Chapter 2
PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement

Vincenzo Esposito Vinzi, Laura Trinchera, and Silvano Amato

Abstract In this chapter the authors first present the basic algorithm of PLS Path Modeling by discussing some recently proposed estimation options. Namely, they introduce the development of new estimation modes and schemes for multidimensional (formative) constructs, i.e. the use of PLS Regression for formative indicators, and the use of path analysis on latent variable scores to estimate path coefficients Furthermore, they focus on the quality indexes classically used to assess the performance of the model in terms of explained variances. They also present some recent developments in PLS Path Modeling framework for model assessment and improvement, including a non-parametric GoF-based procedure for assessing the statistical significance of path coefficients. Finally, they discuss the REBUS-PLS algorithm that enables to improve the prediction performance of the model by capturing unobserved heterogeneity. The chapter ends with a brief sketch of open issues in the area that, in the Authors' opinion, currently represent major research challenges.

2.1 Introduction

Structural Equation Models (SEM) (Bollen 1989; Kaplan 2000) include a number of statistical methodologies meant to estimate a network of causal relationships, defined according to a theoretical model, linking two or more latent complex concepts, each measured through a number of observable indicators. The basic idea is that complexity inside a system can be studied taking into account a causality network among latent concepts, called Latent Variables (LV), each measured by

V. Esposito Vinzi
ESSEC Business School of Paris, Department of Information Systems and Decision Sciences, Avenue Bernard Hirsch - B.P. 50105, 95021 Cergy-Pontoise, Cedex, France
e-mail: vinzi@essec.fr

L. Trinchera and S. Amato
Dipartimento di Matematica e Statistica, Università degli Studi di Napoli "Federico II", Via Cintia, 26 - Complesso Monte S. Angelo, 80126 Napoli, Italy
e-mail: ltrinche@unina.it, silvano.amato@gmail.com

V. Esposito Vinzi et al. (eds.), *Handbook of Partial Least Squares*, Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-540-32827-8_3,

several observed indicators usually defined as Manifest Variables (MV). It is in this sense that Structural Equation Models represent a joint-point between Path Analysis (Tukey 1964; Alwin and Hauser 1975) and Confirmatory Factor Analysis (CFA) (Thurstone 1931).

The PLS (Partial Least Squares) approach to Structural Equation Models, also known as PLS Path Modeling (PLS-PM) has been proposed as a component-based estimation procedure different from the classical covariance-based LISREL-type approach. In Wold's (1975a) seminal paper, the main principles of *partial least squares* for *principal component analysis* (Wold 1966) were extended to situations with more than one block of variables. Other presentations of PLS Path Modeling given by Wold appeared in the same year (Wold 1975b, c). Wold (1980) provides a discussion on the theory and the application of Partial Least Squares for path models in econometrics. The specific stages of the algorithm are well described in Wold (1982) and in Wold (1985). Extensive reviews on the PLS approach to Structural Equation Models with further developments are given in Chin (1998) and in Tenenhaus et al. (2005).

PLS Path Modeling is a component-based estimation method (Tenenhaus 2008a). It is an iterative algorithm that separately solves out the blocks of the measurement model and then, in a second step, estimates the path coefficients in the structural model. Therefore, PLS-PM is claimed to explain at best the residual variance of the latent variables and, potentially, also of the manifest variables in any regression run in the model (Fornell and Bookstein 1982). That is why PLS Path Modeling is considered more as an exploratory approach than as a confirmatory one. Unlike the classical covariance-based approach, PLS-PM does not aim at reproducing the sample covariance matrix. PLS-PM is considered as a *soft modeling* approach where no strong assumptions (with respect to the distributions, the sample size and the measurement scale) are required. This is a very interesting feature especially in those application fields where such assumptions are not tenable, at least in full. On the other side, this implies a lack of the classical parametric inferential framework that is replaced by empirical confidence intervals and hypothesis testing procedures based on resampling methods (Chin 1998; Tenenhaus et al. 2005) such as jackknife and bootstrap. It also leads to less ambitious statistical properties for the estimates, e.g. coefficients are known to be biased but consistent at large (Cassel et al. 1999, 2000). Finally, PLS-PM is more oriented to optimizing predictions (explained variances) than statistical accuracy of the estimates.

In the following, we will first present the basic algorithm of PLS-PM by discussing some recently proposed estimation options and by focusing on the quality indexes classically used to assess the performance (usually in terms of explained variances) of the model (Sect. 2.2). Then, we will present a non-parametric GoF-based procedure for assessing the statistical significance of path coefficients (Sect. 2.3.1). Finally, we will present the REBUS-PLS algorithm that enables to improve the prediction performance of the model in presence of unobserved heterogeneity (Sect. 2.4). This chapter ends with a brief sketch of open issues in the area that, in our opinion, currently represent major research challenges (Sect. 2.5).

2.2 PLS Path Modeling: Basic Algorithm and Quality Indexes

2.2.1 The Algorithm

PLS Path Modeling aims to estimate the relationships among Q ($q = 1, \ldots, Q$) blocks of variables, which are expression of unobservable constructs. Essentially, PLS-PM is made of a system of interdependent equations based on simple and multiple regressions. Such a system estimates the network of relations among the latent variables as well as the links between the manifest variables and their own latent variables.

Formally, let us assume P variables ($p = 1, \ldots, P$) observed on N units ($n = 1, \ldots, N$). The resulting data (x_{npq}) are collected in a partitioned data table $\boldsymbol{X}$:

$$\boldsymbol{X} = \left[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q\right]$$

where $\boldsymbol{X}_q$ is the generic q-th block made of P_q variables.

As well known, each Structural Equation Model is composed by two sub-models: the measurement model and the structural model. The first one takes into account the relationships between each latent variable and the corresponding manifest variables, while the structural model takes into account the relationships among the latent variables.

In the PLS Path Modeling framework, the structural model can be written as:

$$\boldsymbol{\xi}_j = \beta_{0j} + \sum_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j} \beta_{qj}\boldsymbol{\xi}_q + \boldsymbol{\zeta}_j \tag{2.1}$$

where $\boldsymbol{\xi}_j$ ($j = 1, \ldots, J$) is the generic endogenous latent variable, β_{qj} is the generic path coefficient interrelating the q-th exogenous latent variable to the j-th endogenous one, and $\boldsymbol{\zeta}_j$ is the error in the inner relation (i.e. disturbance term in the prediction of the j-th endogenous latent variable from its explanatory latent variables).

The measurement model formulation depends on the direction of the relationships between the latent variables and the corresponding manifest variables (Fornell and Bookstein 1982). As a matter of fact, different types of measurement model are available: the *reflective model* (or outwards directed model), the *formative model* (or inwards directed model) and the *MIMIC model* (a mixture of the two previous models).

In a *reflective model* the block of manifest variables related to a latent variable is assumed to measure a unique underlying concept. Each manifest variable reflects (is an effect of) the corresponding latent variable and plays a role of endogenous variable in the block specific measurement model. In the reflective measurement model, indicators linked to the same latent variable should covary: changes in one indicator imply changes in the others. Moreover, internal consistency has to be checked, i.e. each block is assumed to be homogeneous and unidimensional. It is important to

notice that for the *reflective models*, the measurement model reproduces the factor analysis model, in which each variable is a function of the underlying factor. In more formal terms, in a *reflective model* each manifest variable is related to the corresponding latent variable by a simple regression model, i.e.:

$$\boldsymbol{x}_{pq} = \lambda_{p0} + \lambda_{pq}\boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq} \tag{2.2}$$

where λ_{pq} is the loading associated to the p-th manifest variable in the q-th block and the error term $\boldsymbol{\epsilon}_{pq}$ represents the imprecision in the measurement process. Standardized loadings are often preferred for interpretation purposes as they represent correlations between each manifest variable and the corresponding latent variable.

An assumption behind this model is that the error $\boldsymbol{\epsilon}_{pq}$ has a zero mean and is uncorrelated with the latent variable of the same block:

$$E(\boldsymbol{x}_{pq}|\boldsymbol{\xi}_q) = \lambda_{p0} + \lambda_{pq}\boldsymbol{\xi}_q. \tag{2.3}$$

This assumption, defined as *predictor specification*, assures desirable estimation properties in classical Ordinary Least Squares (OLS) modeling.

As the *reflective* block reflects the (unique) latent construct, it should be homogeneous and *unidimensional*. Hence, the manifest variables in a block are assumed to measure the same unique underlying concept. There exist several tools for checking the block homogeneity and unidimensionality:

(a) *Cronbach's alpha*: this is a classical index in reliability analysis and represents a strong tradition in the SEM community as a measure of internal consistency. A block is considered homogenous if this index is larger than 0.7 for confirmatory studies. Among several alternative and equivalent formulas, this index can be expressed as:

$$\alpha = \frac{\sum_{p\neq p'} cor(\boldsymbol{x}_{pq}, \boldsymbol{x}_{p'q})}{P_q + \sum_{p\neq p'} cor(\boldsymbol{x}_{pq}, \boldsymbol{x}_{p'q})} \times \frac{P_q}{P_q - 1} \tag{2.4}$$

where P_q is the number of manifest variables in the q-th block.

(b) *Dillon-Goldstein's* (or *Jöreskog's*) *rho* (Wertz et al. 1974) better known as *composite reliability*: a block is considered homogenous if this index is larger than 0.7

$$\rho = \frac{(\sum_{p=1}^{P_q} \lambda_{pq})^2}{(\sum_{p=1}^{P_q} \lambda_{pq})^2 + \sum_{p=1}^{P_q}(1 - \lambda_{pq}^2)}. \tag{2.5}$$

(c) *Principal component analysis of a block*: a block may be considered unidimensional if the first eigenvalue of its correlation matrix is higher than 1, while the others are smaller (Kaiser's rule). A bootstrap procedure can be implemented to assess whether the eigenvalue structure is significant or rather due to sampling fluctuations. In case unidimensionality is rejected, eventual

groups of unidimensional sub-blocks can be identified by referring to patterns of variable-factor correlations displayed on the loading plots.

According to Chin (1998), *Dillon-Goldstein's rho* is considered to be a better indicator than *Cronbach's alpha*. Indeed, the latter assumes the so-called tau equivalence (or parallelity) of the manifest variables, i.e. each manifest variable is assumed to be equally important in defining the latent variable. *Dillon-Goldstein's rho* does not make this assumption as it is based on the results from the model (i.e. the loadings) rather than the correlations observed between the manifest variables in the dataset. *Cronbach's alpha* actually provides a lower bound estimate of reliability.

In the *formative model* , each manifest variable or each sub-block of manifest variables represents a different dimension of the underlying concept. Therefore, unlike the reflective model, the formative model does not assume homogeneity nor unidimensionality of the block. The latent variable is defined as a linear combination of the corresponding manifest variables, thus each manifest variable is an exogenous variable in the measurement model. These indicators need not to covary: changes in one indicator do not imply changes in the others and internal consistency is no more an issue. Thus the measurement model could be expressed as:

$$\boldsymbol{\xi}_q = \sum_{p=1}^{P_q} \omega_{pq} \boldsymbol{x}_{pq} + \boldsymbol{\delta}_q \tag{2.6}$$

where ω_{pq} is the coefficient linking each manifest variable to the corresponding latent variable and the error term $\boldsymbol{\delta}_q$ represents the fraction of the corresponding latent variable not accounted for by the block of manifest variables. The assumption behind this model is the following *predictor specification*:

$$E(\boldsymbol{\xi}_q | \boldsymbol{x}_{pq}) = \sum_{p=1}^{P_q} \omega_{pq} \boldsymbol{x}_{pq}. \tag{2.7}$$

Finally, the *MIMIC model* is a mixture of both the reflective and the formative models within the same block of manifest variables.

Independently from the type of measurement model, upon convergence of the algorithm, the standardized latent variable scores ($\hat{\boldsymbol{\xi}}_q$) associated to the q-th latent variable ($\boldsymbol{\xi}_q$) are computed as a linear combination of its own block of manifest variables by means of the so-called *weight relation* defined as:

$$\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} w_{pq} \boldsymbol{x}_{pq} \tag{2.8}$$

where the variables $\boldsymbol{x}_{pq}$ are centred and w_{pq} are the outer weights. These weights are yielded upon convergence of the algorithm and then transformed so as to produce standardized latent variable scores. However, when all manifest variables are

observed on the same measurement scale and all outer weights are positive, it is interesting and feasible to express these scores in the original scale (Fornel 1992). This is achieved by using normalized weights $\tilde{w}_{pq}$ defined as:

$$\tilde{w}_{pq} = \frac{w_{pq}}{\sum_{p=1}^{P_q} w_{pq}} \text{ with } \sum_{p=1}^{P_q} \tilde{w}_{pq} = 1 \ \forall q : P_q > 1. \tag{2.9}$$

It is very important not to confound the *weight relation* defined in (2.8) with a *formative model*. The *weight relation* only implies that, in PLS Path Modeling, any latent variable is defined as a weighted sum of its own manifest variables. It does not affect the direction of the relationship between the latent variable and its own manifest variables in the outer model. Such a direction (inwards or outwards) determines how the weights used in (2.8) are estimated.

In PLS Path Modeling an iterative procedure permits to estimate the outer weights (w_{pq}) and the latent variable scores ($\widehat{\boldsymbol{\xi}}_q$). The estimation procedure is named *partial* since it solves blocks one at a time by means of alternating single and multiple linear regressions. The path coefficients (β_{qj}) are estimated afterwards by means of a regular regression between the estimated latent variable scores in accordance with the specified network of structural relations. Taking into account the regression framework of PLS Path Modeling, we prefer to think of such a network as defining a predictive path model for the endogenous latent variables rather than a causality network. Indeed, the emphasis is more on the accuracy of predictions than on the accuracy of estimation.

The estimation of the outer weights is achieved through the alternation of the *outer* and the *inner* estimation steps, iterated till convergence. It is important to underline that no formal proof of convergence of this algorithm has been provided until now for models with more than two blocks. Nevertheless, empirical convergence is usually observed in practice.

The procedure works on centred (or standardized) manifest variables and starts by choosing arbitrary initial weights w_{pq}. Then, in the outer estimation stage, each latent variable is estimated as a linear combination of its own manifest variables:

$$\boldsymbol{v}_q \propto \pm \sum_{p=1}^{P_q} w_{pq} \boldsymbol{x}_{pq} = \pm \boldsymbol{X}_q \boldsymbol{w}_q \tag{2.10}$$

where $\boldsymbol{v}_q$ is the standardized (zero mean and unitary standard deviation) outer estimate of the q-th latent variable $\boldsymbol{\xi}_q$, the symbol $\propto$ means that the left side of the equation corresponds to the standardized right side and the "$\pm$" sign shows the sign ambiguity. This ambiguity is usually solved by choosing the sign making the outer estimate positively correlated to a majority of its manifest variables. Anyhow, the user is allowed to invert the signs of the weights for a whole block in order to make them coherent with the definition of the latent variable.

In the inner estimation stage, each latent variable is estimated by considering its links with the other Q' adjacent latent variables:

$$\boldsymbol{\vartheta}_q \propto \sum_{q'=1}^{Q'} e_{qq'} \boldsymbol{v}_{q'} \tag{2.11}$$

where $\boldsymbol{\vartheta}_q$ is the standardized inner estimate of the q-th latent variable $\boldsymbol{\xi}_q$ and each inner weight ($e_{qq'}$) is equal (in the so-called *centroid scheme*) to the sign of the correlation between the outer estimate $\boldsymbol{v}_q$ of the q-th latent variable and the outer estimate of the q' latent variable $\boldsymbol{v}_{q'}$ connected with $\boldsymbol{v}_q$. Inner weights can be obtained also by means of other schemes than the centroid one. Namely, the three following schemes are available:

1. *Centroid scheme* (the Wold's original scheme): take the sign of the correlation between the outer estimate $\boldsymbol{v}_q$ of the q-th latent variable and the outer estimate $\boldsymbol{v}_{q'}$ connected with $\boldsymbol{v}_q$.
2. *Factorial scheme* (proposed by Lohmöller): take the correlation between the outer estimate $\boldsymbol{v}_q$ of the q-th latent variable and the outer estimate $\boldsymbol{v}_{q'}$ connected with $\boldsymbol{v}_q$.
3. *Structural or path weighting scheme*: take the regression coefficient between $\boldsymbol{v}_q$ and the $\boldsymbol{v}_{q'}$ connected with $\boldsymbol{v}_q$ if $\boldsymbol{v}_q$ plays the role of dependent variable in the specific structural equation, or take the correlation coefficient in case it is a predictor.

Even though the path weighting scheme seems the most coherent with the direction of the structural relations between latent variables, the centroid scheme is very often used as it adapts well to cases where the manifest variables in a block are strongly correlated to each other. The factorial scheme, instead, is better suited to cases where such correlations are weaker. In spite of different common practices, we strongly advice to use the path weighting scheme. Indeed, this is the only estimation scheme that explicitly considers the direction of relationships as specified in the predictive path model.

Once a first estimate of the latent variables is obtained, the algorithm goes on by updating the outer weights w_{pq}.

Two different *modes* are available to update the outer weights. They are closely related to, but do not coincide with, the *formative* and the *reflective* modes:

- *Mode A* : each outer weight w_{pq} is updated as the regression coefficient in the simple regression of the p-th manifest variable of the q-th block ($\boldsymbol{x}_{pq}$) on the inner estimate of the q-th latent variable $\boldsymbol{\vartheta}_q$. As a matter of fact, since $\boldsymbol{\vartheta}_q$ is standardized, the generic outer weight w_{pq} is obtained as:

$$w_{pq} = cov\left(\boldsymbol{x}_{pq}, \boldsymbol{\vartheta}_q\right) \tag{2.12}$$

  i.e. the regression coefficient reduces to the covariance between each manifest variable and the corresponding inner estimate of the latent variable. In case the manifest variables have been also standardized, such a covariance becomes a correlation.

- *Mode B* : the vector $\boldsymbol{w}_q$ of the weights w_{pq} associated to the manifest variables of the q-th block is updated as the vector of the regression coefficients in the multiple regression of the inner estimate of the q-th latent variable $\boldsymbol{\vartheta}_q$ on the manifest variables in $\boldsymbol{X}_q$:

$$\boldsymbol{w}_q = \left(\boldsymbol{X}_q'\boldsymbol{X}_q\right)^{-1}\boldsymbol{X}_q'\boldsymbol{\vartheta}_q \tag{2.13}$$

  where $\boldsymbol{X}_q$ comprises the P_q manifest variables $\boldsymbol{x}_{pq}$ previously centred and scaled by $\sqrt{1/N}$.

As already said, the choice of the outer weight estimation mode is strictly related to the nature of the measurement model. For a *reflective (outwards directed) model* the *Mode A* is more appropriate, while *Mode B* is better for a *formative (inwards directed) model*. Furthermore, *Mode A* is suggested for endogenous latent variables, while *Mode B* for the exogenous ones.

In case of a one-block PLS model, *Mode A* leads to the same results (i.e. outer weights, loadings and latent variable scores) as for the first standardized principal component in a Principal Component Analysis (PCA). This reveals the reflective nature of PCA that is known to look for components (weighted sums) explaining the corresponding manifest variables at best. Instead, *Mode B* coherently provides an indeterminate solution when applied to a one-block PLS model. Indeed, without an inner model, any linear combination of the manifest variables is perfectly explained by the manifest variables themselves.

It is worth noticing that *Mode B* may be affected by multicollinearity between manifest variables belonging to the same block. If this happens, PLS regression (Tenenhaus 1998; Wold et al. 1983) may be used as a more stable and better interpretable alternative to OLS regression to estimate outer weights in a formative model, thus defining a *Mode PLS* (Esposito Vinzi 2008, 2009; Esposito Vinzi and Russolillo 2010). This mode is available in the PLSPM module of the XLSTAT software [1] (Addinsoft 2009). As a matter of fact, it may be noticed that Mode A consists in taking the first component from a PLS regression, while Mode B takes all PLS regression components (and thus coincides with OLS multiple regression). Therefore, running a PLS regression and retaining a certain number (that may be different for each block) of significant PLS components is meant as an intermediate

[1] XLSTAT-PLSPM is the ultimate PLS Path Modeling software implemented in XLSTAT (http://www.xlstat.com/en/products/xlstat-plspm/), a data analysis and statistical solution for Microsoft Excel. XLSTAT allows using the PLS approach (both PLS Path modeling and PLS regression) without leaving Microsoft Excel. Thanks to an intuitive and flexible interface, XLSTAT-PLSPM permits to build the graphical representation of the model, then to fit the model, to display the results in Excel either as tables or graphical views. As XLSTAT-PLSPM is totally integrated with the XLSTAT suite, it is possible to further analyze the results with the other XLSTAT features. Apart from the classical and fundamental options of PLS Path Modeling, XLSTAT-PLSPM comprises several advanced features and implements the most recent methodological developments.

mode between Mode A and Mode B. This new Mode PLS adapts well to formative models where the blocks are multidimensional but with fewer dimensions than the number of manifest variables.

The PLS Path Modeling algorithm alternates the outer and the inner estimation stages by iterating till convergence. Up to now convergence has been proved only for path diagrams with one or two blocks (Lyttkens et al. 1975). However, for multiblock models, convergence is practically always encountered in practice.

Upon convergence, the estimates of the latent variable scores are obtained according to 2.8. Thus, PLS Path Modeling provides a direct estimate of the latent variable individual scores as aggregates of manifest variables that naturally involve measurement error. The price of obtaining these scores is the inconsistency of the estimates.

Finally, structural (or path) coefficients are estimated through OLS multiple/simple regressions among the estimated latent variable scores. PLS regression can nicely replace OLS regression for estimating path coefficients whenever one or more of the following problems occur: missing latent variable scores, strongly correlated latent variables, a limited number of units as compared to the number of predictors in the most complex structural equation. A PLS regression option for path coefficients is implemented in the PLSPM module of the XLSTAT software (Addinsoft 2009). This option permits to choose a specific number of PLS components for each endogenous latent variable.

A schematic description of the PLS Path Modeling algorithm by Löhmoller (with specific options for the sake of brevity) is provided in Algorithm 1. This is the best known procedure for the computation of latent variable scores and it is the one implemented in the PLSPM module of the XLSTAT software. There exists a second and less known procedure initially proposed in Wold (1985). The Löhmoller's procedure is more advantageous and easier to implement. However, the Wold's procedure seems to be more interesting for proving convergence properties of the PLS algorithm as it is monotonically convergent (Hanafi 2007). Indeed, at present PLS Path Modeling is often blamed not to optimize a well identified global scalar function. However, very promising researches on this topic are on going and interesting results are expected soon (Tenenhaus 2008b; Tenenhaus and Tenenhaus 2009).

In Lohmöller (1987) and in Lohmöller (1989) Wold's original algorithm was further developed in terms of options and mathematical proprieties. Moreover, in Tenenhaus and Esposito Vinzi (2005) new options for computing both inner and outer estimates were implemented together with a specific treatment for missing data and multicollinearity while enhancing the data analysis flavour of the PLS approach and its presentation as a general framework to the analysis of multiple tables.

A comprehensive application of the PLS Path Modeling algorithm to real data will be presented in Sect. 2.4.2 after dealing with the problem of capturing unobserved heterogeneity for improving the model prediction performance.

Algorithm 1 : PLS Path Modeling based on Löhmoller's algorithm with the following options: centroid scheme, standardized latent variable scores, OLS regressions

Input: $X = [X_1, \ldots, X_q, \ldots, X_Q]$, i.e. Q blocks of centred manifest variables;

Output: $w_q, \hat{\xi}_q, \beta_j$;

1: **for all** $q = 1, \ldots, Q$ **do**

2: initialize w_q

3: $v_q \propto \pm \sum_{p=1}^{P_q} w_{pq} x_{pq} = \pm X_q w_q$

4: $e_{qq'} = sign\left[cor\left(v_q, v_{q'}\right)\right]$ following the centroid scheme

5: $\vartheta_q \propto \sum_{q'=1}^{Q'} e_{qq'} v_{q'}$

6: update w_q :

(a) $w_{pq} = cov(x_{pq}, \vartheta_q)$ for Mode A (outwards directed model)

(b) $w_q = \left(\frac{X_q' X_q}{N}\right)^{-1} \left(\frac{X_q' \vartheta_q}{N}\right)$ for Mode B (inwards directed model)

7: **end for**

8: **Steps 1–7 are repeated until convergence** on the outer weights is achieved, i.e. until:

$$max\{w_{pq,\text{current iteration}} - w_{pq,\text{previous iteration}}\} < \Delta$$

where Δ is a convergence tolerance usually set at 0.0001 or less

9: **Upon convergence**:

(1) for each block the standardized latent variable scores are computed as weighted aggregates of manifest variables:

$$\hat{\xi}_q \propto X_q w_q,$$

(2) for each endogenous latent variable ξ_j ($j = 1, \ldots, J$), the vector of path coefficients is estimated by means of OLS regression as:

$$\beta_j = \left(\hat{\Xi}' \hat{\Xi}\right)^{-1} \hat{\Xi}' \hat{\xi}_j,$$

where $\hat{\Xi}$ includes the scores of the latent variables that explain the j-th endogenous latent variable ξ_j, and $\hat{\xi}_j$ is the latent variable score of the j-th endogenous latent variable

2.2.2 The Quality Indexes

PLS Path Modeling lacks a well identified global optimization criterion so that there is no *global fitting function* to assess the goodness of the model. Furthermore, it is a variance-based model strongly oriented to prediction. Thus, model validation mainly focuses on the model predictive capability. According to PLS-PM structure, each part of the model needs to be validated: the *measurement model*, the *structural model* and the overall model. That is why, PLS Path Modeling provides three different fit indexes: the *communality* index, the *redundancy* index and the *Goodness of Fit* (*GoF*) index.

For each q-th block in the model with more than one manifest variable (i.e. for each block with $P_q > 1$) the quality of the measurement model is assessed by means of the *communality* index:

$$Com_q = \frac{1}{P_q} \sum_{p=1}^{P_q} cor^2 \left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q \right) \forall q : P_q > 1. \tag{2.14}$$

This index measures how much of the manifest variables variability in the q-th block is explained by their own latent variable scores $\hat{\boldsymbol{\xi}}_q$. Moreover, the communality index for the q-th block is nothing but the average of the squared correlations (squared loadings in case of standardized manifest variables) between each manifest variable in the q-th block and the corresponding latent variable scores.

It is possible to assess the quality of the whole measurement model by means of the *average communality* index, i.e:

$$\overline{Com} = \frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} P_q Com_q. \tag{2.15}$$

This is a weighted average of all the Q block-specific *communality* indexes (see (2.14)) with weights equal to the number of manifest variables in each block. Moreover, since the *communality* index for the q-th block is nothing but the average of the squared correlation in the block, then the *average communality* is the average of all the squared correlations between each manifest variable and the corresponding latent variable scores in the model, i.e.:

$$\overline{Com} = \frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} cor^2 \left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q \right). \tag{2.16}$$

Let us focus now on the structural model. Although the quality of each structural equation is measured by a simple evaluation of the R^2 fit index, this is not sufficient to evaluate the whole structural model. Specifically, since the structural equations are estimated once the convergence is achieved and he latent variable scores are estimated, then the R^2 values only take into account the fit of each regression equation in the structural model.

It would be a wise choice to replace this current practice by a path analysis on the latent variable scores considering all structural equations simultaneously rather than as independent regressions. We see two advantages in this proposal: the path coefficients would be estimated by optimizing a single discrepancy function based on the difference between the observed covariance matrix of the latent variable scores and the same covariance matrix implied by the model; the structural model could be assessed as a whole in terms of a chi-square test related to the optimized discrepancy function. We have noticed, through several applications, that such a procedure does not actually change the prediction performance of the model in terms of explained

variances for the endogenous latent variables. Up to now, no available software has implemented the path analysis option in a PLS-PM framework.

In view of linking the prediction performance of the measurement model to the structural one, the *redundancy* index computed for the j-th endogenous block, measures the portion of variability of the manifest variables connected to the j-th endogenous latent variable explained by the latent variables directly connected to the block, i.e.:

$$Red_j = Com_j \times R^2\left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\xi_q \to \xi_j}\right). \tag{2.17}$$

A global quality measure of the structural model is also provided by the *average redundancy* index, computed as:

$$\overline{Red} = \frac{1}{J}\sum_{j=1}^{J} Red_j \tag{2.18}$$

where J is the total number of endogenous latent variables in the model.

As aforementioned, there is no overall fit index in PLS Path Modeling. Nevertheless, a global criterion of goodness of fit has been proposed by Tenenhaus et al. (2004): the GoF index. Such an index has been developed in order to take into account the model performance in both the measurement and the structural model and thus provide a single measure for the overall prediction performance of the model. For this reason the GoF index is obtained as the geometric mean of the *average communality* index and the average R^2 value:

$$GoF = \sqrt{\overline{Com} \times \overline{R^2}} \tag{2.19}$$

where the average R^2 value is obtained as:

$$\overline{R^2} = \frac{1}{J} R^2\left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\xi_q \to \xi_j}\right). \tag{2.20}$$

As it is partly based on average communality, the GoF index is conceptually appropriate whenever measurement models are reflective. However, communalities may be also computed and interpreted in case of formative models knowing that, in such a case, we expect lower communalities but higher R^2 as compared to reflective models. Therefore, for practical purposes, the GoF index can be interpreted also with formative models as it still provides a measure of overall fit.

According to (2.16) and (2.20) the GoF index can be rewritten as:

$$GoF = \sqrt{\frac{\sum_{q:P_q>1}\sum_{p=1}^{P_q} Cor^2\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right)}{\sum_{q:P_q>1} P_q} \times \frac{\sum_{j=1}^{J} R^2\left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\xi_q \to \xi_j}\right)}{J}}. \tag{2.21}$$

A normalized version is obtained by relating each term in (2.21) to the corresponding maximum value. In particular, it is well known that in principal component analysis the best rank one approximation of a set of variables $\boldsymbol{X}$ is given by the eigenvector associated to the largest eigenvalue of the $\boldsymbol{X'X}$ matrix. Furthermore, the sum of the squared correlations between each variable and the first principal component of $\boldsymbol{X}$ is a maximum.

Therefore, if data are mean centred and with unit variance, the left term under the square root in (2.21) is such that $\sum_{p=1}^{P_q} cor^2\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right) \leq \lambda^1_{(q)}$, where $\lambda^1_{(q)}$ is the first eigenvalue obtained by performing a Principal Component Analysis on the q-th block of manifest variables. Thus, the normalized version of the first term of the GoF is obtained as:

$$T_1 = \frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \frac{\sum_{p=1}^{P_q} cor^2\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right)}{\lambda^1_{(q)}}. \tag{2.22}$$

In other words, here the sum of the communalities in each block is divided by the first eigenvalue of the block itself.

As concerning the right term under the square root in (2.19), the normalized version is obtained as:

$$T_2 = \frac{1}{J} \sum_{j=1}^{J} \frac{R^2\left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\xi_q \to \xi_j}\right)}{\rho_j^2} \tag{2.23}$$

where ρ_j is the first canonical correlation of the canonical analysis between $\boldsymbol{X}_j$ containing the manifest variables associated to the j-th endogenous latent variable, and a matrix containing the manifest variables associated to all the latent variables explaining $\boldsymbol{\xi}_j$.

Thus, according to (2.21), (2.22) and (2.23), the relative GoF index is:

$$GoF_{rel} = \sqrt{\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \frac{\sum_{p=1}^{P_q} Cor^2\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right)}{\lambda^1_{(q)}} \times \frac{1}{J} \sum_{j=1}^{J} \frac{R^2\left(\hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_{q:\xi_q \to \xi_j}\right)}{\rho_j^2}}. \tag{2.24}$$

This index is bounded between 0 and 1. Both the GoF and the relative GoF are descriptive indexes, i.e. there is no inference-based threshold to judge the statistical significance of their values. As a rule of thumb, a value of the relative GoF equal to or higher than 0.90 clearly speaks in favour of the model.

As PLS Path Modeling is a *soft modeling* approach with no distributional assumptions, it is possible to estimate the significance of the parameters trough cross-validation methods like jack-knife and bootstrap (Efron and Tibshirani 1993). Moreover, it is possible to build a cross-validated version of all the quality indexes

(i.e. of the *communality* index, of the *redundancy* index, and of the GoF index) by means of a *blindfolding* procedure (Chin 1998; Lohmöller 1989).

Bootstrap confidence intervals for both the absolute and the relative Goodness of Fit Indexes can be computed. In both cases the inverse cumulative distribution function (*cdf*) of the GoF ($\boldsymbol{\Phi}_{GoF}$) is approximated using a bootstrap-based procedure. B (usually > 100) re-samples are drawn from the initial dataset of N units defining the bootstrap population. For each of the B re-samples, the GoF^b index is computed, with $b = 1 \cdots B$. The values of GoF^b are then used for computing the Monte Carlo approximation of the inverse *cdf*, $\boldsymbol{\Phi}^B_{GoF}$. Thus, it is possible to compute the bounds of the empirical confidence interval from the bootstrap distribution at the $(1-\alpha)$ confidence level by using the percentiles as:

$$\left[\boldsymbol{\Phi}^B_{GoF}\left(\alpha/2\right), \boldsymbol{\Phi}^B_{GoF}\left(1-\alpha/2\right)\right]. \tag{2.25}$$

Several applications have shown that the variability of the GoF values is mainly due to the inner model while the outer model contribution to GoF is very stable across the different bootstrap re-samples.

2.3 Prediction-Based Model Assessment

In this section we present a non-parametric GoF-based bootstrap validation procedure for assessing the statistical significance of path coefficients (individually or by sub-sets).

In order to simplify the discussion we will refer to a very simple model with only three latent variables: $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ (see Fig. 2.1). The structural relations defined in Fig. 2.1 are formalized by the following equations:

$$\begin{aligned} \boldsymbol{\xi}_2 &= \beta_{02} + \beta_{12}\boldsymbol{\xi}_1 + \zeta_2 \\ \boldsymbol{\xi}_3 &= \beta_{03} + \beta_{13}\boldsymbol{\xi}_1 + \beta_{23}\boldsymbol{\xi}_2 + \zeta_3 \end{aligned} \tag{2.26}$$

where β_{qj} $(q = 1, 2$ and $j = 2, 3)$ stands for the path coefficient linking the q-th latent variable to the j-th endogenous latent variable, and ζ_j is the error term associated to each endogenous latent variable in the model.

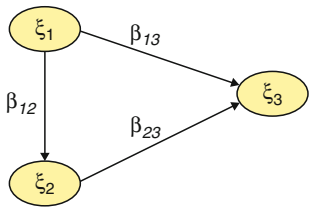

Fig. 2.1 Path diagram of the structural model specified in (2.26)

Equation (2.26) defines a structural model with only three latent variables and with three structural paths. In the following, first we present a non-parametric inferential procedure based on the GoF index to assess the statistical significance of a single path coefficient (Sect. 2.3.1). Then, we discuss the case of an omnibus test on all the path coefficients or on sub-sets of theirs (Sect. 2.3.2).

2.3.1 Hypothesis Testing on One Path Coefficient

Here we want to test if a generic path coefficient β_{qj} is different from 0, i.e.

$$\begin{aligned} \mathrm{H}_0 &: \beta_{qj} = 0 \\ \mathrm{H}_1 &: \beta_{qj} \neq 0 \end{aligned} \tag{2.27}$$

The null hypothesis of $\beta_{qj} = 0$ is tested against the alternative hypothesis that $\beta_{qj} \neq 0$, thus a two-tailed test is performed.

In order to perform this hypotesis testing procedure, we need to define a proper test statistic and the corresponding distribution under the null hypothesis. In particular, the GoF index will be used to test the hypotheses set in (2.27), while the corresponding distribution under the null hypothesis will be obtained by using a bootstrap procedure.

Let GoF_{H_0} be the GoF value under the null hypothesis, $\boldsymbol{\Phi}$ be the inverse cumulative distribution function (cdf) of the GoF_{H_0}, F be the cdf of $\boldsymbol{X}$, and $\boldsymbol{\Phi}^{(B)}$ be the B-sample bootstrap approximation of $\boldsymbol{\Phi}$. In order to approximate $\boldsymbol{\Phi}$ by means of $\boldsymbol{\Phi}^{(B)}$ we need to define a B-sample bootstrap estimate of F under the null hypothesis ($\hat{F}_{H_0{}^{(b)}}$), i.e. such that the null hypothesis is true. Remembering that $\boldsymbol{X}$ is the partitioned matrix of the manifest variables, the sample estimates of F are defined on the basis of $p(\boldsymbol{x}'_n) = \frac{1}{N}$, where $n = 1, 2, \ldots, N$ and $p(\boldsymbol{x}'_n)$ is the probability to extract the n-th observation from the matrix $\boldsymbol{X}$.

Suppose we want to test the null hypothesis that no linear relationship exists between $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$. In other words, we want to test the null hypothesis that the coefficient β_{23} linking $\boldsymbol{\xi}_2$ to $\boldsymbol{\xi}_3$ is equal to 0:

$$\begin{aligned} \mathrm{H}_0 &: \beta_{23} = 0 \\ \mathrm{H}_1 &: \beta_{23} \neq 0 \end{aligned} \tag{2.28}$$

In order to reproduce the model under H_0 the matrix of the manifest variables associated to $\boldsymbol{\xi}_3$, i.e. $\boldsymbol{X}_3$, can be deflated by removing the linear effect of $\boldsymbol{X}_2$, where $\boldsymbol{X}_2$ is the block of manifest variables associated to $\boldsymbol{\xi}_2$. In particular, the deflated matrix $\boldsymbol{X}_{3(2)}$ is obtained as:

$$\boldsymbol{X}_{3(2)} = \boldsymbol{X}_3 - \boldsymbol{X}_2 \left(\boldsymbol{X}'_2 \boldsymbol{X}_2\right)^{-1} \boldsymbol{X}'_2 \boldsymbol{X}_3. \tag{2.29}$$

Thus, the estimate of F under the null hypothesis is $\hat{F}_{[\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_{3(2)}]}$.

Once the estimate of cdf of $\boldsymbol{X}$ under the null hypothesis is defined, the B-sample bootstrap approximation $\boldsymbol{\Phi}^{(B)}$ of $\boldsymbol{\Phi}$ is obtained by repeating B times the following procedure.

For each b: $b = 1, 2 \ldots, B$:

1. Draw a random sample from $\hat{F}_{[X_1, X_2, X_{3(2)}]}$.
2. Estimate the model under the null hypothesis for the sample obtained at the previous step.
3. Compute the GoF value, $GoF_{H_0}^{(b)}$.

The choice of B depends on several aspects such as: the sample size, the number of manifest variables and the complexity of the structural model. Usually, we prefer to choose $B \geq 1000$.

The decision on the null hypothesis is taken by referring to the inverse cdf of GoF_{H_0}. In particular, the test is performed at a nominal size α, by comparing the GoF value for the model defined in (2.26), computed on the original data, to the $(1-\alpha)^{th}$ percentile of $\boldsymbol{\Phi}^{(B)}$. If $GoF > \boldsymbol{\Phi}_{(1-\alpha)}^{(B)}$, then we reject the null hypothesis.

A schematic representation of the procedure to perform a non-parametric Bootstrap GoF-based test on a single path-coefficient is given in Algorithm 2.

Algorithm 2 : Non-parametric Bootstrap GoF-based test of a path-coefficient

Hypotheses on the coefficient β_{qj}:

$$\begin{aligned} \mathrm{H}_0 &: \beta_{qj} = 0 \\ \mathrm{H}_1 &: \beta_{qj} \neq 0 \end{aligned} \tag{2.30}$$

1: Estimate the specified structural model on the original dataset (bootstrap population) and compute the GoF index.
2: Deflate the endogenous block of manifest variable $\boldsymbol{X}_j$: $\boldsymbol{X}_{j(q)} = \boldsymbol{X}_j - \boldsymbol{X}_q \left(\boldsymbol{X}_q' \boldsymbol{X}_q\right)^{-1} \boldsymbol{X}_q' \boldsymbol{X}_j$.
3: Define B large enough.
4: **for all** $b = 1, \ldots, B$ **do**
5: Draw a sample from $\hat{\hat{F}}_{[X_1, X_2, X_{3(2)}]}$.
6: Estimate the model under the null hypothesis.
7: Compute the GoF value named $GoF_{H_0^b}$.
8: **end for**
9: By comparing the original GoF index to the inverse cdf of GoF_{H_0} accept or reject H_0.

2.3.2 Hypothesis Testing on the Whole Set of Path Coefficients

The procedure described in Sect. 2.3.1 can be easily generalized in order to test a sub-set of path coefficients or all of them at the same time. If the path coefficients are tested simultaneously, then this omnibus test can be used for an overall assessment of the model. This test is performed by comparing the default model specified by the user to the so-called baseline models, i.e the *saturated* model and the *independence*

or *null* model. The *saturated* model is the least restrictive model where all the structural relations are allowed (i.e. all path coefficients are free parameters). The *null* model is the most restrictive model with no relations among latent variables (i.e. all path coefficients are constrained to be 0). Following the structure of the model defined in figure 2.1, the null model is the model where : $\beta_{12} = \beta_{13} = \beta_{23} = 0$, while the saturated model coincides with the one in figure 2.1. More formally:

$$\begin{array}{l} \mathrm{H}_0 : \beta_{12} = \beta_{13} = \beta_{23} = 0 \\ \mathrm{H}_1 : \text{At least one } \beta_{qj} \neq 0 \end{array} \tag{2.31}$$

As for the simple case described in Sect. 2.3.1 we need to properly deflate $\boldsymbol{X}$ in order to estimate $\boldsymbol{\Phi}^{(B)}$. In particular, each endogenous block $\boldsymbol{X}_j$ has to be deflated according to the specified structural relations by means of orthogonal projection operators. In the model defined by (2.26), the block of manifest variables linked to $\boldsymbol{\xi}_2$ ($\boldsymbol{X}_2$) has to be deflated by removing the linear effect of $\boldsymbol{\xi}_1$ on $\boldsymbol{\xi}_2$, while the block of the manifest variables linked to $\boldsymbol{\xi}_3$ ($\boldsymbol{X}_3$) has to be deflated by removing the linear effect of both $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. However, since $\boldsymbol{\xi}_2$ is an endogenous latent variable, the deflated block $\boldsymbol{X}_{2(1)}$ has to be taken into account when deflating $\boldsymbol{X}_3$. In other words, the deflation of the block $\boldsymbol{X}_2$ is obtained as:

$$\boldsymbol{X}_{2(1)} = \boldsymbol{X}_2 - \boldsymbol{X}_1 \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1'\boldsymbol{X}_2$$

while, the deflation of the block $\boldsymbol{X}_3$ is obtained as:

$$\boldsymbol{X}_{3(1,2)} = \boldsymbol{X}_3 - \left[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}\right] \left(\left[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}\right]' \left[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}\right]\right)^{-1} \left[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}\right]' \boldsymbol{X}_3.$$

As we deal with a recursive model, it is always possible to build blocks that verify the null hypothesis by means of a proper sequence of deflations.

The algorithm described in Sect. 2.3.1 and in Algorithm 2 can be applied to $\hat{F}_{[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}, \boldsymbol{X}_{3(1,2)}]}$ in order to construct an inverse cdf of $\boldsymbol{\Phi}^{(B)}$ such that H_0 is true. The test is performed at a nominal confidence level α, by comparing the GoF value for the model defined in (2.26) to the $(1-\alpha)^{th}$ percentile of $\boldsymbol{\Phi}^{(B)}$ built upon $\hat{F}_{[\boldsymbol{X}_1, \boldsymbol{X}_{2(1)}, \boldsymbol{X}_{3(1,2)}]}$. If $GoF > \boldsymbol{\Phi}^{(B)}_{(1-\alpha)}$, then the null hypothesis is rejected. By comparing the GoF value obtained for the default model on the bootstrap population with the $GoF^{(b)}_{\mathrm{H}_0}$ obtained from bootstrap samples ($b = 1, 2, \ldots, B$), an empirical p-value can be computed as:

$$\text{p-value} = \frac{\sum_{b=1}^{B} I_b}{B} \tag{2.32}$$

where

$$I_b = \begin{cases} 1 \text{ if } GoF^{(b)}_{\mathrm{H}_0} \geq GoF \\ 0 \text{ otherwise} \end{cases} \tag{2.33}$$

and B is the number of Bootstrap re-samples.

As stated in (2.31), the above procedure tests the null hypothesis that all path coefficients are equal to zero against the alternative hypothesis that at least one of the coefficients is different from zero. By defining a proper deflation strategy, tests on any sub-set of path coefficients can be performed. Stepwise procedures can also be defined in order to identify a set of significant coefficients.

2.3.3 Application to Simulated Data

In this subsection we apply the procedures for testing path coefficients to simulated data.

Data have been generated according to the basic model defined in Fig. 2.2. This model is a simplified version of the one defined in Fig. 2.1.

According to Fig. 2.2, the structural model is specified by the equation:

$$\boldsymbol{\xi}_3 = \beta_{03} + \beta_{13}\boldsymbol{\xi}_1 + \beta_{23}\boldsymbol{\xi}_2 + \boldsymbol{\zeta}_3 \tag{2.34}$$

Three different tests have been performed on the simulated data-set. In particular, we perform a test:

1. On the whole model:
$$\begin{aligned} &\mathrm{H}_0 : \beta_{13} = \beta_{23} = 0 \\ &\mathrm{H}_1 : \text{At least one } \beta_{qj} \neq 0 \end{aligned} \tag{2.35}$$
2. On the coefficient β_{13}
$$\begin{aligned} &\mathrm{H}_0 : \beta_{13} = 0 \\ &\mathrm{H}_1 : \beta_{13} \neq 0 \end{aligned} \tag{2.36}$$
3. On the coefficient β_{23}
$$\begin{aligned} &\mathrm{H}_0 : \beta_{23} = 0 \\ &\mathrm{H}_1 : \beta_{23} \neq 0 \end{aligned} \tag{2.37}$$

2.3.3.1 Simulation Scheme

The following procedure has been used in order to simulate the manifest variables for the model in Fig. 2.2 with a sample size of 50 units:

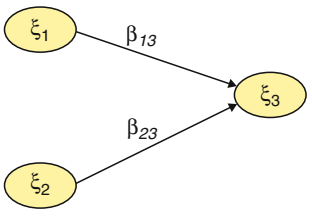

Fig. 2.2 Path diagram of the structural model specified by (2.34)

1. For each exogenous block, three manifest variables have been randomly generated according to a multivariate normal distribution. In particular, the manifest variables linked to the latent variable $\boldsymbol{\xi}_1$ come from a multivariate normal distribution with means equal to 2 and standard deviations equal to 1.5 for every manifest variable. The manifest variables of block 2 come from a multivariate normal distribution with means equal to 0 and standard deviations equal to 1 for every manifest variable.
2. The exogenous latent variables $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ have been computed as a standardized aggregate of the manifest variables obtained in the first step. An error term (from a normal distribution with zero mean and standard deviation equal to $1/4$ of the manifest variables' standard deviation) has been added to both exogenous latent variables.
3. The manifest variables corresponding to the endogenous latent variable $\boldsymbol{\xi}_3$ have been generated as a standardized aggregate of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ plus an error term (from a normal distribution with zero mean and standard deviation equal to 0.25).

2.3.3.2 Results

Table 2.1 reports the path coefficients and the GoF values obtained by running the PLS-PM algorithm on the simulated dataset.

According to the procedure described in Sect. 2.3.2 we need to deflate the data in different ways in order to perform the three different types of tests. Namely, in order to perform the first test ($H_0 : \beta_{13} = \beta_{23} = 0$) we need to deflate the block $\boldsymbol{X}_3$ with regards to $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ (Test 1), while the second test ($H_0 : \beta_{13} = 0$) is performed by deflating the block $\boldsymbol{X}_3$ only with regards to $\boldsymbol{X}_1$ (Test 2) and the last test ($H_0 : \beta_{23} = 0$) is performed by deflating the block $\boldsymbol{X}_3$ with regards to $\boldsymbol{X}_2$ (Test 3).

Under each null hypothesis, bootstrap resampling has been performed to obtain the bootstrap approximation $\boldsymbol{\Phi}^{(B)}$ of $\boldsymbol{\Phi}$. Bootstrap distributions have been approximated by 1,000 pseudo-random samples.

The histograms of the bootstrap approximations of the GoF distributions under the null hypotheses for Test 1, Test 2 and Test 3 are shown in Figs. 2.3–2.5, respectively. These histograms seem to reveal fairly normal distributions.

Table 2.2 reports the values of the critical thresholds computed for test sizes $\alpha = 0.10$ and $\alpha = 0.05$ on the bootstrap distribution for the three different tests. The $p - values$, computed according to the formula in (2.32), are also shown. On this basis, the null hypotheses for Test 1 and Test 2 have been correctly rejected by the proposed procedure. Nevertheless, the proposed test accepts the null hypothesis for Test 3 even if this hypothesis is false. This is due to the very weak value for the corresponding path coefficient, i.e. $\beta_{23} = 0.05$.

Table 2.1 Results from the simulated data-set

β_{13}	0.94
β_{23}	0.05
GoF (Absolute)	0.69

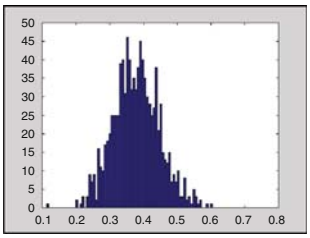

Fig. 2.3 Histogram of the bootstrap approximation of the GoF distribution under the null hypothesis in Test 1

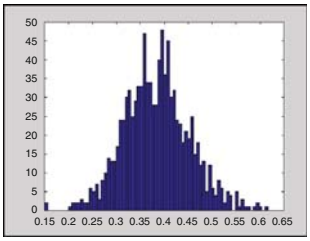

Fig. 2.4 Histogram of the bootstrap approximation of the GoF distribution under the null hypothesis in Test 2

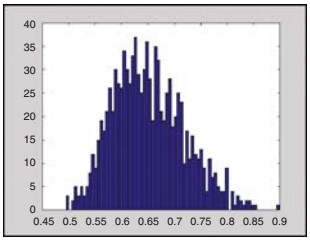

Fig. 2.5 Histogram of the bootstrap approximation of the GoF distribution under the null hypothesis in Test 3

Table 2.2 Thresholds and p-values from bootstrap distributions (1,000 re-samples)

	$\alpha = 0.10$	$\alpha = 0.05$	p-value
Test 1	0.46	0.49	0
Test 2	0.47	0.50	0
Test 3	0.74	0.77	0.27

Further researches are needed to investigate features of the GoF distribution as well as the statistical power of the proposed tests and their sensitivity with respect to the size of the coefficients, the sample size and the complexity of the structural model.

2.4 Heterogeneity in PLS Path Modeling

In this section we discuss how to improve the prediction performance and the interpretability of the model by allowing for unobserved heterogeneity.

Indeed, heterogeneity among units is an important issue in statistical analysis. Treating the sample as homogeneous, when it is not, may seriously affect the quality of the results and lead to biased interpretation. Since human behaviors are complex, looking at groups or classes of units having similar behaviors will be particularly

hard. Heterogeneity can hardly be detected using external information, i.e. using *a priori* clustering approach, especially in social, economic and marketing areas. Moreover, in several application fields (e.g. marketing) more attention is being given to clustering methods able to detect groups that are homogeneous in terms of their responses (Wedel and Kamakura 2000). Therefore, *response-based* clustering techniques are becoming more and more important in statistical literature.

Two types of heterogeneity could be affecting the data: observed and unobserved heterogeneity (Tenenhaus et al. 2010; Hensler and Fassott 2010; Chin and Dibbern 2007). In the first case the composition of classes is known *a priori*, while in the second case information on the number of classes or on their composition is not available.

So far in this paper we have assumed homogeneity over the observed set of units. In other words, all units are supposed to be well represented by a unique model estimated on the whole sample, i.e. *the global model*.

In a Structural Equation Model, the two cases of observed and unobserved heterogeneity match with the presence of a discrete moderating factor that, in the first case is manifest, i.e. an observed variable, while in the second case is latent, i.e. an unobserved variable (Chin and Dibbern 2007).

Usually heterogeneity in Structural Equation Models is handled by first forming classes on the basis of external variables or on the basis of standard clustering techniques applied to manifest and/or latent variables, and then by using the multi-group analysis introduced by Jöreskog (1971) and Sörbom (1974). However, heterogeneity in the models may not be necessarily captured by well-known observed variables playing the role of moderating variables (Hahn et al. 2002). Moreover, *post-hoc* clustering techniques on manifest variables, or on latent variable scores, do not take at all into account the model itself. Hence, while the local models obtained by cluster analysis on the latent variable scores will lead to differences in the group averages of the latent variables but not necessarily to different models, the same method performed on the manifest variables is unlikely to lead to different and well-separated models. This is true for both the model parameters and the means of latent variable scores. In addition, *a priori* unit clustering in Structural Equation Models is not conceptually acceptable since no structural relationship among the variables is postulated: when information concerning the relationships among variables is available (as it is in the theoretical causality network), classes should be looked for while taking into account this important piece of information. Finally, even in Structural Equation Models, the need is pre-eminent for a *response-based* clustering method, where the obtained classes are homogeneous with respect to the postulated model. Dealing with heterogeneity in PLS Path Models implies looking for *local models* characterized by class-specific model parameters.

Recently, several methods have been proposed to deal with unobserved heterogeneity in PLS-PM framework (Hahn et al. 2002; Ringle et al. 2005; Squillacciotti 2005; Trinchera and Esposito Vinzi 2006; Trinchera et al. 2006; Sanchez and Aluja 2006, 2007; Esposito Vinzi et al. 2008; Trinchera 2007). To our best knowledge, five approaches exist to handle heterogeneity in PLS Path Modeling: the Finite Mixture PLS, proposed by Hahn et al. (2002) and modified by Ringle et al.

(2010) (see Chap. 8 of this Handbook), the PLS Typological Path Model presented by Squillacciotti (2005) (see Chap. 10 of this Handbook) and modified by Trinchera and Esposito Vinzi (2006) and Trinchera et al. (2006), the PATHMOX by Sanchez and Aluja (2006), the PLS-PM based Clustering (PLS-PMC) by Ringle and Schlittgen (2007) and the Response Based Unit Segmentation in PLS Path Modeling (REBUS-PLS) proposed by Trinchera (2007) and Esposito Vinzi et al. (2008).

In the following we will discuss the REBUS-PLS approach in detail.

2.4.1 The REBUS-PLS Algorithm

A new method for unobserved heterogeneity detection in PLS-PM framework was recently presented by Trinchera (2007) and Esposito Vinzi et al. (2008). REBUS-PLS is an iterative algorithm that permits to estimate at the same time both the unit membership to latent classes and the class specific parameters of the local models. The core of the algorithm is a so-called *closeness measure* (CM) between units and models based on residuals (2.38). The idea behind the definition of this new measure is that if latent classes exist, units belonging to the same latent class will have similar local models. Moreover, if a unit is assigned to the correct latent class, its performance in the local model computed for that specific class will be better than the performance of the same unit considered as supplementary in the other local models.

The CM used in the REBUS-PLS algorithm represents an extension of the distance used in PLS-TPM by Trinchera et al. (2006), aiming at taking into account both the measurement and the structural models in the clustering procedure. In order to obtain local models that fit better than the global model, the chosen *closeness measure* is defined according to the structure of the Goodness of Fit (GoF) index, the only available measure of global fit for a PLS Path Model. According to the $DmodY$ distance used in PLS Regression (Tenenhaus 1998) and the distance used by Esposito Vinzi and Lauro (2003) in PLS Typological Regression all the computed residuals are weighted by quality indexes: the importance of residuals increases while the quality index decreases. That is why the communality index and the R^2 values are included in the CM computation.

In a more formal terms, the *closeness measure* (CM) of the n-th unit to the k-th local model, i.e. to the latent model corresponding to the k-th latent class, is defined as:

$$CM_{nk} = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[\frac{e_{npqk}^2}{Com\left(\hat{\xi}_{qk},x_{pq}\right)}\right]}{\frac{\sum_{n}^{N}\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[\frac{e_{npqk}^2}{Com\left(\hat{\xi}_{qk},x_{pq}\right)}\right]}{(N-t_k-1)}} \times \frac{\sum_{j=1}^{J}\left[\frac{f_{njk}^2}{R^2\left(\hat{\xi}_j,\hat{\xi}_{q:\xi_q\rightarrow\xi_j}\right)}\right]}{\frac{\sum_{n}^{N}\sum_{j=1}^{J}\left[\frac{f_{njk}^2}{R^2\left(\hat{\xi}_j,\hat{\xi}_{q:\xi_q\rightarrow\xi_j}\right)}\right]}{(N-t_k-1)}}} \tag{2.38}$$

where:
$Com\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_{qk}\right)$ is the communality index for the p-th manifest variable of the q-th block in the k-th latent class;
e_{npqk} is the measurement model residual for the n-th unit in the k-th latent class, corresponding to the p-th manifest variable in the q-th block, i.e. the communality residuals;
f_{njk} is the structural model residual for the n-th unit in the k-th latent class, corresponding to the j-th endogenous block;
N is the total number of units;
t_k is the number of extracted components. Since all blocks are supposed to be reflective, the value of t_k will always be equal to 1.

As for the GoF index, the left-side term of the product in (2.38) refers to the measurement models for all the Q blocks in the model, while the right-side term refers to the structural model. It is important to notice that both the measurement and the structural residuals are computed for each unit with respect to each local model regardless of the membership of the units to the specific latent class. In computing the residual from the k-th latent model, we expect that units belonging to the k-th latent class show smaller residuals than units belonging to the other $(K-1)$ latent classes.

As already said, two kinds of residuals are used to evaluate the closeness between a unit and a model: the measurement or communality residuals and the structural residuals. For a thorough description of the REBUS-PLS algorithm and the computation of the communality and the structural residuals, refer to the original REBUS-PLS papers (Trinchera 2007; Esposito Vinzi et al. 2008).

The choice of the *closeness measure* in (2.38) as a criterion for assigning units to classes has two major advantages. First, unobserved heterogeneity can now be detected in both the measurement and the structural models. If two models show identical structural coefficients, but differ with respect to one or more outer weights in the exogenous blocks, REBUS-PLS is able to identify this source of heterogeneity, which might be of major importance in practical applications. Moreover, since the *closeness measure* is defined according to the structure of the Goodness of Fit (GoF) index, the identified local models will show a better prediction performance.

The CM expressed by (2.38) is only the core of an iterative algorithm allowing us to obtain a *response-based* clustering of the units.

As a matter of fact, REBUS-PLS is an iterative algorithm (see Fig. 2.6). The first step of the REBUS-PLS algorithm involves estimating the global model on all the observed units, by performing a simple PLS Path Modeling analysis. In the second step, the communality and the structural residuals of each unit from the global model are obtained. The number of classes (K) to be taken into account during the successive iterations and the initial composition of the classes are obtained by performing a hierarchical cluster analysis on the computed residuals (both from the measurement and the structural models). Once the number of classes and their initial composition are obtained, a PLS Path Modeling analysis is performed on each class and K provisional local models are estimated. The group-specific parameters computed at the previous step are used to compute the communality and the structural

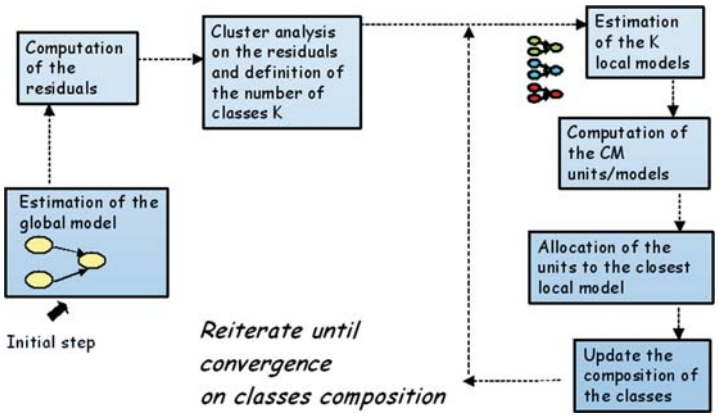

Fig. 2.6 A schematic representation of the REBUS-PLS algorithm

residuals of each unit from each local model. Then the CM of each unit from each local model is obtained according to (2.38). Each unit is, therefore, assigned to the closest local model, i.e. to the model from which it shows the smallest CM value. Once the composition of the classes is updated, K new local models are estimated. The algorithm goes on until the threshold of a stopping rule is achieved.

Stability on class composition from one iteration to the other is considered as a stopping rule. The authors suggest using the threshold of less than 5% of units changing class from one iteration to the other as a stopping rule. Indeed, REBUS-PLS usually assures convergence in a small number of iterations (i.e. less than 15). It is also possible not to define a threshold as a stopping rule and run the algorithm until the same groups are formed in successive iterations. In fact, if no stopping rule is imposed once the "best" model is obtained in the REBUS-PLS viewpoint, i.e. once each unit is correctly assigned to the closest local model, the algorithm provides the same partition of the units at successive iterations.

If the sample size is large, it is possible to have such boundary units that change classes time after time at successive iterations. This leads to obtaining a series of partitions (i.e. of local model estimates) that repeat themselves in successive iterations. In order to avoid the "boundary" unit problem the authors suggest always defining a stopping rule.

Once the stability on class composition is reached, the final local models are estimated. The class-specific coefficients and indexes are then compared in order to explain differences between detected latent classes. Moreover the quality of the obtained partition can be evaluated through a new index (i.e. the *Group Quality Index* - GQI) developed by Trinchera (2007). This index is a reformulation of the *Goodness of Fit* index in a multi-group perspective, and it is also based on residuals. A detailed presentation of the GQI, as well as a simulation study aiming at assessing GQI properties, can be found in Trinchera (2007). The GQI index is equal to the GoF in the case of a unique class, i.e. when $K = 1$ and $n_1 = N$. In other words, the *Group Quality Index* computed for the whole sample as a unique class is equal to

the GoF index computed for the global model. Instead, if local models performing better than the global one are detected, the GQI index will be higher than the GoF value computed for the global model.

Trinchera (2007) performed a simulation study to assess GQI features. In particular, it is suggested that a relative improvement of the GQI index from the global model to the detected local models higher than 25% can be considered as a satisfactory threshold to prefer the detected unit partition to the aggregate data solution. Finally, the quality of the detected partition can be assessed by a permutation test (Edgington 1987) involving T random replications of the unit partition (keeping constant the group proportions as detected by REBUS-PLS) so as to yield an empirical distribution of the GQI index.

The GQI obtained for the REBUS-PLS partition is compared to the percentiles of the empirical distribution to decide whether local models are performing significantly better than the global one. Trinchera (2007) has shown that, in case of unobserved heterogeneity and apart from the outlier solutions, the GQI index computed for the aggregate level is the minimum value obtained for the empirical distribution of the GQI.

If external concomitant variables are available, an *ex-post* analysis on the detected classes can be performed so as to characterize the detected latent classes and improve interpretability of their composition.

So far, REBUS-PLS is limited to reflective measurement models because the measurement residuals come from the simple regressions between each manifest variable in a block and the corresponding latent variable. Developments of the REBUS-PLS algorithm to the formative measurement models are on going.

2.4.2 Application to Real Data

Here, we present a simple and clear example to show the REBUS-PLS ability to capture unobserved heterogeneity on empirical data. We use the same data as in Ringle et al. (2010). This dataset comes from the Gruner&Jahr's Brigitte Communication Analysis performed in 2002 that specifically concerns the Benetton fashion brand. REBUS-PLS has been performed using a SAS-IML macro developed by Trinchera (2007).

The Benetton dataset is composed of ten manifest variables observed on 444 German women. Each manifest variable is a question in the Gruner&Jahr's Brigitte Communication Analysis of 2002. The women had to answer each question using a four-point scale from "low" to "high".

The structural model for Benetton's brand preference, as used by Ringle et al. (2010), consists of one latent endogenous *Brand Preference* variable, and two latent exogenous variables, *Image* and *Character*. All manifest variables are linked to the corresponding latent variable via a reflective measurement model. Figure 2.7 illustrates the path diagram with the latent variables and the employed manifest variables. A list of the used manifest variables with the corresponding meanings is shown in Table 2.3.

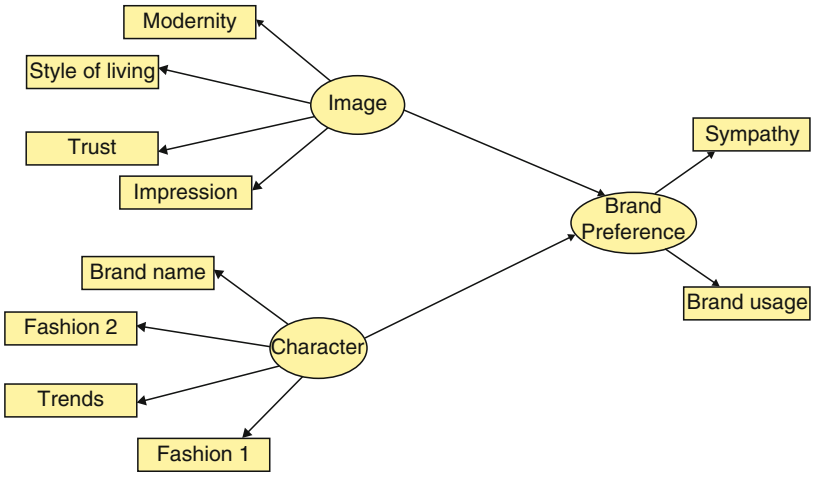

Fig. 2.7 Path diagram for Benetton data

Table 2.3 Manifest (MV) and latent variables (LV) definition for Benetton data

LV Name	MV Name	Concepts
Image	Modernity	It is modern and up to date
	Style of living	Represents a great style of life
	Trust	This brand can be trusted
	Impression	I have a clear impression of this brand
Character	Brand name	A brand name is very important to me
	Fashion 2	I often talk about fashion
	Trends	I am interested in the latest trends
	Fashion 1	Fashion is a way to express who I am
Brand	Sympathy	Sympathy
Preference	Brand usage	Brand usage

A PLS Path Modeling analysis on the whole sample has been performed with standardized manifest variables. As it is obvious, the global model estimates are consistent with the ones obtained by Ringle *et al.* in their study (see Chap. 8). Since all the blocks in the model are supposed to be reflective, then they should be homogeneous and *unidimensional*. Hence, first of all we have to check for block homogeneity and unidimensionality. Table 2.4 shows values of the tools presented in Sect. 2.2.1 for checking the block homogeneity and unidimensionality. According to Chin (1998), all the blocks are considered homogenous, i.e. the *Dillon-Goldstein's rho* is always larger than 0.7. Moreover, the three blocks are unidimensional as only the first eigenvalues for each block are greater than one. Therefore, the reflective model is appropriate.

A simple overview of the global model results is proposed in Fig. 2.8. According to the global model results *Image* seems to be the most important driver for *Brand Preference*, with a path coefficient equal to 0.423. The influence of the

Table 2.4 Homogeneity and unidimensionality of MVs blocks

LV Name	# of MVs	Cronbach's α	D.G.'s ρ	PCA eigenvalues
Image	4	0.869	0.911	2.873
				0.509
				0.349
				0.269
Character	4	0.874	0.914	2.906
				0.479
				0.372
				0.243
Brand preference	2	0.865	0.937	1.763
				0.237

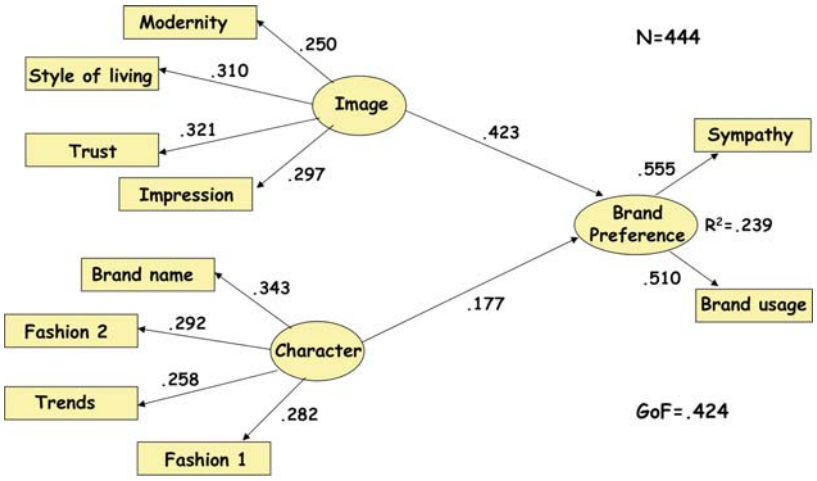

Fig. 2.8 Global model results from Benetton data obtained by using a SAS-IML macro

exogenous latent variable *Character* is considerably weaker (path coefficient of 0.177). Nevertheless, the R^2 value associated with the endogenous latent variable *Brand Preference* is quite low, being equal to 0.239. Ringle et al. (2010) consider this value as a moderate level for a PLS Path Model. In our opinion, an R^2 value of 0.239 has to be considered as unsatisfactory, and could be used as a first sign of possible unobserved heterogeneity in the data. Looking at the measurement models, all the relationships in the reflective measurement models have high factor loadings (the smallest loading has a value of 0.795, see Table 2.5). In Fig. 2.8 the outer weights used for yielding standardized latent variable scores are shown. In the *Brand Preference* block, *Sympathy* and *Brand Usage* have similar weights. Instead, differences arise in both exogenous blocks. Finally, the global model on Benetton data shows a value for the absolute GoF equal to 0.424 (see Table 2.6). The quite low value of the GoF index might also suggest that we have to look for more homogeneous segments among the units.

Table 2.5 Measurement model results for the global and the local models obtained by REBUS-PLS

		Global	Class 1	Class 2	Class 3
Number of units		444	105	141	198
Outer weights	Modernity	0.250	0.328	0.278	0.291
Image	Style of living	0.310	0.264	0.314	0.270
	Trust	0.321	0.284	0.315	0.375
	Impression	0.297	0.292	0.267	0.273
Outer weights	Brand name	0.343	0.342	0.262	0.298
Character	Fashion2	0.292	0.276	0.345	0.314
	Trends	0.258	0.266	0.323	0.335
	Fashion1	0.282	0.314	0.213	0.231
Outer weights	Sympathy	0.555	0.549	0.852	0.682
Brand preference	Brand Usage	0.510	0.637	0.575	0.547
Standardized loadings	Modernity	0.795	0.827	0.810	0.818
Image	Style of living	0.832	0.834	0.860	0.735
	Trust	0.899	0.898	0.890	0.895
	Impression	0.860	0.865	0.840	0.834
Standardized loadings	Brand name	0.850	0.832	0.842	0.822
Character	Fashion2	0.894	0.846	0.929	0.908
	Trends	0.859	0.850	0.902	0.878
	Fashion1	0.801	0.819	0.788	0.762
Standardized loadings	Sympathy	0.944	0.816	0.819	0.855
Brand preference	Brand Usage	0.933	0.867	0.526	0.762
Communality	Modernity	0.632	0.685	0.657	0.668
Image	Style of living	0.693	0.695	0.740	0.541
	Trust	0.808	0.806	0.792	0.801
	Impression	0.739	0.748	0.706	0.696
Communality	Brand name	0.722	0.692	0.709	0.676
Character	Fashion2	0.799	0.715	0.864	0.825
	Trends	0.738	0.722	0.814	0.770
	Fashion1	0.642	0.670	0.620	0.581
Communality	Sympathy	0.891	0.666	0.671	0.730
Brand preference	Brand Usage	0.871	0.752	0.277	0.581

A more complete outline of the global model results is provided in Table 2.5 for the outer model and in Table 2.6 for the inner model. These tables contain also the class-specific results in order to make it easier to compare the segments.

Performing REBUS-PLS on Benetton data leads to detecting three different classes of units showing homogeneous behaviors. As a matter of fact, the cluster analysis on the residuals from the global model (see Fig. 2.9) suggests that we should look for two or three latent classes. Both partitions have been investigated. The three classes partition is preferred as it shows a higher *Group Quality Index*. Moreover, the GQI index computed for the two classes solution ($GQI = 0.454$) is close to the GoF value computed for the global model (i.e. the GQI index in the case of only one global class, $GoF = 0.424$). Therefore, the 25% improvement

Table 2.6 Structural model results for the global model and the local models obtained by REBUS-PLS

		Global	Class 1	Class 2	Class 3
Number of units		444	105	141	198
Path	Image	0.423	0.420	0.703	0.488
Coefficients		[0.331; 0.523]	[0.225; 0.565]	[0.611; 0.769]	[0.314; 0.606]
on brand	Character	0.177	0.274	0.319	0.138
preference		[0.100; 0.257]	[0.078; 0.411]	[0.201; 0.408]	[0.003; 0.311]
Redundancy	Brand preference	0.210	0.207	0.322	0.180
R^2		0.239	0.292	0.680	0.275
Brand preference		[0.166; 0.343]	[0.162; 0.490]	[0.588; 0.775]	[0.195; 0.457]
R^2	Image	0.81	0.67	0.79	0.90
contributions	Character	0.19	0.33	0.21	0.10
GoF value		0.424	0.457	0.682	0.435
		[0.354; 0.508]	[0.325; 0.596]	[0.618; 0.745]	[0.366; 0.577]

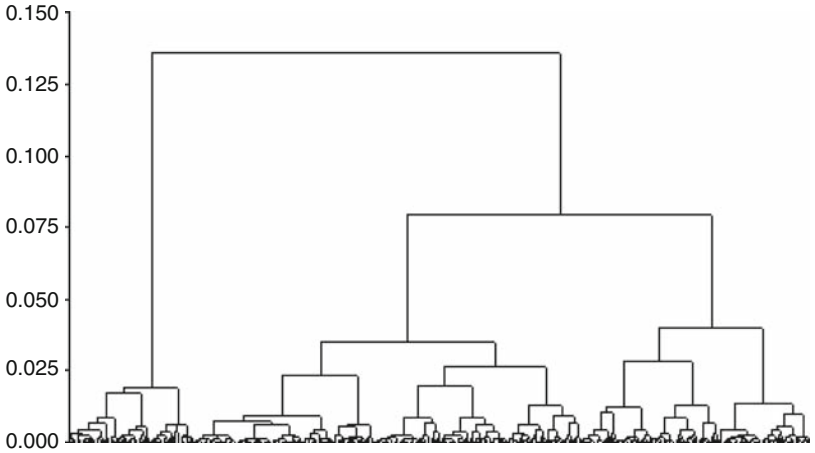

Fig. 2.9 Dendrogramme obtained by a cluster analysis on the residuals from the global model (Step 3 of the REBUS-PLS algorithm)

foreseen for preferring the partition in two classes is not achieved. Here, only the results for the three classes partition are presented.

The first class is composed of 105 units, i.e around 24% of the whole sample. This class is characterized by a path coefficient linking the latent variable *Character* to the endogenous latent variable *Brand Preference* higher than the one obtained for the global model. Moreover, differences in unit behaviors arise also with respect to the outer weights in the *Brand Preference* block, i.e. *Brand Usage* shows a higher weight than *Sympathy*. The GoF value for this class (0.457) is similar to the one for the global model (0.424). Figure 2.10 shows the estimates obtained for this class.

The second class, instead, shows a definitely higher GoF value of 0.682 (see Table 2.6). This class is composed of around 32% of the whole sample, and

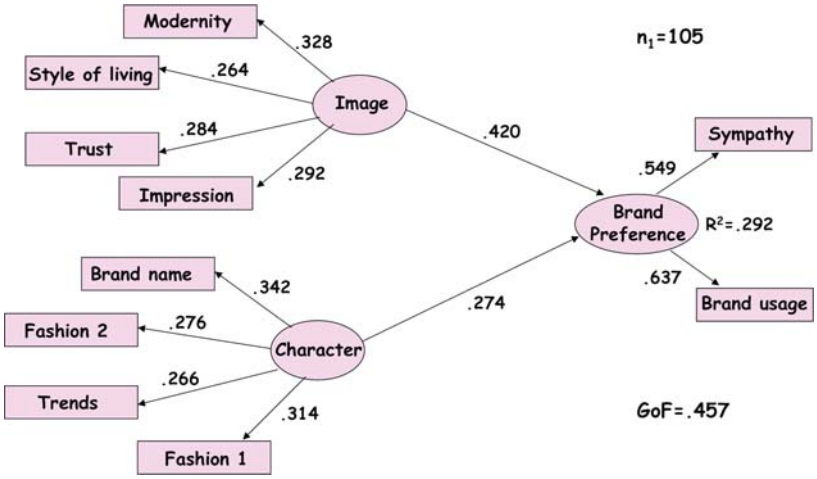

Fig. 2.10 Local model results for the first class detected by the REBUS-PLS algorithm on Benetton data

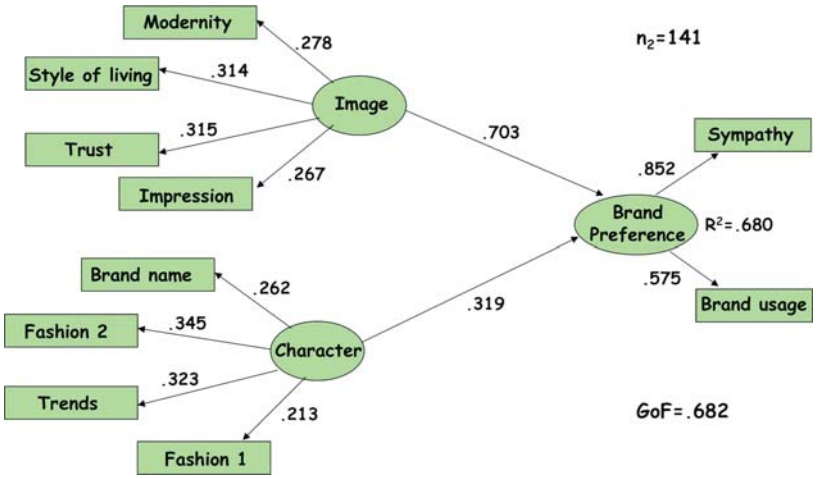

Fig. 2.11 Local model results for the second class detected by the REBUS-PLS algorithm on Benetton data

is characterized by a much higher path coefficient associated to the relationship between the *Image* and the *Brand Preference*. Looking at the measurement model (see Table 2.5), differences arise in the *Brand Preference* block and in the *Character* block. As a matter of fact, the communality index (i.e. the square of the correlation) between the manifest variable *Brand Usage* and the corresponding latent variable *Brand Preference* is really lower than the one obtained for the global model as well as for the first local model described above. Other differences for this second class may be detected by looking at the results provided in Fig. 2.11.

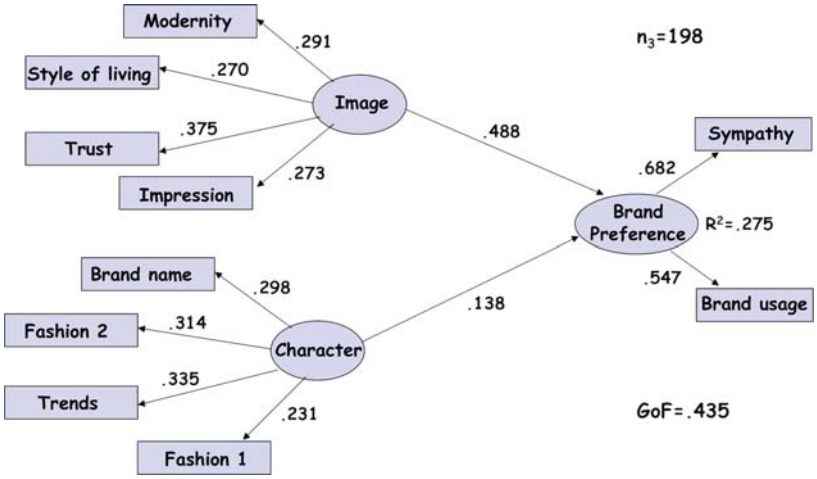

Fig. 2.12 Local model results for the third class by the REBUS-PLS algorithm on Benetton data

Finally, the results for the third class are presented in Fig. 2.12. This class is composed of 198 units., i.e. more than 44% of the whole sample. It is characterized by a very weak relationship between the latent variable *Character* and the endogenous latent variable *Brand Preference*. Moreover, the 95% bootstrap confidence interval shows that this link is close to be non significant as the lower bound is very close to 0 (see Table 2.6). Differences arise also with respect to the measurement model, notably in the *Image* block. As a matter of fact, in this class the manifest variable *Style of living* shows a very low correlation compared with the other models (both local and global).
Nonetheless, the quality index values computed for this third local model are only slightly different from the ones in the global model ($R^2 = 0.275$ and $GoF = 0.435$).

The three classes solution shows a *Group Quality Index* equal to 0.531. In order to validate the REBUS-PLS based partition, an empirical distribution of the GQI values is yielded by means of permutations. The whole sample has been randomly divided 300 times into three classes of the same size as the ones detected by REBUS-PLS. The GQI has been computed for each of the random partitions of the units. The empirical distribution of the GQI values for a three classes partition is then obtained (see Fig. 2.13). As expected, the GQI value from the REBUS-PLS partition is definitely an extremely high value of the distribution thus showing that the REBUS-PLS based partition is better than a random assignment of the units into three classes.

Moreover, in Fig. 2.14, it is possible to notice that the GQI computed for the global model (i.e. the GoF value) is a very small value in the GQI distribution. Therefore, the global model has to be definitely considered as being affected by heterogeneity.

Ringle et al. (2010) apply FIMIX-PLS to Benetton data (see Chap. 8) and identify only two classes. The first one (80.9% of the whole sample) is very similar to

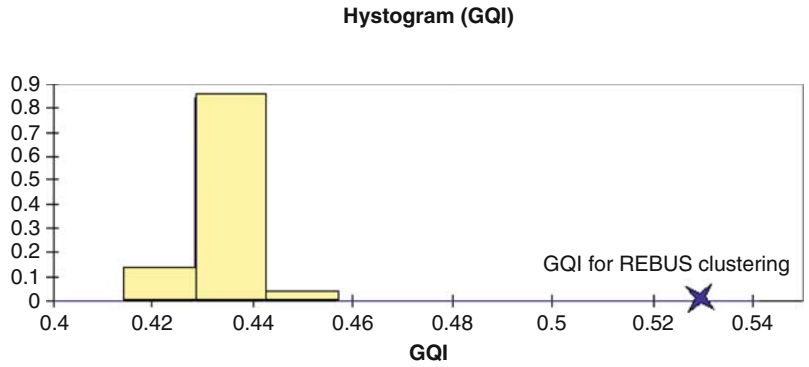

Fig. 2.13 Empirical distribution of the GQI computed on 300 random partitions of the original sample in three classes

Simple Statistics	GQI
No. of observations	301
Minimum	0.421
Maximum	0.531
1° Quartile	0.430
Median	0.433
3° Quartile	0.436
Mean	0.434
Variance (n-1)	0.000
Standard Deviation (n-1)	0.008
Lower bound on mean (95%)	0.424
Upper bound on mean (95%)	0.441

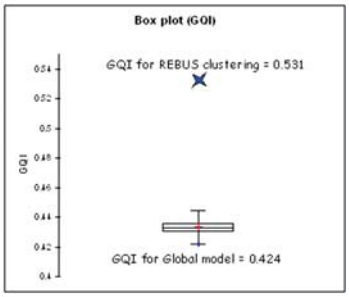

Fig. 2.14 Descriptive statistics for the GQI empirical distribution

the global model results in terms of path coefficients. Nevertheless, the R^2 value associated to the endogenous latent variable *Brand Preference* is equal to 0.108. This value is even smaller than for the global model ($R^2 = 0.239$). The second detected class, instead, is similar to the second class obtained by REBUS-PLS. As a matter of fact, also in this case the exogenous latent variable *Image* seems be the most important driver for *Brand Preference*, showing an R^2 close to 1.

In order to obtain local models that are different also for the measurement model, Ringle et al. (2010) apply a two-step strategy. In the first step they simply apply FIMIX-PLS. Successively they use external/concomitant variables to look for groups overlapping the FIMIX-based ones. Nevertheless, also in this two-step procedure the obtained results are not better than the ones provided by the REBUS-PLS-based partition. As a matter of fact, the R^2 value and the GoF value for the first local model are smaller than for the global model. The local model for the largest class (80% of the whole sample) performs worse than the global model, and worse than all the REBUS-PLS based local models.

The REBUS-PLS algorithm turned out to be a powerful tool to detect unobserved heterogeneity in both experimental and empirical data.

2.5 Conclusion and Perspectives

In the previous sections, where needed, we have already enhanced some of the on going research related to the topics of interest for this chapter. Namely, the development of new estimation modes and schemes for multidimensional (formative) constructs, a path analysis on latent variable scores to estimate path coefficients, the use of GoF-based non parametric tests for the overall model assessment, a sensitivity analysis for these tests, the generalization of REBUS-PLS to capturing heterogeneity in formative models.

We like to conclude this chapter by proposing a short list of further open issues that, in our opinion, currently represent the most important and promising research challenges in PLS Path Modeling:

- Definition of optimizing criteria and unifying functions related to classical or modified versions of the PLS-PM algorithm both for the predictive path model between latent variables and for the analysis of multiple tables.
- Possibility of imposing constraints on the model coefficients (outer weights, loadings, path coefficients) so as to include any information available a priori as well as any hypothesis (e.g. equality of coefficients across different groups, conjectures on model parameters) in the model estimation phase.
- Specific treatment of categorical (nominal and ordinal) manifest variables.
- Specific treatment of non-linearity both in the measurement and the structural model.
- Outliers identification, i.e. assessment of the influence of each statistical unit on the estimates of the outer weights for each block of manifest variables.
- Development of robust alternatives to the current OLS-based PLS Path Modeling algorithm.
- Development of a model estimation procedure based on optimizing the GoF index, i.e. on minimizing a well defined fit function.
- Possibility of specifying feedback relationships between latent variables so as to investigate mutual causality.

The above mentioned issues represent fascinating topics for researchers from both Statistics and applied disciplines.

There is nothing vague or fuzzy about soft modeling;
the technical argument is entirely rigorous
Herman Wold

Acknowledgements The participation of V. Esposito Vinzi to this research was supported by the Research Center of the ESSEC Business School of Paris. The participation of L. Trinchera to this research was supported by the MIUR (Italian Ministry of Education, University and Research) grant "Multivariate statistical models for the ex-ante and the ex-post analysis of regulatory impact", coordinated by C. Lauro (2006).

References

Addinsoft (2009). *XLSTAT 2009*. France: Addinsoft. http://www.xlstat.com/en/products/xlstat-plspm/

Alwin, D. F., and Hauser, R. M. (1975). The decomposition of effects in path. *American Sociological Review, 40*, 36–47.

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

Cassel, C., Hackl, P., and Westlund, A. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics, 26*, 435–446.

Cassel, C., Hackl, P., and Westlund, A. (2000). On measurement of intangible assets: a study of robustness of partial least squares. *Total Quality Management, 11*, 897–907.

Chin, W. W. (1998). The partial least squares approach for structural equation modeling. *in* G. A. Marcoulides (Ed.), Modern methods for business research (pp. 295–236). London: Lawrence Erlbaum Associates.

Chin, W. W., and Dibbern, J. (2007). A permutation based procedure for multigroup PLS analysis: results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between germany and the USA. In: V. Esposito Vinzi, W. Chin, J. Hensler, and H. Wold (Eds.,) *Handbook PLS and Marketing.* Berlin, Heidelberg, New York: Springer.

Edgington, E. (1987). *Randomization test*. New York: Marcel Dekker.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman&Hall.

Esposito Vinzi, V. (2008). The contribution of PLS regression to PLS path modelling: formative measurement model and causality network in the structural model. In: *Joint Statistical Meetings (JSM) 2008*, American Statistical Association, Denver, Colorado, United States of America, August 7th 2008.

Esposito Vinzi, V. (2009). PLS path modeling and PLS regression: a joint partial least squares component-based approach to structural equation modeling. In: *IFCS@GFKL – Classification as a Tool for Research (IFCS 2009)* , University of Technology, Dresden, Dresden, Germany, March 14th 2009 (Plenary Invited Speaker).

Esposito Vinzi, V., and Lauro, C. (2003). PLS regression and classification. In *Proceedings of the PLS'03 International Symposium*, DECISIA, France, pp. 45–56.

Esposito Vinzi, V., and Russolillo, G. (2010). Partial least squares path modeling and regression. In E. Wegman, Y. Said, and D. Scott (Eds.), *Wiley interdisciplinary reviews: computational statistics*. New York: Wiley.

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 (ASMBI), 24*, 439–458.

Fornell, C. (1992). A national customer satisfaction barometer: the swedish experience. *Journal of Marketing, 56*, 6–21.

Fornell, C., and Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS appliead to consumer exit-voice theory. *Journal of Marketing Research, 19*, 440–452.

Hahn, C., Johnson, M., Herrmann, A., and Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review, 54*, 243–269.

Hanafi, M. (2007). PLS path modeling: computation of latent variables with the estimation mode B. *Computational Statistics, 22*, 275–292.

Hensler, J., and Fassott, G. (2010). Testing moderating effects in PLS path models: An illustration of available procedures. In V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang (Eds.), *Handbook Partial Least Squares*. Heidelberg: Springer.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57*, 409–426.

Kaplan, D. (2000). *Structural equation modeling: foundations and extensions*. Thousands Oaks, California: Sage.

Lohmöller, J. (1987). LVPLS program manual, version 1.8, Technical report. Zentralarchiv für Empirische Sozialforschung, Universität Zu Köln, Köln.

Lohmöller, J. (1989). *Latent variable path modeling with partial least squares*. Heildelberg: Physica-Verlag.

Lyttkens, E., Areskoug, B., and Wold, H. (1975). The convergence of NIPALS estimation procedures for six path models with one or two latent variables. Technical report, University of Göteborg.

Ringle, C., and Schlittgen, R. (2007). A genetic algorithm segmentation approach for uncovering and separating groups of data in PLS path modeling. In *'PLS'07: 5th International Symposium on PLS and Related Methods*, Oslo, Norway, pp. 75–78.

Ringle, C., Wende, S., and Will, A. (2005). Customer segmentation with FIMIX-PLS. In T. Aluja, J. Casanovas, V. Esposito Vinzi, A. Morineau, and M. Tenenhaus Eds., *Proceedings of PLS-05 International Symposium*, SPAD Test&go, Paris, pp. 507–514.

Ringle, C., Wende, S., and Will, A. (2010). Finite mixture partial least squares analysis: Methodology and numerical examples. In V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang (Eds.), *Handbook Partial Least Squares*. Heidelberg: Springer.

Sanchez, G., and Aluja, T. (2006). Pathmox: a PLS-PM segmentation algorithm, *in* V. Esposito Vinzi, C. Lauro, A. Braverma, H. Kiers, and M. G.Schmiek (Eds.), *Proceedings of KNEMO 2006*, number ISBN 88-89744-00-6, Tilapia, Anacapri, p. 69.

Sanchez, G., and Aluja, T. (2007). A simulation study of PATHMOX (PLS path modeling segmentation tree) sensitivity. In *5th International Symposium - Causality explored by indirect observation*, Oslo, Norway.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Squillacciotti, S. (2005). Prediction oriented classification in PLS path modelling. In T. Aluja, J. Casanovas, V. Esposito Vinzi, A. Morineau, and M. Tenenhaus, (Eds.), *Proceedings of PLS- 05 International Symposium*, SPAD Test&go, Paris, pp. 499–506.

Tenenhaus, M. (1998). *La Régression PLS: théorie et pratique*. Paris: Technip.

Tenenhaus, M. (2008a). Component-based structural equation modelling. *Total Quality Management & Business Excellence, 19*, 871–886.

Tenenhaus, M. (2008b). 'Structural equation modelling for small samples. *HEC Paris: Jouy-en-Josas,Working paper, no. 885*.

Tenenhaus, M., and Esposito Vinzi, V. (2005). PLS regression, PLS path modeling and generalized procrustean analysis: a combined approach for PLS regression, PLS path modeling and generalized multiblock analysis. *Journal of Chemometrics, 19*, 145–153.

Tenenhaus, M., and Tenenhaus, A. (2009). A criterion-based PLS approach to SEM. In *3rd Workshop on PLS Developments*, ESSEC Business School of Paris, France, May 14th 2009.

Tenenhaus, M., Amato, S., and Esposito Vinzi, V. (2004). A global goodness-of-fit index for PLS structural equation modelling. *Proceedings of the XLII SIS Scientific Meeting*, Vol. Contributed Papers, CLEUP, Padova, pp. 739–742.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., and Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis, 48*, 159–205.

Tenenhaus, M., Mauger, E., and Guinot, C. (2010). Use of ULS-SEM and PLS-SEM to measure a group effect in a regression model relating two blocks of binary variables. In V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang (Eds.), *Handbook Partial Least Squares*. Heidelberg: Springer.

Thurstone, L. L. (1931). *The theory of multiple factors*. Ann Arbor, MI: Edwards Brothers.

Trinchera, L. (2007). Unobserved Heterogeneity in structural equation models: a new approach in latent class detection in PLS path modeling. PhD thesis, DMS, University of Naples.

Trinchera, L., and Esposito Vinzi, V. (2006). Capturing unobserved heterogeneity in PLS path modeling. In *Proceedings of IFCS 2006 Conference*, Ljubljana, Sloveny.

Trinchera, L., Squillacciotti, S., and Esposito Vinzi, V. (2006). PLS typological path modeling : a model-based approach to classification. In V. Esposito Vinzi, C. Lauro, A. Braverma, H. Kiers,

and M. G.Schmiek (Eds.), *Proceedings of KNEMO 2006*, ISBN 88-89744-00-6, Tilapia, Anacapri, p. 87.

Tukey, J. W. (1964). Causation, regression and path analysis. In *Statistics and mathematics in biology*. New York: Hafner.

Wedel, M., and Kamakura, W. A. (2000). , *Market segmentation – conceptual and methodological foundations*, 2 edn. Boston: Kluwer.

Wertz, C., Linn, R., and Jöreskog, K. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34*(1), 25–33.

Wold, H. (1966). Estimation of principal component and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis*, (pp. 391–420). New York: Academic Press.

Wold, H. (1975a). PLS path models with latent variables: the nipals approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Cappecchi (Eds.), *Quantitative sociology: international perspectives on mathematical and statistical modeling*. New York: Academic Press.

Wold, H. (1975b). Modelling in complex situations with soft infromation. *Third World Congress of Econometric Society*, Toronto, Canada.

Wold, H. (1975c). Soft modeling by latent variables: the nonlinear iterative partial least squares approach. In J. Gani (Ed.), *Perspectives in probability and statistics, papers in honor of M. S. Bartlett* (pp. 117–142). London: Academic Press.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In J. Kmenta, & J. B. Ramsey (Eds.), *Evaluation of econometric models*, pp. 47–74.

Wold, H. (1982). Soft modeling: the basic design and some extensions. In K. G. Jöreskog, and H. Wold, (Eds.), *Systems under indirect observation*, Part II (pp. 1–54). Amsterdam: North-Holland.

Wold, H. (1985). Partial least squares. In S. Kotz, and N. L. Johnson, (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 6 (pp. 581–591). New York: Wiley.

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In: A. Ruhe, & B. Kagstrom (Eds.), *Proceedings of the Conference on Matrix Pencils, Lectures Notes in Mathematics*. Heidelberg: Springer.

http://www.springer.com/978-3-540-32825-4

Handbook of Partial Least Squares
Concepts, Methods and Applications
(Eds.)V. Esposito Vinzi; W.W. Chin; J. Henseler; H. Wang
2010, X, 850 p. 238 illus., Hardcover
ISBN: 978-3-540-32825-4