An integrated PLS Regression-based approach for multidimensional blocks in PLS path modeling

Vincenzo Esposito Vinzi, Giorgio Russolillo, Laura Trinchera

AN INTEGRATED PLS REGRESSION-BASED APPROACH FOR MULTIDIMENSIONAL BLOCKS IN PLS PATH MODELING

Vincenzo Esposito Vinzi & Giorgio Russolillo & Laura Trinchera

ESSEC Business School of Paris,
Avenue Bernard Hirsch, B.P. 50105, 95021 Cergy-Pontoise, France,
E-mail: vinzi@essec.edu

Chaire de Statistique appliquée & CEDRIC-CNAM
292 rue Saint Martin, 75141 Paris cedex 03, France,
E-mail: giorgio.russolillo@cnam.fr

SUPELEC, Départment Signaux & Systèmes Électroniques
Plateau de Moulon - 3, rue Joliot-Curie, 91192 Gif-sur-Yvette, France,
E-mail: laura.trinchera@supelec.fr

Mots-clés: Analyse des données - data mining; Problèmes inverses et sparsité

Résumé: L'approche PLS aux modèles à équations structurelles (PLS Path Modeling, PLS-PM) est couramment considérée comme une approche basée sur les composantes. Cette méthode a été récemment revisitée en tant que cadre général pour l'analyse des tableaux multiples. Nous proposons ici deux nouvelles méthodes d'estimation des poids externes dans le cadre de la PLS-PM: le *Mode PLScore* et le *Mode PLScow*. Chaque mode est fondé sur l'utilisation de la régression PLS pour l'étape d'estimation externe. Toutefois, en *Mode PLScore* une régression PLS est exécutée sous les contraintes classiques de la PLS-PM de variance unitaire pour les scores des variables latentes ; tandis que dans le *Mode PLScow* les poids externes sont contraints d'avoir une norme unitaire. Cette dernière contrainte est la contrainte classique de normalisation dans le cadre de la régression PLS. Nous montrons comment les deux nouveaux modes sont liés aux méthodes d'estimation externe classiques de la PLS-PM, c.-à-d. au *Mode A* et au *Mode B*, ainsi qu'au *Nouveau Mode A* récemment proposé par Tenenhaus & Tenenhaus (2009).

Abstract: PLS Path Modeling (PLS-PM) is classically regarded as a component-based approach to Structural Equation Models and has been more recently revisited as a general framework for multiple table analysis. Here we propose two new modes for estimating outer weights in PLS-PM: the *PLScore Mode* and the *PLScow Mode*. Both modes involve integrating a PLS Regression as an estimation technique within the outer estimation phase of PLS-PM. However, in *PLScore Mode* a PLS Regression is run under the classical PLS-PM constraints of unitary variance for the latent variable scores, while in *PLScow Mode* the outer weights are constrained to have a unitary norm thus importing the classical normalization constraints of PLS Regression. Moreover, we show how the newly proposed modes are linked to the standard *Mode A* and *Mode B* outer estimates in PLS-PM as well as to the *New Mode A* recently proposed in a criterion-based approach by Tenenhaus & Tenenhaus (2009).

Introduction

'Partial Least Squares' Path Modeling (PLS-PM by Wold, 1975; Wold, 1982; 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 covariance-based approach to Structural Equation Models (SEMs). PLS-PM has been more recently revisited also as a general framework for multiple table analysis (Tenenhaus & Tenenhaus, 2009; Tenenhaus & Hanafi, 2010).

Here we present a new approach to estimating outer weights in PLS Path Modeling that is fully based on the PLS principle. Indeed, we propose two new modes for estimating the measurement model: *PLScore Mode* with standardized scores and oriented to maximizing correlations between latent variables (LVs); *PLScow Mode* with constrained weights and oriented to maximizing covariances between LVs. Both modes involve integrating a PLS Regression as an estimation technique within the outer estimation phase of PLS-PM. However, in *PLScore Mode* a PLS Regression is run under the classical PLS-PM constraints of unitary variance for the LV scores, while in *PLScow Mode* the outer weights are constrained to have a unitary norm thus importing the classical normalization constraints of PLS Regression.

In the next section, we briefly review the standard PLS Path Modeling algorithm by enhancing some problematic issues when blocks are neither unidimensional nor full dimensional. Then we present the use of PLS Regression to estimate measurement model outer weights while providing the theoretical foundations of this methodological proposal. We show how the newly proposed modes are linked to the standard *Mode A* and *Mode B* outer estimates in PLS-PM as well as to the *New Mode A* recently proposed in a criterion-based approach by Tenenhaus & Tenenhaus (2009).

Brief review of PLS Path Modeling

PLS Path Modeling uses an iterative algorithm to obtain LV estimates through a system of multiple and simple linear OLS 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 standardized weighted aggregate of its own block of manifest variables (MVs), i.e. $\mathbf{v}_j \propto \sum_h w_{hj}\mathbf{x}_{hj} = \mathbf{X}_j\mathbf{w}_j$ where $\mathbf{x}_{hj}$ $(h = 1,\ldots,p; j = 1,\ldots,J)$ is the generic centered and properly scaled manifest variable of the j-th block $\mathbf{X}_j$. Then, in the inner estimation step each LV is obtained as a standardized weighted aggregate ($\mathbf{z}_j$) of the adjacent LVs, i.e. $\mathbf{z}_j \propto \sum_{\mathbf{v}_{j'} \to \mathbf{v}_j} e_{j'j}\mathbf{v}_{j'}$, where the connections between LVs are defined by the user in the path diagram structure. These two steps are iterated till convergence on the outer weights ($\mathbf{w}_j$). Convergence is measured in terms of stability of the numerical values over two successive iterations.

Inner weights ($e_{j'j}$) can be computed according to three different schemes: the centroid scheme, the factorial scheme and the path weighting scheme. In any case, the inner weights are a function of the linear correlation between adjacent LVs. In our proposal we focus on the outer estimation phase as the schemes for the inner estimation remain unchanged.

Computation of the outer weights ($\mathbf{w}_j$), instead, depends on the type of relation between a block of MVs and the underlying LV. We may assume that each block of MVs in the model

is *outwards directed* (also called reflective block in applied research) or *inwards directed* (also called formative block in applied research). In the case of an *outwards directed* block, MVs are assumed to be the reflection in the real word of a latent concept. In other words, the LV is considered as the cause 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, i.e. $w_{hj} = cov(\mathbf{x}_{hj}, \mathbf{z}_j)$ taking into account that $\mathbf{z}_j$ is standardized. This outer estimation is named *Mode A* in PLS-PM literature. In an *inwards directed* scheme, instead, each LV is formed by its own MVs measuring different aspects of the same latent concept. In other words, the LV is caused 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, i.e. $\mathbf{w}_j = (\mathbf{X}'_j\mathbf{X}_j)^{-1}\mathbf{X}'_j\mathbf{z}_j$. This outer estimation is named *Mode B* in PLS-PM literature.

Outwards directed models require blocks to be homogeneous and unidimensional. However, it may happen in many real applications that unidimensionality, as well as homogeneity, are 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 estimation 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 *inwards directed* 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 Regression in PLS Path Modeling

Starting from the above considerations, a new way to compute the outer weights in the case of *inwards directed* blocks has been recently proposed by Esposito Vinzi & Russolillo (2010). This approach uses PLS Regression (PLS-R) (Wold et al., 1983; Tenenhaus, 1998) as a method to estimate the outer weights in the measurement model under the classical PLS-PM constraints of unitary variance of the LVs. We refer to this approach as *PLScore Mode*.

In this approach, we always deal with univariate PLS regressions (PLS1) where the dependent variable is the LV inner estimate $\mathbf{z}_j$ while its own MVs in $\mathbf{X}_j$ play the role of predictors. Then we search for m orthogonal components, $\mathbf{t}_{kj}(k = 1, \cdots, m)$, which are as correlated as possible to $\mathbf{z}_j$ and also explanatory of their own block $\mathbf{X}_j$. Usually, the number m of retained orthogonal components is chosen by cross-validation or defined by the user.

PLScore Mode leads to a compromise between a multiple regression of $\mathbf{z}_j$ on $\mathbf{X}_j$ (*Mode B*) and a Principal Component Analysis of $\mathbf{X}_j$ (*Mode A* for a single block).

The first PLS component ($\mathbf{t}_{1j}$), if $\mathbf{x}_{hj}$ are standardized variables, is defined as:

$$\mathbf{t}_{1j} = \mathbf{X}_j \mathbf{a}_{1j} = \frac{1}{\sqrt{\sum_h cor^2(\mathbf{z}_j, \mathbf{x}_{hj})}} \sum_h cor(\mathbf{z}_j, \mathbf{x}_{hj}) \tag{1}$$

In case the MVs $\mathbf{x}_{hj}$ are not standardized, the correlation is replaced by the covariance. Then the vector $\mathbf{a}_{1j}$ is normalized and a regression of $\mathbf{z}_j$ on $\mathbf{t}_{1j}$ (expressed in terms of $\mathbf{X}_j$) is run. To conclude, the residuals $\mathbf{z}_{j1}$ and $\mathbf{X}_{j1}$ of the regressions of $\mathbf{z}_j$ and $\mathbf{X}_j$ on $\mathbf{t}_{1j}$ are computed as :

$$\mathbf{z}_{j1} = \mathbf{z}_j - c_{1j}\mathbf{t}_{1j} \tag{2}$$

and

$$\mathbf{X}_{j1} = \mathbf{X}_j - \mathbf{t}_{1j}\mathbf{p'}_{1j} \tag{3}$$

where c_{1j} is the regression coefficient from the regression of $\mathbf{z}_j$ on $\mathbf{t}_{1j}$, and $\mathbf{p}_{1j}$ is the vector of regression coefficients from the regression of the variables in $\mathbf{X}_j$ on $\mathbf{t}_{1j}$.

The second component is defined as $\mathbf{t}_{2j} = \mathbf{X}_{j1}\mathbf{a}_{2j} = \mathbf{X}_j\mathbf{a}^*_{2j}$ where $\mathbf{a}^*_{2j}$ differs from $\mathbf{a}_{2j}$ as the former refers to the original variables in $\mathbf{X}_j$ while the latter refers to the residuals and would be very difficult to interpret. Successive orthogonal components are obtained by iterating the procedure described above on residuals from the previous component, and are assessed by means of cross-validation or chosen by the user.

Applying *PLScore Mode* requires to choose a proper number of PLS components in the PLS regression for each block. This allows considering *PLScore Mode* as a fine tuning of the analysis between two extreme cases: classical *Mode A* in case of one PLS-component; classical *Mode B* in case of as many PLS-components as there are MVs in the block. Indeed, a one-component PLS Regression model coincides with running simple linear regressions of each MV on the inner estimate of the corresponding LV while a full-dimensional PLS Regression model coincides with running a multiple regression of the LV inner estimate on its own MVs.

Thus, the *m*-components PLS regression model yielding the weights for the outer estimation $\mathbf{v}_j$ is:

$$\mathbf{z}_j = c_{1j}\mathbf{t}_{1j} + c_{2j}\mathbf{t}_{2j} + \cdots + c_{mj}\mathbf{t}_{mj} + \mathbf{z}_{jm} \tag{4}$$

where $\mathbf{z}_{jm}$ is the residual of $\mathbf{z}_j$ from the *m*-components PLS regression model.
Remembering that the generic *k*-th component for the *j*-th block is defined as $\mathbf{t}_{kj} = \mathbf{X}_{j(k-1)}\mathbf{a}_{kj}$ with $\mathbf{X}_{j(0)} = \mathbf{X}_j$, equation (4) can be rewritten as:

$$\mathbf{z}_j = c_{1j}\mathbf{X}_j\mathbf{a}_{1j} + c_{2j}\mathbf{X}_{j1}\mathbf{a}_{2j} + \cdots + c_{mj}\mathbf{X}_{j(m-1)}\mathbf{a}_{mj} + \mathbf{z}_{jm} \tag{5}$$

that in terms of original variables becomes:

$$\mathbf{z}_j = \mathbf{X}(c_{1j}\mathbf{a}_{1j} + c_{2j}\mathbf{a}^*_{2j} + \cdots + c_{mj}\mathbf{a}^*_{mj}) + \mathbf{z}_{jm} \tag{6}$$

PLS Regression	Mode A standard PLS-PM	New Mode A modified PLS-PM
$\mathbf{a}_j = \mathbf{X}_j'(\mathbf{X}_{j'}\mathbf{a}_{j'})$	$\mathbf{w}_j = \mathbf{X}_j'(\sum_{j'} e_{qq'}\mathbf{X}_{j'}\mathbf{w}_{j'})$	$\mathbf{w}_j = \mathbf{X}_j'(\sum_{j'} e_{jj'}\mathbf{X}_{j'}\mathbf{w}_{j'})$
$\text{norm}(\mathbf{a}_j)$	$\mathbf{t}_j = \mathbf{X}_j\mathbf{w}_j$	$\text{norm}(\mathbf{w}_j)$
$\mathbf{t}_j = \mathbf{X}_j\mathbf{a}_j$	$\text{norm}(\mathbf{t}_j)$	$\mathbf{t}_j = \mathbf{X}_j\mathbf{w}_j$

Table 1: Iteration loop of PLS-R, PLS-PM with *Mode A* and PLS-PM with *New Mode A*. The iteration steps are repeated for each j and j' in $1:J$ with $j \neq j'$. In the case of PLS-R, $J = 2$

Finally, each LV outer estimate $\mathbf{v}_j$ is obtained as the predicted value of $\mathbf{z}_j$ in (6), i.e.:

$$\mathbf{v}_j = w_{1j}\mathbf{x}_{1j} + w_{2j}\mathbf{x}_{2j} + \cdots\cdots\cdots + w_{pj}\mathbf{x}_{pj} \tag{7}$$

where $\mathbf{w}_j = (w_{1j}\cdots w_{pj})$ is the vector of the *PLScore* outer weights. The outer estimates of the LVs are further transformed so as to satisfy the classical normalization constraint: $var(\mathbf{v}_j) = 1$.

More recently, Tenenhaus & Tenenhaus (2009) have presented a Regularized Generalized Canonical Correlation Analysis (RGCCA) as a new approach to multiple table analysis via a modified PLS-PM algorithm. In particular, by using shrinkage estimators for the block covariance matrices, they prove that if the shrinkage constants are all fixed to zero then classical PLS-PM *Mode B* solution is obtained. Conversely, if the shrinkage constants are all fixed to one, *New Mode A* is defined. This new estimation mode has the major advantage, as compared to classical *Mode A*, to maximize a known criterion. Indeed, Tenenhaus & Tenenhaus (2009) have proved that fixing the shrinkage constants to zero (i.e. standardized LV scores) leads to criteria based on maximizing correlations between adjacent LVs while fixing the shrinkage constants to one (i.e. outer weights with unitary variance) leads to criteria based on maximizing covariances between adjacent LVs.

Table 1 summarizes the algorithmic steps for estimating outer weights and LV scores by comparing PLS regression, the standard PLS-PM algorithm and the modified PLS-PM algorithm for *New Mode A*.

Following the same rationale, if we normalize the outer weights to unitary variance at each step of the algorithm *PLScore Mode* and we use a one-component PLS regression as the outer estimation mode, then we have empirically proved to obtain the same solution as the above mentioned *New Mode A*. We refer to this new approach as *PLScow 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).

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 number of components might be then chosen no longer by cross-validation but so as to maximize a specific fitting criterion, such as the *GoF*.

Conclusions and perspectives

In this paper we have proposed an integrated PLS regression-based approach specifically apt to dealing with multidimensional blocks both in component-based SEM and generalized multiple table analysis. By using different constraints, this approach yields two new estimation modes for outer weights in the measurement model: *PLScore Mode* and *PLScow Mode*. Ongoing research aims at investigating their statistical properties as well as at providing guidelines for use and interpretation in applied research. As a promising research avenue, the features of PLS Regression are also fruitfully exploited in the inner estimation phase of PLS-PM and for estimating path coefficients upon convergence of the PLS-PM algorithm when the classical OLS framework is not feasible.

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. (2010) Partial least squares path modeling and regression, in: E. Wegman, Y. Said and D. Scott (eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*, John Wiley and Sons, to appear.

[3] Tenenhaus A., Tenenhaus M. (2009) A criterion based PLS approach to structural equation modelling, presented at the *6th International Conference on Partial Least Squares Methods - PLS 09*, Beijing, China.

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

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

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

[7] Tenenhaus M., Hanafi (2010) A bridge between PLS path modeling and multi-block data analysis, in: V. Esposito Vinzi et al. (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Computational Statistics Handbook series (Vol. II), Springer-Verlag, Europe.

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

[9] 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, pp. 1-54.

[10] 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, pp. 286-293.