UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biomediche, Cliniche e Sperimentali

Dipartimento di Scienze Cliniche e di Comunità

Corso di Dottorato di Ricerca in Statistica Biomedica

Ciclo XXVII – Settore scientifico disciplinare MED/01

TESI DI DOTTORATO DI RICERCA

Partial Least Square Path Modeling Approach in Biomedical Research

Doctoral Dissertation by Valentina Rosato

Advisor: Prof. Adriano Decarli

A.A. 2013/2014

Partial Least Square Path Modeling Approach in Biomedical Research

Doctoral Dissertation
by
Valentina Rosato
(R09565)

Prof. Adriano Decarli
Advisor

Defended in Milan, January 15, 2015

Jury:

Adriano Decarli	Università degli Studi di Milano
Luisa Bernardinelli	Università degli Studi di Pavia
Rocco Micciolo	Università degli Studi di Trento

Abstract

The Partial Least Squares Path Modeling (PLS-PM) is a method meant to estimate a network of causal relationships defined according to a theoretical model. The complexity of the theoretical construct is studied by taking into account the relationships among non measurable indicators (latent variables), represented by a set of observed variables (manifest variables). PLS-PM aims to estimate, through a system of interdependent equations based on simple and multiple regressions, the network of relations among the manifest variables and their own latent variable, and among the latent variables inside the model. The causal relationships among variables are represented through a Path Diagram, in which the latent variables are enclosed in circles and the manifest variables are enclosed in boxes. PLS-PM involves three sets of relations: 1) structural or inner model, 2) measurement or outer model, 3) the weight relations upon which latent variable scores can be calculated. The first model takes into account the relations among the latent variables and the second takes into account the relations between manifest variables and the corresponding latent variable. In the structural model each endogenous (dependent) latent variable is linked to the others by a multiple regression model. The structural design only assumes recursive models, i.e. the path diagram takes the form of a causal chain with no loops. Different types of measurement models exists, depending the kind of relationship: 1) reflective model (observed variables are considered being caused by the latent variable (i.e., indicators reflect the construct; the latent variable is considered as the cause of the manifest variables and each manifest variable is an

effect of the unique corresponding latent variable); 2) formative model (the latent variables are considered as being caused by its manifest variables); and 3) MIMIC model (multiple effect indicators for multiple causes, it represents a mixture of both the reflective and the formative models within the same block of manifest variables). Independently from the type of measurement model, the standardized latent variable scores are computed as a linear combination of its manifest variables and outer weights (the so-called weight relation).

Once the theoretical model is specified, the next phase in PLS-PM is the estimation of the model parameters. The PLS algorithm consists of three stages. The first stage is an iterative procedure of ordinary least squares regressions taking into account the relationships of the structural and measurement model, in order to calculate weights required to give final estimates for each latent variable. This first stage is the "core" stage in the PLS algorithm. Subsequently, the second and third stage involve the non-iterative estimation of the coefficients of the structural and measurement model, respectively. The structural model coefficients (path coefficients) are calculated by ordinary least squares regressions between latent variables. The measurement model coefficients (loading coefficients) are also estimated by regressions but taking into account the kind of mode to be used (reflective or formative).

PLS-PM has been widely used in economical (the customer satisfaction is a typical example) and psychological settings. In biomedical context, the published articles are scanty and generally published in open access journals.

The aim of this study was to apply the PLS-PM in a different field, since it has been widely used in economical (the customer satisfaction is a typical example) and

psychological setting. In biomedical context, the published articles are scanty and generally published in open access journals.

I used the PLS-PM method in order to analyze the adherence of the procedures provided for diagnosis, treatment (surgical and medical), and follow-up of breast cancer through a set of indicators. Indeed, the used approaches in this field since oversimplify the complex problem since they do not consider simultaneously multiple aspects of the diagnostic, therapeutic and follow-up pathways. This method has several strengths, as PLS-PM allows the reduction of dimensionality of several health indicators into a smaller number of latent variables (and more interpretable), and then allows to study causal relationships between these latent variables, representing the different aspects of the diagnostic, therapeutic and follow-up pathways. This method also requires no distributional assumptions with respect to the variables included in the model. The limit of this method is the bias deriving from the *a priori* selection of the relationships among latent variables and of the indicators used to characterize the latent variable. Although the limited sample size makes the analyses explorative-orientated only, the present study represents an unique example of PLS-PM application in the biomedical research, in particular in the evaluation of the adherence of the diagnostic and treatment procedures for breast cancer.

Indice

1. Essentials of Path Modeling

Path Modeling, also known as Structural Equation Modeling (SEM), is a generic term used to designate a set of different statistical techniques that meant to estimate a network of causal relationships defined according to a theoretical model. The concept of SEM refers to cause-effect relationships between variables which can be specified by a series of equations. The concept of Path Modeling refers to a graphical approach in which the relationships between variables (structural equations) are graphically displayed, through in what is known as Path Diagram.

Fornell defined SEM as a second generation of multivariate methods as it allows not only an exploratory approach (data then conceptualization) but also a theory-based approach. This method is used when we are interested in modeling a phenomenon of interest based on a theoretical framework. A theoretical model is imposed on the data, and the strength of the relationships is examined. In summary, path modeling is a useful set of methods that allows the combination of prior knowledge with measured data. The prior knowledge is provided by some theory for a certain

phenomenon of interest, in which a model for the cause-effects relationships among variables is proposed [1].

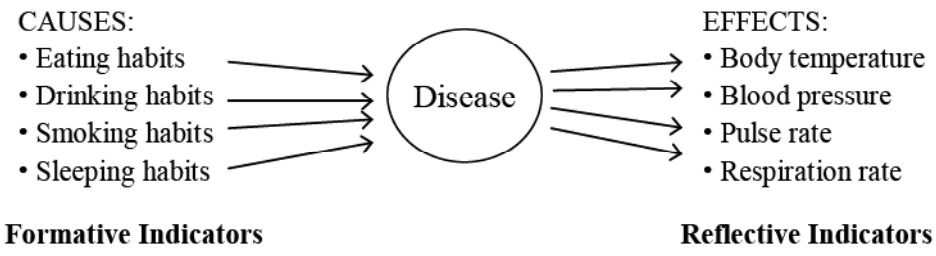
Path Modeling involves latent variables (LVs) which are theoretical variables that cannot be observed nor measured directly. Because these types of variables cannot be observed nor measured explicitly, LVs have to be measured through variables that are perfectly observable and measurable which are known as manifest variables (MVs) or indicators. LVs are very common in social sciences (e.g., psychology, sociology, and economy) in which there are many concepts of theoretical nature such as intelligence, socioeconomic status or industrial development.

1.1 Reflective versus formative indicators

Once we have assumed that LVs can be observed and measured indirectly through MVs or indicators, we need to consider the ways in which LVs are (indirectly) measured: LVs can be observed/measured in two ways: (1) through their consequences or effects reflected on their MVs; (2) through different indicators that are assumed to cause the LVs. In the first case, called *reflective way*, MVs are considered as being caused by the LV, whereas in the second case, called *formative way*, a LV is supposed to be formed by its MVs [2].

Suppose that a doctor is examining a patient trying to determinate the presence or absence of some disease. The doctor might evaluate the patient symptoms (e.g., body temperature, blood pressure, pulse rate, respiration rate, feelings of nausea). The doctor might ask about the patient's lifestyle (e.g., diet, drinking and smoking habits) that might be causing the disease. So, symptoms can be considered as reflective indicators because they *reflect* the disease, whereas lifestyle habits can be seen as formative indicators because they *form* (cause) the disease (Figure 1.1).

Figure 1.1. A latent variable (disease) measured by formative and reflective indicators [3].



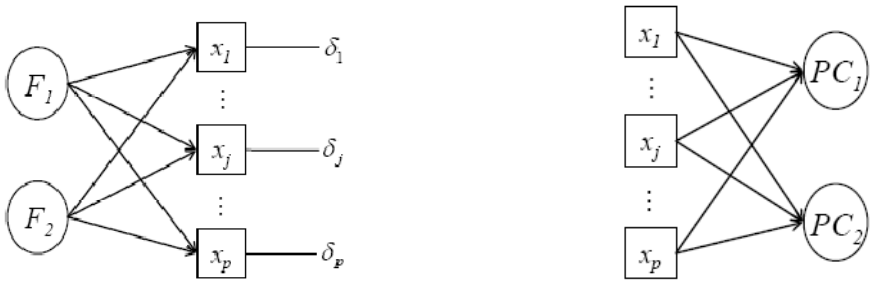
In the reflective way it is expected that different indicators are to be highly correlated because they are measuring the same concept [4]. The same cannot be said about formative indicators because a LV can be caused by two or more MVs mutually uncorrelated.

The difference between reflective and formative indicators is related to the conceptualization of two data analysis methods: factor analysis and principal component analysis. In factor analysis, the latent variables are called factors, and it is assumed that these factors explain the observed variables. In contrast, the LVs in principal component analysis are called components, which are obtained as linear composites of the observed variables. Under the factor analysis point of view, a factor F_j is associated to the observed variables in a reflective form, whereas under the principal component analysis point of view, a component PC_j can be represented as LV with formative indicators (Figure 1.2) [3].

Figure 1.2. Factors from factor analysis (F_j) and components from principal component analysis (PC_j) as latent variables.

Factors regarded as latent variables with reflective indicators.

Principal components regarded as latent variables measured by formative indicators.

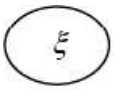
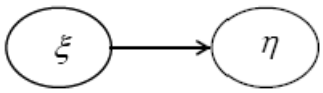


1.2 Path Diagrams

The main characteristic of path modeling technique is its graphical approach. The visual representation of the models are called *path diagrams*. They provide a graphical representation of the relationships among a set of variables, with the special property that they can be translated into a system of simultaneous equations. The great advantage of path diagrams is that they allow for the visualization of the relationships and, in terms of a causal model, its graphical display makes it possible to understand the conceptualization of the model [3].

Path diagrams have a conventional notation (Table 1.1). Briefly, LVs are enclosed in circles and MVs are enclosed in boxes.Variables may be grouped in two classes: endogenous or dependent (variables caused by one or more variables) and exogenous or independent (variables not caused by any other variables). Endogenous LVs are usually represented by η whereas exogenous LVs are represented by the Greek letter ξ. Arrows show causal relationships among variables (either latent or manifest), and the direction of the array defines the direction of the relation, i.e. variables receiving the array are to be considered as endogenous variables in the specific relationship.

Table 1.1. Main path diagram notation

	Manifest variable
	Latent variable
	Exogenous latent variable (ξ) causes endogenous latent variable (η)
	Correlation between two manifest variables

1.3 Path Modeling analysis

The model includes two parts or sub-models: the measurement model and the structural model. The first model represents how each construct is measured by its indicator variables. The structural model involves the causal relations among the constructs and it is represented by a simultaneous system of equations among the latent variables. The path modeling process starts with a theoretical framework that involves the establishment of the theoretical relationships among constructs or latent variables. The subsequent step is deciding how many and which observed variables will be considered as indicators of the constructs. The selection of manifest variables and its number is sometimes a subjective matter and no criterion exists on this point. Regarding the number of indicators some authors suggest to use as many indicators as possible although having too many may present problems with model fitting [3]. Once the relationships of the model are fixed, they can be visualized in the form of a path diagram. The next step involves the mathematical specification of the model, that is, its translation into a system of equations, followed by the estimation phase and the validation of results.

In conclusion, Path Modeling is a methodology for the analysis of indirectly measured cause and effect relationships in complex systems. This analysis can be accomplished under two major approaches: confirmatory and predictive purposes. The confirmatory approach is concerned with theory development and testing by testing whether the assumed theory and hypotheses can be confirmed; the model is analyzed by examining the covariance structure of the data and testing probabilistic assumptions. The predictive approach focuses on making predictions about the

outcome variables of interest and it involves the variability of data in the form of a prediction model of the dependent variables. The path modeling method for confirmatory purposes has the generic name of Covariance Structure Analysis (CSA), also known as LISREL. In turn, the predictive oriented methodology is Partial Least Squares Path Modeling (PLS-PM).

<<When we use a covariance-based SEM approach we implicitly assume that the data is generated by some "true" theoretical model. In this scenario, the goal of CSA is to recover the "true" model that gave rise to the observed covariances. Briefly, when using CSA we are concerned with fitting a model and reproducing the observed covariances. This approach resorts to classical theory of statistical inference and is based on a heavy use of distributional assumptions about the behavior and personality of the data. Consequently, the analyst is forced to move slowly; and the modeling process requires careful thought and stringent justifications that more often than not end up compromising the whole analysis with the bizarre (and sometimes contradictory) principle of the data must follow the model. In contrast, […]PLS-PM models are not considered to be ground truth, but only an approximation with useful predictiveness. In other words, PLS-PM assumes no model by which the data were generated. There is only the data and nothing but the data. In this sense, PLS-PM follows the spirit of a dimension reduction technique that we can use to get useful insight of the data on hand. The ultimate goal in PLS-PM is to provide a practical summary of how the set of dependent variables are systematically explained by their sets of predictors. Besides the description of PLS-PM as an alternative approach to SEM covariance structure analysis, PLS-PM can also be regarded as a technique for analyzing a system of relationships between multiple blocks of variables, or if you

want to put it in simple terms, multiple data tables. […] In summary, we can regard PLS-PM as a coin with the two following faces:

• PLS Path Modeling as a component-based alternative for estimating Structural Equation Models.

• PLS Path Modeling as a method for analyzing a system of linear relationships between multiple blocks of variables.>> [5].

PLS-PM was originally developed as an analytical alternative to CSA for situations where the theory is weak and where the general assumptions of CSA are not met. The overall goal of PLS is to use observed independent variables to predict observed dependent variables. This is realized indirectly by extracting independent and dependent latent variables from observed variables. This is done in such a way that they optimally address one or both of these two goals: explaining response variation and explaining predictor variation. The goal is to predict the dependent variables (both latent and manifest) by minimizing the residual variances of the endogenous (i.e. dependent) variables. In particular, the method of partial least squares balances the two objectives, seeking latent variables that explain both response and predictor variation [3].

2. PLS Path Model (PLS-PM): the Method

The basic idea of Partial Least Squares (PLS) methods is the estimation process used to calculate model parameters. This process is performed by separating the parameters to be estimated in parts (hence the term partial) in order to apply an iterative procedure of least squares regressions to calculate them.

PLS methods are not derived through probabilistic reasoning or numerical optimization. Moreover, PLS has not assumptions about variables and error distributions and for this reason it is called as a “soft modeling”. It doesn’t rely on the classic inferential tradition. Variables can be numerical, ordinal, or nominal.

Partial Least Squares Path Modeling (PLS-PM) is one of the PLS techniques. It is a multivariate technique of second generation by combining causal modeling with data analysis features. PLS-PM is a statistical method that has been developed for the analysis of structural equation models with latent variables, specially designed to

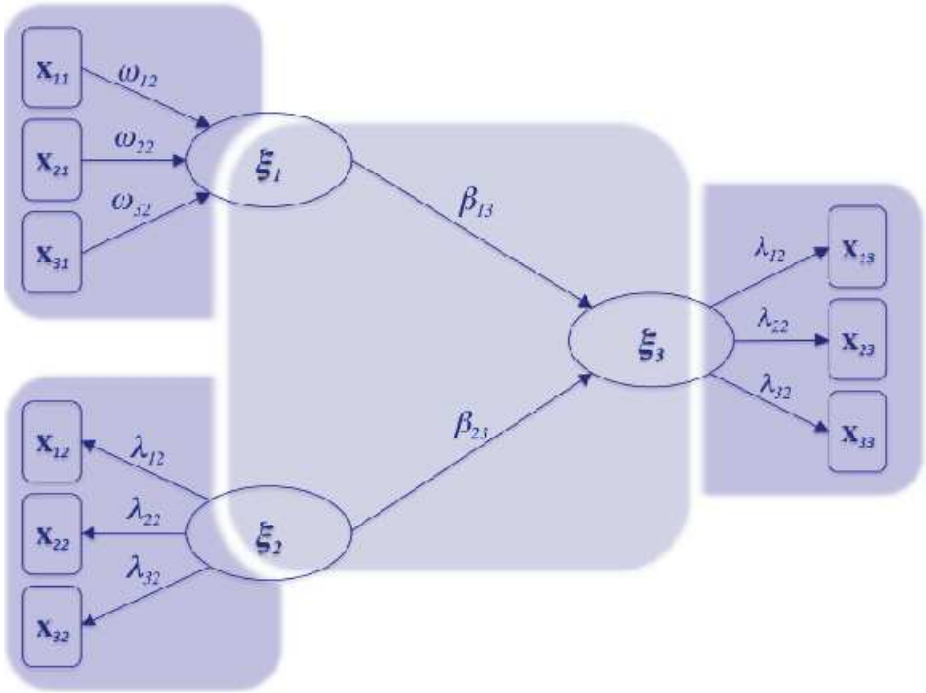
provide an alternative approach to the LISREL models. As opposed to the covariance-based approach, PLS is prediction oriented aiming to obtain estimates of latent variables for prediction purposes, maximizing the variance explained for the dependent variables (both latent and manifest variables).

PLS-PM is a methodology meant to estimate a network of causal relationships defined according to a theoretical model. The complexity of the theoretical construct is studied taking into account the relationships among non measurable indicators (latent variables), represented by a set of observed variables (manifest variables). PLS-PM aims to estimate, through a system of interdependent equations based on simple and multiple regressions, the network of relations among the manifest variables and their own latent variables, and among the latent variables inside the model. The causal relationships among variables are represented through a Path Diagram, in which the latent variables are enclosed in circles and the manifest variables are enclosed in boxes.

PLS-PM involves three sets of relations: 1) structural or inner model, 2) measurement or outer model, 3) the weight relations upon which latent variables scores can be calculated. The first model takes into account the relations among the latent variables and the second one takes into account the relations between manifest variables and the corresponding latent variable. In the structural model each endogenous (dependent) latent variable is linked to the others by multiple regression model. The basic structural design only assumes recursive models, i.e. the path diagram takes the form of a causal chain with no loops. Different types of measurement models exists, depending on kind of relationship: 1) reflective model (the observed variables are considered being caused by the latent variable (i.e.,

indicators reflect the construct; the latent variable is considered as the cause of the manifest variables and each manifest variable is an effect of the unique corresponding latent variable); 2) formative model (the latent variables are considered as being caused by its manifest variables); and 3) MIMIC model (multiple effect indicators for multiple causes, it represents a mixture of both the reflective and the formative models within the same block of manifest variables). Independently from the type of measurement model, the standardized latent variable scores are computed as a linear combination of its manifest variables and outer weights (the so-called weight relation).

Figure 2.1 Path Diagram: The structural model is painted in blue grey, the measurement model in sky blue.



Once the theoretical model is specified, the next phase in PLS-PM is the estimation of the model parameters. The PLS algorithm consists of three stages. The first stage is an iterative procedure of ordinary least squares regressions taking into account the relationships of the structural and measurement model, in order to calculate weights required to give final estimates for each latent variable. Subsequently, the second and third stages involves the non-iterative estimation of the coefficients of the structural and measurement model, respectively. The structural model coefficients (path coefficients) are calculated by ordinary least squares regressions between latent variables. The measurement model coefficients (loading coefficients) are also estimated by regressions but taking into account the kind of mode to be used (reflective or formative).

2.1 Structural model

The structural model (also known as inner model) considers only the latent variables, which are assumed to be linearly interconnected according to a causal-effect relationship model.

The associations among latent variables can be represented by a linear multi-equation system which has to be recursive. Latent variables can play both predictee and predictor roles: a latent variable that is never predicted is called exogenous, otherwise is called endogenous. For simplicity, no distinctions in notation are made between endogenous and exogenous constructs; all latent variables will denote as x. The linear equations take the following form:

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

with predictor specification

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

where the parameter β_{ij} is called the path coefficient (representing the path from the i-th to the j-th latent variable), ζ_i is the inner residual term, and the index i ranges over all predictors of ξ_j. Predictor specification implies that the residuals have zero mean and are uncorrelated with the latent variables.

2.2 Measurement model

The measurement model (also known as outer model) establishes the relation between a block of manifest variables and its latent variable. Each indicator is supposed to be associated with just one latent variable Because the latent variable is an unmeasured variable, it has to be indirectly measured trough the manifest variables, hence the name measurement model. There are three options to establish the connections of the manifest variables to its latent variable:

a) Reflective way

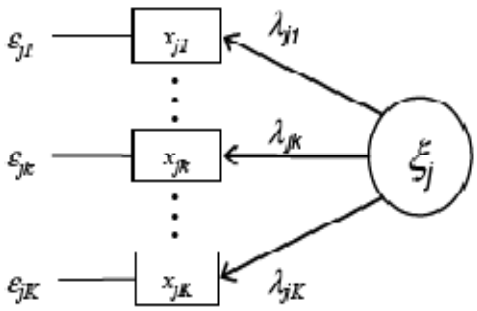
b) Formative way

c) Multiple effect indicators for multiple causes way (MIMIC)

a) Reflective way

In the reflective way the latent construct is considered as the cause of the indicators.

Figure 2.2 Path diagram of reflective way



In this case, the manifest variables can be considered reflects or manifestations of their latent variable. The manifest variable x_{jk} is assumed to be a linear function of its latent variable ξ_j

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

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

When predictor specification is adopted,

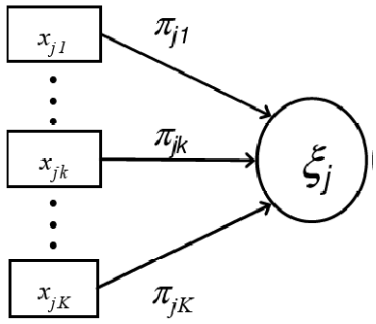
$$E(x_{jk} \mid \xi_j) = \lambda_{0jk} + \lambda_{jk}\xi_j ,$$

which implies that the residuals have zero mean and are uncorrelated with the manifest variables.

b) Formative way

In the formative way the latent construct is considered as being caused by its indicators

Figure 2.3 Path diagram of formative way



The latent variable ξ_j is assumed to be a linear function of its manifest variables x_{jk}

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

assuming predictor specification

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

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

c) MIMIC way

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

In this case there are two linear equations

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

where the index *h* ranges over all reflective manifest variables, and the index *l* ranges over all formative manifest variables, *h+l = k.*

when predictor specification is adopted,

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

Main differences between reflective and formative manifest variables can be summarize as in Table 2.1.

Table 2.1. Main differences between reflective and formative manifest variables.

Reflective	Formative
The direction of causality is from construct to measure	The direction of causality is from measure to construct
It is expected that the measures are correlated between them	It is not expected that the measures are correlated between them. The model does not imply the internal consistency
The elimination of an indicator from the measurement model does not alter the meaning of the construct	The elimination of an indicator from the measurement model can alter the meaning of the construct
The measurement error is taken into account for each item	The measurement error is taken into account for the construct

2.3 Weight relations

Although the measurement model specifies the relations between the latent variables and their set of indicators, this specification is done in a conceptual level. In other words, the outer relations refer to the indicators and the “true” latent variable. However, we do not really know it. For this reason the weight relations must be defined. Latent variable estimates or scores are defined as follows:

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

where $\tilde{w}_{jk}$ are the weights used to estimate the latent variable as a linear combination of their observed manifest variables. “Note that by using weight relations the problem of factor indeterminacy, present in covariance structure models, is avoided in PLS” [3].

2.4 Soft modeling

Predictor specification implies that residual terms have zero mean and are uncorrelated with the independent variables (latent or manifest ones). Moreover, the outer model residuals are uncorrelated with all latent variables and with the inner model residuals. As consequence, the Ordinary Least Squares (OLS) estimates are consistent and the prediction using OLS estimates is consistent with minimum residual variance. It is also important to remark that PLS does not restrict the structure of the residual covariance [3]. The relevant feature of PLS-PM (as well as all the PLS techniques) is that no assumptions need to be made on the data about distribution and observations independently distributed. This means that PLS approach avoids the rigid assumptions of the method of maximum likelihood. For these reasons, PLS approach is more flexible, being known as a soft modeling technique [6].

2.5 PLS-PM algorithm

Once the theoretical model has being specified, the next phase in PLS-PM is the estimation of the parameters carried out by the PLS algorithm. The PLS estimation algorithm proceeds in three stages.

1) The first step consists of an iterative procedure of simple and/or multiple regressions taking into account the relationships of the inner model, the outer model and the weight relations. The result is the estimation of a set of weights which are used to calculate the latent variable scores as linear combinations of their associated manifest variables;
2) the second and third steps involve the non-iterative estimation of the structural model coefficients (path coefficients) and
3) the measurement model coefficients (loadings).

2.5.1 PLS-PM algorithm – stage 1

This first stage is the "core" stage in the PLS algorithm. The goal of this stage is the calculation of weights required to give final estimates for each latent variable ξ_j as a linear combination Y_j of its K_j manifest variables x_{jk}

$$\hat{\xi}_j = Y_j = \sum_k \widetilde{w}_{jk} x_{jk}$$

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

The process to calculate the weights follows an iterative mechanism that takes into account the hypothesized relations of the structural and the measurement models (Table 2.2). For each model (structural and measurement) there is an associated approximation of the latent variables: outside approximation for the measurement model, and inside approximation for the structural model. Several options for performing first stage are available depending on how the relations between latent variables in the structural model are established, and also on how the indicators are associated to their latent variables.

Table 2.2 PLS-PM iterative algorithm

1) Start with arbitrary outer weights w

e.g. w_1=1, w_2=1, ... , w_p=1

2) External approximation

Compute LVs as linear combinations of their MVs

e.g. $Y_1 = w_1x_1 + ... + w_kx_k$

3) Updating inner weights

Take into account the structural relationships between LVs

e.g. e_{12} = cor(Y_1, Y_2), only if LV_1 is connected with LV_2

4) Internal Approximation

Re-compute LVs taking into account their LV neighbors

e.g. $Z_1 = e_{12}Y_2 + e_{13}Y_3$ only if LV_1 is connected with LV_2 and LV_3

5) Updating outer weights

Re-compute w with the LVs form the internal approximation

e.g. under mode A, w_1= cov(x_1, Z_1) or w_1= cor(x_1, Z_1) if MVs are standardized

6) Check for convergence

e.g. $|w_{old} - w_{new1}| < 0.00001$

7) Repeat steps 2 - 6 until convergence

LV= latent variable, MV= manifest variable

Stage 1.1-1.2: External approximation

The iterative process begins with an initial proxy of each latent variable as a linear combination of its manifest variables

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

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

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

The standardized latent variable is finally expressed as:

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

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

The idea behind the outside approximation is to obtain a set of weights to estimate a latent variable accounting for as much variance as possible for the indicators and the constructs. The algorithm begins with an initial outside approximation of the latent variables by using arbitrary weights which are scaled to obtain unit variance for the latent variables.

Stage 1.3-1.4: Internal approximation

In this step the connections among latent variables in the inner model are taken into account in order to obtain a proxy of each latent variable calculated as a weighted aggregate of its adjacent latent variables. The internal estimation Z_j of ξ_j is defined by:

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

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

The connections among latent variables in the inner model are taken into account only when two latent variables are connected by an arrow. In other words, inner weights e_{ji} between two constructs exist only when there is an arrow between ξ_j and ξ_i.

There are three options to calculate the inner weights:

- Centroid scheme. This scheme only considers the sign direction of the correlations between a latent variable and its adjacent (neighboring) latent variables.

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

  Some problems may be present when a correlation is close to zero, causing a sign changes during the iterations from +1 to -1.

- Factor scheme. To avoid the problems of the centroid scheme, the factor scheme uses the correlation coefficient as the inner weight instead of using

only the sign of the correlation. This scheme considers not only the sign direction but also the strength of the paths in the structural model.

- Path scheme. The latent variables are divided in antecedents (predictors) and followers (predictands) depending on the cause-effects relationships between two latent variables. An latent variable can be either a follower, if it is caused by another latent variable, or an antecedent if it is the cause of another latent variable. If ξ_i is a follower of ξ_j then the inner weight is equal to the correlation between Y_i and Y_j. On the other hand, for the antecedents ξ_i of ξ_j, the inner weights are the regression coefficient of Y_i in the multiple regression of Y_j on the Y_i's associated to the antecedents of ξ_j. The path weighting scheme has the advantage of taking into account both the strength and the direction of the paths in the structural model. However, this scheme presents some problems when the latent correlation matrix is singular

The centroid scheme is the Wold's original algorithm scheme, whereas the other two are implemented in Lohmöller's version. The centroid scheme represents the default option in the software R.

In practice, choosing one weighting scheme in particular over the others has little relevance on the estimation process and does not influence the results significantly [7].

Stage 1.5: Updating outer weights

There are three ways of calculating the outer weights w_{ji} (mode A, mode B, and mode C). Each mode corresponds to a different way of relating the manifest variables with the latent variables in the theoretical model. Mode A is used when

the indicators are related to their latent variable through a reflexive way. Instead, mode B is preferred when indicators are associated with their latent variable in a formative way. Mode C is supposed to be used when the indicators of an LV are connected by MIMIC way, and it is rarely used in practice.

- Mode A. In the reflective way, each weight w_{jk} is the regression coefficient of Z_j in the simple regression of x_{jk} on Z_j:

  $$x_{jk} = w_{jk} Z_j.$$

  As $Z_j$is standardized:

  $$w_{jk} = (Z_j^{'} Z_j)^{-1} Z_j^{'} x_{jk} = cov(x_{jk}, Z_j) = cor(x_{jk}, Z_j).$$

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

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

  $$Z_j = \sum_k w_{jk} x_{jk}, \quad \mathrm{w_j} = (X_j' X_j)^{-1} X_j' Z_j,$$

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

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

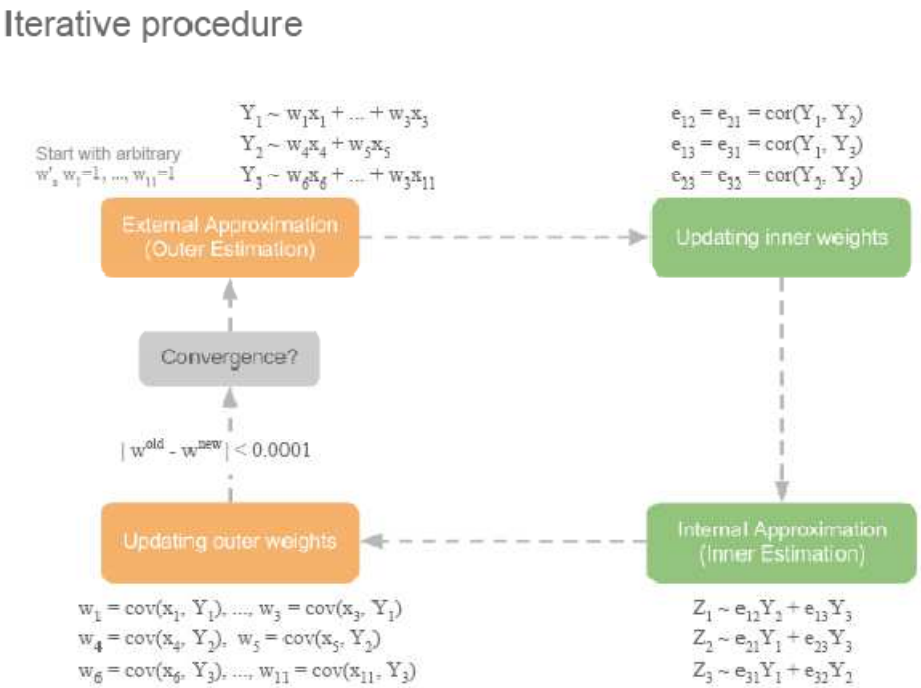
- Mode C. This case is implemented in Lohmöller's version and it is a special case of mode B. The MIMIC way is a kind of mix between reflective and formative ways, so the path coefficients for the *h* manifest variables related in a reflective way are estimated by a simple linear regression: $x_{jh} = p_{jh} Z_j$ and the path coefficients for the *l* manifest variables related in a formative way are estimated by a multiple linear regression: $Z_j = \sum_l g_{jl} x_{jl}$.

Stage 1.6: Check for convergence

In every iteration step (S = 1, 2, 3,…,) convergence is checked comparing the outer weights of step S against the outer weights of step S-1.

Wold proposed $|\tilde{w}_{jk}^{S-1} - \tilde{w}_{jk}^{S}| < 0.00001$ as a convergence criterion [8]. Convergence is not guaranteed although it is always found in practice [9].

Figure 2.4. The first (and core) stage of PLS path modeling algorithm [3].



2.5.2 PLS-PM algorithm – stage 2

The second stage of the algorithm consists in the calculation of the path coefficient estimates $\hat{\beta}_{ji}$, according to the structural or inner model.

The path coefficients are estimated by ordinary least squares in the multiple regression of Y_j on the Y_i's related to it

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

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

The path coefficients can be interpreted as correlation coefficient, if the manifest variables are standardized.

2.5.3 PLS-PM algorithm – stage 3

The third stage of the algorithm consists in the calculation of the loading coefficient estimates $\hat{\lambda}_{ji}$, according to the measurement or outer model.

The loadings are estimated depending on the corresponding way.

In the reflective way, the loading coefficients are the regression coefficients of the simple linear regression of each manifest variable x_{jk} on the corresponding latent variable Y_j:

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

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

In the formative way, the weight coefficients $\hat{\pi}$ coincide with the outer weights obtained in the first stage. This is because we perform the multiple linear regression of Y_j on x_{jk}:

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

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

2.5.4 Location parameters

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

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

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

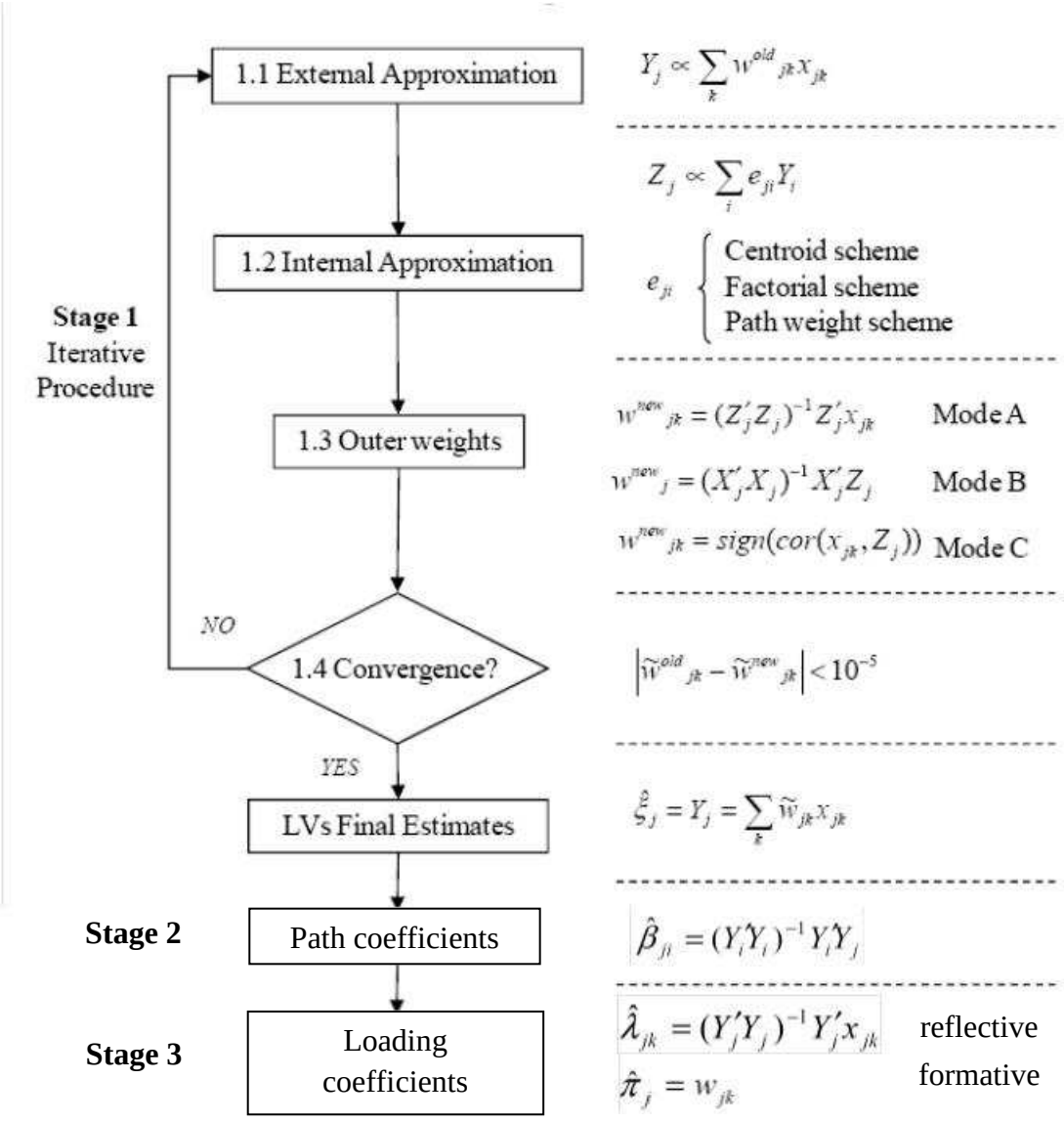
$$E(\xi_j \mid x_{jk}) = \pi_{0j} + \sum_k \pi_{jk} x_{jk} \qquad \text{(formative way measurement model)}$$

These parameters correspond to the location parameters, that is, we take into account the mean of the manifest and latent variables. However, until now, we have only considered standardized manifest variables (zero mean and unit variance). In fact, because it has been imposed that way during the algorithm, the estimated latent variables are also standardized. In order to obtain the location parameters the researcher must consider whether it makes sense to calculate them. This decision concerns data scales which are the key criteria to decide whether to estimate location parameters. We must say that this aspect on scales is not considered in Wold's original algorithm. It was developed by Lohmöller (1989) who extended PLS-PM to applications with mixtures of categorical and interval-scaled data.

2.6 PLS-PM flowchart.

In conclusion, the PLS-PM algorithm can be summarized by the flowchart in Figure 2.5.

Figure 2.5. Flowchart of the Partial Least Square Path Modeling (PLS-PM) algorithm.



2.7 PLS-PM validation

PLS-PM lacks a well identified optimization criterion; however it provides some quality indices or measures. The validation of a PLS-PM requires the analysis and interpretation of both the measurement and the structural model. This order has to be respected because we must first check that we are really measuring what we are assuming to measure, before any conclusions can be drawn regarding the relationships among the latent variables.

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

2.7.1 Measurement model validation: reflective measures

Unidimensionality of indicators

When you have a block of reflective indicators it is supposed that those indicators will reflect, to some extent, the latent variable that they are associated with. Actually, it is assumed that the latent variable is the cause of its indicators. This means that if a construct changes (increases or decreases), then the indicators associated with it will also change in the same direction. Thus, it is logical to suppose that the indicators are closely related in such a way that they are in one dimensional space.

The reflective indicators must be in a space of one dimension since they are practically indicating the same latent variable. In PLS-PM we have three main indices to check unidimensionality:

• Cronbach's alpha.

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

where p is the number of variables.

It evaluates how well a block of indicators measure their corresponding latent construct. In this case, the observed variables are required to be standardized and positively correlated. If the number of variables increases, Cronbach's alpha increases as well. As it is expected. As a rule of thumb, a block is considered as unidimensional when Cronbach’s alpha is larger than 0.7.

• Dillon-Goldstein's rho

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

where t_{j1} is the first principal component of the j-th block of indicators.

As in the case of Cronbach's alpha, the Dillon-Goldstein's rho is also focused on the variance of the sum of variables in the block of interest. As a rule of thumb, a block is considered as unidimensional when Dillon-Goldstein's rho is larger than 0.7. This index is considered to be a better indicator than the Cronbach's alpha because it takes into account to which extent the latent variable explains the block of indicators.

• Check the first eigenvalue of the indicators' correlation matrix

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

Indicators are well explained by their latent variables

Then, we should check that indicators are well explained by their latent variables.

We check it by means of three tools:

• Communality

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

is the communality for the k-th manifest variable of the j-th block.

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

Communality measures how much of the manifest variable variance is explained by its own latent variable. In other words, how well the manifest variables described the related latent variable or the part of variance between a construct and its indicators that is common to both. One expects to have more shared variance between latent variable and manifest variables than error variance, that is: $\lambda^2_{\text{jk}} > var(\varepsilon_{\text{jk}})$ with $\text{var}(\varepsilon_{\text{jk}}) = 1 - \lambda^2_{\text{jk}}$.

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

It is possible to measure the quality of the whole measurement model by mean of the average communality index, that is a weighted average of all the block-specific

communality indexes with weights equal to the number of manifest variables in each j-th block: $\overline{Com(\xi_j, x_{kj})} = \frac{1}{P}\sum_j P_j Com(\xi_j, x_{kj})$.

• Composite reliability

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

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

• Average Variance Extracted

$$\text{AVE} = \frac{\sum \lambda_{jk}^2}{\sum \lambda_{jk}^2 + \sum \text{var}(\varepsilon_{jk})}$$

Average Variance Extracted (AVE) is similar to Jöreskog's composite reliability, but AVE attempts to measure the amount of variance that an latent variable captures from its indicators in relation to the amount of variance due to measurement error. AVE should be larger than 0.50 which means that 50% or more variance of the indicators should be accounted for.

Differentiation between construct

Then, we should assess the degree to which a construct is different from other constructs. This is done by verifying that the shared variance between a construct and its indicators is larger than the shared variance with other constructs. In other words, no indicator should load higher on another construct than it does on the construct it

intends to measure. We calculate the correlations between a construct and other indicator besides its own block. If an indicator loads higher with other constructs than the one it is intended to measure, we might consider its appropriateness because it is not clear which construct or constructs it is actually reflecting.

2.7.2 Measurement model validation: formative measures

Unlike reflective indicators, formative indicators are considered as causing a latent variable. Formative indicators do not necessarily measure the same underlying construct, that is, formative indicators are not supposed to be correlated. For this reason, formative measures cannot be evaluated in the same way of reflective measures; and all the assessment criteria based on the loadings are discarded in the formative measures.

In this way we compare the outer weights of each indicator in order to determine which indicators contribute most effectively to the construct. Attention must be paid in order to avoid misinterpreting relative small absolute values of weights as poor contributions. If we are considering the elimination of some indicator, this should be done based on multicollinearity: the elimination is recommended if high multicollinearity occurs.

2.7.3 Structural model validation

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

• the coefficients of determination R^2

The R^2 is calculated for the endogenous latent variables. R^2 evaluate the quality of each structural equation. For each regression in the structural model we have a R^2 that is interpreted similarly as in any multiple regression analysis: R^2 indicates the amount of variance in the endogenous (dependent) latent variable explained by its independent latent variables.

• the redundancy

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

is the redundancy index for the k-th manifest variable associated to the j-th block, where ξ_j is the j-th endogenous latent variable; x_{jk} is the k-th indicator associated to ξ_j; λ_{jk}^2 is the communality; $R_{j|\xi}^2$ is the R^2 coefficient from the regression between ξ_j and its predictors ξ_i.

Redundancy measures the percent of the variance of indicators in an endogenous block that is predicted from the independent latent variables associated to the endogenous latent variable. Another definition of redundancy is the amount of variance in an endogenous construct explained by its independent latent variables. In other words, it reflects the ability of a set of independent latent variables to explain variation in the dependent latent variable. High redundancy means high ability to predict. In particular, the researcher may be interested in how well the independent latent variables predict values of the indicators' endogenous construct.

A global quality measure of the structural model is provided by the average redundancy index: $\overline{Rd(\xi_j, x_{jk})} = \frac{1}{J}\sum_j Rd(\xi_j, x_{jk})$, where J is the number of endogenous latent variables in the model.

• the Goodness-of-Fit (GoF)

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

where J is the number of latent variables in the model; J* is the number of endogenous latent variables and j* indicates an endogenous block; $\text{cor}^2(\text{x}_{\text{jk}}, \xi_{\text{j}})$ is

the correlation between the k-th manifest variable of the j-th block and the corresponding latent variable; $R^2(\xi_{j*};\ \xi'_{j*}s \rightarrow \text{predicting}\ \xi_{j*})$ is the R^2 value of the regression between the j*-th endogenous latent variable and its associated predictors $\xi'_{j*}s$.

The first term is the average communality of each block which measures the quality of the measurement model. The second term is the average of the determination coefficient for each endogenous construct according to latent variables which explain it. In other words: $\text{GoF} = \sqrt{(\text{Average Communality}) * (\text{Average } R^2)}$. Hence, GoF is a compromise between the quality of the measurement model and the quality of the structural model. Acceptable/good values within the PLS-PM community are GoF >0.7. Since it takes in to account communality, this index is more applicable to reflective indicators than to formative indicators. However, you can also use the GoF index in presence of formative blocks, in which case more importance will be given to the average R^2.

2.7.4 Validation by resampling

Since PLS-PM is a soft modeling approach, that is it does not imply distributional assumptions, significance levels for the parameter estimates (based on normal theory) are not suitable. Instead, it is possible to estimate the significance of the parameters based on resampling procedures are used to obtain information about the variability of the parameter estimates. For example, bootstrapping it is a non-parametric approach for estimating the precision of the PLS parameter estimates. The bootstrap procedure is the following: M samples are created in order to obtain M estimates for each parameter in the PLS model. Each sample is obtained by sampling with replacement from the original data set, with sample size equal to the number of cases in the original data set.

2.8 PLS-PM for non-metric data

PLS-PM is a technique born to handle quantitative variables. However, in the practice categorical indicators could be used to measure complex concepts as well.

To overcome this problem a recent technique has been proposed by G. Russolillo [10], the Non-Metric Partial Least Squares (NM-PLS) algorithm. It consists in a new class of PLS algorithms that allow the PLS iteration to work as an optimal scaling algorithms, calculating iteratively both scaling and model parameters.

In the Non-Metric PLS-PM algorithm the computation of the latent variables starts with an arbitrary choice of their inner estimates $\vartheta_1, \ldots, \vartheta_Q$. Afterwards, a new first step is added in each cycle of the iterative procedure. It is a quantification step, in which each categorical indicator is transformed in a quantitative one; this new quantified indicator x^*_{pq} is obtained as the orthogonal projection of ϑ_q. on the space spanned by the columns of $\tilde{X}_{pq}$. From a computational point of view,

$$\boldsymbol{x}^*_{pq} = \tilde{\boldsymbol{X}}_{pq}\left(\tilde{\boldsymbol{X}}'_{pq}\tilde{\boldsymbol{X}}_{pq}\right)^{-1}\tilde{\boldsymbol{X}}'_{pq}\boldsymbol{\vartheta}_q .$$

The procedure continues with the second and the third steps, i.e. the inner estimation and the outer estimations of each latent variable. Once new outer estimates are computed, the cycle restarts with the quantification step and it is iterated until the convergence between inner and outer estimations is reached.

This procedure yields as output both scaling and model parameters. It assures that quantified indicators show suitable properties in terms of optimality and interpretability. The scaling parameters maximize correlation of the quantified indicator with the inner estimate of the own LV, and as consequence its weight in the construction of the LV in a reflective scheme. Moreover, the weight of each

quantified indicator can be expressed also in terms of part of variability of ϑ_q explained by $\tilde{x}_{pq}$'s modalities. In particular, it is possible to show the following equivalence:

$$\boldsymbol{\rho}_{\boldsymbol{x}^*_{pq},\vartheta_q} = \boldsymbol{\eta}_{\boldsymbol{x}_{pq},\vartheta_q}$$

Hence, the weight of x^*_{pq} reflects the predictive capability of the categories of x_{pq} with respect to ϑ_q, measured by the correlation ratio squared root. It is for this reason that the NM-PLSPM algorithm is very useful to yield reliable weights for building composite and complex indicators from simple indicators observed on a variety of measurement scales.

3. The evaluation of the activities and decisions in health care systems: a cohort study from Lombardy

3.1 Introduction

When evaluating the degree to which the care delivered to oncologic patients − such as woman with breast cancer − is adherent to evidence-based guidelines, it is necessary to consider multiple aspects of the diagnostic, therapeutic and follow-up pathways [11]. In fact, single process indicators are informative variables which allow to concisely evaluate a single aspect of complex phenomena. However, single indicators do not give a complete view of the cure paths and their appropriateness or adherence to the guidelines. Coherent sets of quality indicators have been then developed from evidence-based guidelines, and chosen to measure the different aspects of the clinical pathway suitable for a patient with specific characteristics, such as stage and comorbidities. Nevertheless, it is difficult to simultaneously evaluate many different indicators and a single summary measure it is deemed

necessary, but obtaining a methodologically sound and easy to interpret measure represents a challenge [12,13].

In the literature, different approaches have been proposed [14,15]. Some are along the line of providing a simple summary indicator that can be easily understood from health professionals and patients, such as using the proportion of indicators met by each patient or the all-or-non approach [16]. These approaches have advantages if we want to promote a widespread use of a quality assessment tool. On the other side, it oversimplify a complex problem. The evaluation of the entire diagnostic and therapeutic pathway is complex, as it involves elements of the decision process that cannot be directly measurable or observable. Also, the components of the decisional process are dependent and influenced by each other. For these reasons, other statistical methods are needed. One proposed approach uses a hierarchy or selection of the most relevant indicators, with respect to the outcome, and a system of weights to combine them. Although this approach better reflects complexity, the assigned weights are fairly arbitrary [15,17]. Other proposed methods to summarize a set of indicators into a single, or a few measures use latent variables [18,19].

The aim of the present study is to evaluate the appropriateness of the procedures performed in each step of the breast cancer are pathway (i.e., diagnosis, surgical and medical treatment, and short term follow-up) through the estimation of a summary measure for each of them, and to investigate their relationships. For this purpose, we used a Partial Least Square (or Projection to Latent Structures) path model (PLS-PM) approach. The cohort under study is constituted of incidents breast cancer cases occurring between 2007 and 2009 and followed for at least one year after the diagnosis of cancer in six local health authorities of Lombardy, Northern Italy.

3.2 Material and methods

Population and database

We included all incident breast cancers occurring between 2007 and 2009 in the geographic area corresponding to six local health authorities of Lombardy, Northern Italy (5,320,272 inhabitants on 31 December 2012) [20]. Most of them included the incident breast cancer cases identified by the cancer registry; two local health authorities included all cases derived from a validated algorithm to identify breast cancer cases from hospital discharge records [20]. All cases with another diagnosis or hospitalization for cancer (except for skin cancer) between 1990 and 2009 were excluded from the analysis in order to include cases with only breast cancer and consequently to evaluate the specific diagnostic and therapeutic pathway.

From a systematic review of the literature and considering only indicator that could be calculated with the available information, we defined a preliminary set of indicators.

The indicators were then improved and assessed for their mean and consistency by epidemiologist and breast cancer surgeons, oncologists and radiotherapists through several meetings and the final set was composed by several indicators.

Through deterministic record linkage between each registry and the administrative available databases (hospitalizations, outpatient, pharmaceutical prescription and specific database for anticancer drugs), we calculated for each patient the indicators to measure the appropriateness of the procedures provided for diagnosis, treatment (surgical and medical) and follow-up.

Subsequently we calculated the indicators by each local health authority: the indicators represent the proportion of subjects in this cohort with a specific

caractheristic for each local health authority, e.g., the proportion of patients with ≥85 years. Thus, we have a dataset with 6 observations or rows (6 local health authorities) and more than 20 indicators or columns.

Statistical analysis

To examine our research question we performed a PLS-PM analysis [21]. PLS-PM is a methodology meant to estimate a network of causal relationships defined according to a theoretical model and to represent the causal relationships through a graphic, called diagram. The complexity of the theoretical construct is studied taking into account the relationships among non measurable indicators (latent variables), represented by a set of observed variables (manifest variables). In a path model approach, variables are grouped in two classes: 1) those that are caused by one or more variables (endogenous or dependent variables), and 2) those that are not caused by any other variables in the diagram (exogenous or independent variables).

PLS-PM involves two type of models: 1) the measurement or outer model, taking into account the relations between manifest variables (or indicators) and the corresponding latent variable (or domain); 2) the structural or inner model, taking into account the relations among the latent variables, i.e. between each endogenous (dependent) latent domain and other latent domains. Different types of measurement models exists, depending on the kind of relationship between indicators and domain; we choose a reflective model, i.e. manifest indicators reflect the domain.

Once the hypothesized model is specified, a 3 stages algorithm is used to estimate the parameters. The first stage is an iterative procedure of ordinary least squares (OLS) regressions taking into account the relationships of the structural and measurement

model, to calculate weights required to give final estimates for each latent variable (domain). The second and third stages involves the non-iterative estimation of the coefficients of the structural (path coefficients) and measurement model (loading coefficients), respectively.

As the reflective set of indicators reflects the (unique) domain, it should be homogeneous and unidimensional, i.e., a set of indicators describing a domain are assumed to measure the same unique underlying concept. There exist several tools to check the homogeneity and unidimensionality, including Cronbach's alpha, Dillon-Goldstein's rho, and principal component analysis of a set. A set of indicators is considered homogenous if the Cronbach's alpha or the Dillon-Goldstein's rho are larger than 0.7, while a block is considered unidimensional if the first eigenvalue is higher than 1 and the others are smaller [6]. After assessing the quality of the measurement model, we evaluated the quality of the structural model through the R^2 determination coefficients for each endogenous domain. Instead, for each regression in the structural model we have an R^2 that is interpreted as the amount of variance in the endogenous domain explained by its exogenous domains. In the literature, the explanation of the variance is described as substantial , moderate or weak with reference to thresholds of R^2 equal to 0.67 , 0.33 to 0.19 , respectively .

In the PLS-PM approach, no criterion exist to evaluate the goodness of fit of the global model, so we cannot perform inferentially statistical tests for goodness of fit. Nevertheless, a descriptive index of goodness of fit in both the structural and the measurement models has been proposed, i.e., the GoF index. This index ranges from 0 to 1, with a value close to 1 indicating a good fit [22,6].

The indicators were performed through SAS 9.2, while PLS-PM analysis was conducted using the `plspm` package in R software [23].

3.3 Results

The data set consisted of 18 process indicators measuring adherence to international guidelines for the diagnosis and treatment of breast cancer, 3 indicators for complications, and 3 indicators for patients characteristics. For the PLS-PM analysis, each process indicator was measured as the percentage of patient that did not perform the appropriate diagnostic or therapeutic process; for example, the indicator estimating the percentage of mammographies performed within six months prior to diagnosis − which would be a positive tool for a early diagnosis − is considered as the percentage of patients who did not perform a mammography within six months prior to diagnosis.

We supposed a process characterized by six different domains, plus one domain for patients characteristics and one for complications. A detailed description of the model, domains and indicators is provided in Table 3.1.

The domain called Disadvantage included patients' age and comorbidities, characteristics related to a lower probability of adherence to diagnostic and therapeutic guidelines on primary breast cancer. Six other domains measured lack of adherence to guidelines for primary non metastatic breast-cancer in different phases of the diagnostic and therapeutic pathway: Diagnosis, including indicators measuring lack of adherence to diagnostic guidelines; Timing, including an indicator measuring the proportion of patients with a long interval between diagnosis and breast surgery

and one between surgery and the beginning of medical treatment; Surgical treatment 1, measuring lack of adherence to general surgical guidelines; Surgical treatment 2, measuring lack of use of minimal invasive surgical techniques in stage I patients; Medical treatment, including indicators measuring lack of adherence to medical treatment guidelines; Follow-up, including indicators measuring lack of adherence to guidelines for short-time follow-up. The last domain, Complications, included indicators measuring the proportion of patients experiencing side effects from medical treatment.

We supposed that Disadvantage was connected to all the other domains, except for Surgical treatments 2; Diagnosis was joined to Timing, Surgical and Medical treatments, and Follow-up; Timing was connected to Surgical treatment 1; Surgical treatment 1 was related to Medical treatment and Follow-up; Medical treatment was connected to Complications and Follow-up (Figure 3.1 and 3.2).

The final model selection was based on the clinical meaning and the evaluation of internal consistence that is a high positive correlation between indicators of the same domain (Table 3.2) and quality measures of the measurement models, including Cronbach's alpha, Dillon-Goldstein's rho, and the difference between the two principal components (Table 3.3).

Figure 3.1 and Figure 3.2 show the structural model based on the eight domains as described above. Each domain is inside of an ellipse; the relationships between domains, defined a priori, are represented by arrows. The statistics given in Figure 3.1 and Figure 3.2 represent the values of the path or regression coefficients and the correlation coefficients linking the domains, respectively.

A low adherence to the guidelines for the diagnostic procedures in breast cancer was negatively associated with low adherence to the guidelines of surgical treatment (β = -0.55) and to the timing (β =-0.62) and highly positively associated with low use of minimal invasive surgery (surgical treatment 2, β=0.97). A higher proportion of elderlies and patients with comorbidities was negatively associated to low adherence to the guidelines of medical treatment (β = -0.70) and positively associated to low adherence to the guidelines for the diagnostic procedures (β = 0.62) and surgical treatment (β = 0.69). Moreover, the timing was positively associated with low adherence to the guidelines of surgical treatment (β = 0.64), and a low adherence to the guidelines of surgical treatment was negatively associated with the follow-up (β = -0.87) (Figure 3.1).

In order to take into account all the paths (direct and indirect), Table 3.4 shows the direct, indirect, and total effects. The indirect effects are obtained as the product of the path coefficients by taking an indirect path. For instance, consider the impact of Disadvantage on Surgical treatment2, even though these two domains are not directly related, there is an indirect path from Disadvantage to Surgical treatment2 that goes through Diagnosis. If you multiply the path coefficient of Disadvantage on Diagnosis (0.619) with the path coefficient of Diagnosis on Surgical treatment2 (0.967), you get the indirect effect of Disadvantage on Surgical treatment2: 0.598 = 0.619 x 0.967.

Considering correlation coefficients > 0.5, meaning that the higher the lack of adherence in one domain the higher the lack of adherence in the other, was found between: Diagnosis and Surgical treatment 2 (r= 0.97), Timing and Surgical

treatment 1 (r= 0.67), Disadvantage and Diagnosis (r= 0.62). A negative correlation was found between Disadvantage and Medical Treatment (r =-0.88), Diagnosis and Medical Treatment (r =-0.82), Diagnosis and Timing (r =-0.71), Diagnosis and Surgical treatment 1 (r= -0.59), Disadvantage and Timing (r =-0.53) (Figure 3.2, Table 3.5).

Concerning the structural model, the quality was moderate for the Timing (R^2 = 0.52), Diagnosis (R^2 = 0.38), and Complications (R^2 = 0.36) domains, and substantial for all the other domains (R^2 > 0.68). Furthermore, the quality of the global model was discrete when evaluated in terms of the goodness of fit (GoF = 0.65).

Table 3.6 show the values of the scores of the domains by local health authority. Figure 3.3 is a representation of the values of the scores of the domains by local health authority. Figure 3.4 is a representation of the six local health authorities according to the values of the Diagnosis and Complications domains. Figure 3.5 summarize the scores of all the domains for each local health authority.

Table 3.1. Specification of the measurement model in Partial Least Square Path Modeling (PLS-PM) approach: description of domains and indicators.

Domain	Description of the domain	Indicator	Description of the indicator
Disadvantage	Characteristics related to a lower probability of adherence to diagnostic and therapeutic guidelines on primary breast cancer	Comorbidity	Proportion of patients with chronic cardiovascular disease and/or diabetes at diagnosis
		Advanced age	Proportion of patients ≥ 75 years
		Deprivation	Proportion of patients in the lower two quintiles of deprivation index
Diagnosis	Indicators measuring lack of adherence to diagnostic guidelines on primary breast cancer	D1	Proportion of women aged over 50 who did not receive bilateral mammography 3 months before surgery
		D2	Proportion of women aged 50-69 years who did not have a screening mammography performed in the 3 months preceding diagnosis performed in the 6 months preceding diagnosis
		D3	Proportion of breast cancer women without cytological and/or histological assessment in the 3 months prior surgery
		D4	Proportion of patients in stage I, and not undergoing mastectomy , undergoing bone scanning or thoracic CT or liver US or abdominal CT /MR or tumour markers measurement in the 3 months prior to surgery
Timing	Indicators measuring the proportion of patients with a long interval between diagnosis and breast surgery and between surgery and the beginning of medical treatment	T1	Proportion of patients whose first postoperative treatment was not initiated within 60 days of surgery in the event of chemotherapy and within 90 days in the event of radiotherapy

Domain	Description of the domain	Indicator	Description of the indicator
		T2	Proportion of patients undergoing surgery more than 30 days from mammography.
Surgical treatment1	Indicators measuring lack of adherence to general surgical guidelines on primary breast cancer	S1a	Proportion of stage I and II women who did not undergo breast-conserving surgery
		S2a	Proportion of patients developing lymphedema within two years from breast surgery
		S3a	Proportion of patients undergoing a second surgery within 3 months from the first breast conserving surgery, excluding reconstructions
		S4a	Proportion of patients not undergoing reconstructive surgery within a year among patients who underwent mastectomy
Surgical treatment2	Indicators measuring lack of use of minimal invasive surgical techniques in stage I patients	S1b	Proportion of patients not undergoing SLNB in the setting of breast conserving surgery for T1 tumors
		S2b	Proportion of patients with pathological stage I breast cancer undergoing axillary clearance at first surgery or within 3 months
Medical treatment	Indicators measuring lack of adherence to medical treatment guidelines on primary breast cancer	M1	Proportion of patients with stage III tumors not undergoing neoadjuvant systemic theraphy (either hormonal or chemo)
		M2	Proportion of women who did not receive radiation treatment within a year after breast conserving surgery
		M3	Proportion of breast cancer women > 50 years with pathological stage II-III not receiving adjuvant hormone theraphy or chemotheraphy in the following year
		M4	Proportion of breast cancer women < 50 years with pathological stage II-III not receiving adjuvant

Domain	Description of the domain	Indicator	Description of the indicator
			chemotheraphy in the following year
Complications	Indicators measuring surgical complications and side effects from medical treatment in patients with primary breast cancer	C1	Proportion of patients experiencing side effects requiring hospitalization during chemotherapy
		C2	Proportion of patients experiencing hematological side effects requiring hospitalization during chemotherapy
		C3	Proportion of patients experiencing cardio-vascular side effects requiring hospitalization during chemotherapy
Follow-up	Indicators measuring lack of adherence to guidelines for follow-up after primary treatment of breast cancer	F1	Proportion of patients > 50 years not undergoing mammography within 18 months after surgery
		F2	Proportion of patients not enrolled in palliative care within 6 months of death within 3 months of death

Table 3.2. Correlations between indicators and domains.

	Disadvantage	Diagnosis	Timing	Surg.treat2	Surg.treat1	Medical_treat	Follow-up	Complications
Disadvantage								
Comorbidity	0.9323	0.5874	-0.3516	0.6765	0.170	-0.79067	-0.0867	0.3194
advanced age	0.0808	0.1610	0.1991	0.1096	0.215	0.00463	-0.1613	0.6704
Deprivation	0.9244	0.5320	-0.6833	0.5661	-0.199	-0.86479	0.4464	0.1205
Diagnosis								
D1	-0.3330	0.3992	-0.2120	0.3635	-0.839	0.08100	0.4718	0.1004
D2	0.2680	0.5960	-0.5051	0.5946	-0.385	-0.51164	0.2061	0.1461
D3	-0.7062	0.0759	0.2418	0.0307	-0.396	0.47577	-0.0321	0.1324
D4	0.5877	0.7617	-0.4029	0.7163	-0.227	-0.62661	0.2597	0.8238
Timing								
T1	-0.6210	-0.7629	0.9697	-0.6582	0.676	0.86755	-0.8712	-0.4422
T2	-0.0379	-0.2655	0.6750	-0.1210	0.378	0.40747	-0.5399	-0.1239
Surg.treat2								
S1b	0.4639	0.8079	-0.3235	0.9146	-0.525	-0.50424	0.1436	0.2088
S2b	0.7618	0.9740	-0.7271	0.9421	-0.438	-0.91122	0.4342	0.6904
Surg.treat1								
S1a	0.1290	-0.5983	0.5685	-0.4784	0.924	0.24033	-0.7466	-0.3706
S2a	0.4906	0.1502	0.0571	0.1359	0.426	-0.28156	-0.1670	0.5748
S3a	0.0250	-0.3651	0.4970	-0.3417	0.912	0.17005	-0.7612	0.1345
S4a	-0.1735	-0.6644	0.7856	-0.6081	0.942	0.51094	-0.8010	-0.0772
Medical_treat								
M1	-0.7561	-0.7412	0.4382	-0.7663	0.276	0.69607	-0.2997	-0.5524
M2	-0.4700	-0.6000	0.7404	-0.6136	0.593	0.67058	-0.5359	0.0672
M3	-0.5886	-0.2372	0.7472	-0.1589	0.123	0.70154	-0.5431	-0.0529
M4	-0.6998	-0.6718	0.5930	-0.5717	-0.038	0.82442	-0.2196	-0.8175
Follow-up								
F1	0.1189	0.5043	-0.8590	0.3492	-0.870	-0.48324	0.9931	0.2441
F2	0.2480	0.4434	-0.8887	0.2975	-0.766	-0.55403	0.9934	0.1870
Complications								
C1	0.2730	0.2645	-0.0306	0.1570	0.343	-0.26759	-0.1182	0.8621
C2	0.2854	0.5405	-0.2973	0.4329	0.112	-0.47923	-0.0504	0.8186
C3	0.2584	0.7186	-0.5792	0.5813	-0.580	-0.50206	0.6162	0.7860

Table 3.3. Unidimensionality indices

	#MVs	C.alpha	DG.rho	eig.1st	eig.2nd
Disadvantage	3	0.444	0.704	1.78	1.004
Diagnosis	4	0.497	0.715	1.82	1.046
Timing	2	0.644	0.849	1.47	0.526
Surgical trt 2	2	0.841	0.927	1.73	0.274
Surgical trt 1	4	0.842	0.899	2.79	0.854
Medical trt	4	0.705	0.820	2.15	1.086
Followup	2	0.986	0.993	1.97	0.027
Complications	3	0.763	0.865	2.04	0.655

Figure 3.1. Structural model. The path coefficients represent the direct effects between the domains performed according to the Partial Least Square Path Modeling (PLS-PM) approach.

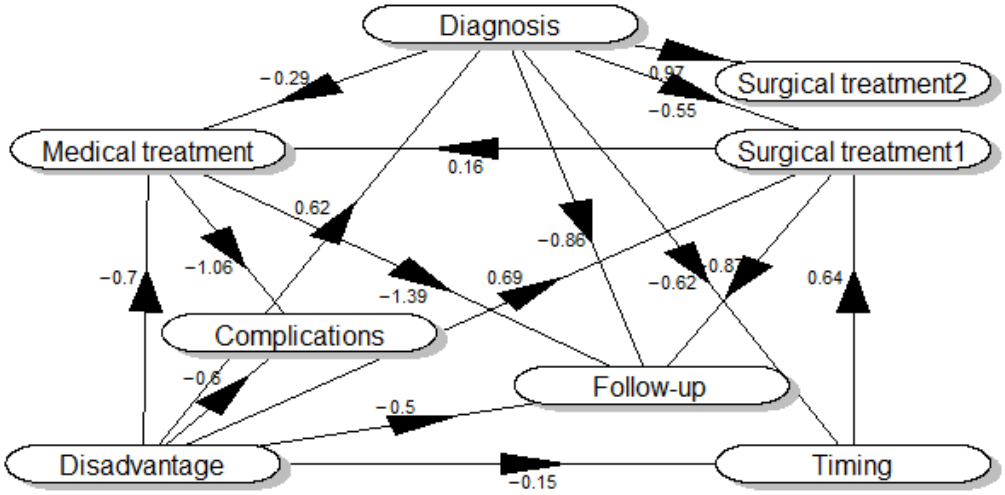


Table 3.4 Direct, indirect, and total effect of the relationship between domains.

	relationships	direct	indirect	total
1	Disadvantage -> Diagnosis	0.619	0.000	0.61879
2	Disadvantage -> Timing	-0.146	-0.386	-0.53105
3	Disadvantage -> surg.trt2	0.000	0.598	0.59817
4	Disadvantage -> surg.trt1	0.689	-0.684	0.00468
5	Disadvantage -> medical_trt	-0.701	-0.181	-0.88243
6	Disadvantage -> followup	-0.504	0.690	0.18531
7	Disadvantage ->complications	-0.603	0.936	0.33282
8	Diagnosis -> Timing	-0.623	0.000	-0.62303
9	Diagnosis -> surg.trt2	0.967	0.000	0.96667
10	Diagnosis -> surg.trt1	-0.554	-0.400	-0.95395
11	Diagnosis -> medical_trt	-0.294	-0.155	-0.44967
12	Diagnosis -> followup	-0.864	1.451	0.58679
13	Diagnosis -> complications	0.000	0.477	0.47698
14	Timing -> surg.trt2	0.000	0.000	0.00000
15	Timing -> surg.trt1	0.643	0.000	0.64272
16	Timing -> medical_trt	0.000	0.105	0.10463
17	Timing -> followup	0.000	-0.702	-0.70158
18	Timing -> complications	0.000	-0.111	-0.11099
19	surg.trt2-> surg.trt1	0.000	0.000	0.00000
20	surg.trt2-> medical_trt	0.000	0.000	0.00000
21	surg.trt2-> followup	0.000	0.000	0.00000
22	surg.trt2-> complications	0.000	0.000	0.00000
23	surg.trt1-> medical_trt	0.163	0.000	0.16280
24	surg.trt1-> followup	-0.865	-0.227	-1.09158
25	surg.trt1 ->complications	0.000	-0.173	-0.17268
26	medical_trt -> followup	-1.392	0.000	-1.39219
27	medical_trt->complications	-1.061	0.000	-1.06072
28	followup -> complications	0.000	0.000	0.00000

Figure 3.2. Structural model. The coefficients represent the correlations between the domains performed according to the Partial Least Square Path Modeling (PLS-PM) approach.

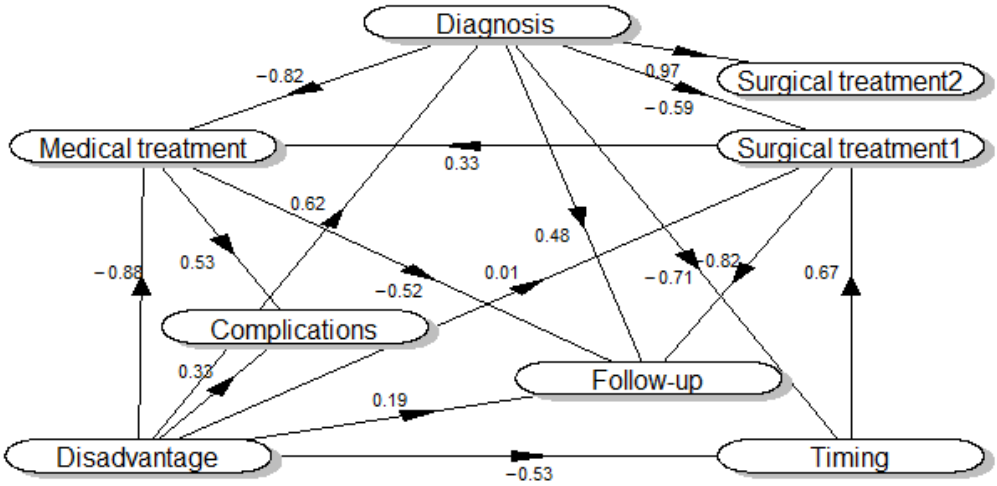


Table 3.5. Correlation between domains

	Disadvantage	Diagnosis	Timing	Surg.trt2	Surg.trt1	Medical_trt	Followup	Complications
Disadvantage	1.0000	0.619	-0.531	0.674	0.0047	-0.882	0.185	0.333
Diagnosis	0.6188	1.000	-0.713	0.967	-0.5858	-0.824	0.477	0.653
Timing	-0.5311	-0.713	1.000	-0.585	0.6718	0.840	-0.880	-0.405
Surg.trt2	0.6742	0.967	-0.585	1.000	-0.5134	-0.782	0.325	0.508
Surg.trt1	0.0047	-0.586	0.672	-0.513	1.0000	0.332	-0.823	-0.101
Medical_trt	-0.8824	-0.824	0.840	-0.782	0.3320	1.000	-0.522	-0.528
Followup	0.1853	0.477	-0.880	0.325	-0.8231	-0.522	1.000	0.217
Complications	0.3328	0.653	-0.405	0.508	-0.1009	-0.528	0.217	1.000

Table 3.6. Scores of the domains by local health authorities

LHA	Disadvantage	Diagnosis	Timing	Surg.trt2	Surg.trt1	Medical_trt	Followup	Complications
1	-0.38	-1.53	1.89	-1.31	2.13	1.21	-1.98	-0.48
2	1.39	1.52	-0.78	1.48	-0.44	-1.31	0.55	1.58
3	0.63	0.90	-0.80	0.81	0.08	-1.05	-0.07	0.68
4	-1.84	-0.38	0.25	-0.74	-0.72	1.16	0.46	0.41
5	0.37	-0.61	-0.97	-0.77	-0.24	-0.43	1.24	-0.75
6	-0.17	0.10	0.40	0.53	-0.82	0.42	-0.19	-1.44

Figure 3.3. Representation of the six local health authorities according to the values of the domains. LHA: local health authority.

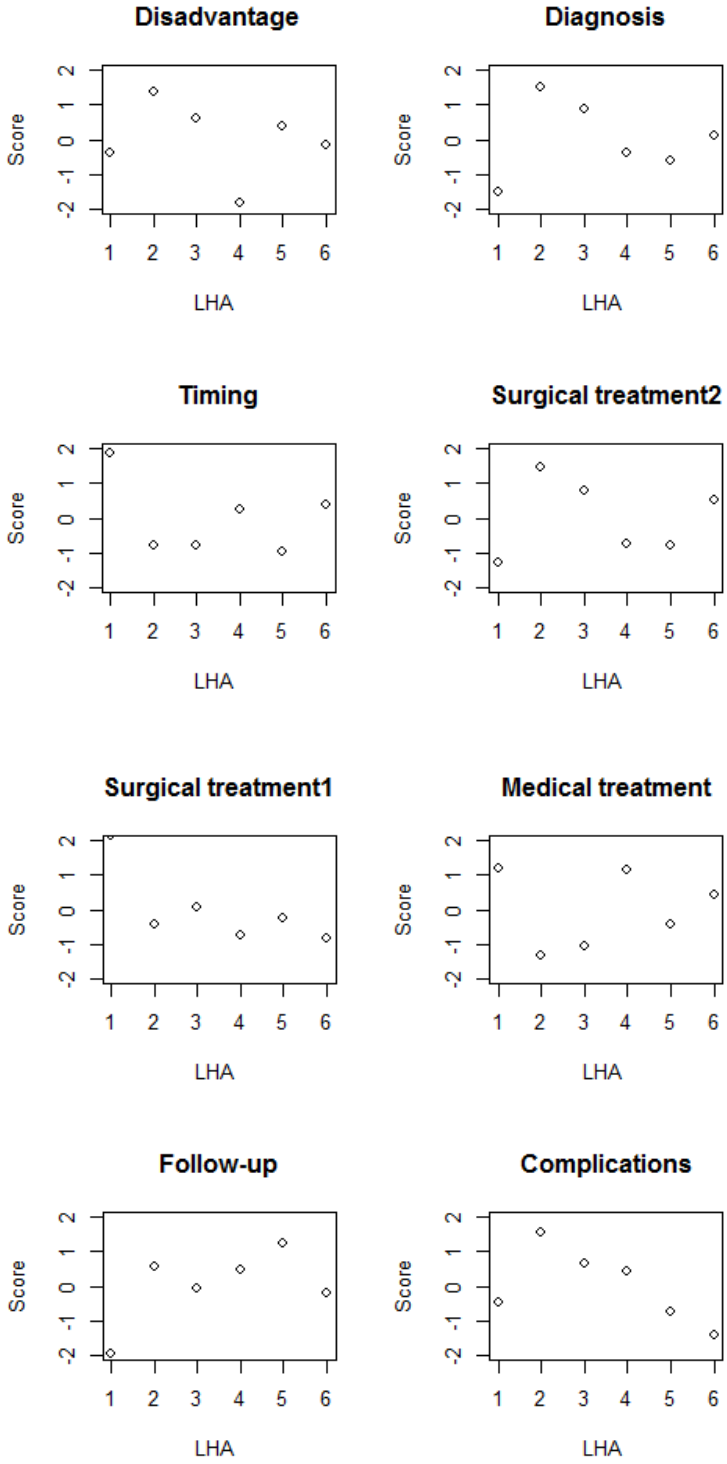


Figure 3.4. Representation of the six local health authorities according to the values of the Diagnosis and Complications domains.

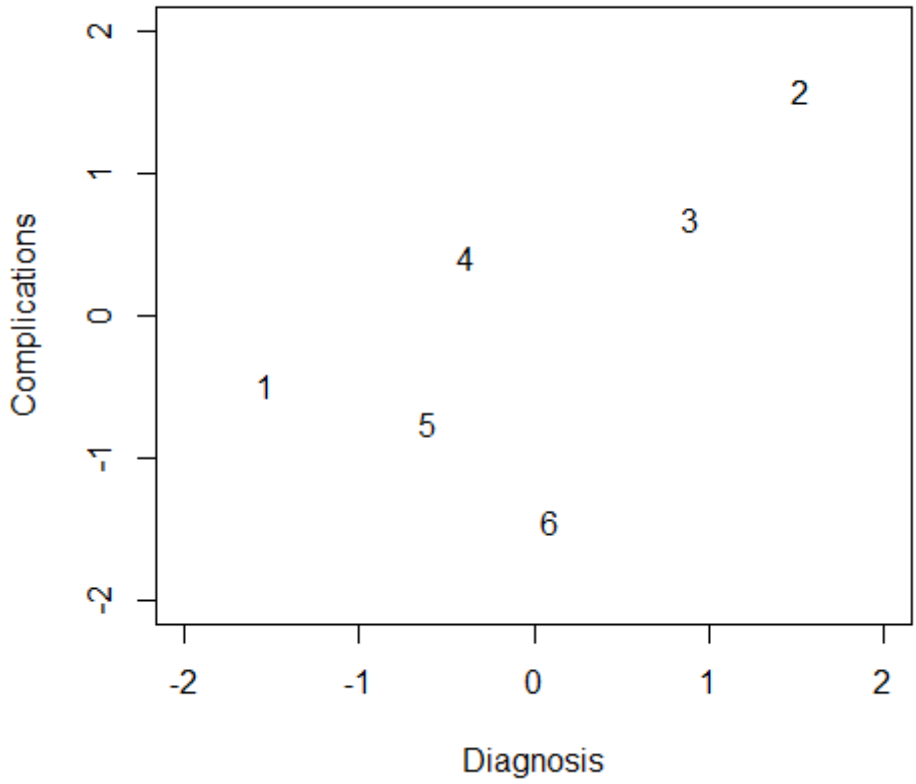
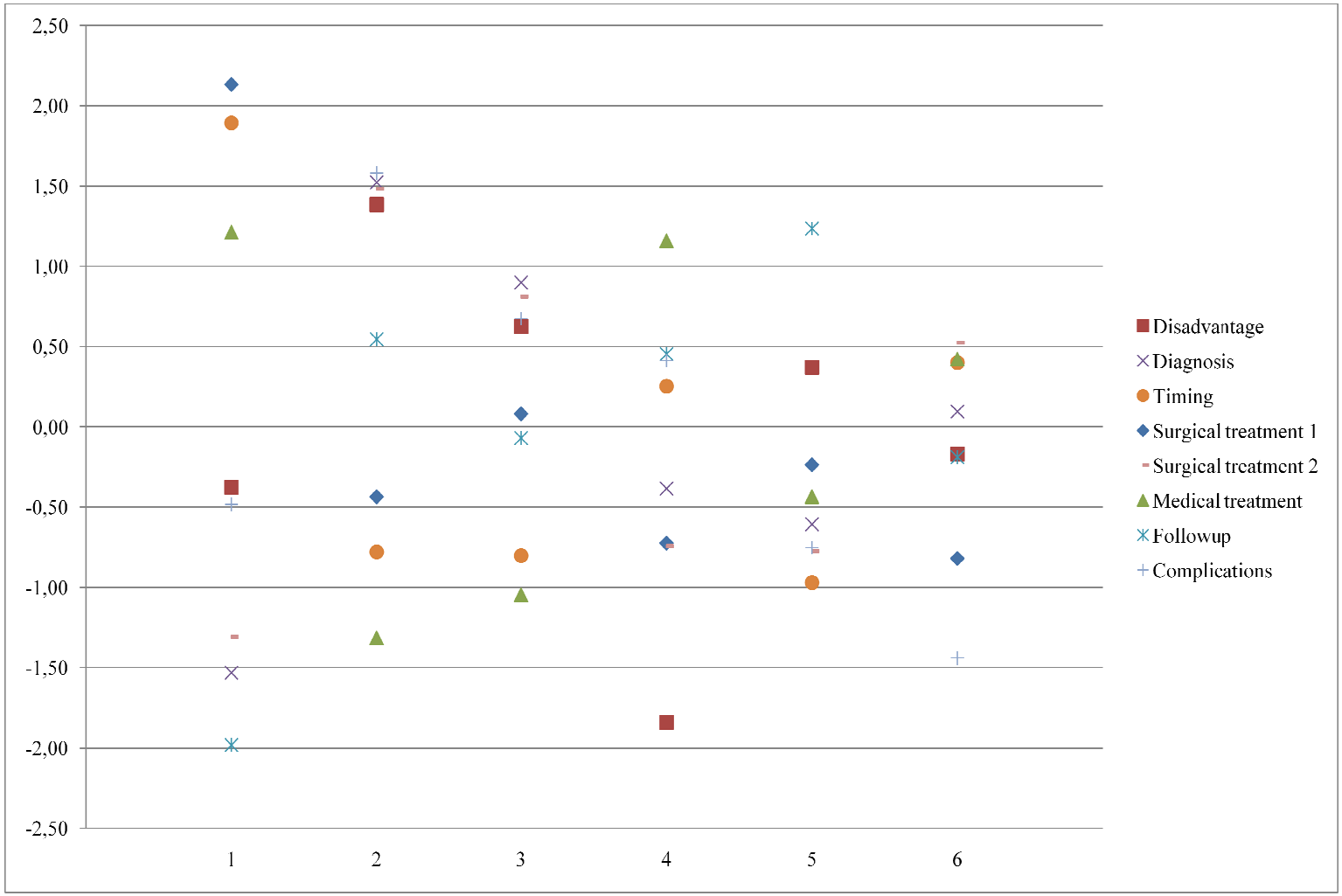


Figure 3.5. Score of the latent variables (domain) by local health authority.



4. Conclusion

The core of the PLS algorithm is the calculation of the weights required to estimate the latent variables. The weights are obtained based on how the structural and the measurement model are specified. This is done by means of an iterative procedure in which two kinds of approximation for the latent variables are alternated until convergence of weight estimates. These two types of approximation, called the inside approximation and the outside approximation, have to do with the inner relations and the outer relations, respectively. The algorithm begins with arbitrary initial weights used to calculate an outside approximation of the latent variables, that is, initial weights are given in order to approximate the latent variables as linear combinations of their manifest variables. Then, the inner relations among latent variables are considered in order to calculate the inside approximations, having the option of choosing between three possible scenarios, called weighting schemes, to perform this approximation: (1) centroid, (2) factor, and (3) path scheme. Once the inside approximations are obtained, the algorithm turns around to the outer relations when new weights are calculated considering how the indicators are related to their constructs: by mode A (reflective), or by mode B (formative). Mode A implies simple linear regressions while mode B implies multiple linear regressions. The

simple/multiple regressions coefficients are then used as new weights for an outside approximation. The process continues iteratively until convergence of the weights is reached. After convergence of the weights, and once the latent variables are estimated, the parameters of the structural and the measurement models can be obtained. The structural coefficients, also known as path coefficients, are calculated by ordinary least square regressions between latent variables. There are as many regressions as endogenous latent variables. The parameters of the measurement model, the loading coefficients, are also estimated by least square regressions but taking into account the kind of mode to be used (reflective or formative). PLS-PM is a more exploratory way of performing structural equation modeling than the popular LISREL approach. The latter approach resorts to classical theory of statistical inference and is based on a heavy use of distributional assumptions about the behavior and personality of the data. LISREL also requires larges samples, in contrast, PLS-PM uses ordinary least squares, which does not make distributional assumptions and can model skewed and ordinal data. Hence, it is more suitable for research with small samples, and non-normal distributions.

I used the PLS-PM method in order to analyze the adherence of the procedures provided for diagnosis, treatment (surgical and medical), and follow-up of breast cancer through a set of indicators. This method has several strengths, as PLS-PM allows the reduction of dimensionality of several health indicators into a smaller number of latent variables (and more interpretable) first, and then allows to study causal relationships between these latent variables. This method also requires no distributional assumptions with respect to the variables included in the model. The

limit of this method is the bias deriving from the selection of the indicators used to characterize the latent variable. Moreover, by using a dataset with 6 observations (6 local health authorities) the sample size is too small to make some inference.

The aim of this study was to apply the PLS-PM in a different field, since it has been widely used in economical (the customer satisfaction is a typical example) and psychological setting. In biomedical context, the published articles are scanty and generally published in open access journals [24-26]. For example, Xue et al. [24] introduced PLSPM to analyze the association between single or multiple SNPs and obesity in the European Prospective Investigation of Cancer (EPIC)-Norfolk study; Vitalino et al. [25] analyzed a theoretical stress model that examined whether relationships of chronic stress, psychophysiology, and coronary heart disease varied between the sex and among users or not users of hormone replacement therapy (among women). Moreover, in the paper by Fischer [26], empirical approaches that applied PLS-PM to decision-making in healthcare were summarized through a systematic literature search. PLS-PM was used as an estimation technique for a structural equation model that specified hypotheses between the components of decision processes and thereasonableness of decision-making. The model was estimated for a sample of 55 coverage decisions on the extension of newborn screening programs in Europe. However, he focused on the economical aspects of the screening programs. Thus, the present study represents an unique example of PLS-PM application in the evaluation of the adherence of the procedures provided for diagnosis, treatment (surgical and medical), and follow-up of breast cancer through a set of health indicators, and to investigate the difference between various

health structures, such as the local health authority, although the limited sample size makes the analyses only explorative-orientated.

References

1. Fornell C (1982) A Second Generation of Multivariate Analysis: Methods, vol I. Praeger Publishers, New York.

2. Diamantopoulos A, Winklhofer H (2001) Index Construction with Formative Indicators: AnAlternative to scale development. Journal of Marketing Research 38 (2):269-278.

3. Sanchez Trujillo G (2009) Doctoral dissertation. Pathmox approach: segmentation trees in partial least squares path modeling. Barcelona

4. De Beuckelaer A (2005) On the nature of constructs and their use in comparative research. In: Aluja T, Casanovas J, Esposito V, Morineau A, Tenenhaus M (eds) Proceedings of the PLS'05 International Symposium. SPAD Test&go, Paris, pp 117-124.

5. Sanchez G (2013) PLS Path Modeling with R. URL http://www.gastonsanchez.com/PLS Path Modeling with R.pdf. California.

6. Vinzi VE, Chin WW, Henseler J, Wang H (2010) Handbook of Partial Least Squares: Concepts, Methods and applications. Springer.

7. Tenenhaus M, Vinzi VE, Chatelin YM, Lauro C (2005) PLS path modeling. Computational Statistics and Data Analysis 48:159-205.

8. Wold H (1982) Systems under Indirect Observation: Causality, Structure, Prediction. Jöreskog, K. G. and Wold, H, North-Holland.

9. Tenenhaus M, Pagès J (2002) Analyse Factorielle Multiple et Approche PLS. Revue de Statistique Appliquée 50 (1):5-33.

10. Russolillo G (2012) Non-Metric Partial Least Squares. Electronic Journal of Statistics 6 1641–1669.

11. Kolfschoten NE, Gooiker GA, Bastiaannet E, van Leersum NJ, van de Velde CJ, Eddes EH, Marang-van de Mheen PJ, Kievit J, van der Harst E, Wiggers T, Wouters MW, Tollenaar RA (2012) Combining process indicators to evaluate quality of care for surgical patients with colorectal cancer: are scores consistent with short-term outcome? BMJ Qual Saf 21 (6):481-489.

12. Jacobs R, Goddard M, Smith PC (2005) How robust are hospital ranks based on composite performance measures? Med Care 43 (12):1177-1184.

13. Smith P (2002) Developing composite indicators for assessing health system efficiency. In: Measuring Up Improving Health System Performance in OECD Countries: Improving Health System Performance in OECD Countries
pp 295-313.

14. Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO (2007) Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. Med Care 45 (6):489-496.

15. Traberg A, Jacobsen P, Duthiers NM (2014) Advancing the use of performance evaluation in health care. J Health Organ Manag 28 (3):422-436.

16. Nolan T, Berwick DM (2006) All-or-none measurement raises the bar on performance. JAMA 295 (10):1168-1170.

17. Chen LM, Staiger DO, Birkmeyer JD, Ryan AM, Zhang W, Dimick JB (2013) Composite quality measures for common inpatient medical conditions. Med Care 51 (9):832-837.

18. Grunkemeier GL, Wu Y (2007) What are the odds? Ann Thorac Surg 83 (4):1240-1244.

19. Shwartz M, Ren J, Pekoz EA, Wang X, Cohen AB, Restuccia JD (2008) Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. Med Care 46 (8):778-785.

20. Russo A, Andreano A, Anghinoni E, Autelitano M, Bellini A, Bersani M, Bizzoco S, Cavalieri d'Oro L, Decarli A, Lucchi S, Mannino S, Panciroli E, Rognoni M, Sampietro G, Valsecchi MG, Villa M, Zocchetti C, Zucchi A (2014) [A set of indicators to monitor the adherence to the guidelines for the diagnosis and treatment of breast cancer]. Epidemiol Prev 38 (1):16-28.

21. Wold H (1982) Soft modeling: the basic design and some extensions. In: Jöreskog KG, Wold H (eds) Systems under Indirect Observation: Causality, Structure, Prediction, vol II. North-Holland: Amsterdam.

22. Tenenhaus M, Amato S, Vinzi EV (2004) A global goodness-of-fit index for PLS structural equation modelling. Proceedings of the XLII SIS Scientific Meeting. CLEUP, Padova:739–742.

23. Sanchez G, Trinchera LR plspm: PArtial Least Squares data analysis methods. R package version 0.4.1 (http://cran.r-project.org/web/packages/plspm/index.html). Accessed 20 December 2013.

24. Xue F, Li S, Luan J, Yuan Z, Luben RN, Khaw KT, Wareham NJ, Loos RJ, Zhao JH (2012) A latent variable partial least squares path modeling approach to regional association and polygenic effect with applications to a human obesity study. PLoS One 7 (2):e31927.

25. Vitaliano PP, Scanlan JM, Zhang J, Savage MV, Hirsch IB, Siegler IC (2002) A path model of chronic stress, the metabolic syndrome, and coronary heart disease. Psychosom Med 64 (3):418-435.

26. Fischer KE (2012) Decision-making in healthcare: a practical application of partial least square path modelling to coverage of newborn screening programmes. BMC Med Inform Decis Mak 12:83.