

An extension of Non-Metric approach to inwards directed PLS Path Models

Una generalizzazione dell'approccio Non-Metrico a modelli PLS con blocchi formativi

Giorgio Russolillo and Laura Trinchera

Abstract Partial Least Squares Path Modeling (PLS-PM) aims at modeling multivariate relations among several blocks of variables observed on the same units. Since its first presentation in 1975 by Herman Wold, several authors have developed specific techniques related to the PLS-PM and investigated its mathematical, numerical and statistical properties. In recent years researchers mainly focused on the optimization criteria behind the PLS-PM algorithm and on modifications of such criteria in order to handle variables independently of the measurement scale on which they have been observed. In this paper we extend the Non-Metric approach for the analysis of variables measured on different scales presented by Russolillo in 2012 to inwards directed PLS Path Models.

Abstract *Il PLS Path Modeling (PLS-PM) è un metodo per l'analisi di tabelle multiple che fu introdotto nel 1975 da Herman Wold. Negli ultimi anni le ricerche sul PLS-PM si sono focalizzate principalmente sulla ricerca dei criteri ottimizzati dall'algoritmo PLS-PM e su come modificare questi criteri per trattare variabili indipendentemente dalla scala di misura sulla quale sono state osservate. In questo paper si presenta una generalizzazione dell'approccio Non-Metric PLS, proposto da Russolillo nel 2012, a modelli PLS con blocchi formativi.*

Key words: Optimal Scaling, Non-Metric Variables, PLS Path Modeling

Giorgio Russolillo

CNAM, 292 Rue Saint-Martin - 75003 Paris, France, e-mail: giorgio.russolillo@cnam.fr

Laura Trinchera

NEOMA Business School, 1 Rue du Marchal Juin BP215 - 76825 Mont Saint Aignan, France
e-mail: laura.trinchera@neoma-bs.fr

1 Introduction

Partial Least Squares Path Modeling [13, 3] (PLS-PM) aims at modeling multivariate relations among several blocks of variables observed on the same units. It is so far the most popular component-based alternative to the classical Analysis of Covariance Structures [7] for estimating the parameters of Path Models with Latent Variables (LVPM), also known as Structural Equation Models. PLS-PM is implemented through a flexible algorithm that, depending on the chosen option, can optimize covariance or correlation based criteria.

Since its first presentation in 1975 by Herman Wold [14], several authors have developed specific techniques related to the PLS-PM and investigated the related mathematical, numerical and statistical properties. In the last years researchers have mainly focus on the optimization criteria behind the PLS-PM algorithm [5, 10, 8, 6, 12], on the use of variables measured on different scales [11] and on new ways to compute the outer weights [3].

In this paper we review the Non-Metric approach to PLS-PM [11], which has been so far limited to covariance based PLS Path Models, and we extend this approach to correlation-based PLS-PM. The rest of the paper is structured as follows. First, we review the main aspect of the PLS-PM (Section 2) and of the Non-Metric PLS approach (Section 3). Then, we extend the Non-Metric PLS approach to the case of PLS-PM with inward measurement model (Section 4).

2 PLS Path Modeling

PLS Path Modeling was first introduced by H. Wold [14] for the analysis of Path Models with Latent Variables. The main idea of such models is that Q latent variables (LVs) $\xi_1, \dots, \xi_q, \dots, \xi_Q$, linked in a network of linear relationships, can be used to study the relations between Q blocks of observed variables (MVs) $\mathbf{X}_1, \dots, \mathbf{X}_q, \dots, \mathbf{X}_Q$. Each model has two levels of relationships. The first level concerns relationships between LVs (structural model); these relations can be stored in a symmetric binary matrix $\mathbf{C}$ of order Q , whose generic element $c_{qq'}$ is one if ξ_q is connected to $\xi_{q'}$ and zero otherwise. The other level of relations concerns the links between each LV and its own block of MVs (measurement model).

PLS-PM assumes that each block of observed variables can be summarized by a latent construct obtained as a linear composite of the observed variables. The system of weights to be applied to the observed variables is obtained through an iterative two-step algorithm. In the first step an outer proxy $\mathbf{t}_q$ of each latent construct is obtained as a linear composite of the observed variables; in the second step an inner proxy $\mathbf{z}_q$ of each latent construct is computed as a linear combination of its adjacent latent constructs estimated using the outer proxies computed at the previous step. The inner proxies are then used to update the system of weights (i.e. the outer weights) to be applied to the observed variables. These two steps are iterated till numerical

convergence on the outer weights. Refer to [3] for a detailed presentation of the algorithm.

The PLS-PM algorithm is extremely flexible. Several options, providing slightly different results, can be used to compute inner and outer proxies of the LVs. In this work we focus on two options (also called *schemes*) to calculate the inner weights for computing the inner proxy $\mathbf{z}_q$ and two options (also called *modes*) to compute the outer weights w_{pq} to be used in computing the outer proxy $\mathbf{t}_q$.

As concerns the *schemes*, for each couple of linked outer proxies $\mathbf{t}_q$ and $\mathbf{t}_{q'}$:

1. the centroid scheme (Wold's original scheme) defines the inner weight ($e_{qq'}$) as:

$$e_{qq'} = \text{sign}(\text{cor}(\mathbf{t}_q, \mathbf{t}_{q'}))$$

2. the factorial scheme defines the inner weight ($e_{qq'}$) as:

$$e_{qq'} = \text{cor}(\mathbf{t}_q, \mathbf{t}_{q'}).$$

With respect to the modes, we will refer to:

1. *New Mode A*, represented in the path diagram by outwards directed arrows joining the LV to the corresponding MVs; in *New Mode A* each inner estimate is modeled as a predictor of corresponding MVs. Hence, the outer weights $\mathbf{w}_q$ associated to each LV are the coefficients of the simple regressions of each MV on the corresponding LV inner proxy. These weights are further transformed such as $\|\mathbf{w}_q\| = 1$
2. *Mode B*, represented in the path diagram by inwards directed arrows joining the MVs to the corresponding LV; in *Mode B* a multiple regression model defines the relation between the latent and manifest variables. The outer weights $\mathbf{w}_q$ are thus the regression coefficients of a multiple regression model of the inner estimate of each LV on its own MVs.

Any combination of the these pairs of *schemes* and *modes* has been proved to converge to different criteria [12].

In particular, when *New Mode A* is used for all the blocks, PLS-PM algorithm converges to the criterion:

$$\arg \max_{\forall \mathbf{w}_q} \sum_q c_{qq'} g(\text{cov}(\mathbf{X}_q \mathbf{w}_q, \mathbf{X}_{q'} \mathbf{w}_{q'})) \quad s.t. \quad \|\mathbf{w}_q\| = 1 \quad (1)$$

where $g()$ is the square function if the factorial scheme is used, and the absolute value function if the centroid scheme is used.

When *Mode B* is used for all the blocks PLS-PM algorithm converges to the criterion:

$$\arg \max_{\forall \mathbf{w}_q} \sum_q c_{qq'} g(\text{cov}(\mathbf{X}_q \mathbf{w}_q, \mathbf{X}_{q'} \mathbf{w}_{q'})) \quad s.t. \quad \|\mathbf{X}_q \mathbf{w}_q\| = 1 \quad (2)$$

Hence, when *Mode New A* is used the PLS-PM maximizes the sum of a function of the covariances between the connected latent variables, while when *Mode B* is

used the sum of a function of the correlations between the connected latent variables is maximized.

3 Non-Metric PLS Path Modeling

PLS-PM has two main assumptions:

- Each observed variable is measured on a interval (or ratio) scale.
- Relationships between variables are linear and, consequently, monotonic.

Hence, PLS-PM only applies to quantitative (metric) data. In many real applications users are interested in analyzing data observed on ordinal or nominal measurement scales.

The Non-Metric PLS (NM-PLS) approach has been recently proposed by Russolillo [11] to extend covariance-based PLS criteria (i.e. solutions to Eq. (1)) to the treatment of non-metric variables and non-linearity. This approach is based on the concept of Optimal Scaling (OS) [2, 4]. The OS principle sees observations as categorical, and represents each observation category by a scaling parameter. This parameter is subject to constraints deriving from the measurement characteristics of the variables. In this process each variable $\mathbf{x}$ is transformed as $\hat{\mathbf{x}} \propto \tilde{\mathbf{X}}\phi$, where $\phi' = (\phi_1, \dots, \phi_K)$ is a vector of numeric values (the scaling parameters) associated to the K different values or categories of the variable $\mathbf{x}$, and the matrix $\tilde{\mathbf{X}}$ defines a space in which constraints imposed by the scaling level are respected. The symbol $\propto$ means that the left side of the equation corresponds to the right side normalized to unitary variance.

In order to apply the OS principle to the PLS Path Modeling, when *New Mode A* is used, the following criterion must be optimized:

$$\arg \max_{\forall \mathbf{w}_q, \phi_{pq}} \sum_q c_{qq'} g(\text{cov}(\hat{\mathbf{X}}_q \mathbf{w}_q, \hat{\mathbf{X}}_{q'} \mathbf{w}_{q'})) \quad s.t. \quad \|\mathbf{w}_q\| = \sqrt{n}, \|\hat{\mathbf{x}}_{pq}\| = \sqrt{n} \quad (3)$$

where $\hat{\mathbf{x}}_{pq} = \tilde{\mathbf{X}}_{pq} \phi_{pq}$ represents the observed variable quantified by means of the scaling parameter ϕ_{pq} .

The criterion in Eq. (3) requires two set of parameters to be optimized: the model parameters and the scaling parameters. Recently a new PLS-type algorithm, named NM-PLS, has been proposed for solving Eq. (3) [11]. This algorithm alternately optimizes criterion (3) with respect to each subset, keeping the other fixed. In particular, Russolillo showed that, keeping fixed PLS parameters $\mathbf{w}_q$, the optimal solution for ϕ_{pq} was given by the quantification function $\mathcal{Q}(\tilde{\mathbf{X}}_{pq}, \mathbf{z}_q)$ which orthogonally projects $\mathbf{z}_q$ on the space spanned by $\tilde{\mathbf{X}}_{pq}$.

4 Non-Metric PLS Path Modeling for Mode B

Here we propose a new algorithm to extend the results in [11] to the *Mode B* and to whatever combination of the modes and schemes presented in Section 2.

To extend the problem in Eq. (2) to the *Mode B* case, the following optimization problem must be solved:

$$\arg \max_{\forall \mathbf{w}_q} \sum_q c_{qq'} g(\text{cov}(\hat{\mathbf{X}}_q \mathbf{w}_q, \hat{\mathbf{X}}_{q'} \mathbf{w}_{q'})) \quad s.t. \quad \|\hat{\mathbf{X}}_q \mathbf{w}_q\| = \sqrt{n}, \|\hat{\mathbf{x}}_{pq}\| = \sqrt{n} \quad (4)$$

It is possible to show [12] that this criterion can be rewritten as:

$$\arg \max_{\forall \mathbf{w}_q, \phi_{pq}} \sum_q \text{cor}(\hat{\mathbf{X}}_q \mathbf{w}_q, \mathbf{z}_q) \quad (5)$$

A NM-PLS algorithm can be used to maximize this criterion. For fixed scaling parameters, the optimization problem in (5) can be solved with respect to $\mathbf{w}_q$ by using the usual PLS-PM iteration. In order to optimize the problem with respect to ϕ_{pq} while keeping $\mathbf{w}_q$ fixed, we propose the following *backfitting procedure* (see [2, 1], among others). Following the back fitting principle, the optimal solution for $\hat{\mathbf{x}}_{pq}$ is obtained by applying the quantification function $\mathcal{Q}(\hat{\mathbf{X}}_{pq}, \mathbf{x}_{pq}^*)$ on the back-fitted variables $\mathbf{x}_{pq}^*$, obtained as:

$$\mathbf{x}_{pq}^* = (1/b_{pq})(\mathbf{z}_q - \sum_{j \neq p} \hat{\mathbf{x}}_{jq} b_j)$$

where $\mathbf{b}_q = (\hat{\mathbf{X}}_q' \hat{\mathbf{X}}_q)^{-1} \hat{\mathbf{X}}_q' \mathbf{z}_q$.

Criterion (5) is solved by alternating a standard PLS iterative loop, a back fitting step and a quantification step. At each step the criterion (5) increases, thus the proposed algorithm is expected to converge to a maximum.

References

1. Breiman, L., Friedman, J. H.: Estimating optimal transformations for multiple regression and correlation. *JASA* **80**, 580–597 (1985)
2. de Leeuw, J., Young, F.W., Takane Y.: Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika* **41**, 471–503 (1976)
3. Esposito Vinzi, V., Russolillo, G.: Partial least squares algorithms and methods. *WIREs Computational Statistics* **5**, 1–19 (2013)
4. Gifi, A.: *Nonlinear Multivariate Analysis*, Wiley, Chichester, UK (1990)
5. Glang, M.: Maximierung der Summe erklärter Varianzen. In: *Linearrekursiven Strukturgleichungsmodellen mit multiple Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahren*, PhD Thesis, Universität Hamburg, Hamburg, Germany (1988)
6. Hanafi, M.: PLS path modeling: computation of latent variables with the estimation mode B. *Computational Statistics* **22**, 275–292 (2007)
7. Jöreskog, K. G.: Analysis of covariance structures. In *Multivariate analysis III*, P. R. Krishnaiah (Ed.), New York: Academic Press, Inc., 263–285 (1973)

8. Krämer, N.: Analysis of high-dimensional data with Partial Least Squares and boosting. PhD Thesis, Technische Universität Berlin, Berlin, Germany (2007)
9. Lohmöller, J.B., Wold, H.: Three-mode path models with latent variables and partial least squares (PLS) parameter estimation. Presented at the European meeting of the Psychometrics Society, Groningen, The Netherlands (1980)
10. Mathes, H.: Global optimisation criteria of the PLS algorithm in recursive path models with latent variables. In: Statistical Modelling and Latent Variables, Haagen, K., et al. (eds), Amsterdam: Elsevier Science (1993)
11. Russolillo, G.: Non-Metric Partial Least Squares. *Electronic Journal of Statistics* **6**, 1641–1669 (2012)
12. Tenenhaus, A., Tenenhaus, M.: Regularized Generalized Canonical Correlation Analysis. *Psychometrika* **76** (2), 257–284 (2011)
13. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS Path Modeling. *Computational Statistics & Data Analysis*, **48**, 159–205 (2005)
14. Wold, H.: Modelling in complex situations with soft information. In: Third World Congress of Econometric Society, Toronto, Canada (1975)
15. Young, F.W.: Quantitative analysis of qualitative data. *Psychometrika* **44** (4), 357–388 (1981)