Universitá degli Studi di Napoli
Federico II

Ph.D. Thesis
in Statistics
XXVII Ciclo

Component-based Path Modeling

Open Issues and Methodological Contributions

Pasquale Dolce

Dipartimento di Scienze Economiche e Statistiche

Component-based Path Modeling

Open Issues and Methodological Contributions

Author:
Pasquale Dolce

Universitá degli Studi di Napoli
Dipartimento di Scienze economiche
e statistiche
Via Cintia - Complesso Monte S.Angelo, 26
email: pasquale.dolce@unina.it

Supervisor:
Carlo Lauro

Committee Members:
Vincenzo Esposito Vinzi
Stella Iezzi
Michele La Rocca
Francisco de A. T. de Carvalho

European Referees:
Tomàs Aluja
Gilbert Saporta

April 2015

a mia madre

Acknowledgements

L’idea di partenza di questo lavoro è scaturita da diversi pomeriggi, intensi e sempre troppo corti, trascorsi con il Prof. Carlo Lauro, il quale è stato molto più di un “tutor”, ha costituito e costituisce la stella polare di idee e suggerimenti, ripensamenti e crisi: ogni diversità di vedute è stata per me occasione di riflessione, a volte di autocritica. È doveroso porre i miei ringraziamenti al Prof. Carlo Lauro, inoltre, per avermi dato la possibilità di poter conoscere altri professori che hanno contribuito non solo alla mia crescita intellettuale ma anche a quella umana.

Tra questi c’è il Prof. Vincenzo Esposito Vinzi che mi ha accolto a Parigi con la generosità, lieve e gratuita, che lo contraddistingue, facendomi sentire a casa sin dal primo giorno. Ritengo che il Prof. Vinzi abbia contribuito in maniera decisiva alla mia crescita professionale e personale e per questo ci tengo a ringraziarlo.

Durante questo soggiorno di studio a Parigi, Laura Trinchera e Giorgio Russolillo mi hanno offerto il bene prezioso della loro amicizia. Ho apprezzato molto la loro disponibilità ad aiutarmi in ogni difficoltà che si incontra quando si va a vivere in un paese straniero. Ricordo con piacere la cena di benvenuto, dove abbiamo incominciato da subito a discutere dei limiti e prospettive dei “nostri” modelli. Dalle loro critiche ho imparato molto e auspico di continuare ad imparare, ma sopratutto spero di continuare a coltivare questa amicizia.

Vorrei inoltre ringraziare la Professoressa Cristina Davino, la quale mi ha seguito costantemente e attentamente nella stesura di una parte di questo lavoro. Il quarto capitolo della tesi è il frutto di un intenso lavoro svolto insieme.

I miei ringraziamenti vanno inoltre alla mia miglior collega Carmela Iorio. In realtà più che una collega si è rilevata essere un’ottima amica a cui confidare problemi e gioie di vita. Per quanto riguarda i problemi, spesso ha trovato la soluzione ”ottima” in tempi più che soddisfacenti.

Intendo infine ringraziare Massimo Aria e Antonio D’Ambrosio, con i quali ho condiviso buona parte di questo percorso fatto di gioie e difficoltà e per il supporto che mi hanno dato in ogni occasione.

I would like to thank Prof. Gilbert Saporta and Prof. Tomàs Aluja who agreed to review this thesis. I benefited a lot from their comments and remarks which have contributed to revise some parts of the thesis.

Contents

Acknowledgements **iv**

Contents **v**

List of Figures **viii**

List of Tables **x**

Introduction **1**

1 Component-based Predictive Path Modeling: Recent Developments and Open Issues **6**

1.1 Introduction 6

1.2 Is PLS-PM an alternative approach for "latent variable" modeling? 8

1.3 Is PLS-PM a method for Structural Equation Modeling? 9

1.4 PLS path modeling 12

1.4.1 Model Specification 12

1.4.2 The Algorithm 13

1.4.3 Optimizing Criteria 14

1.5 Prediction-oriented Component-based Methods 16

1.5.1 In-of-Sample Prediction 17

1.5.2 Out-of-Sample Prediction 18

1.6 Misspecification of the Measurement Model 21

1.7 The Path Direction Incoherence in PLS Path Modeling 22

1.8 A more comprehensive analysis of the Relationships in Component-based Path Modeling 24

1.9 Other critical issues in PLS-PM 26

2 Formative Versus Reflective Measurement Model in Structural Equation Modeling **29**

2.1 Introduction 29

2.2 Nature and Direction of the relationships between latent variables and manifest variables 30

2.2.1 Internal consistency reliability and the issue of Multicollinearity 31

2.2.2 Validity of indicators . . . 32
2.2.3 Interchangeability of the manifest variables . . . 34
2.2.4 The error term in formative measurement model . . . 35
2.2.5 Surplus meaning of the latent variable . . . 36
2.2.6 Criteria for Distinguishing Between Reflective and Formative measurement Models . . . 36
2.2.7 Misspecification of relationships between latent variables and manifest variables . . . 37
2.3 Formative blocks in Covariance-based SEM . . . 38
2.4 A simulation Study . . . 40
2.4.1 Design of the Simulation Study . . . 40
2.4.2 Data Generation and Simulation Results . . . 42
2.5 Conclusion and Future Research . . . 48

3 Non-Symmetrical Component-based Path Modeling 50
3.1 Introduction . . . 50
3.2 Dependence and Interdependence Relationships Between Blocks . . . 52
3.3 PLS-PM incoherence with Path Directions . . . 54
3.3.1 PLS-PM Solutions for a Two-Block Model . . . 54
3.3.2 Illogical form of Redundancy Analysis in PLS-PM . . . 58
3.3.3 PLS-PM solutions for Multi-Block Models . . . 60
3.4 The Proposed Method . . . 62
3.4.1 Model Specification . . . 63
3.4.2 The algorithm . . . 64
3.4.3 Model Assessment . . . 67
3.4.4 A solution to the issue of Multicollinearity . . . 72
3.5 A Comparison with other Component-based approaches . . . 74
3.5.1 Other Component-based approaches for multi-block data . . . 74
3.5.2 Design of the Simulation Study . . . 76
3.5.3 Data Generation and Simulation Results . . . 78
3.6 Conclusion . . . 88

4 Quantile Component-based Path Modeling: proposed methods, performances and interpretations 90
4.1 Introduction . . . 90
4.2 Quantile Regression . . . 92
4.3 Quantile Correlation . . . 96
4.4 Quantile Composite-based Path Modelling . . . 98
4.5 A real data application . . . 103
4.6 Simulation Study . . . 110
4.6.1 Design of the simulation study . . . 111
4.6.2 The Case of Homogeneity . . . 112
4.6.3 The Case of Heterogeneity in the Structural Model . . . 114
4.6.4 The Case of Heterogeneity in the Measurement Model . . . 117
4.7 Conclusion . . . 119

Conclusion **122**

Bibliography **123**

List of Figures

2.1 Different measurement models 31
2.2 Mean of the $RMSE$ in the path coefficients for normal data scenario 43
2.3 Mean of the $RMSE$ in the path coefficients for non-normal data scenario 45

3.1 Commonly used graphical notation in Structural Equation modeling 55
3.2 Two-block model with outwards directed scheme 56
3.3 Two-block model with outwards directed scheme: not oriented arrows 57
3.4 Two-block model with inwards directed scheme 57
3.5 Two-block model with inwards directed scheme: not oriented arrows 58
3.6 Two-block model with inwards and outwards directed scheme: Redundancy analysis 58
3.7 Two-block model for an illogical form of redundancy analysis in PLS-PM 59
3.8 Theoretical Model for Simulated Data 77
3.9 Communalities in PLS-PM(B,B,B) and NSC-PM 79
3.10 Redundancies in PLS-PM(B,B,B) and NSC-PM 80
3.11 R^2 of the structural model in PLS-PM(B,B,B) and NSC-PM 80
3.12 Distributions of path coefficient estimates 81
3.13 Communalities in PLS-PM(A,A,A) and NSC-PM 81
3.14 Redundancies in PLS-PM(A,A,A) and NSC-PM 82
3.15 R^2 of the structural model in PLS-PM(B,B,B) and NSC-PM 83
3.16 Distributions of path coefficient estimates 83
3.17 Communalities in RGCCA(ridge mode) and NSC-PM 84
3.18 Redundancies in RGCCA(ridge mode) and NSC-PM 84
3.19 R^2 of the structural model in RGCCA(ridge mode) and NSC-PM . 85
3.20 Communalities in RGCCA(0,0.5,1) and NSC-PM 85
3.21 Redundancies in RGCCA(0,0.5,1) and NSC-PM 86
3.22 R^2 of the structural model in RGCCA(0,0.5,1) and NSC-PM 86
3.23 Communalities in GSCA and NSC-PM 87
3.24 R^2 of the structural model in GSCA and NSC-PM 88
3.25 Evolution of the criterion as function of iterations for the algorithm 89

4.1 Scatter plot and QC values 97
4.2 Outer model schemes: *Mode A* and *Mode B* 99

4.3 Outer model schemes: *Mode Q* outwards-directed and *Mode Q* inwards-directed . . . 99
4.4 Path diagram of the ACSI model . . . 104
4.5 Maximum quantile value for each MV . . . 105
4.6 Outer weights with respect to the MV averages of the *Expectation* LV . . . 107
4.7 QC-PM path coefficients for a set of quantiles . . . 108
4.8 Theoretical Model for Simulated Data . . . 111
4.9 Path coefficient estimates across quantiles . . . 114
4.10 Theoretical Models for Simulated Data . . . 115
4.11 Simulated Distribution of a generic manifest variable . . . 116
4.12 Path coefficient estimates across quantiles . . . 117
4.13 Theoretical Model for Simulated Data . . . 118
4.14 Outer weights across quantiles . . . 119

List of Tables

2.1 RBias, StD and RMSE of outer weights . . . 45

2.2 RBias, StD and RMSE of outer weights . . . 46

2.3 Inner Paths Coefficients RBias, StD and RMSE . . . 47

2.4 Inner Paths Coefficients RBias, StD and RMSE . . . 47

4.1 The different estimation options for QC-PM . . . 100

4.2 LVs and MVs of ACSI dataset: Means and main quantile values . . . 104

4.3 Outer weights . . . 106

4.4 PLS-PM path coefficients and QC-PM path coefficients . . . 106

4.5 Correlations and Quantile Correlations LV-MVs . . . 108

4.6 Measurement model assessment indexes . . . 109

4.7 R^2 and $pseudoR^2$ in the Structural model . . . 110

4.8 Redundancy measures . . . 110

4.9 Goodness of fit (GoF) indices . . . 110

Introduction

Structural Equation Modeling (SEM) is a powerful multivariate analysis technique that allow us to analyze relationships among several blocks of observed variables, called manifest variables (MV), by summarizing them with a few number of unobserved variables, the so-called latent variables (LV).

In 1970 Karl Jöreskog first proposed to use a covariance-based approach to analyse causal relationships - defined according to a theoretical model - linking two or more latent complex concepts, each measured through a number of observable variables (Jöreskog, 1970). Maximum likelihood method (ML-SEM) is one of the most well-known covariance-based estimation methods for SEM (Jöreskog, 1970, 1973, 1977).

Quite at the same time (in 1975), Herman Wold finalized the so-called *soft modeling* approach for analyzing relationships among several blocks of variables linked by a network of relations specified by a path diagram: the PLS Path Modeling (PLS-PM) (Wold, 1975a,b, 1982).

PLS-PM was originally presented as an alternative approach to the covariance-based SEM (Jöreskog and Wold, 1982b). However, the two approaches belong to two families of statistical methods.

The origin of the Partial Least Squared (PLS) methods goes back to the idea of Herman Wold who in 1996 devised the NILES (Non-linear Iterative Least Squares) (Wold, 1966a,b), an iterative algorithm based essentially on a sequence of simple Ordinary Least Squares (OLS) regressions, and proposed it as an alternative estimation method for Principal Components Analysis (Hotelling, 1933). NILES was later re-named Non-linear Iterative PArtial Least Squares (NIPALS) by the same author (Wold, 1975b) and it was then extended to a more general technique that analyzes several blocks of variables linked by a network of relations specified by a path diagram. Thus, it was proposed to the estimation of SEM parameters, as a Soft Modeling (Wold, 1982) alternative to Jöreskog's approach (Jöreskog, 1970).

This technique is well known with the name PLS Path Modeling (PLS-PM). The acronym PLS (Partial Least Square) has also been interpreted by H. Wold et al. as Projection to Latent Structures. Since this interpretation has a more descriptive meaning we opt for it in this dissertation. The term "Path Modeling" refers to the objective of modeling a network of linear dependence relationships between variables, represented by a system of simultaneous equations.

Nowadays, PLS-PM is commonly used in several subjects where it is common to be associated with hypothetical constructs, defined as a conceptual term used to describe a phenomenon of theoretical interest (e.g., in Marketing studies, Economics, Social and Behavioural Science, Educational Research, Organizational Research, and so forth so on).

PLS-PM is a powerful method because of the minimal demands on measurement scales, sample size, and data distributions. It is particularly applicable for predictive applications and theory building, but it can be also used appropriately for theory confirmation (Chin, 1998; Falk and Miller, 1992).

Even though it is almost unanimously agreed that PLS-PM serves well for predictive purposes, predictive validity is not included as a standard assessment when evaluating path models. The inclusion of predictive validity as an essential part of model assessment in PLS-PM is very important, and further criteria and evaluation techniques should be also considered (Dolce et al., 2015; Sarstedt et al., 2014).

The main differences between Jöreskog's approach and the Wold's approach lie in the definition and the conceptual meaning of the unobserved variables included in the model (Marcoulides et al., 2009) and in the different objectives of the analysis, statistical assumptions, estimation procedures and related outputs. These differences have recently led to a thoughtful discussion (see Bentler and Huang, 2014; Dijkstra, 2014; Henseler et al., 2014; Marcoulides et al., 2009; Rigdon, 2012, 2014; Sarstedt et al., 2014, among others).

A deep study on the relationships between LVs and MVs is of chief importance because there is growing evidence that measurement model misspecification has the potential for poor parameter estimates and misleading conclusions (see Dolce and

Lauro, 2014; Jarvis et al., 2003; MacKenzie et al., 2005, among others). Its effects extend also to the estimates of the path coefficients connected to the misspecified block. We address this issue in the second chapter of the thesis.

Relationships between MVs and LVs can be modeled in two different ways. In the outwards directed scheme (Lohmöller, 1989) or reflective scheme (Fornell and Bookstein, 1982) MVs are considered as being caused by the related LV: variation in LV yields variation in MVs. On the contrary, in the inwards directed scheme (or formative scheme) MVs are viewed as causes of a LV: variation in MVs causes variation in LV.

A common impression found in the literature is that only PLS-PM allows the estimation of SEM including formative blocks. The implication of formative MVs in Covariance-Based framework is a rather difficult task. However, if certain model specification conditions are satisfied the model is identified, and it is possible to estimate a Covariance-Based SEM with formative blocks (Bollen and Davis, 2009; Williams et al., 2003).

Due to the complexity of both SEM estimation methods, we study their relative performance in the framework of the same simulation design, investigating the effects of measurement model misspecification and the implications of formative MVs on both ML-SEM and PLS-PM parameter estimates.

In the third chapter of the thesis we focus on the problem in PLS-PM about its incoherence with the direction of the relationships specified in the structural model. The directions of the links in the structural model do not play a role in the PLS-PM algorithm. In the search for optimally correlated constructs, the estimation process amplifies interdependence among blocks and misses to distinguish between dependent and explanatory blocks in the structural model. As a consequence, there is often a difference between what PLS-PM wants to model and what is actually computed by the PLS-PM algorithm.

We propose a new approach, called Non-Symmetrical Component-based Path Modeling (NSC-PM), based at maximizing the explained variance of MVs of the endogenous blocks by the components of the explanatory blocks (i.e. a new approach based on the optimization of a redundancy-related criterion in a multi-block framework).

The proposed method respects the direction of the relationships specified in the Path diagram (i.e. the path directions), since the directions of the links in the inner model play a role in the algorithm. In particular, bridge LVs (i.e., LVs that appear as both explanatory and dependent LVs in the structural model) are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when play a dependent role.

In order to assess the quality and validity of results, we provide a new goodness-of-fit index based on redundancy criterion and prediction capability together with a classical bootstrap-based inferential approach.

Finally, we show the functioning of the proposed algorithm (implemented in a R code) through a simulation study. The performance of the proposed method in terms of explained variability, predictiveness and interpretation is compared to the classical PLS-PM as well as to other component-based methods such as Regularized Generalized Canonical Correlation Analysis (Tenenhaus and Tenenhaus, 2011) and Generalized Structured Component Analysis (Hwang and Takane, 2004), using artificial data.

Compared to the other component-based methods, NSC-PM seems to be a good compromise between favoring stability (high explained variance) in the blocks and correlation between components.

In chapter four we focus on the particular case where there may be more than a single slope (i.e, the regression coefficient measuring the rate of change) describing the relationship between response variables and predictor variables. This especially occurs in the case of heteroscedastic variance, when dependent variable are highly skewed (as it is typical in subjective measurements), in the presence of outliers, or when the interactions between the factors affecting the dependent variables are very complex and cannot all be measured and accounted for in a model.

In several applications it can be interesting to investigate dependence relationships between variables considering all parts of the response variable distributions. For example, in the business and market research, it can be interesting to evaluate if and how much the impact of consumer preferences on satisfaction is different among highly, medium or low satisfied customers with the objective of differentiating leverages to increase the satisfaction.

A new method, called Quantile Composite-based Path Modelling (QC-PM), introduces both Quantile regression (QR) (Koenker and Basset, 1978) and Quantile correlation (QC) (Li et al., 2014) in the classical PLS-PM algorithm, in order to exploit their features and enhance PLS-PM potentialities when we wish to distinguish regressor effects on different parts of the dependent variable distributions. As a matter of fact, QC-PM accommodates heterogeneity and is able to explore the entire conditional distribution of the response variables. Instead of the only estimation of conditional means it allows the estimation of a set of conditional quantile functions, providing multiple slopes and a more complete picture of the relationships between variables.

QC-PM is advisable as a complementary analysis to the classical PLS-PM results, when heterogeneity in both the measurement and the structural model is expected, and in the case where there is no relationships (or only weak relationships) between LVs or between LVs with their own MVs, even if the underlying theory would suggest the opposite. The exploration of different parts of the dependent variable distributions could highlight significant relationships. It could also be expected that the sign and the size of path coefficients change if the analysis explores not only average effects but the entire conditional distribution of the response variables.

We go through the assessment and the validation of the proposed method extending the goodness of fit measures typically used in PLS-PM.

Finally, the functioning of the QC-PM is shown through a real data application in the area of the American Customer Satisfaction Index and through a Monte Carlo simulation study.

Chapter 1

Component-based Predictive Path Modeling: Recent Developments and Open Issues

1.1 Introduction

PLS-PM is a method aimed at modeling a network of linear dependence relationships between blocks of variables where each block is summarized by a linear composite of its own variables.

PLS-PM was originally presented as an alternative approach to the covariance-based SEM (Jöreskog and Wold, 1982b). However, the two approaches belong to two families of statistical methods.

The main difference between Jöreskog's approach and the Wold's approach lies in the definition and the conceptual meaning of the unobserved variables included in the model (Marcoulides et al., 2009).

The basic idea behind the Jöreskog's method is that the complexity inside a system can be studied taking into account a network of dependence relationships among unobserved variables, called latent variables (LV), each measured by several observed indicators usually defined as manifest variables (MV). Jöreskog's method is commonly referred to as a factor-based (or covariance-based) approach to SEM, as the LVs are defined as common factors, which aim to explain the covariances among their own set of MVs.

PLS-PM, instead, assumes that each block of MVs can be summarized by an unobserved variable defined as a component or a composite (i.e., an exact linear combination of the MVs). Since LVs are defined as components which aim to explain the variances of their own set of MVs, PLS-PM is commonly referred to as a component-based (or variance-based) approach (Lohmöller, 1989; Wold, 1975a,b, 1982).

This difference in the definition of the unobserved variables included in the model has led to a thoughtful discussion on the differences of the two approaches in terms of aims of the analysis (see Bentler and Huang, 2014; Dijkstra, 2014; Henseler et al., 2014; Marcoulides et al., 2009; Rigdon, 2012, 2014; Sarstedt et al., 2014, among others). Furthermore, several authors have compared the two approaches over the years (e.g., Fornell and Bookstein, 1982; Jöreskog and Wold, 1982a, among others).

On the whole, the two approaches differ in the objectives of the analysis, the statistical assumptions, the estimation procedures and the related outputs.

Covariance-based SEM is typically used for performing confirmatory analyses that aim to validate researchers hypotheses on the relations between LVs. If the theoretical model is correct and the standard assumptions underlying covariance-based SEM are satisfied, its estimators are unbiased. PLS-PM estimators lack the accuracy of covariance-based estimators. However, PLS-PM is a powerful method because of the minimal demands on measurement scales, sample size, and data distributions. It is particularly applicable for predictive applications and theory building, but it can be also used appropriately for theory confirmation (Chin, 1998; Falk and Miller, 1992).

Even though it is almost unanimously agreed that PLS-PM serves well for predictive purposes, predictive validity is not included as a standard assessment when evaluating path models. The inclusion of predictive validity as an essential part of model assessment in PLS-PM is very important, and further criteria and evaluation techniques should be also considered (Dolce et al., 2015; Sarstedt et al., 2014).

When predictive ability is interpreted as the ability to explain variance in the MVs of the endogenous blocks, we think that the new approach proposed by Dolce et al. (2015) could be of interest.

The remainder of this chapter proceeds as follows. Firstly, we discuss the distinctive differences and common features of component-based methods and factor-based methods for SEM. In the third section we present the PLS-PM in more details. We focus then on the predictive ability of PLS-PM and on the related evaluation criteria. In the last sections of this chapter we address some inconsistencies and critical issues in PLS-PM. To overcome some of these problems, we propose methodological contributions which are presented in details in the third and fourth chapter of this dissertation.

1.2 Is PLS-PM an alternative approach for "latent variable" modeling?

As underlined by Bollen (2002), the term "latent variable" has multiple meanings and it is commonly used in Statistics (see Muthén, 2003, among others) to refer to a large number of different concepts (i.e., "common factor", "conceptual variable", "construct", "random effects", "missing data", "latent classes", and so on). Moreover, frequently researchers use the term "latent variable" to refer to a "composite" or a "component".

In general, LV refers to a variable whose values can not be directly observed (Jöreskog and Sörbom, 1979). In this optic, any model dealing with unobserved variable could be classified as a "latent variable" model. In our opinion, since either Jöreskog's approach and Wold's approach aim to take into account a network of dependence relationships among unobserved variables, they can both be consider as "latent variable" models. However, a difference arise in the way Jöreskog and Wold deal with the unobserved variables in their respective approaches.

In PLS-PM, unobserved variables are essentially defined as linear composites or weighted sums of MVs (Fornell and Bookstein, 1982; Mathes, 1993; Nooan and Wold, 1982). The composite (or component) belongs to the space spanned by its own MVs (Esposito Vinzi and Russolillo, 2013), and as a consequence it is no longer a "latent variable". At best it can be consider as an approximation of the

LV with some given properties (see Section 1.4 for details). On the contrary, in Jöreskog approach the unobserved variables are included in the model as hidden factors (equivalent to common factors) defining the covariance structure among the MVs. In this perspective, the hidden factors are hypothetical existing entities defined as "latent variables" (Marcoulides et al., 2009).

Whether PLS-PM components can be consider a good approximations of factor-based method LVs or not mainly depends on the magnitude of the measurement error associated to each MV. As underlined by Marcoulides et al. (2009) the higher the measurement error associated to a block of MVs, the less the PLS-PM component will be able to approximate the true LV.

1.3 Is PLS-PM a method for Structural Equation Modeling?

In the literature, Jöreskog's approach is commonly referred to as a factor-based (or covariance-based) approach to SEM, as the LVs are defined as common factors, which aim to explain the covariances among their own set of MVs. PLS-PM, instead, is commonly referred to as a component-based (or composite-based or variance-based) approach, as LVs are defined as components which aim to explain the variances of their own set of MVs (Lohmöller, 1989; Wold, 1975a,b, 1982).

Jöreskog's approach was designed as a confirmatory method for validating researchers' hypotheses on the relations between observed and unobserved variables and among unobserved variables (theory building). Parameter estimates are chosen to minimize overall discrepancy between observed and model-implied covariance matrix. Component-based approaches focus on explaining MV variances and provide unobserved variable scores as a weighted aggregate of its own MVs (i.e., composites or components). PLS-PM is so far the most popular component-based approach for SEM (see Esposito Vinzi et al., 2010a; Tenenhaus et al., 2005, for an overview with recent developments).

PLS-PM currently enjoy widespread popularity in many disciplines while being harshly criticized in some academic literature (e.g., Rigdon, 2012; Rönkko and Evermann, 2013). Rönkko and Evermann (2013) argued that PLS-PM is not truly an SEM method, but it was the misinterpretation of the original articles on PLS-PM that led to incorrectly classify it as an SEM method. In the same direction, Rigdon (2012) suggested to sever every tie between PLS-PM and covariance-based SEM.

Researchers have been using few arguments supporting theses statements. Among them, Rigdon (2012) claimed that we should not consider PLS-PM as an SEM estimator because of the lack of unbiasedness and consistency of PLS-PM estimators and of an overidentification test. Indeed, PLS-PM produces inconsistent and biased estimates (Jöreskog and Wold, 1982b) especially when a small number of MVs is associated to each LV - i.e., "finite item bias" - (Lu, 2004; Lu et al., 2005). However, the bias decreases as the number of observations used and the number of MVs per block increase - i.e. consistence-at-large - (Dijkstra, 1983; Jöreskog and Wold, 1982b; Schneeweiss, 1993). Moreover, Dijkstra (2011) showed that PLS-PM algorithm yield all the ingredients for obtaining CAN (consistent and asymptotically normal) estimations of loadings and LVs squared correlations in a "clean second order factor model".

Furthermore, as Rigdon (2014) himself and Henseler et al. (2014) highlighted, PLS-PM estimates only appear to be biased when interpreted as effects between LVs instead of effects between composites. As the variances of measurement errors decrease in the population model the bias of PLS-PM estimates decreases. If the variances of measurement errors are equal to zero in the population model, then PLS-PM can yield asymptotically unbiased parameter estimates (Becker et al., 2013).

According to Henseler et al. (2014), SEM allows more general model then traditional common factor models. A more general model, called by Henseler et al. (2014) composite factor model, relaxes the strong assumption that the covariances among a set of MVs is explained by a common factor, thus no restriction is imposed on the covariances between MVs of the same block. Furthermore, each block of MVs is summarized by a composite (i.e., an exact linear combination of the MVs). As showed by Henseler et al. (2014), since common factor model has the same restrictions as the composite factor model plus some additional ones, common factor model is nested within the composite factor model.

Hence, if we define SEM as a multivariate analysis technique that allow us to analyze the relationships among theoretical concepts, each one measured by a set of observed variables, then PLS-PM can certainly be considered a *soft modeling* approach to estimate SEM parameters.

Another argument used for supporting the idea that PLS-PM should not be considered a method for SEM is the lack of a probabilistic framework for inference and a global criterion for assessing the fit of the model (Rigdon, 2012). However, in PLS-PM computational inference is commonly used for computing empirical confidence intervals and for hypothesis testing - e.g., blindfolding, permutation and resampling techniques - (Chin, 2010). Furthermore, an on going work at UCLA (Huang, 2013, PhD Dissertation with P. Bentler), proposing a modified PLS-PM suitable for confirmatory research via χ^2 goodness of fit tests and classical inference, seems to be quite promising in this direction.

In our opinion, we should keep looking upon PLS-PM as an alternative method for SEM as well as a descriptive and prediction oriented method. This double nature of PLS-PM seemed to be a natural thing to Wold since its origins. As notes by Dijkstra (2014) “If he had just wanted to extend principal components and canonical variables analysis, there would have been no need to develop the concept of consistency-at-large, that allows one to say when the difference between a factor model and one of PLS-PM modes would be small. In fact, he insisted that a fundamental principle of soft modeling is that all interaction between the blocks of observables is conveyed by the latent variables (Wold, 1981, 1982)”.

Instead of severing every tie between component-based methods and factor-based methods we think that researchers should commit themselves in finding out which approach works best in which circumstances. As Dijkstra (2014) suggests ”let us establish empirically where each works best. For problems in well-established fields highly structured approaches like mainstream SEM may be appropriate, other fields will be well served by highly efficient means of extracting information from high dimensional data”.

However, in any comparison study it is of extremely importance to bear in mind the distinctive statistical characteristics of the two approaches and the objectives of the analysis. For this reasons, the comparison between the two approaches should not ground only on parameter recovery.

If the theoretical model is correct and the standard assumptions underlying covariance-based SEM are satisfied, factor-based approach is expected to outperform PLS-PM in accurately estimating the parameters of the model. PLS-PM estimators lack the accuracy of covariance-based estimators but is a powerful method because of the minimal demands on measurement scales, sample size, and data distributions. It is particularly applicable for predictive applications and theory building, but it can be also used appropriately for theory confirmation (Chin, 1998; Falk and Miller, 1992).

Hence, it could be also interesting, in our opinion, to study into the details the predictive ability of PLS-PM and compare it to the predictive ability of covariance-based SEM.

In conclusion, we think that a continuous dialogue between the community of researchers who works on component-based methods to SEM and the one who works on factor-based methods to SEM it is highly recommended for progress in this area of research. The two approaches should be considered as complementary rather than competitive methods.

1.4 PLS path modeling

PLS Path Modeling aims at studying the relationships among K blocks, $\boldsymbol{X}_1, \ldots,$ $\boldsymbol{X}_k$, of MVs, which are expression of K LVs, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots, \boldsymbol{\xi}_K$, that are essentially defined as components or composites.

1.4.1 Model Specification

As in covariance-based SEM, the general model consist of two sub-models: the structural model and the measurement model. The measurement model relates each MV to its own LV. Each MV, $\boldsymbol{x}_{pk}$, is assumed to be generated as a linear function of its LV, $\boldsymbol{\xi}_k$, and its measurement error variable, $\boldsymbol{\epsilon}_{pk}$,

$$\boldsymbol{x}_{pk} = \lambda_{pk0} + \lambda_{pk}\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{pk} \tag{1.1}$$

where λ_{pk0} is a location parameter and λ_{pk} is the loading coefficient.

The structural model specifies the relationships between LVs. A LV is called endogenous if it is supposed to depend on other LVs in the model and exogenous otherwise. In the structural model a generic endogenous LV, $\boldsymbol{\xi}_m$ $(m = 1...M)$, is linked to corresponding latent predictors by the following multiple regression model:

$$\boldsymbol{\xi}_m = \beta_{m0} + \sum_{k \to m} \beta_{mk} \boldsymbol{\xi}_k + \boldsymbol{\zeta}_m \tag{1.2}$$

where β_{mk} is the so-called path coefficient capturing the effects of the predictor $\boldsymbol{\xi}_k$ on the dependent LV $\boldsymbol{\xi}_m$, and $\boldsymbol{\zeta}_m$ is the inner residual variable.

As a vehicle for the estimation of the model parameters, PLS-PM computes each unobserved variable as a perfect linear combination of its own MVs, i.e.:

$$\hat{\boldsymbol{\xi}}_k = \sum_{p=1}^{P_k} w_{pk} \boldsymbol{x}_{pk} \tag{1.3}$$

where $\boldsymbol{x}_{pk}$ $(p = 1, \ldots, P_k;\ k = 1, \ldots, K)$ is the generic centered and properly scaled MV of the k-th block, P_k is the number of MVs in the same block and w_{pk} is a weight coefficient.

1.4.2 The Algorithm

In PLS-PM the weight vectors, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, \ldots, \boldsymbol{w}_K$, to be associated to each block of MVs, are estimated by an iterative procedure by alternating inner and outer estimation steps.

In the outer estimation step each outer composite is obtained as a standardized weighted aggregate ($\boldsymbol{v}_k$) of its own MVs, i.e. $\boldsymbol{v}_k \propto \sum_h w_{jk} \boldsymbol{x}_{jk} = \boldsymbol{X}_k \boldsymbol{w}_k$ (outer estimation). Then, in the inner estimation step, each inner composite is obtained as a weighted aggregate ($\boldsymbol{z}_k$) of the connected composites, i.e. $\boldsymbol{z}_k \propto \sum_{\boldsymbol{v}_{k'} \to \boldsymbol{v}_k} e_{k'k} \boldsymbol{v}_{k'}$. Two composites are connected if there exists a link between the two blocks: an arrow goes from one LV to the other in the path diagram, independently of the direction (Tenenhaus et al., 2005).

For the computation of both the outer weights, w_{jk}, and the inner weights, $e_{k'k}$, several options are available.

In the outer estimation step, weights are computed by ordinary least-squares regressions. Generally, two different schemes are utilized. In the *Mode A* each MV are regressed on the corresponding instrumental inner composite $\boldsymbol{z}_k$. In the *Mode*

B the weights are computed as the regression coefficients in the multiple regression of the inner composite $\boldsymbol{z}_k$ on its own MVs $\boldsymbol{x}_{pk}$ $(p = 1, ..., P_k)$. Then, the weights are normalized such as $\text{var}(\boldsymbol{X}_k \boldsymbol{w}_k) = 1$.

However, it must be bear in mind that PLS-PM is a composite-based method, thus, whatever measurement scheme we apply, composites are computed as weighted aggregates of their MVs.

In the inner estimation step, there are three options for calculating the inner weights: centroid scheme, factorial scheme, path weighting scheme. In the centroid scheme, the inner weights $e_{kk'}$ are equal to the signs of the correlations between $\boldsymbol{v}_k$ and the $\boldsymbol{v}_{k'}$'s connected to $\boldsymbol{v}_k$. In the factorial scheme, the inner weights $e_{kk'}$ are equal to the correlations between $\boldsymbol{v}_k$ and the $\boldsymbol{v}_{k'}$'s connected to $\boldsymbol{v}_k$. In the path weighting scheme the LVs connected to $\boldsymbol{v}_k$ are divided into two groups: the antecedents of $\boldsymbol{v}_k$, which are LVs explaining $\boldsymbol{v}_k$, and the followers, which are LVs explained by $\boldsymbol{v}_k$, depending on the cause-effects relationships between the blocks of variables specified in the path diagram. If a $\boldsymbol{v}_{k'}$ is a follower of $\boldsymbol{v}_k$ then the inner weight, $e_{kk'}$ is equal to the correlation between $\boldsymbol{v}_k$ and $\boldsymbol{v}_{k'}$. On the other hand, for the antecedents $\boldsymbol{v}_{k'}$ of $\boldsymbol{v}_k$, the inner weights $e_{kk'}$ are the regression coefficient of $\boldsymbol{v}_{k'}$ in the multiple regression of $\boldsymbol{v}_k$ on the all $\boldsymbol{v}_{k'}$'s associated to the antecedents of $\boldsymbol{v}_k$.

These two steps are iterated until numerical convergence of outer weights, $\boldsymbol{w}_k$ $(k = 1, ..., K)$.

The convergence is proven in case of two blocks (Lyttkens et al., 1975), while empirical convergence is observed in most of the real applications with more than two blocks of variables. In 2010 Henseler showed a few examples of non-convergence of the PLS-PM algorithm (Henseler, 2010). However, according to Esposito Vinzi and Russolillo (2013), non convergence seems to be due to model misspecification rather than numerical pitfalls of the algorithm.

1.4.3 Optimizing Criteria

In PLS-PM there is not an overall scalar function optimized. This is mainly due to the different available options in the inner and outer estimation steps, but also to the fact that PLS Path models may differ in number of LVs and in the path of relationships linking them (Esposito Vinzi and Russolillo, 2013). Many researchers have focused on this issue in the last years. Nowadays, the stationary equations for most of the specific models obtained by running a PLS-PM are known and it is possible to show that the PLS-PM generalizes many Multivariate Analysis techniques.

Glang (1988) and Mathes (1993) were among the first who paid attention to the optimization criteria behind the PLS-PM. Let $c_{qq'}$ be the generic element of the Boolean square matrix C of order K, where $c_{kk'} = 1$ if $\boldsymbol{\xi}_k$ is connected to $\boldsymbol{\xi}'_k$ and $c_{kk'} = 0$ otherwise ($c_{kk} = 0$). The Authors showed that the Lagrange equations associated with the optimization of the criterion

$$\sum_{k \neq k'} c_{kk'} g(\text{cor}(\boldsymbol{X}_k \boldsymbol{w}_k, \boldsymbol{X}_{k'} \boldsymbol{w}_{k'})) \tag{1.4}$$

subject to $\|\boldsymbol{X}_k \boldsymbol{w}_k\| = \|\boldsymbol{X}_{k'} \boldsymbol{w}_{k'}\| = 1$, give exactly the stationary equation of PLS-PM algorithm when the weights in all the blocks in the outer estimation step are computed by means of multiple regressions of $\boldsymbol{z}_k$ over its MVs $\boldsymbol{X}_k$ (*Mode B*). $g(.)$ is the absolute value or the square function depending on the option used in the inner estimation step. More recently, Hanafi (2007) proved that the PLS-PM iterative procedure is monotonically convergent to these criteria.

In 2007 Krämer showed that Wold's PLS-PM algorithm with *Mode A* applied to all the blocks, does not lead to a stationary equation related to the optimization of a twice differentiable function (Krämer, 2007). In 2011 Tenenhaus and Tenenhaus have extended the results of Hanafi to a modified *Mode A* in which the outer weights, rather than the components, are normalized to unitary variance at each step of the algorithm (Tenenhaus and Tenenhaus, 2011). Contrary to to classical *Mode A*, this new estimation mode has the major advantage to maximize a known criterion. In particular, the Authors showed that Wold's procedure, applied to a PLS Path model where the *new Mode A* is used in all the blocks for the outer estimation, monotonically converges to the following criterion,

$$\underset{\|\boldsymbol{w}_k\| = \|\boldsymbol{w}_{k'}\| = 1}{\arg\max} \sum_{k \neq k'} c_{kk'} g(\text{cov}(\boldsymbol{X}_k \boldsymbol{w}_k, \boldsymbol{X}_{k'} \boldsymbol{w}_{k'})) \tag{1.5}$$

where $g(.)$ is exactly the same as in equation 1.4.

By comparing equations 1.4 and 1.5 it is easy to notice that the criteria associated to *Mode B* are based on maximizing correlations among adjacent composites, while the ones associated to *New Mode A* are based based on maximizing covariances among composites.

1.5 Prediction-oriented Component-based Methods

Composite-based approaches are necessary as there are many situations where researchers find that the assumptions of factor-models are not fulfilled. Moreover, frequently composite-based methods are preferred to factor-based methods since the objective of the research is to develop a predictive model (Shmueli and Koppius, 2011).

"Factor-based methods are fundamentally unsuitable for prediction, especially for prediction outside the dataset used to estimate the factor model, because of factor indeterminacy" (Rigdon, 2014). PLS-PM is an alternative to factor-based SEM in several applications, but it is also a descriptive/predictive approach and has strengths as a tool for prediction which have not been fully explored and appreciated. We go into further details on this topic in the this section.

Predictive models are developed in order to be able to predict values for individual cases. The aim in predictive analysis is not to test whether the relationships among variables are significant, but instead to accurately predict observations for specific cases that are similar to those in the sample.

PLS-PM is a powerful method for predictive purposes, and it is certainly an important technique deserving a prominent place in research applications when the aims of the analysis is prediction (Becker et al., 2013).

Reproducing model parameters is not the same thing as making valid predictions about individual observations. For these reasons, PLS-PM evaluation cannot focus only on parameter recovery and on the quality of the measurement model and the structural model - in terms of explained variance - indiscriminately.

The PLS-PM evaluation criteria should include the predictive ability and further criteria and evaluation techniques for PLS-PM are needed (Sarstedt et al., 2014). Thus, an interesting topic for further research in PLS-PM is the extension and development of further measures and evaluation criteria for the assessment of PLS path models in terms of predictive capability. Based on the proposed criteria, further extensions and modifications should be made on the basic PLS-PM algorithm in order to improve the predictive capabilities of the model estimation. The non-symmetrical approach for component-based path modelling proposed by Dolce et al. (2015) and presented in the third chapter of this dissertation is an example of work in this direction.

In our opinion, prediction in composite-based methods could refer to different concepts. Predictive ability could be interpreted as either the ability to explain variance in the endogenous LVs or the ability to predict individual observations. Moreover, individual observations may refer to either individual LV score observations or individual observations for MVs of the endogenous blocks. Finally, it should be made a distinction between in-of-sample and out-of-sample prediction.

1.5.1 In-of-Sample Prediction

As said above, the PLS-PM literature offers two main modes in the outer estimation step for computing the outer weights, which are known as *Mode A* and *Mode B*.

Mode B applies multiple linear regression of the inner composite on the corresponding MVs. Thus, it takes into account both the correlation between each MV and the corresponding LV and the intercorrelations among the MVs of the same block. On the contrary, *Mode A* ignores correlations among MVs.

Multiple linear regression adjust for Multicollinearity and gives less weights to more redundant predictors. When assumptions hold, OLS regression coefficients optimize R^2 for the data which are used to estimate the parameters of the model. Hence, we can expect that *Mode B* would perform better in term of in-of-sample prediction of LV.

Furthermore, within the literature on forecasting, it is well established that when the objective is to make as good a forecast as possible then combinations of forecasts can yield improvements in terms of prediction compared to single forecasts (Armstrong, 2001; Bates and Granger, 1969; Makridakis and Hibon, 2000), as each forecast nearly always contains some useful independent information. In this perspective, multiple indicator approaches should have an advantage in prediction over single indicator methods. Thus, *Mode B* in PLS-PM is certainly consistent with this "best practice" from the forecasting literature (Becker et al., 2013).

However, as said above, PLS-PM optimization criteria change depending on the way the outer weights are calculated. When all weights are computing using *Mode B*, PLS-PM maximizes correlation between composites, whereas applying *Mode A* to all blocks maximizes the composite covariances, thus, it takes into account the composite variances as well. *Mode B* in PLS-PM produces higher R^2 in the structural model, providing most accurate in-of-sample prediction for individual endogenous component observations. *Mode A* produces higher R^2 values in the regression of the MVs on their own LV, leading better in-of-sample individual observations prediction of MVs.

As noted by Rigdon (2012), "researchers applying PLS path modeling often assert the 'predictive' nature of their research, though researchers often seem to mean nothing more than aiming to maximize R^2 for dependent variables". However, when the goal of the analysis is prediction of individual score observations, the appropriate metric for assessing the predictive ability of the model is the R^2 in the structural model, but when prediction is to be made for individual observations of MVs of the endogenous blocks, redundancy-based prediction is preferred. Moreover, in either cases above, the metric used for assessing the model regards the in-sample predictive ability.

When prediction is to be made for individual observations of MVs of the endogenous blocks, the fit of the global model can be judged as satisfactory if the average of the redundancy indexes for each block is high enough. The new approach proposed by Dolce et al. (2015) and presented in the third chapter of this thesis is very promising in this case.

1.5.2 Out-of-Sample Prediction

Different outer modes within PLS-PM methodology certainly lead to different out-of-sample predictive capabilities of models as well.

Dana and Dawes (2004) demonstrated, in the context of conventional regression, that correlation weights (which ignore collinearity among the predictors) outperform multiple regression weights for out-of-sample prediction unless sample size is very large. For this reasons, the authors urged researchers to avoid using multiple regression weights for out-of-sample prediction. As noted by Becker et al. (2013) and Rigdon (2012), Dana and Dawes's suggestions would translate into an advantage for *Mode A* estimation of outer weights (which corresponds to the use of correlation weights) over *Mode B* (which corresponds to the use of multiple regression weights). However, further studies are necessary in order to examine this issue into further details.

Dana and Dawes (2004) have also demonstrated that out-of-sample predictive ability depends on sample size. Becker et al. (2013) considered the sample size as an experimental condition in simulation studies aimed at analyzing the out-of-sample prediction capability of PLS-PM. The results of their study showed that if the criterion is out-of-sample predictive ability, PLS-PM perform poorly when sample size is small. Sample sizes that would be adequate for the estimation of the parameters of the model may be highly inadequate for out-of-sample prediction.

Predictive capability of component-based method can be also improved extracting more than one component for each block. PLS-PM generally consider one component for each block of variables. In some case we can lose information in predictor blocks that may be of extremely importance for the predicting endogenous composites or the MVs related to them. Some proposals in this directions has already been introduced. Among the others, we think that future research may focus on the study and improvement of the extended method for PLS-PM proposed by Lohmöller (1989), that allows for more complex methods and, in particular, several components for each block can be simultaneously extracted.

As a measure of out-of-sample predictive relevance, the predictive sample reuse technique as developed by Stone (1974) and Geisser (1975) - the Stone-Geisser's Q^2 - is more appropriate.

The PLS-PM adaptation of this approach follows a blindfolding procedure that proceeds as following. Given a block of n cases and P MVs, e.g. the MVs of the endogenous blocks, the procedures takes out a portion of the considered block during parameter estimations and then attempts to estimate the omitted part using the estimated parameters. To estimate the model, the omitted values are typically replaced with the variable mean, though other imputation techniques may be used (Chin, 1998). Based on the estimated model, the estimates for the omitted values are compared to the observed values, using the squared difference

(E). At the same time, the difference between the variable mean (or otherwise imputed values) and the observed values are also compared using the squared difference (O). This procedure is repeated until every data point has been omitted and estimated. The predictive measure for these MVs is the calculated as:

$$Q^2 = 1 - \frac{\sum_m E_m}{\sum_m O_m} \tag{1.6}$$

where m is the number of times the procedure is repeated to assure that every data point are omitted.

Q^2 represents a measure of how well-observed values are reconstructed by the model and its parameter estimates (Chin, 2010). $Q^2 > 0$ implies the model has predictive relevance whereas $Q^2 < 0$ represents a lack of predictive relevance.

Blindfolding can be done on any set of variables. However, the predictive ability of the model typically concerns the MVs for the endogenous blocks.

Different forms of Q^2 can be obtained based on different procedures for predicting observations from the model. In cross-validated communality Q^2 prediction of observations are made by the computed composite and the estimated loadings. Cross-validated redundancy Q^2 is still based on the estimated loadings but the composite are predicted from the structural model using the estimated path coefficients. Redundancy-based Q^2 is applicable only to observations of MVs of the endogenous blocks, while communality-based Q^2 can be applied to all MVs (Chin, 2010; Evermann and Tate, 2012).

Even though Herman Wold recognized that Stone-Geisser's procedure fits PLS-PM approach "like hand in glove" (Wold, 1982, p. 30), this criterion is seldom reported in PLS-PM studies. Generally, despite the predictive aim of many PLS-PM studies, most of them do not provide appropriate predictive ability metrics (Becker et al., 2013; Hair et al., 2012a,b; Ringle et al., 2012; Sarstedt et al., 2014), and this is surprising considering that PLS-PM is said to be a powerful method for predictive purposes deserving a prominent place in research applications when the aims of the analysis is prediction.

1.6 Misspecification of the Measurement Model

A block of variables is conceptually defined as outwards directed (Lohmöller, 1989) or reflective (Fornell and Bookstein, 1982) if the MVs are considered as being caused by the corresponding LV: variation in LV yields variation in MVs. In this case, MVs should be highly correlated, as they are caused by the same common factor. In other words, the block is expected to be unidimensional and internally consistent (Tenenhaus et al., 2005). The PLS-PM literature has long suggested that the MVs weights in block defined as outwards directed (of reflective) are to be estimated using *Mode A* - i.e., each MV is regressed on the corresponding instrumental composite in the outer estimation step (e.g., Chin, 1998; Esposito Vinzi and Russolillo, 2013; Fornell and Bookstein, 1982; Hair et al., 2011; Henseler et al., 2009).

When each MV is viewed as cause of a LV (i.e., variation in MV causes variation in LV), the block can be conceptually defined as inwards directed or formative. MVs in formative blocks can represent different and weakly correlated ingredients of the underlying concept. In such a case, literature suggests that MVs weights are to be estimated using *Mode B* (i.e., weights are computed as the regression coefficients in the multiple regression of the instrumental composite on its own MVs).

The differences between the two measurement models are not trivial. Since there is no reason to expect high correlation among MVs of a formative block (Tenenhaus et al., 2005), conventional measures used for evaluating the validity and reliability of a LV cannot be applied for formatively-measured LVs (Bollen and Lennox, 1991; Diamantopoulos, 2006). Confirmatory tetrad analysis (CTA) (Bollen and Ting, 2000) is an example of an alternative way for testing construct validity. MVs of a reflective block are interchangeable: dropping an indicator from the measurement model should not alter the meaning of the LV (Bollen and Lennox, 1991). This is not required when considering MVs of formative blocks. As for the nature of the error term in formative blocks, several definitions are found in the literature (Bollen and Lennox, 1991; Diamantopoulos, 2006; Edwards and Bagozzi, 2000). The error term in formative blocks is not a measurement error and it is more properly called as “disturbance”.

In some particular situations determining the real nature of a LV is a difficult task (Edwards and Bagozzi, 2000). Morover, most researchers consider MVs as effects of a LV (reflective scheme) without even questioning their appropriateness for the specific LV at hand.

Measurement model misspecification is fairly common among published research studies, and it is proven that it holds the potential for poor parameter estimates and misleading conclusions (see Dolce and Lauro, 2014; Jarvis et al., 2003; MacKenzie et al., 2005, among others).

Its effects extend also to the estimates of the path coefficients connected to the misspecified block. In covariance-based SEM this is mainly due to the fact that a reflective treatment of a block that should instead be modeled as formative reduces the variance of the LV. The variance of a reflectively-measured LV equals the common variance of its MVs, whereas the variance of a formatively-measured LV encompasses the total variance of its indicators (Fornell et al., 1991). Let us consider the common case of an exogenous formative block misspecified as reflective. If the level of the variance of the endogenous LVs is maintained, the estimates of the path coefficients connected to the misspecified exogenous LV is likely to be substantially inflated (Diamantopoulos et al., 2008).

1.7 The Path Direction Incoherence in PLS Path Modeling

As said above, in the inner estimation step of the PLS-PM algorithm, there are three main options to calculate the inner weights: Centroid scheme, Factorial scheme, Path weighting scheme. The path weighting scheme is said to have the advantage of taking into account both the strength and the direction of the paths in the structural model. However, the path direction is taken into account only in the way the inner weights are computed, but each LV is still defined in the inner step of the algorithm as a function of all the connected LVs. The way the inner weights are calculated leads to some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram, and it does this for all inner schemes. The PLS-PM estimation process amplifies interdependence among blocks, and as a consequence it misses to distinguish between dependent and explanatory blocks.

As for the outer model, generally MV weights are computed using either *Mode B* or *Mode A*. However, beyond the theoretical differences between the two different measurement model schemes (outward directed or inward directed), depending on the way the outer weights are calculated in the measurement model (*Mode A* or *Mode B*), the role of the LVs in the structural model changes. Thus, when choosing which mode to use for computing the outer weights it should be take into account the role of the LVs in the structural model as well. The predictive direction in the structural model is given by the utilized outer scheme, while the directions of the links in the structural model do not play a role in the PLS-PM algorithm.

The only way for giving an explanatory role to a LV is to apply *Mode B*, while applying *Mode A* gives it a role of dependent variable, whatever the path direction is (Dolce et al., 2015). However, in the case of more then two blocks of variables, where some endogenous LVs may appears as both explanatory and dependent LVs, this choice can be a much more complicated matter (Dolce et al., 2015).

As a matter of fact, under conditions of low theoretical knowledge on the conceptual definition of the LVs, a rule of thumb in PLS-PM is to apply *Mode B* to the exogenous block and *Mode A* to the endogenous block (Wold, 1980). However, to the best of our knowledge, there are hardly any studies in the literature that give reasons for following this rule and analyze into details this issue.

PLS-PM does not rigidly adhere to an underlying theoretical model (Chin, 1998), and there is often a difference between what PLS-PM wants to model (the hypothesized model depicted in the path diagram) and what is actually computed by the PLS-PM algorithm. Furthermore, some underlying theoretical models depicted in path diagrams have nonsense in the strict framework of structural equation modeling. Dolce and Hanafi (2015) illustrates this issue by using a simple model, the case of two blocks of variables. The authors shows that Wold (1980) suggestion about using *Mode B* to the exogenous block and *Mode A* to the endogenous block is not just a rule of thumb. Instead, applying *Mode B* for the endogenous block does not make sense in the framework of SEM.

Dolce et al. (2015) propose a new algorithm that takes into account the directions of the links in the structural model and aims at maximizing the explained variance of the MVs of the endogenous blocks, i.e. a new approach based on the optimization of a redundancy-related criterion in a multi-block framework, in order to inherit its prediction oriented objective as well as its non-symmetrical approach that takes the direction of relationships explicitly into account. We present this new method in the third chapter of this dissertation.

1.8 A more comprehensive analysis of the Relationships in Component-based Path Modeling

In some particular case, PLS-PM may give an incomplete picture of the relationships between variables, since the estimates coefficients may not be the same along all parts of the dependent variable distributions.

PLS-PM algorithm is a procedure based on simple and multiple ordinary least squares (OLS) regressions, thus the obtained coefficients measure the rates of change in the mean of the dependent variables (both manifest and latent variables) distributions as a function of changes in the set of predictors. Focusing exclusively on changes in the means may underestimate, overestimate, or fail to distinguish real non-zero coefficient.

This issue may especially occur in the case of heteroscedastic variances, when dependent variables are highly skewed (as it is typical in subjective measurements), in the presence of outliers, or when the interactions between the factors affecting the dependent variables are very complex and cannot all be measured and accounted for in a model.

In these case, there may be more than a single slope (i.e, the regression coefficient measuring the rate of change) describing the relationship between response variables and predictor variables.

For example, when all the factors that may affect an endogenous LV are not included in the models used to investigate relationships between LVs, there may be a weak or no dependence relationship between the mean of the endogenous LV distribution and the corresponding predictive LVs. However, there may be a stronger and useful dependence relationship with other parts of the endogenous LV distribution. The same may happen in the dependence relationships between LV and their own MVs.

In several applications it can be interesting to investigate if the relationships between dependent variables and regressors changes across the different parts of the response variable distributions. For example, in the business and market research, it can be interesting to evaluate if and how much the impact of consumer preferences on satisfaction is different among highly, medium or low satisfied customers with the objective of differentiating leverages to increase the satisfaction.

Quantile regression (QR) (Koenker and Basset, 1978) is an extension of the classical OLS regression for estimating functional relations between variables for all parts of the distribution of the response variable. Instead of the only estimation of the conditional mean it allows the estimation of a set of conditional quantile functions, providing a more complete picture of the relationships between variables. Compared to the OLS regression QR estimates are more robust against outliers.

In this perspective, Li et al. (2014) introduced a correlation measure to examine the linear linear relationships between any two variables for a given quantile, named quantile correlation (QC).

A new method, called Quantile Composite-based Path Modelling (QC-PM) and presented in details in the fourth chapter of this dissertation, introduced both QR and QC in the classical PLS-PM algorithm (Davino and Esposito Vinzi, 2015; Davino et al., 2015a), in order to enhance PLS-PM potentialities when we wish to distinguish regressor effects on the different parts of the dependent variable distribution.

QC-PM accommodates heteroscedastic variances and outliers and is able to explore the entire conditional distribution of the response variables. It is advisable as a complementary analysis to the classical PLS-PM. For example, when path coefficient estimates of classical PLS-PM are not significant, even if the underlying theory would suggest the opposite, the exploration of different parts of the dependent variable distribution could highlight significant relationships. It could also be expected that the sign and the size of the path coefficients change if the analysis explores not only average effects but also the different parts of the dependent variable distributions.

1.9 Other critical issues in PLS-PM

PLS-PM is said to be a powerful method because of the minimal demands on measurement scales, sample size, and data distributions (Chin, 1998; Falk and Miller, 1992).

The topic of minimal requirements on sample size in PLS-PM has been widely debated in recent years (e.g., Hair et al., 2012a; Henseler et al., 2014; Marcoulides and Saunders, 2006; Rönkko and Evermann, 2013) and has been empirically studied in various simulation studies (e.g., Areskoug, 1982; Goodhue et al., 2012; Hulland et al., 2010; Vilares and Coelho, 2013).

In the literature it there seems to be a common belief that sample size issue does not play a role in the application of PLS-PM (Henseler et al., 2014). Many authors follow the "ten times" rule of thumb (Barclay et al., 1995) according to which the sample size should be equal to the larger number of explanatory variables in each particular measurement model and structural model. Namely, the minimum sample size should be equal to the larger number of (1) ten times the largest number of the MVs whose weights are estimated by the inward directed scheme at a particular block, or (2) ten times the largest number of structural paths directed at a particular LV in the structural model.

However, the ten-times rule of thumb does not take into account the magnitude of the relationships, the reliability, the number of indicators, distributional characteristics of the data, or other factors which are known to affect the statistical power. It is only in the case of a strong effect size (and high reliability) that "ten times" rule of thumb may lead to acceptable power (Goodhue et al., 2006), thus, it cannot be applied indiscriminately to all situations (Henseler et al., 2009; Marcoulides and Saunders, 2006).

The distributional characteristics of the data, potential missing data and the properties of the variables examined are also to be considered when deciding on an appropriate sample size to use. Marcoulides and Saunders (2006) noted that "when moderately non-normal data are considered, a markedly large sample size is needed despite the inclusion of highly reliable indicators in the model".

Another issue to take into account in PLS-PM concerns the problem of multicollinearity. Since *Mode B* is based on multiple regression, the stability of the MV outer weights, which reflect the impact of the MVs on the LV, are affected by the strength of the MV intercorrelations as well as the sample size. Therefore, the issue of multicollinearity is particularly important in formative blocks (Albers and Hildebrandt, 2006; Diamantopoulos and Winklhofer, 2001)

On the contrary, under *Mode A*, multicollinearity is not an issue because only simple regressions are involved, and theoretically it is desired.

In the case of a perfect collinearity between two formative MVs (i.e., one MV is a linear combination of another MV), PLS-PM cannot estimates the parameters of the model since the covariance matrix of the formative MVs is singular and cannot be inverted, as requested in the multiple regression when using *Mode B*. In this particular case, we can just drop one of the redundant MV.

Excessive multicollinearity among formative MVs (i.e., any single MV is highly correlated with the others), instead, makes it difficult to separate the distinct influence of the individual MV on the LV. Multicollinarity can inflate bootstrap standard errors leading to type II errors (i.e., it may yields non-significant outer weights when actually the MVs have an effect on the corresponding LV), or else the outer weights may be non-interpretable, having incoherent signs with the correlation with the corresponding LV.

The issue of multicollinearity in formative blocks is still under research and some of solutions found in the literature are not satisfactory. Furthermore, in the literature most of the studies where formative blocks are included in the models do not consider the multicollinearity assessment (Hair et al., 2012a).

A suggestion found in the literature is to simply interpret only the standardized loadings (i.e. correlations between a LV and its own MVs) instead of the outer weights (Cenfetelli and Bassellier, 2009; Hair et al., 2012a), bearing in mind that while the outer weight is a measure the relative contribution of a MV to its LV, the loading can only be used to evaluate the absolute importance of a MV to its LV. However, as Chin (1998) noticed, it makes no sense to compare formative MV loadings with one another as the intraset correlations for each block were never taken into account in the estimation process.

The problem of multicollinearity can be addressed by providing a PLS regression for estimating the outer weights as an alternative to OLS regression (Esposito Vinzi et al., 2010b). This new approach, called *Mode PLS*, can be considered as a fine-tuning between *Mode A* and *Mode B* since it is based on the selection of a certain number of components of the PLS regression[1]. *Mode PLS* adapts well also to formative multidimensional blocks with fewer dimensions than the number of MVs. This new mode is available in the PLSPM module of the XLSTAT software.

While the elimination of a MV with small weight within formative measurement models should be always approached with caution, since this may implies the omission of a substantial and meaningful part of the construct (see Chapter 2), dropping a MV from a formative measurement model in the case of excessive multicollinearity in the block might be recommended.

A possible way to check for multicollinearity in a formative block is computing the "tolerance" of each MV as $1 - R^2$, where the R^2 is the coefficient of determination for the regression of the the specific MV on the other MVs of the block. Obviously, as the tolerance value increases the degree of multicollinearity increases. A measure related to the tolerance is the Variance Inflation Factor (VIF), computed as the inverse of the tolerance ($VIF = 1/TOL$) (Hair et al., 2010). A large VIF value indicates a high standard error of the specific weight due to multicollinearity among the MVs.

As a rule of thumb, the VIF should not exceed a value of 10, but, particularly when samples size is small, the critical value may be smaller then 10 (Hair et al., 2010). In general, the critical value should be defined considering the specific analysis objectives [2].

[1] *Mode A* correspond to taking the first component from a PLS regression, while *Mode B* correspond to taking all the PLS regression components

[2] For some suggested guidelines to follow, see Hair et al. (2010)

Chapter 2

Formative Versus Reflective Measurement Model in Structural Equation Modeling

2.1 Introduction

Research often places great emphasis on explaining causal relationships among LVs but devote little attention to the nature and direction of relationships between LVs and MVs.

Even though there are situations in which MVs are more realistically thought of as causes of a LV (formative scheme), most researchers consider them as effects (reflective scheme) without even questioning their appropriateness for the specific LV at hand.

Furthermore, a common impression found in the literature is that only PLS-PM allows the estimation of SEM including formative blocks. The implication of formative MVs in Covariance-Based framework is a rather difficult task. However, if certain model specification conditions are satisfied the model is identified, and it is possible to estimate a Covariance-Based SEM with formative blocks (Bollen and Davis, 2009; Williams et al., 2003).

A deep study on the relationships between LVs and MVs is of chief importance because there is growing evidence that measurement model misspecification has the potential for poor parameter estimates and misleading conclusions (see Dolce and Lauro, 2014; Jarvis et al., 2003; MacKenzie et al., 2005, among others). Its effects extend also to the estimates of the path coefficients connected to the misspecified block.

Due to the complexity of both SEM estimation methods, we study their relative performance in the framework of the same simulation design, investigating the effects of measurement model misspecification and the implications of formative MVs on both ML-SEM and PLS-PM parameter estimates.

The results presented in this section are based on the paper by Dolce and Lauro (2014).

The simulation results show that the effect of measurement model misspecification is much larger on the ML-SEM parameter estimates. For a model that includes a correctly specified formative block, we find that the inter-correlation level among formative MVs and the magnitude of the variance of the disturbance in the formative block have evident effects on the bias and the variability of the estimates. For high inter-correlation levels among formative MVs, PLS-PM outperforms ML-SEM, regardless of the magnitude of the disturbance variance. For a low inter-correlation level among formative MVs the performance of the two methods depends also on the magnitude of the disturbance variance. For a small disturbance variance, PLS-PM performs slightly better compared to ML-SEM. On the contrary, as the disturbance variance increases ML-SEM outperforms PLS-PM.

2.2 Nature and Direction of the relationships between latent variables and manifest variables

LVs, while not directly observed, are measured by a set of MVs. This observable variables may appear as effects of the LVs, or cause of the LVs, or as both effects and cause. Hence, relationships between MVs and LVs can be modeled in two different ways, depending on the direction of the relationship between the LV and its own MVs.

In the outwards directed scheme (Lohmöller, 1989) or reflective scheme (Fornell and Bookstein, 1982) MVs are considered as being caused by the related LV: variation in LV leads to variation in its MVs (Bollen, 1989). On the contrary, in the inwards directed scheme or formative scheme MVs are viewed as causes of a LV (Blalock, 1971): variation in MVs are assumed to causes variation in LV (see Figure 2.1).

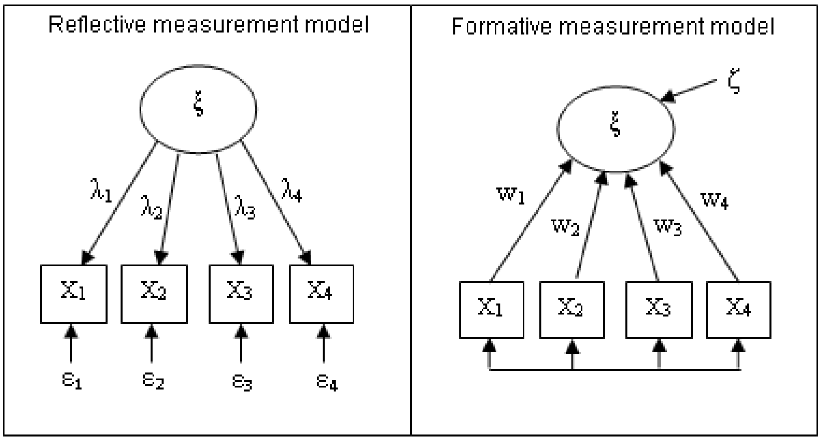


FIGURE 2.1: Different measurement models

The theoretical differences between the two schemes are not trivial and in some particular situations determining the real nature of a LV is a difficult task (Edwards and Bagozzi, 2000).

2.2.1 Internal consistency reliability and the issue of Multicollinearity

In the reflective measurement model MVs are caused by the same common factor, thus variance in each measure is explained by a LV common to all measures and error unique to each measure, and covariance among MVs is attributed to their common cause, the underlying LV. In this respect, blocks of variables thought as outwards directed (of reflective) are expected to be unidimensional and should possess internal consistency - i.e., MVs in each block are supposed to be highly correlated among each other - (Tenenhaus et al., 2005). Many reliability estimates are based on this internal consistency concept (Bollen and Lennox, 1991), and there exist several tools to check the internal consistency (i.e., the unidimensionality) of a block (Tenenhaus et al., 2005).

On the contrary, when blocks are defined as inwards directed (or formative), MVs can represent different ingredients of the underlying concept. In the formative model, the block of MVs can be multidimensional, each MV or each sub-block of

MVs could represent different dimensions of the underlying concept, so these MVs need not to covary.

An important issue for formative blocks is that of multicollinearity (Albers and Hildebrandt, 2006; Diamantopoulos and Winklhofer, 2001). This is because the formative measurement model is based on multiple regression, and therefore the stability of the MV outer weights, which reflect the impact of the MVs on the LV, are affected by the sample size and strength of the MV intercorrelations. Excessive collinearity among MVs thus makes it difficult to separate the distinct influence of the individual MV on the LV. Note that under reflective measurement, multicollinearity is not an issue because only simple regressions are involved (in which the MV serves as the criterion and the LV as the predictor).

In order to overcome the multicollinearity problem in formative measurement model, an alternative approach recently proposed by Esposito Vinzi et al. (2010b) can be used as well. This new approach, called Mode PLS, computes the outer weights applying the PLS regression (Tenenhaus, 1998; Wold et al., 1983). The Mode PLS can be considered as a fine-tuning between Mode A and B since it is based on the selection of a certain number of components of the PLS regression [1].

2.2.2 Validity of indicators

Since in formative scheme the magnitude of the MV correlations is not explained by the model, we cannot say much about the validity of the MVs as a measure of the corresponding LV.

As said above, there is no reason to expect high correlation among MVs of a formative block (Bollen, 1984; Tenenhaus et al., 2005), thus conventional measures used for evaluating the validity and reliability of a LV cannot be applied for formatively-measured LVs (Bollen and Lennox, 1991; Diamantopoulos, 2006). Indeed, as noted by Bollen and Lennox (1991, p. 312), “causal indicators are not invalidated by low internal consistency so to assess validity we need to examine other variables that are effects of the latent construct.” However, there is not recommendations about magnitude of correlations for MVs of formative blocks, because these correlations are explained by factors outside of the model.

[1] Mode A correspond to taking the first component from a PLS regression, while Mode B correspond to taking all the PLS regression components

The literature is unclear as to measure the validity of MVs in formative blocks (Edwards and Bagozzi, 2000), and the assessment of formatively-measured LVs is still an open question and under research.

Given that each MV weight shows the direct relation between the MV and its LV and the impact of the MV on the LV, the magnitudes of the MV weight can be interpreted as validity coefficients (Bollen, 1989). MV with non significant weight could be considered for elimination as they cannot represent valid indicators of the construct. However, removing a MV in formative blocks may implies removing a theoretically meaningful part of the LV and should always be approached with caution. Furthermore, it must be noticed that high multicollinearity among MVs could lead to difficulties in assessing indicator validity on the basis of the magnitude of the MV coefficients (Bollen, 1984; Diamantopoulos and Winklhofer, 2001; MacKenzie et al., 2005).

As for assessing validity at the overall construct level, one common approach is focusing on nomological and criterion-related validity: estimating hypothesized relationships of the LV with theoretically related LVs, checking if the estimated relationships is consistent with the expected direction and significantly different from zero. Diamantopoulos and Winklhofer (2001) stated that "validation along these lines requires (1) that information is gathered for at least one more construct than the one captured by the index, (2) that this other construct is measured by means of reflective indicators, and (3) that a theoretical relationship can be postulated to exist between the constructs".

On this perspective, a way for evaluating formative measurement models could be by testing whether the formatively-measured LV is highly correlated with a reflective measure of the same theoretical concept (Hair et al., 2014). This can be achieved applying a redundancy analysis use the formatively-measured LV as an exogenous LV predicting an endogenous LV measured by one or more reflective MVs, but theoretically both sets of MVs should be tied to the exact same LV. The strength of the path coefficient linking the two LVs is indicative of the validity of the designated set of formative indicators in tapping the LV of interest. Ideally, a magnitude of 0.90 or at least 0.80 and above is desired (Chin, 1998) for the path coefficients between the two LVs.

Diamantopoulos (2006) proposed using the variance of the error term as an indication of construct validity. The error term represents that part of the construct's domain that the set of MVs neglect. Hence, if the set of MVs include all important construct facets, the residual variance should be small, and the construct meaning is validly captured.

Finally, confirmatory tetrad analysis (CTA), (Bollen and Ting, 2000; Gudergan et al., 2008) offered a basic test of LV validity. Although Bollen and Ting (2000, p. 4) originally proposed CTA as "an empirical test of whether a causal or effect indicator specification is appropriate", interpreting evidence supporting the latter as also supporting the LV's validity is reasonable.

2.2.3 Interchangeability of the manifest variables

Another important issue related the measurement model is evaluate the consequence of removing MVs of a unidimensional block. As already said above, MVs should cover all facets of the LV, they need to capture the domain space of the it (Little et al., 1999). If each of our original MV is "representative" of distinct facets of a LV, removing a MV implies removing a theoretically meaningful part of the LV and changing the meaning of the LV (Bollen and Lennox, 1991). Failure to consider all facets of the LV will lead to an exclusion of relevant MVs (Diamantopoulos and Winklhofer, 2001). Furthermore, since the formative measurement model assumes that all the measures have an impact on a single LV, the MVs may be correlated, but the model does not assume or require this. Indeed, it would be entirely consistent for MVs in formative blocks to be completely uncorrelated (Jarvis et al., 2003). This might be the case where a formatively-measured LV is represented by mutually exclusive types of behaviour. What is important to understand is that even if correlated, in formative blocks MVs are not interchangeable.

On the contrary, Reflective MVs are interchangeable: dropping a MV from the measurement model should not alter the meaning of the LV (Bollen and Lennox, 1991). Because all the MVs are assumed to be equally valid indicators of the underlying LV, any two equally reliable effect indicators of an unidimensional construct are interchangeable. Thus, when a MV is dropped the construct validity should be unchanged (Bollen and Lennox, 1991), even if the reliability estimates of the set of MVs will be lower if fewer variables are included in the measurement model.

In summary, for reflective unidimensional block, equally reliable indicators are essentially interchangeable. If many facets of a LV mean many dimensions, then each dimension should be treated separately with its own set of MVs. For formatively-measured LV, excluding a MV may alter the meaning of the LV.

2.2.4 The error term in formative measurement model

As for the nature of the error term in formative blocks, several definitions are found in the literature. In a papers by Bollen and Lennox (1991) and MacCallum and Browne (1993), for example, the error is simply referred to as a "disturbance" with no further elaboration on its nature. Edwards and Bagozzi (2000) stated that "the disturbance term represents that part of the construct [...] that is not explained by the [...] measures and thus may be interpreted as measurement error". (Jarvis et al., 2003), stressed that in formative measurement models "error is represented at the construct level rather than at the individual item level [...] one obtains an estimate of the overall amount of random error in the set of items rather than an estimate attributable to each individual item"

Diamantopoulos (2006) criticized all statements concerning the nature of error in formative measurement, claiming that none of them is completely true: "the type of error involved is not random measurement error; the reliability of the scale cannot be improved by estimating the error term; and the error is not associated either with individual items or the set of items as a whole. In fact, the error term in a formative measurement model tells us hardly anything about the items already used as indicators in the model". Diamantopoulos (2006) showed that, unlike for reflective blocks, the error term in formative blocks is not a measurement error but it is more properly called as "disturbance" term which impacts on the LV and it is uncorrelated with the MVs of the block, "violation of this assumption would result in biased estimates in the [path coefficients] (much in the same way that omission of relevant independent variables which are related to included predictors, would bias the estimates in a multiple regression model)". Thus, the omitted MVs should not be correlated with those included in the formative blocks. He also showed that a correct interpretation of the disturbance may be quite informative regarding MVs not incorporated in the model, thus, can aid in model specification and re-specification.

The magnitude of the error term can be useful in the interpretation of a formative measure model. Setting the disturbance equal to zero means that all possible causes on the LV are accounted for by the MVs in the model. However, since it is very difficult that this occurs in practice, it is a good practice to incorporate the error term in the formative measurement model specification.

2.2.5 Surplus meaning of the latent variable

Another issue to be taken into account is the surplus meaning of LVs. LVs hold surplus meaning beyond that captured by their own MVs used to measure them in both formative and reflective measurement models(Jarvis et al., 2003).

Given that variation in reflective measurement models precedes variation in their own MVs, LVs have surplus meaning because they are assumed to exist independently by of measurement.

On the other hand, since variation in MVs are assumed to causes variation in formatively-measured LV, the latter are inextricably tied to their MVs, thus the nature of their surplus meaning is very different from the reflectively-measured LVs.

Diamantopoulos (2006) claim that "the surplus meaning of formative constructs is directly associated with the error term included in the formative model specification [...] thus, the surplus meaning possessed by a formative construct relates to the influence of unmeasured causes, i.e. indicators not included in the model".

2.2.6 Criteria for Distinguishing Between Reflective and Formative measurement Models

Even though the theoretical differences between the two measurement models are well defined, in some particular situations determining the real nature of a LVs can be a difficult task.

Edwards and Bagozzi (2000) suggested several criteria derived from the literature on causation that might be employed in this regard, including association, temporal precedence, and the elimination of rival causal explanations.

Another way for distinguishing between formative and reflective measurement Models is to perform ”mental experiments,” in which a change in the LV is imagined and then it must be judge whether a subsequent change in all the MVs is reasonable. If so, then this is consistent with a reflective measurement model. On the other hand, if for a change in all the MVs we expect a change in the LV, then this is consistent with a formative measurement model (Bollen, 1989).

Bollen and Ting (2000) suggested that a simple examination of a set of MVs along with a “mental experiment” may be insufficient to make a clear distinction between the two different measurement meodels. For this reason the Authors developed an empirical tool for determining whether the covariance structure among a set of MVs is more consistent with a formative or reflective measurement model based on Vanishing Tetrad Analysis (see Bollen and Ting, 2000; Gudergan et al., 2008, to go into further details).

2.2.7 Misspecification of relationships between latent variables and manifest variables

Conventional measurement model in marketing and business research, psychology and the other social sciences are based by default upon reflective measurement. However, in some situations the measurement models are incorrectly specified as reflective when they should have been as formative.

Even though there are situations in which MVs are more realistically thought of as causes of a LV (formative scheme), most researchers consider them as effects (reflective scheme) without even questioning their appropriateness for the specific LV at hand.

This attitude may lead to misspecified models and there is growing evidence that measurement model misspecification has the potential for poor parameter estimates and misleading conclusions (see Dolce and Lauro, 2014; Jarvis et al., 2003; MacKenzie et al., 2005, among others).

Its effects extend also to the estimates of the path coefficients connected to the misspecified block. In covariance-based SEM this is mainly due to the fact that a reflective treatment of a block that should instead be modeled as formative reduces the variance of the LV. The variance of a reflectively-measured LV equals the common variance of its MVs, whereas the variance of a formatively-measured LV encompasses the total variance of its indicators (Fornell et al., 1991). Let us consider the common case of an exogenous formative block misspecified as reflective. If the level of the variance of the endogenous LVs is maintained, the estimates of the path coefficients connected to the misspecified exogenous LV is likely to be substantially inflated (Diamantopoulos et al., 2008).

2.3 Formative blocks in Covariance-based SEM

Several alternative formulations have been proposed for SEM specification, but a very general formulation was given by Bentler and Weeks (1980). In the Bentler-Weeks approach any variable in the model (MVs, LVs, errors and so on) is either a dependent or an independent variable. The distinction between latent and manifest variables is secondary to the distinction between dependent and independent variables. The covariances among the independent variables can be part of the model parameters while the covariances among the dependent variables, or between the dependent variables and the independent variables, are explained by the model through the so-called free parameters. This model specification permits the inclusion of formative MVs in Covariance-Based SEM.

The general structural equation is written as

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} \tag{2.1}$$

The vector $\boldsymbol{\eta}$ $(m \times 1)$ contains all dependent variables, $\boldsymbol{\eta}' = [\boldsymbol{y}', \boldsymbol{\pi}']$, where $\boldsymbol{y}$ is a vector of reflective MVs and $\boldsymbol{\pi}$ represents the endogenous LVs in the model. The vector $\boldsymbol{\xi}$ $(n \times 1)$ of all independent variables, $\boldsymbol{\xi}^{'} = [\boldsymbol{x}', \boldsymbol{\tau}', \boldsymbol{\zeta}^{'}, \boldsymbol{\varepsilon}']$, contains the formative MVs $\boldsymbol{x}$, the exogenous LVs $\boldsymbol{\tau}$, the disturbances $\boldsymbol{\zeta}$, and measurement errors $\boldsymbol{\varepsilon}$. $\boldsymbol{B}$ $(m \times m)$ and $\boldsymbol{\Gamma}$ $(m \times n)$ contain coefficients capturing the effects of the independent variables on the dependent variables.

To simplify matters, let us consider $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{y}$, $\boldsymbol{x}$ as deviations from their means. Since some of the variables in $\boldsymbol{\eta}$ are measured variables $\boldsymbol{y}$, we obtain them by means of a suitable selection matrix $\boldsymbol{G}$ with elements equal to 0 or 1 such that:

$$\boldsymbol{y} = \boldsymbol{G}_y\boldsymbol{\eta} \tag{2.2}$$

If we have formative MVes $\boldsymbol{x}$ in the vector $\boldsymbol{\xi}$ we extract them by:

$$\boldsymbol{x} = \boldsymbol{G}_x\boldsymbol{\xi} \tag{2.3}$$

$\boldsymbol{G}_x$ and $\boldsymbol{G}_y$ are Boolean matrices with all zero entries except for a single element equal to 1 in each row to select $\boldsymbol{x}$ from $\boldsymbol{\xi}$ and $\boldsymbol{y}$ from $\boldsymbol{\eta}$ respectively.

From the "general structural equation" and the "selection models" follows the "implied covariance matrix", given by the following matrix elements:

$$\begin{aligned}
\mathbf{\Sigma_{yy}} &= \mathbf{G_y(I-B)^{-1}}\gamma\mathbf{\Phi}\gamma^{'}\mathbf{(I-B)^{-1}G_y^{'}} \\
\mathbf{\Sigma_{yx}} &= \mathbf{G_y(I-B)^{-1}}\gamma\mathbf{\Phi G_x^{'}} \\
\mathbf{\Sigma_{xx}} &= \mathbf{G_x\Phi G_x^{'}}
\end{aligned} \tag{2.4}$$

where $\mathbf{\Phi}$ is the covariance matrix of the independent variables $\boldsymbol{\xi}$, and it is not function of other parameters.

When no measured variable is included in $\boldsymbol{\xi}$, there are no formative MVs, thus $\mathbf{\Sigma_{yx}}$ and $\mathbf{\Sigma_{xx}}$ are null, and $\mathbf{\Sigma_{yy}} = \mathbf{\Sigma}$.

Identification of formative measurement models still represents an open problem. Obviously, a necessary but not sufficient condition is the "t-rule" (i.e., the number of free parameters must not exceed the number of elements in the covariance matrix). Regarding the "scaling rule" (i.e. each LV needs to be scaled for the model to be identified), among other options, we can fix a weight from a formative MV to the LV at some non-zero value (MacCallum and Browne, 1993).

To resolve the indeterminacy associated with the LV level error term, a necessary but not sufficient condition is the so-called "2+ Emitted Paths Rule". Every LV with an unrestricted variance or unrestricted error variance must emit at least two directed paths to other variables, when these latter variables have unrestricted error variances (Bollen and Davis, 2009). Another solution is to fix the variance of the disturbance term to zero. However, dropping the residual term implies the theoretical assumption that the formative MVs completely capture the underlying LV and no unexplained variance exists. The obtained variable becomes a "composite variable", not a formatively-measured LV (MacCallum and Browne, 1993).

Another strategy to identification is the so-called “piecewise identification”, based on breaking the model into smaller pieces and establishing the identification of one piece before moving on to the next piece (Bollen and Davis, 2009).

Once a model is identified, we can estimate its parameters by standard estimation procedures (Bentler and Weeks, 1980).

Another important issue that needs to be addressed when modeling formative blocks is what to do with the covariances among MVs in the model (MacCallum and Browne, 1993). Since formative MVs are simply exogenous variables, they may be correlated due to spurious causes not considered in the model, thus it would be more appropriate to free all covariances among them.

Finally, it must be stressed that in the recent SEM literature there is an interesting discussion on the meaning of the formatively-measured LV. Some researchers state that the known solutions for the matter of identification imply interpretation difficulties. A recent paper by Treiblmaier et al. (2011) clearly illustrates this issue. The authors state that a formatively-measured LV is actually a second-order factor that is predicted by some MVs and that explains the correlation of its consequent variables. Without this correlation the formatively-measured LV would disappear, and this is contrary to the idea that the LV is created solely by their exogenous MVs. This is a very interesting topic which needs further investigations.

2.4 A simulation Study

2.4.1 Design of the Simulation Study

The aim of this study is to investigate, within the same simulation design, the performance of both PLS-PM and ML-SEM when a block is modeled as formative.

In order to satisfy the above mentioned identification rules, we considered a formative exogenous block with unrestricted disturbance variance, that emits at least two directed paths to other LVs, and the covariances between the measurement errors of the MVs related to the endogenous LVs were fixed to zero.

A model with this framework is particularly justified when dealing with customer satisfaction data. Indeed in the European Customer Satisfaction Index model

(ECSI, 1998) literature suggests that the LV Image may be formatively-measured. We considered the ECSI model consisting of one formatively-measured exogenous LV, Image ($\boldsymbol{\pi}_1$), and five reflectively-measured endogenous LVs, from $\boldsymbol{\pi}_2$ to $\boldsymbol{\pi}_6$, that represent Customer Expectations, Perceived Quality, Perceived Value, Customer Satisfaction and Customer Loyalty, respectively.

The Monte Carlo simulation was conducted by EQS 6.1 for Windows. The data generation process is consistent with the procedure described by Paxton et al. (2001) for a Monte Carlo SEM study. We first pre-specified the relationships in the SEM and then simulated data for the given parameter values.

The true path coefficient values were chosen in order to be as similar as possible to those commonly encountered in the marketing literature (Vilares et al., 2010). The postulated structural model is:

$$\begin{aligned}
\boldsymbol{\pi}_2 &= 0.9\boldsymbol{\pi}_1 + \zeta_{\mathbf{2}} \\
\boldsymbol{\pi}_3 &= 0.8\boldsymbol{\pi}_2 + \zeta_{\mathbf{3}} \\
\boldsymbol{\pi}_4 &= 0.3\boldsymbol{\pi}_2 + 0.7\boldsymbol{\pi}_3 + \zeta_{\mathbf{4}} \\
\boldsymbol{\pi}_5 &= 0.3\boldsymbol{\pi}_1 + 0.1\boldsymbol{\pi}_2 + 0.4\boldsymbol{\pi}_3 + 0.3\boldsymbol{\pi}_4 + \zeta_{\mathbf{5}} \\
\boldsymbol{\pi}_6 &= 0.3\boldsymbol{\pi}_1 + 0.7\boldsymbol{\pi}_5 + \zeta_{\mathbf{6}} \qquad (2.5)
\end{aligned}$$

For the LV Image, we adopted the following formative model:

$$\boldsymbol{\pi}_1 = 0.4\boldsymbol{x}_1 + 0.25\boldsymbol{x}_2 + 0.15\boldsymbol{x}_3 + 0.1\boldsymbol{x}_4 + 0.1\boldsymbol{x}_5 + \boldsymbol{\zeta}_1 \qquad (2.6)$$

Afterwards, the outer weights were modified in order to obtain the variance of $\boldsymbol{\pi}_1$ equal to one, taking into account the variance of the disturbance $\sigma^2_{\zeta_1}$ as well. In order to focus on the issue of formative blocks in SEM, the loadings between the reflective MVs and the related LVs were set all to 1.

We conducted the simulation setting different variance values of the disturbance $\boldsymbol{\zeta}_1$ in the formative block. In particular, we set four different values of $\sigma^2_{\zeta_1}$, from a small value of 0.05 (yielding a R^2 of 0.95) to a large value of .5 (yielding a R^2 of .5). The values of $\sigma^2_{\zeta_1}$ were chosen to satisfy the equation:

$$R^2 = 1 - \frac{Dev(\boldsymbol{\zeta}_1)}{Dev(\boldsymbol{\pi}_1)} \qquad (2.7)$$

For the model considered in this simulation study, the *2+ emitted path* rule and t rule are met. To satisfy the scaling rule a loading was fixed to 1 in each reflective block, and the first weight in the formative block was fixed to the given parameter value. Furthermore, to confirm the identification of the model we can use the piecewise identification strategy (Bollen and Davis, 2009).

2.4.2 Data Generation and Simulation Results

Once the population parameter values were set, we generated a total of 500 sets of data for three different sample sizes (100, 250, 500), four different disturbance variance values $\sigma^2_{\zeta_1}$ (0.05, 0.2, 0.35, 0.5), three different numbers of MVs in the formative block (3, 5, 7), and three different levels of inter-correlation among MVs in the formative block (0.2, 0.4, 0.6). We did not take into account inter-correlation levels greater than 0.6 to avoid the issue of multicollinearity which might arise when estimating a formatively-measured LV. Given that very often the data do not follow multivariate normal distributions, we also generated data from non-symmetric distributions with different degrees of skewness and kurtosis (0.5, -0.8; 1.5, 2.5; 2.5, 9) following the Vale and Maurelli (1983) technique built in EQS 6.1.

In order to estimate the ML-SEM and PLS-PM parameters, we employed the "ML" Discrepancy function by means of EQS, and the "centroid scheme" by means of the package PLSPM in R, respectively.

Three commonly reported measures were used to assess how well the methods estimate the parameters: the Relative Bias ($RBias$), the Standard Deviation (StD) and the Root Mean Square Error ($RMSE$) of the estimates. $RBias$ is computed as,

$$RBias = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{\theta}_i - \theta)}{\theta} \qquad i = 1, 2, ..., 500 \tag{2.8}$$

where n represents the number of replications in the simulation, $\hat{\theta}_i$ is the parameter estimate for each replication, and θ is the corresponding population parameter. StD is computed as,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - E(\hat{\theta}))^2} \qquad i = 1, 2, ..., 500 \tag{2.9}$$

where $E(\hat{\theta})$ is the mean of the estimates across the 500 simulated datasets. StD provides information on the efficiency of estimates.

Finally, $RMSE$ is computed as,

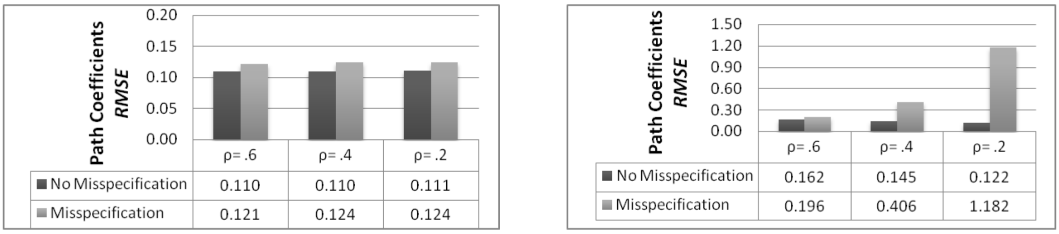
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2} \qquad i = 1, 2, ..., 500 \tag{2.10}$$

Obviously, it holds that $MSE = bias(\hat{\theta})^2 + Var(\hat{\theta})$. Thus $RMSE$ entails information on both bias and variability of the estimates.

In order to understand the effects of measurement model misspecification on the parameter estimates, we compared the performance of the two methods (PLS-PM and ML-SEM), applying both the correct measurement scheme for $\boldsymbol{\pi}_1$ (i.e., formative scheme) and the wrong measurement scheme (i.e., reflective scheme).

For the sake of simplicity, at this step we present only the mean of the $RMSE$ in the path coefficients connected to the exogenous LV $\boldsymbol{\pi}_1$.

Figure 2.2a reports the mean of $RMSE$ of the PLS-PM estimates for both the correctly specified measurement model and the misspecified measurement model, for each inter-correlation level among MVs (0.2, 0.4, 0.6).



(a) PLS-PM

(b) ML-SEM

FIGURE 2.2: Mean of the $RMSE$ in the path coefficients connected to $\boldsymbol{\pi}_1$, for normal data scenario

The $RMSE$ of the PLS-PM estimates slightly increases when the measurement model is misspecified, but the inter-correlation level among MVs does not have any effect on the estimates. We found that the variability of the PLS-PM estimates is very low and almost equal for all these considered experimental conditions, thus different $RMSE$ values are due exclusively to the bias of the estimates. Confirming the expectation, PLS-PM tends to underestimate the path coefficients, and this bias slightly increases when the measurement model is misspecified.

This is not the case for the ML-SEM estimates (Figure 2.2b). The $RMSE$ increases drastically when the measurement model is misspecified. This is due to the fact that ML-SEM overestimates the path coefficients connected to the exogenous misspecified block. As said above, reflective treatment of a block that should instead be modeled using the formative scheme reduces the variance of the LV.

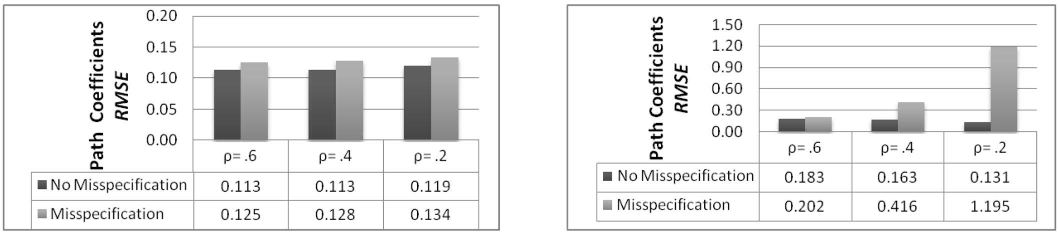
In ML-SEM the MVs inter-correlation level influences the extent of the $RMSE$ in the path coefficients connected to the misspecified block. This is due to the fact that the greater the level of the MVs inter-correlation, the smaller the change in the variance of a LV produced by measurement model misspecification. High MVs inter-correlations yield a less severe misspecification effect.

As regards the variability of the estimates, we found that the StD of the ML-SEM estimates increases when the measurement model is misspecified, for inter-correlation levels among MVs equal to 0.2 and 0.4. When the inter-correlation is on average equal to 0.6, the variability of the estimates is lower in the misspecified measurement model. Even though an inter-correlation level equal to 0.6 - on average - is not extremely high, this result may be due to the issue of multicollinearity. High correlation among MVs of a formative block can be a significant problem for measurement model parameter estimates, while it is a virtue for reflective blocks. However, the $RMSE$ of the estimates is higher when the measurement model is misspecified, for all the inter-correlation levels among MVs.

In keeping with these results, we think that in the case of uncertainty on the real nature of a LV (i.e., the probability of erroneously selecting a measurement scheme is high), researchers should choose PLS-PM rather than ML-SEM, as the $RMSE$ of the ML estimates is much higher when the measurement model is misspecified.

Figure 2.3 reports the mean of the PLS-PM estimates $RMSE$ (3a) and the mean of the ML-SEM estimates $RMSE$ (3b) for both the correctly specified measurement model and misspecified measurement model, for the non-normal data scenario. It does this for each inter-correlation level among MVs (0.2, 0.4, 06). For the sake of simplicity we show only the results for data with the highest degrees of skewness and kurtosis, i.e., 2.5 and 9, respectively.

As we can see in Figure 2.3, these results are not significantly different from those of the normal data scenario, for both the PLS-PM and the ML-SEM estimates. It is well known that PLS-PM is a powerful method because of the minimal demands on distributional assumptions of the variables (Chin, 1998)). However, ML-SEM is



(a) PLS-PM (b) ML-SEM

Figure 2.3: Mean of the *RMSE* in the path coefficients connected to $\boldsymbol{\pi}_1$, for non-normal data scenario

also generally robust against the violation of distributional assumptions (Satorra, 1990). This may explain our simulation results for the non-normal data scenario.

We also compared the performance of the two methods for three different numbers of MVs in the formative block (3, 5, 7), and for three different sample sizes (100, 250, 500). The results were not unexpected and we obtained no interesting findings in the case of measurement model misspecification. On the whole, we found that ML-SEM estimates are sensitive to the sample size and the number of MVs in the formative block, while PLS-PM estimates are extremely robust.

For all the reasons above and for the sake of simplicity, we did not take into account these experimental conditions, which would also complicate the reading of the results. Following we show the results for the ECSI model presented above, with a sample size of 250 and five formative MVs[2].

Table 2.1 shows *RBias*, *StD*, *RMSE*, and their absolute mean, *Mean*(*abs*), for the formative block outer weights (except for the first weight that was fixed in the ML-SEM), for both the smallest value of $\sigma^2_{\zeta_1}$ equal to 0.05 (left hand side) and the largest value equal to 0.5 (right hand side), for an inter-correlation level among the MVs equal to 0.6.

$\sigma_{\zeta_1} = \sqrt{0.05}$, $\rho = .6$							$\sigma_{\zeta_1} = \sqrt{.5}$, $\rho = .6$						
Outer Model	**Bias**		**StD**		**RMSE**		**Outer Model**	**Bias**		**StD**		**RMSE**	
	PLS	ML	PLS	ML	PLS	ML		PLS	ML	PLS	ML	PLS	ML
$\boldsymbol{\pi}_1$,$\mathbf{x}_2$	0.027	0.040	0.079	0.101	0.080	0.102	$\boldsymbol{\pi}_1$,$\mathbf{x}_2$	0.393	0.178	0.140	0.204	0.163	0.208
$\boldsymbol{\pi}_1$,$\mathbf{x}_3$	0.023	0.046	0.080	0.091	0.080	0.092	$\boldsymbol{\pi}_1$,$\mathbf{x}_3$	0.456	0.227	0.139	0.142	0.150	0.145
$\boldsymbol{\pi}_1$,$\mathbf{x}_4$	0.080	0.084	0.084	0.088	0.084	0.089	$\boldsymbol{\pi}_1$,$\mathbf{x}_4$	0.356	0.194	0.141	0.138	0.144	0.139
$\boldsymbol{\pi}_1$,$\mathbf{x}_5$	-0.020	0.018	0.079	0.086	0.079	0.086	$\boldsymbol{\pi}_1$,$\mathbf{x}_5$	0.275	0.151	0.140	0.134	0.142	0.135
Mean(abs)	**0.037**	**0.047**	**0.080**	**0.091**	**0.081**	**0.092**	Mean(abs)	**0.370**	**0.187**	**0.140**	**0.154**	**0.150**	**0.157**

Table 2.1: RBias, StD and RMSE of outer weights, for $\sigma_{\zeta_1} = \sqrt{0.05}$ and $\sigma_{\zeta_1} = \sqrt{.5}$), and high inter-correlation level ($\rho = .6$)

[2] Note that 250 is the common sample size used to estimate an ECSI model, and it is also a large enough number for good parameter estimations in both methods.

When $\sigma^2_{\zeta_1}$ is small, the outer weight estimates are nearly unbiased and with low variability in both methods. As the variance of $\boldsymbol{\zeta_1}$ increases (see the right-hand side of Table 2.1), we found that the bias of both PLS and ML estimates grows, but PLS-PM estimates are by far more biased compared to the ML's. The variability of the estimates increases for both methods, but PL-PM still produces estimates with lower variability. In terms of the $RMSE$ we found that PLS-PM performs slightly better than ML-SEM, regardless of the magnitude of the disturbance variance.

Considering an inter-correlation level equal to 0.2 (see Table 2.2), we found that ML-SEM outperforms PLS-PM in terms of bias of the estimates, regardless of the disturbance variance magnitude. As $\sigma^2_{\zeta_1}$ increases the bias of the PLS estimates grows drastically, while it slightly increases in the ML estimates. The variability of the estimates is almost similar for the two methods when the variance of $\boldsymbol{\zeta_1}$ is small. As the variance of $\boldsymbol{\zeta_1}$ increases the StD of both methods estimates increases, but in this case ML-SEM outperforms PLS-PM also in terms of StD. In terms of the $RMSE$, when the variance of $\boldsymbol{\zeta_1}$ is low the two methods perform almost similar. As the variance of $\boldsymbol{\zeta_1}$ increases ML-SEM outperforms PLS-PM.

$\sigma_{\zeta_1} = \sqrt{0.05}$, $\rho = .2$							$\sigma_{\zeta_1} = \sqrt{.5}$, $\rho = .2$						
Outer	**Bias**		**StD**		**RMSE**		**Outer**	**Bias**		**StD**		**RMSE**	
Model	PLS	ML	PLS	ML	PLS	ML	**Model**	PLS	ML	PLS	ML	PLS	ML
$\boldsymbol{\pi}_1$,**x**$_2$	0.018	0.007	0.055	0.066	0.056	0.066	$\boldsymbol{\pi}_1$,**x**$_2$	0.396	0.041	0.100	0.100	0.147	0.101
$\boldsymbol{\pi}_1$,**x**$_3$	0.017	0.007	0.059	0.063	0.060	0.064	$\boldsymbol{\pi}_1$,**x**$_3$	0.429	0.066	0.101	0.086	0.123	0.087
$\boldsymbol{\pi}_1$,**x**$_4$	0.034	0.016	0.063	0.063	0.064	0.064	$\boldsymbol{\pi}_1$,**x**$_4$	0.358	0.014	0.104	0.082	0.111	0.082
$\boldsymbol{\pi}_1$,**x**$_5$	-0.018	-0.019	0.055	0.056	0.056	0.057	$\boldsymbol{\pi}_1$,**x**$_5$	0.328	0.011	0.103	0.079	0.109	0.080
Mean(abs)	**0.022**	**0.013**	**0.059**	**0.063**	**0.059**	**0.063**	Mean(abs)	**0.378**	**0.033**	**0.102**	**0.087**	**0.123**	**0.087**

TABLE 2.2: RBias, StD and RMSE of outer weights, for $\sigma_{\zeta_1} = \sqrt{0.05}$ and $\sigma_{\zeta_1} = \sqrt{.5}$, and low inter-correlation level ($\rho = .2$)

Let us consider now the path coefficients connected to $\boldsymbol{\pi}_1$ when the inter-correlation level among MVs is equal to 0.6 (Table 2.3). On average, the PLS estimates are more biased compared to those of the ML-SEM, and the difference is more evident when $\sigma^2_{\zeta_1}$ increases. In terms of StD, PLS-PM outperforms ML-SEM by far. As $\sigma^2_{\zeta_1}$ increases the variability of the PLS estimates remains stable, while it increases in the ML estimates. In terms of $RMSE$, PLS-PM outperforms ML-SEM, regardless of the magnitude of $\sigma^2_{\zeta_1}$.

It must be noticed that ML-SEM method extremely overestimates the path coefficient between "Image", ($\boldsymbol{\pi}_1$), and "Customer Satisfaction",($\boldsymbol{\pi}_5$), and it does this systematically. This result is unexpected and needs further investigations.

$\sigma_{\zeta_1} = \sqrt{0.05}$, $\rho = .6$							$\sigma_{\zeta_1} = \sqrt{.5}$, $\rho = .6$						
Outer	**Bias**		**StD**		**RMSE**		**Outer**	**Bias**		**StD**		**RMSE**	
Model	PLS	ML	PLS	ML	PLS	ML	**Model**	PLS	ML	PLS	ML	PLS	ML
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_2$	-0.181	-0.002	0.029	0.150	0.166	0.151	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_2$	-0.400	0.000	0.044	0.275	0.362	0.275
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_5$	-0.253	0.388	0.055	0.200	0.093	0.231	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_5$	-0.644	0.406	0.045	0.438	0.199	0.455
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_6$	-0.181	0.124	0.046	0.099	0.071	0.106	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_6$	-0.540	0.062	0.038	0.144	0.166	0.145
Mean(abs)	**0.205**	**0.171**	**0.043**	**0.150**	**0.110**	**0.162**	Mean(abs)	**0.528**	**0.156**	**0.042**	**0.286**	**0.242**	**0.292**

Table 2.3: Inner Paths Coefficients RBias, StD and RMSE, for $\sigma_{\zeta_1} = \sqrt{0.05}$ and $\sigma_{\zeta_1} = \sqrt{.5}$, and high inter-correlation level ($\rho = .6$)

Considering a low inter-correlation level among MVs (see Table 2.4), we found that ML-SEM outperforms PLS-PM in terms of bias of the estimates, while PLS-PM outperforms ML-SEM in terms of StD, regardless of the magnitude of the disturbance variance. In terms of the $RMSE$, we found that PLS-PM performs slightly better than ML-SEM for a low variance of the disturbance. As the variance of $\boldsymbol{\zeta_1}$ increases ML-SEM outperforms PLS-PM.

$\sigma_{\zeta_1} = \sqrt{0.05}$, $\rho = .2$							$\sigma_{\zeta_1} = \sqrt{.5}$, $\rho = .2$						
Outer	**Bias**		**StD**		**RMSE**		**Outer**	**Bias**		**StD**		**RMSE**	
Model	PLS	ML	PLS	ML	PLS	ML	**Model**	PLS	ML	PLS	ML	PLS	ML
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_2$	-0.190	0.016	0.028	0.086	0.173	0.088	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_2$	-0.400	0.003	0.044	0.155	0.363	0.155
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_5$	-0.285	0.253	0.052	0.162	0.100	0.179	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_5$	-0.645	0.379	0.046	0.370	0.199	0.387
$\boldsymbol{\pi}_1,\boldsymbol{\pi}_6$	-0.139	0.197	0.042	0.081	0.059	0.100	$\boldsymbol{\pi}_1,\boldsymbol{\pi}_6$	-0.541	0.058	0.039	0.114	0.167	0.115
Mean(abs)	**0.205**	**0.155**	**0.041**	**0.110**	**0.111**	**0.122**	Mean(abs)	**0.529**	**0.147**	**0.043**	**0.213**	**0.243**	**0.219**

Table 2.4: Inner Paths Coefficients RBias, StD and RMSE, for $\sigma_{\zeta_1} = \sqrt{0.05}$ and $\sigma_{\zeta_1} = \sqrt{.5}$, and low inter-correlation level ($\rho = .2$)

Overall, our simulation results confirm that PLS estimators lack the parameter accuracy of ML estimators, and this bias is manifested in overestimating the loadings and underestimating the path coefficients, and the larger the disturbance variance the bigger the bias. On the contrary, PLS generally produces estimates with lower variability compared to those obtained using ML estimation method[3].

In order to provide researchers with some guidelines when having to choose between ML-SEM and PLS-PM to estimate formative blocks in SEM, we can take into account the $RMSE$ of the estimates. In keeping with the results we obtained, we think that for a quite high inter-correlation level among formative MVs (ρ= 0.6), researchers should prefer PLS-PM rather than ML-SEM, regardless of the disturbance variance. For low inter-correlation levels among formative MVs the decision depends on the magnitude of the disturbance variance. When $\sigma^2_{\zeta_1}$ is

[3] We do not show the results for the outer loadings in the reflective blocks and for the path coefficients not connected to $\boldsymbol{\pi}_1$ because they showed no interesting findings. Confirming the expectation, we found that PLS estimates present systematically more bias and lower variability compared to those obtained using ML estimation, regardless of the experimental conditions.

small, PLS-PM performs slightly better compared to ML-SEM. On the contrary, as the disturbance variance increases ML-SEM outperforms PLS-PM. Hence, in the latter case, researchers should prefer ML-SEM rather than PLS-PM.

2.5 Conclusion and Future Research

Measurement model misspecification may yield severe effects especially on the ML-SEM parameter estimates. Therefore, it is important to understand the effects of including formative blocks in SEM.

This study attempted to give some insight into this issue, comparing the bias and the variability of the ML-SEM estimates with those of the PLS-PM in the same simulation study.

For a model with a correctly specified formative block, we found that the inter-correlation level among formative MVs and the magnitude of the disturbance variance in the formative block have evident effects on the bias and the variability of the estimates.

In order to merge the information on the bias to the information on the variability of the estimates, we computed the $RMSE$ of the estimates. In terms of the Mean Square Error of the estimates, we found that for high inter-correlation levels among formative MVs, PLS-PM outperforms ML-SEM, regardless of the magnitude of disturbance variance. For low inter-correlation levels the performance of the two methods depends on the magnitude of the disturbance variance. When the disturbance variance is small, PLS-PM performs slightly better compared to ML-SEM. On the contrary, as the disturbance variance increases ML-SEM outperforms PLS-PM.

Different levels of complexity of the inner model were also included in this study. When the values of the population parameters were kept constant the complexity of the inner model did not have any effect on the estimates. On the contrary, the bias and the variability of the estimates were sensitive for different population parameter values. Since in a simulation study the value of the parameters should reflect values commonly encountered in applied research, we think that it would be interesting to run simulation studies considering other well-established models (like the ECSI model), where measurement model misspecification frequently occurs.

Different model specifications can also be considered including an endogenous formatively-measured LV.

Besides the descriptive statistics that we used to summarize and present the simulation results, inferential statistics can be used as well. For example, the experimental conditions can be dummy or effect coded, and main effects and interactions among experimental conditions can be evaluated using standard regression procedures.

Finally, we think that it would also be interesting to look further into the issue of multicollinearity among formative MVs.

Chapter 3

Non-Symmetrical Component-based Path Modeling

3.1 Introduction

PLS-PM is a method aimed at modeling a network of linear dependence relationships between blocks of variables where each block is summarized by a LV (Tenenhaus et al., 2005).

In order to respect the directions of the structural relationships specified in the Path diagram (i.e. the path directions), the estimation process should implicitly assume that there is a network of dependence relationships among LVs. However, it is known that PLS-PM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram.

The directions of the links in the structural model do not play a role in the algorithm apart from the specific case of the so-called path weighting scheme for the inner estimation (Tenenhaus et al., 2005).

In the inner step of the PLS-PM algorithm, each LV is defined as a linear combination of all the connected LVs. Two LVs are connected if there exists a link between the two blocks: an arrow goes from one LV to the other in the Path diagram, independently of the direction. In the path weighting scheme, the path direction is taken into account only in the way the inner weights are computed, but each LV is still defined in the inner step of the algorithm as a function of all the connected LVs.

PLS-PM provides components that are as much correlated as possible to each other while being somehow representative of each corresponding block of MVs. In the search for optimally correlated components, the estimation process amplifies interdependence among blocks and misses to distinguish between dependent and explanatory blocks in the structural model.

As a consequence, there is often a difference between what PLS-PM wants to model and what is actually computed by the PLS-PM algorithm.

We will first illustrate this inconsistency of PLS-PM by using a simple model, the case of two blocks of variables. For the case of more than two blocks of variables, we will look at the different criteria optimized by PLS-PM in order to show this issue.

The role of the LVs in the structural model depends on the way the outer weights are calculated. The only way for giving an explanatory role to a LV is to apply *Mode B*, while applying *Mode A* gives a role of dependent variable, whatever the path direction is (Dolce et al., 2015). However, in the case of more then two blocks, we cannot apply this rule (i.e., *Mode B* to the exogenous block and *Mode A* to the endogenous block), since some endogenous LVs appear only as dependent variable LVs, but others appear as both explanatory and dependent LVs. We defined the latter as “Bridge” LVs.

In this chapter, we propose a more suitable non-symmetrical approach that aims at maximizing the explained variance of the MVs in one block given the others, i.e. a new approach based on the optimization of a redundancy-related criterion in a multi-block framework.

In this new approach, the distinction between reflective and formative measurement model is disregarded. The nature of LVs and the direction of relationships between LVs and MVs is not taken into account. On the contrary, it is placed great emphasis on the dependence relationships between LVs in the structural model. We only make a distinction between explanatory blocks and dependent blocks in the structural model. Bridge blocks are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependents when play a dependent role.

In order to assess the quality and validity of results, we provide a new goodness-of-fit index based on redundancy criterion and prediction capability together with a classical bootstrap-based inferential approach.

Finally, we show the functioning of the proposed algorithm (implemented in a R code) through a simulation study.

The performance of the proposed method in terms of explained vatiability, predictiveness and interpretation is compared to classical PLSPM as well as to other component-based methods such as Regularized Generalized Canonical Correlation Analysis (Tenenhaus and Tenenhaus, 2011) and Generalized Structured Component Analysis (Hwang and Takane, 2004) using artificial data.

3.2 Dependence and Interdependence Relationships Between Blocks

The problem of finding quantitative relationships between groups of variables is central in multivariate analysis.

Multivariate techniques can be categorized as either interdependence or dependence techniques.

Interdependence techniques involve the simultaneous analysis of the relationships among variables in the data set, where variables are not classified as either dependent or explanatory. In the situations where we discard the fact that one block is the predictor and the other the criteria block, the direction of the relationship between the two blocks of variables is symmetrical, and the appropriate multivariate method in this case should predictive in both way, $\boldsymbol{X}_1 \rightarrow \boldsymbol{X}_2$ and $\boldsymbol{X}_2 \rightarrow \boldsymbol{X}_1$.

With dependence technique it is applied a non symmetrical analysis that takes into account a priori information on the different roles of the variables or sets of variables (Lauro and D'Ambra, 1984). The asymmetry is focused on the directional analysis in terms of dependence between the variables. A single variable or a set of variables is identified as the dependent variable to be explained or predicted by other variables known as explanatory or independent variables, and the analysis focus on deriving those combinations of predictors which explain most of the variation in the set of dependent variables. The aim is to develop a quantitative relationship between a predictor matrix $\boldsymbol{X}_1$ and a response matrix $\boldsymbol{X}_2$, that is, the predictive direction of the relationship between the two blocks of variables is asymmetrical, $\boldsymbol{X}_1 \rightarrow \boldsymbol{X}_2$.

The difference between the two techniques is, in fact, very much related to the classical issue of defining correlation versus regression.

In the case of two blocks of variables, canonical correlation analysis (CCA) (Hotelling, 1935, 1936) is one of the most commonly multivariate methods used when the aim of the analysis is to study the symmetrical relationship between two sets of variables.

CCA predicts the "most predictable criterion", which is a purely mathematical criterion and not something that it is determined by the researcher to be worth predicting for substantive reasons (Lohmöller, 1989). Hence, in CCA both blocks are treated in the same way and there is no distinction between predictor and criteria block. Weights for the set of variables $\boldsymbol{X}_1$ and for the set of variables $\boldsymbol{X}_2$ are chosen simultaneously to maximize the correlation between pairs of components (i.e., linear combinations of the original variables, one in each set), the component of $\boldsymbol{X}_1$ and the component of $\boldsymbol{X}_2$.

The problems with canonical correlation relate at least partly to the fact that the linear combinations derived might explain only very little of the variation in the original sets of variables. Furthermore, the correlations between canonical components cannot be interpreted as the degree of relation between the sets of variables. In particular, the squared canonical correlations represent the variance shared by the two canonical components of the same pair but not the variance shared by the two sets of observed variables. Two components might correlate very highly, while the explained variance of the variables is very low, which can lead to difficulties in interpretation.

To overcome the difficulty in using the squared canonical correlations as a measure of the shared variance between the two sets, a non-symmetric redundancy index was proposed by Stewart and (Stewart and Love, 1968). Based on Stewart and Love (1968) index, Wollenberg (1977) proposed an alternative method to CCA, which he refers to as redundancy analysis (RA), that maximizes the redundancy index.

Given two groups of variables RA searches for linear combinations of variables in one group that maximizes the variance of the other group explained by the linear combination. In RA the two blocks are not treated in the same way, one block is the predictor and the other the criteria block.

3.3 PLS-PM incoherence with Path Directions

There is often a difference between what PLS-PM wants to model (the hypothesized model depicted in the path diagram) and what is actually computed by the PLS-PM algorithm.

Generally, the directions of the links in the structural model do not play a role in the algorithm, as a consequence it misses to distinguish between dependent and explanatory LVs.

We first illustrate this issue by using a simple model, the case of two blocks of variables. For the case of more than two blocks of variables, a closer look at the different criteria optimized by PLS-PM will confirm the inconsistency in terms of coherence with the direction of the relationships specified in the path diagram.

3.3.1 PLS-PM Solutions for a Two-Block Model

In a simple hypothetical two-block model, each block of variables is felt to capture an underlying construct represented by a LV. An hypothetical two-block model (and in general all the path models with more than two blocks) can be represented by drawing a picture of it, the so-called *Path Diagram.* The *Path Diagram* provides a graphical representation of the relationships between LVs and between MVs and LVs, with the special property that they can be translated into a system of simultaneous equations.

Figure 3.1 presents the most commonly used graphical notation for the representation in Structural equation modeling.

Specifically, ellipses or circles represent the LVs and rectangles or squares refer to the MVs. Arrows show relatioships among the variables (either latent or manifest),

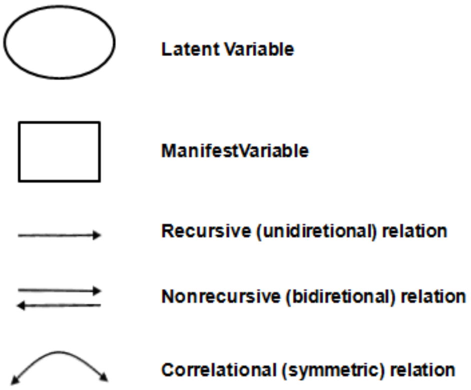

Latent Variable
ManifestVariable
Recursive (unidiretional) relation
Nonrecursive (bidiretional) relation
Correlational (symmetric) relation

FIGURE 3.1: Commonly used graphical notation in Structural Equation modeling

and the direction of the arrow define the direction of the relation, i.e. variables receiving the array are to be considered as dependent variables in the specific relationship. Recursive relation means no reciprocal causation or feedback loops between variables. Nonrecursive relation, on the contrary, means reciprocal causation or feedback loops between variables.

As said above, PLS-PM does not rigidly adhere to an underlying theoretical model depicted in the path diagram Chin (1998), and there is often a difference between what PLS-PM wants to model and what is actually computed by the PLS-PM algorithm.

In the case of two blocks of variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the PLS-PM algorithm converges to three different stationary equations, depending on the way the outer weights are calculated.

As Chin (1998) stated “when modeling path diagrams, it is important to consider the path relations among constructs as well as between constructs and their respective indicators”.

Figure 3.2 depicts a path diagram of two-block model with four MVs per block [1].

In this example the hypothesized relationship between the two LVs in the structural model is asymmetrical, predictive direction is from $\boldsymbol{\xi}_1$ to $\boldsymbol{\xi}_2$, that is, $\boldsymbol{\xi}_1$ is the predictor LV and $\boldsymbol{\xi}_1$ is the dependent LV. This suggests that the aim of the analysis is to seek a quantitative dependence relationship between the two blocks

[1] Path diagram generally shows the relations between all variables, including disturbances and measurement errors. However, in this case we do not consider disturbances and measurement errors since we shall focus on the relationships between LVs and between MVs and LVs

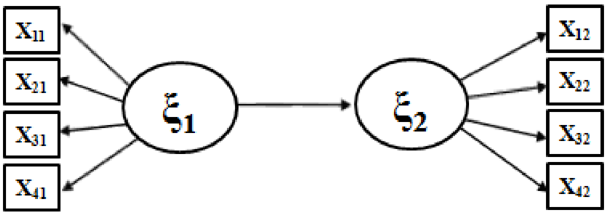


FIGURE 3.2: Two-block model with outwards directed scheme

of variables, in order to predict $\boldsymbol{\xi}_2$ from $\boldsymbol{\xi}_1$. As for the measurement model, both blocks are defined as outwards directed (Lohmöller, 1989) or reflective (Fornell and Bookstein, 1982), since the MVs are considered as being caused by the corresponding LV: variation in LV yields variation in MVs.

The PLS-PM literature has long suggested that the MVs weights in block defined as outwards directed are to be estimated using *Mode A* (e.g., Chin, 1998; Dolce and Lauro, 2014; Esposito Vinzi and Russolillo, 2013; Fornell and Bookstein, 1982; Hair et al., 2011; Henseler et al., 2009). Therefore, for the two-block model depicted in Figure 3.2, weights are computed by using Mode A, obtaining composites such that their are able to explain as much variance as possible in their respective MVs, giving minimum residual variances in the block structure (Wold, 1980). This model is equivalent to Tucker's (1958) inter-battery factor analysis (Chin, 1998; Tenenhaus et al., 2005). As a matter of fact, for this model, PLS-PM algorithm maximizes the covariance between the two composites, thus predictive direction of the structural model (i.e, the direction of the relationship in the structural model) is not explicitly considered in the algorithm.

In order to be coherent between what is depicted in the path diagram and what PLS-PM actually does, we think that the appropriate path diagram for a two-block model estimated by using *Mode A* in both blocks is the one depicted in Figure 3.3. The predictive direction is in both way, $\boldsymbol{\xi}_1 \rightarrow \boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_2 \rightarrow \boldsymbol{\xi}_1$. PLS-PM algorithm aims at obtaining composites such that their are able to explain as much variance as possible in their respective MVs, analyzing and amplifying interdependence among them. Hence, in this case PLS-PM adheres to the underlying theoretical model depicted in the path diagram.

Let's consider now the model in Figure 3.4. The hypothesized relationship between the two LVs in the structural model is asymmetrical, as the model in Figure 3.2. The predictive direction is from $\boldsymbol{\xi}_1$ to $\boldsymbol{\xi}_2$, thus $\boldsymbol{\xi}_1$ is the predictor LV and $\boldsymbol{\xi}_1$ is the dependent LV. Hence, the aim of the analysis should be to seek a quantitative

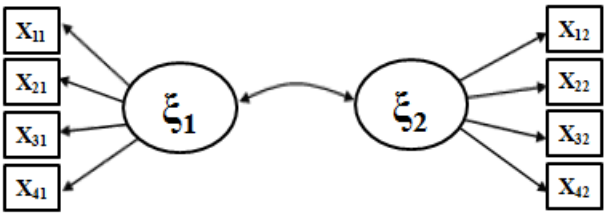


FIGURE 3.3: Two-block model with outwards directed scheme: not oriented arrows

dependence relationship between the two blocks of variables, for the prediction of $\boldsymbol{\xi}_2$ from $\boldsymbol{\xi}_1$. In the measurement model, MVs are viewed as causes of a LV (i.e., variation in MVs causes variation in LV), the block can be conceptually defined as inwards directed or formative. In such a case, literature suggests that MVs weights are to be estimated using *Mode B*.

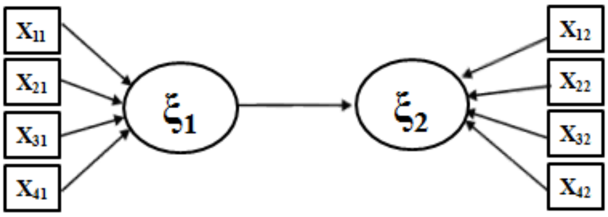


FIGURE 3.4: Two-block model with inwards directed scheme

PLS-PM algorithm using *Mode B* for both blocks computes the outer weights optimally in order to maximize the correlation between the two composites, and no attempt is made to explain the variances of the MVs (Chin, 1998; Tenenhaus et al., 2005; Wold, 1980). In this case, PLS-PM algorithm does converge to the first canonical components of canonical correlation analysis of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, giving the first canonical coefficient as the estimated correlation between the two composites.

Even in this case, the directions of the link in the structural model (i.e., the role of the blocks in the model) are not explicitly considered in the algorithm. The procedure misses to distinguish between dependent and explanatory blocks in the model. Blocks are treated in the same way, that is the direction of the relationship between the two blocks of variables is symmetrical. For this reason, we think that the coherent path diagram for a two-block PLS path model estimated using *Mode B* should consider a predictive direction in both way, $\boldsymbol{\xi}_1 \rightarrow \boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_2 \rightarrow \boldsymbol{\xi}_1$, as the one depicted in Figure 3.5.

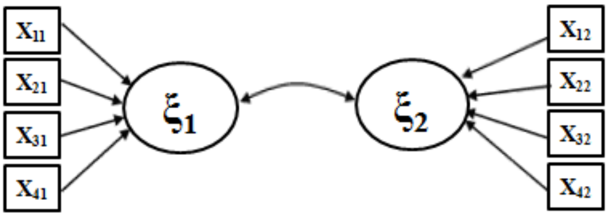


FIGURE 3.5: Two-block model with inwards directed scheme: not oriented arrows

3.3.2 Illogical form of Redundancy Analysis in PLS-PM

As a third case, we consider a two-block model where a block is defined as inwards directed and the other as outwards directed. In particular, the exogenous block is specified as formative, while the endogenous block as reflective (see Figure 3.6). In such a case, in PLS-PM the outer weights of the exogenous LV $\boldsymbol{\xi}_1$ are generally computing by Mode B, while they are usually computing by *Mode A* in the endogenous block. PLS algorithm converges to the same results of the RA of $\boldsymbol{X}_2$ with respect to $\boldsymbol{X}_1$ (Chin, 1998; Tenenhaus et al., 2005; Wollenberg, 1977).

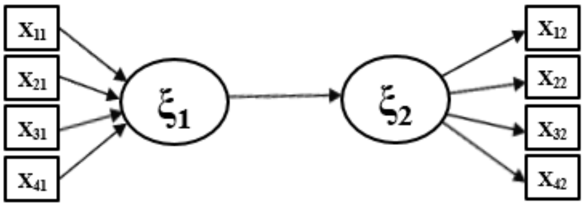


FIGURE 3.6: Two-block model with inwards and outwards directed scheme: Redundancy analysis

Redundancy refers to the mean variance in the endogenous block of variables, being predicted by the exogenous LV, $\boldsymbol{\xi}_1$ (i.e, a linear composite of the MVs of the block $\boldsymbol{X}_1$).

In this example, the hypothesized relationship between the two LVs in the structural model is asymmetrical, predictive direction is from $\boldsymbol{\xi}_1$ to $\boldsymbol{\xi}_2$, thus $\boldsymbol{\xi}_1$ is the predictor LV and $\boldsymbol{\xi}_1$ is the dependent LV. PLS-PM adheres to an underlying theoretical model, being coherent with what is depicted in the path diagram in Figure 3.6.

The choice between using *Mode A* instead of *Mode B* for the computation of the outer weights, mainly depends on the theoretical difference between the two scheme, based essentially on the hypothesized relationships between LVs and their own MVs (see Chapter 2). Under conditions of low theoretical knowledge on the nature of the LVs, a rule of thumb in PLS-PM is to apply *Mode B* to the exogenous block and *Mode A* to the endogenous block (Wold, 1980). However, to the best of our knowledge, there are hardly any studies in the literature that give reasons for following this rule and analyze into details this issue.

Dolce and Hanafi (2015) illustrated this issue by using a simple model, the case of two blocks of variables. The authors showed that Wold (1980) suggestion about using *Mode B* to the exogenous block and *Mode A* to the endogenous block is not just a rule of thumb. Instead, applying *Mode B* for the endogenous block does not make sense in the framework of SEM.

In general, beyond the theoretical differences between the two different measurement model schemes, depending on the way the outer weights are calculated the role of the LV in the structural model changes. The only way for giving an explanatory role to a LV is to apply *Mode B*, while applying *Mode A* gives it a role of dependent variable, whatever the path direction is. Thus, the predictive direction in the structural model is given by the utilized outer mode.

The model in Figure 3.7 shows a two-block model where the hypothesized relationship between the two LVs in the structural model is asymmetrical, predictive direction is from $\boldsymbol{\xi}_1$ to $\boldsymbol{\xi}_2$, thus $\boldsymbol{\xi}_1$ is the predictor LV and $\boldsymbol{\xi}_2$ is the dependent LV.

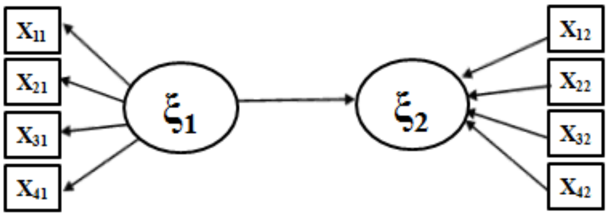


Figure 3.7: Two-block model for an illogical form of redundancy analysis in PLS-PM

However, the exogenous block is specified as reflective (i.e., outwards directed) and the endogenous block as formative (i.e., inwards directed). Thus, in PLS-PM the outer weights of the exogenous LV $\boldsymbol{\xi}_1$ are computing by Mode A, while they are computing by *Mode B* in the endogenous block. In this case, PLS algorithm converges to the same results of the RA of $\boldsymbol{X}_1$ with respect to $\boldsymbol{X}_2$ (the predictive direction is from $\boldsymbol{\xi}_2$ to $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2 \rightarrow \boldsymbol{\xi}_1$). As a consequence, PLS-PM does not adhere to an underlying theoretical model, since it is not coherent with what is depicted in the path diagram in Figure 3.7. What it is depicted in the path diagram in Figure 3.7 can be defined as an illogical form of redundancy analysis in PLS-PM.

As a matter of fact, the only case where PLS-PM adheres to the underlying theoretical two-block model depicted in the path diagram, is for the model in Figure 3.6, that is, when the exogenous block is specified as formative (and the outer weights are computed by *Mode B*), and the endogenous block is specified as reflective (and the outer weights are computed by Mode A), which is equivalent to performing a RA of the endogenous block with respect to the exogenous one.

3.3.3 PLS-PM solutions for Multi-Block Models

Recent works by Hanafi (2007), Krämer (2007) and Tenenhaus and Tenenhaus (2011), proved that the PLS-PM iterative algorithm optimizes different statistical criteria according to the different options chosen for the computation of the outer and inner proxies of the components, also for the case of more than two blocks of variables.

As it was shown in the first chapter, the stationary equations for most of the specific models obtained by running PLS-PM are known and it is possible to show that the PLS-PM generalizes many Multivariate Analysis techniques.

For the sake of easy reference, we show again here the different criteria optimized by PLS-PM.

When all the outer weights are calculated by means of *Mode B*, Hanafi proved that the Wold's PLS-PM algorithm monotonically converges to the the following criterion

$$\underset{||\boldsymbol{X}_k\boldsymbol{w}_k||^2=||\boldsymbol{X}_{k'}\boldsymbol{w}_{k'}||^2=1}{\arg\max} \sum_{k\neq k'} c_{kk'}\, g\Big(cor(\boldsymbol{X}_k\boldsymbol{w}_k, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'})\Big) \tag{3.1}$$

where g is one of the two functions

$$g(x) = \begin{cases} x^2 & \text{if } factorial \\ |x| & \text{if } centroid. \end{cases}$$

In 2007 Kramer showed that the PLS-PM algorithm was not based on a stationary equation related to the optimization of a twice differentiable function when *Mode A* was used for all the blocks in the model. In the same work, Kramer proposed a slight modified version of the classical *Mode A* outer scheme in which a normalization constraint is put on outer weights rather than latent variable scores. If this new scheme - also referred as New *Mode A* by Tenenhaus and Tenenhaus (2011) - is used for all the blocks in the model, PLS-PM iterative algorithm is monotonically convergent to the criterion:

$$\underset{||\boldsymbol{w}_k||^2=||\boldsymbol{w}_{k'}||^2=1}{\arg\max} \sum_{k\neq k'} c_{kk'} g\Big(cov(\boldsymbol{X}_k\boldsymbol{w}_k, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'})\Big) \tag{3.2}$$

where g is defined as above.

Looking at the different optimized criteria, it is clear that PLS-PM algorithm does not focus on directional analysis in terms of dependence relationships between blocks of variables.

Depending on the chosen estimation modes (for the measurement model) and schemes (for the inner model), PLS-PM provides composite scores that are as much correlated as possible to each other while being somehow representative of each corresponding block of manifest variables. The PLS-PM estimation process analyzes symmetrical relationships between blocks, thus, it misses to distinguish between the role of dependent and explanatory blocks in the inner model.

When both *new Mode A* and *Mode B* are used in the same model, Wold's procedure is shown to converge to the criterion

$$\begin{aligned} &\underset{\boldsymbol{w}_k}{\arg\max} \sum_{k\neq k'} c_{kk'} g\Big(cor(\boldsymbol{X}_k\boldsymbol{w}_k, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'}) \times \sqrt{var(\boldsymbol{X}_k\boldsymbol{w}_k)^{\tau_k}}\sqrt{var(\boldsymbol{X}_{k'}\boldsymbol{w}_{k'})^{\tau_{k'}}}\Big) \\ &\text{subject to} \qquad \tau_k||\boldsymbol{w}_k||^2 + (1-\tau_q)||\boldsymbol{X}_k\boldsymbol{w}_k||^2 = 1, \quad k = 1, ..., K. \end{aligned} \tag{3.3}$$

where $\tau_k = 1$ when the block k is estimated by *new Mode A* and $\tau_k = 0$ when the block k is estimated by Mode B, $c_{kk'}$ is the generic element of the Boolean square matrix C of order K, where $c_{kk'} = 1$ if $\boldsymbol{\xi}_k$ is connected to $\boldsymbol{\xi}'_k$ and $c_{kk'} = 0$ otherwise ($c_{kk} = 0$), $g(.)$ is the absolute value or the square function depending on the option used in the inner estimation step.

In the case of two block of variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the redundancy analysis of $\boldsymbol{X}_2$ with respect to $\boldsymbol{X}_1$ maximize the following criterion:

$$
\begin{aligned}
&\underset{\boldsymbol{w}_1,\boldsymbol{w}_2}{\arg\max}\, cor(\boldsymbol{X}_1\boldsymbol{w}_1, \boldsymbol{X}_2\boldsymbol{w}_2) \times var(\boldsymbol{X}_2\boldsymbol{w}_2)^{1/2} \\
&\text{subject to} \qquad ||\boldsymbol{X}_1\boldsymbol{w}_1||^2 = ||\boldsymbol{w}_2||^2 = 1
\end{aligned}
\tag{3.4}
$$

Looking at Equation 3.4 and Equation 3.3 it is clear that the role of the blocks in the structural model depends on the way the outer weights are calculated. The only way for giving an explanatory role to a LV is to apply *Mode B*, while applying *Mode A* gives a role of dependent variable, whatever the path direction is.

However, in the case of more then two blocks, we cannot apply this rule (i.e., *Mode B* to the exogenous block and *Mode A* to the endogenous block), since some endogenous LVs appear only as dependent variable LVs, but others appear as both explanatory and dependent LVs.

3.4 The Proposed Method

We propose a non-symmetrical component-based estimation approach for modeling a network of dependence relationships between blocks of variables where each block is summarized by a LV.

The proposed method, the Non-Symmetrical Component-based Path Modeling (NSC-PM), is based on the optimization of a redundancy-related criterion, and it is more suitable for prediction purposes. It aims at maximizing the explained variance of the MVs in one block given the others. The NSC-PM is applied in a multiblock framework, where relationships among blocks are specified in a path diagram.

In the PLS-PM literature, a LV which never appears as a dependent variable is called as exogenous LV. Otherwise, it is called as endogenous LV. Hence, some endogenous LV appears only as dependent variable while others appears as both explanatory and dependent. We defined the latter as “Bridge” LVs.

Taking into account the two roles that Bridge LVs play into the model, in NSC-PM they are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when play a dependent role.

The distinction between reflective and formative measurement model is disregarded. We only make a distinction between explanatory blocks and dependent blocks in the structural model.

3.4.1 Model Specification

Let us assume that P variables are collected in a partitioned table of standardized data $\boldsymbol{X}$ in K blocks:

$$\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_J, \boldsymbol{X}_{J+1}, ..., \boldsymbol{X}_{J+Q}, \boldsymbol{X}_{J+Q+1}, ..., \boldsymbol{X}_K],$$

where $\boldsymbol{X}_k$ $(k = 1, ..., J)$ are exogenous blocks, $\boldsymbol{X}_k$ $(k = J+1, ..., J+Q)$ are bridge blocks and $\boldsymbol{X}_k$ $(k = J + Q + 1, ..., K\ )$ are endogenous blocks. We denote by $\boldsymbol{\xi}_k$ $(k = 1, ..., K)$ the corresponding components for each block of variables.

As for the PLS-PM, the NSC-PM consists of two sub-models: the structural (or inner) model and the measurement (or outer) model.

In the inner model a generic endogenous LV - or bridge LV - $\boldsymbol{\xi}_m$ $(m = 1, ..., M)$ is linked to corresponding latent predictors by the following multiple regression model:

$$\boldsymbol{\xi}_m = \beta_{m0} + \sum_{m' \to m} \beta_{mm'} \boldsymbol{\xi}_{m'} + \boldsymbol{\zeta}_m \tag{3.5}$$

where $\beta_{mm'}$ is the so-called path coefficient capturing the effects of the predictor $\boldsymbol{\xi}_{m'}$ on the dependent component $\boldsymbol{\xi}_m$ and $\boldsymbol{\zeta}_m$ is the inner residual variable.

In the measurement model each MV $\boldsymbol{x}_{pk}$ is assumed to be generated as a linear function of its component $\boldsymbol{\xi}_k$ and its measurement error variable $\boldsymbol{\epsilon}_{pk}$,

$$\boldsymbol{x}_{pk} = \lambda_{pk0} + \lambda_{pk} \boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{pk}, \qquad k = 1, ..., K. \tag{3.6}$$

where λ_{pk0} is a location parameter and λ_{pk} is the loading coefficient.

As a vehicle for the estimation of the model parameters, the components are estimated as weighted aggregates of their indicators, regardless of the specified measurement model:

$$\hat{\boldsymbol{\xi}}_k = \sum_{p=1}^{P_k} w_{kp}\boldsymbol{x}_{pk}, \qquad k = 1, ..., K. \tag{3.7}$$

where w_{kp} is the outer weight.

3.4.2 The algorithm

The algorithm for estimating the unknown parameters of the model proceeds in two stages. The MVs are treated as deviation from their means and have unit variance.

In the first stage, the outer weight vectors, $\boldsymbol{w}_k$ $(k = 1, ..., K)$, are estimated by an iterative algorithm alternating outer and inner estimation steps, as in PLS-PM. In this stage we distinguish between explanatory and dependent blocks. As said above, in a path model there are LVs that play a role of explanatory variables, LVs that play a role of dependent variables and LVs that play a role of both explanatory and dependent variables (i.e., bridge LVs).

The NSC-PM iterative procedure consider the bridge blocks as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependents when play a dependent role.

In the following, the matrix $\boldsymbol{C} = [c_{kk'}]$ denotes a (K, K) binary lower-triangular matrix, which take into account the link between the latent variables. It is defined from the conceptual structural design of the model. The elements of the matrix $\boldsymbol{C}$ are defined as follows: $c_{kk'} = 1$ if the LV $\boldsymbol{\xi}_k$ depends on the LV $\boldsymbol{\xi}_{k'}$, otherwise $c_{kk'} = 0$. In the algorithm we make use of $\boldsymbol{C}$ and of its transpose $\boldsymbol{C'}$. The element of $\boldsymbol{C'}$ are denoted as $c'_{kk'}$.

The matrix $\boldsymbol{\Theta} = [\theta_{kk'}]$ denotes a (K, K) matrix defined from the correlation matrix, $\boldsymbol{R} = [r(\boldsymbol{X}_k\boldsymbol{w}_k, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'})]$, between the outer approximations of the LVs, $\boldsymbol{X}_k\boldsymbol{w}_k$, $k = 1, ..., K$. The matrix $\boldsymbol{\Theta}$ is used to compute the inner weights in the inner estimation step. We consider two options to calculate the inner weights: centroid scheme and factorial scheme.

If the centroid scheme is applied, $\theta_{kk'}$ is equal to the signs of the correlation between $\boldsymbol{X}_k\boldsymbol{w}_k$ and $\boldsymbol{X}_{k'}\boldsymbol{w}_{k'}$. In the factor scheme, $\theta_{kk'}$ is simply the correlation between $\boldsymbol{X}_k\boldsymbol{w}_k$ and $\boldsymbol{X}_{k'}\boldsymbol{w}_{k'}$.

Using the same formulation as in Hanafi (2007), the first stage of the algorithm is consisted as the following procedure. This procedure is iterate until convergence of the weight vectors $\boldsymbol{w}_k$ $(k = 1, ..., K)$.

A. Initialization

A.1. Choose arbitrary outer weight $\tilde{\boldsymbol{w}}_k^{(0)}$ $(k = 1, ..., K)$

A.2. Weight normalization such as $\boldsymbol{w}_k^{(0)} = \dfrac{\tilde{\boldsymbol{w}}_k^{(0)}}{\|\tilde{\boldsymbol{w}}_k^{(0)}\|}$

B. Inner estimation for dependent block $\boldsymbol{X}_k$, $\quad (k = J+1, ..., J+Q, ..., K)$

B.1. For $(k' = 1, ..., J, J+1, ..., J+Q)$; $\quad (k > k')$

$$\text{Compute } r_{kk'}^{(s)} = \begin{cases} r\big(\boldsymbol{X}_k\boldsymbol{w}_k^{(s)}, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'}^{(s)}\big) & \text{if } 1 < k' \leq J \\ r\big(\boldsymbol{X}_k\boldsymbol{w}_k^{(s)}, \boldsymbol{X}_{k'}\boldsymbol{w}_{k'}^{(s+1)}\big) & \text{if } J < k' \leq J+Q \end{cases}$$

B.2. Compute $\theta_{kk'}$ as,

$$\theta_{kk'}^{(s)} = sign(r_{kk'}^{(s)}) \qquad \text{if centroid weighting scheme}$$
$$\theta_{kk'}^{(s)} = r_{kk'}^{(s)} \qquad \text{if factorial weighting scheme}$$

B.3. Compute $\boldsymbol{z}_k^{(s)} = \sum_{k' \leq J} c_{kk'}\theta_{kk'}^{s}\boldsymbol{X}_{k'}\boldsymbol{w}_{k'}^{(s)} + \sum_{J<k'<k} c_{kk'}\theta_{kk'}^{s}\boldsymbol{X}_{k'}\boldsymbol{w}_{k'}^{(s+1)}$.

C. Outer estimation for dependent block $\boldsymbol{X}_k$, $\quad (k = J+1, ..., J+Q, ..., K)$

C.1. Compute $\tilde{\boldsymbol{w}}_k^{(s+1)} = \boldsymbol{X'}_k \boldsymbol{Z}_k^{(s)}$,

C.2. Compute $\boldsymbol{w}_k^{(s+1)} = \dfrac{\tilde{\boldsymbol{w}}_k^{(s+1)}}{\|\tilde{\boldsymbol{w}}_k^{(s+1)}\|}$

D. Inner estimation for explanatory block $\boldsymbol{X}_k$, $(k = J{+}Q, J{+}Q{-}1, ..., J, ..., 1)$

D.1. For $(k' = K, K-1, ..., J+Q, ..., J)$; $(k < k')$

$$\text{Compute } r_{kk'}^{(s)} = \begin{cases} r\big(\boldsymbol{X}_k \boldsymbol{w}_k^{(s+1)}, \boldsymbol{X}_{k'} \boldsymbol{w}_{k'}^{(s+1)}\big) & \text{if } J < k < J+Q \\ r\big(\boldsymbol{X}_k \boldsymbol{w}_k^{(s)}, \boldsymbol{X}_{k'} \boldsymbol{w}_{k'}^{(s+1)}\big) & \text{if } 1 < k < J \end{cases}$$

D.2. Compute $\theta_{kk'}$ as,

$$\begin{aligned} \theta_{kk'}^{(s)} &= sign(r_{kk'}^{(s)}) && \text{if centroid weighting scheme} \\ \theta_{kk'}^{(s)} &= r_{kk'}^{(s)} && \text{if factorial weighting scheme} \end{aligned}$$

D.3. Compute $\boldsymbol{z}_k^{(s)} = \sum_{k'>k} c'_{kk'} \theta_{kk'}^{s} \boldsymbol{X}_{k'} \boldsymbol{w}_{k'}^{(s+1)}$.

E. Outer estimation for explanatory block $\boldsymbol{X}_k$, $(k = J{+}Q, J{+}Q{-}1, ..., J, ..., 1)$

E.1. Compute $\tilde{\boldsymbol{w}}_k^{(s+1)} = (\boldsymbol{X'}_k \boldsymbol{X}_k)^{-1} \boldsymbol{X'}_k \boldsymbol{Z}_k^{(s)}$,

E.2. Compute $\boldsymbol{w}_k^{(s+1)} = \sqrt{n} \dfrac{\tilde{\boldsymbol{w}}_k^{(s+1)}}{\|\boldsymbol{X}_k \tilde{\boldsymbol{w}}_k^{(s+1)}\|}$

The procedure starts by choosing arbitrary normalized outer weight vectors $\boldsymbol{w}_k$ $(k = 1, ..., K)$. Then it updates the outer weights of the LVs that play a dependent role in the structural model at least in one equation, and subsequently it updates the outer weights of the LVs that play explanatory role in the structural model at least in one equation until convergence of the weights $\boldsymbol{w}_k$ $(k = 1, ..., K)$.

Note that the numerical implementations of the algorithm follows the essence of the multivariate Gauss–Seidel algorithm and, thus, Wold's original algorithm for PLSPM (Krämer, 2007). When computing the inner dependent component $\boldsymbol{z}_k^{(s)}$ at the iteration (s), it takes the weights from the iteration $(s+1)$, $\boldsymbol{w}_{k'}^{(s+1)}$, when $J < k' \leq J + Q$, and the weights from the iteration (s), $\boldsymbol{w}_{k'}^{(s)}$, when $1 < k' \leq J$. When computing the inner explanatory component $\boldsymbol{z}_k^{(s)}$ at the iteration (s) it takes the weights from the iteration $(s+1)$, $\boldsymbol{w}_{k'}^{(s+1)}$, since $k' > k$.

In the second stage, components are computed as weighted aggregates of their indicators:

$$\hat{\boldsymbol{\xi}}_k = \sum_{p=1}^{P_k} w_{kp}\boldsymbol{x}_{pk}, \qquad k = 1, ..., K. \tag{3.8}$$

Note that when convergence is achieved, for the exogenous and bridge blocks the weights $\boldsymbol{w}_k$ used for computing the components $\hat{\boldsymbol{\xi}}_k$ $(k = 1, ..., J, ..., J + Q)$ are the ones computed in E.2, while the weights for the endogenous block, $\boldsymbol{w}_k$ $(k = J + Q + 1, ..., K)$, are the ones computed in C.2.

The loadings are estimated by simple ordinary least squares (OLS) regressions of the manifest variables $\boldsymbol{x}_{kp}$ on the corresponding estimated component scores $\hat{\boldsymbol{\xi}}_k$.

The path coefficients are estimated through OLS simple or multiple regressions among the computed components, according to the equation 3.5.

3.4.3 Model Assessment

Like PLS-PM, the assessment of the quality of the NSC-PM results should take different aspects into account. The quality of the model depends on the goodness of fit of both the outer and the inner models, as it searches for component scores that well explain their own blocks while being related to each other as strongly as possible in accordance with the path diagram.

Moreover, as NSC-PM is based on the maximization of the explained variance of the MVs of the endogenous blocks, it is of extremely importance that the assessment of the quality of the model takes also into account appropriate measures of predictive ability.

Generally, the measures commonly used in PLS-PM can be used.

As in PLSPM, goodness of the inner model depends on the portion of variability of each endogenous components explained by the corresponding exogenous predictors, that can be measure by the multiple linear determination coefficient (R^2).

As for the measurement model, given that each MV $\boldsymbol{x}_{pk}$ is predicted by the corresponding components $\hat{\boldsymbol{\xi}}_k$:

$$\boldsymbol{x}_{pk} = \lambda_{pk}\hat{\boldsymbol{\xi}}_k + \boldsymbol{\epsilon}_{pk} \tag{3.9}$$

it follows that the MVs consist of a systematic part ($\lambda_{pk}\hat{\boldsymbol{\xi}}_k$) and a residual part ($\boldsymbol{\epsilon}_{pk}$). The proportion of the variance of $\boldsymbol{x}_{pk}$ which is reproduced by $\hat{\boldsymbol{\xi}}_k$ is equal to $cor^2(\boldsymbol{x}_{pk}, \hat{\boldsymbol{\xi}}_k)$ that, in the case of standardize MVs, corresponds to $\hat{\lambda}^2_{pk}$. This measures is also called "communality". If all the MVs are standardized, for each block k, the average of the communalities is equal to the average variance extracted (AVE) that expresses the part of variance of the block explained by $\hat{\boldsymbol{\xi}}_k$:

$$Com_k = \frac{1}{P_k}\sum_{p=1}^{P_k} cor^2(\boldsymbol{x}_{pk}, \hat{\boldsymbol{\xi}}_k) = \frac{1}{P_k}\sum_{p=1}^{P_k} \hat{\lambda}^2_{pk} = \frac{\sum_{p=1}^{P_k} \hat{\lambda}^2_{pk}}{\sum_{p=1}^{P_k} var(\boldsymbol{x}_{pk})} = AVE_k \tag{3.10}$$

The weighted average of all the K blocks specific communality indexes, with weights equal to the number of MVs in each block, can be use as a goodness of fit of the whole measurement model.

In NSC-PM communality index is conceptually appropriate just for the endogenous blocks.

For the LVs that appear at least in one equation of the structural model as predictors (i.e., exogenous and bridge LVs), MVs do not necessarily measure the same underlying construct, i.e., they are not supposed to be highly correlated. The components of the blocks that appear only as predictors (i.e, the exogenous blocks) are expected to maximize the explained MVs variance of the related dependent blocks. The components of the bridge blocks are expected to maximize the explained MVs variance of the related dependent blocks while being correlated with its own predictors LVs.

Moreover, since in the NSC-PM algorithm multiple regressions are applied when the outer weights are computed for the explanatory LVs, excessive correlations among MVs is not desired. However, in order to avoid the multicollinearity problem we proposed a solution (see next Section).

The interpretation of exogenous and bridge components should be based on the weights. The weights provide information about the direct relation between the MV and its LV, which reflect the impact of the MVs on its own LV (Bollen, 1989), and a comparison among them gives information about which MVs contribute most effectively to the LV. Loadings can also be used for interpretation, bearing in mind that while the outer weight is a measure of relative contribution of a MV to its LV, the loading can only be used to evaluate the absolute importance of a MV to its LV.

On the contrary, MVs of the endogenous blocks are theoretically expected to be unidimensional and to measure the same construct (i.e., MVs in each block are supposed to be highly correlated among each other). In this case, multicollinearity is not an issue as only simple regressions are involved.

The components of the endogenous blocks are expected to be as much correlated as possible to their predictor LVs, while being somehow representative of each corresponding block of MVs. The interpretation of endogenous components should be based on the loadings.

As a measure of the quality of the global model, the goodness-of-fit (GoF) index proposed by Amato et al. (2005) is not conceptually appropriate for measuring the global quality of NSC-PM. As a matter of fact, Gof index, as proposed by Amato et al. (2005), is computed as the geometric mean of the average communality and the average R^2 of the M linear determination coefficients:

$$GoF = \sqrt{\overline{Com} \times \overline{R^2}} \tag{3.11}$$

Thus, GoF index is partly based on average communality, as a consequence is conceptually appropriate only for the endogenous blocks. For this reason, we cannot use Gof index in NSC-PM.

A way of assessing the global model in NSC-PM may be measuring the amount of variance in the sets of variables of the dependent blocks explained by their own latent predictors. In this direction, we can use the redundancy index which measures the portion of variability of dependent block of MVs explained by its own predictors.

Given two blocks of variables, $\boldsymbol{X}_1 = (\boldsymbol{x}_{11}, ..., \boldsymbol{x}_{P_1 1})$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{12}, ..., \boldsymbol{x}_{P_2 2})$, redundancy index as proposed by Stewart and Love (1968) measures the proportion of the variance in the dependent set $\boldsymbol{X}_2$ that is accounted for by the predictor set $\boldsymbol{X}_1$. The redundancy analysis model, proposed by Wollenberg (1977), searches for the linear combination, $\hat{\xi}_1 = \boldsymbol{X}_1 \boldsymbol{w}_1$ (the so-called first redundancy variate), that maximizes the redundancy index, $R_{\boldsymbol{X}_2}$, defined as

$$R_{\boldsymbol{X}_2} = \sum_{p=1}^{P_2} corr(\hat{\boldsymbol{\xi}}_1, \boldsymbol{x}_{p2})^2 / P_2 \tag{3.12}$$

under the restriction that the variance of $\hat{\xi}_1 = 1$.

In the context of canonical correlation analysis (Hotelling, 1935, 1936), the redundancy index (Equation 3.12) can be written as:

$$R_{\boldsymbol{X}_2} = \rho^2 \sum_{p=1}^{P_2} corr(\hat{\boldsymbol{\xi}}_2, \boldsymbol{x}_{p2})^2 / P_2 \tag{3.13}$$

where ρ is the canonical correlation coefficient and $\boldsymbol{\xi}_2 = \boldsymbol{X}_2 \tilde{\boldsymbol{w}}_2$ is the first canonical component of $\boldsymbol{X}_2$ (Rencher, 1998).

For each endogenous block, in PLS-PM the redundancy index is computed as following,

$$Red_k = Com_k \times R_k^2. \tag{3.14}$$

where Com_k is the average of the communalities in the kth block and R^2_k is multiple linear determination coefficient in the regression model of $\hat{\boldsymbol{\xi}}_q$ on its own predictor LVs.

Looking at the redundancy index from the two different perspectives, it is clear that in PLS-PM the redundancy index is computed as in the context of CCA.

Since NSC-PM aims at maximizing the explained variance of the MVs in one block given the other (i.e., a redundancy-related criterion in a multi-block framework), as a redundancy measure in NSC-PM we propose to computed for each MVs of the endogenous and bridge blocks, the portion of its variability explained by its own predictors represented by the explanatory components as:

$$Red_{\boldsymbol{x}_{pk}} = R^2(\boldsymbol{x}_{pk}, \{\hat{\boldsymbol{\xi}}'_{k'} s\ explaining\ \hat{\boldsymbol{\xi}}_k\} \tag{3.15}$$

that is, as in the context of RA.

For a block k, the redundancy index is defined as

$$Red_k = \sum_{p=1}^{P_k} Red_{\boldsymbol{x}_{pk}} \tag{3.16}$$

Lohmöller gives some advice on evaluating the quality of the model, and it stated that the fit of the global model (outer and inner model) can be judged as satisfactory if the average of the redundancy indexes is high enough. Thus, he considered the redundancy index as an index of Goodness of fit of the global model.

In this perspective, we propose as a global goodness of prediction fit the average of all the $Red_{\boldsymbol{x}_{pk}}$, as it is based on redundancy criterion and prediction capability. If we denote by $\tilde{P}$ the number of MVs of the bridge and endogenous blocks, the global goodness of prediction fit is defined as

$$\overline{Red} = \frac{1}{\tilde{P}} \sum_{k=J+1}^{K} P_k \times Red_k \tag{3.17}$$

Just as with canonical correlations, no generally accepted guidelines have been established for the minimum acceptable redundancy index needed to judge a fit of the model as satisfactory. The researcher must judge the specific research problem being investigated to determine whether the redundancy index is sufficient to justify interpretation.

Model validation regards also the way relations are modeled, in both the structural and the measurement model. In these regards, since NSC-PM does not require any distributional hypothesis on MVs, confidence intervals for model parameters can be obtained by resampling techniques, such as Jackknife and Bootstrap.

NSC-PM is a method for predictive purposes, and could be an important technique deserving a prominent place in research applications when the aims of the analysis is prediction.

For these reasons, NSC-PM evaluation cannot focus only on parameter recovery and on the quality of the measurement model and the structural model - in terms of explained variance - indiscriminately.

In order to evaluate the model in terms of predictive ability the so-called Blindfolding procedure, using the Stone-Geisser's approach to crossvalidation, can be used (Chin, 1998; Geisser, 1975; Stone, 1974) (see Chapter 1).

3.4.4 A solution to the issue of Multicollinearity

As it is shown above, in the NSC-PM algorithm multiple regressions are applied when the outer weights are computed for the explanatory LVs. As a consequence, the stability of the MV outer weights are affected by the strength of the MV intercorrelations. For this reason, multicollinearity should be an important issue to take into account also in NSC-PM.

For the LVs that appear only as dependent variables in the structural model, multicollinearity is not an issue because only simple regressions are involved, and theoretically it is desired.

Excessive multicollinearity among MVs of explanatory LVs makes it difficult to separate the distinct influence of the individual MV on the LV or else the outer weights may be non-interpretable, having incoherent signs with the correlation with the corresponding LV.

A possible way to check for multicollinearity in a block of variables is computing the "tolerance" of each MV as $1 - R^2$, where the R^2 is the coefficient of determination for the regression of the the specific MV on the other MVs of the block (see Chapter 1). A measure related to the tolerance is the Variance Inflation Factor (VIF), computed as the inverse of the tolerance ($VIF = 1/TOL$) (Hair et al., 2010). A large VIF value indicates a high standard error of the specific weight due to multicollinearity among the MVs.

As a rule of thumb, the VIF should not exceed a value of 10, but, particularly when samples size is small, the critical value may be smaller then 10 (Hair et al., 2010). In general, the critical value should be defined considering the specific analysis objectives.

As a preliminary analysis to NSC-PM, multicollinearity is checked in the blocks that appear as explanatory at least in one equation of the structural model.

If excessive multicollinearity occurs in a block, we extract fewer principal components obtained by principal component analysis (PCA) on the specific block of variables, and then we use them instead of the original variables in the outer estimation step when the blocks play an explanatory role. In particular, it is applied a multiple regression of the instrumental inner composite $\boldsymbol{z}_k$ on the extracted principal components and then the outer composite is computed as weighted aggregates of the principal components.

A drawback of this procedure is that PCA creates components to explain the observed variability in the MVs, without considering at all the relationships of this variables with the MVs of the dependent blocks.

An alternative approach could be similar to the one proposed by Esposito Vinzi et al. (2010b) in the PLS-PM algorithm, i.e., providing PLS regression for estimating the outer weights as an alternative to OLS regression. As it is well known, PLS regression does take into account the relationships of the explanatory MVs with the response MVs.

3.5 A Comparison with other Component-based approaches

Among the component-based methods for SEM, PLS-PM is the most utilized (Wold, 1982). However, more recently two component-based methods have been presented as alternative approaches for the analysis of multi-block data.

Hwang and Takane (2004) have proposed a new full information method optimizing a global criterion and named Generalized Structured Component Analysis (GSCA). GSCA can be considered as a generalisation of principal component analysis to the case of several data tables connected by causal links.

More recently, Tenenhaus and Tenenhaus (2011) have presented a Regularized Generalized Canonical Correlation Analysis (RGCCA) as a new approach to multiple table analysis via a modified PLS-PM algorithm.

In this Section we compare the performance of the proposed method in terms of explained vatiability, predictiveness and interpretation to the classical PLS-PM as well as to the RGCCA and GSCA using artificial data.

Each component-based method considered in this simulation study optimizes a criterion and, obviously, it is the best method if we want to optimize this specific criterion. For this reason, the comparison in the simulation study is not made to show which method performs better, but rather to demonstrate how each method behaves in the particular case considered in the simulation study, and respect to the criterion that we are concerned.

3.5.1 Other Component-based approaches for multi-block data

In 2004 Hwang and Takane proposed the (GSCA) as an alternative to PLS-PM (Hwang and Takane, 2004). GSCA used a formulation similar to SEM even if the LVs are defined as weighted components of the MVs.

GSCA positions itself clearly as a component-based approach by defining a LV as a component from the stage of model specification.

The general models involves three sub-models:

- Measurement model: $\mathbf{Z} = \mathbf{C}\gamma + \epsilon$
- Structural model: $\gamma = \mathbf{B}\gamma + \zeta$
- Weighted relation: $\gamma = \mathbf{W}\zeta$

and combines the sub-models into a single one:

$$\begin{aligned}\begin{bmatrix}\mathbf{z}\\ \gamma\end{bmatrix} &= \begin{bmatrix}\mathbf{C}\\ \mathbf{B}\end{bmatrix}\gamma + \begin{bmatrix}\epsilon\\ \zeta\end{bmatrix}\\ \begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\mathbf{z} &= \begin{bmatrix}\mathbf{C}\\ \mathbf{B}\end{bmatrix}\mathbf{W}\mathbf{z} + \begin{bmatrix}\epsilon\\ \zeta\end{bmatrix}\\ \mathbf{V}\mathbf{z} &= \mathbf{A}\mathbf{W}\mathbf{z} + \mathbf{e} \end{aligned} \tag{3.18}$$

The unknown parameters of GSCA are estimated such that the sum of squares of the residuals $\mathbf{e}_i$ is as small as possible.

The single least-squares criterion to be minimized is the following:

$$\Phi = \sum_{i=1}^{N}(\mathbf{V}\mathbf{z}_i - \mathbf{A}\mathbf{W}\mathbf{z}_i)'(\mathbf{V}\mathbf{z}_i - \mathbf{A}\mathbf{W}\mathbf{z}_i) \tag{3.19}$$

More recently Tenenhaus and Tenenhaus (2011) proposed a new method, the RGCCA which is also used for analyzing relationship between blocks MVs. However, RGCCA is based on a monotonically convergent iterative algorithm and rely on an explicit optimization problem.

In RGCCA a continuum is built between the covariance criterion (*new Mode A*) and the correlation criterion (*Mode B*) (see Chapter 1) by means of the tuning parameter $0 \leq \tau \leq 1$, called shrinkage constant (see equation 4). Indeed, Tenenhaus and Tenenhaus (2011) have proved that fixing the tuning parameter to zero (i.e. using standardized LV scores) leads to criteria based on maximizing correlations among adjacent LVs while fixing the tuning parameter to one (i.e. using outer weights with unitary variance) leads to criteria based on maximizing covariances among adjacent LVs.

In the case of $0 < \tau < 1$, called mode Ridge, the determination of the shrinkage constant can be also computed optimally by using the analytical formula proposed by Schäfer and Strimmer (2005).

As already showed in Equation 3.3, the RGCCA optimization problem is :

$$\begin{aligned} &\underset{\boldsymbol{w}_k}{\arg\max} \sum_{k \neq k'} c_{kk'} g\Big(cor(\boldsymbol{X}_k \boldsymbol{w}_k, \boldsymbol{X}_{k'} \boldsymbol{w}_{k'}) \times \sqrt{var(\boldsymbol{X}_k \boldsymbol{w}_k)^{\tau_q}} \sqrt{var(\boldsymbol{X}_{k'} \boldsymbol{w}_{k'})^{\tau_{k'}}} \Big) \\ &\text{subject to} \qquad \tau_k ||\boldsymbol{w}_k||^2 + (1 - \tau_k) ||\boldsymbol{X}_k \boldsymbol{w}_k||^2 = 1, \quad k = 1, ..., K. \end{aligned} \tag{3.20}$$

$\tau_k = 1$ when the block k is estimated by *new Mode A* and $\tau_k = 0$ when the block k is estimated by Mode B.

Equation 3.20 is very interesting from the theoretical point of view and with the introduction of the New *Mode A* PLS-PM seems to be an heuristic approach only when the path weighting scheme is used (Esposito Vinzi and Russolillo, 2013).

However, it is not clear how users should interpret results obtained using a tuning parameter different from 0 or 1 that yields a method maximizing a mixture of correlations and covariances among adjacent LVs.

3.5.2 Design of the Simulation Study

In order to show the functioning and the performance of NSC-PM in terms of explained variability, predictiveness and interpretation, we compare it to the classical PLSPM, RGCCA and GSCA, in the framework of the same simulation design.

The reference model is the one depicted in Figure 3.8.

In the case of strong correlation within-blocks, the results of the most component-based methods are quite similar, because of the strength of the correlations. The same happens for the NSC-PM: the results are almost the same of the PLS-PM ones when correlation within-blocks is high.

For this reason, in order to understand the proprieties of the different component-based methods, the blocks are contaminated.

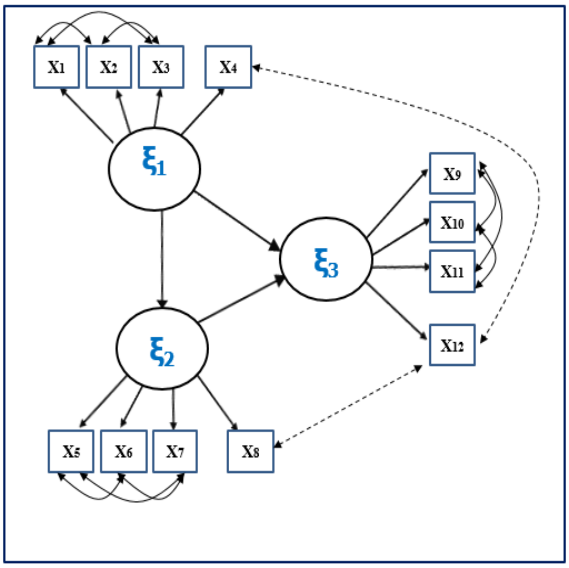


Figure 3.8: Theoretical Model for Simulated Data

In the case of two blocks of variables, when there are two variables, one in each set, which are not characteristic for the whole set, but yet highly correlated with each other, CCA may yields highly correlated, but unimportant components (i.e, the mean explained variance in each block is low). RA could overcome this problem as in the maximization problem it takes into account the variance of the dependent block as well as the correlation between the two components.

In this direction, we contamination the three blocks of variables in the model as following. Three MVs, one in each set, are not characteristic for the whole set, but yet highly correlated among each other. In particular, the variable $\boldsymbol{x}_4$ and the variable $\boldsymbol{x}_8$ are not highly correlated with the variables of their own blocks, but instead are both more correlated with the variable $\boldsymbol{x}_{12}$ (see Figure 3.8).

The mean of the correlations between the three related variables in each block is equal to 0.6. By including the fourth contaminating variable in each block we get a mean correlation level within-blocks equal to 0.35. However, in each block the Cronbach's α is about 0.7, only the first eigenvalue is greater than 1, while the second one is slightly less than 1. Hence, this is an extreme situation where the blocks of variables are generally considered as consistent and unidimensional.

3.5.3 Data Generation and Simulation Results

The data generation process and the subsequent analysis are conducted by EQS for Windows and R. Data are draw from a multivariate distribution with a pre-specified covariance matrix.

We generate 500 Monte Carlo samples for six different levels of correlation averages between-blocks ($\bar{\rho} = 0.05, 0.10, 0.15, 0.2, 0.25, 0.30$), to understand also the effects of different strengths of relationships between blocks.

In order to compare the performance of the different methods, we use three commonly reported measures (communality, redundancy and R^2) as well as the distribution of the path coefficient estimates represented by box plots.

We compare the NSC-PM with the PLS-PM, applying first *Mode B* for the three blocks - we refer this model as PLS-PM(B,B,B) - then for the three blocks we apply *Mode A* - PLS-PM(A,A,A). As for the comparison between NSC-PM and RGCCA, we do not consider the results of the RGCCA setting $\tau = 0$ for the three blocks - referred as RGCCA(0,0,0) - and $\tau = 1$ - referred as RGCCA(1,1,1), as they are very similar to the PLS-PM(B,B,B) results and the PLS-PM(A,A,A) results, respectively. More interesting is the comparison between NSC-PM and RGCCA(0,0.5,1) - RGCCA setting the value of $\tau = 0$ for the exogenous block, $\tau = 0.5$ for the bridge block and $\tau = 1$ for the endogenous block - and between NSC-PM and RGCCA(ridge mode) - RGCCA determining the shrinkage constants by using the Scháfer and Strimmer (2005) formula. Finally NSC-PM is compared to GSCA.

For the sake of simplicity we show only the results for correlation averages between-blocks equal to 0.15, the conditions that represent a middle ground between the case of very low correlation between blocks ($\bar{\rho} = 0.05$) and the case of high correlations between blocks ($\bar{\rho} = 0.3$). Showing the results for all levels of correlation average between-blocks would be redundant, since they showed no more findings and the results for one specific level of correlation average can be generalize for the all correlation levels.

Figure 3.9 reports the communalities of each MVs, for the PLS-PM(B,B,B) and the NSC-PM, when the correlation averages between-blocks is equal to 0.15.

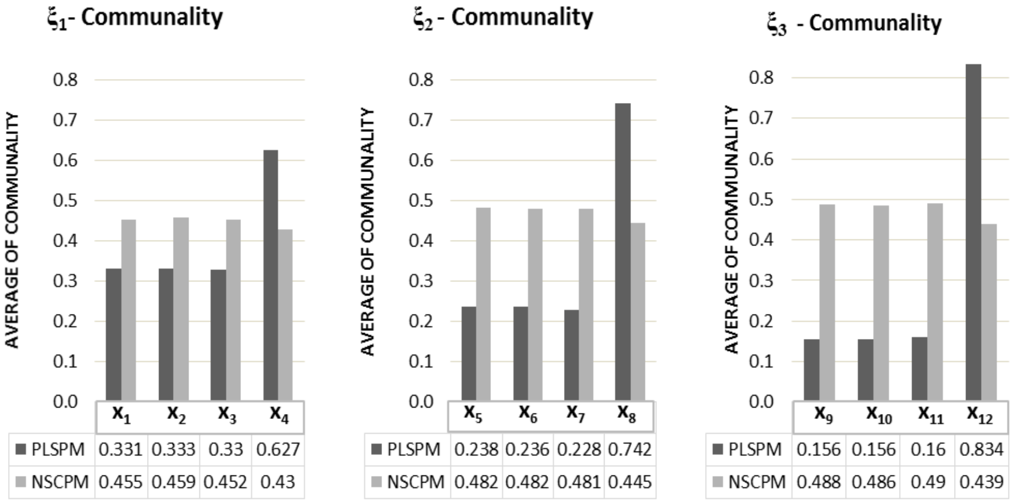


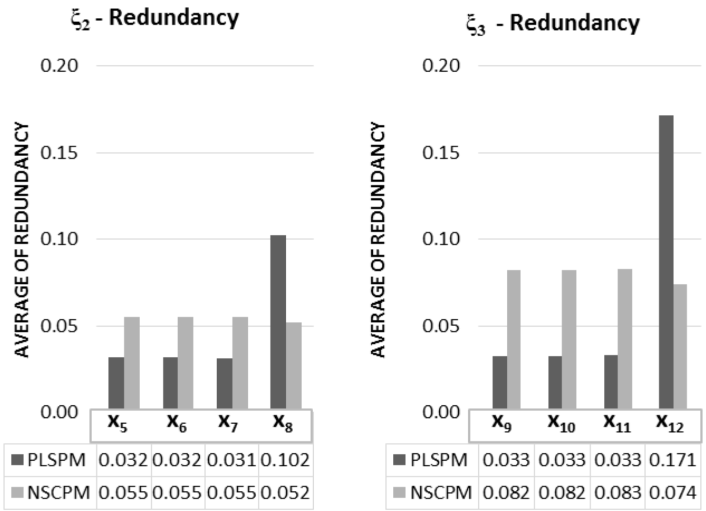
FIGURE 3.9: Communalities in PLS-PM(B,B,B) and NSC-PM

When all the outer weights are calculated by means of *Mode B*, PLS-PM maximizes correlations between components (see Formula 3.1). In the search for optimally correlated LV scores, PLS-PM(B,B,B) components explain better the MVs $\mathbf{x}_4$, $\mathbf{x}_8$ and $\mathbf{x}_{12}$, since they are highly correlated among each others. On the contrary, NSC-PM takes into account also the explained variance of the dependent block. As a consequence, NSC-PM explains on average more of the variations in the original variables compared to the PLS-PM.

In order to measure the the explained variance of the dependent MVs by the explanatory components, we computed the redundancies of the MVs of the endogenous and bridge blocks. As it is shown in Figure 3.10, on average the explained variance of the dependent MVs in one block given the others is larger in NSC-PM.

As for of the portion of variability of each endogenous component explained by the corresponding exogenous predictors, we computed the R^2 in the structural model. As expected the R^2 is higher in PLS-PM(B,B,B) in both regression models of the structural model (see Figure 3.11), since the PLS-PM(B,B,B) maximizes correlations between components.

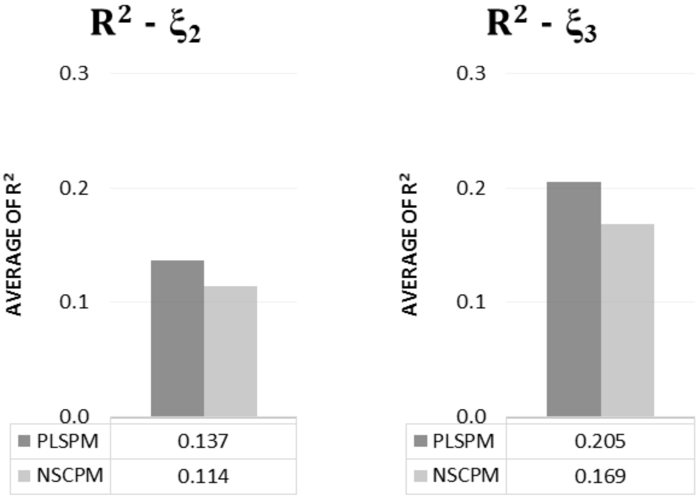
Finally, we compare the distributions of the path coefficient estimates represented by box plots (see Figure 3.12). Path coefficient estimates of PLS-PM(B,B,B) are higher as compared to the NSC-PM ones. In terms of variability of estimates, the two methods perform similarly.



	x_5	x_6	x_7	x_8
PLSPM	0.032	0.032	0.031	0.102
NSCPM	0.055	0.055	0.055	0.052

	x_9	x_{10}	x_{11}	x_{12}
PLSPM	0.033	0.033	0.033	0.171
NSCPM	0.082	0.082	0.083	0.074

FIGURE 3.10: Redundancies in PLS-PM(B,B,B) and NSC-PM



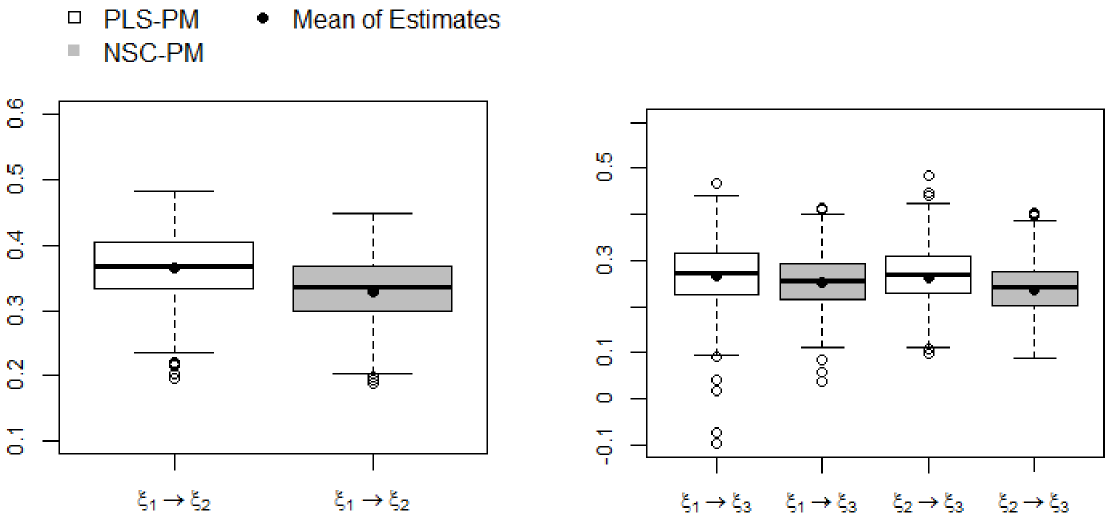
	R^2 - ξ_2	R^2 - ξ_3
PLSPM	0.137	0.205
NSCPM	0.114	0.169

FIGURE 3.11: R^2 of the structural model in PLS-PM(B,B,B) and NSC-PM

To sum up, PLS-PM(B,B,B) provide highly correlated, but less important components compared to the NSC-PM ones. This can also lead to difficulties in the interpretation on the results.

The performance of the NSC-PM is then compared with the performance of the PLS-PM(A,A,A).

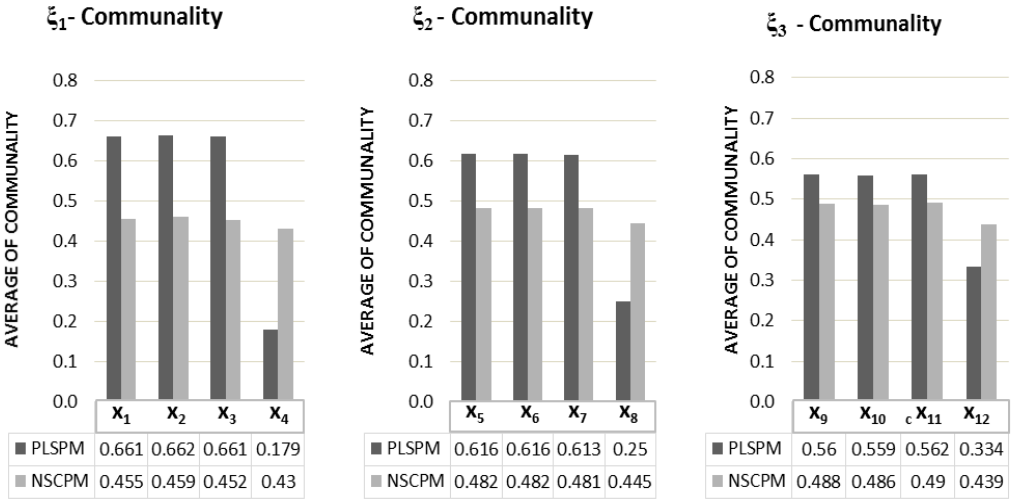
Figure 3.13 reports the communalities of each MVs, for the PLS-PM(A,A,A) and the NSC-PM, when the correlation averages between-blocks is equal to 0.15.



(a) (b)

Figure 3.12: Distributions of path coefficient estimates

When all the outer weights are calculated by means of *Mode A*, PLS-PM algorithm is not based on a stationary equation related to the optimization of a twice differentiable function (Krämer, 2007). However, for the slight modified version of the classical *Mode A*, the so-called *New Mode A*, PLS-PM maximizes covariancec between components (see Formula 3.2), thus it takes into account the variances of the blocks as well as the correlation between components.



	X_1	X_2	X_3	X_4
■ PLSPM	0.661	0.662	0.661	0.179
▪ NSCPM	0.455	0.459	0.452	0.43

	X_5	X_6	X_7	X_8
■ PLSPM	0.616	0.616	0.613	0.25
▪ NSCPM	0.482	0.482	0.481	0.445

	X_9	X_{10}	c X_{11}	X_{12}
■ PLSPM	0.56	0.559	0.562	0.334
▪ NSCPM	0.488	0.486	0.49	0.439

Figure 3.13: Communalities in PLS-PM(A,A,A) and NSC-PM

On average PLS-PM(A,A,A) explains more of the variations in the original variables compared to the NSC-PM. However, PLS-PM(A,A,A) components explain much less the MVs $\mathbf{x}_4$, $\mathbf{x}_8$ and $\mathbf{x}_{12}$. Hence, it may lose in prediction capability as these MVs are important for prediction being highly correlated among each other

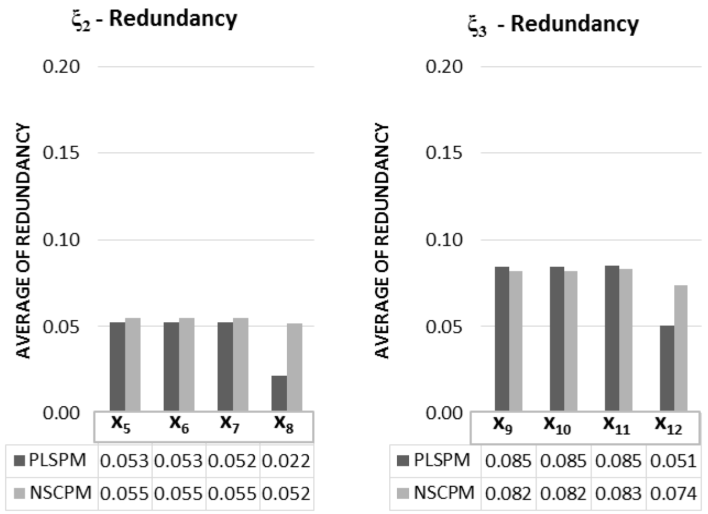


FIGURE 3.14: Redundancies in PLS-PM(A,A,A) and NSC-PM

Even if PLS-PM(A,A,A) explains on average more of the variations in the original variables compared to the NSC-PM, the explained variance of the endogenous and bridge blocks by the explanatory components is larger in NSC-PM.

Figure 3.15 reports the R^2 of the regression models in the structural model for both PLS-PM(A,A,A) and NSC-PM. The portion of variability of each endogenous component explained by the corresponding exogenous predictors, is higher in NSC-PM.

Finally, the distributions of the path coefficient estimates is represented in Figure 3.16). Path coefficient estimates of NSC-PM are higher as compared to the PLS-PM ones, while there is no evident difference in terms of variability of the estimates.

To sum up, in this case, PLS-PM(A,A,A) favours too much stability with respect to correlation, compared to the NSC-PM.

Let consider now the comparison between the NSC-PM and the RGCCA. Firstly we compare the NSC-PM with the RGCCA determining the shrinkage constants (τ) by using the Scháfer and Strimmer (2005) formula - RGCCA(ridge mode). The optimal shrinkage parameters estimated by RGCCA are $\tau = 0.028, 0.028, 0.028$.

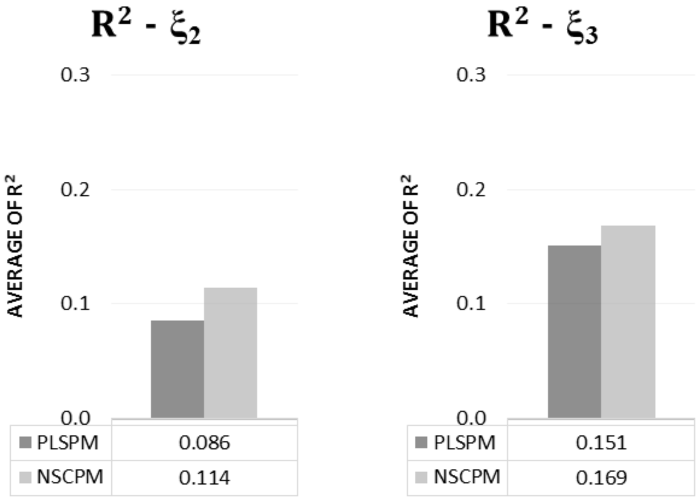


FIGURE 3.15: R^2 of the structural model in PLS-PM(B,B,B) and NSC-PM

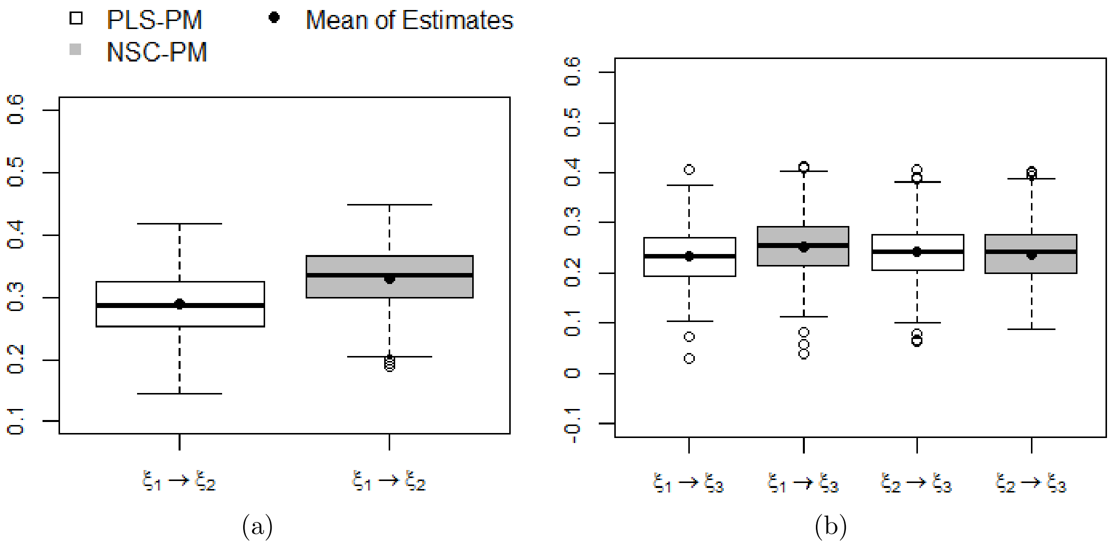


FIGURE 3.16: Distributions of path coefficient estimates

Figure 3.17 reports the communalities of each MVs, for the RGCCA(ridge mode) and the NSC-PM, when the correlation averages between-blocks is equal to 0.15.

Since in this case the optimal shrinkage parameters estimated by RGCCA are $\tau = 0.028, 0.028, 0.028$, it is expected to have results similar to the PLS-PM(B,B,B) ones[2]

[2] As said above, fixing the value of τ equal to zero leads to criteria based on maximizing correlations among adjacent LVs, as in PLS-PM(B,B,B)

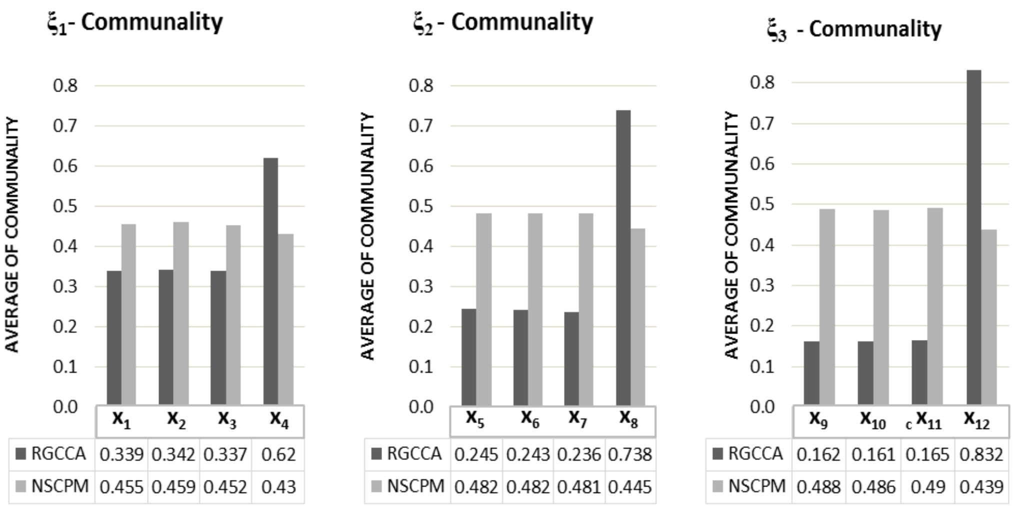


FIGURE 3.17: Communalities in RGCCA(ridge mode) and NSC-PM

In the search for optimally correlated LV scores, RGCCA(ridge mode) components explain better the MVs $\mathbf{x}_4$, $\mathbf{x}_8$ and $\mathbf{x}_{12}$, since they are highly correlated among each others. NSC-PM explains on average more of the variations in the original variables compared to the RGCCA(ridge mode).

As it is shown in Figure 3.18, on average the explained variance of the dependent MVs by the explanatory components is larger in NSC-PM.

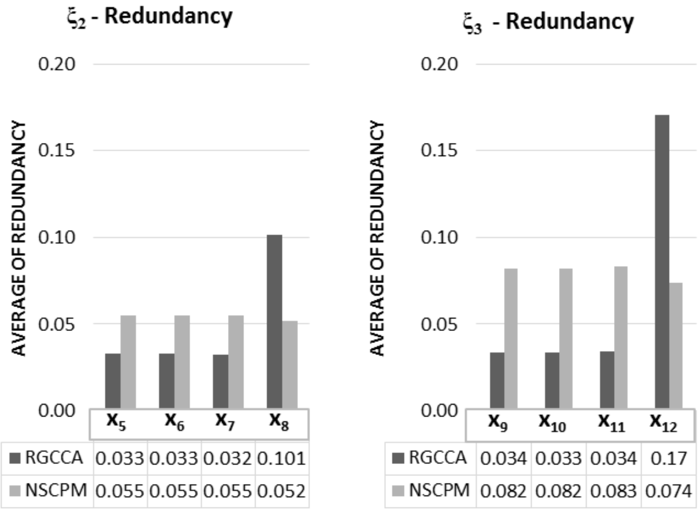


FIGURE 3.18: Redundancies in RGCCA(ridge mode) and NSC-PM

As expected the R^2 is higher in RGCCA(ridge mode) in both regression models of the structural model (see Figure 3.19), since the RGCCA(ridge mode) maximizes correlations between components.

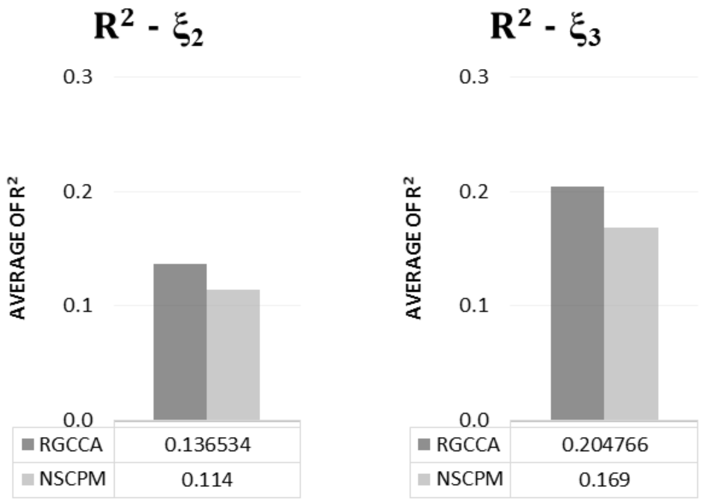


FIGURE 3.19: R^2 of the structural model in RGCCA(ridge mode) and NSC-PM

Let us compare now the NSC-PM with the RGCCA, setting the value of $\tau = 0$ for the exogenous block, $\tau = 0.5$ for the bridge block and $\tau = 1$ for the endogenous block - RGCCA(0,0.5,1).

Figure 3.20 reports the communalities of each MVs, for the RGCCA(0,0.5,1) and the NSC-PM, when the correlation averages between-blocks is equal to 0.15.

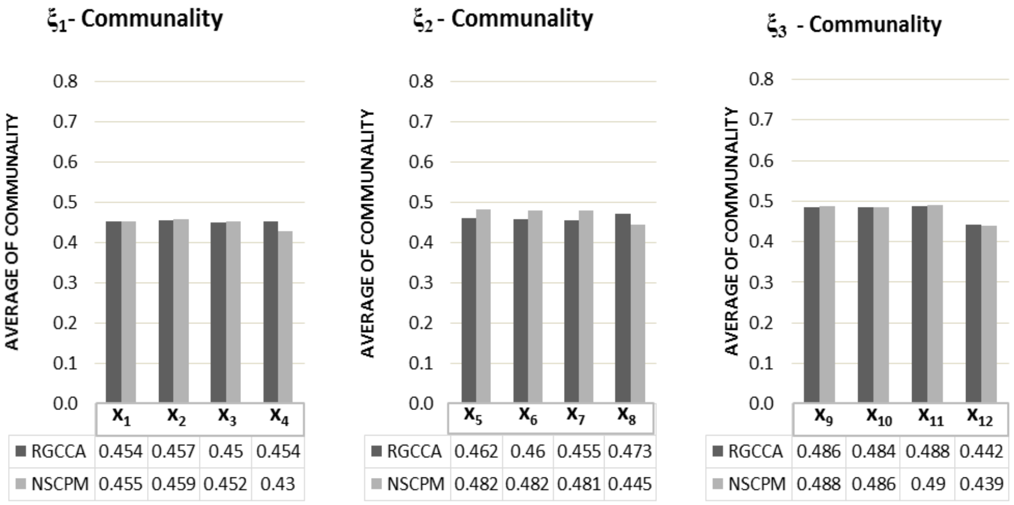
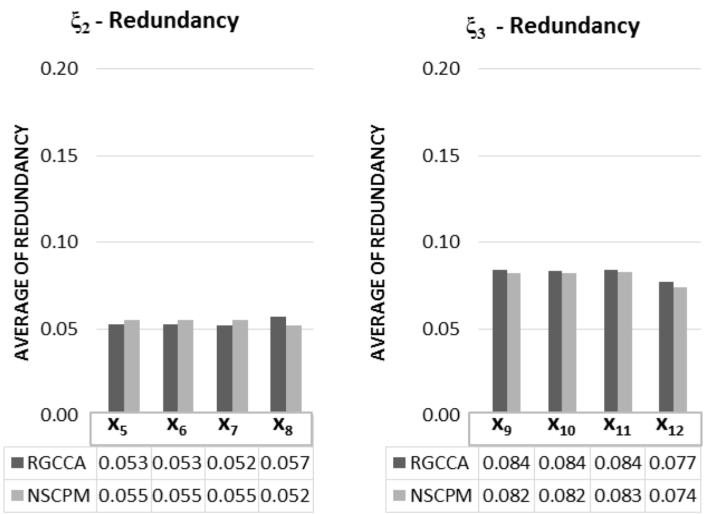


FIGURE 3.20: Communalities in RGCCA(0,0.5,1) and NSC-PM

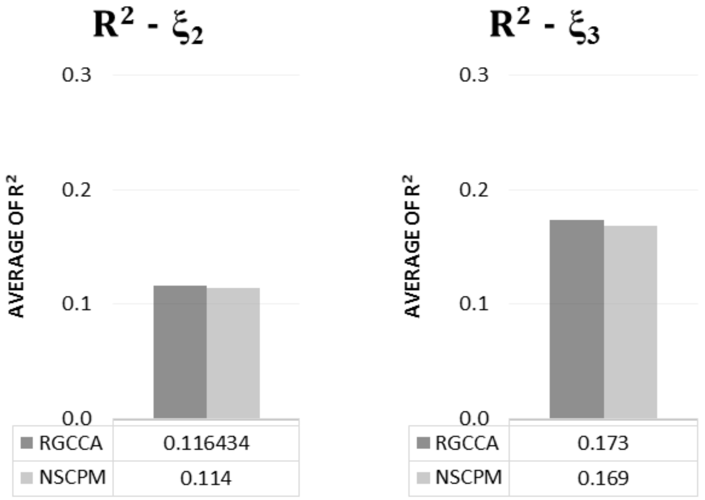
RGCCA(0,.5,1) communalities are very similar to the NSC-PM ones. The same happens for the redundancy indices (see Figure 3.21) and for the R^2 (see Figure 3.22).



	x_5	x_6	x_7	x_8
■ RGCCA	0.053	0.053	0.052	0.057
■ NSCPM	0.055	0.055	0.055	0.052

	x_9	x_{10}	x_{11}	x_{12}
■ RGCCA	0.084	0.084	0.084	0.077
■ NSCPM	0.082	0.082	0.083	0.074

Figure 3.21: Redundancies in RGCCA(0,0.5,1) and NSC-PM



	R^2 - ξ_2
■ RGCCA	0.116434
■ NSCPM	0.114

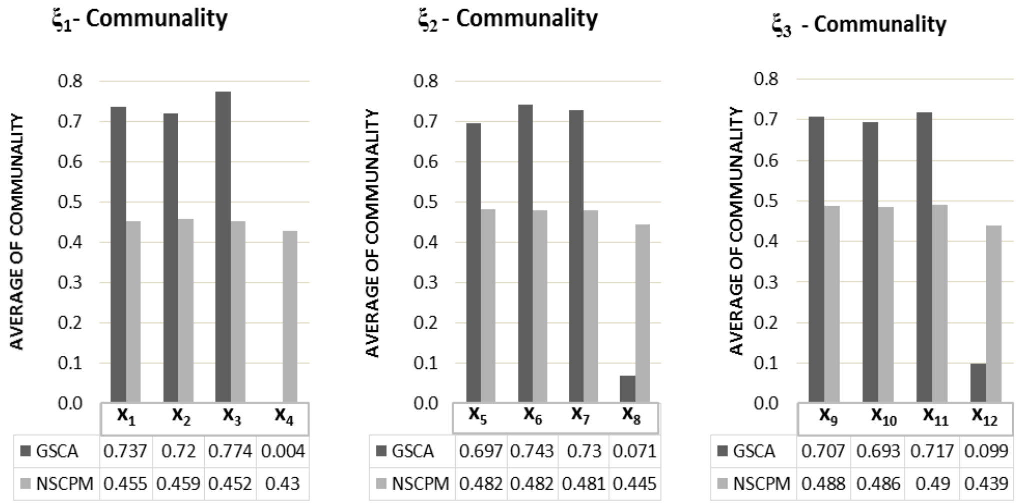
	R^2 - ξ_3
■ RGCCA	0.173
■ NSCPM	0.169

Figure 3.22: R^2 of the structural model in RGCCA(0,0.5,1) and NSC-PM

This similarity between RGCCA(0,0.5,1) results and NSC-PM results is not surprising, if one recall what said above about the dependence of the LVs role on the way the outer weights are computing in PLS-PM. To give a role of explanatory variable to an exogenous LV, outer weights are to be computed applying *Mode B* (that corresponds to fix the value of $\tau = 0$ in RGCCA). On the contrary, outer weights of endogenous LVs are to be computed applying *Mode A* - or *new Mode A* (that corresponds to fix the value of $\tau = 1$ in RGCCA) to give them a role of dependent variables. In the case of more then two blocks, we cannot apply this rule for all blocks of variables since some LVs appear as both explanatory and dependent LVs (the bridge blocks). Setting the value of $\tau = 0.5$ for the bridge blocks might be a valuable alternative to our approach. Clearly, further investigation are needed on this argument.

Finally, we compare the NSC-PM with the GSCA.

As it is shown in Figure 3.23, GSCA explains on average more of the variations in the original variables compared to the NSC-PM. However, GSCA components do not explain the MVs $\mathbf{x}_4$, $\mathbf{x}_8$ and $\mathbf{x}_{12}$, being the commulalities for these MVs close to zero.



	x_1	x_2	x_3	x_4
■ GSCA	0.737	0.72	0.774	0.004
■ NSCPM	0.455	0.459	0.452	0.43

	x_5	x_6	x_7	x_8
■ GSCA	0.697	0.743	0.73	0.071
■ NSCPM	0.482	0.482	0.481	0.445

	x_9	x_{10}	x_{11}	x_{12}
■ GSCA	0.707	0.693	0.717	0.099
■ NSCPM	0.488	0.486	0.49	0.439

Figure 3.23: Communalities in GSCA and NSC-PM

As a consequence, the portion of variability of each endogenous component explained by the corresponding exogenous predictors is higher in NSC-PM compared to the GCSA in both regression models of the structural model (see Figure 3.24).

In this simulation study, GSCA seems to favour too much stability with respect to correlation. GSCA components explain much variations in their own blocks of MVs but the correlations between components are very low. GSCA is the method that favours the more stability in the blocks of variables compared to the other component-based methods, even more than PLSPM (A,A,A).

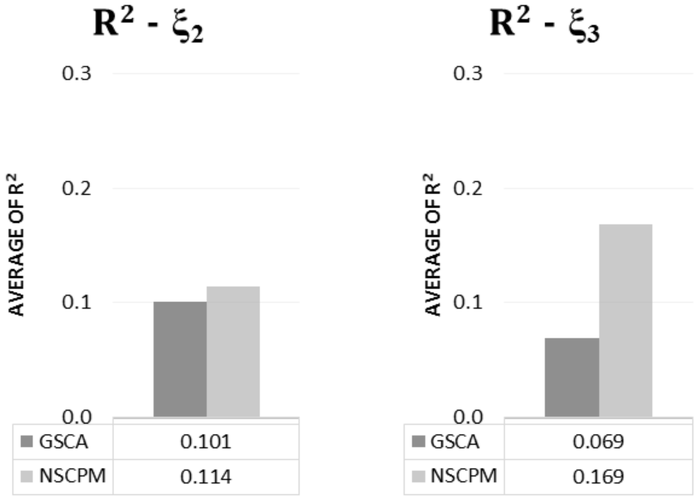


FIGURE 3.24: R^2 of the structural model in GSCA and NSC-PM

3.6 Conclusion

NSC-PM is a non-symmetrical approach that aims at maximizing the explained variance of the MVs of the endogenous and bridge blocks (i.e. an approach based on the optimization of a redundancy-related criterion in a multi-block framework).

The proposed method respects the direction of the relationship specified in the Path diagram (i.e. the path directions), since the directions of the links in the inner model play a role in the algorithm. In particular, bridge LVs (i.e., LVs that appear as both explanatory and dependent LVs in the structural model) are considered as explanatory when they play an explanatory role in the particular step of the algorithm, and as dependent when they play a dependent role.

Compared to the other component-based methods, NSC-PM seems to be a good compromise between favouring stability (high explained variance) in the blocks and correlation between components.

The NSC-PM components of the blocks that appear only as predictors (i.e, the exogenous blocks) are simultaneously stables (i.e., they explain much of variability in their own blocks) and explain as much as possible the variance of the MVs of the related dependent blocks. The components of the bridge blocks explain as much as possible the variance of the MVs of the related dependent blocks while being correlated with its own predictors LVs. The components of the endogenous blocks

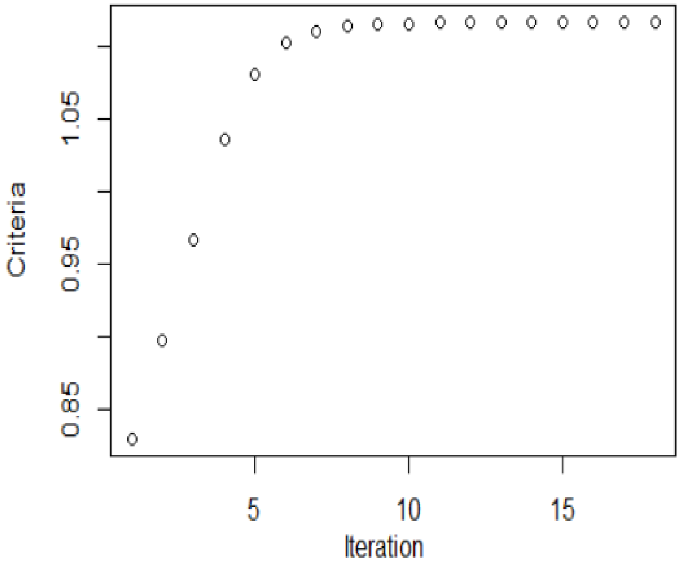


FIGURE 3.25: Evolution of the criterion as function of iterations for the algorithm

are as much correlated as possible to their predictor LVs, while being somehow representative of each corresponding block of MVs.

We have found always the convergence of the algorithm in practice. We looked then at the evolution of different criteria as a function of the iterations and we found that the following criterion increases monotonically (see Figure 3.25):

$$\max_{\mathbf{w}_k,\mathbf{w}_{k'}} \sum_{k\neq k'} c_{kk'}[cor(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_{k'}\mathbf{w}_{k'})var(\mathbf{X}_k\mathbf{w}_k)^{1/2}] \tag{3.21}$$

where $c_{kk'} = 1$ if the k-th block depends on the k'-th block, 0 otherwise.

This is a redundancy-related criterion in a multi-block framework, since for each pair of connected blocks it takes into account the variance of the dependent block as well as the correlation between the two components.

Further research will be carried out to find out if the algorithm optimizes a global criterion. Stability of the algorithm and coherence of the different steps are promising for the investigation of a global optimizing criteria of the procedure.

Chapter 4

Quantile Component-based Path Modeling: proposed methods, performances and interpretations

4.1 Introduction

Since PLS-PM algorithm is a procedure based on simple and multiple ordinary least squares (OLS) regressions, the obtained coefficients measure the rates of change in the mean of the dependent variable distributions as a function of changes in a set of predictor variables. However, in some case, classical OLS regression can give an incomplete picture of the relationships between variables. The single regression coefficient may not be the same along all the dependent variable distribution, and focusing exclusively on changes in the means may underestimate, overestimate, or fail to distinguish real nonzero coefficients.

This is also especially problematic in the case of heteroscedastic variances, when dependent variables are highly skewed (as it is typical in subjective measurements), in the presence of outliers, or when the interactions between the factors affecting the dependent variables are very complex and cannot all be measured and accounted for in a model.

All the factors that may affect an endogenous LV are usually not included in the models used to investigate relationships between LVs. As a consequence, there may be a weak or no dependence relationship between the mean of the endogenous LV distribution and the corresponding predictive LVs. However, there may be a stronger and useful dependence relationship with other parts of the response variable distribution. The same may happen in the dependence relationships between LVs and MVs.

In several applications it can be interesting to investigate dependence relationships between variables considering all parts of the response variable distributions. For example, in the business and market research, it can be interesting to evaluate if and how much the impact of consumer preferences on satisfaction is different among highly, medium or low satisfied customers with the objective of differentiating leverages to increase the satisfaction.

Quantile regression (QR) (Koenker and Basset, 1978) is an extension of the classical OLS regression for estimating functional relations between variables for all parts of the distribution of the response variable. Instead of the only estimation of conditional mean it allows the estimation of a set of conditional quantile functions, providing multiple slopes and a more complete picture of the relationships between variables. Compared to the OLS regression QR estimates are more robust against outliers.

In this perspective, Li et al. (2014) introduced a correlation measure to examine the linear relationships between any two variables for a given quantile, named quantile correlation (QC).

Quantile Composite-based Path Modelling (QC-PM) introduces both QR and QC in the classical PLS-PM algorithm (Davino and Esposito Vinzi, 2015; Davino et al., 2015a), and enhance PLS-PM potentialities when we wish to distinguish regressor effects on the different parts of the dependent variable distribution. As a matter of fact, QC-PM accommodates heteroscedastic variances and outliers and is able to explore the entire conditional distribution of the response variables.

QC-PM is advisable as a complementary analysis to the results deriving from a classical PLS-PM in the cases where there is no relationships (or only a weak relationship) between LVs or between LVs with their own MVs, even if the underlying theory would suggest the opposite. The exploration of different parts of the dependent variable distribution could highlight significant relationships. It could also be expected that the sign and the size of the path coefficients change if the analysis explores not only average effects but the entire conditional distribution of the response variables.

The proposed QC-PM is expected to be of interest in several real applications as the involved methodologies (PLSPM, QR and QC) have attracted researchers from various disciplines, for instance Economics (see Buchinsky, 1994; Fitzenberger et al., 2002; Hendricks and Koenker, 1992, among others), Social and behavioral sciences (see among many (see Davino and Vistocco, 2008; Eide and H.Showalter, 1998; Hsu et al., 2006; Kristensen and Eskildsen, 2010, among others), Sensory analysis (see Davino et al., 2015b; Guinot et al., 2001, among others), Marketing (see Hair et al., 2012b; Henseler et al., 2009; Whittaker et al., 2005, among others), Management of Information Systems (see Huarng, 2014; Ringle et al., 2012, among others), Strategic Management (see Hair et al., 2012a; Li, 2014, among others), Accounting (see Lee et al., 2011, among others).

4.2 Quantile Regression

QR was developed by Koenker and Basset (1978) as an extension of the classical OLS regression for estimating functional relations between variables for all parts of the distribution of the response variable. Instead of the only estimation of the conditional mean it allows the estimation of a set of conditional quantile functions, providing a more complete picture of the relationships between variables.

QR is a suitable solution when the homoschedastic assumption of the classical regression model can not be satisfied (for example because the variability of the dependent variable is not the same at every level of a regressor) or the dependent variable has skewed distribution (this event typically occurs in the evaluation of attitudes and preferences) or data are characterized by outliers (in many applicative contexts, it is often the extremes of the distribution that are most informative).

The estimates are semiparametric in the sense there is no parametric distributional assumption on the random errors of the model, but a parametric form is assumed for the deterministic part of the model.

For a given quantile θ, QR model can be formulated as follows:

$$Q_\theta\left(\hat{\boldsymbol{y}}|\boldsymbol{X}\right) = \boldsymbol{X}\hat{\beta}\left(\theta\right) \tag{4.1}$$

where $\boldsymbol{y}$ is the response variable observed on n individuals, $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{X}_p]$ is a matrix with p regressors and a vector of ones for the intercept estimation, $0 < \theta < 1$ and $Q_\theta(.|.)$ denotes the conditional quantile function for the θ^{th} quantile.

Although different functional forms can be used, here we consider functions of $\boldsymbol{X}$ that are linear in the parameters.

The conditional quantiles denoted by $Q_\theta\left(\hat{\boldsymbol{y}}|\boldsymbol{X}\right)$ are the inverse of the conditional cumulative distribution function of the response variable, $F_\theta^{-1}\left(\hat{\boldsymbol{y}}|\boldsymbol{X}\right)$, where $\theta \in [0, 1]$ denotes the quantiles (Koenker and Machado, 1999). For example, for $\theta = 0.6$, $Q_{0.6}\left(\hat{\boldsymbol{y}}|\boldsymbol{X}\right)$ is the 60th percentile of the distribution of $\boldsymbol{y}$ conditional on the values of a set of variables $\boldsymbol{X}$. Note that for symmetric distributions, the 0.50 quantile (or median) is equal to the mean.

Unconditional quantiles of a variable could be estimated by an optimization function minimizing a sum of weighted absolute deviations, where the weights are asymmetric functions of θ (Fox and Rubin, 1964; Koenker and Basset, 1978). In the same way, the conditional quantile estimator can be estimated as:

$$\hat{\beta}_\theta = argmin_{\beta_\theta} \sum_{i=1}^{n} \rho_\theta\big(y_i - \boldsymbol{x}_i'\beta(\theta)\big) \tag{4.2}$$

where ρ_θ is the so-called check function which weights positive and negative residuals asymmetrically, respectively with weights equal to $(1 - \theta)$ and θ.

For each quantile of interest, the solution of Equation 4.2 provides the related parameter estimates. It follows that, for each quantile, a regression line is estimated and, consequently, a fitted response vector can be obtained. The median regression is a special case of QR with equal weights for positive and negative errors which assures that there is the same number of observations above and below the median line (Koenker and Hallock, 2001).

Parameter estimates in linear quantile regression models have the same interpretation as those in any other linear model. The intercept measures the dependent variable value deriving from setting to zero all the regressors. Each slope coefficient is interpreted as the rates of change of the θth conditional quantile of the dependent variable distribution as a function of changes in a predictor. The parameters vary with θ due to effects of the θth quantile of the unknown error distribution.

Regression quantiles, retain their statistical properties under any linear or nonlinear monotonic transformation of $\boldsymbol{y}$ as a consequence of this ordering property, thus, they are equivariant under monotonic transformation of $\boldsymbol{y}$ (Koenker and Machado, 1999).

The most widespread algorithm for the estimation of the model parameters is the one proposed by Koenker and d'Orey (2001) as a version of the Barrodale and Roberts (1974) simplex algorithm. Although it is theoretically possible to extract infinite quantiles, a finite number is numerically distinct in practice, which is known as the quantile process. A fairly accurate approximation of the whole quantile process can be obtained using a dense grid of equally spaced quantiles in the unit interval $[0, 1]$ (Davino et al., 2013).

QR estimators are asymptotically normally distributed with different forms of the covariance matrix depending on the model assumptions (independent and identically distributed errors or non-identically distributed errors) (Koenker and Basset, 1978, 1982a,b). Resampling methods can represent a valid alternative to the asymptotic inference (Efron and Tibshirani, 1993) because they allow the estimation of parameter standard errors without requiring any assumption in relation to the error distribution. Several bootstrap procedures have been proposed in the QR framework. The simplest and widespread is the xy-pair method or design matrix bootstrap (Parzen et al., 1994). It consists of constructing a given number of samples (B), usually with the same size of the original dataset, where each sample is obtained by a random sampling procedure with replacement from the original dataset. The resampling procedure is simultaneously applied to the regressors and to the dependent variable. B quantile regressions are performed on the bootstrap samples and a vector (or a matrix in case of a multiple regression) of the parameter estimates is retained for each quantile of interest (Davino et al., 2013). The model parameters are estimated through the average of the bootstrap values. The standard error of the vector of parameter bootstrap estimates represents an estimate of the quantile regression standard error useful in confidence intervals and hypothesis tests.

Generally, in quantile regression sampling variation differs among quantiles, and it is usually larger as the value of θ approaches 0 or 1, thus, estimates further from the 50th conditional percentile usually cannot be estimated as precisely. In this case it would be more appropriate to use extreme value testing theory than conventional testing approaches (Chernozhukov and Umantsev, 2001).

The assessment of goodness of fit of QR model is based on the following the idea of the typical R^2 in classical regression analysis. The most common goodness of fit index in the QR framework, is called $pseudoR^2$ citepKoeMac99. For each considered quantile θ, it can be computed a residual absolute sum of weighted differences using the selected model (RASW) (corresponding to the residual sum of squares in classical regression) and a residual absolute sum of weighted differences (TASW) (corresponding to the total sum of squares of the dependent variable in classical regression) using a model with only the intercept (Davino et al., 2013; Hao and Naiman, 2007). Let us consider the simplest regression model with one explanatory variable:

$$Q_\theta(\hat{\boldsymbol{y}}|\boldsymbol{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\boldsymbol{x}. \tag{4.3}$$

For each considered quantile θ, $RASW$ is the corresponding minimizer:

$$\begin{aligned} RASW\,(\theta) = & \sum_{y_i \geq \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} \theta \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right| + \\ & \sum_{y_i < \hat{\beta}_0(\theta)+\hat{\beta}_1(\theta)x_i} (1-\theta) \left| y_i - \hat{\beta}_0(\theta) - \hat{\beta}_1(\theta)x_i \right| \end{aligned} \tag{4.4}$$

where ρ_θ is the so-called check function which weights positive and negative residuals asymmetrically, respectively with weights equal to $(1-\theta)$ and θ.

The $TASW$ is:

$$TASW\,(\theta) = \sum_{y_i \geq \theta} \theta \left| y_i - \hat{\theta} \right| + \sum_{y_i < \theta} (1-\theta) \left| y_i - \hat{\theta} \right|. \tag{4.5}$$

and the obtained $pseudoR^2$ can be computed as follows:

$$pseudoR^2\,(\theta)\,(\boldsymbol{y}, \boldsymbol{x}) = 1 - \frac{RASW\,(\theta)}{TASW\,(\theta)}. \tag{4.6}$$

As $RASW\,(\theta)$ is always less than $TASW\,(\theta)$, the $pseudoR^2\,(\theta)$ ranges between 0 and 1. It is worth noticing that the $pseudoR^2$ the index cannot be considered a measure of the goodness of fit of the whole model because it is related to a given quantile. For each considered quantile, the corresponding $pseudoR^2$ can be evaluated at a local level, thereby indicating whether the presence of the covariates influences the considered conditioned quantile of the response variable.

4.3 Quantile Correlation

In the quantile framework, Li et al. (2014) introduced a correlation measure to examine the linear linear relationships between any two variables for a given quantile $\theta \in (0, 1)$, named quantile correlation (QC). The authors claimed that QC can be used as broadly as the classical correlation in various contexts.

Like the Pearson correlation coefficient, QC is defined just as the ratio between a covariance measure and the the squared root of the product between a measure of variability of the two variables.

For $0 < \theta < 1$, quantile covariance is defined as:

$$\begin{aligned} qcov_{\theta}\left\{\boldsymbol{y}, \boldsymbol{x}\right\} = & \quad cov\left\{I\left(Y - Q_{\theta}(\boldsymbol{y}) > 0\right), \boldsymbol{x}\right\} \\ = & \quad E\left\{\psi_{\theta}\left[\boldsymbol{y} - Q_{\theta}\left(\boldsymbol{y}\right)\right]\left[\boldsymbol{x} - E\left(\boldsymbol{x}\right)\right]\right\} \end{aligned} \tag{4.7}$$

where $Q_{\theta}\left(\boldsymbol{y}\right)$ is the θ^{th} unconditional quantile of $\boldsymbol{y}$, $I\left(\cdot\right)$ is the indicator function and $\psi_{\theta}(u) = \theta - I(u < 0)$.

It follows that the QC can be defined as:

$$qcor_{\theta}\left\{\boldsymbol{y}, \boldsymbol{x}\right\} = \frac{qcov_{\theta}\left\{\boldsymbol{y}, \boldsymbol{x}\right\}}{\sqrt{var\left\{\psi_{\theta}\left[\boldsymbol{y} - Q_{\theta}\left(\boldsymbol{y}\right)\right]\right\} var\left(\boldsymbol{x}\right)}} = \frac{qcov_{\theta}\left\{\boldsymbol{y}, \boldsymbol{x}\right\}}{\sqrt{\left(\theta - \theta^2\right) var\left(\boldsymbol{x}\right)}} \tag{4.8}$$

QC has the same properties as Pearson correlation coefficient. It increases with the slope of the simple linear regression and it lies between -1 and 1. However, It is noteworthy that QC does not enjoy the symmetry property of the classical correlation index. For this reason, it is necessary to identify the role played by the two variables. In Equation 4.8 the $\boldsymbol{y}$ variable is the dependent variable.

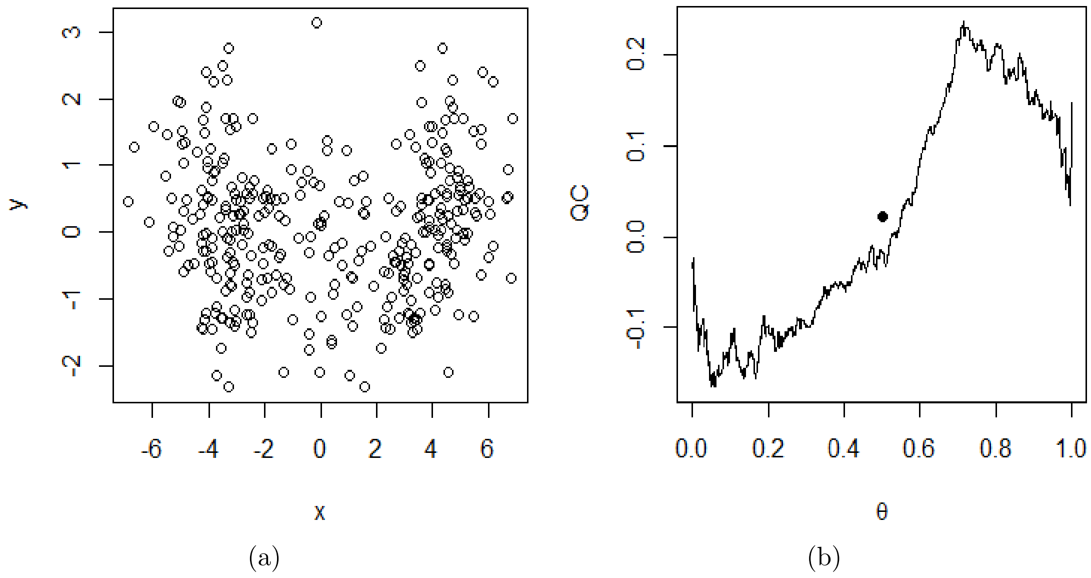
To evaluate the significance of QC, the following distribution convergence can be exploited (Li et al., 2014):

$$\sqrt{n}\left(\widehat{qcor}_{\theta}\left\{\boldsymbol{y},\boldsymbol{x}\right\}-qcor_{\theta}\left\{\boldsymbol{y},\boldsymbol{x}\right\}\right)\rightarrow N(0,\Omega_1) \tag{4.9}$$

where $\widehat{qcor}$ is the sample QC and Ω_1 the asymptotic variance.

As the estimation of Ω_1 is rather complex, a bootstrap approach is proposed for this purpose. The xy-method (Efron and Tibshirani, 1993) used for QR standard error estimates can be also applied in case of QC. Once the B samples have been generated, QC is computed on each sample and the the standard error of the bootstrap QC vector represent an estimate of the QC standard error.

To appreciate strengths and weaknesses of QC, an example based on synthetic data is provided. Let us consider two variables with a low Pearson correlation coefficient equal to 0.022. Figure 4.1a shows the scatter plot of these two variables. Using a dense grid of quantiles (from 0.001 to 0.999 with a step equal to 0.001), Figure 4.1b depicts the trend of the QC values across quantiles. The full circle represents the value of the Pearson correlation coefficient. For the sake of interpretation it was vertically aligned to the median.



(a)

(b)

FIGURE 4.1: A scatter plot of synthetic data (a) and QC values for a set of selected quantiles (b)

As it is shown in Figure 4.1, even though the Pearson correlation coefficients between the two variables is close to zero, the relations between the two variables changes when exploring other parts of the dependent variable distribution. In particular, considering the part of the distribution on the left of the median, the quantile correlations is negative, while it is positive on the right of the median. These two opposite relationships balance out and the Pearson correlation coefficients turn out to be close to zero. This is the case where the investigation of all parts of the response variable distributions gives more interesting information about the relationships between the two variables.

4.4 Quantile Composite-based Path Modelling

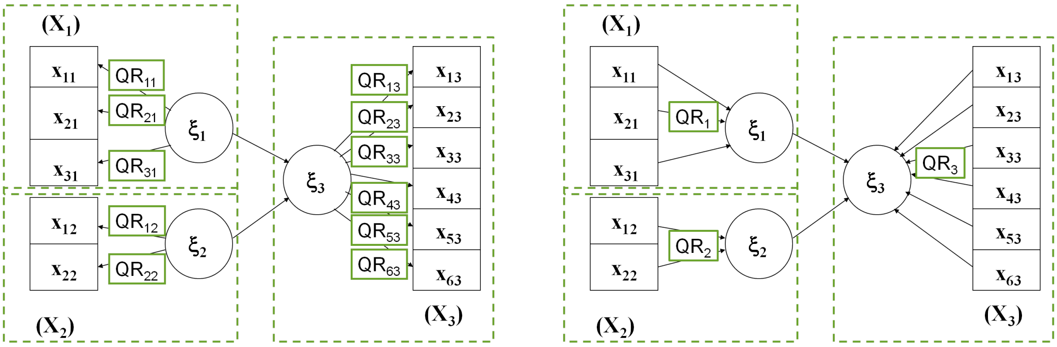
Quantile Composite-based Path Modelling (QC-PM) introduces both QR and QC in the classical PLS-PM algorithm (Davino and Esposito Vinzi, 2015; Davino et al., 2015a).

A very basic approach consists in exploiting QR potentialities a posteriori, after the convergence of a classical PLS-PM and once the LVs scores are estimated. For each quantile θ of interest, QR can be introduced to estimate the path coefficients measuring the impact of LVs on the whole distribution of the endogenous LVs. This quantile approach to PLS-PM can be considered quite basic as it is applied on LVs derived from OLS regressions. However it can be still of interest when we would like to investigate whether differences occur only for path coefficients across quantiles, regardless the measurement model.

A more complex and powerful method is obtained introducing either QR and QC in all the outer and inner estimation steps of the algorithm as well as in the estimation of path coefficients loadings. Hence, for each quantile of interest θ we have estimates for the all model parameters.

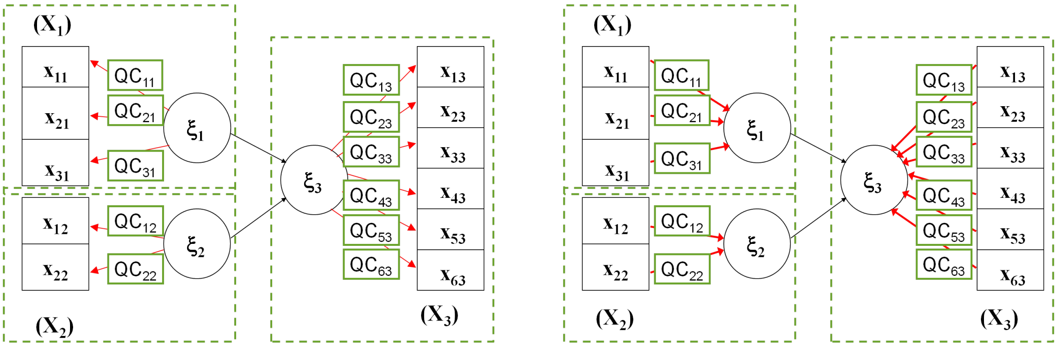
According to the scheme adopted in the various estimation steps, different versions of the QC-PM are available.

In the outer estimation, simple (*Mode A*) or multiple (*Mode B*) QR allows to compute outer weights for each quantile of interest (see Figure 4.2). A new mode (named *Mode Q*) is introduced in the outer estimation. In *Mode Q* weights are obtained by computing QC between LVs and their own MVs (Davino and Esposito Vinzi, 2015). Since QC is an asymmetric correlation coefficient, *Mode Q* allows us to handle both outwards-directed and inwards-directed measurement models (see Figure 4.3).



(a) (b)

FIGURE 4.2: Outer model schemes: *Mode A* (a) and *Mode B* (b)



(a) (b)

FIGURE 4.3: Outer model schemes: *Mode Q* outwards-directed (a) and *Mode Q* inwards-directed (b)

In the inner estimation step, inner weights are computing depending on the adopted weighting scheme. If the path weighting scheme is chosen, the inner weights linking the m^{th} endogenous LV to its predecessors are computed through a QR:

$$Q_\theta\left(\hat{\xi}_m|\Xi_{\rightarrow m}\right) = \Xi_{\rightarrow m}\hat{\beta}\left(\theta\right) \tag{4.10}$$

where $\Xi_{\rightarrow m}$ is the matrix of the ξ_m's predecessor LVs. Instead, the weights among the m^{th} LV and its successor LVs are determined using the QC proposed by Li et al. (2014). Since in the quantile framework even the correlation is a non symmetric measure, the use of QC distinguishes between predecessors and successors. Let ξ_m and $\xi_{q\rightarrow m}$ be respectively a LV and one of its predecessor LVs, the former plays the role of the dependent variable and the latter is the regressor. The QC proposed by Li et al. (2014) and adapted in the QC-PM framework, is defined as:

$$qcor_\theta = \frac{qcov_\theta\left\{\xi_m, \xi_{q\rightarrow m}\right\}}{\sqrt{\left(\theta-\theta^2\right)var\left(\xi_{q\rightarrow m}\right)}} \tag{4.11}$$

where $qcov_\theta\left\{\xi_m, \xi_{q\to m}\right\} = cov\left\{I\left(\xi_{q\to m} - Q_\theta(\xi_{q\to m}) > 0\right), \xi_m\right\}$, $Q_\theta(\cdot)$ is the θ^{th} unconditional quantile and $I\left(\cdot\right)$ is the indicator function.

QC is also proposed as an alternative to the Pearson correlation coefficient if either the centroid or the factorial scheme is adopted.

Once convergence is reached and LV scores are computed, path coefficients are estimated by means of quantile regressions.

Table 4.1 shows twelve proposed QC-PM deriving from the combination of outer and inner schemes. The last column and last row refer to the methodology used in the outer estimation mode and the inner estimation scheme, respectively. Mode Q is the previously defined new option for updating the outer weights.

		Inner Scheme			
		Path Weighting	Factorial	Centroid	Methodology
Outer Mode	Outwards	QCPM1	QCPM2	QCPM3	Simple QR
	Inwards	QCPM4	QCPM5	QCPM6	Multiple QR
	Mode Q - Outwards	QCPM7	QCPM8	QCPM9	QC
	Mode Q - Inwards	QCPM10	QCPM11	QCPM12	QC
	Methodology	QR & QC	QC	QC sign	

Table 4.1: The different estimation options for QC-PM (QR=Quantile Regression, QC=Quantile Correlation)

Some preliminary simulation studies revealed that the use of the path weighting scheme yields convergence problems in case of low correlations within and/or between blocks. Further studies will be devoted in future to this issue. In the following we will not consider QC-PM with the path weighting scheme.

Like PLS-PM, the assessment of the quality of the QC-PM results should take different aspects into account. The quality of the model depends on the goodness of fit of both the outer and the inner models. Moreover, the evaluation of the statistical significance of the coefficients should be carried out.

The assessment of QC-PM is performed exploiting the main indexes proposed in PLS-PM (Davino et al., 2015a): communality and average communality, multiple linear determination coefficient (R^2), redundancy index, average redundancy index and global criterion of goodness of fit (GoF). It is worth noticing that QC-PM is estimated for each quantile θ of interest thus it provides a set of assessment measures for each estimated model.

In PLS-PM, the communality index measures the amount of the variability of a MV explained by its LV, and it is obtained as the square of the correlation between each MV and its LV. Therefore, for a generic $\boldsymbol{x}_{pq}$ MV belonging to the q_{th} block, the communality is equivalent to the R^2 of the simple regression $\boldsymbol{x}_{pq} = \alpha_0 + \alpha_1 \xi_q$. In the quantile framework, we can exploit the $pseudoR^2$ index (Koenker and Machado, 1999), as defined above.

For a generic $\boldsymbol{x}_{pq}$ MV of the q_{th} block and a quantile θ of interest, the communality expresses the quality of each simple regression $\boldsymbol{x}_{pq} = \alpha_0 + \alpha_1 \xi_q$, at the specific quantile, in terms of weighted residuals and can be defined as:

$$Com_{pq}(\theta) = pseudoR^2\left(\theta\right)\left(\boldsymbol{x}_{pq}, \xi_q\right) \tag{4.12}$$

The model assessment can also be carried out for the generic q_{th} block with p_q MVs as:

$$Com_q(\theta) = \frac{1}{p_q} \sum_{p=1}^{p_q} pseudoR^2\left(\theta\right)\left(\boldsymbol{x}_{pq}, \xi_q\right) \tag{4.13}$$

or for the whole measurement part of the model ($\overline{Com}$) through averages respectively of the communalities related to the block and to all the MVs (weighted by the number of MVs in each block):

$$\overline{Com}(\theta) = \frac{1}{\sum_q p_q} \sum_q p_q Com_q(\theta) \tag{4.14}$$

With respect to the structural model, the $pseudoR^2$ index can be computed for each structural equation and each of them measures the amount of variability of a given endogenous LV explained by its predecessor LVs. The average of all the $pseudoR^2$ indexes ($\overline{pseudoR^2}(\theta)$) provides a synthesis of the evaluations regarding the structural part of the model.

Another important measure is the redundancy because it is able to take into account also the contribution of the MVs related to the q_{th} endogenous LV, that is linking the prediction performance of the measurement model to the structural one (Amato et al., 2005). In the QC-PM framework the redundancy of a generic

q_{th} endogenous LV is proposed as:

$$Red_q(\theta) = Com_q(\theta) \times pseudoR^2(\theta)(\hat{\xi}_q; \hat{\Xi}_{\rightarrow} q) \tag{4.15}$$

where $\hat{\Xi}_{\rightarrow} q$ is the matrix of the predictor LVs for the q^{th} LV.

An overall assessment of the quality of the structural part is provided by the average redundancy ($\overline{Red}(\theta)$) obtained as a mean of the redundancies associated to the set of endogenous LVs.

With respect to the goodness-of-fit of the general model, following the global goodness-of-fit index, the GoF, proposed by (Amato et al., 2005), in QC-PM the absolute GoF is obtained as geometric mean of the average communality and the average

$pseudoR^2$:

$$GoF(\theta) = \sqrt{\overline{Com}(\theta) \times \overline{pseudoR^2}(\theta)} \tag{4.16}$$

The first and the second term in Equation 4.16 measure the predictive performance respectively of the measurement and the structural model (Amato et al., 2005) (Esposito Vinzi et al., 2008). GoF is able to take both the measurement and the structural part of the model into account.

Further developments will regard the exploration of different goodness of fit measure in the quantile framework and the adjustment to the QPLS-PM of further assessment indexes proposed in PLS-PM framework (Henseler et al., 2009) (e.g. the average variance extracted (Fornell and Larcker, 1981), the Stone-Geisser's Q^2 using blindfolding procedures (Stone, 1974), the relative GoF (Amato et al., 2005)).

The evaluation of the statistical significance of the coefficients related to the different quantiles can be carried out exploiting the asymptotically normal distribution of the QR estimators as well as the bootstrap approach classically used in PLS-PM and QR.

A bootstrap approach can also applied to obtain a variability measure of the quantile correlation estimates obtained choosing Mode Q in the measurement model and/or factorial or centroid scheme in the structural model.

In future work, a jackknife approach could be explored especially in case of small samples to estimate the standard errors of the parameter estimators and statistical tests could be introduced in a QC-PM to to test if coefficients across quantiles can be considered statistically different Gould (1997).

In the following sections, the functioning and performance of the QC-PM are shown through a real data application in the area of the American Customer Satisfaction Index (ACSI) and through a simulation study.

Since the results of the different QC-PM options are not much different among them, in the real data application we will show the results only for the QC-PM9. As regards to the simulation study we will compare the results of the classic PLS-PM with the basic approach that apply QR using the scores of the classical PLS-PM, referred as QC-PMØ, the QC-PM3 and the QC-PM9.

4.5 A real data application

The proposed method is applied to a real dataset in the area of the ACSI (ACSI, 2000; Anderson and Fornell, 2000) [1]. The results presented in this section are mainly based on the paper by Davino et al. (2015a).

This index was established in 1994 and it is the only national cross-industry measure of customer satisfaction in the United States. The index measures the satisfaction of U.S. household consumers with the quality of products and services offered by both foreign and domestic firms with significant share in U.S. markets.

The real data application refers to the food processing sector including 1617 observations. The customer satisfaction is driven by three factors (customer expectations, perceived value and perceived quality) and has loyalty as outcome. The complaints LV has been excluded because the number of complaints was very small (1%). The relationships among the five LVs are represented in the path diagram in Figure 4.4. Each LV is measured through a set of MVs measured on a scale 1-10 (see Table 4.2).

A preliminary analysis of the MV distribution right tails is advisable before applying QC-PM because data deriving from customer satisfaction surveys are often

[1] http://www.theacsi.org/

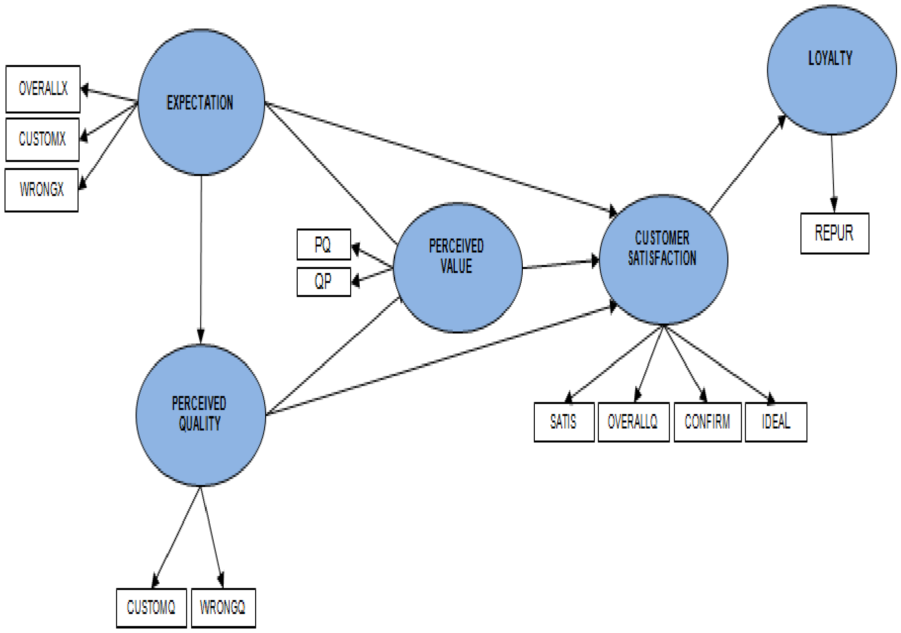


FIGURE 4.4: Path diagram of the ACSI model

LV	MV	Label	Mean	θ=0.1	θ=0.25	θ=0.41	θ=0.5	θ=0.75	θ=0.9
Customer Expectations	Expectations about overall quality	OVERALLX	8	6	8	8	9	10	10
	Expectations about customization	CUSTOMX	9	7	8	9	9	10	10
	Expectation about reliability	WRONGX	8	3	7	9	9	10	10
Perceived Quality	Meeting personal requirements	CUSTOMQ	9	7	8	9	9	10	10
	Things went wrong	WRONGQ	9	6	9	10	10	10	10
Perceived Value	Price given Quality	PQ	8	5	7	7	8	9	10
	Quality given Price	QP	8	6	7	8	8	9	10
Customer Satisfaction	Customer Satisfaction	SATIS	9	7	8	9	9	10	10
	Overall Quality	OVERALLQ	9	7	8	9	9	10	10
	Confirmation to Expectations	CONFIRM	8	5	6	8	8	9	10
	Close to ideal product/service	IDEAL	8	5	7	8	8	9	10
Customer Loyalty	Repurchase Intention	REPUR	8	6	8	9	9	10	10

TABLE 4.2: LVs and MVs of ACSI dataset: Means and main quantile values

characterised by a very high concentration of the responses on the upper values or even the maximum of the used scales. The deriving effect is an absence of variability in a given part of the distribution which is not interesting to explore. This information is highlighted by computing the quantile values (Table 4.2).

In Figure 4.5, the distribution of the maximum quantile value for each MV is shown. In the ACSI dataset all the MVs show a considerable percentage of customers expressing an evaluation equal to 10. We notice, for example, that it is not interesting to explore the variable WRONGQ from the 0.41 quantile forward because all the quantile values will be equal to 10.

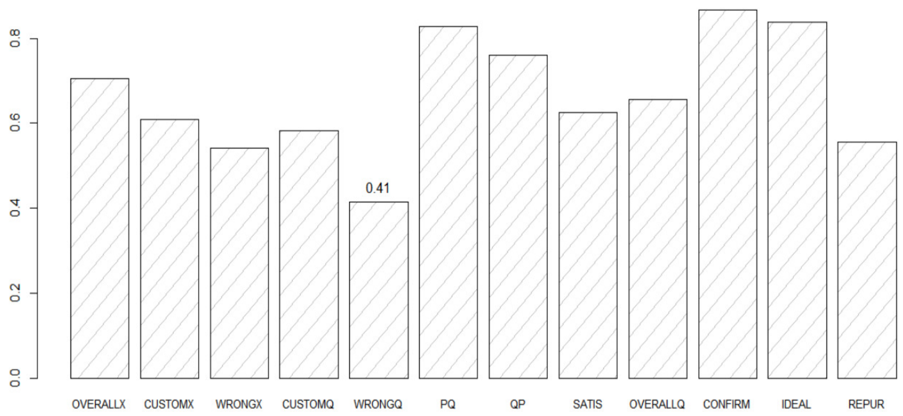


FIGURE 4.5: Maximum quantile value for each MV

Even if the maximum quantile value is different for each MV, QC-PM cannot be performed beyond the minimum threshold quantile which corresponds to 0.41, as for each quantile of interest QC-PM applies regression models simultaneously for all the equations of the model. The requirement to confine the analysis at a lowest maximum quantile value is the most of the case is not a limit, because QC-PM aims at the exploration of the different parts of the dependent variables distribution when they are characterised by different and not constant effects of the regressors. Moreover, in this case, for example, even if the analysis is restricted up to the quantile 0.41, we are still able to model a portion of very satisfied customers as shown in Table 4.2.

For this application we apply QC-PM using the factorial scheme in the inner estimation and the outwards-directed relationship in the outer estimation using the QC (Mode Q). We obtain also the PLS-PM results using the factorial scheme and Mode A.

Table 4.3 shows the obtained outer weights for a selected grid of quantile of interest ($\theta = [0.1, 0.25, 0.41]$). Outer weights that evidently differ across quantiles are in bold.

The differences in the weights across quantiles can be also appreciated using a graphical representation. Figure 4.6 depicts, for the *Expectation* LV, the PLS-PM and QC-PM outer weights with respect to the average values of the corresponding MVs. Labels 10, 25 and 41 refer to QC-PM weights for quantiles equal to 0.10, 0.25 and 0.41, respectively. PLS-PM weights are pointed out with the MV names. QC-PM and PLS-PM weights related to the same MV are vertically aligned with

		Outer weights			
LV	MV	PLSPM	θ=0.1	θ=0.25	θ=0.41
Expectation	OVERALLX	0.478	**0.436**	**0.487**	**0.542**
	CUSTOMX	0.575	**0.674**	**0.532**	**0.471**
	WRONGX	0.228	**0.054**	**0.300**	**0.315**
Quality	CUSTOMQ	0.811	**0.820**	**0.767**	**0.713**
	WRONGQ	0.352	**0.340**	**0.413**	**0.481**
Value	PQ	0.459	0.451	0.477	0.493
	QP	0.633	0.641	0.616	0.602
Satisfaction	SATIS	0.374	0.390	0.375	0.326
	OVERALLQ	0.373	0.377	0.367	0.312
	CONFIRM	0.249	0.221	0.225	0.309
	IDEAL	0.256	0.258	0.286	0.323

Table 4.3: Outer weights

respect to the MV average. According to the PLS-PM results, it is not possible to identify how to improve satisfaction because, for example, WRONGX shows the lowest average values but also the lowest weight. QC-PM complements such a result suggesting that an improvement of WRONGX has a higher impact on the most satisfied customers. As regards to CUSTOMX, the impact of an improvement is more evident on the less satisfied customers.

Table 4.4 shows both the PLS-PM path coefficients and the QC-PM path coefficients for the selected grid of quantile of interest. Path coefficients that evidently differ across quantiles are in bold.

		Path Coefficients			
LV		PLS-PM	θ=0.1	θ=0.25	θ=0.41
Quality	Expectation	0.585	0.748	0.823	0.713
Value	Expectation	0.174	0.154	0.162	0.214
	Quality	0.401	0.479	0.434	0.409
Satisfaction	Expectation	0.252	0.253	0.259	0.238
	Quality	0.435	**0.520**	**0.434**	**0.389**
	Value	0.328	0.342	0.364	0.373
Loyalty	Satisfaction	0.604	**0.903**	**0.828**	**0.687**

Table 4.4: PLS-PM path coefficients and QC-PM path coefficients for a set of quantiles ($\theta = [0.1, 0.25, 0.41]$)

A graphical representation of path coefficients better highlights the differences among PLS-PM and QC-PM results and among QC-PM path coefficients at different quantiles. Figure 4.7 shows the estimated path coefficients of the *Customer Satisfaction* LV across quantiles. Full circles refer to the PLS-PM path coefficients

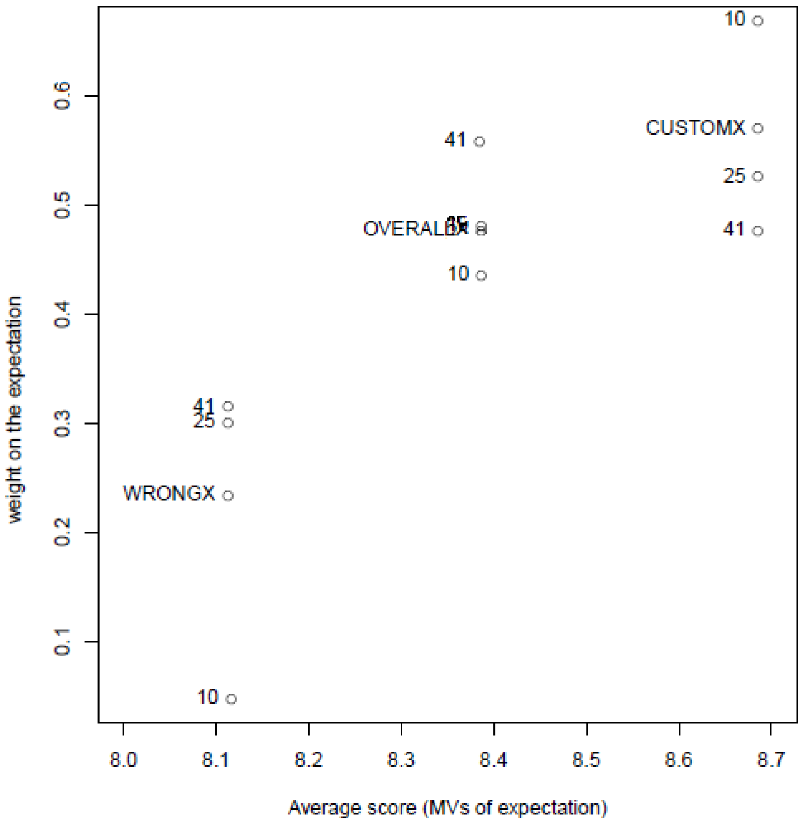


FIGURE 4.6: Outer weights with respect to the MV averages of the *Expectation* LV

while stars represent QC-PM path significant coefficients for each quantile of interest. For the sake of interpretation, PLS-PM results are vertically alligned to the last considered quantile (0.41). It is worth noting that path coefficients vary in the extreme parts of the distribution, meaning that the impact of a given LV changes for either very low and very high satisfied customers. For example, considering the *Quality* LV, its effect decreases moving from the first 10% of the distribution to the last considered quantile.

In order to evaluate the quality of the model, at first, we consider the QC between MVs and LVs. The results are expected to show higher correlations between a LV with its own block of MVs than with other LVs representing different blocks of MV (cross-correlations). The underlying concept of each LV should differ from the other theoretical concepts. In Table 4.5 PLS-PM and QC-PM correlations between MVs and LVs are shown. QC-PM correlations are computed as QC where MVs play the role of dependent variable and LVs the role of explanatory.

The results are satisfactory for all the LVs (for the sake of brevity cross-correlations

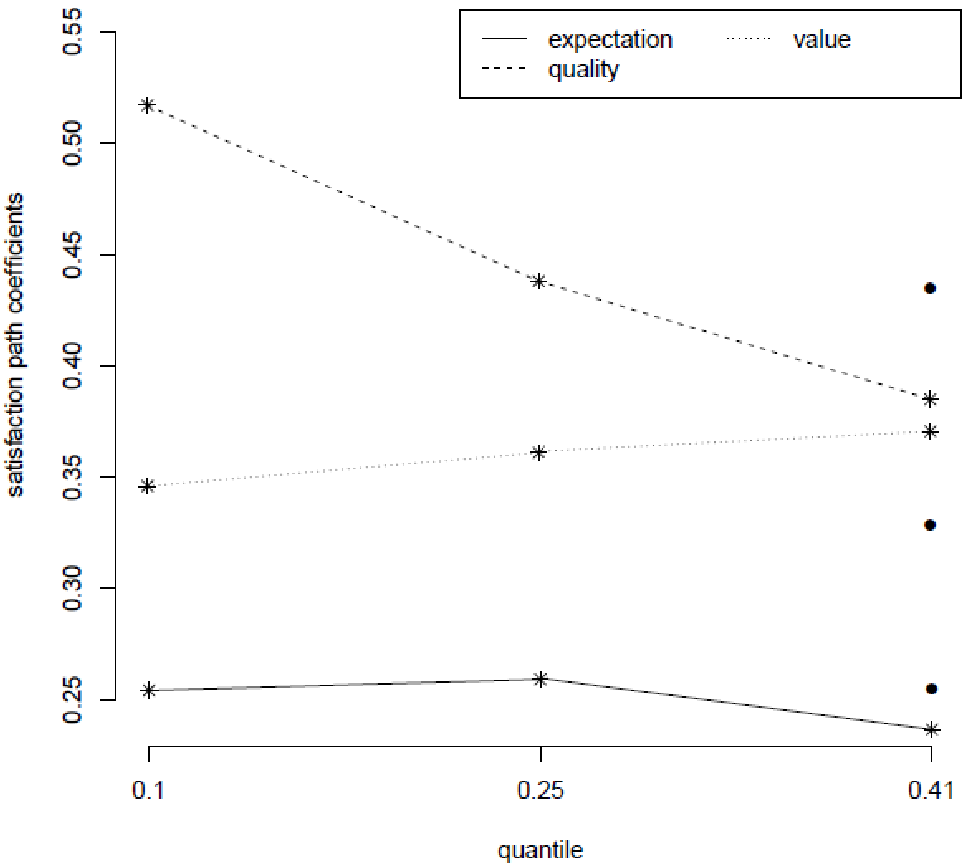


FIGURE 4.7: QC-PM path coefficients for a set of quantiles

		Correlations LV-MVs			
LV	MV	PLS-PM	θ=0.1	θ=0.25	θ=0.41
Expectation	OVERALLX	0.825	0.652	0.770	0.689
	CUSTOMX	0.894	0.908	0.804	0.589
	WRONGX	0.401	0.083	0.491	0.406
Quality	CUSTOMQ	0.945	0.849	0.796	0.595
	WRONGQ	0.661	0.688	0.647	0.593
Value	PQ	0.883	0.763	0.832	0.736
	QP	0.938	0.864	0.793	0.734
Satisfaction	SATIS	0.877	0.884	0.803	0.543
	OVERALLQ	0.846	0.816	0.775	0.509
	CONFIRM	0.697	0.640	0.568	0.629
	IDEAL	0.714	0.712	0.715	0.636

TABLE 4.5: Correlations and Quantile Correlations LV-MVs

are not shown but they are in all cases lower than the correlations). It is worth noting the change of the correlation values across quantiles. For example, the correlation of CUSTOMX to the Expectation LV is higher in the lower part of the distribution (θ=0.1) and even greater than the PLS-PM loading. Considering the

		Communality			
LV	MV	PLS-PM	θ=0.1	θ=0.25	θ=0.41
Expectation	OVERALLX	0.680	0.503	0.494	*0.520*
	CUSTOMX	0.799	0.759	0.639	0.563
	WRONGX	0.161	0.016	0.232	0.209
	$Com_{Expectation}$	0.546	0.426	0.455	0.431
Quality	CUSTOMQ	0.892	0.851	0.768	0.670
	WRONGQ	0.438	0.550	0.464	0.450
	$Com_{Quality}$	0.665	0.701	0.616	0.560
Value	PQ	0.779	0.516	0.587	0.616
	QP	0.881	0.749	0.774	0.731
	Com_{Value}	0.830	0.632	0.681	0.674
Satisfaction	SATIS	0.768	0.617	0.590	0.515
	OVERALLQ	0.716	0.537	0.533	0.462
	CONFIRM	0.486	0.235	0.328	0.385
	IDEAL	0.510	0.356	0.381	0.402
	$Com_{Satisfaction}$	0.620	0.436	0.458	0.441
$\overline{Com}$		0.646	0.517	0.526	0.502

TABLE 4.6: Measurement model assessment indexes

MV WRONGX the correlation is almost equal to zero for a quantile equal to 0.1 while it increases as the quantile increases.

To evaluate the quality of the measurement model communalities and average communalities are computed. Table 4.6 shows the PLS-PM and QC-PM communalities. We recommend not to compare QC-PM communalities to those of the PLS-PM, as they are based on different residuals. We can just compare QC-PM communalities among each other and across quantiles.

As for the quality of the structural model, Table 4.7 shows the PLS-PM R^2 and the QC-PM $pseudoR^2$.

Also for this measure, we recommend not to compare QC-PM $pseudoR^2$ to the PLS-PM R^2, as they are based on different residuals. It is well known that the typical determination index is not a satisfactory assessment index and it is generally smaller than the R^2 (Koenker and Machado, 1999). We can just compare QC-PM $pseudoR^2$ among each other and across quantiles.

Table 4.8 shows the PLS-PM redundancy and the QC-PM redundancy.

	R^2		$pseudoR^2$	
LV	PLS-PM	θ=0.1	θ=0.25	θ=0.41
Quality	0.335	0.240	0.298	0.275
Value	0.250	0.180	0.181	0.153
Satisfaction	0.659	0.502	0.496	0.429
Loyalty	0.364	0.276	0.297	0.282
$MeanofR^2$	0.402	0.299	0.318	0.285

Table 4.7: R^2 and $pseudoR^2$ in the Structural model

			Redundancy		
LV	MV	PLS-PM	θ=0.1	θ=0.25	θ=0.41
GoF	CUSTOMQ	0.299	0.204	0.229	0.184
	WRONGQ	0.146	0.132	0.138	0.124
	$Red_{Quality}$	0.223	0.102	0.136	0.118

Table 4.8: Redundancy measures

An an overall assessment of the quality of general model, we computed the GoF. As it is shown in Table 4.9 the quality of the general model does not much differ across quantiles.

	PLS-PM	θ=0.1	θ=0.25	θ=0.41
GoF	0.510	0.393	0.409	0.378

Table 4.9: Goodness of fit (GoF) indices

4.6 Simulation Study

Due to the complexity of PLS-PM, and consequently of QC-PM, the analysis of its relative performance can hardly be assessed in an analytical form. This is the case where simulation studies come to our aid.

We will perform a simulation study organized in three different scenarios that aim at showing the functioning of QC-PM, studying the QC-PM capabilities in handling the cases where the relationships between variables change across quantiles in both the measurement model and the structural model.

In this simulation study we analyze the artificial data applying the classic PLS-PM, the basic approach that apply QR using the scores of the classical PLS-PM, referred as QC-PMØ, the QC-PM3 and the QC-PM9.

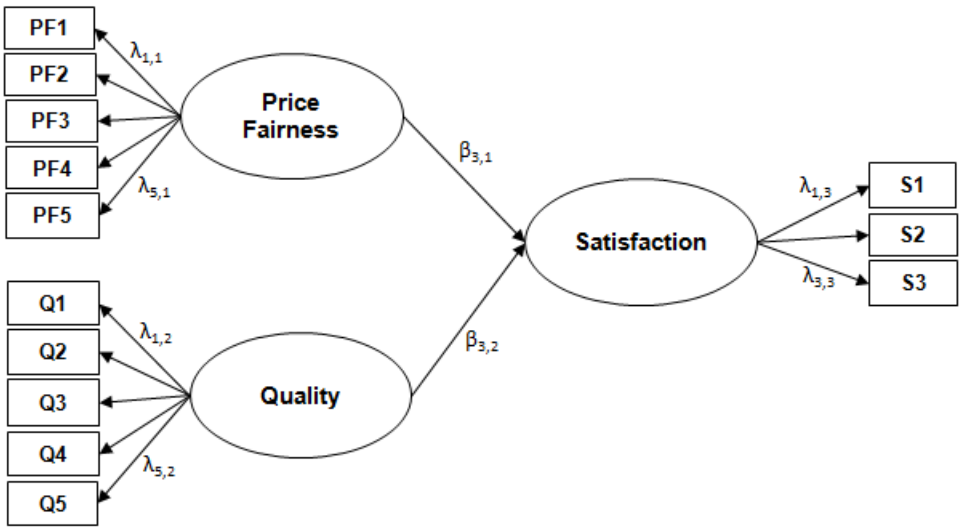


FIGURE 4.8: Theoretical Model for Simulated Data

4.6.1 Design of the simulation study

In order to facilitate the interpretation of the results and to simulate a real data application, we shall perform the study in the field of the customer satisfaction analysis. It will be considered a simple model which it was already used by Esposito Vinzi et al. (2007) and Trinchera (2007) for their simulation studies.

Customer satisfaction is the central variable of this model, having as antecedents or drivers the Price Fairness and Quality. Therefore, the SEM underpinning our design and subsequent analyses consists of one endogenous LV, Customer Satisfaction, and two exogenous LVs, Price Fairness and Quality. Each exogenous LV (i.e., Price Fairness and Quality) are measured by five MVs, and the endogenous LV, Customer Satisfaction, is measured by three MVs (see Figure 4.8). All the blocks are considered as reflective.

In marketing applied research it can be interesting to verify if the effects of consumer preferences on satisfaction differ across different parts of the distribution of the satisfaction variable. As a matter of fact, the impact of consumer preferences on satisfaction may vary as the degree of satisfaction changes. It is very likely that the preferences of satisfied customers are different compared to the preferences of unsatisfied customers. If we have this information we could differentiate leverages to increase satisfaction. This heterogeneity frequently occurs when variables are highly skewed, and this is a typical characteristic of data collected in behavioral research.

The simulation study is organized in three different scenarios. In the first scenario data are generated assuming homogeneity, hence the effects of the variables do not differs across quantiles of the dependent variable distributions. In the second scenario we assume heterogeneity in the structural model. In particular, the path coefficients differ across quantiles of the endogenous LV distribution. In the third scenario we assume heterogeneity in the measurement model. The relationships between MVs and LVs differ across quantiles of the variables distribution.

We intentionally chose the simple model in Figure 4.8 for the simulation study, as the process to generate sample of data from two or more different populations (i.e., customers with different degrees of satisfaction) is complicate, and it would be difficult to control all the factors that can have severe effects on the results in a complex model.

The data generation process is based on the classical covariance-based approach for SEM and it is consistent with the procedure described by Paxton et al. (2001) for a Monte Carlo simulation study for SEM. Once all the model parameters of the SEM are pre-specified, the implied covariance matrix is obtained from the given parameter values, then data were draw from a multivariate distribution with that specific implied covariance matrix. Hence, we assumed that the model parameter values are known. The data generation process and the subsequent analysis were conducted by EQS for Windows and R.

In the next section, we shall present first the simulation study concerning the case of homogeneity in the relatioships between variables (see subsection 4.6.2). Then, QC-PM performance in handling heterogeneity in the relationships of the structural model will be investigated in the second simulation study (see subsection 4.6.3). Finally, we will deal with the case of heterogeneity in the relationships of the measurement model in the third simulation study (cf. subsection 4.6.4).

4.6.2 The Case of Homogeneity

In the first simulation study we assume homogeneity in the model. The data are generated according to the model in Figure 4.8. We conducted the simulation setting different values for both $\boldsymbol{\beta}_{3,1}$ and $\boldsymbol{\beta}_{3,2}$, which are the path coefficients capturing the effect of Price Fairness on Customer Satisfaction and Quality on Customer Satisfaction, respectively. We assume that the relationship between

Price Fairness and Customer Satisfaction and the relationship between Quality and Customer Satisfaction is the same. In particular, we set six different values for the path coefficients ($\boldsymbol{\beta}_{3,1}$=$\boldsymbol{\beta}_{3,2}$ = 0.3, 0.4, 0.5, 0.6). In order to focus only on different levels of correlation between blocks, the loadings between LVs and the corresponding MVs were set all to 1.

Once all the other population parameter values were set, for each path coefficient value a total of 250 sets of multivariate normal data were drawn from a population with the model-implied covariance matrix $\boldsymbol{\Sigma}$ which is a complex function of the model parameters. Each data set has a sample size of 250, a common sample size usually used in marketing research to estimate the customer satisfaction.

In this case we expect that the effect of Price Fairness and Quality on Customer Satisfaction is homogeneous across quantile of Customer satisfaction distribution. For the all considered quantiles, the path coefficient estimates should be for a common parameter, and any deviation among them is simply due to sampling variation.

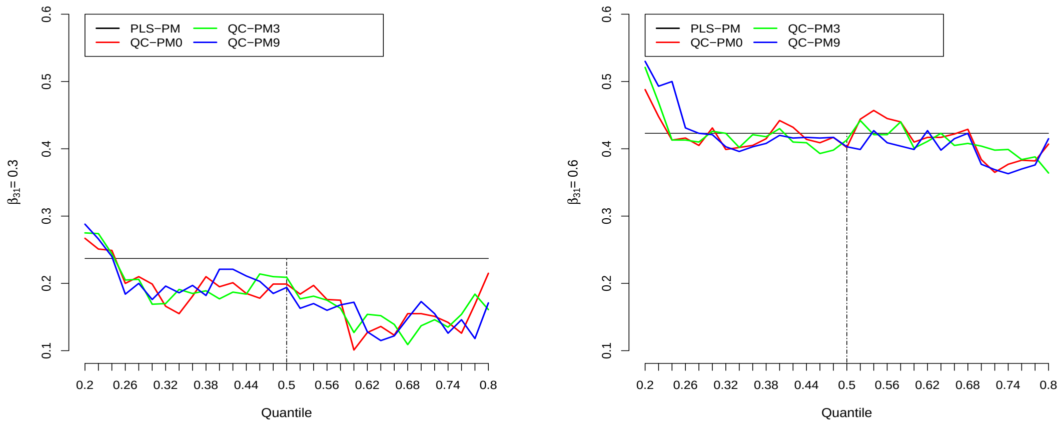
The path coefficient values across quantiles are investigated considering a dense grid of quantiles from 0.2 to 0.8.

Since the path coefficients value are set to be equal ($\boldsymbol{\beta}_{3,1}$=$\boldsymbol{\beta}_{3,2}$), we will show the results only for one path coefficient ($\boldsymbol{\beta}_{3,1}$) as they are almost the same to the results of the other path coefficient. Moreover, for the sake of simplicity we will show only the results for two values of path coefficient, $\boldsymbol{\beta}_{3,1} = 0.3$ and $\boldsymbol{\beta}_{3,1} = 0.6$. Showing the results for all path coefficients values would be redundant, since they showed no more interesting findings.

Figure 4.9 shows the the different QC-PM path coefficient estimates across quantiles as well as the PLS-PM path coefficient estimates.

As a first result, we see that PLS-PM path coefficients are very similar to QC-PM for $\boldsymbol{\theta}$=0.5. Note that for symmetric distributions, the 0.50 quantile (or median) is equal to the mean, thus the quantile regression coefficient is similar to the ordinary least squares regression coefficient. As a consequence we find the same results when comparing QC-PM and PLS-PM when variables distributions are symmetric.

In general, deviations among path coefficients across quantiles are not evident and may be simply due to sampling variation.



(a)

(b)

FIGURE 4.9: Path coefficient estimates across quantiles: $\beta_{31} = 0.3$ (a) and $\beta_{31} = 0.6$ (b)

However, future research must be focused on testing statistical significance of differences among QC-PM estimates at different quantiles, using resampling procedures such as bootstrapping methods (Efron and Tibshirani, 1993; Gould, 1997).

4.6.3 The Case of Heterogeneity in the Structural Model

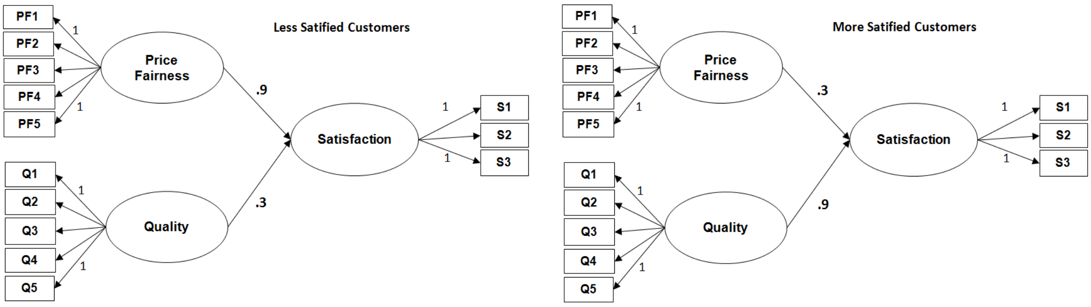
In the second simulation study we assume heterogeneity in the structural model. The exogenous LVs exert both a change in means and a change in variance on the distribution of endogenous LV, hence the path coefficients differ across quantiles.

In order to generate data whit this feature, we suppose that two different populations exist, and for each population the model parameters are different. In particular we divide the customers in two classes. The fist class is represented by the less satisfied customers, while the second class is represented by the more satisfied customers. The two classes of customers have different preferences, leading to two different models.

We suppose that the two groups of customers have the following characteristics:

- less satisfied customers - characterized by a strong relationship between Price Fairness and Customer Satisfaction and a weak relationship between Quality and Customer Satisfaction (see Figure 4.10a);

- more satisfied customers - characterized by a strong relationship between Quality and Customer Satisfaction and a weak relationship between Price Fairness and Customer Satisfaction (see Figure 4.10b).



(a) (b)

FIGURE 4.10: Theoretical Model for Simulated Data. Less satisfied customers (a) and more satisfied customers (b)

For the less satisfied customers, Price Fairness enhance satisfaction more than Quality, while for the more satisfied customers is just the opposite, Quality enhances satisfaction more than Price Fairness.

The simulation procedure was broken down into three steps. Firstly, data were drawn from a multivariate normal population, $\boldsymbol{X} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the implied population covariance matrix derived by the parameters of the model shown in the Figure 4.10a. We generated 250 data set, each one of sample size equal to 250. Given a specific MV of Customer Satisfaction block, for each data set we kept the observations until the 0.6 quantile of this MV, thus once the observations are sorted in non-decreasing order with respect to the values on this MV, we kept the first 60% of the observations (equivalent to 150 units). Then, the MVs of the endogenous block are transformed into new variables such that they take values between 1 and 6.

In the second step, data were drawn from a multivariate normal population, $\boldsymbol{X} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the implied population covariance matrix derived by the parameters of the model shown in the Figure 4.10b. The sample size is equal to 150. We generated 250 date set, each one of sample size equal to 250. For each data set, once the observations are sorted in non-decreasing order with respect to the values on a specific MV of Customer Satisfaction block, we kept the 40% of observation about its mean (equivalent to 100 units), thus 20% of the observations on its left-neighborhood and the other 20% on its right-neighborhood. Then, the

MVs of the endogenous block are transformed into new variables such that they take values between 6 and 10.

In the third step, the two data sets are merged, obtaining an unique data set of sample size equal to 250. Note that the MVs of the exogenous blocks in the two models come from the same population, while the same does not hold for the MVs of the endogenous block.

The distribution of a generic MV of the endogenous block looks like the one depicted in Figure 4.12.

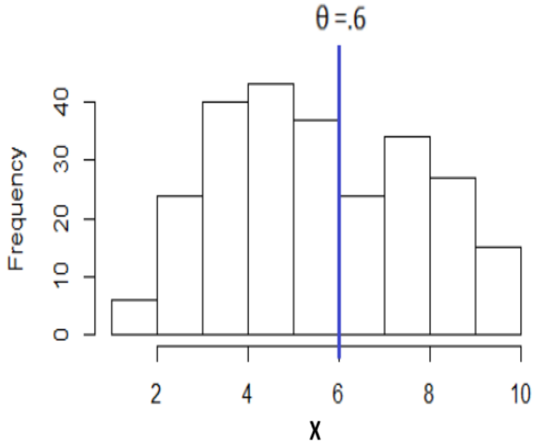


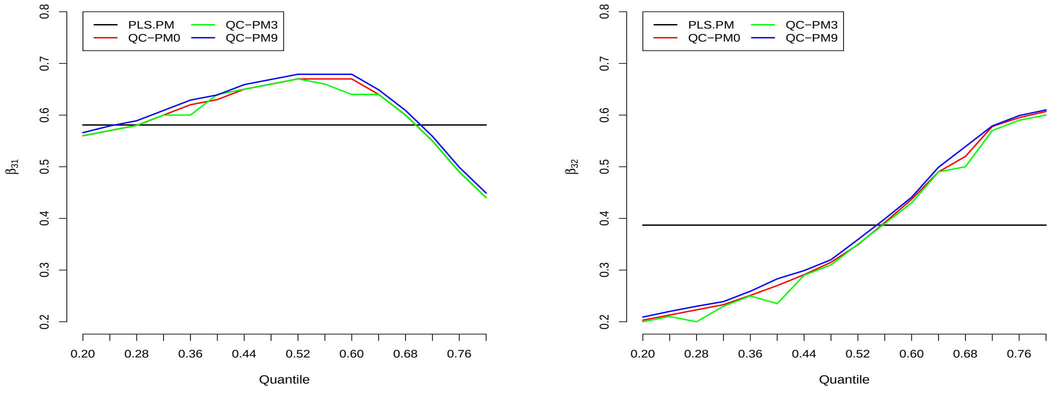
FIGURE 4.11: Simulated Distribution of a generic manifest variable

We expect that for quantiles smaller than 0.6 the model estimates refer to the population parameter of the model shown in the Figure 4.10a, while for quantiles larger than 0.6 the model estimates refer to the population parameter of the model shown in the Figure 4.10b.

Figure 4.12 shows the the different QC-PM path coefficient estimates across quantiles as well as the PLS-PM path coefficient estimates, for the path coefficient β_{31} (a) and β_{32} (b).

Looking at Figure 4.12 it is evident that QC-PM is able to distinguish the different effects in the different parts of the distribution for both β_{31} and β_{32}. β_{31} decreases for quantiles larger than 0.6. To the contrary β_{32} increases for quantiles larger than 0.6.

However, even the basic approach that apply QR using the scores of the classical PLS-PM, the QC-PMØ, is able to is able to distinguish the different effects in the



(a) (b)

FIGURE 4.12: Path coefficient estimates across quantiles: β_{31} (a) and β_{32} (b)

different parts of the distribution in this case, and this is because the heterogeneity appears only in the structural model.

When heterogeneity arises in the measurement model, QC-PMØ is not able to distinguish the different effects of MVs on LVs as the weights of the QC-PMØ are those computed in the classical PLS-PM, which considers only the changes in the means.

In order to highlight also differences in the weights for different parts of the MV distributions, we will show an example with heterogeneity in the measurement model.

4.6.4 The Case of Heterogeneity in the Measurement Model

In the third simulation study we assume heterogeneity in the measurement model. In particular, the relationships between MVs and the corresponding endogenous LV differ across quantiles.

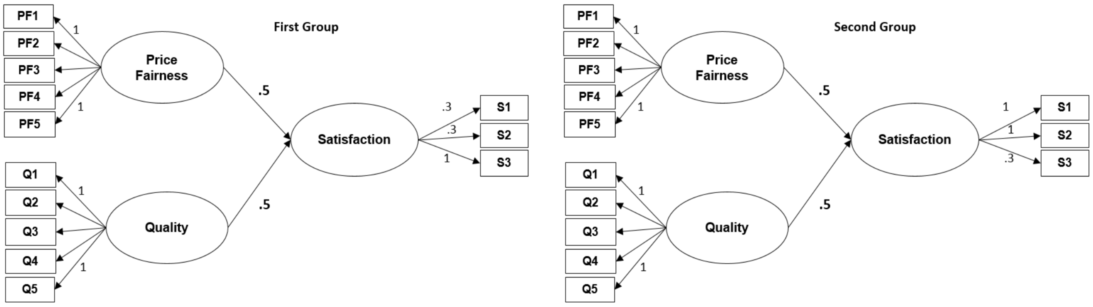
As above, we suppose that two different populations exist, and for each population the model parameters are different.

In this case, we suppose that the two groups of customers have the following characteristics:

- a first group characterized by a weak correlation between the first and second MV of the Customer Satisfaction block;

- in the second group all the correlations between Customer Satisfaction and itw own MVs are the same.

In particular, for the first population, we suppose that the path coefficients between Price Fairness and Customer Satisfaction and between Quality and Customer Satisfaction are the same and equal to 0.5 (in order to focus only on the measurement model). In the Customer Satisfaction measurement model the first and the second loadings are set equal to 0.3, while the third loadings is equal to 1 (see Figure 4.13a). In the second model, instead, the path coefficients are still both equal to 0.5, the first and and the second loadings are set equal to 1 in the Customer Satisfaction measurement model, while the third loading is equal to .3 (see Figure 4.13b).

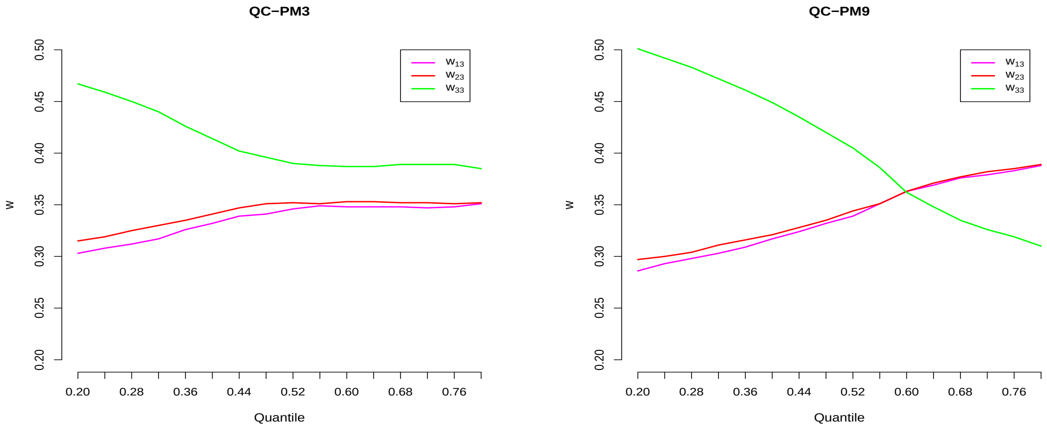


(a) (b)

Figure 4.13: Theoretical Model for Simulated Data

Figure 4.12 shows the outer weights across quantiles of the QC-PM3 (a) and QC-PM9 (b).

Both QC-PM3 and QC-PM9 are able to distinguish differences in the weights for different quantiles. However, QP-PM9 results seem to be more coherent with the simulation design. As a matter of fact, the weights of the first two MVs of the block Satisfaction are smaller before the quantile 0.6 and they increase as the quantile is greater than 0.6. The contrary happens for the weight of the third MV, according to the simulation design.



(a) (b)

FIGURE 4.14: Outer weights across quantiles

4.7 Conclusion

QC-PM enhances PLS-PM potentialities when regressor effects is different for different parts of the dependent variable distributions. As a matter of fact, QC-PM accommodates heteroscedastic variances and outliers and is able to explore the entire conditional distribution of the response variables.

The basic approach that apply QR using the scores of the classical PLS-PM, the QC-PMØ, is able to distinguish different exogenous LVs effects in the different parts of the endogenous LVs distributions. However, QC-PMØ is not able to distinguish the different effects of MVs on LVs as the weights of the QC-PMØ are those computed in the classical PLS-PM.

On the contrary, the more complex QC-PM, that introduces either QR and QC in all the outer and inner estimation steps of the algorithm, is able to distinguish the different effects of MVs on LVs as well as the different exogenous LVs effects in the different parts of the endogenous LVs distributions.

Future researches should concentrate on developing statistical test for evaluating the significance of differences among QPLSPM coefficients across quantiles.

Moreover, the outer inwards scheme in QC-PM should be investigated in more details with the consequent problem of multicollinearity.

Conclusions

In this dissertation we discussed some issues in PLS-PM and proposed methodological contributions to enhance PLS-PM potentialities.

PLS-PM is a component-based method for SEM. Instead of severing every tie between component-based methods and factor-based methods we think that researchers should commit themselves in finding out which approach works best in which circumstances, and a continuous dialogue between the two communities of researchers is highly recommended for progress in this area of research.

In the second chapter, we compared PLS-PM and ML-SEM in the framework of the same simulation design, investigating the effects of measurement model misspecification and the implications of formative MVs on both methods.

The implication of formative blocks in Covariance-Based framework is a rather difficult task. However, if certain model specification conditions are satisfied the model is identified, and it is possible to estimate a Covariance-Based SEM with formative blocks (Bollen and Davis, 2009; Williams et al., 2003).

Measurement model misspecification has the potential for poor parameter estimates and misleading conclusions (see Dolce and Lauro, 2014; Jarvis et al., 2003; MacKenzie et al., 2005, among others). Its effects extend also to the estimates of the path coefficients connected to the misspecified block. Our simulation results showed that misspecification is a severe problem in covariance-based SEM, while it is not a crucial issue in PLS-PM.

This work represent only a first step in this direction of comprehension. Different levels of complexity of the structural model with different population parameter values should be considered for further studies. Since in a simulation study the value of the parameters should reflect values commonly encountered in applied research, we think that it would be interesting to run simulation studies considering other well-established models (like the ECSI model), where measurement model misspecification frequently occurs. Different model specifications can also be considered including an endogenous formatively-measured LV.

Moreover, besides the descriptive statistics that we used to summarize and present the simulation results, inferential statistics can be used as well. For example, the experimental conditions can be dummy or effect coded, and main effects and interactions among experimental conditions can be evaluated using standard regression procedures.

Finally, we think that it would also be interesting to look further into the issue of multicollinearity among MVs in formative blocks.

Besides considering PLS-PM as an alternative method for SEM, PLS-PM is a descriptive and prediction oriented method, deserving a prominent place in research applications when the aims of the analysis is prediction (Becker et al., 2013). For this reasons, further studies on the predictive ability of PLS-PM are needed.

The PLS-PM evaluation criteria should include the predictive ability and further criteria and evaluation techniques for PLS-PM are needed (Sarstedt et al., 2014). Based on the proposed criteria, further extensions and modifications should be made on the basic PLS-PM algorithm in order to improve the predictive capabilities of the model estimation. The non-symmetrical approach for component-based path modelling (NSC-PM) presented in the third chapter of this dissertation is an example of work in this direction.

NSC-PM is a non-symmetrical approach that aims at maximizing the explained variance of the MVs of the endogenous and bridge blocks (i.e. an approach based on the optimization of a redundancy-related criterion in a multi-block framework).

Unlike PLS-PM, which analyzes symmetrically the relationships between LVs, without taking into account the roles of the dependent and explanatory LVs in the structural model, NSC-PM respects the direction of the relationship specified in the path diagram (i.e. the path directions), since the directions of the links in the structural model play a role in the algorithm. Compared to the other component-based methods, NSC-PM seems to be a good compromise between favouring stability (high explained variance) in the blocks and correlation between components. NSC-PM is a new method to consider if prediction is the main purpose.

Further research will be carried out to find out if the algorithm optimizes a global criterion. Stability of the algorithm and coherence of the different steps are promising for the investigation of a global optimizing criteria of the procedure.

In the last chapter of the thesis we presented the Quantile Composite-based Path Modelling (QC-PM). QC-PM exploits both Quantile regression (QR) (Koenker and Basset, 1978) and quantile correlation (QC) (Li et al., 2014), which allow respectively the estimation of a set of conditional quantile functions and a correlation measure to examine the linear relationships between any two variables for different quantiles, providing a more complete picture of the relationships between variables.

QC-PM is advisable as a complementary analysis to the classical PLS-PM, in the case where it is interesting to investigate if the relationships between dependent variables and regressors changes across different parts of the response variable distributions.

Future researches will be needed to develop statistical test for evaluating the significance of differences among QPLSPM coefficients across quantiles. Moreover, the outer inwards scheme in QC-PM should will be investigated in more details with the consequent problem of multicollinearity.

Bibliography

ACSI (2000). *American Customer Satisfaction Index, LLC. Food processing sector.*

Albers, S. and Hildebrandt, L. (2006). Methodological problems in success factor studies - measurement error, formative versus reflective indicators and the choice of the structural equation model. *Zeitschrift für betriebswirtschaftliche Forschung*, 58(2):2–33.

Amato, S., Esposito, V. V., and Tenenhaus, M. (2005). A global goodness-of-fit index for pls structural equation modeling. Technical report, HEC School of Management, France.

Anderson, E. and Fornell, C. (2000). Foundations of the american customer satisfaction index. *Journal of Total Quality Management*, 11(7).

Areskoug, B. (1982). The first canonical correlation: Theoretical pls analysis and simulation experiments,. In Jöreskog, K. and Wold, H., editors, *Systems Under Indirect Observation: Causality, Structure, Prediction*, volume 2, pages 95–118. Amsterdam: North Holland, S.

Armstrong, J. S. (2001). Combining forecasts. In Armstrong, J. S., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 417–439. New York: Springer.

Barclay, D., Higgins, C., and Thompson, R. (1995). The partial least squares (PLS) approach to causal modeling: personal computer adoption and use as an illustration. *Technology Studies*, 2(2):285–309.

Barrodale, I. and Roberts, F. D. K. (1974). Solution of an overdetermined system of equation in the l1 norm. *Communications of the Association for Computing Machinery*, 17:319–320.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20:451–468.

Becker, J.-M., Rai, A., Ringle, C. M., and Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Quarterly*, 37(3):665–694.

Bentler, P. and Huang, W. (2014). On components, latent variables, PLS and simple methods: reactions to ridgon's rethinking of PLS. *Long Range Planning*, 47:138–145.

Bentler, P. and Weeks, D. (1980). Linear structural equations with latent variables. *Psychometrika*, 45(3):289–308.

Blalock, H. (1971). *Causal models in the social sciences.* Chicago, IL: Aldine-Atherton.

Bollen, K. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53:605–634.

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

Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quafity and Quanttty*, 18:377–385.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. and Davis, W. R. (2009). Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):498–522.

Bollen, K. A. and Ting, K. F. (2000). A tetrad test for causal indicators. *Psychological Bulletin*, 5(1):3–22.

Buchinsky, M. (1994). Changes in the u.s. wage structure 1963-1987: application of quantile regression. *Econometrica*, 62:405–458.

Cenfetelli, R. T. and Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Q.*, 33(4):689–707.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, 26(1):271–292.

Chin, W. W. (1998). The partial least squares approach for structural equation modeling. In Marcoulides, G., editor, *Modern Methods for Business Research*, pages 295–336. Lawrence Erlbaum Associates, London.

Chin, W. W. (2010). Bootstrap cross-validation indices for pls path model assessment. In Esposito Vinzi, V., Chin, W. W., Henseler, J., and Wang, H., editors, *Handbook of partial least squares*, pages 83–97. Springer Berlin Heidelberg.

Dana, J. and Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29(3):317–331.

Davino, C. and Esposito Vinzi, V. (2015). Quantile composite-based path modelling. *Submitted*.

Davino, C., Esposito Vinzi, V., and Dolce, P. (2015a). Assessment and validation in quantile composite-based path modelling. In Abdi, H., Esposito Vinzi, V., Russolillo, G., Saporta, G., and Trinchera, L., editors, *Springer Proceedings in Mathematics & Statistics from PLS'14 - 8th International Conference on Partial Least Squares and Related Methods*, Paris.

Davino, C., Furno, M., and Vistocco, D. (2013). *Quantile Regression: Theory and Applications*. Wiley.

Davino, C., Romano, R., and Naes, T. (2015b). The use of quantile regression in consumer studies. *Food Quality and Preference*, 40:230–239.

Davino, C. and Vistocco, D. (2008). Quantile regression for the evaluation of student satisfaction. *Italian Journal of Applied Statistics*, 20:179–196.

Diamantopoulos, A. (2006). The error term in formative measurement models: interpretation and modeling implications. *Journal of Modelling in Management*, 1(1):7–17.

Diamantopoulos, A., Riefler, P., and Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12):1203 – 1218.

Diamantopoulos, A. and Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2):269–277.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics*, 22:67–90.

Dijkstra, T. (2011). Consistent partial least squares estimators for linear and polynomial factor models. Technical report, Working Paper.

Dijkstra, T. (2014). PLS' janus face: response to professor rigdon's 'rethinking partial least squares modeling: in praise of simple methods'. *Long Range Planning*, 47:146–153.

Dolce, P., Esposito Vinzi, V., and Lauro, C. (2015). Non-symmetrical component-based path modeling. In Abdi, H., Esposito Vinzi, V., Russolillo, G., Saporta, G., and Trinchera, L., editors, *Springer Proceedings in Mathematics & Statistics from PLS'14 - 8th International Conference on Partial Least Squares and Related Methods.*

Dolce, P. and Hanafi, M. (2015). Illogical forms in pls path modelling. *Unpublished manuscript.*

Dolce, P. and Lauro, N. (2014). Comparing maximum likelihood and PLS estimates for structural equation modeling with formative blocks. *Quality & Quantity*, pages 1–12.

ECSI (1998). *European Customer Satisfaction Index. Report prepared for the ECSI Steering Committee.*

Edwards, J. and Bagozzi, R. (2000). On the nature and direction of relationships between constructs and measures. *Psychol Methods*, 5(2):155–74.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* CRC Press LLC, New York.

Eide, E. and H.Showalter, M. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, 58:345–350.

Esposito Vinzi, V., Chin, W. W., Henseler, J., and Wang, H. (2010a). *Handbook of Partial Least Squares: Concepts, Methods and Applications.* Springer.

Esposito Vinzi, V., Ringle, C., Squillacciotti, S., and Trinchera, L. (2007). Capturing and Treating Unobserved Heterogeneity by Response Based Segmentationin PLS Path Modeling. A Comparison of Alternative Methods by Computational Experiments. Working paper, ESSEC Business School.

Esposito Vinzi, V. and Russolillo, G. (2013). Partial least squares algorithms and methods. *WIREs Computational Statistics*, 5:1–19.

Esposito Vinzi, V., Trinchera, L., and Amato, S. (2010b). PLS path modeling: Recent developments and open issues for model assessmentand improvement. In Esposito Vinzi, V., Chin, W., Henseler, J., and Wang, H., editors, *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*. Springer, Berlin, Heidelberg, New York, Berlin, Heidelberg, New York.

Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., and Tenenhaus, M. (2008). Rebus-pls: A response-based procedure for detecting unit segments in pls path modeling. *Applied Stochastic Models in Business and Industry*, 24:439–458.

Evermann, J. and Tate, M. (2012). Comparing the predictive ability of pls and covariance models. In *Proceedings of the International Conference on Information Systems*, Orlando, Florida.

Falk, R. F. and Miller, N. B. (1992). *A Primer for soft modeling*. The University of Akron Press, Akron, Ohio, USA.

Fitzenberger, B., Koenker, R., and Machado, J. (2002). *Economic Applications of Quantile Regression*. Series: Studies in Empirical Economics. Physica Verlag, Heidelberg.

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(4):pp. 440–452.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Reseach*, 18:39–50.

Fornell, C., Rhee, B.-D., and Yi, Y. (1991). Direct regression, reverse regression, and covariance structure analysis. *Marketing Letters*, 2:309–320. 10.1007/BF00554134.

Fox, M. and Rubin, H. (1964). Admissibility of quantile estimates of a single location parameter. *Ann. Math. Statist.*, 35(3):1019–1030.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328.

Glang, M. (1988). *Maximierung der Summe erklärter Varianzen in linearrekursiven Strukturgleichungsmodellen mit multiple Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahren.* Phd thesis, Universität Hamburg, Hamburg, Germany.

Goodhue, D. L., Lewis, W., and Thompson, R. (2006). PLS, small sample size, and statistical power in mis research. In *Paper presented at the HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on System Sciences.*

Goodhue, D. L., Lewis, W., and Thompson, R. (2012). Does pls have advantages for small sample size or non-normal data? *MIS Quarterly*, 36(3):981–1001.

Gould, W. (1997). sg70: Interquantile and simultaneous-quantile regression. *Stata Technical Bulletin*, 38:14–22.

Gudergan, S. P., Ringle, C. M., Wende, S., and Will, A. (2008). Confirmatory tetrad analysis in pls path modeling. *Journal of Business Research*, 61:1238 –1249.

Guinot, C., Latreille, J., and Tenenhaus, M. (2001). PLS Path modeling and multiple table analysis. Application to the cosmetic habits of women in the ile-de-France. *Chemometrics and Intelligent Laboratory Systems*, 58:247–259.

Hair, J., Black, W., Babin, B., and Anderson, R. (2010). *Multivariate Data Analysis. A Global Perspective.* Pearson Education, Inc., USA.

Hair, J., Sarstedt, M., Pieper, T., and Ringle, C. (2012a). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, 45:320–340.

Hair, J., Sarstedt, M., Ringle, C., and J.Mena (2012b). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40:414–433.

Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Thousand Oaks, CA: Sage.

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory & Practice*, 19:2:139–152.

Hanafi, M. (2007). PLS path modeling: computation of latent variables with the estimation mode B. *Computational Statistics*, 22:275–292.

Hao, L. and Naiman, D. Q. (2007). *Quantile Regression.* Thousand Oaks, CA: Sage Publications.

Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Statist. Ass.*, 93:58–68.

Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm. *Computational Statistics*, 25 (1):107–120.

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M., and Calantone, R. J. (2014). Common beliefs and reality about pls: Comments on rönkkö and evermann (2013). *Organizational Research Methods*, 17(2):182–209.

Henseler, J., Ringle, C., and Sinkovics, R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20:277–319.

Hotelling, H. (1933). Analysis of a complex of statistical variables into components. *Journal of Educational Psychology*, 24.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26:139 – 142.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hsu, S. H., Chen, W. H., and Hsieh, M. J. (2006). Robustness testing of pls, lisrel, eqs and ann-based sem for measuring customer satisfaction. *Total Quality Management & Business Excellence*, 17:355–372.

Huang, W. (2013). *PLSe: Efficient Estimators and Tests for Partial Least Squares. In progress.* PhD thesis, UCLA.

Huarng, K.-H. (2014). A quantile regression forecasting model for ict development. *Management Decision*, 5.

Hulland, J., Ryan, M. J., and Rayner, R. K. (2010). Modeling customer satisfaction: a comparative performance evaluation of covariance structure analysis versus partial least squares. In Esposito Vinzi, V., Chin, W. W., Henseler, J., and Wang, H., editors, *Handbook of Partial Least Squares: Concepts, Methods and Applications (Springer Handbooks of Computational Statistics Series)*, pages 307–325. Heidelberg, Dordrecht, London, New York: Springer.

Hwang, H. and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, 69:81–99.

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

Jöreskog, K. (1970). A general method for analysis of covariance structure. *Biometrika*, 57:239–251.

Jöreskog, K. (1973). A general method for estimating a linear structural equation system. In Goldberger, A. and Duncan, O., editors, *Structural Equation Models in the Social Sciences*. Academic Press, New York.

Jöreskog, K. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In Krishnaiah, R., editor, *Applications of Statistics*, pages 265–287. Amsterdam: North-Holland.

Jöreskog, K. and Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Abt Books.

Jöreskog, K. and Wold, H. (1982a). The ML and PLS techniques for modeling with latent variables: historical and comparative aspects. In Jöreskog, K. and Wold, H., editors, *Systems Under Indirect Observation*, volume Part I, pages 263–270. North-Holland, Amsterdam.

Jöreskog, K. G. and Wold, H. O. A. (1982b). *Systems under indirect observation : causality, structure, prediction*. North-Holland, Amsterdam.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Basset, G. (1982a). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50:43–61.

Koenker, R. and Basset, G. (1982b). Tests for linear hypotheses and l1 estimation. *Econometrica*, 46:33–50.

Koenker, R. and d'Orey, V. (2001). Computing regression quantiles. *Applied Statistics*, 36:383–393.

Koenker, R. and Hallock, V. (2001). Quantile regression. *Journal of Economic Perspectives*, 15:143–156.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296 – 1310.

Krämer, N. (2007). *Analysis of high-dimensional data with partial least squares and boosting*. Phd thesis, Technische Universität Berlin, Berlin, Germany.

Kristensen, K. and Eskildsen, J. (2010). Design of pls-based satisfaction studies. In Vinzi, V. E., Chin, W., Henseler, J., and Wang, H., editors, *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*. Springer, Berlin, Heidelberg, New York.

Lauro, N. and D'Ambra, L. (1984). L'analyse non symmétrique des correspondances. In Diday, E. and al., editors, *Data Analysis and Informatics, III*. North-Holland.

Lee, L., Petter, S., Fayard, D., and Robinson, S. (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, 12:305–328.

Li, G., Li, Y., and Tsai, C. (2014). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association*, accepted.

Li, M. (2014). Moving beyond the linear regression model. advantages of the quantile regression model. *Journal of Management*, accepted.

Little, T., Lindenberger, U., and Nesselroade, J. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, 4(2):192–211.

Lohmöller, J. (1989). *Latent Variable Path Modeling with Partial Least Squares*. Physica-Verlag, Heildelberg.

Lu, I. R. R. (2004). *Latent Variable Modeling in Business Research: A Comparison of Regression Based on IRT and CTT Scores with Structural Equation Models.* PhD thesis, Carleton University, Canada.

Lu, I. R. R., Thomas, D. R., and Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, 12:263–277.

Lyttkens, E., Areskoug, B., and Wold, H. (1975). The convergence of NIPALS estimation procedures for six path models withone or two latent variables. Technical report, University of Göteborg.

MacCallum, R. and Browne, M. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychol. Bull.*, 114(3):533–541.

MacKenzie, S. B., Podsakoff, P. M., and Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, 90(4):710–730.

Makridakis, S. and Hibon, M. (2000). The m3 competition: Results, conclusions and recommendations. *International Journal of Forecasting*, 16(4):451–476.

Marcoulides, G. A., Chin, W. W., and Saunders, C. (2009). A critical look at partial least squares modeling. *MIS Q.*, 33(1):171–175.

Marcoulides, G. A. and Saunders, C. (2006). Pls: A silver bullet? *MIS Quarterly*, 30(2):III–IX.

Mathes, H. (1993). Global optimisation criteria of the pls-algorithm in recursive path models with latent variables. In Haagen, K., Bartholomew, D., and Deister, M., editors, *Statistical modelling and latent variables.* Elsevier Science: Amsterdam.

Muthén (2003). Beyond SEM: general latent variable modeling. *Behaviormetrika*, 29:81–117.

Nooan, R. and Wold, H. (1982). PLS path modeling with indirectly observed variables: a comparison of alternative estimates for latent variable. In Jöreskog, K. and Wold, H., editors, *Systems Under Indirect Observations: Causality, Structure, Prediction*, pages 75–94. North Holand, Amsterdam.

Parzen, M. I., Wei, L., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 18:341–350.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., and Chen, F. (2001). Monte carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2):287–312.

Rencher, A. (1998). *Multivariate statistical inference and applications.* Wiley Series in Probability and Statistics, New York.

Rigdon, E. E. (2012). Rethinking Partial Least Squares path modeling: in praise of simple methods. *Long Range Planning*, 45:341–358.

Rigdon, E. E. (2014). Rethinking partial least squares path modeling: Breaking chains and forging ahead. *Long Range Planning*, 47:161–167.

Ringle, C., Sarstedt., M., and Straub, D. (2012). A critical look at the use of pls-sem in mis quarterly. *MIS Quarterly*, 36:iii–xiv.

Rönkko, M. and Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods*, 16(3):425–448.

Sarstedt, M., Ringle, C., Henseler, J., and Hair, J. (2014). On the emancipation of PLS-SEM: a commentary on rigdon (2012). *Long Range Planning*, 47:154–160.

Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. *Qual. Quant.*, 24(4):367–386.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Schneeweiss, H. (1993). Consistency at large in models with latent variables. In Haagen, K., Bartholomew, D. J., and Deistler, M., editors, *Statistical Modelling and Latent Variables.* Elsevier, Amsterdam.

Shmueli, G. and Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3):553–572.

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psychol Bull.*, 70(3):160–163.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76 (2)(2):257–284.

Tenenhaus, M. (1998). *La Régression PLS: théorie et pratique.* Technip, Paris.

Tenenhaus, M., Esposito, V. V., Chatelin, Y. M., and Lauro, C. (2005). Pls path modeling. *Computational Statistics & Data Analysis*, 48(1):159–205.

Treiblmaier, H., Bentler, P. M., and Mair, P. (2011). Formative constructs implemented via common factors. *Struct. Equ. Model.: A Multidiscip. J.*, 18(1):1–17.

Trinchera, L. (2007). *Unobserved Heterogeneity in Structural Equation Models: a new approach to latent class detection in PLS Path Modeling.* Phd thesis, Universitá degli Studi di Napoli Federico II, Naples, Italy.

Vale, C. and Maurelli, V. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3):465–471.

Vilares, M., Almeida, M., and Coelho, P. (2010). Comparison of likelihood and pls estimators for structural equation modeling: a simulation with customer satisfaction data. In Esposito Vinzi, V., Chin, W., Henseler, J., and Wang, H., editors, *Handbook of Partial Least Squares*, pages 289–305. Springer-Verlag Berlin Heidelberg.

Vilares, M. J. and Coelho, P. S. (2013). Likelihood and pls estimators for structural equation modeling: an assessment of sample size, skewness and model misspecification effects. In da Silva, J. L., Caeiro, F., Natário, I., and Braumann, C. A., editors, *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, pages 11–33. Berlin, Heidelberg: Springer.

Whittaker, J., Whitehead, C., and Somers, M. (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 54:863–878.

Williams, L. J., Edwards, J. R., and Vandenberg, R. J. (2003). Recent advances in causal modeling methods for organizational and management research. *Journal of Management*, 29(6):903 – 936.

Wold, H. (1966a). Estimation of principal component and related models by iterative leastsquares. In Krishnaiah, P. R., editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York.

Wold, H. (1966b). *Non linear Estimation by Iterative Least Squares procedure.* Research paper in Statistics: Festschift for J. Neyman. F. David.

Wold, H. (1975a). Path models with latent variables: The non-linear iterative partial leastsquares (NIPALS) approach. In Blalock, H. M., Aganbegian, A., Borodkin, F. M., and andVittorio Capecchi, R. B., editors, *Quantitative Sociology: Intentional Perspective on Mathematical and StatisticalModeling*, pages 307–357. Accademic Press, New York.

Wold, H. (1975b). Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In *Gani, J. (Ed.), Perspectives in probability and statistics, papers in honor of M.S. Bartlett*, pages 117–142, London. Academic Press.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In Kmenta, J. and Ramsey, J. B., editors, *Evaluation of Econometric Models*, pages 47–74.

Wold, H. (1981). Comments on the papers by j. b. grubman et al, comments integrated with a briefing of pls (partial least squares) soft modeling. In *ASA-CENSUS-NBER Conference on Applied Time Series Analysis of Economic Data, .*

Wold, H. (1982). Soft modeling: the basic design and some extensions. In Jöreskog, K. and Wold, H., editors, *Systems under Indirect Observation*, volume 2, pages 1–54. North-Holland, Amsterdam.

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In Ruhe, A. and Kagstrom, B., editors, *Proceedings of the Conference on MatrixPencils. Lectures Notes in Mathematics*, Heidelberg. Springer.

Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219.