Università degli Studi di Napoli
Federico II

SEM with ordinal manifest variables

An Alternating Least Squares approach

Daniela Nappo

Tesi di Dottorato di Ricerca in
Statistica

XXII Ciclo

Dipartimento
di Matematica e Statistica
Università degli Studi di Napoli "Federico II"
via Cintia, Monte Sant'Angelo – 80126 Napoli

SEM with ordinal manifest variables

An Alternating Least Squares approach

Napoli, 30 Novembre 2009

Contents

List of Tables

List of Figures

Introduction

Survey data analysis in the marketing research, in the public-opinion survey and in the social research, are often characterized by different typologies of variables, measured on interval scale, nominal and ordinal scale, with a prevalence of the last two.

The simultaneous treatment of variables measured on different scale implies an homogenization problem, that can be solved recoding all variables, numerical and ordinal, to the lowest information level, that at the nominal scale levels such an approach usually implies a loss of information and does not allow to use the more informative quantitative analysis.

A prior quantification of the ordinal and nominal variables based on external optimal scaling technique is largely used in literature. Alternatively optimal scaling approaches can be integrated inside the metodology of data anlysis or modeling.

The choice of the scale level to take in consideration depends by the type of variables that are dominant in the survey and by the method with which the data are analyzed. Researches on Customer Satisfaction, scholastic evaluation, healthcare analysis of a population are typically based on ordinal scale, in this thesis we will refer, especially, to this kind of variables , exploring an approach that drives at the quantification inside a Structural Equation Modeling (SEM) context. The objectives, considered here, are twofold as to identify the latent variables underlying the dimensions explored and their explanation on the base of outer variables that can be latent or manifest ones.

Our aim to design a model that contrary to the explorative approaches, as Principal Component Analysis (numerical variables), Multiple Correspondence Analysis (for nominal variables) and Princals (for mixed variables, including ordinal variables), proposes a confirmative approach that represents a change of paradigm in respect with the approach proposed in the sixties by Benzecrì and the French school, according to which "the models must follow the data", to an approach closed to the soft modeling, developed in the recent years by the psychometric school, for which the hypothesis and the a-priori knowledge of the reality, conceptualized as a model to verify empirically, reversed the Benzecrì principle for which in this case "the data must follow the model and no vice versa".

Unlike the econometric approach in which the models are the product of a theory supported by a large knowledge of the topic to face, the soft modeling calculates at the same time, whether the identification of the latent variables as in the explorative analysis, or the estimation of the relationships between them as parameters of a model generally linear.

In our thesis work we have addressed a particular attention to the ordinal case, in the perspective to introduce a method of quantification of these last variables (ordinal), but that can be extended also to the nominal variables, maintaining unaltered the numerical ones. With this scope we have developed in the well know framework of Structural Equation Models estimation, based on the Partial Least Squares method, an original algorithm that pursues the optimal quantification of ordinal variables and nominal variables according to an Alternating Least Squares (ALS) logic.

Our thesis work starts from the quantification problems presented in a complex survey made by AVSI (Associazione Volontari per il Servizio Internazionale), to which we have participated [32]. The survey has the aim to evaluate the impact on the status of Orphan Vulnerable Children (OVC), residents in three Countries of Africa sudsahariana (Rwanda, Uganda and Kenya) of the supports given to the children, during three years, to their school, healthcare and nutritional aspects as wellas to their family environment. The model supposed on the base of this survey [32] is reported in figure 5.1. This model was

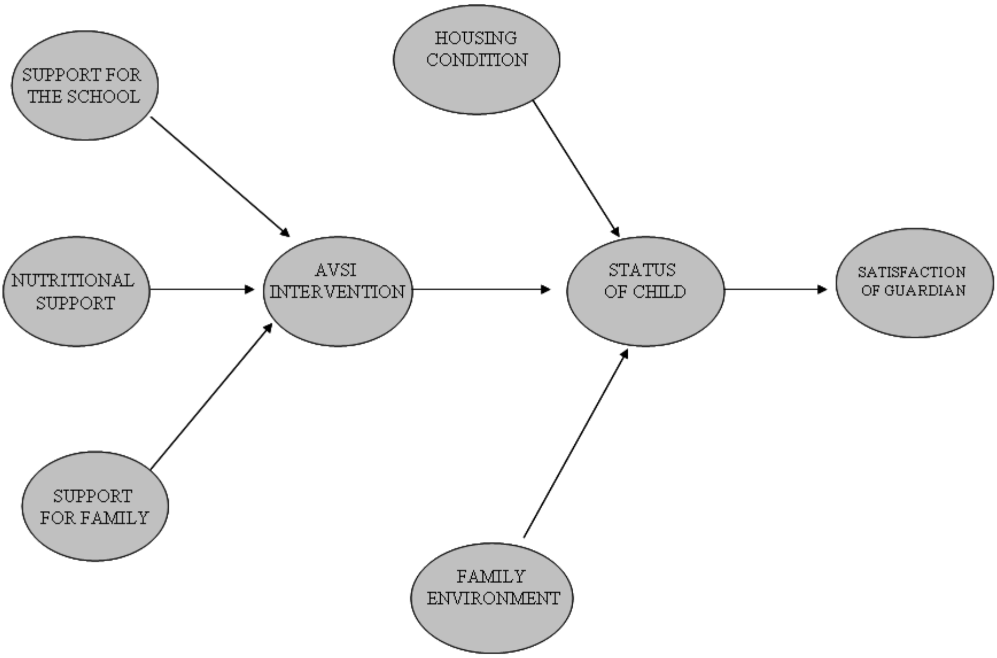


Figure 1: Status of child model

somministrated to 1155 children, of which the manifest variables are all ordinal, except for two variables that are nominal and measure the leaving condition of these children. So the structural model consists of 8 latent variables (six exogenous and 2 endogenous): three latent endogenous block summarize the **Status of child** , **Family characteristics**, **Housing condition** (the characteristics of the house where children live), **Avsi intervention** that is a super block defined by three latent variables, that describe the kind of support offered for the **Family**, for the **School**, and **Nutritional**. An outcome block of the model is associated to the **Guardian satisfaction** depending on the general Status reached by the Child in the year of the Survey.

The manifest variables on which the latent variables rest are quite all (33) ordinal, but for 2.

This is the applicative background on which centres our methodological approach, that has a more general perspective that strictly applicative. Our vision respect to the statistical research is to develop a methodology purposive to real problems and that permits to allure from the reality that lesson based on the integration between methods and knowledge a-priori, rather than a research formal merely that can be redundant sometimes in respect to the reality observed, or do not consider it. In this direction we follow the Tukei's thought:

Exploratory Data Analysis (EDA) is detective work - numerical detective work - or counting detective work - or graphical detective work ... unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider ... [it] can never be the whole story, but nothing else can serve as the foundation stone - as the first step. [Tukey, 1977, p. 1-3]

John Tukey proposed a new approach to data analysis, based heavily on visualization, as an alternative to classical (mathematical) data analysis. Being dependent on graphics, this approach only became practical with the advent of modern computers. However, he proposed the methodology of data exploration, a methodology in which a model of the phenomena might be inferred instead of pre-imposed.

It is this powerful combination that led him to coin the phrase "exploratory data analysis", commonly referred to simply as "EDA".

The exploratory approach is very appropriate for data analysis because it allows you to explore your data with an open mind. Tukey suggests that you think of exploratory analysis as the first step in a two-step process similar to that utilized in criminal investigations. In that first step, you search for evidence using all of the investigative tools that are available. In the second step, that of confirmatory data analysis, you evaluate the strength of the evidence and judge its merits and applicability. It is in this second step that you would likely evaluate the model(s) which you have inferred during your exploration and likely apply the techniques of classical data analysis.

To this philosophy engendered on a soft modeling idea belong our PALSOS-PM approach: similarly to PLS-PM, our methodology has two souls, explorative because it starts with the observation of the data on which one a model is built, and the confermative one because after the use of a statistical technique it tries to confirm the model inferred.

The thesis work goes through four chapters.

The **first chapter** is a presentation of qualitative variables, with a particular attention to the ordinal ones, and of the methods of external quantification of most.

In **second chapter** two approaches to the estimation of a SEM model are presented, LISREL and PLS-PM, showing the difference between them, and for which the problem of ordinal variables is discussed, pointing out the actual proposals for the quantification in the literature.

In the **third chapter**, that represents the core of the thesis, the PALSOS-PM approach is developped. The first part of this chapter is dedicated to explain the process of internal quantification adopted in the ALS algorithms, with the description of Princals and Morals procedures. The central part of the chapter regards the PALSOS-PM algorithm, in which the characteristics of the procedure and the validation process are introduced. An application of PALSOS-PM to a well known customer satisfaction dataset in the literature is estimated, to evaluate the advantage of our approach with resoect to the classical PLS-PM with no scaling options.

In the **fourth chapter** the model and the variables of a the Avsi model are described. The model, called "The status of child", is issued by a database of AVSI. The variables collected are ordinal, and someone are expression of qualitative characteristics (for example the characteristics of house in which the children live). This model estimated with the PALSOS-PM, allows to evidentiate the property of our approach to face very complex data.

A **Conclusion** and perspective ends the thesis to highlight with the aim the main results achieved, the critical aspects and some future development.

Chapter 1

The external quantification of ordinal variables

Hirschfeld, Fisher (with an "appropriate scoring" technique) and Guttman proposed in the thirties to associate real values to the modalities of a nominal variable, in such a way to optimize an external criterion of analysis. The coding of an ordinal variable is the latest and the first significant work was written by Kruskal (1965) on the analysis of monotone variance. So with the term "quantification" (or "scaling" or "scoring") we indicate a transformation of one or several categorical variables, normally ordinary, into numerical ones. The advantage of quantifying non numerical variables consists in the possibility to use classical multivariate techniques such as the Principal Component Analysis, Multiple Regression or Discriminant analysis.

Optimal Scaling is a multidimensional analysis that is based on the association of numerical scores to ordinal variables, across a transformation method. This technique is justified by the necessity to have metodologies capable to elaborate ordinal variables, taking into acount their characteristics and exceeding the classical coding in terms of equidistant intger scores, as collected by the questionnaires.

Bock (1960), that introduced the term "optimal scaling", defines:

"The aim of optimal scaling is to assign numerical values to alternatives or categories, so as to discriminate optimally among the objects, in some sense. Usually it is the least squares sense, and the values are chosen so that the variance between objects after scaling is a maximum with respect to that within objects".

We must choose one of the possible transformations, in order to satisfy some criterion related to the kind of analysis which will be performed afterwards: for example in the regression model it is the maximization of R^2. In this way we have an optimal transformation of variables, taking into account, on the one hand, the aim of analysis, and on the other hand, the nature and the process that the variables originated from. The largest use of scaling has been in psycological and psycometric contest to:

- verify one or more hypothesis on the data
- describe the structure of data, pointing out one or more latent dimensions
- develop a unidimensional scale to assign a score to each subject, variables or both, to use then the new variables in successive analysis

Furthermore, any multidimensional scaling method for qualitative variables (such as correspondence analysis) which gives coordinates for the categories of a set of qualitative variables is in fact a multidimensional quantification technique: the coordinates along an axis are numerical values (scores) to be assigned to the modalities of a qualitative variable.

The quantification process can regard nominal and ordinal variables, with a substantial difference on the method adopted, and on the constraints applied on the method of quantification.

In this chapter, starting from the definition of variables and their classification, we focuse the attention on the methods of quantification of ordinal variables clearing the positive and negative aspects of each proposal.

1.1 The variables and their classification

The knowledge of a phenomenon is obtained across a measurement, that is the allocation of numerical values (mathematical language) to the characteristics, properties and attributes of an object, according to predetermined rules. This procedure generates the variable, expression and measurement of the different aspects of the reality.

In statistics a variable is an operativizied concept, i.e. it is an operativization property of an object, because it is necessary to associate a concept with an object. So a biunivocal correspondence does not exist between the concept and variable, because a concept could be operativizied in different ways. As a consequence we can have different kinds of variables, that can vary between different modalities, correspondent to the different property states (for example the variable "gender" or "educational qualifications").

The social and marketing researches are focused on the study of customers/citizens behavior with respect to some assertions or to evaluate their satisfaction about some services received.

The common scope is to quantify some non quantificable concepts, which can be expressed across a set of observable variables. An important distinction is, however, between the "latent variable" and the "manifest variable". The difference is in the observability of this variable, i.e. the possibility to measure them empirically. The first is a non directly measurable variable , because it represents a general or complex concept, so to operativize it we can use observable variables

Table 1.1: Types of variables

States of property	Procedure of operativization	Type of variable	Characteristics of values	Operations applicable
Not ordered	Classification	Nominal	Names	$=,\neq$
Ordered	Order	Ordinal	Ordinal properties	$\prec,\succ,=,\neq$
Continuous	Measurement-count	Numerical	Numerical properties	four mathematical operations

having a semantic relationship with it.

So a latent variable can be operativizied across some techniques of data analysis and sometimes is defined by empirical data (for example the factors of a factorial analysis). The observed variables are independent given the unobserved variable. All relationships between the observed variables can be " explained" by the latent variable, which is their *common factor*. In predictive terminology the variance of the observed variables can be " explained" by this common factor. In other words all variables measure essentially the same property.

These two kinds of variables are the base of the Structural Equation Models (SEM), in which the aim is exactly the study of the relationships between different latent variables, each expressed by a set of manifest variables.

The manifest variables, being measurable variables, could be nominal, ordinal or numerical. This classification is based on the type of mathematical-logical operations that we can do on them, where the logical operations are the operations of "equal" or "difference", while the others are the four mathematical operations.

This classification establishes the statistical analysis applicable to each kind of variable. This definition also depends on how the measurement of variables, was done.

The variables are essentially distinguished in three classes:

1. *Nominal*: a variable is nominal when it assumes discrete and non ordinable states, i.e. it can assume only a series of finite states,

called categories/modalities. The unique operation applicable to these variables is the classification;

2. *Ordinal*: a variable is ordinal when it assumes discrete and ordinal states (for example the “educational qualifications” that has ordinal values, or the questionnaire questions in which an individual must choose between ordered values);

3. *Numerical*: these variables have the numerical and ordinal property. All kinds of statistical analysis are applicable.

In the SEM models we have generally numerical or ordinal manifest variables, due to the nature of analysis (if the model is created to evaluate the customers satisfaction, we have ordinal variables), or to the typology[1] technique adopted.

The SEM model is a causal model in which the relationships between latent variables and manifest variables are estimated, and that the necessity for numerical variables, as the quantification process for this kind of analysis, is evident.

1.2 The optimal scaling of subjects and variables

The scaling models assign scores to subject, variables and both. The most well known scaling technique is based on the judgements expressed by N subjects on a set of items, concerning their attitudes versus a latent continuum. For example in marketing research the technique that asks N subjects to compare k objects is more frequent,

[1] In particular in this thesis two techniques are considered, LISREL and PLS-PM, for which the problem of the presence of ordinal variables is presented and discussed (see chapter 2).

on the basis of some criterions. In this case the scaling of variables is used . We have talked in the previous section of latent continuum: in respect to its characteristic of unidimensionality or multidimensional, we have a difference in the optimal scaling techniques.
Gordon asserts:
"The theory and the techniques of unidimensional scaling help to select a series of variables or items that, on the basis of empirical evidence, correspond and are attached to a single dimension or latent continuum".

On the base of this definition it is clear that the unidimensional scaling is a set of techniques that assigns subjects and/or variables a numerical score, taking into account that the items are a raw manifestation and codying of the process that generated the data. If we assume, instead, that the latent dimensions are major to 1, the scaling of subjects is Multidimensional, and in this case we have a ranking of subjects respect to one or more latent dimension.
The optimal scaling techniques are divided in three sections:

- methods of optimal scaling drawn through scale construction models
- methods of optimal scaling drawn through an objective function
- methods of optimal scaling obtained simultaneously with the estimation of parameters.

1.2.1 The scale construction

A scale is a *set of coherent elements* (items [2]) *that are considered indicators of a concept more general.* The element is a single component

[2] In the attitude scale the elements are assertations or questions.

(question, assertation, behaviour, attribute), while the scale is the set of these elements.

It is more used in social and psychological research, in which the objective is the study of attitudes of some individuals, but the scale is usable also to assign a score to the stimulus taking into account the responses given by individuals. The scale is applicable also to study the property of other analysis units, as for example to judge the efficiency of the institutions (governments, companies, public corporations).

The foundamental questions are: What is the nature of the variables produced by a scale? Are they nominal, ordinal or numerical?

In the literature and generally also the researcher, it is supposed that the underling dimension is continuous. This continuous property does not determine numerical variables, but quasi-numerical variables, because it is not possible to associate a numerical meaning to the scores of the scaling.

The first measurements of attitudes was made in the twenties by Allport and Hartman, Borgadus, Thurstone (CITARE). In particular, Thurstone made three different proposals (*Paired comparison, Rank order, Equal appearing intervals*). His scale provides that each item has an high number of categories, in which to the central value corresponds a disinterested behaviour, while to the first and last a complete accord or disaccord. In this scale it is supposed that the latent factor is Normal distibuted and that the categories are equal appearing, justifying the codying for them with integer numbers. After the elimination of the redundant items, each subject must choose those items more choerent with his attitude: the final score will be the mean of the scores associated to each item choosen.

Instead Likert made a proposal in the 1932 that had a great success due to its simplicity, following by the proposal of Guttman in the 1944. For many years these three proposals of Thurstone, Likert and Guttman, are being the reference of the scaling. In the recent years it is proposed a new scaling based on a probability approach to measure a continuous property, that was anticipated by Lazarsfeld (1950), applied then by Rash (1960) and Mokken (1971).

The Likert scale

The name of this scale derives from the psychometric Renis Likert that proposed it in the 1932. It is the procedure more used for the measurement of the attitudes, thanks to its simplicity.
The *scale of Likert* is composed by a set of assertation (the questions of a questionnaire) for each of them the individual expresses the degree of accord or disaccord, assigning a total score across the sum of the scores of each question (*additive scale or summarized rating scale*).
In the original proposal the number of alternatives was seven: much agree, agree, partially agree, unsure, partially opposite, opposite, much opposite. Now generally five values are used and sometimes also four. The characteristic of this scale is the partially semantic independence. The scale is built in four phases:

- formulation of questions: on the base of literature the dimensions of the attitude studied are individuated, as the assertations to measure the concept. This phase is important because from it the capacity derives to take the concept;

- administration of questions: the scale is somministrated to a group of individuals;

- item analysis, the measurement of the degree of coherence of the scale: all elements of the scale must be correlated with the same latent concept to analyze. The control is made across the computation of two quantity: the correlation *element-scale* and Cronbach's α. In the first case for each individual a correlation coefficient r is computed between the total score and the score for each item.

Cronbach's α is useful to measure the internal coherence of the scale, that is based on correlation between all elements and on their number:

$$\alpha = \frac{n\bar{r}}{1 + \bar{r}(n-1)} \tag{1.1}$$

It is not a correlation coefficient and ranges between 0 and 1: if the value is 0.7 the scale is coherent. The value of α is sensible to the number of items. The elements with a low correlation element-scale are eliminated, until the value of α increases.

- control of validity and unidimensionality of the scale: the validity is about the applicability of the scale in different researches; even if the third phase evaluates the unidimensionality of the elements, for the unidimensionality of the scale it is necessary to develops a factorial analysis, that individualizes the common factors to the elements, to see if behind to the elements there is a unique concept.

The hypothesises on which is based the scale are:

- the existence of a monotonic relationship between the latent continuum and the scores of the item category

- the quantification for individual j of the continuum η_{ij} is a linear combination of p considered item x_i with equal weights : $\eta_{ij} = \sum_{i=1,...p} x_{ij}$

The advantages of this scale are its simplicity and applicability, but the ordinal elements of the scale are treated as cardinal scale and they are not reproducible (from the score of the scale it is not possible to obtain the single answers given to the questions) and besides the total score is not a cardinal variable. Besides it gives only the scaling of subjects.

The Guttman scale

Guttman's proposal was created as a solution to the problem of unidimensionality of Likert scale. His scale is a sequence of steps, a succession of elements with an increase difficulty.
In this way, if the elements of the scale are scaled perfectly, only some sequences of answers are possible; from this characteristic derives the name of *cumulative scale*. This cumulativity gives the possibility to suppose that a continuum exists of which the elements are indicators. The scale associates value 1 or 0, respectively, for a positive or a negative response, and summarizing the scores of each individual, for each elements, we obtain the total score of each individual on the scale. In this way from the final result it is possible to go back to the answers given by the individual to each object (*reproducibility*).
We can identify, to build the scale, three phases:

- the formulation of the questions: this phase is the same of the Likert scale, with the difference that the responses must be binary and the questions must have an increasing difficulty;

- the administrations of questions: the scale of Guttman, being to binary responses, is more simple and more fast to answer;

- the analysis of results with the elimination of the elements with more errors and the computation of a global index to accept the scale: in this phase the scope is to eliminate the elements not coherent with the model, and to compute an index to accept or reject the scale. The errors in the responses are individuated comparing the observed sequence with theoretical sequence: the *Reproducibility Coefficient* measures the deviation of observed from the theoretical scale:

$$\begin{aligned} C_r &= 1-\frac{\text{n. errors}}{\text{n.total responses}} \\ &= 1-\frac{\text{n. errors}}{\text{n.elements} \times \text{n. of cases}} \\ &= \frac{\text{n. of correct responses}}{\text{n.total responses}} \end{aligned}$$

The scale is accepted if the coefficient is major of 0.9; if the value is inferior the elements with an high number of errors are eliminated and for each elimination it is updated the coefficient. The reproducibility coefficient is a mean of the *Reproducibility Coefficient* for the single elements

Edwards (CITARE)proposed to compute an index called *Minimal Marginal Reproducibility*:

$$MMR = \frac{\sum \text{Proportion of responses in the modal class}}{N} \tag{1.2}$$

where N is the number of elements in the scale.

This index is useful because it signals the minimum value assumes by the reproducibility coefficient; so if the C_r is major of 0.9 and at the same time also of the MMR index, we can claim that the scale has a good reproducibility due to an optimal scalability. After the elimination of the not scalability elements, we can claim the score to each individual, equal to the sum of positive responses, even if there are some errors in the sequence of responses. Guttman suggested to use an high number of elements, showing as a scale of four elements has an high value of C_r even if the elements are all independent.
The Guttman scale is been more important for the development of the scaling technique, even if it has some problems. In particular, it accepts only binary value, coding with 0 and 1; this scale produces a

final value that is ordinal yet and is applicable only when the attitudes are well specified, this characteristic causes the non applicability of this scale to measure the Customer Satisfaction, where it is necessary to investigate the assertion of individuals.
Besides the Guttman scale is a deterministic model, while the reality is perfectly explained only across probabilistic models.

The Rash scale

The probabilistic approach associates a value between 0 and 1 to the probability to give a response. So an individual, that has a position on the continuum of the latent variable, has for example an 80 percent of probability to answer "yes" to a question and 20 percent to answer "no". The model supposes a relationship between the position on the continuum and the probability of response to a particular element of the scale, called *trace line*.

The trace is a curve that describes the probability to answer positively to a certain element respect to the position on the continuum underling.

The curve is not linear but logistic: the probability of a positive response is near to zero, for values of ϑ (the underling dimension) very low. When ϑ increases the probability before increases slowly, then quickly until the 50 percent and more, and then it aims to 1 slowly.

The difficulty of a scale is indicated by the parameter b, correspondent to the value of ϑ on the continuum for which the probability of a positive response is of 50 percent. The traces are not visible so the aim is to build them using the responses of individuals, classifying them on the latent dimension. The curve is representable across a mathematical expression:

$$P(\vartheta) = \frac{e^{\vartheta - b}}{1 + e^{\vartheta - b}} \tag{1.3}$$

where:

- $P(\vartheta)$ is the probability of the subject to give a positive response
- b is the difficulty of the element considered
- ϑ is the position of the subject on the property
- e is equal to 2.718, the base of natural logarithm

The formula expresses the probability that an individual, with a position ϑ, gives a positive response to the elements of difficulty b. The probability of a positive response depends from the difference $(\vartheta - b)$: if the two quantities coincide the probability is 0.5; if $\vartheta \succ b$ the probability of a positive response is major than of negative, instead viceversa if $\vartheta \prec b$.

To estimate the parameters of the model we can use the Maximum Likelihood function. This is the *logistic model with one parameter*, called also *Rash model*: in this model we suppose that the difficulty of the element is due to the characteristic of the element that influences the response. Other models were been proposed in which the trace can assume different shapes in the passage from an element to another[3].

The advantages of this model are two: it is a more realistic description of the mechanism that bears the answers to the item, in respect to the deterministic scale; the variables produced by this scale are numerical, so they have all numerical properties.

1.2.2 The Multidimensional scaling

The Multidimensional Scaling (MDS) starts from a matrix Δ (N*N) of raw dissimilarity δ_{ij} that reproduces the distances of N subjects evaluated on p variables (stimulus).

[3] These studies were developed in the eighties, defining the *item response theory*.

MDS defines an Euclidean space $X(n{*}r)$ r-dimensional (with $r \prec p$) that represents the coordinates of the N points in a space of reduced dimensions, across r latent dimensions. From X it is possible to build a new matrix of distances between N points in r dimensions; the points in the space r-dimensional represent the N subjects, in such a way that, for each dimension, the distances d_{ij}, build in the new space of reduced dimension, reproduce to the best the real raw distances - dissimilatiries δ_{ij} of matrix Δ.

If Δ elements are Euclidean distances we will talk of *metric* scaling (or classical scaling, [57]), if instead the δ_{ij} are only an arbitrary and approximative of the distance between the objects, without metric properties, we will talk of *non metric* scaling ([29, 24]), much used in psycology and marketing researches.

MDS searches simultaneously the r latent continuum, in which the scores coincides with the coordinates of N subjects in the r-dimensional space in such a way it is respected the structure (order) of original distances. So the characteristics of MDS are:

- absence of a statistical causal model
- in the non metric MDS there is a monotonic function I to estimate, so that the transformed data $d_{ij} = I[\delta_{ij}]$ could be performed as Euclidean distances (metric) but with the same rank of raw dissimilarities δ_{ij} (non metric)
- the subject scaling coincides with the configuration X (in r dimensions) obtained by the matrix of Euclidean distances $D = d_{ij} = I[\delta_{ij}]$, derived from the similarity matrix Δ

In the non metric scaling the matrix Δ informs only about the rank between the couple of objects, and not on the real distances, that are instead estimated with the d_{ij}, obtained by the coordinates of the objects in the r dimensional space (X). To estimate the matrix X (the coordinates of the objects in the r-dimensional space), i.e. the scaling of subjects in the r-dimensions, a loss function is minimized Ψ ([31]), with the only constraint that the reproduced distances, expressed in function of the coordinates of the subjects, respect to the r-dimentions, give the monotonicity of dissimilarities δ_{ij}, being measured on a *non metric* scale.

$$d_{ij} = \sqrt{\sum_{s=1}^{r}(x_{is} - x_{js})} \tag{1.4}$$

The constraint on the rank permits to find the optimal position for each subject in the final space (the dimension of the space is choosen by the researcher) in such a way to minimize a loss function Ψ between the distances and inequality, respect to X:

$$\Psi = \sqrt{\frac{\sum_{ij}(d_{ij} - d^*_{ij})^2}{\sum_{ij} d^2_{ij}}} \tag{1.5}$$

If X is such that the distances d_{ij} reflect the monotonicity of d^*_{ij} [4], then $\Psi = 0$.

This technique is generalized also to the case in which m judges compare between them N subjects: in this case we have in input m dissimilarity matrixes.

1.2.3 The optimal scaling with an objective function

The scope of optimal scaling is to transform the nominal and ordinal variables to apply on them the quantitative techniques. Some of the techniques of optimal scaling are based on an objective function to optimize: this function is specific respect to the kind of variables and

[4] The quantity d^*_{ij} is built in such a way to be more similar to d_{ij}.

respect to the function to maximize.

The methods based on Correspondence Analysis, that are the majority, have not a causal structure between the variables to quantify. Starting from an indicator matrix, the optimal scaling obtains the quantification for the categories ω_j for each variable y_j, and the scores z for the subjects.

Consider a population of N subjects described by a set of p variables $y_1, ..., y_p$ with k_j categories, with $k = \sum_j k_j$; so g_{ik_j} is a scalar that assumes the value 1 (0) if the i-simo individual is or not in the category j of variable y_j. If it is done for each indivual we have the vector g_{k_j}. Considering all categories (k_j) of the variable y_j the columns vectors g_{k_j} originate, drawn between them, the indicator matrix $G_j(N * k_j)$. Extending to the p variables, we obtain the indicator matrix $G = (G_1, G_2, ..., G_p)$ the complete indicator matrix of order $(N * k)$.

The vector $\omega_j = (\omega_{j1}, \omega_{jkj}, ..., \omega_{jKj})$ parameterizes the categories of y_j, so they are the values that quantify the categories of y_j, in such a way that the vector ω_j quantifies the categorical variable $h_j(h_j^{os})$:

$$y_j = G_j \omega_j^* \Rightarrow y_j^{os} = \sum_{kj} \omega_{jh} g_{jh} \tag{1.6}$$

The matrix $Y^{os} = (y_1^{os}, ..., y_p^{os})$ is the matrix of p quantified variables (optimal scaled). To obtain unique solutions it is used to standardize the scaling parameters ω_j:

$$1'_{kj}\omega_j = 0$$
$$\omega_j D_j \omega'_j = 1$$

with $D_j = G'_j G_j$ the diagonal matrix embraces the frequence of each category of y_j and 1_{kj} a vector of k_j one.

The procedure of optimal scaling, according to the analysis choosen, computes also the vector z $(N * 1)$ for the optimal transformation of subjects, respect to the dimension of interest, with the normalization constraints:

$$1'_N z = 0$$
$$(z'z) = 1$$

So the methods that estimate ω_j and z, across the maximization of an objective function, are called the optimal scaling methods. Belong to this class the methods as Multiple Correspondence Analysis[5], Canonical Correlation, Principal Component Analysis and Anova.

1.2.4 Optimal scaling with a statistical model

According to Bradley et al. and Kruskal [6, 31] the optimal scaling can not be separated from the model to be estimated. The first analyzes the problem of the scaling of p categorical response variables $y_1, ...y_p$ in the context of Anova, in such a way that the coefficients of experimental factors are significatively different from 0, supposing a-priori that the categorical data depend by the k experimental factors $g_j = (j = 1, ..., k)$.

In particular it is supposed:

- the existence of a monotone transformation I, that transforms each categorical variable in quantitative $s_i = I(x_i)$
- the existence of a linear model in the context of Anova with the transformed response variables
- the function I is choosen in such a way that maximizes the F statistic (the significativity of the k treatments of the linear model on the p response variables

[5] The Optima Scaling hystorically is always associated to the MCA, that has as aim the scoring of subjects and the categories of each item.

The work of Kruskal (CITARE)also searches, in the context of Anova, a monotonic transformation of the categorical data according to the mimimun squares criterion: the transformations I are obtained in such a way that the relationship between g_j and s_i is linear, minimizing the sum of residuals squares of Anova model.

The hypothesis of Kruskal are:

- the existence of a monotone transformation I, that transforms each categorical variable in quantitative $s_i = I(x_i)$
- the existence of a linear model in the context of Anova with the transformed response variables
- the choice of I that maximize the *monotone stress* $S(I, \beta) = \sum_i [s_i - s_i(\beta)]^2$, that is the residual deviance of the linear model:

$$s_i(\beta) = \sum_j g_{ij}\beta_j^* + e_i \tag{1.7}$$

The weights β_j^* are estimated with an initial estimation of I in a previous step. Kruskal was the first to consider the problem to estimate two separate sets of parameters, scaling and structurals, and he is the first to propose a solution in his algorithm. The loss function proposed by Kruskal starts to minimize $S(I^*, \beta)$ respect to the parameters embrace in β, fixed I, and successively it estimates I, minimizing $S(I, \beta^*)$, fixed β. This work has prompted the ALSOS metodology [10, 67] that is based on the alternation of these two steps.

The ALSOS procedure consists in obtaining the optimal quantification of the qualitative variables, across an Optimal Scaling method, and simultaneously to estimate the structural parameters of the statistical model, specified a-priori by the researcher. This metodology, in particular, is useful in all analysis in which the data matrix is composed by mixed variables (numerical, ordinal and nominal): for each typology of variable, this methodology is able to quantify the qualitative variables separately and each according its nature. In this way the parameters of the model specified are estimated, fixed the quantification, taking into account of all observed variables, and across a unique objective function, expressed respect the two sets of parameters [6].

1.3 Comparisons on quantification approaches

In this chapter three different approaches to the quantification of qualitative/ordinal variables have been presented and discussed. The first, the scale construction, is characterized and based on the definition and construction of a scale of values.
Their simplicity and applicability have rendered them the struments more used to measure and quantify the qualitative variables. Despite their simplicity, they have some disavantages that cause the choice, for this work, to use other methods to quantify.
In particular, the Likert and Guttman scales produce respectively ordinal and binary variables: we need, instead, of numerical variables; the Likert scale is an additive scale, with the problem of the correctness to consider an equal distance between the categories of the variable (is equal the distance between 1 and 2 to that 2 and 3, and so on?).
The Guttman scale, proposed as an alternative to the Likert

[6] In the chapter 3 it is possible to have a more detailed description of the ALSOS methodology.

approach, considered deceptive and simplifying, is a method to control if a set of items and the subjects conform to the ideal scale supposed, sooner than a method of quantification.
The approach of Rash produces numerical variables, using a logistic model; this approach does not allow to have a unique function to optimize, in presence of qualitative and quantitative manifest variables. The Rash model does not permit, besides, to estimate simultaneously the quantification and the parameters of the model: the parameters could be estimated only after the process of quantification and across another algorithm. This disavantage is a problem common to the scale construction techniques.
The second approach presented is based on the definition of an objective function coherent with the analysis to develop. This approach has the advantage to obtain the quantification across the maximization of a criterion, but the fucntion optimized does not express a casual model, and so this method does not allow to estimate in a unique function the parameters of a model and the optimal quantification.
The third approach, instead, is the ALSOS algorithms, in which, according to the analysis to develop, the optimal scaling is a step useful to maximize the relationship between the variables optimally quantified, and so for example in the case of a linear regression model, the variables are quantified to maximize the correlation between the variables.
ALSOS takes ispiration from the Kruskal proposal, that transforms the categorical variables supposing a-priori that the relationships, between the quantified variables (dependent and independent), are linear. So this approach has these advantages and properties:

- the optimal scaling in the ALSOS algorithms is used only to quantify the qualitative/ordinal variables, contextualizing the process in the general analysis to develop

- the estimation of the parameters and the quantification are two different steps, alternated, that take ispiration from the non metric ANOVA of Kruskal (1964) and the Factorial Analisys (Kruskal-Shepard, 1974)
- the starting point is the algorithm HOMALS that develops a Multiple Correspondence Analysis, from which other methods are derived, added some constraints on the parameters

Summarizing, the choice of this approach, sooner than the other presented, is due to: i) the possibility to have different kinds of variables in the model; ii) each variable is quantified according to its nature and the analysis to develop; iii) the quantification and the estimation of parameters are two steps of a unique algorithm; iv) we optimize a unique function obtaining as results the optimal scaled variables and the optimal parameters of the model.

1.4 Some remarks

In this chapter we have presented at first the concept of variables and its different typologies (numerical,ordinal and nominal), focusing the attention on the ordinal variables.
Succesively it is faced the problem of quantification for the ordinal variables,that are separable in three categories: the scale construction (the most simple method to use to quantify an ordinal variable), the Multidimensional Scaling and the Optimal

scaling, the last characterized by an objective function to maximize.
A discussion is made for these methodologies, justifying the choice of a method of Optimal scaling to quantify the variables, in particular the ALSOS approach, in our work.
So this chapter is an introduction to the problem that will be developed in the next chapters and for which a new methodological approach is proposed.

Chapter 2

Estimation of a Structural Equation Model with ordinal variables

The study of complex phenomenon is possible across the Structural Equations Model, because they are based on a system of linear equations, each of that represents a casual relationship between two or more latent variables.

These kinds of model are based on the definition of a set of latent variables (concepts non directly measurable) and of a set of observed manifest variables, that are expression of the latent concepts. This structure allows to obtain an important result: the estimation of a variable, through the study of others variables, relationates with it.

These models are much applied in the marketing and social research, in which typically these kinds of variables are analyzed (for example the behaviour of a set of individuals respect some items, or the study of customer satisfaction in respect to a ser-

vice received).
As we can see, from these two examples it is clear that the classical manifest variables used for this analysis, are ordinal, i.e. variables expressed on a scale of values in a definite interval on integer numbers.
This chapter is centered on the presentation of two approaches to the estimation of a SEM model, but in the end, the problem of the treatment of ordinal variables is discussed, showing the proposals in the literature for the two techniques.

2.1 SEM definition and estimation

The *Structural Equation Modeling* (SEM) is a *stochastic model in which each equation represents a casual relationship, and not a simple association* (Goldberger, 1972). With the word SEM a family of statistical techniques for testing and estimating the casual relationship between the latent variables is identified, and in which two concepts are synthesized: the first is the existence of a *model*, that is the formalization of a theory and the second the *structure* of a model across a system of equations that represent the casual relationship.

An advantage of the SEM is the possibility to express complex relationship between the variables, that are characteristics of the social and marketing research, in which the phenomenons analyzed are complex and can not be synthesized and explained by a simple/multiple Regression model.

The structural approach assumes that constructions (latent variables), non directly measurable, can be measure across a system of equation, so that they can be expressed by observable variables (manifest variables), with big measure errors. This technique,then, incorporates and integrates the *Path Analysis* and the *Factor Analysis*:

- the *Path Analysis*, introduced by Sewall Wright in 1921, is an extension of the regression model, used to test the fit of the correlation matrix against two or more casual models which are compared by the researcher. The model is represented by a *path diagram*, and a regression is done for each variable as dependent on others which the *path diagram* indicates as causes. The regression weights predicted by the model are compared with the observed correlation matrix for the variables, and a goodness of fit statistic is calculated, to individuate the best model. The *Path Analysis* has the same assumption as a regression model and is sensitive to model specification, because failure to include relevant causal variables or inclusion of extraneous variables affects the estimation of the path coefficients;
- the *Factor Analysis* introduced by Charles Spearman (1900), is used to uncover the latent structure (dimensions) of a set of variables, reducing the attribute space from a larger number of variables to a smaller number of factors that explain the variance of original variables. The *Factor Analysis* generates a table in which the rows are the observed raw indicator variables and the columns are the factors or latent variables which explain as much of the variance of these variables as possible.

The contribution of the *Path* and *Factor Analysis* is on the one hand, the path diagram as a way to represent a model, and on the other hand the introduction of the concept of latent variable, as a variable not directly measurable. Both aspects are included in the SEM technique and they are the principal characteristics.

So the structural equation modeling process is based on two steps: validating the measurement model and fitting the structural model. The former is accomplished primarily through confirmatory factor analysis, while the latter is accomplished primarily through path analysis with latent variables. The technique starts with the specification of the model on the basis of theory (it is a confirmatory approach: it is suited to theory rather than theory development), in which each variable is conceptualized as latent variable measured by a set of multiple indicators (manifest variables).

For the estimation of a SEM it is possible to follow two different approaches: a confirmatory approach, that characterizes LISREL (Linear Structural Relationship; Joreskog, 197), that estimates the matrix of covariance, according the method of Maximum Likelihood, and an exploratory approach, that characterizes the algorithm of Partial Least Square Path Modeling (H. Wold, 1982), that estimates the best prediction for the latent variables, maximizing the variance between the variables. The first technique is a covariance based model, while the second is a variance based model. Both methods compute simultaneously the path coefficients of the model, but respectively with and without distributional hypothesis on the data, from which the name *hard modeling*, for the first, and *soft modeling* for the second are derived.

In the next section the two approaches will be presented and in particular, the problem of the treatment of ordinal variables, for which an alternative methodology will be proposed and discussed.

2.1.1 The process of estimation

A SEM is composed by two parts, the measurement and specification model, where the first specifies and computes weights expressed by the relationship between the manifest and latent variables, and the second specifies and computes the coefficients of the relationship between the latent variables. The model is built a-priori, so it is necessary to establish the parameters to be estimated, finding those values for which the difference is minimum between the covariance matrix of the model and the covariance matrix of the original data. The phases are:

- model specification
- identificability of parameters
- estimation of parameters
- model validation

The starting point is the raw data collection on which a matrix of variance-covariance [1] S is computed, following the estimation of parameters of the model. The estimation of the structural parameters is obtained by means of an iterative procedure: the algorithm starts with arbitrary values for the parameters (positive and negative) on which the variance-covariance matrix Σ is based, with the aim to minimize the distance between the matrix S and the matrix Σ. The algorithm converges when the change of the values of the parameters does not reduce the distance between the two matrixes, in respect to the previous iteration.

The validation of the model is based on the computation of the residuals, obtained by the difference between the two matrixes: if this difference is major than that imputable to the

[1] In the case of standardized data it is a correlation matrix.

error, the model is rejected. In this case the phase in which the model is modified[2]starts .

The typical path diagram of a SEM (see figure2.1) is composed by three elements: the manifest variables, the latent variables and the path relationship (covariation[3], if the arrows are bent bidirectional, and a casual relationship if the arrows are unidirectional).

The manifest variables are distinguished between endogenous Y, associated to the endogenous latent variable, and exogenous X, associated to the exogenous latent variable, and they are represented by rectangles, connected with the unidirectional arrows[4] to the latent variables, represented by the ellipses. The latent variables are called ξ and η , respectively the exogenous and the endogenous latent variables : the variables ξ are always independent in the model, so the arrows start from them and go to the η, that in the model could be a dependent and independent variable. The residuals of η variables are ζ, the residuals of variables X and Y are respectively δ and ϵ .

The following matrixes of variance-covariance are computed: i) the matrix Φ (symmetric and square matrix) between the variables ξ; ii) the matrix Ψ (symmetric and square matrix) between the residuals ζ; iii) the matrix of residuals ϵ and δ called respectively θ_ϵ and θ_δ.

The correspondence between the path diagram and the analytic model so for each variable on which there is an unidirectional arrow there will be a structural equation, and in this

[2] The changes are based on theoretical hypothesis, and sometimes they are based on the introduction of new relationship, before not considered in the model.

[3] It is the simultaneous variation of two variables in absence of a casual relationship.

[4] The arrows between manifest and latent variables are only reflective.

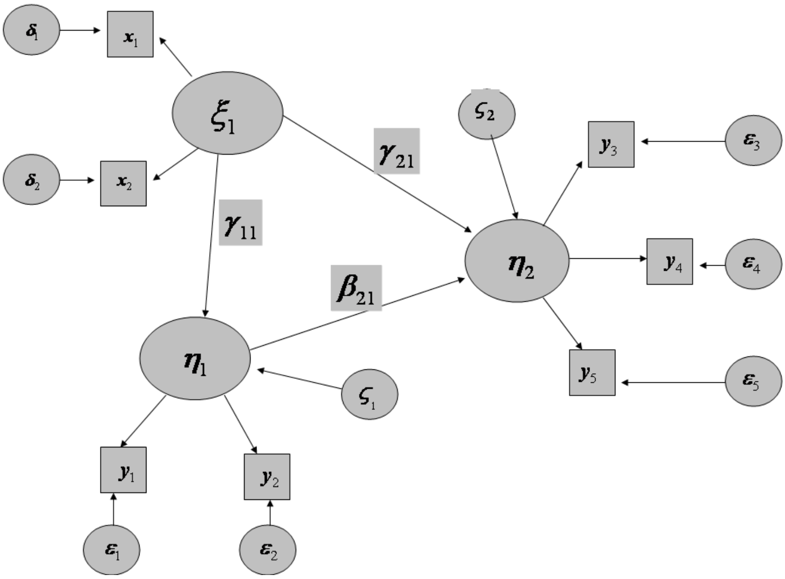


Figure 2.1: Path diagram

equation the dependent variable will be express as the sum of independent variables of equation each of them multiplied for the path coefficient.

The measurement model for the manifest variables Y is:

$$Y = \Lambda_y \eta + \epsilon \tag{2.1}$$

where Y is a vector $(Q \times 1)$, Λ_y is a matrix $(Q \times J)$, η is a vector $(J \times 1)$ and ϵ is a vector $(Q \times 1)$. while for the manifest variables X is:

$$X = \Lambda_x \xi + \delta \tag{2.2}$$

where Λ_x is a rectangular matrix $(P \times M)$ and embraces the structural coefficients between the endogenous and exogenous manifest variables, and the endogenous and exogenous latent variables; ξ is a vector $(M \times 1)$ and δ is a vector $(P \times 1)$.

The structural model for the endogenous latent variables is:

$$\eta = \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \zeta \tag{2.3}$$

where η is the vector of endogenous latent variables, Γ is a rectangular matrix $(J \times M)$ that embraces the path coefficients between the endogenous and exogenous latent variables, ξ is the vector of exogenous latent variables, B is a square matrix $(J \times J)$ that embraces the path coefficients between the endogenous latent variables, with the diagonal elements always zero, ζ is the vector of errors associated to the endogenous latent variables.

2.1.2 Linear Structural Relationship

LISREL is the acronym of Linear Structural Relationship, a software created by Jöreskog [27] in the seventies to estimate the structural coefficients with the method of Maximum Likelihood.

The assumptions made in LISREL are:

- the manifest and latent variables and the errors (both measurement and structural model) are centered;
- the covariance between two errors (both measurement and structural model) is null;
- the covariance between the measurement errors (endogenous and exogenous) and latent variables (exogenous and endogenous) is null;
- the covariance between the structural errors and the exogenous latent variables is null;
- the structural model must not be redundant.

At this point LISREL computes the structural parameters and the distance between the covariance matrix S, computes on raw data, and the covariance matrix Σ, computes by the model. The matrix of variance-covariance Σ of the population is:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{2.4}$$

where Σ_{xx} is the variance-covariance matrix $(P \times P)$ of manifest variables X, Σ_{yx} is the intercovariance matrix $(Q \times P)$ between the endogenous and exogenous manifest variables and Σ_{yy} is the variance-covariance matrix $(Q \times Q)$ of manifest variables Y. It is possible to express the matrix Σ in function of the parameters of the model, rewriting the three matrixes, in such a way that it is possible, across a method of estimation, to make the comparison between the matrix S and Σ. So the matrix of the exogenous manifest variables, in terms of the parameters of the model, is:

$$\begin{aligned}\Sigma_{XX}(\Omega) &= E(xx') = E[(\Lambda_x\xi+\delta)(\Lambda_x\xi+\delta)'] \\ &= \Lambda_x E(\xi\xi')\Lambda_x' + \Lambda_x E(\delta\xi')\Lambda_x' + E(\delta\delta')\end{aligned}$$

assuming that $E(\xi\xi') = \Phi$ and $E(\delta\delta') = \Theta_\delta$ we obtain that:

$$\Sigma_{XX}(\Omega) = \Lambda_x\Phi\Lambda_x' + \Theta_\delta \tag{2.5}$$

The matrix of the endogenous manifest variables is:

$$\begin{aligned}\Sigma_{yy}(\Omega) &= E(yy') = E[(\Lambda_y\eta+\epsilon)(\Lambda_y\eta+\epsilon)'] \qquad (2.6) \\ &= \Lambda_y E(\eta\eta')\Lambda_y' + \Lambda_y E(\epsilon\eta')\Lambda_y' + E(\epsilon\epsilon')\end{aligned}$$

assuming that $E(\epsilon\epsilon') = \Theta_\epsilon$ we obtain

$$\Sigma_{yy}(\Omega) = \Lambda_y E(\eta\eta')\Lambda_y' + \Theta_\epsilon \tag{2.7}$$

The structural model expresses the endogenous variables as:

$$\eta = (I-B)^{-1}(\Gamma\xi+\zeta) \tag{2.8}$$

so assuming $E(\zeta\zeta') = \Psi$ and with some algebraic developments, the matrix Σ_{yy} becomes:

$$\Sigma_{yy}(\Omega) = \Lambda_y[(I-B)^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-B)^{-1}\prime]\Lambda_y' + \Theta_\epsilon \tag{2.9}$$

The matrix of intercovariance, that expresses the relationship between the endogenous and exogenous manifest variables, is rewritable in:

$$\begin{aligned}\Sigma_{XY}(\Omega) &= E(XY') = E[(\Lambda_x\xi+\delta)(\Lambda_y\eta+\epsilon)'] = \\ &= \Lambda_x E(\xi\eta')\Lambda_y' + \Lambda_x E(\xi\epsilon') + E(\delta\eta') + E(\delta\epsilon')\end{aligned}$$

The errors are uncorrelated between them and with the exogenous latent variables, so the equation becomes:

$$\Sigma_{XY}(\Omega) = \Lambda_x E(\xi\eta')\Lambda_y' \tag{2.10}$$

Substituting the equation (2.8) above the matrix Σ_{XY} is equal to:

$$\Sigma_{XY}(\Omega) = \Lambda_x \Phi \Gamma' (I-B)^{-1\prime}\Lambda_y \tag{2.11}$$

So the final matrix is:

$$\Sigma_{YX} = \Sigma_{XY}' = \Lambda_y (I-B)^{-1}\Gamma\Phi'\Lambda_x' \tag{2.12}$$

At this point we can substitute the equation (2.4) with

$$\Sigma(\Omega) = \begin{bmatrix} \Lambda_x\Phi\Lambda_x' + \Theta_\delta & \\ \Lambda_y(I-B)^{-1}\Gamma\Phi'\Lambda_x' & \Lambda_y\left[(I-B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-B)^{-1}{\prime}\right]\Lambda_y' + \Theta_\epsilon \end{bmatrix}$$

The method used by LISREL is the *Maximum Likelihood*[5] (ML),that individualizes, given an observed (in a sample) covariance matrix S, what is the probability that this matrix derives from a theoretic matrix Σ (in the population), computing the values, to associate at the free parameters of the model, to obtain the maximum probability that S derives from Σ.

Assuming that the data follow a multivariate normal distribution the function to minimize is:

$$F_{ML} = log|C| + tr(SC^{-1}) - log|S| - (P+Q) \tag{2.13}$$

where S is the covariance matrix of observed data, C is the var-cov matrix obtained from the model and $(P{+}Q)$ is the number of manifest variables X and Y. The estimators obtained by the ML method are asymptotically correct, consistent and asymptotically efficient. Besides for $N \to \infty$ the distribution of the data leans to a Normal

[5] The ML supposes that the variables has a multivariate normal distribution.

distribution.

Alternative discrepancy functions are:

- The *Unweighted Least Square (ULS)*:

  $$F_{ULS} = \frac{1}{2} tr \left[(S - C)^2 \right]$$

  In the case in which N is big, the ULS estimators are similar to ML estimators, but they are not asymptotically efficient. For $N \rightarrow \infty$ the estimators ULS are consistent, without the necessity to hypothesize a distribution for the manifest variables;

- The *Generalized Minimum Square*:

  $$F_{GLS} = \frac{1}{2} tr(W^{-1}(S - C)^{-1})$$

  The GLS estimators are consistent and for $N \rightarrow \infty$ the distribution leans to a Normal. However this property depends on the choice of W (for $W = I \rightarrow GLS = ULS$). So the matrix W is choosen with these constraints:

  1. the elements of S have to be consistent estimators for the var-cov matrix;
  2. the elements of S have to be asimptotically distributed as a multinormal with mean equal to the corrsipondent variance-covariance matrix and an asintotically covariance between s_{ij} and s_{igh}, equal to $N^{-1}(\sigma_{ig}\sigma_{ih} + \sigma_{ih}\sigma_{jg})$. Generally we have this identity $W = S$.

The validation of the model

The measures of the global fit of the model to the data are functions of the residual, that is the diference between S and C. The problem is the definition of a known distribution to make inferential tests on

the model. So the validation is based on the analysis of residuals of the model, to establish how much of the deviation is due to the sample errors, and how much is due to the difference between the two matrixes. So the steps of the validation are:
→ The global fit of the model to the data;
→ Statistical test on the relationship between the variables (manifest and latent)
It is possible to show that the fitting statistics $f(S, C)$ is distributed as a χ^2 with the degree of freedom equal to:

$$df = \frac{1}{2}(P+Q)(P+Q+1) - t \qquad (2.14)$$

The test based on the calcolous of the statistic T of χ^2[6] is used for the global validation of the model; in particular the test allows to compare the matrix of variance-covariance S with the same matrix C:whereas if the value of χ^2 Statistic is minor to the tabulate one, the null hypothesis is accepted[7] and so is the model. A problem of statistics, that are based on χ^2, is their sensibility to the size of the sample (N): the value of the statistic increases proportionally to the size of the sample. If the size is small, it is possible to accept the model even if it has not a good fit, while if the size is big, it is possible to reject the model even if it has a good fit to data.
Another consequence is the difficulty to compare two statistics T computed on samples of different size. To exceed these limits different alternative measures of the fit of model are proposed; in particular:

- The *Goodness of Fit Index*(GFI):

[6] The χ^2 test is used also to compare two nested models: one model contains a part of the parameters of the other model. Given a model with a statistic T, if we fix some parameters (that is their values are zero) the new model has an higher value of T and of df than the other model.

[7] The null hypothesis is $H_0 : S - \Sigma = 0$

$$GFI = 1 - \frac{T_i}{max(T_i)}$$

where the value of the statistic T is standardized with its max value.
This index ranges between 0 and 1, but sometimes it is possible to observe values outside of this interval. This index is usable with the discrepancy function ML, ULS and GLS; a model is accepted if the value of GFI $\succ$ 0.9. Besides this index allows comparing models on two different samples of different size, but it does not take into account the degrees freedom.

A modified version of the GFI index is

- The *Adjusted Goodness of Fit index* (AGFI):

$$AGFI = 1 - (\frac{k}{df})(1 - GFI)$$

where *df* are the degree of freedom and k is the number of variance-covariance in input, equal to $1/2(p+q)(p+q+1)$.
It ranges between 0 and 1 and is usable only with the discrepancy function ML, ULS and GLS. The model is accepted if AGFI $\succ$ 0.9. It allows comparing two models built on two different samples of different size, taking into account the degrees of freedom, but its distribution is unknown.

The problem of these two indexes is that the distribution is unknown, so we can not make any test for the significativity of the model. The last index computes by LISREL is

- The *Root Mean Square Residuals* (RMR):

$$RMR = \sqrt{\frac{1}{k}\sum(s_{ij} - \sigma_{ij})^2}$$

where k is $1/2(p+q)(p+q+1)$. This is the mean of the square of residuals, that becames 0 when S coincides with C, but respect to the other two indexes, RMR does not have an upper boundary. It has the same problem of the statistic T, so it is useful only to compare different models computed on the same data. This index, however, is not sensible to the size of N, and so in the case of a big sample it is more appropriate to evaluate the fit of model. The statistical distribution is also unknown for this measure .

Other indexes for the measurement of the fit of the model are based on the comparison between two models, the model hypotized in the case of independence, that has as parameters only the variance of manifest variables [8] and the estimated model; the indexes are:

1. *Comparative Fit Index* of Bentler (CFI):

$$CFI = \frac{[(n-1)F_{ind} - df_{ind}] - [(n-1)F - df]}{(n-1)F_{ind} - df_{ind}}$$

where F_{IND} is the minimum of the function for the indipendence model and df_{ind} are the degrees of freedom for the independence model; the model is accepted if CFI $\succ$ 0.9;

2. *Non-Normed Fit Index* of Bentler-Bonnet (NNFI):

$$NNFI = \frac{\frac{F_{ind}}{df_{ind}} - \frac{F}{df}}{\frac{F_{ind}}{df_{ind}} - \frac{1}{n-1}}$$

the model is accepted if NNFI$\succ$ 0.9

There are three possible indexes to choose between two models on the base of information theory; they are:

[8] In this case it is supposed that there are not the relationship between the variables, and it has the maximum number of degrees of freedom.

i $AIC = \chi^2 + 2(n^\circ of parameters)$

ii $ECVI = \frac{\chi^2}{n} + 2(n^\circ parameters/n)$

iii $CAIC = \chi^2 + [1 + ln(n)]\,(n^\circ parameters)$

For each model one of these indexes is computed and the model with the index smaller is accepted.

The improvement of the model

This phase of the LISREL is called the *improvement of the model* because, across some indexes, it is possible to improve the fit of the model to the data.

The model is improvable by: i)exclusion of parameters from the model; ii) introduction of new parameters; iii) riformulation of the model.

The first step consist of ceck the significativity of the parameters, to eliminate those that are not significantly different from zero. In the case the Normal assumption the null hypothesis is rejected if the estimated value of the statistic [9] is major than 1.96.

The following procedure is used to decide whether to eliminate a parameter or not: a parameter is excluded from the model if its statistic, computed as before, is major than 2 ($\approx$ 1.96). This control is made one by one for each parameter estimated by the model, and for each elimination the model is restimated. If a parameter, before being eliminated, becomes significant as a consequence of changes in the model, it is lightlighed by the modification index.

Modification indexes are used to include new sgnificant parameters in the model: for each parameter not included in the model, fixed to zero, it is computed how much the value of the statistic T of χ^2

[9] The statistic of the parameter is $|p|/SE > 1.96$, where SE is the Standard Error.

decreases respect to the case in which this parameter is estimated. The modification index is computed as the ratio between the value of χ^2, of the model in which this parameter is fixed, and the value of χ^2 for the model in which it is free. This index has a χ^2 distribution with one degree of freedom. A parameter is included in the model if the value of the Statistic is $\succ 4$; the inclusion of a new parameter has to have theoretical justification.
The inclusion of new parameters consists on the individuation of the parameters with a higher value of the modification index, and so the new model is estimated with new parameters. The parameters are introduced in the model one at time because the introduction of parameters causes changes in the values of the modification index of the other parameters.

The identification of the model

The starting point of LISREL is a covariance matrix between the observed variables, with the aim to estimation the model parameters. It can sometimes happen that a model is multi-faceted with different sets of parameters, i.e. the same model has different solutions, and this is not admissible: a model has to build in such a way that it has only one solution. This is the problem of identification, so *a model is identified if its parameters are univocally estimated.*
Summarizing we have that:

- a model is perfectly identified when for each parameters , for which the value is not established a-priori, a unique optimal value exists → zero degree of freedom. The problem is that this model has a trivial fit, so the test for the significativity of the model is not interesting;
- a model is over-estimated if there are more equations than parameters, so the degree of freedom are major than zero;

The rules to follow in the specification of a model are:

1. the number of variables excluded from the model in each equation has to be major or equal to the number of equations minus one (necessary condition);
2. some constraints on the rank of the coefficients matrix (necessary and sufficient);
3. a structural model is identified if at least three manifest variables are associated to each latent variables ;
4. a structural model is identified if at least two manifest variables are associated to each latent variables, and each costruction is correlated at least with another construction.

The necessary but not sufficient condition is that the degree of freedom are major of zero:

$$df = \frac{1}{2}(P+Q)(P+Q+1) - t \geq 0 \tag{2.15}$$

where P and Q are respectively the number of manifest variables X and Y, while t is the number of parameters of the model.

2.1.3 The Partial Least Squares - Path Modeling

The *Partial Least Squares - Path Modeling* (PLS-PM) approach [62] to the SEM models is an iterative algorithm that allows to compute the estimation of latent variables and the relationship between them, by means of an interdependent system of alternate elaboration based on multiple and simple regression. The idea is to determine the scores of latent variables through a process that iteratively computes first an outer estimation for them and then an inner estimation.

The way in which this algorithm operates is called *soft modelling*, in contrast to *hard modeling* which identify with the techniques such as LISREL. The name *soft modeling* is due to a set of properties of this algorithm, due also to its larger applicability; the characteristics are: i) it is prediction oriented (no model fitting purposes), i.e. the aim is to obtain the best prediction of the latent variables; ii) it is not theory oriented; iii) the parameters for each block are estimated separately, as in the Path Analysis and by simple/multiple regression; iv) it handles reflective and formative indicators; v) in respect to the LISREL, the PLS-PM has a better estimation of the measurement model, because optimizes the prediction of the latent variables and the relationship between the manifest and latent variables; vi) there is no problem for the identification of the model, because the algorithm estimates the weights separately for each block; vii) the estimates became consistent when the sample size gets larger; viii) it can estimate the model also in presence of multi-collinearity and missing data; ix) it is possible to estimate the model even when the number of observation is smaller than the number of the manifest variables [10].

The absence of distributional hypothesis, different from LISREL, do not allow inferential tests on the parameters and for the model validation. The inference approach, in the PLS-PM, is based on on resampling techniques like *Bootstrap* and *Jacknife*. It permits to obtain empirical distributions of the parameters, and the computation of "empirical indexes" for the global validation of the model.

As in the case of LISREL the PLS-PM also distinguishes between outer and inner estimation: in the outer estimation or *measurement model* the algorithm computes the coefficients of the relationship between the manifest and latent variables, while in the inner or *structural model* it estimates the path coefficients, that express the relationship between the latent variables. The algorithm performs the

[10] It is necessary that the number of latent variables is major than the observations.

estimation of the first step separately for each block [11] and then it updates the estimation of the latent variables, by the inner estimation, usually based on OLS.

In the next subsections the outer-inner estimation of PLS-PM and the validation indexes for the measurement of the fit of the model are described.

The outer and inner estimation

Consider the matrix X $(N \times I)$ of the manifest variables; in the outer estimation phase the algorithm computes the weights w_{ij}, where j represents the j-mo latent block, associated to each manifest variable for the estimation of the latent variable. In the first step the weights are randomly choosen (for example all 1) and the first estimation of a latent variables is equal to the linear combination of its manifest variables, multiplied for the correspondent weight:

$$v_j = \sum x_{ij} w_{ij}$$

The relationship between the manifest and latent variables could be of three different typologies:

- *reflective*: the latent variable is reflected in its manifest variables. In the path diagram the arrows start from the latent variable to the manifest variable;
- *formative*: the latent concept is formed by its manifest variables. In the path diagram the arrows start from the manifest variables to the latent variable;

[11] For this reason it is possible to have a matrix with the number of manifest variables bigger than the observations number.

- *MIMIC*: the latent variable has a formative relationship with some variables, and reflective with the other.

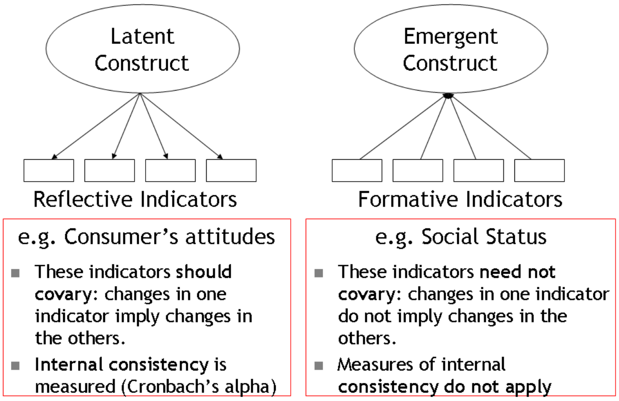

Latent Construct
Emergent Construct
Reflective Indicators
Formative Indicators
e.g. Consumer's attitudes
▪ These indicators should covary: changes in one indicator imply changes in the others.
▪ Internal consistency is measured (Cronbach's alpha)
e.g. Social Status
▪ These indicators need not covary: changes in one indicator do not imply changes in the others.
▪ Measures of internal consistency do not apply

Figure 2.2: The relationship between variables

After the initial step, the algorithm updates the estimation of the latent variables according to the inner estimation, based on the weights $e_{jj'}$ (j' is a generic latent variable associated to the j-ma latent variable). The algorithm allows to choose, as estimation of these weights, three different alternatives:

1. centroid scheme, that computes the weight as :

$$e_{jj'} = sign\left[cor(v_j, v_{j'})\right]$$

2. factorial scheme, that computes the weights as:

$$e_{jj'} = cor(v_j, v_{j'})$$

3. path scheme, that computes the weights as:

$e_{jj'} = cor(v_j, v_{j'})$ if $v_{j'}$ predicts v_j

$e_{jj'} =$ the regression coefficient if $v_{j'}$ is predicted by v_j

The typology of the relationship is important, for the algorithm, when it updates the estimation of the weights w_{ij},in particular if the relationship is reflective the weights are equal to:

$w_{ij} = cov(z_j, x_{ij})$

$w_{ij} = (Z_j' Z_j)^{-1} Z_j' x_{ij}$

where Z is the matrix of the latent variables obtained after the inner estimation. If the variables are standardized, the weights are equal to the correlation between the variables.
If the relationship is formative, the weights are equal to:

$$w_{ij} = (X'X)^{-1} X' z_j$$

In this case the weight is the coefficient of the multiple regression between the latent and manifest variables.
After the new outer estimation, the algorithm proceeds with the control of the convergence [12]: if the weights of the two outer successive estimations are equal the algorithm stops, and then computes the OLS multiple/single regression for the inner estimation of the path coefficients between the latent variables, according to the supposed

[12] The convergence of PLS-PM algorithmis demonstrated for two blocks. In case of more blocks the convergence is demonstrated only empirically.

relationship between them.
So, in synthesis, the steps of the algorithm are:

Algorithm 1. *PLS-PM.*

```
Initialize the algorithm with the matrix $X$ of raw manifest variables
Step1: Compute a first casual vector of weights $w_{ij}$
repeat
Step2: Compute the first estimation of Lvs
    for (j in 1:k)
        $v_j = \sum_{i=1}^{p} w_{ij} x_{ij}$
    endfor
Step3: Update the previous estimation of Lvs
    for (j in 1:k)
        $z_j = \sum_{j=1}^{k} e_{jj'} v_j$
    endfor
Step4: Update the estimation of weights $w_{ij}$
    for (i in 1:p)
        for (j in 1:k )
            $w_{ij} = cov(x_{ij}, z_j)$
            $w_{ij} = (X_j' X_j)^{-1} X_j' z_j$
        endfor
    endfor
Ceck the convergence
    $|w_{ij}^{old} - w_{ij}^{new}| \prec 10^{-5}$ break
```

The validation PLS-PM process

The validation process of the PLS-PM aims at establishing suitable inference about the coefficient of the model and calculating some suitable indexes to measure its predictivity performances and fitting. Due absence of distribution hypothesis on data the PLS-PM inferential tools are usually based on resampling techniques. In the following

we consider the Bootstrap technique that consists in the extraction with replacement of m sample of size n (the size is equal to the original sample), on which the model are computed m times. In this way it is possible to establish an empirical distribution for the parameters whose percentile values allow to obtain suitable confidence interval. This procedure is made for both the parameters of the outer model (weights and loadings), and both the inner model (path coefficients). Intervals including the zero suggest to eliminate manifest or latent variables in the model. To compare the model estimated parameters and the mean of the bootstrap replications, a ratio between their deviation and the standard deviation of the resampling distribution is computed as a classical test statistics.
Thus:

- the estimation of the latent variables are not consistent, because a latent variable is a linear combination of its manifest variables;

- for $n \to \infty$ the manifest variables estimations are consistent (*consistent at large*).

- as in the PCA and CCA, the coefficients are over or under estimated; when all correlation coefficients ρ among the i manifest variables are equal to a certain value s, the bias is :

  $$bias(\lambda) = \sqrt{\frac{s + (1-s)/i}{s}} = \frac{1}{\sqrt{bias(\rho_{ij})}}$$

  the bias of loadings is equal at the reciprocal of bias between the latent variables: major is the bias of the path coefficient, minor is the bias of the loadings (it is a compromise between the measurement and structural model);

The Goodness of Fit of the model is evaluated by some indexes, as the validity, communality, structural and redundancy. These indexes are connected between them in this way:

$$\underbrace{x \rightarrow \xi}_{validity}\underbrace{\underbrace{\rightarrow \eta}_{structural}\quad\underbrace{\rightarrow y}_{communality}}_{redundancy}$$

The *communality* is the capacity of a latent variable to explain the variance of its manifest variables; it is computed as:

$$AVE_j = \frac{\sum_i \lambda_{ij}^2 \times var(\xi_j)}{\sum_i \lambda_{ij}^2 \times var(\xi_j + \sum_i var(\xi_j)}$$

where λ_{ij} is the loading of the variable i associated to the j-ma latent variable. If the AVE is bigger than 0.5, or if the loadings are all major than 0.707, or if the null hypothesis of the test $H_0 : \lambda_{ij} = 0$ is rejected, the latent variable is considered a good predictor for the manifest variables.
If the variables are standardized the $Communality_j$ is equal to :

$$Communality_j = \frac{\sum_i \lambda_{ij}^2}{p_j} \tag{2.16}$$

The R^2 measures the variance of an endogenous latent variable explained by the exogenous latent variables, so this index evaluates the reliability of the structural model.
The *redundancy* is the variance of the manifest variables x (connected with the endogenous latent variables),which is explained by the latent variables of the model, both endogenous and exogenous.

$$RED_{xlk}^2 = \frac{Var(\lambda_{ij}\gamma_h\xi_{hj})}{Var(y_{ij})} \tag{2.17}$$

This index is computable only for the manifest variables connected with an endogenous latent variable. So the redundancy is equal to

the communality multiplied by the R^2, and, for this reason, its value is small.
The cross validation of the Communality and Redundancy, that are descriptive indexes, is made across the Stone-Geisser test that follows a Blindfolding procedure: it repeats (for all data points) omission of a part of data, by row and column, (instead Jacknife only by row) matrix estimating parameters, and then reconstruction of the omitted part by the estimated parameters. This procedure results in:

- a generalized cross validation measure that, in case of a negative value, implies rejection of the related structural equation

- jacknife standard deviations of parameters (but most often this quantities are small and lead to significant parameters)

$$H_j^2 = 1 - \frac{\sum_h \sum_i (x_{jhi} - \bar{x}_{jh} - \hat{\pi}_{jh(-i)} v_{j(-i)})^2}{\sum_h \sum_i (x_{jhi} - x_{j\bar{h}})^2} \tag{2.18}$$

$$F_j^2 = 1 - \frac{\sum_h \sum_i (x_{jhi} - \bar{x}_{jh} - \hat{\pi}_{jh(-i)} Pred(v_{j(-i)}))^2}{\sum_h \sum_i (x_{jhi} - x_{j\bar{h}})^2} \tag{2.19}$$

The *validity or the unidimensionality* is a property to verify for the reflective model. In the reflective case it is important to have for the block the internal consistency, that means that the manifest variables are an expression of a same latent concept. This property is measurable in three different ways:

- by the eigenvalues of a Principal Component Analysis, according the Kaiser rule, the number of significant dimensions is given by the number of eigenvalues greater than one. So for this condition for the unidimensionality of a block we expect that only the first eigenvalue is major than 1;

- the Cronbach α, an index based on the calculation of correlation between the manifest and latent variables:

$$\alpha = \frac{\sum_{j'} cor(x_{ij}, x_{ij'})}{p_j \sum_{j'} cor(x_{ij}, x_{ij'})} \times \frac{p_j}{p_j - 1}$$

- the ρ of Dillon-Goldestein, an index that is based on the correlation computed by the model (loadings):

$$\rho_j = \frac{(\sum_{j=1}^{j_i} \lambda_{ij})^2}{(\sum_{j=1}^{j_i} \lambda_{ij})^2 + \sum_{j=1}^{j_i}(1 - \lambda_{ij}^2)}$$

To have the unidimensionality, the α and ρ indexes have to be major than 0.7. This control is not made in the case in which the relationship between the variables are formative [13].
If the block is not unidimensional there are two alternatives: i)the manifest variables that cause the non-unidimensionality are eliminated from the block; ii) the block is divided into two or more blocks, increasing the number of latent variables.
The *structural prediction* is used to understand if an endogenous latent variable is correctly explained by the endogenous latent variables, that is the R^2 of the inner regression. The change in R^2 is explored to see whether a specific exogenous latent variable has a substantive impact on the R^2(effect size f^2):

$$f^2 = \frac{R^2_{included} - R^2_{excluded}}{1 - R^2_{included}}$$

- if $f^2 \approx 0.02 \rightarrow$ small impact

[13] In this case it is possible to have multicollinearity between variables, because the latent concept is formed by the manifest variables, that measure aspects also correlated between them.

- if $f^2 \approx 0.15 \rightarrow$ medium impact

- if $f^2 \approx 0.35 \rightarrow$ large impact

The *discriminant variability* measures if two latent variables express two different concepts. In this circumstance, the correlation between two latent variables must be significantly lower than 1:

$$H_0 : cor(\xi_j, \xi_{j\prime}) = 1 \text{against the} H_1 : cor(\xi_j, \xi_{j'}) \prec 1$$

Another rule is to build the interval confidence for the correlation at 95 percent: if it does not include the value 1, the null hypothesis is rejected. It is also possible to compare the correlation between two latent variables with the mean of variance of the block:

$$(AVE_j \text{and} AVE_{j'}) \succ cor(\hat{\xi}_j, \hat{\xi}_{j'})$$

this means that the latent variables better explain the manifest variables than the other latent variables.
This index also exists for the manifest variables, called monofactoriality of manifest variables, that have to be more correlated with their latent variable, than others of the model:

$$cor(x_{ij}, \xi_j) \succ cor(x_{ij}, \xi_{j\prime})$$

The discriminant validity is checkable also across a matrix in which the average communality for each latent variable, and the R^2 are reported : if the value of the average communalities of variable j and j' are major than the R^2, it means that the two variables express two different concepts.
The index for the global validation of a SEM model, estimated with the PLS-PM, is the Goodness of Fit index(Gof):

$$\underbrace{\sqrt{\frac{1}{\sum_j p_j} \sum_j \sum_h cor^2(x_{ij}, \xi_j)}}_{\text{validation of the outer model}} \times \underbrace{\sqrt{\frac{1}{\text{Number endogenous LV}} \sum_{\text{endogenous LV}} R^2(\xi_j ; \xi_i \text{explained by} Y_j)}}_{\text{validation of the inner model}}$$

Notice that it is a geometrical mean of average communality, multiplied by an average R^2.

The Gof is a compromise between the quality of the outer model and the quality of the inner model, so the normalized index is obtained reporting each part to its maximum value. In particular for the outer estimation (the first part of the formula is the average communality) for each block the maximum is the first eigenvalue, because the first principal component explains the maximum variability, while for the inner estimation the maximum is given by the first canonical correlation squared.
To verify Gof significativity it is possible to build an interval confidence by the Bootstrap technique as also for the R^2.

An important issue is that in PLS the signs of the latent variables are indeterminated. Since arbitrary sign changes in the parameter estimates of the various bootstrap samples can increase their standard error to a substantial degree, procedures have been developed to correct for sign reversals. The user can choose between two correction procedures: in the first option (individual sign changes), the sign of each individual outer weight is made equal to the corresponding sign in the original sample. Because this procedure does not check for the overall coherence of the model as would be done if mental "reverse coding" [8] were performed, this option should be used with special care.

The second option (construct level changes) compares the loadings for each latent variable with the original loadings and reverses the sign of the weights if the absolute value of the summed difference between the original and the bootstrap loadings is greater than the absolute value of the sum of the original loadings and the bootstrap loadings [55]. However, both procedures do not guarantee that sign changes are properly handled.

In this way it is possible to obtain an empirical distribution for the parameters on which are calculated the values of bounadries for a Confidence Interval at 95 percent of significant. The T-Statistics is computed, instead, as the ratio between the mean and the standard deviation of the replication bootstrap, while for the global validation it's considered the Gof index, used in the PLS-PM.

2.2 Some remarks

In this section we highlight the most important characteristics that differentiate the two approaches, LISREL and PLS-PM, presented in this chapter. They are two different approaches to solve the system of equations.

The two techniques estimate a SEM starting from two different points: the first approach, LISREL, using the Maximum Likelihood to estimate the structural parameters, is considered a covariance based approach, because the objective is the minimization of the distance between the observed and etimated covariance matrix; the second, PLS-PM, is considered a variance based approach, because the aim is the maximization of the variance explained by the manifest variables, in order to obtain the best prediction of the latent variables.

It is clear that the two techniques face the same problem (the estimation of a SEM), but with two different aims: we can confirm a theory, and in this case we use the LISREL approach, that allows to confirm an hypothesis in respect to another, by the inferential tests, we can explore a theory and in this case we use the PLS-PM approach, taht gives the best prediction of the latent variables. We obtain the same results, the estimation of the structural parameters, but with the best prediction of the latent variables if we use the PLS-PM.

Besides, they differes for other two important things: the distributional hypothesis, that are present in the LISREL, but not in the PLS-PM, and the dimension of the raw data matrix. In LISREL it is necessary to have a number of observations higher than the number of variables to can estimate the model, because it estimates the structural parameters simultaneously, while the PLS-PM does nt have this problem because it estimates the parameters separately for each latent block and for each manifest variable (we remember that we have, in the reflective case, simple regression).

These and other differences, explained and presented in this chapter, define our choice to work with the PLS-PM and to introduce in it an internal iterative procedure to quantify the ordinal variables, becuase our aim is to explore and not to confirm a prior theory on the data.

Chapter 3

Alternating Least Squares Optimal Scaling algorithms

The analysis of ordinal data by optimal scaling (see the definition given in the chapter 1) methods leads to the search for a quantification of the categories of the ordinal variables that respects the ordinal structure and maximizes a suitable criterion [12, 19, 67]. Algorithms for the search for optimal scaling, called Alternating Least Squares Optimal Scaling (ALSOS) have been proposed for analysis of variance [10, 30], multiple regression [65], principal component analysis [66], canonical analysis [59], generalized canonical analysis [53, 60, 61] as well as other methods.

The set of scalings of an ordinal variable is a convex polyhedral cone which thus plays an important role in these algorithms. For example monotone regression is the projection of a vector into a convex polyhedral cone, under constraints relative to the nature of the variables.

The monotone regression and the projection into a convex cone are on the base, in particular, of two ALSOS algorithms, called Morals and Princals, that are presented in this chapter and will be used in our proposal.

The two algorithms allow to obtain two important results: Morals estimates the parameters of the model according the nature of the manifest varibles, while Princals can be used to obtain the first estimation of the latent variable,substituting the first step of the classical algorithm of PLS-PM. In this chapter will be presented the characteristics of the Alternating Least Squares algorithms and in particular of Morals and Princals, used in our proposal.

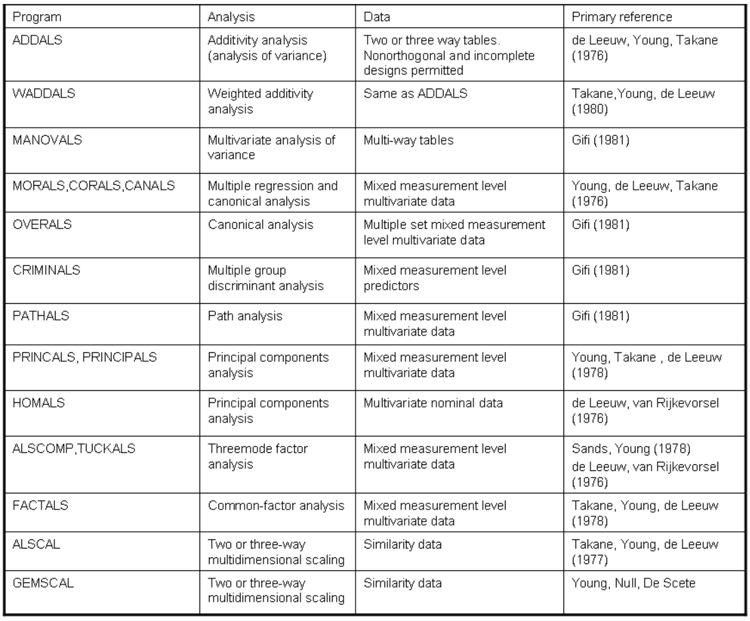
3.1 The ALSOS algorithms

According to the Bock's definition, reported in the first section of chapter 1, Young, De Leeuw and Takane have developed a system of programs to quantify qualitative data (see figure 3.1). The algorithms allow the data to have a variety of measurement characteristics, and allow data analysis with different models.

This system is called ALSOS system since it uses the Alternating Least Squares (ALS) approach to Optimal Scaling (OS). An ALSOS algorithm can be used to obtain a least squares description of qualitative data (having different measurement characteristics).

The ALSOS system includes several programs which quantify qualitative data by applying a)the simple additive model, b)the weighted additive model, c)the multiple regression model, d) the canonical regression model, e) the principal component model or f) the common-factor model, g) the three-mode factor model, or h) the multidimensional scaling model. The data can be defined at the binary, nominal, ordinal and interval levels of measurement, and they can be generated by either a discrete or continuous underlying process.

Each of the ALSOS programs optimizes an objective loss function by using an algorithm based on the alternating least squares and op-



Program	Analysis	Data	Primary reference
ADDALS	Additivity analysis (analysis of variance)	Two or three way tables. Nonorthogonal and incomplete designs permitted	de Leeuw, Young, Takane (1976)
WADDALS	Weighted additivity analysis	Same as ADDALS	Takane,Young, de Leeuw (1980)
MANOVALS	Multivariate analysis of variance	Multi-way tables	Gifi (1981)
MORALS,CORALS,CANALS	Multiple regression and canonical analysis	Mixed measurement level multivariate data	Young, de Leeuw, Takane (1976)
OVERALS	Canonical analysis	Multiple set mixed measurement level multivariate data	Gifi (1981)
CRIMINALS	Multiple group discriminant analysis	Mixed measurement level predictors	Gifi (1981)
PATHALS	Path analysis	Mixed measurement level multivariate data	Gifi (1981)
PRINCALS, PRINCIPALS	Principal components analysis	Mixed measurement level multivariate data	Young, Takane , de Leeuw (1978)
HOMALS	Principal components analysis	Multivariate nominal data	de Leeuw, van Rijkevorsel (1976)
ALSCOMP,TUCKALS	Threemode factor analysis	Mixed measurement level multivariate data	Sands, Young (1978) de Leeuw, van Rijkevorsel (1976)
FACTALS	Common-factor analysis	Mixed measurement level multivariate data	Takane, Young, de Leeuw (1978)
ALSCAL	Two or three-way multidimensional scaling	Similarity data	Takane, Young, de Leeuw (1977)
GEMSCAL	Two or three-way multidimensional scaling	Similarity data	Young, Null, De Scete

Figure 3.1: ALSOS program

timal scaling principles. The ALS principle involves dividing all of the parameters into two mutually exclusive and exhaustive subsets: i) the parameters of the model; ii) the optimal scaling parameters. The algorithm proceeds optimizing the loss function in respect to one subset, then the other, obtaining the least squares estimates of the parameters in one subset while assuming that the parameters of the other subset are constant. This procedure is called conditioned least squares estimate, because the least squares estimation is conditioned by the values of the parameters in the other subset. So alternatively we obtain the conditional estimation for the parameters of the two subsets, until the convergence.

The characteristics of an ALSOS algorithm are:

- it not considers distributional hypothesis on the data, in particular there is not the normal assumption on the latent variables, that produce the observed variables;
- it is not sensible to the nature of observed variables, that is it respects both the process that produces the data (discrete or continuous) and the measurement scale of the data (nominal, ordinal or interval);
- it is possible to choose the method based on the research purpose;
- it estimates the parameters of the model and the optimal scaled variables together ;
- it has one objective function, because the same algorithm estimates the model and computes the optimal quantification

3.1.1 The quantification process

The advantage to use ALS with the Optimal Scaling (OS) is that it is possible to quantify a qualitative variable, apart from the model to estimate. For the optimal scaling we assume that there is a *model space*, represented by a vector whose elements are measured at the cardinal level, and a *data space* [64], represented by a vector of data, and we can assume to know the measurement characteristics of the data. These two spaces are relevant to obtain the *optimal scaling space*.

The goal of OS is to derive an *optimal scaling space* that satisfies two characteristics:

- the measurement characteristics of the data space

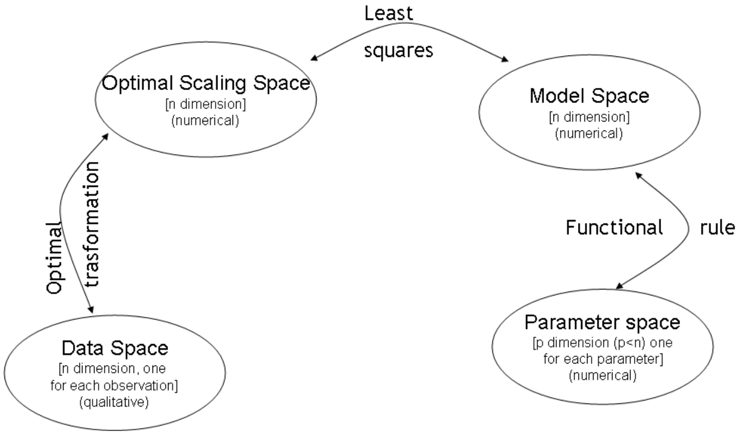


Figure 3.2: The quantification process

- it must have a least squares relationship to the model space, given that the measurement characteristics are perfectly satisfied

Now it is possible to define a vector of raw observations as $\bar{o}$ with general element o_i, and the model vector $\bar{z}$, with general element z_i, and the optimal scaling vector $\bar{z}^*$ with general element z_j^*. The vector $\bar{o}$ is the data space (all observations in a particular category are contiguous), the vector $\bar{z}$ and $\bar{z}^*$ are the model and optimally scaled spaces (the elements of these spaces are organized in a fashion having one to one correspondence with $\bar{o}$). The vector z_j^* is the parameter representing the observation o_i.

The OS problem is to obtain a transformation t of the raw observations which generates optimally scaled observations:

$$t[\bar{o}] = [\bar{z}^*]$$

where t is a function of the measurement characteristics of the observations, and is such that a least squares relationship will exist between $\bar{z}$ and $\bar{z}^*$, maintaining the measurement characteristics. The numerical value assigned to z_j^* is the optimal parameter value for the observation o_i.

The constraints applied to the function t are of two different kinds: measurement level and measurement process. The restrictions measurement concerns the relationships among all the observations within a single category, while the level measurement concerns the relationships among all the observations between different categories. The restrictions are of two types: discrete process and continuous process; if the process is discrete (for example the female and male) all observations of a category should be represented by the same real number obtained across the transformation t^d, while if the process is continuous (for example 97.2 kg), each of the observations of a category should be represented by a real number selected from a closed interval[1]. The figure (3.3) summarizes measurement implications of the constraints.

From a mathematical point of view discrete constraints can be formalized as:

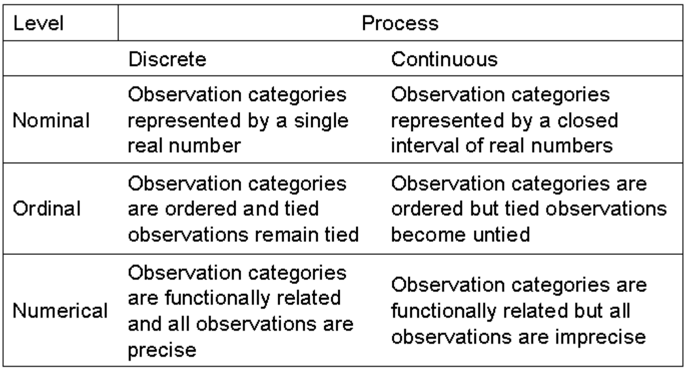
$$t^d : (o_i \sim o_m) \rightarrow (z_i^- = z_m^*) \tag{3.1}$$

where $\sim$ means the empirical equivalence (the observations in the same category have the same real number). The continuous constraints are:

[1] The discrete nature of the process is reflected by the choice of a single number (discrete) to represent all observations in the category; the continuous nature of the process is reflected by the choice of a real number from a closed (continuous) interval of real numbers.

$$t^c : (o_i \sim o_m) \rightarrow (z_i^- = z_m^*) \leq \{z_i^*, z_m^*\} \leq (z_i^* = z_m^*) \tag{3.2}$$

where z_i^- and z_m^* are the lower and upper bounds of the closed interval of real numbers. The implication of the empirical equivalence is that the boundaries of all observations in a particular category are the same. The continuous constraints also imply that, for all the observations of a category, the parameters of optimal scaling have to belong to the same interval, but need not be equal.



Level	Process	
	Discrete	Continuous
Nominal	Observation categories represented by a single real number	Observation categories represented by a closed interval of real numbers
Ordinal	Observation categories are ordered and tied observations remain tied	Observation categories are ordered but tied observations become untied
Numerical	Observation categories are functionally related and all observations are precise	Observation categories are functionally related but all observations are imprecise

Figure 3.3: The measurement characteristics for the types of measurement

The level constraints define the function t to obtain the optimal transformation. For the nominal level there are not constraints because they are specified by the process restraints [2].

For ordinal variables beyond the constraints concerning the process of measurement, it is required that the real numbers, associated to the observation in different categories, respect the original order:

$$t^o : (o_i \prec o_m) \rightarrow (z_i^* \leq z_m^*) \tag{3.3}$$

For numerical variables it is required that the real numbers associated to the observations be functionally related to the observations, for example across a polynomial rule:

$$t^n : z_i^* = \sum_{p=0}^{p} \delta_q o_i^* \tag{3.4}$$

If p=2 we have a quadratic relationship between the raw and the optimally scaled observations, while if p=1 we have the linear relationship.

The (3.4) represents the relationship between the model, data and optimal scaling spaces. In the "problem space", that is composed by model, data and optimal scaling spaces, of dimension n, the data problem analysis is characterized and solved. The parameter space, indeed, is not included in the problem space and it has dimensionality p, one dimension for each of the p parameters, that is much less than n. The parameter and model space are related by a rule called a "combination rule"[64].

In this figure we can see that the model space is represented as vectors while the data space as a cone, and both (vectors and cone) are intersected in the origin of problem space. The elements of the vector $\bar{z}^*$ define a point in the problem space and any point in the optimal scaling space is equivalent to any other point. The data space

[2] It is identifying two types of nominal variables,according to whether the process is discrete or continuous: in the first case we have discrete-nominal variables (for example the the sex of a person), while in the second case we have the continuous-nominal variables (the colors).

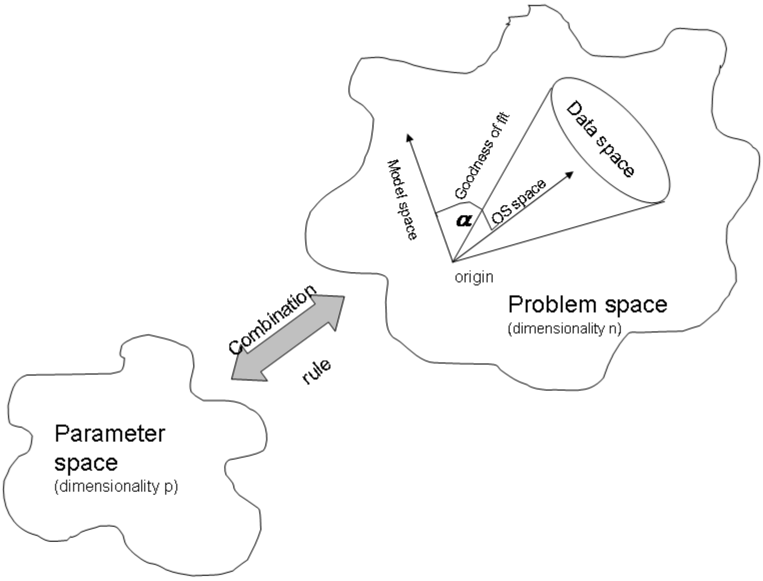


Figure 3.4: Geometrical representation of the ALSOS theory

has this representation because the cone represents the measurement characteristics; so the optimal scaling vector must be contained in the data cone and near the model vector. The model and optimal scaling spaces are as nearly alike as possible in a least squares sense.

Generally the optimal scaling vector is on the surface of the cone, that is the part of the cone which is generally closest to the model space. Instead the angle α represents the goodness of fit between the two spaces: the more smaller is the angle, the better the fit.

3.1.2 The projection of a vector into a convex cone

In this section we will discuss and present the properties of a projection of a vector in a convex cone; some propositions are proved in the Appendix A. The properties and determination of the projection of a vector into a convex polyhedral cone play an important role in optimal scaling algorithm.

A convex polyhedral cone C can be defined as a set of vectors x in $\Re^n$ verifying the condition that $A'x \leq 0$, where A is a matrix with n rows and m columns and A' is its transpose. A convex polyhedral cone C is generated by a finite number of generators: there exist vectors $s_1, ..., s_k$ of $\Re^n$ such that any element x of C can be written as $x = \sum \{\lambda_i s_i; i = 1, ..., k\}$, where $\lambda_1, ..., \lambda_k$ are nonnegative. Conversely, any cone with a finite number of generators is polyhedral. We denote by *C(S)* the conical hull of a set S in $\Re^n$. The polar cone C^p of C is the set of vectors y in $\Re^n$ such that $x'y$ is nonpositive for any vector x of C. The polar cone C^p is polyhedral: it is generated by the columns of the matrix A.

The properties of the projection of a vector of $\Re^n$ into a closed convex set in $\Re^n$ can be applied to the particular case of a convex polyhedral cone.

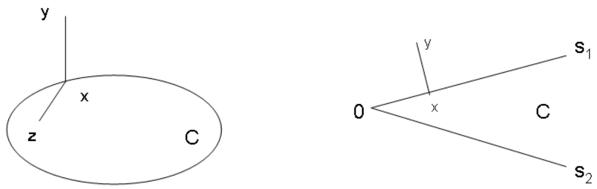


Figure 3.5: Projection of a vector y into a closed set K or a convex polyhedral cone C

Proposition 3.1. *Let y be a vector of* $\Re^n$ *and K a closed convex set in* $\Re^n$*. There exists a unique vector x of K such that* $||y-x|| \leq ||y-z||$ *for any vector z of K. This vector x is the projection of y into K. Furthermore, a vector x is the projection of y into K if and only if* $(y-x)'(z-x) \leq 0$ *for any vector z of K.*

Applying this proposition to the particular case of a convex polyhedral cone, results in Proposition 3.2.

Proposition 3.2. *Let x be the projection of a vector y of* $\Re^n$ *into a convex polyhedral with* $C = C(S)$ *that is generated by a set* $S = s_1, ..., s_k$*. Let R be the set of vectors* s_i *of S orthogonal to* $y-x$*.*

Then the vector x is equal to the projection of y into the subspace $L(R)$ generated by the vectors of R.

The polar cone plays the same role as the orthogonal subspace. This is shown by the following proposition.

Proposition 3.3. *Let y be a vector of $\Re^n$ and C a convex polyhedral cone. The orthogonal decomposition $y = x + z$, where x is in C, z in C^p, and $x'z = 0$, is unique and is obtained from the projection x and z of the vector y into C and C^p respectively.*

Corollary 3.1. *Let y be a vector of $\Re^n$ and $C = C(S)$ be a convex polyhedral cone generated by the set $S = s_1, ..., s_k$. Let R be a subset of S. The projection x of y into $L(R)$ is equal to the projection of y into C if and only if $(y - x)'s \leq= 0$ for any s of $S - R$, where $S - R$ is the set of vectors of S which do not belong to R.*

Corollary 3.2. *Let C be a convex polyhedral cone. A vector y belongs to the polar cone C^p if and only if the projection of the vector y into C is the null vector. The projection of a vector into a convex polyhedral cone possesses optimal properties often used in optimal scaling algorithms.*

Proposition 3.4. *Let C be a convex polyhedral cone, y a vector in $\Re^n$ which does not belong to C^p, x the projection of y into C, and z any vector of C. We have the following results:*

1. *The minimum of $||y - z||/||y||$ over z in C is reached at $z = x$ and is equal to $(1 - cos^2(x, y))^{1/2}$*
2. *The minimum of $||y - z||/||y||$ over z in C is reached at $z = (1/cos^2(x, y))x$ and is equal to $(1 - cos^2(x, y))^{1/2}$*
3. *The maximum of cos(y,x) over z in C is reached at any nonnull vector of $C(x)$. Furthermore, $cos(y, x) = cos(y, z)$ implies that z is any nonnull vector of $C(x)$*

Corollary 3.3. *Let C be a convex polyhedral cone, y a vector of $\Re^n$ not belonging to C^p,S the sphere of radius 1, and x the projection of y into C. Then the projection of y into $C \cap S$ is equal to $x/||x||$.*

This proposition is useful for the study of the convergence of alternating least squares algorithms.

Proposition 3.5. *Let C be a convex polyhedral cone and S the unit sphere. The projection operator A into C is continuous on$\Re^n$, and the projection operator B into $C \cap S$ is continuous on $\Re^n - c^p$.*

Many algorithms have been proposed for the determination of the projection x of a vector y into the convex polyhedral cone $C = C(S)$ generated by the vectors of $S = s_1, ..., s_k$. Quadratic programming can be used to minimize $||y - \sum \alpha_i s_i; i = 1, k||$ subject to the constraints $\alpha_1, ..., \alpha_k \geq 0$. This is done by using standard computational routines for regression. Waterman (1974) proposed to calculate all the possible regressions of y on a subset of S. Deutsch, McCabe and Phillips (1975) improved this procedure by checking the optimality of each subset. Armostrong and Frome (1976) proposed a branch-and-bound algorithm that restrains the number of subsets of S to be examined. More efficient algorithms were been proposed by Lawson and Hanson (1974) and Bremner (1982). The Nonnegative Least Squares algorithm searches the optimal subset R by an iterative multiple regression. At the initialization step R is empty. During the current step, we obtain a subset R of S such that the projection $y_R = \sum \alpha_i s_i; s_i \in R$ of y into $L(R)$ has all its coefficients α_i strictly positive. If $(y - y_R)'s$ is nonpositive for any s belonging to $S - R$, then y_R is the projection of y into C. Otherwise, we add to the subset R either the vector s that maximizes $(y - y_R)'s$ (Lawson and Hanson) or the t statistic of s in the regression of y on $R + s$, the union of sets R and s (Bremner). Now it is possible to find a subset Q of R such that the projection $y_{Q+s} = \sum \{\beta_i s_i; s_i \in Q + s\}$ of y into $L(Q + s)$ has all its coefficients β_i strictly positive and $\|y - y_{Q+s}\| \prec \|y - y_R\|$. The

algorithm converges to the optimal subset, because, at each iteration, the distance between y and y_R strictly decreases and the number of subsets R is finite.
Lawson and Hanson do not explicitly prove the convergence of the NNLS algorithm to the optimal subset R. This can be done quite easily, however, by using only geometrical arguments, as is shown below.

Proposition 3.6. *Let R be a subset of S such that the projection $y_R = \sum \{\alpha_i s_i; s_i \in R\}$ of y into $L(R)$ has all its coefficients α_i strictly positive. Suppose that there exists a vector s of S-R such that $(y - y_R)'s$ is strictly positive. Then there exists a subset Q of R such that the projection $y_{Q+s} = \sum \beta_i s_i; s_i \in Q + s$ of y into $L(Q+s)$ has all its coefficients β_i strictly positive and such that $\|y - y_{Q+s}\| \prec \|y - y_R\|$.*

3.1.3 Methods of quantification

The aim of an OS algorithm is to obtain the best quantification across the definition of a function t that depends on the nature of the process and on the measurement level of the variables, as shown in the figure(3.6).

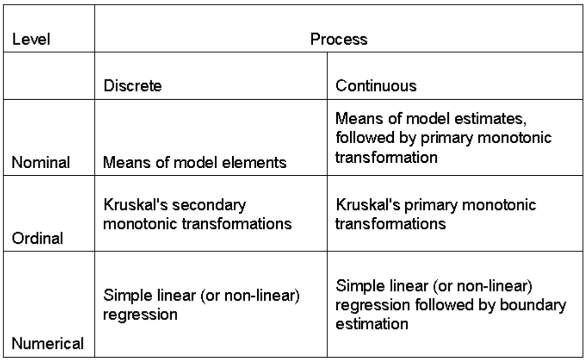
The estimation process is easy, and consists of the definition of an element z_i^* as the mean of all z_i, that corresponds to observation o_i in a category. So z_i^* under the discrete-nominal constraints is:

$$t^d : \bar{z}^* = \bar{G}(\bar{G}'\bar{G})^{-1}\bar{G}'\bar{z} \tag{3.5}$$

where $\bar{G}$ is a binary matrix, whose elements are:

$$g_{ij} = \begin{cases} 1 & \text{if } o_i \text{ is in category c} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

For t^c we have one more requirement that all optimally scaled observations are in the interval, with no restrinction on the determination



Level	Process	
	Discrete	Continuous
Nominal	Means of model elements	Means of model estimates, followed by primary monotonic transformation
Ordinal	Kruskal's secondary monotonic transformations	Kruskal's primary monotonic transformations
Numerical	Simple linear (or non-linear) regression	Simple linear (or non-linear) regression followed by boundary estimation

Figure 3.6: Optimal scaling methods

of the interval.
The two ordinal quantifications t^{do} (discrete-ordinal) and t^{co} (continuous-ordinal) are defined by Kruskal's least squares monotonic transformation. The equation of both transformations is:

$$t^o : \bar{z}^* = \bar{G}(\bar{G}'\bar{G})^{-1}\bar{G}'\bar{z} \tag{3.7}$$

where $\bar{G}$ is a binary matrix. The t^p transformation can be written in matrix notation as

$$t^p : \bar{z}^* = \bar{G}\bar{\delta} \tag{3.8}$$

where $\bar{G}$ is a matrix with a row for each observation and with $p+1$ columns, that being an integer power of the vector $\bar{o}$ of observations.

The transformation t may be considered as though we are regressing the model space $\bar{z}$ onto the observation space $\bar{o}$ in a least squares sense and under the measurement constraints. In particular each t can be considered as the projection operator

$$E = \bar{G}(\bar{G}'\bar{G})^{-1}\bar{G}' \tag{3.9}$$

from which → $\bar{z}^* = \bar{E}\bar{z}$.
So the least squares function can be written as:

$$\phi^2 = (\bar{z}^* - \bar{z})'(\bar{z}^* - \bar{z}) \tag{3.10}$$

and if we define $\bar{F} = 1 - \bar{E}$ we obtain

$$\phi^2 = \bar{z}'\bar{F}\bar{z} \tag{3.11}$$

The transformation can be viewed as optimizing a relationship between the model space, where the linear combination is determined by the measurement restrictions. Geometrically the projection operator projects the model space $\bar{z}$ onto the nearest surface of the data space cone.

3.1.4 Normalization

A trivial and undesirable way to minimize the equation (3.10) is to set the model subspace $\bar{z}$ equal to zero, but in consequence we have $\bar{z}_b$ and ϕ^2 equal to zero, for all transformations.
So the normalization of the solution is made in the ALSOS algorithm to avoid solutions represented by the origin of the problem space or other types of trivial solutions. There are different ways to normalize the solutions, and two of these are:

$$\phi_a^2 = \frac{(\bar{z}_a^* - \bar{z})'(\bar{z}_a^* - \bar{z})}{\bar{z}'\bar{z}} \tag{3.12}$$

or

$$\phi_b^2 = \frac{(\bar{z}_b^* - \bar{z})'(\bar{z}_b^* - \bar{z})}{\bar{z}'\bar{z}} \tag{3.13}$$

where $\bar{z}_b^*$ and $\bar{z}_a^*$ are the normalized versions of $\bar{z}^*$ which optimize ϕ_a^2 and ϕ_b^2, respectively. That is,

$$\bar{z}^a = a\bar{z}^* \tag{3.14}$$

and

$$\bar{z}^b = b\bar{z}^* \tag{3.15}$$

where a and b are two non-negative real numbers.

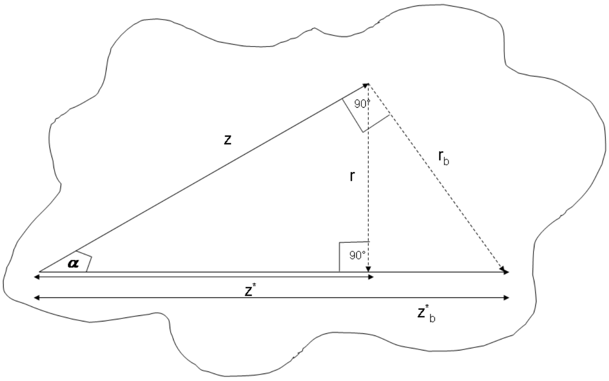


Figure 3.7: Geometrical representation of the normalization aspects of ALSOS algorithms

The figure(3.7) presents a portion of the problem space shown in the figure(3.6). Above the cone's surface the model vector z^* is shown; the orthogonal projection of the model vector onto the surface

of the cone gives z^g the unnormalized optimally scaled data. This projection is obtain by the operator E which minimizes (3.10), the unnormalized index of fit.
The angle α, between $\bar{z}$ and z^g has been minimized by orthogonally projecting $\bar{z}$ onto the cone's surface. The projection defines a right triangle so as follows

$$sin^2\alpha = \frac{r^2}{\bar{z}^2} = \frac{r'r}{\bar{z}'\bar{z}} = \phi_a^2 \tag{3.16}$$

So the orthogonal projection of $\bar{z}$ onto the cone's surface requires a right angle at the surface of the cone.

3.2 Princals and Morals: two ALSOS algorithms

In the next two sections we will present the Morals and Princals algorithms that are used in our proposal to estimate a SEM based on ordinal variables.
Of these two techniques the properties and the steps of algorithms will be explained.

3.2.1 The Princals algorithm

Given a data matrix $n \times m$ of metric variables, Principal Component Analysis (PCA) is a common technique to reduce the dimensionality of the data set, projecting variables into a subspace $\Re_p$ where $p \ll m$. The Eckart-Young theorem states that this classical form of linear PCA can be formulated by means of a loss function.Its minimization leads to a $n \times p$ matrix of component scores and an $m \times p$ matrix of component loadings.

The actual computer programs for PCA impose some restrictions about the completeness of the data matrix and interval measurement of variables. In the social sciences the assumption of interval scales is not usually justified, and often the data matrix is incomplete. In this case it is possible to use the Nonlinear Principal Component Analysis (NPCA), where the term nonlinear pertains to nonlinear transformation of the observed variables [14]. According to the Gifi terminology the NPCA can be defined as homogeneity analysis with restrictions on the quantification matrix Y_J. The ALS algorithm generalizes the approach of PCA to general types of variables.

The Non linear PCA in the ALSOS system is derived as homogeneity analysis with some constraints; the loss function is

$$\begin{aligned}\sigma(X;Y_1,....,Y_J) &= J^{-1}\sum_{j=1}^{J} SSQ(X-G_jY_j) \qquad (3.17)\\ &= J^{-1}tr(X-G_jY_j)'(X-G_jY_j)\end{aligned}$$

with the constraint of rank-one

$$Y_j = q_j\beta_j' \text{with} j \in J \qquad (3.18)$$

The constraints are imposed on the multiple category quantifications, with q_j a l_j-column vector of single category quantifications for variable j, and β_j a p-column vector of weights (component loadings). In this way each quantification matrix Y_j has to be of rank one, so that the quantifications in p dimensional space are proportional to each other. With the introduction of the rank one restrictions it is possible to have multidimensional solutions for object scores with a single quantification for the categories of the variables, and it is possible to introduce the measurement level of the variables in the analysis.

To minimize (3.17) with the constraints (3.18) the algorithm starts to compute Y_j as

$$\hat{Y}_j = D_j^{-1}G_j'X \qquad (3.19)$$

where $D_j = G_j'G_j$ is the diagonal matrix containing the univariate marginals of variable j. The loss function $\sigma(X; Y_1, ..., Y_J)$ is partitioned as follows:

$$\sum_{j=1}^{J} tr(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)])'(X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]) =$$

$$\sum_{j=1}^{J} tr(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) + \sum_{j=1}^{J} tr(Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j)$$

Imposing the rank one restrictions the loss function to minimize, respect to q_j and β_j is:

$$\sum_{j=1}^{J} tr(q_j\beta_j' - \hat{Y}_j)'D_j(q_j\beta_j' - \hat{Y}_j) \tag{3.20}$$

The ALS algorithm alternates over q_j and β_j, which gives for fixed q_j

$$\hat{\beta}_j = (\hat{Y}_j'D_jq_j)/(q_j'D_jq_j) \tag{3.21}$$

and for fixed β_j

$$\hat{q}_j = \hat{Y}_j\beta_j/(\beta_j'\beta_j) \tag{3.22}$$

At this point it is necessary to take into account the restrictions imposed by the measurement level of the variables. This means that we have to project the estimated vector $\hat{q}_j$ on the cone C_j: in the case of ordinal variables the cone C_j is the cone of monotone transformation given by
$C_j = \{q_j | q_j(1) \leq q_j(2) \leq .....q_j(l_j)\}$. So the projection is obtained across a monotone regression in the metric D_j (weights). In the case of numerical data the corresponding cone is a ray given by
$C_j = \{q_j | q_j = \gamma_j + \delta_j s_j\}$, where s_j is a given vector, for example, the original variable quantifications. In this case the projection problem is a regression problem. In the case of nominal variables the cone is the $\Re_j^l$ space and the projection is done by simply setting $q_j = \hat{q}_j$, so

$\hat{Y}_j = \hat{q}_j \hat{\beta}'_j$ and the algorithm proceeds with the estimation of the object scores. This solution is referred in the literature as the Princals solution [15, 19](principal component analysis by means of alternating least squares); if the variables are treated as single numerical the Princals solution corresponds to an ordinary Principal Component Analysis on the s_j variables [15] appropriately standardized, computing the eigenvalues and eigenvectors of the correlation matrix of the s_j variables.

The Princals model allows the data analyst to treat each variable differently; some may be treated as multiple nominal and some others as single nominal, ordinal or numerical. Moreover, with some additional effort (see [41]) one can also incorporate in the analysis categorical variables of mixed measurement level, that is variables with some categories measured on an ordinal scale (e.g. Likert scale) and some on a nominal scale (e.g. categories in survey questionnaires corresponding to the answer “don’t know”).

The steps of Princals are:

Algorithm 2. *Princals.*

```
Initialize $X$ so that $u'X = 0$ and $X'X = NI_p$
repeat
Step1
    for (j in 1:k)
        $\hat{Y}_j = D_j^{-1} G'_j X$
    endfor
Step2
    for (j in 1:k)
        $\hat{\beta}_j = (\hat{Y}'_j D_j q_j)/(q'_j D_j q'_j)$
    endfor
Step3
    for (j in 1:k )
        quantification of the $j^{th}$ variable by a monotone
        or linear regression
    endfor
Step4
    for (j in 1:k )
        Update the quantifications $\hat{Y}_j = \hat{q}_j \hat{\beta}'_j$
    endfor
Step5
    for (j in 1:k )
        $\hat{X} = J^{-1} \sum_{j=1}^{J} G_j Y_j$
        Centering and orthonormalization of the $X$ matrix
    endfor
Ceck the convergence
    if (the objective function is minimum) break
```

Generally the most common options in treating variables in Princals are single ordinal and single numerical; the Gifi loss function can be partitioned into two parts:

$$\sum_{j=1}^{J} tr(X - G_j \hat{Y}_j)'(X - G_j \hat{Y}_j) + \sum_{j=1}^{J} tr(\hat{q}_j \hat{\beta}'_j - \hat{Y}_j)' D_j (\hat{q}_j \hat{\beta}'_j - \hat{Y}_j) \quad (3.23)$$

The first part of the (3.23) can also be written as $N(p - \sum_{j=1}^{J} \sum_{s=1}^{p} \eta_{js}^2)$, called the *multiple loss*, where p is the number of manifest variables. The discrimination measure η_{js}^2 is called *multiple fit* of the variable j in dimension s. Imposing the normalization restrictions $q'_j D_j q_j = N$, and using the fact that $\hat{Y}'_j D_j q_j \beta'_j = N \beta_j \beta'_j$ from the (3.21) the second part of (3.23) can be written as:

$$\sum_{j=1}^{J} tr(\hat{Y}'_j D_j \hat{Y}_j - N \beta_j \beta'_j) = N(\sum_{j=1}^{J} \sum_{s=1}^{p} (\eta_{js}^2 - \beta_{js}^2) \quad (3.24)$$

called the *the single loss*. The quantities β_{js}^2, $s = 1, ..., p$ are called *single fit*, and correspond to squared component loadings (see chapter

4 in [19]).In the single loss part there are two components: the rank-one restrictions, that is the fact that single category quantifications must lie on a straight line in the joint space, and the measurement level restriction, that is the fact that single quantifications may have to be rearranged to be either in the right order (ordinal variables) or equally spaced (numerical variables).

The ALS algorithm is not affected by the presence of missing data, so the (3.2.1)becomes:

$$\sum_{j=1}^{J} tr(X - G_j Y_j)' M_j (X - G_j Y_j) =$$

$$\sum_{j=1}^{J} tr(X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j)))' M_j (X - G_j(\hat{Y}_j + (Y_j - \hat{Y}_j))) =$$

$$\sum_{j=1}^{J} tr(X - G_j \hat{Y}_j)' M_j (X - G_j \hat{Y}_j) + \sum_{j=1}^{J} tr(Y_j - \hat{Y}_j)' D_j (Y_j - \hat{Y}_j)$$

where M_j is a matrix of 0 and 1: 0 if the data is missing and 1 otherwise.

Cone restricted SVD

The loss function of the Princals can be solved in terms of cone restricted Singular Value Decomposition. All the transformations are projections on some convex cone C_j and we look only to the second term of the partitioned loss function (3.23):

$$\sigma(Q, B) = tr(QB' - \hat{Y})' D (QB' - \hat{Y}) \tag{3.25}$$

over Q, the matrix of the vectors of scaling, and B, the matrix of the weights, where $\hat{Y}$ is $k \times p$,Q is $k \times r$, and B is $p \times r$. The first column Q_0 of Q is restricted by $Q_0 \in C$, and Q should also satisfy the normalization condition $u'DQ = 0$ and $Q'DQ = I$.

The basic idea of the algorithm is to apply alternating least squares

with rescaling, minimizing over Q for fixed B and over B for fixed Q. The algorithm does not impose the normalization conditions when it minimizes over Q.
Suppose that $(\hat{Q}, \hat{B})$ is the best solution at present. To improve it we minimize over the nonnormalized Q, satisfying the cone constraints, and keeping B fixed at $\hat{B}$, obtaining $\tilde{Q}$and a corrisponding loss function value $\sigma(\tilde{Q}, \hat{B})$. Clearly

$$\sigma(\tilde{Q}, \hat{B}) \leq \sigma\hat{Q}, \hat{B}) \tag{3.26}$$

but $\tilde{Q}$ is not normalized. Using the Weighted Gram-Schmidt solution $\tilde{Q} = Q^*S$ we update Q with Q^*, where S is the Gram-Schmidt triangular matrix. The first column $\tilde{q}_0$ of $\tilde{Q}$ satisfies the cone constraints, and because of the nature of Gram-Schmidt, so does the first column of Q^*. It's possible to have:

$$\sigma(Q^*, \hat{B}) \succ \sigma(\hat{Q}, \hat{B}) \tag{3.27}$$

This seems to invalidate the usual convergence proof, which is based on a non-increasing sequence of loss function values. But now also adjust $\hat{B}$ to $\bar{B} = \hat{B}(S^{-1})'$. Then $\tilde{Q}\hat{B}' = Q^*\bar{B}'$, and thus

$$\sigma(\tilde{Q}, \hat{B}) = \sigma(Q^*, \bar{B}) \tag{3.28}$$

Finally compute B^* by minimizing $\sigma(Q^*, B)$ over B. Since $\sigma(Q^*, B^*) \leq \sigma(Q^*, \bar{B})$ we have the chain

$$\sigma(Q^*, B^*) \leq \sigma(Q^*, \bar{B}) = \sigma(\tilde{Q}, \hat{B}) \leq \sigma(\hat{Q}, \hat{B}) \tag{3.29}$$

In any iteration the loss function does not increase.

3.2.2 The Morals algorithm

The Morals algorithm [65]optimizes the multiple correlation between a single criterion variable and a set of predictor variables where any of the variables (criterion included) may be nominal, ordinal or interval. The variables do not all have to be measured at the same level nor does the process, which is assumed to have generated data, may be either discrete or continuous.

Morals obtains an optimal scaling for each variable within the restrictions imposed by the regression model, the measurement level, and the generating process. The scaling is optimal in the Fisher [16] sense of optimal scaling: the multiple correlation is maximized. It is based on the minimization of a quadratic function in respect to three parameters. The initialization of the algorithm is based on the assumption that the matrices X and Y (the raw data) are actually the matrices X^* and Y^* (optimally scaled variables). This is equivalent to assuming, for the initialization process, that the raw data are measured on an interval scale (we assign arbitrary values to the observation when a variable is ordinal).

After the initialization step the algorithm proceeds with the estimation phase, in which it computes the regression coefficients. In the optimal scaling phase, the independent variables (ordinal, nominal or numerical), are specified as the product between an indicator matrix G_j $(n * k_j)$ and a vector of the scaling parameters q_j $(k_j * 1)$, that after estimation defines the variables $x_j^{os} = G_j q_j$. This procedure is made also for the dependent variable y, that becomes $y^{os} = G_y t$, where t is the vector of scaling of the dependent variable y. The loss function to optimize is:

$$min_{\beta,q,t} SSQ = (G_y t - Gq\beta) \tag{3.30}$$

with the constraints $u'G_y t = 0$, $t'G'_y G_y t = 1$, $q_j \in C_j$, $t \in C_y$, where u is a vector of 1 and C_j and C_y are the spaces of admissible transformation (they are closed convex cones) for the categories of each variable, taking into account the level of measurement. In particular if the variables are nominal there are no constraints on the values of quantification, while if the variables are ordinal there are order constraints between the categories. So the final object of this technique aim to obtain the best quantification of the qualitative variables and to optimize the regression parameters. Having obtained, in fact, the first estimation of the vectors y^{os} and x^{os} the parameters of multiple regression are then updated using as variables the ones calculated at the previous step. The steps are retereited until the convergence.

The function in the equation (3.30) is rewritable as:

$$SSQ(G_y t - \hat{y}) + SSQ(\hat{y} - Gq^*) + DP \tag{3.31}$$

The first part of equation (3.31) is minimal respect to t projecting $\hat{y}$ on the columns of G_y

$$t = (G_y' G_y)^{-1} G_y' \hat{y} \tag{3.32}$$

So if the variable is nominal the vector of optimal scaling is $t^{os} = t$, otherwise if the variable is ordinal, the vector t^{os} of optimal scaling is obtained though a monotone regression of t [30]. The second part of the equation (3.31) has to be minimized respect to q_j:

$$q_j = (G_j' G_j)^{-1} G_j' (G_y t - v_j) \tag{3.33}$$

where v_j is the estimation of the independent variable without the contribution of variable j :

$$v_j = \sum_{j \in J_p} (\beta_j G_j q_j - \beta_j G_j q_j) \tag{3.34}$$

Therefore we quantify individually each variable of the regression model and then we obtain the new variable quantified as the product between $G_j q_j$. Also in this case, as before, if the variable is nominal no transformation is required on q_j, instead, if it is ordinal there is the constraint on the order between categories, so it is necessary to compute a monotone regression.

The final step consists on the substitution of the new variables (dependent and independent), obtained with the vectors of scaling, and in the regression model compute, to estimate the optimal paths, taking into account the previous quantification of variables. Since the Morals algorithm accepts all kind of variables, and in the case of numerical variables it jumps the step of quantification and computes only the parameters of the regression model.

Summarizing the steps of Morals algorithm are:

Algorithm 3. *Morals.*

```
Consider the matrix $X$ and $Y$ of raw data as $X^*$ and $Y^*$
repeat
Step1
    for (i in 1:p)
        for (j in 1:k)
            $\beta = (X^{*\prime}X^*)^{-1}X^*Y^*$
        endfor
    endfor
Step2
    for (j in 1:k)
        $\hat{Y} = X^*\beta$
        $Y^G = G_Y(G_Y'G_Y)^{-1}G_Y'\hat{Y}$
        $Y^* = Y^G(\frac{||\hat{Y}||}{||Y^G||})$
    endfor
Step3
    for (i in 1:p )
        $\hat{X}_I = \frac{1}{\beta_J}(Q^* - \sum_{I \neq J}\beta_J X_J^*)$
        $x^G = G_I(G_I'G_I)^{-1}G_I'\hat{X}$
        $X_I^* = X_I^G(\frac{||\hat{X}_I||}{||X_I^G||})$
    endfor
Step4
Ceck the convergence
    if (the the $R^2$ is not improved 'enough' from last
        iteration) break
```

The important characteristic of Morals is the division of the quantification process into two steps: one for the single dependent variable and another for the M independent variables. For each variable it computes the model's estimate of the variables, then it uses the estimate with the appropriate indicator matrix to obtain the unnormalized optimally scaled data, and then the normalization, obtaining the variables rescaled.

To assure convergence and to maintain the ALS aspects of an algorithm with non-independent partitions we must immediately replace the previous scaled data with the newly computed (normalized) scaled data.

The monotone regression

In linear regression we fit a linear function $y = \alpha + \beta x$ to a scatter plot of n points (x_i, y_i). We find the parameters α and β by minimizing

$$\sigma(\alpha, \beta) = \sum_{i=1}^{n} w_i (y_i - \alpha - \beta x_i)^2 \tag{3.35}$$

where w_i are known positive weights. In discussing the regression of $\tilde{y}$ on $\tilde{x}$, then, $\tilde{y}$ is a random variable whose values are real numbers. But $\tilde{x}$ may or may not be a random variable. Its values may or may not be real numbers or vectors (ordered k-tuples) of real numbers. In general $\tilde{x}$ ranges over an abstract set X. In defining regression of $\tilde{y}$ on $\tilde{x}$ via least squares, *weights* $w(x)$ associated with the values of $\tilde{x}$ must be used. If $\tilde{x}$ is a random variable, $w(x)$ is the probability that the random variable $\tilde{x}$ will be equal to x, or the density of $\tilde{x}$ at x. In sampling the weights may be proportional to numbers of observations.

In the more general nonlinear regression problem we fit a nonlinear function $\phi_\theta(x)$ by minimizing

$$\sigma(\theta) = \sum_{i=1}^{n} w_i (y_i - \phi_\theta(x_i))^2 \tag{3.36}$$

over the parameters θ. In both cases, consequently, we select the minimizing function from a family of functions indexed by small number of parameters.

In many situations the researcher has no information regarding the mathematical specification of the true regression function. Rather, he can assume a particular shape which can be characterized by certain order restrictions. Typically, this involves that the y_i's increase with the ordered z_i's. Such a situation is called *isotonic regression*; the decreasing case *antitonic regression*, and both case are called *monotonic regression* (see [13]).

Monotone regression is the projection of a numerical variable y into the convex polyhedral cone of the scalings of an ordinal variable x. In data analysis it is used in multidimensional scaling [29, 30] and for methods which describe ordinal data by cardinal methods [64, 65, 66]. An ordinal variable x is observed on n subjects and takes its values on a set $M = \{1, 2, ..., m\}$ of categories provided with the natural order. A scaling δ of the categories of x is a real non-decreasing function which associates a real number δ_j with each category j of M. The scaling x^* of$x$ induced by δ is the numerical variable which associates the scaling $\delta_{x(i)}$ with each subject i.

If we denote by x_j the dummy variable which is equal to one if $x(i) = j$ and zero otherwise, the scaling x^* can be written $x^* = \sum\{\delta_j x_j; j = 1, m\}$. Taking into account the ordinal constraint the scaling δ_j can be written $\delta_j = \alpha_1 + \alpha_2 + ... + \alpha_j$ with $\alpha_2, ..., \alpha_m \geq 0$. Consequently we get $x^* = \sum\{\alpha_j z_j; j = 1, m\}$ where $z_j = \sum\{x_h; h = j, m\}$. So the set of scalings x^* of the ordinal variable x is the convex polyhedral cone $C = L(z_1 \oplus C(z_2, ..., z_m$ which is the direct sum of the subspace generated by z_1 and the convex polyhedral cone generated by $z_2, ..., z_m$.

If a numerical variable y and an ordinal variable x are observed on a set of n subjects, the monotone regression problem consists in looking for the scaling x^* of the ordinal variable x that is as close as possible to the numerical variable y. So we are looking for the projection x^* of the vector y of $\Re^n$ into the convex polyhedral cone C of

the scalings of the ordinal variale x. The particular structure of the cone C allows the construction of fast and simple algorithms based on assembling of adiacent violators, as the Pool Adjacent Violators algorithm [3], described successively . Let $\bar{y}(B)$ denote the mean of the variable y restricted to the subjects i such that $x(i)$ belongs to the subset B of M. A block of categories of M is a subset of M formed by consecutive elements. Let $B_0, B_1, ..., B_r$ be a partition of M into increasing blocks: the largest element of B_h is smaller than the smallest element of B_{h+1}.
By using geometrical arguments, it is possible to show that there exists a partition of M into increasing blocks $B_0, B_1, ..., B_r$ such that the optimal scaling δ takes on the value $\delta(j) = \bar{y}(B_h)$ for any category j in B_h.

Proposition 3.7. [3] *There exists a partition of M into increasing blocks $B_0, B_1, ..., B_r$, such that the optimal scaling δ of the categories of x associated with the monotone regression of y on x takes the value $\delta(j) = \bar{y}(B_h)$ for any category j in B_h.*

The search for the optimal partition of M into increasing blocks $B_0, B_1, ..., B_r$ is equivalent to the search for an optimal subset $J = \{1, j1, ..., jr\}$ of M where each jh is the smallest element of B_h. The Pool Adjacent Violators algorithm gives the optimal blocks $B_0, B_1, ..., B_r$ in an iterative way. At the first iteration each category j is a block. We look at the sequence $y(j)$. If $y(1) \prec y(2) \prec .... \prec y(m)$ the optimal solution has been found. Otherwise, beginning at the first category, we pool together the categories which constitute a monotonically non-increasing run, and this gives another set of blocks on which the procedure is iterated. The optimal solution $B_0, B_1, ..., B_r$ is reached as soon as $\bar{y}(B_0) \prec \bar{y}(B_1) \prec \bar{y}(B_r)$.
This algorithm can be interpreted as a backward stepwise multiple regression of teh dependent variable y on the independent variables

[3] The proof is in Apendix A

$z_2, ..., z_m$. At each step, the variables z_j with a non-positive regression coefficient are suppressed. The optimal subset $J = \{1, j1, ..., jr\}$ is obtained as soon as all the regression coefficients of the variables $z_{j1}, ..., z_{jr}$ are strictly positive. This interpretation of the Pool Adjacent Violators algorithm as a backward stepwise regression comes direclty from the fact that the regression coefficients of the variables z_{jh} are $\bar{y}(B_h) - \bar{y}(B_{h-1})$. Pooling B_{h-1} and B_h when $\bar{y}(B_{h-1}) \geq \bar{y}(B_h)$ is equivalent to ejecting the category jh from J and then to removing the variable z_{jh} from the regression.

3.3 Some remarks

This chapter was focused on the description of the ALSOS algorithms, because we consider their approach to the quantification as a good solution for the problem of the estimation of a SEM.

We have seen that the ALSOS algorithm allow to have simultaneously all kind of variables in the data matrix, and are capable to quantify them separately and each variable according to its nature.

This important characteristics is very useful in the context of a SEM in which we have different latent block, measured by different manifest variables, that in this case can be qualitative/ordinal or numerical.

The approach, followed by this class of algorithms, allows to develop any kind of statistical analysis. In particular we have presented two algorithms Morals and Princals that face two different problems: the first develops a regression model with mixed variables, so it has the scope to predict a dependent variable by a set of dependent variable, while the second develops a PCA with the object to obtain the best synthesis of a set of variables.

Both methods have the characteristics to contestualize the quantification procedure into the scope of the analysis to develop, in such a way to obtain a non arbitrary quantification for the qualitative/ordinal variables.

This is the approach used in our algorithm, that we will present in the next chapter, to obtain the best quantification for the qualitative/ordinal variables.

Chapter 4

The internal approach to the quantification

Social and market researches are mainly based on collecting qualitative and ordinal indicators.

When the aim of the study is the estimation of a casual model built on the relationship between latent concepts as in SEM, the use of qualitative indicators causes some problems due to the meaningfulness of the results. This is because this methodology was created to estimate the relationships between quantitative variables.

The problems are stronger when dealing with Covariance based models (as LISREL) in which some distributional a priori hypotheses about the data are assumed and used for inferential scope This is not the case of Variance based models (as PLS-PM) in which there are no distributional hypotheses and inferential tests to confirm the theory developed across the model, and in which the objective is to explore the data and, when some relationships between latent variables are assumed, to confirm these casual relationships.

Both methods, however, introducing in the model ordinal variables, and in particular the PLS-PM accepts also the nominal variables; when the variables are ordinal, LISREL has a different procedure in respect to the numerical case.

4.1 The LISREL approach

As recalled in the previous section, the LISREL algorithm is applicable to *metric* variables, where *metric* indicates a variable with a standard of measurement. So LISREL is not usable with categorical variables, while for the ordinal variables some alternatives exist that allow their use for the estimation of the model.

The problem regards Pearson's correlation coefficient, that is not computable between two qualitative or qualitative and quantitative variables, unless they are considered as numerical variables. In the case of ordinal variables the statistic used is the matrix R, the matrix of observed correlation, and not the matrix S of covariance-variance, so the variance of these variables is fixed equal to 1 [1].

As an alternative many authors,in literature proposed the use of tetracoric/policoric/poliserial correlation when working with ordinal variables , or mixed data. The idea is that an ordinal variable x can be considered as a raw measurement of an underlying continuous latent variable x^* exists, in such a way if the ordinal variable assumes the values between 1 and 4, we imagine that three thresholds points on the latent variable x^*, called α_1, α_2 and α_3 (where $\alpha_1 \prec \alpha_2 \prec \alpha_3$). So, if x^* is $\prec$ of α_1, then $x = 1$; if x^* is included between α_1 and α_2, then $x = 2$, and so on.

The *tetracoric*[2] *correlation coefficient*, introduced by Pearson (1901), is the estimated correlation coefficient of two continuous variables

[1] Very often the matrix R is not used , but the matrix with the linear correlation coefficient of Pearson, making the assumption of the continuity of ordinal variables.

[2] The correlation between a continuous and dichotomic (polytomic) variable is called *biserial* (*polyserial*) correlation coefficient.

distributed as a Normal, underlying two ordinal variables. The estimation of this correlation can be made in two ways:

- the first method [7], [43], [?], [42] is based on the estimation of the thresholds and after of the tetracoric correlation, across an iterative procedure. Given that

  $$F(\alpha_1) = \int_{-\infty}^{\alpha_1} f(x)dx = p_{1.} \quad F(\alpha_2) = \int_{-\infty}^{\alpha_2} f(x)dx = p_{.1} \quad (4.1)$$

  F is the distribution function and f is the density function. So the thresholds are equal to $\alpha_1 = F^{-1}(p_{1.})$ and $\alpha_2 = F^{-1}(p_{.1})$. Having the thresholds (α_1,α_2), using the bivariate distribution (x_1, x_2), we can extract the tetracoric correlation (ρ_{12}), across an iterative procedure that puts the double integral equal to the probability to observe the cell (1,1) estimated by the relative frequencies:

  $$P(x_1 \leq \alpha_1, x_2 \leq \alpha_2) = \int_{-\infty}^{\alpha_1} \int_{-\infty}^{\alpha_2} F(x_1, x_2, \rho_{12}) dx_1 \dots dx_2 = p_{11}$$

- the second method [43] [28], [?] is based on the simultaneous maximization of the ML compared to the thresholds and the tetracoric correlation.

If the observed ordinal variables are two, it is possible to imagine two underlying continuous latent variables x_1^2 and x_2^2, that form a bivariate normal distribution, in such a way that from the observed ordinal values it is possible to compute the correlation coefficient[3], that have to exist between the two latent variables, to have the joint distribution of the observed variables.

The correlation coefficient between x_1^2 and x_2^2, is called *polycoric*

[3] The latent variables have mean 0 and variance 1, so we do not speak of covariance, but correlation.

correlation. The polycoric correlation coefficient is not a correlation coefficient between ordinal variables, but the estimation of the correlation ρ between the metric latent variables x_1^2 and x_2^2. Its estimation is based on the hypothesis of Normality of underlying latent variables, of which it is supposed the standardized normal distribution $(x_1,\ x_2) \sim N(0,1;0,1;\rho_{12})$, to the qualitative indicators. It is possible to use one of the two procedure for the estimation of the tetracoric correlation.

For the estimation of the *polyserial correlation* [44] it is assumed the continuous normal distribution for the continuous variable x^* (underlying the ordinal variable x) and for the observed continuous variable y. The ML function is defined as:

$$L = \prod_{i=1....n} f(y_i, x_i) = \prod_{i=1....n} f(y_i) f(x_i/y_i) \tag{4.2}$$

The methods for the estimation of polyserial correlation are two:

1. the moments of y are estimated from those samples, the thresholds of X^* from the marginal of X; the polyserial correlation coefficient is obtained across the method of ML, conditioned by the other parameters estimated;
2. the simultaneous estimation of all parameters with the ML, maximizing L respect to the unknown parameters:

$$L = log(L) = \sum_i [log f(y_i) + log f(x_i/y_i)] \tag{4.3}$$

where $f(y)$ is the density function $N(\mu, \sigma)$, and $f(x/y) \approx N(\rho x, (1-\rho^2))$, with $x = (y-\mu)/\sigma$. The distribution $f(x/y)$ is obtained from:

$$Prob(x^* = j/y) = F(x_j^*) - F(x_{j-1}^*)$$

where $x_j^* = (x_j - \rho x)/\sqrt{1-\rho^2}$

The method to estimate the model is no more the ML, but the Weighted Least Square (WLS), with a suitable weighting matrix, because in this way we do not obtain bias estimation.

The process of the estimation is composed of two phases:

- in the first phase the aim is the determination of the correlation matrix of input of LISREL, composed of the policoric-poliserial correlation on the raw data (now this matrix could contain not only the Pearson's coefficient, but also the polyserial correlation coefficient), and after it is computed, also, a covariance asymptotic matrix, useful to the calculation of the weighting least square;
- in the second phase the application of LISREL, with the method of estimation WLS, is considered.

In literature some methods also exist that estimate the parameters of the first phase (thresholds, polycoric correlation) and the parameters of the model simultaneously.
In substance there are four different approaches:

1. the first method, composed of three phases, computes at first the thresholds for each couple of ordinal variables, using the marginal distribution of ordinal variables,under the hypothesis that the marginal cumulated probabilities are equal to the relative marginal frequencies:

   $$x_i = F^{-1}(P_{i.})\text{and } x_j = F^{-1}(P_{.j})$$

   where $P_{i.}$ and $P_{.j}$ are the marginal frequences. Separately for each couple of variables, the polycoric correlation, across the ML estimator is estimated:

   $$\frac{\partial l}{\partial \rho} = \sum_{j=1}^{s} \sum_{i=1}^{v} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \rho} = 0 \tag{4.4}$$

where l is the Log likelihood and π_{ij} is the probability that an observation is in the cell (i, j) of the multinomial distribution.
In the third phase the structural parameters of the model are estimated, by a loss function in which the S matrix is composed of the correlation estimated in the second step. Muthen used a slightly different procedure in three steps . His approach is considered more general then the previous because he does not use the hypothesis of normal distribution of the latent variables, and he divides the parameters (thresholds, correlation and structural coefficients) in three sub models.
So in the first step the thresholds are estimated, while in the second the correlations and in the latest step the structural parameters across the Weighted Least Squares (WLS), with the estimation of the asymptotic covariance matrix R.

2. the second method, consist of only one step by estimating the structural parameters of the model (correlation matrix, loadings) with p dichotomous observed variables; the estimation of the parameters is obtained simultaneously by the ML function. The authors suggest to use the GLS function because the ML computationally problem [4].

3. the third method, consists of two phases. In the first it estimates simultaneously the parameters of the matrix $R(x_i, x_j, \rho)$, across the ML, solving a system of $(v-1)+(s-1)+[(p+q)(p+q+1)]/2$ partial derivative respect to ρ and the thresholds x_i and x_j:

$$\frac{\partial l}{\partial \rho} = \sum_{j=1}^{s} \sum_{i=1}^{v} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \pi_{ij}} = 0 \qquad \frac{\partial l}{\partial x_i} = \sum_{j=1}^{s} \sum_{i=1}^{v} \frac{n_{ij}}{\pi_i} \frac{\partial \pi_{ij}}{\partial x_i} = 0$$

$$\frac{\partial l}{\partial x_j} = \sum_{j=1}^{s} \sum_{i=1}^{v} \frac{n_{ij}}{\pi_j} \frac{\partial \pi_{ij}}{\partial x_j} = 0$$

From this system the estimation of the thresholds and of the

[4] It is necessary to compute the product of N integral.

correlation coefficients are obtained. In the second phase it estimates the structural parameters, minimizing the loss function

4. the fourth method is proposed by Lee-Poon-Bentler [33] for the estimation of polyserial and polycoric correlations. This method is characterized by the simultaneous estimation, for all variables, of correlation and thresholds, across the ML. In the second step the structural parameters are estimated minimizing the GLS loss function. The authors show that the estimates are efficient asymptotically and the asymptotic distribution of estimates is multinormal with a covariance matrix equal to the inverse of information matrix.

In presence of ordinal data we can compute the correlation coefficient, besides by linear coefficient of Pearson (r), also by the polycoric coefficient (ρ), the Spearman's coefficient (ϕ) and τ of Kendall. The studies [43],[7] have demonstrated that the Pearson coefficient under estimates the real correlation between the continuous variables underlying the ordinal variables, respect to the polycoric coefficient.

So the polycoric correlation is the best choice for normal variables underlying ordinal variables, especially when the size of sample is large and also the number of categories (7,9). The disavantage is that the polycoric correlation overestimates the standard errors of the estimation and the values of the statistic χ^2, but it produces parameters with the lowest MSE and no bias.

4.2 The PLS-PM approach for ordinal variables

The PLS-PM, as LISREL, was born for the estimation of a SEM model with metric data, but one of its characteristic is the possibility to introduce also the nominal and ordinal variables. The basic idea of the PLS-PM is to assume the continuity for the ordinal variables so it is possible to treat them as numerical variables; instead when the variables are nominal it is used the binary coding: the problem in this case is that the dimension of the matrix X of manifest variables increases.

Let us now image to introduce one or more binary-coded nominal or categorical variables in one or more latent block. The problem lies in the dimension of the measuring scale which has been adopted for those variables, being it not wide enough to allow to assume continuity for those variables.

This fact has an impact on the estimation of the latent concepts. The risk is to have a final model in which qualitative indicators are eliminated.
At present the qualitative variables are introduced, when possible, in the model or with a binary coding, or across an a-priori transformation, as the Thurstone, Rash transformation or it is used the equidistribution linear normalization, before the model estimation.
In literature a proposal, made by E. Jakobowicz and C. Derquenne (2006), introduces an algorithm, so called Partial Maximum Likelihood (PML), based on the Generalized Linear Models (GLM), estimation. They take into account of variables of different nature (numerical, nominal or ordinal), with final aim the quantification of qualitative variables. The analysis then continues by performing the classical PLS -PM algorithm.

They modify the first step of the PLS-PM algorithm, according to the nature of manifest variables (nominal or ordinal). The authors introduce the concept of reference variable as the initial estimation of the latent variable: it is a manifest variable of any latent block associated to the j-th block that is supposed to better explain the latent concept. The vector of the initial weights will be equal to :

$$w_{jh}^{0} = cov(x_{jh}, x_{i1}) \tag{4.5}$$

where x_{i1} is the reference variable chosen between the blocks associated to j block. The authors propose a series of statistical method-

ologies well known in literature, whose differences are related to the nature of the variable x_{jh} and the reference variable x_{i1}.
In particular:

- *if the reference variable is numerical and is adjusted by a numerical variable*: this is the classical PLS-PM
- *if the reference variable is numeric and the manifest variable is categorical*: an ANOVA model with one effect is used.

$$y_j^{t=0} = \sum_{h=1}^{p_j} \sum_{l=1}^{L_h} w_{jhl}^{t=0} x_{jhl} \tag{4.6}$$

where t=0 indicates the first step of the algorithm, x_{jhl} is a dummy variable with dimension equal to the number of categories L_h $(l = 1, ..., L_h)$ in x_{jh}, and $w_{jhl}^{t=0}$ is the mean of x_{i1} on category l

- *if the reference variable is boolean or ordinal and the manifest variable is numeric*:a simple logit model is used.

$$y_j^{t=0} = \sum_{h=1}^{p_j} w_{jhl}^{t=0} x_{jh} \tag{4.7}$$

In this case $w_{jhl}^{t=0}$ is the logistic regression coefficient of x_{jh} on x_{i1}

- *the reference variable is boolean or ordinal and the manifest variable is categorical*: it is used a logit model with one effect (one group)

$$y_j^{t=0} = \sum_{h=1}^{p_j} \sum_{l=1}^{L_h} w_{jhl}^{t=0} x_{jhl} \tag{4.8}$$

where x_{jhl} is a dummy variable and $w^{t=0}_{jhl}$ is the logistic regression coefficient

- *if the reference variable is nominal and the manifest variable is numeric*: a polytomic logistic model is used

$$y_j^{t=0} = \sum_{h=1}^{p_j} w^{t=0}_{jh(r)} x_{jh} \text{ when } x_{i1} \text{ takes a value r} \qquad (4.9)$$

where $w^{t=0}_{jh(r)}$ is the generalized simple regression coefficient of x_{jh} on x_{i1}

- *if the reference variable is nominal and the manifest variable is categorical*: it is used a generalized logit model with one effect

$$y_j^{t=0} = \sum_{h=1}^{p_j} w^{t=0}_{jhl(r)} x_{jhl} \text{ when } x_{i1} \text{ takes a value r} \qquad (4.10)$$

where x_{jhl} is a dummy variable of category l and $w^{t=0}_{jhl(r)}$ is the logistic regression coefficient

The inner estimation is the same as in the classical PLS-PM algorithm, while for the outer estimation it is important to consider the nature of the manifest variables; in particular if the manifest variable x_{jh} is numeric, the algorithm proceeds in the classical way, while if it is categorical, it is used, to obtain a new estimation of the latent variable, the variance model:

$$y_j^t = \sum_{h=1}^{p_j} \sum_{l=1}^{L_h} w^t_{jhl} x_{jhl} \qquad (4.11)$$

where x_{jhl} is the dummy variable of category l and w^t_{jhl} is the mean of z_i (the inner estimation of the latent variable ξ_i).
In this case for each category of the manifest variable is computed a weight, that corresponds to the coefficient of the variance analysis,

and so the global weight associated to the manifest variable is:

$$\hat{w}_{jhl} = \sqrt{\frac{\sum_{l=1}^{L}(\hat{w}_{jhl} - \bar{w}_{jh})^2}{L_h}} \tag{4.12}$$

This approach has the advantage to make the quantification of the qualitative variables by an internal procedure to the classical algorithm of PLS-PM, with respect to the other approaches that make an external quantification; another important characteristic is the possibility to introduce all kind of variables, and for the qualitative ones, each of them quantified according to its nature.
On the other hand we have two problems:

- we left the properties of the algorithm of PLS-PM (absence of distributional hypothesis, the possibility to apply the technique to matrixes with a number of individuals minor than of number of variables
- we do not estimate in the same way the weights of qualitative and quantitative variables (in the case in which we have mixed variables, the weights of numerical variables are computed as the covariance between manifets and latent variables, instead of the case of rodinal or nominal variables the weights are either the means of the values of these variables or the regression coefficient of logistic regression

Another proposal in the literature is of P.G. Lovaglio (2002), that proposes an algorithm, as an alternative to LISREL, for the estimation of the structural parameters of a model. The nature of the observed variables can be nominal, ordinal or numerical: the algorithm computes a regression model in which there are a set of manifest variables X that are explicative, and a set Y of manifest variables that are dependent and that define a latent variable. The algorithm estimates, based on the alternation two steps: the best quantification for the variables X and Y (in the case in which the sets are composed by qualitative variables) and the best estimation of the parameters of the model.

The steps of this algorithm are:

1. for each y_i is adapted q regression model, specifying the quantitative, nominal or ordinal of the variables in each equation. The output is composed by the regression coefficients, the scores of p explicative variables X_i^{os} in each equation ($i = 1.....q$) and the optimal quantification of q dependent variables $y_1^{os}, ..., y_q^{os}$ that coincides with the prediction obtained from the regression $\hat{y}_i, ..., \hat{y}_q$
2. the principal component $\hat{Y} c_1^*$ is estimated, where $\hat{Y}$ has as elements the prediction of the previous step
3. a regression model is estimated, with as dependent variable $\hat{Y} c_1^*$ and as explicative variables the p regressors (qualitative and quantitative); the outputs of this step are the coefficients of regression b and the optimal transformation of X^{os}
4. to obtain the new estimation of $\hat{y}_i^*, ..., \hat{y}_q^*$, $\hat{Y} c_1^*$ is projected on the space of the columns of the indicator matrix associated to each dependent variable, across the Non Linear Principal Component Analysis (NPCA)

   $$\hat{y}_j^* = G_j D_j^{-1} G_j' \hat{Y} c_1^* \tag{4.13}$$

   where G_j is the indicator matrix of the variable j and D_j^{-1} is the is the diagonal matrix containing the univariate marginals of variable j.
5. with the new estimation of $\hat{y}_j^*$ and the quantification of the regressors X_i^{os}, the algorithm returns to the step 1, with a new iteration of the algorithm

The algorithm proposed belongs to the family of Alternating Least Square Optimal Scaling (ALSOS), and in particular it is based on the join between two approaches: one is the Non Linear Regression of the set Y on X to obtain the optimal quantification for both variables, and the second is the Principal Component Analysis to obtain the estimation of the latent variable as the first component of $\hat{Y}'\hat{Y}$. These two methods, that forms the two steps are alternated until the convergence and the results are the estimation of the regression coefficients and the optimal quantification for the qualitative variables. It is obtained the convergence, because at each iteration the residual is smaller than the previous iteration; the aim is the maximization of the redundance index and of the multiple R^2.
The characteristics of the latent variables estimated according to this method are:

- they do not require distributional hypothesis
- they have not indeterminacy on the scores, because they are linear combination
- the simultaneous estimation of the parameters and of the scores of the latent variables, reach a global optimum
- they have the property of the least square because they are estimated by $\hat{Y}'\hat{Y}$
- in presence of categorical variables, they are coherent with the Kruskal approach, because the algorithm estimates simultaneously the optimal scaling parameters and the parameters of the model

4.2.1 Some observations

The two techniques for the estimation of a SEM allow to cope with ordinal variables in the data matrix. The LISREL approach shows that the assumption of the continuity for the ordinal variables, causes an under-estimation of the real relationship between the underlying continuous latent variables, and that the polycoric correlation produces non biased estimation and with a smaller MSE. The proposal made for LISREL is also consolidated and many software programs introduce ordinal variables in the model now. The only drawback is that it is not possible to use nominal variables.

On the other hand the PLS-PM does not take into constraints, but this can cause some problems on the estimation of the model, as the meaning of the parameters associated to a variable, and in particular the nominal variables cause an increase in the dimension of the data matrix.

However, for the PLS-PM some approaches based on an external or internal quantification exist.
The proposal of Jakobowicz and Derquenne is very interesting, but it referes to by methods based on distributional hypothesis, as the ANOVA or the logistic model, that is beyond the PLS-PM characteristics. As consequence the properties of the PLS-PM as absence of distributional hypothesis, the possibility to model flat matrixes with a number of individuals minor than of number of variables, are lost. The weights of qualitative and quantitative variables are estimated in different way (in the case in which we have mixed variables, the weights of numerical variables are computed as the covariance between manifest and latent variables, instead of the case of ordinal or nominal variables the weights are either the means of the values of these variables or the regression coefficient of logistic regression).

The proposal of Lovaglio [36], instead, is based on the use of the Alternating Least Squares (ALS) algorithms that, as the PLS-PM, have the important property of the absence of distributional hypothesis, and an explorative purpose. However, his proposal has as objective to determine an alternative to the LISREL approach to estimate a latent concept, measured by indicators and causes, using the Non Linear Principal Component Analysis (NPCA).The fundamental characteristic of this algorithm is the simultaneous estimation of the vector of scaling and of the parameters of the model, in this case the regression coefficients. So it has the same characteristics of the PLS-PM with the addition of a unique function to optimize with respect the parameters of the model.

This approach allow to estimate a structural model with mixed variables, using the NPCA algorithm to estimate the latent variables.

As seen, it is possible to distinguish between two different approaches for the PLS-PM: the first consists of an external or a-priori quantification, and the second consists of an iterative quantification internal to the algorithm for the estimation of the model, in such a way to compute the optimal quantification and the optimal estimation of the parameters of the model simultaneously.

4.3 PALSOS-PM: a joint between Alternating Least Squares and PLS-PM

The techniques used in the PLS-PM for the treatment of ordinal variables, we have seen that are oriented on an a-priori quantification, across the Tursthone or Rash scales or across an equidistributional normalization, or an iterative quantification, internal to the algorithm for the estimation of the parameters of the model, as the approach of Derquenne and Jackobowicz.

Aim of this chapter is the development of a procedure, that, saving some characteristics of PLS-PM, could optimally quantify the ordinal manifest variables, and be able to estimate the parameters of the model.

The algorithm is called Partial Alternating Least Squares- Path Modeling (PALSOS-PM), because it has the structure of the algorithm of PLS-PM (the split between the outer and inner estimation and the partial analysis) and it uses, as method of estimation for the parameters of the model, an Alternating Least Square algorithm with the Optimal Scaling to obtain the optimal quantification for the qualitative manifest variables. The principal characteristic of this approach is the absence of distributional hypothesis and the possibility to introduce all kind of variables in the model (in particular ordinal variables[5]) that can be quantified according to their nature and their scale of measurement.

So the PALSOS-PM algorithm has some characteristics of PLS-PM and some of ALS algorithms: of the first it has the basic structure (inner and outer estimation and the Path Analysis) and the inner estimation of the latent variables, of the second it has the the process of quantification, modifying the estimation of the outer weights, because it takes into account the nature of the variables.

So the name PALSOS-PM is due to the presence of:

- the Partial analysis
- the Alternating Least Squares algorithm
- the Optimal Scaling process
- the Path Analysis

Thanks to these caharacteristics, the PALSOS-PM algorithm obtains the best coefficients of the model and the best quantification for the variables by the use of Morals algorithm, that computes simultaneously the parameters of a regression model and the optimal scaling vectors for the manifest variables. This procedure is done for each block of latent variables, separately.

So the join of the two methodologies allows to obtain the same results of the PLS-PM, solving the problem of the treatment of ordinal variables, typical data collected for the customer researches. The algorithm could start or with an arbitrary quantification of the manifest variables (the typical coding of a questionnaire), or with a

[5] This algorithm can estimate a SEM also in presence of nominal variables, but the aim of this thesis is to show the improvements of the quantification for the estimation of structural parameters when the variables are ordinal, so we do not investigate the nominal case.

quantification obtained by Princals algorithm that develops a Principal Component Analysis (PCA). In the latter case we have a first estimation of the latent variable not arbitrary and in line with the classical algorithm of PLS-PM.

The algorithm proceeds with the inner estimation of latent variables, and when returns to the external estimation, it uses Morals to update the outer estimation. We have choosen to use this iterative algorithm for three reasons: the first is the possibility to estimate the relationship between variables in the reflective and formative case; the second is its capability to treat simultaneously different kinds of variables (numerical, nominal or ordinal), because the quantification step is individually; the third is the simultaneous estimation of the relationship between the manifest and latent variables and the best quantification. In synthesis the steps of PALSOS-PM algorithm are:

Algorithm 4. *PALSOS-PM.*

```
Consider the matrix $X$ of manifest variables and define the path diagram
Step1
Compute a first casual vector of weights $w_{ij}$
repeat
Step2
    for (j in 1:k)
        $v_j = \sum_{i=1}^{p} w_{ij} x_{ij}$
        or Princals      endfor
Step3
    for (j in 1:k)
        $z_j = \sum_{j=1}^{k} e_{jj'} v_j$
    endfor
Step4
Update the estimation of weights $w_{ij}$ by Morals
Consider the matrix $X$ and $Z$ of raw data as $X^*$ and $Z^*$
repeat
```

```
Step4.1
    for (i in 1:p)
        for (j in 1:k)
            $\beta = (X^{*\prime}X^{*})^{-1}X^{*}Z^{*}$
        endfor
    endfor
Step4.2
    for (j in 1:k)
        $\hat{Z} = X^{*}\beta$
        $Z^{G} = G_{Z}(G'_{Z}G_{Z})^{-1}G'_{Z}\hat{Z}$
        $Z^{*} = Z^{G}(\frac{||\hat{Z}||}{||Z^{G}||})$
    endfor
Step4.3
    for (i in 1:p )
        $\hat{X}_{I} = \frac{1}{\beta_{J}}(Q^{*} - \sum_{I \neq J} \beta_{J}X^{*}_{J})$
        $x^{G} = G_{I}(G'_{I}G_{I})^{-1}G'_{I}\hat{X}$
        $X^{*}_{I} = X^{G}_{I}(\frac{||\hat{X}_{I}||}{||X^{G}_{I}||})$
    endfor
Step.4.4
Ceck the convergence
    if (the the $R^2$ is not improved 'enough' from last
        iteration) break
Step5
Ceck the convergence
    $|w^{old}_{ij} - w^{new}_{ij}| \prec 10^{-5}$ break
```

When the algorithm returns to the fifth step, a new quantification is obtained by Morals. The PALSOS-PM algorithm besides to estimate a SEM model with ordinal or qualitative variables, allows to estimte a model with all quantitative variables: in this case it computes the parameters by the classical PLS-PM algorithm.

The use of Morals with respect to other methods of quantification has the advantage to obtain simultaneously the best quantification and the best estimation of the relations between the manifest and latent variables.

Another advantage to use Morals is related to the possibility of estimating the model with both reflective and formative manifest variables. In fact the Morals algorithm allows to estimate also the parameters of a multiple regression model.

For the validation of the outer and inner model, we use the bootstrap technique to create suitable the interval confidence, in fact the quantification procedure we use: it not rest on any distributional assumptions. Therefore, information about the variability of the parameter estimates and hence their significance has to been generated by means of resampling procedures.

PALSOS-PM takes into acocunt, during the estimation of the parameters, the problem of the signs and as in the PLS-PM it solves it using the comparison of the signs of the eigenvectors.

4.4 PALSOS-PM: an application on a known dataset (ECSI)

The PALSOS algorithm is here applied, for comparative aims, to a dataset used in the work of Tenenhaus et al. [55], in which they estimate an ECSI model to evaluate the customer satisfaction.

The European Costumer Satisfaction Index (ECSI) is an economic indicator that measures customer satisfaction. A model has been derived specifically for the ECSI. In this model, seven interrelated latent variables are introduced. It is based on well-established theories and approaches in customer behavior and it is to be applicable for a number of different industries. ECSI is an adaptation of the Swedish customer satisfaction barometer (Fornell, 1992) and is compatible with the American customer satisfaction index.

The entire model is important for determining the main target variable, being Customer Satisfaction Index.

The ECSI model is described in figure 4.1. A set of manifest variables is associated with each of the Latent Variables.

This model is applied to a sample of 250 customers of a mobile

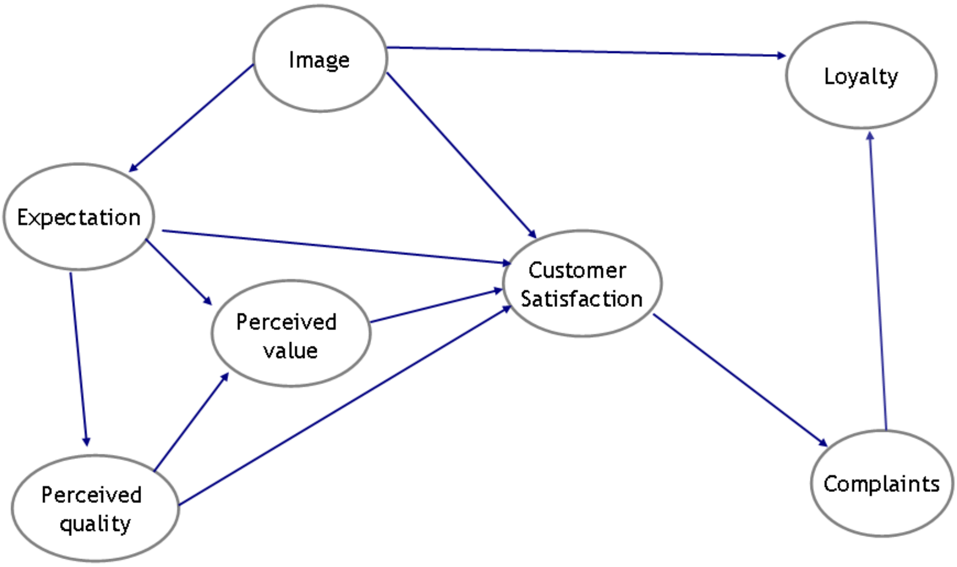


Figure 4.1: The ECSI model

society, to evaluate their satisfaction respect to the services received. The manifest variables are 24 and are so subdivided in the latent blocks:

- five manifest variables for the block *Image*
- three manifest variables for the block *Expectation*
- seven manifest variables for the block *Perceived quality*
- two manifest variables for the block *Perceived value*
- three manifest variables for the block *Customer Satisfaction*
- one manifest variable for the block *Complaints*
- three manifest variables for the block *Loyalty*

All variables are ordinal and express on a scale of ten values, so the PALSOS-PM algorithm is used to obtain an optimal quantification for these variables and to estimate the structural parameters of the model.

The same model was estimated with the PLS-PM algorithm by the software XLSTAT, and with the ML approach by the LISREL software. LISREL estimates an unidentified model, and it suggests to increase the number of fixed parameters in the model; besides the model has a GFI value very low (0.236), so the model supposed is not confirm by the LISREL approach.

As seen in the previous paragraph, the PALSOS-PM algorithm applies a monotone regression to obtain the optimal scaling for these variables. PALSOS-PM algorithm uses some indexes, to validate the model, as in PLS-PM. A first result regards the unidimensionality of the latent blocks: this property is used to understand if the manifest variables really explain the latent concepts to which they are associeted. This is verified by Cronbach's α, ρ of Dillon-Goldstein and the first eigenvalue of a Principal Component Analysis. For both methods this property is verified, being the values of ρ and of first eigenvalue major of 0.7.

However there is a substantial difference: the values of ρ of Dillon-Goldstein, computed with the PALSOS-PM, are bigger than of PLS-PM. This is due to the quantiifcation process that creates mani-

M.vs	Original value	Mean B.	Std. Err	T-Statistic	L. Bound	U. Bound
imag1	0,235	0,237	0,027	8,695	0,182	0,291
imag2	0,238	0,223	0,032	7,446	0,141	0,271
imag3	0,183	0,226	0,035	5,202	0,130	0,299
imag4	0,226	0,242	0,029	7,875	0,186	0,310
imag5	0,241	0,231	0,027	8,786	0,177	0,283
expe1	0,484	0,387	0,042	11,544	0,282	0,473
expe2	0,442	0,382	0,047	9,445	0,285	0,482
expe3	0,356	0,382	0,044	8,143	0,296	0,476
qual1	0,168	0,163	0,020	8,484	0,112	0,204
qual2	0,156	0,154	0,024	6,428	0,095	0,200
qual3	0,187	0,162	0,021	8,753	0,122	0,206
qual4	0,152	0,163	0,016	9,366	0,122	0,199
qual5	0,157	0,161	0,019	8,441	0,120	0,194
qual6	0,167	0,165	0,014	11,659	0,141	0,196
qual7	0,165	0,163	0,018	8,995	0,124	0,204
val1	0,525	0,529	0,029	18,317	0,503	0,618
val2	0,525	0,529	0,029	18,231	0,503	0,617
sat1	0,365	0,361	0,038	9,601	0,308	0,451
sat2	0,468	0,373	0,036	12,869	0,336	0,472
sat3	0,383	0,370	0,038	10,169	0,326	0,464
comp1	1,000	1,000	0,000	2,004	1,000	1,000
loy1	0,357	0,395	0,042	8,519	0,341	0,506
loy2	0,350	0,358	0,044	7,998	0,281	0,455
loy3	0,387	0,391	0,037	10,459	0,331	0,483

Table 4.1: The bootstrap estimation of weights

fest variables maximally correlated with the latent variables. So the unique change obtained at this moment is the improvement of the relationships between the latent variables and the manifets variables.

The two successive tables (4.4 and 4.4) report the results of the outer estimation of weights and loadings, obtained across the PALSOS-PM algorithm.

In each table the weight estimated on the original sample, the mean of bootstrap replications, the Standard Error of the bootstrap replications, the T-Statistic, computed as the ratio between the original value of the parameters and the Standard Error of bootstrap replication, and the interval for the parameters, built considering the value at 0.025 and the value at 0.975 percentile, are reported.

Concerning the results of weights, we can see that the intervals are all positive, however, and significatively different from zero. This consideration is confirmed also by the values of the T-Statistic that

M.vs	Original value	Mean B.	Std. Err	T-Statistic	L. Bound	U. Bound
imag1	0,923	0,872	0,083	11,185	0,636	0,978
imag2	0,934	0,826	0,131	7,148	0,451	0,971
imag3	0,720	0,836	0,134	5,358	0,379	0,967
imag4	0,887	0,889	0,074	12,038	0,644	0,976
imag5	0,946	0,851	0,098	9,650	0,563	0,963
expe1	0,870	0,872	0,095	9,162	0,570	0,971
expe2	0,795	0,859	0,102	7,767	0,572	0,972
expe3	0,639	0,859	0,089	7,167	0,608	0,964
qual1	0,882	0,884	0,097	9,071	0,534	0,972
qual2	0,819	0,835	0,134	6,113	0,444	0,970
qual3	0,982	0,877	0,107	9,202	0,629	0,977
qual4	0,800	0,883	0,097	8,269	0,562	0,972
qual5	0,828	0,873	0,101	8,208	0,597	0,970
qual6	0,878	0,895	0,072	12,136	0,776	0,976
qual7	0,867	0,881	0,078	11,105	0,692	0,986
val1	0,952	0,949	0,046	20,752	0,810	0,994
val2	0,952	0,949	0,046	20,806	0,809	0,994
sat1	0,732	0,889	0,094	7,774	0,612	0,976
sat2	0,938	0,915	0,075	12,462	0,661	0,988
sat3	0,767	0,908	0,082	9,380	0,704	0,987
comp1	1,000	1,000	0,000	1,91E+16	1,000	1,000
loy1	0,893	0,898	0,073	12,185	0,709	0,976
loy2	0,876	0,819	0,111	7,872	0,535	0,954
loy3	0,969	0,889	0,075	12,975	0,659	0,973

Table 4.2: The bootstrap estimation of loadings

are high and all positive .

The loadings, that are the correlation between the manifest and latent variables, also are significatively different from zero. The concordance of the signs between the loadings and weights shows that all manifest variables are expression of the latent concept to which are associated.

We have the same results also with the classical algorithm of PLS-PM, except for the manifest variable *CUSL2* that has a negative interval whether for the weights or for the loadings and a low value of the T-Statistic for both quantities.

Furthermore another aspect to be consider is the values assumed by the loadings: with the PALSOS-PM algorithm the relationships between the manifest variables and latent variables are improved.

The results of the inner estimation

The inner estimation obtained with PALSOS-PM shows a significative improvement in the values of R^2 of regressions, respect to those of PLS-PM. The table 4.4 reports the results for the parameters of the model, in particular the path coefficients computed on the original sample, the mean computed on the bootstrap replication, the Standard Error, the t-statistic of the regression and the interval confidence, built with the percentile (as before for weights and loadings). We can see that three parameters assumes a negative sign for the coefficient: **Expectation** on **Perceived value**, **Perceveid quality** on **Customer Satisfaction**, and **Image** on **Loyalty**. The signs of these path coefficients are different from the signs of the correlation between the latent variables, as we can see from the table. This is due to the fact that the correlation is a simple coefficient only between two variables, while the path coefficient is a partial coefficient of regression, in which is present also the interaction with the other variables in the model.
So the negative interval, that are not centered on the mean value, are justified by the fact that for some samples the relationship between the variables are negative, and for other positive, even if the frequence of the negative values is very low. So if we consider the value of T-statistic and of p-value we can conclude that are significant the following path coefficients:

1. the *Image* on the **Expectation** (the value of T-Statistic is 14.94, with a p-value $2e^{-16}$)
2. the *Expectation* on the **Perceived quality** (the value of T-Statistic is 14.26, with a p-value $2e^{-16}$)
3. the *Perceived quality* on the **Perceived value** (the value of T-Statistic is 20.682, with a p-value $2e^{-16}$)
4. the *Expectation* on the **Customer Satisfaction** (the value of T-Statistic is 13.190, with a p-value $2e^{-16}$)

5. the *Perceived value* on the **Customer Satisfaction** (the value of T-Statistic is 11.941, with a p-value $2e^{-16}$)

In the tables 4.4and 4.4 the results of the inner estimation are reported, allowing to compare the PALSOS-PM with those of the classical PLS-PM. The relationships that are not significant are:

1. the **Expectation** on the **Perceived value** : this latent variable has a low value of T-Statistic, with associated an high p-value. We have the same result also in the PLS-PM
2. the **Perceived quality** on the **Customer Satisfaction**: this path coefficient has not an important impact on the satisfaction; maybe its influence is mediated by the latent variable **Perceived value**, that, instead, has a good impact on the **Customer Satisfaction**.
   Besides, the correlation between the two latent variables **Perceived quality** and **Perceived quality** is high (0.871)
3. the **Image** on the **Customer Satisfaction**: also this relationship is not significant, and also in this case a multicollinearity problem can exist (the variable **Image** has a high correlation with the variable **Expectation**, that also impacts on the **Customer Satisfaction**)
4. the **Image** on the **Loyalty**: this relationship is not significant for the low value of T-Statistic and for the high p-value

Respect to the PLS-PM we can see that PALSOS-PM underlines the problem of multicollinearity between some variables; this effect produces some not significant parameters. It is important to investigate on the real nature of the relationships between the variables (mediator, mediation or direct effect), across other methods as a Partial Least Squares-Regression (PLS).

The PLS Regression is usable in three specific cases in which an OLS regression fails:

Endogenous lvs	Exogenous lvs	Original	Mean boot.	Std.error	T-Statistics	P-value	L. bound	U. bound
Expectation	Image	0,688	0,862	0,074	14,943	0,000	0,672	0,955
P. quality	Expectation	0,671	0,864	0,071	14,262	0,000	0,670	0,964
P. value	Expectation	-0,013	0,297	0,465	-0,312	0,755	-0,525	1,004
	P. quality	0,878	0,562	0,469	20,682	0,000	-0,340	1,315
C. satisfaction	Image	0,074	0,205	0,434	0,611	0,542	-0,804	1,120
	Expectation	0,423	0,130	0,320	13,190	0,000	-0,478	0,691
	P. quality	-0,134	0,467	0,452	-1,375	0,170	-0,350	1,404
	P. value	0,677	0,161	0,291	11,941	0,000	-0,469	0,750
Complaints	C. satisfaction	0,642	0,751	0,146	13,199	0,000	0,398	0,939
Loyalty	C. satisfaction	0,621	0,378	0,409	12,618	0,000	-0,575	1,250
	Image	-0,084	0,487	0,422	-1,772	0,078	-0,466	1,449
	Complaints	0,469	0,074	0,299	14,422	0,000	-0,530	0,616

Table 4.3: Results of inner estimation with PALSOS-PM

Endogenous lvs	Exogenous lvs	Original	Mean boot.	Std.error	T-Statistics	Pr	L. b.	U. b.
Expectation	Image	0,505	0.503	0,052	9,206	0,000	0.382	0,617
P. quality	Expectation	0,557	0,555	0,057	10,568	0,000	0,394	0,646
P. value	Expectation	0.051	0,056	0,088	0,819	0,414	-0,089	0,234
	P. quality	0,557	0,557	0,099	8,982	0,000	0,325	0,730
C. satisfaction	Image	0,179	0,175	0,052	3,214	0,001	0,065	0,299
	Expectation	0,064	0,068	0,044	1,462	0,145	-0,020	0,182
	P. quality	0.513	0,510	0,071	8,413	0,000	0,366	0,651
	P. value	0,192	0,199	0,054	4,266	0,000	0,071	0,308
Complaints	C. satisfaction	0,526	0,529	0,053	9,742	0,000	0,392	0,628
Loyalty	C. satisfaction	0,195	0,209	0,077	2,951	0,003	0,034	0,345
	Image	0.483	0,478	0,071	7,056	0,000	0,321	0,602
	Complaints	0,071	0,073	0,060	1,269	0,206	-0,040	0,199

Table 4.4: Results of inner estimation for PLS-PM (XLSTAT)

Latent variable	Coefficient	Std. Err.	Lower bound	Upper bound
Image	0,254	0,050	0,156	0,351
Expectation	0,231	0,009	0,212	0,249
P. quality	0,242	0,058	0,128	0,355
P. value	0,260	0,051	0,159	0,361

Table 4.5: Results of PLS

- when we have a matrix with more variables than observations
- when there is multicollinearity between the variables (this is our case)
- when there are missing values

In all these cases the PLS is desiderable to use, because the algorithm develops simple regressions, solving the problem of the dimension of the matrix and the problem of multicollinearity, and charges the missing values as belonging to the regression line. It allows to use all variables and to estimate the regression coefficients.

The objective of this approach is to find a number of orthogonal components, maximaly correlated with the dependent variables (prediction) and the most explicative (sinthesis) of the group; then the algorithm proceeds with the regression of the dependent variables on the components. So the PLS is a compromise between an OLS multiple regression and a Principal Component Analysis.

A PLS is used here to investigate if there is a problem of multicollinearity between the manifets variables. In the table 4.4 the estimation of parameters are reported.

This table conteins the regression coefficients of the four latent variables supposed to have an impact on the **Customer Satisfaction**: we have besides the values of the coefficients, also the interval confidence estimated with a bootstrap and the standard deviation. We can see taht using the PLS we obtain positive signs for the path coefficients, and in particular we can note as the impact is similar for each variable of the multiple regression.

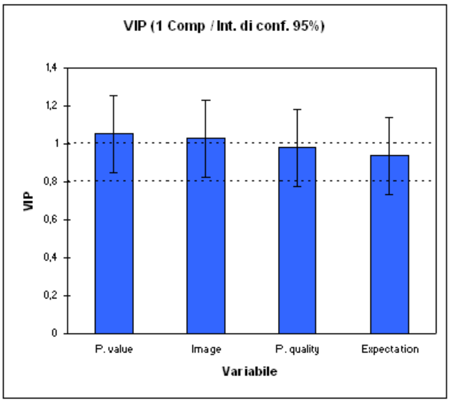


Figure 4.2: The VIP index

The VIP index graphic 4.2 confirm that all variables are important for the definition of the Customer Satisfaction Index, namely **Percevied value** and **Image**. The R^2 of this regression has an high value: 0.817. In the table 4.4 the validation indexes are reported; in particular we have the R^2 of the regressions, the Average Communality and Redundancy, and the Gof index.

We can see as the quantification has produced an improvement in the values of R^2: the variables are strictly correlated and the latent variables are the best obtainable from a given set of manifest variables (the maximization of the correlation coefficients). This improvement is reflected also in the computation of the Gof index, that, as we have seen in the chapter 2, depends from the R^2 and Communality.

in particular we have a significative improvement for the latent block *Image*, for which the value passes from 0.48 to 0.75; for the latent block *Loyalty*, for which the value passes from 0.52 to 0.83. For these latent variables the quantification has producted a significative

Latent variable	R2	A. communality	A. redundancy	Gof index
Image		0,7850		
Expectation	0,4738	0,5988	0,2701	
P.quality	0,4506	0,7514	0,5673	
P.value	0,7554	0,9055	0,7887	
C.Satisfaction	0,8170	0,6679	0,2758	
Complaints	0,4125	1,0000	0,8500	
Loyalty	0,8505	0,8343	0,6378	
				0,7095

Table 4.6: Validation indexes for PALSOS-PM

Latent variable	R2	A. communality	A. redundancy	Gof index
Image		0,4780		
Expectation	0,2550	0,4800	0,1220	
P.quality	0,3110	0,5770	0,1790	
P.value	0,3450	0,8490	0,2920	
C.Satisfaction	0,6800	0,6930	0,4720	
Complaints	0,2770	1,0000	0,2770	
Loyalty	0,4570	0,5170	0,2380	
				0,4710

Table 4.7: Validation indexes for PLS-PM (XLSTAT)

improvement in their definition[6].

Concerning the value of the redundance index, its values for each block, but for the block *C. Satisfaction*, are higher than the one of the model estimated with PLS-PM: the manifest variables and the exogenous latent variables are able to explain more variability of the manifest variables of endogenous latent blocks.

4.5 A review of the programs software for SEM estimation

When it comes to modeling relationships between latent variables, mainly two different methodological approaches can be distinguished: Covariance structure analysis on the one hand and PLS path model-

[6] Remember that a latent variable is obtained as linear combination of its manifest variables with the weights estimation.

ing (not to be confused with PLS regression) on the other. Although both methods emerged roughly at the same time, their development took a rather diverse course. Since the introduction of the first LISREL version in the early 1970s, the software available for covariance structure analysis has experienced substantial progress with respect to ease-of-use and methodological capabilities. Graphical interfaces in programs like AMOS or LISREL have freed the user from having to specify his/her model in matrix or equation form.

PLS path modeling has,until recently, rarely been applied in marketing although its basic algorithms were developed in the 1970s and the first software packages were publicly available in the 1980s (LVPLS [36], PLSPath [49]).Currently, researchers can choose between several alternative software solutions (PLS-GUI, VisualPLS, PLS-Graph, SmartPLS, SPAD-PLS, XLSTAT-PM,PLSPM package of R) which provide a clear improvement especially in terms of user-friendliness.

Against the background of a growing number of PLS software packages and an increasing differentiation in the programs' capabilities, a comprehensive review would help researchers to decide on the specific PLS program to be used in their studies.

In contrast to the former software, these programs are more or less self-contained implementations of the algorithms developed in [62], [63] and [37].It should be noted, that all programs (except LVPLS) are constantly under development and can therefore be expected to offer additional features in the future.

1. the DOS-based program LVPLS 1.8 [37] includes two different modules for estimating path models. Whereas LVPLSC analyzes the covariance matrix of the observed variables, the LVPLSX module is able to process raw data. In order to specify the input file an external editor is necessary. The input specification requires that the program parameters are defined at specific positions in the file - a format which resembles punchcards. Results are reported in a plain text file. The program offers

blindfolding and jackknifing as resampling methods in case raw data has been analyzed. When analyzing covariance/correlation matrices, resampling techniques cannot be applied.

2. the Windows-based PLS-GUI [35] provides a graphical interface for LVPLS which supports both the analysis of raw data (LVPLSX) as well as covariance information (LVPLSC). To specify a model,the user is led through a stepwise procedure which offers a menu at each step. Additional options (e.g., weighting schemes, missing data code) are to be chosen in a separate window. The program finally creates an input file which is processed by the executable file pls.exe of LVPLS. If required, the input file can be modified by the user. The output is the same as for LVPLS. The current version offers a bootstrap option as an additional feature not provided by LVPLS.

3. VisualPLS [17] is a graphical user interface for LVPLS running in the Windows environment which enables the analysis of raw data only. The path model is specified by drawing the latent variables and by assigning the indicators in a pop-up window. Based on the graphical model, the program produces a separate LVPLS input file, which is run by LVPLSX (pls.exe). Different formats of input data are supported. The results are offered as LVPLS output (plain text file) as well as in HTML/Excel format. In addition, a path model showing the estimated parameters is displayed. Beyond blindfolding and jacknifing, bootstrapping has been integrated. Special support for specifying moderating effects and second order factors is offered.

4. PLS-Graph [8] is a Windows-based program which uses modified routines of LVPLS, but only processes raw data (LVPLSX). In order to specify the model, a graphical interface can be used which provides some tools for drawing a path diagram. Different options (e. g., weighting scheme, resampling method) can be chosen from a menu. Although the generated input file is a text file, it can only be processed by PLS-Graph, but not by

LVPLS. Estimation results are presented in ASCII format as well as in a graphical path model; resampling methods include blindfolding, jackknifing, and bootstrapping. SPAD-PLS: This program is part of the comprehensive data analysis software SPAD (running under Windows) which is offered by the French company Test and Go.
SPAD-PLS [51] does not process covariance information but needs raw data instead. Models can be specified with a menu or graphically in a Java applet; the remaining settings may be adjusted in additional menu windows. Different options for handling missing data and multicollinearity are provided. Results are reported both as a path diagram and as text or Excel file; blindfolding, jackknifing, and bootstrapping (including confidence intervals) are available. In the non-graphical manual mode transformations of latent variables (squares, cross-products) can be specified.

5. SmartPLS [46] is a Java software-based. It is independent from the user's operating system. Again, only raw data can be analyzed. The model is specified by drawing the structural model for the latent variables and by assigning the indicators to the latent variables via "drag and drop". The output is provided in HTML, Excel or Latex format, as well as a parameterized path model. Bootstrapping and blindfolding are the resampling methods available. Like in VisualPLS, the specification of interaction effects is supported. A special feature of SmartPLS is the finite mixture routine (FIMIX). Such an option might be of interest if unobserved heterogeneity is expected in the data [39].

6. XLSTAT-PLSPM: The XLSTAT add-in offers a wide variety of functions to enhance the analytical capabilities of Excel, making it the ideal tool for your everyday data analysis and statistics requirements. XLSTAT is compatible with all Excel versions from version 97 to version 2007, and is compatible with

the Windows 9x till Windows Vista systems, as well as with the PowerPC and Intel based Mac systems.The use of Excel as an interface makes XLSTAT a user-friendly and highly efficient software. XLSTAT-PLSPM, that is a module of XLSTAT, implements all methodological features and most recent findings of the PLEASURE (Partial LEAst Squares strUctural Relationship Estimation) technology. This technology has been originally developed as a research tool at the academic level by Y.M. Chatelin and V. Esposito Vinzi in co-operation with C. Lauro and M. Tenenhaus. Thanks to an intuitive and flexible interface, XLSTAT-PLSPM allows to build the graphical representation of the model, then to fit the model, display the results in Excel either as tables or graphical views.

7. PLSPM is a package of the open source language R, more used by statistician researchers. It is published by G. Sanchez in July 2009 and allows to estimate a SEM model with the algorithm of PLS-PM. Respect to the other softwares it does not have a graphical interface, so it is necessary to write the commands, that define the model, in the prompt of R; for the same reason (the absence of a graphical interface) it is not possible to see the path diagram and the results are printed in tables.

Data sets where at least some values of their variables are missing are ubiquitious in empirical research. In order to deal with missing data, several alternative approaches have been proposed. LVPLS offers a specific treatment in the case of missing data which combines mean value imputation and pairwise deletion in the course of the estimation (Lohm¨oller (1984); for a more comprehensive description see Tenenhaus et al. (2005)). This missing data treatment is also provided by the graphical interfaces (PLS-GUI, VisualPLS) as well as by PLS-Graph and SPAD-PLS. In contrast, SmartPLS offers two options equivalent to some data pre-processing which either substitute the mean over all available cases of a variable for the missing values or which delete those cases with missing data (casewise deletion).

Since casewise deletion throws away a lot of useful information and thus leads to lower efficiency, this procedure is not to be recommended. XLSTAT proposes some different solutions, as to use the NIPALS algorithm, to ignore the problem, to eliminate the observation with missing values, to substitute the missing values with the mean or median.

In the package PLSPM of R no proposal is made for the treatment of missing values.

Multi-collinearity can be a problem both for the estimation of indicator weights in the case of formative constructs (mode B) and for the estimation of the relationships among latent variables. SPAD-PLS, XLSTAT and the package PLSPM of R at present are the only programs which address the problem of multi-collinearity by providing a PLS regression routine for estimating weights (Mode PLS) and path coefficients (PLS regression instead of OLS regression).

PLS regression searches for a set of components which decompose the vector y of the endogenous variable and the matrix X of explanatory variables in such a way that the explained covariance between y and X is maximized. Whereas specifying path models in LVPLS is rather inconvenient, all recent programs have made a huge step with respect to ease-of-use, reaching now the same level as the software used in covariance structure analysis. One main methodological improvement is the bootstrap procedure for assessing the significance of parameter estimates, which is now implemented in all software packages and supplements the blindfolding and jacknifing resampling routines of LVPLS. A specific strength of SPAD-PLS, XLSTAT-PLSPM and the R package is the estimation of bootstrap confidence intervals for the parameters. Model validation is another important aspect; although some measures like the goodness-of-fit index [55] are implemented in the recent softwares as SPAD-PLS, XLSTAT-PLSPM and R package. Multi-collinearity is a problem both for the estimation of weights in the case of formative constructs and the estimation path coefficients. To cure this problem, SPAD-PLS,XLSTAT-PLSPM and the R package have implemented a PLS regression routine.

Another important characteristic is the possibility to use qualitative data in the model: only XLSTAT-PLSPM allows to estimate a model with qualitative data, across theri transformation in a binary coded.

Respect to the software the PALSOS-PM algorithm has the disavantage that it does not have a graphical interface, but it is necessary, as for the package of Sanchez, to introduce in the prompt of R the inner model and the division for blocks of the data matrix. On the other hand it allows to estimate the PLSP-PM with numerical variables, as the other softwares presented, to estimate the model with mixed variables, to controll the sign changes and to estimate the model in presence of multicollinearity, across the use of PLS Regression. The problem of missing values is solved across the substitution with the mean value.

4.6 Remarks

In this chapter we have presented a new methodological proposal to estimate a SEM model with ordinal variables, and in general with mixed variables (numerical, nominal and ordinal).

With respect to the other proposals made in the literature, the PALSOS-PM algorithm has the important characteristic to estimate simultaneously the parameters of the model and the vectors of optimal scaling for the qualitative variables, quantifying each variable according to its nature. As the PLS-PM also PALSOS-PM does not have the distributional hypothesis on the data, so it uses the resampling techniques to obtain the empirical distribution. It uses all validation indexes of the PLS-PM to verify the correctness of the supposed model.

PALSOS-PM can be considered as a real alternative to the other programs because:

- it estimates the model with the classical PLS-PM when all variables are numerical;

- it estimates the model with mixed variables;
- for the validation it uses the bootstrap technique (the number of resampling is choose by the researcher);
- it has no problem for the dimension of data matrix;
- it does the control on the signs changes;
- it gives the possibility to develops a PLS if the OLS fails;

In this chapter two important topics are described, that characterizy the Morals algorihm: the projection in a convex cone and the monotone regression,of which the most important properties and theorems are enunciated. To demostrate the importance of the quantification process, this algorithm is tested on a known dataset (the dataset mobile): the results have showed that the process of quantification causes an increase of the values of correlations between the manifest variables and latent variable, but also between the latent variables, being these optimally defined by their manifets variables.

Chapter 5

A model for policy impact analysis: the case study AVSI

A characteristic of social research is the observation of qualitative characteristics on which is not possible to apply the quantitative statistical methods. When the aim is the estimation of casual relationships between latent variables that are measured by qualitative indicators, it is necessary to procede with a quantification of these variables, according the methods described in the previous chapters. This is the case of AVSI research in which the variables are nominal and ordinal. It was defined a SEM model with the objective to measure the impacts of the AVSI program on the status of a child, regarding scholar performance, nutrition, health and social relationships.

In the next sections will be present the model and the results obtained with the PALSOS-PM algorithm.

5.1 The AVSI association

AVSI Foundation is an international, non-profit and non-governmental organization (NGO) founded in Italy in 1972. AVSI has programs in over 40 countries in Africa, Latin America, Eastern Europe, the Middle East, and Asia. AVSI has implemented several programs in education, healthcare, construction, emergency response, water and sanitation, food and nutrition, and psychosocial support for children, adults and even the elderly persons in the community.

AVSI has over 15 years of child support programming, starting with the distance support program (DSP) that developed as a form of specific help directed to an individual child within an AVSI program in developing countries of Africa and beyond. By 2005, AVSI began implementation of a more integrated and comprehensive five year project for Orphans and Vulnerable Children (OVC) in the Great Lakes Region of Uganda, Rwanda, Kenya and recently extended to Ivory Coast with financial support from PEPFAR[1] complemented with private funds from AVSI.

The project that will end in June 2010 supports over 20,000 OVC and their family members through a two-pronged approach; directly support is supplemented with indirect support provided through partners embedded in the community.

AVSI's methodological approach is focused on the centrality of the person: the human being is not reduced to the condition of need, regardless of how great that situation of need might be, but is considered in his/her holistic dignity. AVSI seeks to accompany the individual along a path of self-awareness in order to help him maximize his potential.

This accompaniment requires the figure of an adult who is able to follow the child, to identify his needs and resources and to engage the parents or caretaker. For this reason, AVSI intervenes both directly

[1] PEPFAR is the U.S. President's Emergency Program for AIDS Relief, launched in 2004 by Pres. George W. Bush.

and through local partners. In both cases, trained social workers begin looking at the household level to identify the children and families in greatest need while putting an emphasis not only on the needs but also on the resources.

From the onset, AVSI seeks to ensure that there is active family participation so that the solutions originate from the families and so that external support can be tailored to the individual child and family or community needs, but in way that assists rather than replaces the family and community responsibility. External support should remain only part of the answer.

The breadth of services delivered by AVSI or through local partner organizations includes materials and resources needed for school attendance, inputs to improve the quality of schooling, access to health care, nutritional and psychosocial support including recreational activities, and shelter as needed. In addition, AVSI provides food assistance and economic empowerment support including small investments in income generating activities in order to strengthen the family capacity to care for all their children.

5.2 A model for the impact evaluation of AVSI intervention

The database considered here is relative to an AVSI survey of 2007 done in three Countries of Africa (Uganda, Rwanda and Kenya) to evaluate the services given to the children of program.

Further to the explorative analysis that describes the condition of life in which children live in the three countries, we have built a structural equation model that simultaneously estimates multidimensional concepts, and the path coefficients, that allow to evaluate the power of the relationship between the relative latent variables.

The objective of a model is to extract useful knowledge and to provide valid tool for the decision maker for the improvement of the services supplied; it is no more a simple description of the status of the child, but also the estimation of the effects due to AVSI intervention or to external factors. As consequence of the child status and AVSI intervention the Guardian satisfactionalso evaluates.

The model is build on the basis of a survey-questionnaire where the variables are measured on different scale: some variables are dichotomous, other are expressed on scale of three values, other on scale of five points.

The questionnaire is submitted to the Guardian of the child, that is the person that has his legal custody; the Guardian has to evaluate the serivices received, taking into account the needs of the child.

The model has as central variable the actual status of child, measured across some variables that compare the status before and after the beginning of the program. The model has as outcome the satisfaction of Guardian, so the objective is to study the condition and the health of the child and as reflex the guardian satisfaction for the services received.

In this model two important aspects are evaluated: the **Status of child** that impacts on the **Guardian satisfaction**, that depends by **AVSI intervention**, **Family environment** and **Housing condition**, where the **AVSI intervention** is a multiblock, i.e. it derives from the union of **Support for the school**, **Nutritional support** and **Support for family**.

Finally the **Status of child** of child is a driver for the **Guardian satisfaction**. It is clear that the aim is to understand what is the most important support for the **Status of child**, and as a consequence of Guardian satisfaction, in such a way it is possible to improve the services given by the organization.

In the figure 5.1 is represented the path diagram of the model.

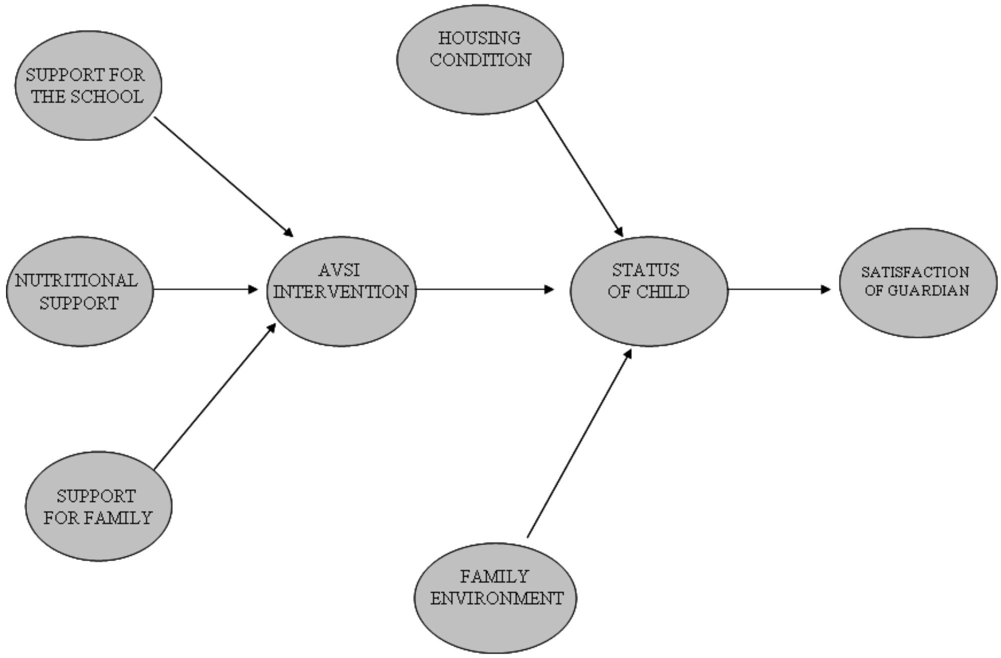


Figure 5.1: The model of "Status of child"

5.2.1 The latent blocks and the manifest variables

The structural model consists of 8 latent variables (six exogenous and 2 endogenous): three latent endogenous block summarize the **Status of child** , **Family characteristics**, **Housing condition** (the characteristics of the house where children live), **Avsi intervention** that is a superblock defined by three latent variables, that describe the kind of support offered for the **Family**, for the **School**, and **Nutritional**. An outcome block of the model is associated to the guardian satisfaction depending on the general Status reached by the Child in the year of the Survey.

So the manifest variables of this model are:

- **Housing condition**[2]: Problems related to the area where the child leaves has the problem of dirtiness, noises, isolation, criminality, House provided of electricity
- **Family environment**: Guardians health, Guardian affected by aids, Guardian handicapped,Guardian drinks alcohol, Effect deseases working capacity, Total income of family,Principal source of income
- **AVSI intervention**: Consists of three kind of support to the children and their families

  1. **Nutritional Support**: Non food items, Healthcare, Nutritional support, Food supply
  2. **Support for the school**: School material, School fees, Recreational activities,After school activities, Emotional support,Vocational training

[2] For this block two manifest variables had a null weights so they were eliminated from the model (*Water obtained from* and House made of).

3. **Support for family**[3] : Healthcare assistance family, Education for parents

- **Status of child**: measured by School attendance, improvement of child health, improvement of child nutrition, improvement of child personality, improvement of the relations with adults, improvement of the relations with friends
- **Satisfaction of guardian**: Expected support from AVSI, AVSI program not adeguacy to the child needs, adeguacy to child needs change, Guardian not consulted, Guardian knowledge of the project

The manifest variables in the model are quite all ordinal and some nominal. In particular the manifest variables of the **AVSI intervention**, (**Nutritional Support**, **Support for the school**, **Support for family**) are all expressed on a dichotomic scale (according if a service has been received or not); the manifest variables of the block **Housing condition** are all ordinal and expressed on a scale of five values, except for the variable that represents the characteristic of the house (i.e. the presence of the electricity). The blocks **Satisfaction of guardian** and **Status of child** are both composed by ordinal variables expressed on an ordinal scale at five levels.

Only the block **Housing condition** is a mixed block and necessity of two different quantifications.

So the quantification process will follow the nature and number of levels of manifets variables. Notice that dichotomous variables describing absence/presence can be treated both as nominal or ordinal. We observed better fitting of the model when considering then as ordinal quantifications.

We estimate the AVSI model with standardized variables.

[3] Two manifest variables of this block had a null weight, so they were eliminated from the model (IGA's for family andOther children of the same family belong to the project).

5.2.2 AVSI outer model estimation

The relationships between the manifest and latent variables are all reflective. The comparison is done between the results of the algorithm PALSOS and PLS-PM, the last estimated with the XLSTAT software. In both programs we estimate the model with standardized variables.

As the relationships between all manifest and latent variables are supposed to be reflective, it is necessary that the manifest variables covary.

A first important result regards the unidimensionality of the latent blocks computed by some indexes. The tables 5.2.2 and 5.2.2 report the results of unidimensionality with the two algorithms.

Observing the results of the table 5.2.2 we can see that all blocks have the property of the unidimensionality, that is all manifest variables explain well the latent concept, in particular it is important to note as the relationships between the manifest variables of the block **Support for the family** improve (0.3307 against 0.8552) after the estimation of the model (we compare the value of Cronbach's alpha and Dillon-Goldtsein's rho); we do not have the same results if we estimate the model with PLS-PM: in this case many manifest variables have a negative sign for the weight and correlation with the respective latent variables. In this case it was been necessary to eliminate the non significative variables[4] from the model and to re-estimate it.

To conclude, the quantification process has caused a significative improvement of the relationships between the latent and manifest variables: the latent variable is maximally correlated with its manifest variables.

We continue the interpretation of the results of the model with the comment of the outer estimation. The weights are normalized (their sum for block is one) and this allows an easier comparison be-

[4] If a variable has a negative weight and a negative correlation with the latent variable means that this variable is not expression of that latent concept.

Latent variables	Mode	C.alpha	DG.rho	F.eigenvalue
Support school	Reflective	0,7166	0,9237	3,54
Nutritional support	Reflective	0,6714	0,9275	3,05
Support family	Reflective	0,3307	0,8552	1,49
AVSI intervention	Reflective	0,8554	0,9494	6,94
Family environment	Reflective	0,7994	0,9463	5,02
Housing condition	Reflective	0,6491	0,8842	3,24
Status of child	Reflective	0,8171	0,9842	5,47
Guardian satisfaction	Reflective	0,7852	0,9857	4,66

Table 5.1: The unidimensionality for PALSOS-PM

Latent variables	Mode	C.alpha	DG.rho	F.eigenvalue
Support school	Reflective			1.74
Nutritional support	Reflective			1.35
Support family	Reflective	0,361		1.22
AVSI intervention	Reflective			2.93
Family environment	Reflective			1.89
Housing condition	Reflective			2.97
Status of child	Reflective	0.576	0,745	1.90
Guardian satisfaction	Reflective	0.913	0,935	4.97

Table 5.2: The unidimensionality for PLS-PM (XLSTAT)

tween them, and assess their importance in the determination of the correspondent latent variable.

The results of PALSOS-PM shows that the parameters (weights and loadings), that express the relationship between a manifest and latent variable, are all positive, unlike the results of XLSTAT [5].

Considering the outer estimation of PALSOS-PM, we can make some consideration about the power and importance of the manifest variables on the latent variable.

In particular, for the block **Support for the school** the variable *School material* (0.262) has an impact major than of the other variables, followed by *Recreational activities* (0.246); in the block

[5] In the **Support for the school** we have an inverse correlation between *school material* and *support for the school* (-0.521); in the block **Family environment** the variables *Guardian handicapped* and *Total income of family* are negative correlated (respectively -0.467 and -0.314); in the block **Housing condition** three variables are negative correlated (*the problem of dirtiness* (-0.564), *noises* (-0.168) and *criminality* (-0.108).

Nutritional support the variables *Healthcare*(0.124) and *No food items*(0.112) have a weight major than the other. In the block **Family environment** the variables *Guardian health*, *Guardian aids* and *Effect deseas* have a major impact than the other; for the block **Housing condition** the most important variables are those regard the environment in which the children leave.

For the measurement of the **Status of child** all manifest variables have the same importance, while for the **Guardian satisfaction** the most important variables are *Expected support from AVSI*(0.213), *Guardian knowledge the project* (0.213), followed by *AVSI program not adeguacy to the child needs* (0.210) and *Adeguacy to child needs change* (0.206).

The Communality and Redundancy are two indexes that explain the power of the relations between the manifest variables and the correspondent latent variable, and the variability explained of the manifest variables of endogenous blocks by all variables of the model. The results obtained with PALSOS-PM shows an increase of these values due to the quantification process that has optimized the correlation between the variables. Again we have a result that confirms the fact that the outer model is well specified.

5.2.3 The PALSOS-PM inner model results

The table 5.2.3 presents the results of the inner model. First of all we observe that on the block **AVSI intervention** the latent variable with a major impact is **Support for the school** (0.474), followed by **Nutritional support**(0.408), being the program of AVSI focused on these two aspects[6], in fact the variable **Support for family** has a low impact (0.199) on the **AVSI intervention**. All relationships are significant, because the T-Statistic has an high value, the p-value

[6] The most of assistance is relative to the school and nutritional support for the child belonging to the program,than the assistance for the family.

Mvs	Normalized Weights	Loadings	Communality	Redundancy
s.fees	0,220	0,781	0,609	0,000
s.material	0,262	0,928	0,861	0,000
recreational.activities	0,246	0,870	0,757	0,000
after s.activities	0,237	0,839	0,704	0,000
emotional support	0,221	0,782	0,611	0,000
no food items	0,297	0,905	0,818	0,000
healthcare	0,300	0,916	0,839	0,000
nutritional support	0,272	0,829	0,688	0,000
food supply	0,275	0,839	0,704	0,000
healthcare family	0,578	0,864	0,747	0,000
educational parents	0,578	0,864	0,747	0,000
s.fees	0,111	0,770	0,592	0,592
s.material	0,121	0,841	0,707	0,707
recreational.activities	0,128	0,889	0,790	0,790
after s.activities	0,101	0,699	0,488	0,488
emotional support	0,111	0,769	0,592	0,592
no food items	0,112	0,776	0,603	0,603
healthcare	0,124	0,860	0,740	0,740
nutritional support	0,117	0,810	0,657	0,657
food supply	0,110	0,766	0,586	0,586
healthcare family	0,120	0,836	0,699	0,699
educational parents	0,101	0,704	0,495	0,495
Guardian health	0,183	0,920	0,847	0,000
Guardian aids	0,183	0,921	0,848	0,000
Guardian handicapped	0,186	0,934	0,872	0,000
Guardian drinks alcohol	0,159	0,797	0,635	0,000
effect deseas	0,165	0,828	0,686	0,000
principal source income	0,159	0,797	0,635	0,000
total income	0,141	0,707	0,500	0,000
dirtiness	0,296	0,959	0,920	0,000
noises	0,297	0,962	0,926	0,000
isolation	0,292	0,948	0,898	0,000
criminality	0,216	0,702	0,492	0,000
electricity	0,020	0,083	0,007	0,000
child attendance	0,159	0,871	0,758	0,703
child health now	0,176	0,964	0,930	0,862
child nutrition now	0,177	0,971	0,943	0,874
child personality now	0,179	0,979	0,959	0,889
child relations adults now	0,177	0,967	0,935	0,867
child relations friends now	0,178	0,975	0,951	0,881
Expected support from AVSI	0,213	0,992	0,984	0,285
AVSI program not adeguacy to the child needs	0,210	0,980	0,961	0,279
Adeguacy to child needs change	0,206	0,958	0,918	0,266
Guardian not consulted	0,194	0,903	0,815	0,237
Guardian knowledge project	0,213	0,992	0,984	0,285

Table 5.3: The outer estimation with PALSOS-PM algorithm

Endogenous lvs	Exogenous lvs	Path	Std.Err	T.Statistics	P-value	L.bound	U.bound
AVSI intervention	Support for the school	0,474	0,045	568,876	0,000	0,383	0,541
	Nutritional support	0,408	0,043	527,693	0,000	0,310	0,467
	Support for family	0,199	0,045	568,876	0,000	0,150	0,229
Status of child -¿	AVSI intervention	-0,176	0,380	-8,400	0,000	-0,531	1,031
	Family environment	0,726	0,393	29,754	0,000	-0,203	1,420
	Housing condition	0,431	0,251	26,891	0,000	-0,472	0,533
Guardian satisfaction	Status of child	0,539	0,120	21,734	0,000	0,452	0,895

Table 5.4: The inner estimation with PALSOS-PM algorithm

is 0 and the confidence intervals do not contein the zero.

The **Status of child** depends by three latent variables: the **AVSI intervention**, **Family environment** and **Housing condition**. From these results emergences that **AVSI intervention** has a negative impact (we have the same result if we use the PLS-PM (-0.230)): according to them the most important variable for the **Status of child** is the **Family environment** (0.726), followed by the **Housing condition** (0.431), while **AVSI intervention** has a negative impact (-0.176))[7] on the **Status of child**, that looks strange being it different in sign with respect to the correlation coefficient between the same variables taht is positive. We have to remember that this is a coefficient of multiple regression so it is the impact of the **AVSI intervention** excluding the effect of the other variables: the change in sign may be consequence of a multicollinearity problem. To verify this hypothesis we can analyze the matrix of correlation between the latent variables (see table 5.2.3): all correlation are positive and high we observe too as the two latent variables **AVSI intervention** and **Family environment** have an high positive correlation (0.924): this correlation can influence the sign of the path coefficient of **AVSI intervention**.

[7] The model estimated by mean of PLS-PM shows some problems in the outer estimation, because some manifest variables have negative weights and loadings, so this cause a bad estimation of the latent variables.

	S.s	N.s	S.f	AVSIi.	F.e.	H.c.	S.c.	G.s.
Support school	1,000							
Nutritional support	0,759	1,000						
Support family	0,801	0,764	1,000					
AVSI intervention	0,943	0,920	0,890	1,000				
Family environment	0,882	0,794	0,912	0,925	1,000			
Housing condition	0,806	0,651	0,860	0,814	0,867	1,000		
Status child	0,787	0,739	0,868	0,846	0,937	0,917	1,000	
Guardian satisfaction	0,586	0,853	0,570	0,741	0,643	0,423	0,539	1,000

Table 5.5: The correlation matrix between the latent variables

Variables	Coefficient	St.error	Lower bound	Upper bound
A. intervention	0,310	0,027	0,256	0,364
F. environment	0,343	0,017	0,309	0,377
H. condition	0,336	0,009	0,317	0,354

Table 5.6: The regression coefficients of PLS-R

This multicollinearity problem can be faced using the Partial Least Squares regression (PLS), instead of OLS estimation in the algorithm PALSOS-PM[8] (in the a PLS regression we have a set of explicative variables and one dependent variable to predict).

The results of the PLS are reported in the table 5.2.3. As for the PLS-PM also in the PLS we do not have distributional hypothesis, so the validation of the coefficients is made across the resampling techniques as bootstrap or jacknife, obtaining the empirical distribution for the parameters. From this table we can see that the PLS, with just one component, solves the problem of multicollinearity, estimating, as expected, a positive regression coefficient for the variable **AVSI intervention**. **Avsi intervention** show an impact (0.310) on the **status of child** quite similar to the other two factors **Family environment** (0.343), and **Housing condition** (0.336), on which AVSI could not intervene[9].

Concerning the interval confidence built with the resampling tech-

[8] The algorithm PALSOS-PM gives the possibility to choose between an OLS or a PLS regression, for the estimation of the path coefficients. The PLS is made using the package pls, written by Mevik and Wehrens

[9] We remember that the manifest variables of **Housing condition** regard the social environment in which the child leaves and the characteristic of his house.

nique, we can see that all intervals not include the null value. The standard deviation asssumes low values, and the coefficients are all significatively different from zero showing stable path estimation in the bootstrap.

The last parameter of the table 5.2.3 expresses, instead, the impact of **Status of child** on the **Guardian satisfaction** (0.539): is quite good as we can see it is significatively different from zero (T-Statistic has an high value and the p-value is very low). As conseguence the Guardian gives a good evaluation of the AVSI program especially if he notes an improvement in the life condition of the child and in his healthcare, confirming the theory on the base of the proposed model.

The estimation of theregression equations describing the structural model one[10]:

1. **Avsi intervention** = 0.000 + 0,474* **Support school** + 0,408* **Nutritional support** + 0,199* **Support for family**
2. **Evolution of child** = 0.000 + 0,336* **Housing condition** + 0,343* **Family environment** + 0,310* **AVSI intervention**
3. **Satisfaction of guardian** = 0.000 + 0,539* **Status of child**

The relationships estimated show that AVSI intervention, produce relevant improvements on the Status of these children. Similarly happens for **Housing condition** and **Family** conditions are good. The last impact suggests to improve the AVSI support to child Family due to the strong impact of it on the child status. So we can read an inderect effect of AVSI.

The next two tables 5.2.3 and 5.2.3 report the results of the 100 bootstrap replications of the weights and loadings. It is worth noticing that due to the high value of the T-Statistic all weights and loadings are significatively.

[10] The 0 intercept regression depends on the assumption of standardization for latent variables.

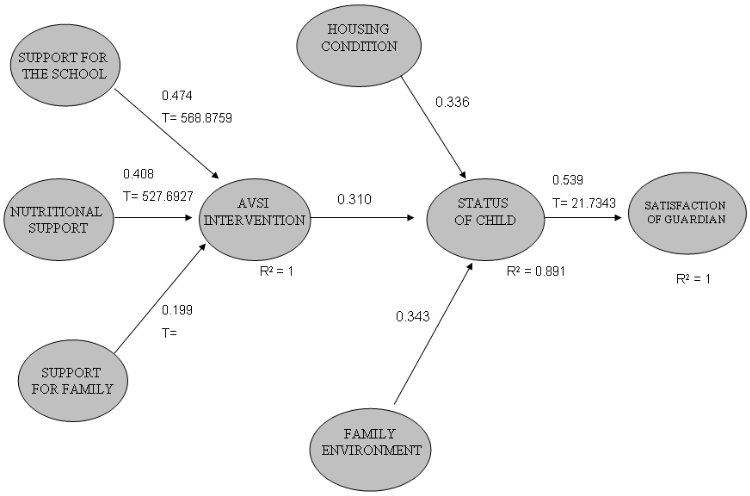


Figure 5.2: The results of the model "Status of child"

The unique exception is the manifest variables *Electricity* of the **Housing condition** block that has a negative confidence interval with a low value for the T-Statistic: we can see that the weight associated is near the zero and so the variable could be excluded from the block.

The standard deviation shows that there is a low variability between the weights estimated in the 100 bootstrap replications it means that the estimation of the weights and loadings are stable.

In the tables 5.2.3 and 5.2.3 a summary of the performance for both PALSOS-PM and PLS-PM are given.

In these tables the latent variables, the type of block (endogenous or exogenous), the values of R^2, the Average communality, Average redundancy indexes and Gof are given. As regard the values of R^2 we can see that for the multiblock it is equal to one, while it assumes an high value for the regression of Status of child on the three latent variables, even if we have substitute this regression with a PLS for the problem of multicollinearity. So for this relationships we consider the validation indexes of a PLS; in particular we can measure the importance of an explicative variable in the definition of a component. The index used for this purpose is the *Variable Importance in the Projection* (VIP): if its value is equal or major of 0.8 the explicative variable is important in the prediction of a dependent variable. In our case all variables are important in the definition of the dependent variable, as we can note from the graphical in figure 5.3.
Another important result is the improvement of the Gof index of the AVSI model that assumes with the PALSOS-PM algorithm the value 0.7482, versus the value 0.345, obtained in the case of non quantified manifest variables with the classical PLS-PM estimation. This big change is due to the improvement obtained by the quantification of the optimal outer estimation. Similarly the R^2 (0.891) of the regression explained the **status of child** latent variable is high too, evidentiating the good performance of the model proposed.

M.vs	Original value	Mean b.	Std,Err	T. Statistic	L.Bound	U. Bound
S.fees	0,220	0,225	0,038	5,845	0,132	0,294
S.material	0,262	0,243	0,038	6,930	0,146	0,314
Recreational activities	0,246	0,256	0,026	9,418	0,206	0,307
After school activities	0,237	0,244	0,035	6,793	0,174	0,314
Emotional support	0,221	0,247	0,029	7,621	0,181	0,312
No food items	0,297	0,301	0,032	9,267	0,254	0,366
Healthcare	0,300	0,306	0,030	9,954	0,258	0,377
Nutritional support	0,272	0,297	0,039	6,992	0,218	0,370
Food supply	0,275	0,290	0,037	7,451	0,231	0,364
Healthcare family	0,578	0,558	0,032	17,837	0,517	0,641
Educational parents	0,578	0,558	0,032	17,837	0,517	0,641
S.fees	0,111	0,104	0,020	5,562	0,049	0,137
S.material	0,121	0,111	0,017	6,935	0,076	0,139
Recreational activities	0,128	0,118	0,013	10,248	0,089	0,140
After school activities	0,101	0,113	0,015	6,550	0,071	0,138
Emotional support	0,111	0,115	0,015	7,587	0,078	0,140
No food items	0,112	0,111	0,015	7,489	0,080	0,144
Healthcare	0,124	0,120	0,010	12,149	0,098	0,141
Nutritional support	0,117	0,112	0,018	6,340	0,062	0,145
Food supply	0,110	0,110	0,015	7,192	0,078	0,139
Healthcare family	0,120	0,117	0,012	9,714	0,088	0,143
Educational parents	0,101	0,110	0,014	7,432	0,078	0,135
Guardian health	0,183	0,184	0,018	9,994	0,148	0,221
Guardian aids	0,183	0,179	0,023	8,038	0,124	0,216
Guardian handicapped	0,186	0,170	0,024	7,778	0,113	0,215
Guardian drinks alcohol	0,159	0,180	0,020	8,056	0,140	0,219
Effect deseas	0,165	0,186	0,023	7,058	0,120	0,227
Principal source income	0,159	0,179	0,028	5,747	0,109	0,216
Total income	0,141	0,175	0,026	5,458	0,127	0,230
Dirtiness	0,296	0,305	0,025	11,927	0,264	0,360
Noises	0,297	0,308	0,027	11,078	0,270	0,373
Isolation	0,292	0,243	0,040	7,380	0,126	0,307
Criminality	0,216	0,296	0,026	8,380	0,261	0,355
House with electricity	0,020	0,000	0,011	1,831	-0,023	0,023
Child attendance	0,159	0,186	0,023	7,043	0,129	0,223
Child health now	0,176	0,192	0,023	7,659	0,148	0,233
Child nutrition now	0,177	0,197	0,020	8,898	0,147	0,232
Child personality now	0,179	0,199	0,026	6,913	0,156	0,250
Child relations adults now	0,177	0,201	0,021	8,512	0,166	0,239
Child relations friends now	0,178	0,201	0,020	8,901	0,167	0,235
Expected support from AVSI	0,213	0,222	0,018	11,638	0,193	0,281
AVSI program and child needs	0,210	0,218	0,016	13,114	0,190	0,258
Adeguacy to child needs change	0,206	0,215	0,014	14,518	0,198	0,249
Guardian not consulted	0,194	0,215	0,013	14,706	0,194	0,255
Guardian knowledge the project	0,213	0,217	0,014	14,947	0,195	0,268

Table 5.7: The bootstrap results for the weights (PALSOS-PM)

M.vs	Original value	Mean B.	Std.Err	T-Statistic	L. Bound	U. Bound
S.fees	0,781	0,753	0,130	5,991	0,406	0,923
S.material	0,928	0,810	0,120	7,749	0,453	0,944
Recreational activities	0,870	0,852	0,079	11,028	0,679	0,947
After school activities	0,839	0,812	0,108	7,790	0,531	0,934
Emotional support	0,782	0,825	0,095	8,266	0,555	0,927
No food items	0,905	0,837	0,083	10,886	0,608	0,950
Healthcare	0,916	0,851	0,073	12,644	0,696	0,950
Nutritional support	0,829	0,827	0,111	7,454	0,514	0,950
Food supply	0,839	0,807	0,100	8,414	0,613	0,942
Healthcare family	0,864	0,900	0,049	17,514	0,780	0,967
Educational parents	0,864	0,900	0,049	17,514	0,780	0,967
S.fees	0,770	0,729	0,135	5,721	0,324	0,893
S.material	0,841	0,781	0,123	6,827	0,497	0,927
Recreational activities	0,889	0,828	0,085	10,427	0,603	0,951
After school activities	0,699	0,791	0,102	6,860	0,506	0,926
Emotional support	0,769	0,805	0,106	7,233	0,523	0,947
No food items	0,776	0,779	0,100	7,798	0,533	0,912
Healthcare	0,860	0,844	0,076	11,258	0,626	0,946
Nutritional support	0,810	0,787	0,127	6,396	0,384	0,928
Food supply	0,766	0,771	0,106	7,244	0,535	0,917
Healthcare family	0,836	0,822	0,083	10,132	0,590	0,927
Educational parents	0,704	0,770	0,098	7,151	0,509	0,904
Guardian health	0,920	0,811	0,082	11,184	0,604	0,937
Guardian aids	0,921	0,790	0,101	9,122	0,514	0,929
Guardian handicapped	0,934	0,748	0,109	8,602	0,488	0,905
Guardian drinks alcohol	0,797	0,793	0,093	8,569	0,613	0,921
Effect deseas	0,828	0,814	0,088	9,407	0,547	0,938
Principal source income	0,797	0,786	0,119	6,676	0,429	0,919
Total income	0,707	0,770	0,111	6,390	0,538	0,920
Dirtiness	0,959	0,905	0,056	17,171	0,768	0,972
Noises	0,962	0,914	0,042	23,056	0,809	0,972
Isolation	0,948	0,726	0,135	7,022	0,358	0,899
Criminality	0,702	0,880	0,080	8,831	0,690	0,963
House with electricity	0,083	0,001	0,043	1,949	-0,088	0,085
Child attendance	0,871	0,801	0,103	8,467	0,525	0,954
Child health now	0,964	0,828	0,086	11,162	0,646	0,948
Child nutrition now	0,971	0,847	0,078	12,401	0,659	0,945
Child personality now	0,979	0,858	0,105	9,365	0,633	0,954
Child relations adults now	0,967	0,866	0,090	10,762	0,651	0,954
Child relations friends now	0,975	0,866	0,074	13,180	0,721	0,950
Expected support from AVSI	0,992	0,936	0,050	20,014	0,771	0,985
AVSI program and child needs	0,980	0,918	0,067	14,579	0,731	0,980
Adeguacy to child needs change	0,958	0,909	0,070	13,747	0,711	0,987
Guardian not consulted	0,903	0,909	0,065	13,845	0,702	0,972
Guardian knowledge the project	0,992	0,916	0,054	18,488	0,781	0,986

Table 5.8: The bootstrap results of loadings (PALSOS-PM)

Latent variables	Type variable	R squared	Average communality	Average Redundancy	Gof index
Support school	Exogenous	0,000	0,709	0,000	
Nutritional support	Exogenous	0,000	0,762	0,000	
Support family	Exogenous	0,000	0,747	0,000	
AVSI intervention	Endogenous	1,000	0,632	0,632	
Family environment	Exogenous	0,000	0,718	0,000	
Housing condition	Exogenous	0,000	0,649	0,000	
Status child	Endogenous	0,927	0,912	0,846	
Guardian satisfaction	Endogenous	0,290	0,932	0,270	
					0,748

Table 5.9: The summary of the model performances with PALSOS-PM

Latent variables	Type variable	R squared	Average communality	Average redundancy	Gof index
S. school	Exogenous		0,345		
N. support	Exogenous		0,319		
F. support	Exogenous		0,592		
A. intervention	Endogenous	1,000	0,251	0,251	
H. condition	Exogenous		0,196		
F. environment	Exogenous		0,204		
S. of child	Endogenous	0,151	0,371	0,055	
G. satisfaction	Endogenous	0,032	0,386	0,012	
					0,345

Table 5.10: The summary of the model performances with PLS-PM (XL-STAT)

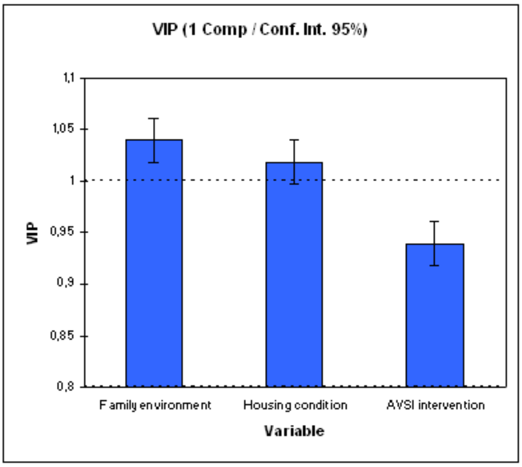


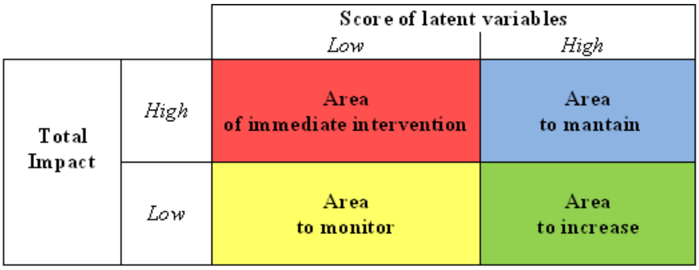
Figure 5.3: The VIP index

The Decision Support Matrix

The Structural Equation Modeling aim at estimating the impact that the exogenous variables, assumed as causes, have on the endogenous ones. In particular in the proposed model we measured the impact of some latent factors such as House condition, **AVSI intervention** and **Family environment**, exercise on the **Status of child** as measured by its manifest variables referred to quality of life, health, behaviour and school performance of OVCs. It is worth noticing that to make a decisional use of the SEM, together with the identification of the latent factors having major impacts on the child status in defining such concepts, we have to take into account the averages scores of the latent variables.

Only a joint lecture of both the information (path coefficients/impacts and average scores) allow to detect the drivers for the Status improvement identifying the critical area on which to intervene, the degree of urgency as well as the set aspects.

A Decision Support Matrix (see the figure 5.4) consists of a suitable Cartesian map reporting on the abscissas axis the average scores of the exogenous latent variables affecting the latent variable measuring the **Status of the Child** and on the ordinates axis the correspondent path coefficients. The reference point (barycentre) of the map is located in the point having as coordinates the mean latent score and the mean path coefficient. By means of this map we can perform a so called swot analysis (strengths - weaknesses - opportunities - threats) by identifying four characteristic areas described as: the area of *weakness*, were variables have an high path coefficient and at the same time a high mean value require immediate intervention; the area of *threats* on which variables have a low value for both the path coefficient and for the mean requires just to be monitored; the area of *opportunities* is the area to promote or to increase, because variables have already an high mean value but a low path coefficient that might be more important in the future; the last area regard the strengths as variables have an high value for the mean and for the path coefficient, is the area to be maintained. In the following (see figure 5.5) we report the map built to identify the critical factors among the ones measured to improve the Status of Child.



		Score of latent variables	
		Low	*High*
Total Impact	*High*	Area of immediate intervention	Area to mantain
	Low	Area to monitor	Area to increase

Figure 5.4: The Decision Support Matrix

We observe that **Housing condition** in order to improve the **Status of Child** is perceived as the factor that requires an immediate intervention as it has an high impact whereas its score is quite low.

The **AVSI intervention** has a low mean, even if it is due in part to the dichotomic nature of the scale of the correspondent manifest indicators and a low path coefficient, so it is in the Area to monitor.

The latent variable **Family environment**, instead is located in the strengths area, nevertheless according a vision of a continuous improvement it might be a concept to increase too.

As we decide to pay more attention to the **AVSI intervention**, a similar map based on the average optimal score for the manifest variables and their weight affecting the multiblock latent variable that is the **AVSI intervention**, can be built (see figure 5.6).

We observe that **Support for the school** in order to improve the **AVSI intervention** is perceived as the factor that requires an

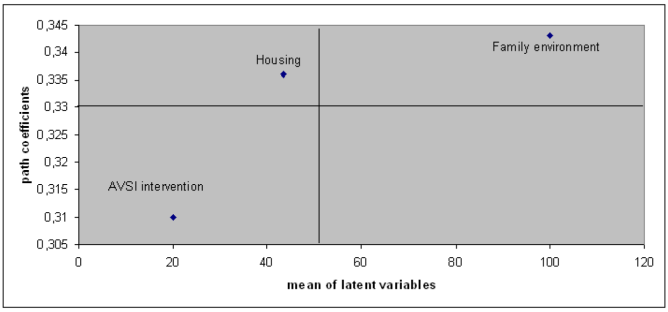


Figure 5.5: The drivers for the Status of child

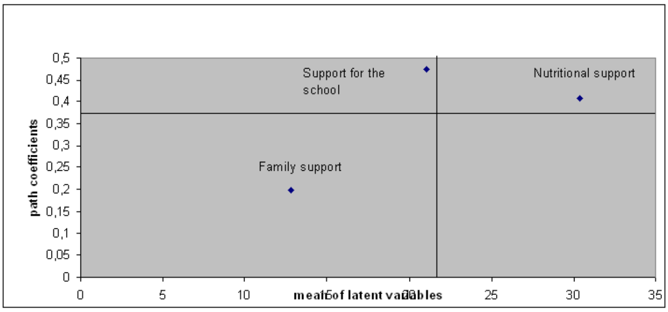


Figure 5.6: The drivers for the AVSI intervention

immediate intervention as it has an high impact whereas its score is quite low.
The **Support for family** has a low mean, which corresponds a low path coefficient: this variable is yhe Area to monitor, so AVSI has to increase the support given to the family of children belonging to the program.

The latent variable **Nutritional support**, instead is located in the strengths area, nevertheless according a vision of a continuous improvement it might be a concept to increase too.

5.3 A discussion

The analysis presented offers a lot of interesting information issued by the use of a multivariate approach to exploit the actual survey data in view of an original evaluation exercise.

The achieved results suggest their potential use in decision making. In conclusions we prefer to mention, among the achieved results, the ones that have a perspective added value, especially in view of the third OVC survey. An important lesson learned from this exercise, different from the data analysis French School view point is that: The data should follow the model and not vice versa.

This imply that a correct and most performing use of research tools should follow a proper design of the survey questionnaire according a conceptual model and the correspondent technology adopted.

For this reason even if the obtained results are very interesting we do not consider them as definitive, but useful in the perspective of the next survey in view of preparing a new questionnaire aimed not only at an explorative description the OVC phenomenon, but also at defining a causal model able to identify the factors having the highest impact on children general status and evolution so to improve the performance of the AVSI intervention in terms of quality, efficacy and efficiency.

From the results of the model is evident that Avsi must improve its support, especially the assistance to the family of children that belong to the program. In fact we have seen as the **Family environment** is the factor most important in the evolution of child condition.

This SEM model toghether with PALSOS-PM algorithm reveal itself as an important tool to take decision. In fact it allows to factors that have a significant impact on the outputs and outcome of an intervention aimed at improving OVCs conditions.

Another important result that we have obtained is the improvement of the estimation performance of the model: the quantification process has produced a better estimation of the latent variables.

On the basis of these two variable results we have also validated questionnaire variables.

This is a social research in which the quantification process was an important tool to reach the objective of the analysis.

Conclusion and future perspectives

In the expression **Structural Equation Modeling** (SEM) two concepts are synthesized: the existence of a model, a formal expression of a theory and analytic methods described by means of a system of equations that represent the casual relationships away latent multidimensional concepts that is structures.

The two approaches for the estimation of this model, LISREL and PLS-PM, are considered as the so called second generation ; they allow to express complex relationships involving non observable variables by a observable or manifest variables.

We can perceive that the SEM was born for the analysis of quantitative variables: the aim of the present thesis is centered on the problem of the treatment of manifest qualitative/ordinal variables.

Both the estimation methods (LISREL and PLS-PM) allow to introduce in the model nominal and ordinal variables, with a substantial difference that LISREL does not accept nominal variables but only ordinal, for which it computes the tetracoric correlations, substituting the Pearson's correlation coefficient. On the other hand the PLS-PM permits to consider for the estimation of the model all kind of variables, nominal and ordinal: for the nominal variables the algorithm builds the dummy variables (in this way the number of manifest variables in the block increases, causing some problems in the estimation of latent variable), while for the ordinal variables there is the assumption of the continuity.

The last hypothesis is very strong and is not correct, in particular, when in the model we have manifest variables with different scale (three, five or ten values): in this case it would be better to quantify the variables in such a way to obtain numerical variables to justify the use of the quantitative technique.

In particular this problem it was faced with a data set composed by only ordinal and nominal variables where the aim was to estimate a SEM model to evaluate the impacts of some variables on a variable of outcome. This is the AVSI model presented in the chapter four: estimating the model considering the variables as numerical we obtain that the path coefficients are not significant as also many manifest variables, because they have negative signs for the weights and loadings.

From this practical problem we have started to develop an algorithm that estimates a SEM model as the PLS-PM, but with the possibility, if it is necessary, to quantify the qualitative variables, introducing them as numerical in the analysis. This is PALSOS-PM algorithm that uses as method of quantification the ALS approach.

In the first chapter of the present thesis the methods of quantification proposed in the literature are discussed with respect to the type of variables produced after the quantification and the objective functions used with the aim to get the optimal quantification. According to the nature of variables and also the objective of the analysis in the fourth chapter after the evolution of the quantification methods that are developed with respect to the technique of analysis used (chapter 2), we propose an original algorithm based on ALSOS (presented in the third chapter) that aims to optimize the estimation of the latent variables and their relationship toghether with an optimal quantification of non numerical manifest variables.

Therefore this approach allow to obtain, in the sense of least squares[11], the best quantification, the best estimation of the parameters of the model, according to the quantified variables, and in our case the best prediction of the latent variables.

The proposed system , PALSOS-PM, seems to be a good solution in all cases in which we have an eterogeneous set of manifest variables, and in which our objective is to estimate a casual model in presence of reflective or formative relationships whereas alternatives present just consider reflective relationships.

In particular with respect to the proposal of Jakobowicz and Derquenne, in our algorithm we use a unique function to obtain both the weights and the vector of scaling; no matter the nature of the variable be numerical, ordinal or nominal, weight in all cases is the covariance between the latent and the manifest variables.

It has been proved that iteratively PALSOS-PM reduces the sum of squares of residuals of the regression in the SEM model until they reach the minimum. It uses all the validation indexes of the PLS-PM to verify the correctness of the supposed model.

Summarizing the most important characteristics of our proposal are:

- the absence of distributional hypothesis on the raw data and in the process of quantification;
- it estimates the model with the PLS-PM algorithm when all variables are numerical, both in case of reflective or formative relationships;
- it estimates the model with mixed variables (nominal, ordinal and numerical) being in this respect more general that algorithms in the literature;
- for the validation it uses the bootstrap technique, the number of resampling is choosen by the researcher: 100 replications is a sufficient replications size;

[11] We remember that it minimizes the squares of the sum of residuals.

- it has no problem for the dimension of data matrix (we can have more variables than observation, without problem in the estimation of the model);
- it allows the control on the signs changes, as made in the XL-STAT and SPAD-PLSPM software;
- it gives the possibility to move to PLS estimators if the classical OLS fail for multicollinearity;
- it computes the best prediction of the latent variables according to the typology of the manifest variables;
- the process of quantification works for typology of relationships between the manifest and latent variables reflective and formative nature;

The results obtained for the Avsi model show the excellent performances of PALSOS-PM approach to cope with very complex situation consisting of an high number of heterogenous variables for their nature, nominal and ordinal, as well as for theri number of categories.

To conclude we remark our thesis was mainly focused on the development of quantification procedures to en in the PLS-PM algorithm capabilities.

In this perspective some questions in future works such as the selection problems regarding the manifets variables,of latent variables as well as the number of components for non unidimensional blocks or in case of multicollinearity. A t the same time the influence of the number of levels for ordinal variables and the number of categories for nominal ones should be particularly investigated as they affect the robustness of both quantifications and model estimations.

Finally the role of the latent variable in a SEM model should be furtherly discussed. It is necessary not only to define the role of the variables (mediator or mediation), but also have to introduce esistence information in the estimation of the model, in such a way that the path coefficients can benefit of such knowledge.

Such situations rise very frequentlyin practical data analysis, especially in social and surveys, tehy require suitable and robust methods to be faced.

Appendix A

The proof of some theorems

In this appendix is reported the proof of the propositions enunciated in the Chapter 3,about the Projection in a Convex Cone.

A.1 Proposition 3.4

The vector x is nonnull since y does not belong to C^p. Result (1) is obvious. We now prove (3). The maximum of cos(y,z) over z in C is reached for some vector $z = x_1$ and is strictly positive since y does not belong to C^p. The cosine being independent of the norm, we may choose a vector x_1 with the same norm as that of x. We then have $\|y - x_1\| \leq \|y - x\|$ and the unicity of the projection into C implies that $x_1 = x$. The same argument is used to show that $cos(y, z) = cos(y, x)$ implies that z belongs to $C(x)$. We now prove (2). The minimum of $\|y - z\| / \|z\|$ is reached for some vector $z = x_2$. We denote by x_3 a vector of $C(x)$ with the norm of x_2. From $\|y - x_3\| / \|x_3\| \geq \|y - x_2\| / \|x_2\|$ we get $cos(y, x_3) \leq cos(y, x_2)$, so x_2 belongs to $C(x)$. Let us calculate the scalar λ such that $x_2 = \lambda x$. The ratio $\|y - \lambda x\| / \|\lambda x\|$ being equal to the minimum of $\|x - \alpha y\| / \|x\|$ with respect to α, we conclude that $\lambda = \|y\|^2 / x' = 1/cos^2(x,y)$. We can deduce (2) from the equalities

$$\frac{\|y - x_2\|^2}{\|x_2\|^2} = \frac{\|x\|^2 - \lambda^{-2}\|y\|^2}{\|x\|^2} \text{and} cos(x,y) = \frac{\|x\|}{\|y\|}$$

A.2 Corollary 3

The minimum of $\|y - u\|$ for u belonging to $C \cap S$ is reached for a unit vector that maximizes cos(y,u).

A.3 Proposition 3.5

This proposition is deduced from the following inequalities:

$$\|Au - Av\| \leq \|u - v\| \,\text{for any u and v in}\Re^n$$

and

$$\|Bu - Bv\| \leq \frac{2}{\|Au\|}\|u - v\| \,\text{for any u and v in}\Re^n - C^p$$

A.4 Proposition 3.6

Let $y_{R+s} = \sum \beta_i s_i; s_i \in R + s$. If all the β_i are strictly positive, then Q=R and the proposition is proved. Otherwise there exists at least one $\beta_i \leq 0$. However, the coefficient β of s is strictly positive, since $\|y - y_R\|^2 \prec \|y - y_{R+s}\|^2 = 2\beta(y - y_R)'s - \|y_R - y_{R+s}\|^2 \succ 0$. Now consider the vector x of the segment $[y_R, y_{R+s}]$ which belongs to C and is as close as possible to y. This vector x is equal to

$\lambda y_R + (1-\lambda) y_{R+s}$ with $\lambda = Max\beta_i/(\beta_i - \alpha_i); \beta_i \leq 0$. We denote by I the set of vectors s_i such that $\lambda\alpha_i + (1-\lambda)\beta_i = 0$. As $\lambda \prec 1$ and $\beta \succ 0$ the vector s does not belong to I. Since vector x belongs to $C(R+s-I)$ we get $\|y_R - y_{R+s-I}\| \leq \|y - x\| \prec \|y - y_R\|$. Let $y_{R+s-I} = \sum \beta'_i s_i; s_i \in R+s-I$ be the projection of y into L(R+s-I). If all the β'_i are strictly positive, then Q=R-I. Otherwise, we iterate the described procedure with x playing the role of y_R and y_{R+s-I} that of y_{R+s}, excluding from R another subset I' such that $\|y_R - y_{R+s-I-I'}\| \prec \|y - y_R\|$. After a finite number of iterations we obtain a subset Q of R with the desired property.

A.5 Proposition 3.7

Proposition 2 implies the existence of a subset R of $\{z_1, z_2, ..., z_m\}$ such that the projection $x^*$of y into $C = L(z_1) \oplus C(z_1, z_2, ..., z_m)$ is equal to the projection of y into the subspace $L(R)$. The vector z_1 belongs to R since $-z_1$ and z_1 belong to C. Let J denote the set of indices j such that z_j belongs to $R : J = \{1, j1, ..., jr\}$. This set J induces a partition of M into increasing blocks: $\{B_h = \{jh, ..., j(h+1)-1\}, h =$ with $j0 = 1$ and $j(r+1) = m+1$. The indicatory variables $i(B_h = \sum\{x_j; j \in B_h\}$of the blocks B_h verify $i(B_h) = z_{jh} - z_{j(h+1)}$ for $h = 0, ..., r-1$ and $i(B_r) = z_{jr}$. Since the subsets $R = \{z_1, z_{j1}, ..., z_{jr}\}$ and $\{i(B_h), h = 0, ..., r\}$ generate the same subspace, we may deduce

$$\begin{aligned} x^* &= \sum\{\bar{y}(B_h)i(B_h); h = 0, r\} \\ &= \bar{y}(B_0)z_1 + \sum\{(\bar{y}(B_h) - \bar{y}(B_{h-1})z_{jh}; h = 1, r\} \end{aligned} \tag{A.1}$$

Appendix B

The code of PALSOS-PM algorithm

In this appendix is reported the R routine. Starting from a matrix in which there are p variables and n units, and partitioned in k latent blocks, the algorithm starts to compute a first estimation of the latent variables as a linear combination of manifest variables multiplied for a casual vector of weights.

Successively the algorithm updates the estimation of the latent variables across the inner weights e_{ij} that are or the correlation between the latent variables, or the signs of these correlations.

Then the last step is the new outer estimation in which the qualitative variables are quantified across Morals. The algorithm stops when the estimation of latent variables is stable.

After the convergence the algorithm develops the inner multiple/single regression and it calculates the validation index used by the PLS-PM algorithm.

The code of PALSOS-PM algorithm

B.1 PALSOS-PM algorithm

```
PALSOS-PM<- function (mat, inner,epsilon,maxiterations)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   PALSOS-PM estimates a SEM model based on numerical, ordinal and      %
%    nominal manifest variables                                          %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      mat: the matrix of the manifest variables. Columns represent p    %
%            variables, while rows contain n individuals                 %
%      inner: the matrix in which the relationships between the latent   %
%              variables are expressed                                   %
%      epsilon: is the criterio of stop for the algorithm, a positive    %
%             value near zero                                            %
%      maxiterations: the number of max iteration                        %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%      weights, loadings: the coefficients of correlation and of         %
%      covariance that express the relationships between                 %
%      the manifest and latent variables                                 %
%                                                                        %
%      correlation: the correlation between                              %
%      the manifest and latent variables                                 %
%                                                                        %
%      path coefficients: the regression coefficients of                 %
%      the inner estimation. They express the relationships              %
%      between the latent variables                                      %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
PALSOSPM <- function(mat, inner, epsilon, maxIterations)
{
weights <- list();
loadings<-list();
w.norm<-list();
l.finali<-list();
v.nostand<-list();
matq<-list();

% These are the commands for the initialization of the vector of weights,
% loadings, normalized weights and standardized loadings
% The matrix matq has the same dimension of the initial
% matrix and will contain the quantified variables
% The v.nostand is the matrix of latent variables not standardized
sigma<-numeric()
iteraz <- 0;
```

```
% Inizialization of the iterations
convergenza<-0;

iniz <- initialstep(mat);
%This function computes the first estimation of the latent variables,
%across a casual vector of weights multiplied for the manifest variables
v<-as.matrix(iniz$v);
% v is the first estimation of the latent variables
v <- as.matrix(scale(v));
vmq<-iniz$vmq;

repeat
%This is the repeat routine in which the algorithm alternates
%the inner and outer estimation until the convergence
    {

    iteraz <- iteraz +1;
    z <- innerestimation(v,inner);

    %This function estimates the inner weights to update the estimation of
    %latent variables

    zeta<-as.matrix(z$zeta);

    %Zeta is the matrix of the latent variables after the inner estimation

    zeta <- scale(zeta);
    stimaEst <- morals(mat,zeta,100,0.0001,natura);

    %This is the Morals algorithm that quantifies the qualitative variables

    if(stimaEst$conv!= -1)
        {
        %This condition is relative to the convergence of Morals algorithm
        %It converges it is possible to proceeds with a new inner estimation

        weights[[iteraz]] <- stimaEst$weights;
        loadings[[iteraz]]<-stimaEst$loadings;
        v <- as.matrix(stimaEst$vardip);
        sigma<-stimaEst$sigma;
        v<-scale(v);
        v.nostand<-stimaEst$vindosnostand;

        if(iteraz != 1)
            {
            %At this step the algorithm verifies if the estimation
            %of latent variables (inner and outer) is stable.
            diffBeta <- betaVet(weights, iteraz) - betaVet(weights, (iteraz -1));

            if(max(abs(diffBeta)) <= epsilon)
                {
                convergenza<-convergenza+1;

                break;
                %At this point the algorithm stops
                }
            }
        }

    if(iteraz==maxIterazioni |stimaEst$conv== -1)
        {

        convergenza<--1;
        }

    if( convergenza == -1) break;
    }

if(convergenza != -1)
    {
    %The algorithm after the convergence computes the normalized
    %weights and the standardized loadings

    w<-weights[[iteraz]];
    zeta<-v
    matq<-stimaEst$varind;

    for(i in 1:dim(zeta)[2])
        {
        w.norm[[i]]<-array(0, dim(matq[[i]])[2]);
        l.finali[[i]]<- array(0, dim(matq[[i]])[2]);

        for(j in 1:dim(matq[[i]])[2])
            {
            w.norm [[i]] [j]<-as.numeric(w[[i]] [j])/sigma[i];
            l.finali[[i]] [j]<-cor(zeta[,i],matq[[i]][,j]);
            }
        }
    }

#==============Inner model=============================================#
% After the convergence the PALSOS-PM function estimates, across
% other functions the final inner estimation and all indexes
% for the validation of the model
% The first function computes the path coefficients and the values of
% T-Statistic, p-value and confidence intervals to verify the significativity
% of the parameters

modi<-modelloInterno(zeta);
r2<-modi$R2;
beta<-modi$beta;
R2<-modi$r2;
t.statistics<-modi$t.statistics;
p.value<-modi$p.value
Corrlv<-modi$corr.latent
Inner.model<-data.frame(Path.coefficients=beta, T.statistics=t.statistics,
P.value=p.value);
rownames(Inner.model)<-path.coefficients;

#==============Communality==============================================#
% This function computes the communality index based on the correlation
% between the latent and manifest variables

com<-com(matq,zeta);
Communality<-unlist(com$communalità,recursive=TRUE);
Average.communality<-com$average.communality;

#=============Redundancy================================================#
% This index is used to measure the part of variability of the manifest
% variables associated to endogenous latent variables explained by the
% other variables of the model

red<-redundancy(matq,modi$R2,com$communalità);
Redundancy<-unlist(red$redundancy,recursive=TRUE);
Average.redundancy<-red$average.redundancy;

#===================Weights-Loadings====================================#
% This is the table of the outer estimation

W<-unlist(w.norm,recursive=TRUE);
L<- unlist(l.finali,recursive=TRUE);
Outer_model<-data.frame(Normalized.Weights=W,Loadings=L,
Communality=Communality, Redundancy=Redundancy);
rownames(Outer_model)<-manifest.names;

#==========================Summary of model=============================#
% In this table there are a summary of the results of previous indexes

Summary.inner_model<-data.frame( Latent.variables= latent.names,
Type.variable =type.variable, R.squared=r2,
Average.communality= Average.communality,
Average.Redundancy=Average.redundancy);
rownames(Summary.inner_model)<-latent.names;
```

```
#============================Unidimensionality=========================#
%This function computes the indexes to verify the
%unidimensionality of the blocks

Uni<-unidimensionality(matq,zeta);
Unidimensionality<-data.frame(Mode=Uni$Mode, C.alpha=Uni$Alpha,
DG.rho=Uni$Rho, F.eigenvalue=Uni$first.eigen, S.eigenvalue=Uni$second.eigen)
rownames(Unidimensionality)<- latent.names;

#==============================Correlation=============================#
Corr.mv<-data.frame(com$correlazioni);
rownames(com$correlazioni)<-manifest.names;
colnames(com$correlazioni)<- latent.names;

output <- list(w.norm=w.norm,l.finali=l.finali,convergenza=convergenza,
zeta=zeta, beta=beta,R2=R2,t.statistics= t.statistics, p.value=p.value,
r2=r2,Outer_model=Outer_model,Summary.inner_model=Summary.inner_model,
Unidimensionality=Unidimensionality,Corr.mv=Corr.mv, Inner.mode=Inner.model,
Corrlv=Corrlv);
output;
}
```

B.2 The functions used in PALSOS-PM

```
initialstep<-function(mat)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   initialstep computes the first estimation of the latent variables    %
%    This step is equal to the first step of PLS-PM algorithm            %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      mat: the original matrix of raw data                              %
%                                                                        %
%                                                                        %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%             v: the first estimation of the latent variables            %
%                                                                        %
%                                                                        %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
initialstep<-function(mat){

weight<-list();
% This is the initialization of the casual vector to multiply for
% the manifest variables

vdo<-list();

n <- dim(mat[[1]])[1];
numLat <- length(mat);
v <- matrix(0, n,numLat);

for(i in 1:numLat)
{
weight[[i]]<-array(1,dim(mat[[i]])[2]);
     vdo[[i]]<-matrix(0,dim(mat[[i]])[1],dim(mat[[i]])[2]);

for(j in 1:dim(mat[[i]])[2])
{
         vdo[[i]][,j]<-as.numeric(mat[[i]][,j])*as.numeric(weight[[i]][j]);
         v[,i]<-rowSums(vdo[[i]]);
         }
     }

v;
```

```
output<-list(v=v);
}
 %---------------------------------------------------------------------------%
 Innerestimation <- function(v, inner, scheme=1)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                         %
%   Innerestimation is the function that updates the                      %
%   outer estimation of the latent variables                              %
%                                                                         %
%                                                                         %
%   INPUT PARAMETERS:                                                     %
%      v: the outer estimation of the latent variables                    %
%                                                                         %
%      inner: the matrix with the inner relationships                     %
%                                                                         %
%      scheme: it indicates how we want computes the inner weights        %
%      It assumes two values: 1 if we use the centroid scheme and 2       %
%      if we use the factorial scheme                                     %
%                                                                         %
%   OUTPUT PARAMETERS:                                                    %
%             zeta: the matrix of latent variables updated                %
%                                                                         %
%                                                                         %
%                                                                         %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


 Innerestimation <- function(v, inner, scheme=1)
 {
     zeta <- matrix(0, dim(v)[1], dim(v)[2]);
     vet<-0;
     for(i in 1:dim(v)[2]) {

           idx <- trovaLegami(inner,i);
           if(schema==1){
           vet<- as.numeric(sign(cor(v[,i],v[,idx])));
           }
           else{
           vet <- as.numeric(cor(v[,i], v[,idx]));
           zeta[,i] <- rowSums(v[,idx] * vet);
           }

     zeta;

     output<-list(zeta=zeta)
}
```

```
%---------------------------------------------------------------------------%

Morals<-function(mat,zeta,itermax,epsilon,natura)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   Morals is the function that quantifies the qualitative               %
%    manifest variables and the weights of the outer estimation          %
%                                                                        %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      mat: the original matrix of raw data                              %
%                                                                        %
%      zeta: the matrix of latent variables after inner estimation       %
%                                                                        %
%      itermax: max number of iterations                                 %
%                                                                        %
%      epsilon: is the criterio of stop for the algorithm,               %
%      a positive value near zero                                        %
%  natura: the vector that explains the                              %
%      typology of variables                                             %
%                                                                        %
%      OUTPUT PARAMETERS:                                                %
%        weights: the matrix of latent variables updated                 %
%        loadings: the correlation between the manifest                  %
%  and latent variables                                            %
%  matq: the matrix of quantified variables                        %
%                                                                        %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Morals<-function(mat,zeta,itermax,epsilon,natura)
{
% This function develops the als algorithm to obtain the optimal quantification
% of qualitative variables

varind<-mat;
vardip<-as.matrix(zeta);

ios<-numeric();
sigma<-numeric();
weights<-list();
loadings<-list();
vdo<-list();
vindos<-list();
vmqs<-list();
%These are the initialization of the quantities computed in this routine
n<-dim(vardip)[2];
```

```
residui<-numeric(n);
vstim<- matrix(0,dim(vardip)[1],dim(vardip)[2]);

iteraz<-0
varianza<-0;
conv<-0;

for(i in 1:dim(zeta)[2])
    {
    vmqs[[i]]<-scale(varind[[i]]);
    }

repeat
    {
    iteraz<-iteraz+1

    for(i in 1:dim(vardip)[2])
        {
        weights[[i]] <- array(, dim(varind[[i]])[2]);
        vdo[[i]]<-matrix(0,dim(varind[[i]])[1],dim(varind[[i]])[2]);
        loadings[[i]]<- array(, dim(varind[[i]])[2]);


        for (j in 1:dim(varind[[i]])[2])
            {
            weights[[i]] [j]<- cov( varind[[i]][,j],vardip[,i])/var(vardip[,i])
            loadings[[i]][j]<-cor(varind[[i]][,j],vardip[,i]);
            vdo[[i]][,j]<-as.numeric(varind [[i]][,j])*as.numeric(weights[[i]][j]);
            }

        vstim[,i]<-rowSums(vdo[[i]]);
        }

    vstim<-scale(vstim);

    for(i in 1:dim(vardip)[2])
        {
        weights[[i]]<- array(, dim(varind[[i]])[2]);
        loadings[[i]]<- array(, dim(varind[[i]])[2]);
        vindos[[i]]<-matrix(0,dim(varind[[i]])[1],dim(varind[[i]])[2]);

        for (j in 1:dim(varind[[i]])[2])
            {
            %This is the quantification process: natura is the vector that establishes the
            %type of variables

            if (natura [[i]][j] == 1)
                {
```

```
            tempo<-as.matrix(varind[[i]] [,j]);
            G<-acm.disjonctif(tempo);
            G<-as.matrix(G);
            temp<-solve(t(G)%*%G)%*%t(G);
            temp<-temp%*%vstim[,i];
            temp<-t(temp);
            ios<-rowSums(as.matrix(G)*as.numeric(temp));
            vindos[[i]] [,j]<-as.numeric(ios);
            }

        if (natura[[i]] [j] == 2)
            {
            ios<-pava(mat[[i]][,j]);
            vindos[[i]] [,j]<-ios;
            }

        if (natura [[i]] [j]==3)
            {
            ios<-varind[[i]] [,j];
            vindos[[i]] [,j]<-as.numeric(ios);
            }

        vindosnostand<-vindos;
        vindos[[i]] [,j]<-scale(vindos [[i]] [,j]);
        weights[[i]] [j]<-cov(vindos[[i]][,j],vardip[,i])/var(vardip[,i]);
        loadings[[i]][j]<-cor(vindos[[i]][,j],vardip[,i]);
        }

    contr<- controllo.segni(weights,vardip,varind);
    %This function makes the control of the signs of weights

    for(k in 1:length(weights))
        {
        if(contr$w.sum[k] == length(weights[[k]]))
            {
            weights[[k]]<-weights[[k]]*(-1);
            }
        }

    for (j in 1:dim(varind[[i]])[2])
        {
    vdo[[i]] [,j]<-as.numeric(vindos[[i]][,j])*as.numeric(weights[[i]][j]);
        }

    vardipe<-vardip;
    vstim[,i]<-rowSums(vdo[[i]]);
    sigma[i]<-sqrt(var(vstim[,i]));
    vstim<-scale(vstim);
```

```
        residui[i]<- sum((vardip[,i]-vstim[,i])^2);
        }

    vardip<-vstim;
    varind<-vindos;
    vtemp<-0;

    for(i in 1:dim(vardip)[2])
        {
        for (j in 1: dim(varind[[i]])[2])
            {
            varianza<-var(vindos[[i]] [,j]);

            if(is.na(varianza) == T )
                {
                conv<--1;
                print(conv)
                }
            }
        }
    if(conv!= -1)
        {
        for( i in 1:dim(vardip)[2])
            {
            if (residui[i]<= epsilon)
                {
                vtemp<-vtemp+1;
                }
            }
        }

         %This is the criterion of convergence for Morals algorithm
    if(vtemp == dim(vardip)[2]|iteraz==itermax |conv== -1)  break;
    }

output<-list(varind=varind,vardip=vardip,weights=weights,loadings=loadings,
residui=residui,iteraz=iteraz,vardipe=vardipe,sigma=sigma,
vindosnostand=vindosnostand, conv=conv);
}
%----------------------------------------------------------------------------------------%
trovaLegami <- function(inner, idx)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   trovaLegami is the function that founds the relationships between    %
%   the latent variables                                                 %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      inner: the matrix with the inner relationships                    %
%                                                                        %
%      idx: indicates the presence of the relationship                   %
%      between two latent variables                                      %
%                                                                        %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%         we have the relationships between the variables on which       %
%          base the path coefficients are computed                       %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

trovaLegami <- function(inner, idx)
{
output <- c(which(inner[,idx] == 1), which(inner[idx,] == 1));
output;
}
%-------------------------------------------------------------------------%
betaVet <- function(weights, iterazione)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   betavet is used to verify the convergence of PALSOS-PM algorithm     %
%   In particular it identifies the weights associated to last iteration %
%   and the previous to compute the difference between them              %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      weights: the covariance between latent and manifest variables     %
%                                                                        %
%      iterazione: the number of iteration                               %
%                                                                        %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%        we have the weights of i-ma iteration to compare with           %
%       the weights of last itration to verify the convergence           %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

betaVet <- function(weights, iterazione)

{
output <- numeric();

nBlocchi <- length(weights[[iterazione]]);

for (i in 1:nBlocchi)
    {
    output <- c(output, weights[[iterazione]][[i]]);
    }

output;
}

%----------------------------------------------------------------------%
controllo.segni<-function(weights,vardip,varind)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   controllo.segni is the function that makes the control on the signs  %
%   of weights in the outer estimation                                   %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      weights: the covariance between latent and manifest variables     %
%                                                                        %
%      vardip: the matrix of dependent variables in Morals               %
%                                                                        %
%      varind: the matrix of explicative variables in Morals             %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%        we change the signs of weights if it is verify                  %
%        the condition imposed                                           %
%                                                                        %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

controllo.segni<-function(weights,vardip,varind)
{
w.temp<-list();
w.sum<-numeric(dim(vardip)[2]);

for(i in 1:dim(vardip)[2])
    {
    w.temp[[i]]<-array(0, dim(varind[[i]])[2]);

    for (j in 1:dim(varind[[i]])[2])
        {
        if( sign(weights[[i]][j])== -1 |is.na(weights[[i]][j]))
            {
            w.temp[[i]][j]<- 1;
            }
        else
            {
            w.temp[[i]][j]<- 0;
            }
        }
    w.sum[i]<- sum(w.temp[[i]]);
    }

output<-list(w.sum=w.sum);
output;
}
%-----------------------------------------------------------------------------------%

com<-function(matq,zeta)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                          %
%   com is the function that estimates the Communality and                 %
%   the Average Communality                                                %
%                                                                          %
%                                                                          %
%   INPUT PARAMETERS:                                                      %
%      matq: the matrix of quantified variables                            %
%                                                                          %
%      zeta: the matrix of latent variables                                %
%                                                                          %
%                                                                          %
%   OUTPUT PARAMETERS:                                                     %
%        this function returns the values of Communality for each block  %
%        and the value of Average communality and the correlation matrix %
%                                                                          %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

com<-function(matq,zeta)
{
communalità<-list();
numLat <- length(matq);
average.communality<-rep(1,numLat);
ptot <- 0;
average<-0
somma<-0;
sommacom<-0;

% p is the total number of manifest variables
p <- array(,numLat);
for (j in 1:numLat)
    {
    p[j] <-dim(matq[[j]])[2];
    ptot <- ptot + p[j];
    }

correlazioni <- matrix(,ptot, numLat);
cp <- c(0,cumsum(p));
for(j in 1:numLat)
```

```
    {
    for(k in 1:numLat)
        {
        for(i in 1: p[j])
            {
            correlazioni[cp[j]+i, k] <- cor(matq[[j]][,i],zeta[,k]) ;
            }

        rownames(correlazioni)<-manifest.names;
        colnames(correlazioni)<- latent.names;
        }
    }

for (j in 1:numLat)
    {
    communalità[[j]]<-array(0,dim(matq[[j]])[2]);
    ptemp <-dim(matq[[j]])[2];

    for (i in 1:ptemp)
        {
        communalità [[j]] [i]<-round((cor(matq[[j]][,i],zeta[,j])^2),4);
        }

    average.communality[j]<-sum(communalità[[j]])/ptemp;
    }

average<-average+sum(average.communality)/length(average.communality);
output<-list(average.communality=average.communality, communalità=communalità,
correlazioni=correlazioni,average=average);
output;
}

%--------------------------------------------------------------------------------------------%
redundancy<-function(matq,R2,communalità)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                         %
%   redundancy is the function that estimates the Redundancy and          %
%   the Average Redundancy                                                %
%                                                                         %
%                                                                         %
%   INPUT PARAMETERS:                                                     %
%      matq: the matrix of quantified variables                           %
%                                                                         %
%      R2: the R squared of the inner regression                          %
%                                                                         %
%      communalità: the values of communality for each block              %
%                                                                         %
%   OUTPUT PARAMETERS:                                                    %
%        this function returns the values of Redundancy for each block   %
%        and the value of Average redundancy                             %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

redundancy<-function(matq,R2,communalità)
{
redundancy<-list();
numLat<-length(matq);
average.redundancy<-rep(1,numLat);

for(j in 1:numLat)
    {
    redundancy[[j]]<-array(0,dim(matq[[j]])[2]);
    ptemp <-dim(matq[[j]])[2];

    for (i in 1:ptemp)
        {
        if(endo[j]==1)
            {
            redundancy[[j]] [i]<-round(communalità[[j]] [i]*R2[j],4);
            }
        else
            {
            redundancy[[j]]<-rep(0, dim(matq[[j]])[2]);
            }
        }

    average.redundancy[j]<-round(sum(redundancy[[j]])/ptemp,4);
    }

output<-list(redundancy=redundancy,average.redundancy=average.redundancy);
}
%-------------------------------------------------------------------------------%
unidimensionality<-function(matq,zeta)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   unidimensionality is the function that verifies if the block         %
%   are unidimensional                                                   %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      matq: the matrix of quantified variables                          %
%                                                                        %
%      zeta: the matrix of latent variables                              %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%        this function returns the values of Cronbach’s Alpha ,         %
%        Rho of Dillon-Goldstein and the first and second eigenvalues    %
%                                                                        %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
unidimensionality<-function(matq,zeta)
{
numLat<-length(matq);
Alpha<-rep(1,numLat);
Rho<-rep(1,numLat);
first.eigen<-rep(1,numLat);
second.eigen<-rep(1,numLat);
Mode<-rep("Reflective",8);

for(j in 1:numLat)
    {
    acp<-princomp(matq[[j]]);
    gof.om<-round(mean(cor(matq[[j]],acp$scores[,1])^2),4);
    a.numerator<-2*sum(cor(matq[[j]])[lower.tri(cor(matq[[j]]))]);
    a.denominator<- var(rowSums(matq[[j]]));
    Alpha[j]<-round((a.numerator/a.denominator)*(ncol(matq[[j]])/ncol(matq[[j]]-1)),4);
    rho.numerator<- colSums(cor(matq[[j]], acp$scores[,1]))^2;
    rho.denominator<- rho.numerator+(ncol(matq[[j]])-colSums(cor(matq[[j]],acp$scores[,1])^2));
    Rho[j]<-round(rho.numerator/rho.denominator,4);
    first.eigen[j]<- round(acp$sdev[1]^2,2);
    second.eigen[j]<-  round(acp$sdev[2]^2,2);
    }

output<-list(Alpha=Alpha, Rho=Rho, first.eigen= first.eigen,
second.eigen=second.eigen,Mode=Mode);
}
%----------------------------------------------------------------------------------------------%
Gof<-function(matq, zeta, average.communality)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   Gof is the function that computes the Goodness of Fit Index          %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%      matq: the matrix of quantified variables                          %
%                                                                        %
%      zeta: the matrix of latent variables                              %
%                                                                        %
%     average.communality: the average communality                       %
%     computes by the com function                                       %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%        this function returns the values of Gof index                   %
%                                                                        %
%                                                                         %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Gof<-function(matq, zeta, inner, average.communality){

numLat<-length(matq);
gof.absolute<-round(sqrt(mean(com$average.communality)*sum(modi$r2)/sum(endo)),4);
gof.outer<-round(mean( com$average.communality/uni$gof.om),4);
Gof<-data.frame(Gof=c("Absolute", "Outer.model"),
value=c(gof.absolute,gof.outer))
output<-list(Gof=Gof);
output;
}
%----------------------------------------------------------------------------------%

modelloInterno <- function(zeta)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                         %
%   modellointerno is the function that estimates the path coefficients   %
%   According to the model it is different, and this is an example        %
%                                                                         %
%                                                                         %
%   INPUT PARAMETERS:                                                     %
%                                                                         %
%      zeta: the matrix of latent variables                               %
%                                                                         %
%   OUTPUT PARAMETERS:                                                    %
%       It returns the path coefficients,                                 %
%       the R squared, the p-value and the T-Statistic                    %
%                                                                         %
%                                                                         %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
modelloInterno <- function(zeta) {
               z2 <- lm(zeta[,2] ~ zeta[,1]);
               z3 <- lm(zeta[,3] ~ zeta[,2]+ zeta[,1]);
               z4 <- lm(zeta[,4] ~ zeta[,2] + zeta[,3]);
               z5 <- lm(zeta[,5] ~ zeta[,1] + zeta[,2]+ zeta[,3]+ zeta[,4]);
               z6 <- lm(zeta[,6] ~ zeta[,5]+zeta[,1]);
               beta <- array(,11);
               r2 <- array(,5);

               beta[1] <- z2$coefficients[-1];
               beta[2:3] <- z3$coefficients[-1];
               beta[4:5] <- z4$coefficients[-1];
               beta[6:9] <- z5$coefficients[-1];
               beta[10:11] <- z6$coefficients[-1];
r2[1] <- summary(z2)$r.squared;
r2[2] <- summary(z3)$r.squared;
r2[3] <- summary(z4)$r.squared;
r2[4] <- summary(z5)$r.squared;
r2[5] <- summary(z6)$r.squared;
R2<-c(0,r2);

output <- list(beta=beta, r2=r2,R2=R2);
output;
}
```

B.3 The bootstrap validation

```
bootstrap.validation<-function(mat,m)
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                                                        %
%   bootstrap.validation is the function that develops m                 %
%   resampling of the initial matrix. On each resampling                 %
%   the model is estimated. In this way we obtain an empirical           %
%   distribution of the structural parameters                            %
%                                                                        %
%                                                                        %
%   INPUT PARAMETERS:                                                    %
%                                                                        %
%      mat: the original matrix                                          %
%      m: the number of resampling                                       %
%                                                                        %
%   OUTPUT PARAMETERS:                                                   %
%     The confidence intervals for outer and inner                       %
% structural parameters, the values of T-Statistics                  %
%                                                                        %
%                                                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 replboot<-m;
 numLat<-length(mat);
 n<-0
 matb<-as.matrix(mat[[1]]);
 weigths.boot<-numeric();
 matlista<-list();
 convergenza<-0;
 sample<-0;
 correlation<-list();
  for (i in 2:numLat){
    matrice<-as.matrix(mat[[i]]);
    matb<-cbind(matb,matrice);
    }
    W.boot<-matrix(NA,m,ncol(matb));
    L.boot<-matrix(NA,m,ncol(matb));
    P.boot<-matrix(NA,m,7);
    R2.boot<-matrix(NA,m,3);

      for (i in 1:numLat){
        for (j in 1:dim(mat[[i]])[2]){
            while(n<m){
              boot.obs<-sample(1:nrow(matb), nrow(matb), replace=TRUE);
              matboot<-as.matrix(matb[boot.obs,]);
              matlista<- creazionelista(matboot)$matlista;
              risultati.boot<-alspmRifl(matlista,inner,0.00001,100);
                  if(risultati.boot$convergenza==1 ){
                     n<-n+1
                     w.boot<- unlist(risultati.boot$w.norm[[1]]);
                     l.boot<- unlist(risultati.boot$l.finali[[1]]);
                     sample<-sample + 1;
  for (i in 2:numLat){
                           weigths.boot<-unlist(risultati.boot$w.norm[[i]]);
                           w.boot<- c(w.boot, weigths.boot);
                           loadings.boot<- unlist(risultati.boot$l.finali[[i]]);
                           l.boot<- c(l.boot, loadings.boot);
                          }
                      W.boot[n,]<-w.boot;
                      L.boot[n,]<- l.boot;
                      mI<- modelloInterno(risultati.boot$zeta);
                      P.boot[n,] <- mI$beta;
                      correlation[[sample]]<-mI$corr.latent;
                      R2.boot[n,] <- mI$r2;
                      }
               }
              }
              }

          t.pb <- matrix(NA, 7, 2);
          t.lb <- matrix(NA,ncol(L.boot), 2);
          l.original<- unlist(prova$l.finali,recursive=TRUE);
          w.original<-unlist(prova$w.norm,recursive=TRUE);
          t.wb <- matrix(NA,ncol(W.boot), 1);
          t.rb <- matrix(NA, 3, 1);

          for (j in 1:ncol(W.boot)){
            t.wb[j,] <- round((w.original[j]/sd(W.boot[,j])),4)
            Weights<-data.frame(Original=w.original,
            Mean.Boot=round(apply(W.boot,2,mean), 4),
            Std.Err=round(apply(W.boot,2,sd),4),t.statis=t.wb[,1],
            intconf.025=round(apply(W.boot,2,function(x) percentili(x, 0.025)) , 4),
            intconf.975=round(apply(W.boot,2,function(x) percentili(x, 0.975)) , 4))
           rownames(Weights)<-manifest.names;
          }
         for (j in 1:ncol(L.boot)){
           t.lb[j,] <- round((l.original[j]/sd(L.boot[,j])),4)
           Loadings <- data.frame(Original=l.original,
           Mean.Boot=round(apply(L.boot,2,mean), 4),
           Std.Err=round(apply(L.boot,2,sd),4), t.statis=t.lb[,1],
           intconf.025=round(apply(L.boot,2,function(x) percentili(x, 0.025)) , 4),
           intconf.975=round(apply(L.boot,2,function(x) percentili(x, 0.975)) , 4))
           rownames(Loadings)<-manifest.names;
```

```
     }
 for (j in 1:ncol(P.boot)){
   Path.coefficient<- data.frame(Original=prova$beta,
   Mean.Boot=round(apply(P.boot,2,mean), 4),
   Std.Err=round(apply(P.boot,2,sd),4), T.Statistics= prova$t.statistics,
   Pr = prova$p.value,
   intconf.025=round(apply(P.boot,2,function(x) percentili(x, 0.025)) , 4),
   intconf.975=round(apply(P.boot,2,function(x) percentili(x, 0.975)) , 4))
   rownames(Path.coefficient)<-path.coefficients;
 }
 for (j in 1:ncol(R2.boot)){
   t.rb[j,] <- round((prova$r2[j]/sd(R2.boot[,j])), 4)
   R2<- data.frame(Original=prova$R2,
   Mean.Boot=round(apply(R2.boot, 2, mean), 3),
   Std.Err=round(apply(R2.boot,2,sd),3), t.statis=t.rb,
   intconf.025=round(apply(R2.boot,2,function(x) percentili(x, 0.025)) , 4),
   intconf.975=round(apply(R2.boot,2,function(x) percentili(x, 0.975)) , 4))
 }

par(mfrow=c(4,4));
hist(P.boot[,1])
hist(P.boot[,2])
hist(P.boot[,3])
hist(P.boot[,4])
hist(P.boot[,5])
hist(P.boot[,6])
hist(P.boot[,7])
output<-list( Weights=Weights, Loadings=Loadings,
Path.coefficient=Path.coefficient,R2=R2);
output;
}
```

B.4 The construction of the model

```
%These are the commands to write in the prompt of R to specify the model,
%the latent blocks and the suddivision of the manifest variables in these blocks.
%It is also specified the names of latent and manifets variables,
%the nature of the manifest variables
%and the type of latent variable (endogenous or exogenous).



mat<- read.table(".txt", header=T,sep="\t")
%This is the data matrix to read and it must be in a txt format

l1<- data.frame(mat[,""].....)

l2<-data.frame(mat[,""],....)

l3<-data.frame(mat[,""],....)

% This is the specification of the latent blocks (for example three):
%in this command it is necessary to specify the
$manifest variables of the mat matrix
mat<- list();
mat[[1]] <- l1;
mat[[2]] <- l2;
mat[[3]] <- l3;

%This is the matrix subdivides in latent blocks

inner <- matrix(0,number of latent variables,number of latent variables);
inner[2,1] =1;
inner[3,2] =1;

%This is the inner matrix in which there are
%specified the relationships between the latent variables

natura<-list()

%The list in which there is specified the nature of
%each manifest variable

latent.names<- c("", .....)
manifest.names<- c("‘"’,....)
path.coefficients<- c("", .....)
type.variable<- c(rep("Exogenous",..), rep("Endogenous",..))
endo<-c(....)
```

```
%These commands are: the vector of names of latent and manifest varisbles,
%the names of pathe coefficients, the typology of latent variables and
%the specification across 0 (exogenous) and 1 (endogenous)
of the typology of the latent variables
```

Bibliografia

The code of PALSOS-PM algorithm

Bibliography

[1] Andersen, EB., (1995) Polytomous Rasch Models and their Estimation. In *G Fischer, I Molenaar (eds.), Rasch models: Foundations, Recent Developments, and Applications*, pp. 271-292. Springer, New York.

[2] Andrich, D., (1978) A Rating Formulation for Ordered Response Categories. In *Psychometrika*, 43, pp. 561-573

[3] Barlow, R.E., Bartholomew, D.J., Bremer, J. M., and Brunk, H.D., (1972) *Statistical inference under order restrinction.* New Tork; Wiley

[4] Baker, F., Kim, SH.(2004)*Item Response Theory.* Marcel Dekker, New York, 2nd edition.

[5] Benzecri, J.P., (1973)*L'analyse des donées: T 2, l'analyse des correspondaces.* Paris: Dunod

[6] Bradley, R.A., Katty, S.K., Coons, I.J., (1962) Optimal scaling for ordered categories. In *Psycometrika*, 27, pp. 355-374

[7] Brown, B., Benedetti, J., (1977), On the Mean and Variance of the Tetrachoric Correlation Coefficient. In *Psychometrika*, 42, pp. 347-355

[8] Chin, W. W., (2000) Frequently Asked Questions-Partial Least Squares and PLS-Graph. In *http://disc-nt.cba.uh.edu/chin/plsfac.htm. Chin, W. W. (2001).* PLS-

Graph User's Guide Version 3.0. C. T. Bauer College of Business, University of Houston, Houston, Texas.

[9] Corbetta, P., (1999) *Metodologia e tecniche della ricerca sociale.* Il Mulino

[10] de Leeuw,J., Young, F. W., Takane, Y.(1976) Additive structure in qualitative data: An alternating least squares method with optimal scaling features. In *Psycometrika*, 41, pp. 471-503

[11] de Leeuw,J.,(1977) Correctness of Kruskal's algorithms for monotone regression with ties. In *Psycometrika*, 42, 1, pp. 141-144.

[12] de Leeuw, J.,(1984) The GIFI system on nonlinear multivariate analysis. In *Data Analysis and Informatics*, E. Diday, M.Jambu, L.Lebart, J.-P.Pagés, and R. Tomassone (Eds.), 3, pp. 415-424. Amsterdam: NorthHolland

[13] de Leeuw, J.,(2005) Monotonic regression. In *BS Everitt, DC Howell (eds.),Encyclopedia of Statistics in Behavioral Science.* Wiley, New York,3, pp. 1260-1261.

[14] de Leeuw,J.(2006) Nonlinear Principal Component Analysis and Related Techniques. In *M.Greenacre, J.Blasius (eds.),Multiple Correspondence Analysis and Related Methods.*Chapman Hall/CRC, Boca Raton, FL, pp. 107-133

[15] de Leeuw,J., van Rijckevorsel, J., (1980), Homals and Princals. Some generalizations of Principal Components Analysis. In *Data Analysis and Informatics*, E . Diday et al. (eds.), North Holland Publishing Company, pp. 231-241

[16] Fisher, R.,(1938) *Statistical methods for research workers* (10th ed.),Edinburg: Oliver and Boyd

[17] Fu, J.R., (2006a). VisualPLS - Partial Least Square (PLS) Regression - An Enhanced GUI for Lvpls (PLS 1.8 PC) Version 1.04. National Kaohsiung University of Applied Sciences, Taiwan, ROC.

[18] Fu, J.R., (2006b). VisualPLS - Partial Least Square (PLS) Regression - An Enhanced GUI for Lvpls (PLS 1.8 PC) Version 1.04. http://www2.kuas. edu.tw/prof/fred/vpls/index.html.

[19] Gifi, A.,(1990) *Nonlinear Multivariate Analysis.* Chichester:Wiley.

[20] Hansohm, J., (2007), Algorithms and error estimations for monotone regression on partially preordered sets. In *Journal of Multivariate Analysis*, 98, pp.1043-1050.

[21] Henseler, J., Ringle, C. M., Sinkovics, R.R., (2009) The use of Partial Least Squares Path Modeling in International Marketing. In *Advances in International Marketing*,20, pp. 277-319

[22] Guttman, L., (1941) The quantification of a class of attributes: a theory and methods of scale construction. In P. Host et al.,*The prediction of personal adjustement*, Soc. Sci. Res. Council, NY.

[23] Guttman, L., (1944) A basis for scaling qualitative data. In *American Sociological Rewiev*, 9, pp. 130-150

[24] Guttman, L., (1968) A general non metric technique for finding the smallest coordinate space for a configuration of points. In *Psycometrika*, 33, pp. 469-506

[25] Guttman, L., (1955) The determinacy of factor score matrices with implications for five others basic problems of common factor theory. In *British Journal of Mathematical and Statistical Psychology*, 8, pp.65-81

[26] Jakobowicz, E., Derquenne C.,(2007) A modified PLS Path Modeling algorithm handling reflective categorical variables and a new model building startegy. In *Computational Statistics and Data Analysis*, 51, 8, pp. 3666-3678

[27] Joreskog, K.G., (1973) A general method for estimating a Linear Structural Equation System. In *Goldberger and Duncan*, pp. 85-112

[28] Joreskog, K.G., (1978) Structural analysis of covariance and correlation matrices. In *Psycometrika*,43, pp.443-477

[29] Kruskal, J.B.,(1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In *Psycometrika*,29, pp. 1-27

[30] Kruskal, J.B.,(1964) Nonmetric multidimensional scaling: a numerical method. In *Psycometrika*, 29, pp. 115-129

[31] Kruskal, J.B.,(1965), Analysis of factorial exeperiments by estimating monotone transformations of the data. In *Journal of Royal Statistical Society*, Series B, 27, pp. 251-263

[32] Lauro, C., Nappo, D., (2008) Rapporto Tecnico *Multivariate Analysis of OVC survey data and Decision Support tools.* OVC Project AVSI second survey

[33] Lee S., Poon W., Bentler P., (1992)Structural Equation Models with continuous and Polytomous variables. In *Psycometrika*, 57, pp. 89-105

[34] Li, Y., (2003) PLS-GUI User Manual - A Graphic User Interface for LVPLS (PLS-PC 1.8) - Version 1.0. University of South Carolina, Columbia, SC.

[35] Li, Y. (2005) PLS-GUI - Graphic User Interface for Partial Least Squares (PLS-PC 1.8) - Version 2.0.1 beta. University of South Carolina, Columbia, SC

[36] Lohmöller, J.B., (1984) LVPLS Program Manual - Version 1.6. Zentralarchiv für Empirische Sozialforschung, Universität zu Köln, Köln.

[37] Lohmöller, J.B., (1987) PLS-PC: Latent Variables Path Analysis with Partial Least Squares - Version 1.8 for PCs under MS-DOS.

[38] Lovaglio, P.G., (2000)*Modelli con variabili latenti e indicatori di tipo misto.* Phd Thesis, University of Trento

[39] McLachlan, J., Peel, D., (2000)*Finite Mixture Models.* Wiley, New York.

[40] Mevik, B. H., Wehrens, R., (2007) The PLS Package: Principal Component and Partial Least Squares Regression in R. In *Journal of Statistical software*, 18, 2, pp. 1-23

[41] Michailidis, G.,de Leeuw, J., (1996) The Gifi system of Nonlinear Multivariate Analysis. In *UCLA Statistical Series Preprints*,204

[42] Muthen, B., (1984) A general Structural Equation Model with dichotomous, oredered categorical and continuos latent variable indicators. In *Psycometrika*, 49, pp. 115-132

[43] Olsson, U., (1979) Maximum Likelihood Estimation of Polychoric Correlation Coefficient. In *Psychometrika*, 44, pp. 443-460

[44] Olsson, U., Drasgow, F., Dorans, N., (1982) The polyserial correlation coefficient. In *Psycometrika*,47,3, pp.337-347

[45] Rasch, G.,(1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen.

[46] Ringle, C. M., Wende, S., and Will, A., (2005)SmartPLS - Version 2.0. Universität Hamburg, Hamburg.

[47] Saporta, G., (1983) Multidimensional Data Analysis and Quantification of Categorical Variables. In *New Trends in Data Analysis and A pplications*,J.Jansen, J.F. Marcotorchino, J.M. Proth (eds.), pp. 73-97

[48] Saporta, G.,(1975) *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.* Phd Thesis, University Pierre et Marie Curie (Paris VI)

[49] Sellin, N., (1989) PLSPATH - Version 3.01. Application Manual. Universität Hamburg, Hamburg.

[50] Takeuchi, K., Yanai, H., Mukherjee, B. N.,(1982) *The foundations of Multivariate Analysis.* Wiley Eastern Limited.

[51] Test and Go (2006). Spad Version 6.0.0. Paris, France.

[52] Tenenhaus, M.,(1977) *Modele Lineaire et Analyse Canonique lorsque les variables sont Heterogenes.* In COREF Note de Travail n.15

[53] Tenenhaus, M.,(1984) L'analyse canonique généralisée de variables numériques, nominales ou ordinales par des méthods de codage optimal. In *Data Analysis and Informatics*, E. Diday, M.Jambu, L.Lebart, J.-P.Pagés, and R. Tomassone (Eds.) Amsterdam: NorthHolland, 3, pp. 71-84

[54] Tenenhaus ,M.,(1988) Canonical Analysis of two convex polyhedral cones and applications. In *Psycometrika*, 53, 4, pp. 503-524

[55] Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., and Lauro, C.,(2005) PLS path modeling. In *Computational Statistics and Data Analysis*, 48,1, pp. 159-205

[56] Thurstone, L.L., (1927) The unity of measurement in educational scales. In *Journal of educational psychology*, 18, pp. 505-524

[57] Torgerson, W.S., (1958) *Theory and methods of scaling.* Wiley, NY.

[58] Tukey, J. W., (1977)*Exploratory Data Analysis.* Addison-Wesley.

[59] van der Burg, E., de Leeuw, J.,(1983) Nonlinear canonical correlation. In *British Journal of Mathematical and Statistical Psychology*, 36, pp. 54-80

[60] van der Burg, E., de Leeuw, J., Verdegall, R., (1984)*Nonlinear canonical correlation with m sets of variables.* Leiden: University of Leiden, Department of Data Theory

[61] Verdegall, R., (1986)*Overals.* Leiden: University of Leiden, Department of Data Theory

[62] Wold, H., (1982) Soft modeling: The basic design and some extensions. In Jöreskog, K. G. and Wold, H., editors, Systems Under Indirect Observation. Part II, pp. 1-54. North-Holland, Amsterdam

[63] Wold, H., (1985) Partial least squares. In Kotz, S. and Johnson, N. L., editors, Encyclopaedia of Statistical Sciences Wiley, New York, 6, pp. 581-591

[64] Young, F.W.,(1975) Methods for describing ordinal data with cardinal models. In *J. Math. Psychol.*, 12 pp. 416-436

[65] Young, F.W.,de Leeuw, J.,Takane, Y.,(1976) Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. In *Psycometrika*,41, 4, pp. 505-529

[66] Young, F.W.,Takane, Y.,de Leeuw, J.,(1978) The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. In *Psycometrika*, 43, pp. 279-281

[67] Young ,F.W.,(1981) Quantitative analysis of qualitative data. In *Psycometrika*, 46, 4, pp. 357-388

[68] Zanella, A.,(2006) *Modelli di misurazione e causalità. Appunti delle lezioni di Statistica matematica (Il modulo)*. Diritto allo Studio, Università Cattolica