Universitat Politècnica de Catalunya

PATHMOX Approach:
Segmentation Trees in Partial Least Squares Path Modeling

Doctoral Dissertation by
Gastón Sánchez Trujillo

Advisor: Tomàs Aluja Banet

Departament d'Estadística i Investigació Operativa

PATHMOX Approach:
Segmentation Trees in Partial Least Squares Path Modeling

Doctoral Dissertation
by
Gastón Sánchez Trujillo

Tomàs Aluja Banet
Advisor

Defended in Barcelona, July 22, 2009

Jury:

Josep Casanovas Garcia	Universitat Politècnica de Catalunya
Roser Rius Carrasco	Universitat Politècnica de Catalunya
Vincenzo Esposito Vinzi	ESSEC Business School
Ludovic Lebart	Centre National de la Recherche Scientifique (CNRS)
Wynne W. Chin	Bauer College of Business, University of Houston

Vist i plau

Director del Departament **Cap d'Administració**

Resolució del Tribunal

Reunit el Tribunal designat a l'efecte, el doctorant / la doctoranda exposa el tema de la seva tesi doctoral

PATHMOX APPROACH: SEGMENTATION TREES IN PARTIAL LEAST SQUARES PATH MODELING

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres del Tribunal, aquest atorga a la tesi esmentada la qualificació d'/de EXCEL·LENT CUM LAUDE

Barcelona, 22 de juliol de 2009

President (Jose Casanovas Garcia) Secretarària (Roser Rius Carrasco)

Vocal (Vicenzo Esposito Vinzi) Vocal (Ludovic Lebart) Vocal (Wynne Chin)

Abstract

Sometimes we must face the fact that the variables and concepts of interest in our models cannot be observed nor measured directly. In these cases we refer to them as theoretical constructs or latent variables. These variables are very common in social sciences. For instance, psychologists speak of satisfaction, sociologists refer to social status, and economists speak of economic development. When researchers work with theoretical concepts they usually conceive of expected relationships between two or more latent variables, they analyze the relationships, and propose theories and models. For such purposes structural equation modeling (SEM) is a statistical methodology with great flexibility and modeling power.

Partial Least Squares Path Modeling (PLS-PM), also known as Structural Equation Modeling by PLS approach, is a methodology of multivariate data analysis that allows for modeling complex cause-effect relationships involving latent and observed variables. Developed by Herman Wold, PLS-PM was designed as a complementary technique to the covariance-based framework of SEM. Currently, typical applications of PLS-PM can be found within marketing and management studies especially those related with customer satisfaction and other types of intangibles measurement.

Traditionally, SEM approaches assume homogeneity over the entire set of observations without considering any group structure. However, this assumption is unrealistic in many cases; for example, in consumer behavior research sources of heterogeneity can be due to customer age or gender. Analysts distinguish between two sources of heterogeneity: observed and unobserved. Heterogeneity is observed if it is possible to define segments based on an observed variable. Heterogeneity is unobserved when the variables that cause heterogeneity in the data are unknown beforehand. If population heterogeneity is not taken into account, conventional analysis may lead the analyst to inadequate results with a serious risk of drawing poor conclusions.

In this dissertation we propose the PATHMOX approach which has been specifically designed to be used when observed sources of heterogeneity are available. Since much of the work on SEM depends on survey-based data, this type of studies usually provide sources of observed heterogeneity that can be used to detect different path models of segments in the population.

Inspired by the segmentation scheme used in decision trees, PATHMOX produces a segmentation tree that has a similar structure to a binary decision tree. The main characteristic of the obtained tree is that each node corresponds to a different segment with its own particular path model. The aim of the algorithm is to select, among a set of segmentation variables (i.e. observed sources of heterogeneity), those having superior discriminant capacity in the sense that they separate the path models as much as possible. The split criterion in this case is used to decide whether two confronted

structural models can be considered to be different. For this purpose, an *F* statistic-based test for assessing the equality of two regression models has been adapted for comparing two structural models by testing the equality of their path coefficients.

In order to evaluate the sensitivity of the split criterion used in PATHMOX we have run a series of simulation studies. The goal is to assess the capabilities of the *F*-test when two path models are compared under different experimental conditions such as sample size, level of noise of disturbance terms, path coefficients, difference in variance of endogenous constructs, and data distributions. The results provide important evidence in favor of the adequate performance of the proposed *F*-test when it is applied for comparing two structural models in presence of non-normal data and skewed distributions. However, unbalanced segments and differences in the variance of the endogenous constructs may affect the sensitivity of the *F*-test.

In regards to the practical aspect, two applications of PATHMOX with real data are described. The first application has to do with customer satisfaction. The second application involves a study on job satisfaction and motivation. We analyze the data using *Visual Pathmox* which is a program specifically designed to provide a graphical interface to calculate PLS path models and PATHMOX segmentation trees. Both analyses show the intuitive scheme, ease of interpretation, and meaningful description of PATHMOX with its potential to identify unexpected models for segments in the population.

Keywords: Partial Least Squares Path Modeling, Structural Equation Modeling, Segmentation Trees, Pathmox Approach.

Acknowledgements

Although this has been probably the easiest page to write in this thesis, it is without a doubt one of the most important parts. Some years ago, different reasons made me decide to look for graduate studies in Statistics without knowing that this intention would turn into a PhD. Likewise, different people contributed in one way or another to make me follow that goal. Now that I have come to the end of my studies, the following people are the ones that have had the most significant influence in the completion of this project:

First of all I want to express my gratitude and appreciation to my advisor, Dr. Tomàs Aluja Banet. I am indebted to him for welcoming me to the Laboratory of Information Analysis and Modeling as well as for the guidance he afforded me during my doctoral research. Not only he was readily available for me, but he always answered my doubts and responded to my enquiries. I deeply appreciate his assistance, trust, and encouragement with this project. He is an awesome advisor and as we say in Mexico: “es un chingón”. To him: *Moltíssimes gràcies.*

I am thankful to Oriol Serch for his invaluable assistance and helpfulness with the development of *Visual Pathmox*. His effort, creativeness, and technical skills have been crucial for the completion of this work.

My thanks go to the graduate students of the Department of Statistics and Operations Research at the Universitat Politècnica de Catalunya. Since the first moment they have been good friends and I have especially enjoyed the several multidisciplinary discussions and coffee meetings we have had a long these years. Thanks for all the shared moments (good and bad ones) and for their friendship.

I am particularly grateful to Jessica Trowbridge, who not only was willing to read most of the manuscripts and provided many corrections, but also for her unconditional love, and for being here even when she was there.

The completion of the final drafts of this dissertation was made easier through a couple of absence periods I took in the last 10 months. Thus I gladly express my gratitude to Dr. John Matsui and all the staff members of the Biology Scholars Program at the University of California Berkeley. Thanks for the hospitality during my two visits, for treating me as a privileged *invader*, and for the space and environment you kindly provided me which have definitely helped me to finish my thesis.

Last but not least, I want to thank my brother Gonzalo for his help in the designing of the cover.

Una imperiosa necesidad que soy incapaz de ignorar me obliga a escribir estas líneas en mi lengua materna. Han pasado ya casi cinco años desde que abandoné todo lo que tenía (y hacía) para emprender un nuevo camino y dedicarme a este proyecto con el cual cierro un ciclo más en mi vida. Este tiempo en Barcelona como estudiante de doctorado deja en mí, tanto personal como profesionalmente, una huella imborrable y un valioso tesoro irremplazable. No todo ha salido como esperaba. Mis planes originales de obtener sólo el Diploma de Estudios Avanzados fueron modificándose poco a poco mientas más crecía en mí el interés y el gusto (¿pasión?) por las técnicas de análisis de datos. No sé en qué momento se torcieron mis planes, pero me alegro de que se hayan torcido de esa manera. Puedo decir con toda certeza que mis expectativas han sido inmensamente sobrepasadas y que me siento totalmente satisfecho. No puedo dedicar esta tesis a nadie más que a mis padres, Beatriz y Raúl, por todo el amor y el cariño que me han dado a lo largo de toda mi vida. No tengo palabras para agradecerles toda su comprensión, así como el apoyo incondicional que han sabido brindarme siempre.

Preface

Success is often the result of taking a misstep in the right direction
(Al Bernstein)

Brick walls are there for a reason. They're not there to keep us out. They are there to give us a chance to show how badly we want something
(Randy Pausch)

The first time I heard about *PLS* was in October 2004 when Professor Tomàs Aluja asked me if I knew the meaning of that term. I didn't want to look like a fool but after a couple of seconds thinking about those three mysterious letters I couldn't find any answer. I have to admit it: I didn't have the slightest idea about PLS. Lucky for me that didn't prevent Professor Aluja from asking me to help him with the organization of the 4^{th} *International Symposium on PLS and Related Methods* held in Barcelona in September of 2005.

In this peculiar way I started my involvement with PLS, and it became more profound in March 2005 when I attended the course on *Advanced Methods of Multivariate Data Analysis* taught by Professor Aluja. We used Michel Tenenhaus' *La Régression PLS* as the textbook of the course, and I remember programming in *R* trying to replicate most of the book's results, while recovering my knowledge of basic French. Again, I was very lucky to be able to attend that course since that was the last time Dr. Aluja had the opportunity to teach it. It was a pity the course was removed from the program syllabus, but that's another story.

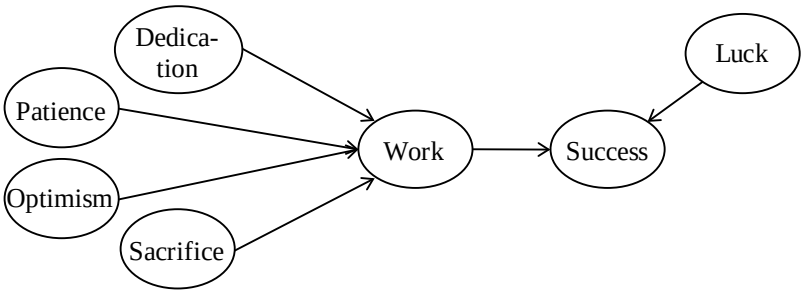
The 2005 PLS symposium reflected the wide dissemination and consolidation of Partial Least Squares as a powerful data analysis and statistical modeling technique. It showed a great number of successful applications in different disciplines, as well as new interesting research developments. These included the scarcely explored topic related to segmentation issues in PLS Path Modeling. This motivating and attractive area offered promising opportunities and, consequently, I began my research project on segmentation issues in *PLS-PM* under the supervision of Dr. Aluja at the Laboratory of Information Analysis and Modeling.

One of the initial goals was the development of an *R* package to perform PLS path modeling analysis. I wasn't really sure about the feasibility of that goal but I started to work on it with nothing but my willpower and imagination. As I was advancing with the programming code and functions, my project was also expanding its scope. Finally, it was decided to use the functions I had in *R* as a baseline for an academic software program in Java with a graphic interface called *Visual Pathmox*.

Almost five years have passed since I started my graduate studies in Statistics and Data Analysis. Since then, many approaches about segmentation, multi-group analysis and latent class detection in PLS-PM have been proposed. This only proves the interest that it has awakened in academics and the enormous relevance in practical applications, being still open for more developments and contributions. Finally, after completing the

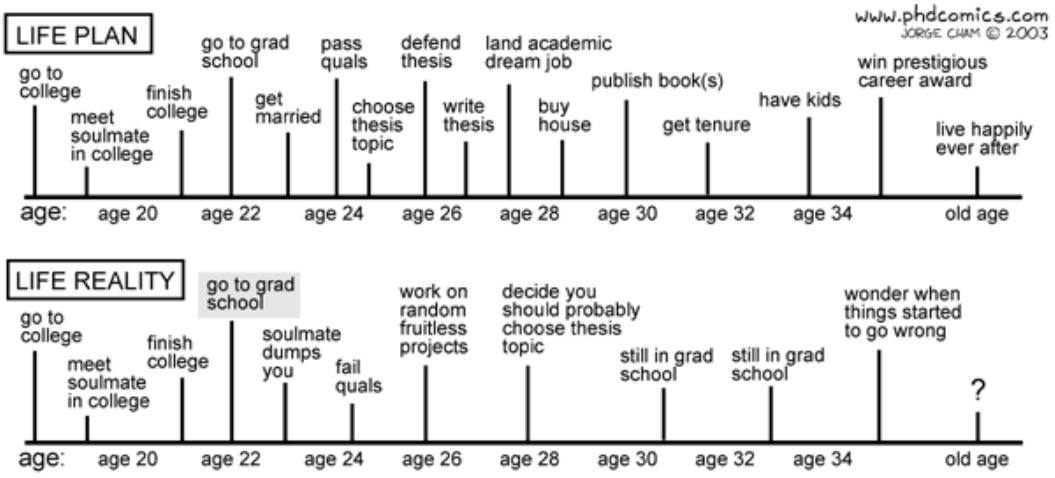
thesis manuscript and following a hunch, I decided to create the plspm package in *R* as a *decision of last minute* in order to collaborate and add my part to the continuously growing PLS community.

From a retrospective point of view I honestly believe I made one of the biggest "missteps" in my life when I took the right direction coming to Barcelona. Finally, "failure after failure", I have finished my doctoral research without ever loosing the interest and enthusiasm. I think my own model would be:



If this work can provide a small contribution to the field of PLS Path Modeling, my *latent* wish by writing this dissertation will be definitely achieved. In turn, my *manifest* wish is to hope for only positive applications (if any) of this work.

Gastón Sánchez Trujillo
Barcelona, May 2009



(*From Náhuatl Language*)

cuahuitl = tree

cuauhmaitl = branch

cuauhtlapalli = leaf

nelhuatl = root

ohtli = path

xexeloa = to divide

moxexeloa = divide in two parts

namox = my book

PATHMOX

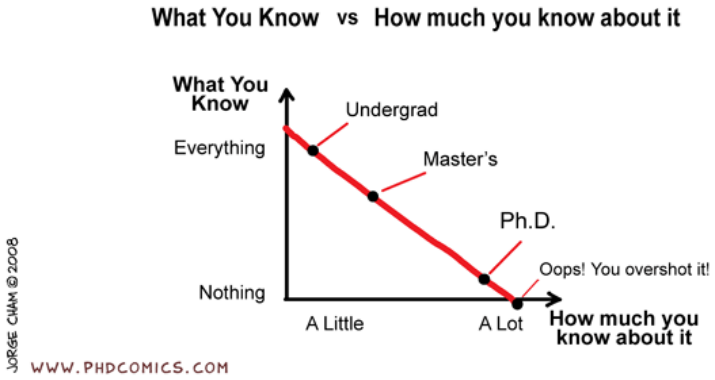


Table of Contents

Chapter 1

Introduction

In the context of statistical data analysis, segmentation is essentially the division of a set of objects (or individuals) into distinct groups. However, it is also a broad term that covers a wide variety of techniques and of problems among which one can find the challenging issue of dealing with segmentation tasks in Partial Least Squares Path Modeling (PLS-PM).

Developed by Herman Wold (1982b, 1985a) Partial Least Squares Path Modeling (Lohmöller, 1989; Tenenhaus *et al*, 2005), also known as Structural Equation Modeling by Partial Least Squares approach, is a methodology of multivariate data analysis that allows for modeling complex cause-effect relationships involving latent (unobserved) and observed variables. Generally speaking, these models seek to analyze the underlying causal process that is assumed to generate some phenomenon of interest.

The great potential of Structural Equation Modeling (SEM) relies on the fact that it can deal with latent variables. These variables correspond to unobserved theoretical concepts that cannot be measured in a direct way. Instead, they have to be indirectly measured through well observed variables. Latent variables are very common in social and behavioral sciences. Psychologists speak of satisfaction and motivation. Sociologists refer to social status. Economists speak of organizational performance. Even in biological sciences it is possible to find latent variables such as the concept of habitat structure. In all cases, researchers conceive of expected causal relationships between latent variables, and structural models are proposed to analyze those relations. In order to estimate the structural models analysts dispose of two main approaches (Tenenhaus, 2008): covariance-based SEM and component-based SEM. These kinds of methods are considered multivariate techniques of a second generation (Fornell, 1982, 1987) that provide an insight scheme by combining causal modeling with data analysis features.

Covariance-based SEM, also known as Covariance Structure Analysis (CSA), has been mainly developed by Karl Jöreskog (1973). CSA (Long, 1983; Bollen, 1989) is usually employed for hypothesis testing and model validation. Component-based SEM is comprised of Partial Least Squares Path Modeling on the one hand and Generalized Structured Component Analysis (GSCA) on the other hand.

PLS-PM was developed as a complementary approach to the covariance-based SEM. In turn, GCSA has been recently proposed by Hwang and Takane (2004) as an alternative to PLS-PM. The common characteristic of both component-based techniques is that latent variables are obtained as linear combinations (as in component analysis) of the observed variables.

Partial Least Squares Path Modeling has become a research topic of enormous interest for many statisticians during the last decade. It has also been adopted as the preferred approach for structural equation modeling among an increasing number of researchers. As a result, PLS-PM has encountered a growing popularity across many disciplines and research areas such as education (Sellin, 1995), sensory analysis (Pagès and Tenenhaus, 2001), operations management (Brown and Chin, 2004; Raymond and St-Pierre, 2005), information technology and systems (Mathieson et al, 2001; Komiak and Benbasat, 2006), marketing (Ashill *et al*, 2005), human resources (Eskildsen *et al*, 2004b; Bontis and Serenko, 2007), and business management (Bontis, 1998; Bart *et al*, 2001; Bontis, 2004; Calvo-Mora *et al*, 2006; Cabrita and Vaz, 2006).

Particularly in marketing, the most typical application has to do with customer satisfaction measurement (Hackl and Westlund, 2000; Martensen *et al*, 2000; Kristensen *et al*, 2001; Westlund *et al*, 2001; López *et al*, 2003; Vilares and Coelho, 2003; Johnson *et al*, 2006). Today, Customer Satisfaction Marketing studies can be considered a landmark for PLS-PM as well as an experimental field, and is becoming the main developmental arena for a number of PLS contributions, proposals and innovations like those found in Cassel *et al* (1999); Stan and Saporta (2003); Eskildsen *et al* (2005).

1.1 Motivation

One of the current research topics in PLS-PM that is mainly motivated by marketing applications is the issue of segmentation. Because of its problem-solving potential, the segmentation challenge has attracted great interest from statisticians and data analysts. For this reason, much of the research work has evolved around the problem of identifying segments. Hence, segmentation approaches in PLS-PM are proving to be one of the most influential methodological developments spawned by PLS researchers to date.

The goal and purpose of segmentation tasks, not only in PLS-PM but also in structural equation modeling, is the same as in any other field: it is used to group individuals into segments with similar characteristics. Hence, segmentation procedures involve examining whether the population (or sample) is homogeneous or heterogeneous. However, one of the common assumptions when estimating structural equation models is to suppose homogeneity over the entire set of individuals. In other words, the analyst treats all individuals alike without considering any group structure. This assumption does not represent a serious concern *per se*, however most of the times it is reasonable to put into doubt sample homogeneity and it is logical to suppose groups having different behavior. For example, consider marketing and consumer behavior research in which potential sources of heterogeneity can be due to brand awareness, product class knowledge, or customer preferences (Dilon, 1990). Moreover, it is very frequent to find heterogeneity defined in terms of demographic and psychographic variables (e.g., gender, groups of age, and marital status).

The problem with failing to considering the possible existence of segments in the population is that conventional SEM may lead the analyst to obtain inaccurate-inadequate results. Since the model for the entire population may be misspecified, the analyst runs the risk of drawing erroneous or poor conclusions (Dilon, 1990; Yung, 1997). Thus, to overcome this situation it is necessary to assume population heterogeneity.

One can distinguish two types of heterogeneity in SEM: observed heterogeneity and unobserved heterogeneity (Lubke and Muthén, 2005). Heterogeneity is observed if subpopulations can be defined based on observed variables. Heterogeneity is unobserved when the variables that cause the heterogeneity in the data are unknown beforehand. For each class of heterogeneity, one can find a general scheme of segmentation: partitioning techniques for observed heterogeneity, and grouping techniques for unobserved heterogeneity. Partitioning approaches consist of dividing the data set into smaller sets according to the observed sources of heterogeneity. In other words, the data set can be divided in segments, and separate models can be estimated for each segment. In contrast, grouping approaches consist of forming groups of elements based on their proximity which is determined by some predefined measure. Since the sources of heterogeneity are assumed to be unknown, some kind of clustering-based procedure is performed to detect population segments.

In this dissertation, we develop a new PLS-PM segmentation approach when heterogeneity is observed. We present the PATHMOX algorithm for obtaining what we call *segmentation trees of PLS path models*. PATHMOX is motivated by the opportunities provided in survey research, and by the need of researchers and practitioners to have automated (or semi-automated) methods to analyze their data. Our rationale is based on the fact that most of the path modeling analyses depend on data from survey-based studies. These surveys usually contain variables that may provide sources of observed heterogeneity. For instance, much of the data for customer satisfaction and other similar analysis (e.g. employee satisfaction) include segmentation variables such as socio-demographic variables (e.g. age, gender, level of studies, etc).

1.2 Contributions

This dissertation presents a new approach to PLS path modeling segmentation called PATHMOX that has specifically been designed to be used when heterogeneity is observed. PATHMOX is inspired from the segmentation scheme used in decision trees. It adapts the basic idea behind binary segmentation processes in order to produce a segmentation tree in which each node has an associated path model with its own set of parameter estimates. The algorithm seeks, among a set of segmentation variables (i.e. observed sources of heterogeneity), those having superior discriminant capacity in the sense that they separate the path models as much as possible.

The designing of an algorithm to produce a segmentation tree of path models requires the development of a split criterion in order to decide whether two confronted structural models can be considered to be different. For this purpose, we propose as split criterion an F statistic to compare structural models. The F statistic is based on a hypothesis test exposed by Lebart, Morineau and Fénelon (1979, 1985) for testing the equality of two regression models. In fact, we have adapted the test to compare two structural models by testing the equality of path coefficients between them. The goal is

to identify path models for different population segments formed with the help of the observed sources of heterogeneity (i.e., the segmentation variables).

The functions to estimate path models and pathmox trees have been programmed to be run in R. However, our main concern was the lack of a user friendly graphical interface in R. Therefore, with the aim of disposing of graphical displays to visualize the PATHMOX results, as well as enabling users to draw path diagrams for their models, an academic software tool has been developed in collaboration with other members of the Laboratory of Information Analysis and Modeling (Barcelona School of Informatics). The result of this collaboration has been the creation of *Visual Pathmox* (Serch, 2008).

Finally, as a derived result of this doctoral project, the R package "plspm" has been created with the purpose of providing a function to perform PLS path modeling analysis in R. The package also contains various methods of the PLS framework such as the NIPALS algorithm, PLS Regression 1 and 2, and PLS Canonical Analysis. It is freely available from CRAN and more methods will be included in new versions of the package.

1.3 Outline

This dissertation is divided into two main parts. Part I (Chapters 2-5) is comprised with a general overview of the structural equation modeling framework with emphasis on Partial Least Squares Path Modeling. Part II (Chapters 6-10) exposes the issue of segmentation in PLS-PM and the presentation of the proposed PATHMOX approach.

Part I proceeds as follows: Chapter 2 is dedicated to presenting an overview of the historical development path modeling and its PLS approach. This review aims to offer a wide panorama of the evolution of PLS-PM that will clarify some confusing topics related to this methodology. Chapter 3 is dedicated to presenting the foundation of path modeling. We describe the fundamental concept of latent variables, and we discuss the ways in which such variables are measured. We also provide the symbols notation and the description of the terminology employed in path modeling. In Chapter 4, the Covariance-based approach of Structural Equation Modeling is examined. We describe the multivariate methods related to the traditional approach in SEM such as Exploratory Factor Analysis, Confirmatory Factor Analysis and Covariance Structure Analysis.

The conceptual background and foundations of Partial Least Squares Path Modeling are described in Chapter 5. Although this chapter is focused on the PLS-PM methodology, some of the basic elements of component-based SEM are discussed. The first sections are dedicated to the review of the method of Principal Components Analysis, the Singular Value Decomposition, and the NIPALS algorithm. In the second part, the PLS-PM methodology is presented and explained in detail. The last section contains a brief overview of the Generalized Structured Component Analysis.

Part II begins with Chapter 6 which aims to provide the state of the art in PLS-PM segmentation approaches. Firstly the background of segmentation in PLS-PM is discussed with its undeniable relationship to marketing segmentation. Then, we describe the approaches for dealing with observed heterogeneity such as the Re-sampling parametric approach (Chin, 2000), the Re-sampling non-parametric approach (Chin, 2003), the Moderation testing approach (Chin *et al*, 1996, 2003), the Henseler's approach (Henseler, 2007), and the Possibilistic PLS-PM approach (Palumbo and Romano, 2008). Finally, the approaches for dealing with unobserved heterogeneity are

revised: Finite Mixture PLS (Hahn *et* al, 2002; Ringle *et al*, 2005), PLS Typological Path Modeling (Squillacciotti, 2005; Squillacciotti *et al*, 2006), Response Based Units Segmentation for PLS-PM (Trinchera *et al*, 2007; Esposito Vinzi *et al*, 2007, 2008), and the PLS Genetic Algorithm Segmentation (Ringle and Schlittgen, 2007).

Chapter 7 is dedicated entirely to the presentation of PATHMOX. In the first sections we provide some basic notions of segmentation trees and binary decision trees. The main part of the chapter contains the explanation of the PATHMOX algorithm as well as a description of the proposed statistic that is used as split criterion. This criterion consists in an F statistics for testing the equality of path coefficients when two path models (obtained from the binary splits) are compared. In order to evaluate the performance of the split criterion used in PATHMOX a series of Monte Carlo simulation studies are described in Chapter 8. The purpose is to assess the capabilities of the F-test when comparing two inner path models under different experimental conditions.

Two applications of the PATHMOX approach with empirical data are examined in Chapter 9. The first application involves a customer satisfaction study, whereas the second application involves a model on employee satisfaction-motivation. Finally, chapter 10 provides the conclusions obtained in this research work, and the possible research lines for further exploration.

In Appendix III, the reader can find a brief description of the R package "plspm" with its main function of the same name (plspm) and a short example with the typical model of the European Customer Satisfaction Index.

Chapter 2

Historical Review

Partial Least Squares Path Modeling (PLS-PM), also known as Structural Equation Modeling by the Partial Least Squares approach, is integrated by two main concepts: (1) the concept of path modeling or structural equation modeling, and (2) the concept of partial least squares. Although the concept of partial least squares appears later than that of structural equation modeling, its history and development can be seen as a process over a long period of time that covers many fields of knowledge such as biometrics, psychometrics, econometrics, and sociology, among others.

The contributions from each discipline have created a comprehensive literature in which is not unusual to find different terms referring to the same concept and, conversely, find identical names used to designate different concepts. This lack of uniformity in terminology is the main cause for the generalized confusion many readers can experience within PLS-PM related literature. In order to clarify some of the doubts and confusions regarding PLS and structural models, while contributing to a better understanding on the topic, this chapter presents a historical review in the evolution process of the structural equation models and the subsequent development of the PLS approach.

We have tried to link the different disciplines and their associated methods with the purpose of providing the context in which they arose, as well as the practical problems they sought to solve. Although we have attempted to offer a review as comprehensive as possible, our presentation does not pretend to be a complete and exhaustive history of PLS; something that is virtually impossible. We have highlighted only some of the historical facts and events that we consider to be the most important or relevant in regards to the development of path modeling.

The first section contains a brief biography of the “father” of PLS: Herman Wold, followed by a description of each of the four different backgrounds related to structural equation modeling. The final sections will cover the antecedents of PLS, describe its evolution, and will mention some recent developments within the PLS field.

2.1 Brief Biography of Herman Wold

Herman Ole Andreas Wold was born at Skien, Norway, on December 25th, 1908; he was the youngest of a family of six brothers and sisters. In 1912 his family moved to a small town near Stockholm and he lived in Sweden for the rest of his life. It was there that he began his higher education at Stockholm University in 1927. He also obtained his doctoral degree in 1938 under the supervision of Professor Harald Cramér (Whittle, 1992). He continued as Docent in actuarial mathematics and mathematical statistics until 1942 when he was offered the chair of Statistics at the University of Uppsala. In the summer before presenting his doctoral dissertation he was appointed by a government committee to perform an econometric analysis of consumer demand from available Swedish statistics. That project took him almost 14 years to complete, from 1938 to 1952.

His interest in path models began by the late 1950s and led him to the organization of the Uppsala Symposium on Psychological Factor Analysis in 1953. However, during the 1950s and early 1960s his research focused on econometric models with directly observed variables, as well as on recursive and non-recursive systems of equations trying to estimate those equations by least squares methods.

While serving as visiting professor at Columbia University from 1958-1959, Wold suffered from “an intellectual turmoil” (Wold, 1982b) and decided to change some of his intellectual ideas and make a new start with causal modeling analysis. He dedicated most of his time to path models with latent variables and to the development of the partial least squares methodology from 1966 till his death. Wold continued in his position as chair at Uppsala until 1970, when he moved to Gothenburg University and became there, the chair of the Statistics Department. Upon retirement in 1975 he returned to live in Uppsala. He kept active doing research for the Volkswagen Foundation and spent some time at the University of Geneva. Herman Wold died on February 16th, 1992.

Professor Wold was Vice-President of the International Statistical Institute from 1957 to 1961, he was elected to membership of the Swedish Academy of Sciences in 1960, received an honorary fellowship of the Royal Statistical Society in 1961, was elected president of the Econometric Society in 1966, member of the Nobel Economic Science Prize Committee from 1968 to 1980, and honorary membership of the American Economic Association and the American Academy of Arts and Sciences in 1978 (Wold, 1982b). During his retirement he also received several honorary doctorates (Hendry and Morgan, 1994).

2.2 Econometrics and Simultaneous Equations

The rapid industrialization process experienced by the United States and some European countries during the last decades of the nineteenth century brought spectacular rates of economic growth together with some periods of crisis. These circumstances led to an increased interest among economists in demand analysis, production theory and business cycle analysis (Gilbert and Qin, 2005). According to Desrosières (2004), from these circumstances two main schools of thought emerged in Economics: one based on economic theories and the other based on observations and records of empirical data. The first, inspired in the model of physical sciences, assumed the existence of *a priori* general principles and laws pretending to describe a deterministic representation of

economic phenomena. Conversely, the second school of thought stated that economical laws could only be extracted from regularities in data. In other words, there were two separated ideologies: the deductive one, and the inductive one.

As a result, a number of quantitative studies of business cycles were developed trying to understand business fluctuations and cycle lengths. However, the unsophisticated methods employed to study economic phenomena gave poor results and only partial explanations were given for the distinct irregularities in business fluctuations (Gilbert and Qin, 2005). Although simple correlation, basic statistics, and regression analysis were known and applied by some, these methods could not fill all the gaps between data and theory. In other words, there was no well established framework for synthesizing data evidence and economic theory. The need for this general mathematical framework to help conceptualize and describe the economy became increasingly clear in the initial decades of the twentieth century when early innovations were proposed as *ad hoc* methods to handle particular problems.

In 1930 the Econometrics Society was founded with the objective of analyzing economic phenomena while integrating statistical methods and mathematical modeling. As Morgan (1990) mentions, there were three main objectives of the Econometrics Society: (1) to make economics more scientific, (2) to express theories more exactly, and (3) to provide stronger, empirically based knowledge. Two years later, Alfred Cowles created the Cowles Commission for Research in Economics in 1932. Cowles was a businessman and an investment counselor, who after the stock market crash of 1929, decided to support academic research in order to understand the workings of the economy (Christ, 1994).

Also in the 1930s, the statistical concepts and the apparatus of hypothesis testing came into econometrics with the Neyman-Pearson method of rejection. This approach inspired econometricians who began combining economic theory, statistical methods, and observed data to model particular aspects of economic behavior, such as the price of food, consumer demand, or consumer income. The idea was to describe economic phenomena as the result of interactions among many agents, i.e., as the result of the *simultaneous* interaction of different agents. These relationships were expressed by a system of simultaneous equations capable of describing the workings of the economy (Christ, 1994).

One of the econometricians inspired by the theory test modeling approach was the Norwegian Trygve Haavelmo (Nobel Prize in 1989). He focused on the formulation of *a priori* theories as a set of admissible hypotheses and the integration of the Maximum Likelihood principle into econometrics. Haavelmo dismissed ordinary least squares (OLS) regression as inconsistent when applied to simultaneous equations and instead recommended estimation by Maximum Likelihood. However, Wold and Bentzel in 1946 distinguished the conditions under which a system of simultaneous equations can be estimated by OLS.

Many important developments in econometrics took place during the 1930s and 1940s due in part to the incorporation of the recently established probabilistic framework, as well as economic events such as the stock market crash of 1929, the implementation of public policies to maintain full employment, the studies on supply-demand analysis, and the studies on rationing policies and regimes of regulated prices during the wartime and after World War II (Wold, 1982b).

Much of the econometricians work was centered in model building with suitable systems of equations, and the development of methods for estimating them. New estimation methods were needed and the maximum likelihood approach took the

attention in the econometrics environment. It is important to note that there had been some early successes with other estimation methods like *path analysis* proposed by Sewall Wright, the son of economist Philip Wright. Unfortunately, this method went unnoticed by most economists.

2.3 Sewall Wright's Path Analysis

As mentioned above, various *ad-hoc* methods were proposed to solve different economic problems. One of those methods, Path Analysis, was presented under circumstances that unfortunately nullified its impact on the econometrics field, having to wait for a couple of decades to be rediscovered by sociologists. The econometrics application of this method appeared in the form of an appendix in a study made by Philip Wright in 1928 about market equilibrium, and supply and demand curves. The appendix was written by his son Sewall Wright, a geneticist at the University of Chicago, who some years earlier had seen that market equilibrium could be analyzed using the method he had developed to study certain problems in heredity (Epstein, 1989).

Path analysis was developed by the American geneticist Sewall Wright in the 1920s as a method for dealing with a system of interrelated variables (Wright, 1972). In 1911, Sewall Wright was a graduate student of Biology at the University of Illinois. In 1915, Wright was working at the Animal Husbandry Division of the United States Department of Agriculture, where he was investigating the role of genetics in the determination of color inheritance (Denis and Legerski, 2006). In particular, he was interested in ascertaining how the genes of the parents (the causes) influenced the offspring's traits (the effects); more concisely he wanted to estimate the sizes of the effects from each parent to the offspring (Murayama, 1998). He came up with a solution by establishing a system of equations in which each equation was expressed in terms of the correlations among the various variables. He started to develop a quantitative method called Path Analysis designed to estimate the degree to which a given effect was determined by each of a number of causes. The first use of this methodology occurred in 1918 in an article titled *On the Nature of Size Factors*. Basically, what Wright proposed in this article was that growth factors of different broadness (e.g., general size, size of skull, size of leg, etc.) may have an effect on the size of diverse bones (skull, tibia, femur, etc.) and induce variability among them (Tomer, 2003) . In concrete, Wright was able to establish the importance of the size factors by calculating the proportion of variation determined by a factor (cause) in the size of a body part or bone (effect).

This method was motivated by the question of whether a set of variables had a causal structure that could be determined from a matrix of simple correlation coefficients. The fundamental contribution of path analysis, and the reason for its name, is the graphical representation of the relationships among the variables by means of a path diagram. As an example of a simple path diagram let us consider a system with three variables X_1, X_2 and X_3, where X_2 is a linear function of X_1, and X_3 is a linear function of X_1 and X_2. This simple system, represented in figure 2.1, is modeled by: $X_2 = bX_1$ and $X_3 = aX_1 + cX_2$.

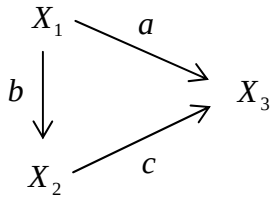


Figure 2.1. *Path diagram example*

The way to think of a path diagram is as if it were the causal flow in a system of interrelated variables. Wright's 1918 article did not contain the conceptualization of path coefficients nor the path-analytic diagrams. It was until 1920 when the notion of path coefficients and path diagrams appeared in a second article written by Wright titled *The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs*. Wright not only used his path analysis for genetics and econometrics applications in the estimation of supply and demand equations, but also in psychometrics factor analysis (Bollen, 1989).

2.4 Psychometrics and Factor Analysis

Psychometrics is a branch of Psychology devoted to the development, evaluation, and application of mental tests with the purpose of measuring knowledge, attitudes, personality traits, and abilities (Rust and Golombok, 1999). One of the statistical data analysis techniques that originated within psychometrics is Factor Analysis. This technique was developed by the British psychologist Charles Spearman at the beginning of the XXth century in order to measure intelligence in an objective way. The idea was to study how mental ability is organized by analyzing correlation matrices for sets of cognitive test variables. Spearman hypothesized that the correlations could be described by a single underlying variable called the factor of "General Intelligence".

Spearman's assumed factor of general intelligence was soon found to be inadequate and the model was then modified in order to include more factors. This modification was not immediate but took place over the four decades following Spearman's initial work. The result was the development of an enormous variety of different factor analysis methods (Kotz and Johnson, 1983). Among the psychologists who contributed to this modification was Louis Thurstone who postulated seven, rather than one, primary mental ability, and developed multiple factor analysis.

Currently, the purpose of factor analysis is to explain most of the variability among a number of variables that are directly observable in terms of a smaller number of unobservable variables called factors. The observable variables are modeled as approximately linear functions of the factors (Loehlin, 1987). One of the main contributions of factor analysis is the concept itself of a factor which is an unobservable variable or latent variable (LV). The interest in LVs is that, although they represent theoretical constructs (unobserved), they can be measured indirectly by observable or manifest variables (MV).

Different estimation procedures were proposed during the 1920s and the 1930s but it was not until 1940 that efficient statistical methods were introduced by Lawley using the method of maximum likelihood (ML). However, there were two problems that had to be solved before ML was to become a feasible method of estimation. First, the

estimating equations for the ML estimates are obtained from the derivatives of matrix functions which are very complex and difficult to treat at the level of scalar algebra. Thus it was necessary to simplify the notations for matrix derivatives. Second, the likelihood equations cannot be solved algebraically, rather they require iterative numerical algorithms. In addition, electronic computers were not available before the 1950s.

In the early 1950s Herman Wold had met with the famous psychometrician Louis Thurstone and his wife Thelma in Stockholm. Wold was so inspired by their work on factor analysis that he decided to organize in 1953 the Uppsala Symposium on Psychological Factor Analysis. A few years later, the young student Karl Jöreskog, after completing his studies to be a grammar school teacher, decided to help out a friend during a summer by replacing him in the statistics laboratory at the University of Uppsala. At the end of the course, due to his aptitude for statistics, professor Wold invited Jöreskog to stay at the university and continue as a PhD student in the statistics department. In 1958 Wold suggested that Jöreskog do a dissertation on factor analysis (Jöreskog, 2004).

By the late 1950s a number of computer programs were available for performing factor analysis however none of these provided the correct maximum likelihood estimates for the common factor model. Major advancements in solving the problems concerning factor analysis and the ML approach came from the efforts of Karl Jöreskog. When he moved to the United States at the Educational Testing Service in the middle 1960s, he developed an efficient algorithm for the computation of the ML estimating equations (Mulaik, 1986).

2.5 Sociology and Causal Models

By the 1950s, the concept of causation had been reintroduced into the main discourse of the philosophy of science through several independent and almost simultaneous publications by different sociologists and economists such as Paul Lazarsfeld, Herbert Simon and Herman Wold (Sobel, 1992). The notions of model and causality attracted many researchers who were trying to model sociology after the physical sciences, that is, they were searching for theoretical universal laws of society that would mimic those of physical sciences. As Nollmand and Strasser (2005) have commented: “There was a wish for sociological scientism… A strong concern with methodology promised to cure sociology’s inferiority complex on its way into academia and to provide equal strength in the competition of scientific disciplines”

Herbert Simon, consultant of the Cowles Commission and Nobel prize winner in Economy 1978, was one of such researchers who thought that the social sciences “needed the same kind of rigor and the same mathematical underpinnings that had made the hard sciences so brilliantly successful” (Simon, 1996). Simon was familiar with Sewall Wright’s path analysis and during the period from 1950 to 1955 he focused on the concept of causal ordering, i.e. the directionality of variables when proposing a model. During this time he came in contact with the work of Herman Wold who impressed him with the use of the concept of causality (Bernert, 1983). Basically, Simon was interested in establishing the conditions under which a correlation between two variables provides sufficient evidence for inferring causal relations between them.

In 1954 Simon published his seminal article *Spurious correlation: a causal interpretation* in the *Journal of the American Statistical Association*. Inspired on

Wright's path analysis, Simon showed how, under certain assumptions, correlation might indicate causation. The idea was to check whether a particular causal model was consistent with a particular pattern of correlations. The importance of this article was its significant impact on sociological thinking regarding issues of causality, and its influence on other members of the sociological community.

Among the various sociologists influenced by Simon's thoughts was Hubert M. Blalock (Tomer, 2003) who was interested in establishing causation among variables by means of using statistical modeling. One of the contributions of Blalock was the expansion of Simon's work by considering more complex models and by examining partial correlations. By the early 1960s, Wold's causal modeling was incorporated into Blalock's theory and he began writing on causal models placing great emphasis on the prior theorizing that it is central to causal analytic methods. In 1964 Blalock published his book on *Causal Inferences in Nonexperimental Research*, considered one of the most influential works in the sociological field. In *Causal Inferences*, Blalock outlined methods for making causal inferences from correlational data as well as outlining problems in confirming these relationships. The concepts of Simon and Blalock were combined in the approach known as the Simon-Blalock method which is a technique of hypothesizing a theory and then testing it with correlational data, serving as an appropriate method for rejecting proposed models (Hackler, 1970).

A somewhat separate evolution of causal method culminated in Dudley Duncan's introduction of path analysis to sociology. In 1963, Otis Dudley Duncan and Robert Hodge were concerned with relating factors of education and occupational mobility. Their research examined antecedents of success in attaining education and jobs, by measuring variables such as social class of the family, past academic achievement, and social support as predictors of success. Their paper titled *Education and Occupational Mobility: A Regression Analysis*, can be considered the first application of Path Analysis in sociology although it was just a "primitive version of a path diagram" (Duncan, 1974).

Later, in 1964, after reading Blalock's book, Duncan decided to study Wright's path analysis seriously. As Duncan (1974) himself recognizes "It occurred to me that the specification of the Simon-Blalock approach was the same as the one Wright used". Duncan started to work in his paper *Path Analysis: Sociological Examples* that was finally published in 1966 which also included some suggestions made by the same Sewall Wright. By this time, Duncan had initiated correspondence with the econometrician Arthur Goldberger who had noted the similarity between sociological causal models and the simultaneous equation models used in econometrics. They began to collaborate and after various discussions they established that there were no real differences between Wright's approach and the one used in econometrics. Moreover, Duncan convinced Goldberger that sociologists were using methods that in fact were in both econometrics and psychometrics. Finally, they showed that path analysis models were closely related to the simultaneous equations models of econometrics, and the confirmatory factor analysis of psychology (Bernert, 1983).

It can be seen that around the 1960s there was a growing interest among sociologists for developing approaches that moved towards the causal understanding and modeling of non-experimental data. As Tomer (2003) indicates "methodological (errors in variables, multiple indicators, causality, testing theories) and philosophical considerations (measurement of theoretical concepts) intertwined" to generate an increasing interest in causal models that allow the development of a methodical framework to these topics.

Different publications on the application of path analysis began to bring together the confirmatory models of sociologists and the simultaneous equations models of econometricians. However, these first applications lacked a global approach for general application (Bentler, 1986).

The overall framework for general applicability was reached by Karl Jöreskog. Together with Duncan and Goldberger, Jöreskog had also seen that generalizations of the confirmatory factor analysis model led to a more general class of models involving analysis of covariance structure models. The collaboration of Duncan and Goldberger involved the organization in 1970 of a conference in Madison, Wisconsin, to which Karl Jöreskog was invited (Wolfe, 2003). It was in this conference that Jöreskog (1973) presented the first formulation of the Covariance Structure Analysis (CSA) for estimating a *linear structural equation system* which came to be known later as LISREL. The papers of the conference were published in Goldberger and Duncan's (1973) *Structural Equations Models in the Social Sciences* which included Jöreskog's (1973) paper in which he unified factor analysis, analysis of covariance structures, and linear structural equations modeling in a single general model (see Mulaik, 1986).

2.6 Development of PLS

As stated before, most of Wold's econometric work was related to estimation methods for simultaneous equations. However, unlike most of his contemporary colleagues, he always preferred to use methods based on least squares rather than maximum likelihood. Wold studied different estimation techniques using iterative procedures from which he developed a special method called the *fix-point algorithm* (Wold, 1973). Generally speaking, the fix-point (FP) method consists of an iterative algorithm of ordinary least squares (OLS) regressions to estimate the parameters of multi-equation systems.

In 1964, after one of his seminars on the FP method at the University of North Carolina, Wold decided to modify his algorithm in order to expand it for calculating principal components. As he himself recognized, this modification was accomplished through the remarks made from one of the participants of a conference, which gave Wold "the clue for computing principal components by an iterative procedure" (Wold, 1982a). Then, as an immediate step he also applied the algorithm to compute Hotelling's canonical correlations. This new method was first called NILES (Nonlinear Iterative Least Squares) (Wold, 1966a) but some time later it was changed to Nonlinear Iterative Partial Least Squares (NIPALS) (Wold, 1973a). With this algorithm Wold showed how to calculate principal components by means of an iterative sequence of simple ordinary least squares (OLS) regressions, as well as how to calculate canonical correlations with an iterative sequence of multiple OLS regressions.

In 1971, inspired by the results presented by Jöreskog on path models with latent variables, he realized that principal components and canonical correlation analysis could also be interpreted as path diagrams. He then started to analyze the possibility of estimating such models by adapting an appropriate generalization of his algorithms for principal components and canonical correlations. It took him five years of experimenting before PLS took its definitive shape which he called *Soft Modeling*. Herman Wold gives the end of 1977 as the birth date of PLS (Geladi, 1988). Wold presented his "soft model basic design" (Wold, 1982b) for the PLS estimation algorithm as an alternative to LISREL avoiding many of the restrictive hypotheses underlying

maximum likelihood estimation techniques, i.e. multivariate normality and large samples (*hard modeling*). The concept of soft modeling technique refers to the overall situation for which PLS Path Modeling (PLS-PM) was conceived: working with non-experimental data, with complex information and data matrices with a large number of variables, with a lack of a solid prior theoretical knowledge and which was intended for causal-predictive modeling (Wold, 1969).

Despite the of advantage flexibility claimed over Covariance Structure Analysis, PLS-PM was not extensively used in econometrics nor in social sciences. Even though both approaches were developed at nearly the same time, their subsequent evolution was rather far from being parallel. The main reason for the divergence between both techniques is related with their software availability. CSA was provided with the LISREL program since the early 1970s and improved versions were released with ease-of-use interfaces. In contrast, PLS-PM lacked of a computer program for many years until Lohmöller's *LVPLS ver1.8* program appeared in 1987.

During the 1980s Jan-Bernd Lohmöller spent many years under Wold's advice and guidance, focused on the study of the PLS-PM methodology and its capabilities. As a result of his research, Lohmöller (1989) published a comprehensive treatise on PLS titled *Latent Variable Path Modeling with Partial Least Squares*. Also in the 1980s, research interests in PLS shifted from social sciences to applications in chemistry into what is now known as chemometrics (application of statistical methods to chemical data). The person responsible for this transition was Svante Wold, the son of Herman Wold. Svante together with Harald Martens developed yet another branch of the PLS techniques in analytical chemistry known as PLS regression (PLS-R).

2.6.1 The American Trend: Bookstein, Fornell, and Chin.

Although Professor Herman Wold visited different American universities and he dictated various seminars, the initial spread of PLS in USA was mainly carried out by Fred Bookstein. Professor Bookstein, an American biometrician, is the principal creator of morphometrics, a specialty that combines techniques of geometry, computer science, and multivariate statistics for analysis of biological shape variation, shape difference, and body parts (Bookstein *et al*, 1985). Being familiar with the work of Sewall Wright, Bookstein got involved with PLS through the study of path analysis, becoming interested in the geometrical aspects of PLS (Bookstein, 1982; 1990) and the singular value decomposition (Bookstein, 1986). He had developed a special PLS method called PLS singular vector analysis, which is considered to be an intermediate technique between PLS-PM and PLS regression.

While working as a professor at the University of Michigan, Bookstein introduced Claes Fornell (Fornell 2007, pp. 24), a business professor, to the PLS-PM methodology. Professor Fornell, a Swedish economist, was already working with structural equation models by CSA approach and he was especially interested in marketing applications. In 1982, Bookstein and Fornell published an article on marketing applications in customer satisfaction with a comparison of LISREL and PLS-PM. Since then, professor Fornell has been working on marketing applications of PLS-PM particularly in the topic of customer satisfaction measurement with the creation of the American Customer Satisfaction Index (ACSI).

Another important researcher related to the evolution of PLS-PM in USA is Professor Wynne Chin, who received his PhD in Computers and Information Systems

(Graduate School of Business Administration) from the University of Michigan. Together with Fornell, Chin is one of the leading references on PLS-PM with applications in marketing and Information Science. He developed the *PLS-Graph* software which is a windows-based package (based on Lohmöller's *LVPLS*) provided with a graphical user interface.

2.6.2 PLS and the French-style Data Analysis

Besides the direction taken by PLS-PM in the USA, a rather different evolution took place in Europe. As previously mentioned, a different version of PLS was developed by the Scandinavian chemometrics community with the development of PLS regression by Svante Wold and Harald Martens among others. Its widespread use and success in the chemical industry, accompanied by a powerful software, eclipsed the PLS structural equation modeling background. The multidimensional approach of PLS-R and its capability to deal with multiple data tables, large number of variables, prediction purposes, and missing data, attracted the attention of French statisticians in the beginnings of the 1990s. Professor Michel Tenenhaus was one of those statisticians and he became the primary researcher involved with the study of PLS techniques. During his research project on PLS, and due to his interests in business and management applications of data analysis techniques, Tenenhaus was referred to the work of Claes Fornell about the applications of PLS-PM in marketing and customer satisfaction. As a result of his amazing research, he published his book *La Régression PLS* in 1998.

The reason that explains the interest in the PLS techniques among the French data analysts is found in the general framework of the so-called French-style data analysis developed by Jean-Paul Benzécri. Data analysis *à la française* was developed based on a philosophical basis that accentuates the development of models that fit the data, rather than the rejection of hypotheses based on the lack of fit. Therefore, there are no statistical significance tests that are customarily applied to the obtained results. French-style data analysis (1) works with large amounts of data (large sets of variables), (2) it focuses not only on the variables but it also emphasizes the importance of individual observations, (3) it has an exploratory approach with geometric tools, and (4) it summarizes the information contained in data into new variables called components. These goals are accomplished by using dimension reduction techniques with multivariate projection methods such as component-based methods and cluster analysis among others. The idea is to reduce the complexity (dimensions) of highly dimensional data and provide the means with which to identify patterns or subjects in the data. In other words, data analysis is based on an inductive scheme. The underlying philosophy of the French standpoint can be summarized in one of the five data analysis principles of Benzécri (1973): "the model must follow the data, not the other way around" (i.e, the model must fit the data, and not vice versa). It is remarkable that these techniques rely exclusively on simple mathematical tools (linear algebra, regarding data from a geometrical point of view, without requiring a probabilistic model).

However, the strong descriptive and exploratory focus of French data analysis was not enough to have a complete understanding of reality. Explanation was also needed and it was important to take into account causal assumptions which are usually extracted from prior experience and/or some previously established theory. Thus, French analysts saw that the gap between exploration and explanation could be covered by PLS techniques without giving up the inductive scheme, geometrical context, and

component-based (projections) methods. Hence, the connection between PLS and the general framework of the French style data analysis (multidimensional exploratory analysis) facilitated the rediscovery of the PLS approach of structural equation models with latent variables. It can be said that this situation gave a new impulse to the PLS techniques which have received further acknowledgement since the late 1990s. This impulse is reflected in the establishment of International Symposia on PLS and Related Methods in Jouy-en-Josas, France, in 1999; Anacapri, Italy, in 2001; Lisbon, Portugal, in 2003; Barcelona, Spain, in 2005; and As, Norway, in 2007. The first Symposium on PLS and Related Methods was organized by Michel Tenenhaus (HEC management school) and Alain Morineau (DECISIA group). Initially it was supposed to be a one time conference but Vincenzo Esposito Vinzi and Carlo Lauro from the University of Naples Federico II proposed to establish the tradition of holding the PLS symposiums every two years. In addition to the PLS symposia, another proof of the impulse on PLS-PM is the existence of several PLS-PM software solutions (*SmartPLS, XLSTAT-PLS, PLS-Graph, SPAD-PLS, PLS-GUI, VisualPLS*) that offer user friendly interfaces and interesting methodological capabilities.

2.7 Some remarks

The development of partial least squares path modeling was a long process, in which different aspects of the method were extracted from other methods and refined over a long period of trial and error. In the 1960s Herman Wold developed his NIPALS algorithm to calculate principal components and canonical correlations, and began applying different versions of the algorithm to various types of problems. This process took several years through which Wold and his research team experimented with different models before acquiring its definitive shape. The term *NIPALS modeling* can be found in the first applications of path models with latent variables (Wold, 1973), however, when the final version was presented at the end of 1977 Wold changed the *NIPALS modeling* term to the term of *Soft Modeling* (Wold, 1982a).

The name NIPALS was then used only to designate the algorithm for calculating principal components, whereas the application of the PLS algorithm to path models with latent variables was called soft modeling. However, during the 1980s some publications on soft modeling appeared under different names like *Partial Least Squares in Structural Modeling* in Fornell and Bookstein (1982), or *Partial Least Squares Path Analysis* in Noonan and Wold (1988). Also in the 1980s, the PLS techniques took a separate development in chemometrics with Heman's son Svante Wold, and Harald Martens (Wold, 2001). The result of this new application was the PLS Regression (PLS-R).

The present term of "PLS Path Modeling" has been accepted following a suggestion from Harald Martens, in order to differentiate PLS-PM from PLS-R. Additionally, Svante (Wold *et al*, 2004) has given another conception to the PLS acronym: "Projection to Latent Structures", in an attempt to clarify what PLS does in a geometrical sense (no matter which method is considered). Nowadays PLS is a highly developed body of methods, all based on the least squares (LS) optimization principle. It comprises of regression and classification tasks as well as dimension reduction techniques and modeling tools. The underlying assumption of all PLS methods is that the observed data is generated by a system (or process) which is driven by a small number of latent variables or factors (not directly observed or measured).

Chapter 3

Basics of Path Modeling

Path Modeling, also known as Structural Equation Modeling (SEM), is one of the major components of multivariate statistical analysis techniques. It provides a flexible and powerful method for analyzing multiple relationships between a set of blocks of variables. Path models are used by economists, educational researchers, marketing researchers, biologists, medical researchers, and a variety of other social and behavioral scientists. The concept of *structural equations* simply refers to the fact that the structure of cause-effect relationships between variables can be specified by a series of equations. In turn, the concept of *path modeling* refers to the graphical display of the structural equations in what is known as a path diagram.

One of the main features of path modeling techniques is the ability to deal with latent variables. Simply stated, latent variables are hypothetical or theoretical variables that cannot be observed nor measured directly. Because these types of variables cannot be measured explicitly, they have to be measured (or constructed) through variables that are perfectly observable/measurable.

In this chapter we provide an introduction to the basic concepts and ideas of path modeling techniques. We begin with the concept of latent variables and the different forms in which they are measured. We discuss the notions of path modeling and its related techniques, as well as the graphical notation for the representation of path models. Finally, a brief description of the main approaches to estimate path models is offered.

3.1 Measuring and Modeling the Unobservable

In an attempt to understand the world around us we take measurements of different things: objects, individuals, processes, and phenomena of interest. On the one hand, for instance, we measure physical properties of objects such as length, temperature, time, and mass. We combine those magnitudes to obtain derived magnitudes like area, volume, speed, acceleration, and density. On the other hand, we measure qualitative attributes of objects such as form, color, odor, quality, and appearance. We measure financial aspects like prices, costs, rates, incomes, and wages. We also measure abstract

concepts and theoretical constructs such as purchasing power, productivity, satisfaction, motivation, loyalty, educational achievement, social status, among others. In summary, we measure a variety of aspects: from quantitative to qualitative characteristics, from physical to abstract properties; from observable to unobservable attributes.

In order to obtain and extract some knowledge, we analyze the measurements we take. We examine possible relationships between them, elaborate on descriptions of reality, and propose hypotheses and theories to be confirmed or discarded. For these purposes, models are of great utility and importance. They enable researchers to have a simplified representation of some part or process of the world, or some phenomenon of interest. We must consider that at times we cannot observe nor measure directly variables and concepts of interest in our models. In these cases we refer to them as theoretical concepts or hypothetical constructs. Michael Sobel (1994) refers to them as unobserved entities. These entities are very common in social and behavioral sciences (e.g., psychology, sociology, economy). Psychologists speak of intelligence, satisfaction, motivation, commitment, and self-esteem. Sociologists refer to cultural values, social structure, social stratification, social status, and ethnicity. Economists speak of utility, economic development, and organizational performance. Unobserved entities are also found in the biological sciences. For instance, in ecological research (Pugesek, 2003) we might find concepts such as soil fertility, territory quality, and habitat physical structure.

When researchers work with theoretical concepts and constructs (i.e., developing theories and models) they conceive expected relationships between two or more unobserved entities. For example, consider the following actions:

- A marketing manager proposes a new product design to fulfill customer requirements
- A human resource manager establishes an employee rotation program to increment employee empowerment
- A group of high school teachers decide to increase the extracurricular activities to improve students' academic achievement
- A team of sociologists recommend a set of policies to encourage social integration of immigrants
- A scientist suggests planting native flowers to regenerate habitat of pollinator insects

The actions above illustrate courses of action that are undertaken as a result of expected relationships between two or more theoretical concepts. The marketing manager believes that changing the actual design of some product will meet customers' needs, who, in turn, are expected to be more satisfied. The teachers believe that setting more extracurricular activities will motivate students and increase their academic performance. The scientist believes that planting native flowers will regenerate the natural habitat of pollinators and hence recover their population.

Each of the expected relationships is causal, which means that one factor influences another. In different fields of knowledge researchers analyze this kind of relationships among constructs and propose theories and models. For these purposes, path modeling (i.e., structural equation modeling) is an excellent statistical methodology with great flexibility and modeling power.

3.2 Latent Variables

Sometimes we must face the fact that the variables of interest in our models cannot be observed nor measured directly. Examples of these kinds of variables are concepts like motivation, confidence, self esteem, and in general, different attitudes and mental abilities related with the psychological theories of human behavior. These variables are known as latent variables. Within the literature related to latent variables we can find synonymous terms like: theoretical concepts, hypothetical variables, constructs, factors, and intangibles.

These types of variables are very common in the social sciences (e.g., psychology, sociology, economy, and politics) in which there are many concepts of theoretical nature such as intelligence, socioeconomic status, industrial development or democracy. In fact it is not a coincidence that many examples of latent variables come from psychology since it was in this discipline where the concept originated. In statistics, latent variables (LVs) are widely used in several data analysis and modeling techniques with applications in many fields of knowledge. Despite its wide use, there is no single general definition of a latent variable. Bollen (2002) discusses the different ways that latent variables can be conceived and he distinguishes among three approaches:

- Latent variables seen as something that comes from the mind of the researcher, that is, LVs are not real but only hypothetical variables or constructs that only exist in the minds of analysts.
- Latent variables considered as real but being unobservable or non-measurable variables.
- Latent variables taken as a data reduction device or factor; that is a means of summarizing a number of variables into many fewer factors aiming to attain a parsimonious description of observed data.

Borsboom, Mellenbergh, and Van Heerden (2003) also analyze three perspectives for considering latent variables: from perspectives of constructivism, realism, and operationalism. These three points of view can be considered similar to those proposed by Bollen:

- Constructivism: regards the latent variable as a construction of the human mind
- Realism: latent variable signifying a real entity, something that exists independently of measurement
- Operationalism: Latent variable is nothing more than a weighted sumscore of its indicators, i.e. a tool for purposes like prediction or data reduction.

Muthén (2002) offers an extended description of the different names and forms of latent variables used in statistics although he considers other terms different from the theoretical constructs that Bollen (2002) and Borsboom *et al* (2003) admit. In the present work, we accept all the aforementioned concepts and we combine them into a single concept of latent variable having the following characteristics:

- They are considered as hypothetical variables (no matter if they are real or not)
- They are impossible to observe or measure (cannot be measured directly)
- They can be regarded as a data reduction device, i.e. a convenient means of summarizing a number of variables into many fewer factors

- They are taken as underlying variables that help explain the association between two or more observable variables

Although the essence of latent variables is that they cannot be measured directly, that does not mean that they are nonsense or useless. To make them operative, latent variables are indirectly measured by means of variables which can be perfectly observed/measured. These types of variables are called manifest variables (MVs), also known as indicators. It is assumed that manifest variables are used as indicators of a latent variable. We assume that each manifest variable contains information which reflects one aspect of the construct; hence we use the information contained in indicators to obtain an approximate representation of the latent variable.

It can be said that latent variables are useful in modeling processes and theory development. They provide a high degree of abstraction that allows researchers to represent and describe relationships among a set of variables for which some structure is assumed. That is to say, behind latent variables is the idea that observable phenomena are influenced by unobservable underlying causes. Paraphrasing Demotes-Mainard (2003), latent variables are rather like the invisible man: known to us only by his bandages or by the traces left by his body on cushions -we have access not to him but to his actions.

3.2.1 Latent variables and Intangibles

Besides its importance in behavioral sciences, latent variables have progressively been acquiring a growing importance among academics due to the intangibles of the so-called *New Economy* also referred to as the *Intangible Economy* (Goldfinger, 1997). The general idea is that the intrinsic value of companies is increasingly based on intangible elements, some of which cannot be measured directly. For instance, concepts like brand image, customer service, and intellectual capital, are some of the intangible elements of the companies. Although there is no broadly accepted definition of intangibles, they can be regarded as the non-physical and non-financial resources of the organization (Lev, 2001).

Currently, organizations employ a wide range of techniques to have a picture of their present status and performance. The most traditional type of measurement is based on financial indicators such as the incomes, the expenditures, the profits and the cash position. However, intangibles are becoming increasingly important because of the rapid growth of service activities over the last two decades. Actually, the existing trend is that the importance of physical goods and equipment is decreasing compared to the whole of the intangible activities within companies (Bounfour, 2003). It does not mean that productivity and monetary performance are not significant anymore. Companies will face the same challenges such as cost efficiency, investments, tangible assets, and physical resources; they are as important as they were before. What is new is the increasing appreciation of intangibles' influence. Intangible concepts such as Research & Development, knowledge creation, corporate identity and advertising, are now considered to be determining factors in the strategic positioning of today's organizations (Daum, 2003). The recognition of intangibles as sources of competitive advantage demand new tools to measure and report them. The concept of intangibles matches perfectly with that of latent variables; this is the main reason to consider the use of

latent variables to measure intangibles, and the use of structural equation modeling to model possible cause-effect relationships between them.

3.2.2 Reflective and Formative measurements

Once we have assumed that latent variables can only be observed and measured indirectly through the use of manifest variables, we need to consider the ways in which latent variables are indirectly measured. Latent variables can be observed/measured in two ways:

- through their consequences or effects reflected on their indicators
- through different indicators that are assumed to cause the latent variables

In the first case, called *reflective way*, manifest variables are considered as being caused by the latent variables. The second case is known as *formative way* because the latent construct is supposed to be formed by its indicators (Diamantopoulos & Winklhofer, 2001). The main difference between the reflective and formative ways has to do with the causal-effect relationships between the indicators and the constructs. To better understand the difference between reflective and formative ways, we will extend the analogy used in Cassel (2006). Suppose that a doctor is examining a patient trying to decide whether the patient is ill or not, namely the doctor is trying to determinate the presence or absence of some disease. She can use two approaches:

- ask about different symptoms
- ask about possible causes of disease

Different symptoms might be evaluated for example: body temperature, blood pressure, pulse rate, respiration rate, feelings of nausea, or headaches. Conversely, the doctor might ask about whether the patient has been consuming a particular kind of food, about the patient's habits (drinking, smoking, sleeping, etc.), or any other pattern behavior that might be causing the disease. Symptoms can be considered as reflective indicators because they *reflect* the disease; patterns of behavior can be seen as formative indicators because they *form* (cause) the disease. The formative and reflective approaches to measure a latent construct are illustrated in figure 3.1 trough the disease example.

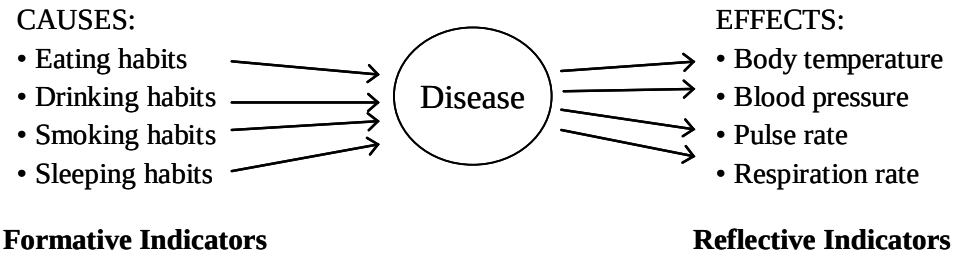


Figure 3.1. *Example of a latent variable measured by formative and reflective indicators*

In the reflective way it is expected that different indicators are to be highly correlated because they are measuring the same concept (De Beuckelaer, 2005). The same cannot

be said about formative indicators because a latent variable can be caused by two or more manifest variables mutually uncorrelated.

A typical example of an LV with formative indicators is socioeconomic status (SES), which is formed as a combination of income, occupation, education level, and residence. To verify the formative aspect of SES, one can see that if any of its indicators measures increase, SES will also increase. If there is still any doubt that it is truly formative, it will suffice to observe the problem conversely, i.e., if a person's SES increases, this would not necessarily imply an increase in all the MVs.

Another example of a formative latent variable is found in biology. Biometricians can be interested in measuring the body size of an animal. However, there is no single measure that accounts for the body size of any animal. For this reason body size can be viewed as a formative construct. For example, a bird's body size can be measured with three indicators: weight, body length, and wing chord length (Pugesek, 2003).

Some authors refer to emergent variables or emergent constructs rather than LVs when formative indicators are considered. For convenience sake, the term latent variable will be used in this work regarding formative or reflective perspectives. Furthermore, if only one manifest variable is used for measuring a construct, then it is assumed that the measured information will reflect only one aspect of the latent variable. It is therefore recommended to use various manifest variables to measure each construct. However, there may be cases in which a single manifest variable is the only available indicator.

3.2.3 Measurements in Principal Components and Factor Analysis

The difference between reflective and formative measurements is intrinsically related to the conceptualization of two different, yet analogous, multivariate data analysis methods: Factor Analysis (FA) and Principal Component Analysis (PCA). Although FA is described in more detail in the following chapter, and PCA is presented in chapter 5, we consider it convenient to outline some of their essential characteristics and how they can be viewed from a latent variable perspective.

In essence, the problem for both factor analysis and component analysis involves the description of a set of observed variables in terms of a reduced number of hypothesized (latent) variables; the use of principal components appeals more to an exploratory data analysis perspective whereas factor analysis is considered as a model building approach. In factor analysis, the latent variables are called factors, and it is assumed that these factors explain the observed variables. In contrast, the latent variables in component analysis are called components, which are obtained as linear composites of the observed variables. This section will present only a brief description of both methods.

Factor Analysis

The main idea behind factor analysis is that observed variables can be explained by a few factors. The aim is to represent the observed variables as functions of the factors. For instance, consider a set of p observed variables $x_1, x_2, \ldots, x_p$. which can be explained by m common factors $F_1,\ldots, F_m$

$$x_j = \lambda_1 F_1 + \lambda_2 F_2 + \ldots + \lambda_m F_m + \delta_j, \qquad j=1,\ldots,p$$

The common factors F_j are solely responsible for the covariation among the observed variables. The δ_j terms, called unique factors, are responsible for variation unique to each respective observed variable.
Under the FA point of view, a factor F_j is associated to the observed variables in a reflective form as shown in the following figure.

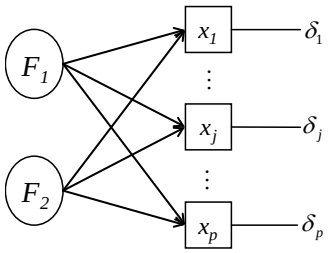


Figure 3.2. *Factors regarded as latent variables with reflective indicators*

Principal Component Analysis
PCA seeks to extract most of the variation present in a set of observed variables. The way in which the components are obtained is as composite variables, to be precise as linear combinations of the observed variables. Data on p observed variables $x_1, x_2, \ldots, x_p$ can be combined in order to obtain a principal component PC as:

$$PC = w_1x_1 + w_2x_2 + \ldots + w_px_p$$

Figure 3.3 diagrams an example of a principal component model with two components. As it can be observed, principal components can be represented as latent variables with formative indicators.

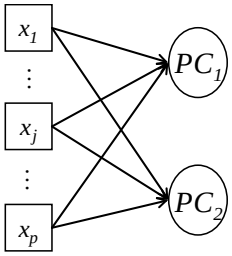


Figure 3.3. *Principal components regarded as latent variables measured by formative indicators*

The difference between FA and PCA from a latent variable standpoint is based on the difference between reflective and formative measurements. Factors are latent variables regarded as hypothetical variables that help explain the association between two or more observed variables. On the other hand, principal components are latent variables obtained as weighted sumscores of their indicators. The main feature shared by both FA and PCA is that they can be used to summarize a set of observed variables aiming to attain a parsimonious description of observed data (i.e. as data reduction techniques).

3.3 Path Modeling

The term path modeling is a very generic term used to designate a set of different statistical techniques that seek to explain the relationships among multiple variables. In some disciplines such as economics or psychology, the term path modeling is better known as structural equation modeling (SEM). In other disciplines like sociology or political sciences, for example, the term causal modeling is more commonly used. In fact, the broad literature related with path modeling is full of several terms that, to some extent, reflect different features of this type of models.

Some examples of statistical methodologies considered as path modeling techniques are:

- Path Analysis
- Exploratory Factor Analysis
- Confirmatory Factor Analysis
- Covariance Structure Analysis
- Linear Structural Relations
- Moments Structure Models

All of them have two main characteristics in common: (1) the use of some prior knowledge (theory) about the relationships among variables, and (2) the graphical representation of those relationships by drawing a picture of the model following a well established set of conventions.

Our decision to employ the term path modeling instead of the others is due to its established use within the PLS community. However, in this work we will also refer to it with the more classical term of structural equation modeling, using both terms interchangeably. The basic difference between them is that path modeling emphasizes on the graphical approach whereas the structural notion focuses on the assumed structure for the relationships between variables.

Claes Fornell (1982) speaks of structural equation models as a second generation of multivariate methods in the sense that they allow not only an exploratory approach (data then conceptualization) but also a theory-based approach. In this case, first generation multivariate methods are those techniques more oriented to an exploratory perspective, namely, conceptualization (theory) comes after data analysis is performed. Hence, second generation methods are extensions and generalizations of first generation methods.

Another argument provided by Fornell is that second generation multivariate methods “have the ability to bring theory and data together”. That is to say, these methods are used when we are interested in modeling a phenomenon of interest based on a theoretical framework. It is assumed that relationships between variables represent a process or system with inputs and outputs; the inputs are referred to as causes and the outputs as its consequences. A theoretical structure (model) is imposed on the data, and the strength of the relationships is examined.

In summary, path modeling is a useful set of methods that allows the combination of prior knowledge with measured data. The prior knowledge is provided by some theory for a certain phenomenon of interest, in which a model for the cause-effects relationships among variables is proposed. Usually, the causal relationships are supposed to be linear and most of the variables represent theoretical or abstract

concepts, some of them unobservable, meaning that they cannot be directly measured in practice.

From the causality perspective, the differences between path modeling methodologies are related to the amount and the abstractness of the *a priori* knowledge about causal relationships among the analyzed variables. By the term "amount" one understands how much theoretical knowledge is hypothesized, from a more exploratory to a more confirmatory sense. "Abstractness" refers to the use of more or less theoretical concepts (constructs or theoretical variables). As seen before, theoretical concepts are represented by variables, some of which usually cannot be observed nor measured in practice, i.e. they are latent variables. Some path modeling methods are specifically conceived to work with these special types of variables.

It is not the intention of the present work to discuss terms like cause, causality, or causation, which have developed great controversy since ancient times and whose debate arena is more in the philosophical field. In this thesis, we define the meaning of cause by saying that a variable *A* causes another variable *B,* refers to that a change in *A* results in a change in *B*.

3.4 Path Diagrams

As aforementioned, one of the main features of path modeling techniques is its graphical approach through a visual representation of the models. Pictures of path models are called path diagrams which are drawn according to well established conventions of terminology and symbols. Path diagrams are very helpful because they provide a graphical representation of the relationships among a set of variables, with the special property that they can be translated into a system of simultaneous equations. The great advantage of path diagrams is that they allow for the visualization of the relationships and, in terms of a causal model, its graphical display makes it possible to understand the conceptualization of the model.

3.4.1 Notation and symbols

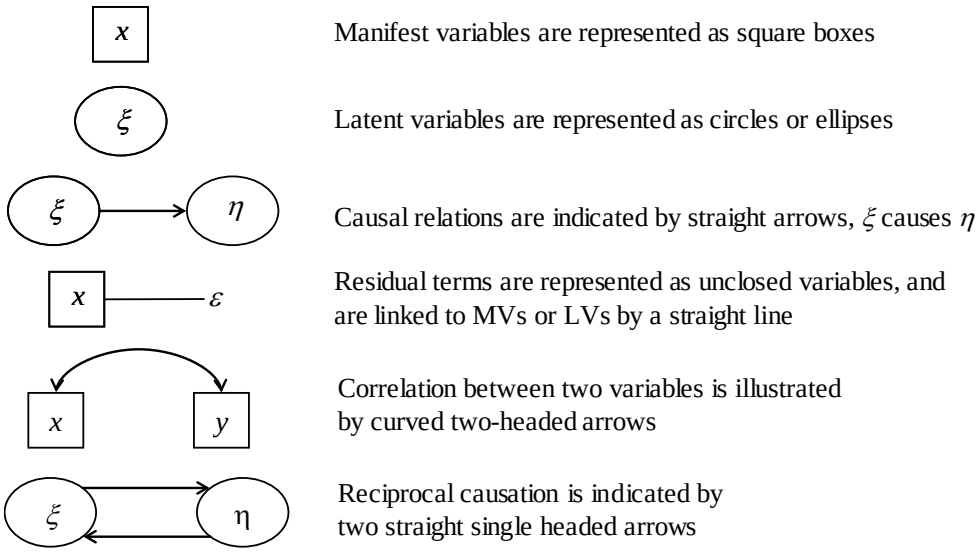
Since the invention of path diagrams in the 1920s by Sewall Wright and the posterior development of path analysis, path diagrams have been acquiring minor adaptations. Its use in causal models, confirmatory factor analysis, and structural equation models have allowed for a general notation.

Variables can be of any kind: manifest variables, latent variables, or residual variables (disturbance terms). Observed variables are enclosed in boxes; latent variables are enclosed in circles/ellipses, and residual terms are maintained unclosed. Relationships also can be of three types: causal links meaning that variable *A* causes variable *B*; correlation links indicating simply a correlation between two variables *A* and *B* without implying causality; or the affection of a residual term ε to some variable *A*. Causal relationships are assumed to be linear, and are represented by straight single-headed arrows, correlations are represented by curved two-headed arrows, and residual affection by straight lines.

In addition, variables may be grouped in two classes: (1) those that are caused by one or more variables, and those that are not caused by any other variables in the diagram. The first class of variables is called endogenous or dependent variables. The second

class is known as exogenous or independent variables. The convention is to use Greek letters for the latent variables, and Italic letters for the manifest ones. Exogenous latent variables are usually represented by the Greek letter ξ (xi), while endogenous latent variables are represented by η (eta).

Table 3.1. *Path diagram notation*



3.4.2 Simple path diagrams

Not all path diagrams contain latent variables. In fact, a simple linear regression model:

$$X = \beta Y + \varepsilon$$

can be represented in a path diagram as follows:

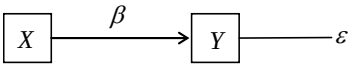


Figure 3.4. *Path diagram of a simple linear regression*

Variable X is the independent variable which is assumed to explain variable Y, and an error term ε is associated to Y. The regression coefficient β is called the path coefficient.

Another example of a path diagram that contains only latent variables can be illustrated as:

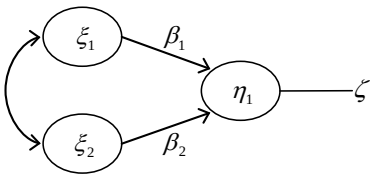


Figure 3.5. *Path diagram of a multiple linear regression with latent variables*

In this case, there are two exogenous latent variables, ξ_1 and ξ_2, interconnected by a curved arrow to indicate that they may be correlated. The endogenous LV η_1 is caused by the exogenous LVs, and an error term ε is associated with it. The path coefficients are indicated by β_1 and β_2.

In causal modeling the typical path diagram contains some structural relations among constructs, each one related to its indicators. For example, consider a model with two LVs ξ causing η, each one is associated to a block of three indicators. The LV ξ is related to their indicators in a formative way and no residual terms are considered. Conversely, the LV η is related to its indicators in a reflective way, so each indicator y_j has its corresponding disturbance term ε_j. The following path diagram illustrates the hypothesized model.

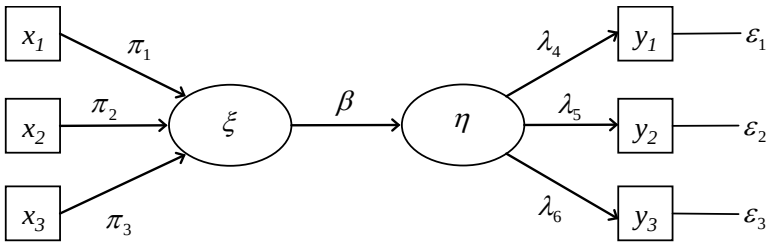


Figure 3.6. *Path diagram of a two block causal model*

3.4.3. Recursive and Non-recursive latent variable path diagrams

Whether reciprocal causations exist or not, path diagrams can be classified in two types: (1) recursive, and (2) non-recursive. In the case that a model contemplates two or more LVs with reciprocal causation the model is said to be non-recursive, i.e. there are some paths going forward and backward. Otherwise is said to be recursive, i.e. the structural relation has the form of a causal chain with no loops.

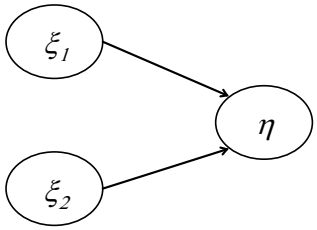


Figure 3.7. *Path diagram of a recursive model*

In figure 3.7 a recursive path diagram is shown with one endogenous LV that is caused by two exogenous LVs. It is clear that there is no reciprocal causation; meaning that there are no loops in the path diagram. Manifest variables have been omitted for the sake of simplicity. In contrast, figure 3.8 shows a reciprocal causation between LVs η_1 and η_2. In this case a feedback loop between the endogenous constructs is formed. That is why this is a non-recursive path diagram.

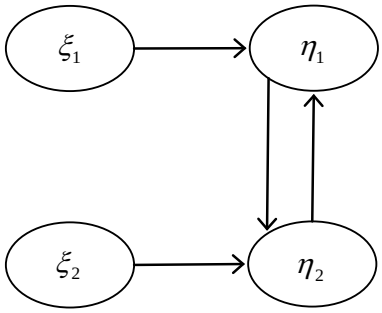


Figure 3.8. *Path diagram of a non-recursive model*

As it has been shown, path modeling methodologies might not contain LVs, and some path models can be non-recursive. In fact, causal relations may not be linear. However, this thesis will be concerned only with recursive path models with latent variables, assuming linear relations as those illustrated in figure 3.7.

3.4.4 Example of path model

The following is a simplified example based on the model described in Young (1998). It consists of a model with three latent constructs: school achievement, socioeconomic status, and classroom learning environment. This model of school achievement takes into account the external influences that appear to affect the student's level achievement. Such external influences include the student's socioeconomic status and the classroom learning environment.

The model, as any general path model, is integrated by two parts or sub-models: one is the measurement model and the other is the structural model. The measurement model represents how each construct is measured by its indicator variables. The structural model involves the causal relations among the constructs and is represented by a simultaneous system of equations among the latent variables. In the present example, the measurement model has three constructs each with their respective manifest variables. Socioeconomic status is measured by four ordinal indicator variables: mother's education, father's education, mother's occupation, and father's occupation. Classroom learning environment is measured by four indicators: teacher support, student involvement, task orientation, and cooperation. Achievement is measured by three continuous variables: math scores, sciences scores, and English scores; likewise, the structural model states that school achievement is affected by socioeconomic status and classroom learning environment. The path diagram of the school achievement model is shown in figure 3.9.

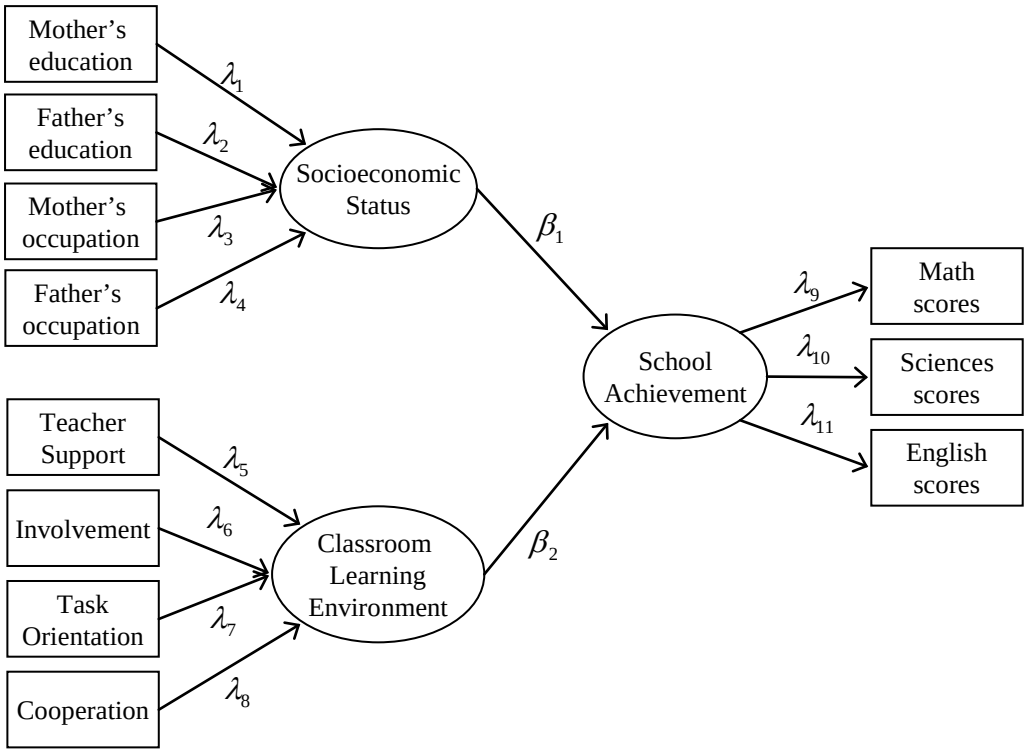


Figure 3.9. *Path diagram of the school achievement model*

Note that both "socioeconomic status" and "classroom learning environment" are measured with formative indicators (the arrowheads are pointing toward the latent variables). Conversely, "school achievement" has reflective indicators. Also, note that the example is a recursive model since there is no reciprocal causation among the latent constructs.

3.5 Path Modeling Approaches

The path modeling process starts at the conceptual level with a theoretical framework; a process that involves the establishment of the theoretical relationships among constructs. The subsequent step is deciding how many and which observed variables will be considered as indicators of the constructs (latent variables). The selection of manifest variables and its number is sometimes a subjective matter and no single criterion exists on this point. Regarding the number of indicators some authors like Bentler (1980) suggest to use as many indicators as possible although having too many may present problems with model fitting.

The form most commonly used for relating each construct to its indicators is the reflective way, in which the construct is taken as a common factor (in the sense of factor analysis) of its indicators. Other times, however, constructs are related to its indicators through formative way, in which the constructs are taken as components or projections (in the sense of principal components analysis). Once the relationships of the model are fixed, they can be visualized in the form of a path diagram. The next step involves the mathematical specification of the model, that is, its translation into a system of equations, followed by the estimation phase and the validation of results.

In essence, Path Modeling is a methodology for the analysis of indirectly measured cause and effect relationships in complex behavioral systems. This analysis can be accomplished under two major approaches (Hulland *et al*, 1996) depending on the desired purposes:

- Confirmatory purposes
- Predictive purposes

The confirmatory approach is concerned with theory development and testing by testing whether the assumed theory and hypotheses can be confirmed. The second approach, as its name implies, focuses on making predictions about the outcome variables of interest. The confirmatory option the model is analyzed by examining the covariance structure of the data and testing probabilistic assumptions. The predictive option has to do with the variability of data in the form of a prediction model of the dependent variables.

The path modeling method for confirmatory purposes receives the generic name of Covariance Structure Analysis, also known as LISREL. In turn, the predictive oriented methodology is Partial Least Squares Path Modeling. However, due to a recent proposal of an alternative method to PLS-PM by Hwang and Takane (2004), the predictive oriented methodology is also referred to as component-based path modeling.

The confirmatory and the predictive oriented approaches imply different notions and protocols in the following aspects:

- assumptions about data
- links between latent variables and indicators
- model specifications
- estimation procedures
- validation techniques

Covariance Structure Analysis (CSA) seeks to determine the extent to which the postulated structure (the postulated theory) is actually consistent with the observed data. This involves performing hypotheses tests to evaluate how well the hypothesized model fits the data. The overall idea consists of calculating a theoretical covariance matrix implied by the specified model and comparing it to the actual covariance matrix based on the empirical data (Diamantopoulos, 1994). To be precise, Covariance Structure Analysis seeks to minimize the difference between the empirical data covariance matrix and the theoretical covariance matrix deduced by the estimated parameters. The obtained model is used to explain the co-variability of the observed variables. In general, CSA procedure assesses whether a sample covariance or correlation matrix is consistent with a hypothetical matrix implied by the model and specified by the researcher (Kumar and Deregowska, 2002). Indeed, CSA is designed to maximize and test the degree of fit and consistency between the model and the data.

Partial Least Squares Path Modeling (PLS-PM) was originally developed as an analytical alternative to Covariance Structure Analysis for situations where the theory is weak and where the general assumptions of CSA are not met. The overall goal of PLS is to use observed independent variables to predict observed dependent variables. This is achieved indirectly by extracting independent and dependent latent variables from observed variables. This is done in such a way that they optimally address one or both of these two goals: explaining response variation and explaining predictor variation. The goal is to predict the dependent variables (both latent and manifest) by minimizing the residual variances of the endogenous (i.e. dependent) variables. In particular, the

method of partial least squares balances the two objectives, seeking latent variables that explain both response and predictor variation. (Kumar and Deregowska, 2002).

Generalized Structured Component Analysis (GSCA) is an alternative method to PLS-PM and it has been recently developed by Heungsun Hwang and Yoshio Takane in 2004. Because PLS-PM does not solve a global optimization problem for parameter estimation, there is no single criterion minimized or maximized to estimate model parameters. To overcome this situation, Hwang and Takane (2004) proposed a new method that avoids the major drawbacks of PLS-PM.

Although the present dissertation is focused on the Partial Least Squares approach, it is convenient to provide a general panorama of Covariance Structure Analysis techniques in order to have a better understanding of PLS as well as Path Modeling methods. The following chapter is dedicated to Path Modeling based on Covariance Structure Analysis. In turn, chapter 5 contains the conceptual background and the discussion of the PLS-PM algorithm.

Chapter 4

Path Modeling Based on Covariance Structure Analysis

Covariance Structure Analysis (CSA) or Covariance-Based Structural Modeling is the most widely used approach in structural equation modeling. In the broadest sense, covariance structure analysis is a set of techniques that aims to explain the structure or pattern among a set of latent variables, each measured by one or more manifest indicators. CSA seeks to reproduce the covariance of the observed variables that are supposed to be explained by some hypothesized structure among the latent variables. The name implies that the relationships among the observed variables, contained in the covariance matrix Σ, are characterized by the covariance of the latent variables. As a result, the model assumes that the latent variables generate the pattern or structure among the observed variables.

We have decided to begin this chapter with the presentation of two precursor methods of the CSA approach: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). In fact, both methods can be considered special cases of the general CSA framework. We consider it convenient to offer a description of EFA and CFA in order to have a better understanding of the CSA approach. In the last section of this chapter, the discussion of the Covariance Structure Analysis is provided.

4.1 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA), better known as Factor Analysis, attempts to explain a set of observable variables by means of a reduced number of hypothesized (latent) variables called factors. The idea behind this method is that observed variables are intercorrelated and that the intercorrelations can be explained by a few number of common factors. In other words, in EFA it is assumed that correlations among the observed variables occur because they are determined in part by common but unobserved variables, the factors. Exploratory Factor Analysis is the oldest latent variable modeling technique. It was proposed by the psychologist Charles Spearman in the early 1900s who give it the name of Factor Analysis. Spearman (1904) attempted to

model human intelligence from the analysis of matrix correlations of cognitive tests. He stated that the correlations could be described by a single underlying variable called "factor of general intelligence". However, the single factor hypothesis was soon found to be inadequate and the model was then modified in order to include more variables or factors. This modification was not immediate but took place over the four decades following Spearman's initial work. Most of the generalizations were developed under Thurstone's psychometric school in the 1930s and 1940s. Until the 1960s, factor analysis was understood merely as a tool for exploring the unknown dimensionality of the manifest variables, hence the reason of the term "exploratory". The idea behind the exploratory notion is to find a set of few factors explaining the underlying (unknown) structure of the observed variables.

The presentation in this section is based on the works of Cooper (1983), Harman (1968), Lawley and Maxwell (1962, 1971), Jöreskog and Sörbom (1979), Cuadras (1981), and Basilevsky (1994).

4.1.1 Exploratory Factor Analysis Model

Let X_{nxp} be a matrix of n-individuals and p-variables. The columns (or variables) of X are represented by $x_1, x_2, \ldots, x_p$, so we may consider X to be a (px1) vector of variables. The idea of factor analysis is to find a new set of $q+p$ variables $\xi_1,\ldots,\xi_q, \delta_1,\ldots, \delta_p$ allowing to express the observed variables as linear combinations of these new variables denominated factors. It is assumed that all information shared in common by the observed variables is due to the contribution of the factors. The linear relation between variables and factors is:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \ldots + \lambda_{1q}\xi_q + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \ldots + \lambda_{2q}\xi_q + \delta_2 \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \ldots + \lambda_{pq}\xi_q + \delta_p \end{aligned} \tag{4.1}$$

where:
- $\xi_1,\ldots,\xi_q$ are called common factors because it is supposed they are common for all the manifest variables;
- $\delta_1,\ldots,\delta_p$ are called unique factors because they only have influence with its corresponding manifest variable;
- λ_{jk} are coefficients denominated factor loadings and show the magnitude of the dependence of each x_j with the common factors ξ_k, $j = 1,..,p$ and $k = 1,\ldots,q$.

The factor model can be expressed in matrix notation as:

$$X = \Lambda\xi + \delta \tag{4.2}$$

where:
- ξ is the vector (qx1) of common factors
- Λ is the matrix (pxq) of factor loadings
- δ is the vector (px1) of unique factors

It is assumed that

- X, ξ and δ have zero means
- since factors are unobserved we can suppose them standardized, i.e. $Var(\xi_k) = 1$
- unique factors $\delta_1, \ldots, \delta_p$ are uncorrelated with each other and with the common factors $\xi_1, \ldots, \xi_q$
- covariance matrix among observed variables is denoted by

$$E(XX') = \Sigma \tag{4.3}$$

- covariance matrix among the factors is denoted by

$$E(\xi\xi') = \Phi \tag{4.4}$$

- covariance matrix among unique terms is denoted by

$$E(\delta\delta') = \Theta \tag{4.5}$$

Model assumptions imply that the covariance matrix Σ between the p observed variables can be expressed in the form:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \tag{4.6}$$

Equation 4.6 can be translated as:

Total Variance = Common Variance + Unique Variance

and is deduced as:

$$\begin{aligned}
\Sigma &= cov(\Lambda\xi + \delta) = E(XX') \\
&= E\left[(\Lambda\xi + \delta)(\Lambda\xi + \delta)'\right] \\
&= \left[\Lambda\xi\xi'\Lambda' + \delta\xi'\Lambda' + \Lambda\xi\delta' + \delta\delta'\right] \\
&= \Lambda E(\xi\xi')\Lambda' + E(\delta\xi')\Lambda' + \Lambda E(\xi\delta') + E(\delta\delta') \\
&= \Lambda E(\xi\xi')\Lambda' + E(\delta\delta') \\
&= \Lambda\Phi\Lambda' + \Theta
\end{aligned} \tag{4.7}$$

where $\Theta = diag(var(\delta_1), var(\delta_2), \ldots, var(\delta_p))$

The goal of factor analysis is to estimate matrices Λ, Φ and Θ to make Σ close to the empirical covariance matrix S of the observed variables. In other words, by estimating Λ, Φ and Θ we aim to fit a covariance matrix Σ that is in some sense close to the observed covariance matrix S.

Figure 4.1 illustrates a path diagram of a factor analysis model with two common factors and five manifest variables. For a better understanding, only a few factor loadings are shown.

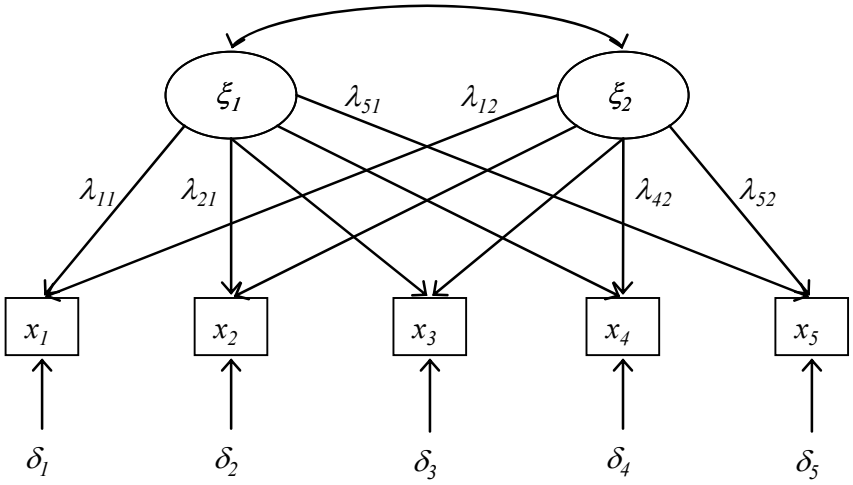


Figure 4.1. *Path diagram of a FA model with two common factors and five variables*

4.1.2 Orthogonal factors

For practical reasons it is often assumed that common factors are independent of each other (i.e, they are orthogonal). By doing this, the covariance matrix Φ becomes the identity matrix I of dimension pxp, that is:

$$E(\xi\xi') = I_{pxp} \tag{4.8}$$

Then Σ takes the form:

$$\Sigma = \Lambda\Lambda' + \Theta \tag{4.9}$$

Examining the variance of each variable x_j we have

$$Var(x_j) = Var(\lambda_{j1}\xi_1 + \lambda_{j2}\xi_2 + \ldots + \lambda_{jq}\xi_q + \delta_j) \tag{4.10}$$

and since we assume uncorrelated common factors

$$Var(x_j) = \lambda_{j1}^2 Var(\xi_1) + \ldots + \lambda_{jq}^2 Var(\xi_q) + Var(\delta_j)$$

$$Var(x_j) = \lambda_{j1}^2 + \ldots + \lambda_{jq}^2 + Var(\delta_j)$$

$$Var(x_j) = \sum_{k=1}^{q} \lambda_{jk}^2 + Var(\delta_j) \tag{4.11}$$

We see from equation 4.11 that the variance of variable x_j is composed by two parts. The first part is the term $\sum_{k=1}^{q} \lambda_{jk}^2$ called *communality* and represents the portion of the total variance attributed to the common factors. Each term λ_{jk}^2 is the contribution of the common factor ξ_k to the total variability of x_j. The second term, $Var(\delta_j)$, is known as the *uniqueness* or unique variance which is the portion of the variance in x_j not shared with the common factors. In other words, $Var(\delta_j)$ is the contribution of the unique factor to the variance of x_j.

Communality is then expressed as:

$$c_j^2 = \lambda_{j1}^2 + \ldots + \lambda_{jq}^2 = \sum_{k=1}^{q} \lambda_{jk}^2 \tag{4.12}$$

The decomposition of the variance of the manifest variable x_j as the sum of communality and uniqueness is given by:

$$Var(x_j) = c_j^2 + Var(\delta_j) \tag{4.13}$$

In addition, the covariance σ_{ij} of two variables x_i and x_j is

$$Cov(x_i, x_j) = \sigma_{ij} = \sum_{k=1}^{q} \lambda_{ik}\lambda_{jk} \tag{4.14}$$

With orthogonal factors, when they are represented in a path diagram they are not linked anymore with a curved arrow. The same example of figure 4.1 is shown in figure 4.2 but in this case common factors are not associated with one other.

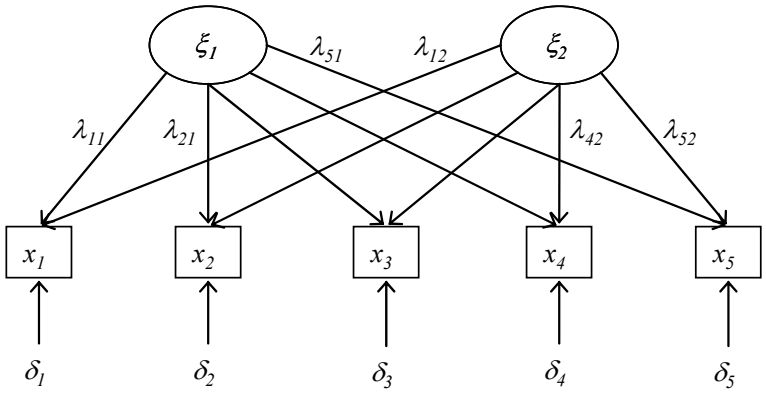


Figure 4.2. *Orthogonal Factor Analysis model with two common factors and five variables*

4.1.3 Factor Model with Correlation Matrix

If we make an additional assumption and consider manifest variables to be standardized then the covariance matrix becomes a correlation matrix and equation 4.6 changes as follows:

$$R = \Lambda\Lambda' + \Theta \tag{4.15}$$

where R is the correlation matrix.

Factor loadings λ's then become correlations between the manifest variables and factors

$$\begin{aligned} \text{cor}\,(x_j,\ \xi_k) &= \text{cov}(x_j,\ \xi_k) = E(x_j\ \xi'_k) = E\left[(\lambda_{j1}\xi_1 + \ldots + \lambda_{jk}\xi_k + \ldots + \lambda_{jq}\xi_q)\xi'_k\right] \\ \text{cor}\,(x_j,\ \xi_k) &= \lambda_{j1}E(\xi_1\xi'_k) + \ldots + \lambda_{jk}E(\xi_k\xi'_k) + \ldots + \lambda_{jq}E(\xi_q\xi'_k) \\ &= \lambda_{jk}E(\xi_k\xi'_k) \\ &= \lambda_{jk} \end{aligned} \tag{4.16}$$

The correlation matrix R takes the form:

$$R = \begin{pmatrix} 1 & \lambda_{12}^2 & \cdots & \lambda_{1q}^2 \\ \lambda_{21}^2 & 1 & \cdots & \lambda_{2q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1}^2 & \lambda_{p2}^2 & \cdots & 1 \end{pmatrix} + \begin{pmatrix} Var(\delta_1) & 0 & \cdots & 0 \\ 0 & Var(\delta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Var(\delta_p) \end{pmatrix} \tag{4.17}$$

A special matrix of interest is the reduced correlation matrix R^* obtained as

$$R^* = R - \Theta = \Lambda\Lambda' \tag{4.18}$$

$$R^* = \begin{pmatrix} 1-Var(\delta_1) & \lambda_{12}^2 & \cdots & \lambda_{1q}^2 \\ \lambda_{21}^2 & 1-Var(\delta_2) & \cdots & \lambda_{2q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1}^2 & \lambda_{p2}^2 & \cdots & 1-Var(\delta_p) \end{pmatrix}$$

$$= \begin{pmatrix} c_1^2 & \lambda_{12}^2 & \cdots & \lambda_{1q}^2 \\ \lambda_{21}^2 & c_2^2 & \cdots & \lambda_{2q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1}^2 & \lambda_{p2}^2 & \cdots & c_p^2 \end{pmatrix} \tag{4.19}$$

Note that diagonal elements of R^* contain communalities. This matrix (or its equivalent reduced covariance matrix S^*) will be used for purposes of estimation discussed in section 4.2.

4.1.4 Model Indeterminacy

The problem of indeterminacy comes from the fact that two representations, (Λ,ξ), and (Λ^*,ξ^*), will be equivalent if $\Lambda\xi = \Lambda^*\xi^*$. In this case, the model can be explained with the same precision by correlated factors as well as uncorrelated (orthogonal) factors. As we shall see, indeterminacy implies the existence of an infinite number of solutions providing the same results.

Correlated factors:
If we suppose a set of orthogonal common factors ξ, then an equivalent structure (Λ^*,ξ^*) with correlated factors can be obtained if we take any matrix H non-singular such that

$$X = \Lambda HH^{-1}\xi + \delta \tag{4.20}$$

where $\Lambda^*=\Lambda H$ and $\xi^*=H^{-1}\xi$

Uncorrelated factors:
If we take any matrix T such that $V = TT'$ where V is positive definite; i.e. $T^{-1}V(T^{-1})'= I$, an equivalent structure (Λ^*,ξ^*) of uncorrelated factors can be obtained from a set of correlated common factors ξ as follows

$$X = \Lambda T T^{-1} \xi + \delta \tag{4.21}$$

where $\Lambda^* = \Lambda T$ and $\xi^* = T^{-1}\xi$

Some of the indeterminacy in the model can be eliminated by assuming common factors to be uncorrelated, that is, $E(\xi\xi') = I$. However, the indeterminacy is not completely eliminated because any orthogonal matrix Q, i.e. $QQ' = I$, applied to a structure (Λ, ξ) will generate an identical structure:

$$X = \Lambda Q Q^{-1} \xi + \delta \quad \text{and} \quad X = \Lambda \xi + \delta \quad \text{are identical} \tag{4.22}$$

Both models contain orthogonal common factors and covariance matrix $\Phi = I$. This kind of indeterminacy is known as "factor rotation indeterminacy".

4.1.5 Model Identification

In addition to the issue of indeterminacy we must consider other concerns about factor analysis modeling. First of all, note that common factors ξ's, loadings λ's, and unique factors δs are unknown. Moreover, the number q of common factors is unknown.

Given a covariance matrix Σ and a number of q factors, the main concern is whether there exist matrices Λ, Φ and Θ satisfying $\Sigma = \Lambda\Phi\Lambda' + \Theta$, and if so, whether or not they are unique. The situation in which model and data constraints are insufficient to identify (locate or determine) unique estimates is known as the "identification problem". The solution to the identification problem is to place further theoretical or data constraints on the parameters.

Without any restrictions, the factor model does not impose sufficient constraints on Λ, Φ and Θ to ensure that Σ will be found to satisfy $\Sigma = \Lambda\Phi\Lambda' + \Theta$. If we examine this last equation carefully, Factor Analysis implies reproducing the $p(p+1)/2$ variances and covariances of Σ in terms of pq elements of Λ, $q(q+1)/2$ elements of Φ, and p elements of Θ. Indeed, we need to impose certain restrictions to find the appropriate matrices. The usual procedure when imposing restrictions, as seen in the previous section, is to let $\Phi = I$ in order to reduce some of the indeterminacy in the model. This reduces the number of conditions to $p(q+1) = pq + q(q+1)/2 + p - q(q+1)/2$.

With respect to the number of factors, q, a necessary and sufficient condition that there exist q common factors is that there is a matrix Θ such that the matrix $\Sigma - \Theta$ is positive semi-definite and of minimal rank q. In fact, since we let $\Phi = I$, we have that $\Sigma - \Theta = \Lambda\Lambda'$. The matrix $\Lambda\Lambda'$ has the same off-diagonal elements as Σ.

Regarding Λ, it must satisfy $q(q-1)/2$ independent conditions (the number of unique elements in the orthogonal matrix Q). In addition, to define Λ uniquely it is also convenient to add one of the following restrictions: make $\Lambda\Lambda'$ to be diagonal or make $\Lambda'\Phi^{-1}\Lambda$ to be diagonal. As we can see from the above conditions, the effective number of unknown parameters is not $p(q+1)$ but $p(q+1) - q(q-1)/2$.

4.2 EFA Parameters Estimation

The estimation problem in factor analysis is to fit a matrix Σ of the form as in 4.9 to the matrix S (or R). The objective when estimating a FA model is not to estimate the factors

but to obtain values of Λ, Θ, and Ψ, that is, to estimate the parameter matrices. There are a great number of parameters estimation procedures in factor analysis which can be classified according to different characteristics.

Historically, the first estimation techniques were essentially algebraic. Examples of such methods are the principal factor analysis, and the centroid method (Cuadras, 1981). Another classification consists of scale free methods and non-scale free methods (Timm, 2002). Non-scale free methods are based on the least squares principle which minimizes the sum of squares of the elements of $S - \Lambda\Lambda' - \Psi = S - \Sigma$. Principal component factor, principal factor analysis, and iterative principal factor analysis are examples of this type of procedure. With respect to the scale free methods, the most famous procedure is Maximum Likelihood, developed in the 1940s by Lawley (1940), and discussed in Lawley and Maxwell (1962, 1971).

For the purpose of the present work, we will only focus on three estimation techniques:

- Maximum Likelihood (ML) estimation
- Generalized Least Squares (GLS) estimation
- Unweighted Least Squares (ULS) estimation

These estimation techniques are based on fitting a function to the sample covariance matrix. The fitting function can be developed in terms of the latent roots and vectors of certain matrices, and in a suitable form for numerical analysis by computer programs. Knight (1978) presents a general description of the estimation process that is used in these techniques. The general steps for these methods are the following:

1) Determine the fitting function to be minimized (ML, GLS or ULS)
2) The chosen fitting function yields partial derivatives with respect to the model parameters. The model parameters, together with the diagonal constraint to eliminate the rotation problem, may be expressed in terms of latent roots and vectors of a matrix involving S and Θ. A conditional solution for Λ for given Ψ is obtained.
3) Substitution of this conditional solution in the fitting function then gives a conditional function $f(\Theta)$ in terms of latent roots, which is to be minimized. Estimates of Θ are obtained, and substituted back to iterate again until convergence is reached, that is to find the overall minimum.

The numerical evaluation of the latent roots and vectors of the relevant matrix is followed by this iterative procedure to minimize the fitting function by evaluating it and its first and second derivatives.

4.2.1 Maximum Likelihood

The Maximum Likelihood approach for estimating factor analysis models was introduced by Lawley (1940). It was presented at a time when there was a very limited computing capacity making this method an unpractical option. It was necessary to wait for the development of modern digital computers that make it possible to perform the computations required in the maximum likelihood method. Jöreskog (1967) contributed advanced procedures which could use a more powerful analysis method such as the Newton-Raphson iterations. With the computer developments and Jöreskog's

contributions, maximum likelihood factor analysis became quite feasible (Jöreskog and Lawley, 1968).

For the sake of simplicity, we will assume uncorrelated factors. This implies no loss of generality since the factors can still be obliquely rotated anytime after extraction. Maximum Likelihood Factor Analysis attempts to estimate Λ and Θ in such a way that the likelihood of obtaining the observed sample covariance matrix is maximized. The maximum likelihood estimation thus maximizes a likelihood function, which indicates how likely it is that the actual covariances would emerge if the population parameters were equal to Λ and Θ.

ML estimation requires assuming that variables in X are multivariate normally distributed with mean vector μ and variance-covariance matrix $\Sigma=\Lambda\Lambda' + \Theta$. In this case the density of $X \sim N_p(\mu, \Sigma = \Lambda\Lambda' + \Theta)$ is:

$$N_p(\mu,\Sigma) = (2\pi)^{-p/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \tag{4.23}$$

We are interested in finding a model covariance matrix Σ that best reproduces the sample covariance matrix S. For doing so, we may begin by examining the distribution of S whose probability density was calculated by Wishart in 1928, hence called the Wishart distribution.

Assume that a random sample of N individuals is obtained from a multivariate normal population having variance-covariance matrix Σ and mean μ. The sample covariance matrix S with general terms $[s_{ij}]$ is defined by

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ij}-\bar{x}_j)' = \frac{1}{n}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ij}-\bar{x}_j)' \tag{4.24}$$

where: $n = N - 1$, and $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$.

It can be proved (Ghosh and Sinha, 2002) that nS is $W(\Sigma,n)$, that is, nS follows a Wishart distribution given as:

$$W(\Sigma,n) = \frac{\exp\left\{-\frac{n}{2}tr(S\Sigma^{-1})\right\}\left|nS\right|^{1/2(n-p-1)}}{\left|\Sigma\right|^{n/2}2^{np/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma((n+1-i)/2)} \tag{4.25}$$

If we join all the constant terms (not depending on Σ) and we call this combined value Q, then eq. 4.25 can be expressed as

$$W(S;\Sigma,n) = \frac{\exp\left\{-\frac{n}{2}tr(S\Sigma^{-1})\right\}}{\left|\Sigma\right|^{n/2}}Q \tag{4.26}$$

A statistical function is developed using the likelihood ratio (Neyman and Pearson, 1928) which compares any theoretical model with a perfect fitting model. The test involves a comparison of the maximum value the likelihood can take when the parameters of the model are allowed to take any values (null hypothesis H_0), and the maximum value of the likelihood when the parameters are restricted by the hypothesis

in the parameter space (alternative hypothesis H_1). In other words, it implies the "distance" between the hypothesized model and the theoretically perfect model.

The likelihood ratio is defined by:

$$\text{Likelihood ratio} = \frac{\text{likelihood for any given model under } H_0}{\text{likelihood with a perfectly fitting model under } H_1}$$

$$\text{likelihood ratio} = \frac{\exp\left\{-\frac{n}{2}tr(S\Sigma^{-1})\right\}|\Sigma|^{-n/2}Q}{\exp\left\{-\frac{n}{2}tr(SS^{-1})\right\}|S|^{n/2}Q} \tag{4.27}$$

We can observe that Σ has been replaced by *S* in the denominator of eq. 4.27, because this represents a perfect model.

$$\text{likelihood ratio} = \exp\left\{-\frac{n}{2}tr(S\Sigma^{-1})\right\}|\Sigma|^{-n/2}\exp\left\{\frac{n}{2}tr(SS^{-1})\right\}|S|^{n/2} \tag{4.28}$$

taking the natural logarithm of both sides of the equation

$$\text{log likelihood ratio} = -\frac{1}{2}tr(S\Sigma^{-1}) - \frac{n}{2}\log|\Sigma| + \frac{1}{2}tr(SS^{-1}) + \frac{n}{2}\log|S|$$

$$= -\frac{n}{2}\left\{tr(S\Sigma^{-1}) + \log|\Sigma| - tr(SS^{-1}) - \log|S|\right\} \tag{4.29}$$

S and S^{-1} are square matrices containing $p+p$ rows and columns. Hence, SS^{-1} is a pxp identity matrix whose trace is *p*, then

$$\text{log likelihood ratio} = -\frac{n}{2}\left\{tr(S\Sigma^{-1}) + \log|\Sigma| - \log|S| - p\right\} \tag{4.30}$$

Maximizing eq. 4.30 is equivalent to maximizing eq. 4.27, but if we discard the minus sign and the constant term, maximizing 4.27 is equivalent to minimizing the following function

$$F_{ML}(\Lambda,\Theta) = tr(S\Sigma^{-1}) + \log|\Sigma| - \log|S| - p \tag{4.31}$$

which is known as the fit function.

For the purpose of minimizing the fit function F_{ML} we need to take its partial derivates with respect to the elements of Λ and the diagonal elements of Θ. The partial derivates equations are

$$\frac{\partial F_{ML}}{\partial \Lambda} = 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\Lambda \qquad \text{and} \qquad \frac{\partial F_{ML}}{\partial \Theta} = diag(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}) \tag{4.32}$$

We can see that maximum likelihood estimates are determined as the solution of two matrix equations. However, these equations cannot be solved algebraically; instead it is necessary to use an iterative procedure that evaluates at each step evaluates the function and its derivatives. Computer programs make use of an algorithm due to Fletcher and Powell (1963). It is beyond our scope to discuss this algorithm. For the interested

reader, a detailed and technical description of the ML method is given in Lawley and Maxwell (1971), and Jöreskog and Goldberger (1972). Clarke (1970) describes a minimization method based on the Newton-Raphson approach as an alternative to the Fletcher and Powell algorithm. Rubin and Thayer (1982, 1983) propose the Expectation Maximization (EM) algorithm to estimate Maximum Likelihood Factor Analysis.

Hypothesis testing
Lawley and Maxwell (1971) developed the usual likelihood ratio test using maximum likelihood estimators. The criterion for testing the hypothesis H_q that there are q common factors is

$$U_q = \left[n - (2p+5)/6 - 2q/3\right] F_{ML}(\hat{\Lambda}, \hat{\Theta}) \tag{4.33}$$

U_q is distributed approximately as χ^2 if H_q is true with degrees of freedom given by $\upsilon = \frac{1}{2}\left[(p-q)^2 - (p+q)\right]$, where q must be such that υ is positive for the hypothesis to be non-trivial. The hypothesis of exactly q factors would be rejected at the α level if $U_q \geq \chi^2_{\alpha,\upsilon}$. The procedure is to fit a model with a small number of factors. If this model is rejected, then the number of factors is increased one at a time until the corresponding value of U is not significantly larger at the chosen probability.

4.2.2 The Least Squares Approach

We know that Factor Analysis aims to explain common variance, which is the source of correlations among the observed variables. Thus, a factor analysis model should obtain the best possible reproduction of these correlations. The principle of least squares can be applied for such a purpose in the following way: minimizing the sum of squared residuals from the differences between the observed and the reproduced correlations. We present two approaches that are based on the least squares criterion:

- Generalized Least Squares (GLS)
- Unweighted Least Squares (ULS)

Generalized Least Squares (GLS)
Perhaps the most applied estimation method based on the least squares principle is the Generalized Least Squares (GLS) function developed by Jöreskog and Goldberger (1972). They modified Aitken's generalized least squares principle in order to obtain parameter estimates to minimize the following function:

$$F_{GLS} = tr\left\{\left[\Sigma_T^{-1}(S-\Sigma)\right]^2\right\} \tag{4.34}$$

where Σ_T is the true population covariance matrix. As this matrix is unknown, it is substituted by the sample covariance matrix S as its estimate. This leads to the following function:

$$F_{GLS} = tr\left\{\left[S^{-1}(S-\Sigma)\right]^2\right\} \tag{4.35}$$

As we can see, the matrix of residual covariances, $(S - \Sigma)$, is weighted by the inverted sample covariance matrix. The fitting function to be minimized is given as:

$$F_{GLS} = tr\left\{\left[I - S^{-1}\Sigma)\right]^2\right\} \tag{4.36}$$

Solution for minimum F_{GLS} involves two steps: the first step is a conditional solution for the estimated factor matrix, Λ, and is dependent on values of the uniqueness; the second step is an overall solution for the estimated uniqueness matrix Θ. When the observed variables are assumed to be multinormally distributed the behavior of the GLS estimates is very similar to the ML estimates. In this case, the estimates obtained have the same asymptotic properties as the maximum likelihood estimates. This leads to the same hypothesis test of q factors as for the maximum likelihood estimation, where GLS is also evaluated at the minimum, since both minima are asymptotically equivalent.

Unweighted Least Squares (ULS)

The simplest criterion for the goodness-of-fit of the reproduced correlation matrix to the observed correlation matrix is the Unweighted Least Squares (ULS) criterion. The criterion to be minimized is the sum of the squares of the difference between corresponding elements of S and the estimated Σ, that is:

$$F_{ULS} = tr\left[(S - \Sigma)^2\right] \tag{4.37}$$

A test of significance for the hypothesis of k factors like the one used in the maximum likelihood estimation, is not available because the condition for ofefficacy of the estimators is not fulfilled. In addition, the ULS fitting function is not scale free. This means that the ULS method is usually applied to the correlation matrix R instead of the covariance matrix S.

4.3 Confirmatory Factor Analysis

In the previous sections Factor Analysis was presented as an exploratory device to analyze a hypothetical, but unknown, covariance structure in a set of observed variables. As we have seen, the goal in exploratory factor analysis is to find a set of few common latent variables that explain the relationships between the manifest variables. EFA is a "primitive" causal model in the sense that there is scarcity of knowledge about the phenomena of interest and, aside from the model hypotheses, no extra assumptions can be made about the structure of the latent variables and how manifest variables depend on them.

Without much *a priori* knowledge about the studied phenomenon, factor analysis is well suited for the early stages of experimentation. As more information is obtained about the model, the analyst may shift from an exploratory perspective to a confirmatory one. When the analyst has previous knowledge about the model regarding the structure of the factors, then it is possible to propose a more sophisticated model than the one defined in an exploratory approach. The analyst begins with a hypothesis about the model specifying not only the number of factors but also which variables will be correlated with which factors, as well as which factors are correlated among them. In this case the Confirmatory Factor Analysis (CFA) is applied.

CFA is considered as a theory-testing model as apposed to a theory-generating method like EFA. Model assumptions are based on a strong theoretical and empirical foundation, and the analyst can test some suggested hypotheses. Rather than exploration, the goal is confirmation: the specified model is tested in order to determine the goodness of fit with the covariance of the observed data.

4.3.1 Confirmatory Factor Analysis Model

In section 4.1.1 the specification of the EFA model was presented. The model definition is reproduced below:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \ldots + \lambda_{1q}\xi_q + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \ldots + \lambda_{2q}\xi_q + \delta_2 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \ldots + \lambda_{pq}\xi_q + \delta_p \end{aligned}$$

In matrix notation we have:

$$X = \Lambda\xi + \delta$$

where:
- ξ is the vector ($qx1$) of common factors
- Λ is the matrix (pxq) of factor loadings
- δ is the vector ($px1$) of unique factors or errors of measurement

In EFA, every manifest variable loads on every factor in the model. It is assumed that they are common for all the manifest variables. However, when there is a prior knowledge about the factors structure, more restrictions can be imposed on the model defined in equation 4.2. Based on that previous knowledge, the researcher can specify a model that is uniquely determined. For example, the exploratory factor model in figure 4.3 could be now modified as in figure 4.4.

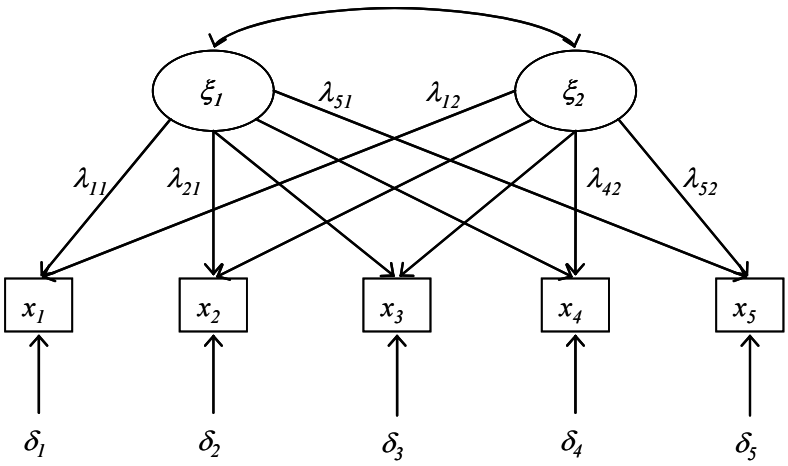


Figure 4.3. *Path diagram of an EFA model with two common factors and five variables*

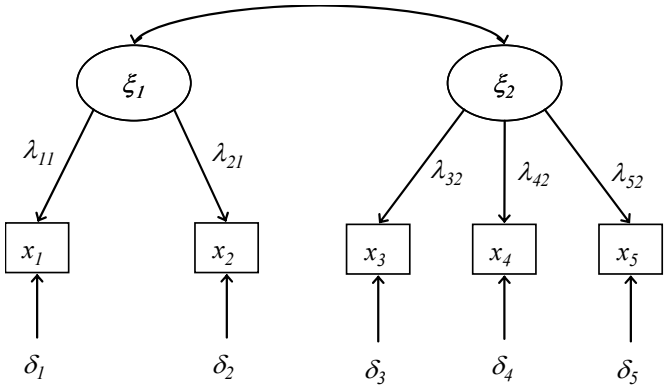


Figure 4.4. *Path diagram for a CFA model with two factors and five variables*

The new model is expressed as follows

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \quad 0 \quad + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \quad 0 \quad + \delta_2 \\ x_3 &= \quad 0 \quad + \lambda_{32}\xi_2 + \delta_3 \\ x_4 &= \quad 0 \quad + \lambda_{42}\xi_2 + \delta_4 \\ x_5 &= \quad 0 \quad + \lambda_{52}\xi_2 + \delta_5 \end{aligned} \tag{4.38}$$

As it is noted, there are some terms that have been eliminated from the model; those terms stand for zero loadings and are referred to as fixed parameters. The rest of parameters which have to be estimated are called free parameters. In CFA a simple structure is imposed on the pattern of factor loadings.

The usual process to perform a CFA starts with the specification of the model taking into account the following aspects:

- The number of factors present in the model
- The influence of factors over the indicators
- The interrelation among factors

The model is determined by fixing or freeing specific parameters such as the factor loadings, the correlations among factors, and the variance/covariance of the error of measurement. It is supposed that these parameters are set according to the *a priori* knowledge and theoretical assumptions of the researcher. The parameter settings can be used by the researcher to: (1) establish the validity of single factor model; (2) test the significance of a specific factor loading; (3) test the relationship between two or more factor loadings; (4) test whether a set of factors is correlated or uncorrelated; and (5) compare the ability of two different models to account for the same set of data.

In matrix notation the confirmatory factor analysis model implies:

$$\begin{aligned} \Sigma &= E(XX') = E\left[(\Lambda\xi + \delta)(\Lambda\xi + \delta)'\right] \\ \Sigma &= \Lambda E(\xi\xi')\Lambda' + \Theta \\ \Sigma &= \Lambda\Phi\Lambda' + \Theta \end{aligned} \tag{4.39}$$

where the matrix Φ represents the correlation matrix between the factors, which in the case to be assumed as orthogonal, this matrix will be the identity. This equation implies

reproducing the $p(p+1)/2$ variances and covariances of Σ in terms of pq elements of Λ, $q(q+1)/2$ elements of Φ, and $p(p+1)/2$ elements of Θ.

4.3.2 Identification

As in Exploratory Factor Analysis, the matter of identification is one of the most important aspects when specifying a CFA model. The identification problem involves verifying whether a unique solution exists for the covariance equation, $\Sigma = \Lambda\Phi\Lambda' + \Theta$, in terms of the parameters Λ, Φ and Θ. This verification has to be done before the estimation phase.

First of all, it is important to see what is understood by parameter identification:

- A parameter can be identified either by fixing it to a particular value in the specification or by solving for it as a function of parameters with fixed values.
- A parameter is unidentified if it can take on multiple values given the pattern of fixed parameters in the model.

There are three situations for model identification:

- A model is said to be under-identified if one or more free parameters are unidentified. An under-identified model implies that the covariance equation, $\Sigma = \Lambda\Phi\Lambda' + \Theta$, has no solution.
- In case that all free parameters are identified and their unique estimate can be obtained through only one manipulation of the fixed parameters, then the model is said to be just identified. That is, the covariance equation has a unique solution for each of the parameters Λ, Φ and Θ.
- If one or more free parameters can be obtained through multiple manipulations of the fixed parameters, then the model is said to be over-identified. It means that the covariance equation has more than one solution for the parameters.

It is obvious that without any restriction on the elements of Λ, Φ and Θ, the model is under-identified. To overcome this problem some constraints on the parameters must be imposed. Thus, before estimating the parameters we need to determine whether the model is identifiable or not. It is important to mention that model identification becomes an individual task for each analyzed model.

Bollen (1989) describe four ways of evaluating the identification status of the CFA model:

- The number of free parameters in the model must not exceed the number of elements in the covariance matrix. Although failure to satisfy this criterion ensures an underidentified model, meeting the criterion does not guarantee an identified model.
- If each factor has at least three indicators and the loadings of at least one indicator per factor is fixed at a nonzero value or the variance of each factor is fixed to a nonzero value, then the model is identified.
- A measurement model will be identified if every latent construct is associated with at least two indicators and every construct is correlated with at least one other construct
- Rules 2 and 3 are no longer definitive when covariances between uniqueness are allowed to covary.

Vinacua (1986) classifies the conditions to evaluate the identification of a CFA model in three the different types:

- Necessary Condition
  a) This condition refers to the number of parameters to be estimated which has to be less than the number of variances-covariances of matrix S, that is: $pq + q(q+1)/2 + p(p+1)/2 < p(p+1)/2$
- Sufficient Conditions
  a) Is a symmetric matrix and positive definite with elements in the diagonal equal to 1.
  b) Is a diagonal matrix
  c) Has at least (q-1) values fixed to 0 in each column
- Necessary and sufficient condition
  a) In general, the best way to show that a particular model is identifiable is to show that each parameter can be solved in terms of the variances-covariances of the observed variables. This is done through a series of algebraic operations taking into account the covariance equations from the model. If this condition is met, the model is identifiable, otherwise it is not.

4.3.3 CFA Parameters Estimation

Once we have specified the model and we have obtained a suitable covariance matrix, we can proceed with the estimation of the free parameters. The objective is to find the values of the free parameters that, when substituted into the equations, most precisely recover the observed variances and covariances. In other words, the goal of the estimation procedure is to estimate the unknown parameters in such a way that the difference between the sample covariance matrix and the implied covariance matrix is minimized in a certain sense.
As in the case of EFA, CFA has three different fitting functions are the most applied in practice:

- Maximum Likelihood
- Unweighted Least Squares
- Generalized Least Squares

Maximum Likelihood (ML)
Using similar assumptions and arguments as those presented in the exploratory approach, the maximum likelihood fitting function F_{ML} is:

$$F_{ML}(\Lambda, \Phi, \Theta) = \log|\Sigma| + tr(S\Sigma^{-1}) - \log|S| - p \qquad (4.40)$$

but now we have that $\Sigma = \Lambda\Phi\Lambda' + \Theta$

The function (4.40) is considered as a function of the free parameters: those in Λ, those in the lower half of Φ including the diagonal, and those in the diagonal of Θ. The function F_{ML} is minimized with respect to the free parameters (first derivatives with respect to the fixed parameters are zero). Here, however, the matrix $\Lambda'\Theta^{-1}\Lambda$ is not in general diagonal and the function F_{ML} cannot apparently be arranged in terms of the eigenvalues and eigenvectors of a matrix, thus a two stage procedure as in ML-EFA is

not possible. This means that the function F_{ML} has to be minimized simultaneously with respect to all free parameters using a modified method of Fletcher and Powell.

Generalized Least Squares
The parameter estimators from GLS are obtained by minimizing the following function:

$$F_{GLS} = \text{tr}[(S-\Sigma)S^{-1}]^2 \quad (4.41)$$

The difference between matrices S and Σ in GLS is weighted by the elements of S^{-1}. If X has a normal distribution, GLS provides similar estimates to those obtained by ML.

Unweighted Least Squares
The parameter estimators from ULS will be those values that minimize the function

$$F_{ULS} = \text{tr}[(S-\Sigma)^2] \quad (4.42)$$

ULS minimizes the differences between matrices S and Σ. This function differs from the others in that it is not built on an assumption of multivariate normality in the data. As a result, this discrepancy function does not lead to estimated standard errors or an overall chi-square fit statistic.

4.3.4 CFA Model Fit

Once the parameters have been estimated, the next step is to assess how well the estimated model matches the observed data, that is, how well the estimated model may be fit to the empirical covariance matrix. Unfortunately, there is no single measure to evaluate the adequacy between the estimated model and the data. Instead, a large number of tests have been proposed for determining the overall model fit. The most common test, available in all statistical packages, is the χ^2 statistic which is used to test the null hypothesis (H_0) that the implied covariance matrix $\hat{\Sigma}_0$ is equivalent to (H_1) the observed covariance matrix S:

$$H_0: \Sigma = \hat{\Sigma}_0$$

$$H_1: \Sigma = S$$

However, the χ^2 test is widely recognized to be problematic because it is influenced by sample size. But also because it may be invalid when distributional assumptions are violated, causing good models to be rejected while some bad models could be retained. This has led to the proliferation of alternative indices of fit that are based on the chi-square statistic. Marsh, Balla and McDonald (1988) review a compilation of goodness of fit indices in confirmatory analysis

Chi-square: Is the most frequently used index

$$\chi^2 = tr(\Sigma^{-1}S - I) - \log(\Sigma^{-1}S) \quad (4.43)$$

A test with a chi-square statistic can be used with GLS and ML in order to verify the null hypothesis that a given model adequately fit the data. The model fit is obtained by

comparing the observed variance-covariance matrix, *S*, to the estimated variance-covariance matrix, $\hat{\Sigma}$. The larger the difference between these two matrices, the greater the value of χ^2 and thus the worst the fit of the model to the data, indeed the test is a simultaneous test that all residuals (calculated by taking the difference between all model implied covariances and the observed sample covariances) are zero. Conversely, the less the difference between *S* and $\hat{\Sigma}$, the lower the value of χ^2, and consequently the fit of the model will be better.

Goodness-of-fit index (GFI) and Adjusted Goodness-of-fit index (AGFI)
The goodness of fit index is a measure of the relative amount of variances and covariances jointly accounted for by the model (Jöreskog and Sörbom, 1986). The closer the GFI is to 1 the better is the fit of the model to the data.
The AGFI is based on a correction for the number of degrees of freedom in a less restricted model obtained by freeing more parameters. Both the GFI and the AGFI are less sensitive to sample size than the chi-square statistic.

Word of caution
If the model does not fit the data, the proposed model is rejected as a possible candidate for the causal structure underlying the observed variables. If the model cannot be rejected statistically, it is a plausible representation of the causal structure. It is important to mention that finding a model that adequately fits the data does not imply that the model is the only one for that data. Moreover, it is convenient to use more than one fit index to compare results in order to have more support to accept or reject a proposed model.

The main advantage of confirmatory factor analysis over exploratory factor analysis is that CFA allows the researcher to test several competing hypotheses regarding the factors underlying the data.

4.4 Covariance Structure Analysis

We have seen that Factor Analysis (EFA and CFA) aims to explain the correlations among a set of observed variables in terms of fewer unobserved variables called factors. Conversely, Covariance Structure Analysis aims to explain the relationships among a set of unobservable variables (latent variables), each measured by one or more observed variables. CSA can be seen as a generalization of the factor analysis model into a more complex and sophisticated one. For this reason it is said that CSA is a second generation multivariate technique (Fornell, 1982, 1987). CSA is also referred to as Structural Equation Modeling, Linear Structural Relations, Latent Variable Structural Equation Modeling, Moments Structure Models, and Causal Modeling, among others.

A CSA model is integrated by two parts or sub-models: one is the measurement model and the other is the structural model. The measurement model, in fact, is a confirmatory factor analysis model (it represents the relationships of the observed variables to their constructs, or in other words, how each construct is measured by one or more observed variables). The structural model is a simultaneous system of equations among the latent variables (it represents the causal relations among the constructs without considering the MVs). The measurement model in conjunction with the structural model enables a comprehensive, confirmatory assessment for theory testing.

In summary, we can say that a general model for CSA combines a confirmatory factor analysis referred to as the measurement model, and a system of equations among latent variables referred to as the structural model. The main difference between CSA and Factor Analysis (either EFA or CFA) is that CSA involves a set of relationships among the latent variables modeled in terms of a system of simultaneous equations. CSA seeks to describe the variances and covariances of a set of latent constructs in terms of a smaller number of structural parameters; hence the name Structural Equation Modeling.

4.4.1 Model Specification

We describe the specification of the general covariance structural model as it appears in most structural equation modeling textbooks such as Long (1983), Visauta (1986), Hayduk (1987), Bollen (1989), Hoyle (1995), Murayama (1998), Kaplan (2000), and Byrne (2001). Related literature can be found in Jöreskog (1977), Jöreskog and Wold (1982), and Hoyle (2000).

We define the structural model as follows:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \tag{4.44}$$

where η is an mx1 vector of endogenous latent variables, ξ is a kx1 vector of exogenous latent variables, B is an mxm matrix of structural coefficients relating the endogenous variables to each other, Γ is an mxk matrix of structural coefficients relating endogenous variables to exogenous variables, and ζ is an mx1 vector of disturbance terms.

We can rewrite equation (4.44) as

$$\underset{(m\times 1)}{\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_m\end{bmatrix}} = \underset{(m\times m)}{\begin{bmatrix}0 & \beta_{12} & \cdots & \\ \beta_{21} & 0 & & \\ \vdots & & \ddots & \\ \beta_{m1} & & & 0\end{bmatrix}} \underset{(m\times 1)}{\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_m\end{bmatrix}} + \underset{(m\times k)}{\begin{bmatrix}\gamma_{11} & \gamma_{12} & \cdots & \\ \gamma_{21} & \gamma_{22} & & \\ \vdots & & \ddots & \\ \gamma_{m1} & & & \gamma_{mk}\end{bmatrix}} \underset{(k\times 1)}{\begin{bmatrix}\xi_1\\ \xi_2\\ \vdots\\ \xi_n\end{bmatrix}} + \underset{(m\times 1)}{\begin{bmatrix}\zeta_1\\ \zeta_2\\ \vdots\\ \zeta_m\end{bmatrix}} \tag{4.45}$$

Note that matrix B contains zeros in the diagonal because an endogenous construct cannot affect itself.

Structural Model

Considering the structural model of equation 4.44 it can be re-expressed as follows:

$$\eta - \mathrm{B}\eta = \Gamma\xi + \zeta$$

$$(I - \mathrm{B})\eta = \Gamma\xi + \zeta$$

$$\ddot{\mathrm{B}}\eta = \Gamma\xi + \zeta \tag{4.46}$$

If we assume $\ddot{\mathrm{B}}$ is non singular, the reduced form is:

$$\eta = \ddot{\mathrm{B}}^{-1}\Gamma\xi + \ddot{\mathrm{B}}^{-1}\zeta \tag{4.47}$$

The reduced form is an expression in which the endogenous latent variables are expressed in terms of the exogenous latent variables and the disturbance terms.

Measurement Model

The definition of the measurement model is identical to the one presented in the case of the Exploratory Factor Analysis:

$$y = \Lambda_y \eta + \varepsilon \tag{4.48}$$

$$x = \Lambda_x \xi + \delta \tag{4.49}$$

where Λ_y and Λ_x are pxm and qxk matrices of factor loadings, respectively, and ε and δ are px1 and qx1 vectors of error terms.

We can rewrite equation (4.48) as

$$\underset{(p\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}} = \underset{(p\times m)}{\begin{bmatrix} \lambda_{11}^y & \lambda_{12}^y & \cdots & \\ \lambda_{21}^y & \lambda_{22}^y & & \\ \vdots & & \ddots & \\ \lambda_{p1}^y & & & \lambda_{pm}^y \end{bmatrix}} \underset{(m\times 1)}{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_m \end{bmatrix}} + \underset{(p\times 1)}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}} \tag{4.50}$$

and equation (4.49) as

$$\underset{(q\times 1)}{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix}} = \underset{(q\times k)}{\begin{bmatrix} \lambda_{11}^x & \lambda_{12}^x & \cdots & \\ \lambda_{21}^x & \lambda_{22}^x & & \\ \vdots & & \ddots & \\ \lambda_{q1}^x & & & \lambda_{qk}^x \end{bmatrix}} \underset{(k\times 1)}{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_k \end{bmatrix}} + \underset{(q\times 1)}{\begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_q \end{bmatrix}} \tag{4.51}$$

The following assumptions are made:

- all variables are recorded as deviations from their means
  $E(x) = E(\delta) = 0 \qquad E(\xi) = 0$
  $E(y) = E(\varepsilon) = 0 \qquad E(\eta) = 0$
- latent variables and error terms are uncorrelated
  $E(\xi\delta') = 0 \quad$ or $\quad E(\delta\xi') = 0$
  $E(\xi\varepsilon') = 0 \quad$ or $\quad E(\varepsilon\xi') = 0$
  $E(\eta\delta') = 0 \quad$ or $\quad E(\delta\eta') = 0$
  $E(\eta\varepsilon') = 0 \quad$ or $\quad E(\varepsilon\eta') = 0$
- disturbance terms are also uncorrelated
  $E(\delta\varepsilon') = 0 \quad$ or $\quad E(\varepsilon\delta') = 0$

In addition, there are four covariance matrices of interest:

- covariance matrix among the exogenous variables
  $E(\xi\xi') = \Phi$
- covariance matrix among structural errors ζ
  $E(\zeta\zeta') = \Psi$
- covariance matrix among errors ε
  $E(\varepsilon\varepsilon') = \Theta_\varepsilon$
- covariance matrix among errors δ
  $E(\delta\delta') = \Theta_\delta$

The goal of covariance structure analysis is to find an implied-model covariance matrix Σ that corresponds to the actual observed covariances among the observed variables contained in S.

The variance-covariance matrix of the manifest variables can be expressed as

$$\Sigma = \underset{(m+n)\times(m+n)}{\left[\begin{array}{c|c} E(YY') & E(YX') \\ \hline E(XY') & E(XX') \end{array}\right]} \tag{4.52}$$

which must be examined by sub-blocks in order to be re-expressed in terms of the four matrices of structural coefficients B, Γ, Λ_x, Λ_y, and the four covariance matrices of the error terms Φ, Ψ, Θ_ε, and Θ_δ.

First, we examine the variance-covariance matrix among the *x*'s: $E(XX')$

$$\begin{aligned} E(XX') &= E\left[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)'\right] \\ E(XX') &= E\left[\Lambda_x\xi\xi'\Lambda_x' + \delta\xi'\Lambda_x' + \Lambda_x\xi\delta' + \delta\delta'\right] \\ E(XX') &= \Lambda_x E(\xi\xi')\Lambda_x' + E(\delta\xi')\Lambda_x' + \Lambda_x E(\xi\delta') + E(\delta\delta') \\ E(XX') &= \Lambda_x E(\xi\xi')\Lambda_x' + E(\delta\delta') \\ E(XX') &= \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{aligned} \tag{4.53}$$

Then, we examine the variance-covariance matrix among the *y*'s: $E(YY')$

$$\begin{aligned} E(YY') &= E\left[(\Lambda_y\eta + \varepsilon)(\Lambda_y\eta + \varepsilon)'\right] \\ E(YY') &= E\left[\Lambda_y\eta\eta'\Lambda_y' + \varepsilon\eta'\Lambda_y' + \Lambda_y\eta\varepsilon' + \varepsilon\varepsilon'\right] \\ E(YY') &= \Lambda_y E(\eta\eta')\Lambda_y' + E(\varepsilon\eta')\Lambda_y' + \Lambda_y E(\eta\varepsilon') + E(\varepsilon\varepsilon') \\ E(YY') &= \Lambda_y E(\eta\eta')\Lambda_y' + E(\varepsilon\varepsilon') \\ E(YY') &= \Lambda_y E(\eta\eta')\Lambda_y' + \Theta_\varepsilon \end{aligned} \tag{4.54}$$

In order to determine the covariance matrix for the endogenous variables η's it is necessary to use the reduced form $\eta = (I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)$

$$\begin{aligned} E(\eta\eta') &= E\left(\left[(I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)\right]\cdot\left[(I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)\right]'\right) \\ E(\eta\eta') &= E\left((I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)(\xi'\Gamma' + \zeta')(I - \mathrm{B})^{-1'}\right) \\ E(\eta\eta') &= (I - \mathrm{B})^{-1}E\left[\Gamma\xi\xi'\Gamma' + \zeta\xi'\Gamma' + \Gamma\xi\zeta' + \zeta\zeta')\right](I - \mathrm{B})^{-1'} \\ E(\eta\eta') &= (I - \mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - \mathrm{B})^{-1'} \end{aligned} \tag{4.55}$$

So, the covariance matrix among the y's is

$$E(YY') = \Lambda_y (I - \mathrm{B})^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - \mathrm{B})^{-1'} \Lambda_y' + \Theta_\varepsilon \tag{4.56}$$

Finally, we examine the covariance matrix between x's and y's

$$E(XY') = E\left[(\Lambda_x \xi + \delta)(\Lambda_y \eta + \varepsilon)'\right]$$

$$E(XY') = E\left[\Lambda_x \xi \eta' \Lambda_y' + \Lambda_x \xi \varepsilon' + \delta \eta' \Lambda_y' + \delta \varepsilon'\right] \tag{4.57}$$

substituting η by its reduced form we have

$$E(XY') = E\left[\Lambda_x \xi (\xi' \Gamma' + \zeta')(I - \mathrm{B})^{-1'} \Lambda_y' + \Lambda_x \xi \varepsilon' + \delta (\xi' \Gamma' + \zeta')(I - \mathrm{B})^{-1'} \Lambda_y' + \delta \varepsilon'\right]$$

$$E(XY') = \Lambda_x E(\xi \xi') \Gamma' (I - \mathrm{B})^{-1'} + E(\xi \zeta')(I - \mathrm{B})^{-1'} \Lambda_y'$$

$$E(XY') = \Lambda_x \Phi \Gamma' (I - \mathrm{B})^{-1'} \Lambda_y' \tag{4.58}$$

Thus, the variance-covariance matrix of the manifest variables

$$\Sigma = \left[\begin{array}{c|c} E(YY') & E(YX') \\ \hline E(XY') & E(XX') \end{array}\right]$$

can be expressed as

$$\Sigma = \left[\begin{array}{c|c} \Lambda_y \left[(I - \mathrm{B})^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - \mathrm{B})^{-1'} \Lambda_y' + \Theta_\varepsilon\right] & \Lambda_y (I - \mathrm{B})^{-1} \Gamma \Phi \Lambda_x' \\ \hline \Lambda_x \Phi \Gamma' (I - \mathrm{B})^{-1'} \Lambda_y' & \Lambda_x \Phi \Lambda_x' + \Theta_\delta \end{array}\right] \tag{4.59}$$

Table 5.1. Summary for Measurement Model

Matrix	*Mean*	*Covariance*	*Description*
ξ	0	$\Phi = \mathrm{E}(\xi\xi')$	Exogenous LVs
X	0	$\Sigma_{XX} = \mathrm{E}(XX')$	Manifest variables
Λ_x	-	-	Loadings of X on ξ
δ	0	$\Theta_\delta = \mathrm{E}(\delta\delta')$	Disturbance terms for X
η	0	$\mathrm{cov}(\eta) = \mathrm{E}(\eta\eta')$	Endogenous LVs
Y	0	$\Sigma_{YY} = \mathrm{E}(YY')$	Manifest variables
Λ_y	-	-	Loadings of Y on η
ε	0	$\Theta_\varepsilon = \mathrm{E}(\varepsilon\varepsilon')$	Disturbance terms for Y

4.4.2 CSA Model Identification

The logical procedure for confirming identification consists of expressing the unknown parameters in terms of the elements of the covariance matrix by means of algebraic operations. That is, an equation for each observed variance/covariance term is written as a function of the unknown parameters. Then, the equations have to be rearranged so that each unknown parameter is expressed in terms of the known parameters and the known

covariances. The goal is to solve the equations so that each parameter to be estimated can be calculated from the available information.

A complementary strategy to determine whether or not a covariance structure model is identified is to separately examine the identifiability of the measurement model and the structural model. In order to have an identified model both parts (the measurement and the structural) must be identified. The reason for this strategy is that it is easier to check identification separately rather than jointly. However, the identification of models may become enormously difficult as the model increases in complexity, thus identifying all parameters may not be an easy task. As a consequence, researchers have developed several simple methods or rules of thumb that can be applied to various covariance structure models for demonstrating identification without spending a lot of considerable effort in proving direct solutions.

A complete review of the variety of identification conditions exceeds the scope of the present work. As Murayama (1998, p. 191) recognizes, there is disagreement about how to establish necessary and sufficient conditions for identification of covariance structure models. Instead, we will only mention a basic condition: the so called t-rule. As cited by Marcoulides and Hershberger (1997), it is one of the most frequently used necessary identification rules.

The t-rule ***(necessary but not sufficient)***

Basically, the t-rule for identification says that the number of non-redundant elements in the covariance matrix of the observed variables must be greater than or equal to the number of unknown parameters. If t is the number of unknown parameters in a covariance structure model, the necessary condition for identification is

$$t \leq \frac{1}{2}(p+q)(p+q+1) \tag{4.60}$$

where $p + q$ is the number of observed variables. If the number of unknowns, t, exceeds the number of equations $\frac{1}{2}(p+q)(p+q+1)$, then the identification of the model is not possible. The general consensus is that this condition must precede any estimation attempt.

In a more practical oriented approach, we can take into account the outputs of the software to have some insight into the identifiability of models. If the software fails to converge to a solution or if it calculates absurd results, it is very likely that the model is under identified. However, obtaining reasonable results does not prove model identification.

Jöreskog and Sörbom (1978) say that the information matrix calculated in maximum likelihood estimation can be used to see whether a model is identified or not. The information matrix is the matrix of second order derivatives of the fit or discrepancy function with respect to all the free parameters of the model. If the information matrix is positive definite it is almost certain that the model is identified. Nevertheless, this is not a sufficient condition for identification because even if the information matrix is positive definite, the model could be unidentified.

Hayduk (1987, p. 143-144) remarks three interesting points:

- First, no general conditions have been enumerated that guarantee identifiability of the diverse types of models
- Second, the procedures for checking identification are not standardized.
- Third, the non-standardized steps require lengthy hand calculations.

Unfortunately, there are no easily applicable sufficient, or necessary and sufficient, conditions for the full covariance structure model. The general consensus is that in order to have an identified model involves imposing theoretical constraints on the model and/or data constraints on the parameters. Further reading for a more detailed description about the issue of identification and identifiability conditions can be found in Bollen (1989), Hayduk (1987), Long (1983), Murayama (1998), and Pugesek *et al* (2003).

4.4.3 Parameters Estimation

Once we have an identified model, the next step is the estimation of the model. The overall estimation process begins by considering the covariance matrix *S* of the manifest variables. We can think of *S* as a partitioned matrix:

$$S = \left[\begin{array}{c|c} \text{cov}(X) & \text{cov}(X,Y) \\ \hline \text{cov}(Y,X) & \text{cov}(Y) \end{array}\right] \tag{4.61}$$

In terms of parameters estimates we have:

$$S = \hat{\Sigma} = \left[\begin{array}{c|c} \hat{\Lambda}_y \hat{\hat{B}}^{-1}(\hat{\Gamma}\hat{\Phi}\hat{\Gamma}' + \hat{\Psi})\hat{\hat{B}}'^{-1}\hat{\Lambda}_y + \hat{\Theta}_\varepsilon & \hat{\Lambda}_y \hat{\hat{B}}'^{-1}\hat{\Gamma}\hat{\Phi}\hat{\Lambda}_x \\ \hline \hat{\Lambda}_x \hat{\Phi}\hat{\Gamma}'\hat{\hat{B}}'^{-1}\hat{\Lambda}_y & \hat{\Lambda}_x \hat{\Phi}\hat{\Lambda}_x' + \hat{\Theta}_\delta \end{array}\right] \tag{4.62}$$

The objective is to find values of $\hat{\hat{B}}$, $\hat{\Gamma}$, $\hat{\Phi}$, $\hat{\Psi}$, $\hat{\Lambda}_y$, $\hat{\Lambda}_x$, $\hat{\Theta}_\varepsilon$, and $\hat{\Theta}_\delta$, in such a way that the resulting $\hat{\Sigma}$ can reproduce the values of the observed covariance matrix *S* as much as possible. In other words, we want to find those values of the parameters that minimize the differences between the observed covariance matrix *S* and the predicted covariance $\hat{\Sigma}$. In essence, the problem of estimation involves to measuring how close $\hat{\Sigma}$ is to *S*.

The covariance structure model can be estimated by any of the methods discussed in EFA and CFA: maximum likelihood (ML), generalized least squares (GLS), and unweighted least squares (ULS). The fitting functions are shown below.

$$F_{ML} = tr(S\Sigma^{-1}) + \log|\Sigma| - \log|S| - (p+q)$$

$$F_{GLS} = \text{tr}[(S\text{-}\Sigma)S^{-1}]^2$$

$$F_{ULS} = \text{tr}[(S\text{-}\Sigma)^2]$$

4.4.4 CSA Model Fit

Once we have obtained the parameters estimates, the next step is the assessment of fit between the hypothesized model and the observed data. We have seen that a covariance structure model leads to the estimation of a variance-covariance matrix Σ that seeks to reproduce the variance-covariance matrix *S* of the manifest variables. The question now is how to evaluate the extent to which the matrices *S* and Σ differ. In fact, answering this

question implies looking for a fitness measure between *S* and Σ. The expected criterion will be a measure that allows us to conclude that the model represents the observed data reasonably well if the difference between S and Σ is negligible. Conversely, if the difference is large, the desired measure will allow us to conclude that the proposed model is not consistent with the observed data.

There is a large collection of indexes but all have the same goal: to provide a synthetic measure of the overall fit of the model to the available data. Although there are various fit indexes, there is no agreement about a single optimal test or even a set of optimal tests. Bollen and Long (1993) have published an entire book dedicated to model testing. We will present the most widely used indexes (listed below) which are included in practically all statistical software for calculating covariance structure models.

- Chi-Square
- Goodness-of-Fit (GFI)
- Adjusted Goodness-of-Fit (AGFI)
- Root Mean Square Residuals (RMR)
- Comparative Fit Index (CFI)
- Normed Fit Index (NFI)
- Root Mean Square Error of Approximation (RMSEA)
- Akaike's Information Criterion (AIC)
- Consistent version of the AIC (CAIC)
- Bayes Information Criterion (BIC)

Chi-Square

A chi-square test for the hypothesized model can be used in the ML and GLS estimation methods. The chi-square test is based on the maximum likelihood function F_{ML} which is asymptotically distributed as a χ^2 with degrees of freedom equaling the differences between the number of free parameters in the models represented in the denominator and numerator of the ratio.

$$\chi^2 = -2\left\{-\frac{n}{2}\left(tr(S\Sigma^{-1}) + \log|\Sigma| - \log|S| - (p+q)\right)\right\} \tag{4.63}$$

$$\chi^2 = n\left\{tr(S\Sigma^{-1}) + \log|\Sigma| - \log|S| - (p+q)\right\} = nF_{ML} \tag{4.64}$$

Degrees of freedom for the chi-square test are $df = ½(p+q)(p+q+1) - t$ where t is the number of independent parameters being estimated. Large values of the chi-square relative to degrees of freedom indicate that the model does not provide an adequate fit of the data.

Goodness-of-fit (GFI)

GFI is defined as:

$$GFI = 1 - \frac{tr(\Sigma^{-1}S - I)^2}{tr(\Sigma^{-1}S)^2} \tag{4.65}$$

GFI is analogous to a squared multiple correlation; it indicates the proportion of the observed covariance explained by the model covariance. This index ranges from 0 to 1, with a value close to 1 indicating a good fit. GFI is sensitive to large sample sizes leading to high values. It also tends to favor complexity in models.

Adjusted Goodness-of-Fit (AGFI)
The AGFI is a variant of the GFI which includes an adjustment for model complexity, i.e. AGFI adjusts for the number of degrees of freedom in the model:

$$AGFI = 1 - \frac{d_0}{d_m}(1 - GFI) \tag{4.66}$$

In order to overcome the drawbacks of the GFI, AGFI is obtained by using a correction under two competing models: a model consisting of the data observations with d_0 degrees of freedom and an alternative structural model with d_m degrees of freedom. As with the GFI, the AGFI varies from 0 to 1, with values grater than 0.9 considered as well fitting. However, the AGFI has not performed well in some computer simulations and is less popular than the GFI.

Root Mean Square Residual
Root Mean Square Residual is calculated as the square root of the average squared amount by which the observed variances and covariances differ from their estimates, obtained under the assumption that the model is correct.

$$RMR = \sqrt{\frac{\left\{\sum_{i=1}^{p}\sum_{j=1}^{p}(s_{ij} - \sigma_{ij})\right\}}{p}} \tag{4.67}$$

The RMR indicates the average discrepancy between the elements in the sample and hypothesized covariance matrices. Standardized RMR represents the average value across all standardized residuals. In a well-fitting model, the value will be small (less than 0.05). However, this index must be used carefully because wrong models may also have low values.

Comparative Fit Index (CFI)
This index compares the existing model fit with a null model, also known as independence model, in which all variables are assumed to be uncorrelated. CFI is calculated as

$$CFI = 1 - \frac{(\chi_m^2 - d_m^2)}{(\chi_0^2 - d_0^2)} \tag{4.68}$$

CFI ranges from 0 to 1. A value close to 1 indicates a good fit.

Normed Fit Index (NFI)
This index was developed as an alternative to CFI and represents the proportion in the improvement of the overall fit of the existing model relative to a null model

$$NFI = \frac{\chi_a^2 - \chi_b^2}{\chi_n^2} \tag{4.69}$$

where *a* and *b* are alternative models and *n* is the null model. NFI compares the fit of two different models to the same data set. Usually, one of the models is a baseline or null model. NFI ranges from 0 to 1, with 1 indicating a perfect fit. Values below 0.9 express the need to re-specify the model.

Root Mean Square Error of Approximation (RMSEA)
This index is based on the covariance residuals which are the differences between the observed covariances and the hypothesized model covariances. RMSEA is a standardized summary of the average covariance residuals. RMSEA has the following advantages: (1) It does not require comparisons to a null model, (2) it has a known distribution related to the non-central chi-square distribution, and (3) it is less affected by sample size

$$RMSEA = \sqrt{\frac{F_0}{d_m}} \tag{4.70}$$

Akaike's Information Criterion (AIC)
The AIC addresses the issue of parsimony in the assessment of model fit. This index is used in the comparison of two or more models, with smaller values representing a better fit of the hypothesized model. The AIC is a modification of the standard goodness of fit Chi-square that includes a penalty for complexity. AIC is defined as:

$$AIC = -2\log(L) + 2q \tag{4.71}$$

where log (L) is the log-likelihood function and q is the number of parameters in the model. Adding additional parameters (and increasing the complexity of a model) will always improve the fit of a model. The AIC provides a quantitative method for model selection, whether or not models are hierarchical; however, it may not improve the fit enough to justify the added complexity. The AIC is computed for each candidate model and the one with the lowest AIC is selected.

Consistent version of the AIC (CAIC)
The CAIC is a modification of the AIC that takes into account the sample size, it assigns greater penalty to model complexity than the AIC. The AIC and the CAIC assume that no true model exists, but tries to find the best one among those being considered.

Bayes Information Criterion (BIC)
The BIC, assigns more penalty for complexity than the AIC and CAIC, hence it has a greater tendency to pick parsimonious models. The BIC has been compared, in Monte Carlos simulations directly to the AIC and has been found to perform comparably. The BIC assumes that it is in the set of candidate models and that the goal of model selection is to find the true model. This requires that the sample size to be very large.

$$BIC = -2\log(L) + \log(n)q \tag{4.72}$$

Due to its penalization, BIC benefits a more parsimonious models than AIC. However, there is no rationale to use AIC and BIC simultaneously (Saporta, 2008). Although AIC and BIC have similar formulas they originate from different theories: AIC comes from the Information theory and is an approximation of the Kullback-Leibler divergence between the true model and the estimated one. BIC comes from the Bayesian theory and is based on the maximization of the posterior probability of the model, given the data.

Chapter 5

Partial Least Squares Path Modeling (PLS-PM)

Partial Least Squares Path Modeling (PLS-PM) is one of the PLS techniques with a great modeling power. It is considered as a multivariate technique of second generation, thus providing an insight scheme by combining causal modeling with data analysis features. PLS-PM is a statistical method that has been developed for the analysis of structural equation models with latent variables, specially designed to provide an alternative approach to the most well-known LISREL models. As opposed to the covariance-based approach, PLS is prediction oriented aiming to obtain estimates of latent variables for prediction purposes.

The theoretical foundations of PLS are strongly related to the Principal Components framework, especially with the algorithms used to solve principal components based problems. In the PLS field we find the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm developed to calculate principal components. For this reason, PLS-PM can be viewed as a component-based approach to structural equation modeling in the sense that the goal is to estimate the latent variables as components of the manifest variables.

This chapter begins with an overall review of the PLS conceptual backgrounds, particularly a brief description of the Principal Components Analysis followed by the explanation of the NIPALS algorithm. Then, the PLS-PM method is presented in a detailed way. In the last section we briefly discuss the Generalized Structure Component Analysis (GSCA) which has been recently proposed as an alternative to PLS-PM in the component-based approaches to SEM.

5.1 Principal Components Analysis

The conceptual roots of the Partial Least Squares method can be traced back to the Principal Components Analysis (PCA) and the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm. The relationship between PCA and PLS is analogue to that of Factor Analysis and Covariance Structure Analysis models. Principal Components Analysis (PCA) is a multivariate technique in which a number of observed variables are

transformed into a set of new variables, the principal components, which are uncorrelated and explain the variation of the observed variables. Usually, just a few principal components explain most of the variance of the original set of variables. For this reason, PCA can be seen as a method for reduction of dimensionality in which a reduced number of principal components retain most of the variation present in all the observed variables. PCA has its origins in the early work of Karl Pearson (1901) and its posterior development by Harold Hotelling (1933). For many years both techniques had limited applications due to the lack of computational devices appeared until the 1950s.

For the purpose of the thesis, the exposure of the PCA given in the present work is merely descriptive, and it is focused on the basic results of the method. An extensive review of the topic can be found in most textbooks of multivariate techniques such as in Escofier and Pagès (1998), Lebart, Morineau and Piron (2000), Timm (2002), Saporta (2006) and Tenenhaus (2006). For a more detailed presentation and discussion excellent references are Jackson (1991), Aluja and Morineau (1999), and Jolliffe (2002).

5.1.1 Method of Principal Components

Let X_{nxp} be a matrix of n-observations and p-variables of rank a. The columns (or variables) of X are represented by $x_1, x_2, \ldots, x_p$. Without loss of generality the variables are supposed to be centered. The idea is to find a new set of a variables, named principal components, $t_1, \ldots, t_a$, expressed in terms of linear combinations of the observed variables. The linear relation between variables and principal components is:

$$\begin{aligned} t_1 &= u_{11}x_1 + u_{21}x_2 + \ldots + u_{p1}x_p \\ t_2 &= u_{12}x_1 + u_{22}x_2 + \ldots + u_{p2}x_p \\ &\vdots \\ t_a &= u_{1a}x_1 + u_{2a}x_2 + \ldots + u_{pa}x_p \end{aligned} \tag{5.1}$$

In matrix notation we have

$$T = XU \tag{5.2}$$

where T is an nxa matrix of principal components and U is an axp matrix of scores or directional vectors of the principal axis. Figure 5.1 shows a graphical representation of a PCA problem in path diagram notation:

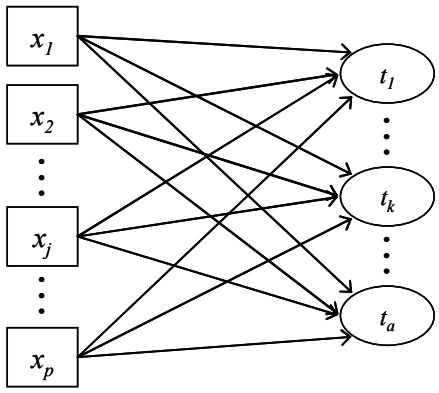


Figure 5.1. *Representation of a PCA problem with a path diagram*

The way to find the principal components (PCs) is to construct them as linear combinations of the original variables following some constraints. The first principal component t_1 is the linear combination $t_1=Xu_1$ which has maximum variance, where u_1 is a vector of p-coefficients. The second principal component t_2 is the linear combination Xu_2, uncorrelated with t_1, which has maximum variance, and so on, so that the k-th component $t_k=Xu_k$ is found having maximum variance subject to being uncorrelated with the previous components $t_1, t_2, ..., t_{k-1}$. Up to $a \leq p$ PCs could be found, but it is hoped that most of the variation in X will be accounted for by m PCs ($m<<a$).

5.1.2 Derivation of the Principal Components

Consider the linear combination $t_1=Xu_1$. The condition of t_1 with maximum variance implies to look for u_1 which maximizes $\text{var}(Xu_1) = u'_1Su_1$, where S is the variance-covariance matrix. It is clear that the maximum will not be achieved for finite u_1, so a normalization constraint must be imposed. The goal is to maximize u'_1Su_1 subject to $u'_1u_1=1$. Using Lagrange multipliers one must maximize:

$$u'_1Su_1 - \lambda(u'_1u_1-1) \tag{5.3}$$

Differentiation with respect to u_1 gives

$$Su_1 - \lambda u_1 = 0 \tag{5.4}$$

or

$$Su_1 = \lambda u_1 \tag{5.5}$$

It is clear that λ is an eigenvalue of S and u_1 is the corresponding eigenvector. Remind that the quantity to be maximized is

$$u'_1Su_1 = u'_1\lambda u_1 = \lambda u'_1u_1 = \lambda \tag{5.6}$$

Maximization implies that λ must be as large as possible, that is, u_1 is the eigenvector corresponding to the largest eigenvalue, λ_1, of S.

As mentioned above, the k-th component $t_k=Xu_k$ is found having maximum variance subject to being uncorrelated with the previous components $t_1, t_2, ..., t_{k-1}$. Then, the process of computing PCs consists in finding the eigenvalues and eigenvectors of the covariance matrix S. This process is based on a key result from matrix algebra which stays that a pxp symmetric, non-singular matrix, may be reduced to a diagonal matrix L by pre-multiplying and post-multiplying it by an orthogonal matrix U such that (Lebart *et al*, 1979)

$$U'SU = L \tag{5.7}$$

The diagonal elements of L, (λ_1, λ_2, …, λ_p) are called the characteristic roots or eigenvalues of S. The columns of U, (u_1, u_2, …,u_p) are called the characteristic vectors or eigenvectors of S.

5.1.3 The Eigenvalue and Eigenvector Problem

From a narrow point of view, computation of principal components reduces to finding the eigenvalues and eigenvectors of a positive-semidefinite matrix such as a covariance

(or correlation) matrix S. The eigenvalue problem has to do with finding nontrivial solutions of the expression $Sx = \lambda x$, which is equivalent to find the roots of the following equation known as the characteristic equation:

$$|S - \lambda I_p| = 0 \tag{5.8}$$

This equation produces a p-th degree polynomial in λ from which the values $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the desired eigenvalues. Thus, finding eigenvalues reduces to finding the roots of a p-th degree polynomial. However, direct computation of the roots from the characteristic equation cannot be obtained in a numerically stable way.

The calculation of the roots, in finite precision, for a p-th degree polynomial has been a problem that has called the attention of many mathematicians for many centuries. In fact, before 1832, attempts for solving polynomials focused on smaller degree equations from quadratics (2^{nd} degree) to quartics (4^{th} degree). General formula for cubics and quartics using arithmetic operations and radicals were developed in the 16th century by Italians mathematicians. Then, for the next 300 years all the attempts made for solving polynomials of degree greater than 4 in terms of radicals failed. Abel in 1827 proved the nonexistence of such a formula for polynomials of degree grater than 4. Finally, a way of deciding whether a given polynomial can be solved in radicals came from the theory developed by the French mathematician Evariste Galois. Brief summaries on the history of solving polynomials may be found in Pan (1997) and Bouchard-Côté (2004).

Having no general formula for the solution of the roots of a polynomial of degree greater than 4, all the efforts to solve nonlinear equations have been focused on using approximate methods. Usually, these methods are based on the idea of successive approximation or on linearization, that is, these are iterative methods: starting from one or more initial approximations to the root, a sequence $x_0, x_1, x_2, \ldots$ is produced which presumably converges to the desired root (Dahlquist *et al*,1974).

5.1.4 The Power Method

Among all the set of methods which can be used to find eigenvalues and eigenvectors, one of the basic algorithms is the Power Method which was already proposed and explained in the seminal paper of Hotelling (1933). In its simplest form, the power method is used to find *the largest* eigenvalue and its corresponding eigenvector. The main interest to explain this algorithm is its relationship with the partial least squares methodology, specifically with the NIPALS algorithm which will be explained in the next section.

The basic idea of the power method is to choose a vector v_0 and applying a variance-covariance matrix S to it repeatedly to form the sequence:

$$\begin{aligned} v_1 &= Sv_0 \\ v_2 &= Sv_1 = S^2v_0 \\ v_3 &= Sv_2 = S^3v_0 \\ &\vdots \end{aligned}$$

In practice one must rescale the vector at each step in order to avoid an eventual overflow or underflow, and to be able to judge whether the sequence is converging. Assuming a reasonable scaling strategy, the sequence of iterates will usually converge to an eigenvector of S.

The following explanation of the convergence is based on Watkins (1982). Suppose matrix S has eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ with $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_p|$. Also assume S has p linearly independent vectors $u_1, ..., u_p$ ordered in such way that u_i corresponds to λ_i. The initial vector v_0 may be expressed as a linear combination of $u_1, ..., u_p$

$$v_0 = a_1 u_1 + a_2 u_2 + \ ... + a_n u_p \tag{5.9}$$

At every step of the iterative process the vector v_m is given:

$$S^m v_0 = a_1 \lambda^m{}_1 u_1 + a_2 \lambda^m{}_2 u_2 + \ ... + a_n \lambda^m{}_p u_p , \qquad m = 1, 2, 3, ... \tag{5.10}$$

Since λ_1 is the *dominant* eigenvalue, the component in the direction of u_1 becomes relatively greater than the other components as m increases. If λ_1 were known in advance, one could rescale at each step by dividing by it to get

$$S^m v_0 / \lambda^m{}_1 = a_1 u_1 + a_2 (\lambda_2 / \lambda_1)^m u_2 + \ ... + a_n (\lambda_p / \lambda_1)^m u_p \tag{5.11}$$

which converges to the eigenvector $a_1 u_1$, provided that a_1 is nonzero. Although in real problems this scaling strategy is unavailable, the eigenvector is determined only up to a constant multiple; the important thing is the direction not the length. To obtain λ_1, when u_1 is the first eigenvector of S, the corresponding eigenvalue is given by the Rayleigh quotient:

$$\lambda_1 = \frac{\langle Su_1, u_1 \rangle}{\langle u_1, u_1 \rangle} \tag{5.12}$$

In words of Gene Golub (2000) "The power method is based on the idea that if a given vector is repeatedly applied to a matrix, and is properly normalized, then ultimately, it will lie in the direction of the eigenvector associated with the eigenvalues which are the largest in absolute value". The power method provides a simple algorithm for finding the first eigenvalue and its associated eigenvector. The speed of convergence depends on how bigger is λ_1 with respect to λ_2 and on the choice of the initial vector v_0. If λ_1 is not much larger than λ_2 the convergence will be slow (Jolliffe, 2002).

One of the advantages of the power method is that it is a sequential method and one can obtain u_1, u_2, and so on, so that if the only first k vectors, ($u_1, u_2, ..., u_k$) are needed, the procedure may be finished at that point. Once the first u_1 eigenvector has been computed, to find the second vector the matrix S must be reduced by the amount explained by the first principal component. This operation of reduction is called *deflation* and the residual matrix is obtained as

$$S - t_1 t_1' \tag{5.13}$$

where $t_1 = \sqrt{\lambda_1}\, u_1$

In order to calculate the second eigenvalue and its corresponding eigenvector, one would operate on $S - t_1 t_1'$ in the same way as the operations on S to obtain u_1. The quantity $t_2' t_2$ will be the amount of variation explained by the second PC.

Another important feature of the power method is that it can be applied not only to symmetric positive-semidefinite matrices (like the variance-covariance matrices), but also to matrices of any given size. Obviously, in case of non-square matrices, the values and vectors obtained will not be eigenvalues nor eigenvectors. In this general case of an nxp matrix X, the power method is used to obtain the Singular Value Decomposition.

5.1.5 Singular Value Decomposition

The Singular Value Decomposition (SVD) is a special factorization technique which may be used as an alternative form of obtaining eigenvalues and eigenvectors. The SVD factorizes a matrix X into the product $V\Lambda U'$ of a unitary matrix V, a diagonal matrix Λ, and another unitary matrix U'. The way in which SVD relates to the obtained principal components of a matrix X is that it is decomposed into a product of the eigenvectors of $X'X$, the eigenvectors of XX', and a diagonal matrix containing the associated eigenvalues. All three matrices are obtained in one operation without having to obtain a covariance (or correlation) matrix (Jackson, 1991).

Given an arbitrary matrix X_{nxp}, it can be written as

$$X = V\Lambda U' \tag{5.14}$$

where V and U are nxa and pxa matrices respectively, each of which has orthogonal columns so that $U'U = I_a$, and $V'V = I_a$. Λ is an axa diagonal matrix; a is the rank of X; and I is the identity matrix of dimension a. The singular values of X are equal to the positive square roots of the eigenvalues of the symmetric matrices $X'X$ and XX'.

The factorization technique of Singular Value Decomposition has the importance of providing an alternative computationally efficient way of computing PCs without the need of a covariance matrix. The reason is that using U and Λ one can compute the eigenvectors and the square roots of the eigenvalues of $X'X$ and hence the PCs for the covariance matrix S.

$$X'X = U\Lambda^{1/2}\Lambda^{1/2}U' \tag{5.15}$$

$$XX' = V\Lambda^{1/2}\Lambda^{1/2}V' \tag{5.16}$$

5.2 The NIPALS Algorithm

The term NIPALS is the acronym of Nonlinear Iterative Partial Least Squares and it is an algorithm to find principal components. It was developed by Herman Wold (1966a) and presented initially under the name NILES (Nonlinear Iterative LEast Squares). The NIPALS algorithm is at the very core of the PLS *building* and its study is one of the keys to understand PLS methods. The following description of NIPALS is based on the exposures given by Tenenhaus (1998, 1999).

Let X_{nxp} be a matrix of n-individuals and p-variables of rank a. The columns (or variables) of X are represented by $x_1, x_2, \ldots, x_p$. Without loss of generality the variables are assumed to be centered. The decomposition formula of matrix X in terms of the principal components is given by $X = TU'$ or

$$X = \sum_{h=1}^{a} t_h u_h' \tag{5.17}$$

where $t_1, \ldots, t_a$ are the principal components and $u_1, \ldots, u_a$ are the score coefficients or directional vectors of the principal axis. In fact, the principal components $t_h=(t_{h1}, \ldots, t_{hn})'$ and the principal axis $u_h=(u_{h1}, \ldots, u_{hp})'$ are column vectors of length n and p respectively.

There are two equivalent ways of considering the decomposition of matrix X: in terms of its columns (variables) and in term of its rows (individuals):

By columns: $x_j = \sum_{h=1}^{a} u_{hj} t_h, \qquad j = 1,\ldots,p$ (5.18)

By rows: $x_i = \sum_{h=1}^{a} t_{hi} u_h, \qquad i = 1,\ldots,n$ (5.19)

By considering X from this double perspective (columns and rows) the principal component factorization can be seen as a double regression model. One can consider u_{hj} as the regression coefficient of t_h in the regression of x_j on t_h; in a similar form t_{hi} can be considered as the regression coefficient of u_h in the regression without constant term of x_i on u_h. These situations are illustrated in the following figure.

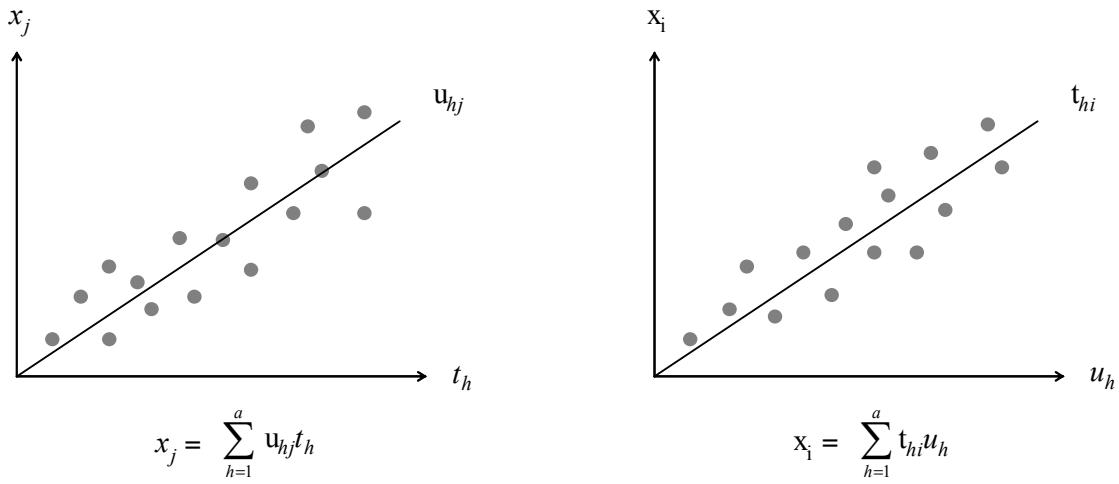


Figure 5.2. *Illustration of the regression lines: by columns (x_j) and by rows (x_i)*

Equation 5.17 can be viewed as a model with two types of parameters, t_h and u_h to be estimated. Moreover, this model can be considered as a nonlinear model because of the product term $t_h u_h'$. This is the reason for the *Nonlinear* word in NIPALS, and hence one of the objectives of NIPALS as a method for linearization of models.

To get the linearization, Wold developed a cleaver procedure by separating the initial nonlinear model into two regression models by fixing one of the parameters in each regression. In the regression for the variables (columns) the fixed parameter is the principal component t_h while the u_{hj} coefficients are the parameters to be estimated. Conversely, in the regression for the individuals (rows) the fixed parameter is the principal axis u_h while the t_{hi} coefficients are the parameters to be estimated. Because both parameters, t_h and u_h, are unknown at the beginning, one must start with some arbitrary value of one of them.

The estimation process follows an iterative procedure alternating simple least squares regressions between the data and one subset of the parameters set, hence the meaning of Partial Least Squares.

5.2.1 Description of the NIPALS algorithm

As said before, the idea of the algorithm is to estimate parameters t_h and u_h by an iterative process of least squares regressions. Although the algorithm is designed to find the complete set of a principal components, it may be adapted to find any desired number PCs, that is, it is not necessary to find all the a PCs, but only a few of them. The steps of the NIPALS algorithm are the following.

NIPALS algorithm steps:

1) $X_0 = X$

2) For h = 1, 2, …, a

2.1) t_h = first column of X_{h-1}

2.2) Repeat until convergence of u_h

2.2.1) $u_h = X'_{h-1}t_h / t_h't_h$

2.2.2) Normalize u_h to 1

2.2.3) $t_h = X_{h-1}u_h / u_h'u_h$

2.3) $X_h = X_{h-1} - t_hu'_h$

For every loop ($h=1,2,...,a$), one must start the iterative procedure giving an initial arbitrary value for t_h. For practical reasons, the initial value is the first column of X_{h-1}. The iterative process combines the computation of both parameters t_h and u_h. Every element u_{hj} of vector u_h represents, before normalization, the regression coefficient of t_h in the regression of variable $x_{h-1,j}$ on the component t_h. On the other hand, the element t_{hi} of vector t_h represents the regression coefficient of u_h in the regression without constant term of $x_{h-1,i}$ on u_h.

After convergence is reached, step 2.3 consists of a deflation step to obtain a residual matrix X_h which will be used to obtain the next ($h+1$) principal component and principal axis. The cyclical relations of step 2.2 when convergence is achieved, shows that t_h and u_h verify the following equations:

$$X'_{h-1}X_{h-1}u_h = \lambda_h u_h \tag{5.20}$$

$$X_{h-1}X'_{h-1}t_h = \lambda_h t_h \tag{5.21}$$

where λ_h is the largest eigenvalue of both matrices. The normalization step of u_h gives the $\lambda_h = t_h't_h$.

It can be seen that the NIPALS algorithm has the important property of following the power method. The estimation of principal components is obtained by means of a series of simple regressions.

Through these first sections we have seen different aspects for obtaining principal components which in turn involves an eigenvalue-eigenvector problem. We have seen how to solve this problem by using the power method and its connection to the NIPALS algorithm. The most important part is the mechanism used in NIPALS to obtain parameters. The general idea and the essence of PLS methods is the estimation process used to calculate model parameters. This process is performed by separating the parameters to be estimated in parts (hence the term *partial*) in order to apply an iterative

procedure of least squares regressions to calculate them. It is important to have in mind these ideas which are the conceptual background of Partial Least Squares.

5.3 PLS Path Modeling Literature Review

As it can be appreciated in the historical review (see chapter 2), the development of PLS path modeling took around a decade to be presented in a well established version. The first developments appeared in 1966 (Wold, 1966a, 1966b) followed by some publications in the early 1970s (Wold, 1973a, 1973b, 1975). An academic paper of 1979 (Wold, 1979), published one year later in 1980 (Wold, 1980) can be considered the "formal" presentation of the PLS approach to latent variable path models. In October of the same year a workshop at Cartigny (Switzerland) took place and it was focused on the two structural equation modeling approaches: the distribution-based maximum likelihood (LISREL), and the distribution-free least squares approach (PLS). The results from that workshop were published by Jöreskog and Wold (1982) in two proceedings volumes: Part I: LISREL and Part II: PLS. Three years later, another Wold's main reference about the PLS algorithm appeared in the *Encyclopedia of Statistical Sciences* (Wold, 1985a). Also in that same year, a less cited work on different aspects about PLS appeared as contributions in *Measuring the Unmeasurable* (Wold, 1985c).

On the computational side, the first and unique available software for many years was LVPLS 1.8 developed by Jan-Bernd Lohmöller. During the 1980s he worked under the supervision of Herman Wold making a comprehensive research on PLS techniques and developing a set of computer programs to make feasible PLS path modeling analyses. In 1989 Lohmöller published the book *Latent Variable Path Modeling with Partial Least Squares* (Lohmöller, 1989) which is considered by many researchers as a hard to understand monograph. The book contains his research results presenting a detailed and wide-ranging account of the capabilities of PLS-PM. In his book statistical, modeling, algorithmic, and programming aspects of the PLS methodology are treated in great depth. He also extended the basic PLS algorithm in various directions to show the scope of problems that can be handled with PLS.

During the first half of the 1990s some interesting works on PLS were published containing descriptions and explanations of PLS path modeling with its practical issues: Falk and Miller (1992), Fornell and Cha (1994), and Barclay *et a*l (1995). Without complex mathematical notation, they provide the logic behind the technique and discuss the differences with other modeling methodologies. Other important research on PLS has been developed by Wynne Chin. Two of his papers on PLS approach to structural equation modeling (Chin, 1998, 1999) are considered as two of the main references, especially oriented for marketing and business researchers. Chin gives a detailed description and explanation of practical and theoretical issues for the application of the PLS-PM methodology. He has also developed the PLS-Graph 3.0 software, which contains a user-friendly graphical interface and cross-validation techniques by jackknife and bootstrap.

In the more theory oriented side, the research work performed by Michel Tenenhaus is of enormous interest and value. His book on PLS Regression (Tenenhaus, 1998) together with other papers (1999, 2001a, 2001b, 2002) show the relationships between PLS and multi-block data analysis methods. Tenenhaus has studied a wide set of multi-block techniques as particular cases of the PLS path modeling, thus showing its great potentiality. A more recently and obliged reference is found in Tenenhaus, Esposito

Vinzi, Chatelin, and Lauro (2005), which gives a complete analysis and description of PLS-PM with a discussion of its extensions and some comparisons with the estimation of structural equation modeling by means of the maximum likelihood approach. In addition, a comprehensive overview of PLS will be available in the *Handbook of Partial Least Squares* (not yet published) with specific works focused on marketing applications and a variety of current developments and perspectives. Finally, there are also the proceedings of the international symposia on PLS and related Methods.

5.4 Specification of a PLS Path Model

PLS path modeling is a statistical method that has been developed for the analysis of latent variable structural models. As opposed to the covariance-based approach (e.g. LISREL), the goal of PLS is to obtain scores of the latent variables for predictive purposes without using the model for explaining the covariation of all the indicators. According to Chin (1998), parameter estimates are obtained based on the ability to minimize the residual variances of all dependent variables (both latent and observed). The key feature of PLS is the explicit estimation of the latent variables, i.e. the so-called scores, by means of least squares methods. We do not pretend here to give the "ultimate" version of PLS-PM. Instead, we only wish to show a unified version as complete as possible, being conscious that PLS-PM has experienced several modifications since its origins, and knowing also that it will probably experience future modifications and adaptations.

As in any structural equation modeling analysis, the first step consists of explicitly specifying a path model which in turn is composed by the structural model and the measurement model. The path model is supposed to be based on a theoretical basis and is graphically represented by means of a path diagram following the rules and notation described in chapter 3. As an example of a path diagram, in figure 5.3 a simple recursive model is shown assuming three latent variables (two exogenous and one endogenous) each of which is supposed to be measured by three reflective indicators.

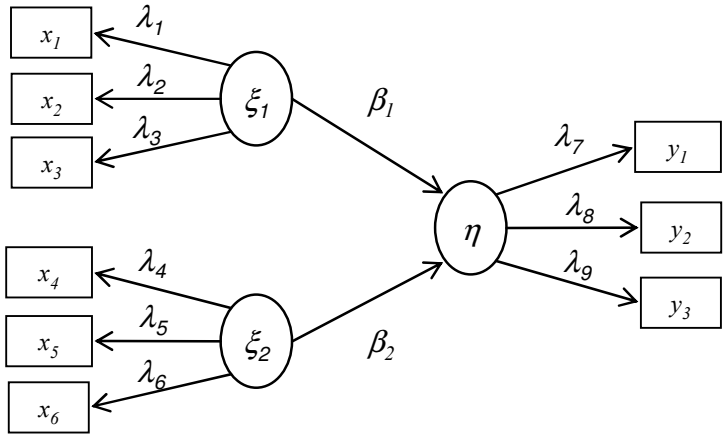


Figure 5.3. *A simple path diagram example: Recursive model with reflective indicators*

As Winn (1998) recognizes, it is necessary to say that in a conceptual level the path diagram within the PLS methodology is used a little bit different from that of a

covariance-based analysis. For the covariance approach, the path diagram indicates the paths by which indicators are associated in a correlational sense. Instead, the path diagram in the PLS approach is used to determine the set of parameters in order to maximize the variance explained for the dependent variables (both latent and manifest variables).

Two more aspects have to be taken into account when specifying a PLS path model. One has to do with the type of path model handled in the PLS approach because the basic design only assumes recursive models, i.e. the path diagram takes the form of a causal chain with no loops. The other aspect has to do with the measurement model, specifically with the relationships between the indicators and their latent variables. In the measurement model the arrows relating indicators to their latent variables can go inwards or outwards, that is, the relationships between a construct and its indicators can be reflective or formative.

About whether the indicators have a reflective or formative nature, the researcher must consider the theoretical frame of the phenomenon under analysis to specify the adequate type of relation. In the reflective way the observed variables are considered being caused by the latent variable (i.e., indicators reflect the construct). In the formative way the latent variables are considered as being caused by the observed variables.

Together with the structural model and the measurement model, the path models in PLS comprise three sets of relations: (1) the *inner model*, which makes reference to the structural model and specifies the relationships between latent variables; (2) the *outer model*, which makes reference to the measurement model and specifies the relationships between constructs and their associated indicators; and (3) the *weight relations* upon which latent variables scores can be calculated.

5.4.1 Inner Model

The Inner model (also known as inner relations) considers only the LVs, which are assumed to be linearly interconnected according to a causal-effect relationship model. The associations among LVs can be represented by a linear multi-equation system which has to be recursive. LVs can play both predictand and predictor roles: an LV that is never predicted is called an exogenous variable, otherwise is called an endogenous variable. For the sake of simplicity, no distinctions in notation are made between endogenous and exogenous constructs; we will denote all latent variables as ξ's. The linear equations take the form:

$$\xi_j = \beta_{0j} + \sum_i \beta_{ji}\xi_i + \zeta_j \tag{5.22}$$

with predictor specification

$$E(\xi_j \mid \xi_i) = \beta_{0j} + \sum_i \beta_{ji}\xi_i \tag{5.23}$$

where the parameter β_{ji} is called the path coefficient (representing the path from the i-th LV to the j-th LV), ζ_j is the inner residual term, and the index i ranges over all predictors of ξ_j. Predictor specification implies

$$E(\zeta_j) = E(\xi_i \zeta_j) = 0 \tag{5.24}$$

which means that the residuals have zero mean and are uncorrelated with the LVs.

5.4.2 Outer Model

The Outer model (also known as outer relations) establishes the relation between a block of MVs and its LV. Because the LV is an unmeasured variable, it has to be indirectly measured trough the MVs, hence the name measurement model. There are three options to establish the connections of the MVs to its LV:

- Reflective way
- Formative way
- MIMIC way (multiple effect indicators for multiple causes)

i) Reflective way
In the reflective way the latent constructs are considered as the cause of the indicators (see figure 5.4)

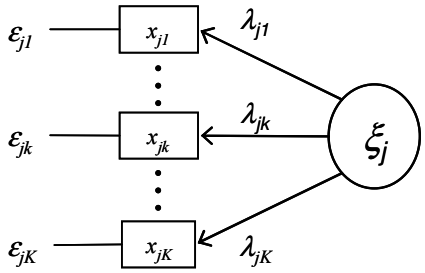


Figure 5.4. *Path diagram of reflective way*

In this sense the MVs can be considered reflects or manifestations of their LV. The MV x_{jk} is assumed to be a linear function of its LV ξ_j

$$x_{jk} = \lambda_{0jk} + \lambda_{jk}\xi_j + \varepsilon_{jk} \tag{5.25}$$

where λ_{jk} is the loading coefficient and ε_{jk} is the outer residual term.
Predictor specification is adopted,

$$E(x_{jk} \mid \xi_j) = \lambda_{0jk} + \lambda_{jk}\xi_j \tag{5.26}$$

which implies that

$$E(\varepsilon_{jk}) = E(\xi_j \varepsilon_{jk}) = 0 \tag{5.27}$$

the residuals have zero mean and are uncorrelated with the MVs.

ii) Formative way
In the formative way the latent constructs are considered as being caused by its indicators (see figure 5.5)

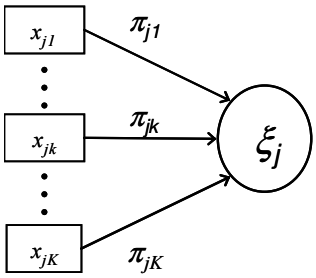


Figure 5.5. *Path diagram of formative way*

The LV ξ_j is assumed to be a linear function of its MVs x_{jk}

$$\xi_j = \pi_{0j} + \sum_k \pi_{jk} x_{jk} + \delta_j \tag{5.28}$$

assuming predictor specification

$$E(\xi_j \mid x_{jk}) = \pi_{0j} + \sum_k \pi_{jk} x_{jk} \tag{5.29}$$

which means that the residuals have zero mean and are uncorrelated with the MVs

$$E(\delta_j) = E(\xi_j \delta_j) = 0 \tag{5.30}$$

iii) MIMIC way

MIMIC way can be considered as a mix of reflective and formative ways.

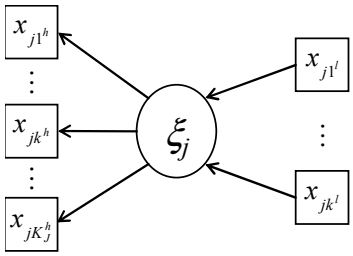


Figure 5.6. *Path diagram of the MIMIC way*

In this case there are two linear equations

$$x_{jh} = \lambda_{0jh} + \lambda_{jh}\xi_j + \varepsilon_{jh} \quad \text{and} \quad \xi_j = \pi_{0j} + \sum_l \pi_{jl} x_{jl} + \delta_j \tag{5.31}$$

where the index h ranges over all reflective MVs, and the index l ranges over all formative MVs, $h+l = k$. Predictor specification is adopted,

$$E(x_{jh} \mid \xi_j) = \lambda_{0jh} + \lambda_{jh}\xi_j \quad \text{and} \quad E(\xi_j \mid x_{jl}) = \pi_{0j} + \sum_l \pi_{jl} x_{jl} \tag{5.32}$$

5.4.3 Weight Relations

Although the outer model specifies the relations between the latent variables and their set of indicators, this specification is done in a conceptual level. In other words, the

outer relations refer to the indicators and the "true" LV. However, we do not really know the true LV. For this reason the weight relations must be defined for completeness. As mentioned before, one of the main characteristics of the PLS approach is the possibility to estimate case values for the LVs (or scores). LV estimates are defined as follows:

$$\hat{\xi}_j = \sum_k \tilde{w}_{jk} x_{jk} \tag{5.33}$$

where $\tilde{w}_{jk}$ are the weights used to estimate the LV as a linear combination of their observed MVs. Note that by using weight relations the problem of factor indeterminacy, present in covariance structure models, is avoided in PLS.

5.4.4 Soft Modeling and Predictor Specification

At first sight, model specification of the structural and measurement models seems to have the same specifications as in the covariance based approach. However, when we analyze the consequences of predictor specification, we note that there are no linear relationships between the predictors and the residuals in any of the linear equations shown previously. In fact, predictor specification takes the form of a linear conditional expectation relationship between a dependent variable and the independent variables (Wold, 1985). It implies that residual terms have zero mean and are uncorrelated with the independent variables (latent or manifest ones). Moreover, the outer model residuals are uncorrelated with all LVs and with the inner model residuals. As consequence, we have that: (1) the Ordinary Least Squares estimates are consistent, and (2) the prediction using OLS estimates is consistent with minimum residual variance. It is also important to remark that PLS does not restrict the structure of the residual covariance. The reason in PLS for not being concerned with the residual covariance is due to LS estimation, which instead is focused on minimizing the residual variance terms.

The relevant implications related to all PLS methodologies are that no assumptions need to be made on the data about some specific multivariate distribution and observations independently distributed. This means that PLS approach avoids the *rigid* assumptions ML estimate methods demand. As opposed to the rigid assumptions referred to as *hard modeling* (i.e. testing hypothesis, need of sound theory), PLS approach is more *flexible*, being known as a *soft modeling* technique (i.e. estimate of latent variables, just prediction relationships needed).

5.5 PLS Algorithm

Once the model has being specified, the next phase in PLS path modeling is the estimation phase carried out by the PLS algorithm. By means of this algorithm, estimation of both the latent variables and the parameters are obtained. The PLS estimation algorithm proceeds in three stages. The first step consists of an iterative procedure of simple and/or multiple regressions taking into account the relationships of the inner model, the outer model and the weight relations. The result is the estimation of a set of weights which are used to calculate the LV scores as linear combinations of their associated manifest variables. Once the LV estimates are obtained, the second and

third steps involve the non-iterative estimation of the structural model coefficients and the measurement model coefficients, respectively.

In its purest essence, the PLS algorithm is nothing more than a series of simple and multiple ordinary least squares regressions, and as Tenenhaus (1998) remarks, the algorithm is of a great simplicity. Nevertheless, beginner users are easily confused with all the multiple options of input settings available during the estimation phase.

Data

The observed manifest variables are generally assumed to be grouped into non-overlapping blocks, each of which represents one latent variable. That is, each indicator is supposed to be associated with just one latent variable. In our case, illustrated in figure 5.7, we will consider the set of manifest variables observed on N cases; arranged into J disjoint blocks $X_1, X_2, ..., X_j, ..., X_J$, each block $X_j = \{x_{j1}, ..., x_{jKj}\}$ of dimension (NxK_j) containing K_j MVs with general term x_{jkn}, $j = 1,\ldots,J$; $k = 1,\ldots,K_j$; $n = 1,\ldots,N$. For the sake of simplicity, the general terms x_{jkn} will be omitted and only the variables x_{jk} will be considered. Unless stated explicitly, variables are supposed to be centered and standardized (zero mean and unit variance).

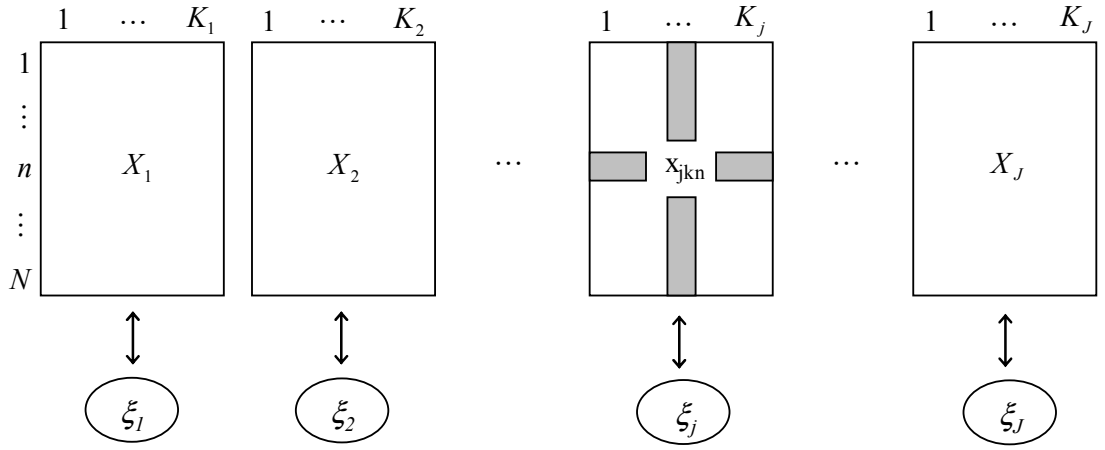


Figure 5.7. *J blocks of manifest variables, having* x_{jkn} *for general term, related with their latent variables*

5.5.1 PLS Algorithm Stage 1

The goal of this stage is the calculation of weights required to give final estimates for each LV ξ_j as a linear combination (Y_j) of its K_j manifest variables, x_{jk},

$$\hat{\xi}_j = Y_j = \sum_k \tilde{w}_{jk} x_{jk} \tag{5.34}$$

where $\tilde{w}_{jk}$ are called outer weights, scaled to give Y_j unit variance. This standardization is done to avoid scale ambiguity of the LV. Since they are unknown, some standardization is required to avoid such scale ambiguity.

Emphasis must be done to remark the fact that this first stage is the "core" stage in the PLS algorithm. The process to calculate the weights follows an iterative mechanism that takes into account the hypothesized relations of the structural and the measurement models. For each model (inner and outer) there is an associated approximation of the LVs: (1) outside approximation for the measurement mode, and (2) inside

approximation for the structural model. Several options for performing first stage are available depending on how the relations between LVs in the structural model are established, and also on how the indicators are associated to their LVs.

Stage 1.1: Outside Approximation

Step 1 of first stage is the outside approximation, also called external estimation. In this step the iterative process begins with an initial proxy of each LV as a linear combination (or weighted aggregate) of its MVs

$$\hat{\xi}_j = Y_j = \pm f_j \sum_k w_{jk} x_{jk} \tag{5.35}$$

where f_j is a scalar that gives Y_j unit variance, and the sign ambiguity $\pm$ is solved by choosing the sign so that the majority of the x_{jk} is positively correlated with Y_j

$$sign\left[\sum_k sign\{cor(x_{jk} Y_j)\}\right] \tag{5.36}$$

The standardized LV is finally expressed as:

$$Y_j = \sum_k \tilde{w}_{jk} x_{jk} \tag{5.37}$$

where the $\tilde{w}_{jk}$ are called the outer weights.

The idea behind the outside approximation is to obtain a set of weights to estimate a latent variable accounting for as much variance as possible for the indicators and the constructs. The algorithm begins with an initial outside approximation of the LVs by using arbitrary weights which are scaled to obtain unit variance for the LVs. Following Chin's (1999) suggestion, we can set initial weights with equal value to perform a first approximation of the latent variable as a simple sum of its indicators. This option is based on the scenario where the researcher, having no additional information, would obtain the best first approximation of the LV as a summation of the MVs.

Stage 1.2: Inside Approximation

After the external approximation step, the next step is the inside approximation, also called internal estimation. In this step the connections among LVs in the inner model are taken into account in order to obtain a proxy of each latent variable calculated as a weighted aggregate of its adjacent LVs. The internal estimation Z_j of ξ_j is defined by:

$$Z_j = \left(\sum_{\substack{i:\beta_{ij}\neq 0,\\ \beta_{ji}\neq 0}} e_{ji} Y_i \right) \tag{5.38}$$

where e_{ji} are the inner weights which are assumed to be scaled so that the variable in parentheses is standardized.

The connections among LVs in the Inner Model are taken into account only when two LVs are connected by an arrow. For example, consider the next path diagram (see figure 5.8):

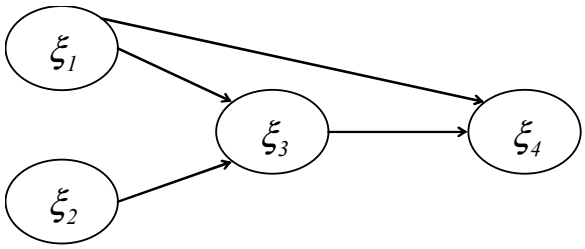


Figure 5.8. *Illustrative path diagram for Inner Model connections*

ξ_4 is joined to ξ_1 and ξ_3 but not to ξ_2. However ξ_2 affects ξ_4 indirectly by means of ξ_3. Inner weights e_{ji} between two constructs exist only when there is an arrow between ξ_j and ξ_i; $e_{ji} \Leftrightarrow Y_j$ and Y_i are adjacent.

There are three options to calculate the inner weights:

- Centroid scheme
- Factor scheme
- Path scheme.

i) Centroid scheme: The centroid scheme is the Wold's original algorithm scheme, whereas the other two are implemented in Lohmöller's version. The inner weights are defined as:

$$e_{ji} = \begin{cases} sign\{cor(Y_j, Y_i)\} & \xi_j, \xi_i \_ adjacents \\ 0 & otherwise \end{cases} \tag{5.39}$$

This scheme only considers the sign direction of the correlations between a LV and its adjacent (neighboring) LVs. It does not consider the direction nor the strength of the paths in the structural model. Some problems may be present when a correlation is close to zero, causing a sign changes during the iterations from +1 to -1.

ii) Factor scheme: The inner weights are taken as

$$e_{ji} = \begin{cases} cor(Y_j, Y_i) & \xi_j, \xi_i \_ adjacents \\ 0 & otherwise \end{cases} \tag{5.40}$$

To avoid the problems of the centroid scheme, Lohmöller proposed the factor scheme which uses the correlation coefficient as the inner weight instead of using only the sign of the correlation. This scheme considers not only the sign direction but also the strength of the paths in the structural model.

iii) Path scheme: In this case the LVs are divided in antecedents (predictors) and followers (predictands) depending on the cause-effects relationships between two LVs. An LV can be either a follower, if it is caused by another LV, or an antecedent if it is the cause of another LV. For example, consider again the diagram in figure 5.8. LV ξ_3 has two antecedents, ξ_1 and ξ_2, and a follower ξ_4. If ξ_i is a follower of ξ_j then the inner weight is equal to the correlation between Y_i and Y_j. On the other hand, for the

antecedents ξ_i of ξ_j, the inner weights are the regression coefficient of Y_i in the multiple regression of Y_j on the Y_i's associated to the antecedents of ξ_j.

$$e_{ji} = cor(Y_j, Y_i) \qquad \text{if } \xi_j \text{ is explained by } \xi_i \tag{5.41}$$

$$Y_j = \sum_i e_{ji} Y_i \qquad \text{coeffs. } e_{ji} \text{ in the regression of } Y_j \text{ on the } Y_i\text{'s} \tag{5.42}$$

The path weighting scheme has the advantage of taking into account both the strength and the direction of the paths in the structural model. However, this scheme presents some problems when the LV correlation matrix is singular

In practice, choosing one weighting scheme in particular over the others has little relevance on the estimation process. As Tenenhaus *et al* (2005) mention, it has been observed that they do not influence the results significantly. However, in a more theoretical level, they are of a great importance to understand how PLS-PM can be applied to different techniques of multiple table analysis.

Stage 1.3: Updating Outer Weights

We can conceive the inside approximation as a stage where the information contained in the inner relations is incorporated in the estimation process of the latent variables. Once the inside approximation is done, the internal estimates Z_j must then be considered with regard their indicators. This is done by updating the outer weights

There are basically two ways of calculating the outer weights w_{ji}: (1) mode A, and (2) mode B. However, there is also a third option called mode C, which is rarely used in practice. Each mode corresponds to a different way of relating the MVs with the LVs in the theoretical model. Mode A is used when the indicators are related to their latent variable through a reflexive way. Instead, mode B is preferred when indicators are associated with their latent variable in a formative way. Mode C is supposed to be used when the indicators of an LV are connected by MIMIC way.

i) Mode A: In the reflective way, each weight w_{jk} is the regression coefficient of Z_j in the simple regression of x_{jk} on Z_j, i.e. the simple regression $x_{jk} = w_{jk}Z_j$ where:

$$w_{jk} = (Z_j' Z_j)^{-1} Z_j' x_{jk} = cor(x_{jk}, Z_j) \tag{5.43}$$

as Z_j is standardized (see figure 5.9)

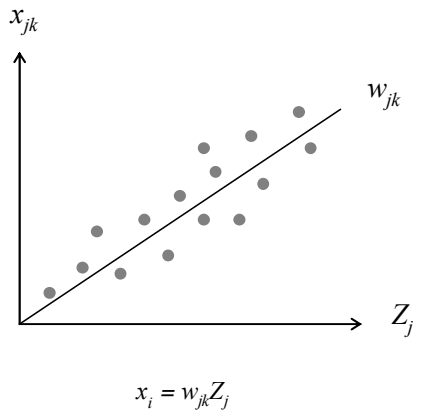


Figure 5.9. *Illustration of a simple regression of x_{jk} on Z_j in mode A*

Note that the covariance between variable x_{jk} and the latent variable Z_j is used without considering how x_{jk} is related to other variables in block X_j. In other words, it does not matter if variables in block X_j are highly correlated, mode A guarantees statistical stabilization of Y_j in the outside approximation.

ii) Mode B: In the formative way, Z_j is regressed on the block of indicators related to the latent construct ξ_j, and the vector w_j of weights w_{jk} is the regression coefficient in the multiple regression

$$Z_j = \sum_k w_{jk} x_{jk} \tag{5.44}$$

defined by

$$w_j = (X_j' X_j)^{-1} X_j' Z_j \tag{5.45}$$

where X_j is the matrix with columns of MVs x_{jk}.

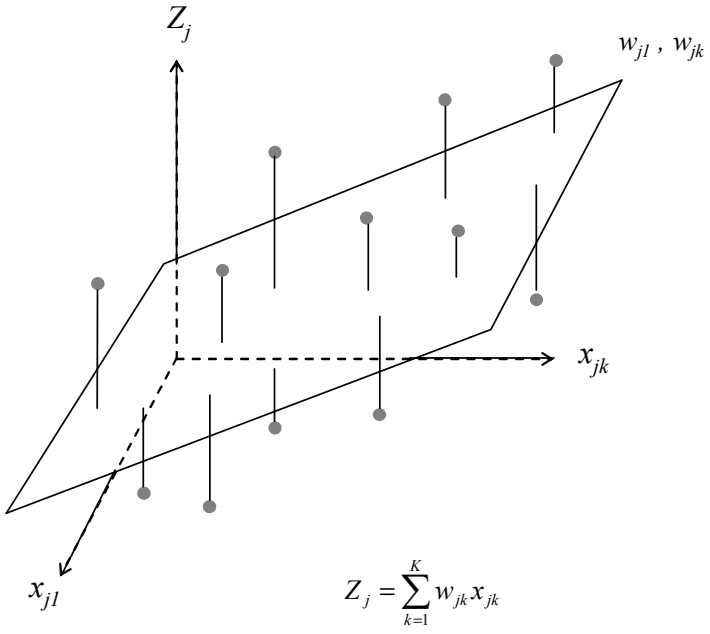


Figure 5.10. *Z_j is regressed on the block of MVs related to the construct ξ_j in mode B*

In this case, we might have some problems when variables x_{jk} in X_j are highly correlated, causing the estimation process to become unstable.

Beside the theoretical reasons to use mode A with reflective way and mode B with formative way, there are also some practical reasons to prefer some mode than other. As Tenenhaus *et al* (2001a) say, if we have more variables than cases in one block X_j, or if the variables x_{jk} of one block X_j are very correlated, the applied option will be mode A. Additionally, they also suggest the use of PLS regression instead of multiple regression.

iii) Mode C: The mode C is implemented in Lohmöller's version and it is a special case of mode B. The MIMIC way is a kind of mix between reflective and formative ways, so the path coefficients for the *h* MVs related in a reflective way are estimated by a simple

linear regression: $x_{jh} = p_{jh}Z_j$; an the path coefficients for the l MVs related in a formative way are estimated by a multiple linear regression:

$$Z_j = \sum_l g_{jl} x_{jl} \; ; \qquad h + l = k \tag{5.46}$$

Stage 1.4: Check for convergence

In every iteration step, say S = 1, 2, 3,…, convergence is checked comparing the outer weights of step S against the outer weights of step *S-1*. For example Wold (1982a) proposed $\left|\tilde{w}^{s-1}{}_{jk} - \tilde{w}^{s}{}_{jk}\right| < 10^{-5}$ as a convergence criterion. It is important to say that, as Tenenhaus *et al* (2002) mention, convergence is not guaranteed although it is always found in practice.

The following figure represents a diagram summarizing the steps of the first stage in the PLS algorithm.

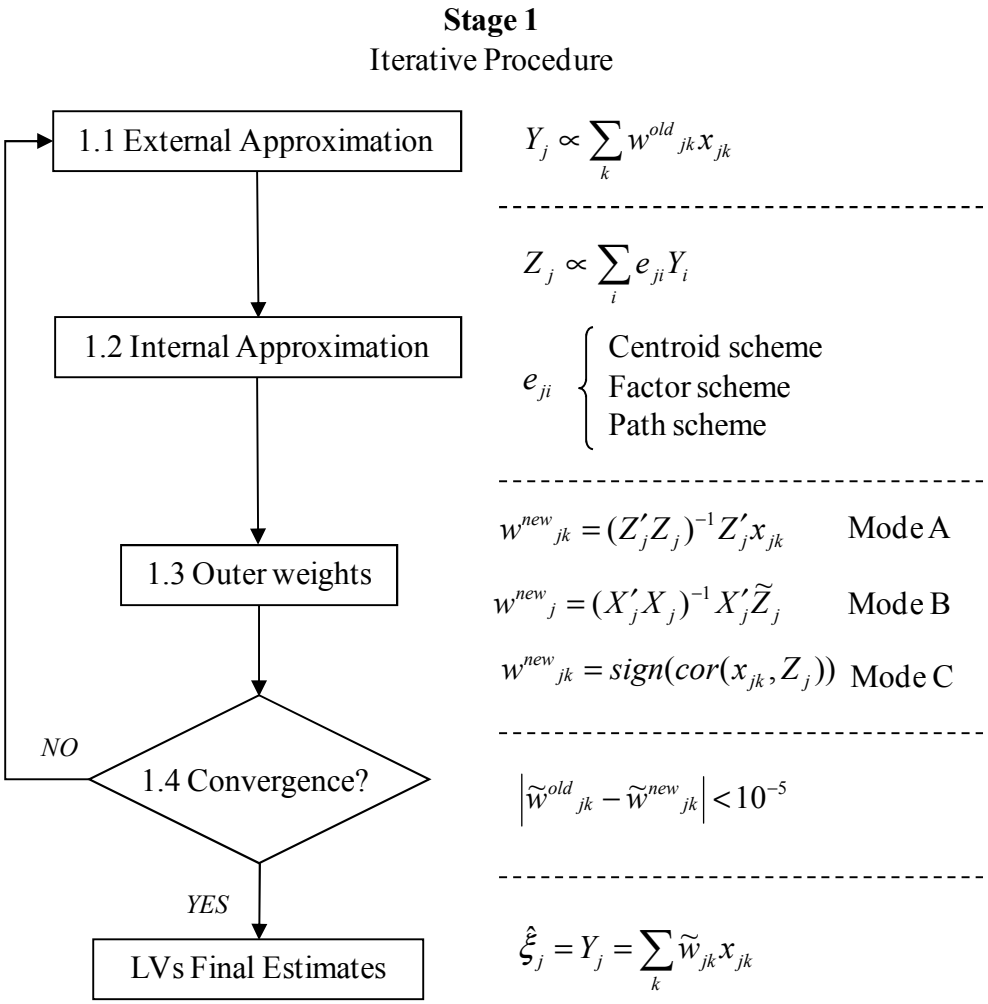


Figure 5.11. *Diagram of the Iterative Process in Stage 1*

5.5.2 PLS Algorithm Stage 2

The second stage of the algorithm consists in the calculation of the path and loading coefficient estimates, $\hat{\beta}_{ji}$ and $\hat{\lambda}_{jk}$, according to the inner and outer models.

For the structural model the loading coefficients are estimated by ordinary least squares in the multiple regression of Y_j on the Y_i's related to it,

$$Y_j = \sum_i \hat{\beta}_{ji} Y_i \tag{5.47}$$

$$\hat{\beta}_{ji} = (Y_i' Y_i)^{-1} Y_i' Y_j \tag{5.48}$$

For the measurement model the path coefficients are estimated depending on the corresponding way. In the reflective way, the loading coefficients are the regression coefficients of the simple linear regression of x_{jk} on Y_j,

$$x_{ij} = \hat{\lambda}_{jk} Y_j \tag{5.49}$$

$$\hat{\lambda}_{jk} = (Y_j' Y_j)^{-1} Y_j' x_{jk} \tag{5.50}$$

In the formative way, the weight coefficients π*'s* coincide with the outer weights obtained in the first stage. This is because we perform the multiple linear regression of Y_j on the x_{jk},

$$Y_j = \sum_k \hat{\pi}_{jk} x_{jk} \tag{5.51}$$

$$\hat{\pi}_{jk} = (X_j' X_j)^{-1} X_j' Y_j = w_{jk} \tag{5.52}$$

Hence, the weight coefficients $\hat{\pi}_{jk}$ are equal to the outer eights w_{jk}.

Scales and Metrics

If we look at the predictor specification equations (shown below) we can observe three more parameters that we have not estimated at all: β_{0j}, λ_{0jk} (in reflective way), and λ_{0j} (in formative way)

$$E(\xi_j \mid \xi_i) = \beta_{0j} + \sum_i \beta_{ji} \xi_i \qquad \text{(structural model)} \tag{5.53}$$

$$E(x_{jk} \mid \xi_j) = \lambda_{0jk} + \lambda_{jk} \xi_j \qquad \text{(reflective way measurement model)} \tag{5.54}$$

$$E(\xi_j \mid x_{jk}) = \pi_{0j} + \sum_k \pi_{jk} x_{jk} \qquad \text{(formative way measurement model)} \tag{5.55}$$

These parameters correspond to the location parameters, that is, we take into account the mean of the manifest and latent variables. However, until now, we have only considered standardized manifest variables (zero mean and unit variance). In fact, because it has been imposed that way during the algorithm, the estimated latent variables are also standardized. In order to obtain the location parameters the researcher must consider whether it makes sense to calculate them. This decision concerns data scales which are the key criteria to decide whether to estimate location parameters.

We must say that this aspect on scales is not considered in Wold's original algorithm. It was developed by Lohmöller (1989) who extended PLS-PM to applications with

mixtures of categorical and interval-scaled data. He presented four different treatments for MVs standardization based on three criteria: (1) comparability between variable scales; (2) interpretability of the means; and (3) importance of the variables reflected on their variances. As a result, he introduced a standardization parameter called *metric* which is described in detail in Tenenhaus *et al* (2005). However, among PLS-PM researchers it is admitted that Lohmöller's standardization options lead to some confusion (i.e., it is not always clear which metric option should be applied). In fact, only two options for data standardization are employed in practice according to the following criteria:

- If the scales of the indicators are not comparable, then they should be standardized (zero mean, unit variance).
- If (1) the scales of the indicators are comparable, (2) their means are interpretable and (3) their variance reflect their importance, then no standardization is needed (use raw data).

For example, consider two variables 'temperature' and 'age'. Assume that temperature is measured in Celsius degrees and age is measured in years. Since both variables are not comparable and their difference is not interpretable, the location parameters are meaningless. Hence, data standardization must be used. Conversely, if all manifest variables represent preference attitudes measured in the same scale, it is perfectly assumable that they are comparable. Moreover, we can assume that their means are interpretable and their variances reflect their importance. In this case raw data is kept without any standardization, and location parameters can be estimated following the next procedure:

Before calculating location parameters, means for the LVs estimates are obtained as:

$$\hat{m}_j = \sum_k \widetilde{w}_{jk} \bar{x}_{jk} \tag{5.56}$$

$$\hat{\xi}_j = Y_j + \hat{m}_j \tag{5.57}$$

Location parameters estimates are obtained as follows

$$\hat{\beta}_{0j} = b_{0j} = \hat{m}_j - \sum_i b_{ji} \hat{m}_i \tag{5.58}$$

$$\hat{\lambda}_{0jk} = \bar{x}_{jk} - \hat{\lambda}_{jk} \hat{m}_j \tag{5.59}$$

$$\hat{\pi}_{0j} = \hat{m}_j - \sum_k \hat{\pi}_{jk} \bar{x}_{jk} \tag{5.60}$$

Another point of interest is related when all MVs are measured in the same scale. In this situation, recommended by Fornell (1992) and comented by Bayol *et al* (2000) and Tenenhaus *et al* (2005), it helps the analysts to have LVs expressed in the original scale. To do this, the outer weights $\widetilde{w}_{jk}$ must be positive, and the LVs $\hat{\xi}_j^*$ are expressed in the original scale using normalized outer weights $\hat{w}_{jk}$

$$\hat{w}_{jk} = \frac{\widetilde{w}_{jk}}{\sum_k \widetilde{w}_{jk}} \tag{5.61}$$

$$\hat{\xi}_j^* = \sum_k \hat{w}_{jk} x_{jk} = \frac{\sum_k \tilde{w}_{jk} x_{jk}}{\sum_k \tilde{w}_{jk}} \tag{5.62}$$

The relationship between $\hat{\xi}_j$ and $\hat{\xi}_j^*$ is

$$\hat{\xi}_j^* = \frac{\hat{\xi}_j}{\sum_k \tilde{w}_{jk}} \tag{5.63}$$

Consistency at large

One of the drawbacks in PLS, explicitly recognized by Wold (1982a, p.25), is that it can lead to biased parameter estimates. This bias is manifested in higher estimates for loadings and lower estimates for path coefficients. However, the estimates will approach the "true" latent variable scores as both the number of indicators per block and the sample size increase. This situation is known as "consistency at large". Consistency at large, as employed in the statistical sense, implies that estimation error decreases as sample size increases. In other words, any estimated PLS coefficients will converge on the parameters of the model as both sample size and number of indicators in the model become "infinite". Hui and Wold (1982) indicate that PLS "estimates will in the limit tend to the true values as the sample size increases indefinitely, while at the same time the block sizes increase indefinitely but remain small relative to the sample size". According to Chin (1998), parameter estimates "will be asymptotically correct under joint conditions of consistency (i.e, large sample size) and consistency at large (i.e, the number of indicators per latent variable becomes large).

Recapitulation and Comments

As it can be seen, the goal of PLS is to obtain score values of latent variables for prediction purposes. The idea is to calculate estimates of latent variables as linear combinations of their associated indicators using a *special* linear combination. We look for a linear combination in such a way that the obtained latent variables take into account the relationships of the structural and the measurement models in order to maximize the explained variance of the dependent variables (both latent and observed variables). Figure 5.12 represents a general view of the PLS algorithm steps

The core of the PLS algorithm is the calculation of the weights (for the linear combination) required to estimate the latent variables. The weights are obtained based on how the structural and the measurement model are specified. This is done by means of an iterative procedure in which two kinds of approximation for the latent variables are alternated until convergence of weight estimates. These two types of approximation, called the inside approximation and the outside approximation, have to do with the inner relations and the outer relations, respectively.

The algorithm begins with arbitrary initial weights used to calculate an outside approximation of the latent variables, that is, initial weights are given in order to approximate the LVs as linear combinations of their MVs. Then, the inner relations among LVs are considered in order to calculate the inside approximations, having the option of choosing between three possible scenarios, called weighting schemes, to perform this approximation: (1) centroid, (2) factor, and (3) path scheme. Once the

inside approximations are obtained, the algorithm turns around to the outer relations when new weights are calculated considering how the indicators are related to their constructs: by mode A (reflective), or by mode B (formative). Mode A implies simple linear regressions while mode B implies multiple linear regressions. The simple and/or multiple regressions coefficients are then used as new weights for an outside approximation. The process continues iteratively until convergence of the weights is reached.

After convergence of the weights, and once the latent variables are estimated, the parameters of the structural and the measurement models can be obtained. The structural coefficients, also known as path coefficients, are calculated by ordinary least squares regressions between LVs. There are as many regressions as endogenous latent variables. The parameters of the measurement model, the loading coefficients, are also estimated by least squares regressions but taking into account the kind of mode to be used (reflective or formative).

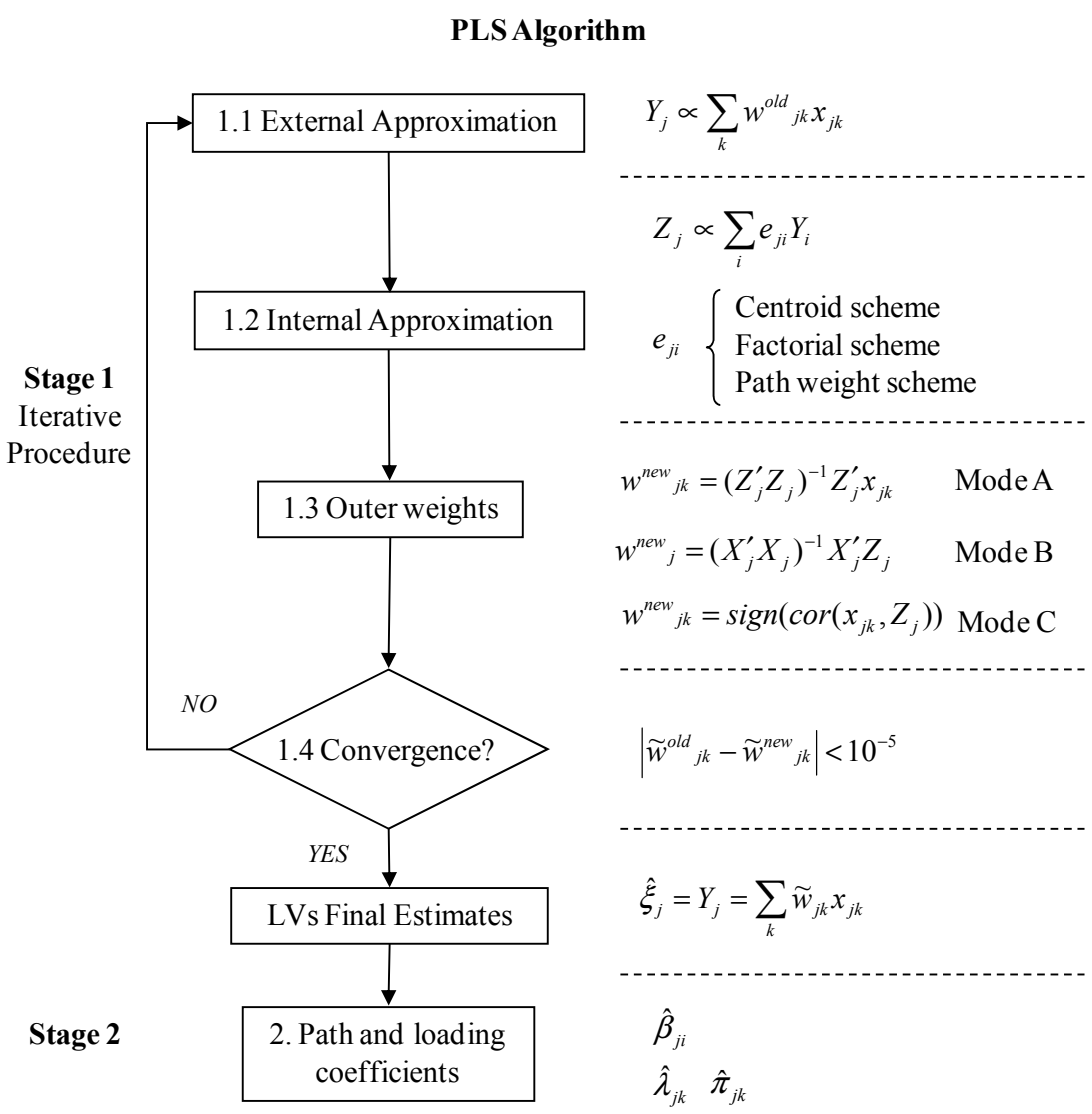


Figure 5.12. *General view of PLS Algorithm*

5.6 PLS Path Model Validation and Diagnosis

We know that a PLS path model consists of a structural model and a measurement model. Then, the validation of a PLS path model requires the analysis and interpretation of both the structural and the measurement model. This validation can be considered as a two-stage process: (1) the assessment of the measurement model, and (2) the assessment of the structural model. This order has to be respected because we must first check that we are really measuring what we are assuming to measure, before any conclusions can be drawn regarding the relationships among the latent variables.

One remarkable aspect is that no single criterion exists within the PLS framework to measure the overall quality of a model, so we cannot perform inferential statistical tests for goodness of fit. As an alternative, non-parametrical tests can be applied for the assessment of the structural model.

5.6.1 Measurement Model Validation: Reflective Measures

It is important to differentiate the assessment of the measurement model depending on whether the indicators' nature is reflective or formative. In the case of reflective measures as illustrated in figure 5.13, they are supposed to measure the same underlying latent variable or construct, that is, they are reflections of the construct.

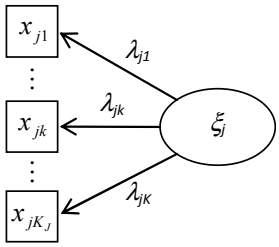


Figure 5.13. *Path diagram of reflective indicators*

Basically, we must evaluate three aspects of reflective measures:

- Unidimensionality of the indicators
- Check that indicators are well explained by its latent variable
- Assess the degree to which a given construct is different from other constructs

Unidimensionality of indicators

In the reflective way the MVs are considered as being caused by their LV. This means that if a construct changes (increases or decreases), then the indicators associated with it will also change in the same direction. Thus, it is logical to suppose that the indicators are closely related in such a way that they are in one dimensional space. As it is mentioned in Tenenhaus *et al* (2002), there is no lack of generality by thinking in this form and, at the theoretical level, the MVs can always be built in this way. However, in practice the unidimensionality of a block has to be confirmed due to the possibility that one or more indicators may be uncorrelated with the rest.

In order to assess the extent to which a block is unidimensional different methods can be applied. Usually, they are defined taking into account one of the following

criteria: (1) in terms of unit rank; (2) percentage of variance explained by the first principal component, (3) and type of correlation. Nonetheless, Hattie (1984) recognizes the lack of an effective index of the unidimensionality of a block of indicators.

Some recent tools have been proposed to evaluate unidimensionality of PLS-PM reflective blocks (Sahmer *et al*, 2005), but the most common methods employed for this purpose are the following three indices:

- Check the first eigenvalue of the MVs correlation matrix
- Calculate the Cronbach's alpha
- Calculate the Dillon-Goldstein's ρ

i) First eigenvalue

The use of principal components analysis is based on the importance of the eigenvalues, and it is related to the so called Kaiser's criterion. If a block is unidimensional, then the first eigenvalue of the correlation matrix of the MVs should be ("much more") larger than one whereas the second eigenvalue should be smaller than 1. In this way, the assessment of the first eigenvalue differs from the Kaiser's criterion since it is not used to extract the number of components (which is considered one of the least accurate methods for deciding which components to extract from a PCA). The evaluation of the first eigenvalue is performed in regards to the rest of the eigenvalues in order to have an idea of how unidimensional is a block of indicators.

ii) Cronbach's alpha coefficient

Cronbach's alpha coefficient (Cronbach, 1951) is another criterion used to assess a block's unidimensionality; it evaluates how well a block of indicators measure their corresponding latent construct. In this case, the observed variables are required to be standardized and positively correlated.

Assuming that the *j*-th block has *p* variables, the variance of the sum of variables in the block is calculated as:

$$Var\left(\sum_{k=1}^{p} x_{jk}\right) = \sum_{k=1}^{p} \text{var}(x_{jk}) + \sum_{k\neq k'} \text{cov}(x_{jk}, x_{jk'}) = p + \sum_{k\neq k'} cor(x_{jk}, x_{jk'}) \qquad (5.64)$$

Note that the larger $\sum_{k\neq k'} cor(x_{jk}, x_{jk'})$, the more the block is unidimensional, because we expect variables to be highly correlated. Now, consider the following ratio:

$$\alpha' = \frac{\sum_{k\neq k'} cor(x_{jk}, x_{jk'})}{p + \sum_{k\neq k'} cor(x_{jk}, x_{jk'})} \qquad (5.65)$$

In the extreme case that all pair-wise correlations cor(x_{jk},$x_{jk'}$) are equal to 1, we have

$$\sum_{k\neq k'} cor(x_{jk}, x_{jk'}) = \binom{p}{2} = \frac{p(p-1)}{2} \qquad (5.66)$$

So, the maximum value of α' is equal to $\frac{p-1}{p+1}$, obtained when all the cor(x_{jk},$x_{jk'}$) are equal to 1:

$$\max \alpha' = \frac{\frac{p(p-1)}{2}}{p+\frac{p(p-1)}{2}} = \frac{p(p-1)}{2p+p^2-p} = \frac{p(p-1)}{p^2+p} = \frac{p(p-1)}{p(p+1)} = \frac{p-1}{p+1} \tag{5.67}$$

Cronbach's alpha is then obtained by diving α' by its maximum value:

$$\alpha = \frac{\sum_{k\neq k'} cor(x_{jk}, x_{jk'})}{p+\sum_{k\neq k'} cor(x_{jk}, x_{jk'})} \times \frac{p-1}{p+1} \tag{5.68}$$

One can see from formula 5.68 that if the number of variables increases, Cronbach's alpha increases as well. As it is expected, if the average inter-variable correlation is low, alpha will be low. Conversely, if the average inter-variable correlation increases, Cronbach's alpha increases as well. As a rule of thumb, a block is considered as unidimensional when Cronbach's alpha is larger than 0.7.

iii) The Dillon-Goldstein's ρ

As in the case of Cronbach's alpha, the Dillon-Goldstein's ρ is also focused on the variance of the sum of variables in the block of interest, but in this case the measurement model specification in equation 5.25 is used. That is, observed variables are defined in terms of their corresponding construct according to:

$$x_{jk} = \lambda_{0k} + \lambda_{jk}\xi_j + \varepsilon_{jk} \tag{5.69}$$

The variance of the sum of indicators is expressed as:

$$Var\left(\sum_{k=1}^{p} x_{jk}\right) = Var\left(\sum_{k=1}^{p}\left[\lambda_{0k} + \lambda_{jk}\xi_j + \varepsilon_{jk}\right]\right) \tag{5.70}$$

If we assume independence of residual terms ε_k then

$$Var\left(\sum_{k=1}^{p} x_{jk}\right) = \left(\sum_{k=1}^{p} \lambda_{jk}\right)^2 \text{var}(\xi_j) + \sum_{k=1}^{p} \text{var}(\varepsilon_{jk}) \tag{5.71}$$

Recall that loadings λ_k in reflective way are regression coefficients of the simple linear regression $x_k = \lambda_k\xi$, that is: $\lambda_k = (\xi'\xi)^{-1}\xi' x_k$

Moreover, as both latent and manifest variables are standardized, λ_k is the correlation between the construct and its indicator. Thus, the larger the loadings, the more a block is unidimensional. Dillon-Goldstein ρ is defined by

$$\rho = \frac{\left(\sum_{k=1}^{p} \lambda_{jk}\right)^2 \text{var}(\xi_j)}{\left(\sum_{k=1}^{p} \lambda_{jk}\right)^2 \text{var}(\xi_j) + \sum_{k=1}^{p} \text{var}(\varepsilon_{jk})} \tag{5.72}$$

Because in practice we do not know the real values of λ_{jk} and ξ_j, an estimate of Dillon-Goldstein's ρ is needed. The approximation of the latent variable is achieved by using the first principal component t_{j1} of the j-th block of indicators; the approximation of the

loading coefficient is taken as the correlation between t_1 and the observed variable x_k, cor(t_1,x_k); the term var(ε_k) is approximated by 1-cor^2(t_1,x_k). Then, estimate of Dillon-Goldstein's ρ is given by:

$$\hat{\rho} = \frac{\left[\sum_{k=1}^{p} cor(x_{jk}, t_{j1})\right]^2}{\left[\sum_{k=1}^{p} cor(x_{jk}, t_{j1})\right]^2 + \sum_{k=1}^{p}\left(1 - cor^2(x_{jk}, t_{j1})\right)} \tag{5.73}$$

As a rule of thumb, a block is considered as unidimensional when Dillon-Goldstein's ρ is larger than 0.7. This index is considered to be a better indicator than the Cronbach's alpha because it takes into account to which extent the latent variable explains the block of indicators.

Indicators well explained by their latent variables

We check it by means of three tools:

- Communality
- Composite reliability
- AVE

i) Communality

Communality is calculated with the purpose to check that indicators in a block are well explained by its latent variable. It is supposed that each indicator represents an error measurement of its construct. The relation: $x_{jk} = \lambda_{jk}\xi_j + \varepsilon_{jk}$, implies that the latent variable explains its indicator, so we have to evaluate how well indicators are explained by its latent variables. To do this, we examine the loadings which indicate the amount of variance shared between the construct and its indicators.

The communality for the k-th manifest variable of the j-th block is calculated as:

$$Com(\xi_j, x_{kj}) = \text{cor}^2(\xi_j, x_{kj}) = \lambda^2_{jk} \tag{5.74}$$

Communality measures how much of a given manifest variable's variance is reproducible from the latent variable. In other words, the part of variance between a construct and its indicators that is common to both. One expects to have more shared variance between LV and MVs than error variance, that is:

$$\lambda^2_{jk} > \text{var}(\varepsilon_{jk}) \tag{5.75}$$

with var(ε_{jk}) = 1 – λ^2_{jk}

Indicators with low communality are those for which the model is "not working" and the researcher may use this information to drop such variables from the analysis.

In addition, the analyst can calculate the *mean communality*, which is the average of all the block communalities.

ii) Composite reliability of Jöreskog

$$\rho_c = \frac{\left(\sum \lambda_{jk}\right)^2}{\left(\sum \lambda_{jk}\right)^2 + \sum \text{var}(\varepsilon_{jk})} \tag{5.76}$$

where λ_{jk} is the component loading of the *k*-th indicator in the *j*-th block, and $var(\varepsilon_{jk}) = 1 - \lambda^2_{jk}$

iii) Average Variance Extracted

Average Variance Extracted (AVE) of Fornell and Sha, is similar to Jöreskog's composite reliability, ρ_c, but AVE attempts to measure the amount of variance that an LV captures from its indicators in relation to the amount of variance due to measurement error.

$$\text{AVE} = \frac{\sum \lambda^2_{jk}}{\sum \lambda^2_{jk} + \sum \text{var}(\varepsilon_{jk})} \tag{5.77}$$

AVE should be larger than 0.50 which means that 50% or more variance of the indicators should be accounted for.

Differentiation between constructs

We evaluate the extent to which a given construct differentiates from the others. This is done by verifying that the shared variance between a construct and its indicators is larger than the shared variance with other constructs. In other words, no indicator should load higher on another construct than it does on the construct it intends to measure. We calculate the correlations between a construct and other indicator besides its own block. If an indicator loads higher with other constructs than the one it is intended to measure, we might consider its appropriateness because it is not clear which construct or constructs it is actually reflecting.

5.6.2 Measurement Model Validation: Formative Measures

Unlike reflective indicators, formative indicators are considered as causing (i.e. forming) a latent variable (see Figure 5.14). Formative indicators do not necessarily measure the same underlying construct. In this case, any change experienced by a construct does not imply a change in all its indicators; that is, formative indicators are not supposed to be correlated. For this reason, formative measures cannot be evaluated in the same way of reflective measures; and all the assessment criteria based on the loadings are discarded in the formative measures.

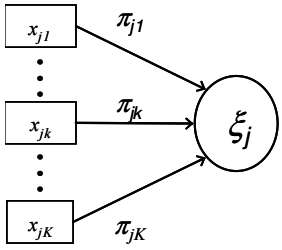


Figure 5.14. *Path diagram of formative indicators*

Considering the weights is equivalent to consider the outer weights. In this way we compare the weights of each indicator in order to determine which indicators contribute

most effectively to the construct. Attention must be paid in order to avoid misinterpreting relative small absolute values of weights as poor contributions. If we are considering the elimination of some indicator, this should be done based on multicollinearity: the elimination is recommended if high multicollinearity occurs. This implies a strong consensus among experts (based on theory) about how the latent variable is formed.

5.6.3 Structural Model Validation

The quality of the structural model is evaluated examining three indices or quality measures:

- the R^2's
- the redundancy index
- the Goodness-of-Fit (GoF)

Coefficients of determination R^2

The R^2 are the coefficients of determination of the endogenous latent variables. For each regression in the structural model we have an R^2 that is interpreted similarly as in any multiple regression analysis. R^2 indicates the amount of variance in the endogenous latent variable explained by its independent latent variables.

Redundancy

Redundancy measures the percent of the variance of indicators in an endogenous block that is predicted from the independent latent variables associated to the endogenous LV. Another definition of redundancy is the amount of variance in an endogenous construct explained by its independent latent variables. In other words, it reflects the ability of a set of independent latent variables to explain variation in the dependent latent variable. The redundancy index for the *k*-th manifest variable associated to the *j*-th block is:

$$Rd(\xi_j, x_{jk}) = \lambda_{jk}^2 \times R_{j|\xi_i}^2 \tag{5.78}$$

where:

- ξ_j is the *j*-th endogenous latent variable;
- x_{jk} is the *k*-th indicator associated to ξ_j;
- λ^2_{jk} is the communality;
- $R^2_{j|\xi_i}$ is the R^2 coefficient from the regression between ξ_j and its predictors ξ_i's.

High redundancy means high ability to predict. In particular, the researcher may be interested in how well the independent latent variables predict values of the indicators' endogenous construct. Analogous to the communality index, one can calculate the *mean redundancy*, that is, the average of the redundancy indices of the endogenous blocks.

Goodness-of-Fit

The Goodness-of-Fit (GoF) is a global criterion developed by Amato, Esposito Vinzi and Tenenhaus (Amato *et al*, 2004). The formula of GoF is given by:

$$\text{GoF} = \sqrt{\frac{\sum_{j=1}^{J}\left(\frac{1}{p_j}\sum_{k=1}^{p_j} cor^2(x_{kj},\xi_j)\right)}{J} \times \frac{\sum_{j^*=1}^{J^*} R^2\left(\xi_{j^*};\xi_j's \rightarrow predicting\,\xi_{j^*}\right)}{J^*}} \qquad (5.79)$$

where:
- J is the number of latent variables in the model;
- J^* is the number of endogenous latent variables; j^* indicates an endogenous block;
- $\text{cor}(x_{kj}, \xi_j)$ is the correlation between the k-th manifest variable of the j-th block and the corresponding latent variable;
- $R^2(\xi_{j^*};, \xi_j$'s $predicting\,\xi_{j^*})$ is the R^2 value of the regression between the j^*-th endogenous LV and its associated predictors ξ_j's.

The first term is the average communality of each block which measures the quality of the measurement model. The second term is the average of the determination coefficient for each endogenous construct according to latent variables which explain it. In other words: GoF^2 = (Average Communality) x (Average R^2). Hence, GoF is a compromise between the quality of the measurement model and the quality of the structural model.

5.6.4 Validation by resampling methods

Since PLS-PM does not rest on any distributional assumptions, significance levels for the parameter estimates (based on normal theory) are not suitable. Instead, resampling procedures such as blindfolding, jackknifing, and bootstrapping are used to obtain information about the variability of the parameter estimates. Actually, derivation of valid standard errors or *t*-values by means of bootstrapping is superior to the other two resampling methods (Temme *et al*, 2006)

Bootstrapping
It is a non-parametric approach for estimating the precision of the PLS parameter estimates. The bootstrap procedure is the following: M samples are created in order to obtain M estimates for each parameter in the PLS model. Each sample is obtained by sampling with replacement from the original data set, with sample size equal to the number of cases in the original data set (Chin, 1998)

5.7 Generalized Structured Component Analysis

Generalized Structured Component Analysis (GSCA) is an alternative method to PLS-PM recently proposed by Hwang and Takane (2004). As in PLS-PM, GSCA calculates latent variables as exact linear combinations of manifest variables. Hence, GSCA is regarded as a component-based approach to structural equation modeling. Unlike PLS-PM, however, GSCA provides a global optimization procedure for parameter estimation based on an alternating least squares algorithm. GSCA is thus equipped with an overall goodness of fit measure to evaluate how well the model fits to the data and also to compare it with alternative models. A brief introduction to GSCA is presented.

Data
In order to describe the GSCA method, we have tried to maintain the same notation used for PLS. The manifest variables are observed on N cases; arranged into J disjoint blocks $X_1, X_2, \ldots, X_j, \ldots, X_J$, each block $X_j = \{x_{j1}, \ldots, x_{jKj}\}$ of dimension (NxK_j) containing K_j indicators. We have a total number of P manifest variables, $P = K_1 + \ldots + K_J$. All variables are supposed to be centered and standardized (zero mean and unit variance).

Weight relations
PLS-PM and GSCA share the characteristic that the latent variables are expressed as linear combinations of their indicators.

$$\xi_j = \sum_k w_{jk} x_{jk} \tag{5.80}$$

where w_{jk} are the weights used to estimate the latent variable as a linear combination of its manifest variables. Using weight relations the problem of factor indeterminacy is also avoided in GSCA.

Structural Model
The associations among LVs can be represented by a linear multi-equation system. LVs can play both predictand and predictor roles. For the sake of simplicity, no distinctions in notation are made between endogenous and exogenous constructs. The linear equations take the form:

$$\xi_j = \sum_i \beta_{ji} \xi_i + \zeta_j \tag{5.81}$$

where the parameter β_{ji} is the path coefficient from the i-th LV to the j-th LV, ζ_j is the structural residual term, and the index i ranges over all predictors of ξ_j.

Measurement Models
Two measurement options are considered:

- Reflective indicators
- Formative indicators

i) Reflective indicators
In a measurement with reflective indicators the latent constructs are considered as the cause of the indicators (see figure 5.15)

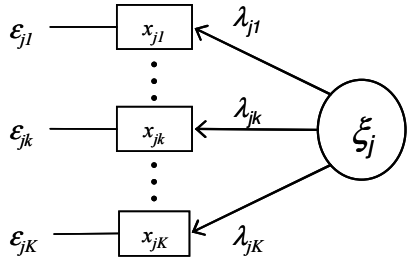


Figure 5.15. *Path diagram of reflective indicators*

In this sense the MVs can be considered reflects or manifestations of their LV. The MV x_{jk} is assumed to be a linear function of its LV ξ_j

$$x_{jk} = \lambda_{jk} \xi_j + \varepsilon_{jk} \tag{5.82}$$

where λ_{jk} is the loading coefficient and ε_{jk} is a residual term.

ii) Formative indicators

With formative indicators the latent constructs are considered as being caused by its manifest variables (see figure 5.16)

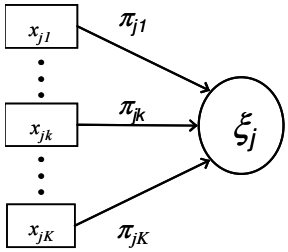


Figure 5.16. *Path diagram with formative indicators*

The construct ξ_j is assumed to be a linear function of its MVs x_{jk}

$$\xi_j = \sum_k \pi_{jk} x_{jk} \tag{5.83}$$

We know that latent variables are linear combinations of manifest variables. Thus, with formative indicators the component weights coincide with the weight coefficients, that is $\pi_{jk} = w_{jk}$. We have that

$$\xi_j = \sum_k \pi_{jk} x_{jk} = \sum_k w_{jk} x_{jk} \tag{5.84}$$

In GSCA the path diagrams shown in figure 5.17 are equivalent.

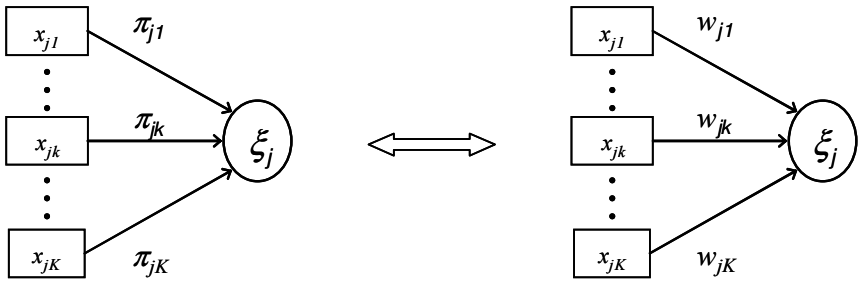


Figure 5.17. *Equivalence of path diagrams with formative indicators*

5.7.1 Algebraic formulation of the GSCA model

One of the fundamental aspects of a GSCA model is its representation in a unified algebraic formulation. Hwang and Takane achieve to express all the dependent variables in the model in terms of the latent variables. The "dependent" variables refer

to the reflective manifest variables and the endogenous latent variables. In other words, the dependent variables refer to those variables that have residual terms (ε_{jk} for the reflective MVs, and ζ_j for the endogenous LVs). The model is expressed as in matrix notation as:

$$XV = XUA + E \tag{5.85}$$

where:
- X is a matrix of dimension NxP of manifest variables
- V is a matrix of dimension PxT of component weights associated with the dependent variables (T is the total number of dependent variables)
- U is a matrix of dimension PxJ of component weights associated with the latent variables
- A is a super-matrix of dimension JxT comprised by two other matrices: C and B. C is a matrix of loadings relating latent variables to their reflective indicators. B is the matrix of path coefficients.
- E is a matrix of dimension NxT of residuals (for both the reflective indicators and the endogenous latent variables)

Let $\Omega = XV$, and $\Gamma = XU$. We can re-express the model $XV = XUA + E$ in a compact way as:

$$\Omega = \Gamma A + E \tag{5.86}$$

where:
- Ω is a matrix of dimension NxT of all dependent variables (reflective indicators and endogenous constructs)
- Γ is a matrix of dimension NxJ of latent variables

To illustrate the model in equation 5.87, we present two cases of a simple path model with only latent variables, ξ_1 and ξ_2. The first case is shown in figure 5.18, in which both latent variables have reflective indictors. The second case is displayed in figure 5.19, in which the exogenous construct ξ_1 has formative indicators.

Case 1
The first example consists in a simple path model with two latent variables: one exogenous, and one endogenous. Both constructs are measured with reflective indicators.

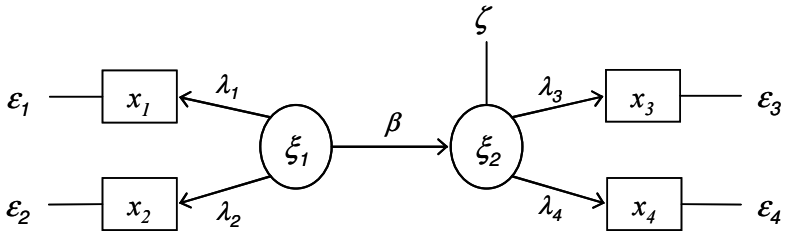


Figure 5.18. *Simple path model with reflective indicators*

Since all manifest variables in the measurement model are reflective indicators this implies that each indicator has associated to it a residual term. The equations for the reflective manifest variables are:

$$\begin{aligned} x_1 &= \lambda_1 \xi_1 + \varepsilon_1 \\ x_2 &= \lambda_2 \xi_1 + \varepsilon_2 \\ x_3 &= \lambda_3 \xi_2 + \varepsilon_3 \\ x_4 &= \lambda_4 \xi_2 + \varepsilon_4 \end{aligned} \tag{5.87}$$

With respect to the structural model, there is only one residual term since there is only one endogenous variable. The equation in the structural model is:

$$\xi_2 = \beta \xi_1 + \zeta \tag{5.88}$$

The definition of the GSCA model consists in expressing the equations of the dependent variables (i.e., the variables with residual terms) in a single algebraic formulation. Such formulation can be expressed in matrix form as:

$$\left[\begin{array}{cccc|c} x_1 & x_2 & x_3 & x_4 & \xi_2 \end{array}\right] = \left[\begin{array}{cc} \xi_1 & \xi_2 \end{array}\right] \left[\begin{array}{cccc|c} \lambda_1 & \lambda_2 & 0 & 0 & \beta \\ 0 & 0 & \lambda_3 & \lambda_4 & 0 \end{array}\right] + \left[\begin{array}{cccc|c} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \zeta \end{array}\right] \tag{5.89}$$

From the weight relations we know that $\xi_1 = w_1 x_1 + w_2 x_2$, and $\xi_2 = w_3 x_3 + w_4 x_4$. If we consider the weight relations and we take $X = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]$ we can express the model as follows:

$$X \left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & w_3 \\ 0 & 0 & 0 & 1 & w_4 \end{array}\right] = X \left[\begin{array}{cc} w_1 & 0 \\ w_2 & 0 \\ 0 & w_3 \\ 0 & w_4 \end{array}\right] \left[\begin{array}{cccc|c} \lambda_1 & \lambda_2 & 0 & 0 & \beta \\ 0 & 0 & \lambda_3 & \lambda_4 & 0 \end{array}\right] + \left[\begin{array}{cccc|c} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \zeta \end{array}\right] \tag{5.90}$$

Let

$$V = \left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & w_3 \\ 0 & 0 & 0 & 1 & w_4 \end{array}\right], \quad U = \left[\begin{array}{cc} w_1 & 0 \\ w_2 & 0 \\ 0 & w_3 \\ 0 & w_4 \end{array}\right], \quad A = \left[\begin{array}{cccc|c} \lambda_1 & \lambda_2 & 0 & 0 & \beta \\ 0 & 0 & \lambda_3 & \lambda_4 & 0 \end{array}\right], \text{ and}$$

$$E = \left[\begin{array}{cccc|c} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \zeta \end{array}\right] \tag{5.91}$$

It can be seen that A is formed by two matrices C and B

$$C = \left[\begin{array}{cccc} \lambda_1 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & \lambda_4 \end{array}\right] \qquad \text{and} \qquad B = \left[\begin{array}{c} \beta \\ 0 \end{array}\right] \tag{5.92}$$

Then, the model has the required structure $XV = XUA + E$ and can be expressed in the compact form $\Omega = \Gamma A + E$ of equation 5.87.

Case 2

The second example is similar to the first case. There are two latent constructs: one exogenous and one endogenous. Each construct is measured with two indicators. However, in this case the exogenous variables have formative indicators (see figure 5.19). The equations for the reflective manifest variables are:

$$\begin{aligned} \xi_1 &= w_1 x_1 + w_2 x_2 \\ x_3 &= \lambda_3 \xi_2 + \varepsilon_3 \\ x_4 &= \lambda_4 \xi_2 + \varepsilon_4 \end{aligned} \tag{5.93}$$

The equation in the structural model is

$$\xi_2 = \beta \xi_1 + \zeta \tag{5.94}$$

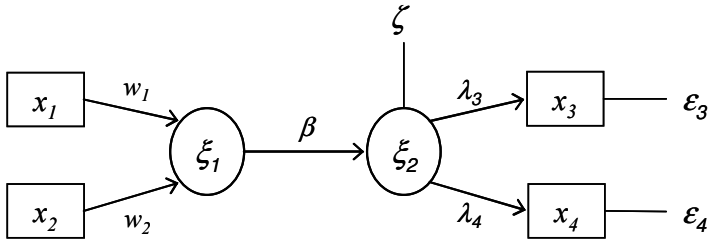


Figure 5.19. *Simple path model with formative and reflective indicators*

As in Case 1, the definition of the GSCA model for case 2 consists in expressing the equations of the dependent variables (i.e., the variables with residual terms) in a single algebraic formulation. Such formulation can be expressed in matrix form as:

$$\left[x_3 \quad x_4 \mid \xi_2\right] = \left[\xi_1 \quad \xi_2\right] \left[\begin{array}{cc|c} 0 & 0 & \beta \\ \lambda_3 & \lambda_2 & 0 \end{array}\right] + \left[\varepsilon_3 \quad \varepsilon_4 \mid \zeta\right] \tag{5.95}$$

From the weight relations we know that $\xi_1 = w_1 x_1 + w_2 x_2$, and $\xi 2 = w_3 x_3 + w_4 x_4$. If we consider the weight relations and we take $X = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]$ we can express the model in following form:

$$\mathrm{X} \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & w_3 \\ 0 & 1 & w_4 \end{array}\right] = \mathrm{X} \left[\begin{array}{cc} w_1 & 0 \\ w_2 & 0 \\ 0 & w_3 \\ 0 & w_4 \end{array}\right] \left[\begin{array}{cc|c} 0 & 0 & \beta \\ \lambda_3 & \lambda_2 & 0 \end{array}\right] + \left[\varepsilon_3 \quad \varepsilon_4 \mid \zeta\right] \tag{5.96}$$

Let

$$\mathrm{V} = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & w_3 \\ 0 & 1 & w_4 \end{array}\right], \; \mathrm{U} = \left[\begin{array}{cc} w_1 & 0 \\ w_2 & 0 \\ 0 & w_3 \\ 0 & w_4 \end{array}\right], \; \mathrm{A} = \left[\begin{array}{cc|c} 0 & 0 & \beta \\ \lambda_3 & \lambda_2 & 0 \end{array}\right], \; \text{and} \; \mathrm{E} = \left[\varepsilon_3 \quad \varepsilon_4 \mid \zeta\right] \tag{5.97}$$

It can be seen that A is formed by two matrices C and B

$$C = \begin{bmatrix} 0 & 0 \\ \lambda_3 & \lambda_4 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \beta \\ 0 \end{bmatrix} \tag{5.98}$$

Then, the model has the required structure $XV = XUA + E$ and can be expressed in the compact form $\Omega = \Gamma A + E$ of equation 5.87.

5.7.2 Parameter Estimation

We have seen that PLS lacks of a single criterion to be minimized or maximized for parameter estimation. Although some proposals such as the GoF index have been developed to provide an overall quality measure, there is no global optimization function to be solved. GSCA, instead, provides a global least squares optimization criterion, which is consistently minimized to estimate model parameters. The estimation of the unknown parameters V, U and A is done in such a way that the sum of squares of the residuals, $E = XV - XWA = \Omega - \Gamma A$, is as small as possible.
Let

$$\begin{aligned} f &= SS(XV - XWA) \\ &= SS(\Omega - \Gamma A) \end{aligned} \tag{5.99}$$

where $SS(Y) = \text{trace}(Y'Y)$.

The estimation of the parameters implies minimizing *f* with respect to V, U, and A. The elements in Ω and/or Γ are subject to normalization for identification purposes, for instance, $\xi'_1\xi_1 = 1$.

Since equation 5.101 cannot be solved in an analytic way, Hwang and Takane have developed an alternating least squares (ALS) algorithm (de Leeuw, Young, and Takane, 1976) to minimize *f*. The proposed algorithm consists in two main steps:

- In the first step, A is updated for fixed V and U.
- In the second step, V and U are updated for fixed A.

The two steps are repeated until convergence is reached, that is, until the decrease in the function value falls below a certain threshold value (for a detailed description of the ALS algorithm refer to Hwang and Takane, 2004; Hwang *et al*, 2007).

A few comments about the algorithm are necessary. Although the algorithm is convergent, it does not guarantee a global minimum. However, the problem of convergence to a non-global minimum may be avoided in two ways. One way is by choosing "good" initial values so that the function value is likely to start near the global minimum. The other way is to repeat the ALS algorithm with several random initial values. The obtained results are compared, and the solution associated with the smallest value is chosen.

5.7.3 Overall goodness of fit

GSCA is provided with a mechanism to evaluate the overall goodness of fit of the model. The overall measure of model fit is evaluated by the total variance of all the

dependent variables explained by the specified model predictions. This is given by the index:

$$\text{FIT} = 1 - \text{SS}(\Omega\text{-}\Gamma\text{A}) / \text{SS}(\Omega) \tag{5.100}$$

The FIT index ranges from 0 to 1. The larger the value of FIT, the more variance of the dependent variables is accounted by the model. It is a function of the sum of squared residuals that summarizes the discrepancies between the model and the data. However, FIT is affected by model complexity. Thus, in order to take into account this complexity, an adjusted index, called AFIT, is given by:

$$\text{AFIT} = 1 - (1 - \text{FIT}) \frac{d_0}{d_1} \tag{5.101}$$

where:
- $d_0 = NJ$ is the degrees of freedom for the null model (U=0 and A=0)
- $d_1 = NJ - Q$ is the degrees of freedom for the model being tested, where Q is the number of free parameters.

In addition to the FIT and the AFIT indexes, the evaluation of the model is done by checking individual parameter estimates. It is important to examine the loadings and the squared multiple correlations for individual manifest variables to assess the adequacy of the parameters. Other methods such as jackknife and bootstraps can be used to calculate standard errors of parameter estimates.

Remarks on GSCA

GSCA allows the incorporation of linear constraints into the model. These constraints are specified by reparametrization. It also enables to handle categorical variables and high-order constructs (i.e., latent variables formed by other latent variables). It is said that GSCA offer similar results to those obtained from PLS. However, as Hwang and Takane (2004) recognize, further studies are necessary for more careful comparisons between both techniques. In addition, further analyses are necessary to study robust estimation since GSCA may not be robust against outliers as far as it is based on solving a simple (unweighted) least squares criterion.

Chapter 6

PLS Path Modeling Segmentation

One of the most challenging issues within PLS-PM is related with tasks of segmentation and the detection of groups and classes of elements in the population. In last four years the importance of segmentation in PLS-PM has received a considerable amount of attention. As a result, various proposals have been developed in an attempt to deal with this promising branch of the PLS methodology. We have decided to begin this chapter by providing a brief introduction on the topic that gives rise to path modeling segmentation: market segmentation in customer satisfaction measurement. In the second section we present some considerations about path modeling segmentation and the issue of population heterogeneity. Then, the different segmentation approaches are described and discussed in the following sections.

6.1 Background for Segmentation in PLS-PM

Over the last ten years many of the successful applications of PLS-PM have been performed within marketing and management studies especially those related with measuring customer satisfaction. Examples can be found in Fornell (1992); Anderson and Fornell (2000); Anderson, Fornell, and Lehmann, (1994); Hackl and Westlund (2000); Martensen, Gronholdt and Kristensen (2000); Kristensen, Juhl and Ostergaard (2001); Westlund *et al* (2001); López, Fernández and Mariel (2003); and Vilares and Coelho (2003). It is impossible to deny that customer satisfaction measurement has played a key role on the applicability of PLS-PM. It has become the reference field of application when introducing and presenting the PLS approach of structural equation models in academic courses, conferences, seminars and papers.

The concept of customer satisfaction has been for years one of the central research topics in marketing and management studies. A simple search on "customer satisfaction" and/or "consumer satisfaction" is enough to find a number of articles in some of the most recognized marketing journals such as the Journal of Marketing, the Journal of Marketing Research, Marketing Science, the European Journal of Marketing, the International Journal of Research in Marketing, and Total Quality Management, among others.

Analysis and measurements on customer satisfaction are not a new topic. However, during the last decade customer satisfaction has received more attention than ever. The reasons are many but, as Helgesen (2000) mentions, some of them can be related to the increased interest concerning total quality management and national quality awards.

Closely related to the topic of customer satisfaction is the increasing area of customer relationship management, and the development and implementation of national customer satisfaction indices. Among the large number of available measures of customer satisfaction the first customer satisfaction index at a national level appeared in 1989 with the development of the Customer Satisfaction Barometer in Sweden (Fornell, 1992). The architect of this project was Professor Claes Fornell, Doctor of Economics by the University of Lund, Sweden, and current professor of Business Administration and the Director of the National Quality Research Center at the University of Michigan.

Professor Fornell also supervised the conduct of the preliminary analysis of the American Customer Satisfaction Index (ACSI) in 1993 (Fornell *et al* 1996). He introduced the idea of a Customer Satisfaction Index after exploring Sweden's use of a National Customer Satisfaction Barometer to measure customer satisfaction across industries (Fornell, 1992). This index, developed in 1994 by the National Quality Research Center at the Stephen M. Ross Business School (University of Michigan), is an adaptation of the Swedish Customer Satisfaction Barometer, with some revisions, modifications and improvements.

The European Union, inspired by the success of the Swedish and the American Customer Satisfaction Indices, was interested in the development and installation of a comparative system of national satisfaction indices. The result was the establishment of the European Customer Satisfaction Index (ECSI) founded by the European Organization for Quality (EOQ), the European Foundation for Quality Management (EFQM) and the European Academic Network for Customer Oriented Quality Analysis (IFCF), with the support of the European Commission and the collaboration of the CSI university network integrated by 8 European universities (Kristensen *et al*, 2001). The theoretical ECSI model constitutes a modified adaptation of the ACSI/Fornell's model, and considers the European economy as a whole, and thus, customer satisfaction indices can be compared with each other and with the European average (Grigoroudis and Siskos, 2004).

Today, Customer Satisfaction Marketing Studies can be considered a landmark for PLS-PM as well as an experimental field, and is becoming the main developmental arena for PLS contributions, proposals and innovations like those found in Cassel, Hackl and Westlund (2000); Stan and Saporta (2003); Eskildsen, Kristensen and Juhl (2005).

6.1.1 Market Segmentation in Customer Satisfaction Measurement

Closely related to Customer Satisfaction Marketing Research is the topic of Customer Market Segmentation. Market segmentation is the process of partitioning markets into groups of potential customers with similar needs or characteristics who are likely to exhibit similar purchase behavior (Wedel and Kamakura, 2000). In other words, segmentation can be considered as the process of dividing a set of elements into distinct subsets "the segments" that behave in the same way or have characteristics in common that make them have a similar conduct. For example, customers can be divided

according to their communication preferences or their lifestyle choices, and employees according to what stage they have reached in their careers.

Since its introduction by Smith (1956), market segmentation has become a fundamental concept in marketing theory and practice. Smith stated that "market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to the desires of consumers for more precise satisfaction of their varying wants".

The interest in market segmentation comes from the fact that the buyers of a product or a service are not a homogeneous group. Different customers/employees have different needs, and it rarely is it possible to satisfy all customers/employees by treating them alike. In these cases is important to consider other alternatives than the *one-size-fits-all* solution, and the need for segmentation becomes a challenging task.

6.2 Segmentation and Heterogeneity

The purpose of segmentation tasks, not only in PLS-PM but also in covariance-based SEM, is the same as in any other field: it is used to group individuals into segments with similar characteristics. Hence, segmentation procedures involve examining whether the population (or sample) is homogeneous or heterogeneous. However, one of the common assumptions when estimating structural equation models is to suppose homogeneity over the entire set of individuals. In other words, the analyst treats all individuals alike without considering any group structure. Moreover, it is taken for granted that a single model will adequately represent all the individuals, which implies that the same set of parameter values applies to all individuals (Muthén, 1989). This assumption, however, is unrealistic in many cases and it is reasonable to put into doubt sample homogeneity. Consider, for example, marketing and consumer behavior research. Potential sources of heterogeneity can be due to brand awareness, product class knowledge, product usage rate, customer preferences, desire for specifics and benefits (Dilon, 1990). In survey research studies heterogeneity can be expected among different subgroups defined by gender, groups of age, ethnicity, and marital status.

Heterogeneity is not only present in marketing studies but also in other social and behavioral sciences. In educational research, assuming homogeneity among a sample of students with varying instructional background is unrealistic. In natural sciences a sample may consist explicitly of groups such as experimental and control groups. Test results on a medical test may reflect two types of patients in the sample: those with a disease and those who are healthy.

The problem of not considering the possible existence of segments in the population is that conventional structural equation modeling techniques may lead the analyst to obtain inaccurate-inadequate results. Since the model for the entire population may be miss-specified, the analyst runs the risk of drawing erroneous or poor conclusions (Dilon, 1990; Yung, 1997). Thus, to overcome this situation it is necessary to assume population heterogeneity with groups having different behavior (Jedidi, Jagpal, and DeSarbo, 1997). In these cases, heterogeneity may imply that more than one single set of parameter estimates is needed in order to adequately characterize the phenomenon under analysis (Dilon and Kumar, 1994).

6.2.1 Types of heterogeneity

According to Lubke and Muthén (2005) the sources of population heterogeneity may be known beforehand but they can also be unknown. Thus, we can distinguish two types of heterogeneity:

- Observed heterogeneity
- Unobserved heterogeneity

Heterogeneity is observed if subpopulations can be defined based on an observed variable (i.e., the data can be divided into groups). Typical examples of these variables are demographic variables such as gender, groups of age, or income level. For instance, observed heterogeneity can be defined in terms of "marital status" (e.g., single, married, divorced, widowed). In contrast, Heterogeneity is unobserved when the variables that cause the heterogeneity in the data are not known beforehand. In this case, it is not possible to *a priori* divide the observations into groups. For this reason, subpopulations must be inferred from the data in the structural models.

Observed Heterogeneity
In the context of observed heterogeneity, the elements can be *a priori* divided in segments, and separate models can be estimated for each segment. It is presupposed that each element can be uniquely assigned to a single group based on one or more segmentation variables (i.e., the observed sources of heterogeneity). The goal is to compare the formed groups by comparing the separate models. In particular, the analyst may wish to compare models in terms of: (1) the assumed cause-effect relationships of the latent variables; (2) the model parameters (e.g., loadings, structural coefficients, or R^2 coefficients); (3) the adequacy of indicators in the measurement of constructs; and (4) the mean scores of the latent variables between models.

In order to perform these comparisons one may apply methods of multi-group analysis which are used for testing model equivalences and for investigating invariance hypotheses. For example, one analyst may be interested in evaluating whether latent means are equal across groups. Another analyst may be concerned about whether the magnitude of a structural relationship is the same in different groups, and she can examine the difference of path coefficients among the groups.

Unobserved Heterogeneity
In the context of unobserved heterogeneity, it is not possible to *a priori* divide the observations into groups. Although it is supposed that the data consists of different segments, it is not known beforehand which observation belongs to which of the segments. Then, the subpopulation membership of the observations has to be inferred from the data by applying some type of clustering-based procedure. The methods for unobserved heterogeneity include traditional clustering techniques, latent class analysis, and mixture models. These methods are designed to detect a given data set clusters of observations with similar response patterns on a set of variables.

6.2.2 Comparing path models

Before providing the different segmentation approaches within PLS-PM, it is important to discuss the issue in which path models can be compared. In fact, path modeling

segmentation implies that one is able to compare structural models (in some particular way or by some particular criterion) in order to establish their differences. The main question could be stated as follows: Given two structural models, how can they be compared? To answer this question, it is important to know the various ways in which path models may be different:

- They may differ at the causal network level: this refers to differences in the assumed network linking the latent variables. While two latent variables may be positively correlated for some consumers, they may also be orthogonal for another group of consumers.
- Differences at the structural level: this involves differences of magnitude in the structural coefficients. For instance, there may be two groups of customers: in one group satisfaction is driven by the perceived value of the product they consume, whereas satisfaction in the other group is driven by the perceived quality.
- Differences at the measurement level: this refers to the way in which the latent variables are defined by their indicators. While one indicator may be appropriate for some construct in one model, the same indicator may not be appropriate in other model.
- Models may differ at the latent variables level: this implies that latent means across models may be different. For example, the mean of Satisfaction may vary across groups in terms of marital status (e.g, single, married, divorced, etc)

Within PLS path modeling, the common standpoint among analysts for the purpose of segmentation is to assume that heterogeneity is exhibited in the structural parameters (i.e., path coefficients). In other words, there is an emphasis on comparing models by taking into account differences only at the structural level, despite of the type of heterogeneity. The rationale for focusing exclusively on the structural coefficients has to do with the modeling goals in path models with latent variables: researchers are interested in the complex system of cause-effect relationships among constructs. Thus, attention is paid to looking for differences in the magnitude of the cause-effect relationships among constructs.
This standpoint is not without its limitations. One limitation is that no differences at the measurement level are considered, and hence the set of indicators has to be the same across groups. However, it is important to recognize that there could be some research contexts in which this assumption may not hold (Baumgartner and Steenkamp, 1998). For example, while one indicator may be appropriate for some construct in a particular culture/country, the same indicator may not be appropriate in another culture/country. Another general limitation involves discarding possible differences in the causal network. In this cases the number of underlying constructs, and the causal relationships between them, have to be the same across groups. Finally, regarding the option of comparing models at the latent variables level (i.e., evaluating possible differences in the mean scores of the latent variables), this aspect is often relegated to subsequent analyses of a more exploratory/descriptive nature.

6.2.3 Segmentation approaches in PLS-PM

Based on the type of heterogeneity, we provide the following classification of the segmentation approaches in PLS-PM.

PLS-PM approaches in observed heterogeneity

The approaches in observed heterogeneity involve methods of multi-group analysis which are used for testing hypotheses and model equivalences in terms of path coefficients. In Henseler (2007) it is possible to find a classification of PLS-PM multi-group analysis in four approaches: 1) Re-sampling parametric approach (Chin, 2000); 2) Re-sampling non-parametric approach (Chin, 2003); 3) Moderation testing approach (Chin, Marcolin, and Newsted, 2003; Henseler and Fassot, 2005; Tenenhaus, Mauger, and Guinot, 2007); and 4) Henseler's approach (Henseler, 2007). There are, however, two additional methods that can be used to analyze observed heterogeneity: the PATHMOX (PLS Path Modeling Segmentation Tree) approach (Sánchez and Aluja, 2006, 2007), and the recently proposed Possibilistic PLS Path Modeling (Palumbo and Romano, 2008).

Usually, the number of groups in which the observations can be divided is known a priori. Indeed, most of the applications in the related literature are based on one or two nominal segmentation variables as observed sources of heterogeneity. An example would be groups defined in terms of gender (male, female) and education level (basic education, college degree). A group could be formed entirely of males or it could be formed entirely of males with a college degree. Another group could be formed by females with a college degree. If we cross all the categories among the two segmentation variables, the total number of groups would be 4 (females with basic education; females with college degree, males with basic education; males with college degree). Most of the segmentation approaches are designed to deal with these kinds of schemes in which the number of segments is known beforehand.

There is the case, however, in which the number of groups is not defined a priori. This situation may happen when we have several segmentation variables with many potential groups that can be formed according to the cross-classification of their categories. For such a scheme, we have developed the PATHMOX algorithm, which will be described in detail in the following chapter.

PLS-PM approaches in unobserved heterogeneity

Perhaps the most naive and simplest approach of unobserved heterogeneity consists of a sequential two-step procedure; One that combines cluster analysis in the first step with a multi-group analysis in the second step. Firstly, groups are formed by performing typical cluster analysis on the data (on the manifest variables or on the latent variables). Then, a multi-group analysis is performed on the separate models for each cluster. However, this approach has been criticized by many researchers (Jedidi *et al*, 1997; Görz *et al*, 2000; Lubke and Muthén, 2005) because the resulting groups may not produce differentiated path models, and because this approach does not take into account the hypothesized structural relationships among variables.

To overcome the shortcomings of the two-step procedure, a variety of proposals have been developed that have a model-based focus approach. Generally speaking, model-based techniques involve some clustering-based procedure that takes into consideration the cause-effect structure of models. Particularly in covariance-based SEM, the methods for dealing with unobserved heterogeneity are known as latent class models (Vermunt and Magidson, 2000). In a strict sense, latent classes refer to unobservable (latent) segments that which are represented by the distinct categories of a nominal latent variable. However, the important thing behind the notion of latent class is the ability to conceptualize unobserved heterogeneity.

These approaches include Finite Mixture Partial Least Squares (FIMIX-PLS) approach (Hahn, Johnson, Herrmann, and Huber, 2002; Ringle, Wende, and Will, 2005); PLS Typological Path Modeling (Squillacciotti, 2005; Squillacciotti, Trinchera, and Esposito Vinzi, 2006), and more recently the Response Based Procedure for detecting Unit Segments (REBUS) approach (Trinchera *et al*, 2007; Esposito Vinzi *et al*, 2007, 2008), and the PLS Genetic Algorithm Segmentation (PLS-GAS) approach (Ringle and Schlittgen, 2007)

The following figure shows the classification of the segmentation approaches taking into account the sources of population heterogeneity.

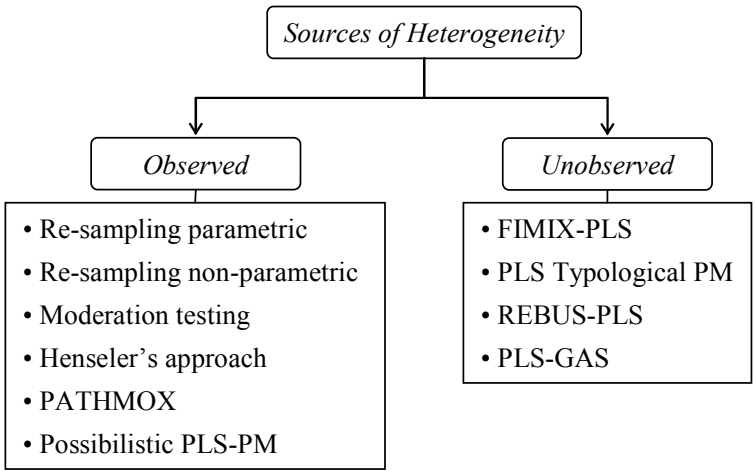


Figure 6.1. *Classification of segmentation approaches according to the sources of population heterogeneity*

If we consider whether the number of segments is known *a priori*, we can classify the segmentation approaches in a slightly different way as shown in figure 6.2.

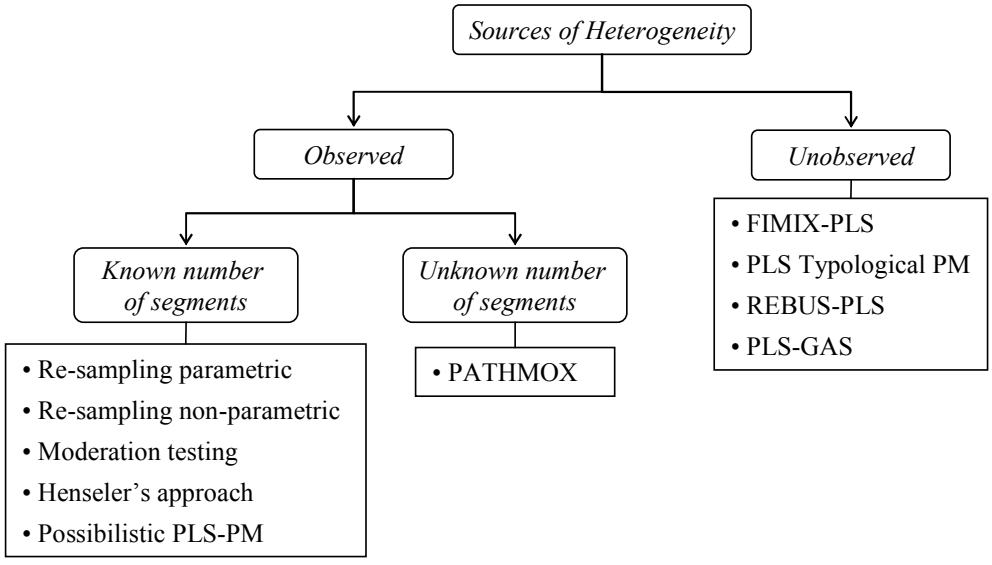


Figure 6.2. *Classification of segmentation approaches according to the a priori knowledge of the number of segments*

The difference between the classification displayed in figure 6.1 and the one displayed in figure 6.2 is due to the PATHMOX approach. As it has been mentioned, PATHMOX is specifically suited for cases with observed sources of heterogeneity; Cases in which the number of segments in the data is not defined beforehand. In such cases, there are several categorical variables but the analyst cannot be certain about the number of segments in the population.

6.3 Segmentation approaches with observed heterogeneity

The case of segmentation with observed sources of heterogeneity has to do with typical applications in cross-national/cross-cultural studies and also in the so-called multi-group analysis or subgroup analysis. It is assumed that the same path model is estimated for each group, and the idea is to compare models by means of examining the difference of path coefficients among groups. The way in which differences are assessed is what differentiates one approach from others. We can distinguish six types of segmentation approaches with observed sources of heterogeneity:

- Re-sampling parametric approach
- Re-sampling non-parametric approach
- Moderation testing approach
- Henseler's approach
- PATHMOX approach
- Possibilistic PLS-PM approach

6.3.1 Re-sampling parametric approach

This approach is suggested by Chin (2000) as a parametric approach for testing the statistic of the difference in path coefficients between segments. The procedure consists of separating the data into groups (i.e. segments) and of running bootstrap re-samplings for each group. Path coefficients are calculated in each re-sampling and the standard error estimates are treated in a parametric sense via *t*-tests.

For instance, in the case of having two groups we need to calculate the pooled estimator *Sp* for the variance, which is:

$$Sp = \sqrt{\frac{(m-1)^2}{(m+n-2)} SE^2_{sample\_1} + \frac{(n-1)^2}{(m+n-2)} SE^2_{sample\_2}} \tag{6.1}$$

where *SE* denotes the standard error, and *m* and *n* are the sample sizes of group 1 and 2, respectively. Then we need to take the difference between the paths for the two samples and divide it by the product $Sp \times (1/m + 1/n)^{1/2}$, that is:

$$t = \frac{Path_{sample\_1} - Path_{sample\_2}}{\left[\sqrt{\frac{1}{m} + \frac{1}{n}}\right]\left[\sqrt{\frac{(m-1)^2}{(m+n-2)} SE^2_{sample\_1} + \frac{(n-1)^2}{(m+n-2)} SE^2_{sample\_2}}\right]} \tag{6.2}$$

This would follow a *t*-distribution with $m+n-2$ degrees of freedom.

It is important to mention that this *t*-test relies on the assumption that the underlying weights in the formation of constructs for each group are approximately equivalent. Chin also remarks that this approach works reasonably well if the two samples are not too non-normal and/or the two variances are not too different from one another. A series of articles about moderation testing with categorical moderator variables with PLS applications in Information Technology are given below:

- Gefen and Straub (1997) studied the influences of gender on Information Technology diffusion and the Technology Acceptance Model (TAM).
- Keil *et al* (2000) use this approach to test the influence of national culture on risk taking and the willingness to continue a project.
- Venkatesh and Morris (2000) investigated gender differences in the context of individual adoption and sustained usage of technology in the workplace.
- Ahuja and Thatcher (2005) study the influence of gender on the relationship between perceptions of the work environment and trying to innovate with IT.

In the case of marketing applications there are some interesting studies:

- Eberl (2005) conducts a multi-group analysis in Corporate-Level Marketing to test differences between stakeholder groups with a model for corporate marketing activities affecting corporate reputation and customer loyalty.
- Choe *et al* (2007) analyze moderating effects on online consumer behavior.

6.3.2 Re-sampling non-parametric approach

The re-sampling non-parametric or permutation approach is suggested by Chin (2003) as an alternative distribution free approach to the re-sampling parametrical one. The procedure applies an approximate randomization (i.e. permutation) test to assess the differences in path coefficients across segments. The basic premise of this kind of tests is to use the assumption that it is possible that all of the groups are equivalent, and that every member of the group is the same before sampling began. From this, one can calculate a statistic and then observe the extent to which this statistic is special by seeing how likely it would be if the group assignments had been jumbled/mixed.

To illustrate the basic idea of a permutation test, suppose we have two groups G_1 and G_2 whose path coefficients are B_{G1} and B_{G2}, and that we want to test whether they come from the same population, at 5% significance level. Let n_1 and n_2 be the sample size corresponding to each group. The permutation test is designed to determine whether the observed difference between the path coefficients is large enough to reject the null hypothesis H_0 that the two groups can be considered identical. The test proceeds as follows:

1. First, a test statistic is calculated for the data. In our case the test statistic is the difference of path coefficients between the two groups.
2. Then the observations of groups G_1 and G_2 are combined into a single large group.
3. Next, the data are permuted (divided or rearranged) repeatedly in a manner consistent with the random assignment procedure. Each permutation implies:
   - dividing data in two groups of size n_1 and n_2

- estimating the PLS models for each group
- calculating and recording the test statistic

The set of these calculated differences is the distribution of possible differences under the null hypothesis that group label does not matter.

4. Then, we sort the recorded differences and we check if the original test statistic is contained within the middle 95% of the sorted values. If it does not, we reject the null hypothesis of identical groups at the 5% significance level.

The permutation approach is a distribution free test that requires no parametric assumptions concerning statistical distributions.

6.3.3 Moderation testing approach

The comparison of models in this case is performed by taking the segmentation variables as moderator variables. It is possible to find different definitions of moderator variables but we have decided to mention the most accepted in the reviewed literature:

- According to Baron and Kenny (1986) a moderator variable is a variable that "affects the direction and/or strength of the relationship between an independent or predictor variable (X) and a dependent or criterion variable (Y)"
- James and Brett (1984) define a moderator variable Z "when the relationship between two (or more) other variables, say X and Y, is a function of the level of Z".
- Cortina (1993) says that "moderation occurs when the effect of one variable, X, on another, Y, depends on the level of some third variable Z".

Depending on the type of moderator variable we can find categorical variables and continuous variables. In both cases, the moderation testing approach is based on the study of interaction terms which are especially well suited for continuous metric variables. An interest reference for moderating effects is found in Henseler and Fassot (2005). Tenenhaus *et al* (2006) shows an application of this approach.

The procedure to measure interaction effects is proposed by Chin, Marcolin and Newsted (1996, 2003) using the product-indicator approach of Kenny and Judd (1984). In their work, Chin *et al* (2003) use a model relating a response block η to a predictor block ξ with a main effect A and an interaction term $\xi * A$ added to the model as shown in figure 6.3.

In this case not only are the predictor ξ and the dependent η variables latent constructs, but they are also the moderator variable A and the interaction term $\xi * A$. The indicators for the interaction latent variable are created by multiplying the indicators from the predictor and the moderator variables. Note that product indicators are developed creating all possible products from the two sets of indicators. In the figure above, there are three manifest variables reflecting the predictor ξ and two manifest variables for the moderator variable A, there would be six indicators reflecting the interaction term.

The process for testing interaction effects consists of comparing the results of the model with the interaction construct versus the results of the model without the interaction construct. For the analysis with the interaction term, it is necessary to include the two main effects terms to assess how the moderator construct influences the

impact of ξ on η. The path coefficient from ξ to η is interpreted as the amount of influence of ξ on η when the moderator term is equal to zero. In a similar way, the path coefficient of moderator variable A to η is interpreted as the amount of direct influence of A on η when ξ is equal to zero. The interaction construct represents how a change in the level of the moderator term A would change the influence of ξ on η.

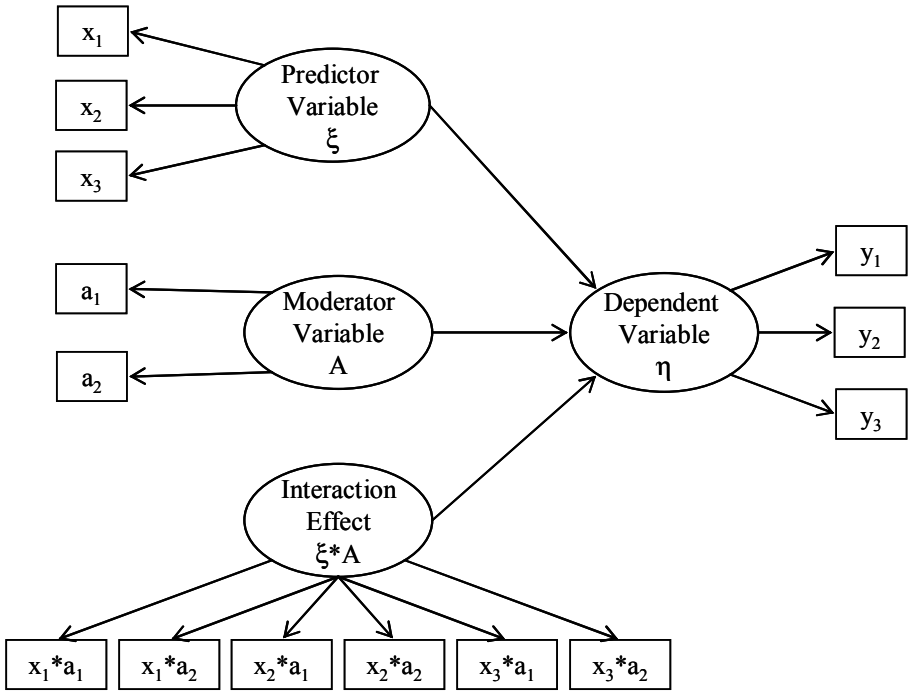


Figure 6.3. *Path Diagram illustrating an interaction effect*

Chin *et al* (2003) suggest assessing the interaction effect by comparing the squared multiple correlation (R^2) of the model with the interaction construct versus the R^2 of the model without the interaction term. The difference between multiple correlation coefficients is used to evaluate the global effect size f^2 for the interaction construct calculated as:

$$f^2 = \frac{R^2\,\text{int} - R^2 no\,\text{int}}{1 - R^2 no\,\text{int}} \tag{6.3}$$

where R^2int is the R^2 of the model with the interaction construct while R^2*no*int is the R^2 of the model without the interaction construct. Reference values of f^2 are 0.02 for small size effect, 0.15 for moderate size effect, and 0.35 for large size effect; however the authors acknowledge that small values of f^2 do not imply an unimportant effect.

6.3.4 Henseler's Approach

A more recent proposal was presented by Henseler (2007). This approach is based on bootstrapping and it consists of obtaining the empirical cumulative distribution of the parameters under consideration. Bootstrap samples are taken in order to obtain the empirical cumulative distribution function for the path coefficients.

To illustrate this proposal suppose we have two groups G_1 and G_2 whose path coefficients are B_{G1} and B_{G2}. The required steps are the following:

1. Take J bootstrap samples for each group.
2. For each bootstrap sample calculate the J estimates for the path coefficients
3. Then, build all the possible combinations (J^2) of the bootstrap parameters across groups (i.e. all possible comparisons of bootstrap parameters have to be made).
4. From all the combinations, count the number of times in which the path coefficient of one group exceeds the path coefficient of the other group.

The idea is that the relative frequency of these counts reflects the error probability, that is, the probability that the path coefficient obtained for group one is larger than the obtained for group two:

$$P(\beta_{G1} > \beta_{G2}) = \frac{1}{J^2}\sum_{i=1}^{J}\sum_{j=1}^{J}\Theta(b_{1j} - b_{2i}) \tag{6.4}$$

where Θ is the unit step function, which has a value of one if its argument exceeds zero, and has a value of zero otherwise.

This method does not require distributional assumptions. Nevertheless, increasing the number of groups directly increases the number of bootstrap samples combinations that have to be drawn. To calculate the probability of differences in subpopulation parameters, this method uses the empirical cumulative distribution functions provided by bootstrap re-sampling.

6.3.5 PATHMOX

In most cases, path modeling data comes from surveys or researches that contain more information (i.e. observed heterogeneity) than the one that is used for the path models' establishment. We refer to this information as external information because it is outside of the model. For instance, in many marketing studies, like those of consumer satisfaction, it is common to collect socio-demographic variables and psycho-demographic variables such as age, gender, social-status, or consume habits that are not part of the path model but that can be extremely useful for segmentation purposes.

Thus, in order to incorporate the available external variables, also known as segmentation variables, the PATHMOX algorithm has been developed. The idea behind PATHMOX is to build a tree having a structure similar to a binary decision tree with specific path models for different segments in each of its nodes. The detailed description of PATHMOX is presented in the next chapter (see Chapter 7).

6.3.6 Possibilistic PLS-PM

To the best of our knowledge, Possibilistic PLS path modeling is the most recent proposal within PLS-PM segmentation approaches. It has been proposed by Palumbo and Romano (2008) as an approach to multi-group comparison based on fuzzy regression. The idea is to estimate path models for each group using Fuzzy Possibilistic Regression to obtain fuzzy (interval) estimated path coefficients. Then, the comparison

between models is done by comparing distances of the fuzzy path coefficients. The entire procedure consists in three steps:

- Step 1: Estimate local fuzzy structural models for each group
- Step 2: Compare the estimated models on the basis of their fuzzy path coefficients.
- Step 3: Display distances for fuzzy/interval data by hierarchical trees or pyramids

The first step combines Fuzzy Possibilistic Regression (Tanaka *et al*, 1982) with PLS-PM. The estimation of fuzzy path coefficients uses a two stage estimation procedure:
Stage 1: Latent variables are estimated according to the PLS path model
Stage 2: Fuzzy possibilistic regression on the estimated latent variables is performed to obtain fuzzy intervals for the path coefficients.

In the second step, the comparison of the estimated models is done by evaluating the distances of the fuzzy path coefficients between models. For this purpose, a Euclidean metric is adopted.
The third step involves the visualization of distances using a fuzzy classification algorithm for interval data. In particular, the employed algorithm is the HIPYR hierarchical classification algorithm (Brito, 2000) which uses a Hausdorff metric. The advantage of using the HIPYR algorithm is that it can build both hierarchical and pyramidal classification trees. Hierarchical trees lead disjunctive clusters whereas Pyramidal trees lead to fuzzy clusters.

6.4 Segmentation Approaches with Unobserved Heterogeneity

Sometimes it is not possible to form sub-groups or segments in an a priori way due to the lack of sources of observed heterogeneity. Sometimes, even when disposing of categorical variables to form sub-groups, the analyst may be interested in looking for solutions under a framework of unobserved heterogeneity. This second alternative may be based on the researcher's knowledge or belief that the observed sources of heterogeneity cannot provide a feasible or adequate segmentation. Moreover, the researcher may be interested in exploring unexpected configurations of groups.

There are three approaches of segmentation in unobserved heterogeneity:

- FIMIX-PLS
- PLS Typological PM
- REBUS-PLS
- PLS-GAS

6.4.1 FIMIX-PLS

The finite mixture approach supposes that population is a mixture of two or more subpopulations (a finite number of subpopulations); each is characterized by a distribution, and mixed in different proportions. The finite mixture modeling approach, also known as latent class, was first developed within the covariance-based analysis approach of structural equation modeling. It was proposed independently by Arminger

and Stein (1997), Dolan and Var der Maas (1997), and Jedidi, Jagpal and DeSarbo (1997).

Within the PLS approach, the extension of Finite Mixture to the PLS approach was developed in 2002 by Hahn, Johnson, Herrmann and Huber (2002). Ringle, Wende and Will (2005) have continued the research of finite mixture PLS with its implementation in the SmartPLS software and its application in marketing studies (Ringle, 2006).

Model Formulation

An inner model from a path model with latent variables can be formulated according to the following conditions.

Let

η: the matrix of endogenous variables in the inner model

ξ: the matrix of exogenous variables in the inner model

The inner relations can be expressed by:

$$\eta = B^s\eta + \Gamma^s\xi + \zeta \tag{6.5}$$

where B^s is the matrix of path coefficients for the endogenous variables, Γ^s is the matrix of path coefficients for the exogenous variables, and ζ is the matrix of residuals. Rearranging equation X.1:

$$(I - B^s)\eta - \Gamma^s\xi = \zeta \tag{6.6}$$

Let $B=(I-B^s)$ and $\Gamma = -\Gamma^s$, we get:

$$B\eta + \Gamma\xi = \zeta \tag{6.7}$$

Finite mixture distribution

Now, suppose we have G classes or segments. For each individual i (i=1,…,N) in the structural model for the g-th class we have:

$$B_g\,\eta_i + \Gamma_g\xi_i = \zeta_{ig} \tag{6.8}$$

The main point of the finite mixture approach is the assumption that the data is a mixture of a finite number of subpopulations each characterized by a distribution. Thus, we assume that η_i is distributed as a finite mixture of conditional multivariate normal densities, $f_{i|g}$ (.):

$$\eta_i \sim \sum_{g=1}^{G} \rho_g f_{i|g}\left(\eta_g \mid \xi_i, \mathrm{B}_g, \Gamma_g, \Psi_g\right) \tag{6.9}$$

where Ψ_g is the covariance matrix for g-th class, and ρ_g are the mixing proportions, of the finite mixture such that $\rho_g > 0$ and $\sum_{g=1}^{G} \rho_g = 1$.

Replacing the conditional multivariate normal densities $f_{i|g}$ (.) in (6.9):

$$\eta_i \sim \sum_{g=1}^{G} \rho_g \left[\frac{\left|\mathrm{B}_g\right|}{(2\pi)^{Q/2}\left|\Psi_g\right|^{1/2}} \exp\left\{ -\frac{1}{2}\left(\mathrm{B}_g\eta_i + \Gamma_g\xi_i\right)' \Psi_g^{-1}\left(\mathrm{B}_g\eta_i + \Gamma_g\xi_i\right)\right\}\right] \tag{6.10}$$

Suppose the η_i vectors are independent, the likelihood function for the N vectors is given by:

$$L = \prod_{i=1}^{N}\left(\sum_{g=1}^{G}\rho_g\left[\frac{|\,\mathrm{B}_g\,|}{(2\pi)^{Q/2}\,|\,\Psi_g\,|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathrm{B}_g\eta_i+\Gamma_g\xi_i\right)'\Psi_g^{-1}\left(\mathrm{B}_g\eta_i+\Gamma_g\xi_i\right)\right\}\right]\right) \qquad (6.11)$$

The mixing proportions ρ_g can be obtained as prior probabilities of any individual belonging to the G classes.

Estimation via the EM Algorithm

The likelihood of the model previously developed (see eq. 6.11) can be maximized by means of the EM algorithm (Dempster, Laird and Rubin, 1977; McLahlan and Krishnan, 1997). The algorithm involves an expectation step (E-step) and a maximization step (M-step).

For an EM algorithm formulation, it is necessary to include an indicator variable z_{ig} that indicates the membership of the i-th individual to the g-th class,

$$z_{ig} = \begin{cases} 1 & \text{if observation } i \text{ belongs to class } g \\ 0 & 0 \text{ otherwise} \end{cases}$$

Note that for every individual its associated vector $z_i = (z_{i1},\ldots,z_{iG})$' is non-observed. For a particular individual i, the non-observed vector z_i is independently and identically multinomially distributed with probabilities ρ_g. The joint likelihood of η_i and z_i is

$$L_i(\eta_i, z_i; \xi_i, \mathrm{B}_g, \Gamma_g, \Psi_g, \rho_g) = \prod_{g=1}^{G}\left[\rho_g f\left(\eta_i \mid \xi_i, \mathrm{B}_g, \Gamma_g, \Psi_g\right)\right]^{z_{ig}} \qquad (6.12)$$

The likelihood for all individuals is

$$L = \prod_{i=1}^{N}\prod_{g=1}^{G}\left[\rho_g f\left(\eta_i \mid \xi_i, \mathrm{B}_g, \Gamma_g, \Psi_g\right)\right]^{z_{ig}} \qquad (6.13)$$

$$L = \prod_{i=1}^{N}\left(\prod_{g=1}^{G}\rho_g\left[\frac{|\,\mathrm{B}_g\,|}{(2\pi)^{Q/2}\,|\,\Psi_g\,|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathrm{B}_g\eta_i+\Gamma_g\xi_i\right)'\Psi_g^{-1}\left(\mathrm{B}_g\eta_i+\Gamma_g\xi_i\right)\right\}\right]\right)^{z_{ig}} \qquad (6.14)$$

and the log-likelihood is

$$\ln(L) = \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\,\ln\left(f(\eta_i \mid \xi_i, \mathrm{B}_g, \Gamma_g, \Psi_g)\right) + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\,\ln(\rho_g) \qquad (6.15)$$

Because vectors z_i are non-observed, the matrix $Z=(z_1,\ldots,z_N)$ is considered as missing data.

E-Step

The EM algorithm begins with an E-step, where the expectation of $\ln(L)$ is computed over the conditional distribution of the non-observed data Z given the predicted values of η_i and ξ_i and the provisional estimates (B^*, Γ^*, Ψ^*, and ρ^*) of the parameters B, Γ, Ψ, and ρ respectively. The important thing to note is that the elements z_{ig} of matrix

$Z=(z_1,\ldots,z_N)$ are considered missing data. In order to start the E-step, it is required to have classes already formed. The initialization of the classes is done either by: (1) randomly assigning individuals into the G classes; or by (2) assigning individuals according to prior knowledge or analyst's assumptions about the groups.

Once individuals have been assigned to one class, the mixing proportions ρ_g can be calculated as a priori probabilities of each individual belonging to the g-th class. Also, provisional estimates (B^*, Γ^*, Ψ^*, and ρ^*) can be calculated.

The expectation of the likelihood is

$$E[\ln(L)] = \sum_{i=1}^{N}\sum_{g=1}^{G} E\left(z_{ig};\xi_i,\rho^*,\mathrm{B}^*,\Gamma^*,\Psi^* \mid \eta_i\right)\ln\left(f\left(\eta_i \mid \xi_i,\rho_g^*,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*\right)\right) + \sum_{i=1}^{N}\sum_{g=1}^{G} E\left(\xi_i,\rho^*,\mathrm{B}^*,\Gamma^*,\Psi^* \mid \eta_i\right)\ln(\rho_g^*) \tag{6.16}$$

The conditional expectation of z_{ig} can be obtained as

$$E\left[z_{ig};\xi_i,\rho^*,\mathrm{B}^*,\Gamma^*,\Psi^* \mid \eta_i\right] = \frac{\rho_g^* f(\eta_i \mid \xi_i,\rho_g^*,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*)}{\sum_{g=1}^{G}\rho_g^* f(\eta_i \mid \xi_i,\rho_g^*,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*)} \tag{6.17}$$

The mixing proportions ρ_g can be obtained as prior probabilities of any individual belonging to the G classes. The posterior probability of membership for individual i in class g, P^*_{ig}, can be calculated using Bayes' theorem, conditional on the estimates of the class-specific parameters ρ^*_g, B^*_g, Γ^*_g, Ψ^*_g via:

$$P_{ig}^* = \frac{\rho_g^* f_{i|g}\left(\eta_i \mid \xi_i,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*\right)}{\sum_{g=1}^{G}\rho_g^* f_{i|g}\left(\eta_i \mid \xi_i,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*\right)} \tag{6.18}$$

If we compare (6.18) with (6.17) we see that

$$P_{ig}^* = E\left(z_{ig};\xi_i,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^* \mid \eta_i\right)$$

The non-observed data in matrix Z are replaced by the posterior probabilities calculated on the base of provisional estimates. Thus, equation 6.16 becomes

$$E[\ln(L)] = \sum_{i=1}^{N}\sum_{g=1}^{G} P_{ig}^* \ln\left(f\left(\eta_i \mid \xi_i,\rho_g^*,\mathrm{B}_g^*,\Gamma_g^*,\Psi_g^*\right)\right) + \sum_{i=1}^{N}\sum_{g=1}^{G} P_{ig}^* \ln(\rho_g^*) \tag{6.19}$$

M-Step

In the M-step equation 6.15 is maximized with respect to the parameters subject to the restriction $\rho_g > 0$ and $\sum_{g=1}^{G}\rho_g = 1$, conditional on the new provisional estimates of z_{ig}, in order to obtain improved parameter estimates.

Initially, new mixing proportions ρ_g are calculated as the average of adjusted expected values that are obtained in the previous E-step. Then, new parameters for B, Γ and Ψ are estimated by different OLS regressions, one for each regression in the inner model. Recall that we have an OLS regression for each endogenous variable, so the M-

step involves the computation OLS regressions for each endogenous variable using the Maximum Likelihood estimators of the coefficient and the variance -which are assumed to be identical to OLS predictions.

With the M-step we obtain new mixing proportions ρ_g together with the results from each independent regression which serve as new provisional estimates for the next E-step iteration to improve the outcomes for P_{ig}. The E-step and the M-step are applied successively until no further improvement in ln(L) is possible according to some previously specified convergence criterion. However, convergence is only guaranteed at to at least a local optimum solution, hence different starting values of the parameters must be tried to determine the potential occurrence of local optimum.

The most important results in FIMIX are the probability of membership P_{ig}, the mixing proportions ρ_g, class-specific estimates B_g and Γ_g of the inner model and Ψ_g for the regression variances. A final consideration with FIMIX-PLS consists of the identification of an appropriate number of segments. The problem is that there is no satisfactory solution. So, in order to have a better idea of the right number of classes, FIMIX-PLS algorithm must be repeatedly executed with consecutive numbers of segments G (e.g., from 2 to 10). Because the algorithm converges for any given number of segments, the usual criterion for deciding on an adequate number of segments is segment size. The argument is that, at some certain point, an additional segment will have a small size, explaining a marginal proportion of heterogeneity in the overall set of data.

6.4.2 PLS Typological Path Modeling

A second approach named PLS Typological Path Modeling (TPM) has been developed by Squillacciotti (2005) and Squillacciotti, Trinchera and Esposito (2006) as an extension of PLS Typological Regression (Esposito Vinzi & Lauro, 2003). Unlike FIMIX-PLS, TPM seeks to optimize the predictive capacity of the segments' models without requiring distributional assumptions about the latent or manifest variables. Instead, TPM consists of assigning units to segments based on a unit-model distance.

This procedure starts with calculating the global model; according to its results, population segments are defined. Local models are estimated for each segment and a measure of the distance between each individual and each local model is computed. Individuals are then re-assigned to the segment corresponding to the closest local model. An iterative algorithm re-estimates local models until no change in the composition of the segments is observed.

The first problem with this method is defining a measure to calculate the distance of an individual from the model. To do so, Squillacciotti extends the distance index $DModX.N_g$ used in PLS Typological Regression, which is based on a special distance used in PLS Regression: the *DModX.N* index (Tenenhaus and Esposito, 2005; Bastien *et al*, 2005; Tenenhaus, 1998). Although we will not talk about PLS Regression, we can say that *DModX.N* shown in (6.20) is a normalized distance of each observation, x_i, from its estimate, $\hat{x}_i$, yielded by the model.

$$DModX.N = \sqrt{\frac{d^2(x_i, \hat{x}_i)}{\frac{1}{n}\sum_{i=1}^{n} d^2(x_i, \hat{x}_i)}} \tag{6.20}$$

The distance proposed in PLS Typological Path Modeling requires the identification of a highlighted endogenous latent variable called "target" latent variable. The target endogenous LV is supposed to be the construct at the end of the causal flow and it is assumed to be the most important construct for the predictive purposes of the model.

A simple example of a target endogenous latent variable is shown in the following figure (exogenous indicators are represented only by empty squares).

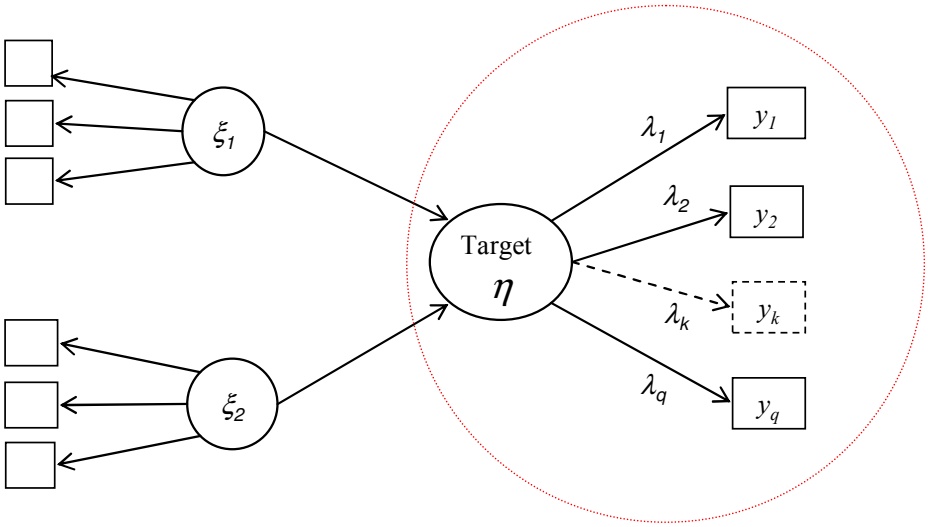


Figure 6.4. *Example of a target endogenous latent variable in a path diagram*

The target block containing the target construct, η, is enclosed in a big circle. As we can see, the target endogenous latent variable is the endogenous construct that is found at the end of the causal network of a path model. Briefly stated, the target latent variable is the construct, among all the endogenous constructs, that does not predict any other latent variable.

The proposed distance index, D_{ig}, is a measure for the i-th individual to the g-th class; its expression is

$$D_{ig} = \sqrt{\frac{\frac{\sum_{k=1}^{q}\left[e_{gik}^2 / Rd(\eta, y_k)\right]}{(q - t_g)}}{\frac{\sum_{i=1}^{N}\sum_{k=1}^{q}\left[e_{gik}^2 / Rd(\eta, y_k)\right]}{(n_g - t_g - 1)(q - t_g)}}} \qquad (6.21)$$

where:
- y_k, are the indicators of the target endogenous construct, k=1,...,q (it is supposed that there are q indicators measuring the final endogenous LV);
- e^2_{gik} is the square of the i-th residual on the k-th indicator for the g-th group, from the regression of the k-th indicator over the LVs explaining the target endogenous LV;
- $Rd(\eta,y_k)$ is the redundancy index (see eq 5.79, chapter 5) for the k-th manifest variable associated to the target block in the g-th group;
- i ranges from 1 to N; k ranges from 1 to q; g ranges from 1 to G.
- n_g is the number of units in group g;
- t_g is the number of exogenous LVs in the local model for group g;

- q is the number of manifest variables associated to the target variable (the number of variables in the target block);

We know from chapter 5 that Redundancy is calculated for each manifest variable in an endogenous block. It measures the portion of the variance for a manifest variable that is explained by the predictor latent variables. Redundancy represents the extent to which the manifest variables of the target construct are explained by its explanatory latent variables. Thus, the ratio $e^2_{gik} / Rd(\xi,x_k)$ reflects the part of the i-th observation that is not explained by the redundancy of the k-th indicator for the g-th model.

The weighting terms attempt to take into account relevant issues for each model: (q - t_g) considers the difference between the number of final endogenous manifest variables and the number of explanatory endogenous constructs; (n_g - t_g -1) is related to the number of observations being included in each group; and the double summation term can be considered as the explanatory power provided by the g-th model.

Algorithm Description

The algorithm begins with the estimation of the global model over all the individuals. Then, a hierarchical clustering over the communality residuals e^2_{Gik} is computed for the global model in order to determine the number of groups and the initial unit assigned to them. In the iterative loops of the algorithm, individuals are assigned to the group corresponding with the closest local model, according to the index D_{ig} defined in 6.21. Stability of results in terms of groups' compositions is considered as a stopping criterion.

Another issue of concern within typological path modeling the definition of the segments and its number because they have to be determined by the user usually in one of two ways: they can be provided *a priori* or determined by performing a cluster analysis of the results in the PLS global model. An additional drawback is directly related to the chosen target endogenous latent variable because the whole process is focused on this construct which may not be explained as well as other latent variables. In addition, only the target block is taken into account instead of the entire structural model. That is, focusing on only one construct, which may be poorly explained, involves disregarding the whole structural relations of the model.

6.4.3 REBUS-PLS

Following the research line established in Typological Path Modeling and aiming to improve its limitations, Trinchera *et al* (2007) and Esposito Vinzi *et al* (2007, 2008) have presented the Response Based Units Segmentation (REBUS) approach for PLS-PM. REBUS is developed with the purpose to overcome the implied need in TPM to identify a target latent variable by making use of the entire inner model. In addition, REBUS also takes into account the measurement model.

In REBUS the proposed pseudo distance or *closeness measure* is composed of two elements: one element to assess the quality of the measurement model, the other element to assess the quality of the structural model. The combination of the two elements in a single measure seeks to assigns individuals to the group whose model fit is the best. Broadly speaking, REBUS, like TPM, is designed to identify local models that have a better fit than the global model. Unlike TPM, REBUS has the added feature of taking into account both the inner model and the outer model.

The basis of the closeness measure used in REBUS concerns the Goodness of Fit index "GoF" (see eq 5.80, chapter 5).

$$\text{GoF} = \sqrt{\frac{\sum_{j=1}^{J}\left(\frac{1}{p_j}\sum_{k=1}^{p_j} cor^2(x_{kj},\xi_j)\right)}{J} \times \frac{\sum_{j^*=1}^{J^*} R^2\left(\xi_{j^*};\xi_j's \rightarrow predicting\,\xi_{j^*}\right)}{J^*}} \qquad (6.22)$$

where:
- J is the number of latent variables in the model;
- J^* is the number of endogenous latent variables; j^* indicates and endogenous block;
- cor(x_{kj},ξ_j) is the correlation between the k-th manifest variable of the j-th block and the corresponding latent variable;
- $R^2(\xi_{j^*};,\xi_j$'s $predicting\,\xi_{j^*})$ is the R^2 value of the regression between the j^*-th endogenous LV and its associated predictors ξ_j's.

We know that GoF^2 = (Average Communality) x (Average R^2). Hence, the GoF index is a compromise between the quality of the measurement model and the quality of the structural model.

The idea behind REBUS is to use the communality and the R^2 measures as in the GoF index to form a pseudo-distance called CM_{ig} that represents the closeness of an individual i to the class g:

$$CM_{ig} = \sqrt{\frac{\sum_{j=1}^{J}\sum_{k=1}^{p_j}\left[e_{ikjg}^2 / Com(\xi_{jg},x_{kj})\right]}{\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{p_j}\left[e_{ikjg}^2 / Com(\xi_{jg},x_{kj})\right]} \times \frac{\sum_{j^*=1}^{J^*}\left[f_{ij^*g}^2 / R^2\left(\xi_{j^*};\xi_j's \rightarrow predicting\,\xi_{j^*}\right)\right]}{\sum_{i=1}^{N}\sum_{j^*=1}^{J^*}\left[f_{ij^*g}^2 / R^2\left(\xi_{j^*};\xi_j's \rightarrow predicting\,\xi_{j^*}\right)\right]}} \qquad (6.23)$$

where:
- $Com(\xi_{jg},x_{kj})$ is the communality index for the k-th manifest variable of the j-th block in the g-th group;
- e_{ikjg} is the measurement model residual for the i-th individual in the g-th group, corresponding to the k-th manifest variable in the j-th block (the communality residuals);
- f_{ij^*g} is the structural model residual for the i-th individual in the g-th group, corresponding to the j^*-th endogenous block;
- n_g is the number of individuals belonging to the g-th group;
- m_g is the number of extracted components. Since all blocks are assumed to be reflective, this figure will always be equal to 1.

The measurement residual e_{ikjg} (communality residual) is obtained for each individual i belonging to the g-th group as the difference between the observed value of the k-th manifest variable in the j-th block, x_{ikjg}, and its corresponding estimated value (obtained by regressing x_{kj} on the j-th latent variable computed for the g-th group, ξ_{jg})

$$e_{ikjg} = x_{ikjg} - \hat{x}_{ikjg} \tag{6.24}$$

with $\hat{x}_{ikjg} = \lambda_{kjg}\hat{\xi}_{ijg}$ where λ_{kjg} is the loading coefficient associated to the *k*-th variable of the *j*-th block in the *g*-th group, and $\hat{\xi}_{ijg}$ is the *j*-th latent construct obtained as

$$\hat{\xi}_{ijg} = \sum_{k=1}^{p_j} w_{kjg} x_{ikj} \tag{6.25}$$

where w_{kjg} is the outer weight associated to the *k*-th manifest variable of the *j*-th block in the *g*-th group.

The structural residual f_{ij*g} is obtained for each individual *i* in the *g*-th group for the *j**-th endogenous construct, as the difference between the endogenous latent variable and its corresponding estimated value

$$f_{ij*g} = \xi_{ij*g} - y_{ij*g} \tag{6.26}$$

with $y_{ij*g} = \sum_{j=1}^{J's \to j*} \beta_{jj*g}\xi_{ijg}$ where β_{jj*g} is the path coefficient associated to the *j*-th explanatory latent variable of the *j**-th endogenous construct in the *g*-th group.

REBUS Algorithm steps:

- Step 1: Estimation of the global PLS Path Model;
- Step 2: Computation of the communality and structural residuals of all individuals from the global model, according to equations (6.24) and (6.26);
- Step 3: Perform a hierarchical classification on the residuals computed in step 2;
- Step 4: Choice the number of classes (G) according to the dendrogram obtained in step 3;
- Step 5: Assignment of the individuals to each group according to the cluster analysis results;
- Step 6: Estimation of the G local models
- Step 7: Computation of the closeness measure CM for each individual with respect to each local model
- Step 8: Assignment of each individual to the closest local model;

  If stability of group membership *Then* step 9 *else* go back to step 6;
- Step 9: Description of the obtained groups according to differences among the local models.

REBUS algorithm begins with estimating the global model (over all the individuals). Then, a hierarchical classification of the communality residuals e_{ikjg} and structural residuals f_{ij*g} is calculated for the global model; In order to determine the number of groups and the initial unit assignment to them. In the iterative loops of the algorithm, individuals are assigned to the group corresponding with the closest local model,

according to the closeness measure CM_{ig}. The stopping criterion takes into account the stability of results in terms of groups' compositions.

On one hand, it is supposed that if two models show identical structural coefficients but different outer weights, REBUS will be able to detect this difference. On the other, the identified local models will exhibit higher values in the communalities and in the R^2 coefficients.

6.4.4 PLS-GAS

The PLS Genetic Algorithm Segmentation (PLS-GAS) approach has been recently proposed by Ringle and Schlittgen (2007) as a new contribution for PLS path modeling segmentation. PLS-GAS is another PLS-PM based clustering procedure that is based on a genetic algorithm.

Similar to FIMIX-PLS, PLS-TPM and REBUS-PLS, the PLS-GAS approach is also based on the assignment of data to a predefined number of groups or segments, by performing several assignment trials for different numbers of groups in order to find the best solution, i.e. the best fitting number of segments according to some criteria. However, the characteristic feature of PLS-GAS is that it uses a genetic algorithm to search for the best number of segments.
PLS-GAS follows a two-stage procedure:

- In the first stage, a non-deterministic genetic algorithm is used to find the best possible optimum starting partition by strolling through the search space and thereby coming close to many local optima.
- The second stage involves the best partition that was found for a deterministic hill-climbing approach to improve (if possible) the local partition for the final best segmentation.

Hence, Ringle and Schlittgen state that PLS-GAS is a genetic/hill-climbing, clustering hybrid algorithm.

Because the appropriate number of segments is unknown *a priori*, in each PLS-GAS run a fixed number of segments is used. Consequently, the PLS-GAS procedure must be tested with consecutive numbers (e.g. 2 to 10) of pre-specified clusters. In each execution, cases are assigned to a pre-determined number of clusters. The PLS path model estimates for each formed cluster are evaluated to assess the quality of the segmentation. The best segmentation with the most appropriate number of groups, is obtained after examining the fitting segmentation results. However, since it is limited by only one available reference, there is no information about the model comparison process that is used to select the best number of segments.

Chapter 7

PATHMOX Approach: Path Modeling Segmentation Trees

PATHMOX emerges as an approach on segmentation tasks in PLS-PM, inspired by the segmentation scheme used in decision trees. It is motivated by the opportunities provided in survey research, and by the need of researchers and practitioners to have automated (or semi-automated) methods to analyze their data. PATHMOX adapts the basic idea behind binary segmentation processes in order to produce a segmentation tree with different path models in each of the obtained nodes. Unlike decision trees, there is no prediction purpose in PATHMOX but rather has the goal of identification to detect different path models. This chapter begins with the motivation behind PATHMOX and the practical needs it seeks to meet. The second section contains a brief review of some of the segmentation tree concepts that we consider the most helpful to clearly understand the underlying "philosophy" in path modeling segmentation trees. Finally, the PATHMOX algorithm and its description are presented in detail.

7.1 Motivation of PATHMOX

Much of the work on structural equations modeling depends on survey research studies and survey-based data. These types of studies involve gathering information using different data collection methods such as mail questionnaires, personal interviews, telephone interviews, online surveys, and mobile-based surveys. Among all the available survey data collection methods today, most professional research organizations (e.g, academic, governmental, and commercial) employ some form of computerized data collection (De Leeuw and Nichols, 1996). Moreover, there is an increased interest in applying online (i.e., surveys through the internet) data collection techniques (Ilieva, Baron, and Healy, 2002). These trends and advances in data collection practices benefit survey research studies and provide challenges and opportunities for path modeling analysis. As the quantity of data available in surveys and questionnaires increases, there is a need to use all the available information. In addition, academics and practitioners are attempting to use automated (or semi-

automated) methods to analyze their data while taking into account as much information as possible. The main motivation behind path modeling segmentation trees is to use “additional” information included in survey-based studies to perform segmentation tasks. We argue that the majority of path modeling analyses are carried out with survey-based studies, which contain external information that can be taken into account in the process of segments identification.

It is well known that survey questionnaires not only contain the manifest variables that integrate the path models, but also other additional observable variables that do not form part of the models. For instance, much of the data for customer satisfaction and other similar analysis (e.g. employee satisfaction) include well-known observable variables such as socio-demographic, geographic, psychographic and behavioral variables. Examples of socio-demographic variables may be age, gender, family size, occupation, education, and social class. In the case of geographic variables we may observe regions by continent, country, state, neighborhood, size of metropolitan area, climate, etc. Psychographic factors might be lifestyle, values or personality. Behavioral variables refer to usage rate, seeking profits, readiness to buy, and user status. These types of variables are rarely included in path models. However, if we consider them as segmentation variables, they can be used to detect different path models of segments in the population.

We propose to take the segmentation variables as sources of observed heterogeneity in population. In order to meet some of the needs of the PLS-PM analysts and practitioners, PATHMOX is proposed as a segmentation approach that attempts to deal with the background mentioned above. Inspired by the segmentation scheme used in binary decision trees, the idea behind PATHMOX is to build a tree having a structure similar to a binary decision tree with different population segments in each of its nodes. However, PATHMOX cannot be considered a true decision (or classification) tree because there are no predefined classes. There is no prediction purpose in PATHMOX but rather a goal to identify and detect different path models.

7.2 Basics of segmentation trees

Segmentation trees, also known as decision trees, are one the most popular and intuitive data mining tools (Tufféry, 2007). These methods are used in classification tasks to solve problems of discrimination and regression by detecting a set of rules allowing the analyst to assign the elements in data into a set of predefined groups or segments. The first approaches were proposed by Sonquist and Morgan (1964) and Sonquist, Baker and Morgan (1971) with their methodology called Automatic Interaction Detection (AID). A different scheme was developed by Kass with the CHAID algorithm (Kass, 1980). New impulse to segmentation tasks was given by the work of Breiman, Friedman, Olshen and Stone (1984) with the CART (Classification And Regression Tree) algorithm, together with the work of Ross Quinlan and his ID3 algorithm (Quinlan, 1986). Successors of the ID3 algorithm have been developed with the C4.5 algorithm (Quinlan, 1993) and the C5 algorithm (Quinlan, 1998).

A segmentation tree is a special form of tree structure. A tree structure is a way of representing the hierarchical nature of a structure in a graphical form. It is called a "tree structure" because its graph has the appearance of a tree (although the tree is generally displayed upside down). Decision trees have three types of nodes: root, internal, and leaf nodes.

- There is exactly one **root node** that contains the entire set of elements in the sample; the root node has no incoming branches, and it may have zero, two, or more outgoing branches.
- **Internal nodes**, each of which is hoped to contain as many elements as possible belonging to one class; each internal node has one incoming branch and two (or more) outgoing branches.
- **Leaf or terminal nodes** which, unlike internal nodes, have no outgoing branches.

The root node can be regarded as the starting node. The branches are the lines connecting the nodes. Every node that is preceded by a higher level node is called a "child" node. A node that gives origin to two (or more) child nodes is called a parent node. Terminal nodes are nodes without children. The most common way of visually representing a tree is by means of the classical node-link diagram that connects nodes together with line segments (see figure 7.1).

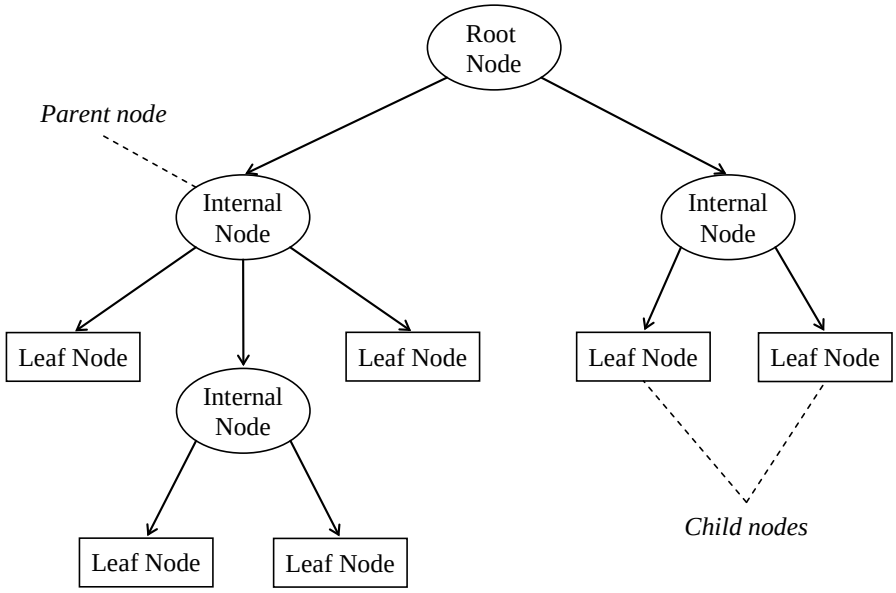


Figure 7.1. *Graphical representation of a segmentation tree structure with its different types of nodes*

The top-down approach, also referred to as top-down induction, is the most commonly used method for building a decision tree (Martin, 1997; Murthy, 1998). Top-down heuristics start from the entire data, somehow partitioning it into subsets, and recursively applying the partitioning procedure. The initial state of a decision tree is the root node which contains all the objects to be classified. It is assumed that data contains one dependent variable indicating to which class an object belongs to. The rest of the variables are the predictor variables, also known as "segmentation variables". If it is the case that all objects in the root node belong to the same class, then no further partitions need to be made to split the data, and the solution is complete. If objects at this node belong to two or more classes, then a test is made at the node that will result in a split. The split implies selecting a segmentation variable which provides the best division of the objects into different subsets. The divisions refer to the nodes, each of which is hoped to contain as most elements as possible belonging to one class. The process is

recursively repeated for each of the new internal nodes until the subsets cannot be partitioned any further (Apté and Weiss, 1997).

7.2.1 Binary Segmentation Trees

One special kind of segmentation trees are binary segmentation trees. They are called "binary" because the nodes are only partitioned in two (see figure 7.2). The usual process to build an ordinary binary decision tree is as follows. First, a search of all the segmentation variables is done in order to find the single variable which best splits the data in two groups. In other words, one must find the segmentation variable that split the data in two subsets so that they are as different as possible with respect to the response variable. This process is applied separately to each subgroup, and so on recursively until the subgroups either reach a minimum size or until no improvement can be made. In each dichotomy the two parts are the most contrasting according to the response variable.

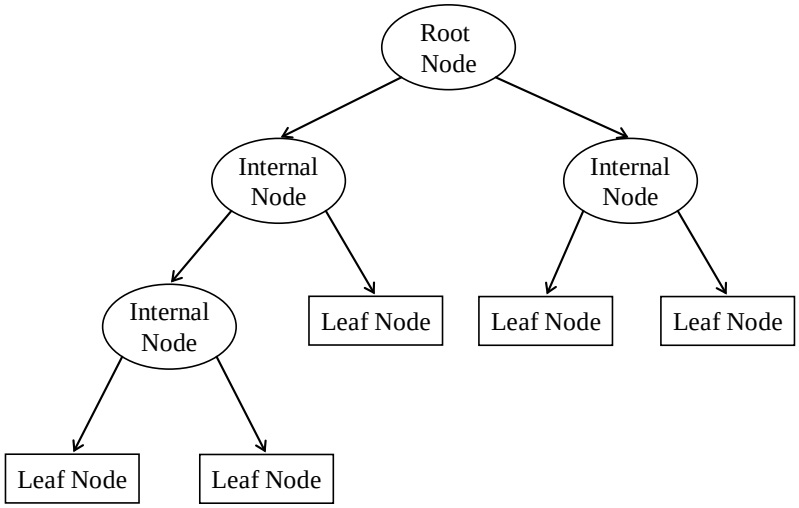


Figure 7.2. *Example of a binary segmentation tree and its different types of nodes*

The tree building process requires three main components (Breiman *et al*, 1984):

- Establish the set of admissible divisions for each node
- Define a criterion to find the "best" node split
- Define a rule to declare a given node as internal node or leaf node

Establishing the set of admissible splits for each node is related with the nature of the segmentation variables in the data. The number and type of possible binary splits will vary depending on the scale (e.g., binary, nominal, ordinal, and continuous) of the segmentation variables. Further discussion is given in the next section.

The second aspect to build a segmentation tree is the definition of a node splitting rule. This rule is necessary to find at each internal node a test for splitting the data into subgroups. Since finding a split involves finding the segmentation variable that is the most discriminating, the splitting test helps to rank variables according to their discriminating power.

The last component of the tree building process has to do with the definition of a node termination rule. In essence, there are two different ways to deal with this issue: 1)

pre-pruning and 2) post-pruning. Pre-pruning implies that the decision of when to stop the growth of a tree is made prospectively. Post-pruning refers to reducing the size of a fully expanded tree by pruning some of its branches retrospectively.

Pre-pruning methods establish stopping rules for preventing the growth of those branches that do not seem to improve the predictive accuracy of the tree. Indeed, the pre-pruning criteria are needed to avoid the construction of large segmentation trees with uninteresting and insubstantial nodes. Three stop rules are the most employed restrictions: a minimum number of elements in a node, a threshold based on statistical significance, and the number of final subgroups in the tree. Unlike pre-pruning methods, in post-pruning methods the trees are grown even when it seems worthless and are then retrospectively pruned of those branches that seem superfluous with respect to predictive accuracy.

Of the previous components, we will only focus on the establishment of a set of admissible divisions for one node. An entire presentation of the variety of binary decision trees building process is beyond the scope of this work. The discussion of different criteria to choose the best splits, and the description of rules for declaring some node as internal or terminal, can be found in Safavian and Landgrebe (1991), Esposito *et al* (1997), Martin (1997), Murthy (1998), Nakache and Confais (2003), and Tan, Steinbach and Kumar (2006).

7.2.2 Binary partitions of segmentation variables

In binary segmentation trees, the set of admissible splits for each node depends on the nature of the segmentation variables in the data. Segmentation variables can be of different type. In practice, most segmentation variables are of categorical nature although sometimes continuous variables may be present. We can divide them in four classes:

- Binary variables
- Ordinal variables
- Nominal variables
- Continuous variables

The number of possible binary splits for each type of segmentation variable is described below.

Binary variables

Because binary variables have two values, they only generate one binary partition. For example, the variable “Gender” with two values, female and male, can only de divided in one possible way.

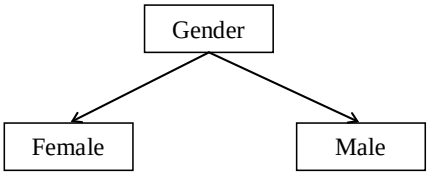


Figure 7.3. *Example of a binary partition for a binary variable*

Nominal Variables

A nominal variable may have many categories and the number of binary splits depends on the number of distinct categories for the corresponding variable. The total number of possible binary partitions for a nominal variable with *k* categories is 2^{k-1}-1. For example, suppose a variable "Color" with three categories: Green, White and Red. The number of binary splits is 2^{3-1}-1 = 3.

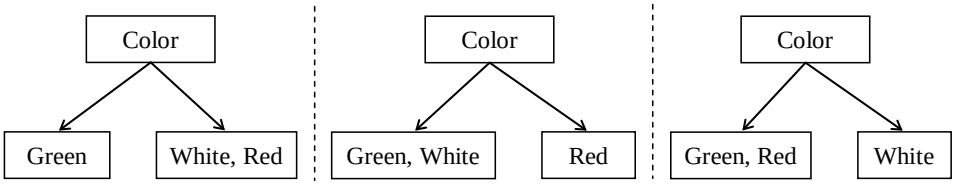


Figure 7.4. *Example of binary partitions of a nominal variable with 3 categories*

Ordinal variables

Binary splits for ordinal variables must respect the order property of the categories, that is, the grouping of categories must preserve the order among the values. The total number of binary partitions for an ordinal variable with *k* categories is *k*-1. For example, the "apparel size" may have four categories: small (S), medium (M), large (L) and extra-large (XL). In this case, the number of binary splits will be 4-1=3

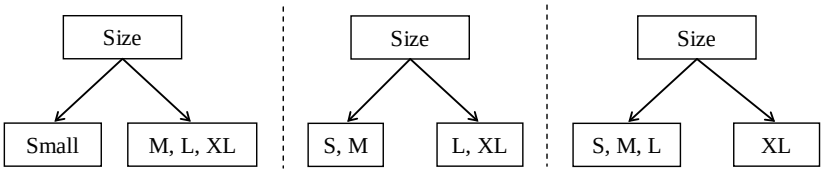


Figure 7.5. *Three binary partitions of an ordinal variable with 4 categories*

Continuous variables

The treatment for continuous variables depends on the number of different values the variable takes. Suppose a continuous variable with *k* different values. If we consider that *k* is "adequate" in some sense, we can treat the continuous variable like an ordinal variable. If we consider *k* to be large we may group its values and reduced their number in *k** (*k**<*k*). In both cases, the variable is treated as an ordinal variable and the number of binary splits will be *k*-1 or *k**-1. For example, imagine the continuous variable "Age" (measured in years) with five values 5, 7, 8, 9 and 10. The number of binary splits is 5 - 1 = 4

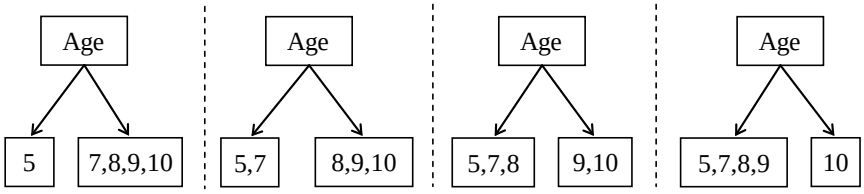


Figure 7.6. *Binary partitions of a continuous variable with 5 values*

The following table summarizes the number of two-way partitions for the different types of segmentation variables.

Table 7.1. *Number of binary splits according to the type of variable*

Type of Variable	*Binary splits*
Binary	1
Nominal	2^{k-1}-1
Ordinal	k-1
Continuous	k-1

7.3 Path Modeling Segmentation Trees

Based on the segmentation approach used in decision trees, we propose the concept of Path Modeling Segmentation Trees. For this purpose, we have developed the PATHMOX algorithm with the idea of building segmentation trees that have a similar tree structure to binary decision trees. The main characteristic of a path modeling segmentation tree, and the reason for its name, is that each node has an associated path model. The aim is to select among a set of segmentation variables those having a superior discriminant capacity in the sense of separating path models as much as possible according to a predefined criterion. We seek to identify and detect path models for different subgroups in the population.

7.3.1 PATHMOX Algorithm

The general purpose of PATHMOX is to construct a segmentation tree by means of successively dividing the set of elements in a binary way. This idea of partitioning is the reason for the name PATHMOX which is composed by two terms: PATH and MOX. PATH refers to path modeling while MOX refers to *Moxexeloa* which is a Náhuatl (the Aztec language) word that means “divide in two groups”.

The resulting tree has the aspect of a binary decision tree. However, each of the nodes in the path modeling segmentation tree is associated to a path model for some subgroup in the population. PATHMOX cannot be considered a true decision (or classification) tree because there is no intention of prediction, rather there is the goal to identify and detect different path models. We assume the existence of segmentation variables in data which are taken as sources of observed heterogeneity. Although this type of heterogeneity implies the definition of subpopulations based on observed variables, there are no predefined classes to be predicted. Hence PATHMOX is not prediction-oriented but rather exploratory-oriented.

The segmentation process seeks to find a first segmentation variable having a particular binary split whose associated segments give place to path models that are as different as possible. The procedure is applied to each of the generated segments while looking for a second segmentation variable, and so on. Like any decision tree, a segmentation tree of path models has three types of nodes:

- A root node that contains the path model for the entire population (e.g. the global model). The root node has no incoming branches and it can have zero or two outgoing branches.

- Internal nodes, each of which contains one local path model; each internal node has one incoming edge and two outgoing branches.
- Leaf or terminal nodes, each of which contains one local path model; unlike internal nodes leaf nodes have no outgoing branches.

Like any tree building process, a segmentation tree of path models requires three main components:

- Establish the set of admissible divisions for each node
- Define a criterion to choose the "best" node split
- Define a rule to declare a given node as internal node or leaf node

As it has been seen in section 7.2.2 the set of admissible splits for each node depends on the scale of the segmentation variables. With respect to the definition of a node splitting rule, we propose a test based on an *F*-statistic for splitting the data into subgroups. This test helps to rank the segmentation variables according to their discriminating power. The split criterion and the node termination rules are discussed below.

How should the individuals be split?
In order to split a path model associated to a given node, the tree-growing process selects the best partition from the available segmentation variables. The selection implies investigating all possible binary splits for each available segmentation variable. The best partition, according to an *F*-test (Fisher, 1935), of each segmentation variable is considered as a candidate for best split. Among the candidates, the partition with the lowest *p*-value is chosen as the best split.

How should the splitting procedure stop?
Since a stopping condition is needed to finish the tree-growing process we have adopted a pre-pruning approach. The main reason to adopt pre-pruning methods is because the produced trees will be of moderate complexity. Moreover, with a pre-pruning approach we avoid creating "large" trees with impractical and uninteresting nodes. The selected stopping conditions are a significance threshold for the *p*-value, and the number of elements inside a node. By establishing a *p*-value threshold we avoid obtaining partitions with a low discriminant capacity. By asking for a minimum number of elements inside a node we avoid small size nodes, which might be impractical and uninteresting.

7.3.2 Algorithm Description

The algorithm starts with the estimation of the path model for the entire population which is considered as the global model contained in the root node. If we take the root node as a parent node, the next steps are carried out according to the following process:

- For each segmentation variable we calculate all the possible binary partitions
- For each partition we split the set of latent variables in two segments
- We calculate the inner model for each segment and we obtain their corresponding sets of path coefficients

- We compare the inner models by performing the test of equality for path coefficients. This involves calculating the *F* statistic and its associated *p*-value.
- Once we have computed all the *F* statistics for all the binary splits for all the segmentation variables, we select the best split as the one with the lowest *p*-value.
- If the *p*-value does not exceed some pre-established significance threshold, we estimate new path models for the elements in the obtained child nodes. This means that the obtained child nodes from the parent node under analysis will contain their own path model.

One node is labeled as a leaf node when:

- The analyzed parent node has no more segmenting categories from which binary splits could be performed. In other words, the segmentation variables for that parent node have been exhausted and no more splits are feasible.
- The selected best split has an associated *p*-value exceeding the significance threshold.
- The number of elements in one node is less than or equal to some predefined minimum number of elements.

The algorithm stops when:

- There are no leafs with significant *p*-value (i.e., uninteresting nodes).
- There are no remaining nodes to segment, that is, when all remaining nodes are leaf nodes.
- The tree has grown up to some predefined depth level (optional parameter).

The algorithm has three parameters that have to be defined by the analyst:

- The significance *p*-value threshold. This parameter prevents selecting “best” splits which are not significant.
- The minimum node size (the minimum number of elements for a node). This parameter avoids fragmentation of small size nodes.
- The grow tree depth level (optional parameter). This stops the segmentation process which might derivate in an intractable number of segments.

PATHMOX Algorithm

Step 1: Start with the global PLS path model at the root node
Step 2: Root node division (Node partition routine*****)
Step 3: ***If*** (Stop-condition = false) ***then***
For each new child node != leaf node
Node partition routine*
Else
Stop algorithm

(Stop-condition: All child nodes are leaf nodes)
(Stop-condition: Optional depth level)

*****Node partition routine:
Step 1: ***If*** (stop_criterion1 = false) ***then***
For each segmentation variable of the node
Calculate all possible binary splits
Test the equality of path coefficients of inner models (*F*-test)
Select the best split (lowest *p*-value)
Else
Label node as leaf node
Step 2: ***If*** (stop_criterion2 = false) ***then***
Re-estimate PLS path models in the child nodes
Else
Label node as leaf node

(Stop-criterion1: Number of elements inside a node)
(Stop-criterion2: *p*-value larger than significance level threshold)

Figure 7.7. *Pseudo code of the PATHMOX algorithm*

7.3.3 Models Comparison in PATHMOX

Once we have seen the description of the algorithm, it is important to analyze how path models are compared in PATHMOX and how the *F*-test is applied as the split criterion. To do this, let us consider a parent node containing n elements. According to the binary split produced by the segmentation variable we divide the elements in two subsets. For each partition (see figure 7.8) we have two child nodes, each one with its corresponding path model. One node contains n_A elements and the other node contains n_B observations ($n = n_A + n_B$). The strategy to select the best split is based on the comparison of path models for each pair of segments.

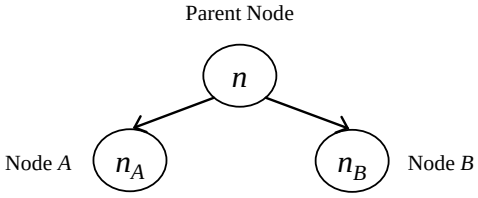


Figure 7.8. *Binary split of a parent node*

The comparison of two models within PATHMOX approach focuses on the structural relationships by taking into account the path coefficients of the structural part of the models. In regards to the measurement part of the models, it can freely fluctuate without imposing any kind restrictions on it.

Comparing two structural models can be put in terms of comparing two regression models. Comparing two regression models, in turn, can be performed by testing the equality of coefficients among regressions using for that purpose a hypothesis test introduced by Lebart, Morineau and Fénelon (1979; 1985, pp. 212) that is an *F*-type based statistic (Fisher, 1935; Scheffé, 1959; Seber, 1966). This test is similar to the one introduced by Chow (1960), and discussed in Ghilagaber (2004) and Moreno *et al* (2005), for testing the equality between sets of coefficients in two linear regressions.

In our case, instead of comparing two regression models we are comparing two structural or inner models. Actually, a structural model is nothing other than a set of regressions among latent variables; one regression for each endogenous variable. For this reason, we have adapted the *F* statistic to test the equality of regression coefficients to the structural case for testing the equality of path coefficients.

The description of the test of equality of path coefficients and its related *F* statistic is done by considering two cases of structural models. The first case consists of a simple structural model with only one endogenous latent variable, that is, with only one regression model. The second case is the generalization for more than one endogenous latent construct which is the common situation in real path modeling applications. Before providing the discussion of the test, it is convenient to present two lemmas (described in Lebart *et al*, 1985) based on the multiple regression model.

The classical multiple linear model can be written as

$$y = XB + \varepsilon \tag{7.1}$$

where:
- y is a (nx1) vector of observations on the response variable
- X is a (nxp) data matrix of observations on the p explanatory variables
- B is a (px1) vector of regression coefficients
- ε is a (nx1) vector of errors

The residuals are assumed to be normally distributed with zero mean and finite variance, that is, $E(\varepsilon) = 0$ and $V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n$ (with σ^2 unknown).

Lemma 1 (Lebart *et al* 1985, pp.201)
Let ε be a normally distributed vector in R^n with $E(\varepsilon)=0$ and $V(\varepsilon)=\sigma^2 I$

1) Let Q be an (nxn) symmetric and idempotent matrix. Then, the quadratic form $\varepsilon' Q_0 \varepsilon / \sigma^2$ follows a Chi-square distribution with υ degrees of freedom (υ is the rank of Q).

2) Let L be a matrix such that $LQ = 0$. Then, the vectors $L\varepsilon$ and $Q\varepsilon$ follow independent normal distributions. In particular, the vector $L\varepsilon$ and the variable $\varepsilon' Q_0 \varepsilon / \sigma^2$ are independents.

Lemma 2 (Lebart *et al* 1985, pp.208)
Let ε be a normally distributed vector in R^n with $E(\varepsilon)=0$ and $V(\varepsilon)=\sigma^2 I$. Consider two matrices X and X_0 where X_0 is defined as X_0=XA (for any matrix A).

From X and X_0 the following matrices Q and Q_0 are given as:

$$Q = I_n - X(X'X)^{-1}X' \tag{7.2}$$

$$Q_0 = I_n - X_0(X_0'X_0)^{-1}X_0' \tag{7.3}$$

It can be shown that the quotient

$$F = \frac{(\varepsilon' Q_0 \varepsilon - \varepsilon' Q \varepsilon) \big/ (\upsilon_0 - \upsilon)}{\varepsilon' Q \varepsilon \big/ \upsilon} \tag{7.4}$$

follows an F distribution with (υ_0-υ) and υ degrees of freedom, where:
- υ is the rank of Q ;
- υ_0 is the rank of Q_0 ;

If we consider the vector of residuals $e = y - \hat{y}$, we can define

- SS_{H0} as the sum of squares of residuals associated to Q_0; $SS_{H0} = \left(\sum_{i=1}^{n} e_i^2 \right)_{H_0} = \varepsilon' Q_0 \varepsilon$;

- SS_{H1} as the sum of squares of residuals associated to Q; $SS_{H1} = \left(\sum_{i=1}^{n} e_i^2 \right)_{H_1} = \varepsilon' Q \varepsilon$.

The F statistic can be re-expressed as follows:

$$F = \frac{(\varepsilon' Q_0 \varepsilon - \varepsilon' Q \varepsilon) \big/ (\upsilon_0 - \upsilon)}{\varepsilon' Q \varepsilon \big/ \upsilon} = \frac{(SS_{H0} - SS_{H1}) \big/ (\upsilon_0 - \upsilon)}{SS_{H1} \big/ \upsilon} \tag{7.5}$$

7.3.4 Models Comparison: Simple Case

Adaptation of the *F*-statistic to structural models: Simple Case

In usual PLS path modeling applications we are in presence of several endogenous latent variables. However, in order to describe the hypothesis test with the *F*-statistic we begin with the simple case of considering only one endogenous latent variable which implies that we have a structural model with only one structural equation (i.e. one regression).

Let us suppose a simple path model with three latent constructs: two exogenous, ξ_1 and ξ_2, and one endogenous, η, like shown in Figure 7.9.

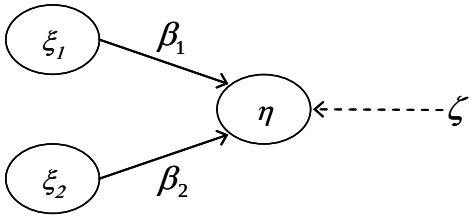


Figure 7.9. *Path diagram for a simple path model with one endogenous latent variable*

The model can be expressed as follows:

$$\eta = \beta_1\xi_1 + \beta_2\xi_2 + \zeta \tag{7.6}$$

The error term ζ is assumed to be normally distributed with zero mean and finite variance, that is, $E(\zeta) = 0$ and $V(\zeta) = \sigma^2 I$. This implies that $\eta \sim N(\beta_1\xi_1 + \beta_2\xi_2, \sigma^2 I)$. Define

$\xi = [\xi_1\ \xi_2]$ a (nx2) matrix with the explicative latent variables (7.7)

$B = [\beta_1\ \beta_2]'$ the vector of path coefficients (7.8)

We can re-express the model in matrix notation as:

$$\eta = \xi B + \zeta \tag{7.9}$$

We suppose that the structural model is associated to the parent node's model containing a total number of n observations. We also assume that the parent node is divided in two child nodes or segments, one segment containing n_A elements and the other one containing n_B observations. Each segment is represented by its own structural model:

$$\begin{aligned} &\text{Segment } A: \quad \eta^A = \xi^A B^A + \zeta^A \\ &\text{Segment } B: \quad \eta^B = \xi^B B^B + \zeta^B \end{aligned} \tag{7.10}$$

with $\zeta^A \sim N(0, \sigma_A^2 I)$ and $\zeta^B \sim N(0, \sigma_B^2 I)$.

The purpose is to test the equality of both models by means of testing the equality of the path coefficients B^A and B^B. Thus, in the null hypothesis the path coefficients are assumed to be equal: $B^A = B^B = B$, that is:

$$H_0: \quad \eta^A = \xi^A B + \zeta^A, \qquad \eta^B = \xi^B B + \zeta^B \tag{7.11}$$

In the alternative hypothesis the path coefficients are considered to be different ($B^A \neq B^B$):

$$H_1: \quad \eta^A = \xi^A B^A + \zeta^A, \qquad \eta^B = \xi^B B^B + \zeta^B \tag{7.12}$$

In matrix notation:

Under H_0:
$$\begin{bmatrix} \eta^A \\ \eta^B \end{bmatrix} = \begin{bmatrix} \xi^A \\ \xi^B \end{bmatrix} [B] + \begin{bmatrix} \zeta^A \\ \zeta^B \end{bmatrix} \tag{7.13}$$

Under H_1:
$$\begin{bmatrix} \eta^A \\ \eta^B \end{bmatrix} = \begin{bmatrix} \xi^A & 0 \\ 0 & \xi^B \end{bmatrix} \begin{bmatrix} B^A \\ B^B \end{bmatrix} + \begin{bmatrix} \zeta^A \\ \zeta^B \end{bmatrix} \tag{7.14}$$

Define the following matrices:

$$X_0 = \underset{(n,p)}{\begin{bmatrix} \xi^A \\ \xi^B \end{bmatrix}} \qquad X = \underset{(n,2p)}{\begin{bmatrix} \xi^A & 0 \\ 0 & \xi^B \end{bmatrix}} \qquad A = \underset{(2p,p)}{\begin{bmatrix} I_p \\ I_p \end{bmatrix}} \tag{7.15}$$

where p is the number of explicative latent variables for η (in this example p=2) and I_p is the identity matrix of order p.

In this situation we have two matrices, X and X_0, where X_0 is defined as X_0=XA (for a given matrix A). Hence, we can apply the previous lemma 2. Substituting matrices X, X_0 and A, we have the following equality:

$$\begin{bmatrix} \xi^A \\ \xi^B \end{bmatrix} = \begin{bmatrix} \xi^A & 0 \\ 0 & \xi^B \end{bmatrix} \begin{bmatrix} I_p \\ I_p \end{bmatrix} \tag{7.16}$$

The matrices Q and Q_0 are defined by

$$Q = I_n - X(X'X)^{-1}X' \quad \text{and} \quad Q_0 = I_n - X_0(X_0'X_0)^{-1}X_0' \tag{7.17}$$

When H_0 is true and $\sigma_A^2 = \sigma_B^2 = \sigma^2$, the quotient

$$F = \frac{(\zeta' Q_0 \zeta - \zeta' Q \zeta) \big/ (\upsilon_0 - \upsilon)}{\zeta' Q \zeta \big/ \upsilon} \tag{7.18}$$

follows an F distribution with (υ_0-υ) and υ degrees of freedom, where:

- $\zeta = \begin{bmatrix} \zeta^A \\ \zeta^B \end{bmatrix}$
- υ is the rank of Q; and υ_0 is the rank of Q_0.

To obtain the degrees of freedom we need to calculate the rank of the denominator and the numerator. By the lemma 1, $\zeta' Q \zeta / \sigma^2$ follows a Chi-square distribution with degrees of freedom equal to the rank of Q:

$$\begin{aligned} \text{rank}(Q) &= \text{tr}(Q) = \text{tr}(I_n - X(X'X)^{-1}X') = \\ &= \text{tr}(I_n) - \text{tr}((X'X)^{-1}X'X) = \text{tr}(I_n) - \text{tr}(I_{2p}) = n - 2p \end{aligned} \tag{7.19}$$

To obtain the degrees of freedom of the numerator we have to examine the expression:

$$\zeta' Q_0 \zeta - \zeta' Q \zeta = \zeta'(Q_0 - Q)\zeta \tag{7.20}$$

We know that both Q_0 and Q are symmetric, therefore $Q_0 - Q$ is symmetric. We also know that Q_0 and Q are orthogonal projectors and, hence, $Q_0 - Q$ is also an orthogonal projector. In order to see that $Q_0 - Q$ is idempotent a geometrical proof is given as follows:

The linear space V(X) spanned by the columns of X (all vectors written Xu) includes the linear subspace V(X_0) spanned by the columns of X_0=XA (since XAv can be expressed as Xu, with u=Av). Therefore the orthogonal space $V^+(X)$ of V(X) is included in the orthogonal $V^+(X_0)$ of V(X_0). As a consequence, the projector Q_0 onto V(X_0) has no effect on the vectors of Q(X). This means that:

$$Q_0 Q = Q Q_0 = Q \quad \text{(since both matrices are symmetric)} \tag{7.21}$$

Then

$$(Q_0 - Q)^2 = Q_0 - Q_0 Q - Q Q_0 + Q = Q_0 - Q \tag{7.22}$$

Thus, the rank of $Q_0 - Q$ equals its trace:

$$\begin{aligned} \text{rank}(Q_0 - Q) &= \text{tr}(Q_0 - Q) = \text{tr}(Q_0) - \text{tr}(Q) = \\ &= \text{rank}(Q_0) - \text{rank}(Q) = (n - p) - (n - 2p) = p \end{aligned} \tag{7.23}$$

In this way, the degrees of freedom of the numerator are $(\upsilon_0\text{-}\upsilon) = p$, and the degrees of freedom of the denominator are $\upsilon = n - 2p$. Then, the F statistic can be expressed as:

$$F = \frac{(\zeta' Q_0 \zeta - \zeta' Q \zeta) \big/ (\upsilon_0 - \upsilon)}{\zeta' Q \zeta \big/ \upsilon} = \frac{(\zeta' Q_0 \zeta - \zeta' Q \zeta) \big/ p}{\zeta' Q \zeta \big/ (n - 2p)} \tag{7.24}$$

If we consider the vector of residuals $z = \eta - \hat{\eta}$, we can define

- SS_{H0} as the sum of squares of residuals associated to Q_0; $SS_{H0} = \left(\sum_{i=1}^{n} z_i^2 \right)_{H_0} = \zeta' Q_0 \zeta$;

- SS_{H1} as the sum of squares of residuals associated to Q; $SS_{H1} = \left(\sum_{i=1}^{n} z_i^2 \right)_{H_1} = \zeta' Q \zeta$.

Note that SS_{H1} can be decomposed into the sum of $SS_{H1}^{A} = \left(\sum (z_i^{A})^2 \right)$ and $SS_{H1}^{B} = \left(\sum (z_i^{B})^2 \right)$.

For convenience, we express the F statistic in terms of SS_{H0} and SS_{H1} as

$$F = \frac{(SS_{H0} - SS_{H1}) \big/ p}{SS_{H1} \big/ (n - 2p)} \tag{7.25}$$

which follows an F distribution with p and $(n - 2p)$ degrees of freedom, where
- p is the number of explicative latent variables for the endogenous construct η
- $n = n_A + n_B$, (number of elements in the model containing the two nodes)
- SS_{H0} is the residual sum of squares under the null hypothesis
- SS_{H1} is the residual sum of squares under the alternative hypothesis

7.3.5 Models Comparison: General Case

Adaptation of the *F*-statistic to structural models: General case

Now, let us suppose a case in which we have more than one endogenous variable. For the sake of simplicity we show the case with two latent constructs but its generalization into more than two LVs is straightforward. Consider a path model with three latent variables as in figure 7.10 with two endogenous latent variables, η_1 and η_2.

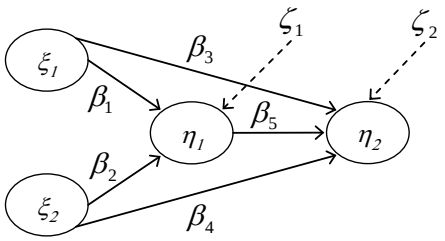


Figure 7.10. *Path diagram of a path model with two endogenous latent variables*

The structural equations for both endogenous constructs are

$$\eta_1 = \beta_1\xi_1 + \beta_2\xi_2 + \zeta_1 \tag{7.26}$$

$$\eta_2 = \beta_3\xi_1 + \beta_4\xi_2 + \beta_5\eta_1 + \zeta_2 \tag{7.27}$$

The disturbance terms ζ_1 and ζ_2 are assumed to be normally distributed with zero mean and finite variance, that is, $E(\zeta_1) = E(\zeta_2) = 0$ and $V(\zeta_1) = V(\zeta_2) = \sigma^2 I$. It is also assumed that $\text{cov}(\zeta_1, \zeta_2) = 0$.

We define new matrices

$X_1 = [\xi_1\ \xi_2]$	a column matrix with the explicative latent variables of η_1
$B_1 = [\beta_1\ \beta_2]'$	a vector of path coefficients for the regression of η_1
$X_2 = [\xi_1\ \xi_2\ \eta_1]$	a column matrix with the explicative latent variables of η_2
$B_2 = [\beta_3\ \beta_4\ \beta_5]'$	a vector of path coefficients for the regression of η_2

Then, the structural equations are expressed as:

$$\eta_1 = X_1 B_1 + \zeta_1 \tag{7.28}$$

$$\eta_2 = X_2 B_2 + \zeta_2 \tag{7.29}$$

We suppose that the structural model is associated to the parent node's model containing a total number of n observations. We also assume that the parent node is divided in two child nodes or segments, one segment containing n_A elements and the other one containing n_B observations. Each segment is represented by its own structural model:

$$\text{Segment } A: \quad \eta_1^A = X_1^A \mathrm{B}_1^A + \zeta_1^A \quad \text{and} \quad \eta_2^A = X_2^A \mathrm{B}_2^A + \zeta_2^A \tag{7.30}$$

$$\text{Segment } B: \quad \eta_1^B = X_1^B \mathrm{B}_1^B + \zeta_1^B \quad \text{and} \quad \eta_2^B = X_2^B \mathrm{B}_2^B + \zeta_2^B \tag{7.31}$$

with $\zeta_1^A \sim \mathrm{N}(0, \sigma_A^2 I)$, $\zeta_2^A \sim \mathrm{N}(0, \sigma_A^2 I)$, $\zeta_1^B \sim \mathrm{N}(0, \sigma_B^2 I)$ and $\zeta_2^B \sim \mathrm{N}(0, \sigma_B^2 I)$.

We can define a null hypothesis for each structural equation:

$$H_0 \text{ for } \eta_1: \quad \underset{(n,1)}{\begin{bmatrix} \eta_1^A \\ \eta_1^B \end{bmatrix}} = \underset{(n,p_1)}{\begin{bmatrix} X_1^A \\ X_1^B \end{bmatrix}} \underset{(p_1,1)}{[B_1]} + \underset{(n,1)}{\begin{bmatrix} \zeta_1^A \\ \zeta_1^B \end{bmatrix}} \tag{7.32}$$

$$H_0 \text{ for } \eta_2: \quad \underset{(n,1)}{\begin{bmatrix} \eta_2^A \\ \eta_2^B \end{bmatrix}} = \underset{(n,p_2)}{\begin{bmatrix} X_2^A \\ X_2^B \end{bmatrix}} \underset{(p_2,1)}{[B_2]} + \underset{(n,1)}{\begin{bmatrix} \zeta_2^A \\ \zeta_2^B \end{bmatrix}} \tag{7.33}$$

where $n = n_A + n_B$ is the number of elements in the model containing the two nodes; p_j is the number of explicative latent variables for each j-th endogenous construct j=1,…,J (in this example J=2).

The corresponding alternative hypotheses are

$$H_1 \text{ for } \eta_1: \quad \underset{(n,1)}{\begin{bmatrix} \eta_1^A \\ \eta_1^B \end{bmatrix}} = \underset{(n,2p_1)}{\begin{bmatrix} X_1^A & 0 \\ 0 & X_1^B \end{bmatrix}} \underset{(2p_1,1)}{\begin{bmatrix} B_1^A \\ B_1^B \end{bmatrix}} + \underset{(n,1)}{\begin{bmatrix} \zeta_1^A \\ \zeta_1^B \end{bmatrix}} \tag{7.34}$$

$$H_1 \text{ for } \eta_2: \quad \underset{(n,1)}{\begin{bmatrix}\eta_2^A\\ \eta_2^B\end{bmatrix}} = \underset{(n,2p_2)}{\begin{bmatrix}X_2^A & 0\\ 0 & X_2^B\end{bmatrix}} \underset{(2p_2,1)}{\begin{bmatrix}B_2^A\\ B_2^B\end{bmatrix}} + \underset{(n,1)}{\begin{bmatrix}\zeta_2^A\\ \zeta_2^B\end{bmatrix}} \tag{7.35}$$

If we take into account the n_A observations belonging to segment A and the n_B observations belonging to segment B, and we concatenate them together forming two groups, then each pair of hypotheses can be combined in more compact expressions:

$$H_0: \quad \underset{(2n,1)}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{(2n,p_1+p_2)}{\begin{bmatrix}X_1^A & 0\\ 0 & X_2^A\\ X_1^B & 0\\ 0 & X_2^B\end{bmatrix}} \underset{(p_1+p_2,1)}{\begin{bmatrix}B_1\\ B_2\end{bmatrix}} + \underset{(2n,1)}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{7.36}$$

$$H_1: \quad \underset{(2n,1)}{\begin{bmatrix}\eta_1^A\\ \eta_2^A\\ \eta_1^B\\ \eta_2^B\end{bmatrix}} = \underset{(2n,2p_1+2p_2)}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \underset{(2p_1+2p_2,1)}{\begin{bmatrix}B_1^A\\ B_2^A\\ B_1^B\\ B_2^B\end{bmatrix}} + \underset{(2n,1)}{\begin{bmatrix}\zeta_1^A\\ \zeta_2^A\\ \zeta_1^B\\ \zeta_2^B\end{bmatrix}} \tag{7.37}$$

Assuming that the error terms associated to the endogenous latent variables are uncorrelated, the matrices X_0, X and A required in the lemma 2 to obtain the F statistic can be defined as:

$$X_0 = \underset{(2n,p_1+p_2)}{\begin{bmatrix}X_1^A & 0\\ 0 & X_2^A\\ X_1^B & 0\\ 0 & X_2^B\end{bmatrix}} \qquad X = \underset{(2n,2p_1+2p_2)}{\begin{bmatrix}X_1^A & 0 & 0 & 0\\ 0 & X_2^A & 0 & 0\\ 0 & 0 & X_1^B & 0\\ 0 & 0 & 0 & X_2^B\end{bmatrix}} \qquad A = \underset{(2p_1+2p_2,p_1+p_2)}{\begin{bmatrix}I_{p_1} & 0\\ 0 & I_{p_2}\\ I_{p_1} & 0\\ 0 & I_{p_2}\end{bmatrix}} \tag{7.38}$$

where I_{p_j} is the identity matrix of order p_j. We can see that $X_0 = XA$.

The matrices Q and Q_0 are defined by

$$Q = I_{2n} - X(X'X)^{-1}X' \text{ and } \qquad Q_0 = I_{2n} - X_0(X_0'X_0)^{-1}X_0' \tag{7.39}$$

The sum of squares of residuals under the null hypothesis, SS_{H0}, can be decomposed into the sum of squares of each structural equation as:

$$SS_{H0} = SS_{H0|\eta_1} + SS_{H0|\eta_2} \tag{7.40}$$

where $SS_{H0|\eta_1}$ and $SS_{H0|\eta_2}$ are the residual sum of squares of η_1 and η_2 under H_0, respectively. Likewise, the sum of squares of residuals under the alternative hypothesis, SS_{H1}, can be decomposed as:

$$SS_{H1} = SS_{H1}^A + SS_{H1}^B \tag{7.41}$$

where SS_{H1}^A and SS_{H1}^B are the residual sum of squares of segments A and B under H_1, respectively. Note that:

$$SS_{H1}^{A} = SS_{H1|\eta_1}^{A} + SS_{H1|\eta_2}^{A} \quad \text{and} \quad SS_{H1}^{B} = SS_{H1|\eta_1}^{B} + SS_{H1|\eta_2}^{B} \tag{7.42}$$

Substituting the sum of squares of each segment in SS_1 we have that:

$$SS_{H1} = (SS_{H1|\eta_1}^{A} + SS_{H1|\eta_2}^{A}) + (SS_{H1|\eta_1}^{B} + SS_{H1|\eta_2}^{B})$$

$$SS_{H1} = (SS_{H1|\eta_1}^{A} + SS_{H1|\eta_1}^{B}) + (SS_{H1|\eta_2}^{A} + SS_{H1|\eta_2}^{B}) \tag{7.43}$$

If we join the terms belonging to η_1 and η_2, respectively, then

$$SS_{H1} = SS_{H1|\eta_1} + SS_{H1|\eta_2} \tag{7.44}$$

When H_0 is true and $\sigma_A^2 = \sigma_B^2 = \sigma^2$, the quotient F given by:

$$F = \frac{(SS_{H0} - SS_{H1}) \big/ (\upsilon_0 - \upsilon)}{SS_{H1} \big/ \upsilon} \tag{7.45}$$

follows an F distribution with (υ_0-υ) and υ degrees of freedom, where:
- υ is the rank of Q
- υ_0 is the rank of Q_0

Substituting the degrees of freedom (υ_0-υ) and υ we have:

$$F = \frac{(SS_{H0} - SS_{H1}) \big/ (p_1 + p_2)}{SS_{H1} \big/ [2n - 2(p_1 + p_2)]} \tag{7.46}$$

General Formula
The formula of the F statistic for the general case is given as

$$F = \frac{(SS_{H0} - SS_{H1}) \big/ (\upsilon_0 - \upsilon)}{SS_{H1} \big/ \upsilon} \tag{7.47}$$

where:

- $\upsilon_0 - \upsilon = \sum_{j=1}^{J} p_j$;

- $\upsilon = Jn - 2\sum_{j=1}^{J} p_j$;

- p_j is the number of explicative latent variables for each j-th endogenous latent variable j=1,…,J;
- J is the total number of endogenous latent variables.

Substituting the degrees of freedom (υ_0-υ) and υ, the F statistic is expressed as:

$$F = \frac{(SS_{H0} - SS_{H1}) \Big/ \sum_{j=1}^{J} p_j}{SS_{H1} \Big/ \sum_{j=1}^{J} (n - 2p_j)} = \frac{\left(Jn - 2\sum_{j=1}^{J} p_j \right)(SS_{H0} - SS_{H1})}{\left(\sum_{j=1}^{J} p_j \right) SS_{H1}} \tag{7.48}$$

with $\sum_{j=1}^{J} p_j$ and $Jn - 2\sum_{j=1}^{J} p_j$ degrees of freedom.

7.3.6 Hypothesis test considerations

As we have seen in the general case with two or more endogenous latent variables, the test of equality between two sets of path coefficients requires the assumption of equal variances of the error terms. Consider for example the structural model described in section 7.3.5 with two endogenous latent variables, η_1 and η_2. It is assumed that the corresponding error terms have equal variances: var(ζ_1) = var(ζ_2) = $\sigma^2 I$. Then, when a binary partition splits the elements in two segments (*A* and *B*), the endogenous constructs are split in $\eta_1^A, \eta_1^B, \eta_2^A$, and η_2^B. Under the null hypothesis, the associated error terms are supposed to be normally distributed with equal variances among segments, that is:

$$\zeta_1^A \sim \mathrm{N}(0, \sigma^2 I),\ \zeta_2^A \sim \mathrm{N}(0, \sigma^2 I), \text{ and } \zeta_1^B \sim \mathrm{N}(0, \sigma^2 I),\ \zeta_2^B \sim \mathrm{N}(0, \sigma^2 I).$$

An important concern arises over the assumption of equal variances of the error terms. It is clear that this condition is very unlikely to be found in practice, and consequently, one might consider this assumption as stringent and unrealistic. Thus, a less rigorous assumption is to suppose inequality of variances of the error terms, that is, var(ζ_1) ≠ var(ζ_2).

Let us suppose var(ζ_1) = σ_1^2, and var(ζ_2) = σ_2^2, with $\sigma_1^2 \neq \sigma_2^2$. From equation 7.40, the residual sum of squares in the null hypothesis SS_{H0} can be decomposed as:

$$SS_{H0} = SS_{H0|\eta_1} + SS_{H0|\eta_2}$$

Unbiased estimates of σ_1^2 and σ_2^2 under H_0 can be calculated as:

$$\text{estimate of } \sigma_1^2 : \frac{SS_{H0|\eta_1}}{(n - p_1)}, \quad \text{and} \quad \text{estimate of } \sigma_2^2 : \frac{SS_{H0|\eta_2}}{(n - p_2)} \tag{7.49}$$

Expressing $SS_{H0} = SS_{H0|\eta_1} + SS_{H0|\eta_2}$ in terms of the estimates of σ_1^2 and σ_2^2 gives:

$$SS_{H0} = (n - p_1)\sigma_1^2 + (n - p_2)\sigma_2^2 \tag{7.50}$$

Likewise, from eq. 7.44, the residual sum of squares SS_{H1} in H_1 can be decomposed as follows:

$$SS_{H1} = SS_{H1|\eta_1} + SS_{H1|\eta_2}$$

Unbiased estimates of σ_1^2 and σ_2^2 under H_1 can be obtained as:

$$\text{estimate of } \sigma_1^2: \ \frac{SS_{H1|\eta_1}}{n-2p_1}, \qquad \text{and} \quad \text{estimate of } \sigma_2^2: \ \frac{SS_{H1|\eta_2}}{n-2p_2} \tag{7.51}$$

Expressing $SS_{H1} = SS_{H1|\eta_1} + SS_{H1|\eta_2}$ in terms of the estimates of σ_1^2 and σ_2^2 gives:

$$SS_{H1} = (n-2p_1)\sigma_1^2 + (n-2p_2)\sigma_2^2 \tag{7.52}$$

Regarding the numerator of F, when H_0 is true, the difference of the residual sum of squares ($SS_{H0} - SS_{H1}$) gives:

$$SS_{H0} - SS_{H1} = \left\{(n-p_1)\sigma_1^2 + (n-p_2)\sigma_2^2\right\} - \left\{(n-2p_1)\sigma_1^2 - (n-2p_2)\sigma_2^2\right\}$$

$$SS_{H0} - SS_{H1} = \sigma_1^2(n-p_1-n+2p_1) + \sigma_2^2(n-p_2-n+2p_2)$$

$$SS_{H0} - SS_{H1} = p_1\sigma_1^2 + p_2\sigma_2^2 \tag{7.53}$$

The F-statistic is written in the following form:

$$F = \frac{(SS_{H0} - SS_{H1}) \big/ (p_1+p_2)}{SS_{H1} \big/ \left[2n-2(p_1+p_2)\right]} = \frac{\dfrac{(p_1\sigma_1^2 + p_2\sigma_2^2)}{(p_1+p_2)}}{\dfrac{(n-2p_1)\sigma_1^2 + (n-2p_2)\sigma_2^2}{2n-2(p_1+p_2)}}$$

$$F = \frac{\left(\dfrac{p_1}{p_1+p_2}\right)\sigma_1^2 + \left(\dfrac{p_2}{p_1+p_2}\right)\sigma_2^2}{\left(\dfrac{n-2p_1}{2n-2(p_1+p_2)}\right)\sigma_1^2 + \left(\dfrac{n-2p_2}{2n-2(p_1+p_2)}\right)\sigma_2^2} \tag{7.54}$$

General Formula

In the general case, the formula is given by:

$$F = \frac{\sum_{j=1}^{J} p_j\sigma_j^2 \Big/ \sum_{j=1}^{J} p_j}{\left(\sum_{j=1}^{J}(n-2p_j)\sigma_j^2\right) \Big/ \left(Jn - 2\sum_{j=1}^{J} p_j\right)}$$

where:

- p_j is the number of explicative latent variables for each j-th endogenous latent variable j=1,…,J;
- J is the total number of endogenous latent variables.
- $\hat{\sigma}_j^2$ is the unbiased estimate of the error variance of the j-th endogenous variable

Let us suppose however that the concatenated endogenous latent variable $\tilde{\eta} = \left[\eta_1^A \quad \eta_2^A \,\vdots\, \eta_1^B \quad \eta_2^B\right]'$, (i.e. vertical concatenation of the endogenous constructs) has error term with an average variance $\tilde{\sigma}^2$. Hence, looking at equation 7.54 we can assume that we are applying the hypothesis test with regard to an artificial endogenous latent variable $\tilde{\eta}$ with an artificial error term whose variance is a weighted average of the variances of the original error terms. The variance $\tilde{\sigma}^2$ is a pooled average of the variances of the latent endogenous constructs.

The variance $\tilde{\sigma}^2$ behaves as a simple average as the number of observations *n* becomes large, that is:

$$\tilde{\sigma}^2 = \frac{\sigma_1^2 + \sigma_2^2}{2}$$

Thus, the *F*-statistic behaves as if we had an "artificial" endogenous variable $\tilde{\eta}$ whose variance is an average variance of the latent error terms. Regardless of the variances in the error terms, the important issue is that we can think about an artificial model which has an error term that is a weighted average of the error variances.

7.3.7 Validation of Segments

Once a PATHMOX tree has been constructed and the final nodes have been obtained, it is necessary to identify the differences among segments. Since the PATHMOX approach is based on the inner structural model, we focus only on the path coefficients. This implies comparing the path coefficients of the different segments.

We propose the use of bootstrapping to validate the results of the final segments. The bootstrap samples are built by resampling with replacement from the original sample. The samples consist of the same number of units as in the original sample, and the number of resamples is fixed to 100. Moreover, bootstrap confidence intervals of the path coefficients can be obtained from the resampling procedure. Hence, confidence intervals allow identification of those coefficients in a segment that may be different to the rest of the segments. With this information, the user can identify those structural relationships in which some path model differs from the rest of the segments.

Chapter 8

Simulation studies

In order to evaluate the sensitivity of the split criterion used in PATHMOX (i.e., the *F*-test) we run a series of Monte Carlo simulation analysis divided in three major studies. Hence, our purpose is threefold. First, we are interested in evaluating the hypothesis test when the variances of the endogenous error terms are different. In second place, we aim to assess the capabilities of the *F*-test when comparing two structural models with normal data under different experimental conditions. In third place, we aim to assess the capabilities of the test when comparing two structural models with non-normal data.

Our simulations are mainly based on the exposures of Chin and Newsted (1999), Cassel *et al* (1999, 2000), Westlund *et al* (2001), Chin *et al* (2003), Goodhue *et al* (2006), and Ringle (2006). Broadly speaking, Monte Carlo simulation techniques refer to procedures for generating artificial data following a specific statistical model. As mentioned in Chin *et al* (2003) and Goodhue *et al* (2006), Monte Carlo simulation studies are commonly used by structural equation modeling researchers for different purposes. Some of those purposes include investigating the robustness of statistical estimators, comparing alternative estimation approaches, and examining the behavior of different path modeling procedures. In our case, we have adopted a simulation framework with simple path models generated under different experimental conditions.

We begin this chapter with the simulation design to study the first issue (different variances of the endogenous error terms). The second section provides the description of the simulation design for the second and third issues (comparing two structural models under different experimental conditions). The simulation analysis with normal data is presented in the third section. Likewise, the simulation analysis with non-normal data is discussed in the fourth section.

8.1 Simulation study concerning different variances of the endogenous error terms

The first simulation study consists of assessing the performance of the hypothesis test (when the null hypothesis H_0 is true) in presence of different variances of the endogenous error terms. This simulation is related to the discussion presented in section 7.3.6 of the previous chapter.

As it has been seen, the *F*-test requires the assumption of equality of variances in the endogenous error terms, that is: $\zeta_j \sim N(0, \sigma^2 I)$, where ζ_j is associated to the *j*-th endogenous latent variable.

Although it has been argued in section 7.3.6 that this assumption does not represent a serious problem, we have decided to carry out a simulation study in which different variances of the endogenous error terms are considered. The objective is to provide evidence that the *F*-test still performs well when each ζ_j has its own particular σ_j^2. We consider a simple structural model with three latent variables as shown in figure 8.1. The structural model has one exogenous latent variable, ξ, and two endogenous latent variables, η_1 and η_2. The structural equations are defined as

$$\eta_1 = \beta_1 \xi + \zeta_1$$
$$\eta_2 = \beta_2 \xi + \beta_3 \eta_1 + \zeta_2 \qquad (8.1)$$

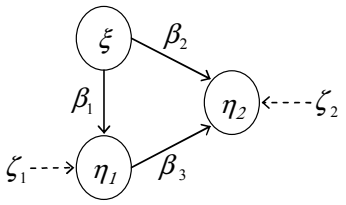


Figure 8.1. *Structural model of the first simulation study*

Two conditions have been manipulated to investigate the performance of the test: the variance of the endogenous error terms, and the sample size. Simulated observations of the exogenous latent variable ξ are generated from a standard normal distribution by the 'rnorm' function of R software (R Development Core Team, 2008). The error term ζ_1 is generated from a normal distribution with zero mean and different variances which can take four possible values {0.35, 0.30, 0.25, 0.20}. Likewise, the error term ζ_2 is generated from a normal distribution with zero mean and variances taking four possible values {0.05, 0.10, 0.15, 0.20}. Then, the endogenous variables η_1 and η_2 are generated following the proposed model. The selected values for the variances of the error terms are supposed to reflect four situations:

- large difference between variances (σ_1^2 =0.35, σ_2^2 =0.05),
- medium difference between variances (σ_1^2 =0.30, σ_2^2 =0.10),
- small difference between variances (σ_1^2 =0.25, σ_2^2 =0.15),
- equality of variances (σ_1^2 =0.20, σ_2^2 =0.20).

The values of the path coefficients are taken as β_1=0.8, β_2=0.7, and β_3=0.5. Finally, the selected values of the sample sizes are 100, 200, 500, and 1000.

In this way, there are 4x4=16 experimental combinations (4 sample sizes x 4 pairs of variances). We are interested in examining the behavior of the hypothesis test by taking into account the possibility of having endogenous error terms with different variances. Given that we are assuming the null hypothesis to be true, the observations are randomly split in two subsets of equal size in order to produce two "virtual" segments. Then, a structural model is estimated for each partition, and the hypothesis test to evaluate the equality of path coefficients is applied. This process is repeated 100 times for each model experiment condition.

The mean and median of the *p*-values within each set of 100 replications are calculated. In addition, the proportion of the 100 replications that yielded significant results (*p*-value < 0.05) is also recorded. These summary results are displayed in table 8.1. It can be seen that the average of *p*-values ranges from 0.407 (for sample size 500, and medium variances difference) to 0.508 (for sample size 100, and equal variances). In the case of the median values, the smallest one is associated to sample size 100 with large variance difference. Conversely, the largest median value 0.555 is found in the sample size of 200 with equal variances. The last sub-table contains the proportion of *p*-values<0.05, that is, the proportion of significant results that lead to the rejection of the null hypothesis H_0 at a 0.05 significance threshold. The column of large variance differences shows the largest proportions of significant results. In the column of equal variances, in contrast, there is the smallest proportions of significant results.

Table 8.1. Summary results of the simulation study with different variances in the endogenous error terms

Average of p-values

	Difference of endogenous error variances			
Sample Size	*large*	*medium*	*small*	*equal*
100	0.507	0.454	0.445	0.509
200	0.500	0.468	0.404	0.461
500	0.472	0.511	0.483	0.509
1000	0.494	0.470	0.419	0.534

Median of p-values

	Difference of endogenous error variances			
Sample Size	*large*	*medium*	*small*	*equal*
100	0.516	0.444	0.392	0.530
200	0.543	0.427	0.330	0.453
500	0.476	0.489	0.433	0.530
1000	0.519	0.423	0.369	0.516

Proportion of p-values < 0.05

	Difference of endogenous error variances			
Sample Size	*large*	*medium*	*small*	*equal*
100	0.10	0.05	0.07	0.04
200	0.05	0.08	0.07	0.04
500	0.10	0.10	0.05	0.05
1000	0.13	0.08	0.07	0.02

Graphical results for each set of 100 repetitions are contained in figures 8.2 – 8.5. The results of the model experiment conditions with sample size 100 are shown in figure 8.2. The figure contains the boxplots of the *p*-values for each variance condition (e.g., large, medium, small, and equal), and a bar-chart of the proportions of the *p*-values that are smaller than a threshold of 0.05.

Looking at the four boxplots, there is some difference between them. The first quartile of the first box-plot is clearly above the rest of the other first quartiles. On the contrary, the third quartile of the medium-variance's boxplot is above the other third quartiles. Regarding the bar-chart of the proportions of *p*-values<0.05, the condition of large variance difference has the tallest bar, whereas the bar of the equal variance condition is the shortest one. In summary, the graphical information of figure 8.2

reflects the fact that the *F*-test performs reasonably well even when the endogenous error terms have different variances, assuming that the null hypothesis is true.

Figures 8.3, 8.4 and 8.5, present similar patterns in the case of equal variances. One can see that the proportion of *p*-values in this case is always below the five percent. With respect to the boxplots of large differences, it is not possible to appreciate a clear trend in order to differentiate it from the rest of boxplots. However, we note that the larger proportions of *p*-values smaller than 0.05 occur with large variance differences. In fact, the proportion of *p*-values<0.05 decreases as the difference of endogenous error variances become smaller. In summary, the results provide evidence in order to support the good performance of the hypothesis test among the different experimental conditions. Hence, breaking the assumption of equal variances in the error terms of the endogenous latent variables may not be a serious concern.

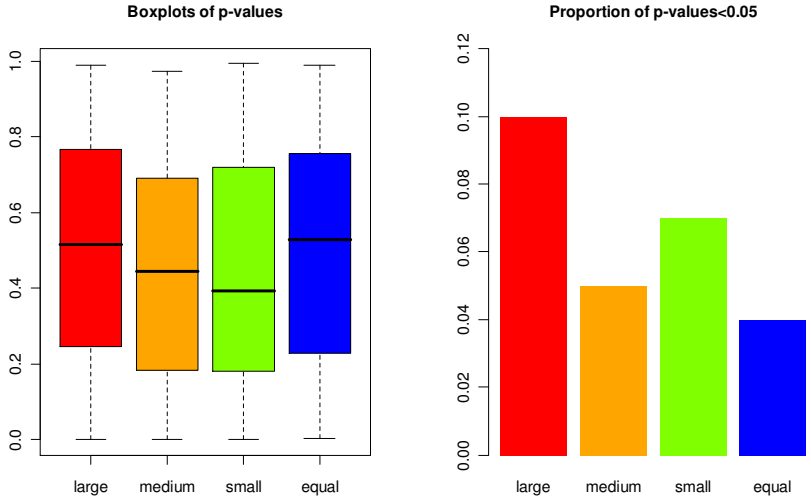


Figure 8.2. *Simulation results of the first study with sample size 100*

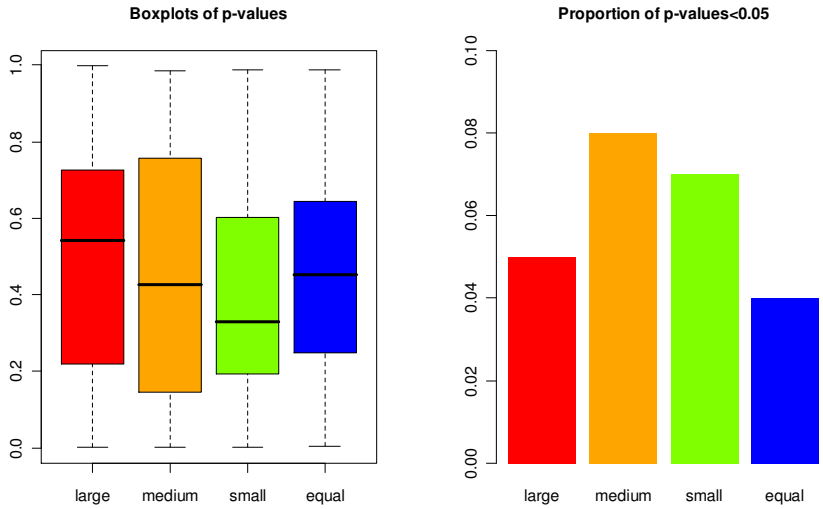


Figure 8.3. *Simulation results of the first study with sample size 200*

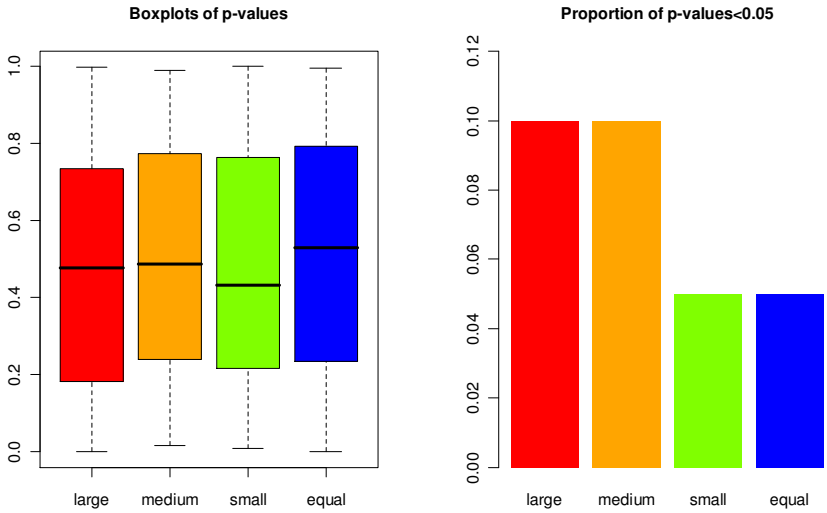


Figure 8.4. *Simulation results of the first study with sample size 500*

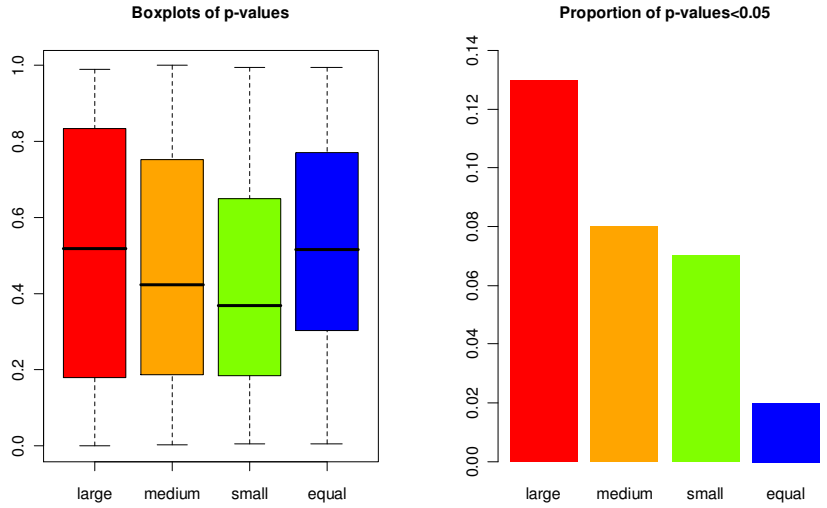


Figure 8.5. *Simulation results of the first study with sample size 1000*

8.2 Simulation study concerning the comparison of two structural models

Monte Carlo simulation techniques refer to procedures for generating artificial data following a specific statistical model. As mentioned in Chin *et al* (2003) and Goodhue *et al* (2006), Monte Carlo simulation studies are commonly used by structural equation modeling researchers for different purposes. Some of those purposes include investigating the robustness of statistical estimators, comparing alternative estimation approaches, and examining the behavior of different path modeling procedures. In our case, we have adopted a simulation framework with simple path models generated under different experimental conditions (e.g., data distributions, sample size, path coefficients, disturbance terms for the endogenous construct, and measurement errors for the indicators).

The main objective behind the simulation analyses is to evaluate the performance of the *F*-test when two path models are compared. Since the *F*-test is intended to test the equality of path coefficients between two models, we want to generate data that follow different path models. Thus, the idea is to simulate two path models which are assumed to reproduce a binary partition of some parent node. In other words, we want to generate two path models that play the role of two child nodes (i.e., two segments).

In order to generate the data we follow one of the approaches for data generation in structural equation modeling according to Reinartz *et al* (2005). The approach consists of first generating data on the latent variables following the relationships specified in the structural model and then, data is generated on the manifest variables from the latent variables. According to Reinartz *et al* (2005), this data generation procedure is "better suited to generate data with distributional characteristics imposed by the model".

Since we are interested in assessing the capabilities of the *F*-test to detect different segments under different experimental conditions, we need to control the amount of difference between segments. The way in which this is done is by determining the path coefficients of the path models. Hence, the overall plan is to generate couples of path models under different data generating conditions. We compare each pair of generated path models among each other by applying the *F*-test. In addition, we put special interest in the distance of path coefficients between models (i.e., average squared Euclidean distance). Although the distance of path coefficients is not enough to compare models, we use it as a means to observe the behavior of the *F* statistic.

To illustrate the process for generating the series of segments (generated in pairs) assume a simple path model with only two latent variables, ξ and η, as the one shown in figure 8.6.

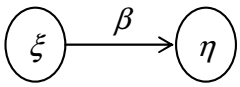


Figure 8.6. *Simple path model with two latent variables*

The scheme to generate the couples of path models is the following.

Step 1: We begin by generating a first pair of path models (segments *A* and *B*) with identical path coefficients, that is, the path coefficient of segment *A* is identical to the path coefficient of segment *B* (i.e, $\beta^A = \beta^B$)

Step 2: The next pair of segments is generated by keeping fixed the path coefficient of segment *A* while changing the path coefficient of segment *B*. This modification of the path coefficient in segment *B* is made in a relatively "low" magnitude with respect to the value of the path coefficient in segment *A*.

Further steps: The subsequent pairs of segments are generated in the same way as in step two. While the path coefficient of segment *A* keeps fixed, the path coefficient of segment *B* is modified. The change in the path coefficient is done in such a way that the difference of path coefficients between segments increases gradually at each step.

A graphical representation of the segments generating process is shown in figure 8.7.

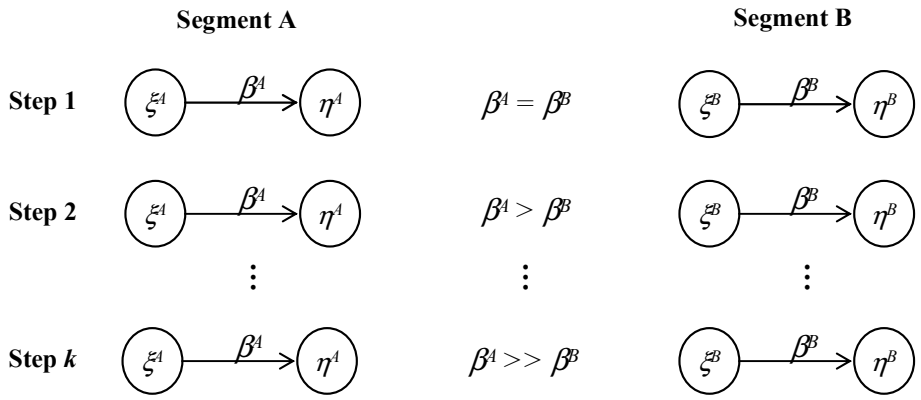


Figure 8.7. *Example of the process for generating the pairs of segments*

We seek to examine the evolution of the *p*-values associated to the *F*-tests when the distance of path coefficients between models increases. This evaluation is performed under different experimental conditions such as sample size, the error terms for both latent and manifest variables, and the data distribution. We have established two groups of simulation analyses: (1) simulation with normal data, and (2) simulation with non-normal data. The two groups of simulations are described in the next sections.

8.3 Comparing two structural models with normal data

The data are generated according to the structural model shown in figure 8.8. The structural model has two exogenous latent variables, ξ_1 and ξ_2, and one endogenous latent variable η. The inner model is defined as

$$\eta = \beta_1 \xi_1 + \beta_2 \xi_2 + \zeta \tag{8.2}$$

where β_1 and β_2 are the structural path coefficients, and ζ is a random disturbance term. The measurement model is assumed to be reflexive and the equations for the exogenous latent variables are defined as

$$x_{j1} = \lambda_{j1} \xi_j + \varepsilon_{j1} \tag{8.3}$$

$$x_{j2} = \lambda_{j2} \xi_j + \varepsilon_{j2} \tag{8.4}$$

$$x_{j3} = \lambda_{j3} \xi_j + \varepsilon_{j3} \tag{8.5}$$

for $j = 1, 2$. The equations for the endogenous latent variable are

$$x_{31} = \lambda_{31} \eta + \varepsilon_{31} \tag{8.6}$$

$$x_{32} = \lambda_{32} \eta + \varepsilon_{32} \tag{8.7}$$

$$x_{33} = \lambda_{33} \eta + \varepsilon_{33} \tag{8.8}$$

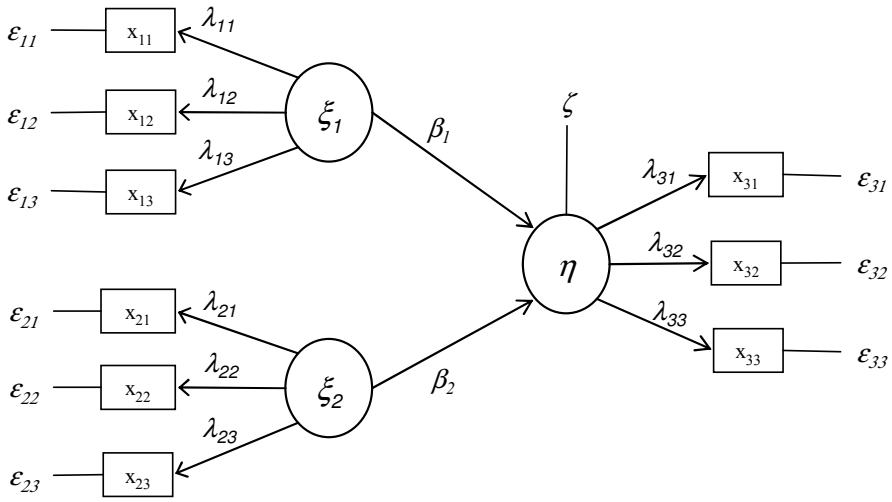


Figure 8.8. *Path diagram of the structural model in the simulation study*

The exogenous latent constructs ξ_1 and ξ_2 are normally distributed with mean zero and unit variance. The error term ζ follows a normal distribution with expectation zero and variance having three levels. The levels are chosen such that the variance of ζ accounts for 10%, 30%, and 50% of the total variance of η. These three levels of noise (10%, 30%, and 50%) represent scenarios of low, medium and high noise, respectively.

The way in which var(ζ) is obtained is as follows. We know that:

$$\text{var}(\eta) = \text{var}(\beta_1\xi_1 + \beta_2\xi_2 + \zeta) \tag{8.9}$$

Assuming independence between latent variables and between the error term ζ we have:

$$\text{var}(\eta) = \text{var}(\beta_1\xi_1) + \text{var}(\beta_2\xi_2) + \text{var}(\zeta) \tag{8.10}$$

$$\text{var}(\eta) = \beta_1^2 + \beta_2^2 + \text{var}(\zeta) \tag{8.11}$$

In the case that var(ζ) accounts for q% of the variance of η we have

$$\text{var}(\zeta) = (q/(1\text{-}q))\ (\beta_1^2 + \beta_2^2) \tag{8.12}$$

For instance, if $\beta_1 = \beta_2 = 0.5$, and q=0.10, we can calculate var(ζ) as

$$\text{var}(\zeta) = (0.10/0.90)\ (0.25 + 0.25) = 0.0555 \tag{8.13}$$

Indicator loadings are specified with $\lambda_{j1} = 0.75$, $\lambda_{j2} = 0.80$, and $\lambda_{j3} = 0.85$, for j=1, 2, 3. Following the exposure of Goodhue *et al* (2006), we select the unequal values for the indicator loadings in order to have similar values of those found in real applications, in which differences among loadings are present.

The error terms ε_{ji}'s for the manifest variables are normally distributed with expectation zero and variance having three levels. As in the case for the error term ζ, the levels of variance for ε are chosen such that the variance of ε accounts for 10%, 30%, and 50% of the total variance of x_{ji}. These three levels represent scenarios of low, medium and high noise, respectively. The way in which var(ε_{ji}) is obtained is as follows. We know that:

$$\text{var}(x_{ji}) = \text{var}(\lambda_{ji}\xi_j + \varepsilon_{ji}) \tag{8.14}$$

Assuming independence between the manifest variables and the error term we have:

$$\text{var}(x_{ji}) = \lambda_{ji}^2 + \text{var}(\varepsilon_{ji}) \quad (8.15)$$

In the case that $\text{var}(\varepsilon_{ji})$ accounts for q% of the variance of x_i we have

$$\text{var}(\varepsilon_{ji}) = (q/(1-q))\ (\lambda_{ji}^2) \quad (8.16)$$

For instance, if $\lambda_{ji} = 0.80$, and q=0.10, we can calculate $\text{var}(\varepsilon_{ji})$ as

$$\text{var}(\varepsilon_{ji}) = (0.10/0.90)\ (0.64) = 0.0711 \quad (8.17)$$

Finally, the values of the path coefficients for each pair of segments (*A* and *B*) are determined in the following way. Segment *A*, as said before, is kept fixed with values for its path coefficients of $\beta^A_1 = \beta^A_2 = 0.5$. Segment *B* will vary its path coefficients, starting with identical values to the first segment (i.e. $\beta^B_1 = \beta^B_2 = 0.5$). The next values of the path coefficients for segment *B* are β^B_1=0.55 and β^B_2=0.45. The coefficient β^B_1 will increase 0.05 at each step. Conversely, coefficient β^B_2 will decrease 0.05 at each step. The final values of the path coefficients in segment B are β^B_1=0.90 and β^B_2=0.10. The list of nine different path coefficients for segment *A* and *B* is contained in table 8.2. A graphical display of the comparison between path models is in figure 8.9.

Table 8.2. List for the nine pairs of path coefficients for segments *A* and *B*.

Num	*Segment A*		*Segment B*	
1	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.50$	$\beta_2 = 0.50$
2	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.55$	$\beta_2 = 0.45$
3	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.60$	$\beta_2 = 0.40$
4	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.65$	$\beta_2 = 0.35$
5	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.70$	$\beta_2 = 0.30$
6	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.75$	$\beta_2 = 0.25$
7	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.80$	$\beta_2 = 0.20$
8	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.85$	$\beta_2 = 0.15$
9	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.90$	$\beta_2 = 0.10$

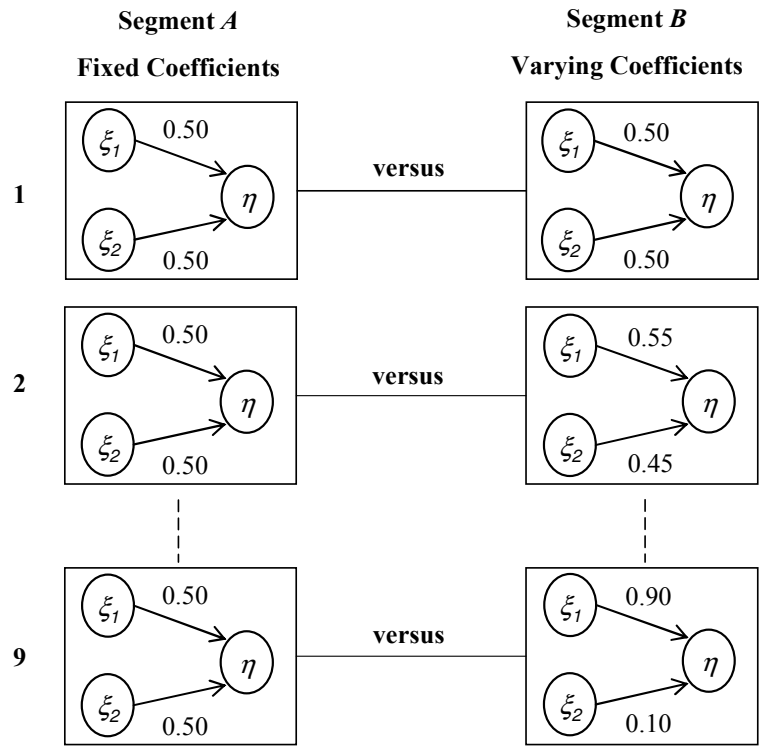


Figure 8.9. *Comparison (by couples) of path models between segments A and B*

8.3.1 Balanced segments (equal proportions)

We assume segments of equal size. This means that we have balanced proportions of 0.50 for each segment. Four sample sizes for the total number of cases are considered: {100, 200, 500, and 1000}. This implies having segments of sizes {50, 100, 250, and 500}.

In total, we have 4x3x3=36 scenarios which is the number of possible combinations of sample sizes and noise levels (4 sample sizes, 3 noise levels for the endogenous construct variance, and 3 noise levels for the indicators variance). For every possible scenario, we compare 9 pairs of inner models which correspond to the nine sets of path coefficients. In each comparison we run 100 repetitions, that is, we generate 100 path models. From the 100 repetitions we calculate the average of the *F*-test's *p*-value, and the mean value of the distance of path coefficients between models.

For example, one possible scenario can be specified with a segment's sample size of 100, low level of noise for the endogenous construct, and medium level of noise for the indicators. The first pair of path models is made with identical segments *A* and *B* (i.e., $\beta^A_1 = \beta^A_2 = 0.5$, $\beta^B_1 = \beta^B_2 = 0.5$). This comparison is repeated 100 times, and it is expected to obtain high *p*-values of the *F*-test in each repetition. In other words, we expect the *F*-test to not detect any difference between segments since they are supposed to be identical.

Under the same hypothetical scenario, the second pair of path models is made by keeping fixed segment *A* (i.e., $\beta^A_1 = \beta^A_2 = 0.5$), but changing segment *B* (i.e., $\beta^B_1 = 0.55$, $\beta^B_2 = 0.45$). Once again, this comparison is repeated 100 times, but unlike the previous comparison, in this case we expect to obtain some of the *p*-values of the *F*-test to be

low. This process is performed subsequently with the rest of the seven pairs of path models, and with the rest of the possible scenarios.

For the computational aspects, we have employed the statistical software R version 2.7.0. We have used its pseudo-random generator function 'rnorm' for normal data, and we have programmed the required algorithms for calculating PLS path models as well as the PATHMOX approach.

Results on the aggregate data level

First we analyze how the significance of the *F*-test is affected by different factors: the difference of path coefficients between inner path models, the different sample sizes (100, 200, 500, and 1000), the different levels in error variance terms of the endogenous construct, and the different levels in error variance terms for indicators. We observe four trends affecting the sensitivity of the *F*-test:

- The more different the path coefficients between segments, the more sensitive is the test
- The larger the sample size, the more sensitive
- The larger the level of noise (error variance) of the endogenous construct, the less sensitive
- The larger the level of noise (error variance) for indicators, the less sensitive.

These results are graphically illustrated in Figure 8.10. There are four plots which represent each of the trends mentioned above. For ease of interpretation we have included in each plot the lowess (Cleveland, 1979) regression line of the *p*-value with respect to the evaluated experimental condition. In the first plot it is possible to observe how the *p*-values decrease (i.e., they become more significant) as the distance of path coefficients between models become more different. The influence of the sample size can be appreciated in the second plot in which the *p*-values decrease as the sample size increases. In the third plot we can see that a higher sensitivity occurs when the level of noise in the variance of the endogenous error term becomes lower. The same effect is seen in the fourth plot with the levels of noise in the variance of the manifest disturbance terms.

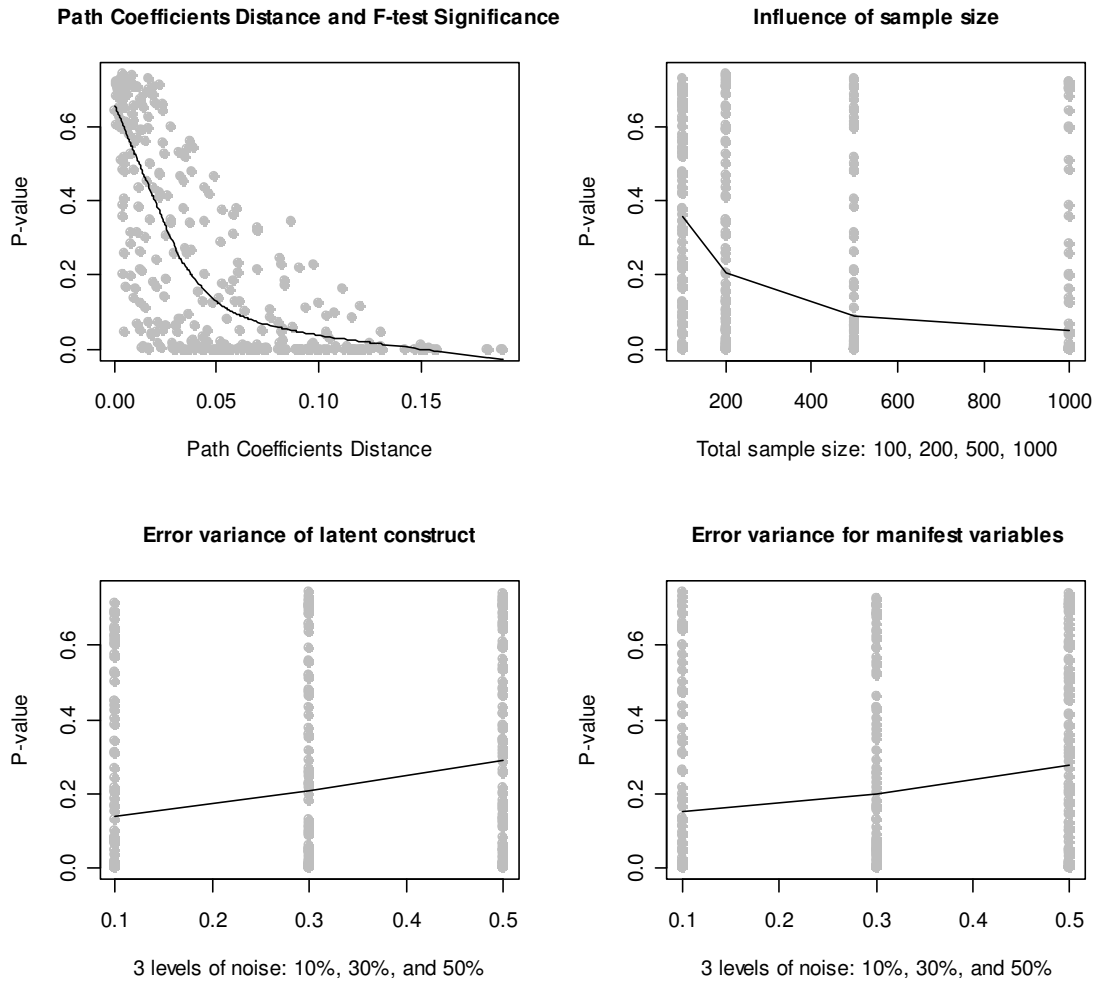


Figure 8.10. *Influence of different data generating conditions on the significance of the F-test at the aggregated level*

Summarized results: tables and charts

The summarized results from the simulations with normal data are contained in the tables 8.3, 8.4, 8.5, and 8.6. Each table corresponds to a different sample size. For instance, table 8.2 shows the summarized results from the 100 repetitions for each of the 9 pairs of compared path models with sample size of 100.

Each table is divided in three horizontal blocks. Each horizontal block, in turn, is divided in three sub-blocks. The horizontal blocks refer to the different levels of variance in the endogenous disturbance term. The sub-blocks refer to the levels of variance in the manifest variables. In each sub-block are displayed the following results: (1) the average distance of path coefficients, (2) the average value of the *p*-values, and (3) the number of *p*-values less than 0.05. We have established the value 0.05 as a threshold in order to have a reference value of the test significance.

For example, consider the first sub-block of the first horizontal block in table 8.3. This sub-block contains the results with low levels of noise for both the endogenous variable and the manifest variables. Each row contains the results of the comparisons between 100 of simulated segments. Examining the first cell of the first row, it can be

seen that it has the value 0.0027. This value is the average distance of path coefficients between the comparisons of the first pair of path models. It is obtained from the 100 repetitions in which two identical segments are compared. The second cell of the first row contains a value of 0.6499 which is the average of the *p*-values from the 100 comparisons. The third cell with a value of 1 corresponds to the number of times that the *p*-value is less than 0.05 in the 100 comparisons. These obtained results are in accord to our expectations. Because we are comparing 100 pairs of identical segments, the distance of path coefficients between the segments must be very small. The *p*-values have a high (non-significant) value. Moreover, from the 100 comparisons, only in one occasion the *p*-value was significant.

Unlike the results in the first row, those of the ninth row show completely different values. In this case, we are comparing 100 of segments with very different path coefficients. Thus, the average distance of path coefficients is large and the all the *p*-values are zero (i.e., highly significant). Hence, all of the 100 comparisons have *p*-values smaller than 0.05.

The examination of the sub-blocks is made for the four tables. In all the sub-blocks, as the comparisons move from the first to the ninth pair of segments we detect three general trends: First, the average distance of the path coefficients between segments becomes larger. Second, the average *p*-value goes from high values to small values. Third, the number of times that the *p*-values are less than 0.05 becomes larger. Hence, we can state that the more different the path coefficients between path models, the greater the significance of the *F*-test. In addition, due to the increment of the variance in the endogenous and the manifest variables, it is possible to observe two different patterns. In the one hand, when the segments are similar or their difference is relatively small (pairs 1,2,3 and 4), the average distance of path coefficients is larger compared to those cases with low levels of noises. In the other hand, when the differences between segments is large (pairs 5 to 9) the average distance of path coefficients is smaller compared to those cases with low levels of noise.

Table 8.3. Simulation results for normal data and balanced nodes. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p*-value (mean)	Number of *p*-values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p*-value (mean)	Number of *p*-values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p*-value (mean)	Number of *p*-values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0076	0.5765	7	0.0100	0.6155	4	0.0126	0.6705	2
2	0.0117	0.4367	10	0.0113	0.5679	4	0.0164	0.6257	4
3	0.0257	0.1876	46	0.0218	0.4232	13	0.0244	0.5244	5
4	0.0395	0.0643	64	0.0382	0.2683	30	0.0337	0.4509	11
5	0.0630	0.0111	94	0.0592	0.1291	57	0.0439	0.3434	20
6	0.1035	0.0014	100	0.0803	0.0713	69	0.0703	0.2156	36
7	0.1303	0.0002	100	0.1040	0.0376	83	0.0835	0.1701	48
8	0.1566	0.0001	100	0.1310	0.0079	97	0.1077	0.0996	63
9	0.1904	0.0000	100	0.1546	0.0041	99	0.1165	0.0837	64
30% of noise in the variance of the endogenous disturbance term									
1	0.0114	0.7051	0	0.0139	0.6567	1	0.0170	0.6977	2
2	0.0142	0.6481	3	0.0145	0.6830	1	0.0175	0.6912	2
3	0.0223	0.4724	7	0.0240	0.5537	6	0.0203	0.6582	1
4	0.0337	0.3769	18	0.0323	0.4633	11	0.0349	0.5173	9
5	0.0582	0.1837	48	0.0524	0.2887	18	0.0376	0.4798	9
6	0.0766	0.0868	61	0.0611	0.2321	34	0.0581	0.3608	18
7	0.1038	0.0468	83	0.0936	0.1097	57	0.0809	0.2453	31
8	0.1252	0.0165	89	0.1043	0.0915	64	0.0818	0.2260	35
9	0.1504	0.0125	94	0.1303	0.0467	84	0.0979	0.2259	38
50% of noise in the variance of the endogenous disturbance term									
1	0.0189	0.6848	1	0.0165	0.7276	0	0.0218	0.7087	2
2	0.0197	0.6611	1	0.0183	0.6825	0	0.0238	0.6568	1
3	0.0278	0.5998	5	0.0237	0.6397	4	0.0194	0.7050	2
4	0.0319	0.5309	6	0.0385	0.5440	5	0.0375	0.5601	3
5	0.0467	0.4159	19	0.0444	0.4327	15	0.0354	0.5361	4
6	0.0587	0.3086	22	0.0527	0.3739	15	0.0489	0.4628	8
7	0.0902	0.2183	42	0.0703	0.3177	26	0.0599	0.3788	13
8	0.1001	0.1256	51	0.0832	0.1834	35	0.0704	0.3265	17
9	0.1203	0.1155	70	0.1117	0.1631	43	0.0866	0.3454	25

Table 8.4. Simulation results for normal data and balanced nodes. Total sample size of 200 (100 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0027	0.6499	1	0.0042	0.6891	0	0.0068	0.6858	2
2	0.0080	0.3140	19	0.0077	0.5268	6	0.0090	0.6039	2
3	0.0212	0.0716	74	0.0203	0.2421	38	0.0158	0.4512	12
4	0.0424	0.0056	97	0.0320	0.0805	68	0.0257	0.3061	23
5	0.0654	0.0001	100	0.0522	0.0207	88	0.0411	0.1528	47
6	0.0888	0.0000	100	0.0751	0.0050	97	0.0555	0.0820	69
7	0.1216	0.0000	100	0.1010	0.0007	100	0.0815	0.0161	90
8	0.1581	0.0000	100	0.1315	0.0000	100	0.0978	0.0153	93
9	0.1899	0.0000	100	0.1525	0.0000	100	0.1253	0.0031	98
30% of noise in the variance of the endogenous disturbance term									
1	0.0044	0.7421	0	0.0065	0.7078	1	0.0071	0.7213	1
2	0.0093	0.5539	5	0.0091	0.6334	2	0.0084	0.7317	0
3	0.0175	0.3490	20	0.0191	0.4331	13	0.0152	0.5594	5
4	0.0360	0.0965	58	0.0289	0.2567	29	0.0224	0.4108	13
5	0.0456	0.0533	73	0.0433	0.1307	62	0.0344	0.2690	28
6	0.0705	0.0144	91	0.0613	0.0516	74	0.0484	0.1235	50
7	0.0949	0.0013	100	0.0825	0.0237	88	0.0639	0.1026	66
8	0.1215	0.0008	100	0.1085	0.0103	96	0.0827	0.0549	73
9	0.1505	0.0003	100	0.1266	0.0033	97	0.1052	0.0315	88
50% of noise in the variance of the endogenous disturbance term									
1	0.0086	0.7051	1	0.0099	0.7079	2	0.0092	0.7374	0
2	0.0099	0.6543	1	0.0131	0.5899	2	0.0089	0.7157	0
3	0.0176	0.5007	10	0.0138	0.5913	6	0.0169	0.5922	6
4	0.0279	0.3383	23	0.0278	0.3470	22	0.0220	0.4691	8
5	0.0440	0.2002	46	0.0356	0.2564	29	0.0265	0.4143	16
6	0.0510	0.1365	53	0.0570	0.1574	46	0.0374	0.3411	25
7	0.0724	0.0497	74	0.0618	0.1291	59	0.0531	0.1758	40
8	0.0967	0.0130	93	0.0817	0.0486	80	0.0609	0.2053	43
9	0.1141	0.0080	98	0.0875	0.0580	77	0.0766	0.1079	58

Table 8.5. Simulation results for normal data and balanced nodes. Total sample size of 500 (250 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0013	0.6019	0	0.0022	0.6227	3	0.0023	0.6436	2
2	0.0054	0.1672	48	0.0050	0.4030	16	0.0050	0.4998	3
3	0.0168	0.0025	99	0.0164	0.0745	73	0.0133	0.2103	43
4	0.0368	0.0000	100	0.0315	0.0038	98	0.0241	0.0495	79
5	0.0628	0.0000	100	0.0520	0.0000	100	0.0382	0.0205	91
6	0.0892	0.0000	100	0.0742	0.0000	100	0.0534	0.0003	100
7	0.1213	0.0000	100	0.0981	0.0000	100	0.0751	0.0000	100
8	0.1499	0.0000	100	0.1240	0.0000	100	0.0924	0.0001	100
9	0.1834	0.0000	100	0.1501	0.0000	100	0.1151	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0023	0.6844	3	0.0025	0.7032	2	0.0032	0.7078	1
2	0.0048	0.4770	9	0.0047	0.5187	3	0.0057	0.5928	6
3	0.0150	0.1130	58	0.0130	0.2153	40	0.0115	0.3143	23
4	0.0282	0.0176	91	0.0243	0.0617	73	0.0187	0.1827	50
5	0.0514	0.0002	100	0.0414	0.0073	96	0.0311	0.0462	76
6	0.0725	0.0000	100	0.0563	0.0013	100	0.0431	0.0122	93
7	0.0960	0.0000	100	0.0819	0.0000	100	0.0612	0.0027	99
8	0.1217	0.0000	100	0.0944	0.0000	100	0.0737	0.0021	99
9	0.1475	0.0000	100	0.1182	0.0000	100	0.0815	0.0001	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0032	0.7295	1	0.0033	0.7231	1	0.0046	0.7100	2
2	0.0046	0.6481	4	0.0042	0.6721	2	0.0059	0.6096	2
3	0.0124	0.2638	27	0.0119	0.3840	19	0.0098	0.4823	11
4	0.0251	0.0657	74	0.0203	0.1683	50	0.0150	0.3004	23
5	0.0385	0.0256	87	0.0299	0.0664	66	0.0247	0.1436	53
6	0.0524	0.0028	99	0.0437	0.0339	90	0.0323	0.0821	68
7	0.0737	0.0003	100	0.0626	0.0020	100	0.0457	0.0192	88
8	0.0833	0.0000	100	0.0690	0.0016	100	0.0542	0.0141	90
9	0.1082	0.0000	100	0.0913	0.0001	100	0.0651	0.0100	97

Table 8.6. Simulation results for normal data and balanced nodes. Total sample size of 1000 (500 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0005	0.6397	2	0.0008	0.6823	0	0.0013	0.7141	0
2	0.0047	0.0484	83	0.0044	0.2022	41	0.0039	0.3854	15
3	0.0168	0.0000	100	0.0137	0.0033	98	0.0111	0.0693	66
4	0.0348	0.0000	100	0.0302	0.0000	100	0.0226	0.0040	98
5	0.0597	0.0000	100	0.0501	0.0000	100	0.0370	0.0002	100
6	0.0895	0.0000	100	0.0727	0.0000	100	0.0529	0.0000	100
7	0.1197	0.0000	100	0.0952	0.0000	100	0.0729	0.0000	100
8	0.1530	0.0000	100	0.1248	0.0000	100	0.0894	0.0000	100
9	0.1826	0.0000	100	0.1517	0.0000	100	0.1124	0.0000	100
10% of noise in the variance of the endogenous disturbance term									
1	0.0010	0.7186	2	0.0012	0.7102	1	0.0017	0.7015	4
2	0.0045	0.2593	34	0.0042	0.3573	18	0.0033	0.5088	9
3	0.0144	0.0087	95	0.0131	0.0510	83	0.0081	0.1987	40
4	0.0289	0.0000	100	0.0229	0.0034	98	0.0186	0.0399	84
5	0.0473	0.0000	100	0.0392	0.0000	100	0.0293	0.0044	98
6	0.0698	0.0000	100	0.0573	0.0000	100	0.0441	0.0000	100
7	0.0916	0.0000	100	0.0807	0.0000	100	0.0567	0.0000	100
8	0.1169	0.0000	100	0.0949	0.0000	100	0.0725	0.0000	100
9	0.1426	0.0000	100	0.1173	0.0000	100	0.0917	0.0000	100
10% of noise in the variance of the endogenous disturbance term									
1	0.0016	0.7178	0	0.0018	0.7036	1	0.0021	0.6995	1
2	0.0038	0.4816	12	0.0031	0.5932	6	0.0033	0.5965	5
3	0.0100	0.1361	48	0.0101	0.1654	50	0.0081	0.2853	29
4	0.0213	0.0153	95	0.0185	0.0412	80	0.0130	0.1239	56
5	0.0332	0.0012	100	0.0283	0.0046	97	0.0209	0.0473	80
6	0.0505	0.0000	100	0.0425	0.0000	100	0.0322	0.0099	95
7	0.0686	0.0000	100	0.0553	0.0000	100	0.0423	0.0012	100
8	0.0841	0.0000	100	0.0669	0.0000	100	0.0490	0.0001	100
9	0.1046	0.0000	100	0.0822	0.0000	100	0.0619	0.0001	100

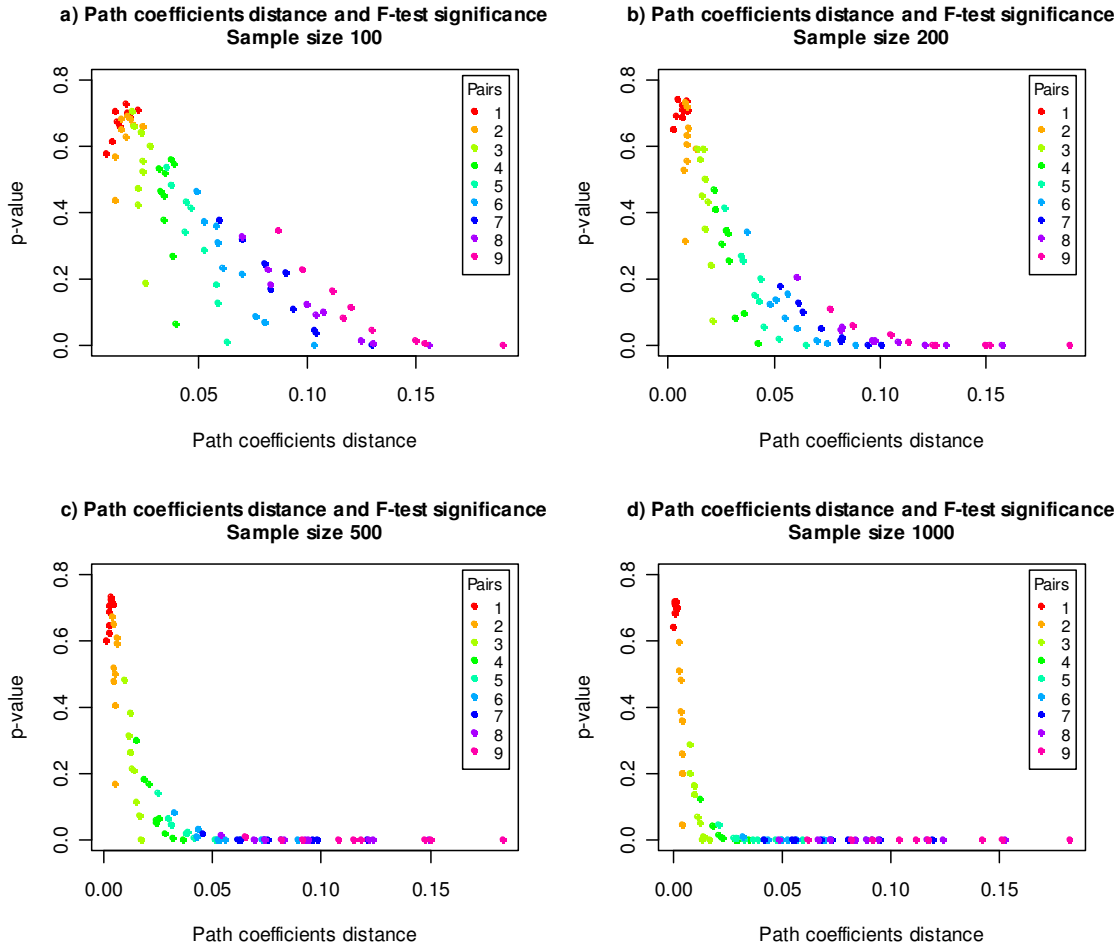


Figure 8.11. *Path coefficients distance and p-value (mean value).*

Charts in figure 8.11 help to see the general trends of tables 8.3 – 8.6. In all the carts, we are not making distinction between the different levels of noise. We only take into account the nine pairs of segments in each sub-block (identified with a different color). The first chart (a) corresponds to the results in table 8.3. It is possible to observe the general trend in which the larger the difference of path coefficients, the more significant the *p*-values. The second plot (b) corresponds to the results of table 8.4, whereas charts (c) and (d) correspond to tables 8.5 and 8.6, respectively. The overall impression is that the *F*-test works very well under conditions of normality. It has been seen how its performance can be affected by the level of noise in the variance of the endogenous and the manifest variables. However, the *F*-test has a good capability to detect different segments even when the level of noise in the endogenous and the manifest variables is relatively high.

The plots in figure 8.12 illustrate the general trends of tables 8.3 – 8.6 in a different manner than the charts in figure 8.11. The first plot (8.12.a) corresponds to table 8.3, the plot in (8.12.b) corresponds to table 8.4, (8.12.c) to table 8.5, and (8.12.d) to table 8.6. All the plots show bar charts (for the nine pair of segments) of the average proportion of *p*-values smaller than 0.05. For instance, in (8.12.a) we can appreciate nine different bar charts numbered from 1 to 9.

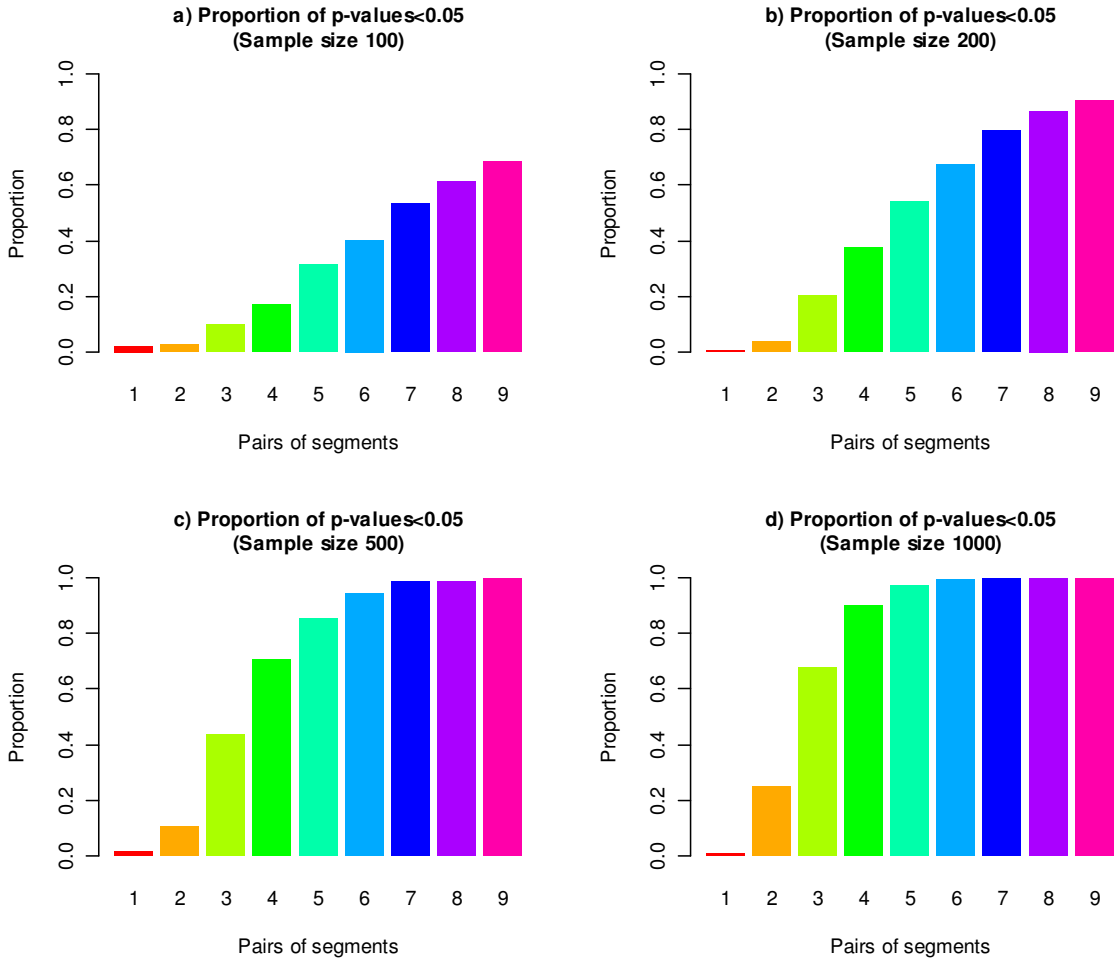


Figure 8.12. *Box-plots of the number of p-values>0.05 by pair of segments.*

The first chart is associated to the first pair of compared models in each experimental condition. Since the first pair of models have identical path coefficients, the average proportion of *p*-values<0.05 is relatively small. For this reason, the bar in this case is very tiny. In contrast, the bar associated to the ninth pair of models has a larger longitude since these models are supposed to have the most distinct path coefficients. Hence, there is a large number of times that the *p*-value is less than 0.05. It can be seen that the larger the sample size, and the more similar the path coefficients, the smaller the longitude of the bars. Conversely, the larger the sample size, and the more different the path coefficients, the bars become larger.

8.3.2 Unbalanced segments (distinct proportions)

A deeper evaluation of the capabilities of the *F*-test requires considering segments of different proportions. In this case, for the sake of simplicity we have only taken into account a total sample size of 200 with four mixing proportions. The first mixing proportion is of (0.40,0.60) which means that one segment has 40% of elements from the total sample size, whereas the other segment has 60% of the elements, that is: 80

elements in one segment, and 120 elements in the other segment. The other three mixing proportion are (0.30,0.70), (0.20,80), and (0.10,0.90).

We have used the same structural model as the one used in the previous simulations. In total, we have 4x3x3=36 scenarios which is the number of possible combinations of mixing proportions and noise levels (4 mixing proportions, 3 noise levels for the endogenous construct variance, and 3 noise levels for the indicators variance). For every possible scenario, we compare 9 pairs of inner models. In each comparison we run 100 repetitions, and from the 100 repetitions we calculate the average of the *F*-test's *p*-value, and the mean value of the distance of path coefficients between models.

In this case we obtain similar results to those obtained with balanced proportions. At the aggregate level one can see how the significance of the *F*-test is affected by the difference of path coefficients between inner models. However, it can also be seen that the more different the mixing proportions, the less sensitive is the test (see figure 8.13).

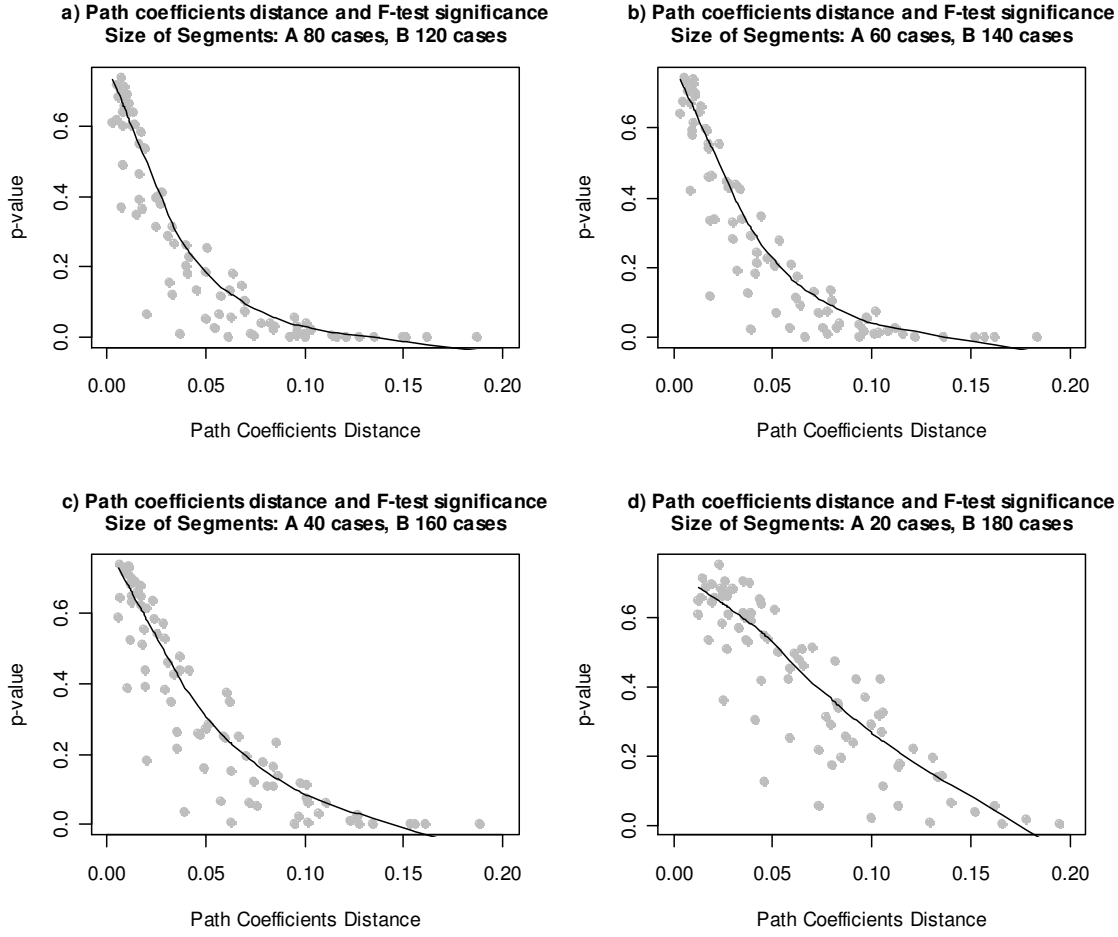


Figure 8.13. *Plots for each mixing proportion: distance of path coefficients and average p-values*

Summarized results (unbalanced segments)

The summarized results from the simulations with normal data for the first mixing proportion are contained in tables 8.7. The results of the other mixing proportions are contained in tables 8.8, 8.9 and 8.10. As in the balanced case, table 8.7 shows the summarized results from the 100 repetitions for each of the 9 pairs of compared path models with sample size of 200.

Table 8.7. Simulation results for normal data and unbalanced nodes. (Segment A 80 cases, Segment B 120 cases).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0031	0.6094	3	0.0051	0.6184	2	0.0054	0.7176	1
2	0.0076	0.3702	17	0.0085	0.4909	9	0.0085	0.6420	4
3	0.0201	0.0628	67	0.0153	0.3459	20	0.0180	0.3658	12
4	0.0370	0.0077	95	0.0330	0.1217	56	0.0250	0.3152	23
5	0.0611	0.0002	100	0.0547	0.0237	89	0.0407	0.1790	47
6	0.0926	0.0000	100	0.0742	0.0050	97	0.0566	0.0649	64
7	0.1165	0.0000	100	0.1006	0.0005	100	0.0851	0.0317	86
8	0.1514	0.0000	100	0.1354	0.0000	100	0.0962	0.0175	91
9	0.1869	0.0000	100	0.1622	0.0000	100	0.1142	0.0028	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0056	0.6849	0	0.0064	0.7098	0	0.0094	0.6759	1
2	0.0079	0.6026	3	0.0088	0.6568	2	0.0084	0.6876	1
3	0.0165	0.3905	13	0.0165	0.4646	8	0.0127	0.6007	3
4	0.0318	0.1544	49	0.0312	0.2892	29	0.0252	0.3954	18
5	0.0500	0.0515	76	0.0458	0.1332	60	0.0343	0.2661	23
6	0.0732	0.0074	98	0.0633	0.0540	79	0.0420	0.2265	41
7	0.0957	0.0026	99	0.0845	0.0214	87	0.0621	0.1346	54
8	0.1205	0.0015	99	0.1000	0.0065	96	0.0831	0.0400	80
9	0.1496	0.0000	100	0.1280	0.0008	100	0.0966	0.0291	87
50% of noise in the variance of the endogenous disturbance term									
1	0.0086	0.7118	2	0.0072	0.7403	0	0.0107	0.6901	0
2	0.0110	0.6661	5	0.0114	0.6417	0	0.0135	0.6420	3
3	0.0165	0.5490	6	0.0143	0.6036	1	0.0174	0.5845	4
4	0.0269	0.3757	18	0.0256	0.4000	12	0.0192	0.5361	7
5	0.0402	0.2018	34	0.0335	0.3115	19	0.0277	0.4123	14
6	0.0577	0.1152	57	0.0503	0.1823	43	0.0404	0.2611	25
7	0.0782	0.0393	77	0.0695	0.1015	61	0.0508	0.2522	28
8	0.1011	0.0390	85	0.0701	0.0737	77	0.0635	0.1817	45
9	0.1035	0.0216	92	0.0949	0.0575	76	0.0679	0.1438	56

Table 8.8. Simulation results for normal data and unbalanced nodes. (Segment A 60 cases, Segment B 140 cases).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0034	0.6405	3	0.0051	0.6738	0	0.0074	0.7050	1
2	0.0084	0.4186	13	0.0090	0.5818	4	0.0103	0.6163	3
3	0.0187	0.1167	55	0.0184	0.3330	21	0.0180	0.4572	11
4	0.0387	0.0200	92	0.0377	0.1261	54	0.0298	0.2835	26
5	0.0665	0.0003	100	0.0587	0.0242	84	0.0417	0.2139	36
6	0.0939	0.0000	100	0.0780	0.0088	95	0.0618	0.1129	65
7	0.1219	0.0000	100	0.1010	0.0069	96	0.0836	0.0396	82
8	0.1524	0.0000	100	0.1366	0.0005	100	0.0936	0.0355	79
9	0.1838	0.0000	100	0.1623	0.0000	100	0.1082	0.0166	88
30% of noise in the variance of the endogenous disturbance term									
1	0.0053	0.7467	0	0.0065	0.7304	1	0.0106	0.6937	1
2	0.0091	0.5943	4	0.0085	0.6721	2	0.0133	0.6435	2
3	0.0204	0.3376	18	0.0191	0.4641	12	0.0176	0.5540	5
4	0.0321	0.1911	43	0.0298	0.3296	24	0.0317	0.4376	19
5	0.0521	0.0690	73	0.0410	0.1824	35	0.0389	0.2892	23
6	0.0755	0.0270	87	0.0641	0.0926	68	0.0472	0.2232	33
7	0.0962	0.0154	94	0.0824	0.0265	87	0.0706	0.1293	55
8	0.1219	0.0014	100	0.1038	0.0145	94	0.0780	0.0728	61
9	0.1568	0.0002	100	0.1162	0.0086	95	0.0979	0.0550	74
50% of noise in the variance of the endogenous disturbance term									
1	0.0091	0.7268	0	0.0098	0.7391	2	0.0105	0.7009	1
2	0.0102	0.7067	2	0.0100	0.7298	0	0.0141	0.6615	2
3	0.0176	0.5427	6	0.0161	0.5980	3	0.0172	0.5942	1
4	0.0275	0.4295	14	0.0268	0.4475	11	0.0228	0.5543	6
5	0.0420	0.2407	31	0.0342	0.3376	16	0.0335	0.4228	12
6	0.0512	0.2017	43	0.0511	0.2089	41	0.0445	0.3461	18
7	0.0734	0.0677	67	0.0624	0.1713	45	0.0532	0.2765	27
8	0.0978	0.0506	82	0.0800	0.1025	65	0.0597	0.2070	39
9	0.1124	0.0259	89	0.1018	0.0747	72	0.0790	0.1356	46

Table 8.9. Simulation results for normal data and unbalanced nodes. (Segment A 40 cases, Segment B 160 cases).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0060	0.5865	7	0.0066	0.6428	1	0.0091	0.7136	0
2	0.0107	0.3843	14	0.0117	0.5237	7	0.0127	0.6441	2
3	0.0205	0.1821	48	0.0199	0.3890	12	0.0180	0.5086	3
4	0.0395	0.0357	80	0.0357	0.2141	41	0.0295	0.3814	10
5	0.0628	0.0054	97	0.0578	0.0647	76	0.0471	0.2513	33
6	0.0950	0.0006	100	0.0762	0.0515	80	0.0626	0.1516	46
7	0.1276	0.0000	100	0.1019	0.0052	97	0.0813	0.1095	57
8	0.1560	0.0000	100	0.1342	0.0017	99	0.1011	0.0753	71
9	0.1884	0.0000	100	0.1609	0.0009	100	0.1269	0.0279	86
30% of noise in the variance of the endogenous disturbance term									
1	0.0069	0.7347	1	0.0109	0.7042	2	0.0110	0.7304	0
2	0.0130	0.6293	3	0.0121	0.6807	2	0.0171	0.6193	4
3	0.0196	0.4378	11	0.0186	0.5517	4	0.0238	0.5830	3
4	0.0353	0.2609	28	0.0324	0.3474	17	0.0258	0.5379	7
5	0.0492	0.1615	49	0.0463	0.2575	28	0.0337	0.4259	13
6	0.0723	0.0607	70	0.0743	0.1205	60	0.0501	0.2701	21
7	0.0970	0.0233	88	0.0843	0.1092	62	0.0665	0.2485	33
8	0.1232	0.0092	97	0.1073	0.0298	83	0.0845	0.1635	45
9	0.1538	0.0015	99	0.1260	0.0177	88	0.0983	0.1153	53
50% of noise in the variance of the endogenous disturbance term									
1	0.0114	0.7289	0	0.0160	0.6587	1	0.0128	0.6980	0
2	0.0146	0.6902	4	0.0206	0.6137	1	0.0172	0.6755	0
3	0.0190	0.6104	3	0.0175	0.6444	1	0.0236	0.6334	4
4	0.0311	0.4565	12	0.0297	0.5250	6	0.0285	0.5693	9
5	0.0518	0.2835	26	0.0369	0.4371	11	0.0368	0.4754	7
6	0.0598	0.2464	28	0.0589	0.2483	27	0.0416	0.4367	9
7	0.0790	0.1757	47	0.0705	0.1923	38	0.0625	0.3460	20
8	0.1020	0.0594	72	0.0864	0.1368	48	0.0604	0.3726	19
9	0.1108	0.0631	72	0.1011	0.1147	55	0.0856	0.2330	34

Table 8.10. Simulation results for normal data and unbalanced nodes. (Segment A 20 cases, Segment B 180 cases).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0127	0.6074	5	0.0124	0.6478	1	0.0163	0.6846	2
2	0.0176	0.5357	10	0.0139	0.6543	0	0.0207	0.6562	1
3	0.0251	0.3572	23	0.0265	0.5080	13	0.0274	0.6092	1
4	0.0461	0.1222	51	0.0413	0.3024	17	0.0368	0.5353	6
5	0.0730	0.0532	81	0.0587	0.2506	34	0.0587	0.4517	15
6	0.0997	0.0214	87	0.0804	0.1722	52	0.0791	0.2910	26
7	0.1295	0.0048	99	0.1132	0.0535	79	0.0907	0.2374	30
8	0.1659	0.0006	100	0.1520	0.0371	84	0.1135	0.1698	41
9	0.1949	0.0004	100	0.1785	0.0147	91	0.1358	0.1406	48
30% of noise in the variance of the endogenous disturbance term									
1	0.0148	0.7133	1	0.0194	0.6958	1	0.0268	0.6594	2
2	0.0195	0.6446	2	0.0236	0.6836	6	0.0243	0.6596	1
3	0.0248	0.5814	2	0.0329	0.5678	5	0.0387	0.5917	4
4	0.0445	0.4168	15	0.0377	0.5288	6	0.0356	0.6000	1
5	0.0734	0.2136	37	0.0581	0.4205	18	0.0705	0.5112	8
6	0.0843	0.1922	50	0.0768	0.3121	25	0.0611	0.4930	9
7	0.1060	0.1100	63	0.0873	0.2527	34	0.0833	0.3497	18
8	0.1405	0.0636	70	0.1141	0.1744	46	0.1038	0.3159	24
9	0.1621	0.0540	85	0.1335	0.1352	45	0.0965	0.3695	20
50% of noise in the variance of the endogenous disturbance term									
1	0.0259	0.7022	1	0.0229	0.7507	0	0.0355	0.7050	2
2	0.0273	0.6722	3	0.0300	0.6804	3	0.0386	0.6999	2
3	0.0354	0.6135	3	0.0388	0.6098	8	0.0437	0.6512	5
4	0.0474	0.5390	6	0.0459	0.5450	5	0.0440	0.6380	0
5	0.0527	0.4978	10	0.0631	0.4775	10	0.0509	0.6193	3
6	0.0829	0.3360	25	0.0653	0.4592	5	0.0650	0.5079	7
7	0.1054	0.2695	31	0.0826	0.3500	20	0.0813	0.4738	10
8	0.1208	0.2205	38	0.1002	0.2878	23	0.0920	0.4224	15
9	0.1314	0.1941	41	0.1057	0.3263	24	0.1048	0.4183	13

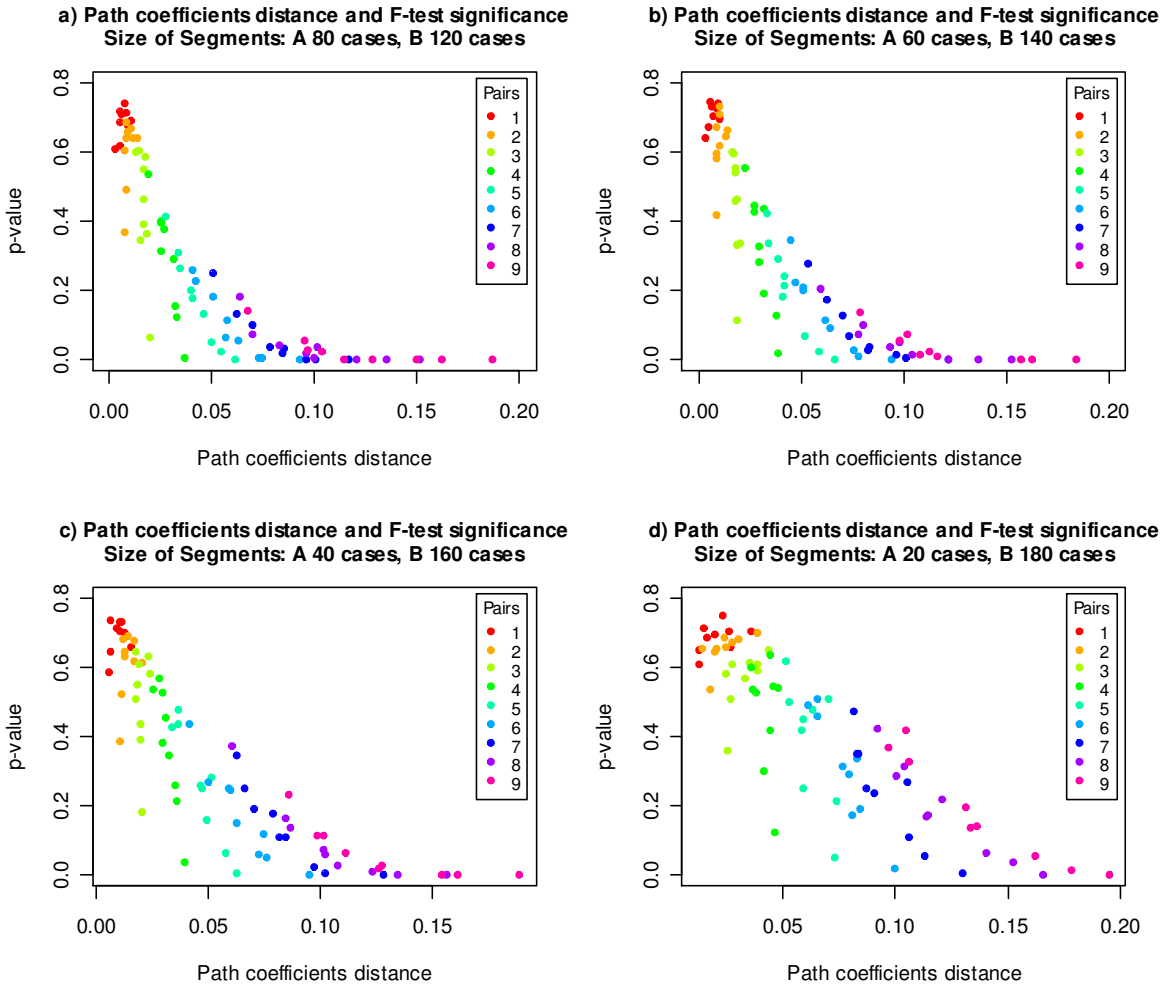


Figure 8.14. *Path coefficients distance and p-value*

Charts in figure 8.14 illustrate the trends of tables 8.7 – 8.10. No distinction between the different levels of noise is made. Only the nine pairs of segments in each sub-block (identified with a different color) are taken into account. The first chart (a) corresponds to the results in table 8.7. The second plot (b) corresponds to the results of table 8.8, whereas charts (c) and (d) correspond to tables 8.9 and 8.10, respectively. We can distinguish an overall pattern in which the larger the difference of path coefficients, the more significant the *p*-values. However, it is also possible to see the effect of comparing models of different sample sizes. Looking at figure 8.14.d) we see that the test still continues to detect different models although it is slightly affected in a negative manner. In conclusion, the *F*-test performs adequately even though it is affected when the compared segments have extremely different sample sizes.

8.4 Simulation comparing structural models with non normal data

The third type of simulation analysis is carried out with non-normal data. Non-normality is assumed in many studies and different values of skewness and kurtosis of the manifest variables are specified in order to control the degree of non-normality (Mattson, 1997). The interest in skewed distributions is due to the observed frequency distributions in empirical data such as the obtained from customer satisfaction research. This kind of data is usually measured with 7-point or 10-point scales, with values ranging from "completely disagree" to "completely agree", or from "very dissatisfied" to "very satisfied", for example. In these cases, virtually all the indicators are moderately skewed, which means that most of the respondents answer with high scores values (Fornell, 1992). In fact, there can be two classes of skewness: positive and negative. A positive skew implies that most respondents are dissatisfied. Conversely, negative skewness means that most respondents are satisfied. Empirical evidence and generalizations of skewness in frequency distributions of customer satisfaction studies can be found in Fornell (1995). For a more practical discussion (with non-technical aspects) of skewed distributions on customer satisfaction data see Fornell (2007).

In our case, the simulation study with non-normal data is based in the analysis carried out by Cassel *et al* (1999) who studied the robustness of the PLS-PM method when researchers are in presence of data with skewed distributions. They generated data following a beta distribution with different parameters in order to have three cases: (1) symmetric distributions, (2) moderately right-skewed distributions, and (3) highly right-skewed distributions. We use the same structural model as the one used in the simulations with normal data. The path diagram is shown below.

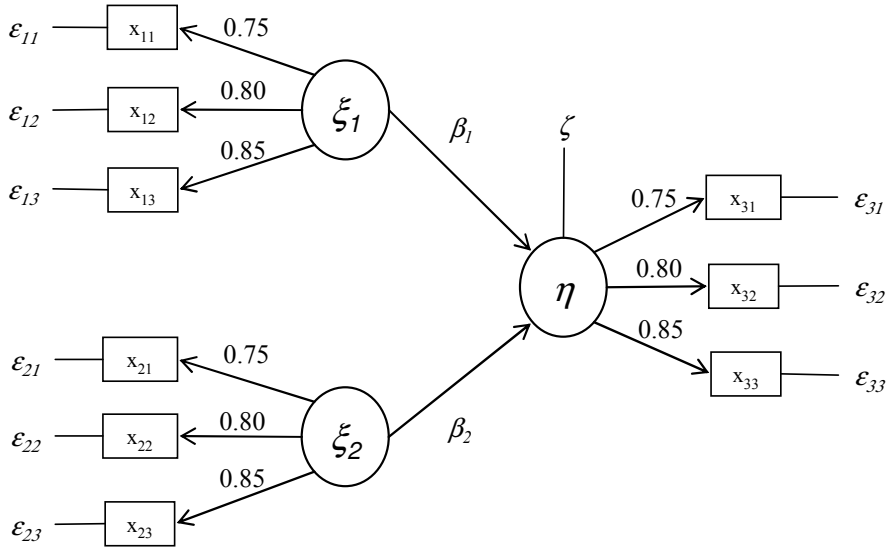


Figure 8.15 *Path Diagram of the structural model in the simulation study*

Unlike the data generating procedure carried out in the simulations with normal data, in the case of non-normality we generate the exogenous constructs ξ_1 and ξ_2 as realizations from a beta distribution $\beta(u,v)$. In order to take into account both symmetry and

skewness in distributions for the latent variables, three cases of parameters *u* and *v* for the beta distribution are considered:

- B(6,6) symmetric case
- B(9,4) moderately right-skewed case
- B(9,1) right-skewed case

Figure 8.16 shows an example of distribution for each of the three cases of skewness.

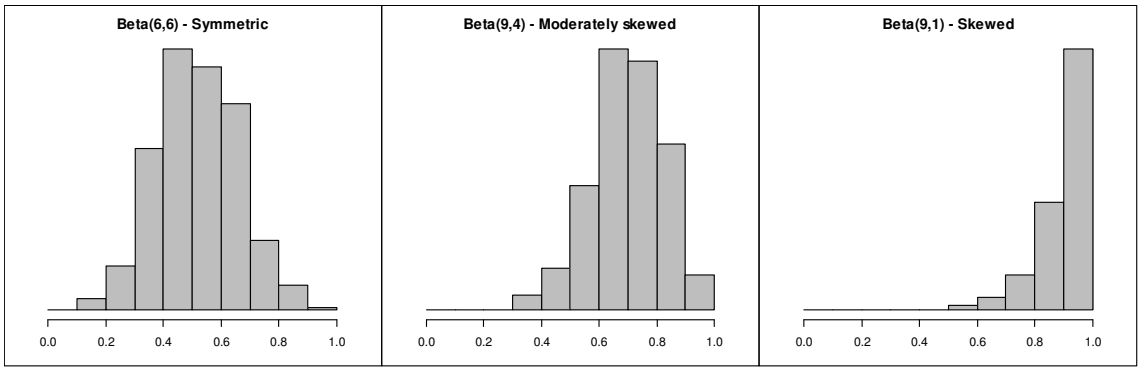


Figure 8.16. *Three types of skewness for Beta distributions*

With respect to the error term ζ, it follows a uniform distribution U(*a,b*) with expectation zero and variance having three levels. The levels are chosen such that the variance of ζ accounts for 10%, 30%, and 50% of the total variance of η. To obtain the required parameters *a* and *b* for the uniform distribution we begin by examining var(η):

$$\text{var}(\eta) = \text{var}(\beta_1\xi_1 + \beta_2\xi_2 + \zeta) \tag{8.18}$$

Assuming independence between latent variables and between the error term ζ we have:

$$\text{var}(\eta) = \beta_1^2\ \text{var}(\xi_1) + \beta_2^2\ \text{var}(\xi_2) + \text{var}(\zeta) \tag{8.19}$$

In the case that var(ζ) accounts for *q*% of the variance of η we have

$$\text{var}(\zeta) = (q/(1\text{-}q))\ (\beta_1^2\ \text{var}(\xi_1) + \beta_2^2\ \text{var}(\xi_2)) \tag{8.20}$$

The problem is to find values for the parameters *a* and *b* so that var(ζ) accounts for *q*% of the variance of η. Since ζ has expectation zero, we have that $b = -a$. Substituting *a* by $-b$ we have:

$$\text{var}(\zeta) = \frac{(b-a)^2}{12} = \frac{(b-(-b))^2}{12} = \frac{(2b)^2}{12} = \frac{b^2}{3} \tag{8.21}$$

In the case that var(ζ) accounts for *q*% of the variance of η, the desired parameter *b* is

$$b = \sqrt{\frac{q}{(1-q)} 3 \sum_{j=1}^{2} \beta_j^2\ \text{var}(\xi_j)} \tag{8.22}$$

For instance, if ξ_1 and ξ_2 follow a beta distribution β(6,6); $\beta_1 = \beta_2 = 0.5$; and q=0.10, we can calculate b and a as

$$b = \sqrt{\frac{0.10}{0.90} \times 3 \times \left(0.5^2 \times \frac{36}{1872} + 0.5^2 \times \frac{36}{1872} \right)} = 0.0566 \quad \text{and} \quad a = -0.0566 \quad (8.23)$$

Thus, ζ follows a uniform distribution U(-0.0566, 0-0566) with expectation zero and variance 0.00106.

Indicator loadings are specified with $\lambda_{j1} = 0.75$, $\lambda_{j2} = 0.80$, and $\lambda_{j3} = 0.85$, for j=1, 2, 3. The error terms ε_{ji} for the manifest variables are also uniformly distributed U(a,b) with expectation zero and variance having three levels. As in the case for the error term ζ, the levels of variance for ε_{ji} are chosen such that the variance of ε_{ji} accounts for 10%, 30%, and 50% of the total variance of x_{ji}. In order to obtain var(ε_{ji}) we must examine the variances of the manifest variables as follows:

$$\text{var}(x_{ji}) = \text{var}(\lambda_{ji}\xi_j + \varepsilon_{ji}) \qquad \text{for } i = 1,2,3 \quad \text{and } j = 1,2 \qquad (8.24)$$

$$\text{var}(x_{3i}) = \text{var}(\lambda_{3i}\eta + \varepsilon_{3i}) \qquad \text{for } i = 1,2,3 \qquad (8.25)$$

Assuming independence between the latent variables and the error term we have:

$$\text{var}(x_{ji}) = \lambda_{ji}^2 \text{ var}(\xi_j) + \text{var}(\varepsilon_{ji}) \qquad \text{for } i = 1,2,3 \text{ and } j = 1, 2), \qquad (8.26)$$

and

$$\text{var}(x_{3i}) = \lambda_{3i}^2 \text{ var}(\eta) + \text{var}(\varepsilon_{3i}) \qquad \text{for } i = 1,2,3 \qquad (8.27)$$

In the case that var(ε_{ji}) accounts for q% of the variance of x_{ji} we have

$$\text{var}(\varepsilon_{ji}) = (q/(1-q)) \, (\lambda_{ji}^2 \text{ var}(\xi_j)) \qquad (8.28)$$

Since ε_{ji} is uniformly distributed U(a,b) its variance is:

$$\text{var}(\varepsilon_{ji}) = \frac{(b-a)^2}{12} \qquad (8.29)$$

Thus, if var(ε_{ji}) accounts for q% of the variance of x_{ji}, the desired parameters b and a are:

$$b = \sqrt{\frac{q}{(1-q)} 3\lambda_{ji}^2 \text{ var}(\xi_j)} \qquad \text{and} \qquad a = -b \qquad (8.30)$$

Values for the path coefficients for each pair of segments are considered as in the simulations with normal data, that is, the first segment keeps fixed with values for its path coefficients of $\beta_1 = \beta_2 = 0.5$. The second segment will vary its path coefficients, starting with identical values to the first segment (i.e. $\beta_1 = \beta_2 = 0.5$), and then being gradually modified in order to increase the difference between the path coefficients of the first segment. The list of nine different path coefficients for the second segment is in table 8.11. We consider four sample sizes as the total number of cases: {100, 200, 500 and 1000}. This implies having segments of sizes {50, 100, 250 and 500}.

Table 8.11. List of path coefficients for segments *A* and *B*

Num	*Segment A*		*Segment B*	
1	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.50$	$\beta_2 = 0.50$
2	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.55$	$\beta_2 = 0.45$
3	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.60$	$\beta_2 = 0.40$
4	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.65$	$\beta_2 = 0.35$
5	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.70$	$\beta_2 = 0.30$
6	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.75$	$\beta_2 = 0.25$
7	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.80$	$\beta_2 = 0.20$
8	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.85$	$\beta_2 = 0.15$
9	$\beta_1 = 0.50$	$\beta_2 = 0.50$	$\beta_1 = 0.90$	$\beta_2 = 0.10$

In total, we have 4x3x3x3=108 scenarios which is the number of possible combinations of sample sizes, beta distributions, and noise levels (4 sample sizes, 3 beta distributions, 3 noise levels for the endogenous construct variance, and 3 noise levels for the indicators variance).

For every possible scenario, we compare 9 pairs of inner models which correspond to the nine sets of path coefficients. For each set of path coefficients, we run 100 repetitions in which we generated two segments or child nodes. One segment has fixed path coefficients of $\beta_1 = \beta_2 = 0.5$. The other segment varies its path coefficients according to the nine sets of values. In each repetition we apply the *F*-test to see whether the path coefficients are identical.

As mentioned before, we have employed the statistical software R for the computational aspects. In order to simulate the required distributions we have used the pseudo-random generator functions ‘rbeta’ and ‘runif’. We generate the exogenous variables with the function ‘rbeta’. The error terms for the endogenous and the manifest variables are generated with the function ‘runif’.

In order to visualize the different types of skewness in the generated data we present three figures. Figure 8.17 shows an example of the variables’ distributions when the exogenous variables follow a distribution Beta(6,6). Figure 8.18 represents the case in which exogenous variables follow a distribution Beta(9,4). Finally, figure 8.19 shows an example for exogenous variables following a distribution Beta(9,1).

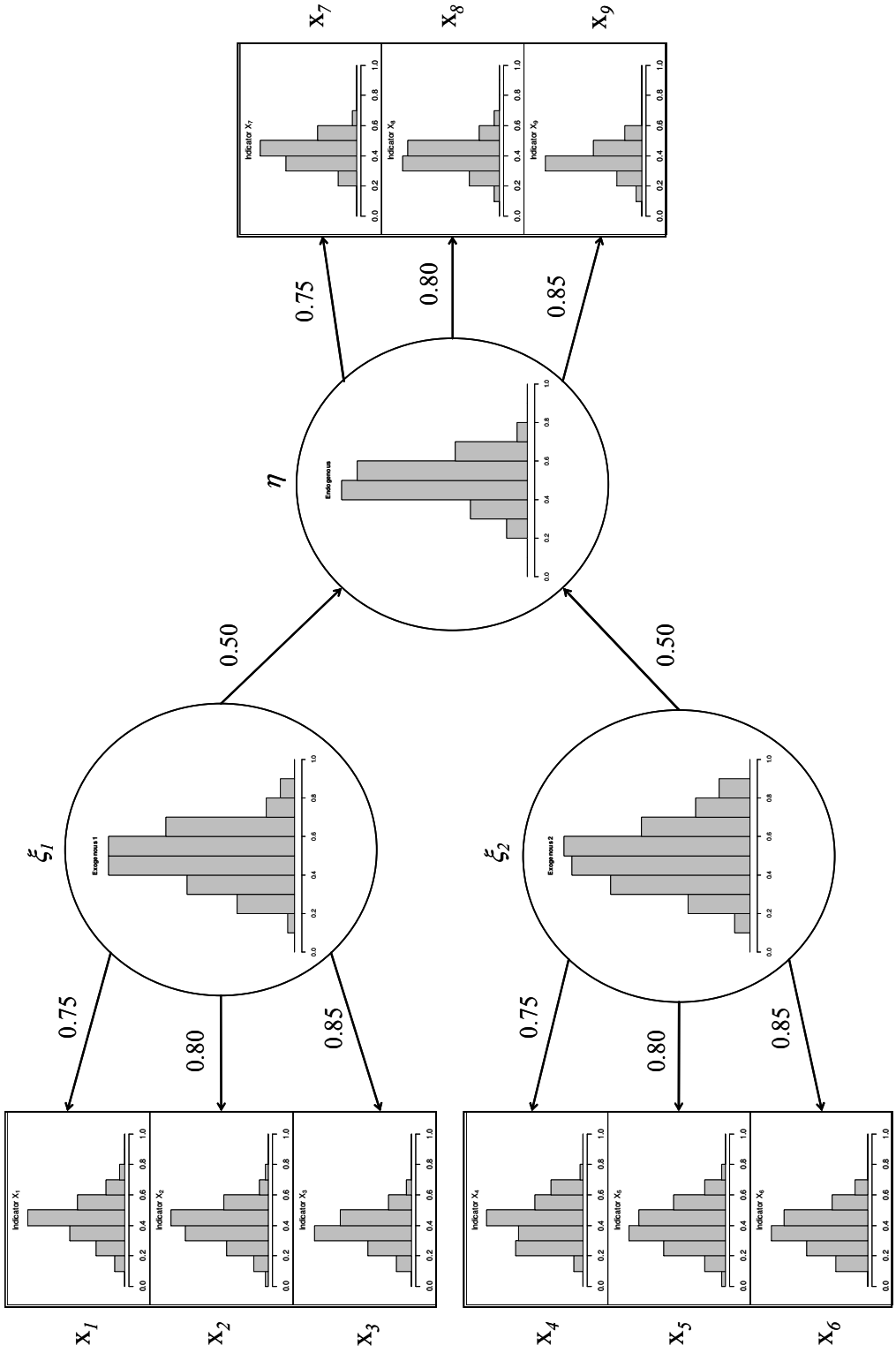


Figure 8.17. *An example of the variables distribution. Exogenous constructs generated with a Beta(6,6), low level of noises uniformly distributed, sample size of 100 cases.*

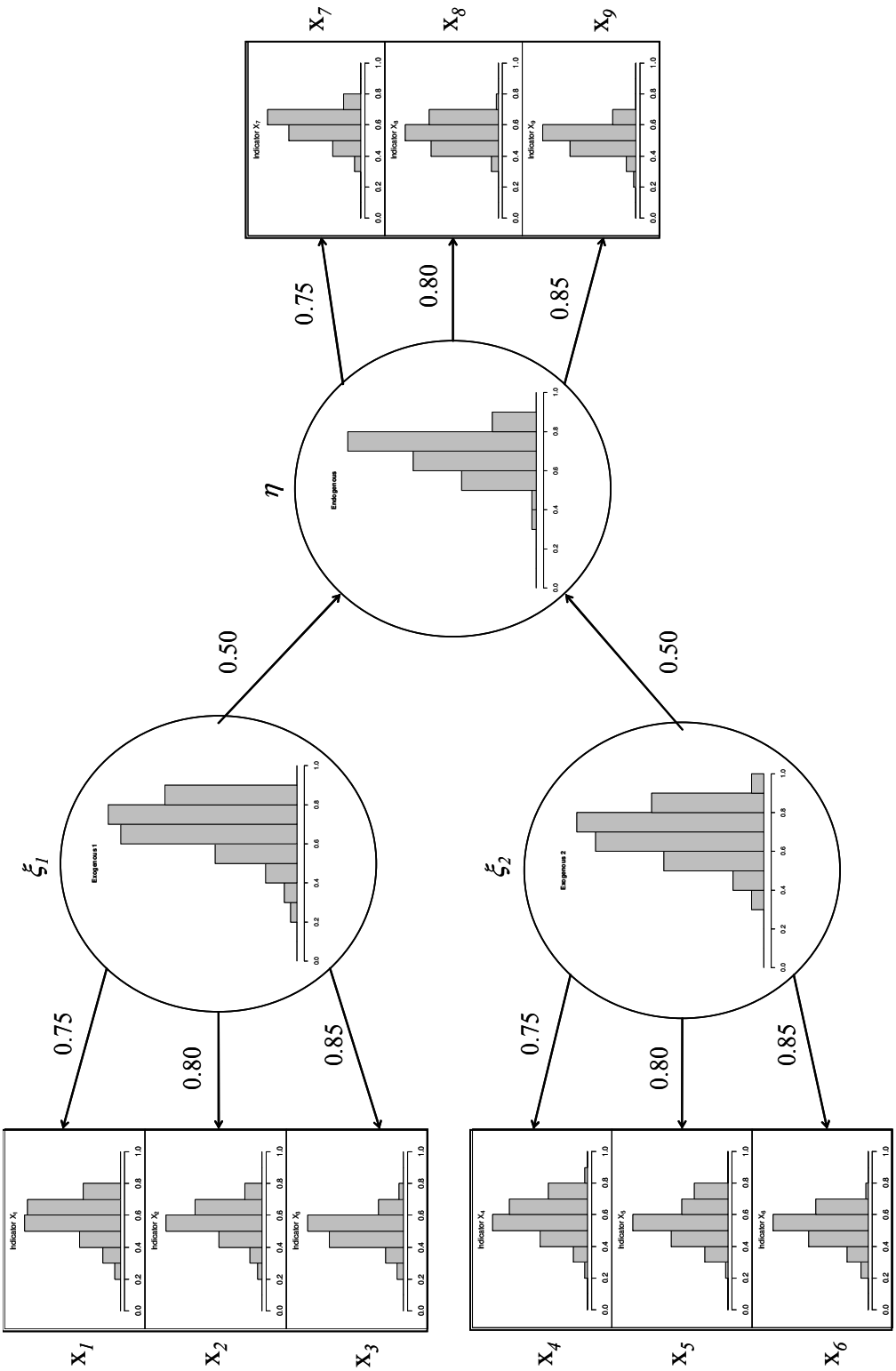


Figure 8.18. *An example of the variables distribution. Exogenous constructs generated with a Beta(9,4), low level of noises uniformly distributed, sample size of 100 cases.*

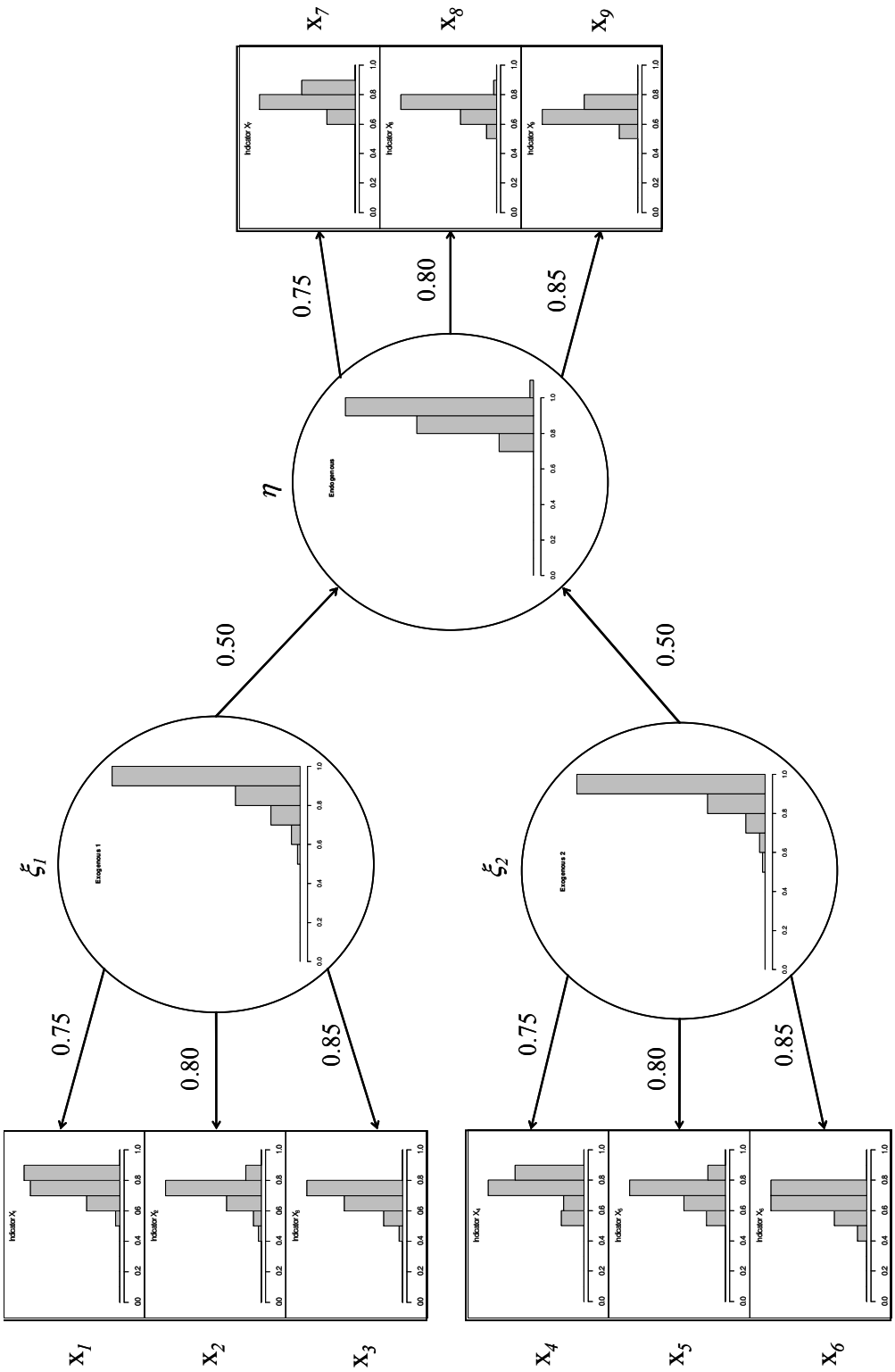


Figure 8.19. *An example of the variables distribution. Exogenous constructs generated with a Beta(9,1), low level of noises uniformly distributed, sample size of 100 cases.*

8.4.1 Results for the symmetric case

Similar results to those of the simulation studies with normal data are obtained in the case of symmetric beta distribution. On the aggregate data level we can appreciate how significance of the *F*-test is affected by the difference of path coefficients between inner path models, by different sample sizes (100, 200, 500 and 1000), by different levels in error variance terms of the endogenous construct, and by different levels in error variance terms for indicators.
The four trends detected in the normal case that affect the sensitivity of the *F*-test are observed again:

- The more different the path coefficients between segments, the more sensitive is the test
- The larger the sample size, the more sensitive
- The larger the level of noise (error variance) of the endogenous construct, the less sensitive
- The larger the level of noise (error variance) for indicators, the less sensitive.

Figure 8.20 illustrates these four trends

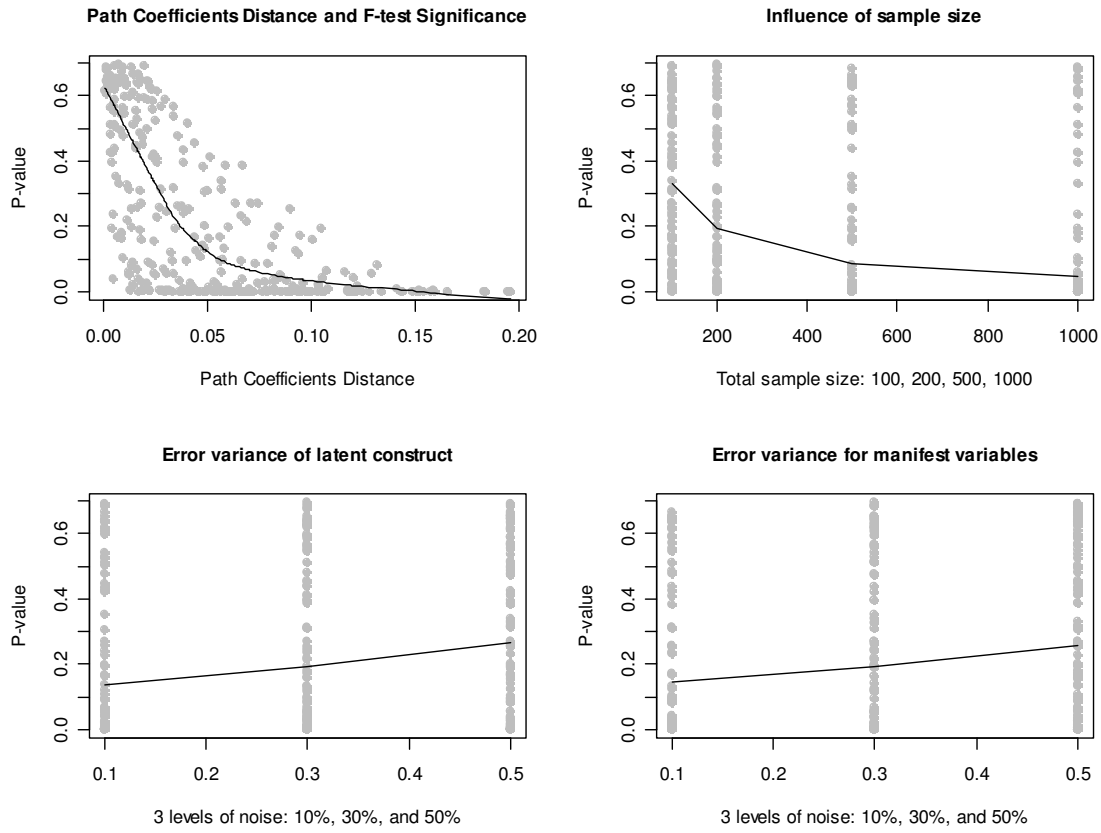


Figure 8.20. *Influence of different data generating conditions on the significance of the F-test at the aggregated level (Non-normal symmetric distributions)*

8.4.2 Summarized results: tables and charts in the symmetric case

The summarized results from the simulations with symmetric non-normal data are contained in the tables 8.12, 8.13, 8.14, and 8.15. As in the tables for normal data, each table corresponds to a different sample size. Each table is divided in three horizontal blocks. Each horizontal block, in turn, is divided in three sub-blocks. The horizontal blocks refer to the different levels of variance in the endogenous disturbance term. The sub-blocks refer to the levels of variance in the manifest variables. In each sub-block are displayed the following results: (1) the average distance of path coefficients, (2) the average value of the *p*-values, and (3) the number of *p*-values less than 0.05.

The examination of the sub-blocks is made for the four tables. The same general trends detected for the normal data are detected for the symmetric non-normal distributions. First, the average distance of the path coefficients between segments becomes larger. Second, the average *p*-value goes from high values to small values. Third, the number of times that the *p*-values are less than 0.05 becomes larger. In addition, the patterns due to the increment of the variance in the endogenous and the manifest variables are observed. In the one hand, when the segments are similar or their difference is relatively small (pairs 1,2,3 and 4), the average distance of path coefficients is larger compared to those cases with low levels of noises. In the other hand, when the differences between segments is large (pairs 5 to 9) the average distance of path coefficients is smaller compared to those cases with low levels of noise.

The charts in figures 8.21 help to see the general trends of tables 8.12 - 8.15. In these charts we do not distinguishing the different levels of noise. We only take into account the nine pairs of segments in each sub-block, each one marked with a different color. In each plot it is possible to observe the general trend in which the larger the difference of path coefficients, the more significant the *p*-values. Figure 8.22 shows the bar charts (for the nine pairs of segments) of the proportion of *p*-values which are smaller than 0.05. The chart in figure 8.22.a corresponds to the results of table 8.12. The rest of the bar plots (figures 8.22.b – 8.22.d) correspond to the results of tables 8.13 – 8.15.

Once again, the overall impression is that the *F*-test works reasonably well under conditions of non-normality and symmetry distributions. Although the *F*-test is influenced by the level of noise in the endogenous and the manifest variables, it has a good capability to detect different segments.

Table 8.12. Simulation results for data with symmetric non-normal distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0105	0.6499	1	0.0108	0.6404	1	0.0145	0.666	2
2	0.0147	0.3140	19	0.0162	0.5269	7	0.0171	0.6368	3
3	0.0270	0.0716	74	0.0243	0.4212	7	0.0251	0.5259	6
4	0.0404	0.0056	97	0.0386	0.2421	28	0.0385	0.4319	10
5	0.0642	0.0001	100	0.0511	0.1633	45	0.0489	0.3056	20
6	0.0961	0.0000	100	0.0855	0.0515	75	0.0605	0.196	32
7	0.1341	0.0000	100	0.1033	0.0216	85	0.0811	0.1386	50
8	0.1655	0.0000	100	0.1226	0.0149	89	0.1073	0.0653	73
9	0.1958	0.0000	100	0.1586	0.0025	98	0.1206	0.0444	79
30% of noise in the variance of the endogenous disturbance term									
1	0.0166	0.6422	1	0.0138	0.6865	1	0.0227	0.6252	2
2	0.0170	0.5938	7	0.0200	0.5931	1	0.0235	0.6244	4
3	0.0257	0.4058	12	0.0238	0.5571	9	0.0238	0.5828	4
4	0.0413	0.2541	27	0.0387	0.3955	15	0.0281	0.566	7
5	0.0582	0.1324	48	0.0501	0.3105	24	0.0438	0.4553	11
6	0.0790	0.0857	62	0.0693	0.215	35	0.0590	0.3872	21
7	0.1053	0.0414	80	0.0942	0.1153	59	0.0710	0.2711	32
8	0.1411	0.0091	97	0.1167	0.0616	75	0.0945	0.173	38
9	0.1525	0.0028	98	0.1294	0.0559	77	0.1006	0.1597	41
50% of noise in the variance of the endogenous disturbance term									
1	0.0180	0.6532	4	0.0215	0.6423	2	0.0196	0.6921	1
2	0.0172	0.6657	1	0.0229	0.6167	2	0.0217	0.6434	2
3	0.0213	0.5906	2	0.0261	0.6135	6	0.0298	0.5895	6
4	0.0485	0.3803	16	0.0406	0.5159	13	0.0340	0.5657	3
5	0.0570	0.3138	23	0.0359	0.4765	7	0.0340	0.5385	8
6	0.0664	0.2324	29	0.0569	0.3408	21	0.0518	0.4124	15
7	0.0826	0.1721	49	0.0640	0.2694	28	0.0671	0.3868	15
8	0.0930	0.1242	51	0.0845	0.1969	39	0.0742	0.2723	25
9	0.1317	0.0823	71	0.1046	0.1934	48	0.0900	0.2539	32

Table 8.13. Simulation results for data with symmetric non-normal distributions. Total sample size of 200 (100 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	p-value (mean)	Number of p-values<0.05	Path Coeffs Distance	p-value (mean)	Number of p-values<0.05	Path Coeffs Distance	p-value (mean)	Number of p-values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0053	0.6171	1	0.0061	0.6528	2	0.0056	0.6920	1
2	0.0096	0.4746	8	0.0089	0.5400	11	0.0105	0.5968	0
3	0.0217	0.0891	65	0.0186	0.2688	31	0.0157	0.4475	11
4	0.0388	0.0068	96	0.0324	0.1194	55	0.0260	0.2281	36
5	0.0662	0.0001	100	0.0547	0.0188	88	0.0466	0.1073	62
6	0.0851	0.0000	100	0.0784	0.0047	99	0.0569	0.0606	69
7	0.1216	0.0000	100	0.1046	0.0002	100	0.0747	0.0254	88
8	0.1597	0.0000	100	0.1278	0.0001	100	0.0932	0.0095	94
9	0.1944	0.0000	100	0.1575	0.0000	100	0.1193	0.0042	98
30% of noise in the variance of the endogenous disturbance term									
1	0.0060	0.6579	1	0.0076	0.6951	3	0.0091	0.6765	2
2	0.0094	0.5854	4	0.0100	0.5634	3	0.0122	0.5969	3
3	0.0202	0.3641	17	0.0157	0.4391	14	0.0189	0.4527	9
4	0.0364	0.1322	59	0.0309	0.2454	38	0.0236	0.4119	19
5	0.0580	0.0226	91	0.0447	0.1224	58	0.0344	0.1909	35
6	0.0791	0.0095	95	0.0674	0.0393	80	0.0486	0.1318	51
7	0.0989	0.0032	98	0.0819	0.0106	96	0.0575	0.0996	61
8	0.1280	0.0001	100	0.1037	0.0023	99	0.0820	0.0346	81
9	0.1534	0.0000	100	0.1178	0.0044	97	0.0964	0.0250	92
50% of noise in the variance of the endogenous disturbance term									
1	0.0083	0.6549	1	0.0103	0.6530	1	0.0100	0.6856	0
2	0.0157	0.5525	5	0.0111	0.6306	3	0.0107	0.6353	1
3	0.0179	0.4856	14	0.0191	0.4704	8	0.0141	0.6090	2
4	0.0280	0.3152	26	0.0262	0.3942	23	0.0209	0.4969	10
5	0.0440	0.1688	52	0.0360	0.2625	33	0.0343	0.3173	20
6	0.0564	0.0926	66	0.0482	0.1429	46	0.0381	0.2600	28
7	0.0776	0.0282	81	0.0664	0.0945	59	0.0486	0.2016	36
8	0.0915	0.0150	94	0.0842	0.0381	83	0.0566	0.1658	48
9	0.1083	0.0088	96	0.0880	0.0338	86	0.0729	0.0848	62

Table 8.14. Simulation results for data with symmetric non-normal distributions. Total sample size of 500 (250 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0020	0.6341	1	0.0020	0.6827	0	0.0025	0.6395	2
2	0.0067	0.1683	47	0.0062	0.3526	17	0.0063	0.5039	6
3	0.0189	0.0024	99	0.0165	0.0700	71	0.0140	0.1960	41
4	0.0382	0.0000	100	0.0273	0.0024	99	0.0248	0.0555	77
5	0.0619	0.0000	100	0.0495	0.0000	100	0.0390	0.0062	96
6	0.0939	0.0000	100	0.0691	0.0000	100	0.0543	0.0010	99
7	0.1207	0.0000	100	0.0983	0.0000	100	0.0713	0.0004	100
8	0.1525	0.0000	100	0.1243	0.0000	100	0.0936	0.0000	100
9	0.1839	0.0000	100	0.1490	0.0000	100	0.1050	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0033	0.6359	1	0.0035	0.6257	4	0.0037	0.6519	1
2	0.0055	0.4380	14	0.0053	0.5130	12	0.0057	0.5501	3
3	0.0140	0.0792	66	0.0118	0.2179	37	0.0122	0.3148	17
4	0.0306	0.0037	99	0.0252	0.0493	83	0.0186	0.1576	42
5	0.0463	0.0002	100	0.0421	0.0076	97	0.0296	0.0680	73
6	0.0684	0.0000	100	0.0598	0.0003	100	0.0432	0.0142	93
7	0.0952	0.0000	100	0.0785	0.0000	100	0.0570	0.0029	98
8	0.1221	0.0000	100	0.1002	0.0000	100	0.0784	0.0001	100
9	0.1442	0.0000	100	0.1187	0.0000	100	0.0828	0.0001	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0038	0.6423	3	0.0038	0.6497	0	0.0039	0.6622	2
2	0.0051	0.5695	4	0.0051	0.5671	2	0.0054	0.5894	2
3	0.0139	0.2544	32	0.0115	0.3283	19	0.0081	0.4928	6
4	0.0242	0.0794	69	0.0175	0.1871	44	0.0164	0.2592	29
5	0.0370	0.0338	89	0.0304	0.0543	74	0.0244	0.1521	51
6	0.0548	0.0021	99	0.0419	0.0154	92	0.034	0.0555	76
7	0.0710	0.0003	100	0.0593	0.0048	96	0.0421	0.0252	82
8	0.0891	0.0000	100	0.0658	0.0010	100	0.0506	0.0133	95
9	0.1029	0.0000	100	0.0883	0.0014	98	0.0704	0.0022	99

Table 8.15. Simulation results for data with symmetric non-normal distributions. Total sample size of 1000 (500 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0011	0.6163	1	0.0013	0.6056	3	0.0013	0.6892	0
2	0.0052	0.0377	82	0.0041	0.1916	40	0.0036	0.4249	13
3	0.0173	0.0000	100	0.0133	0.0040	98	0.0123	0.0595	72
4	0.0368	0.0000	100	0.0298	0.0000	100	0.0225	0.0052	98
5	0.0588	0.0000	100	0.0493	0.0000	100	0.0379	0.0000	100
6	0.0907	0.0000	100	0.0766	0.0000	100	0.0529	0.0000	100
7	0.1210	0.0000	100	0.0952	0.0000	100	0.0706	0.0000	100
8	0.1511	0.0000	100	0.1213	0.0000	100	0.0943	0.0000	100
9	0.1831	0.0000	100	0.1545	0.0000	100	0.1064	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0014	0.6435	2	0.0019	0.6193	2	0.0018	0.6770	2
2	0.0050	0.2368	40	0.0040	0.3935	19	0.0037	0.4820	10
3	0.0146	0.0194	93	0.0125	0.0501	76	0.0091	0.1793	44
4	0.0298	0.0000	100	0.0221	0.0018	99	0.0183	0.0299	89
5	0.0453	0.0000	100	0.0408	0.0000	100	0.0291	0.0012	99
6	0.0679	0.0000	100	0.0609	0.0000	100	0.0418	0.0001	100
7	0.0941	0.0000	100	0.0790	0.0000	100	0.0561	0.0000	100
8	0.1180	0.0000	100	0.0993	0.0000	100	0.0722	0.0000	100
9	0.1434	0.0000	100	0.1182	0.0000	100	0.0885	0.0000	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0021	0.6137	1	0.0020	0.6833	2	0.0021	0.6568	1
2	0.0042	0.4230	8	0.0038	0.5090	13	0.0035	0.5624	8
3	0.0097	0.1014	61	0.0089	0.1633	37	0.0075	0.3314	21
4	0.0209	0.0112	92	0.0169	0.0379	82	0.0143	0.0902	63
5	0.0336	0.0010	100	0.0284	0.0051	98	0.0215	0.0240	85
6	0.0480	0.0000	100	0.0421	0.0013	99	0.0320	0.0037	98
7	0.0679	0.0000	100	0.0580	0.0000	100	0.0402	0.0017	99
8	0.0828	0.0000	100	0.0705	0.0000	100	0.0529	0.0006	99
9	0.1053	0.0000	100	0.0861	0.0000	100	0.0633	0.0000	100

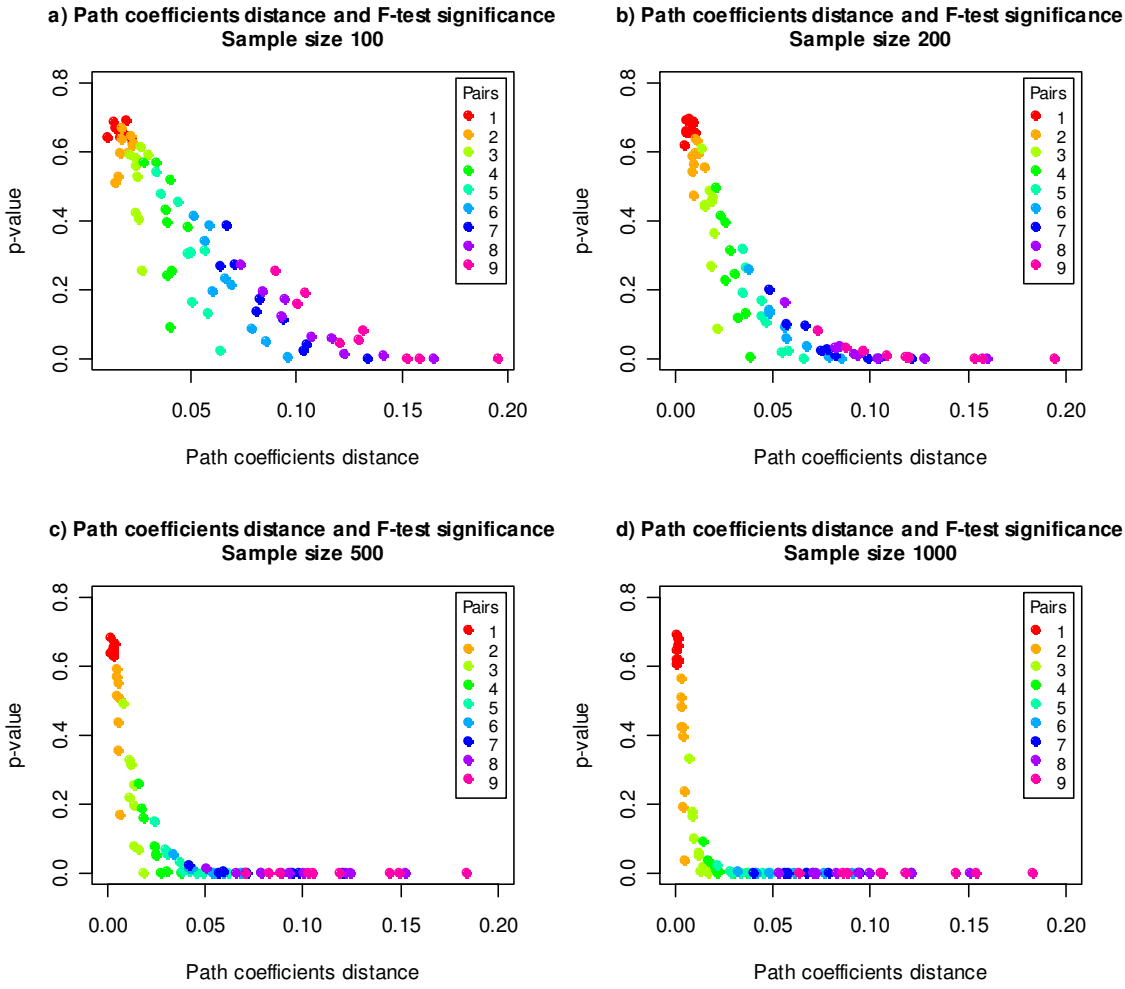


Figure 8.21. *Path coefficients distance and p-value (mean)*
Symmetric non-normal data with distribution B(6,6)

Figure 8.22 contain bar charts (for the nine pair of segments) of the average proportion of *p*-vlaues<0.05. For instance, the first bar in figure 8.22.a is associated to the first pair of compared models in each experimental condition. Since the first pair of models have identical path coefficients, the average proportion of *p*-values<0.05 is very small. Consequently, the size of the bar is very small. Likewise, as we move along the horizontal axis, the size of the bars increases as the pairs of segments become more dissimilar.

Another important aspect is the effect of the sample size which is also reflected through the four plots in figure 8.22. As the number of observations in each model increases, the length of the bars becomes larger. In summary, the general feeling is that the *F*-test works reasonably well under conditions of non-normality and symmetry distributions. Although the *F*-test is influenced by the level of noise in the endogenous and the manifest variables, it has a good capability to detect different segments.

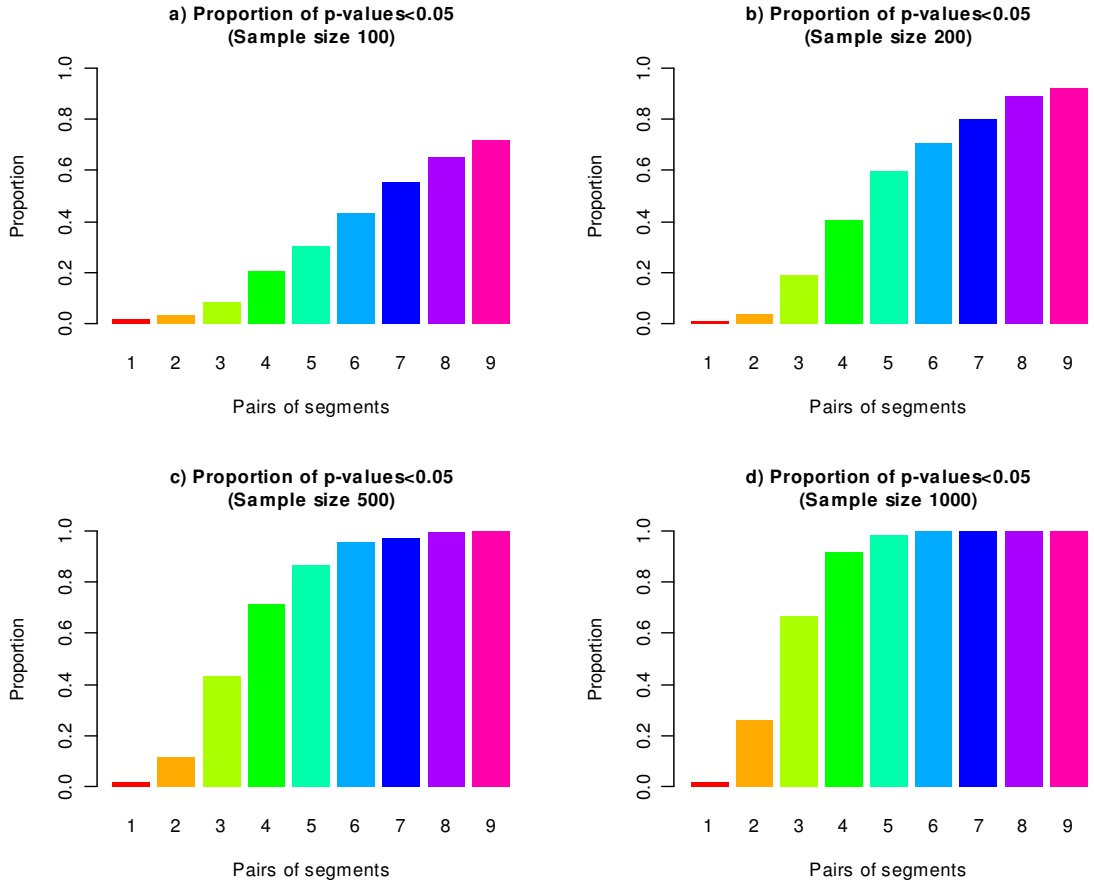


Figure 8.22. *Path coefficients distance and p-value (mean) Symmetric non-normal data with distribution B(6,6)*

8.4.3 Results for the moderately right-skewed case

The summarized results from the simulations with moderately right-skewed case are contained in the tables 8.16, 8.17, 8.18, and 8.19. As in the tables for normal data, each table corresponds to a different sample size. Each table is divided in three horizontal blocks. Each horizontal block, in turn, is divided in three sub-blocks. In each sub-block are displayed the following results: (1) the average distance of path coefficients, (2) the average value of the *p*-values, and (3) the number of *p*-values less than 0.05.

The examination of the sub-blocks is made for the four tables. The same general trends detected for the normal data and the non-normal symmetric data are detected for the moderately right-skewed distributions. First, the average distance of the path coefficients between segments becomes larger. Second, the average *p*-value goes from high values to small values. Third, the number of times that the *p*-values are less than 0.05 becomes larger. The patterns due to the increment of the variance in the endogenous and the manifest variables are also observed: (1) when the segments are similar or their difference is relatively small (pairs 1,2,3 and 4), the average distance of path coefficients is larger compared to those cases with low levels of noises; (2) when the differences between segments is large (pairs 5 to 9) the average distance of path coefficients is smaller compared to those cases with low levels of noise. Figure 8.23 help to see the general trends of tables 8.16 - 8.19.

Table 8.16. Simulation results for data with moderately skewed distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0110	0.6288	2	0.0146	0.5956	4	0.0156	0.6724	2
2	0.0157	0.4175	8	0.0180	0.5393	2	0.0175	0.6516	2
3	0.0300	0.2358	34	0.0322	0.4309	17	0.0291	0.5192	7
4	0.0534	0.0718	74	0.0414	0.2236	33	0.0412	0.3682	17
5	0.0684	0.0135	94	0.0669	0.1362	56	0.0541	0.2620	23
6	0.0971	0.0044	98	0.0812	0.0472	77	0.0623	0.2247	27
7	0.1295	0.0007	100	0.1159	0.0196	91	0.0878	0.1584	54
8	0.1486	0.0000	100	0.1310	0.0091	96	0.0933	0.1010	60
9	0.1914	0.0000	100	0.1471	0.0021	99	0.1173	0.0781	70
30% of noise in the variance of the endogenous disturbance term									
1	0.0155	0.6667	4	0.0141	0.7262	0	0.0173	0.6476	3
2	0.0165	0.5883	3	0.0182	0.6088	2	0.0216	0.6116	2
3	0.0256	0.4755	7	0.0221	0.5948	5	0.0238	0.5884	3
4	0.0391	0.3173	18	0.0376	0.3930	17	0.0326	0.4737	4
5	0.0510	0.1923	42	0.0556	0.2783	17	0.0396	0.4414	12
6	0.0793	0.0929	67	0.0743	0.1909	44	0.0519	0.3422	16
7	0.0980	0.0406	81	0.0857	0.1003	61	0.0720	0.2327	25
8	0.1368	0.0118	92	0.1210	0.0537	76	0.0899	0.2174	35
9	0.1576	0.0060	98	0.1246	0.0379	82	0.0986	0.1525	46
50% of noise in the variance of the endogenous disturbance term									
1	0.0189	0.6148	2	0.0259	0.5883	3	0.0230	0.6240	2
2	0.0238	0.6131	5	0.0238	0.6159	6	0.0253	0.5932	2
3	0.0318	0.5014	4	0.0253	0.5674	5	0.0269	0.6266	1
4	0.0334	0.4651	13	0.0354	0.5408	8	0.0350	0.5342	7
5	0.0541	0.3610	25	0.0460	0.3777	17	0.0431	0.5143	7
6	0.0703	0.2450	29	0.0617	0.3221	19	0.0447	0.4759	8
7	0.0793	0.1671	39	0.0628	0.3837	22	0.0456	0.4180	7
8	0.1059	0.1186	58	0.0905	0.1931	35	0.0812	0.2775	22
9	0.1320	0.0536	72	0.1079	0.1517	48	0.0929	0.2344	32

Table 8.17. Simulation results for data with moderately skewed distributions. Total sample size of 200 (100 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0051	0.6490	1	0.0072	0.6696	2	0.0082	0.6828	4
2	0.0094	0.3881	14	0.0094	0.5070	4	0.0094	0.6364	5
3	0.0214	0.1006	63	0.0216	0.2711	32	0.0185	0.3811	16
4	0.0424	0.0072	97	0.0352	0.0667	69	0.0305	0.2298	34
5	0.0602	0.0002	100	0.0509	0.0110	94	0.0403	0.1037	51
6	0.0942	0.0000	100	0.0775	0.0014	100	0.0667	0.0375	79
7	0.1150	0.0000	100	0.0993	0.0003	100	0.0808	0.0357	87
8	0.1533	0.0000	100	0.1234	0.0000	100	0.0848	0.0182	90
9	0.1853	0.0000	100	0.1546	0.0004	100	0.1194	0.0037	99
30% of noise in the variance of the endogenous disturbance term									
1	0.0083	0.6262	3	0.0071	0.7091	0	0.0102	0.6655	4
2	0.0095	0.5492	5	0.0094	0.5912	4	0.0111	0.6119	4
3	0.0192	0.2692	19	0.0197	0.4429	18	0.0176	0.5025	8
4	0.0293	0.1200	51	0.0282	0.2455	32	0.0239	0.3435	16
5	0.0508	0.0320	82	0.0438	0.0972	58	0.0372	0.2536	30
6	0.0769	0.0131	95	0.0638	0.0546	74	0.0493	0.1546	52
7	0.0941	0.0023	99	0.0781	0.0167	92	0.0655	0.0716	68
8	0.1257	0.0002	100	0.0981	0.0052	97	0.0816	0.0447	81
9	0.1498	0.0001	100	0.1301	0.0005	100	0.0907	0.0261	92
50% of noise in the variance of the endogenous disturbance term									
1	0.0106	0.6650	2	0.0115	0.6744	2	0.0101	0.6689	0
2	0.0119	0.5827	4	0.0100	0.6056	1	0.0126	0.6132	0
3	0.0203	0.4141	15	0.0171	0.5209	6	0.0138	0.6118	3
4	0.0272	0.3633	24	0.0263	0.3963	19	0.0219	0.4478	10
5	0.0409	0.1535	43	0.0379	0.2765	31	0.0312	0.3152	20
6	0.0567	0.0794	68	0.0476	0.1446	46	0.0416	0.2379	33
7	0.0782	0.0280	86	0.0653	0.1019	62	0.0441	0.2478	32
8	0.0913	0.0140	93	0.0866	0.0369	81	0.0633	0.1182	53
9	0.1126	0.0121	95	0.0920	0.0343	88	0.0677	0.1011	63

Table 8.18. Simulation results for data with moderately skewed distributions. Total sample size of 500 (250 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0024	0.6850	2	0.0029	0.6227	3	0.0029	0.6591	2
2	0.0068	0.2134	46	0.0058	0.3770	16	0.0044	0.5790	3
3	0.0183	0.0031	99	0.0149	0.0627	70	0.0131	0.2221	36
4	0.0376	0.0000	100	0.0320	0.0010	100	0.0243	0.0656	73
5	0.0610	0.0000	100	0.0517	0.0000	100	0.0394	0.0142	96
6	0.0895	0.0000	100	0.0727	0.0000	100	0.0567	0.0004	100
7	0.1179	0.0000	100	0.1000	0.0000	100	0.0773	0.0000	100
8	0.1517	0.0000	100	0.1199	0.0000	100	0.0917	0.0000	100
9	0.1797	0.0000	100	0.1522	0.0000	100	0.1146	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0021	0.6755	1	0.0035	0.6181	1	0.0038	0.6276	1
2	0.0063	0.3379	19	0.0057	0.5264	9	0.0060	0.5946	4
3	0.0148	0.1143	60	0.0133	0.2109	45	0.0115	0.3593	22
4	0.0297	0.0083	97	0.0281	0.0254	88	0.0173	0.1991	45
5	0.0511	0.0001	100	0.0410	0.0055	97	0.0334	0.0424	80
6	0.0737	0.0000	100	0.0619	0.0005	100	0.0418	0.0222	92
7	0.0964	0.0000	100	0.0811	0.0000	100	0.0584	0.0010	100
8	0.1209	0.0000	100	0.0984	0.0000	100	0.0706	0.0008	100
9	0.1511	0.0000	100	0.1189	0.0000	100	0.0810	0.0002	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0042	0.6187	5	0.0037	0.6311	1	0.0048	0.6866	1
2	0.0058	0.5478	5	0.0058	0.5536	5	0.0061	0.5841	2
3	0.0123	0.2851	31	0.0095	0.4243	13	0.0110	0.4056	15
4	0.0215	0.1203	58	0.0197	0.1625	51	0.0150	0.2652	25
5	0.0362	0.0140	92	0.0304	0.0490	74	0.0228	0.1657	47
6	0.0545	0.0033	98	0.0446	0.0082	95	0.0317	0.0732	71
7	0.0700	0.0006	100	0.0531	0.0029	99	0.0466	0.0241	89
8	0.0867	0.0000	100	0.0694	0.0008	99	0.0473	0.0133	95
9	0.1011	0.0000	100	0.0837	0.0000	100	0.0684	0.0049	97

Table 8.19. Simulation results for data with symmetric non-normal distributions. Total sample size of 1000 (500 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0011	0.6706	1	0.0012	0.7019	0	0.0016	0.6686	2
2	0.0052	0.0360	85	0.0042	0.2057	46	0.0039	0.3951	16
3	0.0170	0.0000	100	0.0149	0.0014	100	0.0110	0.0716	76
4	0.0357	0.0000	100	0.0310	0.0000	100	0.0217	0.0057	96
5	0.0595	0.0000	100	0.0508	0.0000	100	0.0375	0.0000	100
6	0.0881	0.0000	100	0.0708	0.0000	100	0.0564	0.0000	100
7	0.1215	0.0000	100	0.0985	0.0000	100	0.0695	0.0000	100
8	0.1495	0.0000	100	0.1255	0.0000	100	0.0904	0.0000	100
9	0.1860	0.0000	100	0.1530	0.0000	100	0.1088	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0017	0.6455	2	0.0014	0.6659	1	0.0020	0.6363	5
2	0.0040	0.2490	31	0.0045	0.3335	20	0.0037	0.4949	12
3	0.0146	0.0036	99	0.0120	0.0858	62	0.0088	0.2264	40
4	0.0303	0.0000	100	0.0228	0.0048	97	0.0180	0.0233	89
5	0.0476	0.0000	100	0.0412	0.0000	100	0.0275	0.0069	97
6	0.0700	0.0000	100	0.0577	0.0000	100	0.0428	0.0009	99
7	0.0943	0.0000	100	0.0759	0.0000	100	0.0566	0.0000	100
8	0.1162	0.0000	100	0.0992	0.0000	100	0.0709	0.0000	100
9	0.1448	0.0000	100	0.1163	0.0000	100	0.0869	0.0000	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0018	0.6729	3	0.0020	0.6296	2	0.0025	0.6257	3
2	0.0047	0.3656	19	0.0040	0.4706	8	0.0040	0.4806	4
3	0.0111	0.1134	57	0.0105	0.1540	58	0.0081	0.2733	31
4	0.0205	0.0109	93	0.0182	0.0390	85	0.0140	0.1142	61
5	0.0336	0.0009	99	0.0286	0.0059	96	0.0211	0.0377	86
6	0.0533	0.0000	100	0.0404	0.0003	100	0.0310	0.0039	98
7	0.0711	0.0000	100	0.0535	0.0000	100	0.0457	0.0002	100
8	0.0858	0.0000	100	0.0668	0.0000	100	0.0551	0.0003	100
9	0.1026	0.0000	100	0.0854	0.0000	100	0.0629	0.0001	100

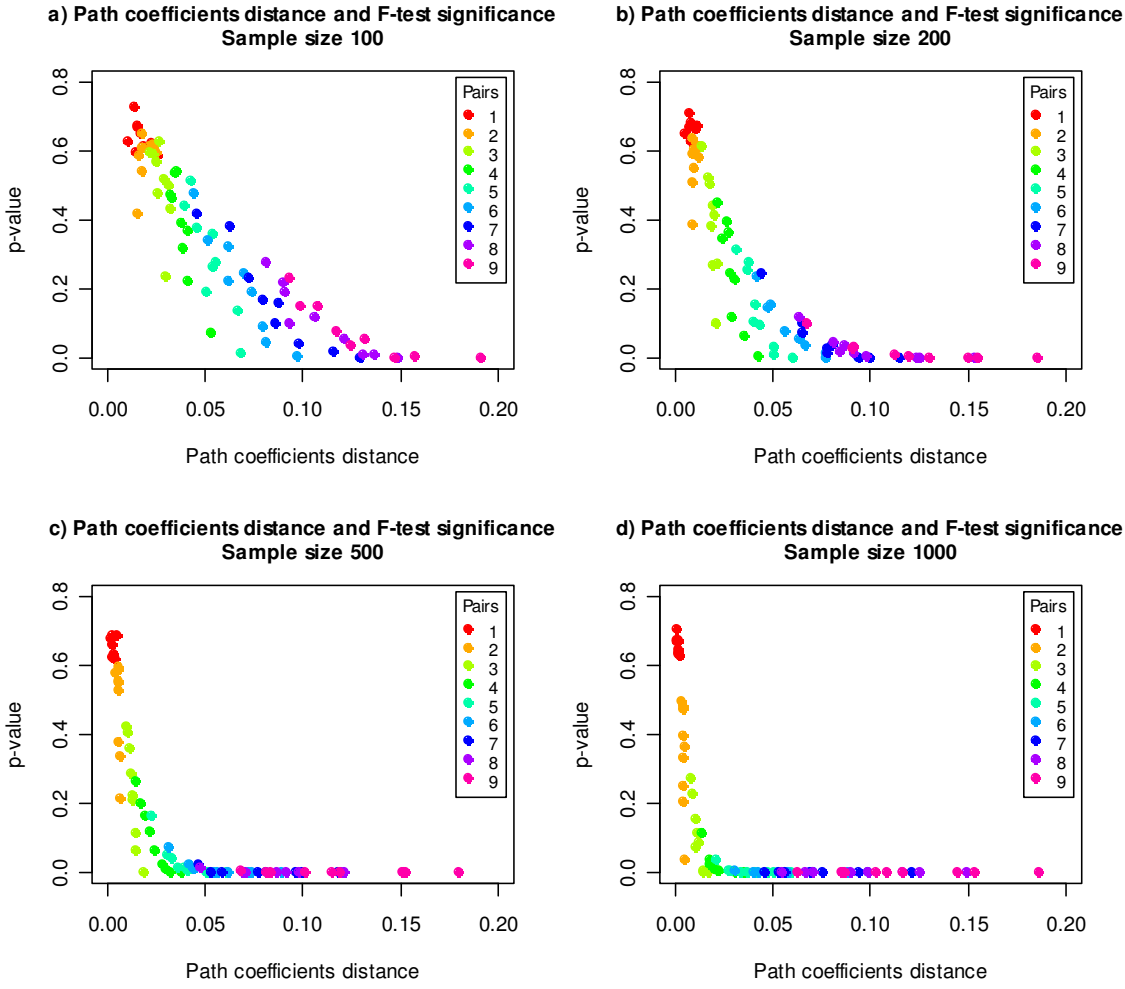


Figure 8.23. *Path coefficients distance and p-value.*
(Data following moderate right-skewed distributions)

Figure 8.24 contains the bar charts of the average proportion of *p*-values smaller than 0.05. The pattern of the charts presents the same behavior as in the non-normal symmetric distributions. Basically, the global impression is that the *F*-test has a good performance under conditions of non-normality and moderately right-skewed distributions. This performance is very important because of the fact that many observed distributions in empirical data have skewed distributions. Hence, although the *F*-test is influenced by the level of noise in the endogenous and the manifest variables, it has a good capability to detect different segments in the presence of non-normal non-symmetric distributions.

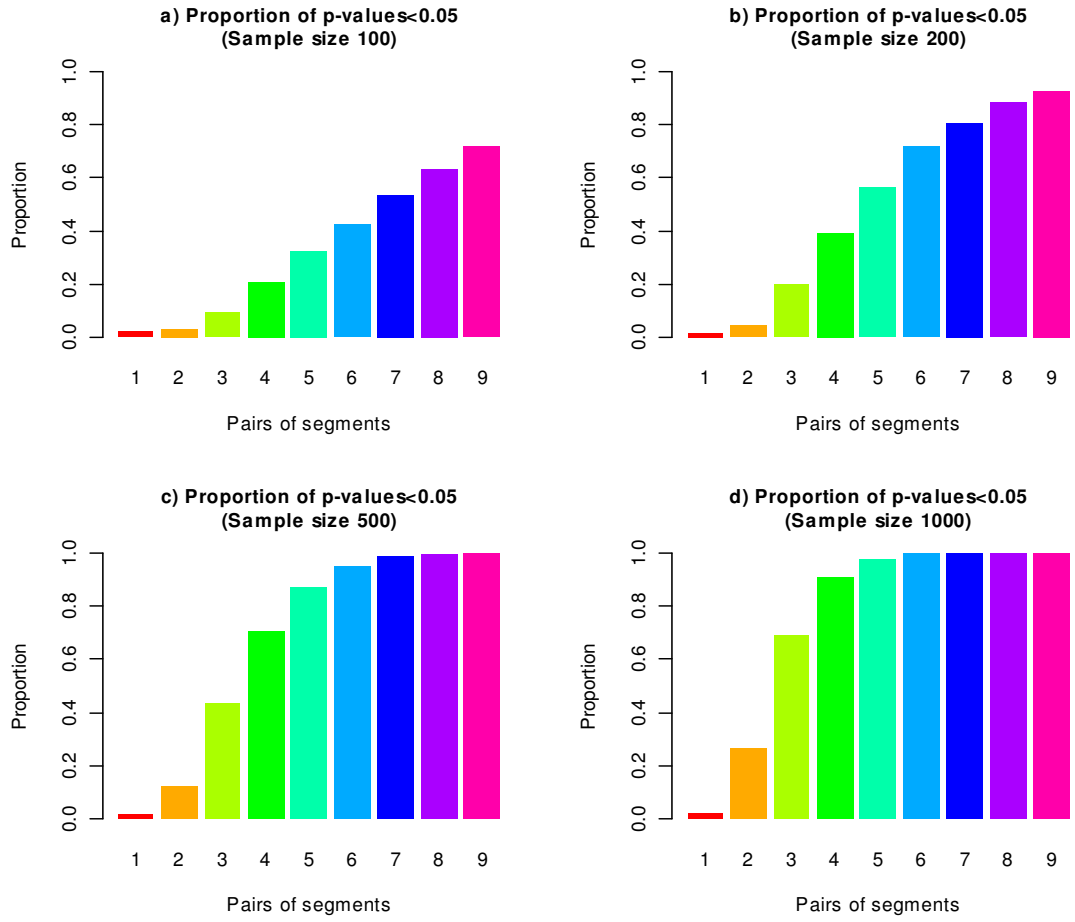


Figure 8.24. *Boxplots of number of p-values>0.05 by pair of segments. (Data following moderate right-skewed distributions)*

8.4.4 Results for the right-skewed case

The summarized results from the simulations with right-skewed distributions are contained in the tables 8.20, 8.21, 8.22 and 8.23. As in the tables for normal data, each table corresponds to a different sample size and it is divided in three horizontal blocks, each of which is divided in three sub-blocks.

Similar trends to the normal data are detected for the moderately right-skewed distributions. First, the average distance of the path coefficients between segments becomes larger. Second, the average *p*-value goes from high values to small values. Third, the number of times that the *p*-values are less than 0.05 becomes larger. The patterns due to the increment of the variance in the endogenous and the manifest variables are also observed: (1) when the segments are similar or their difference is relatively small (pairs 1,2,3 and 4), the average distance of path coefficients is larger compared to those cases with low levels of noises; (2) when the differences between segments is large (pairs 5 to 9) the average distance of path coefficients is smaller compared to those cases with low levels of noise. Plots in figure 8.25 show the results of tables 8.20 – 8.23.

Table 8.20. Simulation results for data with skewed distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0133	0.6929	2	0.0170	0.6857	2	0.0178	0.6636	1
2	0.0209	0.5177	9	0.0189	0.6372	0	0.0156	0.6460	3
3	0.0291	0.2619	25	0.0283	0.4596	9	0.0308	0.5524	6
4	0.0533	0.0756	75	0.0451	0.2462	27	0.0431	0.3966	17
5	0.0746	0.0173	88	0.0659	0.1452	51	0.0501	0.3347	21
6	0.0993	0.0027	99	0.0883	0.0492	76	0.0616	0.2293	35
7	0.1447	0.0009	99	0.1088	0.0324	87	0.0870	0.1450	50
8	0.1597	0.0001	100	0.1275	0.0201	93	0.1044	0.1007	61
9	0.1959	0.0000	100	0.1613	0.0023	99	0.1230	0.0487	69
30% of noise in the variance of the endogenous disturbance term									
1	0.0165	0.6541	1	0.0179	0.7076	1	0.0202	0.6956	3
2	0.0277	0.5594	6	0.0245	0.6257	2	0.0245	0.6093	2
3	0.0310	0.4777	9	0.0321	0.4980	13	0.0267	0.5861	4
4	0.0394	0.3365	21	0.0401	0.4142	12	0.0367	0.4863	9
5	0.0572	0.1833	35	0.0525	0.2768	23	0.0446	0.3846	14
6	0.0855	0.0865	68	0.0750	0.1514	48	0.0630	0.3416	25
7	0.1093	0.0425	82	0.0898	0.1383	51	0.0694	0.2246	34
8	0.1246	0.0194	93	0.1101	0.0826	73	0.0900	0.2131	34
9	0.1560	0.0093	94	0.1358	0.0349	80	0.1122	0.1190	46
50% of noise in the variance of the endogenous disturbance term									
1	0.0222	0.6101	4	0.0209	0.6502	2	0.0255	0.6574	0
2	0.0220	0.6677	4	0.0273	0.5996	3	0.0209	0.6923	0
3	0.0317	0.5538	8	0.0268	0.5808	1	0.0245	0.6458	2
4	0.0343	0.4666	6	0.0320	0.4977	5	0.0347	0.5460	5
5	0.0501	0.3665	18	0.0499	0.4243	12	0.0351	0.5343	6
6	0.0777	0.2397	31	0.0586	0.3741	21	0.0452	0.4656	6
7	0.0794	0.1931	41	0.0721	0.2803	29	0.0667	0.3736	18
8	0.1055	0.1430	54	0.0941	0.1987	34	0.0766	0.2953	17
9	0.1236	0.0674	74	0.0999	0.1823	40	0.0786	0.2705	27

Table 8.21. Simulation results for data with skewed distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0091	0.6315	2	0.0082	0.6834	1	0.0098	0.6424	1
2	0.0117	0.3895	19	0.0111	0.5429	4	0.0116	0.6092	3
3	0.0218	0.1102	61	0.0212	0.2958	27	0.0182	0.4671	6
4	0.0448	0.0013	100	0.0375	0.0779	64	0.0296	0.2631	29
5	0.0630	0.0002	100	0.0508	0.0274	89	0.0417	0.1471	42
6	0.0943	0.0000	100	0.0781	0.0017	100	0.0476	0.0873	62
7	0.1286	0.0000	100	0.0969	0.0005	100	0.0736	0.0274	85
8	0.1442	0.0000	100	0.1279	0.0000	100	0.0958	0.0078	95
9	0.1861	0.0000	100	0.1569	0.0000	100	0.1196	0.0024	99
30% of noise in the variance of the endogenous disturbance term									
1	0.0094	0.6712	3	0.0102	0.6719	0	0.0100	0.6118	1
2	0.0130	0.5153	9	0.0112	0.5669	1	0.0101	0.6632	2
3	0.0202	0.3471	14	0.0205	0.4122	12	0.0165	0.5133	8
4	0.0389	0.0937	61	0.0292	0.2589	32	0.0245	0.3871	18
5	0.0490	0.0472	80	0.0429	0.1039	59	0.0361	0.2635	33
6	0.0716	0.0114	93	0.0651	0.0622	80	0.0472	0.1627	46
7	0.1012	0.0016	99	0.0755	0.0193	88	0.0605	0.0820	60
8	0.1278	0.0002	100	0.1044	0.0081	97	0.0728	0.0739	70
9	0.1568	0.0000	100	0.1204	0.0015	100	0.0981	0.0259	87
50% of noise in the variance of the endogenous disturbance term									
1	0.0131	0.6216	1	0.0106	0.6958	2	0.0129	0.6622	0
2	0.0145	0.5628	4	0.0120	0.6202	6	0.0129	0.6286	2
3	0.0213	0.4196	13	0.0173	0.5085	5	0.0160	0.5769	3
4	0.0294	0.3341	22	0.0295	0.3290	25	0.0228	0.4574	15
5	0.0495	0.1484	50	0.0328	0.2866	23	0.0292	0.3913	20
6	0.0574	0.0995	66	0.0526	0.1656	48	0.0339	0.2911	26
7	0.0778	0.0311	82	0.0622	0.1044	68	0.0461	0.2230	38
8	0.0923	0.0183	92	0.0671	0.0744	70	0.0600	0.1389	50
9	0.1078	0.0088	96	0.0938	0.0288	91	0.0800	0.0984	66

Table 8.22. Simulation results for data with skewed distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*30% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	*50% of noise for indicators' variance* Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0030	0.6505	2	0.0030	0.6538	2	0.0030	0.7011	0
2	0.0072	0.2087	36	0.0058	0.3990	14	0.0068	0.5389	5
3	0.0215	0.0021	99	0.0165	0.0574	72	0.0141	0.2410	33
4	0.0378	0.0000	100	0.0318	0.0037	98	0.0245	0.0639	80
5	0.0595	0.0000	100	0.0514	0.0000	100	0.0443	0.0061	97
6	0.0925	0.0000	100	0.0720	0.0000	100	0.0545	0.0006	100
7	0.1168	0.0000	100	0.1023	0.0000	100	0.0716	0.0000	100
8	0.1519	0.0000	100	0.1244	0.0000	100	0.0945	0.0000	100
9	0.1789	0.0000	100	0.1521	0.0000	100	0.1108	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0036	0.6384	3	0.0045	0.6452	1	0.0046	0.6310	0
2	0.0069	0.3982	16	0.0059	0.4993	5	0.0056	0.5706	7
3	0.0153	0.1147	66	0.0135	0.2254	42	0.0116	0.3280	18
4	0.0306	0.0090	98	0.0239	0.0427	76	0.0220	0.1238	54
5	0.0492	0.0001	100	0.0404	0.0130	94	0.0327	0.0412	80
6	0.0743	0.0000	100	0.0586	0.0004	100	0.0417	0.0091	95
7	0.0958	0.0000	100	0.0744	0.0001	100	0.0528	0.0029	99
8	0.1236	0.0000	100	0.0938	0.0000	100	0.0750	0.0005	100
9	0.1486	0.0000	100	0.1152	0.0000	100	0.0937	0.0001	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0041	0.6435	3	0.0037	0.7035	0	0.0044	0.6882	0
2	0.0071	0.4906	12	0.0070	0.5567	4	0.0062	0.5725	4
3	0.0140	0.2413	34	0.0124	0.3465	19	0.0097	0.4687	10
4	0.0229	0.1223	64	0.0181	0.1677	38	0.0171	0.2523	29
5	0.0371	0.0135	93	0.0316	0.0757	76	0.0238	0.1256	54
6	0.0543	0.0077	97	0.0493	0.0026	100	0.0336	0.0872	67
7	0.0662	0.0003	100	0.0596	0.0013	100	0.0465	0.0183	89
8	0.0869	0.0000	100	0.0756	0.0002	100	0.0521	0.0121	94
9	0.1027	0.0000	100	0.0831	0.0001	100	0.0613	0.0073	98

Table 8.23. Simulation results for data with skewed distributions. Total sample size of 100 (50 cases per segment).

Number of Pair	*10% of noise for indicators' variance*			*30% of noise for indicators' variance*			*50% of noise for indicators' variance*		
	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05	Path Coeffs Distance	*p* -value (mean)	Number of *p* -values<0.05
10% of noise in the variance of the endogenous disturbance term									
1	0.0019	0.6466	1	0.0015	0.6450	2	0.0023	0.6988	3
2	0.0052	0.0548	81	0.0055	0.2027	41	0.0045	0.3463	12
3	0.0203	0.0000	100	0.0151	0.0021	99	0.0115	0.0735	69
4	0.0355	0.0000	100	0.0312	0.0000	100	0.0221	0.0056	97
5	0.0585	0.0000	100	0.0490	0.0000	100	0.0360	0.0001	100
6	0.0897	0.0000	100	0.0737	0.0000	100	0.0566	0.0000	100
7	0.1196	0.0000	100	0.0974	0.0000	100	0.0740	0.0000	100
8	0.1534	0.0000	100	0.1240	0.0000	100	0.0924	0.0000	100
9	0.1865	0.0000	100	0.1549	0.0000	100	0.1089	0.0000	100
30% of noise in the variance of the endogenous disturbance term									
1	0.0018	0.6285	2	0.0019	0.6470	3	0.0022	0.6269	1
2	0.0058	0.2544	33	0.0040	0.3899	12	0.0037	0.5407	5
3	0.0152	0.0108	94	0.0117	0.0545	75	0.0094	0.1627	43
4	0.0306	0.0000	100	0.0236	0.0017	100	0.0181	0.0299	80
5	0.0485	0.0000	100	0.0424	0.0000	100	0.0294	0.0017	99
6	0.0684	0.0000	100	0.0559	0.0000	100	0.0438	0.0000	100
7	0.0896	0.0000	100	0.0752	0.0000	100	0.0566	0.0000	100
8	0.1169	0.0000	100	0.1014	0.0000	100	0.0703	0.0000	100
9	0.1412	0.0000	100	0.1215	0.0000	100	0.0858	0.0000	100
50% of noise in the variance of the endogenous disturbance term									
1	0.0023	0.6254	4	0.0025	0.6452	2	0.0022	0.6477	1
2	0.0041	0.4466	13	0.0043	0.4990	9	0.0037	0.5545	6
3	0.0105	0.1274	57	0.0094	0.1878	43	0.0084	0.2768	27
4	0.0219	0.0083	97	0.0180	0.0370	82	0.0145	0.0980	64
5	0.0338	0.0002	100	0.0323	0.0042	98	0.0217	0.0218	88
6	0.0529	0.0000	100	0.0415	0.0002	100	0.0303	0.0060	99
7	0.0689	0.0000	100	0.0559	0.0000	100	0.0404	0.0031	98
8	0.0843	0.0000	100	0.0703	0.0000	100	0.0534	0.0000	100
9	0.1055	0.0000	100	0.0819	0.0000	100	0.0626	0.0000	100

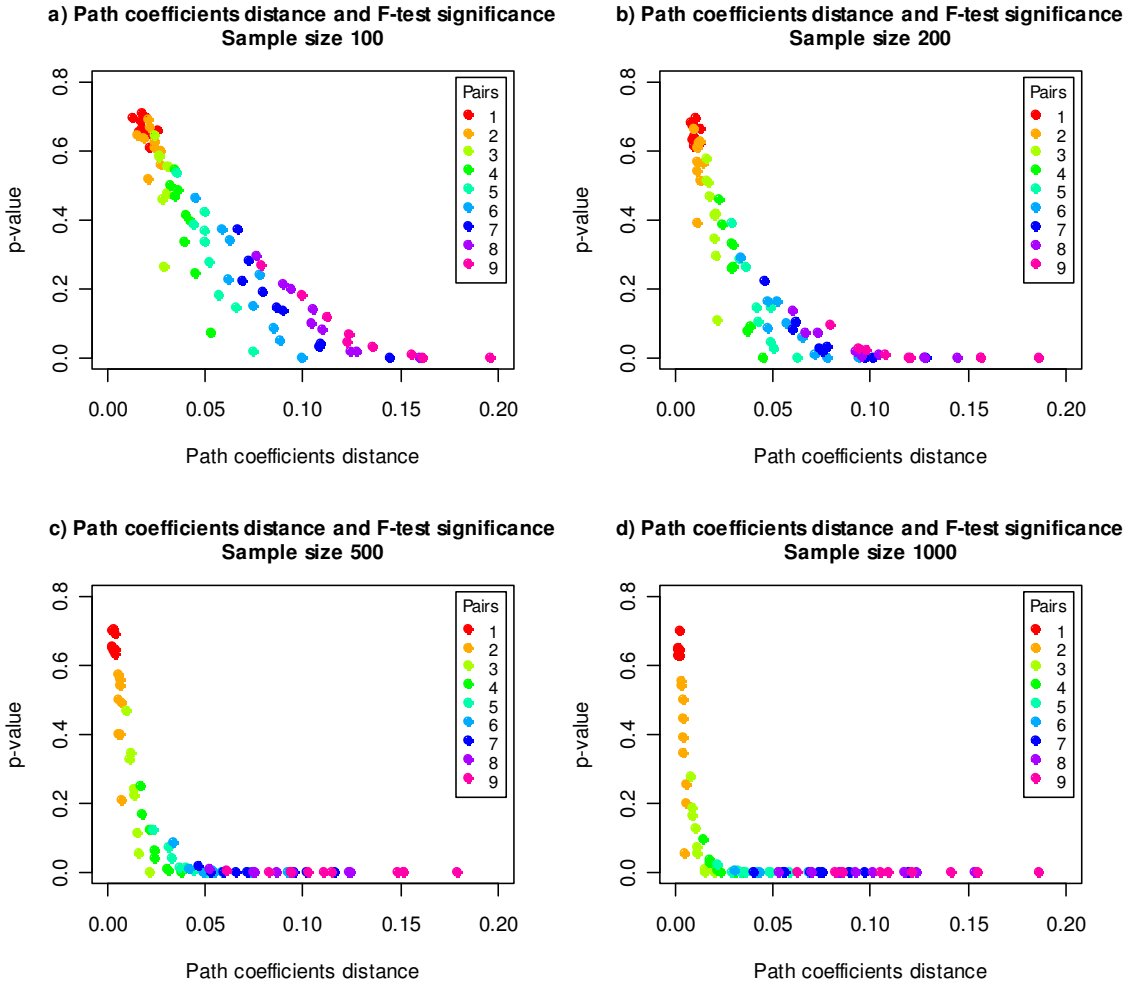


Figure 8.25 *Path coefficients distance and p-value*
(Data following right-skewed distributions)

As in the previous cases of symmetric non-normal distribution and moderately right-skewed distribution, we observe the same pattern of dispersion in the dot plots of figure 8.25. Hence, no distributional effect influences the proposed test.

Bar charts in figure 8.26 correspond to the average proportion of *p*-values that are smaller than 0.05. Once again, the pattern of the charts presents the same structure as in the other normal and non-normal experiments. In summary, the broad sense is that the *F*-test manages plausibly well under circumstances of non-normality and skewed distributions in data. The *F*-test possesses a good capacity to distinguish different segments even when the latent and manifest variables have skewed distributions, and sample sizes are relatively small.

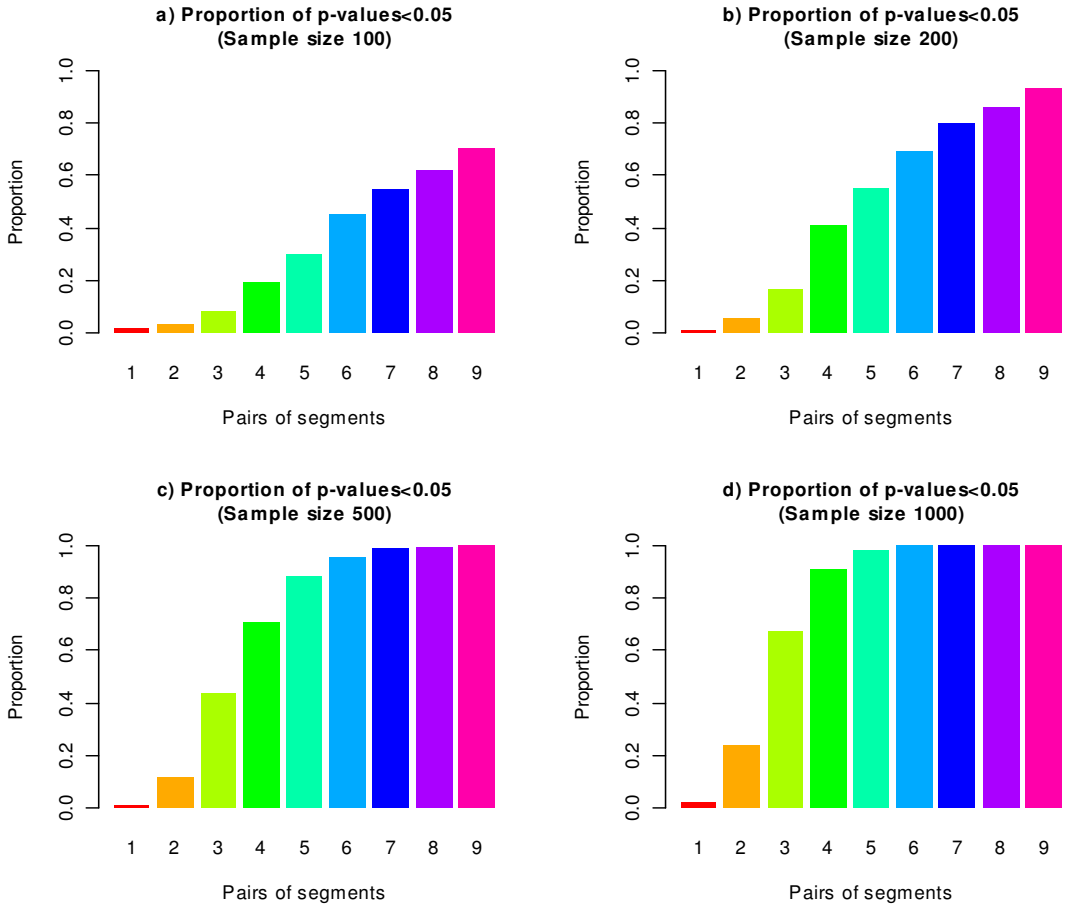


Figure 8.26. *Boxplots of number of p-values>0.05 by pair of segments (Data following right-skewed distributions)*

8.4.5 Results for the combined non-normal cases

In order to have a better appreciation from the non-normal simulation results (symmetric, moderately skewed, and skewed), a similar plot to the one contained in figure 8.10 is shown in figure 8.27. In this case it is possible to observe the similar trends influencing the sensitivity of the *F*-test detected for the normal case:

- The more different the path coefficients, the more sensitive is the test
- The larger the sample size, the more sensitive
- The larger the level of noise (error variance) of the endogenous construct, the less sensitive
- The larger the level of noise (error variance) for indicators, the less sensitive.

In the first plot we can see how the *p*-values decrease as the distance of path coefficients increases. Likewise, the effect of the sample size is reflected in the second plot in which the *p*-values decrease as the sample size becomes larger. In the third plot we can see that a higher sensitivity occurs when the level of noise in the variance of the endogenous error term becomes lower. The same effect is seen in the fourth plot with the levels of noise in the variance of the manifest disturbance terms.

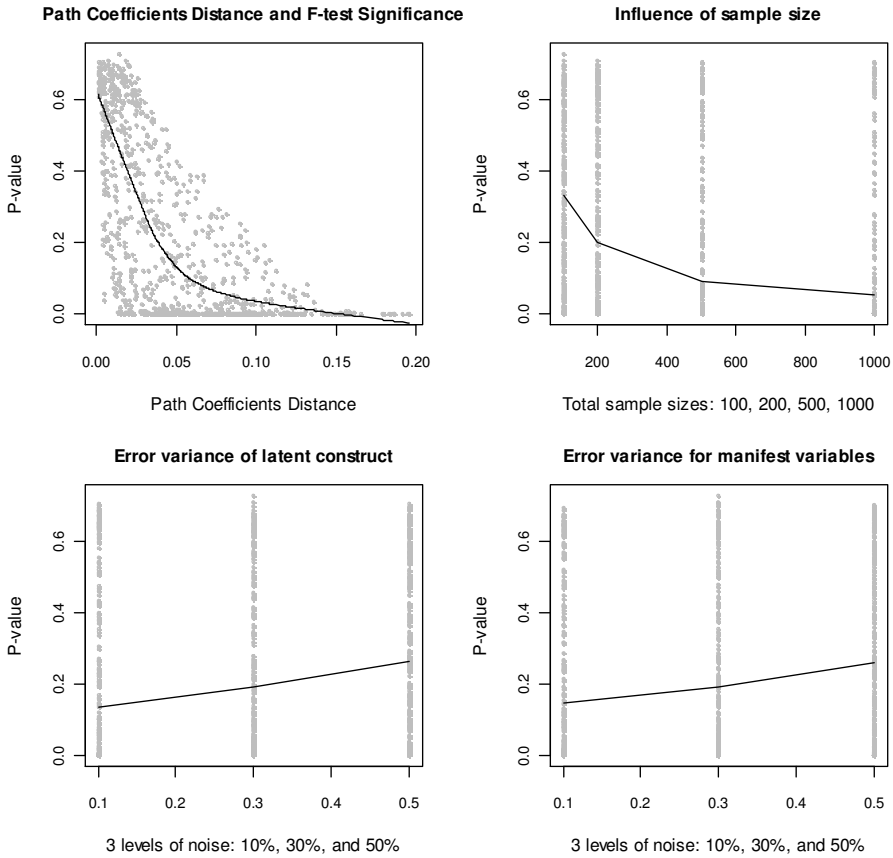


Figure 8.27. *Influence of different data generating conditions on the significance of the F-test at the aggregated level for the non-normal distributions*

If we distinguish by type of beta distribution, the first plot in fig. 8.27 becomes:

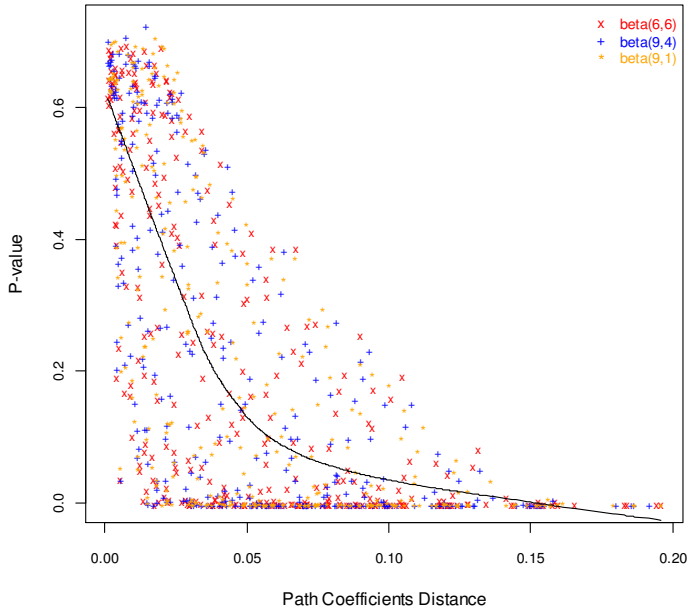


Figure 8.28. *Path coefficients distance and p-value (mean value) for the non-normal distribution results*

The same distinction is made for the other plots (figures 8.29, and 8.30).

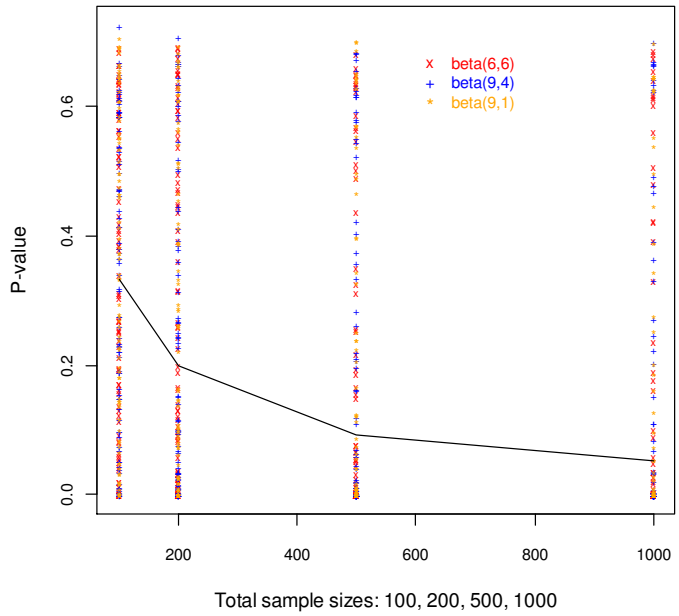


Figure 8.29. *Sample sizes and p-value (mean value) for the non-normal distribution results*

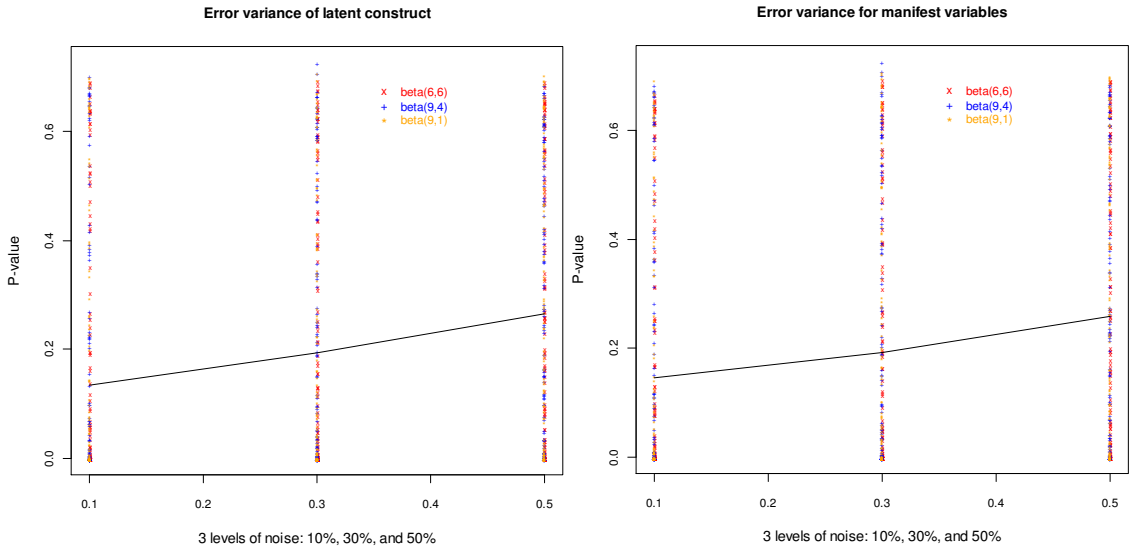


Figure 8.30. *Levels of noise and p-value (mean value) for the non-normal distribution results*

Note that the simulation results have the same aspect as in the normal plots. Again, on the aggregated data level with the three types of beta distributions, the *p*-values from the test present identical behaviors.

To be sure that there is no difference between the symmetric, the moderately skewed and the right skewed distributions, an additional plot with the significance of the *F*-test and the data distributions, as well as the lowess regression line, is given in figure 8.31. No difference is obtained between the beta distributions.

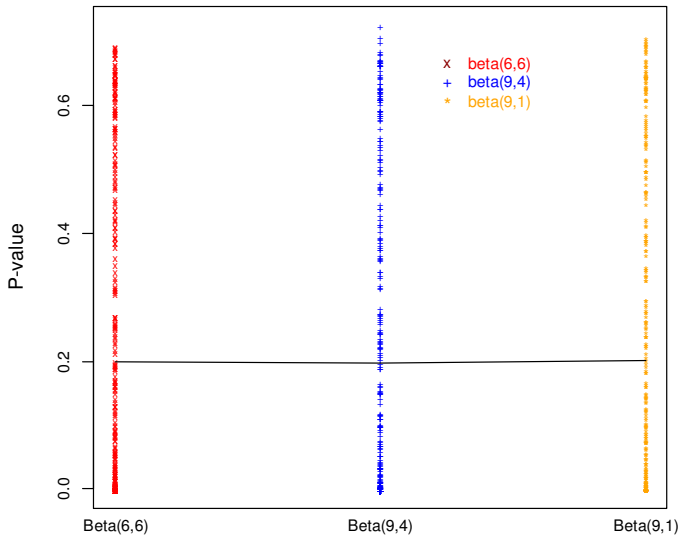


Figure 8.31. *Beta distributions and p-value (no difference detected)*

8.5 Conclusions

To evaluate the performance of the proposed hypothesis test as a split criterion of PATHMOX, a series of simulation studies with experimental data have been undertaken. Different conditions have been investigated such as sample sizes, level of noise of the disturbance terms, number of elements in the models, distinct path coefficients, and distributions in data.

In the first simulation study we have analyzed the performance of the hypothesis test when the endogenous disturbance terms have different variances. Given that the test is based on the assumption of equality of variances in the error terms, the first study has focused on the more realistic situation of having unequal variances. Thus, when the null hypothesis is supposed as being true, the obtained results have shown that the test has an adequate performance even when the variances of the endogenous disturbance terms are not equal.

The second and third simulation studies have been designed to evaluate the sensitivity of the proposed hypothesis test when two structural models are confronted. In particular, the second study examines the behavior of the *F*-test with normal data, whereas the third study is based on models with non-normal data. With regard to the simulation studies with normal data, we have also investigated the case of having different sample sizes in the confronted models. On the other hand, the simulations with non-normal data have been carried out with symmetric distributions, moderate skewed distributions, and skewed distributions. In fact, the interest on the skewed distributions makes emphasis on real-life data scenarios: observed frequency distributions of empirical data are likely to be skewed.

In general, the obtained results provide important evidence in favor of the adequate performance of the proposed *F*-test when it is applied to compare two structural models. The results show that the test has an adequate performance not only in presence of normal distributions, but also with non-normal skewed distributions. From the different experimental conditions, the level of noise of the disturbance terms, and the sample size of the models, are the factors that influence the most the sensitivity of the test. As we

have noticed from the different tables and plots, large level of noises and extreme differences in sample sizes negatively affect the capability of the test in order to detect difference between confronted path models. However, even in the worst experiment conditions, the test has been able to distinguish those models that are very different. In this way, we have confidence in its use as a split criterion in the PATHMOX approach.

Chapter 9

PATHMOX Applications with Real Data

In this chapter we present two applications of PATHMOX with real data. The first application involves a customer satisfaction measurement. The second application involves a study on job satisfaction and motivation. We analyze the data using *Visual Pathmox*, software program specifically designed to provide a graphical interface to calculate PLS path models and PATHMOX segmentation trees. The first section briefly describes the motivation to develop *Visual Pathmox* as well as a description of its basic capabilities. In the second section we discuss the application of PATHMOX with a path model on customer satisfaction. In the third section, an application in a job satisfaction/motivation model is described.

9.1 Visual Pathmox

Visual Pathmox is a *Java*-based software program specifically designed to make PLS path modeling with PATHMOX approach easily accessible for academic research purposes. *Visual Pathmox* has been developed by Oriol Serch (2008) in collaboration with Tomàs Aluja and Gastón Sánchez. Its design and implementation is part of a project from the *Laboratory of Information Analysis and Modeling -LIAM-* at the Barcelona School of Informatics, (Universitat Politécnica de Catalunya).

The original code of the functions to estimate PLS path models as well as the PATHMOX algorithm have been programmed to be run in R (R Development Core Team, 2008). R is a free software based on S language for data manipulation, calculation, and graphical display. It was originally created by Ross Ihaka and Robert Gentleman, from the Auckland University (New Zealand), to provide a statistical environment to their laboratory in 1992. Later on in 1995, with the help of Martin Mächler, they released the R software as a free open-source software (Ihaka and Gentleman, 1996). Currently, R is developed and maintained by the R Development Core Team. R is freely available and its sources, binaries and documentation can be downloaded via CRAN, the "Comprehensive R Archive Network" (www.r-project.org).

With R, analysts can carry out statistical analyses, create sophisticated high level graphics, and run simulations. Users can write their own code to build their own statistical tools, and they also can interact with other programming languages. However, the use of R implies learning a command line language which it is not very user friendly. Thus, in order to provide a user-friendly graphical interface to the code in R of PLS-PM and PATHMOX, it was decided to develop the Java-based program *Visual Pathmox*. The link between Java and R is achieved by means of the Java library JRI (Java to R Interface). This library enables the declaration of an "REngine" in a Java program which encapsulates treatments that can be sent to be computed in R, and then the results can be retrieved in the Java program.

Visual Pathmox enables users to:

- perform standard PLS-PM analysis in a simplified way
- directly draw path diagrams
- specify reflective and formative manifest variables in an easy way
- have segmentation variables as supplementary variables
- use bootstrap method to estimate the standard errors of parameter estimates
- calculate PATHMOX segmentation trees
- interpret results with a variety of outputs (tables, charts, and visual displays)

9.2 Application in Customer Satisfaction

The first application of the PATHMOX approach involves a study on customer satisfaction which is the most typical application of PLS-PM in marketing research (Fornell, 1992; Fornell *et al*, 1996; Anderson *et al*, 2000; Hackl and Westlund, 2000; Martensen *et al*, 2000; O'Loughlin and Coenders, 2004; Cassel, 2006; Hsu *et al*, 2006).

9.2.1 The Model

The analyzed structural model is based on the models presented in Westlund *et al* (2001), Kristensen *et al* (2001) and Tenenhaus *et al* (2005). The model includes six constructs: Image, Expectations, Perceived Quality, Perceived Value, Customer Satisfaction, and Loyalty. The model is designed to measure the cause-effect relationships of the antecedents of customer satisfaction to its consequences. The antecedents of customer satisfaction (Image, Expectations, Perceived Quality, and Perceived Value) are the drivers that affect customer satisfaction, while the consequences of customer satisfaction (Loyalty) are performance indicators. The definitions of the theoretical constructs of the model are given below. The path diagram of the structural model drawn in *Visual Pathmox* is shown in figure 9.1.

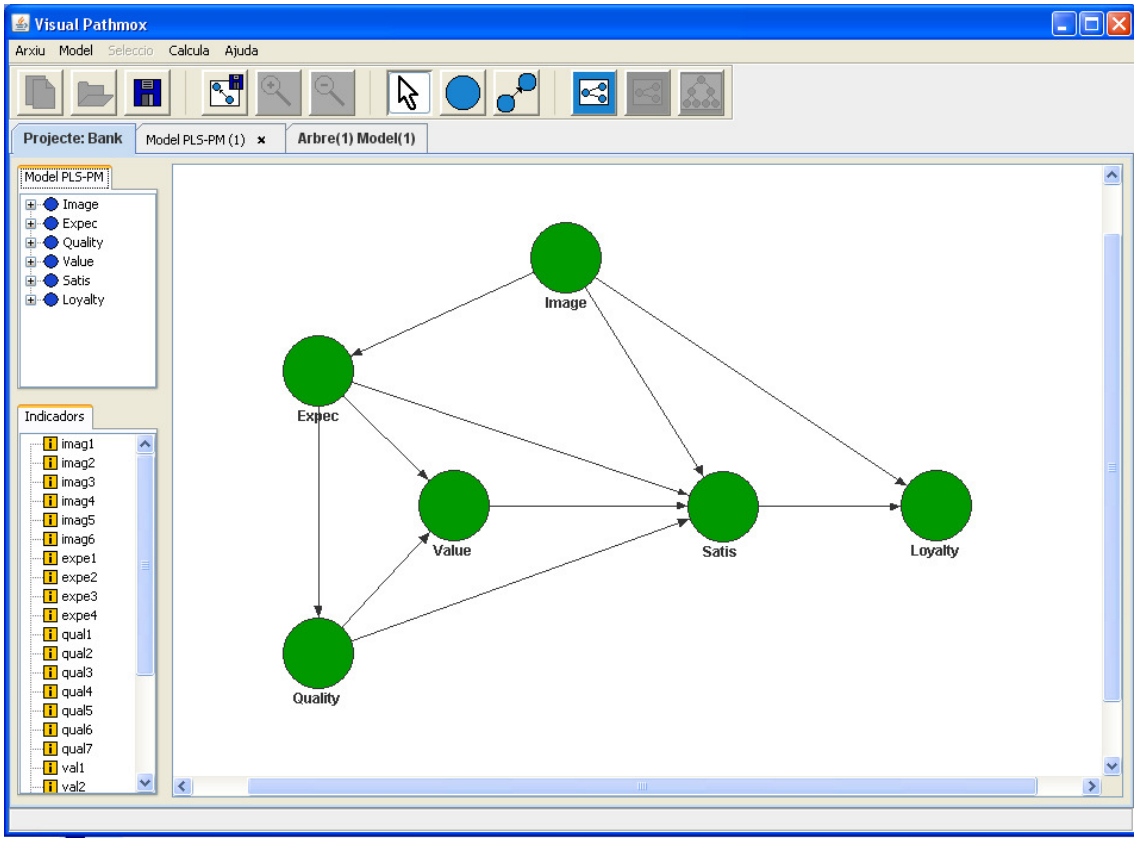


Figure 9.1. *Path diagram of the customer satisfaction model drawn in Visual Pathmox*

Customer Satisfaction

Customer satisfaction is defined as an overall evaluation of a firm's post-purchase performance or utilization of a service (Fornell, 1992). It is modeled taking into account its antecedents and consequences.

Antecedents of customer satisfaction

Image: Image refers to the brand name and the kind of associations customers get from the product/brand/company (Andreassen and Lindestad, 1998). It is expected that image will have a positive effect on customer satisfaction and loyalty. In addition, image is also expected to have a direct effect on expectations.

Expectations: Information based on, not actual consumption experienced but, accumulated information about quality from outside sources, such as advertising, word of mouth, and general media (Anderson *et al*, 1994).

Perceived Quality: Perceived quality comprises product quality (hardware) and service quality (software/humanware). Perceived product quality is the evaluation of recent consumption experience of products. Perceived service quality is the evaluation of recent consumption experience of associated services like customer service, conditions of product display, range of services and products, etc. (Fornell *et al*, 1996; Kristensen *et al*, 1999). Perceived quality is expected to affect satisfaction.

Perceived Value: It is the perceived level of product quality relative to the price paid of the "value for the money" aspect of the customer experience (Anderson *et al*, 1994).

Perceived value is expected to have a direct impact on satisfaction and to be positively affected by perceived quality.

Consequence of customer satisfaction
Loyalty: Customer loyalty refers to the intention repurchase and price tolerance of customers. It is the ultimate dependent variable in the model and it is expected that the better image and higher customer satisfaction should increase customer loyalty.

9.2.2 Data

The data for this application comes from a marketing research study performed in 2008 of one of Spain leading firms in providing retail financial services. Due to confidentiality reasons, the complete details of the survey-based study will not be provided. Only the description of the variables is given. The data is integrated by a total of 32 variables, measured on 1707 clients. The set of 32 variables are divided in two groups. One group is formed by 27 indicator variables for the structural model, and the other group is formed by 5 segmentation variables.

Manifest variables
The questionnaire used in the study was based on the ECSI model questions (see table 9.1); the customers were asked to provide measures on a 11-point ordinal scale ranging from very satisfied (10) to very dissatisfied (0).

Table 9.1. *Description of the manifest variables for each of the latent constructs*

Construct	*Description of Indicators*
Image	1) Bank's reputation
	2) Trustworthiness
	3) Bank's solidness
	4) Innovation and forward looking
	5) Bank's emphasis on public affairs
	6) Caring about the customer's needs
Expectations	1) Providing products and services to meet customer's needs
	2) Providing customer service
	3) Providing solutions to daily banking steps
	4) Expectations for the overall quality
Perceived Quality	1) Reliable products and services
	2) Range of products and services
	3) Degree to which customer feels well informed
	4) Personal advice
	5) Customer service
	6) Overall rating of perceived quality
	7) Clarity and transparency of operations and transactions
Perceived Value	1) Beneficial services and products
	2) Valuable investments
	3) Quality relative to price
	4) Price relative to quality

Cont. Table 9.1. *Description of the manifest variables for each of the latent constructs*

Construct	*Description of Indicators*
Customer Satisfaction	1) Overall rating of satisfaction 2) Fulfillment of expectations 3) Rating the performance relative to customer's ideal bank
Loyalty	1) Propensity to choose the same bank if the customer would had to choose again 2) Propensity to switch to other banks when they offered better transactions 3) Customer's intention to recommend the bank to friends or colleagues

Segmentation variables
The five segmentation variables, which serve as observed sources of heterogeneity, are: Gender, Age, Education Level, Occupation, and Region.

9.2.3 Results of the global PLS path model

Measurement Model
The measurement relationships of the latent variables and their blocks of indicators were taken in a reflective way. Since all manifest variables use the same scale, no standardization was used. The measurement model is assessed in three aspects:

- Unidimesionality of blocks of indicators
- Reliability of indicators
- Differentiation between latent variables

Unidimensionality of block of indicators
In a reflective measurement model it is assumed that manifest variables are considered as being caused by their latent constructs. Hence, it is assumed that the indicators are closely related; in such a way that they are in one dimensional space. In order to assess the extent to which a block is unidimensional three indices are the most common methods employed for this purpose: 1) principal component analysis, 2) Cronbach's α, and 3) Dillon-Goldstein's ρ.

The use of principal components analysis is based on the importance of the eigenvalues. If a block is unidimensional, then the first eigenvalue of the correlation matrix of the indicators should be larger than one whereas the second eigenvalue should be smaller than 1. Cronbach's α coefficient evaluates how well a block of indicators measure their corresponding latent construct. As a rule of thumb, a block is considered unidimensional when Cronbach's α is larger than 0.7. Like Cronbach's α, the Dillon-Goldstein's ρ is also focused on the variance of the sum of variables in the block of interest. As a rule of thumb, a block is considered as unidimensional when Dillon-Goldstein's ρ is larger than 0.7. The eigenvalues from the principal component analysis as well as the Cronbach's α values, and the Dillon-Goldstein's ρ, are shown in table 9.2. The first eigenvalue of the correlation matrix of the manifest variables of each construct is larger than one, and the second one is smaller than one. All the Cronbach's a met the threshold value of 0.70. Similarly, Dillon-Goldstein's ρ values are also above 0.70 for each construct. All three tools support unidimensionality of blocks of indicators.

Table 9.2. *Different measures to assess unidimensionality of blocks of indicators*

Constructs	*Number of indicators*	*First Eigenvalue*	*Second Eigenvalue*	*Cronbach's* α	*Dillon-Goldstein* ρ
Image	6	3.63	0.82	0.87	0.90
Expectations	4	2.63	0.58	0.83	0.89
Quality	7	4.32	0.67	0.90	0.92
Value	4	2.81	0.58	0.86	0.90
Satisfaction	3	2.50	0.34	0.90	0.94
Loyalty	3	2.19	0.58	0.81	0.89

Reliability of indicators

The reliability of each manifest variable is assessed by examining the loadings of the indicators with their respective latent constructs. The loadings (see table 9.3) are simple correlation coefficients between the indicators and their respective latent variables. Communalities are squared loadings. Moreover, a broader way to assess the reliability of each indicator is by examining their correlations with their associated latent variables (see table 9.4).

Table 9.3. *Measurement model: outer weights, loadings, and communalities of each manifest variable*

Indicator	*Outer Weight*	*Loading*	*Communality*
imag1	0.154	0.783	0.614
imag2	0.188	0.819	0.671
imag3	0.130	0.723	0.523
imag4	0.176	0.792	0.627
imag5	0.128	0.678	0.459
imag6	0.223	0.836	0.700
expe1	0.231	0.796	0.633
expe2	0.238	0.793	0.629
expe3	0.269	0.816	0.666
expe4	0.262	0.836	0.699
qual1	0.121	0.673	0.453
qual2	0.153	0.792	0.628
qual3	0.147	0.838	0.703
qual4	0.164	0.850	0.722
qual5	0.132	0.772	0.596
qual6	0.161	0.830	0.688
qual7	0.122	0.723	0.523
val1	0.294	0.880	0.775
val2	0.215	0.813	0.662
val3	0.225	0.819	0.670
val4	0.266	0.815	0.665
sat1	0.341	0.928	0.860
sat2	0.333	0.929	0.864
sat3	0.326	0.883	0.779
loy1	0.385	0.920	0.846
loy2	0.232	0.698	0.487
loy3	0.383	0.921	0.847

Another index that is used to assess how well the indicators are explained by their latent variables is Communality. It measures how much of a given manifest variable's variance is reproducible from the latent variable. That is to say, it measures the part of variance between a construct and its indicators that is common to both. The fourth column of table 9.3 contains the communality indices of each manifest variable. Communality should be larger than 0.50 which means that 50% or more variance of the indicators should be accounted for. It can be observed that three indicators (imag5, qual1, and loy2) are below the recommended 0.5 value. However, we do not consider their communality indices to be so low as to be eliminated.

Table 9.4. *Correlations between the manifest variables and the latent constructs*

	Image	Expectations	Perceived Quality	Perceived Value	Customer Satisfaction	Loyalty
imag1	**0.783**	0.472	0.592	0.539	0.546	0.515
imag2	**0.819**	0.539	0.660	0.593	0.618	0.576
imag3	**0.723**	0.451	0.541	0.499	0.466	0.429
imag4	**0.792**	0.480	0.544	0.555	0.552	0.519
imag5	**0.677**	0.353	0.381	0.420	0.397	0.364
imag6	**0.837**	0.519	0.581	0.620	0.625	0.564
expe1	0.505	**0.795**	0.590	0.478	0.474	0.430
expe2	0.469	**0.793**	0.539	0.479	0.451	0.417
expe3	0.468	**0.816**	0.593	0.468	0.496	0.451
expe4	0.550	**0.836**	0.692	0.543	0.555	0.508
qual1	0.526	0.524	**0.673**	0.495	0.460	0.428
qual2	0.592	0.606	**0.792**	0.582	0.601	0.540
qual3	0.553	0.566	**0.838**	0.506	0.556	0.521
qual4	0.601	0.595	**0.850**	0.565	0.625	0.583
qual5	0.501	0.534	**0.772**	0.484	0.545	0.499
qual6	0.645	0.720	**0.830**	0.643	0.667	0.616
qual7	0.491	0.542	**0.723**	0.481	0.488	0.456
val1	0.637	0.580	0.638	**0.880**	0.738	0.695
val2	0.582	0.527	0.625	**0.812**	0.626	0.605
val3	0.605	0.514	0.588	**0.818**	0.614	0.579
val4	0.535	0.419	0.471	**0.817**	0.596	0.552
sat1	0.671	0.581	0.692	0.752	**0.928**	0.755
sat2	0.639	0.564	0.674	0.704	**0.930**	0.733
sat3	0.617	0.527	0.615	0.668	**0.882**	0.647
loy1	0.614	0.512	0.618	0.698	0.743	**0.920**
loy2	0.395	0.377	0.444	0.435	0.489	**0.698**
loy3	0.615	0.520	0.623	0.687	0.719	**0.920**

Differentiation between constructs

The third aspect evaluated in a measurement model is the extent to which a given construct differentiates from the others. This is done by verifying that no indicator loads higher on another construct than the one it intends to measure. To do this, we examine

the correlations between the manifest variables and all the latent variables (shown in table 9.4). If a manifest variable loads higher with other constructs than the one it is intended to measure, we might question its appropriateness because it is not clear which construct it is actually reflecting. In table 9.4, the correlations of the indicators with their associated construct are highlighted in bold. In each row, it is expected that the highlighted correlation is the highest value. As it can be seen, all indicators have their highest value on the latent variables they are intended to measure.

Structural model

The results of the structural model involve the estimates of the path coefficients, the R^2 values for each endogenous construct, and the correlations among latent variables. The indirect effects and the total effects of the relationships among latent variables can also be calculated. In addition, the redundancy for each endogenous block of indicators as well as the Goodness-of-Fit (GoF) index can also be obtained to evaluate the quality of the model.

Path coefficients

First, we examined the path coefficients of the structural model (see figure 9.2 and table 9.5). In order to do this, we have to take into account the theoretical framework underlying the causal model. For instance, it is expected that image will have a positive effect on expectations, customer satisfaction and loyalty.

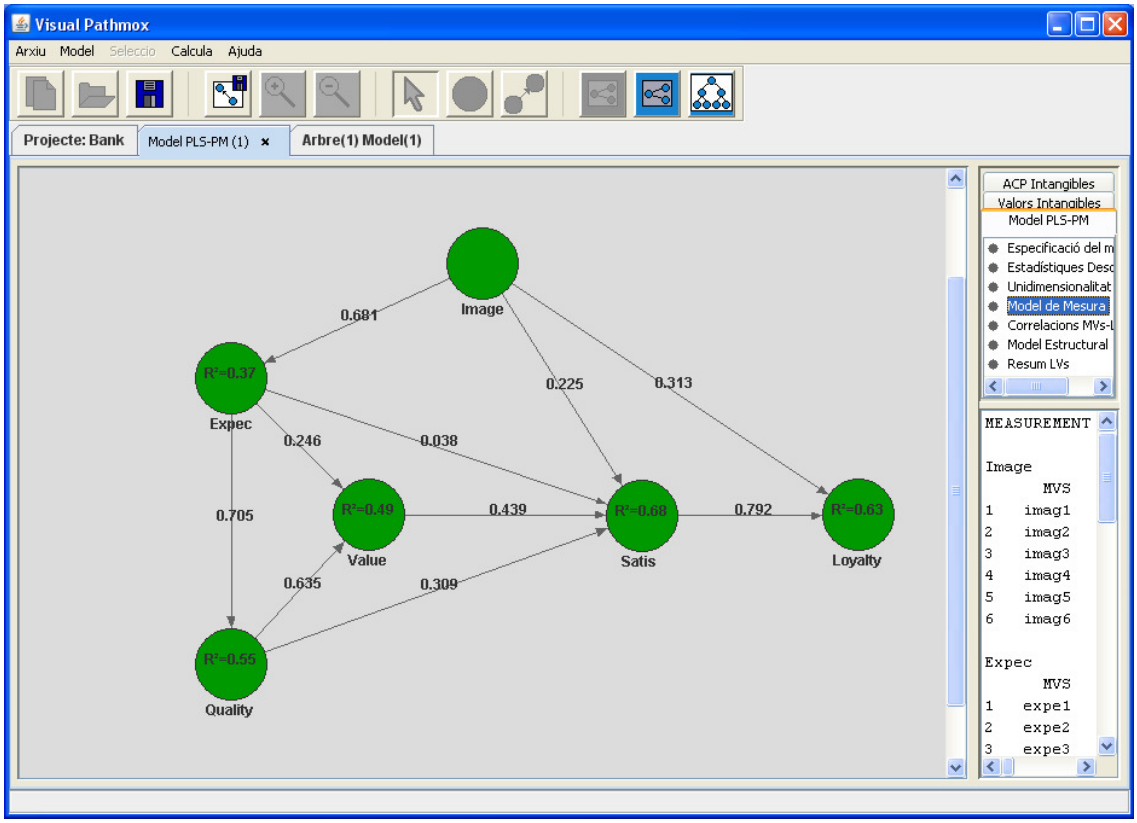


Figure 9.2. *Results of the structural model in Visual Pathmox. Path coefficients shown on the arrows. R^2 values inside the latent variables*

Table 9.5. *Results of the structural model (R^2 and path coefficients)*

Constructs		*Estimate*	*Std. Error*	*t value*	*Pr(>|t|)*
Expectations	$R^2 = 0.37$				
	Image	0.681	0.021	31.935	0.000
Quality	$R^2 = 0.55$				
	Expectations	0.705	0.015	45.976	0.000
Value	$R^2 = 0.49$				
	Expectations	0.246	0.030	8.291	0.000
	Quality	0.636	0.031	20.290	0.000
Satisfaction	$R^2 = 0.68$				
	Image	0.225	0.027	8.364	0.000
	Expectations	0.038	0.023	1.607	0.108
	Quality	0.309	0.029	10.839	0.000
	Value	0.439	0.020	21.552	0.000
Loyalty	$R^2 = 0.63$				
	Image	0.313	0.032	9.702	0.000
	Satisfaction	0.792	0.026	30.582	0.000

By examining the path coefficients of those constructs affected by image it can be seen that the previous assumptions are confirmed since all the coefficients are positive. In particular, we are interested in examining the path coefficients affecting customer satisfaction. Perceived value is the latent variable that has the strongest direct influence on satisfaction with a value of 0.439. The second highest effect is due to perceived quality with a value of 0.309. In the third place we found image with a path coefficient of 0.225. Finally, Expectations has a very low direct impact on customer satisfaction. In the case of the path coefficients related to loyalty, satisfaction has a value of 0.792 while image has a value of 0.313.

R^2: Predictive power

The measures of the predictive power of the model are the R^2 values of the endogenous latent variables. R^2 coefficients, shown inside the latent variables in figure 9.2, are interpreted in the same manner as the R^2 obtained in regression analysis. R^2 indicates the amount of variance in the construct explained by the model. For instance, the results in the model indicate that 37 percent of the variance in Expectations is explained. Similarly, 55 percent of the variance in Perceived Quality, and 49 percent of the variance in Perceived Value are explained. Customer Satisfaction, in turn, has an R^2 of 68 percent of the variance whereas Loyalty has an $R^2$of 63. These values are considered as adequate values for good predictive power of the model.

Bootstrap validation

In order to validate the precision of the estimates obtained in the structural model re-sampling methods are applied. In our case we used bootstrapping validation with 100 samples. The results of the bootstrap validation of the path coefficients are shown in table 9.6. The second column contains the original estimate of the path coefficients. The third column is the mean value of the 100 bootstrap samples. The standard deviation appears in the fourth column. The last two columns have the 5 and 95 percentiles of the

bootstrap estimates. Note that the interval for the percentile of expectations on satisfaction contains the zero. This means that expectations can be considered as having no influence on satisfaction.

Table 9.6. *Bootstraps results of path coefficients (direct effects)*

Path relations	*Original*	*Mean.Boot*	*Std.Dev*	*perc05*	*perc95*
Image on Expectations	0.681	0.683	0.024	0.641	0.721
Expectations on Quality	0.705	0.709	0.020	0.677	0.741
Expectations on Value	0.246	0.240	0.037	0.183	0.299
Quality on Value	0.635	0.643	0.034	0.591	0.698
Image on Satisfaction	0.225	0.220	0.040	0.155	0.280
Expectations on Satisfaction	0.038	0.031	0.026	-0.014	0.070
Quality on Satisfaction	0.309	0.316	0.032	0.268	0.366
Value on Satisfaction	0.439	0.439	0.036	0.380	0.499
Image on Loyalty	0.313	0.310	0.041	0.251	0.373
Satisfaction on Loyalty	0.792	0.798	0.033	0.740	0.848

Direct effects, indirect effects and total effects

An interesting aspect of the relationships among latent variables in the structural model involves the indirect and total effects. Indirect effects comprise the indirect paths from one variable to another. Total effects are the combination of direct plus indirect effects. For example, although there is no arrow (i.e., path) from image to perceived quality, image can affect quality indirectly by means of expectations. The indirect effect of image on quality is calculated by multiplying the path coefficient of image-on-expectation times the path coefficient of expectations-on-quality, that is, 0.681 x 0.705 = 0.480. The total effect of expectations on perceived value is 0.694 which is obtained as the sum of the direct effect (0.246) plus the indirect effect (0.448). Table 9.7, contains the three types of effects in the path relations of the structural model.

Table 9.7. *Direct, indirect, and total effects among latent constructs*

Path relations	*Direct*	*Indirect*	*Total*
Image on Expectations	0.681	0.000	0.681
Expectations on Quality	0.705	0.000	0.705
Expectations on Value	0.246	0.448	0.694
Quality on Value	0.635	0.000	0.635
Image on Satisfaction	0.225	0.381	0.606
Expectations on Satisfaction	0.038	0.443	0.481
Quality on Satisfaction	0.309	0.279	0.588
Value on Satisfaction	0.439	0.000	0.439
Image on Loyalty	0.313	0.480	0.793
Satisfaction on Loyalty	0.792	0.000	0.792

A graphical representation of the total effects on satisfaction and loyalty are presented in figures 9.3 and 9.4, respectively. Note that perceived value is the latent variable that has the strongest direct influence on satisfaction with a direct effect of 0.439. However, the direct and indirect influences of image on satisfaction have the highest combined

effect with a value of 0.605. In turn, perceived quality is the second largest influence on satisfaction with a total effect of 0.588. In third place we found the total effect of expectations on satisfaction with a value of 0.480.

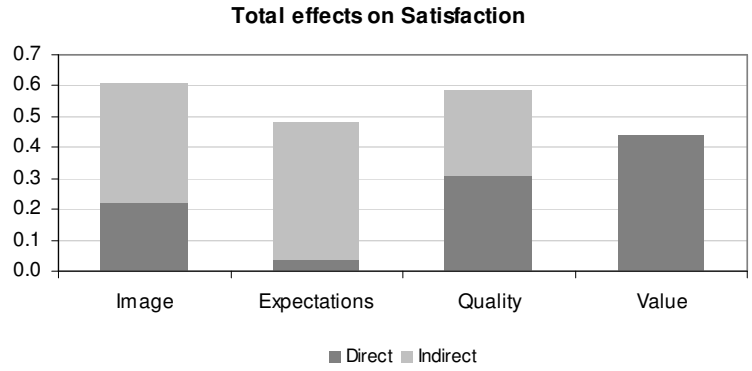


Figure 9.3. *Total effects on customer satisfaction*

In regards to loyalty, image has a total effect similar to that of satisfaction with 0.793. In third place we have the total effect of perceived quality (0.466), followed by the total effects of expectations (0.443) and perceived value (0.348).

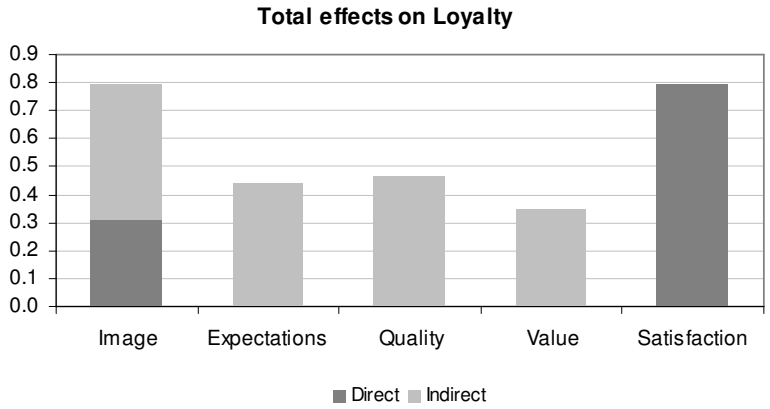


Figure 9.4. *Total effects on loyalty*

Correlations among latent variables

Table 9.8 contains the correlations among the latent constructs. As expected from the theoretical framework, we can see that all variables are highly positive correlated.

Table 9.8. *Correlations between latent variables in the Customer Satisfaction model*

	Image	Expectations	Quality	Value	Satisfaction	Loyalty
Image	1					
Expectations	0.612	1				
Quality	0.714	0.74	1			
Value	0.704	0.61	0.685	1		
Satisfaction	0.704	0.61	0.723	0.775	1	
Loyalty	0.650	0.56	0.668	0.730	0.779	1

Redundancy
Redundancy measures the amount of variance in an endogenous construct block that is explained by its independent latent variables. It reflects the ability of a set of independent latent variables to explain variation in the manifest variables of a dependent latent variable. Table 9.9 shows the redundancy measures for each block of indicators. It can be observed that the block of expectations has smaller redundancy values since expectation has the lowest R^2. Conversely, manifest variables in satisfaction and loyalty have the largest redundancy values, which is in accordance to their respective R^2.

Table 9.9. *Redundancy of indicators*

Indicator	*Redundancy*	*Indicator*	*Redundancy*
expe1	0.237	val1	0.379
expe2	0.235	val2	0.323
expe3	0.249	val3	0.328
expe4	0.262	val4	0.327
qual1	0.251	sat1	0.589
qual2	0.347	sat2	0.591
qual3	0.389	sat3	0.533
qual4	0.400		
qual5	0.330	loy1	0.530
qual6	0.381	loy2	0.305
qual7	0.289	loy3	0.531

9.2.4 PATHMOX segmentation tree

In order to calculate the PATHMOX segmentation tree, it is necessary to specify the scale (e.g., binary, ordinal, or nominal) of the segmentation variables. This codification, shown in table 9.10, is needed in order to determine the number of possible binary splits of each variable.

Table 9.10. *Codification of segmentation variables depending on their type of scale*

Segmentation Variable	*Scale*
Gender	Binomial
Age	Ordinal
Education Level	Ordinal
Occupation	Nominal
Region	Nominal

In addition, we have to determine the parameters and stop conditions of the algorithm:

- p-value significance threshold = 0.05
- minimum number of individuals inside a node = 10% (expressed as a percentage of the total population)
- depth level of the tree (level up to which the tree is allowed to grow) = 2

The specified parameters are somewhat arbitrary. We have chosen those settings for demonstration purposes only of a PATHMOX application. We have decided to establish a value of 0.05 for the threshold of the *p*-value to look for those partitions that are highly significant. Given that we have a total sample of 1707 customers, a number of 170 customers (10% of total sample) seems us to be a reasonable minimum number to stop the growth of a node. The depth level has been selected with the aim to obtain a simple segmentation tree with a possible maximum number of four final segments. Despite the reasons outlined to choose the parameters of the algorithm, it is assumed that in a real life application the user or the analyst must try different setting options. The knowledge and advice of an expert in the field of application is also encouraged to guide the process of segments identification.

The obtained PATHMOX tree is shown in figure 9.5. It is a tree with a total number of seven nodes. The root node is assumed to be associated with the global model which is calculated over the entire sample of 1707 customers. As it can be seen in figure 9.5, the number of elements in each node, as well as its proportion with respect to the total sample size (shown in parenthesis), appear insideeach circle nodes.

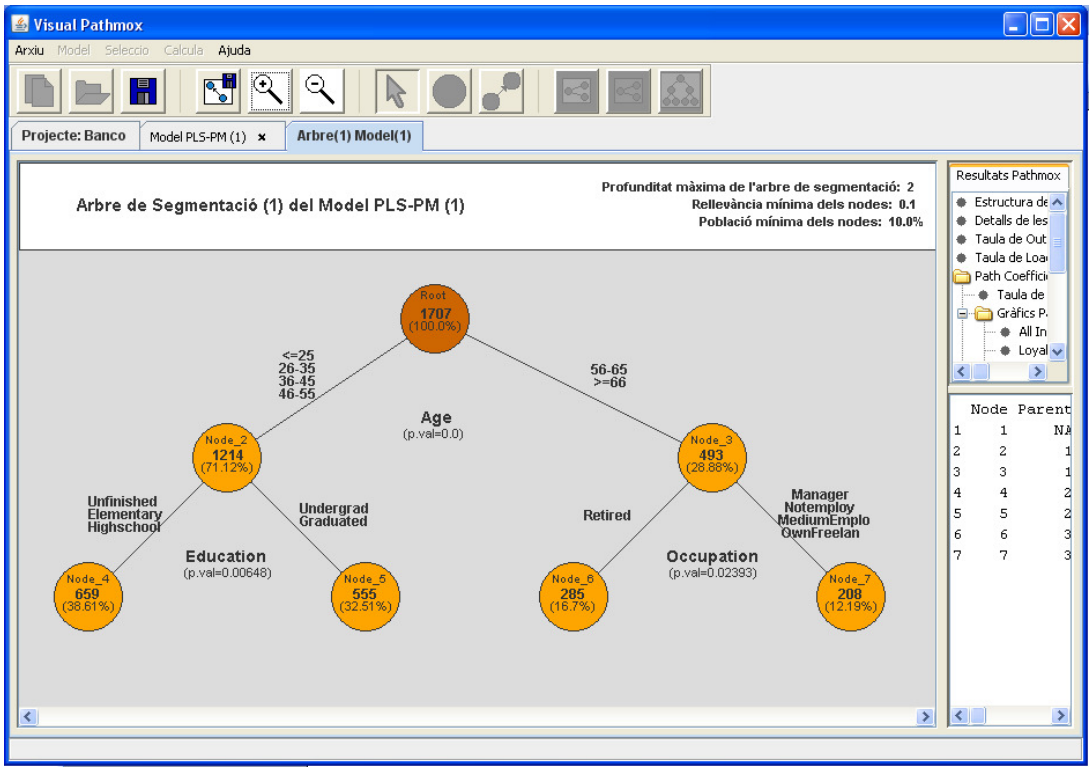


Figure 9.5. *PATHMOX segmentation tree in Visual Pathmox*

The first split results from the segmentation variable 'Age' which has a corresponding *F*-statistic's *p*-value of zero. This split divides the 1707 customers of the root node in two subsets. One subset is formed by 1214 customers with an age of 55 years or younger (node 2). The other subset consists of 493 customers with an age of 56 years or older (node 3). These first nodes correspond to the first depth level of the tree.

In the second level of the tree we have the partitions of nodes 2 and 3. In regards to node 2, it is split by 'Education Level' in two nodes: node 4 represents the path model

of 659 customers that have a high school degree as a maximum level of education; node 5 represents the path model of the segment with 555 customers with an undergraduate or graduate degree. In contrast, node 3 is divided by 'Occupation' in nodes 6 and 7. Node 6 represents the path model of 285 retired customers. Node 7 represents the path model of the segment with 208 customers that are not retired.

The *p*-values associated to the three splits indicate a high significance level in their corresponding *F* tests. For instance, the partition of node 2 has a *p*-value of 0.0054 which is regarded highly significant at a 0.05 level. This implies that the binary split of node 2 according to the segmentation variable 'Education' with 'Unfinished-Elementary-HighSchool' categories versus 'Undergraduate-Graduate' categories, generates the pair of structural models that are the most different among all the possible binary splits. The path coefficients of the structural model for each final segment, as well as the path coefficients of the global model, are listed in table 9.11. The total effects of the structural relationships are displayed in table 9.12. In turn, the path diagrams of the corresponding structural models are in Appendix I.

Table 9.11. *Path coefficients of the path models (root node and the terminal segments)*

Path Relations	*Global*	*Node4*	*Node5*	*Node6*	*Node7*
Image on Expectations	0.681	0.659	0.704	0.733	0.610
Expectations on Quality	0.705	0.713	0.718	0.688	0.686
Expectations on Value	0.246	0.268	0.239	0.321	0.080
Quality on Value	0.636	0.624	0.721	0.491	0.555
Image on Satisfaction	0.225	0.198	0.258	0.210	0.196
Expectations on Satisfaction	0.038	-0.009	0.097	0.028	0.003
Quality on Satisfaction	0.309	0.345	0.208	0.355	0.371
Value on Satisfaction	0.439	0.431	0.490	0.393	0.360
Image on Loyalty	0.313	0.379	0.218	0.329	0.260
Satisfaction on Loyalty	0.792	0.723	0.899	0.748	0.798

Table 9.12. *Total Effects of the path relations corresponding to the root node and the terminal segments*

Path Relations	*Global*	*Node4*	*Node5*	*Node6*	*Node7*
Image on Expectations	0.681	0.659	0.704	0.733	0.610
Image on Quality	0.480	0.470	0.505	0.504	0.418
Image on Value	0.472	0.470	0.533	0.483	0.281
Image on Satisfaction	0.606	0.557	0.692	0.599	0.454
Image on Loyalty	0.793	0.781	0.841	0.777	0.622
Expectations on Quality	0.705	0.713	0.718	0.688	0.686
Expectations on Value	0.694	0.713	0.757	0.659	0.461
Expectations on Satisfaction	0.480	0.452	0.575	0.407	0.310
Expectations on Loyalty	0.443	0.393	0.555	0.397	0.338
Quality on Value	0.635	0.624	0.721	0.491	0.555
Quality on Satisfaction	0.587	0.614	0.561	0.548	0.571
Quality on Loyalty	0.465	0.444	0.505	0.410	0.455
Value on Satisfaction	0.439	0.431	0.490	0.393	0.360
Value on Loyalty	0.348	0.312	0.441	0.294	0.287
Satisfaction on Loyalty	0.792	0.723	0.899	0.748	0.798

We have decided to show bar charts of the total effects compared to the values obtained for the global model (see figures 9.6-9.10) in order to provide a visual method to appreciate the differences between segments.

Figure 9.6 shows the total effects of Image on Expectations of the four final segments using the effects of the global model as reference values. It can be observed that the Expectations of the customers with a basic-to-medium education level (node 4) are less driven by Image than the global model. The same effect, but in a stronger way, can be appreciated in the segment of "older" customers that are not retired. Conversely, segments of nodes 5 and 6 show path coefficients larger than the global model, meaning that they are more driven by image than the global model.

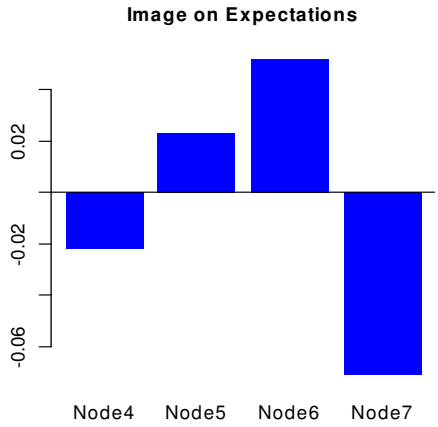


Figure 9.6. *Total effects on Expectations: Comparison with respect to the global model*

The total effects on Quality of the four final segments are shown in figure 9.7.

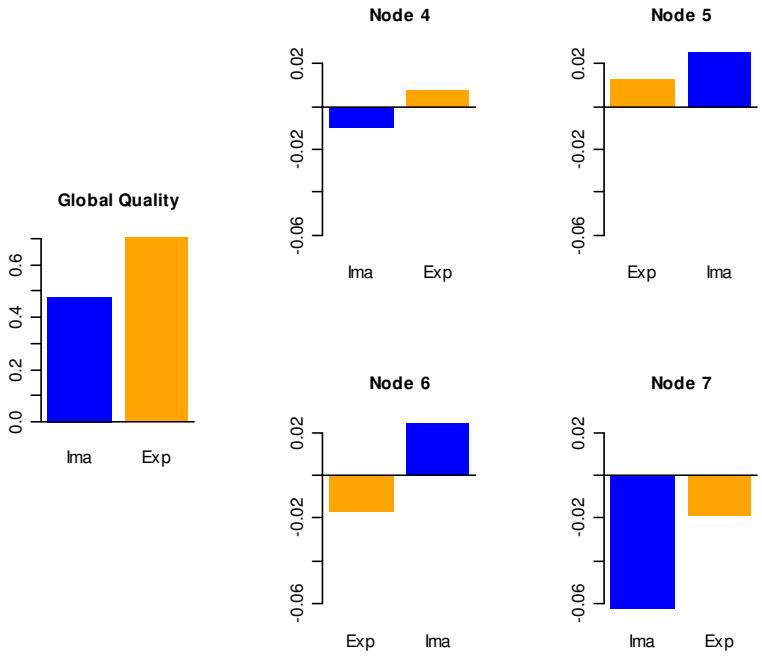


Figure 9.7. *Total effects on Quality: Comparison of the four segments with respect to the global model*

Node 4 presents total effects very similar to the global model. Customers aged 55 years or younger with higher education (node 5) have total effects of Expectations and Image on Quality greater than that of the global model. In the case of node 6 (retired customers) Expectations show a value smaller than the root node, however the total effect of Image is larger than in the global model. About node 7 (customers with ages of 56 years or older which are still working), both Image and Expectations present a total effect smaller than in the global model.

Figure 9.8 presents the effects on the Perceived Value. We can see that Node 4 has similar values to those of the global model. In contrast, the segment of customers with an age of 55 years or younger with a high degree of education (node 5) differs from the rest of the nodes because it has the stronger influences on Value. In the case of the retired customers (node 6) the Perceived Value is the least driven by Quality than the rest of the segments. Finally, node 7 is characterized by showing smaller direct effects on Value in all of the latent constructs.

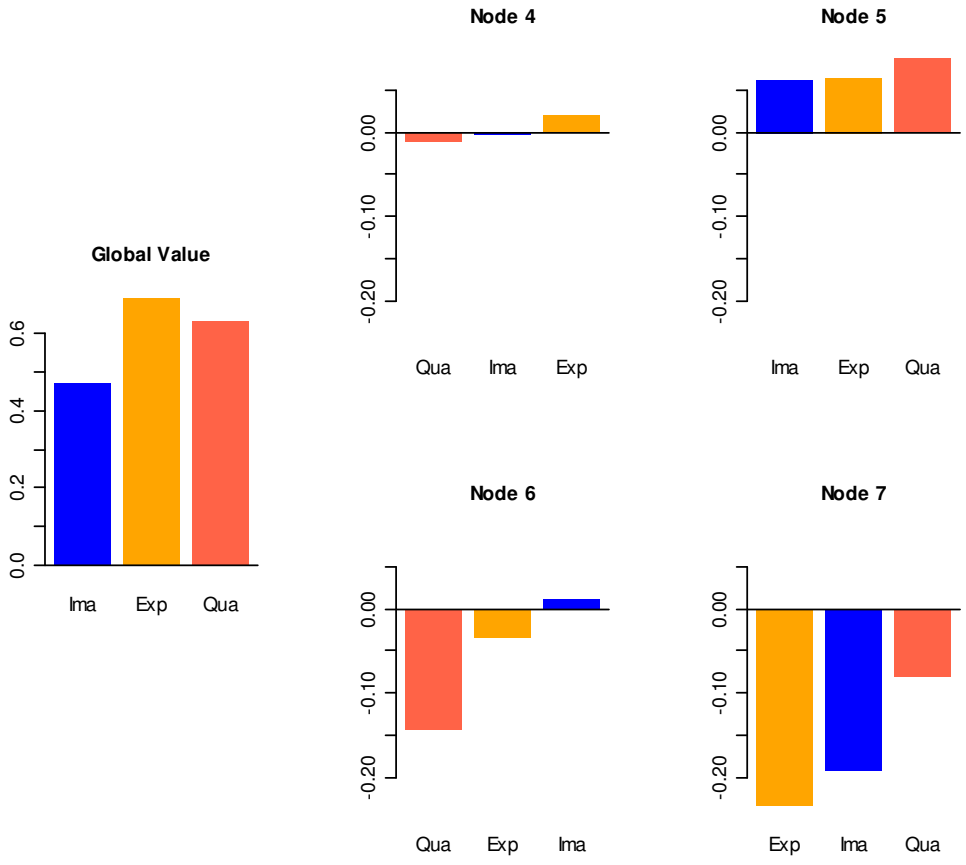


Figure 9.8. *Total effects on Value: Comparison of the four segments with respect to the global model*

The total effects on Satisfaction (see figure 9.9) display an analogous pattern as in the effects of the Perceived Value. The segment of customers with an age of 55 years or younger without higher education shows total effects that behave similarly to the global model. Instead, the path coefficients in node 5 (i.e., customers with a college or university degree) indicate that these customers differ from the rest of the segments because their Satisfaction is the one with higher effects.

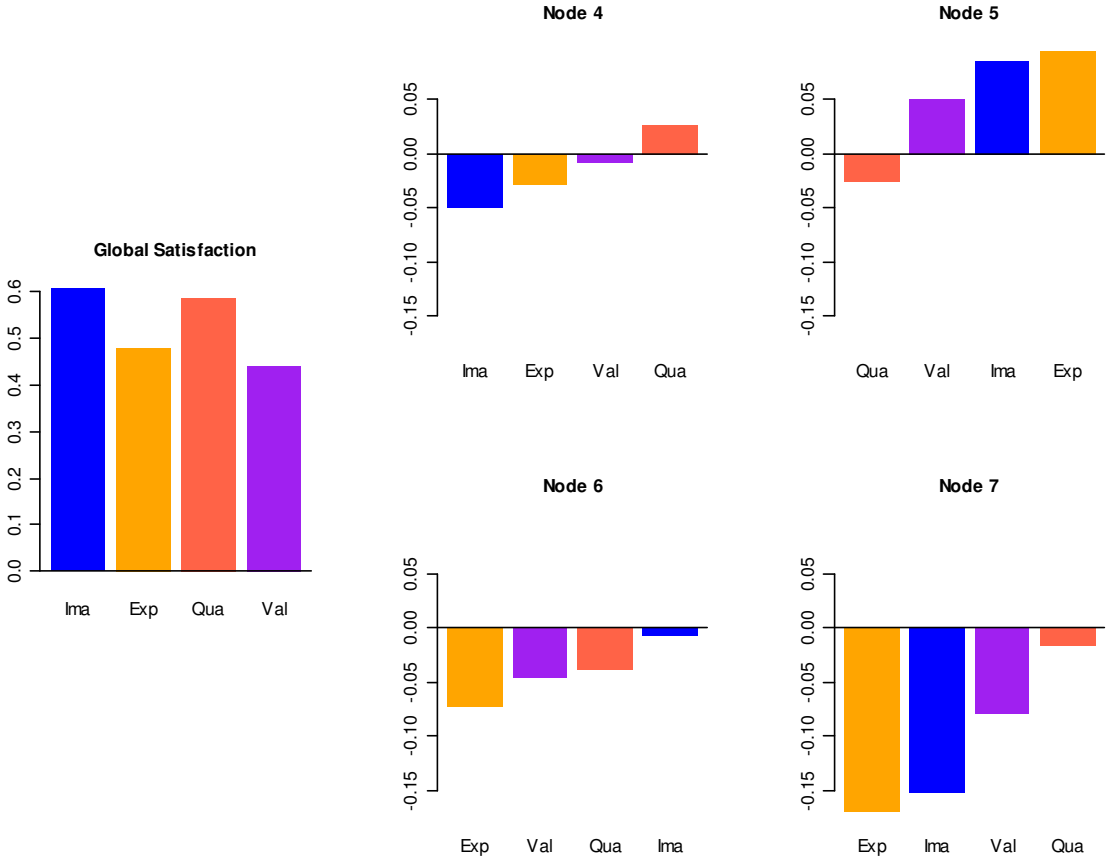


Figure 9.9. *Total effects on Satisfaction: Comparison of the four segments with respect to the global model*

The Satisfaction in the segment of retired customers is less influenced by Expectations than in the global model. The influence of Image remains similar to that of the global model. In the case of the customers with an age of 56 years or older that are not retired their Satisfaction contain the weakest effects of all segments. Specifically, this node is the least affected by Expectations and Image.

The total effects on Loyalty are shown in figure 9.10. Compared to the global model, the Loyalty of customers in node 4 has a smaller influence on Satisfaction. In the case of the customers that have a high level of education they have larger direct effects on Loyalty than the global model. As it can be seen, this segment has the strongest influence of Expectations, Satisfaction and Perceived Value on Loyalty. This implies that the Loyalty in these customers, compared to the rest of segments, is strongly affected by Expectations and Satisfaction. Total effects on Loyalty in node 6 show similar patterns to those of node 4 but in a lower magnitude.

Finally, the segment of older customers that are not retired has a Loyalty less driven by Image and Expectations.

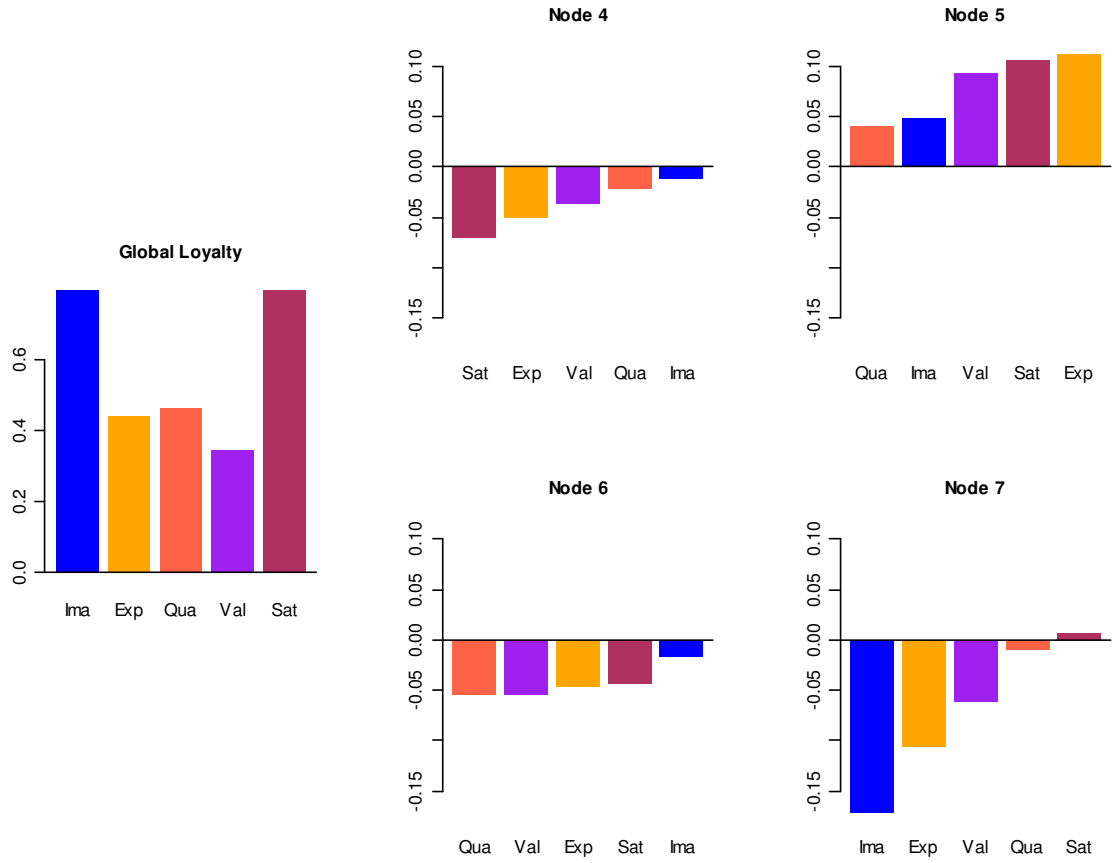


Figure 9.10. *Total Effects on Loyalty: Comparison of the four segments with respect to the global model*

Bootstrap validation

In order to evaluate the precision of the estimates obtained in each final segment we have applied bootstrapping with 100 samples. The results of the mean values of the bootstrap path coefficients are shown in table 9.13. Bootstrap confidence intervals in the fifth percentile and the 95th percentile are listed in table 9.14.

Table 9.13. *Bootstraps results of path coefficients (mean values)*

Path Relations	*Node4*	*Node5*	*Node6*	*Node7*
Image on Expectations	0.658	0.700	0.734	0.612
Expectations on Quality	0.709	0.721	0.685	0.687
Expectations on Value	0.266	0.238	0.312	0.091
Quality on Value	0.622	0.722	0.510	0.558
Image on Satisfaction	0.207	0.249	0.232	0.208
Expectations on Satisfaction	-0.007	0.090	0.029	-0.006
Quality on Satisfaction	0.337	0.213	0.347	0.368
Value on Satisfaction	0.425	0.494	0.377	0.361
Image on Loyalty	0.385	0.221	0.339	0.285
Satisfaction on Loyalty	0.716	0.898	0.747	0.788

Table 9.14. *Bootstrap Confidence Intervals for path coefficients (5 and 95 percentiles)*

Path Relations	*Node4*		*Node5*		*Node6*		*Node7*	
Image on Expec	0.599	0.713	0.633	0.774	0.647	0.822	0.515	0.700
Expec on Quality	0.631	0.768	0.664	0.765	0.624	0.753	0.599	0.787
Expec on Value	0.152	0.358	0.144	0.332	0.166	0.444	-0.040	0.203
Quality on Value	0.533	0.717	0.627	0.812	0.353	0.635	0.431	0.686
Image on Satisfac	0.103	0.287	0.156	0.337	0.032	0.406	0.070	0.338
Expec on Satisfac	-0.081	0.075	0.014	0.152	-0.085	0.142	-0.095	0.092
Quality on Satisfac	0.256	0.428	0.130	0.314	0.216	0.453	0.240	0.493
Value on Satisfac	0.350	0.496	0.419	0.558	0.232	0.511	0.262	0.461
Image on Loyal	0.310	0.452	0.119	0.314	0.215	0.482	0.157	0.451
Satisfac on Loyal	0.631	0.785	0.801	0.993	0.614	0.858	0.613	0.914

9.3 Application in Employee Satisfaction

The second example is an application in job satisfaction-motivation analysis from an organizational survey study of a Spanish banking entity. The measurement of employee satisfaction and employee motivation has a long tradition among psychologists, sociologists and human resources researchers (Thierry, 1998). These topics occupy an important place in different fields such as industrial-organizational psychology, social psychology, organizational behavior, personnel and human resource management, and organizational management (Drenth *et al*,1998; Anderson *et al*, 2002; Furnham, 2005).
To the best of our knowledge, the first application of PLS-PM to study a model related with employee satisfaction was carried out by Igbara and Greenhaus (1992). They proposed a model of turnover intentions among management information systems (MIS) employees. More applications of PLS-PM within turnover of Information Technology workers are exposed in Thatcher, Stepina and Boyle (2003), Taylor and Chin (2004), and Thatcher *et al* (2006). In a more organizational and psychological approach, a model entirely focused on job satisfaction is found in Staples and Higgins (1998). Money and Graham (1999) tested a causal model of salesperson performance and satisfaction using data collected in Japan and the United States. They considered national culture as a moderator construct and they compared parameter estimates across the cultural groups.

Another remarkable application of PLS-PM related with job satisfaction-motivation analysis involves the European Employee Index^{TM} (EEI) (Eskildsen *et al*, 2003, 2004a, 2004b; Kristensen *et al*, 2004). It can be regarded as the counterpart of the ECSI model in customer satisfaction analysis. However, the EEI is not yet considered a standard model for job satisfaction and it is not well established as the ECSI model. The EEI was developed in a joint research project between the market research companies MarkedsConsult A/S and CFI Group. It is an improved version of the Danish Employee Index which was a pilot study conducted in Denmark in 2000. In 2001 the index was widened to include Sweden, Norway and Finland, and it was introduced as the Nordic Employee Index. Currently, the European Employee Index is under further expansion to include more European countries. An interesting adaptation of the European Performance Satisfaction Index (EPSI) and the European Employee Index can be found in an article of Känd and Rekor (2005) who proposed a model to study job satisfaction among nurses in Estonia.

9.3.1 The Model

The proposed structural model is an adaptation of the models presented in Eskildsen *et al* (2003), Eskildsen *et al* (2004a,b), Kristensen and Westlund (2004), and Känd and Rekor (2005). Similar models are found in Gaertner, (1999), Currivan, (1999), and Kim (1999). The model aims to measure employee satisfaction, employee motivation, and loyalty as well as their antecedents. According to Eskildsen *et al* (2004b) predictors of job satisfaction, motivation and loyalty can be grouped in to four main characteristics of the job and work environment:

- Organizational vision: is the area that focuses on the cultural/ethical aspects of the organization, and the ability of corporate management to make sound decisions.
- Superiors: this aspect focuses on the relationship that the employee has to the immediate manager (i.e., perceived professional and leadership skills of the manager)
- Conditions of work: this area focuses on the job content, the physical work environment, the pay and the benefit package.
- Corporate Image: refers to the brand name and the kind of associations employees get from working in the company.

The model includes eight constructs five of which are exogenous and three are endogenous. The exogenous constructs are Empowerment, Image, Pay, Work Conditions, and Leadership. These latent variables are assumed to cover the four characteristics of the job and work environment. 'Empowerment' is associated to Organizational vision; 'Image' is related to Corporate Image, 'Pay' and 'Work Conditions' are associated to Conditions of work; and 'Leadership' is related to Superiors. The endogenous constructs are: Employee Satisfaction, Employee Motivation, and Loyalty. Table 9.15 contains the description of each the eight latent constructs of the model.

Table 9.15. *Definitions of Constructs in the Employee Satisfaction-Motivation Model*

Construct	*Definition*
Empowerment	Perceptions of autonomy, initiative, responsibility, recognition
Image	Degree to which an employee feels identified with the organization
Pay	Remuneration for work performed in an organization
Work Conditions	Perceptions of the workplace conditions and facilities
Leadership	Degree of consideration expressed from an employee in a subordinate position
Satisfaction	Degree to which an employee has positive emotions toward the work role
Motivation	Extent to which an employee is willing to perform the work role
Loyalty	Degree to which an employee feels loyalty to the organization

The proposed causal model is shown in figure 9.11.

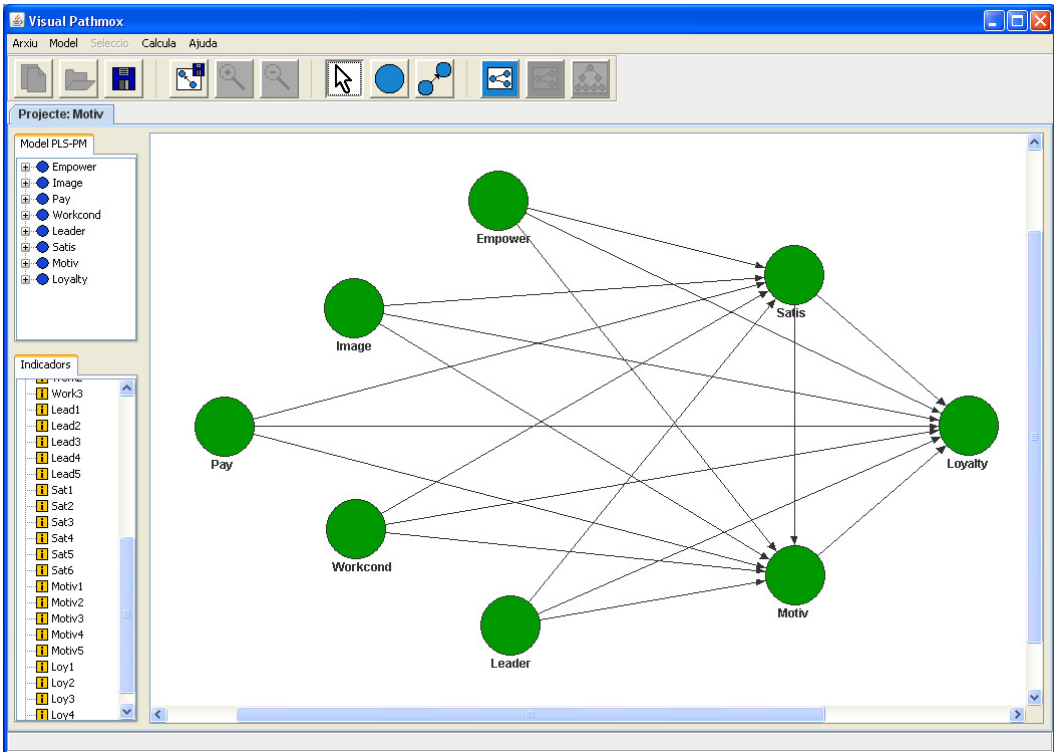


Figure 9.11. *Path diagram of the employee satisfaction-motivation model*

The model captures employee satisfaction, motivation and loyalty based on the organization's image, empowerment, pay, work conditions, and leadership. By letting all the exogenous latent variables be related to the three endogenous constructs, the aim is to analyze the influence of the different drivers of job satisfaction and motivation. Beyond proposing a causal model on job satisfaction-motivation, our purpose is to show the application of PATHMOX. The emphasis on this study is on the performance of the PATHMOX algorithm, aiming to detect different segments with their own path models.

9.3.2 Data

The data used in this example was obtained from an organizational study of a Spanish financial institution performed in 2001. However, due to confidentiality reasons, the complete details of the survey-based study will not be mentioned. Only the description of the variables is given. A total of 44 variables, measured on 8020 employees, were selected from the original database. The set of 44 variables are divided in two groups. One group is formed by 5 segmentation variables, and the other group is formed by 39 indicator variables for the structural model.

Manifest Variables

The 39 manifest variables used in this model are described in table 9.16. The variables are measured on a five-point Likert scale ranging from (1) strongly disagree to (5) strongly agree.

Table 9.16. *Description of the manifest variables for each of the latent constructs*

Construct	*Description of Indicators*
Empowerment	1) Participating in decision-making process
	2) Opinion of employees is taken into consideration
	3) Recognition by the performed work
	4) Considering employees as responsible workers
	5) Empower team work
	6) Autonomy is favored
	7) Confidence in performed tasks
	8) Creativity and initiative are favored
Image	1) Organization's reputation
	2) Organization's values
	3) Organization's customer relationships
	4) Organization's internal relationships
	5) Organization's external projection
Pay	1) Salary
	2) Social Benefits
	3) My salary is in accordance with my duties
	4) My salary is in accordance with my effort
Work Conditions	1) Enough personnel in the office
	2) Enough time to perform the tasks
	3) Conditions and tools to perform the work
Leadership	1) Agenda and planning
	2) Receptiveness
	3) Encouraging
	4) Communication
	5) Celebrating success
Satisfaction	1) Overall rating of satisfaction
	2) Tasks in accordance to employee capabilities
	3) Possibility to know efficiency
	4) Possibility to learn new things
	5) Usefulness of performed job
	6) Fulfillment of expectations
Motivation	1) Wake up feeling all's going well
	2) Feeling cheerful after job
	3) Not feeling frustrated with job
	4) Not being irritable at the job
	5) Overall rating of motivation
Loyalty	1) Not willing to leave in case of finding another alternative
	2) Committed with the organization's success
	3) Willingness to make an extra effort without remuneration

Segmentation variables
The five segmentation variables, which serve as observed sources of heterogeneity, are: Gender, Age, Job Level, Seniority, and Type. Job Level refers to the position in the employee rank hierarchy. Seniority indicates an employee's length of service. Type is a special variable whose real meaning has been "masked" due to confidentiality reasons.

9.3.3 Results of the global PLS path model

Measurement Model
The measurement relationships of the latent variables and their blocks of indicators were taken in a reflective way. As in the model of customer satisfaction, we assessed the measurement in the following three aspects:

- Unidimesionality of blocks of indicators
- Reliability of indicators
- Differentiation between latent variables

Unidimensionality of block of indicators
As we have before, unidimensionality is checked by means of three indices: 1) principal component analysis, 2) Cronbach's α, and 3) Dillon-Goldstein's ρ. Table 9.17 shows the eigenvalues from the principal component analysis, the Cronbach's α values, and the Dillon-Goldstein's ρ. The first eigenvalue of the manifest variables correlation matrix in each construct is larger than one. The second eigenvalue is smaller than one except in the case of Empowerment. However, when examining all the Cronbach's α it is observed that they meet the threshold value of 0.70. Similarly, Dillon-Goldstein's ρ values are also above 0.70 for each construct. Thus, unidimensionality of blocks of indicators is supported.

Table 9.17. *Different measures to assess unidimensionality of blocks of indicators*

Constructs	*Number of indicators*	*First Eigenvalue*	*Second Eigenvalue*	*Cronbach's* α	*Dillon-Goldstein* ρ
Empowerment	8	4.19	1.15	0.87	0.90
Image	5	2.88	0.92	0.79	0.86
Pay	4	2.40	0.81	0.77	0.85
Work Conds	3	1.86	0.67	0.69	0.83
Leadership	5	3.68	0.43	0.91	0.93
Satisfaction	6	3.36	0.66	0.84	0.88
Motivation	5	2.80	0.66	0.80	0.86
Loyalty	5	1.77	0.70	0.65	0.81

Reliability of indicators
The reliability of each manifest variable can be assessed by examining: 1) the loadings of the indicators with their respective latent constructs (see table 9.18); 2) the correlations of the indicators with their associated latent variables (see table 9.19); and 3) the communalities of the indicators (fourth column of table 9.18) It can be seen from the communalities that eight indicators (empo4, empo5, empo6, ima2, ima4, pay2,

motiv2 and motiv4) are below the recommended 0.5 value. However, their communality indices are so close to 0.5 that we do not consider them to be so low as to be eliminated.

Table 9.18. *Results of the measurement model: outer weights, loadings, and communalities of each manifest variable*

Indicator	*Outer Weight*	*Loading*	*Commu nality*	*Indicator*	*Outer Weight*	*Loading*	*Commu nality*
empo1	0.207	0.733	0.538	lead1	0.241	0.851	0.725
empo2	0.212	0.746	0.557	lead2	0.212	0.833	0.694
empo3	0.215	0.718	0.515	lead3	0.255	0.899	0.808
empo4	0.137	0.668	0.447	lead4	0.227	0.842	0.708
empo5	0.158	0.706	0.499	lead5	0.230	0.861	0.742
empo6	0.135	0.693	0.480				
empo7	0.153	0.738	0.544				
empo8	0.168	0.753	0.567	sat1	0.248	0.715	0.511
				sat2	0.223	0.794	0.630
ima1	0.574	0.744	0.553	sat3	0.223	0.717	0.514
ima2	0.218	0.641	0.411	sat4	0.226	0.756	0.572
ima3	0.219	0.717	0.513	sat5	0.208	0.748	0.560
ima4	0.195	0.688	0.473	sat6	0.210	0.755	0.570
ima5	0.198	0.721	0.520				
				motiv1	0.213	0.709	0.502
pay1	0.319	0.807	0.651	motiv2	0.254	0.706	0.498
pay2	0.344	0.655	0.428	motiv3	0.317	0.821	0.675
pay3	0.271	0.761	0.579	motiv4	0.201	0.674	0.454
pay4	0.367	0.848	0.719	motiv5	0.335	0.815	0.664
work1	0.402	0.776	0.602	loy1	0.544	0.802	0.643
work2	0.448	0.835	0.697	loy2	0.362	0.723	0.523
work3	0.420	0.749	0.561	loy3	0.397	0.762	0.580

Differentiation between constructs

The third aspect consists in evaluating the extent to which a given construct differentiates from the others. Table 9.19 shows the correlations between the indicators and the latent variables. The correlations of the indicators with their associated construct are highlighted in bold. In each row of the table, it is expected the highlighted correlation is the highest value. As it can be appreciated, all indicators have their highest value on the latent variables they intend to measure.

Table 9.19. *Correlations between the manifest variables and the latent constructs*

	Empo wer	Image	Pay	Work Conds	Leader ship	Satis faction	Motiva tion	Loyalty
empo1	**0.733**	0.274	0.258	0.380	0.450	0.632	0.457	0.463
empo2	**0.746**	0.267	0.230	0.384	0.449	0.653	0.468	0.468
empo3	**0.718**	0.278	0.329	0.406	0.522	0.617	0.508	0.502
empo4	**0.669**	0.221	0.198	0.300	0.262	0.342	0.345	0.356
empo5	**0.706**	0.267	0.215	0.316	0.375	0.431	0.375	0.392
empo6	**0.693**	0.222	0.203	0.267	0.298	0.369	0.316	0.342
empo7	**0.738**	0.257	0.210	0.312	0.326	0.406	0.371	0.383
empo8	**0.753**	0.264	0.227	0.333	0.354	0.450	0.393	0.432
ima1	0.419	**0.744**	0.307	0.341	0.209	0.501	0.421	0.552
ima2	0.127	**0.641**	0.125	0.112	0.097	0.157	0.165	0.234
ima3	0.160	**0.717**	0.125	0.126	0.100	0.175	0.162	0.223
ima4	0.139	**0.688**	0.096	0.115	0.093	0.149	0.145	0.204
ima5	0.145	**0.721**	0.112	0.131	0.096	0.153	0.144	0.208
pay1	0.206	0.195	**0.807**	0.201	0.106	0.335	0.147	0.242
pay2	0.260	0.256	**0.655**	0.263	0.147	0.304	0.205	0.296
pay3	0.239	0.175	**0.761**	0.213	0.153	0.257	0.156	0.218
pay4	0.304	0.201	**0.848**	0.339	0.184	0.342	0.232	0.285
work1	0.354	0.225	0.266	**0.776**	0.219	0.365	0.331	0.304
work2	0.407	0.223	0.283	**0.835**	0.225	0.389	0.398	0.326
work3	0.366	0.251	0.247	**0.749**	0.190	0.376	0.320	0.350
lead1	0.463	0.176	0.191	0.272	**0.851**	0.386	0.375	0.298
lead2	0.448	0.150	0.116	0.200	**0.833**	0.338	0.338	0.251
lead3	0.500	0.178	0.169	0.237	**0.899**	0.425	0.376	0.318
lead4	0.457	0.178	0.181	0.224	**0.842**	0.381	0.321	0.296
lead5	0.467	0.175	0.166	0.214	**0.861**	0.372	0.351	0.288
sat1	0.509	0.405	0.361	0.409	0.305	**0.715**	0.568	0.570
sat2	0.524	0.283	0.322	0.385	0.335	**0.794**	0.486	0.470
sat3	0.609	0.272	0.244	0.366	0.390	**0.717**	0.436	0.460
sat4	0.538	0.309	0.256	0.383	0.363	**0.757**	0.481	0.495
sat5	0.463	0.343	0.241	0.304	0.279	**0.749**	0.467	0.497
sat6	0.508	0.256	0.393	0.284	0.321	**0.755**	0.404	0.476
motiv1	0.314	0.239	0.171	0.321	0.212	0.346	**0.709**	0.348
motiv2	0.394	0.286	0.183	0.352	0.271	0.422	**0.706**	0.409
motiv3	0.511	0.280	0.222	0.344	0.352	0.610	**0.821**	0.513
motiv4	0.330	0.189	0.107	0.308	0.255	0.325	**0.674**	0.303
motiv5	0.533	0.356	0.207	0.347	0.400	0.590	**0.815**	0.574
loy1	0.590	0.361	0.278	0.391	0.360	0.665	0.543	**0.802**
loy2	0.309	0.417	0.235	0.210	0.181	0.391	0.392	**0.723**
loy3	0.398	0.379	0.269	0.318	0.197	0.406	0.396	**0.762**

Structural model

The results of the structural model involve the estimates of the path coefficients, the R^2 values for each endogenous construct, and the correlations among latent variables. The indirect effects and the total affects of the relationships among latent variables are calculated as well.

Path coefficients

First, we examined the path coefficients of the structural model (see figure 9.12 and table 9.20). They can be divided in regard to the three endogenous latent variables. When examining the path coefficients on Employee Satisfaction it can be appreciated that Empowerment has the largest value (0.498) whereas Leadership has the smallest value (0.082).

Motivation is influenced by Satisfaction (0.397), and followed by Empowerment (0.154) and Work Conditions (0.145). In this case, Pay has a negative value (-0.064) although it is a very small influence. Finally, Loyalty is mainly affected by Satisfaction (0.300) and Motivation (0.219) which is in accordance with the theoretical framework (i.e. Loyalty is the psychological response to Satisfaction and Motivation).

However, the interesting aspect consists of evaluating the influences of the characteristics of the job and work environment. Among all job characteristics, Image is the one with the highest coefficient (0.217). In second place we find Empowerment (0.167). Finally, it can be noted that Leadership has a negative path coefficient (-0.035), although it has the lowest influence in absolute value.

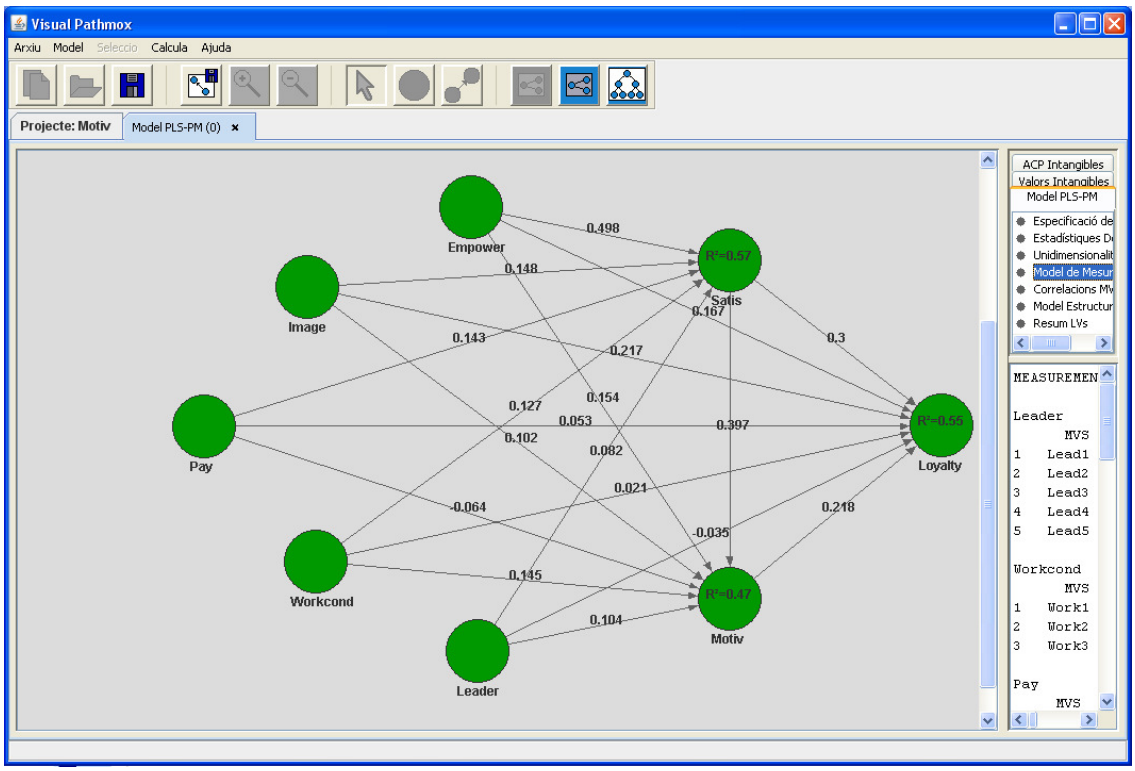


Figure 9.12. *Results of the structural model in Visual Pathmox. Path coefficients shown on the arrows. R^2 values inside the latent variables*

Table 9.20. *Direct, indirect, and total effects among latent constructs*

Construct		*Estimate*	*Std. Error*	*t value*	*Pr(>|t|)*
Satisfaction	$R^2 = 0.57$				
	Empowerment	0.498	0.010	50.057	0.000
	Image	0.148	0.008	18.376	0.000
	Pay	0.143	0.008	17.726	0.000
	Work Conds	0.127	0.009	14.689	0.000
	Leadership	0.082	0.009	9.365	0.000
Motivation	$R^2 = 0.47$				
	Empowerment	0.154	0.013	12.197	0.000
	Image	0.102	0.009	11.216	0.000
	Pay	-0.064	0.009	-7.083	0.000
	Work Conds	0.145	0.010	15.044	0.000
	Leadership	0.104	0.010	10.641	0.000
	Satisfaction	0.397	0.012	32.116	0.000
Loyalty	$R^2 = 0.55$				
	Empowerment	0.167	0.012	14.268	0.000
	Image	0.217	0.008	25.770	0.000
	Pay	0.053	0.008	6.271	0.000
	Work Conds	0.021	0.009	2.350	0.019
	Leadership	-0.035	0.009	-3.843	0.000
	Satisfaction	0.300	0.012	24.787	0.000
	Motivation	0.219	0.010	21.232	0.000

R^2: Predictive power

The measures of the predictive power of the model are the R^2 values of the endogenous latent variables. Satisfaction has an R^2 of 0.57, which indicates that 57 percent of the variance in Satisfaction is explained by its predictive constructs. Motivation in turn has an R^2 of 0.47 (i.e., 47 percent of the variance in Motivation is explained by the model). Finally, 63 percent of the variance in Loyalty is explained by the rest of latent variables. We consider the R^2 values of the model as adequate values of its predictive power.

Direct effects, indirect effects and total effects

Besides path coefficients we can also look at the indirect and total effects. Recall that indirect effects comprise the indirect paths from one variable to another. The total effects are the sum of direct and indirect effects. Table 9.21, contains the three types of effects in the path relationsships of the structural model.

Among the indirect effects on motivation, the effect of Empowerment is the largest one (0.198), followed by Image (0.059) and Pay (0.057). Leadership has the smallest indirect effect (0.033). In respect to Loyalty, the main indirect influence is due to Satisfaction (0.087) although Pay and Image have similar influences to satisfaction.

Table 9.21. *Direct, indirect, and total effects among latent constructs*

Path relations	*Direct*	*Indirect*	*Total*
Empowerment on Satisfaction	0.498	0.000	0.498
Image on Satisfaction	0.148	0.000	0.148
Pay on Satisfaction	0.143	0.000	0.143
Work Conds on Satisfaction	0.127	0.000	0.127
Leadership on Satisfaction	0.082	0.000	0.082
Empowerment on Motivation	0.154	0.198	0.352
Image on Motivation	0.102	0.059	0.161
Pay on Motivation	-0.064	0.057	-0.008
Work Conds on Motivation	0.145	0.050	0.196
Leadership on Motivation	0.104	0.033	0.136
Satisfaction on Motivation	0.397	0.000	0.397
Empowerment on Loyalty	0.167	0.226	0.394
Image on Loyalty	0.217	0.079	0.297
Pay on Loyalty	0.053	0.041	0.094
Work Conds on Loyalty	0.021	0.081	0.102
Leadership on Loyalty	-0.035	0.054	0.020
Satisfaction on Loyalty	0.300	0.087	0.387
Motivation on Loyalty	0.219	0.000	0.219

The three types of effects on satisfaction, motivation and loyalty are represented in figures 9.13, 9.14, and 9.15, respectively. In the case of satisfaction there are no indirect effects since it is only affected by exogenous latent variables. As we have seen from the path coefficients Satisfaction is mainly driven by Empowerment. In contrast, Leadership has the smallest influence on Satisfaction.

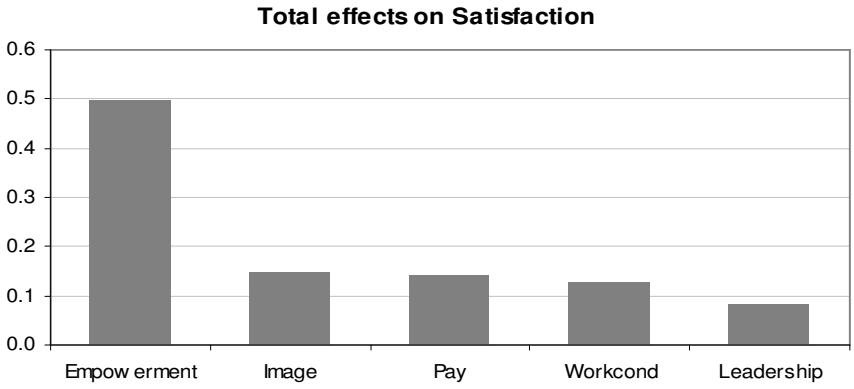


Figure 9.13. *Total effects on employee satisfaction (direct effects)*

In figure 9.14 we can appreciate that Motivation is mostly driven by Satisfaction. However, taking into consideration the indirect effects, Empowerment is the second latent variable that drives Motivation. An interesting situation occurs with Pay which has positive indirect effect of 0.057. However, it has a negative direct influence of -0.064, resulting in an almost null but negative total effect of -0.008.

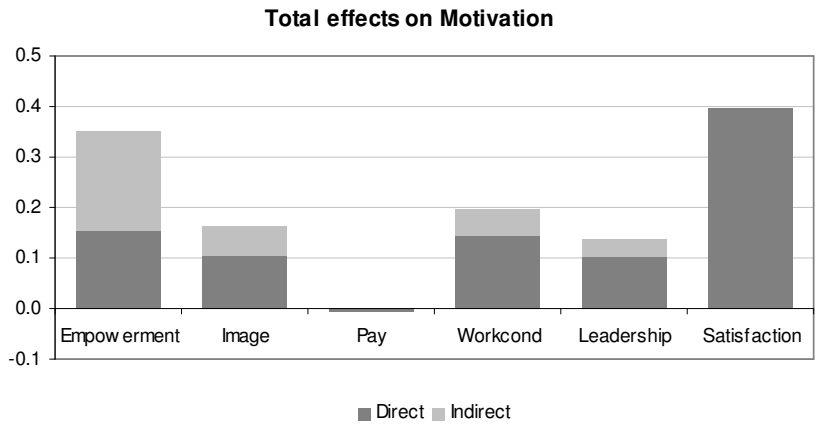


Figure 9.14. *Total effects on employee motivation (direct and indirect effects)*

Among the indirect effects on Loyalty (figure 9.15), the most important is due to Empowerment, followed by Work Conditions and Image. In fact, the total effect of Empowerment is slightly larger than that of Satisfaction. Note that Leadership has a negative direct influence of -0.035, but a positive indirect influence of 0.054. Hence, Leadership has a positive total effect of 0.020, which is the smallest one.

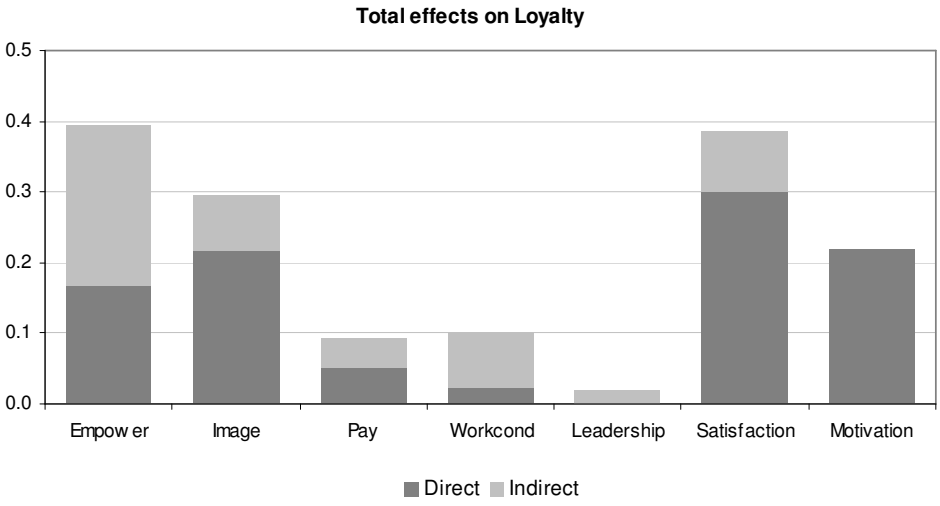


Figure 9.15. *Total effects on Loyalty (direct and indirect effects)*

Bootstrap validation

To validate the obtained path coefficients we used bootstrapping with 100 samples. The results are contained in table 9.22. The column 'Original' refers to the original estimate of the path coefficients. The column 'Bootstrap' contains the mean value of the 100 bootstrap samples. The standard deviation appears in the fourth column. The last two columns have the 5 and 95 percentiles of the confidence interval. Note that the bootstrap interval of 'Pay on Motivation' does not contain the zero, which indicates that Pay has a (low) negative influence on Motivation. The same can be said about 'Leadership on Loyalty' with a negative value.

Table 9.22. *Bootstraps results of path coefficients (direct effects)*

Path relations	*Original*	*Mean.Boot*	*Std.Dev*	*perc05*	*perc95*
Empower on Satisfac	0.498	0.498	0.010	0.483	0.514
Image on Satisfac	0.148	0.147	0.009	0.133	0.159
Pay on Satisfac	0.143	0.143	0.010	0.128	0.158
Work Conds on Satisfac	0.127	0.127	0.010	0.110	0.144
Leadership on Satisfac	0.082	0.081	0.010	0.066	0.096
Empower on Motiv	0.154	0.153	0.016	0.131	0.181
Image on Motiv	0.102	0.101	0.010	0.084	0.119
Pay on Motivation	-0.064	-0.065	0.010	-0.081	-0.046
Work Conds on Motiv	0.145	0.145	0.011	0.130	0.164
Leadership on Motiv	0.104	0.104	0.010	0.089	0.121
Satisfac on Motiv	0.397	0.398	0.016	0.372	0.423
Empower on Loyalty	0.167	0.169	0.014	0.149	0.196
Image on Loyalty	0.217	0.218	0.010	0.200	0.233
Pay on Loyalty	0.053	0.052	0.010	0.037	0.069
Work Conds on Loyalty	0.021	0.021	0.011	0.004	0.038
Leadership on Loyalty	-0.035	-0.036	0.009	-0.053	-0.022
Satisfac on Loyalty	0.300	0.299	0.015	0.276	0.321
Motiv on Loyalty	0.219	0.218	0.011	0.200	0.236

Correlations among latent variables

The correlations among the latent variables are listed in table 9.23.

Table 9.23. *Correlations between latent variables in the Customer Satisfaction model*

	Empower ment	Image	Pay	Work Conds	Leader ship	Satis faction	Motiva tion	Loyalty
Empowerment	1							
Image	0.359	1						
Pay	0.332	0.271	1					
WorkConds	0.478	0.296	0.337	1				
Leadership	0.545	0.200	0.193	0.268	1			
Satisfaction	0.704	0.419	0.407	0.479	0.445	1		
Motivation	0.575	0.370	0.245	0.445	0.411	0.638	1	
Loyalty	0.591	0.498	0.343	0.415	0.339	0.664	0.594	1

Redundancy

Another measure of the predictive power of the model is given by the Redundancy. It reflects the ability of a set of independent latent variables to explain variation in the indicators of a dependent latent variable. Table 9.24 contains the redundancy measures for each endogenous block. It can be observed that the block of motivation has the lowest redundancy values, which is in accordance with its R^2. Conversely, manifest variables in satisfaction and loyalty have the highest redundancy values.

Table 9.24. *Redundancy of indicators*

Indicator	*Redun.*	*Indicator*	*Redun.*	*Indicator*	*Redun.*
sat1	0.290	motiv1	0.236	loy1	0.311
sat2	0.358	motiv2	0.234	loy2	0.150
sat3	0.292	motiv3	0.318	loy3	0.208
sat4	0.325	motiv4	0.214		
sat5	0.318	motiv5	0.313		
sat6	0.324				

9.3.4 PATHMOX segmentation tree

The application of the PATHMOX algorithm requires defining the codification of the segmentation variables based on their type of scale (see table below). It also requires determining the parameters of the stopping rules:

- *p*-value = 0.05
- Minimum number of individuals inside a node = 10%
- Depth level of the tree = 3

Table 9.25. *Codification of segmentation variables according to their type of scale*

Segmentation Variable	*Scale*
Gender	Binomial
Age	Ordinal
Job level	Ordinal
Seniority	Ordinal
Type	Binomial

The obtained PATHMOX tree appears in figure 9.16. The root node is associated to the global model which is calculated over the entire sample of 8020 employees. The first split is caused by the segmentation variable 'Job Level' with a corresponding *F*-statistic's *p*-value of 0. It separates the 2650 Managers (in node 2) from the rest of the 5370 employees (in node 3).

In regards to node 2, it is split by gender into 272 female managers (node 4), and 2378 male managers (node 5). The number of female managers represents a proportion of 3.4% of the total sample. Since this proportion is below the predefined 10%, this node is declared as a leaf node. Node 5, in contrast, continues splitting in nodes 10 and 11 in terms of 'Seniority'. Node 10 represents the path model of male managers segment that have been working since 1993 or earlier (1563 employees). Node 11 represents the path model of the segment with male managers that have been working since 1994 or later (825 employees).

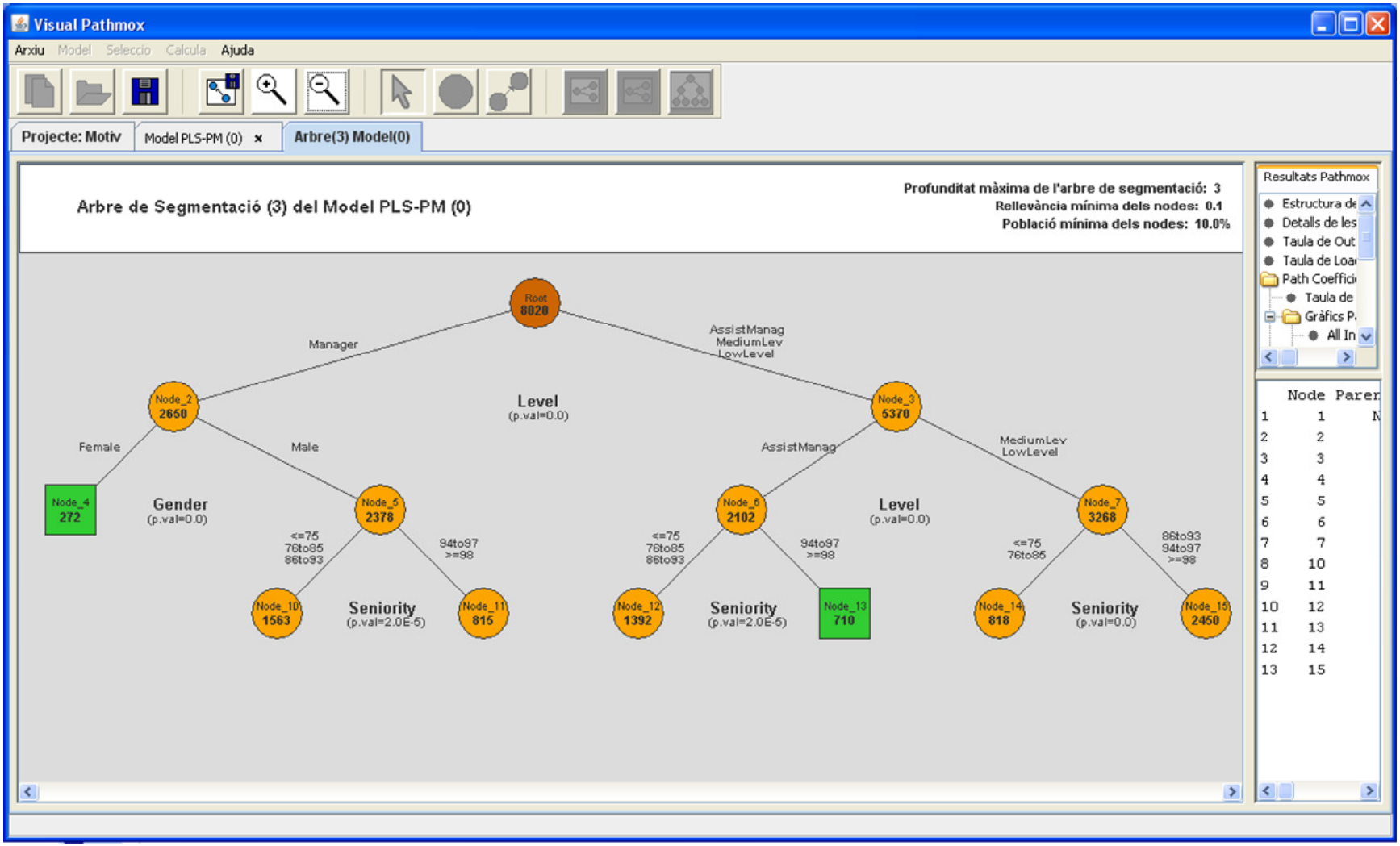


Figure 9.16. *PATHMOX segmentation tree in Visual Pathmox*

With respect to node 3, it is divided into nodes 6 and 7 by the segmentation variable 'Job Level'. Node 5 is associated with the path model of assistant managers (2102 employees) whereas node 6 is associated with the path model of medium and low level employees (3268 employees). The *p*-vale of the *F*-test in this split is zero. Node 6 splits assistant managers in nodes 12 and 13 by dividing those assistants that have been working since 1993 or earlier (1392 employees) from those that have been working since 1994 or later (710 employees). Since the sample size of node 13 represents a proportion of 8.7% of the total population, this node is declared as a leaf node.

Finally, Node 7 is also partitioned in terms of seniority as in the cases of nodes 5 and 6. Node 14 represents the path model of medium and low level employees with a seniority level of 1985 or earlier (818 employees). Node 15 represents the path model that is associated with the medium and low level employees that have been working since 1986 or later (2450 employees). The final nodes in the tree are considered as the final segments: Node 4, Node 10, Node 11, Node 12, Node 13, Node 14, and Node 15.

The path coefficients of the structural model for each final segment, as well as the path coefficients of the model for the total sample (global model), are listed in table 9.26. In turn, the path diagrams of the corresponding structural models are in Appendix II.

Table 9.26. *Path coefficients of the path models corresponding to the root node and the terminal segments*

Path Relations	*Global*	*Node4*	*Node10*	*Node11*	*Node12*	*Node13*	*Node14*	*Node15*
Empo on Sat	0.498	0.389	0.440	0.459	0.508	0.396	0.526	0.544
Imag on Sat	0.148	0.217	0.207	0.197	0.161	0.143	0.139	0.080
Pay on Sat	0.143	0.025	0.126	0.090	0.149	0.137	0.099	0.111
Work on Sat	0.127	0.235	0.143	0.184	0.096	0.240	0.162	0.128
Lead on Sat	0.082	0.116	0.066	0.049	0.067	0.070	0.114	0.122
Empo on Mot	0.154	0.170	0.158	0.079	0.115	0.168	0.077	0.142
Imag on Mot	0.102	-0.023	0.086	0.113	0.074	0.080	0.149	0.096
Pay on Mot	-0.064	-0.048	-0.058	-0.012	-0.028	0.007	-0.021	-0.022
Work on Mot	0.145	-0.108	0.060	0.053	0.146	0.134	0.170	0.143
Lead on Mot	0.104	0.232	0.115	0.106	0.100	0.165	0.081	0.099
Sat on Mot	0.397	0.434	0.384	0.370	0.394	0.325	0.424	0.436
Empo on Loy	0.167	0.289	0.139	0.074	0.199	0.156	0.117	0.194
Imag on Loy	0.217	0.203	0.175	0.174	0.200	0.254	0.228	0.228
Pay on Loy	0.053	-0.180	0.099	0.094	0.081	0.054	0.151	0.042
Work on Loy	0.021	0.061	-0.037	0.015	-0.038	-0.043	0.037	0.051
Lead on Loy	-0.035	0.081	0.038	0.041	-0.002	-0.011	-0.098	-0.069
Sat on Loy	0.300	0.231	0.331	0.337	0.318	0.304	0.291	0.228
Mot on Loy	0.218	0.191	0.195	0.227	0.143	0.151	0.195	0.268

The total effects of the structural relationships are displayed in table 9.27. We have decided to use these total effects in order to graphically illustrate the differences among segments. To accomplish this goal, bar charts of the total effects compared to the values obtained for the global model are displayed in figures 9.17-9.19.

Table 9.27. *Total Effects of the path relations corresponding to the root node and the terminal segments*

Path Relations	*Root*	*Node4*	*Node10*	*Node11*	*Node12*	*Node13*	*Node14*	*Node15*
Empo on Sat	0.498	0.389	0.440	0.459	0.508	0.396	0.526	0.544
Imag on Sat	0.148	0.217	0.207	0.197	0.161	0.143	0.139	0.080
Pay on Sat	0.143	0.025	0.126	0.090	0.149	0.137	0.099	0.111
Work on Sat	0.127	0.235	0.143	0.184	0.096	0.240	0.162	0.128
Lead on Sat	0.082	0.116	0.066	0.049	0.067	0.070	0.114	0.122
Empo on Mot	0.352	0.339	0.327	0.249	0.315	0.297	0.300	0.379
Imag on Mot	0.161	0.071	0.165	0.186	0.137	0.126	0.208	0.131
Pay on Mot	-0.007	-0.037	-0.010	0.021	0.031	0.052	0.021	0.026
Work on Mot	0.195	-0.006	0.115	0.121	0.184	0.212	0.239	0.199
Lead on Mot	0.137	0.282	0.140	0.124	0.126	0.188	0.129	0.152
Sat on Mot	0.397	0.434	0.384	0.370	0.394	0.325	0.424	0.436
Empo on Loy	0.359	0.411	0.318	0.267	0.389	0.296	0.314	0.382
Imag on Loy	0.274	0.271	0.259	0.257	0.260	0.304	0.280	0.256
Pay on Loy	0.108	-0.172	0.150	0.132	0.137	0.102	0.188	0.080
Work on Loy	0.070	0.135	0.021	0.092	-0.002	0.042	0.098	0.095
Lead on Loy	-0.003	0.117	0.065	0.062	0.023	0.014	-0.055	-0.027
Sat on Loy	0.387	0.314	0.406	0.421	0.374	0.353	0.374	0.345
Mot on Loy	0.218	0.191	0.195	0.227	0.143	0.151	0.195	0.268

Figure 9.17 shows the total influences of the exogenous constructs on Satisfaction for the seven segments. The total effects of the global model appear in the first bar chart, which are taken as reference values in order to compare them with the rest of the segments. As it can be seen, the charts of the final segments present some of their bars below, and others above, the horizontal axis. This help us to see which constructs are below the global model, as well as which are above.

It can be observed that the Satisfaction of the female managers segment (node 4) is less driven by Empowerment and Pay. In contrast, female managers' satisfaction is more influenced by Image and Work Conditions. Satisfaction in the segment of male managers that have been working since 1993 or earlier (node 10) has an important influence because of Image. However, influences on Satisfaction due to Empowerment, Pay and Leadership, are below the global reference values.

Among the other segments, it is remarkable that the Satisfaction of the assistant managers that have been working since 1994 or later (node 13) is less influenced by Empowerment. Instead, they present a high influence from Work Conditions. The other constructs (Image, Pay, and Leadership) affecting satisfaction are very similar to the global model. Node 12, which is the segment formed by assistant managers that have been working since 1993 or earlier, can be considered the node most similar to the global model in terms of the total effects on Satisfaction.

The segment of medium or low levels of senior employees (node 14) is characterized by having more influence of Work Conditions than that observed for the global model. However, the effect of Pay is clearly inferior. Workers in node 15 (integrated by medium or low level employees working since 1986 or later) consider Image and Pay less important to be satisfied. Conversely, they give more importance to Leadership and Empowerment.

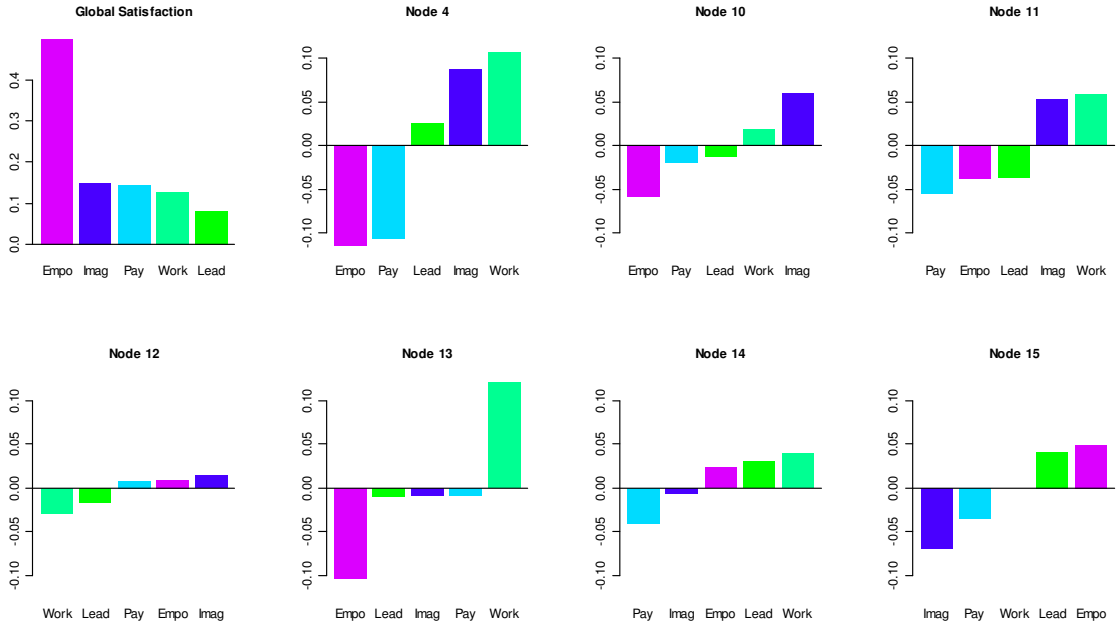


Figure 9.17. *Effects on Satisfaction: Comparison of the seven segments with respect to the global model*

In the case of Motivation, the segment of female managers has four of its exogenous total effects below those values of the global model. Indeed, the total effect which mainly distinguishes this segment is due to Work Conditions. Nevertheless, it also has the largest effect above the global values with the Leadership construct.

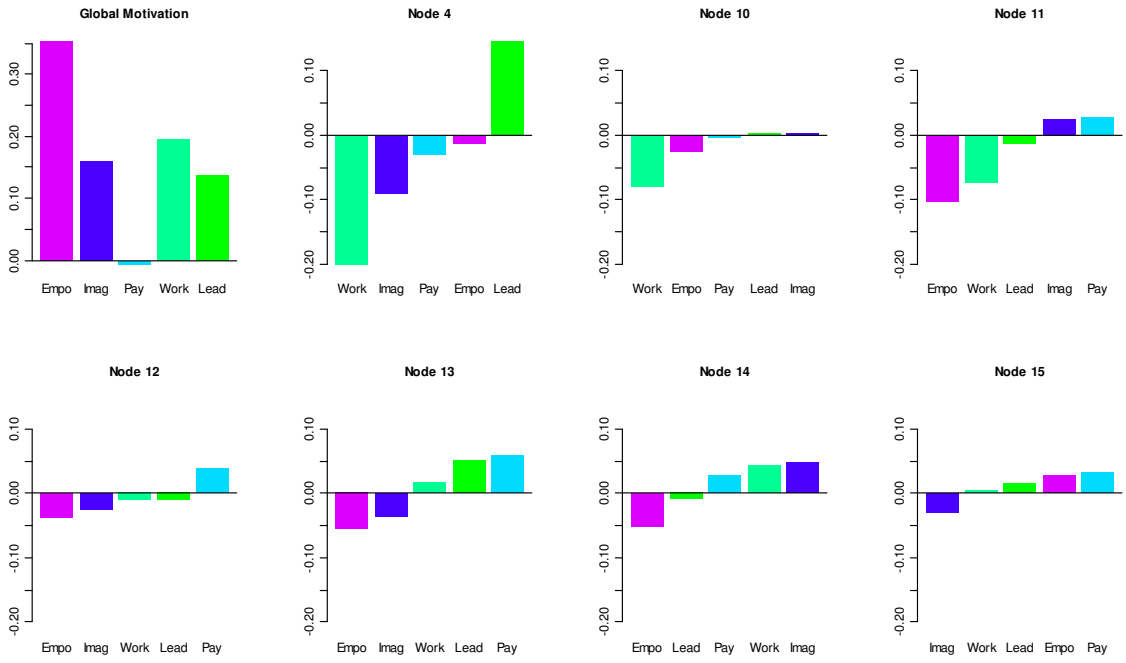


Figure 9.18. *Effects on Motivation: Comparison of the seven segments with respect to the global model*

The segment of male managers that have been working since 1993 or earlier (node 10) is also differentiated from the global model in the effect of Work Conditions. Although

it has a total effect below the horizontal axis, it is not as serious as node 4. In a similar way, the male managers that have been working since 1994 or later (node 11), have their motivation less influenced by the Empowerment and Work Conditions. To a lesser extent, the senior assistant managers coincide with nodes 10 and 11.

Regarding node 13, its Motivation is affected more by Leadership and Pay, but less by Empowerment and Image. The segment of medium or low levels of senior employees (node 14) is characterized by having more influence of Work Conditions and Image than that observed for the global model, but less than Empowerment. Finally, node 15 shows total effects of Pay and Empowerment greater than the global model. However, its Image is below the reference value.

Regarding the comparative bar charts among segments in terms of Loyalty (figure 9.19), the segment of female managers is the most differentiated from the global model. Taking into account the total effects of the exogenous constructs, its Loyalty is mainly affected by Leadership, and to a less extent, Work Conditions and Empowerment. However, it is also negatively affected by Pay.

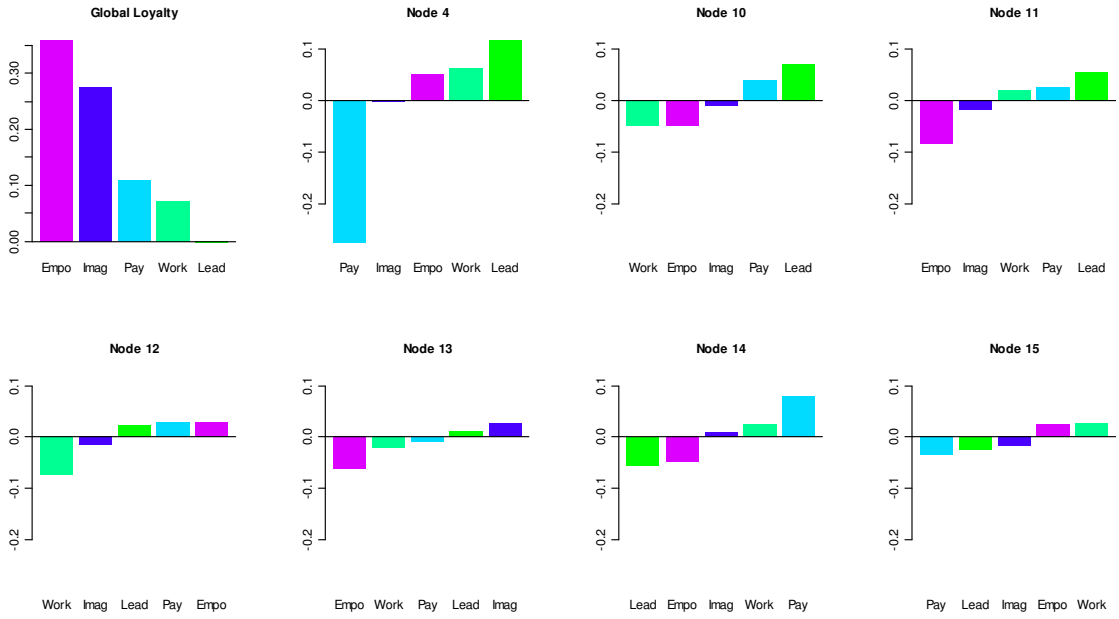


Figure 9.19. *Effects on Loyalty: Comparison of the seven segments with respect to the global model*

The segment of male managers with more seniority (node 10) presents a negative path coefficient of Work Conditions on Loyalty. In contrast, the influences of Pay and Leadership over Loyalty are above the values of the global model. Similar characteristics are found in the complementary segment of male managers (node 11). This segment, in particular, has the smallest path coefficient of Empowerment on Loyalty.

The segments of assistant managers (nodes 12 and 13) have lower influences on Loyalty of Work Conditions and Empowerment, respectively. With respect to the segment of medium and low level employees with more seniority (node 14), the influence of Pay has the most important effect on Loyalty. However, Leadership and

Empowerment influence below the global model values. Finally, the segment of medium and low employees with less seniority (node 15) shows a similar behavior of the total effects as in the global model.

Bootstrap validation

In order to evaluate the precision of the estimates obtained in each of the final segment we have applied bootstrapping with 100 samples. The results of the mean values of the bootstrap path coefficients are shown in table 9.28. Bootstrap confidence intervals within the 5 percentile and the 95 percentile are listed in table 9.29.

Table 9.28. *Bootstraps results of path coefficients (mean values)*

Path Relations	*Node4*	*Node10*	*Node11*	*Node12*	*Node13*	*Node14*	*Node15*
Empo on Sat	0.384	0.440	0.461	0.508	0.395	0.522	0.548
Imag on Sat	0.236	0.209	0.202	0.163	0.141	0.142	0.079
Pay on Sat	0.036	0.123	0.089	0.151	0.135	0.102	0.108
Work on Sat	0.234	0.146	0.186	0.097	0.248	0.167	0.127
Lead on Sat	0.107	0.068	0.044	0.065	0.071	0.111	0.122
Empo on Mot	0.174	0.159	0.085	0.115	0.167	0.079	0.142
Imag on Mot	-0.018	0.089	0.113	0.070	0.079	0.145	0.096
Pay on Mot	-0.050	-0.061	-0.014	-0.027	0.005	-0.019	-0.023
Work on Mot	-0.107	0.058	0.048	0.150	0.131	0.172	0.144
Lead on Mot	0.230	0.117	0.106	0.097	0.156	0.085	0.099
Sat on Mot	0.429	0.384	0.376	0.395	0.334	0.421	0.434
Empo on Loy	0.292	0.130	0.085	0.200	0.158	0.118	0.197
Imag on Loy	0.200	0.181	0.173	0.199	0.252	0.232	0.231
Pay on Loy	-0.179	0.098	0.099	0.082	0.054	0.151	0.037
Work on Loy	0.060	-0.039	0.012	-0.038	-0.038	0.035	0.054
Lead on Loy	0.081	0.040	0.034	-0.004	-0.015	-0.099	-0.069
Sat on Loy	0.229	0.334	0.333	0.311	0.305	0.290	0.227
Mot on Loy	0.190	0.197	0.227	0.150	0.154	0.193	0.265

Table 9.29. *Bootstrap Confidence Intervals for path coefficients (5 and 95 percentiles)*

Path Relations	*Node4*		*Node10*		*Node11*		*Node12*		*Node13*		*Node14*		*Node15*	
Empo on Sat	0.293	0.476	0.387	0.486	0.399	0.512	0.468	0.544	0.332	0.446	0.477	0.561	0.515	0.577
Imag on Sat	0.165	0.306	0.173	0.240	0.174	0.241	0.129	0.199	0.101	0.196	0.104	0.174	0.055	0.106
Pay on Sat	-0.061	0.129	0.090	0.163	0.041	0.142	0.117	0.184	0.076	0.181	0.059	0.141	0.080	0.130
Work on Sat	0.136	0.327	0.109	0.183	0.140	0.239	0.063	0.125	0.194	0.300	0.118	0.216	0.095	0.155
Lead on Sat	0.021	0.201	0.032	0.113	0.002	0.098	0.038	0.094	0.024	0.114	0.072	0.151	0.092	0.146
Empo on Mot	0.069	0.302	0.103	0.216	0.008	0.156	0.061	0.166	0.088	0.237	0.019	0.147	0.103	0.187
Imag on Mot	-0.111	0.057	0.055	0.121	0.054	0.164	0.032	0.113	0.021	0.130	0.095	0.192	0.072	0.124
Pay on Mot	-0.148	0.045	-0.110	-0.019	-0.078	0.039	-0.070	0.018	-0.056	0.051	-0.079	0.038	-0.053	0.006
Work on Mot	-0.214	0.003	0.013	0.098	-0.015	0.105	0.116	0.186	0.075	0.189	0.129	0.223	0.113	0.176
Lead on Mot	0.121	0.326	0.063	0.155	0.053	0.162	0.060	0.142	0.101	0.218	0.033	0.137	0.064	0.126
Sat on Mot	0.303	0.543	0.326	0.427	0.305	0.449	0.349	0.443	0.248	0.415	0.346	0.494	0.388	0.470
Empo on Loy	0.203	0.386	0.093	0.171	0.025	0.143	0.141	0.252	0.111	0.209	0.037	0.189	0.155	0.238
Imag on Loy	0.109	0.285	0.146	0.217	0.122	0.212	0.162	0.239	0.194	0.300	0.172	0.278	0.206	0.258
Pay on Loy	-0.254	-0.100	0.064	0.131	0.043	0.155	0.037	0.120	-0.008	0.121	0.093	0.205	0.008	0.066
Work on Loy	-0.041	0.142	-0.079	-0.002	-0.051	0.064	-0.073	-0.001	-0.096	0.022	-0.019	0.082	0.029	0.081
Lead on Loy	0.011	0.156	0.001	0.075	-0.025	0.081	-0.037	0.026	-0.065	0.034	-0.147	-0.044	-0.099	-0.036
Sat on Loy	0.094	0.336	0.290	0.380	0.263	0.400	0.260	0.365	0.235	0.382	0.214	0.381	0.186	0.263
Mot on Loy	0.087	0.292	0.164	0.233	0.175	0.273	0.108	0.189	0.094	0.213	0.132	0.252	0.234	0.301

9.4 Some remarks

Path modeling segmentation trees are conceptually simple and easy to interpret. They offer a tool a meaningful description of population segments. Moreover, they meet some of the needs of the PLS-PM analysts and practitioners by providing a method to analyze their models while taking into account as much information as possible. One of the main characteristics of PATHMOX is its capability to simplify a complex reality for which no a priori classification is imposed, revealing hidden (unexpected) forms and isolating characteristics models. Another valuable aspect of PATHMOX is the ordering of possible binary splits according to their *p*-values in a given node. This order of the splits can contribute to understanding the behavior of different segments. We emphasize the use of PATHMOX as a data mining approach to identify unexpected models for segments of the population.

Chapter 10

Conclusions and Future Work

In this dissertation, a novel segmentation approach has been proposed for Partial Least Squares Path Modeling when observed sources of heterogeneity are available. The PATHMOX approach allows for obtaining segmentation trees of path models with the purpose of identifying population segments in the analyzed data.

10.1 Summary

To achieve the purpose of providing a new segmentation approach in Partial Least Squares Path Modeling, we have:

- Adapted the basic idea behind binary segmentation processes in order to produce a tree having a structure similar to a binary decision tree with different path models in each of the obtained nodes,
- Developed a binary segmentation algorithm which is capable of identifying segmentation variables in such a way that their binary splits give place to path models with as much difference as possible (in terms of path coefficients),
- Proposed a test for comparing structural models that consists of testing the equality of path coefficients between structural models; this has been done by adapting a test for assessing the equality of coefficients between two regression models,
- Used the adapted test as a split criterion in the binary segmentation algorithm in order to identify the most significant binary splits,
- Evaluated the performance of the test through a series of simulations by comparing two structural models under different experimental conditions (e.g., data distributions, sample size, path coefficients, disturbance terms for the endogenous construct, and measurement errors for the indicators),
- Launched the project to develop the software program *Visual Pathmox* which provides a graphical interface and makes the calculation of segmentation trees of path models feasible,

- Demonstrated the practical application of the PATHMOX approach using real data examples with a model of customer satisfaction and a model of employee satisfaction-motivation.

10.2 Conclusions and Contributions

One of the biggest challenges we faced in designing the PATHMOX algorithm was the development of a split criterion in order to decide if two confronted structural models could be considered to be different. To overcome this problem we have adapted a test for assessing the equality of two regression models (Lebart, 1985; Chow, 1960). In the present work, our goal has been the application of the test with the purpose of comparing two structural models by testing the equality of the path coefficients between path models.

To evaluate the performance of the proposed test, a series of simulation studies with experimental data have been undertaken. Different conditions have been investigated such as sample size, level of noise of the disturbance terms, number of elements in the model, distinct path coefficients, and distributions in data. The obtained results have shown that the test has an adequate performance, and that it can identify different structural models reasonably well.

Although the test relies on some assumptions about the distribution (normality) in the disturbance terms, and equality of variances, it has also been shown that the split criterion works reasonably well even when ignoring such assumptions. However, unbalanced segments and differences in the variance of the endogenous constructs may affect the sensitivity of the *F*-test.

In essence, all of the obtained results provide important evidence in favor of the adequate performance of the proposed *F*-test when it is applied to comparing two structural models. In other words, the results have shown that the test has an adequate performance, and that it can identify different structural models reasonably well. In this way, we have confidence in its use as a split criterion in the PATHMOX approach.

The PATHMOX algorithm breaks down the elements' (individuals') attributes making it easier to take a large customer base and form them into smaller, more specific segments according to their similarities. Being an automated method, PATHMOX does not capture expert's information that might guide the analysis of segments. In this sense, this approach is not intended to be used as a substitute for user knowledge. However, if no *a priori* knowledge is assumed about the existence of segments, PATHMOX can be of great help for experts in order to discover "unexpected" segments.

From the examples presented in this work, it is possible to observe that path modeling segmentation trees are conceptually simple and easy to interpret. One of the primary attributes of PATHMOX is the ability to convert a complex reality, for which no *a priori* classification is entailed, into a simpler one by detecting hidden patterns and isolating characteristics models. We believe that PATHOX is able to meet some of the practitioners' needs by offering a method to analyze their models while taking into account as much information as possible. The ordering of possible binary partitions according to their *p*-values in a given node may be helpful to understand the behavior of different segments. Thus, we recommend the utilization of PATHMOX as a data mining approach to identify unexpected models for segments of the population under analysis.

Pros

Among the advantages of PATHMOX we can identify the following:

- We emphasize the use of PATHMOX as a data mining approach to identify unexpected models for segments of the population and providing an intuitive scheme, with ease of interpretation and meaningful description
- Meet researcher's, analyst's and practitioner's needs to have an automated tool to analyze data
- Useful for managerial purposes: To help in decision making process, the segmentation trees obtained in PATHMOX are fully functional
- Accompanied by *Visual Pathmox* provides a graphical user-friendly interface

Contributions

We believe that this dissertation lays the foundation for a new approach in automated segmentation techniques in PLS-PM with observed heterogeneity. This dissertation makes several key contributions to the challenging topic of segmentation tasks in Partial Least Squares Path Modeling:

- Detection of segments with the help of observed sources of heterogeneity
- It presents a technique and tools for automated interpretation
- It shows that manageable segments can be identified with ease of interpretation. (i.e., examining the results of the pathmox tree the analyst obtains the characterization of the identified segments)
- It provides a validation tool by means of bootstrapping in order to obtain confidence intervals of the path coefficients in the final segments.

10.3 Future Research

This dissertation represents an important step towards enabling practitioners to identify segments in PLS path modeling with a scheme of observed heterogeneity. However, the work presented in this dissertation is not without its limitations. A great deal of work can be done to overcome these limitations and extend PATHMOX's capabilities.

- Further simulation studies should be undertaken to investigate the behavior of the *F* statistic with more complex models.
- A natural line of research consists of comparing PATHMOX with other PLS-PM segmentation methods.
- An interesting future direction would be to extend the PATHMOX algorithm to include Generalized Structured Component Analysis and covariance-based Structural Equation Modeling by Unweighted Least Squares (SEM-ULS) estimation. Since the *F*-test applied to compare path models only involves the path coefficients, we believe that it is possible to adapt the test to GSCA and SEM-ULS. Expanding PATHMOX trees to other structural equation modeling approaches will enable path modeling segmentation trees to evolve and have a wider application
- An important issue that has been side-stepped not only by PATHMOX, but by most PLS-PM segmentation approaches, is "how to handle the comparison of path models

taking into consideration both the measurement and the structural models?" An interesting future direction will be to investigate new criteria to make comparisons of path models in a feasible way, both in the measurement and the structural parts.

- A desirable extension of this work is to continue the development of *Visual Pathmox* to include more methodological capabilities and more exploitable features.

Despite the recent abundance of segmentation approaches in PLS-PM, this topic continues to be a challenging and open field for further contributions. Given the great complexity of structural equation models, more developments are needed to improve the processes of segments identification. We are sure that this area is still full of opportunities for new proposals.

Appendix I

Structural Models of the Segments in the Customer Satisfaction example (Chapter 9)

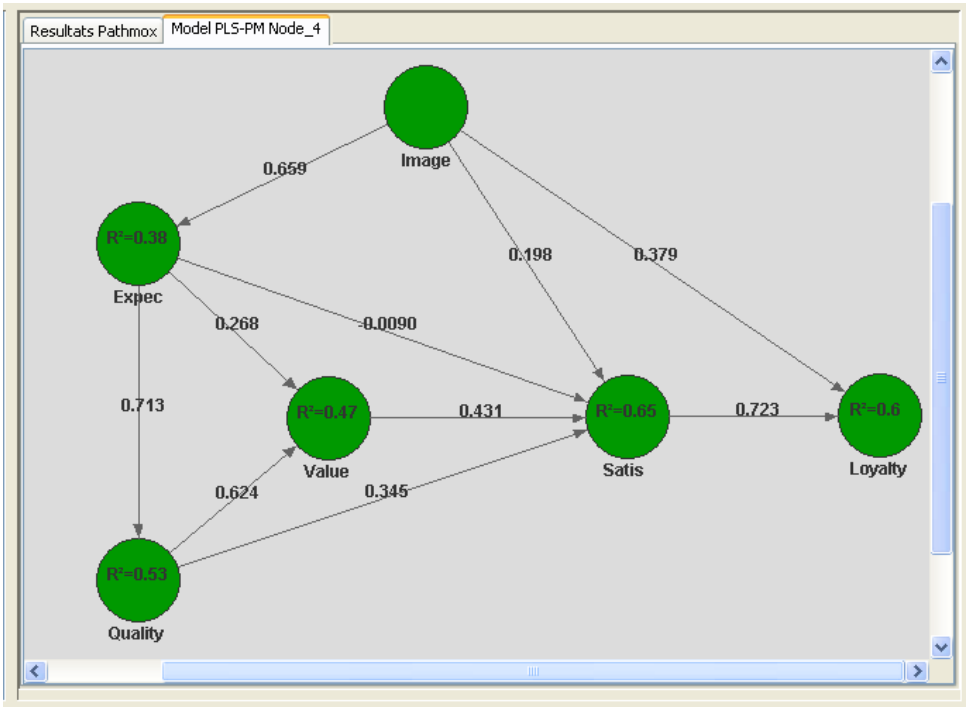


Figure A1. *Results of the structural model calculated in Visual Pathmox for node 4*

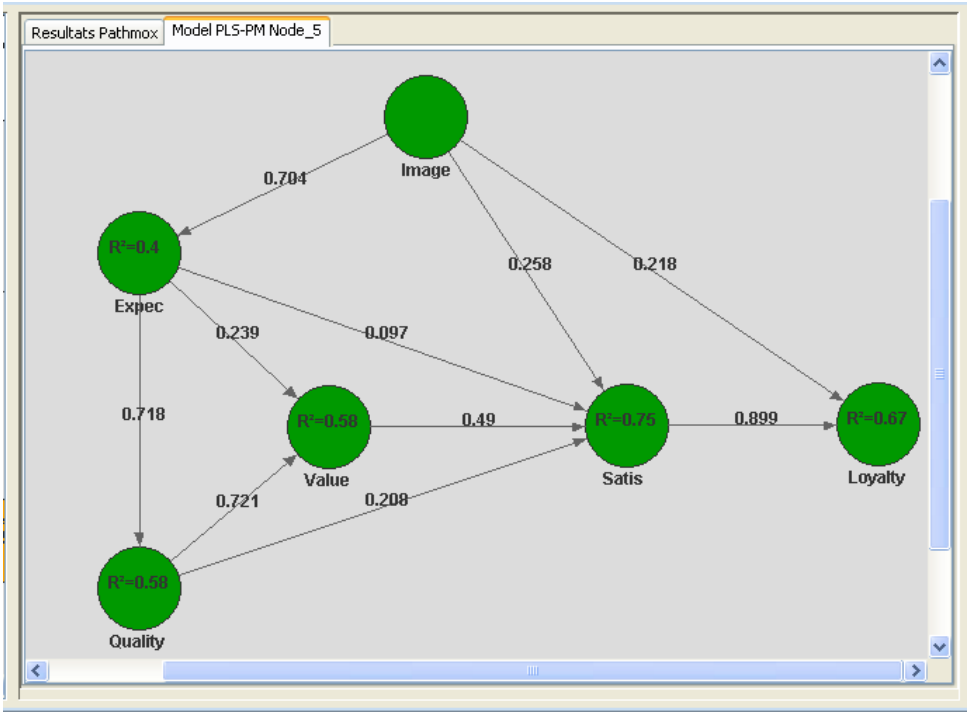


Figure A2. *Results of the structural model calculated in Visual Pathmox for node 5*

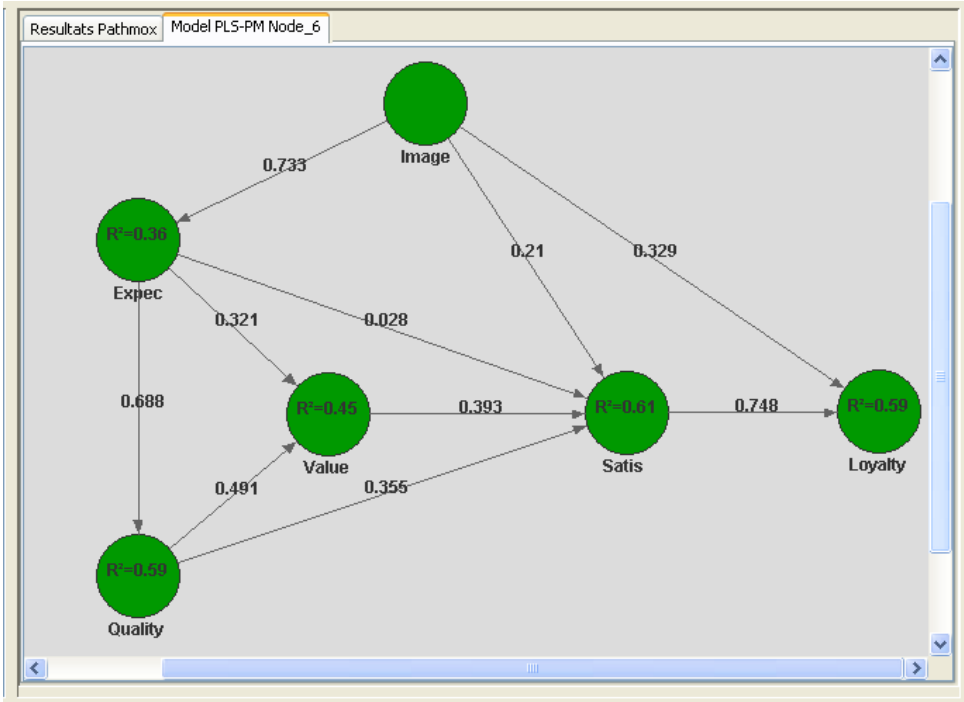


Figure A3. *Results of the structural model calculated in Visual Pathmox for node 6*

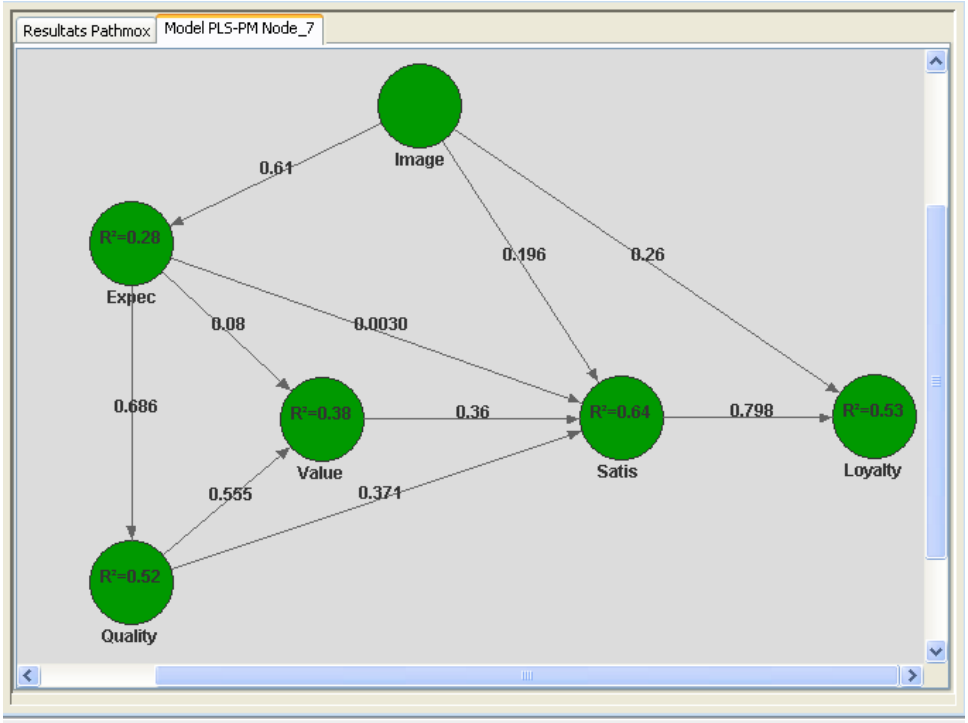


Figure A4. *Results of the structural model calculated in Visual Pathmox for node 7*

Appendix II

Structural Models of the Segments in the Employee Satisfaction-Motivation example of Chapter 9

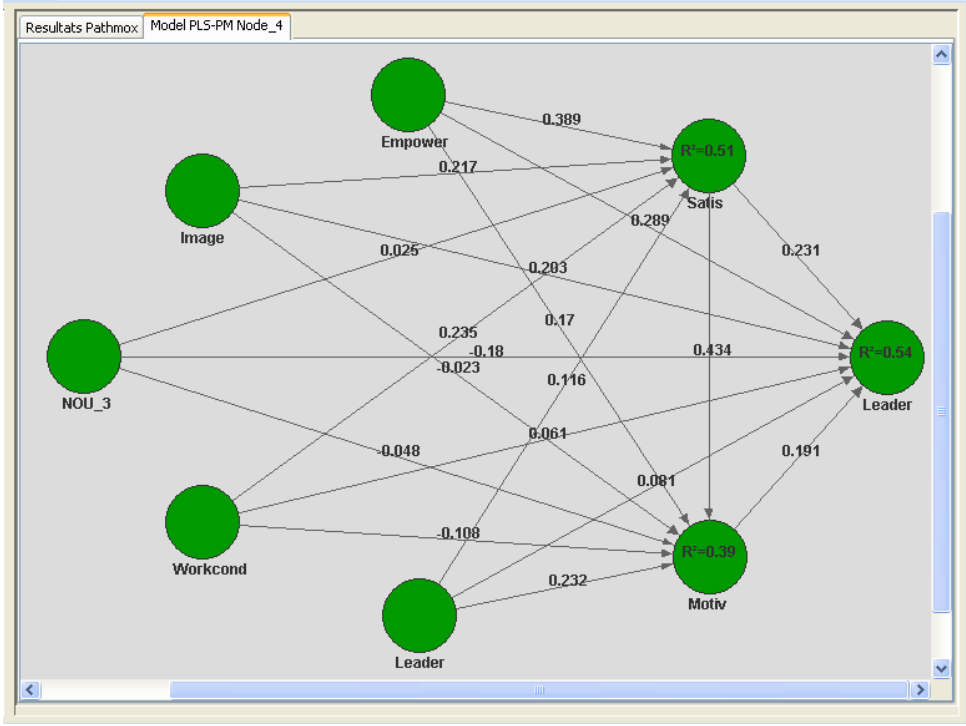


Figure B1. *Results of the structural model calculated in Visual Pathmox for node 4*

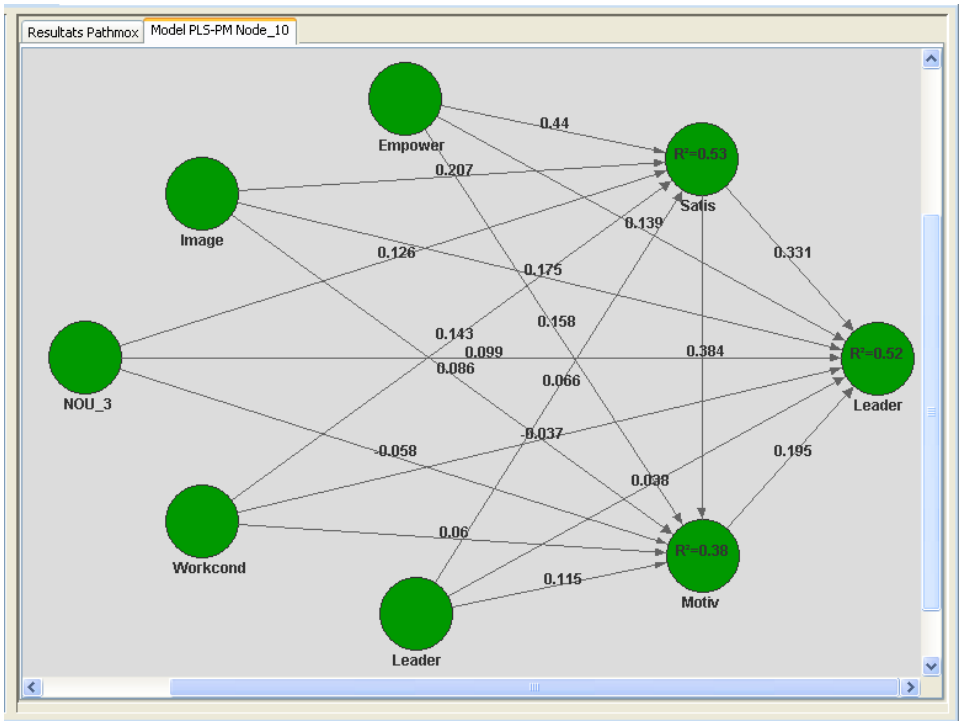


Figure B2. *Results of the structural model calculated in Visual Pathmox for node 10*

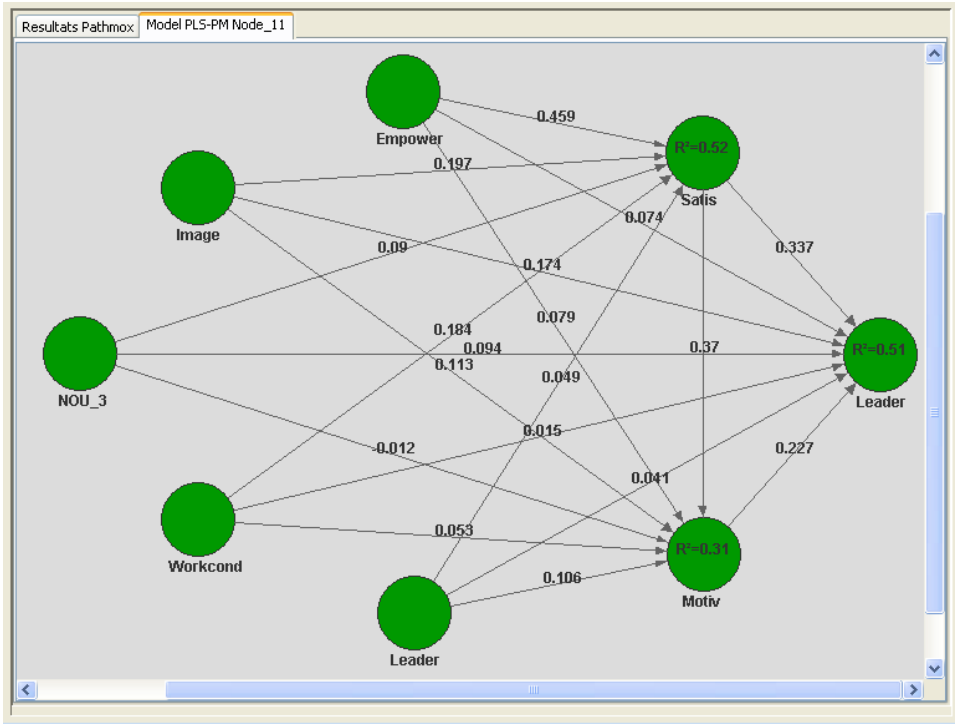


Figure B3. *Results of the structural model calculated in Visual Pathmox for node 11*

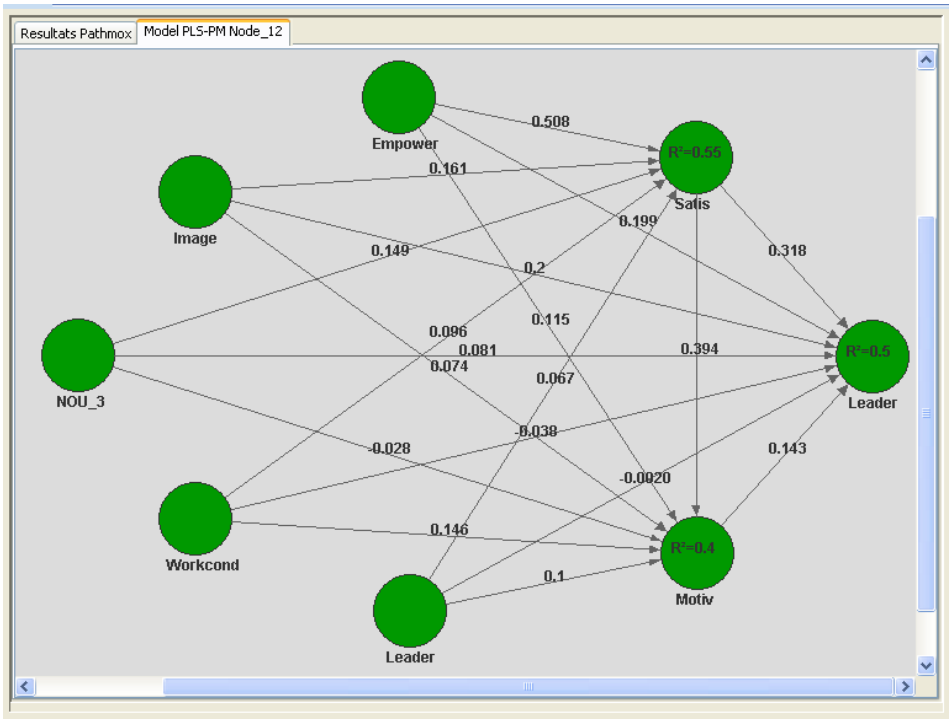


Figure B4. *Results of the structural model calculated in Visual Pathmox for node 12*

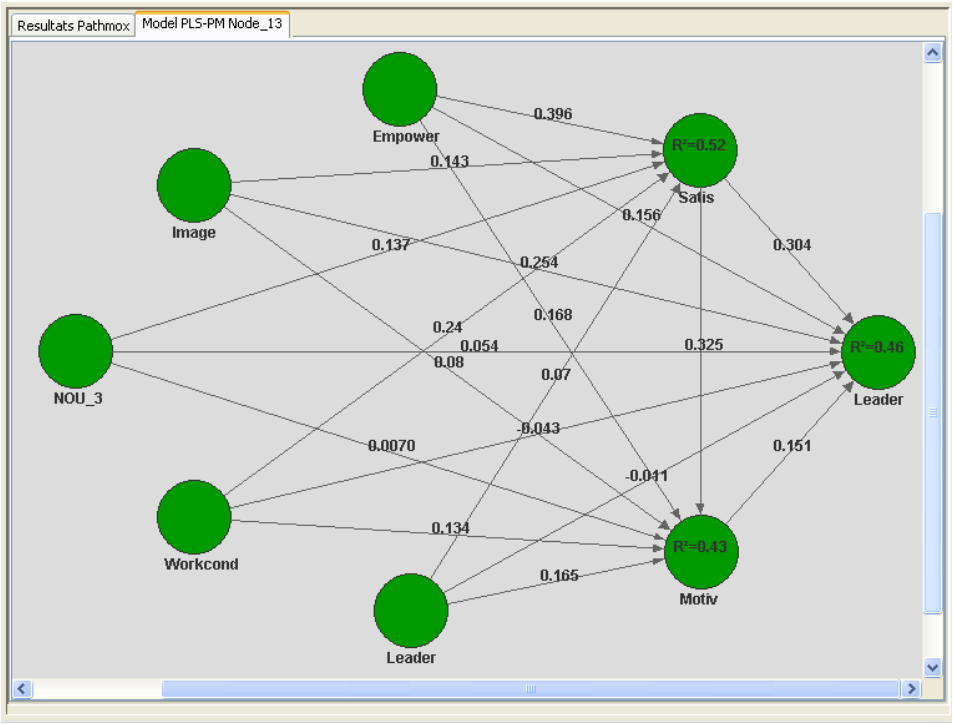


Figure B5. *Results of the structural model calculated in Visual Pathmox for node 13*

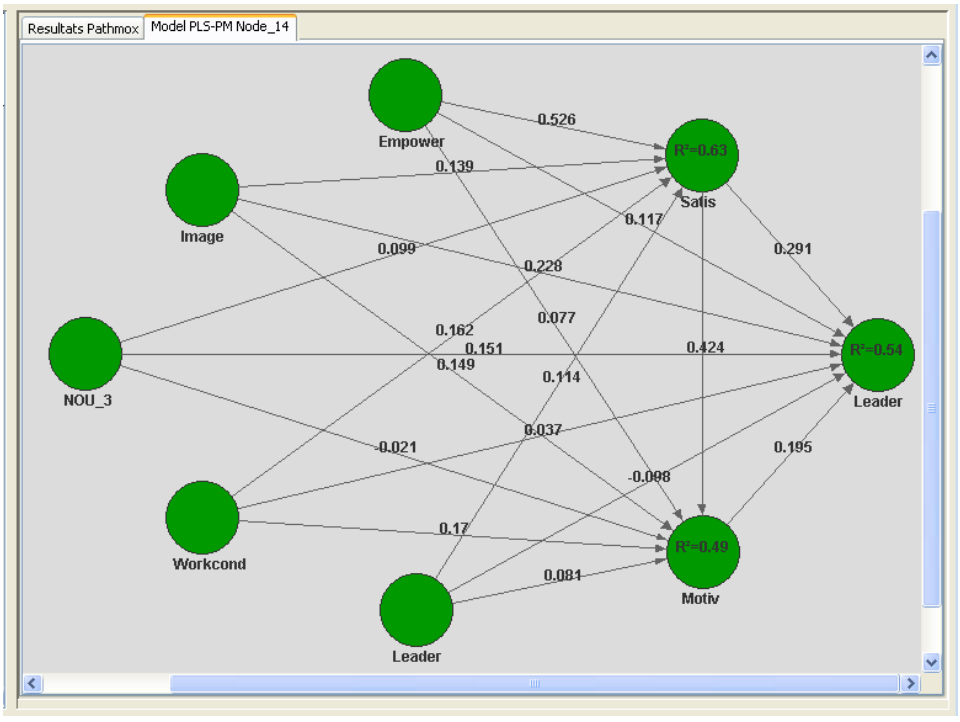


Figure B6. *Results of the structural model calculated in Visual Pathmox for node 14*

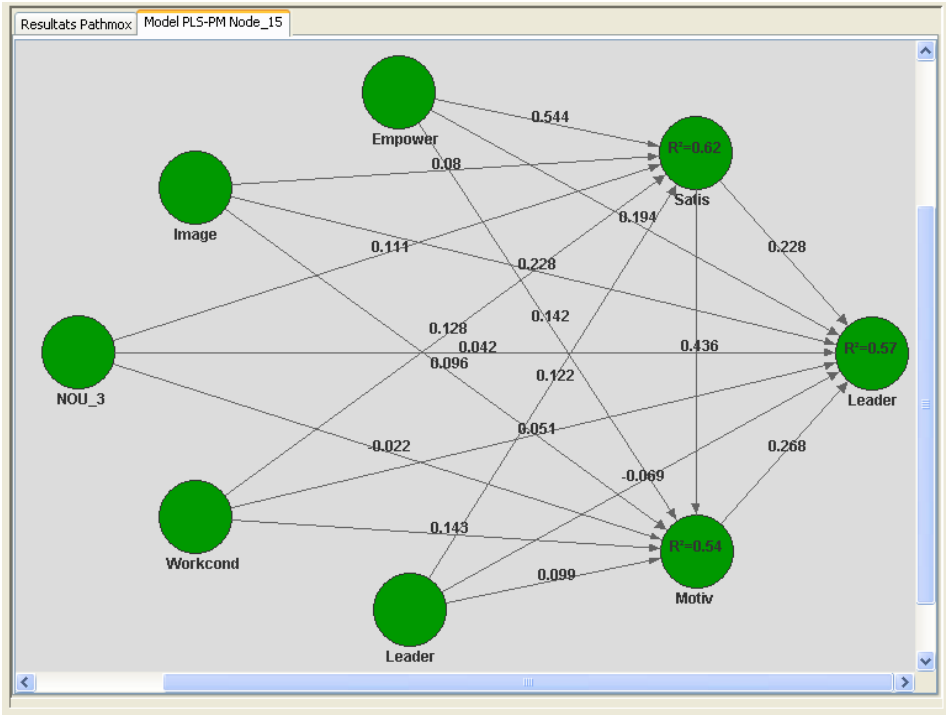


Figure B7. *Results of the structural model calculated in Visual Pathmox for node 15*

Appendix III: The R package "plspm"

The day Dr. Aluja proposed me to do a research project on PLS-PM, he also encouraged me to use R for programming the functions and routines to perform PLS-PM analysis. The proposal consisted of developing an R package that would allow us to estimate simple path models. As the doctoral project started to expand its scope to path modeling segmentation trees, we decided to use the programming code I had in R as a baseline for what will became Visual Pathmox. However, at the end of my research project I realized that there was something missing and that it could be worthwhile to create an R package for PLS-PM.

The plspm package is a new option to the list of existing programs that perform PLS-PM analysis such as *LVPLS*, *PLS-Graph*, *PLS-GUI*, *VisualPLS*, *SPAD-PLS*, *SmartPLS*, and *XLSTAT-PLSPM*. Although plspm lacks of a graphic interface to draw path diagrams, its main advantage is that one can complement its use with all the data analysis options and programming capabilities of R. The plspm package is freely available from the CRAN http://cran.r-project.org/.

The plspm package does not contain the PATHMOX algorithm which is still under a beta version for experimental purposes at LIAM. However, the it does contain various methods of the PLS data analysis framework such as NIPALS, PLS Regression (1 and 2), PLS Canonical Analysis, and PLS-PM.

The "plspm" function

The main function of the package is also called plspm. This function has eight arguments:

1. x: a matrix or data frame containing the manifest variables
2. inner.mat: the inner design matrix that indicates the relationships among latent variables
3. sets: a list of vectors that contain the column indices of x to form the different blocks of variables
4. modes: a character vector indicating the type of measurement (reflective or formative) for each latent variable
5. scheme: the inner weighting scheme to be used
6. scaled: a logical value indicating whether the data must be standardized
7. boot.val: a logical value indicating whether bootstrap validation must be performed
8. plsr: a logical value indicating whether pls regression must be used to calculate path coefficients.

The path model has to be specified with the use of the arguments inner.mat, sets, and modes; that is, using a matrix, a list, and a vector. To be exact, the structural part (inner model) is specified with the argument inner.mat while the measurement part (outer model) is specified by means of the arguments sets and modes. The different options of

the inner weighting schemes are set with the argument scheme. The value returned by the function plspm is an object of class "plspm", which can be printed and summarized by the print and summary methods.

The first three arguments must be provided by the user. The rest of the arguments have default values and the function can be run without the need to specify them.

- The argument x must be a numeric matrix or data frame and no missing values are allowed.
- The inner.mat argument is a special kind of matrix. This is the matrix that indicates the structural relationships among latent constructs. The function plspm requires the inner.mat argument to be defined as a squared matrix with only zeros or ones. In fact, since PLS-PM only works with recursive models, this matrix must be a lower triangular matrix which means that the entries in the diagonal and above the diagonal are zero.
- The argument sets is a list of length equal to the number of latent variables. The elements of sets are vectors which contain the column indices of x; that is, the indices of the manifest variables that form the different blocks.
- The argument modes is an optional argument. By default it is a character vector of length equal to the number of latent variables, and it contains as many letters "*A*" as latent variables in the model. This means that the latent variables are measured in a reflective way. If any LV is supposed to be measured in a formative way, the user has to specify the argument modes indicating which blocks are formative by a letter "*B*".
- The argument scheme is an optional parameter. By default it is set to be the character string *"factor"*, which means that inner weights are calculated according to the factor scheme. The other possible values are "*centroid*" and "*path*".
- The argument scaled refers to a logical value indicating whether data should be standardized.
- The last argument is the optional logical argument boot.val. It is FALSE by default, meaning that no bootstrap validation is performed.

The function plspm produces a list with the following results:

1. unidim: results for checking the unidimensionality of blocks. Includes: first and second eigenvalues, Cronbach's alpha, and Dillon-Goldstein's rho.
2. outer.mod: results of the outer (measurement) model. Includes: outer weights, standard loadings, communalities, and redundancies.
3. inner.mod: results of the inner (structural) model. Includes: path coefficients and R-squared for each endogenous latent variable.
4. latents: matrix of standardized latent variables (variance=1).
5. scores: matrix of re-scaled latent variables when scaled=FALSE. If scaled=TRUE then scores are equal to latents.
6. out.weights: vector of outer weights.
7. loadings: vector of standardized loadings (i.e. correlations with LVs).
8. path.coefs: matrix of path coefficients; this matrix has a similar form as inner.mat.
9. r.sqr: vector of R-squared coefficients.
10. outer.cor: correlations between the latent variables and the manifest variables (also called cross-loadings).

11. inner.sum: summarized results by latent variable. Includes: type of measurement, number of indicators, R-.squared, average communality, average redundancy, and average variance extracted.
12. gof: Table with indexes of Goodness-of-Fit. Includes: absolute GoF, relative GoF, outer model GoF, and inner model GoF.
13. effects: table of path effects from the structural relationships. Includes: direct, indirect, and total effects.
14. boot.mod: List of bootstrapping results; only available when argument boot.val=TRUE.

4 An example with a Customer Satisfaction study

To illustrate the usage and outputs of plspm, we use the satisfaction dataset and the typical model of the European Customer Satisfaction. The dataset refers to customers' perceptions about the service provided by a Spanish credit institution. The data set contains 27 variables observed on 250 individuals:

Variables of block Image:	columns 1 to 5
Variables of block Expectations:	columns 6 to 10
Variables of block Quality:	columns 11 to 15
Variables of block Value:	columns 16 to 19
Variables of block Satisfaction:	columns 20 to 23
Variables of block Loyalty:	columns 24 to 27

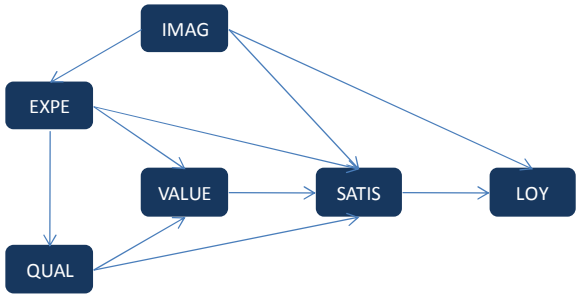


Fig 1 *Path diagram of the ECSI Model*

Since the structural part of the path model has to be specified in matrix form, the inner design matrix is defined as follows:

	IMAG	EXPE	QUAL	VAL	SAT	LOY
IMAG	0	0	0	0	0	0
EXPE	1	0	0	0	0	0
QUAL	0	1	0	0	0	0
VAL	0	1	1	0	0	0
SAT	1	1	1	1	0	0
LOY	1	0	0	0	1	0

Fig 2 *Inner design matrix for expressing the structural relationships*

Note that the inner design matrix is a lower triangular matrix. The zeros in the diagonal indicate that a latent variable cannot affect itself. The causal relationships among latent constructs are established in a top-down direction, that is: columns affecting rows. The

first column of the matrix implies that IMAG affects EXPE, SAT and LOY; the second column implies that EXPE affects QUAL, VAL, and SAT; the third column implies that QUAL affects VAL and SAT; and so on. The code in R to perform a PLS-PM analysis with the satisfaction data is shown below (figure 4).

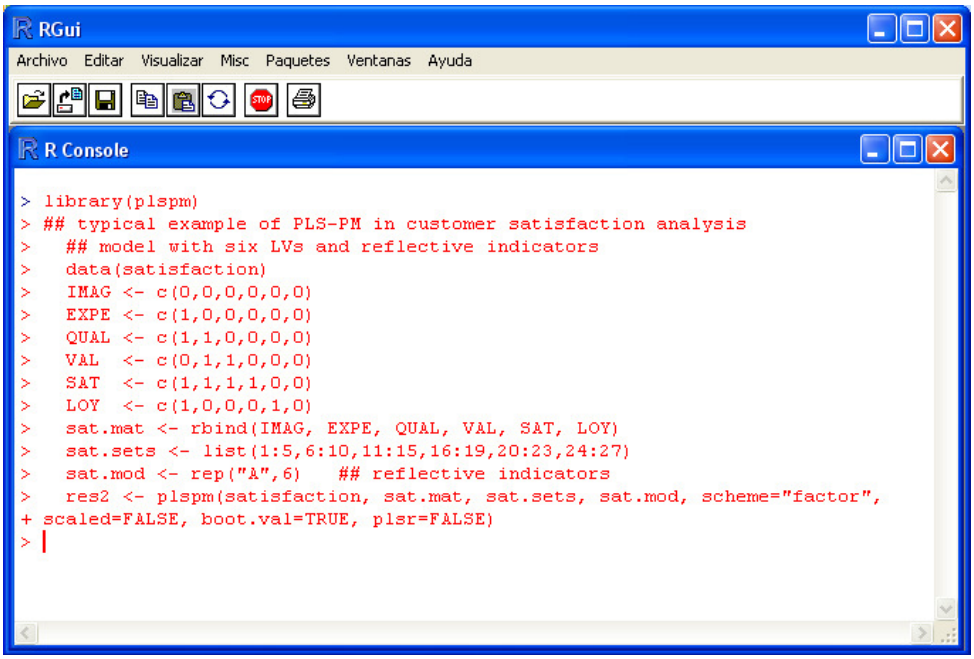


Fig 4 *Screen display in R with the instructions to perform PLS-PM analysis with the satisfaction data*

The selected scheme to calculate the inner weights is the factor scheme. Since the manifest variables are measured in the same scale, the argument scaled is set as FALSE (i.e. no standardization required). In addition, the argument boot.val is set as TRUE (i.e. bootstrap validation is performed). The measurement relationships for the latent variables and their manifest variables are indicated with a character vector containing *A*'s and/or *B*'s. An "*A*" is used to indicate a reflective block of manifest variables while a "*B*" is used for a formative block of indicators. In this example all the latent variables are considered in reflective form. Thus, the argument modes is a character vector with six elements given as:

> modes <- rep("A", 6)

The print method gives the following display:

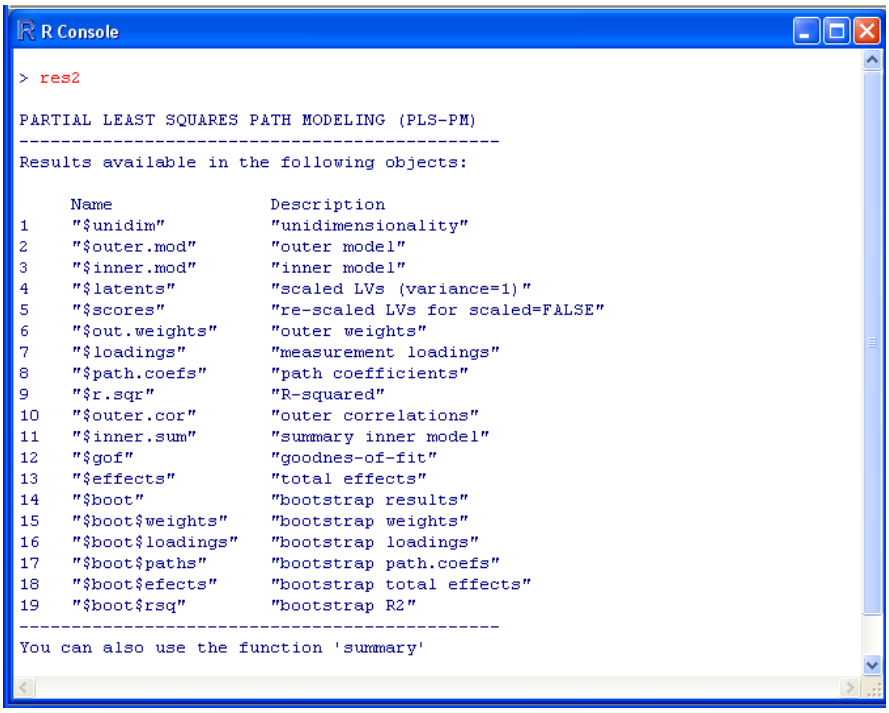


Fig 5 *Printing display of the of an object of class "plspm" showing the list of results*

In order to see a general output of the model, the summary method provides the following results. The model specification contains the number of cases (250) of the analyzed data, the number of latent variables in the model (6), the number of used manifest indicators (27), the scale of the data (Raw Data), the inner weighting scheme (factor), and the bootstrap validation option (TURE).
The definition of the blocks shows the latent variables, their type (exogenous or endogenous), the number of indicators in each block and the type of measurement relationships.

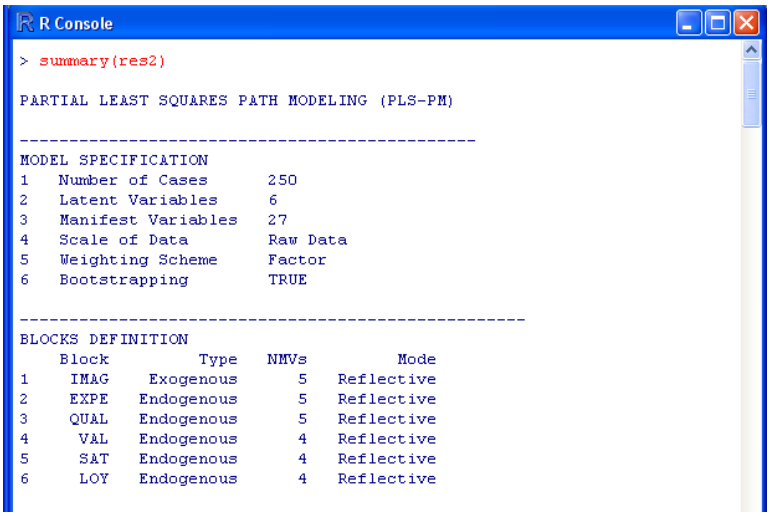


Fig 6 Model specification and definition of the blocks of variables

The next results in the summary of the object "plspm" are the indexes used to check the unidimensionality of the blocks. Since all measurement relationships of the model are reflective, it is meaningful to analyze whether the blocks can be considered to be unidimensional. In order to assess the extent to which a block is unidimensional, plspm provides three indexes: the first and second eigenvalues of the MVs correlation matrix, the Cronbach's alpha, and the Dillon-Goldstein's ρ (see figure 7).

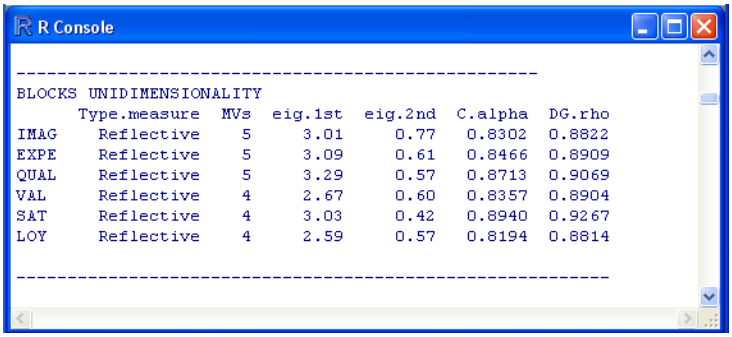


Fig 7 Indexes used to assess blocks' unidimensionality

In fourth place, the results of the outer model are provided: outer weights, standard loadings, communality, and redundancy.

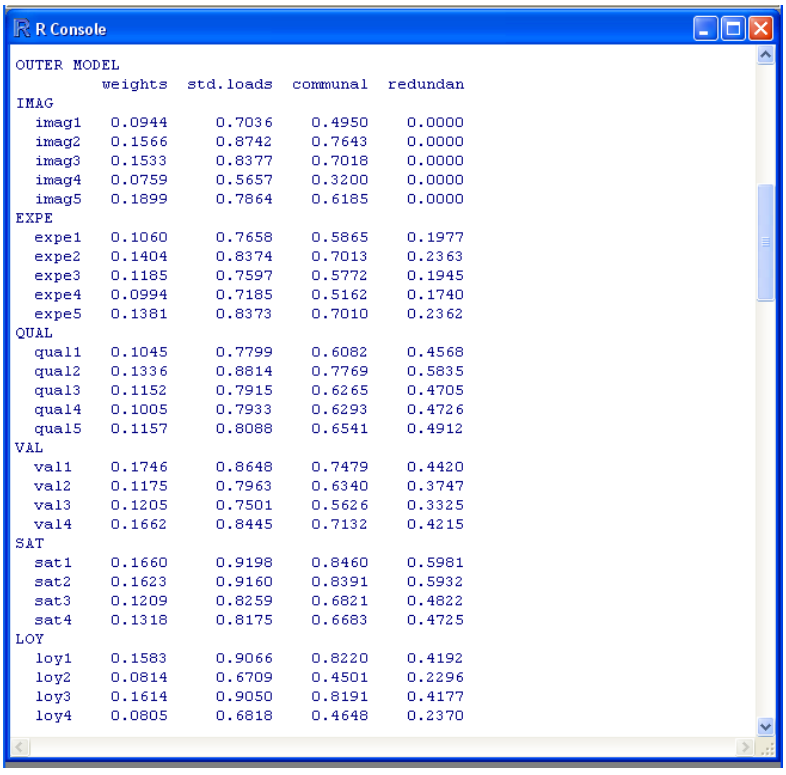


Fig 8 Results of the outer model: outer weights, standard loadings, communality, and redundancy

Then, the correlations between latent variables and manifest variables are listed.

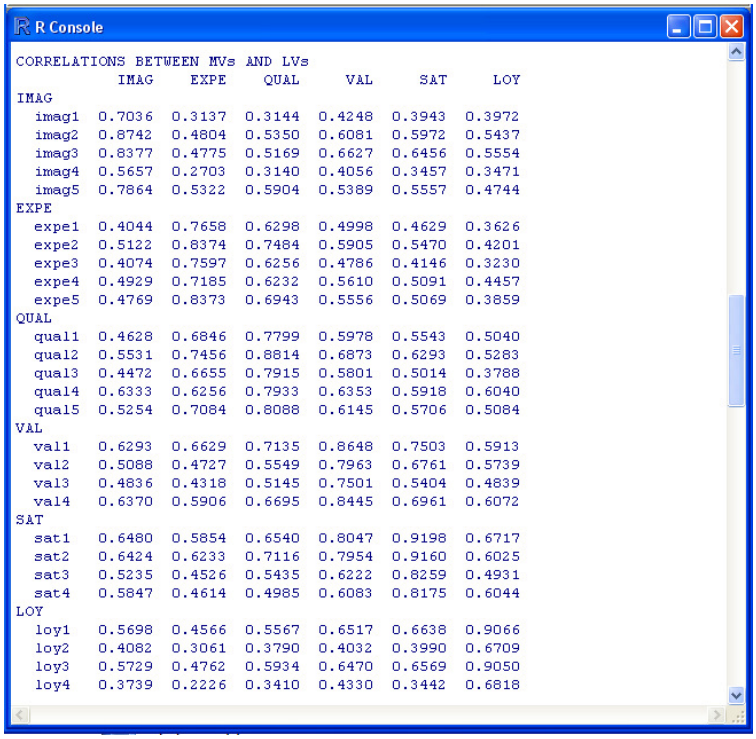


Fig 9 Correlations between latent variables and manifest variables (i.e. cross-loadings)

In sixth place, the output of the inner model is displayed in a list with the results for each endogenous latent variable: R^2 coefficient, intercept term, and path coefficients.

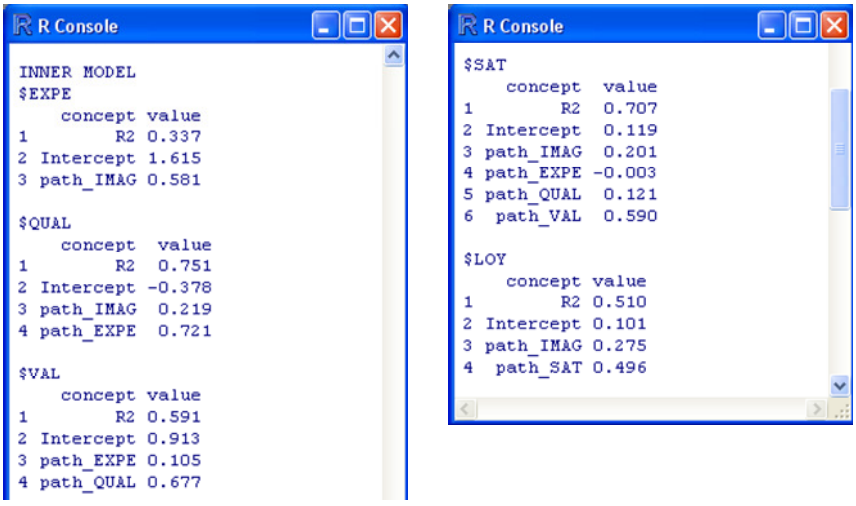
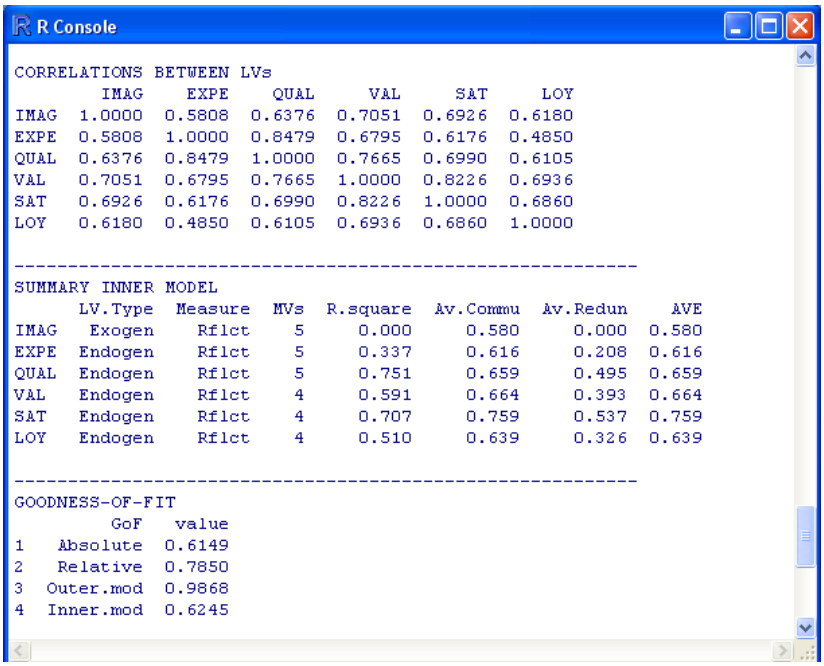


Fig 10 Inner model results: R^2, intercept term, and path coefficients

The next results are the correlations between latent variables. Then, a summary table for the inner model is presented with the average communality, the average redundancy, and the average variance extracted index. In ninth place, the Goodness-of-Fit (GoF) are displayed (absolute gof index, relative gof index, outer model gof, and inner model gof).



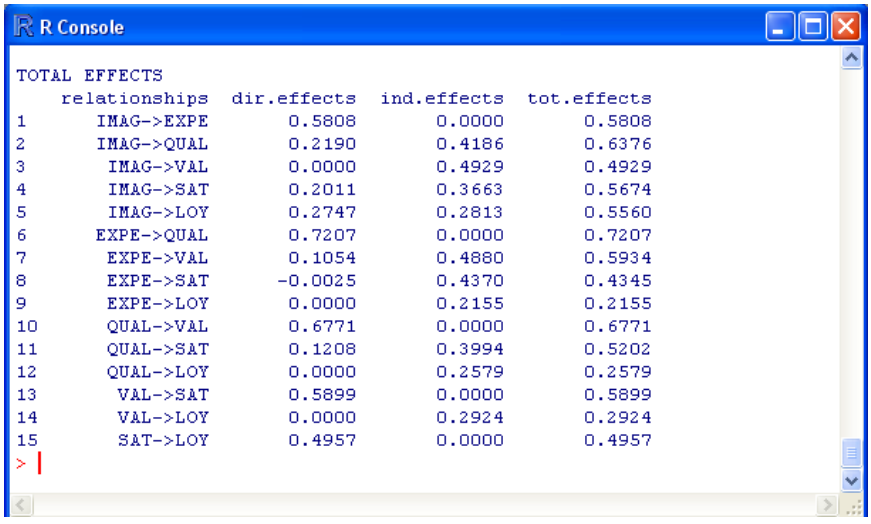
```
CORRELATIONS BETWEEN LVs
        IMAG    EXPE    QUAL     VAL     SAT     LOY
IMAG  1.0000  0.5808  0.6376  0.7051  0.6926  0.6180
EXPE  0.5808  1.0000  0.8479  0.6795  0.6176  0.4850
QUAL  0.6376  0.8479  1.0000  0.7665  0.6990  0.6105
VAL   0.7051  0.6795  0.7665  1.0000  0.8226  0.6936
SAT   0.6926  0.6176  0.6990  0.8226  1.0000  0.6860
LOY   0.6180  0.4850  0.6105  0.6936  0.6860  1.0000

------------------------------------------------------------
SUMMARY INNER MODEL
      LV.Type  Measure  MVs  R.square  Av.Commu  Av.Redun    AVE
IMAG   Exogen     Rflct    5     0.000     0.580     0.000  0.580
EXPE  Endogen     Rflct    5     0.337     0.616     0.208  0.616
QUAL  Endogen     Rflct    5     0.751     0.659     0.495  0.659
VAL   Endogen     Rflct    4     0.591     0.664     0.393  0.664
SAT   Endogen     Rflct    4     0.707     0.759     0.537  0.759
LOY   Endogen     Rflct    4     0.510     0.639     0.326  0.639

------------------------------------------------------------
GOODNESS-OF-FIT
         GoF   value
1   Absolute  0.6149
2   Relative  0.7850
3  Outer.mod  0.9868
4  Inner.mod  0.6245
```

Fig 11 Inner model correlations, summary table, and GoF index

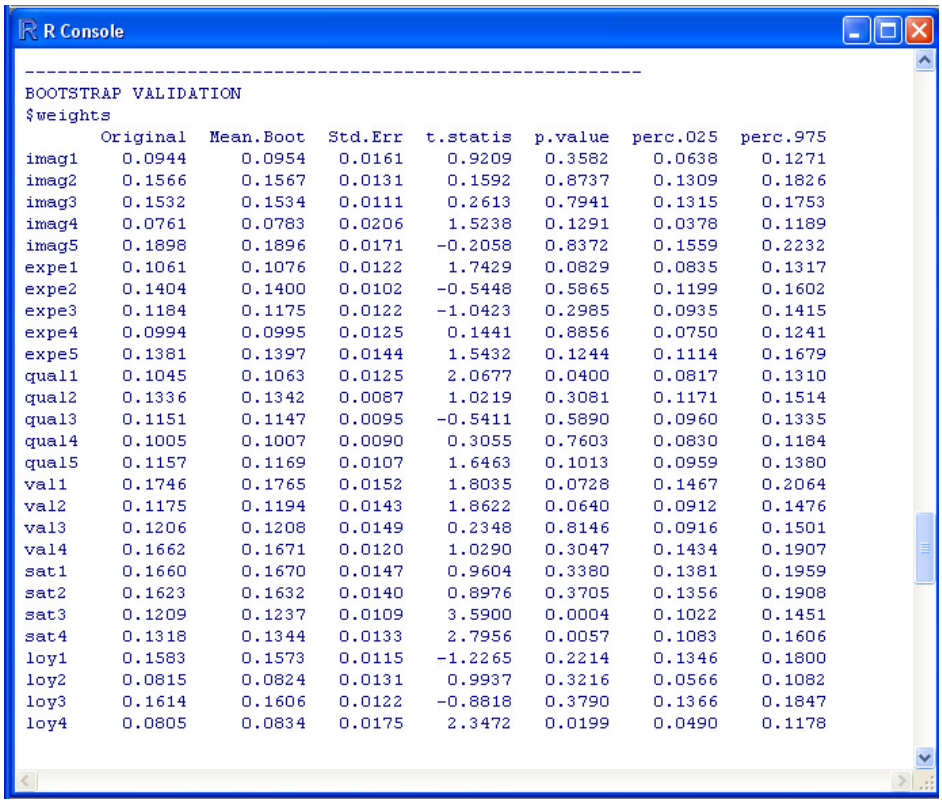
Figure 12 presents the table with the path relations effects.



```
TOTAL EFFECTS
    relationships  dir.effects  ind.effects  tot.effects
1      IMAG->EXPE       0.5808       0.0000       0.5808
2      IMAG->QUAL       0.2190       0.4186       0.6376
3       IMAG->VAL       0.0000       0.4929       0.4929
4       IMAG->SAT       0.2011       0.3663       0.5674
5       IMAG->LOY       0.2747       0.2813       0.5560
6      EXPE->QUAL       0.7207       0.0000       0.7207
7       EXPE->VAL       0.1054       0.4880       0.5934
8       EXPE->SAT      -0.0025       0.4370       0.4345
9       EXPE->LOY       0.0000       0.2155       0.2155
10      QUAL->VAL       0.6771       0.0000       0.6771
11      QUAL->SAT       0.1208       0.3994       0.5202
12      QUAL->LOY       0.0000       0.2579       0.2579
13       VAL->SAT       0.5899       0.0000       0.5899
14       VAL->LOY       0.0000       0.2924       0.2924
15       SAT->LOY       0.4957       0.0000       0.4957
>
```

Fig 12 Table of path effects: direct, indirect and total effects

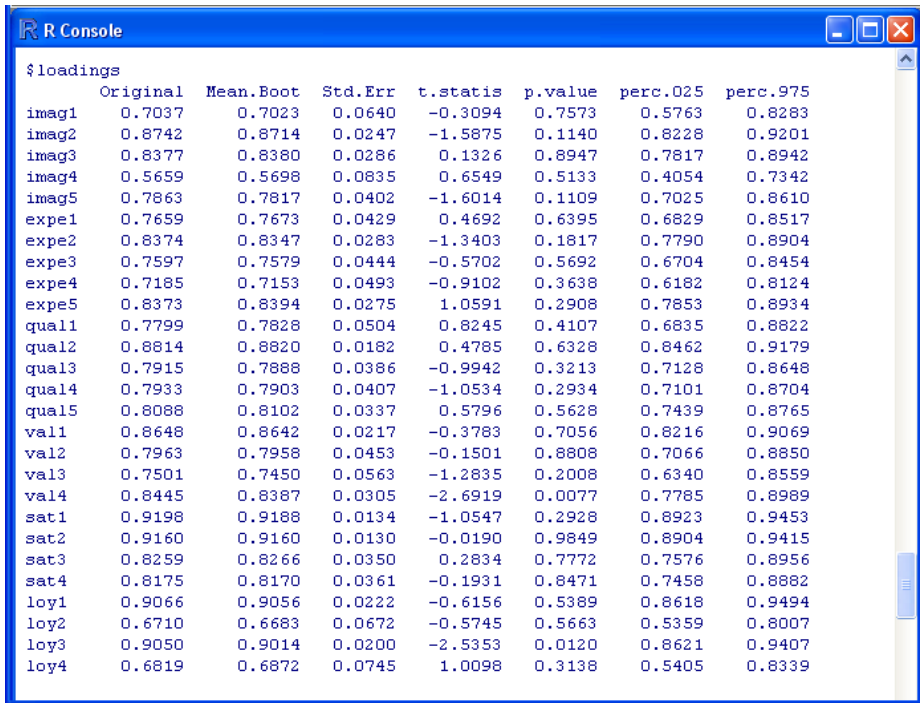
Finally, the results of the bootstrap validation are shown in figures 13 to 16.



R Console

```
------------------------------------------------------------
BOOTSTRAP VALIDATION
$weights
       Original  Mean.Boot  Std.Err  t.statis  p.value  perc.025  perc.975
imag1    0.0944     0.0954   0.0161    0.9209   0.3582    0.0638    0.1271
imag2    0.1566     0.1567   0.0131    0.1592   0.8737    0.1309    0.1826
imag3    0.1532     0.1534   0.0111    0.2613   0.7941    0.1315    0.1753
imag4    0.0761     0.0783   0.0206    1.5238   0.1291    0.0378    0.1189
imag5    0.1898     0.1896   0.0171   -0.2058   0.8372    0.1559    0.2232
expe1    0.1061     0.1076   0.0122    1.7429   0.0829    0.0835    0.1317
expe2    0.1404     0.1400   0.0102   -0.5448   0.5865    0.1199    0.1602
expe3    0.1184     0.1175   0.0122   -1.0423   0.2985    0.0935    0.1415
expe4    0.0994     0.0995   0.0125    0.1441   0.8856    0.0750    0.1241
expe5    0.1381     0.1397   0.0144    1.5432   0.1244    0.1114    0.1679
qual1    0.1045     0.1063   0.0125    2.0677   0.0400    0.0817    0.1310
qual2    0.1336     0.1342   0.0087    1.0219   0.3081    0.1171    0.1514
qual3    0.1151     0.1147   0.0095   -0.5411   0.5890    0.0960    0.1335
qual4    0.1005     0.1007   0.0090    0.3055   0.7603    0.0830    0.1184
qual5    0.1157     0.1169   0.0107    1.6463   0.1013    0.0959    0.1380
val1     0.1746     0.1765   0.0152    1.8035   0.0728    0.1467    0.2064
val2     0.1175     0.1194   0.0143    1.8622   0.0640    0.0912    0.1476
val3     0.1206     0.1208   0.0149    0.2348   0.8146    0.0916    0.1501
val4     0.1662     0.1671   0.0120    1.0290   0.3047    0.1434    0.1907
sat1     0.1660     0.1670   0.0147    0.9604   0.3380    0.1381    0.1959
sat2     0.1623     0.1632   0.0140    0.8976   0.3705    0.1356    0.1908
sat3     0.1209     0.1237   0.0109    3.5900   0.0004    0.1022    0.1451
sat4     0.1318     0.1344   0.0133    2.7956   0.0057    0.1083    0.1606
loy1     0.1583     0.1573   0.0115   -1.2265   0.2214    0.1346    0.1800
loy2     0.0815     0.0824   0.0131    0.9937   0.3216    0.0566    0.1082
loy3     0.1614     0.1606   0.0122   -0.8818   0.3790    0.1366    0.1847
loy4     0.0805     0.0834   0.0175    2.3472   0.0199    0.0490    0.1178
```

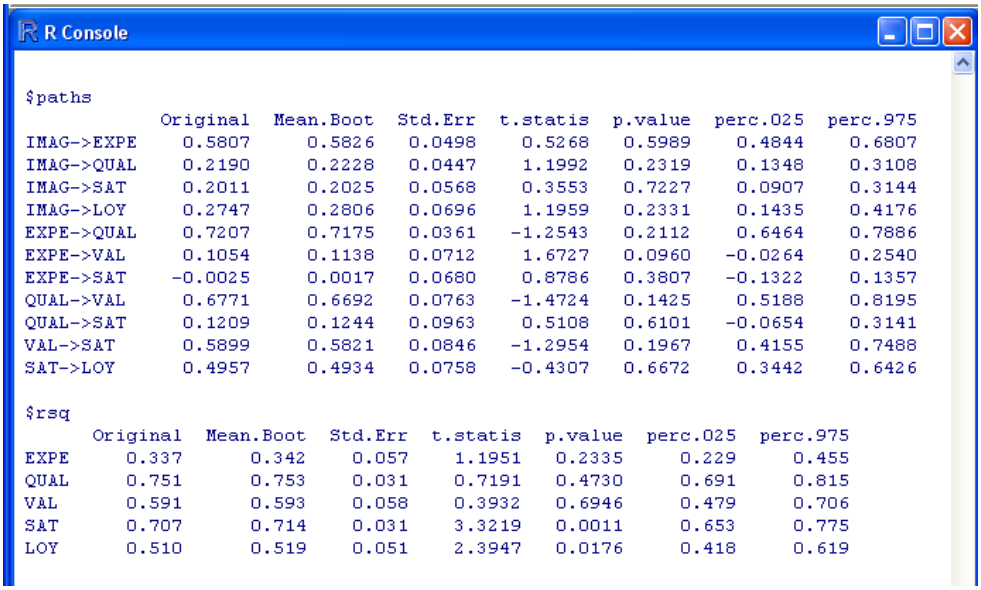
Fig 13 Table of bootstrap results for the outer weights



R Console

```
$loadings
       Original  Mean.Boot  Std.Err  t.statis  p.value  perc.025  perc.975
imag1    0.7037     0.7023   0.0640   -0.3094   0.7573    0.5763    0.8283
imag2    0.8742     0.8714   0.0247   -1.5875   0.1140    0.8228    0.9201
imag3    0.8377     0.8380   0.0286    0.1326   0.8947    0.7817    0.8942
imag4    0.5659     0.5698   0.0835    0.6549   0.5133    0.4054    0.7342
imag5    0.7863     0.7817   0.0402   -1.6014   0.1109    0.7025    0.8610
expe1    0.7659     0.7673   0.0429    0.4692   0.6395    0.6829    0.8517
expe2    0.8374     0.8347   0.0283   -1.3403   0.1817    0.7790    0.8904
expe3    0.7597     0.7579   0.0444   -0.5702   0.5692    0.6704    0.8454
expe4    0.7185     0.7153   0.0493   -0.9102   0.3638    0.6182    0.8124
expe5    0.8373     0.8394   0.0275    1.0591   0.2908    0.7853    0.8934
qual1    0.7799     0.7828   0.0504    0.8245   0.4107    0.6835    0.8822
qual2    0.8814     0.8820   0.0182    0.4785   0.6328    0.8462    0.9179
qual3    0.7915     0.7888   0.0386   -0.9942   0.3213    0.7128    0.8648
qual4    0.7933     0.7903   0.0407   -1.0534   0.2934    0.7101    0.8704
qual5    0.8088     0.8102   0.0337    0.5796   0.5628    0.7439    0.8765
val1     0.8648     0.8642   0.0217   -0.3783   0.7056    0.8216    0.9069
val2     0.7963     0.7958   0.0453   -0.1501   0.8808    0.7066    0.8850
val3     0.7501     0.7450   0.0563   -1.2835   0.2008    0.6340    0.8559
val4     0.8445     0.8387   0.0305   -2.6919   0.0077    0.7785    0.8989
sat1     0.9198     0.9188   0.0134   -1.0547   0.2928    0.8923    0.9453
sat2     0.9160     0.9160   0.0130   -0.0190   0.9849    0.8904    0.9415
sat3     0.8259     0.8266   0.0350    0.2834   0.7772    0.7576    0.8956
sat4     0.8175     0.8170   0.0361   -0.1931   0.8471    0.7458    0.8882
loy1     0.9066     0.9056   0.0222   -0.6156   0.5389    0.8618    0.9494
loy2     0.6710     0.6683   0.0672   -0.5745   0.5663    0.5359    0.8007
loy3     0.9050     0.9014   0.0200   -2.5353   0.0120    0.8621    0.9407
loy4     0.6819     0.6872   0.0745    1.0098   0.3138    0.5405    0.8339
```

Fig 14 Table of bootstrap results for the loadings

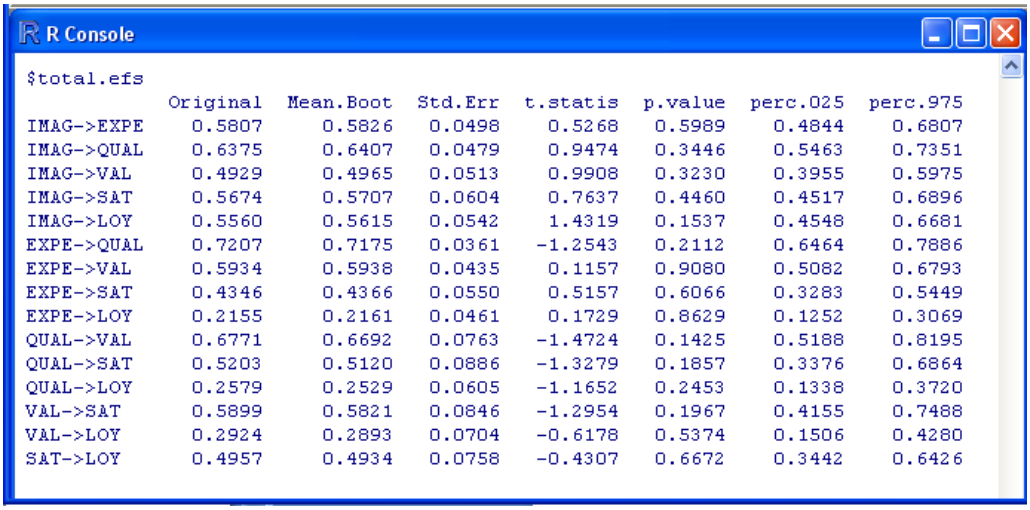


R Console

```
$paths
            Original  Mean.Boot  Std.Err  t.statis  p.value  perc.025  perc.975
IMAG->EXPE    0.5807     0.5826   0.0498    0.5268   0.5989    0.4844    0.6807
IMAG->QUAL    0.2190     0.2228   0.0447    1.1992   0.2319    0.1348    0.3108
IMAG->SAT     0.2011     0.2025   0.0568    0.3553   0.7227    0.0907    0.3144
IMAG->LOY     0.2747     0.2806   0.0696    1.1959   0.2331    0.1435    0.4176
EXPE->QUAL    0.7207     0.7175   0.0361   -1.2543   0.2112    0.6464    0.7886
EXPE->VAL     0.1054     0.1138   0.0712    1.6727   0.0960   -0.0264    0.2540
EXPE->SAT    -0.0025     0.0017   0.0680    0.8786   0.3807   -0.1322    0.1357
QUAL->VAL     0.6771     0.6692   0.0763   -1.4724   0.1425    0.5188    0.8195
QUAL->SAT     0.1209     0.1244   0.0963    0.5108   0.6101   -0.0654    0.3141
VAL->SAT      0.5899     0.5821   0.0846   -1.2954   0.1967    0.4155    0.7488
SAT->LOY      0.4957     0.4934   0.0758   -0.4307   0.6672    0.3442    0.6426

$rsq
     Original  Mean.Boot  Std.Err  t.statis  p.value  perc.025  perc.975
EXPE    0.337      0.342    0.057    1.1951   0.2335     0.229     0.455
QUAL    0.751      0.753    0.031    0.7191   0.4730     0.691     0.815
VAL     0.591      0.593    0.058    0.3932   0.6946     0.479     0.706
SAT     0.707      0.714    0.031    3.3219   0.0011     0.653     0.775
LOY     0.510      0.519    0.051    2.3947   0.0176     0.418     0.619
```

Figure 15 Table of bootstrap results for the path coefficients



R Console

```
$total.efs
            Original  Mean.Boot  Std.Err  t.statis  p.value  perc.025  perc.975
IMAG->EXPE    0.5807     0.5826   0.0498    0.5268   0.5989    0.4844    0.6807
IMAG->QUAL    0.6375     0.6407   0.0479    0.9474   0.3446    0.5463    0.7351
IMAG->VAL     0.4929     0.4965   0.0513    0.9908   0.3230    0.3955    0.5975
IMAG->SAT     0.5674     0.5707   0.0604    0.7637   0.4460    0.4517    0.6896
IMAG->LOY     0.5560     0.5615   0.0542    1.4319   0.1537    0.4548    0.6681
EXPE->QUAL    0.7207     0.7175   0.0361   -1.2543   0.2112    0.6464    0.7886
EXPE->VAL     0.5934     0.5938   0.0435    0.1157   0.9080    0.5082    0.6793
EXPE->SAT     0.4346     0.4366   0.0550    0.5157   0.6066    0.3283    0.5449
EXPE->LOY     0.2155     0.2161   0.0461    0.1729   0.8629    0.1252    0.3069
QUAL->VAL     0.6771     0.6692   0.0763   -1.4724   0.1425    0.5188    0.8195
QUAL->SAT     0.5203     0.5120   0.0886   -1.3279   0.1857    0.3376    0.6864
QUAL->LOY     0.2579     0.2529   0.0605   -1.1652   0.2453    0.1338    0.3720
VAL->SAT      0.5899     0.5821   0.0846   -1.2954   0.1967    0.4155    0.7488
VAL->LOY      0.2924     0.2893   0.0704   -0.6178   0.5374    0.1506    0.4280
SAT->LOY      0.4957     0.4934   0.0758   -0.4307   0.6672    0.3442    0.6426
```

Figure 16 Table of bootstrap results for the total effects

The "plspm" package offers the main results of a PLS-PM analysis as most of the available PLS-PM software programs. To conclude, plspm provides a flexible easy-to-use function which allow practitioners and researchers from different disciplines to compute, interpret, and analyze path models from a pls approach.

References

1. Ahuja, M.K., and Thatcher, J.B. (2005) Moving Beyond Intentions and Toward the Theory of Trying: Effects of Work Environment and Gender on Post-Adoption Information Technology Use. *Management Information Systems Quarterly*, 29(3): 427-459.

2. Aluja, T., and Morineau, A. (1999) *Aprender de los Datos: El Análisis de Componentes Principales - Una Aproximación desde el Data Mining*. Barcelona: EUB.

3. Amato, S., Esposito Vinzi, V., and Tenenhaus, M. (2004) A global Goodness-of-Fit for PLS structural equation modelling. Oral communication to PLS Club, HEC School of Management, France, March 24.

4. Anderson, N., Ones, D.S., Handan, K.S., and Chockalingam, V. (2002) *Handbook of Industrial, Work and Organizational Psychology Volume 2*. New York: Sage Publications.

5. Anderson, E., Fornell, C., and Lehmann, D.R. (1994) Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. *Journal of Marketing*, 58(3): 53-66.

6. Anderson, E., and Fornell, C. (2000) Foundations of the American Customer Satisfaction Index. *Total Quality Management*, 11(7): 869-882.

7. Andreasen, T.W., and Lindestad, B. (1998) The effects of corporate image in the formation of customer loyalty. *Journal of Service Marketing*, 1: 82-92.

8. Apté, C., and Weiss, S. (1997) Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 13(2-3): 197-210.

9. Arminger, G., and Stein, P. (1997) Finite mixture of covariance structure models with regressors: loglikehood function, distance estimation, fit indices, and a complex example. *Sociological Methods and Research*, 26: 146-182.

10. Ashill, N.J., Carruthers, J., and Krisjanous, J. (2005) Antecedents and outcomes of service recovery performance in a public heath-care environment. *Journal of Services Marketing*, 19(5): 293-308.

11. Barclay, D., Higgins, C., and Thompson, R. (1995) The Partial Least Squares (PLS) Approach to Causal Modelling: Personal Computer Adoption and Use as an Illustration. *Technology Studies*, 2(2): 285-309.

12. Baron, R.M., and Kenny, D.A. (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51: 1173-1182.

13. Bart, C.K., Bontis, N., and Taggar, S. (2001) A model of the impact of mission statements on firm performance. *Management Decision*, 38(1): 19-35.

14. Basilevsky, A. (1994) *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: John Wiley & Sons.

15. Bastien, P., Tenenhaus, M., and Esposito Vinzi, V. (2005) PLS generalized linear regression. *Computational Statistics & Data Analysis*, 48(1): 17-46.

16. Bayol, M.P., De La Foye, A., Tellier, C., and Tenenhaus, M. (2000) Use of PLS Path modeling to estimate the European Consumer Satisfaction Index (ECSI) model. *Statistica Applicata: Italian Journal of Applied Statistics*, 12(3): 361-375.

17. Bentler, P. M. (1980) Multivariate Analysis with Latent Variables: Causal Modeling. *Annual Revue of Psychology*, 31: 419-456.

18. Bentler, P.M. (1986) Structural Modeling and Psychometrika: An Historical Perspective on Growth and Achievements. *Psychometrika,* 51(1): 35-51.

19. Benzécri, J.P. (1973) *L'Analyse des Données. Volume II: L'Analyse des Correspondence.* Paris: Dunod.

20. Bernert, C. (1983) The Career of Causal Analysis in American Sociology. *The British Journal of Sociology,* 34(2): 230-254.

21. Bollen, K.A. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

22. Bollen, K.A. (2002) Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53: 605-634.

23. Bollen, K.A., and Long, J.S. (1993) *Testing structural equation models*. California: Sage Publications.

24. Bontis, N. (1998) Intellectual capital: an exploratory study that develops measures and models. *Management Decision*, 36(2): 63-76.

25. Bontis, N. (2004) National Intellectual Capital Index: A United Nations initiative for the Arab Region. *Journal of Intellectual Capital*, 5(1): 13-39.

26. Bontis, N., and Serenko, A. (2007) The moderating role of human capital management practices on employee capabilities. *Journal of Knowledge Management*, 11(3): 31-51.

27. Bookstein, F.L. (1982) The Geometric Meaning of Soft Modeling with Some Generalizations. In: *Systems under indirect observation: Causality, structure, prediction. Part II*, 55-74. K.G. Jöreskog & H. Wold (Eds). Amsterdam: North Holland.

28. Bookstein, F.L. (1986) The Elements of Latent Variable Models: A Cautionary Lection. In: *Advances in Developmental Psychology. Volume: 4*, 204-230. M.E. Lamb, A.L. Brown, B. Rogoff (Eds). New Jersey: Lawrence Erlbaum Associates.

29. Bookstein, F.L. (1990) Least Squares and Latent Variables. *Multivariate Behavioral Research*, 25(1): 75-80.

30. Bookstein, F.L., Chernoff, B., Elder, R., Humpries, J., Smith, G., and Strauss, R. (1985) *Morphometrics in Evolutionary Biology. The Geometry of Size and Shape Change, with Examples from Fishes*. Philadelphia: Academy of Natural Sciences of Philadelphia.

31. Borsboom, D., Mellenbergh, G.J., and Van Heerden, J. (2003) The Theoretical Status of Latent Variables. *Psychological Review*, 110(2): 203-219.

32. Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984) *Classification and Regression Trees*. California: Chapman & Hall.

33. Brito, P. (2000) Hierarchical and Pyramidal Clustering with Complete Symbolic Objects. In: *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, 312-341. Heidelberg: Springer.

34. Brown, S.P., and Chin, W.W. (2004) Satisfying and Retaining Customers through Independent Service Representatives. *Decision Sciences*, 35(3): 527-550.

35. Bouchard-Côté, A. (2004) Solving Polynomials: The Roots of Modern Algebra. McGill University, personal paper.

36. Bounfour, A. (2003) *The Management of Intangibles: The organization's most valuable assets*. London: Routledge.

37. Byrne, B.M. (2001) *Structural Equation Modeling with AMOS*. New Jersey: Lawrence Earlbaum Associates.

38. Cabrita, M.R., and Vaz, J.L. (2006) Intellectual Capital and Value Creation: Evidence form the Portuguese Banking Industry. *Electronic Journal of Knowledge Management*, 4(1): 11-20.

39. Calvo-Mora, A., Leal, A., and Roldán, J.L. (2006) Using enablers of the EFQM model to manage institutions of higher education. *Quality Assurance in Education*, 14(2): 99-122.

40. Cassel, C., (2006) Measuring Customer Satisfaction, a methodological guidance. *Statistiska Centralbyran,* Eurostat report. Available from: http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47143266/CUSTOMER%20SATISFACTION%20SURVEYS_SE_2006_EN_1.PDF. Accessed 3 September 2008.

41. Cassel, C., Hackl, P., and Westlund, A.H. (1999) Robustness of partial least squares method for estimating latent variable quality structures. *Journal of Applied Statistics,* 26(4): 435-446.

42. Cassel, C.M., Hackl, P., and Wsetlund, A.H. (2000) On measurement of intangible assets: a study of robustness of partial least squares. *Total Quality Management,* 11(7): 897-907.

43. Chin, W.W. (1998) The partial least squares approach to structural equation modeling. In: *Modern Methods for Business Research*, 295-336. Marcoulides, G.A. (Ed). London: Lawrence Erlbaum Associates.

44. Chin, W.W. (2000) Frequently Asked Questions – Partial Least squares & PLS-Graph. Available from: http://disc-nt.cba.uh.edu/chin/plsfaq/plsfaq.htm (accessed 17-10-07).

45. Chin, W.W. (2003) A Permutation Based Procedure for Multi-Group Comparison of PLS Models. In: *Proceedings of the PLS'03 International Symposium*, 33-43. M. Vilares., M. Tenenhaus, P. Coelho, V. Esposito Vinzi, A. Morineau (Eds), Decisia.

46. Chin, W.W., Marcolin, B.L., and Newsted, P.R. (1996) A Partial Least Squares Latent Variable Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study. In: *Proceedings of the Seventeenth International Conference on Information Systems*, 21-41. J.I. DeGross, S. Jarvenpaa, A. Srinivasan (Eds).

47. Chin, W.W., Marcolin, B.L., and Newsted, P.R. (2003) A Partial Least Squares Latent Variable Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study. *Information Systems Research*, 14(2): 189-217.

48. Chin, W.W., and Newsted, P.R. (1999) Structural Equation Modeling Analysis with Small Samples using Partial Least Squares. In: *Statistical Strategies for Small Sample Research*, 307-341. Hoyle R. (Ed). London: Sage Publications.

49. Choe, Y.C., Hwang, D.R., Kim, M., and Moon, J. (2007) Product heterogeneity: Moderating effect on online consumer behaviour. Proceedings of the 40th Hawaii International Conference on System Sciences – 2007.

50. Chow, G.C. (1960) Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3): 591-605.

51. Christ, C.F. (1994) The Cowles Commission's Contributions to Econometrics at Chicago, 1939-1955. *Journal of Economic Literature* 32(1): 30-59.

52. Clarke, M.R.B. (1970) A rapidly convergent method for maximum-likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23: 43-52.

53. Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

54. Cooper, J.C.B. (1983) Factor Analysis: An Overview. *The American Statistician*, 37(2): 141-147.

55. Cortina, J.M. (1993) Interaction, Nonlinearity, and Multicollinearity: Implications for Multiple-regression. *Journal of Management*, 19: 915-922.

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

57. Cuadras, C.M., (1981) *Métodos de Análisis Multivariante*. Barcelona: Editorial Universitaria de Barcelona.

58. Currivan, D.B. (1999) The Causal Order of Job Satisfaction and Organizational Commitment in Models of Employee Turnover. *Human Resource Management Review*, 9(4): 495-524.

59. Dahlquist, A., Björck, A., and Anderson, N. (1974) *Numerical Methods*. New Jersey: Prentice-Hall.

60. Daum, J.H. (2003) *Intangible Assets and Value Creation*. West Sussex: John Wiley & Sons.

61. De Beuckelaer, A. (2005) On the nature of constructs and their use in comparative research. In: *Proceedings of the PLS'05 International Symposium*, 117-124. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds), SPAD Test&go.

62. De Leeuw, E., and Nichols, W. (1996) Technological Innovations in Data Collection: Acceptance, Data Quality and Costs. *Sociological Research Online*, 1(4). Available at: http://www.socresonline.org.uk/socresonline/1/4/leeuw.html. Accessed 3 September 2008.

63. de Leeuw, J., Young, F.W., and Takane, Y. (1976) Additive Structure in Qualitative Data: An Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, 41: 471-503.

64. Demotes-Mainard, M. (2003) Statistical Information on Intangibles. *Information Society Statistics, Voorburg Group on Service Statistics (October, 2003)* Available from: http://www.stat.go.jp/english/info/meetings/voorburg/pdf/mag_stat.pdf . Accessed 3 September 2008.

65. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM-algorithm. *Journal of the Royal Statistical Society: Series B*, 39: 1-38.

66. Denis, D.J., and Legerski, J. (2006) Causal Modeling and the Origin of Path Analysis. *Theory and Science*, 7(1). Available from: http://theoryandscience.icaap.org/content/vol7.1/denis.html. Accessed 3 September 2008.

67. Desrosières, A. (2004) *La política de los grandes números. Historia de la razón estadística*. Barcelona: Melusina.

68. Diamantopoulos, A. (1994) Modelling with LISREL: A Guide for the Uninitiated. *Journal of Marketing Management,* 10: 105-136.

69. Diamantopoulos, A., and Winklhofer, H. (2001) Index Construction with Formative Indicators: An Alternative to scale development. *Journal of Marketing Research*, 38(2): 269-278.

70. Dijkstra, T. (1983) Some Comments on Maximum Likelihood and Partial Least Squares Methods. *Journal of Economics*, 22: 67-90.

71. Dilon, W.R. (1990) Unmixing Models for Analyzing Marketing Research Data Having Heterogeneous Components. In: *First Annual Advanced Research Techniques Forum (Conference Proceedings)*, 278-316. William D. Neal (Ed.) American Marketing Association, Chicago, Illinois.

72. Dilon, W.R., and Kumar, A. (1994) Latent Structure and Other Mixture Models in Marketing: An Integrative Survey and Overview. In: *Advanced Methods of Marketing Research*, 295-351. Richard P. Bagozzi (Ed.), Cambridge, Mass: Blackwell Business.

73. Dolan, C.V., and Van der Maas, H.L.J. (1997) Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, 63: 227-253.

74. Drenth, P., Thierry, H., and De Wolff, C. J. (1998) What is Work and Organizational Psychology? In: *Handbook of Work and Organizational Psychology Volume 1: Introduction to Work and Organizational Psychology*, 1-7. Drenth, P., Thierry, H., and De Wolff, C. J. (Eds), East Sussex: Psychology Press.

75. Duncan, O.D. (1974) Autobiography for the National Academy of Sciences. Available at: http://personal.psc.isr.umich.edu/yuxie-web/files/duncan/Autobio1974.pdf. Accessed 3 September 2008

76. Eberl, M. (2005) An Application of PLS in Multi-Group Analysis: The need for differentiated corporate-level marketing in the mobile communications industry. In: *Proceedings of the PLS'05 International Symposium*, 203-210. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&go.

77. Epstein, R.J. (1989) *A History of Econometrics*. Amsterdam: North Holland.

78. Escofier, B., and Pagès, J. (1998) *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*. Paris: Dunod.

79. Eskildsen, J.K., Westlund, A., and Kristensen, K. (2003) The predictive power of intangibles. *Measuring Business Excellence*, 7(2): 46-54.

80. Eskildsen, J.K., Kristensen, K., and Westlund, A. (2004a) Work motivation and job satisfaction in the Nordic countries. *Employee Relations*, 26(10): 122-136.

81. Eskildsen, J.K., Westlund, A.H., and Kristensen, K. (2004) Measuring Employee Assets: The Nordic Employee Index. *Business Process Management Journal*, 10(5): 537-550.

82. Eskildsen, J., Kristensen, K., and Juhl, H.J. (2005) The consequences of different model specifications when estimating national customer satisfaction indices using PLS. In: *Proceedings of the PLS'05 International Symposium*, 291-298. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go.

83. Esposito, F., Malerba, D., and Semeraro, G. (1997) A Comparative Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5): 476-491.

84. Esposito Vinzi, V., and Lauro, C. (2003) PLS Regression and Classification. In: *Proceedings of the PLS'03 International Symposium*, 45-56. M. Vilares., M. Tenenhaus, P. Coelho, V. Esposito Vinzi, A. Morineau (Eds.), Decisia.

85. Esposito Vinzi, V., Ringle, C.M., Squillacciotti, S., and Trinchera, L. (2007) Capturing and Treating Unobserved Heterogeneity by Response Based Segmentation in PLS Path Modeling: A

Comparison of Alternative Methods by Computational Experiments. ESSEC Working Papers-DR 07019. ESSEC Business School.

86. Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., and Tenenhaus, M. (2008) REBUS-PLS, A Response-based procedure for detecting unit segments in PLS Path Modelling. *Applied Stochastic Models in Business and Industry*, 24(5): 439-459.

87. Falk, R.F., and Miller, N.B. (1992) *A Primer for Soft Modeling*. Ohio: The University of Akron Press.

88. Fletcher, R., and Powell, M.J.D. (1963) A Rapidly Convergent Descent Method for Minimization. *The Computer Journal*, 6(2): 163-168.

89. Fisher, R.A. (1935) *The design of experiment*. Edinburgh, Scotland: Oliver & Boyd.

90. Fomiak, S.Y.X., and Benbasat, I. (2006) The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *Management Information Systems Quarterly*, 30(4): 941-960.

91. Fornell, C. (1982) *A Second Generation of Multivariate Analysis: Methods, Volume I*. New York: Praeger Publishers.

92. Fornell, C. (1987) A Second Generation of Multivariate Analysis: Classification of Methods and Implications for Marketing Research. In: *Review of Marketing*, 407-450. M.J. Houston (Ed). Chicago: American Marketing Association.

93. Fornell, C. (1992) A National Customer Satisfaction Barometer: The Swedish Experience. *Journal of Marketing*, 56(1): 6-21.

94. Fornell, C. (1995) The Quality of Economic Output: Empirical Generalizations about its Distribution and Relationship to Market Share. *Marketing Science*, 14(3): 203-211.

95. Fornell, C. (2007) The Satisfied customer: winners and losers in the battle for buyer preference. New York: Palgrave Macmillan.

96. Fornell, C., and Bookstein, F.L. (1982) Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research*, 19: 440-452.

97. Fornell, C., and Cha, J. (1994) Partial Least Squares. In: *Advanced Methods of Marketing Research*, 52-78. Bagozzi R.P. (Ed). Massachusetts: Blackwell.

98. Fornell, C., Johnson, M.D., Anderson, E.W., Cha, J. and Everitt, B (1996) The American Customer Satisfaction Index: Nature, Purpose and Findings. *Journal of Marketing,* 60(4): 7-18.

99. Furnham, A. (2005) The *Psychology of Behaviour at Work: The Individual in the Organisation*. Psychology Press, New York, USA.

100. Gaertner, S. (1999) Structural Determinants of Job Satisfaction and Organizational Commitment in Turnover Models. *Human Resource Management Review*, 9(4): 479-493.

101. Gefen, D., and Straub, D.W. (1997) Gender Differences in the Perception and Use of E-Mail: An Extension to the Technology Acceptance Model. *Management Information Systems Quarterly*, 21(4): 389-400.

102. Geladi, P. (1988) Notes on the History and Nature of Partial Least Squares (PLS) Modeling, *Journal of Chemometrics* 2: 231-246.

103. Ghilagaber, G. (2004) Another Look at Chow's Test for the Equality of Two Heteroscedastic Regression Models. *Quality & Quantity*, 38: 81-93.

104. Ghosh, M., and Sinha, B.K. (2002) A Simple Derivation of the Wishart Distribution. *The American Statistician*, 56(2): 100-101.

105. Gilbert, C.L., and Qin D. (2005) The First Fifty Years of Modern Econometrics. Department of Economics, University of London. Working Paper No. 544. Available from: http://www.econ.qmul.ac.uk/papers/doc/wp544.pdf. Accessed 3 September 2008.

106. Goldberger, A.S., and Duncan, O.D. (1973) *Structural Equation Models in the Social Sciences*. New York: Seminar.

107. Goldfinger, C. (1997) Intangible Economy and its Implications for Statistics and Statisticians. *International Statistical Review*, 65(2): 191-220.

108. Golub, G.H. (2000) Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123: 35-65.

109. Goodhue, D., Lewis, W. and Thompson, R. PLS, Small Sample Size, and Statistical Power in MIS Research. (Proceedings of the 39th Hawaii International Conference on System Sciences - 2006, HICSS'06, Track 8, 2006).

110. Görz, N., Hildebrandt, L., and Annacker, D. (2000) Analyzing multigroup data with structural equation models. In: *Proceedings of the 23rd Annual Conference of the Gfkl*, 312-319. Berlin: Springer.

111. Grigoroudis, E., and Siskos, Y. (2004) A survey of customer satisfaction barometers: Some results from the transportation-communications sector. *European Journal of Operational Research*, 152(2): 334-353.

112. Hackl, P. and Westlund, A.H. (2000) On structural equation modelling for customer satisfaction measurement. *Total Quality Management,* 11(4,5&6): 820-825.

113. Hackler, J.C. (1970) Testing a Causal Model of Delinquency. *The Sociological Quarterly* 11(4): 511-522.

114. Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C. (1998) *Multivariate Data Anlalysis*. New Jersey: Prentice-Hall.

115. Hahn, C., Johnson, M.D., Herrmann, A., and Huber, A. (2002) Capturing Customer Heterogeneity Using a Finite Mixture PLS Approach. *Schmalenbach Business Review*, 54: 243-269.

116. Harman, H.H. (1968) *Modern Factor Analysis*. Chicago: The University of Chicago Press.

117. Hattie, J. (1984) An Empirical Study of Various Indices for Determining Unidimensionality. *Multivariate Behavioral Research*, 19: 49-78.

118. Hayduk, L.A. (1987) *Structural Equation Modeling with LISREL*. Baltimore: The Johns Hopkins University Press.

119. Helgesen, O. (2000) Are loyal Customers Profitable? Customer Satisfaction, Customer Loyalty and Customer Profitability and the Individual Level. Centre for Fisheries Economics, Working Paper No. 07/2000.

120. Hendry, D.F., and Morgan, M.S. (1994) The ET Interview: Professor H.O.A. Wold 1908-1992. *Economic Theory,* 10: 419-433.

121. Henseler, J. and Fassot, G. (2005) Testing Moderating Effects in PLS Path Models. In: *Proceedings of the PLS'05 International Symposium*, 371-377. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds), SPAD Test&Go.

122. Henseler, J. (2007) A New and Simple Approach to Multi-Group Analysis in Partial Least Squares Path Modeling. In: *Proceedings of the PLS'07 International Symposium*, 104-107. H. Martens and T. Naes. (Eds). Matforsk, As, Norway.

123. Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24: 417-441.

124. Hoyle, R.H. (1995) *Structural Equation Modeling: Concepts, Issues, and Applications*. California: Sage Publications.

125. Hoyle, R.H. (2000) Confirmatory Factor Analysis. In: *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, 465-497. H.E.A. Tinsley and S. D. Brown (Eds). New York: Academic Press.

126. Hsu, S.H., Chen, W.H., and Hsueh, J.T. (2006) Application of Customer Satisfaction Study to Derive Customer Knowledge. *Total Quality Management*, 17(4): 349-454.

127. Hui, B.S., and Wold, H. (1982) Consistency and Consistency at Large in Partial Least Squares Estimates. In: *Systems Under Indirect Observation: Causality, structure, prediction, Part II*, 119-130. K. Jöreskog and H. Wold (Eds). Amsterdam: North Holland.

128. Hulland, J., Chow, Y.H., and Lam, S. (1996) Causal Models in Marketing Research: A Review. *International Journal of Research in Marketing*, 13: 181-197.

129. Hwang, H., and Takane, Y. (2004) Generalized Structured Component Analysis. *Psychometrika*, 69(1): 81-99.

130. Hwang, H., DeSarbo, W.S., and Takane, Y. (2007) Fuzzy Clusterwise Generalized Structured Component Analysis. *Psychometrika*, 72: 181-198.

131. Igbara, M., and Greenhaus, J.H. (1992) Determinants of MIS Employees' Turnover Intentions: A Structural Equation Model. *Communications of the ACM*, 35(2): 35-49.

132. Ilieva, J., Baron, S., and Healy, N.M. (2002) Online surveys in marketing research: pros and cons. *International Journal of Market Research*, 44(3): 361-376.

133. Jackson, J.E. (1991) *A User's Guide to Principal Components*. New York: John-Wiley & Sons.

134. James, L.R., and Brett, J. (1984) Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69: 307-321.

135. Jedidi, K., Jagpal, H.S., and DeSarbo, W.S. (1997) Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science*, 16(1): 39-59.

136. Johnson, M.D.; Herrmann, A., and Huber, F. (2006) The Evolution of Loyalty Intentions. *Journal of Marketing*, 70: 122-132.

137. Jolliffe, I.T. (2002) *Principal Component Analysis*. Second Edition. New York: Springer.

138. Jöreskog, K.G. (1967) Some Contributions to Maximum Likelihood Factor Analysis. *Psychometrika*, 32(4): 443-482.

139. Jöreskog, K.G. (1973) A General Method for Estimating a Linear Structural Equation System. In: *Structural Equation Models in Social Sciences*, 85-112. A.S. Goldberger and O.D. Duncan (Eds). London: Academic Press.

140. Jöreskog, K.G. (1977) Structural Equation Models in the Social Sciences: Specification, Estimation and Testing. In: *Applications of Statistics*, 265-287. R. Krishnaiah (Ed). Amsterdam: North-Holland.

141. Jöreskog, K.G. (2004) Factor Analysis and Its Extensions. Factor Analysis at 100: Historical Developments and Future Decisions. Department of Psychology, University of North Carolina. May 13-15, 2004. Available from: http://www.fa100.info/joreskog.pdf. Accessed 3 September 2008.

142. Jöreskog, K.G., and Goldberger, A.G., (1972) Factor Analysis by generalized least squares. *Psychometrika*, 37: 243-260.

143. Jöreskog, K.G., and Lawley, D.N. (1968) New Methods in Maximum Likelihood Factor Analysis. *Statistical Psychology*, 21: 85-96.

144. Jöreskog, K.G., and Sörbom, D. (1978). *LISREL IV: analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

145. Jöreskog, K.G., and Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Massachusetts: Abt Books.

146. Jöreskog, K.G., Sörbom, D. (1986) *PRELIS: A Program for Multivariate Data Screening and Data Summarization*. Scientific Software, Mooresville, IL.

147. Jöreskog, K.G., and Wold, H. (1982) *Systems under indirect observation: Causality, structure, prediction, Part I and Part II*. Amsterdam: North Holland.

148. Jöreskog, K.G., and Wold, H. (1982) The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. In: *Systems under indirect observation: Causality, structure, prediction. Part I*, 263-270. K.G. Jöreskog & H. Wold (Eds). Amsterdam: North Holland.

149. Känd, M., and Rekor, M. (2005) Perceived Involvement in Decision Making and Job Satisfaction: The Evidence from a Job Satisfaction Survey among Nurses in Estonia. SSE Riga Working Papers, 6, Stockholm School of Economics in Riga. (Available from: http://www.sseriga.edu.lv/library/working_papers/FT_2005_6.pdf)

150. Kaplan, D. (2000) *Structural Equation Modeling: Foundations and Extensions*. California: Sage Publications.

151. Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 129(2): 119-127.

152. Keil, M., Tan, B.C.Y., Wei, K.K., Saarinen, T., Tuunainen, V., and Wassenaar, A. (2000) A Cross-Cultural Study on Escalation on Commitment Behavior in Software Projects. *Management Information Systems Quarterly*, 24(2): 299-325

153. Kenny, D.A., and Judd, C.M. (1984) Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulleting*, 96: 201-210.

154. Kim, S. (1999) Behavioral Commitment Among the Automobile Workers in South Korea. *Human Resource Management Review*, 9(4): 419-451.

155. Knight, A. (1978) Common Factor Analysis: Some Recent Developments in Theory and Practice. *The Statistician*, 27(1): 27-42.

156. Kotz, S., and Johnson, N.L. (1983) Factor Analysis. *Encyclopedia of Statistical Sciences* 3: 2-8.

157. Kristensen, K., Martensen, A., and Gronholdt, L. (1999) Measuring the impact of buying behaviour on customer satisfaction. *Total Quality Management*, 10(4&5): 602-614.

158. Kristensen, K., Juhl, H.J., and Ostergaard, P. (2001) Customer satisfaction: some results for European Retailing. *Total Quality Management*, 12(7&8): 809-897.

159. Kristensen, K., and Westlund, A.H. (2004) Performance Measurement and Business Results. *Total Quality Management*, 15(5–6): 719–733.

160. Kumar, V., and Deregowska, D. (2002) Latent Structural Modeling: Choosing Between Structural Equation Modeling and Partial Least Squares. ASAC 2002, Winnipeg, Manitoba. Available from: http://luxor.acadiau.ca/library/ASAC/v23/230201.pdf. Accessed 3 September 2008.

161. Lawley, D.N. (1940) The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60: 64-82.

162. Lawley, D.N., and Maxwell, A.E. (1962) Factor Analysis as a Statistical Method. *The Statistician*, 12(3): 209-229.

163. Lawley, D.N., and Maxwell, A.E. (1971) *Factor Analysis as a Statistical Method*. London: Butterworth.

164. Lebart, L., Morineau, A., and Fénelon J.P. (1979) *Traitement des données statistiques*. Paris: Dunod.

165. Lebart, L., Morineau, A,. and Fénelon J.P. (1985) *Tratamiento estadístico de datos*. Barcelona: Marcombo.

166. Lebart, L., Morineau, A., and Piron, M. (2004) *Statistique Exploratoire Multidimensionnelle*. Paris: Dunod.

167. Lev, B. (2001) *Intangibles: Management, Measurement, and Reporting*. Washington: Brookings Institution Press

168. Loehlin, J.C. (1987) *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Hillsdale, NJ: Erlbaum.

169. Lohmöller, J. B. (1989) *Latent Variable Path Modeling with Partial Least Squares*. Heidelberg: Physica-Verlag.

170. Long, J.S. (1983) *Covariance Structure Models: An Introduction to LISREL*. California: Sage Publications.

171. López, C., Fernández, K., and Mariel, P. (2003) Spanish Customer Satisfaction Indices by Cumulative Panel Data. A Marketing Application for Automobile Industry. In: *Proceedings of the PLS'03 International Symposium*, 97-100. M. Vilares., M. Tenenhaus, P. Coelho, V. Esposito Vinzi, A. Morineau (Eds), Decisia.

172. Lubke, G.H., and Muthén, B. (2005) Investigating Population Heterogeneity with Factor Mixture Models. *Psychological Methods*, 19(1): 21-39.

173. Marcoulides, G.A., and Hershberger, S.L. (1997) *Multivariate Statistical Methods: A First Course*. New Jersey: Lawrence Erlbaum Associates.

174. Marsh, H.W., Balla, J.R., and McDonald, R.P. (1988) Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size. *Psychological Bulletin*, 103(3): 391-410.

175. Martens, H. (2001) Reliable and relevant modeling of real world data: a personal account of the development of PLS Regression. *Chemometrics and Intelligent Laboratory Systems,* 58(2): 85-95.

176. Martensen, A., Gronholdt, L., and Kristensen, K. (2000) The drivers of customer satisfaction and loyalty: cross-industry findings from Denmark. *Total Quality Management*, 11(4/5&6): 544-553.

177. Martin, J.T. (1997) An Exact Probability Metric for Decision Tree Splitting and Stopping. *Machine Learning*, 28: 257-291.

178. Mathieson, K., Peacock, E., and Chin, W. W. (2001) Extending the Technology Acceptance Model: The Influence of Perceived User Resources. *The Data Base for Advances in Information Systems*, 32(3): 86-112.

179. McLahlan, G.J., and Krishnan, T. (1997) The EM-Algorithm and Extensions. New York. Wiley.

180. Money, R.B., and Graham, J.L. (1999) Salesperson Performance, Pay, and Job Satisfaction: Tests of a Model Using Data Collected in the United States and Japan. *Journal of International Business Studies*, 30(1): 149-172.

181. Moreno, E., Torres, F., and Casella, G. (2005) Testing equality of regression coefficients in heteroscedastic normal regression models. *Journal of Statistical Planning and Inference*, 131: 117-134.

182. Morgan, M.S. (1990) *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

183. Mulaik, S.A. (1986) Factor Analysis and Psychometrika: Major Developments. *Psychometrika*, 51(1): 23-33.

184. Murayama, G.M. (1998) *Basics of Structural Equation Modeling*. Thousand Oaks, California: SAGE Publications.

185. Murthy, S.K. (1998) Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2: 345-389.

186. Muthén, B.O. (1989) Latent Variable Modeling in Heterogeneous Populations. *Psychometrika*¸ 54: 557-585.

187. Muthén, B.O. (2002) Beyond SEM: General Latent Variable Modeling. *Behaviormetrika*, 29(1): 81-117.

188. Nakache, J.P. and Confais, J. (2003) *Statistique explicative appliquée*. Paris: Editions Technip.

189. Neyman, J. and Pearson, E.S: (1928) On the use and interpretation of certain test criteria for purposes of statistical inference (Part I). *Biometrika*, 20A: 175-240.

190. Neyman, J. and Pearson, E.S: (1928) On the use and interpretation of certain test criteria for purposes of statistical inference (Part II). *Biometrika*, 20A: 263-294.

191. Nollmann, G., and Strasser, H. (2005) The History of Sociology: The European Perspective. Available from: http://soziologie.uni-uisburg.de/personen/nollmann/The_history_of_sociology.pdf. Accessed 3 September 2008.

192. Noonan, R., and Wold, H. (1982) PLS Path Modeling with Indirectly Observed Variables: A Comparison of Alternative Estimates for the Latent Variable. In: *Systems under indirect observation: Causality, structure, prediction*. Part II, 75-94. K.G. Jöreskog and H. Wold (Eds). Amsterdam: North Holland.

193. Noonan, R.D., and Wold, H. (1988) Partial Least Squares Path Analysis. In: *Educational Research, Methodology, and Measurement: An International Handbook*, 710-717. John P. Keeves (Ed). Oxford: Pergamon Press.

194. O'Loughlin, C., and Coenders, G. (2004) Estimation of the European Customer Satisfaction Index: Maximum Likelihood versus Partial Least Squares. Application to Postal Services. *Total Quality Management*, 15(9-10): 1231-1255.

195. Pagès, J., and Tenenhaus, M. (2001) Multiple factor analysis combined with PLS path modeling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and Intelligent Laboratory Systems*, 58: 261-273.

196. Palumbo, F., and Romano, R. (2008) Possibilistic PLS Path Modeling: A New Approach to the Multigroup Comparison. In: *Proceedings in Computational Statistics*, 303-314. Paula Brito (Ed), Heidelberg: Physica-Verlag.

197. Pan, V.Y. (1997) Solving Polynomial Equation: Some History and Recent Progress. *Society for Industrial Applied Mathematics*, 39(2): 187-220.

198. Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2: 559-572.

199. Pugesek, B.H. (2003) Concepts of structural equation modeling in biological research. In: *Latent Variables Analysis: Applications for Developmental Research*, 36-42. Alexander Von Eye and Clifford C. Clogg (Eds). California: Sage Publications.

200. Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, 1: 81-106.

201. Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. California: Morgan Kauffman.

202. Quinlan, J.R. (1998) C5/See5 Software.

203. R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

204. Raymond, L., and St-Pierre, J. (2005) Antecedents and performance outcomes of advanced manufacturing systems sophistication in SMEs. *International Journal of Operations & Production Management*, 25(6): 514-533.

205. Reinartz, W.J., Echambadi, R., and Chin, W.W. (2002) Generating Non-normal Data for Simulation of Structural Equation Models Using Mattson's Method. *Multivariate Behavioral Research*, 37(2): 227-244.

206. Ringle C.M., Wende S., and Will, A. (2005) Customer Segmentation with FIMIX-PLS. In: *Proceedings of the PLS'05 International Symposium*, 507-514. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go.

207. Ringle, C.M. (2006) Segmentation for Path Models and Unobserved Heterogeneity: The Finite Mixture Partial Least Squares Approach. Research Papers on Marketing and Retailing, University of Hamburg, No. 35, November 2006.

208. Ringle, C.M., and Schlittgen, R. (2007) A Genetic Algorithm Segmentation Approach for Uncovering and Separating Groups of Data in PLS Path Modeling. In: *Proceedings of the PLS'07 International Symposium*, 75-78. H. Martens and T. Naes. (Eds.). Matforsk, As, Norway.

209. Rubin, D.B., and Thayer, D.T. (1982) EM algorithms for ML factor analysis. *Psychometrika*, 47: 69–76.

210. Rubin, D.B., and Thayer, D.T. (1983) More on EM for ML factor analysis. *Psychometrika*, 48: 253-257.

211. Rust, J., and Golombok, S. (1999) *Modern Psychometrics: The Science of Psychological Assessment*. New York: Routledge.

212. Safavian, S.R., and Landgrebe, D. (1991) A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3): 660-674.

213. Sánchez, G., and Aluja, T. (2006) PATHMOX: A PLS-PM Segmentation Algorithm. In: Electronic Proceedings of the Workshop on Knowledge Extraction and Modeling (KNEMO), V. Esposito Vinzi, C. Lauro, A. Braverman, H.A.L. Kiers, M.G. Schimek (Eds).

214. Sánchez, G., and Aluja, T. (2007) A Simulation Study of PATHMOX (PLS Path Modeling Segmentation Tree) Sensitivity. In: *Proceedings of the PLS'07 International Symposium*, 33-36. H. Martens and T. Naes. (Eds), Matforsk, As, Norway.

215. Saporta, G. (2006) *Probabilités, analyse de données et statistique*. Paris: Editions Technip.

216. Saporta, G. (2008) Models for Understanding versus Models for Prediction. In: *Proceedings in Computational Statistics*, 315-322. Paula Brito (Ed), Heidelberg: Physica-Verlag.

217. Scheffé, H. (1959) *The Analysis of variance*. New York: Wiley.

218. Seber, G.A.F. (1966) *The linear Hypothesis: A general theory.* Griffin's Statistical Monographs and Courses. London: Charles Griffini and Co.

219. Serch, O. (2008) *Sistema de Visualització de models PLS-PM*. Projecte Final de Carrera. Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya. Enero, 2008.

220. Sellin, N. (1995) Partial Least Square Modeling in Research on Educational Achievement. In: *Reflections on Educational Achievement; Papers in Honour of T. Neville Postlethwaite*, 256-267. Wilifred Bos and Rainer H. Lehmann (Eds), New York: Waxmann Munster.

221. Simon, H.A. (1996) *Models of My Life*. New York: The MIT Press.

222. Smith, W.R. (1956) Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, 21(1), 3-8.

223. Sobel, M.E. (1992) Review: The American Occupational Structure and Structural Equation Modeling in Sociology. *Contemporary Sociology* 21(5): 662-666.

224. Sobel, M.E. (1994) Causal Inference in Latent Variable Models. In: *Latent Variables Analysis: Applications for Developmental Research*, 3-35. Alexander Von Eye and Clifford C. Clogg (Eds). California: Sage Publications.

225. Sonquist, J.A., and Morgan, J.N. (1964) *The Detection of Interaction Effects*. Institute for Social Research, University of Michigan.

226. Sonquist, J.A., Baker, E.L., and Morgan, J.N. (1971) *Searching for Structure*. Institute for Social Research, University of Michigan.

227. Spearman, C. (1904) General Intelligence, objectively determined and measured. *American Journal of Psychology*, 15: 201-293.

228. Squillacciotti, S. (2005) Prediction oriented classification in PLS Path Modelling. In: *Proceedings of the PLS'05 International Symposium*, 499-506. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go.

229. Squillacciotti, S., Trinchera, L., and Esposito Vinzi, V. (2006) PLS Typological Path Modeling: a model-based approach to classification. In: Electronic Proceedings of the Workshop on Knowledge Extraction and Modeling (KNEMO), V. Esposito Vinzi, C. Lauro, A. Braverman, H.A.L. Kiers, M.G. Schimek (Eds).

230. Stan, V., and Saporta, G. (2003) Conjoint use of variables clustering and PLS structural equations modelling. . In: *Proceedings of the PLS'05 International Symposium*, 133-140. T. Aluja, J. Casanovas, V. Esposito, A. Morineau, M. Tenenhaus (Eds.), SPAD Test&Go.

231. Staples, S., and Higgins, C.A. (1998) A study of the impact of factor importance weightings on job satisfaction measures. *Journal of Business and Psychology*, 13(2): 211-233.

232. Tan, P., Steinbach, M. and Kumar, V. (2006) *Introduction to Data Mining*. Addison Wesley, Boston, USA.

233. Tanaka, H., Uejima, S., and Asai, K. (1982) Linear regression analysis with fuzzy model. *IEEE Transactions Systems Man Cybernet*, 12: 903:907.

234. Taylor, D.S., and Chin, W.W. (2004) Drivers of Turnover Decisions of Information Systems Personnel: Job Satisfaction Versus Job Fit with Quality of Life Goals. *Proceedings of the Academy of Information and Management Sciences*, 8(2): 51-54.

235. Temme, D., Kreis, H., and Hildebrandt, L. (2006) PLS Path Modeling: A Software Review. SFB 649 Discussion Paper 2006-084, Economic Risk.

236. Tenenhaus, M. (1998) *La Régression PLS*. Paris: Éditions Technip.

237. Tenenhaus, M. (1999) L'approche PLS. *Revue de Statistique Appliquée,* 47(2): 5-40.

238. Tenenhaus, M. (2006) *Statistique*. Paris: Dunod.

239. Tenenhaus, M. (2008) Component-based Structural Equation Modelling. *Total Quality Management & Business Excellence*, 19(7&8): 871-886.

240. Tenenhaus, M., Chatelin, Y.M., and Esposito Vinzi, V. (2002) State-of-art on PLS Path Modelling through the available software Research Papers series of HEC School of Management, Jouy-en-Josas, July, 2002. Available at: www.esisproject.com/pdf/PLS_State_of_art.pdf. Accessed 3 September 2008.

241. Tenenhaus, M., and Esposito Vinzi, V. (2005) PLS regression, PLS path modeling and generalizaed Procrustean analysis: a combined approach for multiblock analysis. *Journal of Chemometrics*, 19(3): 145-153.

242. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., and Lauro, C. (2005) PLS path modeling. *Computational Statistics & Data Analysis*, 48: 159-205.

243. Tenenhaus, M., Guinot, C., and Latreille, J. (2001a) PLS path modelling and multiple table analysis. Application to the cosmetics habits of women in Ile-de-France. *Chemometrics and Intelligent Laboratory Systems*, 58: 247-259.

244. Tenenhaus, M., Mauger, E., and Guinot, C. (2006) Test of a group effect in a regression model relating two blocks of binary variables with ULS-SEM and SEM-PLS. In: Electronic Proceedings of the Workshop on Knowledge Extraction and Modeling (KNEMO), V. Esposito Vinzi, C. Lauro, A. Braverman, H.A.L. Kiers, M.G. Schimek (Eds).

245. Tenenhaus, M., and Naes, T. (2001) Foreword. *Chemometrics and Intelligent Laboratory Systems,* 58(2): 77-81.

246. Tenenhaus, M., and Pagès, J. (2001b) Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between psychometrical variables, sensory profiles and hedonic judgments. *Chemometrics and Intelligent Laboratory Systems*, 58: 261-273.

247. Tenenhaus, M., and Pagès, J. (2002) Analyse Factorielle Multiple et Approche PLS. *Revue de Statistique Appliquée*, 50(1): 5-33.

248. Thatcher, J.B., Stepina, L.P., and Boyle, R.J. (2003) Turnover of Information Technology Workers: Examining Empirically the Influence of Attitudes, Job Characteristics, and External Markets. *Journal of Management Information Systems*, 19(3): 231-261.

249. Thatcher, J.B., Liu, Y., Stepina, L.P., Goodman, J.M, and Treadway, D.C. (2006) IT Worker Turnover: An Empirical Examination of Intrinsic Motivation. *The DATA BASE for Advances in Information Systems*, 37(2&3): 133-146.

250. Thierry, H. (1998) Motivation and Satisfaction. In: *Handbook of Work and Organizational Psychology Volume 4: Organizational Psychology*, 253-285. Drenth, P., Thierry, H., and De Wolff, C. J. (Eds), East Sussex: Psychology Press.

251. Timm, N.H., (2002) *Applied Multivariate Analysis*. New York: Springer-Verlag.

252. Tomer, A. (2003) A Short History of Structural Equation Models. In: *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology,* 85-124. Bruce H. Pugesek, Adrian Tomer & Alexander Von Eye (Eds). Cambridge: Cambridge University Press.

253. Trinchera, L., Squillacciotti, S., Esposito Vinzi, V. and Tenenhaus, M. (2007) PLS path modeling in presence of a group structure: REBUS-PLS, a new response-based approach. In: *Proceedings of the PLS'07 International Symposium*, 79-82. H. Martens and T. Naes. (Eds.),Matforsk, As, Norway.

254. Tufféry, S. (2007), *Data Mining et statistique décisionnelle*. Paris: Editions Technip.

255. Venkatesh, V., and Morris, M.G. (2000) Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *Management Information Systems Quarterly*, 24(1): 115-139.

256. Vilares, M.J. and Coelho, P.S. (2003) The employee-customer satisfaction chain in the ECSI model. *European Journal of Marketing*, 37: 1703-1722.

257. Vinacua, B.V. (1986) *Modelos Causales: Técnicas de Investigación Social*. Barcelona: Hispano Europea.

258. Watkins, D.S. (1982) Understanding the QR algorithm. *Society for Industrial and Applied Mathematics*, 24(4): 427-440.

259. Wedel, M., and Kamakura, W. (2000) *Market Segmentation: Conceptual and Methodological Foundations*. Springer, USA.

260. Westlund, A.H., Cassel, C.M, Eklöf, J., and Hackl, P. (2001) Structural analysis and measurement of customer perceptions, assuming measurement and specifications errors. *Total Quality Management*, 12(7&8): 873-881.

261. Whittle, P. (1992) Obituary: Professor Herman Wold. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 155(3): 466-469.

262. Wold, H. (1966a) Nonlinear estimation by iterative least squares procedures. In: *Research Papers in Statistics, Festschrift for J. Neyman*, 411-444. F. David (Ed). New York: Wiley.

263. Wold, H. (1966b) Estimation of Principal Components and Related Models by Iterative Least Squares. In: *Multivariate Analysis, Proceedings of an International Symposium,* 391-420. P. R. Krishnaiah (Ed). New York: Academic Press.

264. Wold, H. (1969) Model Building and Scientific Method: A Graphic Introduction. In: *Mathematical Model Building in Economics and Industry*, 143-158. London: Charles Griffin & Co. LTD

265. Wold, H. (1973a) Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In: *Multivariate Analysis III, Proceedings of the Third International Symposium on Multivariate Analysis*, 383-407. P.R. Krishnaiah (Ed). New York: Academic Press.

266. Wold, H. (1973b) Aspects opératoires des modèles économétriques et sociologiques. Développement actuel de l'estimation "F.P." (Point fixe) et de la modélisation "NIPALS" (linéarisation par itération de moindres carrés partiels). *Économie Appliquée* N° 2-3-4: 389-421.

267. Wold, H. (1975) Soft Modelling by latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. In: *Perspectives in Probability and Statistics: Papers in honour of M. S. Bartlett on the occasion of his sixty-fifth birthday*, 117-142. J. Gani (Ed). Sheffield, Eng.: Applied Probability Trust.

268. Wold, H. (1979) Model construction and evaluation when theoretical knowledge is scarce. An example of the use of Partial Least Squares. Cahier 79:106. Department of Econometrics, University of Geneva.

269. Wold, H. (1980) Model construction and evaluation when theoretical knowledge is scarce: On the theory and application of Partial Least Squares. In: *Model Evaluation in Econometrics*, 47-74. J. Kmenta, and J. Ramsey. (Eds). New York: Academic Press.

270. Wold, H. (1982a) Soft modeling: The Basic Design and Some Extensions. In: *Systems under indirect observation: Causality, structure, prediction. Part II*, 1-54. K.G. Jöreskog & H. Wold (Eds). Amsterdam: North Holland.

271. Wold, H. (1982b) Models for Knowledge. In *The Making of Statisticians*, 89-212. J. Gani (Ed). Berlin: Springer-Verlag.

272. Wold, H. (1985a) Partial least squares. In: *Encyclopedia of Statistical Sciences, Vol. 6*, 581-591. Kotz, S., and Johnson, N.L. (Eds). New York: Wiley.

273. Wold, H. (1985b) Specification, Predictor. In: *Encyclopedia of Statistical Sciences, Vol. 8,* 587-599. Kotz, S., and Johnson, N.L. (Eds). New York: Wiley.

274. Wold, H. (1985c) Systems Analysis by Partial Least Squares. In: *Measuring the Unmeasurable*, 221-251. P. Nijkamp, H. Leitner, and N. Wrigley (Eds). Boston: Martinus Nijhoff Publishers.

275. Wold, S. (2001) Personal memories of the early PLS development. *Chemometrics and Intelligent Laboratory Systems,* 58(2): 83-84.

276. Wold, S. Eriksson, L., Trygg, J., and Kettaneh, N. (2004) The PLS method: partial least squares projections to latent structures and its applications in industrial RDP (research, development, and production). Available from: http://www.umetrics.com/pdfs/events/prague%200408%20__%20PLS_text_wold.pdf. Accessed 3 September 2008.

277. Wolfe, L.M. (2003) The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography. *Structural Equation Modeling,* 10(1): 1-34.

278. Wright, S. (1972) Path Coefficients and Path Regressions: Alternative or Complementary Concepts? In: *Causal Models in the Social Sciences*, 101-114. H. M. Blalock (Ed). Chicago: Aldine Atherton.

279. Young, D.J. (1998) Ambition, Self-Concept, and Achievement: A Structural Equation Model for Comparing Rural and Urban Students. *Journal of Research in Rural Education,* 14(19): 34-93.

280. Yung, Y.F. (1997) Finite Mixtures in Confirmatory Factor Analysis Models. *Psychometrika*, 62(3): 297-330.