NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

P-IRLS-PM

A new approach to non-linear formative consctructs

Francisco de Sousa Gago Prata Lourenço

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

P-IRLS-PM

A NEW APPROACH TO NON-LINEAR FORMATIVE CONSCTRUCS

by

Francisco de Sousa Gago Prata Lourenço

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information Analysis and Management

Advisor: Jorge Morais Mendes

May 2018

ABSTRACT

The traditional approach to PLS-PM estimated the scores of the formative latent variables as exact linear combinations of their associated manifest variables, not allowing for modelling other relationships besides the linear ones. The present study intends to overcome this limitation, introducing the P-IRLS-PM. The P-IRLS-PM is a new approach to the variance based structural equations models, which intends to widen the spectra of how the formative latent scores are represented by its associated manifest variables. Throughout this work the core concepts of the algorithm P-IRLS-PM will be introduced, followed by a Monte Carlo experiment comparing the new approach with the traditional PLS-PM.

KEYWORDS

Partial Least Squares Path Modeling; Generalized Additive Models; Non-linear Relationships; Thin-Plate Regression Splines; Formative Measurement Model

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

PLS-PM	Partial Least Squares Path Modeling
PLS-SEM	Partial Least Squares Structural Equation Modeling
P-IRLS-PM	Penalized Iteratively Re-Weighted Least Squares Path Modeling
SEM	Structural Equation Modeling
COR	Correlation
MSE	Mean Squared Error
OBS	Observations
ITER	Iterations
MV	Manifest Variable
LV	Latent Variable
OLS	Ordinary Least Squares

1. INTRODUCTION

In the field of Structural Equation Models the popularity of Partial Least Squares Path Modeling (Wold, 1975) or just PLS-PM has been increasing in the last years (Ringle et al., 2012). According to Hair et al. (2013) Accounting, International Marketing, Management Information Systems, Marketing and Operations Management are examples of areas where the phenomenon described above is taking place.

Structural equation modelling is a multivariate statistical analysis, which enables researchers to model unobservable variables measured indirectly by indicator variables (Hair et al. , 2014). Within this analysis, there are two main approaches that should not be seen as independent but as complementary techniques (Hair et al., 2011). Such techniques are the Covariance-Base Structural Equation Modeling (Jöreskog, 1978) and Partial Least Squares Path Modeling (Wold, 1975). Depending on the objective of the work, the researcher should take in consideration the purpose of each methodology. In the early stages of theory development, when the researcher wants to predict latent variable relationships, but there is none or little prior knowledge on how the variables are related, the PLS-PM is more suited (Hair et al., 2011). Otherwise, when the objective has more emphasis in confirmation than in exploratory research, it should be used the CB-SEM (Reinartz et al., 2009).

The PLS-PM is known by two terms that reflect its characteristic, variance-based approach to SEM and "soft modeling". The first, comes from how the algorithm works, PLS-PM "estimates model parameters to maximize the variance explained for all endogenous constructs intermodal through a series of ordinary least squares (OLS) regressions" (Reinartz et al., 2009, p. 332), while the covariance base "attempts to minimize the difference between the sample covariance and those predicted by the theoretical model.…Therefore, the parameter estimation process attempts to reproduce the covariance matrix of the observed measures" (Chin & Newsted, 1999, p. 309). The second term comes from the model assumptions are less strictly than the alternative developed by Jöreskog, known as "hard modeling" (Hair et al., 2014) and pointed out in the study (Sosik et al., 2009), "Soft modeling" allows for greater flexibility from a practical point of view. In other words, the PLS-PM is more suitable (Hair et al., 2011), when the CB-SEM assumptions do not hold or this method reach is limit. This is experienced when the available data are not normally distributed and sample sizes are small (Sosik et al., 2009). Among these strengths is important to highlight the fact the PLS-SEM provides higher levels of statistical power compared with CB-SEM for theory testing (Hair et al., 2011) and formative constructs are easier to incorporate in variance-based than covariance-based approach (Hair et al., 2013). On the other side, the PLS-PM has its drawbacks. One of these shortcomings is related with the focus on maximizing partial model structures. Thus, it is required to firstly examine the measurement model characteristics, before assess the structural model (Hair et al., 2011).

As described by Monecke and Leisch (2012), the scores of the latent variables, are estimated as exact linear combinations of their associated manifest variables. These hypotheses can be hard to hold, particularly when faced with nonlinearity and asymmetric data. For example, modeling an attribute assuming a symmetric and linear relationship with a variable like customer satisfaction, leads to misestimates when that relationship is in fact a asymmetric and nonlinear (Anderson & Mittal, 2000).

The main objective of the present work is the improvement of how the latent variables are represented by the manifest variables, when faced with formative models and provide a way to overcome the misspecifications mention in the previous paragraph. To reach such target, the outer estimation of the measurement model will be switched from the traditional linear regression to a thin-plate regression spline (Wood, 2003), creating a new approach to the Partial Least Squares Path Modeling the P-IRLS-PM.

Bearing in mind the growing popularity of PLS-PM (Becker et al., 2012) especially of reflective constructs (Becke et al., 2012), the present thesis has a considerable importance due to its focus on the less popular- formative constructs. Moreover, the proposed changes to the algorithm given throughout the following chapters, will strengthen the PLS-PM against criticisms. The present study uses formative constructs with nonlinear specification, a topic which should receive more attention due its relevance. Furthermore, the computational implementation will be done in an open source programming language and software (*R*), aiming to instigate the study in PLS-PM.

The present work is structured in a Literature Review (chapter 2), where the roots of the PLS-PM are discussed, and a brief overview of the advances made for this model are introduced. Chapter 3 describes the methodology. Chapter 4 presents the theoretical background of the traditional PLS-PM alongside with the presentation of the new approach. In chapters 5 the two main algorithms are compared, discussed and assessed. Lastly, Chapters 6 and 7 presents conclusions, limitations of present work and recommendations to further research.

2. LITERATURE REVIEW

When trying to find the roots of the Structural Equations Models or of its creators, the researcher is not confronted with a straightforward answer, nonetheless it is possible to find some mutual outlines. According to Bollen (1989), Kline (2015) and Sanchez (2013), part of its origins date to early days of the twentieth century with the development of the Path Analysis (Wright, 1921) and Factor Analysis (Spearman, 1904). A few years later, these two methods were integrated into the same model in the work of Karl Jöreskog making the Covariance Based Structural Equation Modeling (Jöreskog, 1978). Around the time period, the covariance approach was in its first stages, Herman Wold influenced the adaptation of the Nonlinear Iterative Partial Least Squares modeling to approach SEM models (Sanchez, 2013), creating the Partial Least Squares Path Modeling (Wold, 1975, 1980). The two main methodologies of Structural Equation Modeling were developed in the seventies, by Karl Jöreskog and Hernan Wold. The latter developed the variance-based approach, also known as soft modeling or PLS-PM, however, it was the method developed by Karl Jöreskog that became the most popular, the CB-SEM. Two possible factors that have triggered this preference were that variance-based software was only later developed (Hair et al., 2012b, p. 312) and the fact that the literature is not unanimous on issues regarding its legitimacy and usefulness of soft modeling. There are studies (e.g., Evermann & Tate, 2010; Hwang et al., 2010) who question the model technical value even suggesting its discontinuity (Rönkkö et al., 2016). However, other studies (Hair et al., 2011; Hair et al., 2012a; Ringle et al., 2012; Sosik et al., 2009) pointing out that some misunderstanding and criticism is often related to the lack of knowledge that prevents researchers to fully make use of the method's capabilities, sometimes even using it incorrectly. For example, in a review by Hair et al. (2012b), PSL-SEM was misused in three main points: model specification issues, data characteristics, and model assessment.

The outer model of the variance-based SEM has two main ways to measure non-observable variables Hair et al. (2014), the reflective and formative constructs (Monecke and Leisch (2012). More than forty years ago, the formative measurement models were introduced in the literature and the debate about their methodological advances has been increasing since the nineties (Diamantopoulos et al., 2008). However, the use of such measurement models in empirical studies is still scarce (Diamantopoulos & Siguaw, 2006). A reason behind this trend relates to the fact that there is a lack of practical guidelines on how to create, estimate and validate formative models (Suoniemi et al., 2012, p. 1648) and a significant number of researchers engaging in PLS-SEM might still be unaware of the potential of these indicators (Bollen, 2002). The traditional PLS-PM, enables the researcher to create and estimate models without imposing additional and limiting constraints (Hair et al., 2012a). As Hair at a. (2013) states this is the reason behind the increasing popularity of PLS-PM in a wide range of disciplines.

Over time, soft modeling has witnessed improvements in its methodology. These include, the advanced models introduced by Lohmöller (Lohmöller, 1989), a method that allows distinguishing a formative indicator specification from a reflective indicator specification (Gudergan et al., 2008), improvements on assessing hierarchical component models (Becker et al., 2012), PLS-SEM-specific data segmentation techniques (Ringle et al., 2010; Sarstedt, 2008) and developments in assessing the robustness of the outer and inner models (Chin & Dabber, 2010, Chapter 2; Shmueli et al., 2016).

Regarding the works on nonlinear effects (e.g. Dijkstra & Henseler, 2011; Hackl & Westlund, 2000; Ingrassia & Trinchera, 2008; Jakobowicz & Saporta, 2002), there are already some studies that focus on the PLS-PM´s inner model with non-linear relationships, which is the case of Henseler et al., 2012. However, it is possible to say that there was no empirical evidence found regarding any previous work that covers how formative measurement models can capture the non-linear relationships within the PLS-PM outer models. As such, the present work will embrace this challenge and introduce a new approach to the PLS-PM and its methodology will be discussed in the next chapter.

3. METHODOLOGY

The present study seeks to improve the latent scores obtained from formative specifications of the PLS-PM measurement model. To achieve this, the study's workflow (Figure 1) followed the methodology of Hair et al. (2014, p. 25) which can be summarized in three main stages. The first stage comprises four essential tasks: specifying the structural and measurement models, data collection and examination, while the second stage refers to the estimation of the PLS-PM model. Finally, the last stage encompasses the model assessment and its results analysis.

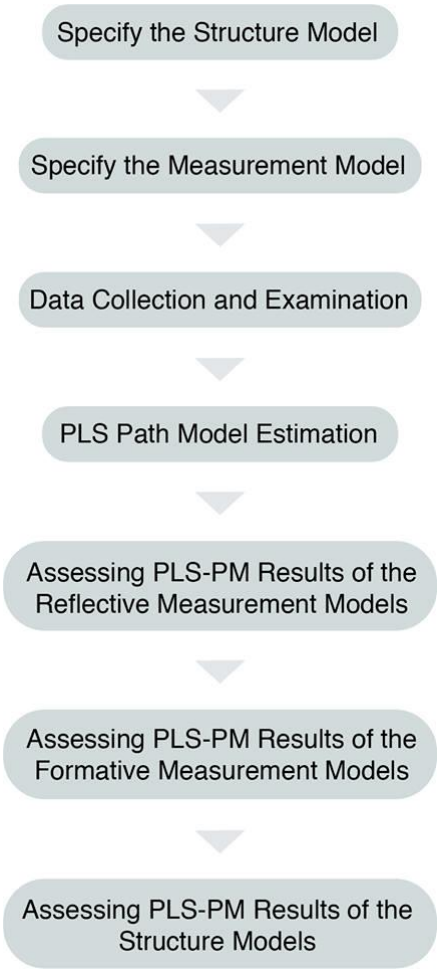

Figure 1 – Systematic Procedure for Applying PLS-PM (J. F. J. Hair et al., 2014)

The theoretical framework is composed by a recall of the traditional Partial Least Squares Path Modeling (Monecke and Leisch, 2012), alongside the introduction of the Penalized Iteratively Re-Weighted Least Squares Path Modeling (P-IRLS-PM), providing the core concepts of both models. Moreover, the notation used in the present study follows the description of Tenehaus (2005).

The comparison between both algorithms was conducted with the use of a synthetic dataset specifically designed for this study composed by several types of non-linear relationships, which follows the recommendation of (Henseler et al., 2012, p. 110) stating that further research could strive to analyze the impact, on PLS-PM, of the quadratic function alongside with other commonly used non-linear functions such as logarithmic or exponential functions.

The model used to develop the dataset consists in two endogenous latent variables and one exogenous latent construct. The endogenous variables have formative specifications and the exogenous is defined by a reflective model. Each measurement model is made of five indicators. After defining the underlying true model, a dataset of 1000 observations were generated, with a script in *R* language and the package *MASS*. Following a Monte Carlo procedure, 1000 samples of 100, 250, 750 and 900 observations were drawn and used for assessment of both algorithms. The results are then compared. The outputs of the traditional PLS-PM computations, obtain through the package *plspm*, were used as benchmark. Then, with the same samples, the new approach was estimated, but at the time this paper was written, none of the available PLS-PM software would allow to introduce the changes needed to estimate the new algorithm, so was developed a script in the *R* programing language, where the splines of P-IRLS-PM were estimated by the package *mgcv*. The script to run the P-IRLS-PM in R, is available in the appendix.

To compare the performance of the traditional approach with the P-IRLS-PM parameter accuracy, prediction accuracy, model fit and computational cost were assessed. Inspired by Henseler et al.(2012), the parameter accuracy was measure by the mean squared error and the prediction accuracy by the correlation between estimated and true values. This two metrics were computed in two levels, the first was for the latent scores and the second for the partial residuals of the latent scores of LV1 and LV2. The R^2 of the structural model was used to assess the model fit and the number of iterations to assess the computational cost. To provide a more clear and complete analysis of the partial residuals were introduced the plots of this relationships from one sample with 750 observations.

4. A NEW APROACH TO PLS-PM

This section introduces a new approach to Partial Least Squares Path Modeling, in which the standard formative measurement model was adjusted in order to account for non-linear relationships between the latent indicators and its block of variables. Penalized Iteratively Re-Weighted Least Squares Path Modeling (P-IRLS-PM) can be distinguished from earlier models through how the formative construct specification is made.

Herman Wold's methodology (Wold, 1980) uses multiple linear regressions to model formative constructs, leading to measure these relationships as linear. Such approach can be difficult to follow in practical experiments and researchers' attention has shifted toward non-linear models (Henseler et al., 2012). Taking into account the previous statements, there is a need to improve how the formative latent scores are represented. This can be achieved by replacing the linear regression on this type of constructs by a thin-plate regression spline (Wood, 2003). Instead of specifying a detailed parametric relationship using a linear regression, the model is solely characterized by smooth functions, allowing for a more flexible specification of the latent variable on its manifest variables. This spline basis provides a knot free approach that can be applied for any number of manifest variables. Apart from the latter major change and few adjustments in the algorithm, the components and methodology of the PLS-PM and P-IRLS-PM are very similar.

The present section is characterized by a stepwise overview of the subjacent theoretical concepts in P-IRLS-PM alongside with a recall of the traditional PLS-PM approach. For a more detailed literature on the traditional model review (Hair et al., 2014; Monecke, A. and Leisch, 2012; Sanchez, 2013; Tenenhaus et al., 2004)

4.1. THE COMPONENTS

The structural model, the measurement model and the weighting scheme are the three components of PLS-PM. The first two are presented in all kinds of structural equation models with latent constructs, but the weighting scheme is specific to the PLS-PM approach. Adding to the previous point, is only allowed recursive relationships and such relations can be expressed in path digraphs. Like the traditional approach, P-IRLS-PM has the same components and similar guidelines, being the specification of the formative constructs where the difference between the two models lies.

4.1.1. The Structural Model

The relationships between latent variables[1] are presented is the structural or inner model and the structural paths can only head in a single direction. Such variables are split in two classes, exogenous and endogenous constructs. The exogenous variables do not have any predecessor in the inner model, all the others are endogenous. For the benefit of simplicity, the notation used dismisses the difference

[1] Latent variables are measure concepts that are abstract and cannot be directly measured, also called factors.

between these kinds of variables. The structural model, with n number of latent variables, can be expressed by:

$$LV_j = \beta_0 + \sum_i \beta_{ji} LV_i + \nu_j \qquad i = 1, \dots, n \wedge i \neq j \tag{1}$$

where LV denotes a latent variable, the weights of the structural model are represented by β and the error term is represented by υ.

4.1.2. The Measurement Model

The measurement or outer model is described by the connection between each block of manifest variables[2] and its corresponding latent variable (LV). Additionally, each manifest variable (MV) is only allowed to be connected with one LV and within all MV's related to one LV all the arrows must point in the same direction. This model subdivides in reflective constructs, formative constructs and MIMIC[3].

4.1.2.1. Reflective measurement models

In these type of constructs, the causality is from the latent variable to its measures. The indicators presented in reflective models, share the properties that they are all cause by the same construct and they should by interchangeable, as long the construct has sufficient reliability (Hair et al., 2014). Hence the MV's must have a high correlation. Each reflective indicator is related to its latent construct by:

$$MV_h = w_h LV + \varepsilon_h, \tag{2}$$

where MV represent a manifest variable, the LV is its corresponding latent variable, w denotes the loadings and ε the error term. For the above model to be valid, it is necessary to confirm the hypothesis that the error term was zero mean and is uncorrelated with the manifest variables linked to latent variable (Tenenhaus et al., 2004).

[2] Manifest variables are the directly measured variables also called raw data or indicators.
[3] MIMIC models is a mixture of formative and reflective constructs.

4.1.2.2. Formative measurement models

The formative way assumes that the indicators cause the latent variable, thus the latter change due to variations in the manifest variables. An important characteristic of the indicators of these models is, each indicator captures a specific feature of the construct's domain, thus these variables are non-interchangeable, like they are in the reflective constructs.

These relationships in the traditional approach are defined by multiple linear regressions, where the latent variable is a linear function of its manifest variables plus a residual term (Tenenhaus et al., 2004) and the loadings are estimated using the Ordinary Least Squares. In the standard linear case, formative constructs are characterized by the following:

$$LV = \sum_{h} w_h MV_h + \delta, \tag{3}$$

where LV denotes a latent variable, MV_h represents the manifest variable h, w_h are the outer weights and δ the error term. This specification implies the hypothesis that the residual term has mean equal to zero and is uncorrelated with the indicators related to its latent variable (Tenenhaus et al., 2004).

The above model infers that the relationships in the formative constructs are supposed to be linear, restraining the used of nonlinear models. However, this hypothesis looks to be too restrictive in some cases. The P-IRLS-PM suggested an alternative technique, where the model specification is made through "smooth functions" terms. The new approach uses a thin-plate regression spline (Wood, 2003) to model formative constructs ,that can be represented by:

$$LV = \sum_{h} f_h(MV_h) + \delta \tag{4}$$

where LV denotes a latent variable, f_h represents the smooth function term of the manifest variable (MV) h and δ is the error term.

4.1.2.3. Weighting scheme

In the early days of PLS-PM, Herman Wold introduced the centroid weighting scheme, being developed later two more schemes by Lohmöller, the factorial weighting scheme and the path weighting scheme. Despite differences between these three methods being minor and usually does not leading to significantly changes the interpretations of the results (Garson, 2016), the present study will focus is attention in the scheme develop by Herman Wold.

Centroid weighting scheme:

In the scheme developed by Herman Wold, the sign of the correlation between a latent variable and its adjacent latent variables is used to populate the matrix of inner weights (matrix E), by following the next expressions:

$$C = D + D^T \tag{5}$$

$$R = COR(LV_j, LV_i) \qquad j, \mathrm{i} = 1, \dots, \mathrm{n} \tag{6}$$

$$e_{ji} \begin{cases} sign(r_{ji}) & , c_{ji} = 1 \\ 0 & , else \end{cases} \tag{7}$$

The inner design matrix is denoted by D, in which the structural model is expressed in an upper triangular matrix that is populated with the value 1 when LV_j is a predecessor of LV_i in the entry d_{ji}, otherwise is placed a 0). C represents an auxiliary matrix, LV denotes a latent variable, R is the empirical correlation matrix and the inner weights are represented by e_{ij}.

4.2. The Algorithm

As previously mentioned, the new approach follows the method of the traditional PLS-PM, and the algorithm is not an exception. Both algorithm have 8 steps, which are divided into 3 stages. The first stage encompasses the setup of the data, the prerequisites and the initialization of the algorithm. Within the second stage, namely steps two to six, the latent variables scores are estimated following a six-step iterative process. The last stage is composed of the step 7, where the estimates for the structural path coefficients are computed. The steps in which the outer approximation and the factor scores are computed were adjusted to allow the use of a thin-plate regression spline in the algorithm. A detailed explanation of these steps will be presented below, explaining the procedure for both models.

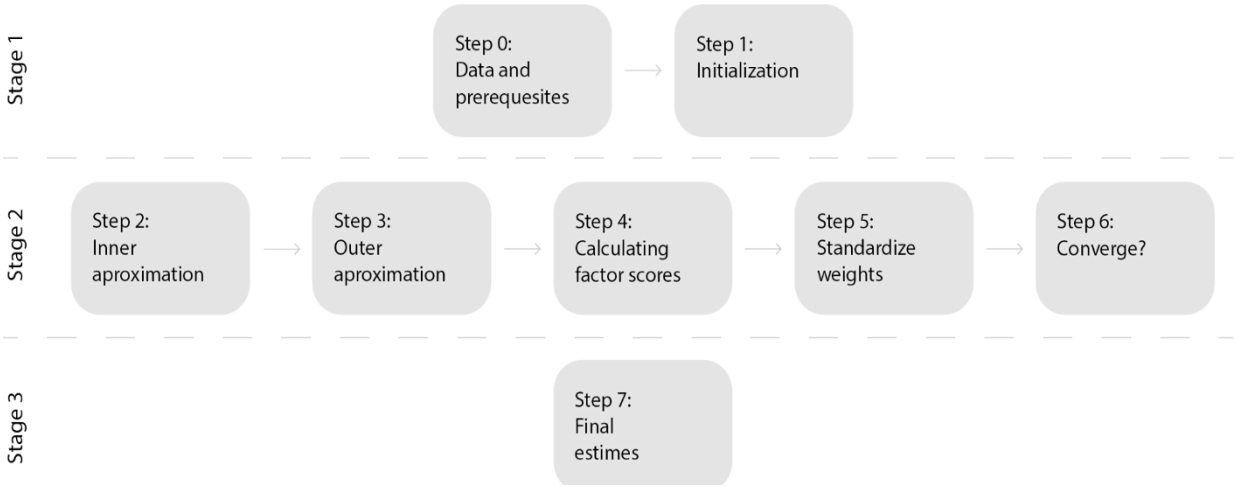

Figure 2 – Algorithm steps

Step 0 (Data and Prerequisites):

The PLS-PM and the P-IRLS-PM are gifted with flexible assumptions, nevertheless should be made an exploratory analysis of the dataset. Such analyses comprise a first examination to the dataset following (Chatfield, 2006) and which distribution the data follows. "Although the PLS-PM is a nonparametric statistical method is important to verify that the data are not too far from normal as extremely non-normal data." (Hair et al., 2014).

After this procedure, the (standardized) manifest variables should be placed in the matrix X. Hereafter, this matrix already presents the standardized data, which will be used as an input to the algorithm.

After defining a subjacent theory for the structural and measurement models, collected the data and the weighting scheme was been chosen is time to pass this information into a matrix form. Starting with the relationship between latent variables described in the structural model, the researcher passes such information into the inner design matrix (matrix D). This task is done by creating a squared triangular matrix with dimensions equal to the number of latent variables, where is placed a 0 when the latent variables are not linked in the structural model, otherwise is placed a 1.

On the other hand, the matrix M describes the measurement model, in columns have the latent variables and in rows the manifest variables, if the latent variable and the manifest variable are linked in the outer model is placed 1 and 0 otherwise. The matrix W, has the same characteristics as the matrix M but the initial weights are replaced by its corresponding estimated outer weights.

Step 1 (Initialization):

Grounded by the measurement model, each LV is initialized based on a weighted sum of their MVs, and can be written as,

$$\widehat{LV}_j = \sum_h m_{jh} MV_{jh} \qquad \mathrm{j} = 1, \dots, \mathrm{g} \ \wedge \ \mathrm{h} = 1, \dots, \mathrm{c} \tag{8}$$

$$\widehat{LV}_j = \frac{\widehat{LV}_j}{\sqrt{VAR(\widehat{LV}_j)}} \qquad \mathrm{j} = 1, \dots, \mathrm{g} \tag{9}$$

The number of MV's in each block is denoted by c, g is the number of LVs, m represents the initial outer weights and $\widehat{LV}$ the latent score obtained from the initialization.

Step 2 (Inner approximation):

In the inner approximation, the LVs are reconstructed by its relationship with its neighboring LVs. The estimation of the inner weights depends on the weighting scheme and on the previous defined inner design matrix:

$$\overline{LV_j} = \sum_i e_{ji} \cdot \widehat{LV_i} \qquad \text{j,i} = 1, \dots, \text{g} \wedge i \neq j \tag{10}$$

$$\overline{LV_j} = \frac{\overline{LV_j}}{\sqrt{VAR(\overline{LV_j})}} \qquad \text{j} = 1, \dots, \text{g} \tag{11}$$

where g symbolizes the number of latent variables, the inner weights are represented by e and $\overline{LV_j}$ is the latent score j obtain from the inner approximation.

Step 3 (Outer approximation):

In the traditional approach, this step is characterized by the search of the best linear combination to express each LV by means of its MVs. For such purpose, the Ordinary Least Squares is used to as the estimator. It is important to highlight the use of OLS makes the model to restrict the search for a linear relationship between each block of variables and its corresponding latent variable. As previously mentioned, the measurement model has two main specifications. For the reflective models is applied a simple linear regression by each MV and its corresponding LV:

$$MV_h = w_h \overline{LV_j} \qquad \text{j} = 1, \dots, \text{g} \ \wedge \ \text{h} = 1, \dots, \text{c} \tag{12}$$

When faced with a formative construct, the traditional model uses a multiple linear regression, with the latent variables as response and its block of manifest variables as regressors, computed by:

$$\overline{LV_j} = \sum_h w_h MV_h \qquad \text{j} = 1, \dots, \text{g} \ \wedge \ \text{h} = 1, \dots, \text{c} \tag{13}$$

Through the OLS estimator, the outer weights (w_h) are obtain based on the standardized latent scores ($\dot{L}\ddot{V}$) from the previous step and its indicators (MV).

For the reflective models, the P-IRLS-PM follows the PLS-PM methodology, however for the formative constructs have a different specification. Instead of applying the linear regression, the P-IRLS-PM uses

a thin-plate regression spline, which is estimated by the penalized iterative re-weighted squares (P-IRLS) and can be expressed by:

$$\overline{LV_j} = \sum_h f_h(MV_h) \qquad \mathrm{j} = 1, \dots, \mathrm{g} \wedge \mathrm{h} = 1, \dots, \mathrm{c} \tag{14}$$

the LV denotes a latent variable, f_h represents the smooth function term of the manifest variable (MV).

Step 4 (Calculating Factor scores):

Like the previous step, the computation of the factor scores, in the P-IRLS-PM, had to be adjusted to allow the use of splines in the formative outer models.

The PLS-PM used the weights estimated in the former step, to reconstruct the latent variables as a weighted sum or a linear combination of its manifest variables. The procedure used is similar to the one applied in the first step, but instead of the initial weights, the outer weights from the fifth step are used. For the formative models, the factor scores are equal to the fitted values obtained from the OLS estimator.

$$\widetilde{LV_j} = \sum_h \widetilde{w}_h \quad MV_h \qquad \mathrm{j} = 1, \dots, \mathrm{g} \wedge h = 1, \dots, c \tag{15}$$

$$\widetilde{LV_j} = \frac{\widetilde{LV_j}}{\sqrt{VAR(\widetilde{LV_j})}} \qquad \mathrm{j} = 1, \dots, \mathrm{g} \tag{16}$$

Where $\widetilde{w}$ denotes the weights computed in the previous steps, g is the number of latent variables and c is the number of manifest variables linked in each latent construct.

The P-IRLS-PM method follows the standard approach for the reflective measurement models and for the formative constructs uses the fitted values obtained in the previous step for computing the factor scores of the formative latent variables, and can be denoted by:

$$\widetilde{LV_j} = \sum_h f_h(\widetilde{MV_h}) \qquad \mathrm{j} = 1, \dots, \mathrm{g} \wedge h = 1, \dots, c \tag{17}$$

$$\widetilde{LV_j} = \frac{\widetilde{LV_j}}{\sqrt{VAR(\widetilde{LV_j})}} \qquad \mathrm{j} = 1, \dots, \mathrm{g} \tag{18}$$

Where $\widetilde{LV}$ represents the fitted values for the latent variable j, f_h represents the smooth function term of the manifest variable h.

Step 5 (Standardize weights):

As has been shown after each step, the latent variables are standardized by dividing for the standard deviation, and this procedure also is applied to the outer weights obtained in the step 4.

Step 6 (Converge?):

The PLS-PM algorithm is designed to run until the outer weights stabilize, therefore the five steps in stage two are repeated until the sum of the outer weights changes between two iterations drops below than a predefined value or when it reaches a maximum number of iterations. In the traditional approach studies (Hair et al., 2014; Hair et al., 2011) recommend the use of the threshold value of 10^{-5} to ensure the algorithm converges and a maximum number of three hundred iterations and the new approach will follow the same recommendations. To compute the tolerance between iterations is used:

$$\left|\frac{\widetilde{w}_{jh}^{old} - \widetilde{w}_{jh}^{new}}{\widetilde{w}_{jh}^{new}}\right| < tolerance \quad \forall\, j = 1, \dots, g \;\wedge\; h = 1, \dots, c \tag{19}$$

The number of MVs in each block is denoted by c, g is the number of LVs and w represents outer weights.

The P-IRLS-PM algorithm follows the same technique and recommendations descried above, but this procedure is applied to the latent scores, computed in the step 4, instead of the outer weights.

Step 7 (Final estimates):

Once the algorithm converges, the path coefficients can be estimated thru the OLS, according to the structural model:

$$\widetilde{LV_j} = \beta_0 + \sum_i \beta_{ji}\widetilde{LV_i} + v_j \qquad i = 1, \dots, n \;\wedge\; i \neq j \tag{20}$$

Where $\widetilde{LV}$ denotes a estimated latent variable, the path coefficients of the structural model are represented by β and the error term is represented by v.

5. COMPARING THE NEW AND THE TRADITIONAL PLS-PM

A Monte Carlo experiment was conducted, when comparing the traditional PLS-PM with P-IRLS-PM*, in order to* bring forth generalized patterns. The purpose of this computational procedure is to elucidate different performance approaches, when formative constructs are faced with non-linear relationships. To conduct the Monte Carlo experiment, an underlying model was defined with the same components of a Partial Least Squares Path Model and the formative measurement embody non-linear relationships. Graphically, this model is represented by the following path diagram(LV represents a latent variable and MV a manifest variable):

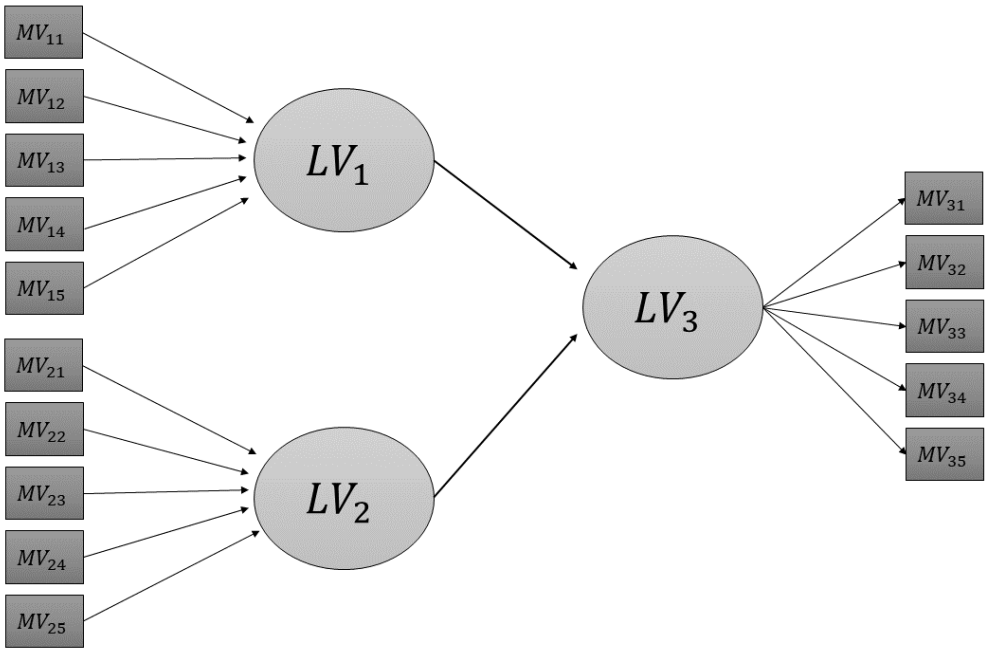

Figure 3 - Path diagram of the underlying model used in the Monte Carlo experiment

The structural model was a linear specification, with two exogenous and one endogenous latent variable and are expressed by:

$$LV_3 = 0.6 \cdot LV_1 + 0.7 \cdot LV_2 + \upsilon \qquad \upsilon \sim N(0\,,\,0.2\,) \tag{21}$$

The measurement model is made by two formative constructs (LV_1 and LV_2) and one reflective construct (LV_3). Each block of formative variables is made up of five manifest variables, which were generated from a normal multivariate distribution. The reflective block of variables has also five indicators. The underlying outer model is described by the following equations:

The formative constructs:

$$LV_1 = \frac{1}{1+e^{-MV_{11}}} + \log(MV_{12}^2) + \sqrt{|MV_{13}|} + MV_{14}^2 + MV_{15} + \delta_1 \qquad \delta_1 \sim N(0\,,0.2\,) \tag{22}$$

$$LV_2 = 3 \cdot cos(MV_{21}) + \frac{1}{2} \cdot sin(MV_{22}) + sin(MV_{23}) \cdot cos(MV_{23}) + cos(2 \cdot MV_{24}) + arctan(MV_{25}) + \delta_2 \tag{23}$$

$$\delta_2 \sim N(0\,,0.2\,)$$

The reflective construct:

$$MV_{31} = 0.8 \cdot LV_3 + \varepsilon_{31} \qquad \varepsilon_{31} \sim N(0\,,0.2\,) \tag{24}$$

$$MV_{32} = 0.8 \cdot LV_3 + \varepsilon_{32} \qquad \varepsilon_{32} \sim N(0\,,0.2\,) \tag{25}$$

$$MV_{33} = 0.8 \cdot LV_3 + \varepsilon_{33} \qquad \varepsilon_{33} \sim N(0\,,0.2\,) \tag{26}$$

$$MV_{34} = 0.8 \cdot LV_3 + \varepsilon_{34} \qquad \varepsilon_{34} \sim N(0\,,0.2\,) \tag{27}$$

$$MV_{35} = 0.8 \cdot LV_3 + \varepsilon_{35} \qquad \varepsilon_{35} \sim N(0\,,0.2\,) \tag{28}$$

Based on the model described above, a population of 10000 observations was generated. From this synthetic dataset, 1000 samples from each subset of 150, 250, 750 and 900 observations were withdrawn for posterior analysis. The traditional PLS-PM and P-IRLS-PM were estimated for all the samples with the outputs being evaluated afterwards. The input used for the computation of both models was based on the original values of the manifest variables. From the outcome of each model, the saved output for this study consisted on the estimated latent scores, the number of iterations until each model converges and the R^2 of the structural model. Moreover, as previously mentioned in the methodology, MSE and the correlation between the estimated and true values were computed for each sample.

METRIC	OBS	P-IRLS-PM	PLS-PM
MSE	150	0.19	1.305
	250	0.143	1.279
	750	0.116	1.278
	900	0.116	1.281
COR	150	0.904	0.345
	250	0.928	0.359
	750	0.942	0.361
	900	0.942	0.359

Table 1 – LV_1 Results

		MSE		COR	
VAR	OBS	P-IRLS-PM	PLS-PM	P-IRLS-PM	PLS-PM
MV11	150	2.027	2.007	-0.020	-0.010
	250	2.067	2.043	-0.037	-0.025
	750	2.112	2.120	-0.058	-0.061
	900	2.123	2.129	-0.063	-0.066
MV12	150	0.750	2.008	0.623	-0.011
	250	0.724	2.036	0.637	-0.022
	750	0.717	2.160	0.641	-0.082
	900	0.716	2.144	0.642	-0.073
MV13	150	1.832	1.938	0.078	0.025
	250	1.815	1.948	0.089	0.022
	750	1.780	2.004	0.109	-0.003
	900	1.783	2.012	0.108	-0.007
MV14	150	1.134	1.884	0.429	0.052
	250	1.101	1.849	0.447	0.072
	750	1.088	1.838	0.455	0.080
	900	1.087	1.843	0.456	0.077
MV15	150	1.878	1.887	0.055	0.050
	250	1.898	1.902	0.047	0.045
	750	1.908	1.910	0.045	0.044
	900	1.912	1.913	0.043	0.043

Table 2 – LV_1 Partial Residuals Results

When inspecting the results concerning LV1 (Table 1), they show that the P-IRLS-PM provided the lowest values of MSE and the strongest correlations between estimated and true values of the latent scores, revealing a higher parameter and prediction accuracy compared to the traditional PLS-PM. Although both approaches exhibit slight changes with the increase of sample size, there is a substantial difference between these two models. In the MSE, the values of PLS-PM are at least 6 times bigger compared to the P-IRLS-PM. Regarding the correlation results, the PLS-PM attained values lower than 0.4 while the new approach showed an improvement reaching values higher than 0.9.

In order to identify the reasons that led to the differences pointed in the previous paragraph, the partial residuals of LV_1 were analyzed and their results shown in table 2. The manifest variables represented by the logistic (MV_{15}) and linear (MV_{11}) functions have very similar results in both models, having the highest errors and lowest correlations, with the variable MV_{11} showing the worst performance. When comparing the two prior variables with MV_{13}, the latter shows an MSE and correlation improvement, presenting better results in P-IRLS-PM than in PLS-PM. The quadratics effect on MV_{14} is the second-best variable within this block of variables, presenting a considerable improvement in terms of the correlation values, passing from 0.07 to 0.45, as well as an increase in parameter accuracy. The two major improvements occur in the logarithm of the squared values of the MV_{12}. Firstly, in the new approach this indicator presents the lowest MSE of its block of variables, half of the value reached by

the traditional approach. Secondly, the prediction accuracy registered a change from very weak to strong correlation values.

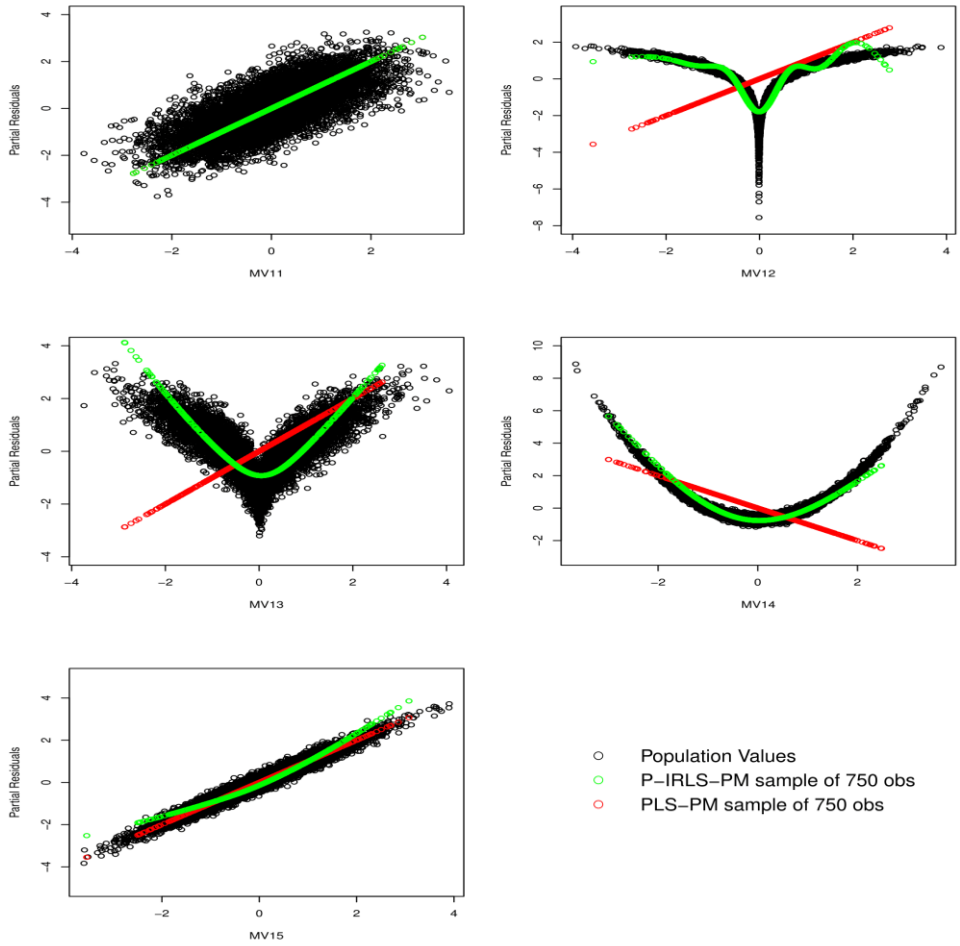

Figure 4 - Parial Residuals Plots of LV_1

The plots of the partial residuals were also inspected to provide more complete analysis. In the graphs presented above, the traditional approach leads to a misleading representation of variables MV_{12}, MV_{13} and MV_{14}. Although the results of the linear function (MV_{15}) and logistic function (MV_{11}) in table 2 exhibit high MSE values comparing to the other indicators of this block of variables, both models display the proper relationships; Furthermore, the P-IRLS-PM shows a noteworthy improvement on how the variables MV_{12}, MV_{13} and MV_{14} are represented, expressing these variables closer to its true relationship.

METRIC	OBS	P-IRLS-PM	PLS-PM
MSE	150	0.264	1.108
	250	0.171	1.08
	750	0.077	1.061
	900	0.069	1.06
COR	150	0.867	0.444
	250	0.914	0.459
	750	0.961	0.469
	900	0.965	0.47

Table 3 – LV_2 Results

		MSE		COR	
VAR	OBS	P-IRLS-PM	PLS-PM	P-IRLS-PM	PLS-PM
MV21	150	0.956	1.956	0.519	0.015
	250	0.929	1.997	0.534	-0.003
	750	0.909	2.081	0.545	-0.042
	900	0.909	2.087	0.545	-0.045
MV22	150	2.222	2.241	-0.119	-0.128
	250	2.265	2.288	-0.137	-0.149
	750	2.346	2.383	-0.175	-0.193
	900	2.351	2.394	-0.177	-0.198
MV23	150	1.953	2.015	0.017	-0.014
	250	1.948	2.048	0.022	-0.028
	750	1.888	2.123	0.055	-0.063
	900	1.888	2.147	0.055	-0.075
MV24	150	1.614	1.977	0.187	0.005
	250	1.547	1.985	0.223	0.004
	750	1.484	2.024	0.257	-0.013
	900	1.485	2.065	0.256	-0.034
MV25	150	2.015	2.038	-0.014	-0.026
	250	2.023	2.045	-0.016	-0.027
	750	2.037	2.053	-0.020	-0.028
	900	2.042	2.058	-0.022	-0.030

Table 4 - LV_2 Partial Residuals Results

The computed results regarding LV_2 (table 3) show that when sample size increases, both models improve theirs results, however the outcomes of P-IRLS-PM achieved a better performance compared to that of PLS-PM. Moreover, the MSE values of the PLS-PM are at least 4 times bigger than the values documented in P-IRLS-PM, while the correlations passed from moderate in the traditional approach to strong values in the new model.

The partial residuals of LV_2 on table 4 were evaluated, following the same procedure applied to LV_1. Within this block, MV_{22} and MV_{25} were the variables that attained the highest MSE values and the lowest prediction accuracy. In the new approach, the variable MV_{23} show a slight improvement in the parameter accuracy. When comparing the traditional model with P-IRLS-PM, the results that stand out the most are related to MV_{24} and MV_{21}. On the new model, these variables presented the lowest MSE and the highest correlation values within its block of variables. The differences observed between both approaches are largely due to the performance of MV_{21} and MV_{24} and in a lesser extent to MV_{23}.

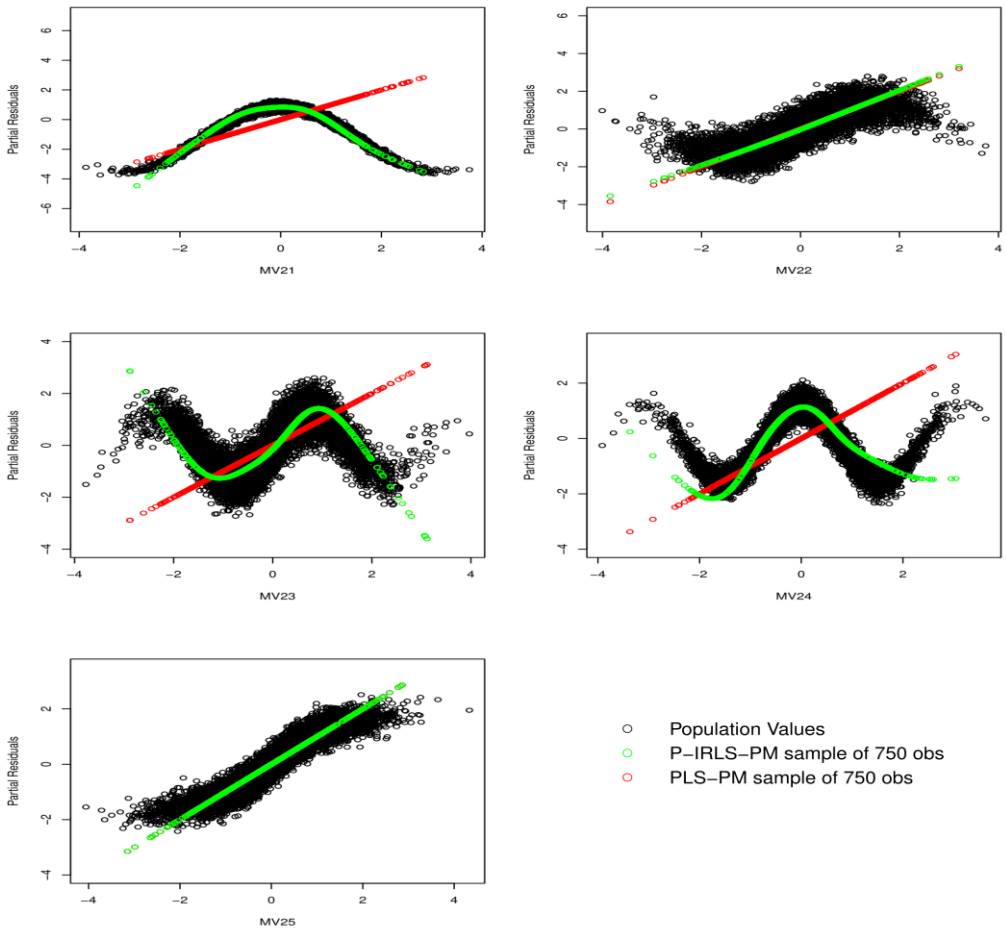

Figure 5 - Parial Residuals Plots of LV_2

A joint analysis of table 4 with the partial residuals plots of LV_2, two clear patterns stand out: when the traditional PLS-PM applies a linear regression to capture these relationships, it tends to provide misleading results, despite the moderate prediction accuracy values in table 3; and the P-IRLS-PM can fit a model with a considerably higher parameter accuracy than the traditional approach, although it presents problems in the tails distribution of MV_{22}, MV_{23}, MV_{24} and MV_{25}. Furthermore, it's also possible to say that MV_{21}, MV_{23} and MV_{24} are well represented when applied to the new algorithm, contrary to the plots of the traditional approach. Despite the moderate results on table 4, MV_{23} shows a substantial improvement on how is represented by the new approach.

		150	250	750	900
R^2	**P-IRLS-PM**	0.868	0.876	0.896	0.898
	PLS-PM	0.212	0.185	0.162	0.16
ITER	**P-IRLS-PM**	8.394	7.514	6.373	6.658
	PLS-PM	3.037	3.003	3	3

Table 5 – Model Fit and Computational Cost

The model fit assessment measured by the R^2 showed that the P-IRLS-PM values are at least 4 times higher than the values recorded for the PLS-PM. The computational cost of the new model is the double of the traditional approach, needing twice the iterations that the traditional approach requires to converge in order to obtain the final estimates.

METRIC	OBS	P-IRLS-PM	PLS-PM
MSE	**150**	0.003	0.003
	250	0.003	0.003
	750	0.003	0.003
	900	0.003	0.003
COR	**150**	0.999	0.999
	250	0.999	0.999
	750	0.999	0.999
	900	0.999	0.999

Table 6 – LV_3 Results

The results of the reflective exogenous variable (LV_3) on table 6 shows that, regardless of the parameter and prediction accuracy on formative constructs or the R^2 of the inner model, the approaches and samples used, had almost indistinguishable results. Taking this into account, it is important to highlight that for the results obtained in the present work, the model or sample size do not lead to a change in how the models capture the exogenous reflective variable.

6. CONCLUSIONS

Since the introduction of the PLS-PM by Herman Wold, the variance approach has experienced several advances, however some topics have been left aside, such is the case of how to model non-linear formative constructs. In order to fill this gap in the literature, the present study introduces the P-IRLS-PM. This model proposes an alternative specification of the formative measurement model of the PLS-PM, with the main objective of improving how the latent variables are represented by its manifest variables, when its relationship is non-linear.

The Monte Carlo experiment used to compare the two models reveals substantial differences between the approaches. The P-IRLS-PM has a better overall performance in terms of parameter accuracy, and prediction accuracy when faced with non-linear relationships, although it has a higher computational cost. Both approaches exhibit similar results, when faced with a linear relationship within the formative constructs framework. By analyzing the partial residuals plots of the formative constructs, has observed that the traditional approach misrepresents the non-linear relationships, and the splines fitted by P-IRLS-PM shown an improvement comparing to the PLS-PM. It was also spotted higher R^2 values of the structural model in the new approach. Despite the previous differences between the two approaches, the reflective exogenous variable has similar results in both models.

Putting together the developments presented through this study with an implementation of the algorithm in an open-source programming language, the present works aims to instigate the scientific community to go one step further and bring new advances into the variance approach.

7. LIMITATIONS AND RECOMMENDATIONS

While this study intrudes a new way to work with non-linear formative indicators for the variance approach, it has its limitations and opens up further avenues for future research.

Since this study is restricted to the selected sample sizes and a single population, further insights may be gained from using different sample sizes and a different populations. To compare the two models conducted a month Carlo experiment, altghouth the achieved results are only valid within the boundaries of the scenarios explore, and they only can be applied to the theoretical model on which the mount Carlo simulation were based, further research could strive to analyze the impact of an alternative design, by using a different number of latent constructs and manifest variables, combined with other non-linear functions that were not used in this study. Another limitation of the present study, concerns that in practical cases often the available data are integrated and/or categorical and this work does not cover this topic.

Future research might also investigate the impact of different P-IRLS-PM architecture settings (e.g. weighting scheme and spline basis) and use different metrics to compare the two models, like the root-mean-square error (RMSE), normalized root-mean-square deviation or the mean absolute relative error (Reinartz et al., 2009)

8. BIBLIOGRAPHY

Anderson, E. W., & Mittal, V. (2000). Strengthening the Satisfaction-Profit Chain. *Journal of Service Research*, *3*(2), 107–120. https://doi.org/10.1177/109467050032001

Becker, J.-M., Klein, K., & Wetzels, M. (2012). Hierarchical Latent Variable Models in PLS-SEM: Guidelines for Using Reflective-Formative Type Models. *Long Range Planning*, *45*(5), 359–394. https://doi.org/10.1016/j.lrp.2012.10.001

Bollen, K. A. (1989). Structural equations with latent variables. *Wiley Series in Probability and Mathematical Statistics*, 528. https://doi.org/10.1007/s13398-014-0173-7.2

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*, 605–627. https://doi.org/10.1146/annurev.psych.53.100901.135239

Chatfield, C. (2006). Initial Data Analysis. In *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471667196.ess0309.pub2

Chin, W. W., & Dibbern, J. (2010). *Handbook of Partial Least Squares*. *Handbook of Partial Least Squares*. https://doi.org/10.1007/978-3-540-32827-8

Chin, W. W., & Newsted, P. R. (1999). Structural Equation Modeling Analysis with Small Samples Using Partial Least Square. In *In Rick Hoyle (Ed.), Statistical Strategies for Small Sample Research, Sage Publications* (pp. 307–341).

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, *17*(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Dijkstra, T. K., & Henseler, J. (2011). Linear indices in nonlinear structural equation models: Best fitting proper indices and other composites. *Quality and Quantity*, *45*(6), 1505–1518. https://doi.org/10.1007/s11135-010-9359-z

Evermann, J., & Tate, M. (2010). Testing Models or Fitting Models? Identifying Model Misspecification in PLS. *International Conference on Information Systems (ICIS)*, 20.

Garson, G. D. (2016). *Partial Least Squares: Regression & Structural Equation Models*.

Gudergan, S. P., Ringle, C. M., Wende, S., & Will, A. (2008). Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research*, *61*(12), 1238–1249. https://doi.org/10.1016/j.jbusres.2008.01.012

Hackl, P., & Westlund, A. H. (2000). On structural equation modelling for customer satisfaction measurement. *Total Quality Management*, *11*(4–6), 820–825. https://doi.org/10.1080/09544120050008264

Hair, J. F. J., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. *Long Range Planning* (Vol. 46). https://doi.org/10.1016/j.lrp.2013.01.002

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2012b). Partial Least Squares: The Better Approach to Structural Equation Modeling? *Long Range Planning*, *45*(5–6), 312–319. https://doi.org/10.1016/j.lrp.2012.09.011

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. *Long Range Planning*, *46*(1–2), 1–12. https://doi.org/10.1016/j.lrp.2013.01.001

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (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*(5–6), 320–340. https://doi.org/10.1016/j.lrp.2012.09.008

Henseler, J., Fassott, G., Dijkstra, T. K., & Wilson, B. (2012). Analysing quadratic effects of formative constructs by means of variance-based structural equation modelling. *European Journal of Information Systems*, *21*(1), 99–112. https://doi.org/10.1057/ejis.2011.36

Hwang, H., Malhotra, N. K., Kum, Y., Tomiuk, M. A., & Hong, S. (2010). A comparative study on parameter recover of three approaches to sturctureral equation modeling. *American Markering Association*, *48*(August), 699–712.

Ingrassia, S., & Trinchera, L. (2008). Some Remarks on Nonlinear Relationships. *Statistica Applicata*, *20*(3–4).

Jakobowicz, E., & Saporta, G. (2002). A non linear PLS path modeling based on monotonic B-splines transformations.

Jöreskog, K. G. (1978). STRUCTURAL ANALYSIS OF COVARIANCE AND CORRELATION MATRICES. *Psychometrika*, *43*(4), 443–477.

Kline, R. B. (2015). *Principles and Practices of Structural Equation Modelling*. *Methodology in the social sciences*.

Lohmöller, J.-B. (1989). *Latent Variable Path Modeling with Partial Least Squares*.

Monecke, A. and Leisch, F. (2012). semPLS : Structural Equation Modeling Using Partial Least Squares. *Journal of Statistical Software*, *48*(3), 1–32.

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, *26*(4), 332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001

Ringle, C. M., Sarstedt, M., & Schlittgen, R. (2010). Response-Based Segmentation Using Finite Mixture Partial Least Squares. Theoretical Foundations and an Application to American Customer Satisfaction Index Data. In *Data Mining, Annals of Information Systems*. https://doi.org/10.1007/978-3-642-01044-6_15

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). A Critical Look at the Use of PLS-SEM in MIS Quarterly. *MIS Quarterly (MISQ)*, *36*(1).

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management*, *47–48*, 9–27. https://doi.org/10.1016/j.jom.2016.05.002

Sanchez, G. (2013). *PLS Path Modeling with R*. *R Package Notes*. https://doi.org/citeulike-article-id:13341888

Sarstedt, M. (2008). A review of recent approaches for capturing heterogeneity in partial least squares path modelling. *Journal of Modelling in Management*, *3*(2), 140–161. https://doi.org/10.1108/17465660810890126

Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research*, *69*(10), 4552–4564. https://doi.org/10.1016/j.jbusres.2016.03.049

Sosik, J. J., Kahai, S. S., & Piovoso, M. J. (2009). Silver Bullet or Voodoo Statistics? A Primer for Using the Partial Least Squares Data Analytic Technique in Group and Organization Research. *Group & Organization Management*, *34*(1), 5–36. https://doi.org/10.1177/1059601108329198

Spearman, C. E. (1904). "General Intelligence ," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292.

Suoniemi, S., Terho, H., & Olkkonen, R. (2012). The Measurement of Endogenous Higher-Order Formative Composite Variables in PLS-SEM: An Empirical Application from CRM System Development. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *6*(12), 1648–1652.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2004). PLS path modeling. *Computational Statistics and Data Analysis*, *48*(1), 159–205. https://doi.org/10.1016/j.csda.2004.03.005

Wold, H. (1975). Soft modelling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach. In *Perspectives in Probability and Statistics* (pp. 117–142).

Wold, H. (1980). Soft Modeling: Intermediate between Traditional Model Building and Data Analysis. *Mathematical Statistics*.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585. https://doi.org/10.1017/CBO9781107415324.004

9. APPENDIX

P-IRLS-PM R script:

```
library(mgcv)

#  Model used:
#  * This script is written to have 3 Latent Variables and 5 Manifest Variables
#     per bock of varibles, but is easily changeable.
#  * 3 Latente variables: -2 Formative variables
#                         -1 Reflective variable
#  * Structural model
#    LV3 = beta1 * LV1 + beta2 * LV2 + error
#  * Measurement model:
#    LV1 = f(MV11)+f(MV12)+f(MV13)+f(MV14)+f(MV15)+ error
#    LV2 = f(MV21)+f(MV22)+f(MV23)+f(MV24)+f(MV25)+ error
#    MV31 = w31 * LV3 + error
#    MV32 = w32 * LV3 + error
#    MV33 = w33 * LV3 + error
#    MV34 = w34 * LV3 + error
#    MV35 = w35 * LV3 + error
#

###################################
# Step 0 (Data and Prerequisites) #
###################################

# Define a data.frame with the Manifes variables:
#   Block_variables_of_lv1 have the 5 manifest variables in the columns
#   dim(data) = 150 rows and 15 columns
#   data <- data.frame( block_variables_of_lv1,
#                       ,block_variables_of_lv2
#                       ,block_variables_of_lv3)
X <- data.frame(data)

# Define the number of observations:
obs <- 150

# Define the number of Latent variables:
numb_lv <- 3

# Define the total number of Manifest variables:
numb_mv <- 15

# W is the initial outer weights matrix:
W <-matrix( c( rep(c(1,0,0), times=5)  # 5 is the number of MVs linked with LV1
              ,rep(c(0,1,0), times=5)  # 5 is the number of MVs linked with LV2
              ,rep(c(0,0,1), times=5)) # 5 is the number of MVs linked with LV3
            , nrow= numb_mv, ncol= numb_lv, byrow = T)


# Create the matrix to store the tolerance values:
t0 <- matrix(rep(1, times = obs),obs, numb_lv)

# D is the Inner Disign matrix:
#  (If LV is linked with other LV put 1, otherwise is 0;
#   And if LV is exogenous put 1)
D <- rbind( c(1,0,1)  # LV1 is a exogenous variable and is linked with LV3
           ,c(0,1,1)  # LV2 is a exogenous variable and is linked with LV3
           ,c(0,0,0)) # LV3 is a endogenous varible
C <- (D + t(D))

X <- scale(X, center = TRUE, scale = TRUE)

tolerance <- 10000
itera <- 0
```

```
# E is the inner weights matrix:
E <- matrix(1, ncol = numb_lv, nrow = numb_lv)

############################
# Step 1 (Initialization) #
############################
Y <- X%*%W
Y <- scale(Y, center = TRUE, scale = TRUE)

while (tolerance > 1e-06){
  R <- cor(Y, method = "pearson") # Used Weighting scheme: Centroid
  for(j in 1:numb_lv){
    for(i in 1:numb_lv){
                         if(C[i,j] == 1){ E[i,j] <- sign(R[i,j])
                         }else{ E[i,j] <- 0}
                        }
                       }
 ################################
 # Step 2 (Inner approximation) #
 ################################
 Yold <- Y
 Y <- Y%*%E
 Y <- scale(Y, center = FALSE, scale = TRUE)

 ################################
 # Step 3 (Outer approximation) #
 ################################
 # For Formative LV1
 fit_lv1 <- gam(Y[,1] ~ s(X[,1], bs="tp")+s(X[,2], bs="tp") # New
                +s(X[,3],bs="tp")+s(X[,4],bs="tp")
                +s(X[,5],bs="tp"))
 # For Formative Latent Variable 2
 fit_lv2 <- gam(Y[,2] ~ s(X[,6], bs="tp")+s(X[,7], bs="tp") # New
               +s(X[,8],bs="tp")+s(X[,9],bs="tp")
               +s(X[,10],bs="tp"))
 # For refletive Latent Variable 3
 # Is used the same approach of traditional PLS-PM
 outer_coef_LV3 <-c( coef(lm(X[,11]~Y[,3]))[2]
                    ,coef(lm(X[,12]~Y[,3]))[2]
                    ,coef(lm(X[,13]~Y[,3]))[2]
                    ,coef(lm(X[,14]~Y[,3]))[2]
                    ,coef(lm(X[,15]~Y[,3]))[2])
 mv3<- X[,c(11:15)]

 ####################################
 #Step 4 (Calculating factor scores) #
 ####################################
 Y1<-  fit_lv1$fitted.values
 Y2<-  fit_lv2$fitted.values
 Y3 <- mv3%*%outer_coef_LV3
 Y <- cbind(Y1,Y2,Y3)
 Y <- scale(Y, center = FALSE, scale = TRUE)

 ####################################
 # Step 5 (Standardize coefficients) #
 ####################################
 outer_coef_LV1 <- fit_lv1$coefficients
 sd_Y1 <- rep(sd(Y[,1]), times= length(outer_coef_LV1))
 outer_coef_LV1 <- outer_coef_LV1/sd_Y1

 outer_coef_LV2 <- fit_lv2$coefficients
 sd_Y2 <- rep(sd(Y[,2]), times= length(outer_coef_LV2))
 outer_coef_LV2 <- outer_coef_LV2/sd_Y2

 outer_coef_LV3
 sd_Y3 <- rep(sd(Y[,3]), times= length(outer_coef_LV3))
 outer_coef_LV3 <- outer_coef_LV3/sd_Y3
```

```
 ######################
 # Step 6 (Converge?) #
 ######################
 for(j in 1:3){for(i in 1:150){t0[i,j] <- abs((Yold[i,j] - Y[i,j])/Y[i,j])}}
 tolerance <- max(t0)
 itera=itera+1
 if(itera > 300 ) break
}

############################
# Step 7 (Final estimates) #
############################
inner_final_fit <- lm(Y[,3] ~ Y[,1]+Y[,2])

r2 <- summary(inner_final_fit)$r.squared

gam_inner_coef <- inner_final_fit$coefficients

p_irls_pm <- list( tolerance, itera, Y
                ,outer_coef_LV1, outer_coef_LV2, outer_coef_LV3
                ,gam_inner_coef
                ,r2)
nomes <- c( "Tolerance", "itera", "Y"
           ,"outer_coef_LV1", "outer_coef_LV2", "outer_coef_LV3"
           ,"gam_inner_coef"
           ,"r2")
names(p_irls_pm) <- nomes

p_irls_pm
```