now defined can be further improved by leaving all the indicators free to load on one of the two latent constructs.

[5]The likelihood ratio χ^2_v test was applied to all models in this analysis and it was found always statistically significantly different from zero, thus discarding the proposed model.

[6]The dimension *Civic engagement*(7) is characterized by only the indicator *Voter turnout* which has been resulted not statistically significant in OECD proposal, hence, this indicator was left free to load on one of the two latent constructs MLC or QL.

[7]Let $\Delta BIC = BIC_1 - BIC_2$ the difference between BIC of model M_1 and BIC of model M_2; the evidence that model M_2 is better than model M_1 is: "Weak" if $0 < \Delta BIC \leq 2$; "Positive" if $2 < \Delta BIC \leq 6$; "Strong" if $2 < \Delta BIC \leq 10$; "Very Strong" if $\Delta BIC \geq 10$.

Table 2.5. Analysis of different hierarchical two level factor analysis models for defining two dimensions of well-being: material living conditions and quality of life.

Column		1	2	3	4	5	6	7	8	9
Dimensions		2 Dim. OECD MLC_{OECD}	2 Dim OECD QL_{OECD}	Std error Pr(p>|Z|)	2 Dim Constr. MLC_{HNDFA}	2 Dim Constr. QL_{HNDFA}	Std error Pr(p>|Z|)	2 Dim Uncons. MLC_{HNDFA}	2 Dim Uncons. QL_{HNDFA}	Std error Pr(p>|Z|)
1. Housing	1	0.6106		0.135815 (0.000126)		0.6149	0.135249 (0.000109)	0.6365		0.132273 (0.000052)
	2	-	-	-	-	-	-	-	-	-
	3	0.8038		0.102030 (0.000000)	0.8131		0.099842 (0.000000)	0.8216		0.097771 (0.000000)
2. Income	4	0.9464		0.055407 (0.000000)	0.9485		0.054307 (0.000000)	0.9209		0.066863 (0.000000)
	5	0.7470		0.114014 (0.000000)	0.7431		0.114762 (0.000000)	0.7058		0.121491 (0.000003)
3. Jobs	6	0.6306		0.133104 (0.000064)		0.8511	0.090032 (0.000000)		0.8473	0.091090 (0.000000)
	7	0.2726*		0.165006 (0.118878)		0.4158	0.155967 (0.014446)		0.4457	0.153526 (0.008255)
	8	0.3333*		0.161694 (0.054098)		0.3925	0.157738 (0.021681)		0.4105	0.156383 (0.015889)
	9	0.9438		0.056679 (0.000000)	0.9412		0.057950 (0.000000)	0.9631		0.046170 (0.000000)
4. Community	10		0.6857	0.124834 (0.000008)		0.6032	0.136784 (0.000159)	0.5635		0.141673 (0.000520)
5. Education	11		0.6217	0.134321 (0.000087)		0.5764	0.140142 (0.000360)		0.5370	0.144676 (0.001059)
	12		0.6759	0.126396 (0.000012)		0.5584	0.142274 (0.000600)	0.4393		0.154067 (0.009345)
	13		0.6151	0.135217 (0.000108)		0.5297	0.145462 (0.001274)		0.5389	0.144460 (0.001007)
6. Environment	14		0.5632	0.141718 (0.000525)		0.5185	0.146649 (0.001683)		0.4782	0.150617 (0.004228)
	15		0.7816	0.106973 (0.000000)		0.9540	0.051430 (0.000000)		0.9902	0.023952 (0.000000)
7. Civic engagement	16		-	-	-	-	-	-	-	-
	17		0.2066*	0.167797 (0.240983)	0.4281		0.154985 (0.011534)	0.4276		0.155026 (0.011643)
8. Health	18		0.5864	0.138913 (0.000268)		0.5525	0.142949 (0.000704)	0.6986		0.122707 (0.000004)
	19		0.2901*	0.164123 (0.096048)	0.5163		0.146870 (0.001772)	0.5331		0.145102 (0.001171)
9. Life Satisfaction	20		0.4384	0.154141 (0.009505)		0.6218	0.134309 (0.000086)	0.7111		0.120588 (0.000002)
10. Safety	21		0.6339	0.132641 (0.000057)	0.3813		0.158543 (0.026089)	0.3927		0.154524 (0.010377)
	22		0.6687	0.127509 (0.000015)		0.4207	0.155587 (0.013242)		0.3879	0.157719 (0.023400)
11. Work-Life Balance	23		0.6702	0.127285 (0.000015)		0.5479	0.143468 (0.000796)		0.5132	0.147196 (0.001912)
	24		0.5393	0.144423 (0.000998)		0.4020	0.157027 (0.018419)	0.3673		0.159510 (0.032596)
	CI	0.8864	0.8909		0.9094	0.8772		0.8834	0.8805	
Communality		3.9463	4.8719		3.5941	5.2274		5.6142	3.2868	
Cronbach's alpha		0.87	0.87	-	0.88	0.85		0.90	0.82	
Unidimensionality		No 1.876920	No 2.160364	-	No 2.885669	No 1.130160		No 2.2414	No 1.7969	
BIC		6725.06		-	4958.65			4741.67		
AIC		6618.21			4851.8			4634.83		
Discrepancy		192.498			140.544			134.163		
Total Communality		10.3976			10.4180			10.4567		

Therefore, HDNFA was newly applied for the third time, but without constraints on the indicators. The best solution over 100 runs was retained and the results were reported in Table 2.5, columns 7, 8 and 9. The values of BIC and AIC still reduce to 4741.67 and 4634.83, even though not so strongly as in the second application of HDNFA. However, the difference of BIC ($BIC_{HDNFA,Con} - BIC_{HDNFA,Unc} = 216.98$) is still very relevant. In the unconstrained model the *Life Satisfaction*, *Community* (*Quality of support network*), *Students Skills* and *Time devoted to leisure and personal care*, differently from the OECD model, load on MLC.

Furthermore, *Life expectancy* - the other MI of the *Health* dimension - loads on MLC. Therefore, the MLC construct of the unconstrained HDNFA model includes also *Life Satisfaction*, *Community* and part of the *Work-Life Balance*. It is worth to observe that the clusters of specific indicators proposed by OECD to define the 11 dimensions are frequently "incoherent" in the sense that they load on different latent constructs. For example, the two indicators *Employees working very long hours* and *Time devoted to leisure and personal care* that define the dimension *Work-Life Balance* actually load the first on MLC and the second on QL. The same happens for *Safety*, *Education* and *Jobs*. So, this suggests that the proposed indicators when considered all together are not coherent with the hypothesized theoretical construct.

The analysis was repeated for different values of H in order to find the solution with the lowest BIC and AIC in order to understand it there are more than two pillar dimensions (MLC and QL), and actually to confirm if there are the 11 specific theoretical dimensions indicated by OECD. For $H = 5$, BIC and AIC were the lowest with 4007.78 and 3791.04, respectively. The results of this HDNFA is reported in Table 2.6[8].

This time the five factors are almost all unidimensional since the variance of the second component of each of the five clusters is not strongly larger than 1 and in four clusters is actually lower (1.024, 0.619, 0.770, 0.795, 0.261). The Cronbach's α is equal to 0.88, 0.86, 0.82, 0.77, 0.85 respectively, thus confirming the good internal consistency of factors.

The first latent construct is mainly MLC including in addition *Civic Engagement* (*Voter turnout*), *Health* (*Life expectancy*), *Life Satisfaction* and *Work-Life Balance*. This factor is mainly formed by a cluster of indicators defining the MLC in the 2 factors unconstrained solution of HDNFA. It is interesting to observe that there is another factor for MLC, the fifth, dedicated to the dimension *Jobs* with *Job insecurity* (reversed) and *Long-term unemployment rate* (reversed). The other three factors of the HDNFA actually split the QL into specific factors. In fact, the second factor of HDNFA is formed by *Education* (*Educational attainment*, *Student skills*) and *Safety* (*Assault rate*, *Homicide rate* both reversed) and therefore measures the level of safety and education in the *Society* as a relevant aspect of the QL. The third factor measures the *Quality of the Society* in terms of *Community* (*Quality of social support network*), *Environment* (*Water quality*), *Years in education* and *Employment rate*. Therefore, there is a quality factor on fundamental aspects for the *Society* such the quality of the social network support, the quality of the water, the quality of the education and the quality of the working life. The fourth factor measures the *Quality of Habitat for Humanity* based on *Dwellings without basic facilities* (reversed), *Air pollution* (reversed), *Self-reported health* and *Employees working very long hours* (reversed). Thus, it measures the quality of the habitat for the humanity based on decent and affordable housing, good air quality, the

[8]Note that indicators (2) and (16) were re-inserted in the HDNFA analysis; however, these increase drastically BIC and AIC and therefore were discarded again.

Table 2.6. Analysis of different hierarchical two level factor analysis models for defining five dimensions of well-being.

Column		1	2	3	4	5	6
Dimensions		MLC_{HDNFA}	ES_{HDNFA}	QS_{HDNFA}	$QSHH_{HDNFA}$	J_{HDNFA}	Std error Pr(p>|Z|)
1. Housing	1				0.9411		0.132273 (0.000052)
	2	-	-	-		-	-
	3	0.8068					0.097771 (0.000000)
2. Income	4	0.9238					0.066863 (0.000000)
	5	0.7188					0.121491 (0.000003)
3. Jobs	6			0.8409			0.091090 (0.000000)
	7					0.8596	0.153526 (0.008255)
	8					0.8596	0.156383 (0.015889)
	9	0.9673					0.046170 (0.000000)
4. Community	10			0.5634			0.141673 (0.000520)
5. Education	11		0.6573				0.144676 (0.001059)
	12		0.7579				0.154067 (0.009345)
	13			0.5456			0.144460 (0.001007)
6. Environment	14				0.5808		0.150617 (0.004228)
	15			0.9977			0.023952 (0.000000)
7. Civic engagement	16		-			-	-
	17	0.4410					0.155026 (0.011643)
8. Health	18	0.6943					0.122707 (0.000004)
	19				0.4863		0.145102 (0.001171)
9. Life Satisfaction	20	0.7081					0.120588 (0.000002)
10. Safety	21		0.9249				0.154524 (0.010377)
	22		0.7591				0.157719 (0.023400)
11. Work-Life Balance	23				0.7037		0.147196 (0.001912)
	24	0.3496					0.159510 (0.032596)
	CI	0.9244	0.3451	0.6489	0.8215	0.7697	
Communality		4.2569	2.4380	2.3176	1.9546	1.4779	
Cronbach's alpha		0.88	0.86	0.82	0.77	0.85	
Unidimensionality		No 1.024	Yes 0.619	Yes 0.770	Yes 0.795	Yes 0.261	
BIC		4007.78					
AIC		3791.04					
Discrepancy		104.17					
Total Communality		12.445 + 2.6619 = 15.1070					

perceived health in the habitat and the working stability of the habitat.

2.7 Conclusion

In the OECD (2004), the Factor Analysis (FA) methodology is proposed as a weighting method used to combine observed indicators.

In order to use FA with this aim, four steps are necessary:

1. choose H and extract H factors;
2. rotate factors to obtain a simpler structure easier to interpret;
3. apply soft thresholding and compute the normalization of the truncated loadings;
4. aggregate the H loadings by a linear combination with weights equal to the normalized truncated loadings.

The use of FA has the advantage to define loadings that best reconstruct the observed indicators according to the common factor model estimated by Maximum Likelihood Estimation (MLE) method. Therefore, weights are not subjective, but statistically estimated summarizing the observed common relation among data. However, in this procedure there are some subjective choices that complicate the CI construction: in step 2 the choice of type of rotation (varimax, equimax, orthomax, etc); in step 3 the selection of the size of the threshold for the loadings that identifies the correlation structure of the data. Furthermore, the final CI is not a general latent construct that best reconstruct the observed indicators and estimated in the MLE context, but it is the weighted mean - which is known to be subject to compensability problems - of the H specific factors. The choice of the type of rotation and the use of soft thresholds are already discussed in Section 1.3.

An alternative method is the Hierarchical Confirmatory Factor Analysis, that in a unique model estimates the specific factors and the general factor representing the CI. However, this methodology needs a lot of a priori information on the most relevant relations in the hierarchy associated with the CI. To avoid the subjective choices necessary to apply FA and a large amount of a priori hypotheses on the CI structure, Hierarchical Disjoint Non-Negative Factor Analysis has been proposed. It allows estimating, by using MLE, the complete system of weights for the composite indicator that best reconstruct the data according to the two level Hierarchical Factor Analysis model. The methodology is scale invariant, therefore standardization, min-max normalisation or other linear transformation of the data used to make the data commensurable can be used without producing differences in the results.

It is worth to remark that HDNFA is a new method to model composite indicators generalizing both hierarchical factor analysis and non-orthogonal disjoint factor analysis with the addictional constrains of non-negativity of loadings. The paper provides a new two-level hierarchical structure for FA, a recalling of disjoint models and then the complete model with its estimations. According to the distributional assumptions, the maximum likelihood estimation of the model allow us to make inference on the parameters. The estimation of all parameters is simultaneous following a coordinated descendent algorithm, since this is a discrete and continuous problem that cannot be solved by a quasi-Newton type algorithm. The algorithm has been implemented in a MATLAB routine.

In our research, some further challenges for HDNFA are foreseen. Our goal is develop the inclusion of cross-loadings in the SSM in order to increase the fit of the model using the work made by Vichi (2017) and the inclusion of a simultaneous clustering model into HDNFA in order to classify the statistical units (e.g., countries) as well. Another important achievement is to extend the model with a time component in order to conduct analysis through several occasions (e.g., years). Finally, we are working in order to make the MATLAB routine available for free.

Chapter 3

Models, properties and features for tracking coherent policy conclusions

3.1 Introduction

The issue of the construction of CIs has been discussed many times by many authors. In the latter chapter, a model-based approach for the construction of CIs based on a hierarchical factor model has been proposed. A statistical model is required to simplify and synthesize the complexity of the reality by means of a mathematically-formalized reconstruction of the observed data and their main relations. Models are needed for investigating phenomena in order to acquire new knowledge based on empirical evidence and are also needed in the development process used to specify the appropriate CI model for the studied phenomenon. The use of a model-based CI classifies the approach as *statistical* because hypotheses formulated in the specification of the CI are empirically tested. The methodology wishes to be objective and transparent, avoiding - as far as possible - the use of arbitrary choices that cannot be tested.

Moreover, it is necessary to introduce a body of properties and features that a good CI should have in order to avoid poorly constructed indexes that induce simplistic and misleading policy conclusions. CIs (e.g., Human Development Index (HDI) and Multiple Poverty Index (MPI)) are often built without evaluating how much they reconstruct the MIs (i.e., manifest data), and without taking into account some important properties (e.g., unidimensionality, non-compensability, scale-invariance, etc.).

The *reflective* measurement model for CIs, presented in Chapter 2, is a hierarchical factor model with disjoint clusters of MIs and non-negativity constraint on the loadings. The current chapter is framed in the same setting and provides useful tools in order to build well-done CIs. Hence, we propose an alternative estimation for the HDNFA model (in particular, we consider Least-Squares Estimation - LSE), and, in order to assess the methodology of construction of a CI, a set of properties is considered. Our proposal includes the most frequently used approaches in the literature of CIs, and it provides an evaluation of the model to assess their performances. An extensive study of our GCI compares its perfomances with the ones of some well-known CIs such as the HDI and the MPI.

The advantage of using a model-based approach is that the researcher can assess the model at each level of the hierarchy. This characteristic is very important because the model should fit the objectives and intentions of the user, such that, it must be the most appropriate

tool for expressing the set of objectives that motivate the whole exercise. The choice of the sub-indicators which should be used, their division into clusters, whether a normalization method (and which one) has to be used or not, the choice of the weighting method, and how information is aggregated; all these features stem from a certain specific perspective of the user and its decisions on how to model the issue. However, all these choices on the model must be empirically tested in order to verify if the used model correctly describes the reality under study. The model of the phenomenon will reflect not only the characteristics of the real system but also the choices made by the researchers on how to observe the reality and how to represent it. Thus, the empirical testing phase is needed to verify if the decisions are actually distorting the reality.

The paper is organised as follows. The Least-Squares Estimation of Hierarchical Disjoint Non-Negative Factor Analysis model is introduced in Section 3.2. The goodness of fit of a CI is shown in Section 3.3. Section 3.4 introduces an overwiew of confimatory, exploratory and mixed model selection of a model-based CI. A set of important properties are considered in Section 3.5. Section 3.6 presents two applications: HDI and MPI. A final discussion completes the paper in Section 3.7.

3.2 Model-Based CI: Least-Squares Estimation of HDNFA

3.2.1 Model and LS estimation

In the Chapter 2, a FA model has been proposed with its ML estimation. The model has been used to construct the GCI via the detection of SCIs as specific factors, the model can be consider as a nested factorial model with a disjoint structure where the estimation of all parameters is simultaneous and not sequential.

Thus, in order to identify the SCI, we consider the Disjoint Factor Analysis model, given by Vichi (2017), using the same notation present in Chapter 2

$$\mathbf{X} = \mathbf{Y}\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{X}} \tag{3.1}$$

where $\mathbf{V}$ is the $(J \times H)$ membership matrix specifying the relations between the J MIs and H SCIs, whereas the $(J \times J)$ diagonal matrix $\mathbf{B}$ specifies on the weights associated with the MIs. In case model 3.1 is estimated according to MLE, this corresponds to evaluate a FA with the constraint that the loading matrix has a simple form of the type $\mathbf{BV}$; thus, each MI has a relationship only with one SCI, which corresponds to suppose that not relevant cross-loadings are observed in the data.

Let us now formalize the most used formulation of model 3.1, when $\mathbf{V}$ and $\mathbf{B}$ are fixed, that is when a theory known by the researcher formalizes the relations between MIs and SCIs and therefore it is only necessary to estimate the intensities of these relations. The LS estimation of model 3.1, with $\mathbf{V}$ and $\mathbf{B}$ fixed (indicated by $\widehat{\mathbf{V}}$,$\widehat{\mathbf{B}}$) minimizes $||\mathbf{X} - \mathbf{Y}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2$ with respect to $\mathbf{Y}$, by

$$\widehat{\mathbf{Y}} = \mathbf{X}(\widehat{\mathbf{V}}'\widehat{\mathbf{B}})^{+} \tag{3.2}$$

hence, each SCI (column of $\mathbf{Y}$), $\mathbf{y}_h = \mathbf{x}_1 v_{1h} b_{11} + \mathbf{x}_2 v_{2h} b_{22} + \cdots + \mathbf{x}_J v_{Jh} b_{JJ}$, is the weighted sum of MIs (columns of $\mathbf{X}$) $\mathbf{x}_1, \ldots, \mathbf{x}_J$, selected by $\widehat{\mathbf{V}}$ and weighted according $\widehat{\mathbf{B}}$. If we normalize $\widehat{\mathbf{B}}$ such that $\mathrm{tr}(\widehat{\mathbf{B}}) = 1$, the columns of $\mathbf{Y}$ are the *weighted arithmetic means* of the MIs assigned to the SCIs. When MIs are standardized weights can be equal to the proportion of MIs defining each SCI.

We can now consider the second level models, which is nested in the previous one, by considering

$$\mathbf{Y} = \mathbf{g}\mathbf{c}' + \mathbf{E}_{\mathbf{Y}} \tag{3.3}$$

and including it into the model 3.1. Thus, we have the GCI model presented in Chapter 2 (Cavicchia and Vichi 2020a),

$$\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B} + \mathbf{E} \tag{3.4}$$

where $\mathbf{g}$ is the $(n \times 1)$ vector of the GCI, which we hypothesize standardized, $\mathbf{g}\mathbf{c}'$ is the $(n \times H)$ matrix of the H SCI, $\mathbf{V}$ is the $(J \times H)$ binary and row stochastic matrix, $\mathbf{B}$ is the $(J \times J)$ diagonal matrix. $\mathbf{E} = \mathbf{E}_{\mathbf{Y}}\mathbf{V}'\mathbf{B} + \mathbf{E}_{\mathbf{X}}$ is the residual matrix which has two typologies of error, one depending on the SCI and the other regarding the MIs. Model in (3.4) has been estimated by using MLE in Chapter 2; we now consider LS estimation.

Vector $\mathbf{c}$, and matrices $\mathbf{V}$, $\mathbf{B}$, are the parameters of the CI model 3.4. The elements of $\mathbf{c}$ represent the weight between each SCI and the GCI; whereas, the elements of the matrix $\mathbf{BV}$ are the weights between each MI and the related SCI. The least-squares estimation of model 3.4 corresponds to minimize

$$F(\mathbf{g}, \mathbf{c}, \mathbf{V}, \mathbf{B}) = ||\mathbf{X} - \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B}||^2 \rightarrow \min_{\mathbf{g},\mathbf{c},\mathbf{V},\mathbf{B}} \tag{3.5}$$

Let matrix $\mathbf{V}$, $\mathbf{c}$, and $\mathbf{B}$, be fixed by the researcher, that is, a cluster of MIs specified by columns of $\widehat{\mathbf{V}}$ describes a SCI with weights reported in $\widehat{\mathbf{B}}$ and each SCI characterizes the GCI with the weight indicated by $\widehat{\mathbf{c}}$.

Thus, we have to minimize with respect to $\mathbf{g}$

$$F(\mathbf{g}) = ||\mathbf{X} - \mathbf{g}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 \rightarrow \min_{\mathbf{g}} \tag{3.6}$$

The minimization problem is equivalent to the solution of a multivariate regression problem given by

$$\widehat{\mathbf{g}} = \mathbf{X}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})^{+} = \mathbf{X}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1} \tag{3.7}$$

For $\widehat{\mathbf{c}} = \mathbf{1}_H$ and $\widehat{\mathbf{B}} = \mathbf{I}_J$, and any partition $\widehat{\mathbf{V}}$, the GCI in (3.7) reduces to

$$\widehat{\mathbf{g}}_M = \mathbf{X}(\mathbf{1}_H'\widehat{\mathbf{V}}')^{+} = \mathbf{X}\mathbf{1}_J'^{+} = \frac{1}{J}(\mathbf{x}_1 + \cdots + \mathbf{x}_J) \tag{3.8}$$

the arithmetic mean of the MIs, $\mathbf{x}_1, \ldots, \mathbf{x}_J$, that is, the columns of the matrix $\mathbf{X}$. The GCI is the linear combination with equal weights $(\frac{1}{J})$ of the MIs. This GCI has been frequently used in the literature of the CI, also for the first releases of the Human Development Index. It is important to notice that the use of the GCI in (3.8) deletes the hierarchy of the model, the GCI is the arithmetic mean of the MIs for any matrix $\widehat{\mathbf{V}}$, thus the intermediate levels of the hierarchy have not to be defined.

If $\widehat{\mathbf{g}}$ in (3.7) is multiplied by the constant $(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\mathbf{1}_J)^{-1}$, we have

$$\widehat{\mathbf{g}}_{WM} = \mathbf{X}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\mathbf{1}_J)^{-1} \tag{3.9}$$

the weighted arithmetic mean of the MIs, $\mathbf{x}_1, \ldots, \mathbf{x}_J$. One of the most famous example of GCI computed as weighted mean of the MIs is the Multidimensional Poverty Index.

Thus, $\widehat{\mathbf{g}}_{WM}$ and $\widehat{\mathbf{g}}$ are related by

$$\widehat{\mathbf{g}}_{WM} = \frac{\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}}{\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\mathbf{1}_J}\widehat{\mathbf{g}} \tag{3.10}$$

And, a strict relation between $\widehat{\mathbf{g}}$ and $\widehat{\mathbf{Y}}$ exists

$$\widehat{\mathbf{g}} = \widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1} \tag{3.11}$$

Let us now consider a less restricted by hypotheses CI, where also $\mathbf{c}$ and $\mathbf{B}$ have to be estimated. Therefore, we hypothesize that a theory specifies the relations between MIs and SCIs, that is, only $\mathbf{V}$ is fixed a priori. Thus, we leave $\mathbf{c}$ and $\mathbf{B}$ free to be estimated [1], minimizing

$$F(\mathbf{c},\mathbf{B}) = ||\mathbf{X} - \widehat{\mathbf{g}}\mathbf{c}'\widehat{\mathbf{V}}'\mathbf{B}||^2 \rightarrow \min_{\mathbf{c},\mathbf{B}} \tag{3.12}$$

Thus, the solution of (3.12) for $\mathbf{c}$ corresponds to a multivariate regression problem

$$\widehat{\mathbf{c}} = \widehat{\mathbf{Y}}'(\widehat{\mathbf{g}}')^{+} = (\widehat{\mathbf{B}}\widehat{\mathbf{V}})^{+}\mathbf{X}'(\widehat{\mathbf{g}}')^{+} \tag{3.13}$$

and the solution of (3.12) for $\mathbf{B}$ is equivalent to a solution of a *Parafac* (Harshman 1970; Bro 1997) problem

$$\widehat{\mathbf{B}} = \mathrm{diag}(\widehat{\mathbf{U}}^{+}\mathrm{vec}(\mathbf{X})) \tag{3.14}$$

where $\widehat{\mathbf{U}}$ is the $(nJ \times J)$ matrix having the Kronecker products of the corresponding columns of $\mathbf{I}_J$ and $\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'$ as columns (J.M.F. ten Berge 2005). By examining (3.14), we can rewrite each element of the diagonal matrix $\widehat{\mathbf{B}}$ as

$$\widehat{\mathbf{b}}_j = \mathbf{P}_j^{+}\mathbf{X}_j \tag{3.15}$$

where $\mathbf{P} = \widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'$ and $\mathbf{P}_j$, $\mathbf{X}_j$ are the j-th columns of matrices $\mathbf{P}$ and $\mathbf{X}$, respectively.

Model 3.4 has not a unique solution. In fact, each weight for the MIs is the product of an element of $\mathbf{B}$ and an element of $\mathbf{c}$, so the same weight for a MI could be given by an infinite number of combinations of two values. To solve this problem, we propose to introduce the constraints of orthonormality on the loadings $\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}} = \mathbf{I}_H$ and $\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}} = 1$, the same considered in PCA (Pearson 1901; Hotelling 1933).

It is important to notice that the minimization of $F(\mathbf{g},\mathbf{c},\mathbf{V},\mathbf{B})$, (3.5), corresponds to maximize $f = \mathrm{tr}(\mathbf{BVcc}'\mathbf{V}'\mathbf{BX}'\mathbf{XBVcc}'\mathbf{V}'\mathbf{B})$, since the following deviance decomposition holds:

$$\begin{aligned} SS_{tot} &= SS_{mod} + SS_{res} \\ &= ||\mathbf{X}||^2 = ||\mathbf{gc}'\mathbf{V}'\mathbf{B}||^2 + ||\mathbf{X} - \mathbf{gc}'\mathbf{V}'\mathbf{B}||^2 \\ &= \mathrm{tr}(\mathbf{X}'\mathbf{X}) = \mathrm{tr}(\mathbf{BVcg}'\mathbf{gc}'\mathbf{V}'\mathbf{B}) + \mathrm{tr}((\mathbf{X} - \mathbf{gc}'\mathbf{V}'\mathbf{B})'(\mathbf{X} - \mathbf{gc}'\mathbf{V}'\mathbf{B})) \end{aligned} \tag{3.16}$$

Let us consider $\mathbf{A} = \mathbf{BV}$, f can be simplified:

$$f = \mathrm{tr}(\mathbf{Acc}'\mathbf{A}'\mathbf{X}'\mathbf{XAcc}'\mathbf{A}') = \mathrm{tr}(\mathbf{c}'\mathbf{A}'\mathbf{Acc}'\mathbf{A}'\mathbf{X}'\mathbf{XAc}) = \mathrm{tr}(\mathbf{c}'\mathbf{A}'\mathbf{X}'\mathbf{XAc}) \tag{3.17}$$

Let us consider the transformations $\mathbf{A}^* = \mathbf{AT}$ and $\mathbf{c}^* = \mathbf{T}'\mathbf{c}$ for any arbitrary orthogonal matrix $\mathbf{T}$. This transformation does not change the function f (3.17) since

$$f^* = \mathrm{tr}((\mathbf{c}^*)'(\mathbf{A}^*)'\mathbf{X}'\mathbf{XA}^*\mathbf{c}^*) = \mathrm{tr}(\mathbf{c}'\mathbf{TT}'\mathbf{A}'\mathbf{X}'\mathbf{XATT}'\mathbf{c}) = \mathrm{tr}(\mathbf{c}'\mathbf{A}'\mathbf{X}'\mathbf{XAc}) = f \tag{3.18}$$

However, the transformation $\mathbf{A}^*$ with an arbitrary orthogonal matrix $\mathbf{T}$ produces a non-admissible loadings matrix that does not satisfy the constraint $\mathbf{A} = \mathbf{BV}$. The only admissible matrix $\mathbf{T}$ is a permutation matrix or a permutation matrix with negative elements as in a reflection transformation. Thus, constraints of orthonormality on the loadings are sufficient restrictions to uniquely factorize.

[1] The relations between MIs and SCIs are known. The magnitude of the relations between MIs and SCIs and between SCIs and GCI has to be estimated.

3.2.2 The correlation matrix of the model

Let us now consider the correlation matrix of model 3.4, given the idempotent centering matrix $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$,

$$\Sigma_{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{J}\mathbf{X} = \mathbf{B}\mathbf{V}[\mathbf{c}(\frac{1}{n}\mathbf{g}'\mathbf{J}\mathbf{g})\mathbf{c}' - \frac{1}{n}\mathbf{E}'_{\mathbf{Y}}\mathbf{E}_{\mathbf{Y}}]\mathbf{V}'\mathbf{B} + \frac{1}{n}\mathbf{E}'_{\mathbf{X}}\mathbf{E}_{\mathbf{X}} = \mathbf{B}\mathbf{V}\Sigma_{\mathbf{Y}}\mathbf{V}'\mathbf{B} + \Psi_{\mathbf{X}} \quad (3.19)$$

where

$$\Sigma_{\mathbf{Y}} = \mathbf{c}\mathbf{c}' + \Psi_{\mathbf{Y}} \quad (3.20)$$

It is assumed that $\mathbf{g} \sim N(0,1)$ and $\mathbf{E}_{\mathbf{Y}} \sim N_H(\mathbf{0}, \Psi_{\mathbf{Y}})$, where $\Sigma_{\mathbf{E}_{\mathbf{Y}}} = \Psi_{\mathbf{Y}}$ is the H dimensional diagonal positive definite variance-covariance matrix of the error of SCIs and $\mathbf{E}_{\mathbf{X}} \sim N_J(\mathbf{0}, \Psi_{\mathbf{X}})$, where $\Sigma_{\mathbf{E}_{\mathbf{X}}} = \Psi_{\mathbf{X}}$. In addition, it is assumed that errors in the two models are uncorrelated $\Sigma_{\mathbf{E}_{\mathbf{X}}\mathbf{E}_{\mathbf{Y}}} = 0$; and errors and factors are uncorrelated, i.e.,$\Sigma_{\mathbf{g}\mathbf{E}_{\mathbf{X}}} = 0$ and $\Sigma_{\mathbf{g}\mathbf{E}_{\mathbf{Y}}} = 0$[2].

It is important to notice that the choice of the weights can be related to the correlation structure of $\mathbf{X}$. Thus, values into the vector $\mathbf{c}$ represent correlations between SCIs and GCI, whereas values on the diagonal of matrix $\mathbf{B}$ are correlations between MIs and SCIs. Therefore, the correlation matrix $\Sigma_{\mathbf{X}}$ has on the main diagonal blocks representing the correlation within the clusters of MIs and outside the main diagonal the correlation between the clusters of MIs. By constraining each value of $\mathbf{B}$ to be larger than each value of $\mathbf{c}$, we can guarantee that the correlation into the clusters of MIs is higher than outside and we can guarantee the presence of the hierarchy.

According to the content of Section 1.3, the measurement model 3.4 is reflective since SCIs are considered as causing the MIs and GCI is considered as causing the SCIs.

3.3 Goodness of fit of the GCI model

With the proposed model formulation of the CIs, we can now evaluate the Goodness of Fit (GoF) of the CI model (3.4). GoF provides a measure of the extent to which the observed MIs are reconstructed by the CI model, based on the proportion of total variation of MIs explained by the CI model (3.4).

$$R^2_{GCI} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\widehat{\mathbf{g}}'\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})}{\text{tr}(\mathbf{X}'\mathbf{X})} \quad (3.21)$$

where

$$SS_{tot} = ||\mathbf{X}||^2 = \text{tr}(\mathbf{X}'\mathbf{X}) \quad (3.22)$$

$$SS_{mod} = ||\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 = \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\widehat{\mathbf{g}}'\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \quad (3.23)$$

$$\begin{aligned} SS_{res} = ||\mathbf{X} - \widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{X}'\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\widehat{\mathbf{g}}'\mathbf{X}) + \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\widehat{\mathbf{g}}'\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\widehat{\mathbf{g}}'\widehat{\mathbf{g}}\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \end{aligned} \quad (3.24)$$

If $\widehat{\mathbf{V}}\widehat{\mathbf{c}} = \mathbf{1}^+_J$ with $\widehat{\mathbf{c}} = \mathbf{1}_H$ and $\widehat{\mathbf{B}} = \mathbf{I}_J$, the GCI is $\widehat{\mathbf{g}}_M$ (i.e., the arithmetic mean of the MIs) and the GoF is

$$R^2_{GCI} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{1}'_J\mathbf{X}'\mathbf{X}\mathbf{1}'_J{}^+)}{\text{tr}(\mathbf{X}'\mathbf{X})} \quad (3.25)$$

[2]Same assumptions of HDNFA.

where

$$SS_{modM} = ||\widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 = \text{tr}(\mathbf{1}_J'\mathbf{X}'\mathbf{X}\mathbf{1}_J'^{+}) \tag{3.26}$$

$$\begin{aligned} SS_{resM} = ||\mathbf{X} - \widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 &= \text{tr}((\mathbf{X} - \widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})'(\mathbf{X} - \widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})) = \text{tr}((\mathbf{X} - \widehat{\mathbf{g}}\mathbf{1}_J')'(\mathbf{X} - \widehat{\mathbf{g}}\mathbf{1}_J')) \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) + \text{tr}(\mathbf{1}_J'\mathbf{X}'\mathbf{X}\mathbf{1}_J'^{+}) - \text{tr}(\mathbf{X}'\mathbf{X}\mathbf{1}_J^{+}\mathbf{1}_J') - \text{tr}(\mathbf{X}\mathbf{1}_J'^{+}\mathbf{1}_J'\mathbf{X}') \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{1}_J'\mathbf{X}'\mathbf{X}\mathbf{1}_J'^{+}) \end{aligned} \tag{3.27}$$

Thus, we can prove for both cases the following decomposition

$$\begin{aligned} ||\mathbf{X}||^2 &= ||\widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 + ||\mathbf{X} - \widehat{\mathbf{gc}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 \\ SS_{tot} &= SS_{mod} + SS_{res} \end{aligned} \tag{3.28}$$

Therefore, it follows that

$$R^2_{GCI} = \frac{SS_{mod}}{SS_{tot}} \tag{3.29}$$

is the proportion of variance due to the CI model.

GoF is minimum when $R^2_{GCI} = 0$, that is, when $SS_{res} = SS_{tot}$, and $SS_{mod} = 0$; whereas GoF is maximum when $R^2_{GCI} = 1$, that is when $SS_{mod} = SS_{tot}$, and $SS_{res} = 0$.

Is is also useful to assess the proportion of variance accounted for each SCI, with respect to model 3.1

$$R^2_{SCI} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}})}{\text{tr}(\mathbf{X}'\mathbf{X})} \tag{3.30}$$

where

$$SS_{tot} = ||\mathbf{X}||^2 = \text{tr}(\mathbf{X}'\mathbf{X}) \tag{3.31}$$

$$SS_{modY} = ||\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 = \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \tag{3.32}$$

$$\begin{aligned} SS_{resY} = ||\mathbf{X} - \widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{X}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\mathbf{X}) + \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) \end{aligned} \tag{3.33}$$

Also in this case, if $\widehat{\mathbf{V}}\widehat{\mathbf{c}} = \mathbf{1}_J^+$ with $\widehat{\mathbf{c}} = \mathbf{1}_H$ and $\widehat{\mathbf{B}} = \mathbf{I}_J$, we have the arithmetic mean of the MIs of each SCI and the GoF is

$$SS_{modYM} = ||\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 = \text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}) = \text{tr}(\widehat{\mathbf{V}}\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\widehat{\mathbf{V}}') \tag{3.34}$$

$$\begin{aligned} SS_{resYM} = ||\mathbf{X} - \widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}}||^2 &= \text{tr}((\mathbf{X} - \widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}})'(\mathbf{X} - \widehat{\mathbf{Y}}\widehat{\mathbf{V}}'\widehat{\mathbf{B}})) = \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) + \text{tr}(\widehat{\mathbf{V}}^+\mathbf{X}'\mathbf{X}\widehat{\mathbf{V}}) - \text{tr}(\mathbf{X}'\mathbf{X}\widehat{\mathbf{V}}'^{+}\widehat{\mathbf{V}}') - \text{tr}(\widehat{\mathbf{V}}\mathbf{X}'\mathbf{X}\widehat{\mathbf{V}}'^{+}) \\ &= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\widehat{\mathbf{V}}^+\mathbf{X}'\mathbf{X}\widehat{\mathbf{V}}) \end{aligned} \tag{3.35}$$

Each SCI has its own R^2

$$R^2_{SCI_h} = 1 - \frac{SS_{resY_h}}{SS_{tot_h}} = 1 - \frac{\text{tr}(\mathbf{X}_h'\mathbf{X}_h) - \text{tr}(\widehat{\mathbf{B}}_h\widehat{\mathbf{V}}_h\widehat{\mathbf{Y}}_h'\widehat{\mathbf{Y}}_h\widehat{\mathbf{V}}_h'\widehat{\mathbf{B}}_h)}{\text{tr}(\mathbf{X}_h'\mathbf{X}_h)} \tag{3.36}$$

where $\mathbf{X}_h$ is the sub-matrix of the MIs corresponding to the h-th SCI, and $\mathbf{B}_h$ is the diagonal matrix with the weight of the MIs on the diagonal corresponding to the h-th SCI.

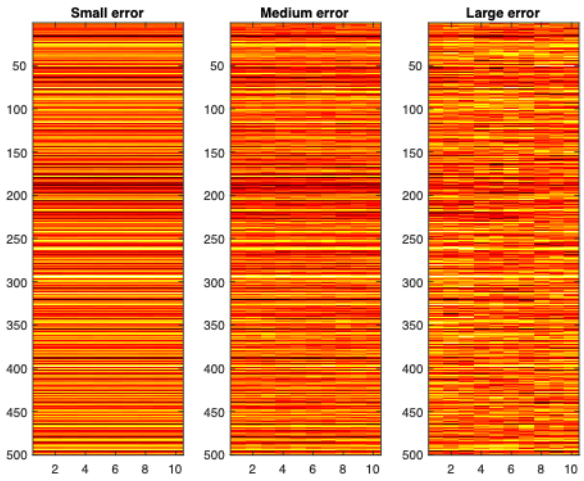


Figure 3.1. Heatmap of 3 data matrices $\mathbf{X}$ with (left) small, (center) medium and (right) large errors.

Example 1. *Let us consider the vector* $\mathbf{g}$ *as a normalized normal random vector of* 15 *statistical units, the vector* $\mathbf{c}$ *as a vector of* 3 *weights equal to* 1 *and the matrix* $\mathbf{B}$ *as an identity matrix of order* 10*, that is with all weights equal to* 1*. So, we have* 10 *MIs explained by* 3 *SCIs with the constraint that each MI can be explained by one SCI, only.*

Let us consider the matrix $\mathbf{V}'$ *given by:*

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

which describes the relationships between MIs and SCIs.

We can generate the data starting from the model given by (3.4): $\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B} + \mathbf{E}$*, and hypothesizing* 3 *levels of the error* $\mathbf{E}$*: small, medium and large. The corresponding matrices* $\mathbf{X}$ *are compared in Figure 3.1. In Figure 3.1 (left) the small error does not modify the columns of* $\mathbf{X}$ *that can be well distinguished. The same does not apply to the matrix* $\mathbf{X}$ *in Figure 3.1 (right) with high error.*

First of all, we can measure the goodness of fit of the model by R^2_{GCI}

Error	R^2_{GCI}
Small	0.974
Medium	0.622
Large	0.131

We can see how the smaller the error, the better the values of R^2_{GCI} *will be. Thus, we can compute the* $R^2_{SCI_h}$ *for each of the three SCIs in every situation of error:*

Error	$R^2_{SCI_1}$	$R^2_{SCI_2}$	$R^2_{SCI_3}$
Small	0.988	0.988	0.989
Medium	0.778	0.837	0.855
Large	0.624	0.539	0.672

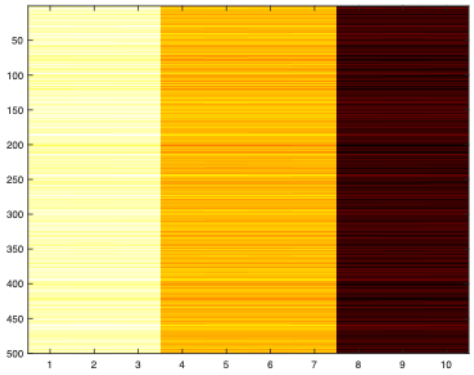


Figure 3.2. Heatmap of matrix **X** formed by 3 blocks of variables.

Here, we can see that the values of $R^2_{SCI_h}$ are greater than R^2_{GCI} for each error and for each SCI.

Example 2. *Let us consider, now, a different situation where the data matrix* **X** *is divided in three blocks: the first three variables are normal with mean equal to* 0 *and variance equal to* 1*, the variables from forth to seventh are normal with mean equal to* 3 *and variance equal to* 1 *and finally, the last three variables are normal with mean equal to* 7 *and variance equal to* 1*. This matrix* **X** *is generated considering a small error and its heatmap is reported in Figure 3.2.*

We can see that if we use the same weight system as before ($\mathbf{c} = \mathbf{1}_3$ *and* $\mathbf{B} = \mathbf{I}_{10}$*) the fit of the model will be relatively bad* $R^2_{GCI} = \frac{SS_{mod}}{SS_{tot}} = 0.548$*, because the arithmetic mean is not a good measure when variables are divided in blocks or are very different each other (i.e., the example with large error). However, we can see that the values of* $R^2_{SCI_h}$ *are very good*

$R^2_{SCI_1}$	$R^2_{SCI_2}$	$R^2_{SCI_3}$
0.999	0.999	1

Thus, in this situation the researcher should avoid using the GCI and considers the SCIs. The case of the dimensions of well-being: it is more appropriate to stop the analysis at the SCIs' level such as Material Living Conditions and Quality of Life (Stiglitz, Sen, and Fitoussi 2009).

Example 3. *Let us consider the setting given in the Example 1 with the difference that the vector* **c** *has three values not equal to* 1*. Thus, we have three clusters of MIs with different weights (i.e., importance) for GCI but where into each cluster any MI has the same weight (i.e., equal to* 1*). Thus, the GCI is a weighted mean of the MIs and all MIs belonging to clusters corresponding to a high value of* **c** *(e.g., higher weight) are more important.*

Let us consider $\mathbf{c} = [0.80.50.6]'$ *and try to measure the goodness of fit of the model by* R^2_{GCI} *and* $R^2_{SCI_h}$ *for all levels of error*

We can see that the values of $R^2_{SCI_h}$ *are greater than* R^2_{GCI} *for each error and for each SCIs also in this situation, that's why the generated matrix of data is divided in three blocks and we have already seen arithmetic mean is not a good estimator when data are divided in different blocks of variables. Let us now consider B as a diagonal matrix (not equal to the*

Error	Small	Medium	Large
R^2_{GCI}	0.934	0.577	0.269
$R^2_{SCI_1}$	0.984	0.848	0.679
$R^2_{SCI_2}$	0.954	0.754	0.610
$R^2_{SCI_3}$	0.972	0.832	0.739

identity matrix). Here, each MI has a different weight also into every single cluster. Thus, for example: $\mathbf{B} = dg([0.90.70.80.80.60.90.70.70.60.8]')$ *and* $\mathbf{c} = [0.70.60.5]'$*; and let us try to measure the goodness of fit of the model by* R^2_{GCI} *and* $R^2_{SCI_h}$ *for all levels of error*

Error	Small	Medium	Large
R^2_{GCI}	0.909	0.431	0.196
$R^2_{SCI_1}$	0.968	0.699	0.599
$R^2_{SCI_2}$	0.936	0.722	0.554
$R^2_{SCI_3}$	0.934	0.554	0.513

Results highlights the same situation previously studied: the arithmetic mean is a good GCI only when the MIs are similar.

3.4 Confirmatory, exploratory and mixed model selection of the model-based CI

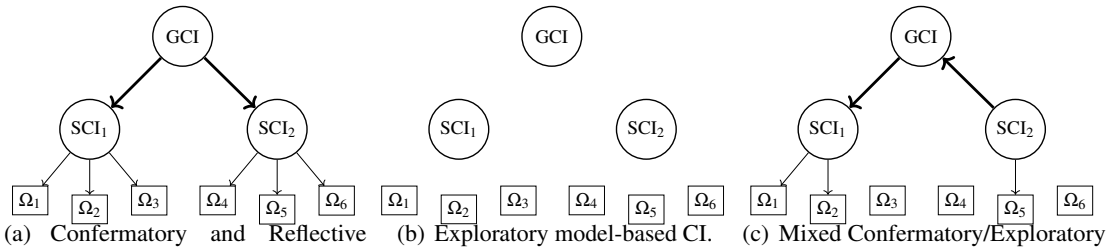
When a theoretical framework is available for the phenomenon described by the CI, the researcher can make conjectures (hypotheses) on:

- the MIs used to characterize the SCIs;
- the SCIs necessary to characterize the GCI;
- the types of relations (e.g., reflective or formative) useful to describe the phenomenon.

Thus, the model is fully selected (Figure 3.3(a)) and specified. Hence, predictions (reconstruction of the data) can be derived from the model-based CI as logical consequence. Then, carrying out an empirical observation of the MIs the selected CI model is tested, in order to confirm with this investigation, whether the phenomenon under study behaves as predicted by the hypotheses. Therefore, in the confirmatory model selection approach MIs are all selected; the relationships between MIs and CIs (SCI and/or GCI) are all, a priori, specified; and the typology of relationships is defined.

Often a theoretical framework is not available for the phenomenon described by the CI, or simply the framework has not been confirmed by empirical evidences. Thus, an exploratory approach can be used. In the case of a fully exploratory approach:

- the MIs used to characterize a SCI are not known and, generally, a larger cluster of possible candidates of MIs are considered;
- the SCIs necessary to characterize GCI are not a priori known (often neither the number);



(a) Confermatory and Reflective model-based CI.

(b) Exploratory model-based CI.

(c) Mixed Confermatory/Exploratory Reflective/Formative model-based CI.

Figure 3.3. Two-Levels Models for CI

- the typologies of relations (e.g., formative or reflective) useful to describe the phenomenon are not known (Figure 3.3(b)).

The distinction between the confirmatory and the exploratory approach is explained in the context of Factor Analysis to highlight differences between CFA and EFA by Ullman (2006).

More frequently, the theoretical framework is available only in part, or, only in part it is confirmed by the analysis. In this case a mixed approach, in part confirmatory and in part exploratory, can be adopted to avoid not optimized modifications of the model. In other terms, the model selection is an artisanal skill of the researcher, but it should be guided by a systematic selection of the most relevant indicators and relationships in the model.

Obviously, in each approach the level of relationships (e.g., the magnitude of correlations) has to be estimated by the model. This estimation is very important also for the confirmatory and the mixed approaches because the theory might suggest weak relationships as valid. The inclusion of not statistically significant MIs into the model could reduce the performance in terms of the goodness of fit. The role of the researcher is very important to explain why an indicator has been excluded from the model and why no SCI might explain that MI into the framework. It is also possible that the model is misspecified, so the researcher has to probe the cause of it.

3.5 Properties

3.5.1 Scale-invariant model-based CI and data transformation

Normalizations

The MIs used to build a CI generally have different units of measurement. The CI is a form of linear or non-linear combination of the MIs, so, it is necessary to avoid that the units of measurement of MIs may influence the building of the SCIs and the final GCI. For this reason, data are normalized to allow the comparison and the combination of the MIs into the SCIs and GCI. It is particularly worth to observe that when in the same GCI, indicators measured in a different scale do coexist, the normalization method used should properly remove the scale effect.

Given the data matrix $\mathbf{X}$ of MIs, generally, three different kinds oif normalization are alternatively applied (Table 3.1), which are linear transformations of the data $\mathbf{X}$, therefore they do not affect the correlation between MIs. Anyway, each normalization affects the

Table 3.1. Types of normalizations

Standardization	Generic element of $\mathbf{Z}$	Characteristics of $\mathbf{Z}$
$\mathbf{Z} = \mathbf{JX}(dg(\Sigma))^{-\frac{1}{2}}$ with $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$	$z_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}$	Mean zero and unitary variance
Min-max normalization (unit-based)		
$\mathbf{Z} = \frac{\mathbf{X}-\mathbf{1}_n \min \mathbf{X}.}{\mathbf{1}_n \max \mathbf{X}-\mathbf{1}_n \min \mathbf{X}}$	$z_{ij} = \frac{x_{ij}-\min(\mathbf{x}_j)}{\max(\mathbf{x}_j)-\min(\mathbf{x}_j)}$	Values are between 0 and 1
Normalized dispersion		
$\mathbf{Z} = \mathbf{JX}\text{diag}(\mu_{\mathbf{X}})^{-1}$ with $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$	$z_{ij} = (x_{ij} - \mu_j)/\mu_j$	Mean zero and standard deviation equal to the coefficient of variation

MIs in its own way, for instance, it is important to observe how the variance of the MIs changes depending on which method of transformation we use. By using *Standardization*, the variance of MIs becomes equal to 1, using *Min-max* the value of variance decreases consistently and with *Normalized dispersion* it becomes equal to the coefficient of variation.

Example 4. *Let us generate a sample of* 5000 *vectors (of* 1000 *units) from a standard normal distribution and let us compute the empirical distribution of variance for each type of normalizations presented previously. The average of the sample variance for each type of normalization is reported:*

Raw Data	Standardized	Min-max	Norm dispersion
1.04	1	0.02	24229.95

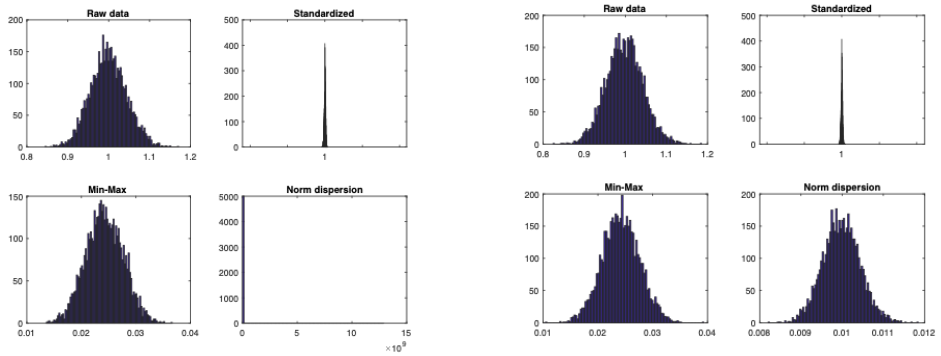
It is useful to recall that for standardized data variance is constantly equal to one for all MIs thus keeping under control the effect of the variability while maintaining the data centered; that is why this is the most used method of normalization. The Min-max method keeps constant the range of values of the data, however, the value of the variance is reduced toward zero. This might be considered a problem because different MIs tend to be compressed reducing differences. The variance of data with Normalized dispersion method depends on the value of the mean of the raw data (i.e. they are inversely proportional), so if the mean is close to zero the value of the variance is very big (e.g., in the last example) otherwise the variance decreases consistently.

Let us generate a sample of 5000 *vectors (of* 1000 *units) from a normal distribution with mean equal to* 10 *and variance equal to* 1 *and let us compute the empirical distribution of variance for each type of normalization previously presented. The average of the sample variance for each type of normalization is reported:*

Raw Data	Standardized	Min-max	Norm dispersion
0.95	1	0.03	0.01

In Figure 3.4, it is possible to observe the difference among empirical distribution of variances according to different types of normalization via histograms.

The normalization of data is a very relevant step for the analysis since it allows for comparing MIs with different units and measurement scales. Thus, a very useful property



(a) Empirical distribution of variances for each type of normalization from the standard normal distribution.

(b) Empirical distribution of variances for each type of normalization from the normal distribution with mean equal to 10 and variance equal to 1.

Figure 3.4. Comparison among empirical distribution of variances.

for a CI is to be scale-invariant, such that, the CI is not sensitive to linear transformations of the data such as those reported above. In other terms, the goal is to have a CI that does not change by using different methods of normalization on the MIs. Hence, a CI model-based on the reconstruction of the correlation matrix $\mathbf{R_X}$ of the MIs is invariant under the above defined normalizations.

Logarithmic transformation

A very useful transformation of the MIs is the *logarithmic* one which is used:

- to transform a highly skewed variable into one that is more approximately normal (symmetric);
- to handle non-linear relations between MIs and the SCIs or GCI.

Since the logarithmic transformation is monotone, the ordering of units is preserved

$$z_{ij} = \log(x_{ij}) \tag{3.37}$$

where for $\log(a)$ the natural logarithm of a is considered. It is simple to see that the arithmetic mean $\widehat{\mathbf{g}}_{\log}$ of $\mathbf{z}_j = [z_{1j}, \ldots, z_{nj}]'$, $j = 1, \ldots, J$ is the logarithm of the geometric mean $\widehat{\mathbf{g}}_{GM}$ of original MIs (e.g., $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$).

$$\widehat{\mathbf{g}}_{\log} = \frac{1}{J}(\mathbf{z}_1 + \cdots + \mathbf{z}_J) = \log(\prod_{j=1}^{J} \mathbf{x}_j^{\frac{1}{J}}) = \log(\widehat{\mathbf{g}}_{GM}) \tag{3.38}$$

The Human Development Index is the most popular index computed as a geometric mean. (3.38) shows that the propriety of non-compensability often conferred to geometric means, fails when logarithmic transformation is considered.

The type of aggregation employed is strongly related with the method used to normalize raw data; the normalization method used should properly remove the scale effect. In both linear and geometric aggregations, weights express trade-offs between indicators: the idea is that deficits in one dimension can be offset by surplus in another one (Munda and Nardo 2009).

Power transformation

The family of functions that are used in order to create a monotonic transformation of data using power functions is known with the name *power transform*. The aim of this technique is to stabilize the variance of the data, to make more valid the measures of association as the correlation between MIs and usually to make the data more normal distribution-like. In the field of power transformations, the most famous one is the Box-Cox transformation (Box and Cox 1964) which takes the following form:

$$\mathbf{z}_j = \begin{cases} \frac{\mathbf{x}_j^{\lambda}-1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \log(\mathbf{x}_j) & \text{if} \quad \lambda = 0 \end{cases} \tag{3.39}$$

where λ is the power parameter and $(\mathbf{x}_1, \ldots, \mathbf{x}_J)$ are the original MIs (columns of the data matrix $\mathbf{X}$). In the same paper, the authors proposed also an extended form which could consider negative variables, and throughout several years many modifications have been proposed by many other authors. Many researchers have studied Box-Cox transformation for its properties, e.g. R. Carroll and Ruppert (1981). The power transformation family is often used for transforming the data into a normal linear model, e.g. in regression parameter estimation. Thus, the aim of the Box-Cox transformations is to ensure the usual assumptions of normality of the MIs, but the first step when using it is to estimate the parameter λ in order to proceed into the analysis by considering the parameter as known.

In an inferential context, Bickel and Doksum (1981) proved that the asymptotic marginal (unconditional) variance of the parameter of interest (e.g. β) is inflated by a very large factor over the conditional variance, for fixed λ. One partial remedy to the problem is to use the scaled form of the Box-Cox transformation:

$$\mathbf{z}_j = \begin{cases} \frac{\mathbf{x}_j^{\lambda}-1}{\lambda(\bar{X}_{GM})^{\lambda-1}} & \text{if} \quad \lambda \neq 0 \\ \bar{X}_{GM}\log(\mathbf{x}_j) & \text{if} \quad \lambda = 0 \end{cases} \tag{3.40}$$

where $\bar{X}_{GM}$ is the geometric mean of the J MIs. Such scaling can effectively control the size of transformed responses and can also reduce the conditional variance for the parameter of interest (e.g. β).

3.5.2 Non-Compensable and Non-Negative model-based CI

A non-compensable CI is an indicator where the positive relations among MIs are not compensated by the effect of the negative relations among them. In other terms, all MIs are concordantly related to the CI, so that, increments of CI correspond to increments of the MIs, and vice-versa. Hence, non-compensability and non-negativity of the weights used to combine MIs are strictly connected. Ebert and Welsch (2004) proved that the use of linear aggregations of the MIs yields a meaningful CI only if data are all expressed in a partially comparable interval scale or in a fully comparable interval scale and are concordant.

Compensability is constant when a linear aggregation is used, whereas it is lower when the CI contains indicators with low values if the aggregation is geometric. The geometric mean is not a good solution to comply the non-compensability of a CI since, like we have shown in (3.38), the logarithm of the geometric mean is equal to the arithmetic mean of the logarithm of MIs. It means that if we use the geometric mean as GCI, we have the same problems of compensability of the arithmetic mean or any linear aggregations.

Weights represent the importance (e.g., correlations in FA) of each MI into the definition of the related latent construct (i.e., SCI). Thus, weights must be positive for reasons of interpretation; the presence of a negative weight means that the MI reflects negatively the latent construct. For instance, the MIs *life expectancy* and *child mortality* might be caused by the same latent concept (e.g., *well-being*) but they contributes to the latent concept positively and negatively, respectively. If these two MIs are combined together with positive weights, they will compensate each other. So, it is crucial to properly define the latent concepts we want to measure and the MIs which are supposed to be included in their definition, by establishing whether a change of polarity is needed. In order to comply the non-compensability of a CI, our proposal is to constrain all the weights to be strictly positive and reverse all the MIs with negative weights.

Moreover, in a FA model in which all the loadings (i.e., correlations between factors and MIs) are positive, rankings of latent factors are concordant among each other. Finally, rankings of each SCI and of the GCI are concordant. If some weights are negative, it means that the MIs are negative measures of the latent construct, for this reason the MIs should be edited (e.g., reversed). Therefore, if all the loadings are positive, the composite indicator is non-compensable and the correlations between MIs are positive.

Let us first recall the following necessary and sufficient conditions for a matrix to be a covariance or correlation matrix. A $(J \times J)$ square matrix $\Sigma_{\mathbf{X}}$ is the covariance matrix of J random variables of $\mathbf{X}$ if, and only if, $\Sigma_{\mathbf{X}}$ is symmetric and positive semidefinite. A $(J \times J)$ square matrix $\mathbf{R}_{\mathbf{X}}$ is the correlation matrix of J random variables if, and only if, $\mathbf{R}_{\mathbf{X}}$ is symmetric, positive semidefinite and has all the entries on the main diagonal equal to 1.

In order to have a $(J \times J)$ square matrix $\mathbf{R}_{\mathbf{X}}$ that is a correlation matrix, it is necessary for all (3×3) sub-matrices of $\mathbf{R}_{\mathbf{X}}$ to be correlation matrices.

Thus, any given a (3×3) matrix A equal to

$$\begin{bmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{bmatrix}$$

is a correlation matrix if, and only if, $-1 \leq a, b, c \leq 1$ and $\det(\mathbf{A}) \geq 0$. Thus, let us consider a and b as correlations (i.e., between -1 and 1); then, c must to be into the range (Budden, Hadavas, and Hoffman 2008):

$$[ab - \sqrt{(1-a^2)(1-b^2)}, ab + \sqrt{(1-a^2)(1-b^2)}]$$

Example 5. *Suppose* $a = 0.8$ *and* $b = 0.3$. *In order to guarantee a valid correlation matrix,* c *must be into the range:* $[-0.332364, 0.812364]$. *In fact, for different values of* c *the eigenvalues are:*

$c = 0.9$	$c = 0.8$	$c = -0.2$	$c = -0.4$
-0.0651	0.0087	0.0658	-0.0369
0.7024	0.7000	1.1274	1.2293
2.3627	2.2913	1.8069	1.8076

Thus, values of c *chosen outside of the given range, e.g.,* $c = 0.4$ *or* $c = 0.9$, *give a matrix* $\mathbf{A}$ *that is not positive semi-definite, since it has at least one negative eigenvalue.*

3.5.3 Polarity

The study of the polarity of the MIs is crucial to determinate the correlation structure among MIs and between MIs and SCIs. A general approach to the correlation analysis is proposed by Jöreskog (1978). A very important step of the CIs building is the selection of MIs to be considered into the model for complying the fact that the indicator is non-compensable; hence, it is important to find a cluster of MIs all positively correlated among them. During the construction of a composite indicator, some MIs could be negative indicators for the character of interest: they should be edited.

If all MIs are normalized with the Min-max method and we use a factor analysis model, we can change the polarity of the "negative" MI using the formula: $\tilde{x}_{ij} = 1 - x_{ij}$. In this way, the correlation between this MI and the factors does not change in intensity but only in sign (it becomes positive).

By facing this problem for each MI, we can change many correlations inside the model and we might find the final solution when changing the polarity of one MI does not generate some other negative correlations. It is important to notice how the performance of the model is not affected at all by changing the polarity of the MI with this strategy. This tool could be very useful also to understand better the meaning of some MIs considered into the framework. If a MI has no positive correlation with any factor (i.e., SCI), both pre-and-post editing, this MI could be non-statistically significant or it can have a different meaning into the framework. Furthermore, we can understand the single role of each MI into the framework and how to work on a cluster of MIs to have effects also on all the others.

Example 6. *Given a* (100×3) *raw data matrix* $\mathbf{X}$ *with correlation matrix* $\mathbf{R}$*:*

$$\mathbf{R} = \begin{bmatrix} 1 & 0.45 & -0.5 \\ 0.45 & 1 & -0.8 \\ -0.5 & -0.8 & 1 \end{bmatrix}$$

After a normalization of $\mathbf{X}$*, we can edit the third MI with the formula* $\tilde{x}_{ij} = 1 - x_{ij}$*. In this way, we would obtain a new correlation matrix with all positive values:*

$$\tilde{\mathbf{R}} = \begin{bmatrix} 1 & 0.45 & 0.5 \\ 0.45 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{bmatrix}$$

It is important to notice how this operation does not change the absolute value of correlations (i.e., intensity, level or magnitude).

3.5.4 Reliability of a CI

Reliability is a crucial property for a well-done composite indicator, and it is well described for example in R.E. Zinbarg, Revelle, et al. (2005), Vinzi et al. (2010), and Edwards (2011). Reliability of a CI is the global consistency of MIs based on the correlations between different MIs related to the same CI (i.e., latent construct). It is frequently called *internal consistency* and it is usually measured by Cronbach's α, which can be seen as the expected correlation between all pairs of MIs used to specify the latent concept compiled by the CI. Cronbach's α ranges between negative infinity and one. A commonly accepted rule of thumb for describing internal consistency has been given by George and Mallery (2003): $\alpha \geq 0.9$,

Excellent; $0.9 > \alpha \geq 0.8$, Good; $0.8 > \alpha \geq 0.7$, Acceptable; $0.7 > \alpha \geq 0.6$, Questionable; $0.6 > \alpha \geq 0.5$, Poor; $0.5 > \alpha$, Unacceptable.

Cronbach's α has been demonstrated many times to attain quite high values even when the set of items measures several unrelated latent variables (Green, Lissitz, and S.A. Mulaik 1977; Schmitt 1996). Thus, an alternative index in the reliability analysis is the Dillon-Goldstein's (or Jöreskog's) ρ (Werts, Linn, and Jöreskog 1974) better known as *composite reliability*: a set of indicators (i.e., the cluster of MIs explained by the related SCI) is considered reliable if this index is larger than 0.7 (Fornell and Larcker 1981). The ρ is less sensitive to the number of items analysed (Fornell and Larcker 1981); this is also shown by Chaouachi and Rached (2012). Several other measures of reliability are reported in Revelle and R. Zinbarg (2009).

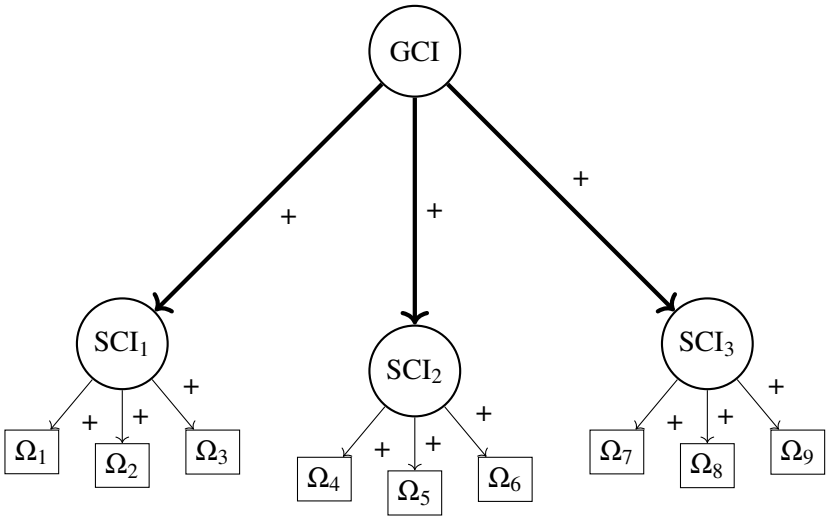
3.5.5 Unidimensionality and Presence of a General Factor

Unidimensionality evaluates to which extent a single latent indicator, generally a SCI, has been measured with a cluster of MIs. Revelle and R. Zinbarg (2009) observe that whereas unidimensionality represents the ideal of measurement (McNemar 1946), some phenomena - that consist of related yet discriminable SCIs that differ and are not themselves unidimensional - exist, and so it would be unrealistic to expect measures of them to be unidimensional. Thus, unidimensionality is more realistic for SCIs, whereas Revelle and R. Zinbarg (2009) hypothesize that there is a general factor, i.e., a GCI that can be tested by nested confirmatory factor models or detected by exploratory disjoint factor models. In other words, each latent construct (i.e., dimension) is measured by a cluster of MIs and each SCI measure only one single latent construct (J. Anderson and Gerbing 1982). That is, the clusters of MIs defining each construct are unidimensional (Aaker and R. Bagozzi 1979; Hayduk 1980; Jöreskog 1970).

The Kaiser rule (Guttman 1954; Kaiser 1960; Yeomans and Golder 1982) is used as unidimensional check on the eigenvalues of the correlation matrix of the cluster related to the SCI. The first has to be greater than 1, whereas the others have to be smaller. Thus, a measure of unidimensionality for each SCI might be the variance of the second component (i.e., the component corresponding to the second largest eigenvalue) of the cluster (i.e., the cluster of MIs explained by the related SCI). A SCI is considered unidimensional if the aforementioned variance is lower the 1.

A problem that occurs when unidimensionality is not guaranteed is interpretational confounding (Hayduk 1980; Burt 1973; Burt 1976) and misspecification of the model (J. Anderson and Gerbing 1982). It is not possible to interpret the meaning of or assign a name to a SCI when the conditions for unidimensionality are not met (J. Anderson and Gerbing 1982; Burt 1976).

Example 7. *Let us suppose that data have this hierarchical structure:*



If we try to estimate this situation with a model with only 2 *factors, it will be evident how two aspects are considered together into one single factor:*

	Factor 1	Factor 2
Unidimensionality	2.737	0.556
Reliability	0.526	0.794

The Factor 1 *explains the first two clusters of MIs, and its measure of unidimensionality is higher than* 1*. We can also see how its Cronbach's* α *is not acceptable.*

If we try to consider one more factor, the measures change essentially:

	Factor 1	Factor 2	Factor 3
Unidimensionality	0.400	0.618	0.556
Reliability	0.781	0.781	0.794

Here the values of Cronbach's α *are all good and the unidimensionality is verified per each factor (i.e., the values of the variance of the second component of the cluster are lower than* 1*).*

A very common error is to interpret the Cronbach's α as a measure of unidimensionality: strictly speaking, it is not. It is a measure of internal consistency, whereas unidimensionality involves the homogeneity of a set of items. The key to understand how internal consistency and unidimensionality are jointed is the fact that internal consistency is certainly necessary for homogeneity, but it is not sufficient. Thus, reliability might be considered as a necessary condition for unidimensionality (Aaker and R. Bagozzi 1979). We can see that improving the internal consistency leads to an improvement of unidimensionality as well, but we cannot use the Cronbach's α to measure both (e.g., in the Example 7).

3.6 Applications

Development and *poverty* are non-observable multidimensional concepts which can be defined as latent constructs. They can be measured in a great variety of ways, the widest used approach being the aggregation of a set of individual indicators. The most important examples are Human Development Index (HDI), developed by Pakistani economist Mahbub ul Haq for the UNDP in 1990 and reformulated by Alkire (2010), and the Multidimensional Poverty Index (MPI) (Alkire and Foster 2011; Alkire, Foster, et al. 2015).

The Human Development Index (HDI) used by the United Nations Development Programme represents an attempt to place the emphasis on human welfare rather than on the progress of national economy. This index has the merit to have introduced an important alternative to the traditional unidimensional measure of *development* (i.e. the gross domestic product). However, the HDI contains many conceptual flaws (Trabold-Nübler 1991; Welzel, Inglehart, and Kligemann 2003; Ronald and Welzel 2005). Moreover, Sagar and Najam (1998) showed that the HDI fails to include any ecological considerations and, over the years, the related reports seem to have become stagnant, repeating the same rhetoric without necessarily increasing the HDI's utility.

Over time, many authors studied and reviewed the HDI suggesting some improvements on the components of the index, proposing a different structure for the index itself and attempting to establish whether the developed indices based on the components of the HDI have the expected properties of an index, or whether they are 'redundant' as proposed by other literature (Noorbakhsh 1998; Despotis 2005; Martínez 2012).

Poverty has been usually measured as the lack of basic capabilities, and Sen (1981), Sen (1985), and Sen (1992) was the first to underline its multidimensional aspect. The MPI was developed in 2010 by the Oxford Poverty & Human Development Initiative (OPHI) and the United Nations Development Programme to measure the level of *poverty* by aggregating its different dimensions. Alkire and Foster (2011) proposed a method which involves counting different types of deprivation that individuals experience at the same time. The analysis of these deprivation profiles is used to construct a multidimensional index of *poverty*.

Neubourg, Milliano, and Plavgo (2014) reviewed the insights of various contributions from research into multidimensional *poverty* and deprivation, by combining them into an internally consistent framework. The author explained the importance of avoiding the "loss of dimensions" when reducing multiple dimensions into a multidimensional poverty index.

Ultimately, Alkire, Foster, et al. (2015) provides an in-depth account of multidimensional poverty comparison methodologies, with a particular focus on the Alkire-Foster method and on the MPI.

3.6.1 Human Development Index

The Human Development Index (HDI) (Alkire 2010) is a CI of 3 specific indicators: *life expectancy*, *education* and *income per capita* indicators, which are used to rank countries into four tiers of human development.

In its 2010 Human Development Report, the UNDP began using a new method of calculating the HDI by considering only three dimensions: *life expectancy at birth*, *education index* and *GNI per capita*. Each dimension is represented by a specific index (normalized with an own method): *Life Expectancy Index* (LEI), *Education Index* (EI) and *Income Index* (II).

The HDI is the geometric mean of the previous three normalized indices. Thus, following our model-based approach, we can measure the goodness of fit of the HDI by considering that the logarithm of the geometric mean is equal to the arithmetic mean of the logarithm of MIs.

Let us consider $\widehat{\mathbf{B}}$, $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{c}}$ given: $\widehat{\mathbf{B}} = \widehat{\mathbf{V}} = \mathbf{I}_3$ and $\widehat{\mathbf{c}} = \mathbf{1}_3$.

$$R^2_{\text{HDI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{\text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\log(\widehat{\mathbf{g}}'_{\text{HDI}})\log(\widehat{\mathbf{g}}_{\text{HDI}})\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})}{\text{tr}((\log(\mathbf{Z}))'(\log(\mathbf{Z})))} = 0.901 \tag{3.41}$$

where $\log(\mathbf{Z})$ is a matrix where each column is the logarithmic transformation of the respectively normalized column of $\mathbf{X}$.

It is important to see how the three indices are normalized and how these transformations have a role on the goodness of the HDI:

- *Life Expectancy Index* (LEI) is normalized according to the formula: $Z = \frac{(X-20)}{65}$, where X is *life expectancy at birth*;
- *Education Index* (EI) is the composition (i.e. the arithmetic mean) of two MIs: *Expected years of schooling* (X_1) and *Mean years of schooling* (X_2), where the first one is normalized by the formula: $Z_1 = \frac{\min(X_1,18)}{18}$ and the second one accordin to the formula: $Z_2 = \frac{X_2}{15}$. Thus, the *Education Index* is computed by: $Z = \frac{Z_1+Z_2}{2}$;
- *Income Index* (II) is normalized according to the formula: $Z = \frac{\log(X)-\log(100)}{\log(75000)-\log(100)}$, where X is *GNI per capita*.

In this specific case, researchers have chosen three different ways to normalize the data. This might be a dangerous approach because each normalization has a specific influence on the final index. Indeed, the transformation used for normalization has the unique purpose to allow comparability among MIs, eliminating the influence of the different unit of measurements without introducing other artificial distortions on the data that for example produce a better fit of the model; otherwise, there might be the suspect that normalization is used to artificially increase the fit of the model.

Let us see what the result of the HDI is if we use a unique normalization for the three MIs, for example the Min-max normalization. In this way we can evaluate how much influent are the different normalizations with respect to a unique one.

Let us consider $\widehat{\mathbf{B}}$, $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{c}}$ given: $\widehat{\mathbf{B}} = \widehat{\mathbf{V}} = \mathbf{I}_3$ and $\widehat{\mathbf{c}} = \mathbf{1}_3$, as before, and let us define $\widehat{\mathbf{g}}_{\text{MinMaxHDI}}$ as the Min-max normalized HDI, so we consider: *Life Expectancy Index* equal to "Min-max normalized *life expectancy at birth*", *Education Index* equal to the arithmetic mean of "Min-max normalized *expected years of schooling*" and "Min-max normalized *mean years of schooling*" and *Income Index* equal to "Min-max normalized *GNI per capita*".

$$R^2_{\text{HDI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{\text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}}\log(\widehat{\mathbf{g}}'_{\text{MinMaxHDI}})\log(\widehat{\mathbf{g}}_{\text{MinMaxHDI}})\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}})}{\text{tr}((\log(\mathbf{X}))'(\log(\mathbf{X})))} = 0.632 \tag{3.42}$$

where $\log(\mathbf{X}))$ is a matrix where each column is the logarithmic transformation of the respective column of $\mathbf{X}$.

The increase of the 27% of R^2_{HDI} with respect to $R^2_{\text{MinMaxHDI}}$ has to be imputed to the use of different normalizations in the HDI. Therefore, it is important to understand that different normalizations of MIs must be strongly motivated.

Furthermore, by analyzing the intensity of the correlation between each dimension and the HDI we have found levels much higher than we expected. It is reasonable expecting a significant and positive correlation, but values so high empower the question by McGillivray (1991): is it useful to create another indicator that provides little more information than some traditional indicator like the GNI? We proposed a reflective model where the dimensions have to be correlated among themselves; but could it be more interesting to consider more dimensions into the model by including more MIs into the framework?

Correlation	LEI	EI	II
HDI	0.90	0.95	0.94

By considering the transformations adopted in the construction of the HDI and the levels of correlation that they lead, the performance of the goodness of fit in terms of R^2_{HDI} is explained.

3.6.2 Multidimensional Poverty Index

The global Multidimensional Poverty Index (MPI) (Alkire and Foster 2011; Alkire, Foster, et al. 2015) is an international measure of acute poverty covering over 100 developing countries produced by Oxford Poverty & Human Development Initiative (OPHI) and the United Nations Development Programme in 2010.

The index uses the same three dimensions of the Human Development Index: *health*, *education*, and *standard of living*. However, these are measured using ten indicators split in three dimensions:

- *Education* (each indicator is weighted equally at $\frac{1}{6}$):
  - *Years of schooling*: deprived if no household member has completed six years of schooling;
  - *Child school attendance*: deprived if any school-aged child is not attending school up to class 8.
- *Health* (each indicator is weighted equally at $\frac{1}{6}$)
  - *Child mortality*: deprived if any child has died in the family in past 5 years;
  - *Nutrition*: deprived if any adult or child for whom there is nutritional information is stunted.
- *Standard of Living* (each indicator is weighted equally at $\frac{1}{18}$)
  - *Electricity*: deprived if the household has no electricity;
  - *Sanitation*: deprived if the household's sanitation facility is not improved (according to MDG guidelines), or it is improved but shared with other households;
  - *Drinking water*: water: deprived if the household does not have access to safe drinking water or safe drinking water is more than a 30-minute walk from home roundtrip;
  - *Floor*: deprived if the household has a dirt, sand or dung floor;
  - *Cooking fuel*: deprived if the household cooks with dung, wood or charcoal;

– *Assets ownership*: deprived if the household does not own more than one of: radio, TV, telephone, bike, motorbike or refrigerator and does not own a car or truck.

The MPI is the average of the three arithmetic means of the MIs of each dimension. It can be seen as a weighted mean of MIs with weights $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}$. Therefore, by exploiting the deviance decomposition (3.28), we can measure the goodness of fit of the MPI.

Let us consider $\widehat{\mathbf{B}}$, $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{c}}$ given:

$$\widehat{\mathbf{B}} = \text{diag}(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

$$\widehat{\mathbf{V}}' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\widehat{\mathbf{c}}' = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

$$R^2_{\text{MPI}} = \frac{SS_{mod}}{SS_{tot}} = \frac{\text{tr}((\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})\widehat{\mathbf{g}}'_{\text{MPI}}\widehat{\mathbf{g}}_{\text{MPI}}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1})}{\text{tr}(\mathbf{X}'\mathbf{X})} = 0.515 \quad (3.43)$$

Let us compare the MPI with the composite indicator as solution of our algorithm. The partition of the MIs is different only for one MI out of ten; indeed, the MI *Nutrition* defines a SCI by itself (called *Health*) whereas the MI *Child mortality* is included into the cluster that defines the SCI called *Standard of Living*. The estimated weights are different from the weights proposed by OPHI also because the normalization method is different. The proposed weights sum 1, whereas the square of the estimated weights sum 1.

Thus, the best solution for our model is:

$$\widehat{\mathbf{B}} = \text{diag}(0.71, 0.71, 0.37, 1, 0.39, 0.38, 0.38, 0.37, 0.39, 0.36)$$

$$\widehat{\mathbf{V}}' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\widehat{\mathbf{c}}' = \text{diag}(0.85, 0.43, 0.30)$$

$$R^2 = \frac{SS_{mod}}{SS_{tot}} = \frac{\text{tr}((\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1}(\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})\widehat{\mathbf{g}}'\widehat{\mathbf{g}}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}(\widehat{\mathbf{c}}'\widehat{\mathbf{V}}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}\widehat{\mathbf{c}})^{-1})}{\text{tr}(\mathbf{Z}'\mathbf{Z})} = 0.884 \quad (3.44)$$

where $\mathbf{Z}$ is a matrix where each column is the standardized column of $\mathbf{X}$, respectively.

If we compare the partition $\widehat{\mathbf{V}}$of our model with the partition proposed by OPHI, we will obtain an Adjusted Rand Index (Hubert and Arabie 1985) equal to 0.687.

Thus, the MI *Child mortality* seems more correlated with the MIs which define the SCI called *Standard of Living*, and the best solution found by our model seems to be better than the solution with the proposed weights in terms of R^2.

3.7 Conclusion

Composite indicators are useful and powerful tools for providing support to decision making. The CI field of study is very wide and proposing good CIs is an interesting challenge for a researcher. Moreover, it is important to test the available theoretical knowledge we have at the moment of the analysis without taking it for granted. A cluster of MIs is not always a system of indicators and not always we have to include all of them into the analysis. The construction of a CI is made by the many choices of the researcher.

We proposed a model-based approach which tries to build CIs that are objective and transparent, reducing as much as possible the arbitrary choices. In our opinion, the main goal of a CI, from a reflective point of view, is measuring a phenomenon of interest by attempting to reconstruct the relations present into the manifest data.

The entire process of *Dimensionality Reduction* is important in order to understand the phenomenon of interest, a single indicator might not be sufficient to represent it because of its complexity and multidimensionality. Our novelty proposal could be extremely useful to measure the multidimensional nature of the phenomenon of study, even if the number of MIs is large.

This paper has proposed both a model-based approach for the construction of CIs and a body of properties and characteristics that a good CI should have in order to produce well-constructed indexes that induce realistic and useful policy conclusions. We have used two famous indices, HDI and MPI, in order to show how the model-based approach can be used to assess their importance. Behind these two indices there are many theoretical arguments that ensure their relevance; however, several additional statistical properties can be included and by using a confirmatory approach the theory hypothesized can be statistically evaluated. On the basis of the results reported in this paper, the following conclusions have emerged:

a The use of a model-based CI allows to test empirically the hypotheses formulated in the specification of the CI; a model-based CI wishes to be objective and transparent, avoiding - as much as possible - the use of arbitrary choices that cannot be tested.

b The CI is a form of linear or non-linear combination of the MIs, so, it is necessary to avoid that the units of measurement of MIs influence the building of the SCIs and the final GCI. The choice of the scale transformation of MIs has many consequences on the performance of SCIs and GCI.

c A CI has to be non-compensable. That is, positive relations between MIs are not compensated by negative relations between MIs. In a reflective model like FA, a solution is to impose the non-negativity constraint on the loadings.

d The selection of MIs to consider in the model is a fundamental step. If the SCI is seen as a common concept to a cluster of MIs (reflective relation between SCI and MIs) these last must be correlated with the SCI. Therefore, statistically significant relations must be observed. In addition, the correct polarity must be chosen.

e The reliability of a CI is the global consistency of MIs based on the correlations between different MIs on the same CI. On the one hand, the reliability is connected to the concept of unidimensionality, which, on the other hand, evaluates to which extent a single latent indicator has been measured with a cluster of MIs. Reliability

and unidimensionality are more realistic for SCIs, while, when considering a GCI, we have to hypothesize the presence of a general factor.

f HDI is the most famous indicator for development and we evaluated how the different methods of normalization used for the four MIs are influent in terms of goodness of fit.

g We have found a slightly different 'formulation' of MIs for MPI by including "child mortality" into the cluster that defines the SCI called "Standard of Living". A consistent increase of goodness of fit has been registered.

The HDFA model and its non-negative version guarantee a simple, two-level hierarchy with SCIs and GCI with all the aforementioned characteristics. This paper proposed the LS estimations of the parameters for the model.

As we already argued, a lot of famous CIs are built without testing crucial properties like reliability and unidimensionality, and, very often without testing the theoretical framework proposed by the literature as well. Hence, during the construction of a CI, some salient aspects could seem controversial; for example, the study of MIs' polarity gets often undervalued.

Finally, it seems very important to take care about all the researcher's choices and to not take them for granted since the CI's effectiveness is strictly linked to each individual choice taken during the analysis.

Chapter 4

A composite indicator for waste management in EU

4.1 Introduction

We are more than 7.5 billion people on our planet, and we are producing waste every day. The constant expanding of population implies an increasing generation of waste, and although the management of the waste keeps improving in the EU, many estimates tell us that half of that waste is not collected, treated or safely disposed of. That is why policymakers need consistent and useful tools to measure and monitor waste management. The challenge is composed by two different aspects: sustainable consumption and smarter waste management. In order to plan a coordinate action through the EU countries, a reliable measure of "good" waste management is needed.

A multidimensional phenomenon like waste management is described by a huge quantity of information useful for making strategical decisions and the demand for statistics on waste generation and treatment has grown considerably in recent years. This amount of information needs to be synthesized by studying relationships among manifest indicators (MIs). It is important to find the relationships among dimensions and MIs in order to synthesize the information and have a response on the conduct of each country to achieve the priority goals set by Europe, reducing waste generation and maximizing recycling and re-using. Identifying these relationships could be fundamental to understand where each country should focus its actions and what impacts each action could have. A "good" waste management is vital for global sustainable development, it is connected with Sustainable Development Goals (SDGs). The SDGs constitute the core of the 2030 Agenda for Sustainable Development *Transforming our world: the 2030 Agenda for Sustainable Development* (2015), and their aim is to guide global, regional and national actions regarding development for the next 15 years. The United Nations Industrial Development Organization (UNIDO) contributes to the achievement of the SDGs by supporting Member States in achieving inclusive and sustainable development. Since the interconnected nature of the SDGs, many of UNIDO's activities contribute to more than one goal. In our specific field of interest the Goal 12, named "Ensure sustainable consumption and production patterns", has as targets by 2030 among others: achieve the sustainable management and efficient use of natural resources, substantially reduce waste generation through prevention, reduction, recycling and reuse, and halve per capita global food waste at the retail and consumer level, and reduce food

losses along production and supply chains. Other Goals are involved by waste management, for instance, plastics are devastating for the planet and its inhabitant, especially marine environment, in many areas of our planet it is common to find waste burned instead of collected properly, it has a huge impact on methane and CO_2 emissions and thus, on climate change.

The quantity of information and statistics about waste management are more and more consistent but so far, few studies are available in this field. The aim of this work is producing a model-based measure of "good" waste management, in order to provide a useful tool for policy-makers and institutions.

A usual way to synthesize a big amount of information is using Composite Indicators (CIs), that is, non-observable latent variables, linear combinations of MIs (OECD 2004). These have been considered to have several advantages: they can summarize multidimensional situations and facilitate evidence-based decision-making; it can be easier to interpret than a list of separate indicators; they facilitate communication among policy-makers, the media and the general public. However, CIs also have some potential disadvantages: their construction is particularly difficult as it requires both a sound scientific base and political consensus; they may send misleading messages about policy if poorly constructed or misinterpreted; they may lead to over-simplistic conclusions, on behalf of both the general public and political actors.

In this paper we propose a CI for waste management in Europe by using a model-based approach. The model has a hierarchical structure formed by factors associated with subsets of MIs with positive loadings. This approach guarantees to comply with all the good properties on which a composite indicator - summarizing a multidimensional phenomenon - should be based. Such properties are: model-based, statistically estimated, with a hierarchical structure, scale-invariant, uni-dimensional, reliable and non-compensable[1].

It is worth to remind that, in a FA model, if all the loadings (i.e., correlations between factors and MIs) are positive then rankings of latent factors are concordant among each other. Therefore, rankings of each SCI and of the GCI are concordant. In other words, if all the loadings are positive, the CI is non-compensable and the correlations among MIs are positive.

The general goal is to find, via statistical data modeling, the hierarchically aggregated index that best represents the waste management and its parameters are estimated according to the maximum likelihood estimation method (MLE) in order to make inference on the parameters and on the validity of the model.

In this paper, the *Hierarchical Disjoint Non-negative Factor Analysis* model presented in Chapter 2 (Cavicchia and Vichi 2020a) with its estimations is not recalled. Thus, the paper is organized as follows. In Section 4.2, the MIs about waste management and recycling are introduced. The results of the waste management are presented in Section 4.3. A final discussion completes the paper in Section 4.4.

4.2 Data

Waste management is a complex phenomenon, described by a huge quantity of information and its importance is crucial for the environmental and for the human life in general. It

[1] Following the definition in Section 3.5.2, a non-compensable CI is an indicator where the positive relations among MIs are not compensated by the effect of the negative relations among them.

Table 4.1. List of MIs.

Code	Name	Code	Name
1	Generation of waste*	6	Share of material recovered and fed back into the economy (%)
2	Generation of recyclable waste*	7	Imports from non-EU countries of recyclable raw materials*
3	Private investments as value added at factor cost related to circular economy (€ per capita)	8	Imports intra EU countries of recyclable raw materials*
4	Jobs as persons employed related to circular economy (% persons employed)	9	Recycling rate of municipal waste (% generated waste)
5	Patents related to recycling and secondary raw materials (number per capita)	10	Recycling rate of all waste excluding major mineral waste (% generated waste)

* tonnes per capita

is more and more important to find the way to measure it in order to provide support for decision making. The number of statistics and measures related to waste collection, recycling and circular economy is expanding every year and the need of build aggregated index to monitor the countries' behavior in terms of policy is even more important.

Since the aim of our analysis is identifying a system of non-negative loadings in order to define a two-levels hierarchy with the general composite indicator as root, a preliminary analysis is needed to identify potential MIs that measure as negative components of the general latent construct (i.e., waste management) and to dectect the ones that are not statistically significant. If variables (both manifest and latent) have positive loadings, they contribute positively to the construction of the general composite indicator. However, if variables (both manifest and latent) have negative or null loadings, they must be reversed and it is important to investigate about their importance. It is crucial for the definition of the general composite indicator (GCI) and it avoids a compensation effect among MIs. Another essential part of the preliminary analysis consists in detecting not statistically significant MIs into the model in order to discard the latter ones from the analysis because they are not relevant; they can assume a confounding role for the analysis and could be removed without incurring in a significant loss of information.

After the preliminary steps of the analysis, the final data-set considered in this paper is composed by 10 MIs (Table 4.1) and 28 units (i.e., countries) (Table 4.2). Many MIs about the characteristic of countries have been considered in order to help the interpretation of the results. The MIs into the data-set come from different sources: Eurostat, Joint Research Centre, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (DG GROW) and the European Patent Office, and they are regularly updated and available for free on Eurostat website.

Table 4.2. List of countries.

Name	Abbreviation	Country	Abbreviation
Belgium	(BE)	Bulgaria	(BL)
Czech Republic	(CZ)	Denmark	(DK)
Germany	(DE)	Estonia	(EE)
Ireland	(IE)	Greece	(GR)
Spain	(ES)	France	(FR)
Croatia	(HR)	Italy	(IT)
Cyprus	(CY)	Latvia	(LV)
Lithuania	(LT)	Luxembourg	(LU)
Hungary	(HU)	Malta	(MT)
Netherlands	(NL)	Austria	(AT)
Poland	(PL)	Portugal	(PT)
Romania	(RO)	Slovenia	(SI)
Slovakia	(SK)	Finland	(FI)
Sweden	(SE)	United Kingdom	(UK)

4.3 Results

4.3.1 A composite indicator for waste management

The *Hierarchical Disjoint Non-Negative Factor Analysis* model (Chapter 2) has been applied on the data-set composed by 10 MIs for the 28 EU countries. Some MIs are taken into consideration during the analysis in order to enrich the information about countries and their performance in waste management (e.g., population, density of population, GDP pro capite, etc). Population has resulted as being the most appropriate element to normalize some MIs (see Table 4.1).

In order to formalize and analyze, in a general framework, the waste management indicator, we propose a hierarchically aggregated index that best represents the waste management in EU, via the statistical identification of reliable and uni-dimensional SCIs.

The SCIs, which represent dimensions, measure specific concepts contributing into the definition of waste management. Furthermore, the general CI reconstructs the MIs via a set of composite indicators according to reflective relations. Few missing data were present in the studied data-set, they were MCAR (Missing Completely at Random). Hence, under the assumption that each missing value could be approximated by the closest values to it, they have been imputed by the K-nearest neighbors method by setting $K = 10$ and by using the euclidean distance.

In order to set the right polarity of the MIs, the HDFNA model has been applied many times by following a recursive strategy with increasing number of dimensions. Since, the aim of this analysis is to get a CI for waste management in Europe through H reliable dimensions, the number of dimensions is not known a priori; and the best model (i.e., the model with the optimal number of dimensions) has been selected via the evaluation of the *BIC* (Schwarz 1978).

The optimal model is the model with 3 dimensions and the related factors, thus the one associated at the lowest value of *BIC* equal to 362.866, the screeplot (Cattell 1966) of the discrepancy function confirms the optimal number of dimensions is equal to 3. The partition

Table 4.3. Results of the optimal model for defining dimensions of waste management.

MIs	RCE	GW	PII	Std error	$Pr(p > \|Z\|)$
1	0	0.682	0	0.138	0.000
2	0	0.876	0	0.091	0.000
3	0	0	0.841	0.102	0.000
4	0	0.461	0	0.168	0.014
5	0	0	0.841	0.102	0.000
6	0.640	0	0	0.145	0.000
7	0.401	0	0	0.172	0.030
8	0	0.417	0	0.172	0.027
9	0.869	0	0	0.093	0.000
10	0.770	0	0	0.120	0.000
Uni-dimensionality	Yes	Yes	Yes		
Cronbach's α	0.751	0.705	0.828		
CI	0.964	0.385	0.338		

and the loadings are reported in Table 4.3. It is worth to observe that all the MIs result statistical significant into the model, all the factors are reliable (Cronbach's α higher than 0.70, Cronbach (1951)) and uni-dimensional (the variance of the second component of each class is lower than 1).

As reported in Table 4.3, the first factor is characterized by the presence of four MIs and the most important one is "Recycling rate of municipal waste", this factor is named *Recycling and Circular Economy* (RCE); the second factor is defined by four MIs as well and the most important is "Generation of recyclable waste", the second factor is called *Generation of Waste* (GW); whereas the third factor, denominated *Private Investments and Innovation* (PII), is characterized by the presence of two MIs: "Private investments as value added at factor cost related to circular economy" and "Patents related to recycling and secondary raw materials" with equal loadings.

The factor RCE is the most important in the construction of the general composite indicator *Waste Management* (WM), with loading equal to 0.964, wheareas the others two have similar loadings (0.385 for GW and 0.338 for PII). The contribution of the countries' performance in recycling activities is crucial in the definition of their "Waste Management" performance. It is possible to observe the reflective hierarchical structure in Figure 4.1.

An interesting result to underline is the low correlation between GW and PII, it seems that the countries that produce most waste are also the ones which less invest privately into circular economy. We can explain it by observing that the correlation between the MIs "Private investments as value added at factor cost related to circular economy" and "Generation of recyclable waste" is equal to 0.023.

4.3.2 Rankings and Specific Composite Indicators

For each SCI it is possible to rank the 28 EU countries and finally it is possible to observe the final ranking based on the CI "Waste Management", Table 4.4. Netherlands results to be the best country in terms of recycling performances, it is reflected on RCE's ranking and on the final CI's ranking (see Figure 4.2). Since the Spearman's correlations among SCIs are not high (as reported in Table 4.5), the behavior of countries is not equal for each aspect of WM. If we take into account only the group of the best countries (i.e., in green in Table 4.4)

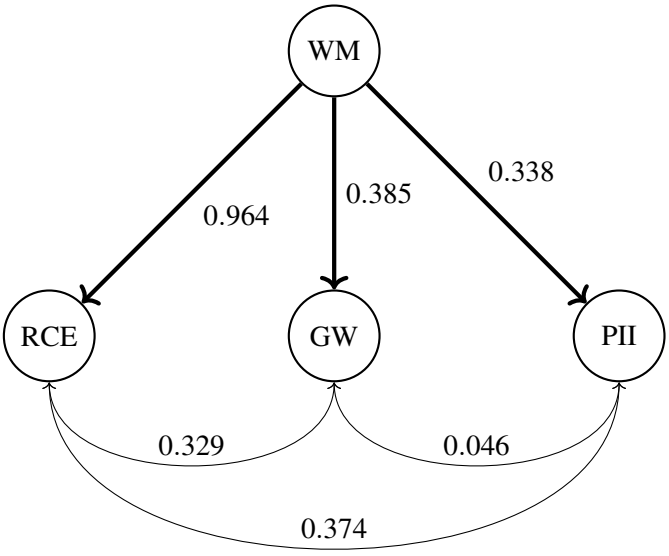


Figure 4.1. Path diagram of two-levels hierarchy

of each ranking, no country is present in all of them. Germany is present in the group of the best countries for RCE, PII and for the WM, but it is present in the group of intermediate countries (i.e., in yellow in Table 4.4) for GW. All the others, which are present in the group of the best countries for WM, are present in the group of the best countries only for one SCI. By taking into consideration the group of the worst countries (i.e., in red in Table 4.4) for WM, we can observe that seven countries (Portugal, Hungary, Slovakia, Croatia, Greece, Malta and Cyprus) are present in the group of the worst countries for all SCIs whereas only three (Ireland, Bulgaria and Romania) are present in the group of the worst countries for two rankings out of three.

Therefore, we can observe that the behavior of the best countries is quite different according to the three main aspects of WM whereas it is more stable for the countries which do not perform well in terms of WM.

4.4 Conclusion

The constant increasing of population and the related increasing generation of waste rises the necessity to develop consistent and useful tools to measure and monitor waste management in order to light the way of countries' policies. The idea of this work is to understand in which underlying dimensions the multidimensional construct Waste Management roots and measure the EU countries' performances both for each dimensions and for the general construct.

It is possible to detect two different goals for institutions' actions: sustainable consumption and smarter waste management. Thus, a plan of a coordinated actions through the EU countries is more and more crucial, our proposal represents a reliable measure of "good" waste management and we consider it needed.

In this paper, we propose a hierarchically aggregated CI for WM, by detecting three

Table 4.4. Rankings based on SCIs and GCI according to these thresholds, green: normalized score > 0.60, yellow: normalized score > 0.30 and < 0.60, red: normalized score < 0.30

RCE	GW	PII	WM
NL	LU	DE	NL
BE	FI	FR	BE
SI	BE	UK	DE
DE	EE	IT	LU
AT	SE	ES	SI
IT	AT	MT	FR
LU	NL	PL	AT
UK	BL	NL	UK
FR	UK	IE	IT
DK	DK	CZ	SE
SE	FR	AT	DK
LT	DE	FI	FI
PL	RO	BE	PL
CZ	SI	DK	LT
ES	SK	SE	CZ
LV	IT	LU	ES
FI	PT	RO	EE
IE	CZ	HU	IE
PT	IE	EE	PT
HU	GR	SI	HU
HR	PL	PT	BL
SK	MT	BL	SK
EE	ES	SK	LV
BL	HU	GR	HR
CY	HR	HR	RO
GR	LT	LT	GR
RO	CY	LV	MT
MT	LV	CY	CY

Table 4.5. Spearman's correlations among SCIs and with WM

	RCE	GW	PII
RCE	1	0.43	0.48
GW	0.43	1	0.32
PII	0.48	0.32	1
WM	0.95	0.64	0.56

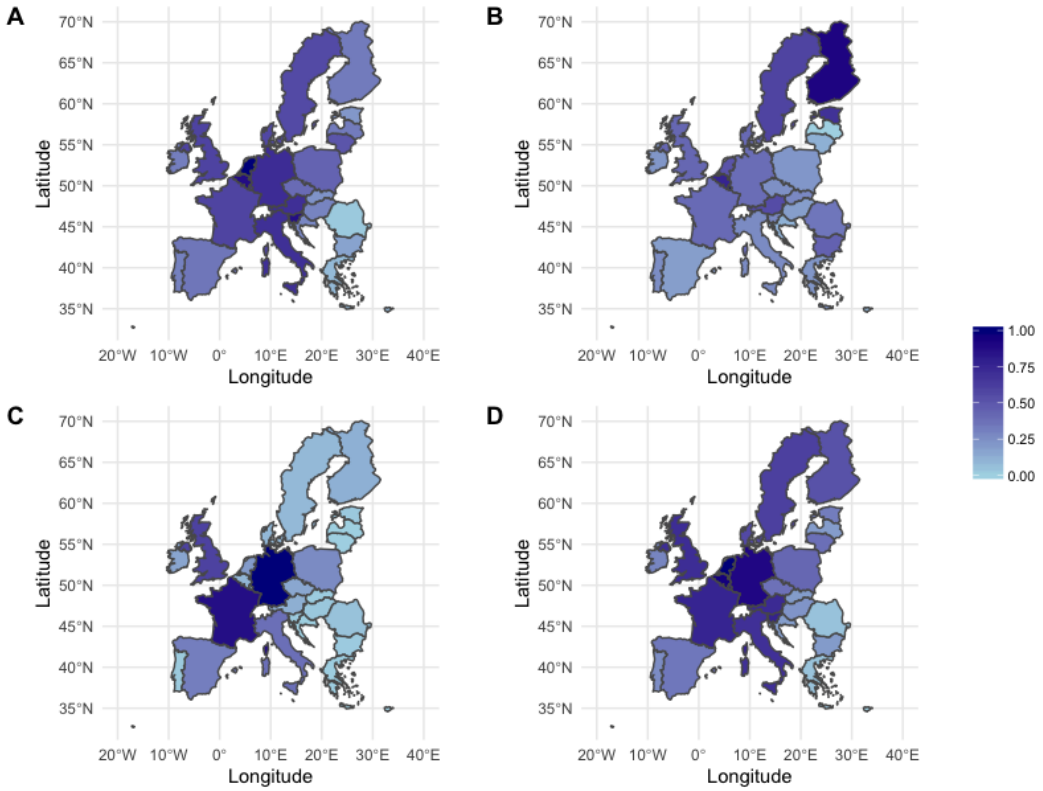


Figure 4.2. Normalized scores of Eu Countries: **A**) RCE, **B**) GW, **C**) PII and **D**) WM

important SCIs which represent dimensions, in order to provide a useful tool for policy-makers and institutions. The GCI is a general latent construct that best reconstruct the MIs. This approach has several advantages because the CI computed respects important properties: scale invariant, unidimensional, reliable and non-compensable, Section 3.5. This model-based approach limits the choices of the researcher which often are based on not verified theory.

The three SCIs which characterize WM are: RCE, GW and PII. The most important is RCE, GW and PII contribute less and with almost equal importance. The GCI is important in order to compare the countries' behaviors in terms of WM. Moreover it is possible to evaluate countries' performances for each specific aspect in order to understand why some countries perform better than others and in which way it is possible to improve their performances.

In conclusion, this study provides a useful tool to measure the "goodness" of WM in EU with its main aspects by identifying the most important relationships among MIs. The goal is provide a support for EU countries' actions and policies. The goal for future developments is to include new MIs into the framework and to extent the same conceptual model to the Italian case in order study the regions' behaviour.

Chapter 5

The ultrametric correlation matrix for modelling hierarchical latent concepts

5.1 Introduction

The dimensionality of a manifold of different phenomena is currently growing with the complexity of the reality as well as the relationships between their many facets. Multidimensional concepts often include more specific ones, highlighting the existence of a hierarchical structure to represent the multifaceted phenomena. Well-being, sustainable development, poverty, climate change, global innovation, are relevant examples of phenomena composed of hierarchical latent concepts (as already explained in Chapter 1). For instance, according to the United Nations Development Programme definition of the Multidimensional Poverty Index, poverty is described as a general concept resulting from the interaction of three key dimensions: *health*, *education* and *standard of living* (see Section 3.6.2). They are in turn defined by more specific concepts, as *nutrition* and *child mortality* for *health*; *years of schooling* and *school attendance* for education; *sanitation*, *housing*, etc. for *standard of living*. Thus, concepts are nested to fully describe the phenomenon and are latent because not directly observable and they need more MIs to be defined.

Therefore, data useful to accurately describe the studied phenomenon are "best" reconstructed at different levels of syntheses by means of some specific dimensions that reconstruct in turn some other concepts of a lower level or a subsets of observable, manifest indicators in the lowest level; thus, each dimension (or latent concept) is directly or indirectly specified by a set C_h of MIs. Assuming to pinpoint H MIs groups (clusters), two features characterise C_h ($h = 1,\ldots,H$): the *internal consistency of the latent concept*, i.e., the concordant relations observed within C_h that assess the *reliability* of the concept, and the *correlation between concepts*, i.e., the concordant and discordant relations between C_h and C_q ($h,q = 1,\ldots,H, h \neq q$). The former represents the development of MIs classes at different levels - each one associated with a dimension -, whereas the latter identifies all the hierarchical relationships existing between them.

If we hypothesise a hierarchical factorial structure among concepts we can observe that the design of this structure can be *crossed* or *nested*. In the former, all possible combinations of concepts between two levels are considered, whereas in the latter, higher levels of concepts

reconstruct only some (nested) lower levels of concepts or manifest indicators. The nested form has a graphical configuration of a tree, which is in turn associated with a hierarchical partition of manifest indicators. It is defined by sets of manifest indicators with different internal consistencies and some others which are broader MIs groups obtained by pairwise possible amalgamations of the initial ones according to their correlation, i.e., the magnitude of the relationships between concepts.

The study of hierarchical aggregation of factors is not new in many fields of research, for instance Revelle (1979) introduced a cluster analysis method very useful to detect clusters of MIs in a hierarchical approach. In this paper, we propose a new model to investigate in depth the hierarchical data correlation structure through an ultrametric correlation matrix in order to describe a general construct (i.e., concept) which needs a set of nested specific concepts to be fully defined and reconstruct the observed MIs. Our proposal can be considered into the *Dimensionality Reduction* framework for its ability of summarizing a big quantity of information by way of many steps of aggregation.

The model herein gives rise to a hierarchical partition which can be associated with the hierarchy of latent concepts defining a general broader and multidimensional concept (phenomenon). Thus, it provides a parsimonious hierarchical partition of MIs represented by parsimonious trees, which are composed by a limited number of internal nodes. This choice is strengthen by the assumption that only a restricted part of the data matrix contains relevant information for the analysis. Thus, to better represent the main features of the model dimensionality reduction from the original data is needed (Gordon 1999).

This model can be meant as an extension of the two level hierarchical model presented in Chapter 2, called HDNFA, for its capacity to detect all the $(2H-1)$ levels of a parsimonious hierarchy from H main concepts to the most general one. However, it is not the "natural" extension of HDNFA because Cavicchia, Vichi, and Zaccaria (2020b) aims to reconstruct a positive data correlation matrix $\mathbf{R}$ via an ultrametric correlation matrix, whereas HDNFA is supposed to model a covariance structure in order to build a CI through a reduced number of SCIs.

The introduction of an ultrametric correlation matrix aims to reconstruct the right order of aggregation (in case it is present) of concepts. The idea is detecting a general concept which is defined by a set of nested specific concepts through sequential aggregations starting from reliable and consistent concepts. This approach results to be proper for reflective CIs, where the general concept is reflected by specific concepts along the hierarchy with different levels of aggregation.

By considering correlation as a measure of similarity among MIs, one interpretation of our model can be as clustering method of MIs. Like Vichi (2008) solved the cluster analysis problem of partitioning a set of objects from dissimilarity data via a statistical model-based approach of fitting the “closest” classification matrix to the observed dissimilarities, our proposal has the same purpose by working on MIs and their similarities. Usually, a hierarchical clustering algorithm (for instance, the one proposed by Ward (1963) for units) is applied on dissimilarities and subsequently a partition is chosen in according to a visual inspection of the dendrogram. Our model is able to get the optimal partition of MIs starting from their correlation matrix $\mathbf{R}$, and on this partition it is able to build the entire hierarchy.

Therefore, our model has a double application: it is able to investigate the correlations among MIs in order to detect a reduced number of reliable concepts and it merges them sequentially according to their similarities. The ultimate goal is to have a general concept which reflects this aggregation process.

The model 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 studied general concept. Furthermore, to avoid that specific concepts compensate in the definition of a latent dimension, the correlation matrix of the data is assumed to be non-negative. Even if restrictive, the latter assumption turns out to be realistic in a manifold of real applications.

The paper is organised as follows. In Section 5.2, the notation used in all sections and some basic notions on hierarchical partitions and ultrametric matrices are given to allow the reader to follow the specification of the model herein. An in-depth description of the proposed methodology is provided in Section 5.3. Section 5.4 is dedicated to the non-parametric least-squares estimation of the model together with a coordinate descent algorithm. Section 5.5 discusses the results of a simulation study to assess the model; a real application is illustrated in Section 5.6 and a final discussion completes the paper in Section 5.7.

5.2 Notation and basic notions

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

J, H	number of manifest indicators, number of classes of the partition of manifest indicators.
C	set (partition) $\{C_1, \ldots, C_H\}$ of H classes of manifest indicators.
$\mathbf{R} = [r_{jl}]$	$(J \times J)$ data correlation matrix, where r_{jl} is the correlation between manifest indicators j and l $(j, l = 1, \ldots, J, l \neq j)$.
$\mathbf{R}_W = [{}_W r_{hh}]$	$(H \times H)$ within-concept consistency (diagonal) matrix, where ${}_W r_{hq} = 0$ for all $h \neq q$; ${}_W r_{hh} > 0$ $(h = 1, \ldots, H)$, represents the consistency within the h^{th} group of manifest indicators.
$\mathbf{R}_B = [{}_B r_{hq}]$	$(H \times H)$ between-concept correlation matrix, where ${}_B r_{hq}$ $(h, q = 1, \ldots, H, h \neq q)$ denotes the correlation between latent concepts associated with the classes of manifest indicators $\{C_h, C_q\} \subset C$; ${}_B r_{hh} = 1$ for all $h = 1, \ldots, H$.
$\mathbf{V} = [v_{jh}]$	$(J \times H)$ membership matrix, where $v_{jh} = 1$ if the j^{th} MI belongs to the h^{th} class C_h; $v_{jh} = 0$ otherwise. It pinpoints a partition C of manifest indicators in H classes, $C_1, \ldots, C_H$; thus, $\mathbf{V}$ is binary and row-stochastic, i.e., with one non-zero element per row.
$\mathbf{I}_J, \mathbf{I}_H$	identity matrix of order J and H, respectively.
$\mathbf{E} = [e_{jl}]$	$(J \times J)$ error matrix.

The *internal consistency* of a set C_h $(h = 1, \ldots, H)$ of manifest indicators is the extent to which all manifest indicators in C_h measure the same latent concepts (see Section 3.5.4). It is measured as a function of the correlations between manifest indicators in C_q. These values, one per group, can be arranged on the main diagonal of a diagonal matrix $\mathbf{R}_W$, where ${}_W r_{hq} = 0$ for all $h \neq q$, and ${}_W r_{hh}$ is the measure of the internal consistency of class C_h, $h = 1, .., H$. In other terms, the *internal consistency* of a set C_h $(h = 1, \ldots, H)$ of manifest indicators is the ability of all manifest indicators to measure the same latent concept. Many measure of *internal consistency* are reviewed by Revelle and R. Zinbarg (2009).

Given two sets C_h and C_q $(h \neq q)$ of manifest indicators, the *correlation between* them measures the extent to which manifest indicators in C_h are concordant or discordant with

manifest indicators in C_q and, therefore, measures the correlation between latent concepts. It is a function of correlations between manifest indicators, one belonging to C_h and the other belonging to C_q. For H classes of MIs C_h ($h = 1,..,H$), $H(H-1)$ correlations between pairs of classes - each one representing a latent concept - can be computed.

These values can be arranged in a square correlation matrix $\mathbf{R}_B = [{}_Br_{hq}]$ of order H, where ${}_Br_{hq}$ denotes the correlation between manifest indicators in classes C_h and C_q ($h,q = 1,\ldots,H, h \neq q$); hence, ${}_Br_{hh} = 1$ for all H.

For simplicity, let us suppose that the H classes of manifest indicators form a partition. This implies that each MI contributes to specify only one latent dimension. We now need to define a hierarchy of latent concepts and to specify the properties of an ultrametric matrix.

Definition 1. *A hierarchy of latent concepts is a set* $\mathscr{H}_H = \{C_h, (h = 1,\ldots,H), C_{H+1},\ldots,C_{2H-1}\} = \{C, C_{H+1},\ldots,C_{2H-1}\}$, *composed of* $2H-1$ *classes, each one representing a latent concept, where the first* $C_1,\ldots,C_H$ *stand for the subsets of an initial partition* C *of MIs with internal consistency* ${}_Wr_{hh}$ $(h = 1,\ldots,H)$*; the remaining classes, i.e.* $C_{H+1},\ldots,C_{2H-1}$*, are obtained by* $H-1$ *pairwise possible amalgamations of subsets of* C *with correlation between classes* ${}_Br_{hq}$ $(h,q = 1,\ldots,H, h \neq q)$*. Thus,* $C_h, C_k \subset \mathscr{H}_H \Leftrightarrow C_h \cap C_k \subset \{C_h, C_k, \emptyset\}$, $h,k = 1,\ldots,2H-1$.

Definition 2 ((Dellacherie, Martinez, and San Martin 2014))**.** *Given a set of objects I, a non-negative matrix* $\mathbf{R}$ *is said to be ultrametric if*

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

(ii) $r_{jj} \geq \max\{r_{kj} : k \in I\}$ *for all* $j \in I$ *(column point-wise 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 non-negative correlation matrix $\mathbf{R}$ satisfies properties (i) and (ii), i.e., it is symmetric and column point-wise diagonal dominant.

5.3 The model

Starting from the observation of a $(J \times J)$ non-negative data correlation matrix $\mathbf{R}$, the hierarchical statistical problem we want to deal with can be formalised as follows

$$\mathbf{R} = \mathbf{R}_u + \mathbf{E} \tag{5.1}$$

where the $(J \times J)$ matrix $\mathbf{R}_u$ represents the hierarchical structure of the latent concepts, and $\mathbf{E}$ is the random error matrix, i.e., the residual matrix. The non-negativity assumption of $\mathbf{R}$ and, consequently, of $\mathbf{R}_u$ is needed to pinpoint a non-compensatory hierarchical structure of

the latent concepts. Therefore, in evaluating the feasible bottom-up relationships between dimensions, the appraisal of their correlation or of their absolute magnitude leads to the same hierarchical solutions.

The proposed model to build an ultrametric correlation matrix for modeling hierarchical latent concepts is formally specified as follows

$$\mathbf{R}_u = \mathbf{V}(\mathbf{R}_B - \mathbf{I}_H)\mathbf{V}' + \mathbf{V}\mathbf{R}_W\mathbf{V}' - \operatorname{diag}\big(\operatorname{dg}(\mathbf{V}\mathbf{R}_W\mathbf{V}')\big) + \mathbf{I}_J \tag{5.2}$$

subject to constraints

$$\mathbf{V} = [v_{jh} \in \{0,1\} : j = 1,\ldots,J, h = 1,\ldots,H] \tag{5.3}$$

$$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J \quad \text{i.e.} \quad \sum_{h=1}^{H} v_{jh} = 1 \quad j = 1,\ldots,J \tag{5.4}$$

$$\mathbf{R}_B \text{ is an ultrametric correlation matrix (Definition 2)} \tag{5.5}$$

$$\min\{{}_W r_{hh} : h = 1,\ldots,H\} \geq \max\{{}_B r_{hq} : h,q = 1,\ldots,H, h \neq q\} \tag{5.6}$$

In (5.2), $\operatorname{dg}(\cdot)$ gives rise to a vector, whereas $\operatorname{diag}(\cdot)$ produces a diagonal matrix. Thus, $\operatorname{diag}\big(\operatorname{dg}(\mathbf{V}\mathbf{R}_W\mathbf{V}')\big)$ represents a diagonal matrix with diagonal elements equal to those of $\mathbf{V}\mathbf{R}_W\mathbf{V}'$.

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

The ultrametricity constraint (5.5) implies that the following $O(H^3)$ constraints on the triplets of $\mathbf{R}_B$ hold:

$$\begin{cases} {}_B r_{hq} \geq \min({}_B r_{hk}, {}_B r_{kq}) \\ {}_B r_{hk} \geq \min({}_B r_{hq}, {}_B r_{qk}) \\ {}_B r_{qk} \geq \min({}_B r_{hq}, {}_B r_{hk}) \end{cases} \qquad h = 1,\ldots,H, h = h,\ldots,H, k = h,\ldots,H \tag{5.7}$$

Proposition 1. *The number of different off-diagonal elements of $\mathbf{R}_u$ ($n_{\mathbf{R}_u}$) is*

$$\begin{cases} 1 \leq n_{\mathbf{R}_W} \leq H \\ 1 \leq n_{\mathbf{R}_B} \leq H-1 \end{cases} \quad \Rightarrow \quad 2 \leq n_{\mathbf{R}_u} \leq 2H-1 \tag{5.8}$$

where $n_{\mathbf{R}_W}$, $n_{\mathbf{R}_B}$ are the number of different diagonal elements of $\mathbf{R}_W$ and the number of different off-diagonal elements of $\mathbf{R}_B$, respectively.

Definition 3. *A $(2H-1)$-ultrametric correlation matrix is a square ultrametric matrix of order J with diagonal elements equal to one and off-diagonal elements that can assume one of at most $(2H-1)$ different values ${}_B r_{hq}$, ${}_W r_{hh}$, such that $0 \leq {}_B r_{hq} \leq {}_W r_{hh} \leq 1$.*

Lemma 1. *A hierarchy of $\mathscr{H}$ latent concepts, i.e. $\mathscr{H}_H$, - starting from J manifest indicators - with within-concept correlations, i.e., the class consistency, ${}_W r_{hh}$ $(h = 1,\ldots,H)$ and between-concept correlations ${}_B r_{hq}$ $(h,q = 1,\ldots,H,\ h \neq q)$, with $0 \leq {}_B r_{hq} \leq {}_W r_{hh} \leq 1$, is one-to-one associated with a $(2H-1)$-ultrametric correlation matrix $\mathbf{R}$.*

Proof. Considering a hierarchy of latent concepts $\mathscr{H}_H$, it can be stated that the j^{th} $(j = 1,\ldots,J)$ MI belongs to only one group C_h $(h = 1,\ldots,H) \subset \mathscr{H}_H$. The latter statement implies that any triplet (i,j,l) of manifest indicators certainly falls into one of the following scenarios:

(a) all the elements of the triplet belong to a single cluster C_h $(h = 1,\dots,H)$; (b) the elements of the triplet belong to two distinct clusters C_h and $C_q \subset \mathscr{H}_H$ $(h,q = 1,\dots,H, h \neq q)$; (c) all the elements of the triplets belong to different clusters $C_h, C_q, C_k \subset \mathscr{H}_H$ $(h,q,k = 1,\dots,H, k \neq q \neq h)$. According to the Definition 2-3 it can be gathered that (a), (b) and (c) correspond to the following correlation triplets: $({}_W r_{hh}, {}_W r_{hh}, {}_W r_{hh})$, $({}_W r_{hh}, {}_B r_{hq}, {}_B r_{hq})$ and $({}_B r_{hq}, {}_B r_{hk}, {}_B r_{qk})$, respectively. Furthermore, considering a hierarchy of latent concepts $\mathscr{H}_H$, all the triplets previously pinpointed verify the ultrametric inequality by definition (i.e. condition (iii) and (iii') of the Definition 2) and they identify an equilateral or an obtuse isosceles triangle. Thus, the $(2H-1)$-ultrametric matrix $\mathbf{R}$ has H levels ${}_W r_{hh}$ $(h = 1,\dots,H)$ corresponding to the subsets of C, while the remaining $H-1$ levels ${}_B r_{hq}$ $(h,q = 1,\dots,H,$ $h \neq q)$ match the subsets $C_{H+1},\dots,C_{2H-1}$.

Conversely, each $(2H-1)$-ultrametric correlation matrix $\mathbf{R}$ leads to a hierarchy of latent concepts, because for each pair of manifest indicators (j,l) they belong to the same class $C_h \subset C$ if their correlation is ${}_W r_{hh}$, or to different classes $C_h, C_k \subset C$ if their correlation is ${}_B r_{hk}$. Moreover, $C_{H+1},\dots,C_{2H-1}$ may contain only the triplets previously defined that are associated with non-overlapping groups since the ultrametricity constraint on $\mathbf{R}$. □

Theorem 1. *The matrix $\mathbf{R}_u$ defined in (5.2) is a $(2H-1)$-ultrametric correlation matrix.*

Proof. The matrix $\mathbf{R}_u$ defined in (5.2) can be rewritten as

$$\mathbf{R}_u = \mathbf{V}(\mathbf{R}_B - \mathbf{I}_H + \mathbf{R}_W)\mathbf{V}' - \operatorname{diag}\big(\operatorname{dg}(\mathbf{V}\mathbf{R}_W\mathbf{V}')\big) + \mathbf{I}_J \tag{5.9}$$

According to constraints (5.5)-(5.6) and the ultrametricity of $\mathbf{R}_W$, it is easily to demonstrate that $\mathbf{P} = \mathbf{R}_B - \mathbf{I}_H + \mathbf{R}_W$ turns out to be ultrametric. Indeed, $\mathbf{P}$ is symmetric as well as $\mathbf{R}_B$, since constraint (5.5) holds; it is column point-wise diagonal dominant since its diagonal elements are those of $\mathbf{R}_W$ and the constraint (5.6) holds; the ultrametric inequality holds for $\mathbf{P}$ observing that its possible triplets are contained in $\mathbf{R}_B$ which is assumed to be ultrametric. Thus, $\mathbf{V}(\mathbf{R}_B - \mathbf{I}_H + \mathbf{R}_W)\mathbf{V}'$ is in turn ultrametric because it may contain the only triplets defined in Lemma 1, and the remaining addends of (5.9) affect only the diagonal of $\mathbf{R}_u$, whose elements turn out to be unitary. Thus, the conditions of the Definition 2-3 are satisfied and the $(2H-1)$-ultrametricity of $\mathbf{R}_u$ is proved.

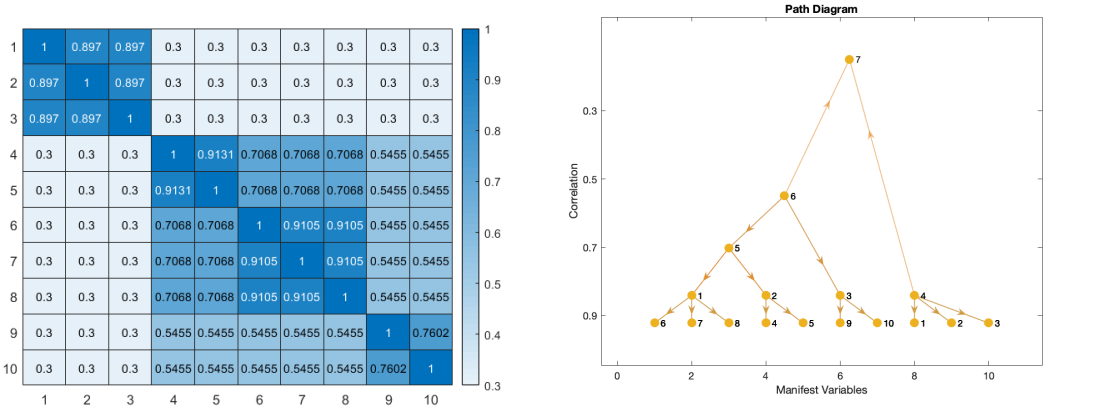
Moreover, a symmetric diagonally dominant real matrix with non-negative diagonal entries is positive semi-definite (Brouwer and Haemers 2012). The matrix $\mathbf{R}_u$ is a symmetric diagonally dominant real matrix since it is an ultrametric matrix and it has unitary diagonal elements. Thus, if $\mathbf{R}_u$ is defined according to (5.2) subject to constraints (5.3)-(5.6), then it is a correlation matrix. □

An example of the $(2H-1)$-ultrametric correlation matrix $\mathbf{R}_u$ and its corresponding hierarchical representation - one-to-one correspondence defined in Lemma 1 - is shown in Figure 5.1.

In the next section, the estimates of the proposed model defined in (5.2) are provided according to a non-parametric least-squares approach.

5.4 Least-Squares estimation of the model and Algorithm

The Least-Squares estimation of the model (5.2), when the correlation matrix is the $(2H-1)$-ultrametric matrix identifying a hierarchical latent concept model (Theorem 1), is defined



(a) $(2H-1)$-ultrametric correlation matrix. (b) Hierarchy of latent concepts: graphical representation.

Figure 5.1. Relationship between a $(2H-1)$-ultrametric correlation matrix and a corresponding hierarchy of latent concepts.

as the minimisation of the following quadratic constrained problem with respect to $\mathbf{R}_W, \mathbf{R}_B$ and $\mathbf{V}$

$$\begin{aligned} F(\mathbf{R}_W, \mathbf{R}_B, \mathbf{V}) =& \| \mathbf{R} - \mathbf{V}(\mathbf{R}_B - \mathbf{I}_H)\mathbf{V}' - \mathbf{V}\mathbf{R}_W\mathbf{V}' + \operatorname{diag}\big(\operatorname{dg}(\mathbf{V}\mathbf{R}_W\mathbf{V}')\big) - \mathbf{I}_J \|^2 \\ =& \sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{\substack{l=1 \\ l\neq j}}^{J}(r_{jl} - {}_W r_{hh})^2 v_{jh}v_{lh} \\ &+ \sum_{h=1}^{H}\sum_{\substack{q=1 \\ q\neq h}}^{H}\sum_{j=1}^{J}\sum_{\substack{l=1 \\ l\neq j}}^{J}(r_{jl} - {}_B r_{hq})^2 v_{jh}v_{lq} \quad \rightarrow \quad \min_{\mathbf{R}_W, \mathbf{R}_B, \mathbf{V}} \end{aligned} \tag{5.10}$$

subject to constraints (5.3)-(5.6).

5.4.1 Estimation of $\mathbf{R}_W$

The estimators of the elements of $\mathbf{R}_W$ are computed by differentiating (5.10) with respect to ${}_W r_{hh}$ $(h = 1, \ldots, H)$ for a fixed $\widehat{\mathbf{V}}$:

$${}_W\hat{r}_{hh} = \frac{\sum_{j=1}^{J}\sum_{\substack{l=1 \\ l\neq j}}^{J} r_{jl}\hat{v}_{jh}\hat{v}_{lh}}{\sum_{j=1}^{J}\sum_{\substack{l=1 \\ l\neq j}}^{J} \hat{v}_{jh}\hat{v}_{lh}} \qquad h = 1, \ldots, H \tag{5.11}$$

For fixed $\widehat{\mathbf{R}}_B$ and $\widehat{\mathbf{V}}$, the loss function (5.10) can be rewritten as

$$F(\mathbf{R}_W, \widehat{\mathbf{R}}_B, \widehat{\mathbf{V}}) = \| \tilde{\mathbf{R}} - \widehat{\mathbf{V}}\mathbf{R}_W\widehat{\mathbf{V}}' + \operatorname{diag}(\operatorname{dg}(\widehat{\mathbf{V}}\mathbf{R}_W\widehat{\mathbf{V}}')) \|^2 \tag{5.12}$$

where $\tilde{\mathbf{R}} = \mathbf{R} - \widehat{\mathbf{V}}(\widehat{\mathbf{R}}_B - \mathbf{I}_H)\widehat{\mathbf{V}}' - \mathbf{I}_J$ is the known residual matrix. (5.12) is minimised by

$$\widehat{\mathbf{R}}_W = \operatorname{diag}\big(\operatorname{dg}(\widehat{\mathbf{V}}'(\mathbf{R} - \mathbf{I}_J)\widehat{\mathbf{V}})\big)[(\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^2 - \widehat{\mathbf{V}}'\widehat{\mathbf{V}}]^+ \tag{5.13}$$

where $\mathbf{A}^+$ denotes the Moore-Penrose inverse of a matrix $\mathbf{A}$. The estimates in (5.11) are equivalent to the diagonal elements of $\widehat{\mathbf{R}}_W$ in (5.13).

It is worth to notice that a strict relationships between (Equation 5.11) and Cronbach's α (Cronbach 1951) holds:

$$\alpha_h = \frac{n_h \, {}_W\hat{r}_{hh}}{1+(n_h-1)_W\hat{r}_{hh}} \quad h=1,\ldots,H \tag{5.14}$$

where α_h is the Cronbach's α and n_h is the number of MIs related to the h-th concept, respectively.

5.4.2 Estimation of $\mathbf{R}_B$

The estimators of the elements of $\mathbf{R}_B$ are computed by differentiating (5.10) with respect to ${}_Br_{hq}$ $(h,q=1,\ldots,H,\, h\neq q)$ for a fixed $\widehat{\mathbf{V}}$ and subject to constraint (5.5):

$${}_B\hat{r}_{hq} = \frac{\sum_{j=1}^{J}\sum_{\substack{l=1\\ l\neq j}}^{J} r_{jl}\hat{v}_{jh}\hat{v}_{lq}}{\sum_{j=1}^{J}\sum_{\substack{l=1\\ l\neq j}}^{J} \hat{v}_{jh}\hat{v}_{lq}} \quad h,q=1,\ldots,H, h\neq q \tag{5.15}$$

For fixed $\widehat{\mathbf{R}}_W$ and $\widehat{\mathbf{V}}$, the loss function (5.10) can be rewritten as

$$F(\widehat{\mathbf{R}}_W,\mathbf{R}_B,\widehat{\mathbf{V}}) = \| \, \tilde{\mathbf{R}} - \widehat{\mathbf{V}}\mathbf{R}_B\widehat{\mathbf{V}}' \, \|^2 \tag{5.16}$$

where $\tilde{\mathbf{R}} = \mathbf{R} - \widehat{\mathbf{V}}\widehat{\mathbf{R}}_W\widehat{\mathbf{V}}' + \mathrm{diag}(\mathrm{dg}(\widehat{\mathbf{V}}\widehat{\mathbf{R}}_W\widehat{\mathbf{V}}')) - \mathbf{I}_J + \widehat{\mathbf{V}}\mathbf{I}_H\widehat{\mathbf{V}}'$ is the known residual matrix. The minimisation of (5.16) is a Penrose multivariate regression problem with the following matricial solution

$$\widehat{\mathbf{R}}_B = (\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^{-1}\widehat{\mathbf{V}}'\tilde{\mathbf{R}}\widehat{\mathbf{V}}(\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^{-1} \tag{5.17}$$

Since constraint (5.5) must be satisfied, the estimate of $\mathbf{R}_B$ is the closest - in the LS sense - ultrametric matrix to (5.17). To find this ultrametric solution an average linkage UPGMA algorithm is computed on the dissimilarity matrix $\tilde{\mathbf{D}}_B = \mathbf{1_H 1_H}' - \widehat{\mathbf{R}}_B$, whose elements belong to the interval $[0,1]$ since the non-negativity of $\mathbf{R}_B$. With an inverse transformation, i.e., $\widehat{\mathbf{R}}_B = \mathbf{1}_H\mathbf{1}'_H - \tilde{\mathbf{D}}_B$, we obtain the ultrametric estimate of $\mathbf{R}_B$.

5.4.3 Estimation of V

The estimators of the elements of $\mathbf{V}$ are computed minimising (5.10) row by row for each $\mathbf{v}_j$ $(j=1,\ldots,J)$ of $\mathbf{V}$, when all the remaining rows, $\widehat{\mathbf{R}}_W$ and $\widehat{\mathbf{R}}_B$ are fixed. Specifically, the j^{th} MI is assigned to the q^{th} class $(v_{jq}=1)$ if (5.10) reaches its minimum with respect to $\mathbf{V}$. Formally, each row $\mathbf{v}_j$ of $\mathbf{V}$, $j=1,\ldots,J$, is estimated as follows

$$\begin{cases} \hat{v}_{jh}=1 & \text{if} \quad \underset{h=1,\ldots,H}{\arg\min} F(\widehat{\mathbf{R}}_W,\widehat{\mathbf{R}}_B,[\hat{\mathbf{v}}_1,\ldots,\mathbf{v}_j=\mathbf{i}_h,\ldots,\hat{\mathbf{v}}_J]') \\ \hat{v}_{jh}=0 & \text{otherwise} \end{cases} \tag{5.18}$$

where $\mathbf{i}_h$ is the h^{th} row of the identity matrix of order H.

5.4.4 Algorithm for detecting the LS Ultrametric Correlation Matrix

Given H, a coordinate descent algorithm for the estimation of the model (5.2) can be described by two basic steps which are sequentially repeated until a stopping rule is satisfied.

Step 0 [*Initialization*] A small non-negative convergence tolerance value ε is fixed and an initial partition $\mathbf{V}^{(0)}$ of manifest indicators in non-empty groups is computed, according to constraints (5.3)-(5.4).

Step 1 [*Update* $\mathbf{R}_W$ *and* $\mathbf{R}_B$] For the current matrix $\mathbf{V}^{(t-1)}, t = 1,\ldots,$ $\mathbf{R}_W^{(t)}$ and $\mathbf{R}_B^{(t)}$ are updated according to (5.13) and (5.17), respectively. The UPGMA average linkage algorithm is applied to the dissimilarity matrix $\tilde{\mathbf{D}}_B^{(t)}$, computed by $\mathbf{R}_B^{(t)}$, to obtain a new $\tilde{\mathbf{D}}_B^{(t)}$ satisfying the ultrametric inequality on distance matrices. Then, the ultrametric matrix $\tilde{\mathbf{D}}_B^{(t)}$ is transformed into the ultrametric correlation matrix by $\mathbf{R}_B^{(t)} = \mathbf{1}_H\mathbf{1}_H' - \tilde{\mathbf{D}}_B^{(t)}$. If constraint (5.6) does not hold for some ${}_Br_{hq}$, we substitute these values with $\min\{{}_Wr_{hh}, h = 1,\ldots,H\}$.

Step 2 [*Update* $\mathbf{V}$] For the current matrices $\mathbf{R}_W^{(t)}$ and $\mathbf{R}_B^{(t)}$, $\mathbf{V}^{(t)}$ is updated solving the assigned problem (5.18). Each row j^{th} of $\mathbf{V}^{(t)}$ has a unique non-zero element, specifically equal to 1, which is set in the column (class of manifest indicators) that minimises the loss function F. Passing to the next row, the loss function never increases and generally decreases.

Stopping rule The function $F(\mathbf{R}_W^{(t)}, \mathbf{R}_B^{(t)}, \mathbf{V}^{(t)})$ is computed. If the decrease of the loss function is larger than ε (set in **Step 0**), the algorithm continues to iterate **Step 1** and **Step 2**, otherwise the algorithm stops and the process is considered to have converged. Thus, $\widehat{\mathbf{R}}_W = \mathbf{R}_W^{(t)}$, $\widehat{\mathbf{R}}_B = \mathbf{R}_B^{(t)}$, $\widehat{\mathbf{V}} = \mathbf{V}^{(t)}$.

The algorithm is NP-hard as clustering problems described by Křivánek and Morávek (1986), it is computationally efficient since it rapidly converges to a stationary point. Moreover, the parsimony property guarantees a low complexity. Thus, it is worth to notice that the computational time and space are reduced relative to an algorithm on a data matrix, since at most $2H - 1 + J$ elements per iteration are stored.

5.5 Simulation

In order to assess the performances of our model defined by (5.1), we have implemented a simulation study by following the ultrametric correlation structure determined in (5.2). Two different scenarios have been taken into account to test the model with a small scale correlation structure, characterised by $J = 30$ and $H = 4$, and with a larger one, with $J = 100$ and $H = 4$.

The error matrix $\mathbf{E}$ has been generated starting from a MultiVariate Normal distribution with a zero-mean vector and a diagonal covariance matrix (J uncorrelated dimensions), i.e., $\mathbf{E} \sim \sigma_E\, MVN_J(\mathbf{0_J}, \mathbf{I}_J)$. Two standard deviations σ_E are fixed to consider a small and a large level of error: $\sigma_E = 0.03$ and $\sigma_E = 0.15$, respectively. Then, the matrix $\mathbf{E}$ has been symmetrized and its diagonal has been set to zero. For each generated matrix $\mathbf{R}$ according to (5.1), we have verified if a proper correlation matrix (i.e., positive semi-definite and with values between 0 and 1) has been determined.

Table 5.1. Simulation study results.

	Scenario 1		Scenario 2	
	Small error	Large error	Small error	Large error
MSE($\mathbf{R}_W$)	0.0021	0.0131	0.0018	0.0644
MSE($\mathbf{R}_B$)	0.0475	0.0541	0.0214	0.0693
% ARI= 1	100%	69.0%	100%	62.5%
ARI Mean	1	0.9576	1	0.9455

Table 5.2. List of Bechtoldt Manifest Indicators.

Factor	Manifest Indicators	ID	Factor	Manifest Indicators	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

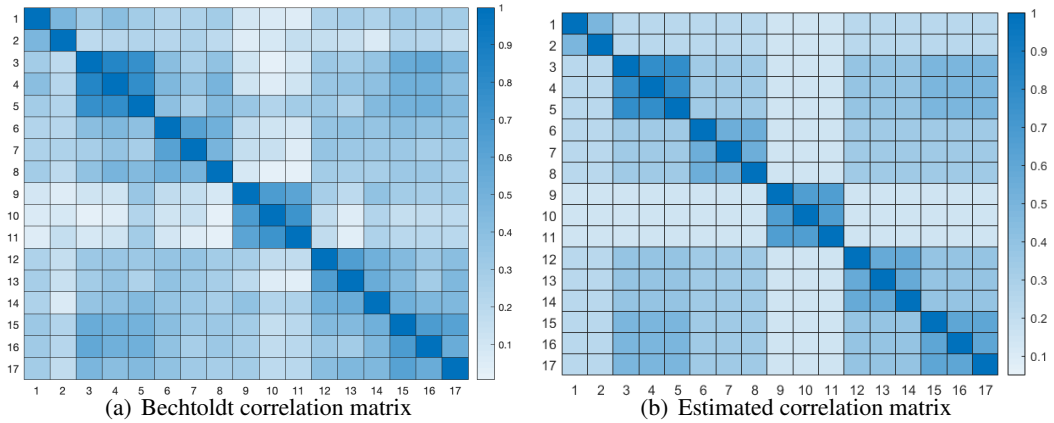
The simulated model is evaluated according to the ARI, which compares the generated membership matrix $\mathbf{V}$ with the estimated one $\widehat{\mathbf{V}}$. Furthermore, the Mean Squared Error (MSE) of the within-concept consistency matrix $\mathbf{R}_W$ and the between-concept correlation matrix $\mathbf{R}_B$ is computed. All the simulation study results are shown in Table 5.1.

We have generated 200 correlation matrices for each scenario and each level of error. The MI 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 is high it masks the generated correlation structure, and the detection is not always correct (69.0% and 62.5% of samples with ARI equal to 1 for the first and the second scenario, respectively). Nevertheless, the mean of the ARI is high in all cases and the values of the MSE for $\mathbf{R}_W$ and $\mathbf{R}_B$ are extremely good.

Moreover, the number of random starts has been set equal to 50. The algorithm has shown relevant performances in terms of running time and it converges in few steps. Nevertheless, it is advisable to increase the initial random starts when the correlation structure is not clearly determined.

5.6 Application

The Bechtoldt data set contains 17 manifest indicators, shown in Table 5.2, with 6 proposed latent factors (Bechtoldt 1961; Kano 1997). Our goal is to correctly estimate the MI 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



(a) Bechtoldt correlation matrix (b) Estimated correlation matrix

Figure 5.2. Comparison between heat maps of observed and estimated correlation matrices.

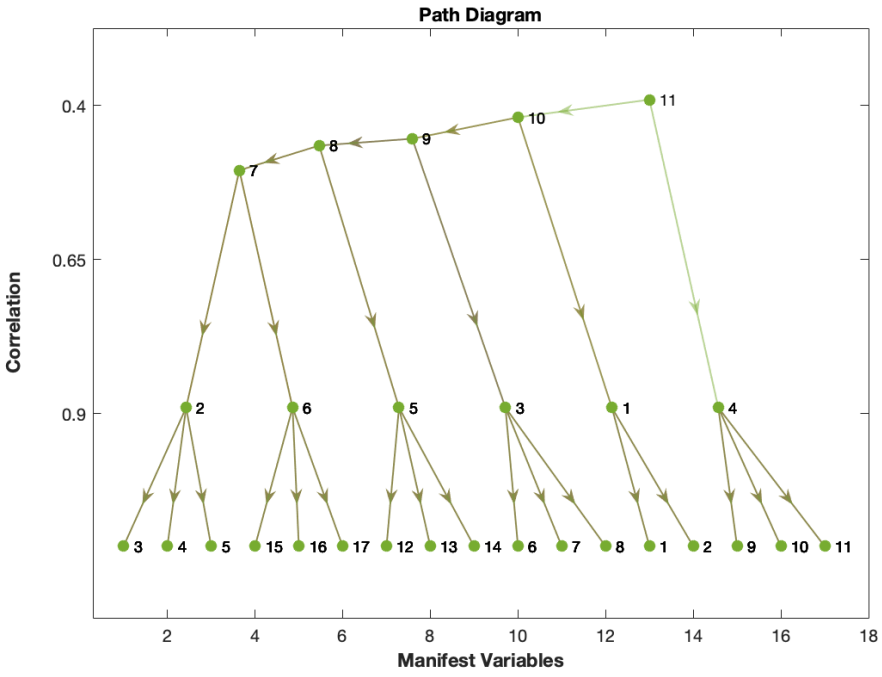


Figure 5.3. Aggregation of latent concepts in Benchtoldt hierarchical structure.

understand the relationships among the manifest indicators groups.

The data set is presented as a correlation matrix with all non-negative values (Figure 5.2(a)). The algorithm run with 50 random starts has found the right partition in 6 main groups (Figure 5.2(b)) corresponding to the theoretical ones. All the clusters of manifest indicators are reliable except the first, i.e. *Memory*, whose Cronbach's α is equal to 0.631.

The hierarchical structure pinpointed by the model starts from the first aggregation of two groups, corresponding to the latent concepts *Verbal* and *Reasoning*, which enlarges from the bottom of the hierarchy upwards as shown in Figure 5.3. It is worth to notice that even if the graphic representation in Figure 5.3 shows a constant trend in the definition of broader dimensions, we could characterise three main concepts: (i) *Intellectual Abilities* composed of the concepts *Verbal, Reasoning, Number, Word*; (ii) *Memory* and (iii) *Space* which are defined according to the theoretical groups of the original manifest indicators. 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 aggregation in Figure 5.3 it can be observed the existence of a general concept, which is clearly evident and it may represent the dimension *Mental Abilities*.

5.7 Conclusion

A novel for correlation matrices is proposed. It investigates in depth the non-compensatory hierarchical structure of the latent concepts and of the MIs through an ultrametric correlation matrix. The argumentation is focused on the non-negativity property of the studied correlation matrices, which is common in psychometric studies and in marketing applications, just to give some of the many fields of application.

A correlation matrix $\mathbf{R}$ is reconstructed via a ultrametric hierarchical structure which highlights two crucial characteristic regarding clusters of MIs: the *internal consistency* and the *correlation between clusters*. In order to detect the ultrametric structure of the latent concepts, it is important to investigate in depth the reliability of each cluster of MIs and all the relations among them. In other terms, for correlation matrices $\mathbf{R}$ which are composed only of non-negative elements, as common in psychometric applications, it has been presented a model that considers two main matrices, the first one which contains non-zero element only on the diagonal, that is the *internal consistency* measure for the related cluster, and the second one which is a correlation matrix with the off-diagonal elements that represent the *correlation between clusters*. The Dimensionality Reduction model with a hierarchical structure that goes from disjoint sets of MIs to the General Concept (GC) is given by detecting consistent clusters and by following correlation between them.

The ultrametric correlation structure of the model allows us to disclose a parsimonious hierarchy from the manifest indicators up to the most general concept, composed of all indicators. A LS estimation 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 application provided show the extremely good performances of the model. The coordinate descent algorithm proves very fast and stable.

The goal for future studies is a development of our model to general correlation matrices, relaxing the non-negativity constraint in order to have a model also when compensation among concepts is required, assessing the performance of the model through a more exhaus-

tive simulation study. It is worth to develop both an alternative way to avoid that specific concepts compensate in the definition of a latent dimension and a objective criterion to choose H.

Chapter 6

Discussion

The construction of model-based Composite Indicators (CIs) represents an interesting methodological challenge for researchers. This thesis wishes to contribute to the current research on hierarchical latent variable models for *Dimensionality Reduction* with a specific focus on the construction of CIs.

Our society is developing a huge quantity of data in many applicative fields (e.g., socio-economic, psychometrics, marketing, etc.) and this work hopes to represent an important step forward into the literature presenting model-based approaches with the crucial goal to acquire new knowledge from the empirical evidence in order to provide useful support. The motivation for the proposed work is grounded into the necessity to provide model-based CIs which are built according to a *statistical* approach avoiding subjective choices (e.g., normative weights). It is worth to underline that many valid methods to construct a CI and to estimate weights of aggregation are present in literature. For instance, Hierarchical Confirmatory Factor Analysis (HCFA, Holzinger (1944), Jöreskog (1966), Jöreskog (1969), Jöreskog (1978), and Jöreskog (1979)) allows to estimate the magnitude of the correlation between factors and the related MIs when a theoretical framework is available and the goal is to confirm it. However, a theoretical framework can be not available or available only in part, hence, an exploratory approach can be adopted in order to reconstruct the latent structure of the data. An example of exploratory analysis is Hierarchical Factor Analysis (HFA) (Thompson 1951; Wherry 1959; Wherry 1975; Wherry 1984) which is a sequential strategy to compose factors step by step.

The aim of this dissertation is to provide methodological proposals to build CIs via simultaneously estimations when the theoretical framework is not available (i.e., at least in part). Starting form this point of view, the thesis presents an introduction (Chapter 1) and four main chapters (2-5) which contain both methodological proposals and applications on real data.

Chapter 1 provides an overall introduction to the definition of CIs and their main characterizing aspects. The most used models, in Multidimensional Data Analysis (MDA) framework, for the construction of CIs are presented in order to show the starting point of the analysis. The discussion is focused on the two different types of *measurement model* (i.e., formative and reflective), Factor Analysis (FA) and its hierarchical extensions (Section 1.3).

Chapters 2-5 discuss hierarchical latent variable models for *Dimensionality Reduction* with related applications on CIs. The whole thesis is characterized by the assumption

to build CIs according to a reflective approach by considering CIs as latent constructs (i.e., factors). In particular, Chapter 2 presents a new latent factor model that could be used for modelling CIs. It might be seen as a hierarchical extension of the DFA by Vichi (2017), since an unknown Simple Structure Model (SSM) is hypothesised for the data and each cluster of Manifest Indicators (MIs) must be internally consistent and reliable, that is, MIs related to the factor measure "consistently" a unique theoretical construct. This implies that MIs are concordant with the related factor and loadings must be positive. This last requirement is included as a constraint in the new methodology that for this reason is named Hierarchical Disjoint Nonnegative Factor Analysis (HDNFA). The aim of the model is to investigate the hierarchical structure of the data in order to define a General Composite Indicator (GCI) based on a set of H Specific Composite Indicators (SCIs), by generalizing non-orthogonal DFA with additional constraints. It is worth to underline that both Hierarchical Factor Analysis (HFA, Section 2.2) and HDNFA (Section 2.4) are novel, the first one has to be considered as a two level hierarchical extension of FA whereas the second one is its disjoint version with non-negative loadings. Although the model has been presented with an application about well-being, it has wide applicability in other fields of study, in particular, it has been tested on datasets about sport, psychometrics, economics, and so on and so forth.

Chapter 3 extends the argumentation about CIs by stressing out their use for tracking coherent policy conclusions. The approach presented in Chapter 3 aims to underline features, properties and rules that should be respected by the research in order to build "good" CIs. For instance, it is crucial to test the theoretical framework without taking it for granted, checking whether the MIs compose a system of indicators or not, which MIs should be excluded from the analysis and which ones it is worth including. Chapter 3 recalls the model presented in Chapter 2 by proposing an alternative estimation (i.e., least-squares estimation) of it. The achievement is to provide a model-based approach for constructing transparent and objective CIs which can be used as support for policy-making. A set of properties (i.e., scale-invarince, non-compensability, reliability and unidimensionality) has been used in order to underline some problems which affect well-known indicators such as HDI and MPI.

Chapter 4 presents a hierarchical model for the multidimensional phenomenon *Waste Management* (WM) in European Union (EU). It is presented as an application of HDNFA (Chapter 2). WM is increasing its importance through the years and the number of statistics dedicated to it is getting larger and larger, therefore the construction of a CI able to reduce the dimensionality of the phenomenon and to highlight the main dimensions of it is useful in order to provide support to EU countries' action and policies. Three main reliable dimensions have been detected in order to define *Waste Management*: *Recycling and Circular Economy* (RCE), *Generation of Waste* (GW) and *Private Investments and Innovation* (PII). RCE is the most important, whereas GW and PII contribute less and with almost equal importance. This analysis allows us to understand the countries' performances for each specific aspect (i.e, dimension) in order to understand why some countries perform better than others and in which way it is possible to improve their performances. This study provides a useful tool to measure the "goodness" of WM in EU with its main aspects by identifying the most important relationships among MIs and the goal for future is to include other MIs into the framework and to extent the same conceptual model to the Italian case in order study the regions' behaviour.

Chapter 5 introduces a model for positive data correlation matrices which aims to detect of H realiable concepts and to build the $(2H-1)$ levels hierarchy from them to the most

general one. HDNFA is supposed to model the covariance structure in order to build a CI through the detection of a reduced number of reliable latent concepts, whereas the model presented in Section 5.3 aims to reconstruct a positive data correlation matrix via an ultrametric correlation matrix which is characterized by two main matrices: the one (a positive definite diagonal matrix) which has the *internal consistency* measures on its diagonal, and the one (ultrametric correlation matrix) which has the *correlations between clusters* as off-diagonal elements. This model is illustrated via an application to Bechtoldt data set which is composed by 17 MIs divided into 6 clusters defining 6 latent concepts. The data set is presented as a correlation matrix with all non-negative values. A future goal is to extend the model to a generic correlation matrix by developing an alternative way to avoid the compensation among concepts.

6.1 Open problems and future research

The models introduced in this thesis are all developed in the context of hierarchical latent variable models for *Dimensionality Reduction* with the goal of building CIs or, at least, of detecting latent concepts which respect certain properties (e.g., reliability).

It is of interest for future research to extend the aforementioned models by relaxing constraints, for instance the row stochasticity of the membership matrix adding some cross-loadings (Chapter 2,3,4) and the non-negativity of the correlation matrix (Chapter 5).

Vichi (2017) already included the possibility to add cross-loadings in order to increase the fit of DFA, by following the step-wise strategy proposed by Raftery (1993) for SEM selection, that might be followed also for cross-loading identification in HDNFA. Therefore, after HDNFA identifies the best estimation for the parameters, the strategy should apply a forward search approach to identify the most relevant cross-loading that greatly improves the SSM. The inclusion of cross-loadings into the model implies that a MI might reflect several specific factors at the same time instead of just one. This approach, thus, involves starting from the best SSM without cross-loadings, followed by the decision of adding a cross-loading based on a model selection criterion (e.g., BIC). The model adds the cross-loading that most decreases the criterion, and repeats this process until a relevant decrease of the criterion is observed.

Another important problem that stays unsolved is the inclusion of a clustering model into HDNFA in order to detect simultaneously the clusters of observation and the hierarchy to build the CI. This aspect is different from applying a clustering method on the CI's scores, since the model should be estimated simultaneously according to a coordinate descent algorithm. This idea recalls what Vichi and Saporta (2009) did in order to simultaneously cluster MIs and observations in a PCA and LS framework. However, a further step is possible by considering the inclusion of a hierarchical clustering method instead of a partitioning one. This latter could have an important applicability in the visualization of the results.

Finally, the most important development with regards to the HDNFA model is the introduction of a time dependence in order to provide a model able to fit the phenomenon of study throughout the years. CIs are often used to compare performances of different observations in time and to detect their trend. In order to include the time dependence, two ideas are being developed: the first consists of encompassing a time dependence in the loading matrices, as already done by Maruotti and Vichi (2014) on the centroids matrices in their time-varying extension of k-means. The second idea consists of the inclusion of a

Hidden Markov Model (HMM) in order to follow the evolution of MIs depending on both the specific factors (i.e., SCIs) and on the general one (i.e., GCI).

By considering the model presented in Chapter 5, the most important future work regards the relaxing of the non-negativity constraint of the correlation matrix, which might allow us to explore general correlation matrices in order to avoid them any kind of transformation.

All these models and methods aim at being very applicable, hence it is of the upmost importance to make these tools available for free. For this reason, we are developing a MATLAB and Octave package containing all the routines used in this thesis. CIs are widely used and it is crucial to provide new models for the their construction and the related software.

Bibliography

Aaker, D. and R. Bagozzi (1979). "Unobservable Variables in Structural Equation Models With an Application in Industrial Selling". In: *Journal of Marketing Research* 16, p. 147.

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, pp. 199–213.

Alkire, S. (2010). *Human Development: Definitions, Critiques, and Related Concepts*. Human Development Research Papers (2009 to present). Human Development Report Office (HDRO), United Nations Development Programme (UNDP).

Alkire, S. and J. Foster (2011). "Counting and multidimensional poverty measurement". In: *Journal of Public Economics* 95.7, pp. 476–487.

Alkire, S., J. Foster, et al. (2015). *Multidimensional Poverty Measurement and Analysis*. Oxford: Oxford University Press.

Anderson, J.C. and D.W. Gerbing (1982). "Some Methods for Respecifying Measurement Models to Obtain Unidimensional Construct Measurement". In: *Journal of Marketing Research* 19.4, pp. 453–460.

Anderson, T.W. and H. Rubin (1956). "Statistical inferences in factor analysis". In: *Proceedings of the Third Symposium on Mathematical Statistics and Probability* 5, pp. 111–150.

Bartholomew, D.J. (1981). "Posterior analysis of the factor model". In: *British Journal of Mathematical and Statistical Psychology* 34.1, pp. 93–99.

Bartlett, M.S. (1937). "The statistical conception of mental factors". In: *British Journal of Psychology. General Section* 28.1, pp. 97–104.

Bechtoldt, H.P. (1961). "An empirical study of the factor analysis stability hypothesis". In: *Psychometrika* 26.4, pp. 405–432.

Bickel, P.J. and K.A. Doksum (1981). "An Analysis of Transformations Revisited". In: *Journal of the American Statistical Association* 76.374, pp. 296–311.

Blalock, H.M. (1964). *Causal Inferences in Nonexperimental Research*. Norton, New York, NY.

Bollen, K.A. (1984). "Multiple indicators: Internal consistency or no necessary relationship?" In: *Quality and Quantity* 18.4, pp. 377–385.

— (1989). *Structural Equations with Latent Variables*.

— (2001). "Indicator: Methodology". In: *International Encyclopedia of the Social and Behavioral Sciences*, pp. 7282–7287.

— (2011). "Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models". In: *MIS Quarterly* 35.2, pp. 359–372.

Bollen, K.A. and S. Bauldry (2011). "Three Cs in measurement models: causal indicators, composite indicators, and covariates". In: *Psychological methods* 16.3, pp. 265–284.

Bollen, K.A. and R. Lennox (1991). "Conventional Wisdom on Measurement: A Structural Equation Perspective". In: *Psychological Bulletin* 110, pp. 305–314.

Bollen, K.A. and K. Ting (2000). "A Tetrad Test for Causal Indicators". In: *Psychological methods* 5, pp. 3–22.

Box, G.E.P. and D.R. Cox (1964). "An Analysis of Transformations". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2, pp. 211–252.

Bro, R. (1997). "PARAFAC. Tutorial and applications". In: *Chemometrics and Intelligent Laboratory Systems* 38.2, pp. 149–171.

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

Browne, M.W. (2001). "An Overview of Analytic Rotation in Exploratory Factor Analysis". In: *Multivariate Behavioral Research* 36.1, pp. 111–150.

Brüggemann, R. and G. Patil (2011). *Ranking and Priotization for Multi-indicator System*. New York, NY: Springer International Publishing.

Budden, M., P. Hadavas, and L. Hoffman (2008). "On The Generation Of Correlation Matrices". In: *Applied Mathematics E-Notes,* 8, pp. 279–282.

Burt, R.S. (1973). "Confirmatory Factor-Analysis Structures and the Theory Construction Process". In: *Sociological Methods and Research* 2, pp. 131–187.

— (1976). "Interpretational Confounding of Unobserved Variables in Structural Equation Models". In: *Sociological Methods and Research* 5, pp. 3–52.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates Publishers.

Carlsen, L. and R. Brüggemann (2013). "An Analysis of the 'Failed States Index' by Partial Order Methodology". In: *Journal of Social Structure* 14.

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

Carroll, R.J. and D. Ruppert (1981). "On Prediction and the Power Transformation Family". In: *Biometrika* 68.3, pp. 609–615.

Cattell, R.B. (1957). *Personality and motivation structure and measurement*. Oxford, England: World Book Co.

— (1966). "The Scree Test For The Number Of Factors". In: *Multivariate Behavioral Research* 1.2, pp. 245–276.

Cattell, R.B. and D.K. Khanna (1977). "Principles and procedures for unique rotation in factor analysis". In: *Mathematical methods for digital computers*. Ed. by K. Enslein and A. Ralston. Wiley - Interscience. Chap. 9.

Cavicchia, C., P. Sarnacchiaro, and M. Vichi (2020). "A composite indicator for the waste management in the EU via Hierarchical Disjoint Non-Negative Factor Analysis". Manuscript submitted for publication.

Cavicchia, C. and M. Vichi (2020a). "Hierarchical Disjoint Non-Negative Factor Analysis". Manuscript submitted for publication.

— (2020b). "Statistical Model-based Composite Indicators for Complex Socio-Demographic and Economic Phenomena: Models, Properties and Features for tracking coherent policy conclusions". Manuscript submitted for publication.

Cavicchia, C., M. Vichi, and G. Zaccaria (2020a). "Hierarchical Disjoint Principal Component". Manuscript submitted for publication.

— (2020b). "The ultrametric correlation matrix for modelling hierarchical latent concepts". Manuscript submitted for publication.

Chaouachi, S.G. and K.S.B. Rached (2012). "Perceived Deception in Advertising: Proposition of a Measurement Scale". In: *Journal of Marketing Research and Case Studies*.

Chen, S.X., L. Zhang, and P. Zhong (2010). "Tests for High-Dimensional Covariance Matrices". In: *Journal of the American Statistical Association* 105.490, pp. 810–819.

Clarkson, D.B. and R.I. Jennrich (1988). "Quartic rotation criteria and algorithms". In: *Psychometrika* 53.2, pp. 251–259.

Cronbach, L.J. (1951). "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16.3, pp. 297–334.

d'Aspremont, A. et al. (2007). "A Direct Formulation for Sparse PCA Using Semidefinite Programming". In: *SIAM Review* 49.3, pp. 434–448.

Davey, B.A. and H.A. Priestley (2002). *Introduction to Lattices and Order*. Cambridge: Cambridge University Press.

Dellacherie, C., S. Martinez, and J. San Martin (2014). *Inverse M-Matrices and Ultrametric Matrices*. Lecture Notes in Mathematics. Springer International Publishing.

Despotis, D.K. (2005). "A reassessment of the human development index via data envelopment analysis". In: *Journal of the Operational Research Society* 56.8, pp. 969–980.

DeYoung, C.G., J.B. Peterson, and D.M. Higgins (2002). "Higher-order factors of the big five predict conformity: Are there neuroses of health?" In: *Personality and Individual Differences* 33.4, pp. 533–552.

Diamantopoulos, A. and J.A. Siguaw (2006). "Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration". In: *British Journal of Management* 17.4, pp. 263–282.

Digman, J.M. (1997). "Higher-order factors of the big five". In: *Journal of Personality and Social Psychology* 73, pp. 1246–1256.

Ebert, U. and H. Welsch (2004). "Meaningful environmental indices: a social choice approach". In: *Journal of Environmental Economics and Management* 47.2, pp. 270–283.

Edwards, J.R. (2011). "The Fallacy of Formative Measurement". In: *Organizational Research Methods* 14.2, pp. 370–388.

Edwards, J.R. and R.P. Bagozzi (2000). "On the Nature and Direction of the Relationship between Constructs and Measures". In: *Psychological Methods* 5.2, pp. 155–174.

Escofier, B. and J. Pages (1998). *Analyses Factorielles Simples et Multiples : Objectifs, Methodes et Interpretation*. Vol. 3e ed. Sciences sup. Mathematiques. Dunod.

Eysenck, H.J. (1967). *The biological basis of personality*. Thomas, Springfield.

Eysenck, H.J. and H.T. Himmelweit (1947). *Dimensions of personality: a record of research carried out in collaboration with H.T. Himmelweit [and others]*. London: Routledge Kegan Paul.

Fattore, M. (2016). "Partially Ordered Sets and the Measurement of Multidimensional Ordinal Deprivation". In: *Social Indicators Research* 128.2, pp. 835–858.

— (2017). "Synthesis of Indicators: The Non-aggregative Approach". In: *Complexity in Society: From Indicators Construction to their Synthesis*. Ed. by F. Maggino. 1st ed. Social Indicators Research Series. Springer International Publishing, pp. 193–212.

Fattore, M. and F. Maggino (2014). "Partial Orders in Socio-economics: A Practical Challenge for Poset Theorists or a Cultural Challenge for Social Scientists?" In: *Multi-indicator Systems and Modelling in Partial Order*. Ed. by R. Brüggemann, L. Carlsen, and J. Wittmann. New York, NY: Springer New York, pp. 197–214.

Ferrara, C., F. Martella, and M. Vichi (2016). “Dimensions of Well-Being and Their Statistical Measurements”. In: *Studies in theoretical and applied statistics*. Ed. by G. Alleva and A. Giommi, pp. 85–99.

— (2018). “Probabilistic Disjoint Principal Component Analysis”. In: *Multivariate Behavioral Research*.

Fornell, C. and D.F. Larcker (1981). “Evaluating Structural Equation Models with Unobservable Variables and Measurement Error”. In: *Journal of Marketing Research* 18.1, pp. 39–50.

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

Gordon, A.D. (1999). *Classification*. 2nd ed. Chapman and Hall/CRC.

Grbovic, M., C.R. Dance, and S. Vucetic (2012). “Sparse Principal Component Analysis with Constraints”. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI 12. Toronto, Ontario, Canada: AAAI Press, pp. 935–941.

Green, S.B., R.W. Lissitz, and S.A. Mulaik (1977). “Limitations of Coefficient Alpha as an Index of Test Unidimensionality”. In: *Educational and Psychological Measurement* 37.4, pp. 827–838.

Guttman, L. (1954). “Some necessary conditions for common-factor analysis”. In: *Psychometrika* 19.2, pp. 149–161.

— (1955). “The determinacy of factor score matrices with implications for five other basic problems of common-factor theory”. In: *British Journal of Statistical Psychology* 8.2, pp. 65–81.

Harman, H.H. (1976). *Modern Factor Analysis*. The University of Chicago Press, Chicago.

Harshman, R.A. (1970). “Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis”. In: *UCLA Working Papers in Phonetics* 16.

Hauser, R.M. and A.S. Goldberger (1971). “The treatment of unobservable variables in path analysis”. In: *Sociological Methodology* 3, pp. 81–117.

Hayduk, L.A. (1980). *Causal Models in Marketing*. New York: John Wiley and Sons, Inc.

— (1987). *Structural Equation Modeling with LISREL: Essentials and Advances*. Baltimore, MD, USA: Johns Hopkins University.

Holzinger, K.J. (1944). “A simple method of factor analysis”. In: *Psychometrika* 9, pp. 257–262.

Hooper, D., J. Coughlan, and M.R. Mullen (2008). “Structural Equation Modelling: Guidelines for Determining Model Fit”. In: *The Electronic Journal of Business Research Methods* 6.1, pp. 53–60.

Horn, J.L. and R.B. Cattell (1966). “Refinement and test of the theory of fluid and crystallized general intelligences”. In: *Journal of Educational Psychology* 57.5, pp. 253–270.

— (1982). “Whimsy and misunderstanding of gf-gc theory: A comment on Guilford”. In: *Psychological Bulletin* 91.3, pp. 623–633.

Horst, P. (1965). *Factor Analysis of Data Matrices*.

Hotelling, H. (1933). “Analysis of a Complex of Statistical Variables Into Principal Components”. In: *Journal of Educational Psychology* 24.6, 417–441 and 498–520.

Hubert, L. and P. Arabie (1985). “Comparing Partitions”. In: *Journal of Classification* 2.1, pp. 193–218.

Jarvis, C.B., S.B. MacKenzie, and P.M. Podsakoff (2003). "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research". In: *Journal of Consumer Research* 30.2, pp. 199–218.

Jennrich, R.I. and P.F. Sampson (1966). "Rotation for simple loadings". In: *Psychometrika* 31.3, pp. 313–323.

Jolliffe, I.T., N.T. Trendafilov, and M. Uddin (2003). "A modified principal component technique based on the LASSO". In: *Journal of Computational and Graphical Statistics* 12.3, pp. 531–547.

Jöreskog, K.G. (1966). "Testing a simple structure hypothesis in factor analysis". In: *Psychometrika* 31.2, pp. 165–178.

— (1969). "A general approach to confirmatory maximum-likelihood factor analysis". In: *Psychometrika* 34.2, pp. 183–202.

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

— (1978). "Structural analysis of covariance and correlation matrices". In: *Psychometrika* 43.4, pp. 443–477.

— (1979). *A general approach to confirmatory maximum likelihood factor analysis with addendum. In Jöreskog, K.G. and Sörbom, D. [Eds.] Advances in factor analysis and structural equation models.*

Jöreskog, K.G. and A.S. Goldberger (1975). "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable". In: *Journal of the American Statistical Association* 70.351a, pp. 631–639.

Jöreskog, K.G. and D. Sörbom (1984). *LISREL VI, analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Scientific Software, Inc.

Kaiser, H.F. (1958). "The varimax criterion for analytic rotation in factor analysis." In: *Psychometrika* 23.3, pp. 187–200.

— (1960). "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* 20.1, pp. 141–151.

Kano, Y. (1997). "Exploratory Factor Analysis with a Common Factor with Two Indicators". In: *Behaviormetrika* 24.2, pp. 129–145.

Kaplan, D. (2000). *Structural equation modeling: foundations and extensions*.

Kline, P. (2000). *The handbook of psychological testing*. Second edition. London; New York: Routledge.

Křivánek, M. and J. Morávek (1986). "NP-hard Problems in Hierarchical-tree Clustering". In: *Acta Informatica* 23.3, pp. 311–323.

Land, K. (1970). "On the estimation of path coefficients for unmeasured variables from correlations among observed variables". In: *Social Forces* 48, pp. 506–511.

Lawson, C.L. and R.J. Hanson (1974). *Solving least squares problems*. Pratice-Hall, Englewood.

Le Dien, S. and J. Pages (2003). "Analyse Factorielle Multiple Hierarchique". In: *Revue de statistique applique* 51.2, pp. 47–73.

MacCallum, R.C. and M.W. Browne (1993). "The use of causal indicators in covariance structure models: Some practical issues." In: *Psychological Bulletin* 114.3, pp. 533–541.

Madden, D. (2010). "Ordinal and cardinal measures of health inequality: an empirical comparison". In: *Health Economics* 19.2, pp. 243–250.

Maggino, F., ed. (2017). *Complexity in Society: From Indicators Construction to their Synthesis*. 1st ed. Vol. 70. Social Indicators Research Series. Springer International Publishing.

Maggino, F. and B.D. Zumbo (2012). "Measuring the Quality of Life and the Construction of Social Indicators". In: *Handbook of Social Indicators and Quality of Life Research*. Ed. by K.C. Land, A.C. Michalos, and M.J. Sirgy. Springer Netherlands, pp. 201–238.

Maraun, M.D. (1996). "Metaphor Taken as Math: Indeterminancy in the Factor Analysis Model". In: *Multivariate Behavioral Research* 31.4, pp. 517–538.

Marsh, H.W., J.R. Balla, and R.P. McDonald (1988). "Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size". In: *Psychological Bulletin* 103.3.

Martínez, R. (2012). "Inequality and the new human development index". In: *Applied Economics Letters* 19.6, pp. 533–535.

Maruotti, A. and M. Vichi (2014). "Time-varying clustering of multivariate longitudinal observations". In: *Communications in Statistics - Theory and Methods* 45.

Mazziotta, M. and A. Pareto (2013). "Methods for constructing composite indices: One for all or all for one?" In: *Rivista Italiana di Economia Demografia e Statistica* 67.2, pp. 67–80.

McDonald, R.P. (1974). "The measurement of factor indeterminacy". In: *Psychometrika* 39, pp. 203–222.

McGillivray, M. (1991). "The human development index: Yet another redundant composite development indicator?" In: *World Development* 19.10, pp. 1461–1468.

McNemar, Q. (1946). In: *Psychological Bulletin* 43.4, pp. 289–374.

Mulaik, S. (2009). *Foundations of Factor Analysis*. New York, NY: Chapman and Hall/CRC.

Mulaik, S. and R.P. McDonald (1977). "The effect of additional observables on factor indeterminancy in model with a single common factor". In: *Psychometrika* 4, pp. 177–192.

Munda, G. and M. Nardo (2009). "Noncompensatory/nonlinear composite indicators for ranking countries: a defensible setting". In: *Applied Economics* 41.12, pp. 1513–1523.

Nardo, M., M. Saisana, A. Saltelli, and S. Tarantola (2005). *Tools for Composite Indicators Building*. report EUR 21682. European Commission (Join Research Centre, Ispra, Italy).

Nardo, M., M. Saisana, A. Saltelli, S. Tarantola, et al. (2005). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Statistics Working Papers 2005/3. OECD Publishing.

Neubourg, C. de, M. de Milliano, and I. Plavgo (2014). "Lost (in) Dimensions: Consolidating Progress in Multidimensional Poverty Research". In: *Innocenti Working Papers*.

Noorbakhsh, F. (1998). "The human development index: some technical issues and alternative indices". In: *Journal of International Development* 10.5, pp. 589–605.

OECD (2001). *The Well-being of Nations: The Role of Human and Social Capital*. Tech. rep. OECD, Paris.

— (2004). *The OECD-JRC Handbook on Practices for Developing Composite Indicators, paper presented at the OECD Committee on Statistics*.

OECD-JRC (2008). *Handbook on Constructing Composite Indicators. Methodology and User Guide*. Paris: OECD Publisher.

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". In: *Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.

Radloff, L.S. (1977). "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population". In: *Applied Psychological Measurement* 1.3, pp. 385–401.

Raftery, A.E. (1993). "Bayesian model selection in structural equation models". In: *Testing structural equation models*. Ed. by K.A. Bollen and Long J.S. Sage, Newbury Park, pp. 163–180.

— (1995). "Bayesian Model Selection in Social Research". In: *Sociological Methodology* 25, pp. 111–163.

Revelle, W. (1979). "Hierarchical Cluster Analysis and the Internal Structure of Tests". In: *Multivariate Behavioral Research* 14.1, pp. 57–74.

Revelle, W. and R. Zinbarg (2009). "Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sijtsma". In: *Psychometrika* 74, pp. 145–154.

Ronald, I. and C. Welzel (2005). *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge: Cambridge University Press.

Rozeboom, W. W. (1996). "What might common factors be?" In: *Multivariate Behavioral Research* 31, pp. 555–570.

Sagar, A.D. and A. Najam (1998). "The human development index: a critical review". In: *Ecological Economics* 25.3, pp. 249–264.

Composite indicators – the controversy and the way forward (2004). OECD World Forum on Key Indicators.

Schmid, J. and J.M. Leiman (1957). "The development of hierarchical factor solutions". In: *Psychometrika* 22.1, pp. 53–61.

Schmitt, N. (1996). "Uses and abuses of coefficient alpha". In: *Psychological Assessment* 8.4, pp. 350–353.

Schönemann, P.H. (1973). "The minimum average correlation between equivalent sets of uncorrelated factors". In: *Psychometrika* 38.1, pp. 149–149.

Schönemann, P.H. and J.H. Steiger (1978). "On the validity of indeterminate factor scores". In: *Bulletin of the Psychonomic Society* 12.4, pp. 287–290.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *Annals of Statistics* 6.2, pp. 461–464.

Sen, A.K. (1981). *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Clarendon Press.

— (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.

— (1992). *Inequality Reexamined*. Oxford: Clarendon Press.

Sherin, R.J. (1966). "A matrix formulation of Kaiser's varimax criterion". In: *Psychometrika* 31.4, pp. 535–538.

Stiglitz, J.E., A.K. Sen, and J.P. Fitoussi (2009). *Report by the commission on the measurement of economic performance and social progress*. Tech. rep. Commission on the Measurement of Economic Performance and Social Progress, Paris.

ten Berge, J.M.F. (1984). "A joint treatment of varimax rotation and the problem of diagonalizing symmetric matrices simultaneously in the least-squares sense". In: *Psychometrika* 49.3, pp. 347–358.

— (2005). *Least squares optimization in multivariate analysis*. DSWO Press.

ten Berge, J.M.F, D.L. Knol, and H.A.L. Kiers (1988). "A treatment of the Orthomax rotation family in terms of diagonalization, and a re-examination of a singular value approach to Varimax rotation". In: *Computational Statistics Quarterly* 4.3, pp. 207–217.

Thompson, G.H. (1934). "The meaning of "i" in the estimate of "g"". In: *British Journal of Psychology* 25.1, pp. 92–99.

— (1951). *The factorial analysis of human ability (5th ed.)* New York: Houghton Mifflin.

Thurstone, L.L. (1947). *Multiple Factor Analysis*. Chicago [etc.] : The University of Chicago press.

Trabold-Nübler, H. (1991). "The human development index—A new development indicator?" In: *Intereconomics: Review of European Economic Policy* 26.5, pp. 236–243.

Transforming our world: the 2030 Agenda for Sustainable Development (2015). United Nations – Sustainable Development knowledge platform.

Trinchera, L. and G. Russolillo (2010). "On the use of Structural Equation Models and PLS Path modeling to build composite indicators". Working paper no. 30, Università Degli Studi Di Macerata.

Ullman, J. (2006). "Structural Equation Modeling: Reviewing the Basics and Moving Forward". In: *Journal of personality assessment* 87, pp. 35–50.

UNDP (2018). *Human Development Indices and Indicators 2018: Statistical update*. UN, New York.

Vichi, M. (2008). "Fitting semiparametric clustering models to dissimilarity data". In: *Advances in Data Analysis and Classification* 2.2, pp. 121–161.

— (2017). "Disjoint factor analysis with cross-loadings". In: *Advances in Data Analysis and Classification* 11.3, pp. 563–591.

Vichi, M. and G. Saporta (2009). "Clustering and disjoint principal component analysis". In: *Computational Statistics & Data Analysis* 53.8, pp. 3194–3208.

Vinzi, V.E. et al. (2010). *Handbook of Partial Least Squares*. Springer.

Ward, J.H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244.

Welzel, C., R. Inglehart, and H.D. Kligemann (2003). "The theory of human development: A cross-cultural analysis". In: *European Journal of Political Research* 42.3, pp. 341–379.

Werts, C.E., R.L. Linn, and K.G. Jöreskog (1974). "Intraclass Reliability Estimates: Testing Structural Assumptions". In: *Educational and Psychological Measurement* 34.1, pp. 25–33.

Wherry, R.J. (1959). "Hierarchical factor solutions without rotation". In: *Psychometrika* 24.1, pp. 45–51.

— (1975). "Underprediction from overfitting: 45 years of shrinkage". In: *Personnel Psychology* 28.1, pp. 1–18.

— (1984). *Contributions to correlational analysis*. New York: Academic Press.

Wilson, E.B. (1996). "A Review of 'The Abilities of Man, Their Nature and Measurement' by C. Spearman". In: *Science* 67, pp. 244–248.

Yeomans, K. and P. Golder (1982). "The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 31.3, pp. 221–229.

Zinbarg, R.E. and D.H. Barlow (1996). "Structure of anxiety and the anxiety disorders: A hierarchical model". In: *Journal of Abnormal Psychology* 105.2, pp. 181–193.

Zinbarg, R.E., D.H. Barlow, and T.A. Brown (1997). "Hierarchical structure and general factor saturation of the anxiety sensitivity index: Evidence and implications". In: *Psychological Assessment* 9.3, pp. 277–284.

Zinbarg, R.E., W. Revelle, et al. (2005). "Cronbach's [Alpha], Revelle's [Beta], and McDonald's [Omega][sub H]: Their Relations with Each Other and Two Alternative Conceptualizations of Reliability". In: *Psychometrika* 70.1, pp. 123–133.

Zou, H., T. Hastie, and R. Tibshirani (2006). "Sparse Principal Component Analysis". In: *Journal of Computational and Graphical Statistics* 15.2, pp. 265–286.