The Non-Metric Partial Least Squares approach

Giorgio Russolillo

THE NON-METRIC PARTIAL LEAST SQUARES APPROACH

Giorgio Russolillo

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

Résumé: Dans ce travail, nous montrons comment les algorithmes PLS, correctement ajustés, peuvent travailler comme des algorithmes de Codage Optimal. Cette nouvelle fonctionnalité du PLS, qui avait été jusqu'à maintenant totalement inexplorée, nous a permis de mettre au point une nouvelle série de méthodes PLS: les méthodes Non-Metric PLS (NM-PLS).

Abstract: In this work we find out how PLS algorithms, properly adjusted, can work as optimal scaling algorithms. This new feature of PLS, which had until now been totally unexplored, allowed us to devise a new suite of PLS methods: the Non-Metric PLS (NM-PLS) methods.

Mots-clès: Analyse des données - data mining, Problèmes inverses et sparsité

Introduction

Partial Least Squares (PLS) methods embrace a suite of data analysis techniques based on algorithms belonging to the PLS family. These algorithms consist of various extensions of the Nonlinear estimation by Iterative PArtial Least Squares (NIPALS) algorithm, which was proposed by Herman Wold (1966) as an alternative algorithm for implementing a Principal Component Analysis (PCA). Wold proposed NIPALS also to analyze causal relations between several blocks of variables (Wold, 1975): this is the PLS approach to Structural Equation Modeling, later called PLS-Path Modeling (PLS-PM, Tenenahus *et al.*, 2005). Svante Wold, Herman's son, perceived that the PLS approach could be used in order to implement a regularized component-based regression, called PLS-Regression (PLS-R) (Wold *et al.*, 1983).

PLS techniques, as all quantitative methods, were born to handle data sets forming metric spaces. This involves all the variables embedded in the analysis being observed on interval or ratio scales. In this work, variables measured at ratio or interval scale level will be referred to as *numeric* or *metric* variables.

Unfortunately, in many fields where PLS methods are applied (*e.g.* genomics, sensorial analysis, consumer analysis, marketing) researchers are also interested in analyzing sets of variables measured on a non-metric scale, *i.e.* ordinal and nominal variables.

Nowadays, "*among the open issues that currently represent the most important and promising research challenges in PLS-PM,*" there is the "*specific treatment of categorical (nominal and ordinal) variables and specific treatment of non linearity*" (Esposito Vinzi *et al.*, 2010).

This work focuses on new methodological proposals to make PLS techniques able to handle jointly metric and non-metric data. Next, we briefly introduce three PLS methods: NIPALS, PLS-R and PLS-PM, focusing on their algorithmic flow. Then, a suite of corresponding algorithms working as optimal scaling methods, called Non-Metric PLS (NM-PLS), is proposed.

PLS methods: algorithmic flows and optimization criteria

NIPALS algorithm performs a PCA on a set $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_p \dots \mathbf{x}_P]$ of variables observed on N individuals. The weight w_p of variable $\mathbf{x}_p$, is calculated in such a way to maximize the squared correlation of the variable with a linear combination $\mathbf{t} = \mathbf{X}\mathbf{w}$ of all the variables. Such a weight leads to the first PC, that satisfies the criterion

$$\underset{||\mathbf{w}||=1}{\arg\max} \left\{ \text{var}(\mathbf{X}\mathbf{w}) \right\}. \tag{1}$$

PLS-R analyzes the dependence between a set $\mathbf{X}_1 = [\mathbf{x}_{11} \dots \mathbf{x}_{1p} \dots \mathbf{x}_{1P}]$ of predictors and a set $\mathbf{X}_2 = [\mathbf{x}_{21} \dots \mathbf{x}_{2p} \dots \mathbf{x}_{2P}]$ of response variables. PLS-R extracts a suite of orthogonal components in the predictor space aimed to explain well both predictors and response variables. The first PLS-R component satisfies the criterion

$$\underset{||\mathbf{w}_1||=||\mathbf{w}_2||=1}{\arg\max} \left\{ \text{cov}^2(\mathbf{X}_2\mathbf{w}_2, \mathbf{X}_1\mathbf{w}_1) \right\} \tag{2}$$

In PLS-R, in order to satisfy criterion (2), Russolillo (2009) demostrated that the weight for each variable in a set is calculated in such a way to maximize the squared correlation with a linear combination (the score vector) of the variables belonging to the other set.

PLS Path Modeling (Tenenhaus *et al.*, 2005) aims to estimate the relationships among Q blocks $\mathbf{X}_1, \dots, \mathbf{X}_q, \dots, \mathbf{X}_Q$ of manifest variables (MVs), which are expression of Q unobservable constructs $\xi_1, \dots, \xi_q, \dots, \xi_Q$, that are usually called latent variables (LVs). The corresponding conceptual model can be represented by a path diagram. In the PLS path model external weights w_{pq}, linking each MV to the corresponding LV, are estimated by an iterative procedure in which the latent variable scores are obtained through the alternation of *outer* ($\mathbf{t}_q$) and *inner* estimations of the LVs. This procedure is referred to as the PLS Path Modeling (PLS-PM) algorithm. In PLS-PM the outer weights can be calculated in two ways, called *Mode A* and *Mode B*. When *Mode A* is used, in each iteration, the outer weight w_{pq} of variable $\mathbf{x}_{pq}$ is calculated in such a way to maximize the squared correlation of the variable with a weighted sum (the inner estimation of ξ_q) of linear combinations (the outer estimations) of the variables $\xi_{q'}$ belonging to connected blocks $\mathbf{X}_{q'}$. Full *Mode A* PLS Path Model does not seem to optimize any criterion, as Kramer (2007) showed that Wold's *Mode A* algorithm is not based on stationary equations related to the optimization of a twice differentiable function. However, Tenenhaus *et al.* (2009) have recently proved that a slightly adjusted *Mode A*, in which a normalization constraint is put on the weights is used in all the blocks, monotonically converges to the criterion

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

NIPALS iteration	PLS-R iteration	new Mode A PLS-PM iteration
$\mathbf{w} = \mathbf{X}'(\mathbf{X}\mathbf{w})$	$\mathbf{w}_q = \mathbf{X}_q'(\mathbf{X}_{q'}\mathbf{w}_{q'})$	$\mathbf{w}_q = \mathbf{X}_q'(\sum_{q'} e_{qq'}\mathbf{X}_{q'}\mathbf{w}_{q'})$
$\text{norm}(\mathbf{w})$	$\text{norm}(\mathbf{w}_q)$	$\text{norm}(\mathbf{w}_q)$
$\mathbf{t} = \mathbf{X}\mathbf{w}$	$\mathbf{t}_q = \mathbf{X}_q\mathbf{w}_q$	$\mathbf{t}_q = \mathbf{X}_q\mathbf{w}_q$

Table 1: Iterative steps of NIPALS algorithm, PLS-R algorithm and PLS-PM algorithm with Mode A. In PLS-PM the iteration steps are repeated for each q and q' in $1:Q$ with $q \neq q'$. In the case of PLS-R iteration $Q = 2$. $e_{qq'}$ is the generic element of a squared matrix of order Q that is null if $\xi_{q'}$ is connected to ξ_q; otherwise, it represents the corresponding inner weight.

where c_{qq} is the generic element of the binary matrix $\mathbf{C}$ defining the path diagram and function $g(\cdot)$ depends on the scheme used for the inner estimation of the LVs.

The algorithmic core of all these methods is an iterative process with which vectors of scores for each component (or latent variable) are obtained as a weighted sum of the corresponding block of indicators (or manifest variables). Each iteration of a PLS algorithm can be resumed in three steps. In the first step the weights are updated, in the second one they are used for building the score vector(s), and in the third step the score vector(s) or the weight vector(s) (depending on the algorithm we take into account) are properly normalized. In this work we focus on this iterative process, used for the extraction of the first order component, in NIPALS, PLS-R and PLS-PM algorithms. The strong analogy of these algorithms can be observed in table 1. It is noteworthy that in all these methods, in order to optimize criteria (1), (2) and (3), PLS weights are worked out in such a way to maximize the squared correlation of the corresponding variables with a latent construct, that from here we will call Latent Criterion (LC). The LC is an unknown vector of order N, centered by construction. For each PLS method different LCs are considered:

- In NIPALS, the LC to bear in mind is the first PC.
- In PLS-R we have to keep into account as LCs the vector scores in predictor and in response spaces.
- In *new Mode A* PLS-PM framework, a LC is considered for each block of manifest variables, *i.e.* the corresponding outer estimate.

Next, we will refer to all of these LCs with the generic notation $\gamma = f(\mathbf{w})$. Through the concept of LC, we can resume all the PLS criteria (equations (1), (2), (3)) in a general criterion expressed as a function of γ:

$$\underset{||\mathbf{w}_q||=1}{\arg\max} \left\{ \sum_p g\left(\text{cor}(\mathbf{x}_{pq}, \gamma_q)\right) \right\} \forall q \tag{4}$$

Limits of the PLS algorithms: a non-metric solution

In order to satisfy criterion (4), in NIPALS, PLS-R and *new Mode A* PLS-PM, when working on standardized variables, optimal weights are calculated as Pearson's product-moment correlation coefficients between each variable and the LC. This leads to two basic hypotheses underlying PLS models:

- Each variable is measured on a interval (or ratio) scale.
- Relations between variables and latent constructs are linear and, consequently, monotone.

As a consequence, standard PLS methods cannot handle data which are measured on a scale which does not have metric properties.

There is a simple way to overcome this problem: replacing each non-metric variable with the corresponding indicator matrix. Most of the software currently used to perform PLS analyses use such a coding in order to handle categorical variables; however, this is not a valid solution to the problem (Russolillo, 2009).

We propose an alternative approach, aimed to adjust the PLS iteration to devise an Optimal Scaling (OS) procedure, calculating iteratively scaling and model parameters. This new PLS procedure leads to a new class of algorithms which implement methods that generalize the standard PLS methods. We call them Non-Metric PLS (NM-PLS) methods (Russolillo, 2009), because they are able to provide optimally scaled data ($\widehat{\mathbf{x}}$) with a new metric structure, which does not depend on the metric properties of the raw data $\mathbf{X}^*$), *i.e.* data to be scaled. In other words, NM-PLS methods yield a metric to non-metric data, and a new metric to metric data, making relationships between variables and latent constructs linear, as required by the hypothesis of standard PLS models. These methods could be named non-linear PLS methods as well, since they discard the intrinsic linearity hypothesis of the standard PLS methods. However, by naming them Non-Metric PLS methods, we prefer to highlight their ability to work just on non-metric features of data.

Similarly to other OS approaches (Gifi, 1990), the aim of NM-PLS algorithms is to optimize criterion

$$\underset{\text{var}(\widehat{\mathbf{x}})=1,||\mathbf{w}_q||=1}{\arg\max} \left\{\text{cor}^2(\widehat{\mathbf{x}}, \gamma)\right\} \tag{5}$$

under two sets of parameters: the model parameters and the scaling parameters, constrained to the restrictions due to the scaling level chosen for each raw variable $\mathbf{x}^*$. In NM-PLS algorithms, model and scaling parameters are alternately optimized in a modified PLS loop where a quantification step is added. In standard PLS steps the model parameters are optimized for given scaling parameters. In the quantification step, instead, the scaling parameters are optimized for given model parameters: raw variables are properly transformed thought scaling functions $\mathscr{Q}(\cdot)$, then they are normalized to unitary variance (see table 2).

NM-PLS methods satisfy (5) under three possible levels of scaling analysis: nominal, ordinal and functional. To each level of scaling analysis, it corresponds an *ad hoc* scaling function.

NM-NIPALS iteration	NM-PLSR iteration	NM-PLSPM iteration
$\widehat{\mathbf{x}}_p \propto \mathscr{Q}(\mathbf{x}^*_p, \mathbf{t})$	$\widehat{\mathbf{x}}_{pq} \propto \mathscr{Q}(\mathbf{x}^*_{pq}, \mathbf{t}_{q'})$	$\widehat{\mathbf{x}}_{pq} \propto \mathscr{Q}(\mathbf{x}^*_{pq}, \mathbf{t}_{q'})$
$\mathbf{w} = \widehat{\mathbf{X}}'(\widehat{\mathbf{X}}\mathbf{w})$	$\mathbf{w}_q = \widehat{\mathbf{X}}'_q(\widehat{\mathbf{X}}_{q'}\mathbf{w}_{q'})$	$\mathbf{w}_q = \widehat{\mathbf{X}}'_q(\sum_{q'} e_{qq'}\widehat{\mathbf{X}}_{q'}\mathbf{w}_{q'})$
$\text{norm}(\mathbf{w})$	$\text{norm}(\mathbf{w}_q)$	$\text{norm}(\mathbf{w}_q)$
$\mathbf{t} = \widehat{\mathbf{X}}\mathbf{w}$	$\mathbf{t}_q = \widehat{\mathbf{X}}_q\mathbf{w}_q$	$\mathbf{t}_q = \widehat{\mathbf{X}}_q\mathbf{w}_q$

Table 2: Iterative loops of NM-NIPALS algorithm, NM-PLSR algorithm and NM-PLSPM algorithm. Symbol $\propto$ implies equality but for a normalization constant

In a nominal analysis, a variable is quantified as the orthogonal projection of γ in the space spanned by the columns of the indicator matrix $\tilde{\mathbf{X}}$ generated by the K categories of $\mathbf{x}^*$:

$$\tilde{\mathscr{Q}}(\mathbf{x}^*, \gamma) : \widehat{\mathbf{x}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\gamma. \tag{6}$$

Quantification function $\tilde{\mathscr{Q}}$ maximizes (5) under the grouping constraint that, for each pair of observations i and i',

$$(x^*_i \sim x^*_{i'}) \Rightarrow (\hat{x}_i = \hat{x}_{i'}), \tag{7}$$

where the symbol $\sim$ indicates membership in the same category. When this scaling function is used, the relation between γ and $\mathbf{x}^*$ in terms of linear correlation can be expressed as Pearson's correlation ratio $\eta_{\gamma|\mathbf{x}^*}$

$$\text{cor}(\gamma, \widehat{\mathbf{x}}) = \eta_{\gamma|\mathbf{x}^*}. \tag{8}$$

If $\mathbf{x}^*$ is an (almost) ordinal variable, to be quantified at an ordinal scale level, the following ordering scaling function (Young, 1975) is used:

$$\tilde{\tilde{\mathscr{Q}}}(\mathbf{x}^*, \gamma) : \widehat{\mathbf{x}} = \tilde{\tilde{\mathbf{X}}}(\tilde{\tilde{\mathbf{X}}}'\tilde{\tilde{\mathbf{X}}})^{-\infty}\tilde{\tilde{\mathbf{X}}}'\gamma, \tag{9}$$

where $\tilde{\tilde{\mathbf{X}}}$ is built by Kruskal's secondary least squares monotonic transformation (Kruskal, 1964). The vector of the regression coefficient $(\tilde{\tilde{\mathbf{X}}}'\tilde{\tilde{\mathbf{X}}})^{-1}\tilde{\tilde{\mathbf{X}}}'\gamma$ contains the unnormalized optimal scaling values which preserve the order of the categories of $\mathbf{x}^*$, as required by the condition

$$(x^*_i \sim x^*_{i'}) \Rightarrow (\hat{x}_i = \hat{x}_{i'}) \text{ and } (x^*_i \prec x^*_{i'}) \Rightarrow (\hat{x}_i \leq \hat{x}_{i'}). \tag{10}$$

where symbol $\prec$ indicates empirical order. In this case $\text{cor}(\gamma, \widehat{\mathbf{x}})$ can be interpreted as a measure of the approaching monotonicity of the relation between $\mathbf{x}^*$ and the LC; it equals the unity if there exists a perfect increasing monotonicity and it is equal to -1 when there exists a perfect decreasing monotonicity.

With the functional scaling we suppose that we know the degree of a polynomial relation between a raw numerical variable and the LC. Following Young (1981), optimal parameters for the polynomial transformation are found by projecting γ in the conic space spanned by the

columns of matrix $\dot{\mathbf{X}}$. Matrix $\dot{\mathbf{X}}$ is built with a row for each observation and with $D+1$ columns, each column being an integer power of the vector $\mathbf{x}^*$:

$$\dot{\mathscr{Q}}(\mathbf{x}^*, \gamma) : \widehat{\mathbf{x}} = \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\gamma \tag{11}$$

If we suppose that the variable and the LC are linked by a linear relation, we just have to put $D = 1$. If this is the case for all of the variables, NM-PLS methods provide the same results of the standard PLS methods applied on standardized data.

Scalings provided by NM-PLS methods are shown to be optimal (Russolillo, 2009), as:

- They optimize the same criterion of the method in which they are involved.
- They respect the constraints defining which properties of the original measurement scale we want to preserve.

References

[1] Esposito Vinzi, V., Trinchera, L. & Amato, S. (2010), PLS path modeling: Recent developments and open issues for model assessment and improvement, in V. Esposito Vinzi *et al.* eds, *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*, Springer.

[2] Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Chichester, UK: Wiley

[3] Krämer, N. (2007), *Analysis of high-dimensional data with partial least squares and boosting*, Phd thesis, Technischen Universität Berlin, Berlin, Germany

[4] Kruskal, J. (1964), Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29 (1), 1-27

[5] de Leeuw, J., Young, F. & Takane, Y. (1976), Additive structure in qualitative data: an alternating least squares method with optimal scaling features, *Psychometrika* 41, 471-503.

[6] Russolillo G. (2009), *PLS methods for Non-Metric data*, PhD thesis, Università degli Studi di Napoli "Federico II", Italy.

[7] Tenenhaus M. & Tenenhaus M. (2009), A criterion based PLS approach to structural equation modelling, presented at *6th International Conference on Partial Least Squares Methods*.

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

[9] Wold, H. (1966), Estimation of principal component and related models by iterative least squares, in P. R. Krishnaiah, ed., *Multivariate Analysis*, Academic Press, New York, 391-420.

[10] Wold, H. (1975), Path models with latent variables: The non-linear iterative partial least squares (NIPALS) approach, in H. M. Blalock *et al.* eds, *Quantitative Sociology: Intentional Perspective on Mathematical and Statistical Modeling*, Accademic Press, 307-357.

[11] 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 Mathematic*, Ruhe A. and Kagström B., eds., Springer-Verlag, Heidelberg, 286-293.

[12] Young, F. [1975], Methods for describing ordinal data with cardinal models, *Journal of Mathematical Psychology* 12, 416-436.

[13] Young, F. [1981], Quantitative analysis of qualitative data, *Psychometrika* 44 (4), 357-388.