Component-based Predictive and Exploratory Path Modeling and Multi-block Data Analysis

Laura Trinchera[1], Vincenzo Esposito Vinzi[2,3]
[1]Rouen Business School, Mont Saint-Aignan, FRANCE
[2]ESSEC Business School of Paris, Cergy, FRANCE
[3]Corresponding author: V.E. Vinzi, e-mail: *vinzi@essec.edu*

Abstract

This discussion paper will focus on the predictive modeling of relationships between latent variables in a multi-block data framework. We will refer to component-based methods such as Partial Least Squares Path Modelling, Generalized Structured Component Analysis as well as to some of their recent variants and other alternatives. We will compare these approaches by paying particular attention to the statistical criteria optimized by each of them. To conclude we will present some interesting developments needed to cope with new challenges (e.g. big data, regularization, feature and variable selection, multidimensionality of latent variables and so on) raised by the complex data structures available nowadays.

1. Introduction

Several statistical methodologies have been developed to analyze the relations among several blocks of variables observed on the same statistical units by summarizing them with a few number of unobserved variables. However, all methods based on Canonical Correlation do not allow to take into account a specific pattern of directed relations among the observed blocks of variables. In 1970 Karl Jöreskog first proposed to use Structural Equation Models (SEMs) to estimate the causal relationships, defined according to a theoretical model, linking two or more latent complex concepts, each measured through a number of observable indicators (Jöreskog, 1970).

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

The main difference between Jöreskog's approach and the Wold's one lies in the definition (actually, even the conceptual meaning) of the unobserved variables included in the model. The basic idea behind a SEM is that the complexity inside a system can be studied taking into account a network of causal relationships among unobserved variables, called latent variables (LV), each measured by several observed indicators usually defined as manifest variables (MV). PLS-PM, instead, assumes that each block of manifest variables can be summarized by an observed variable defined as a component or a composite. This "slight" difference in the definition of the unobserved variables included in the model leads to major differences in terms of aims of the analysis. As a matter of fact, SEMs are confirmatory models that aim to validate a researcher's hypotheses on the relations between the observed manifest variable, while the PLS-PM is more an exploratory and predictive approach than a confirmatory one, though it is often used to validate theories in a few applied disciplines.

Several authors have compared the two approaches over the years; see, for example, Jöreskog & Wold (1982), Fornell & Bookstein (1982). The two approaches differ in the objectives of the analysis, the statistical assumptions, the estimation procedures and the related outputs.

In this paper we discuss the PLS-PM and other component-based approaches to the analysis of relations among several blocks of variables, such as the Generalized Structured Component Analysis (GSCA) by Hwang and Takane (2004) and the Regularized Generalized Canonical Correlation Analysis (RGCCA) by Tenenhaus and Tenenhaus (2011). We will compare these approaches by paying particular attention to the statistical criteria optimized by each of them. To conclude we will present some recent developments and open issues in the component-based approaches to structured multi-block analysis.

2. Component-based approaches for multi-block data

Partial Least Squares Path Modeling (Tenenhaus *et al.*, 2005; Esposito Vinzi *et al.*, 2010 for an overview with recent developments) is so far the most popular component-based alternative (Tenenhaus, 2008) to the classical SEM.

PLS Path Modeling aims at studying the relationships among Q blocks $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q$ of manifest variables (MVs), which are expression of Q unobservable constructs $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_q, \ldots, \boldsymbol{\xi}_Q$, that are usually called latent variables (LVs) but that are indeed block components. In PLS Path Modeling an iterative procedure allows us to define a system of weights, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q, \ldots, \boldsymbol{w}_Q$, to be associated to each block of manifest variables in order to obtain each $\boldsymbol{\xi}_q$ as a linear combination of the original variables, i.e. as:

$$\boldsymbol{\xi}_q = \sum_{j=1}^{J_q} w_{jq} \boldsymbol{x}_{jq} \tag{1}$$

where $\boldsymbol{x}_{jq}$ ($j = 1, \ldots, J_q$; $q = 1, \ldots, Q$) is the generic centered and properly scaled manifest variable of the q-th block and J_q is the number of MVs in the same block.

The iterative algorithm works by alternating inner and outer estimates of the LVs. In particular, in the outer estimation step each LV is obtained as a standardized weighted aggregate ($\mathbf{v}_q$) of its own block of manifest variables, i.e. $\mathbf{v}_q \propto \sum_h w_{jq} \boldsymbol{x}_{jq} = \mathbf{X}_q \boldsymbol{w}_q$. Then, in the inner estimation step each LV is obtained as a standardized weighted aggregate ($\mathbf{z}_q$) of the adjacent LVs, i.e. $\mathbf{z}_q \propto \sum_{\mathbf{v}_{q'} \to \mathbf{v}_q} e_{q'q} \mathbf{v}_{q'}$, where the connections between LVs are defined by the user in the path diagram structure. For both the outer weights, w_{jq}, and the inner weights, $e_{q'q}$, several options are available. For more details on the outer estimation modes and the inner estimation schemes in PLS-PM refer to Esposito Vinzi *et al.* (2010).

These two steps are iterated till numerical convergence on the outer weights ($\boldsymbol{w}_q$). This convergence is proven in case of two blocks (Lyttkens *et al.*, 1975). Empirical convergence is observed in most of the real applications. However, in 2010 Henseler showed a few examples of non-convergence of the PLS-PM algorithm (Henseler, 2010). According to Vinzi and Russolillo (2013), non convergence seems to be due to model misspecification rather than numerical pitfalls of the algorithm.

Indeed, there is not an overall scalar function optimized by PLS-PM. This is mainly due to the different available options in the inner and outer estimation steps, but also to the fact that PLS Path Models may differ in number of LVs and in the path of relationships linking them. Many researchers have paid attention to this issue in the last years. Nowadays, the stationary equations for most of the specific models obtained by running a PLS-PM are known and it is possible to show that the PLS-PM generalizes many Multivariate Analysis techniques.
Glang (1988) and Mathes (1993) were among the first who paid attention to the optimization criteria behind the PLS-PM. In particular, they showed that the Lagrange equations associated with the optimization of the criterion

$$\sum_{q \neq q'} c_{qq'} g(\mathrm{cor}(\boldsymbol{X}_q \boldsymbol{w}_q, \boldsymbol{X}_{q'} \boldsymbol{w}_{q'})) \tag{2}$$

subject to $\|\boldsymbol{X}_q \boldsymbol{w}_q\| = 1$, give exactly the stationary equation of PLS-PM algorithm when the weights in all the blocks in the outer estimation step are estimated by means of multiple regressions of $\boldsymbol{z}_q$ over its manifest variables $\boldsymbol{X}_q$ (the so-called *Mode B* outer estimation); $g(.)$ is the absolute value or the square function depending on the option used in the inner estimation step. More recently, Hanafi (2007) proved that the PLS-PM iterative procedure is monotonically convergent to these criteria.
In 2007 Krämer showed that Wold's PLS-PM algorithm with the outer estimation based, for each block, on simple linear regression of the variables in $\boldsymbol{X}_q$ on $\boldsymbol{z}_q$ (the so-called *Mode A* outer estimation), does not lead to a stationary equation related to the optimization of a twice differentiable function. Until now, PLS Path Models with *Mode A* applied to all the blocks do not optimize any known criterion. In 2011 Tenenhaus & Tenenhaus have extended the results of Hanafi to a modified *Mode A* in which the outer weights, rather than the latent variable scores, are normalized to unitary variance at each step of the algorithm. This new estimation mode has the major advantage, as compared to classical *Mode A*, to maximize a known criterion. In particular, they showed that Wold's procedure, applied to a PLS Path Model where the *new Mode A* is used in all the blocks for the outer estimation, monotonically converges to the criterion

$$\underset{\|\boldsymbol{w}_q\|=1}{\arg\max} \sum_{q \neq q'} c_{qq'} g(\mathrm{cov}(\boldsymbol{X}_q \boldsymbol{w}_q, \boldsymbol{X}_{q'} \boldsymbol{w}_{q'})) \tag{3}$$

where $g(.)$ is exactly the same as in equation (2).
By comparing equations (2) and (3) it is easy to notice that the criteria associated to *Mode B* are based on correlations while the ones associated to *New Mode A* are based on covariances. In the same paper Tenenhaus & Tenenhaus (2011) proposed a new method, the Regularized Generalized Canonical Component Analysis (RGCCA), where a continuum is built between the covariance criterion (*new Mode A*) and the correlation criterion (*Mode B*) by means of the tuning parameter $0 \leq \tau \leq 1$ (see equation 4). Indeed, Tenenhaus & Tenenhaus (2011) have proved that fixing the tuning parameter to zero (i.e. using standardized LV scores) leads to criteria based on maximizing correlations among adjacent LVs while fixing the tuning parameter to one (i.e. using outer weights with unitary variance) leads to criteria based on maximizing covariances among adjacent LVs.

$$\underset{\|\boldsymbol{w}_q\|=1}{\arg\max} \sum_{q \neq q'} c_{qq'} g\left(\left[\mathrm{cor}(\boldsymbol{X}_q \boldsymbol{w}_q, \boldsymbol{X}_{q'} \boldsymbol{w}_{q'}) \sqrt{\mathrm{var}(\boldsymbol{X}_{q'} \boldsymbol{w}_{q'})^{\tau_{q'}}} \sqrt{\mathrm{var}(\boldsymbol{X}_q \boldsymbol{w}_q)^{\tau_q}}\right]\right) \tag{4}$$

Equation (4) is very interesting from the theoretical point of view and the introduction of the *New Mode A* reduces the cases where the PLS-PM seems to be an heuristic approach at the case when the inner estimation takes explicitly into account the direction of the path model (the so-called path weighting scheme). However, it is not clear how users should interpret results obtained using a tuning parameter different from 0 or 1 that yields a method maximizing a mixture of correlations and covariances among adjacent LVs.

Starting from the above considerations, two new ways to compute the outer weights have been recently proposed by Esposito Vinzi & Russolillo (2011): the *PLScore Mode* and the *PLScow Mode*. These approaches use PLS Regression (PLS-R) (Wold et al., 1983; Tenenhaus, 1998) as a method to estimate the outer weights in the measurement model. In these approaches, we always deal with univariate PLS regressions (PLS1) where the dependent variable is the LV inner estimate $\mathbf{z}_q$ while its own MVs in $\mathbf{X}_q$ play the role of predictors. In the *PLScore Mode* the classical PLS-PM constraints of unitary variance of the LVs are kept. In the *PLScow Mode* the outer weights are constrained to unitary variance at each step of the algorithm as in RGCCA.

Applying *PLS Modes* requires to choose a proper number of PLS components in the PLS regression for each block. This allows considering both the *PLS Modes* as a fine tuning of the analysis between two extreme cases. Classical *Mode A* in case of one PLS-component and classical *Mode B* in case of as many PLS-components as there are MVs in the block are the extreme cases of *PLScore Mode*. *New Mode A* is an extreme case of *PLScow Mode*, when one-component PLS regression is used as outer estimation mode. If more components are considered (while keeping the normalization constraint on the outer weights), *PLScow Mode* yields a new range of solutions between *New Mode A* (one PLS component) and a *New Mode B* (as many PLS components as there are MVs in a block). In such a framework, we certainly prefer talking of composites rather than latent variables. The criterion, if any, being optimized by the multi-component solutions of *PLScow Mode* still needs to be investigated. However, we have empirically shown that *New Mode B* performs very close to classical *Mode B* in terms of correlations between adjacent LVs and, in any case, better in terms of covariances.

The *PLS Modes* are very useful from the user's point of view. As a matter of fact, in the case of *Mode A* the MVs in a block are assumed to be the reflection in the real word of a unique unobserved concept. In other words, the unobserved variable is considered as the source of the covariance between the MVs within the block. As a consequence, the generic outer weight used in the outer estimate of the LV is the regression coefficient of the simple linear regression of each MV on the inner estimate of the corresponding LV. In *Mode B*, instead, each unobserved variable is formed by its own MVs. In other words, the unobserved variable is generated by its own indicators. In this case, the outer weights are the regression coefficients from a multiple regression model of the inner estimate of each LV on its own MVs.

Mode A requires blocks to be homogeneous and unidimensional. However, it may happen in many real applications that unidimensionality, as well as homogeneity, is not verified. In such cases, in standard PLS-PM applications users might pragmatically switch from *Mode A* to *Mode B*. Statistically speaking, switching from simple regressions to a multiple regression implies considering the block of MVs as full dimensional, i.e. the LV being formed by as many dimensions as there are MVs in a block. Indeed, this is also a quite rare situation in real practice as most often blocks are neither unidimensional nor full dimensional. Due to a certain degree of multicollinearity in each block of MVs, it is important to consider just a few dimensions, i.e. we need an estimate of the measurement model capable to yield solutions somewhere between the classical *Mode A* and *Mode B*. The same rationale applies when the user specifies *Mode B* blocks that happen to violate the independence hypothesis of the classical multiple regression model and are affected by multicollinearity problems possibly leading to wrongly non significant or non interpretable outer weights with incoherent signs between the weight of a MV and its correlation with the corresponding LV. The current PLS-PM literature suggests to circumvent such problems in a simplistic and unsatisfactory way by interpreting only the standardized loadings (i.e. correlations between a LV and its own MVs) and not the outer weights in case *Mode B* is used.

PLS Modes allow the user to define multi-dimensional composites in a more proper way. Moreover, standard PLS-R plots and statistics can be useful for interpretation purposes.

To conclude, in 2004 Hwang and Takane proposed the Generalized Structured Component Analysis (GSCA) as an alternative to PLS-PM. GSCA used a formulation similar to SEM even if the latent variables are defined as weighted components of the observed variables. In GSCA all the manifest variables, as well as all the latent variables defined according to equation (1), are included in a supervector $\boldsymbol{u}_i$ of dimension $(J+Q)$ by 1. Moreover all the parameters of the model (i.e. the loadings and the path coefficients) are included in the squared matrix $\boldsymbol{A}$ of dimension $(J+Q)$. This allows the authors to identify a unique function to maximize:

$$\vartheta = \sum_{i=1}^{n} (\boldsymbol{u}_i - \boldsymbol{A}\boldsymbol{u}_i)' (\boldsymbol{u}_i - \boldsymbol{A}\boldsymbol{u}_i) \tag{5}$$

with respect to the component weights and $\boldsymbol{A}$ and under the constraint that the latent variable scores are normalized to unit variance. An Alternating Least Squares (ALS) algorithm is used so as to minimize equation (5), thus is not assured that the convergence is reached in a global minimum. GSCA has the main advantage to provide a unique formulation for both the structural and the measurement models. However, the objective function in equation (5) favors the measurement model part as compared to the structural one. Therefore, GSCA provides results (in terms of loadings) that are most often close to the ones obtained by applying a PCA to each block of MVs. To conclude, among the Goodness of Fit indexes proposed by Hwang and Takane some are based on the discrepancy between the observed covariance matrix and the "implied" one, exactly as in classical SEMs. However, it is not clear why GSCA goodness of fit should be measured in terms of such a discrepancy.

In these last years several authors presented PLS-PM developments, or more generally new component-based approaches, to cope with the challenges raised by the complex data structures available nowadays. In particular, besides the methods presented above, new methods have been presented to include categorical (nominal and ordinal) manifest variables in the PLS-PM, to deal with both unobserved or observed group structure and to model different kinds of structural relations among the blocks (i.e. hierarchical models, predictive flow models, etc..). Further open issues that, in our opinion, currently represent the most important and promising research challenges in PLS Path Modeling and Multi-Tables Analysis include manifest variables selection, the possibility of imposing constraints on the model coefficients, the development of a model estimation procedure based on optimizing the *GoF* index, and finally the possibility to automatically define the structure of the model in terms of structural and/or measurement relations.

References

[1] Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H. (2010) *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Computational Statistics Handbook series (Vol. II), Springer-Verlag, Europe.

[2] Esposito Vinzi, V., Russolillo, G. (2013) "Partial least squares algorithms and methods", *WIREs Computational Statistics*, 5, 1-19.

[3] Fornell, C, Bookstein, F.L. (1982) "Two structural equation models: LISREL and PLS applied to consumer exit-voice theory", *Journal of Marketing Research*, XIX, 440-452.

[4] Glang, M. (1988) "Maximierung der Summe erklärter Varianzen in linear-rekursiven Strukturgleichungsmodellen mit multiple Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahren", PhD Thesis, Universität Hamburg, Hamburg, Germany.

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

[6] Hwang, H., Takane, Y. (2004) "Generalized Structured Component Analysis", *Psychometrika*, 69, 81-99.

[7] Henseler, J. (2010) "On the convergence of the partial least squares path modeling algorithm", *Computational Statistics*, 25 (1), 107-120.

[8] Jöreskog, K.J. (1970) "A general method for analysis of covariance structure", *Biometrika*, 57, 239-251.

[9] Jöreskog K.G., Wold H, (1982) "The ML and PLS techniques for modeling with latent variables: historical and competitive aspects", in K.G. Jöreskog and H. Wold (eds), *Systems under indirect observation*, Part 1, North-Holland, Amsterdam, 263-270.

[10] Krämer, N. (2007) "Analysis of high-dimensional data with Partial Least Squares and boosting", PhD Thesis, Technische Universität Berlin, Berlin, Germany.

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

[12] Mathes, H. (1993) "Global optimisation criteria of the PLS algorithm in recursive path models with latent variables", in: Haagen, K., et al. (eds), *Statistical Modelling and Latent Variables*, Amsterdam: Elsevier Science.

[13] Tenenhaus, A., Tenenhaus, M. (2011) "Regularized Generalized Canonical Correlation Analysis", *Psychometrika*, 76 (2), 257-284.

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

[15] Tenenhaus, M. (2008) "Component-based structural equation modelling", *Total Quality Management & Business Excellence*, 19, 871-886.

[16] Tenenhaus. M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C. (2005) "PLS Path Modeling", *Computational Statistics and Data Analysis*, 48, 159-205.

[17] Wold, H. (1975) "Modelling in complex situations with soft information", in *Third World Congress of Econometric Society*, Toronto, Canada.

[18] Wold, H. (1982) "Soft modeling: the basic design and some extensions", in K.G. Jöreskog et al. (eds.), *Systems under Indirect Observation, Part 2*, North-Holland, Amsterdam, 1-54.

[19] Wold, S., Martens, H., Wold, H. (1983) "The multivariate calibration method in chemistry solved by the PLS method", in: A. Ruhe and B. Kagström (eds.), *Proc. Conf. Matrix Pencils. Lecture Notes in Mathematics*, Springer-Verlag, Heidelberg, 286-293.