

A Non Linear Regularized Component-based Approach to Structural Equation Modeling

Giorgio Russolillo¹

Dip. di Matematica e Statistica - Università degli Studi di Napoli “Federico II”
e-mail: giorgio.russolillo@unina.it

Laura Trinchera

Dip. di Studi sullo Sviluppo Economico - Università degli Studi di Macerata
e-mail: laura.trinchera@unimc.it

Vincenzo Esposito Vinzi

ESSEC Business School of Paris
e-mail: vinzi@essec.fr

Abstract: Structural Equation Models (SEMs) are widely used to model complex causal relations as the ones defining human behaviors. Several techniques exist to estimate SEM parameters. Among them the PLS Path Modeling (PLS- PM) algorithm is the most widely used technique. In particular, PLS-PM allows taking into account *formative* blocks of manifest variables. A new way to compute outer weights in the case of *formative* block of manifest variables has been recently proposed by Esposito Vinzi et al. (2009). This approach involves using PLS Regression (PLS-R) in order to compute outer weights even in the case of multicollinearity among the manifest variables of the same block. However, PLS Regression supposes linearity in relations between variables. Following the work by Esposito Vinzi et al. (2009), we decide to use a non-linear approach to PLS-R in order to estimate measurement model parameters in a non-linear PLS-PM approach to SEMs.

Keywords: PLS Path Modeling, PLS Regression, Non-linear relations

1. Introduction

Structural Equation Models (SEMs) (Bollen, 1989) are widely used to model complex causal relations as the ones defining human behaviors. As a matter of fact, SEMs are very suitable to describe and estimate conceptual structures where some latent variables (LVs), linked by linear relations, are measured by sets of manifest variables (MVs). A double level of relations characterizes each SEM: the first involves relations among the LVs (structural model), while the other considers the links between each LV and its own block of MVs (measurement model).

Several techniques exist to estimate SEM parameters. Among them, PLS Path Modeling (PLS- PM) algorithm (Tenenhaus et al., 2005) is the most widely used technique, thanks to its features. Actually, PLS-PM does not require distributional hypotheses and is more prediction-oriented than classical covariance-based methods, like the LISREL one. Moreover, according to Tenenhaus (2008), it provides systematic convergence of the algorithm

¹ Address of correspondence: Dipartimento di Matematica e Statistica - Università degli Studi di Napoli “Federico II”, Via Cintia, 27 - Complesso Monte S. Angelo, 80126 Napoli (Italy)

due to its simplicity; it allows data to be managed with a small number of individuals and a large number of variables; it provides a practical interpretation of the LV estimates; and it represents a general framework for multi-block analysis. This is of main importance in several application domains, especially when modeling human behaviors.

However, PLS-PM only allows taking into account linear relations both in measurement and structural models. Especially in social sciences is of main importance to model also non-linear relations. Here a modified PLS Path Modeling algorithm, able to model also non-linear relations, is proposed.

In the next section we briefly review the standard PLS Path Modeling algorithm and the use of PLS Regression to estimate measurement model parameters (see section 2.). Then, non-linear approaches to PLS Regression are presented and the original proposition is detailed (see section 3.).

2. Brief review of PLS Path Modeling and PLS Regression

The PLS approach to SEMs (PLS Path Modeling, PLS-PM) uses an iterative algorithm to obtain LV estimates through a system of multiple and simple linear regressions. 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 weighted linear combination of its own MVs, while in the inner estimation step each LV is obtained as a weighted linear combination of the connected LVs (according to the path diagram structure).

Different ways exist to compute outer and inner weights to be used in the two steps of the iterative algorithm. Namely, the inner weights, i.e. the weights associated to each LV in the inner estimation step, can be computed according to three different schemes: the centroid scheme, the factorial scheme and the path weighting scheme. In all of these schemes, the inner weights are a function of the linear correlation between the LVs.

Computation of the outer weights, instead, depends on the nature of the LVs. As a matter of fact, in SEMs it is possible to assume that each block of MVs in the model is of *reflective* or *formative* type. In the case of a *reflective* block, the manifest variables are assumed to be the reflection in the real word of a latent concept. In other words, the LV is considered as a predictor of the MV. 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 a *formative* scheme, instead, each LV is formed by its own MVs measuring different aspect of the same latent concept. In other words, the LV is a function of its own indicators. In this case, a multiple linear regression model defines the relation between the latent and manifest variables. Hence, in a formative scheme the outer weights are the regression coefficients of a multiple regression model of the inner estimate of each LV on its own MVs. However, in the case of multicollinearity among some of the MVs a formative block violates the independence hypothesis of the classical multiple regression model. In this case variance of OLS estimators exponentially increases, leading to regression coefficients never significant. Moreover, multicollinearity may lead to hardly interpretable weights because of the difference in sign between the regression coefficient of a MV and its correlation with the LV. In particular, not interpretable weights may occur if manifest predictors are very correlated and one of the following conditions occurs:

1. there is a weak correlation between the manifest predictors and the regression external criterion (i.e. the LV).

2. the LV is strongly correlated with all the predictors; in this case removing a MV does not change the multiple determination index (redundancy effect).
3. there is some suppressor variable highly correlated with other MVs but very poorly correlated to the LV: if we remove it, the multiple determination index significantly decreases.

Starting from these considerations, a new way to compute outer weights in the case of a *formative* block has been recently proposed by Esposito Vinzi et al. (2009). This approach involves using PLS Regression (PLS-R) (Wold et al., 1983; Tenenhaus, 1998) in order to compute significant outer weights. In particular, Esposito Vinzi et al. (2009) propose to calculate at each iteration the outer weights as coefficient of a PLS1 Regression of the LV inner estimate on the MVs.

The PLS Regression (PLS-R) methodology has been extensively described in the literature (Wold et al., 1983; Tenenhaus, 1998). PLS-R is a linear regression technique that allows relating a set of predictor variables to one (PLS1) or several (PLS2) response variables. PLS-R shrinks the predictor matrix by sequentially extracting orthogonal components which at the same time summarise the explanatory variables and allow modelling and predicting the response variables. Finally, it provides a classical regression equation, in which the response is estimated as a linear combination of the predictor variables. De Jong (1995) showed that the sequence of PLS coefficient estimators forms a suite of vectors whose length strictly increases with the number of components. The upper bound of this suite is the length of least squares estimator. Due to its usefulness in reducing the variance of estimators, PLS-R has become a standard tool to face multicollinearity problems such as the ones previously described.

Following the work by Esposito Vinzi et al. (2009) we decide to use non-linear PLS-R in order to estimate measurement model parameters in a non-linear PLS Path Modeling with formative indicators. In the following section non-linear approaches to PLS Regression are investigated and the new proposal is explained.

3. Non-linear PLS Regression in PLS Path Modeling

Several approaches have been proposed to provide non-linear models that retain the properties of a linear PLS Regression, whereby there appears to be at least three different approaches to non-linear modeling. The first one is based on a non-linear relation linking the predictor PLS components with the response PLS components. The second approach, instead, is based on the maximisation of different criteria, maintaining the properties of a PLS Regression. The last way to cope with non-linearity is based on the transformation of the explanatory variables.

All methods following the latter approach are based on a priori transformation function of the predictors. According to the Durand's proposal (1997, 2001) predictors are transformed by spline functions, while Berglund's (2001) proposal involves the transformation of quantitative variables to a set of dichotomous variables similarly to the coding of qualitative variables. In both these approaches the transformed variables are used as predictors in a linear PLS Regression. All these techniques strongly increase the number of predictors and so there are some difficulties in their application when the number of original predictors is high.

In order to overcome this problem, we propose to use another approach for handling non linearity in PLS Regression: the PLS algorithm for CAtegorical Predictors (PLS-CAP)

(Russolillo, 2009). PLS-CAP was born to handle categorical predictors in a PLS regression. It involves an optimal quantification of predictors, finding out scaling parameters that maximize covariance between the first PLS component and the response variable. PLS-CAP assigns a numerical value to each class of the variable properly discretized. This implies that the number of predictors does not change when variables are transformed. Another advantage of this technique is that the transformation of the variables is internal to the algorithm and not a priori. As a consequence, transformations are more prediction oriented and coherent to the model.

Starting from these considerations, we propose to use such non-linear PLS Regression instead of linear PLS Regression to compute outer weights in the case of *formative* blocks to analyse non linear relations between a LV and its formative MVs. In particular, we propose to perform at each outer estimation step a PLS-CAP Regression between the MVs (as predictors) and the LV (as response variable). This approach is compared to linear PLS-PM and to the approach by Berglund et al. (2001) as estimation methods of formative measurement models.

References

- Berglund A., Kettaneh N., Uppgard L., Wold S. and et al. (2001) The GIFI approach to non-linear PLS modelling, *Journal of Chemometrics*, 15, 321–336.
- Bollen K.A. (1989) *Structural equations with latent variables*, Wiley, New York.
- De Jong S. (1995) Pls shrinks, *Journal of Chemometrics*, 9, 323–326.
- Durand J. (2001) Local polynomial additive regression through PLS and splines: PLSS, *Chemometrics and Intelligent Laboratory Systems*, 58, 235–246.
- Durand J. and Sabatier R. (1997) Additive splines for partial least squares regression, *Journal of the American Statistical Association*, 92, 1546–1554.
- Esposito Vinzi V. and Russolillo G. (2009) Partial least squares path modeling and regression, in: *Wiley Interdisciplinary Reviews: Computational Statistics*, Wegman E., Said Y. and Scott D., eds., John Wiley and Sons, to appear.
- Russolillo G. (2009) A proposal for handling categorical predictors in pls regression framework, in: *Post-Conference Proceedings of First joint meeting of the Societe Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society - Series "Studies in Classification, Data Analysis, and Knowledge Organization"*, June, 11 -13th 2008 Caserta Italy, Springer, in press.
- Tenenhaus M. (1998) *La Régression PLS: théorie et pratique*, Technip, Paris.
- Tenenhaus M. (2008) Component-based structural equation modelling, *Total Quality Management & Business Excellence*, 19, 871–886.
- Tenenhaus M., Esposito Vinzi V., Chatelin Y. and Lauro C. (2005) PLS path modeling, *Computational Statistics and Data Analysis*, 48, 159–205.
- Wold S., Martens H. and Wold H. (1983) The multivariate calibration method in chemistry solved by the pls method, in: *Proc. Conf. Matrix Pencils. Lecture Notes in Mathematics*, Ruhe A. and Kagström B., eds., Springer-Verlag, Heidelberg, 286293.