Chemometrics and Intelligent Laboratory Systems 120 (2013) 116–125

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A family of regression methods derived from standard PLSR

Jean-Claude Boulet [a,*], Dominique Bertrand [b,1], Gérard Mazerolles [a], Robert Sabatier [c], Jean-Michel Roger [d]

[a] INRA, UMR1083 Sciences pour l'oenologie, F-34060 Montpellier, France
[b] INRA, UR1268 Biopolymères interactions assemblages, F-44316 Nantes, France
[c] UM1, EA2415 Biostatistique épidémiologie recherche clinique, F-34093 Montpellier, France
[d] IRSTEA, UMR ITAP Information et technologie pour les agroprocédés, F-34191 Montpellier, France

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form 30 October 2012
Accepted 3 November 2012
Available online 13 November 2012

Keywords:
PLSR
Metric
Calibration
Regression
Orthogonal
Oblique
Projection

ABSTRACT

The standard PLSR is presented from a geometric point of view consisting of two projections. In the first, the scores are obtained after an oblique projection of the spectra onto the loadings. In the second, the vector of response values is projected orthogonally onto the scores. A metric is introduced for the oblique projection and a new algorithm for the calculation of the loadings into the variables space is proposed. This work also develops a new parameter, a vector, whose different values lead to different regression models with their own abilities of prediction; one of them is the exact form of the standard PLSR. Two applications are described to illustrate the performance of the proposed method called VODKA regression, which is also a way to build least square regressions by introducing additional knowledge into the models.

1. Introduction

Many current analytical methods are based on spectroscopic techniques such as near infrared, mid infrared or Raman spectroscopy. The data generated are sets of observations, e.g. spectra acquired from several samples, and sets of the corresponding analytical results obtained using generally time-consuming analytical techniques. The observations form a first matrix. The analytical results form a second matrix containing the quantitative amounts of one or several compounds of interest, or response variables. Response variables are predicted using the observations associated to a calibration method. This is an important goal for the development of on-line, fast and non-destructive analytical methods. The most popular of the proposed methods is partial least square regression or projection to latent structures regression (PLSR). PLSR is a linear indirect calibration method. Terms of PLSR-1 and PLSR-2 are respectively associated with the prediction of one or several response variables. This work addresses only PLSR-1, called PLSR hereafter for simplification.

The non linear iterative partial least squares (NIPALS) algorithm was proposed by H. Wold for principal component analysis (PCA) calculations [1]. Modifications of this algorithm led H. and S. Wold and H. Martens to the first PLSR algorithm [2], here called standard PLSR [3,4] to avoid confusion with NIPALS for PCA. Other algorithms have since been proposed, including non orthogonalized scores PLSR by Martens [5,6] and SIMPLS by De Jong [3]. The goal of these algorithms is to give results close to the standard PLSR, at least for PLSR-1. As a consequence, Andersson [7] compared the respective performances of nine PLSR algorithms for the two following criteria: speed and numerical stability. The standard PLSR belonged to the four most stable algorithms, and thus confirmed its status as a reference.

The standard PLSR has been presented from different points of view, e.g. an application of the Heisenberg uncertainty principle [8], statistical modeling [9], its geometry [4], or the algorithm itself [2,7,10,11] with the calculation of the different parameters: loadings **P** and **c**, weights **W**, scores **T**. The various properties of PLSR have been reviewed elsewhere and are beyond the scope of this article. Our aim is to show that the same algorithm can be written in such a way that only two parameters are necessary: a metric **M** and the loadings **P**; the geometric properties are highlighted.

VODKA regressions, a new family of regression methods, are derived from this new presentation of standard PLSR. The vector $\mathbf{r} = \mathbf{X}'\mathbf{y}$ is considered as a parameter, which can be replaced by any other vector of the same dimension for the calculation of the loadings. Each value of **r** is associated with a different regression model whose accuracy depends strongly on a relevant choice for **r**. Several approaches are proposed for the choice of **r** and two applications illustrate the proposed method.

* Corresponding author. Tel.: +33 499613148; fax: +33 499612857.
E-mail addresses: bouletjc@supagro.inra.fr (J.-C. Boulet), domibertrand@free.fr (D. Bertrand), mazeroll@supagro.inra.fr (G. Mazerolles), sabatier@univ-montp1.fr (R. Sabatier), jean-michel.roger@irstea.fr (J.-M. Roger).
[1] Present address: data_frame, 25 rue Stendhal, F-44300 Nantes, France.

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.11.002

2. Theory

The theory is divided into three parts: definitions and notations; a rewriting of standard PLSR including a new algorithm for the calculation of the loadings; and the proposal of a new regression method.

2.1. Definitions and notations

Vectors are noted in bold lowercase, all matrices, but projectors, in bold uppercase, projectors in calligraphic, scalars in normal uppercase and variables in normal characters. A spectrum is represented as a column vector, but several spectra form the rows of a matrix, e.g. in $\mathbf{X}$ or $\mathbf{X}_G$. Vectors generated by calculations form the columns of the matrices which include them, e.g. $\mathbf{P}$ or $\mathbf{W}$. The transposed forms of vector $\mathbf{m}$ and matrix $\mathbf{M}$ are respectively noted $\mathbf{m}'$ and $\mathbf{M}'$. Terms of metric (also used for pseudo-metric), orthogonal and oblique projectors, antiprojectors, and the main notations are given in Table B.1.

The linear model is written in terms of matrices and vectors:

$$\mathbf{y}_{raw} = \mathbf{1}_N b_0 + \mathbf{X}_{raw}\mathbf{b}_{raw} + \mathbf{e}_{raw} \tag{1}$$

where ($\mathbf{X}_{raw}$, $\mathbf{y}_{raw}$) is the raw dataset containing the values of the explanatory variables and the explained variable for N observations, b_0 the offset, $\mathbf{b}_{raw}$ the regression vector and $\mathbf{e}_{raw}$ the error. The problem of the offset can be fixed by adding a column of ones to $\mathbf{X}_{raw}$ [6]. Moreover, if other pretreatments are necessary, e.g. centering, smoothing, orthogonal projection, they should be applied previously to the raw dataset. They yield the calibration dataset ($\mathbf{X}$, $\mathbf{y}$) with $\mathbf{X}$ of dimensions ($N\times Q$), N observations of Q variables. Eq. (1) can be simplified to:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

The aim of a regression is to estimate the vector $\hat{\mathbf{b}}$ that minimizes the residual error.

2.2. A new presentation of the standard PLSR algorithm

The standard PLSR algorithm is described in Appendix A. Weights $\mathbf{W}$, and loadings $\mathbf{P}$ and $\mathbf{c}$ are calculated then used to build the regression vector of b-coefficients $\hat{\mathbf{b}}$, see Eq. (A.6). PLSR decomposes a matrix $\mathbf{X}$ into matrices $\mathbf{T}$, $\mathbf{P}$ and a residual matrix $\mathbf{E}$ such that: $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \mathbf{X}^{\mathbf{U}} + \mathbf{E}$ [12], where $\mathbf{X}^{\mathbf{U}}$ represents the information from $\mathbf{X}$ which is useful for the prediction of $\mathbf{y}$, the column vectors of $\mathbf{P}$ form a basis of the useful space, $\mathbf{T}$ contains the scores of the observations in this basis and $\mathbf{E}$ is the error. Various properties of standard-PLSR are reported in Appendix A.2. Let $\mathbf{M}$ be the Moore-Penrose pseudo-inverse of $\mathbf{X}'\mathbf{X}$, and $\mathbf{P}$ be the oblique projector with an inner product $\mathbf{M}$ onto the space spanned by the columns of $\mathbf{P}$. A new expression of $\mathbf{X}^{\mathbf{U}}$ is obtained with the expression of $\mathbf{T}$ from property 5:

$$\mathbf{X}^U = \mathbf{T}\mathbf{P}' = \mathbf{X}\mathbf{M}\mathbf{P}\left(\mathbf{P}'\mathbf{M}\mathbf{P}\right)^{-1}\mathbf{P}' = \mathbf{X}\mathcal{P}' \tag{2}$$

The geometry of PLSR is highlighted: the useful information extracted from $\mathbf{X}$ is obtained by a M-oblique projection of $\mathbf{X}$ onto the loadings $\mathbf{P}$. As PLSR is also an orthogonal projection (or regression) of the reference values $\mathbf{y}$ onto the scores $\mathbf{T}$ [4], a new expression of $\hat{\mathbf{b}}$ is obtained (see property 6):

$$\hat{\mathbf{b}} = \mathbf{M}\mathbf{P}\left(\mathbf{P}'\mathbf{M}\mathbf{P}\right)^{-1}\mathbf{P}'\mathbf{M}\mathbf{X}'\mathbf{y} \tag{3}$$

$$\begin{aligned} &= \mathcal{P}'\mathbf{M}\mathbf{X}'\mathbf{y} \\ &= \mathbf{M}\mathcal{P}\mathbf{X}'\mathbf{y} \end{aligned} \tag{4}$$

The calculation of $\mathbf{M}$ is straightforward with the function *pinv* of Matlab or Scilab, based on a singular value decomposition of $\mathbf{X}'\mathbf{X}$; this is however slow, so faster methods have been proposed for large and rank-deficient matrices, e.g. geninv [13] and CGS-MPi [14]. Thus, $\mathbf{P}$ remains the only parameter to be calculated to obtain a PLSR model. According to Eq. (A.5), the deflation of $\mathbf{X}$ at step i is performed into $\mathbb{R}^N$, so the calculation of the loadings $\mathbf{p}_i$ implies successive steps into $\mathbb{R}^N$ and into $\mathbb{R}^Q$. However, property 7 and Eq. (A.9) show that the deflation of $\mathbf{X}$ at step i can be performed only into $\mathbb{R}^Q$ following:

$$\mathbf{X}_{1:i} = \mathcal{T}^{\perp}_{1:i}\mathbf{X} = \mathbf{X}\mathcal{P}'^{\perp}_{1:i}$$

So the steps into $\mathbb{R}^N$ are no longer necessary, and it becomes possible to rewrite the calculation of the loadings of standard PLSR into $\mathbb{R}^Q$ only and also independently of the parameters $\mathbf{T}$, $\mathbf{W}$ and $\mathbf{c}$. An algorithm is thereby obtained and described in Appendix B.

Once the model is built, for a new observation $\mathbf{x}_v$, $\hat{y}_v$ is deduced from Eq. (3):

$$\hat{y}_v = \mathbf{x}'_v\mathbf{M}\mathbf{P}\left(\mathbf{P}'\mathbf{M}\mathbf{P}\right)^{-1}\mathbf{P}'\mathbf{M}\mathbf{X}'\mathbf{y} = \mathbf{x}'_v\mathcal{P}'\mathbf{M}\mathbf{X}'\mathbf{y} = \mathbf{x}'^{U}_v\mathbf{M}\mathbf{X}'\mathbf{y} \tag{5}$$

Thus the prediction is a two-step process. The first step is an M-oblique projection of $\mathbf{x}_v$ onto $\mathbf{P}$, yielding the useful part $\mathbf{x}^U_v = \mathcal{P}\mathbf{x}_v$; the scores of the observations associated to the basis $\{\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_A\}$ are: $\mathbf{t}_v = \mathbf{x}'_v\mathbf{M}\mathbf{P}(\mathbf{P}'\mathbf{M}\mathbf{P})^{-1}$. In the second step, $\hat{y}_v$ is the M-inner product between $\mathbf{x}^U_v$ and $\mathbf{X}'\mathbf{y}$. The term $\mathbf{X}'\mathbf{y}$ appears in the regression coefficients of PLSR, PCR and OLSR [4,6]. However, it also has a particular place in PLSR when building the loadings, as seen in Appendix B. This allows us to develop the following method.

2.3. VODKA regression, an outcome of the new presentation of standard PLSR

PLSR aims at determining scores that maximize $(\mathbf{t}'_i\mathbf{y})^2$ under the condition: $\|\mathbf{w}_i\| = 1$ [12]. Using property 3 and the normalization of $\mathbf{t}_i$ in the proposed algorithm, this constraint can be switched from $\mathbb{R}^N$ to $\mathbb{R}^Q$ and expressed as: maximizing $\mathbf{p}'_i\mathbf{M}\mathbf{X}'\mathbf{y}$ under the condition $\mathbf{p}'_i\mathbf{M}\mathbf{p}_i = 1$. The question is: is $\mathbf{X}'\mathbf{y}$ really the best vector? If, as suggested by Helland [15], PLSR models can theoretically be improved, there may be another vector $\mathbf{r}$ which is more representative of the relevant information from $\mathbf{X}$ that explains $\mathbf{y}$.

This issue has been discussed in contexts other than PLSR. The net analyte signal (NAS) [16] is the most condensed spectral information about the compound to be predicted; it is also the basis of the principle of direct calibration, e.g. [17]. Two definitions of the NAS have been proposed [18]: (1) *the NAS for a component is the part of its pure spectrum which is orthogonal to the pure spectra of the other constituents*; (2) *the NAS is the part of the gross spectrum that is useful for prediction*. According to the first definition, if the pure spectrum $\mathbf{k}$ of the compound to be quantified is known, and if all other influences have been characterized as spectra or loadings and merged into the matrix $\mathbf{D}$, the NAS can be estimated: $\mathbf{s}_{\mathrm{nas}} = (\mathbf{I}_Q - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}')\mathbf{k}$. The regression coefficients obtained from PLSR or other regression models are also estimations of the NAS [18]. Moreover, in certain conditions, the regression vector of PLSR can be the NAS exactly [9]. Therefore, if a good approximation of the NAS can be obtained with additional information, the NAS can be used as the value of $\mathbf{r}$.