A family of regression methods derived from standard PLSR

J.C. Boulet, D. Bertrand, G. Mazerolles, Rodolphe Sabatier, J.M. Roger

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

A family of regression methods derived from standard PLSR

Jean-Claude Boulet[a,1], Dominique Bertrand[b,2], 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*

Abstract

We present a new regression method derived from standard PLSR which has a geometric point of view and consists of two projections. In the first, 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 calculating the loadings into the variable space is proposed. This work also puts forward a new parameter, a vector, whose different values lead to different regression models with their own prediction abilities, and one of them is the exact form of standard PLSR. This method (called vector orientation decided through knowledge assessment, or VODKA regression) is another way to build least squares regressions using only a few latent variables. We propose two

Email addresses: bouletjc@supagro.inra.fr (Jean-Claude Boulet), domibertrand@free.fr (Dominique Bertrand), mazeroll@supagro.inra.fr (Gérard Mazerolles), sabatier@univ-montp1.fr (Robert Sabatier), jean-michel.roger@irstea.fr (Jean-Michel Roger)

[1] corresponding author

[2] present address: Data_ Frame, 25 rue Stendhal, F-44300 Nantes, France

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

applications to illustrate its performance capabilities.

Keywords: PLSR; metric; orthogonal; oblique; projection; Vodka

1. Introduction

Many current analytical methods are based on spectroscopic techniques such as near-infrared (NIR), mid-infrared(MIR) or raman spectroscopy. The data consist of a set of observations, such as spectra acquired from several samples, and a set of the corresponding analytical results obtained using generally time-consuming analytical techniques. The observations form a first matrix, and the analytical results form a second matrix, which contains the quantitative amounts of one or more compounds of interest, the response variables. The prediction of the response variables can be done using the observations associated with the calibration method. This is a main goal for the development of online, fast and non-destructive analytical methods. Among the proposed methods, partial least squares regression or projection to latent structures regression (PLSR) is the most popular. PLSR is a linear indirect calibration method. PLSR-1 and PLSR-2 are associated with the prediction of one or several response variables respectively. This work concerns only PLSR-1 which is noted PLSR for simplification.

The nonlinear iterative partial least squares (NIPALS) algorithm was proposed by H.Wold for principal component analysis (PCA) calculations [1]. Modifications of this algorithm led H.Wold, S.Wold and H.Martens to develop the first PLSR algorithm [2], which is referred to as ¨ standard PLSR ¨ [3, 4] to avoid confusion with NIPALS for PCA. Later, other algorithms were proposed,

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

such as non-orthogonalized scores PLSR by Martens [5] [6], and SIMPLS by De Jong [3]. The goal of most PLSR algorithms including PLSR-1 has been to produce results close to standard PLSR, at least for PLSR-1. As a consequence, Andersson [7] compared the respective performances of nine PLSR algorithms for speed and the numerical stability. Standard PLSR belonged to the four most stable algorithms, and thus confirmed its status as a reference method.

Standard PLSR has been presented from different points of view, such as an application of the Heisenberg uncertainty principle [8], statistical modeling [9], and its geometry [4]. However, most presentations have concerned the algorithm itself [2, 7, 10, 11] and the calculation of different parameters, such as loadings $\mathbf{P}$ and $\mathbf{c}$, weights $\mathbf{W}$ and scores $\mathbf{T}$. The parameters $\{ \mathbf{w}_i, \mathbf{t}_i, \mathbf{p}_i, c_i \}$ are calculated simultaneously for a latent variable i, then $\{ \mathbf{w}_{i+1}, \mathbf{t}_{i+1}, \mathbf{p}_{i+1}, c_{i+1} \}$ are calculated for the latent variable $i+1$, and so on. We show that the same algorithm can be written differently such that each item is calculated separately. A metric $\mathbf{\Sigma}$ is defined as the Moore-Penrose pseudo-inverse of $\mathbf{X}'\mathbf{X}$. The loadings $\mathbf{P}$ are obtained using two elements: a square matrix $\mathbf{X}'\mathbf{X}$ and a vector $\mathbf{r} = \mathbf{X}'\mathbf{y}$, but in a manner different from the Krylov sequence. Then, the matrix $\mathbf{T}$ and the regression vector $\mathbf{b}$ are obtained with $\mathbf{P}$ and $\mathbf{\Sigma}$. Neither $\mathbf{W}$ nor $\mathbf{c}$ are necessary, so they are not calculated.

VODKA regressions, which comprise 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 to be a parameter, which can be replaced by any other vector of the same dimension to calculate the loadings. Each value of $\mathbf{r}$ is associated with a

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

different regression model whose accuracy depends strongly on a relevant choice for $\mathbf{r}$. Several approaches are proposed for choosing $\mathbf{r}$ and two applications illustrate the proposed method.

2. Theory

The theory is divided into three parts: the standard PLSR algorithm; a rewriting of the standard PLSR including a new algorithm for the calculation of the loadings into the variable space; and the proposal of a new regression method. Vectors are noted in bold lowercase, matrices in bold uppercase, scalars in normal uppercase, 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 $\mathbf{X}_G$. On the other hand, vectors issued from calculations form the columns of the matrices which gather 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 $\mathbf{M}'$. Table B.1 summarizes the main notations. In the general case, if optional pretreatments are necessary, e.g. centering, smoothing, orthogonal projection, they should be applied to the raw data previously to yield the calibration dataset $(\mathbf{X}, \mathbf{y})$.

A data $\mathbf{X}$ of dimension $(N \times P)$ can be explained into the $\mathbb{R}^N$ space spanned by its P column vectors of dimensions $(N \times 1)$, or into the $\mathbb{R}^P$ space spanned by its N line vectors of dimensions $(P \times 1)$. We will focus on this second issue. A metric represented by a square and symmetrical matrix $\mathbf{\Sigma}$ of dimensions $(P \times P)$ is associated with $\mathbb{R}^P$ to form a vectorial space, and is used to calculate inner (dot) products and distances as well as to perform projections. To simplify the

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

notations, the term metric is also used for pseudo-metric, that is, when $\mathbf{\Sigma}$ is not of full rank. The usual Euclidian space is associated with the identity: $\mathbf{\Sigma} = \mathbf{I}_P$ and with *orthogonal* projections. Oblique Euclidian spaces are associated with $\mathbf{\Sigma} \neq \mathbf{I}_P$ and with *oblique* projections. Two types of projectors can handle both orthogonal and oblique projections: *projectors onto* a subspace, and *anti-projectors to* a subspace. For example, the oblique $\mathbf{\Sigma}$ *projector onto* the subspace spanned by the column vectors of $\mathbf{M}$ is: $\mathcal{P}_M = \mathbf{\Sigma M}(\mathbf{M}'\mathbf{\Sigma M})^{-1}\mathbf{M}'$, and the oblique $\mathbf{\Sigma}$ *anti-projector to* the subspace spanned by the column vectors of $\mathbf{M}$ is: $\mathcal{P}_M^{\perp} = \mathbf{I}_P - \mathcal{P}_M$. The terms $\mathcal{P}$ and $\mathcal{T}$ designate projectors; $\mathcal{P}^{\perp}$ and $\mathcal{T}^{\perp}$ designate anti-projectors. Due to the context, $\mathcal{P}$ and $\mathcal{P}^{\perp}$ are defined into $\mathbb{R}^P$ with the metric $\mathbf{\Sigma}$, whereas $\mathcal{T}$ and $\mathcal{T}^{\perp}$ are defined into $\mathbb{R}^N$ with the usual Euclidian metric.

2.1. The standard PLSR algorithm

Standard-PLSR has been described several times, for instance by Geladi [11]. It aims at building a model for A latent variables. To start, $\mathbf{X}_{1:0} = \mathbf{X}$ and $\mathbf{y}_{1:0} = \mathbf{y}$. Then a loop calculates the PLSR parameters at each iteration. For $i = 1, 2, 3...A$:

$$\mathbf{w}_i = \mathbf{X}'_{1:i-1}\mathbf{y}_{1:i-1} \quad (1)$$

$$\| \mathbf{w}_i \| = 1 \quad (2)$$

$$\mathbf{t}_i = \mathbf{X}_{1:i-1}\mathbf{w}_i \quad (3)$$

$$c_i = \mathbf{y}'_{1:i-1}\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \quad (4)$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

$$\mathbf{p}_i = \mathbf{X}'_{1:i-1}\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \quad (5)$$

$$\mathbf{X}_{1:i} = \mathbf{X}_{1:i-1} - \mathbf{t}_i\mathbf{p}'_i = (\mathbf{I}_N - \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i)\mathbf{X}_{1:i-1} \quad (6)$$

$$\mathbf{y}_{1:i} = \mathbf{y}_{1:i-1} - \mathbf{t}_i\mathbf{c}'_i = (\mathbf{I}_N - \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i)\mathbf{y}_{1:i-1} \quad (7)$$

then the algorithm returns to equation (1) incrementing i by 1. After A iterations, the A weight vectors $\mathbf{w}_i$, the A loadings-for-$\mathbf{X}$ vectors $\mathbf{p}_i$, the A score vectors $\mathbf{t}_i$ and the A loadings-for-$\mathbf{y}$ scalars c_i are gathered into the respective matrices and vector $\mathbf{W}$, $\mathbf{P}$, $\mathbf{T}$ and $\mathbf{c}$. The calibration model for A latent variables is represented by a regression vector of b-coefficients $\mathbf{b}$ which verifies: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, with a vector of errors $\mathbf{e}$. Let $\widehat{\mathbf{b}}$ be an estimation of $\mathbf{b}$, and let $\widehat{\mathbf{y}}$ be the estimation of $\mathbf{y}$ using $\widehat{\mathbf{b}}$. Thus, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\mathbf{b}}$, with:

$$\widehat{\mathbf{b}} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{c} \quad (8)$$

2.2. Rewriting standard PLSR

PLSR decomposes a matrix $\mathbf{X}$ into matrices $\mathbf{T}$, $\mathbf{P}$ and a residual matrix $\mathbf{E}$ such that: $\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \mathbf{X}^U + \mathbf{E}$, where $\mathbf{X}^U$ represents the information from $\mathbf{X}$ which is useful for the prediction of $\mathbf{y}$. Several properties of standard PLSR, recalled or demonstrated, are reported in Appendix A. Let $\mathbf{\Sigma}$ be the Moore-Penrose pseudo-inverse of $(\mathbf{X}'\mathbf{X})$; $\mathbf{\Sigma} = (\mathbf{X}'\mathbf{X})^+$. In Property 5, Eq. A.6 gives a new expression of $\mathbf{T}$, that is: $\mathbf{T} = \mathbf{X\Sigma P}(\mathbf{P}'\mathbf{\Sigma P})^{-1}$. So:

$$\mathbf{X^U} = \mathbf{TP}' = \mathbf{X\Sigma P}(\mathbf{P}'\mathbf{\Sigma P})^{-1}\mathbf{P}' \quad (9)$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

This means that the useful information from $\mathbf{X}$ is obtained by an oblique projection of $\mathbf{X}$ onto the loadings $\mathbf{P}$ into $\mathbb{R}^P$. The matrix $\mathbf{T}$ contains the scores of the observations in the basis of the useful space spanned by the column-vectors of $\mathbf{P}$. PLSR is also a regression (or orthogonal projection into $\mathbb{R}^N$) of the reference values $\mathbf{y}$ onto the scores $\mathbf{T}$, i.e. $\widehat{\mathbf{y}} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}\mathbf{y}$ [4]. The new expression of $\mathbf{T}$ leads to a new expression of $\widehat{\mathbf{b}}$ (see property 6, appendix A):

$$\widehat{\mathbf{b}} = \mathbf{\Sigma}\mathbf{P}(\mathbf{P}'\mathbf{\Sigma}\mathbf{P})^{-1}\mathbf{P}'\mathbf{\Sigma}\mathbf{X}'\mathbf{y} \tag{10}$$

From a geometric point of view, standard PLSR consists of two projections (see Fig.B.1):

- First, an oblique projection into the variable space $\mathbb{R}^P$

  The vector $\mathbf{p}_i$ is the i^{th} element of a basis of the subspace of $\mathbb{R}^P$ which contains the relevant information for the prediction of $\mathbf{y}$. The A first vectors $\mathbf{p}_i$ form the loading matrix $\mathbf{P}$ of dimensions $(P \times A)$. The information from $\mathbf{X}$ which is useful for the prediction of $\mathbf{y}$ is extracted by an oblique $\mathbf{\Sigma}$ projection of $\mathbf{X}$ onto $\mathbf{P}$, yielding $\mathbf{X}^U$; the scores of its observations into the basis $\{\ \mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_A\}$ are given by $\mathbf{T}$.

- Then, an orthogonal projection into the observation space $\mathbb{R}^N$

  The predicted vector $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathbf{T}$. The regression vector $\widehat{\mathbf{b}}$ is deducted from this last equation.

The calculation of $\mathbf{\Sigma}$ is straightforward. Thus, $\mathbf{P}$ remains the only parameter to calculate to obtain a PLSR model. According to Eq. 6, the deflation of $\mathbf{X}$ at

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

step i is performed into $\mathbb{R}^N$ by multiplying the anti-projector orthogonal to $\mathbf{T}_{1:i}$ by $\mathbf{X}$. Thus, the calculation of the loadings $\mathbf{p}_i$ implies successive steps into $\mathbb{R}^N$ and into $\mathbb{R}^P$. However, it is shown in Property 7 and Eq. A.7 that the deflation of $\mathbf{X}$ at step i can also be performed into $\mathbb{R}^P$:

$$\mathbf{X}_{1:i} = \mathcal{T}_{1:i}^{\perp}\mathbf{X} = \mathbf{X}\mathcal{P}_{1:i}^{\perp}$$

Thus, the calculation steps into $\mathbb{R}^N$ are no longer mandatory, and it becomes possible to rewrite standard PLSR as a calculation of the loadings only into $\mathbb{R}^P$ and independently from the parameters $\mathbf{T}$, $\mathbf{W}$ or $\mathbf{c}$. An algorithm is obtained and described in Appendix B. This new approach for the calculation of the loadings raises the following issues.

- Geometric building of the $\mathbf{p}_i$

  Let z be a point of $\mathbb{R}^P$ and $\mathbf{q}$ the vector from the origin of $\mathbb{R}^P$ to z. Let γ be a positive scalar. The set of points z which verify: $\mathbf{q}'\mathbf{X}'\mathbf{X}\mathbf{q} = \gamma$ form an ellipsoid. The value of γ tunes the size of the ellipsoid. We are more interested in the shape of the ellipsoid, whose main directions are the eigenvectors of $\mathbf{X}'\mathbf{X}$. According to Fig. B.2 adapted from [4], the direction of $\mathbf{p}_1$ is the result of a tangent rotation of $\mathbf{X}'\mathbf{y}$ towards the main direction of the ellipsoid, i.e. the eigenvector associated with the largest eigenvalue of $\mathbf{X}'\mathbf{X}$. The other loadings $\mathbf{p}_i$ are calculated similarily using the deflated forms of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ (see Appendix B). They are compromises between the main spectral information and the spectral information that explains

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

$\mathbf{y}$.

- The Krylov sequences

  The Krylov sequences can yield the matrix $\mathbf{P}$ directly [12]. A sequence is obtained by the multiplication of two terms, $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ for the loadings, the first being risen to power i for loading i. For example, $\mathbf{p}_{K,1} = \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y}$, $\mathbf{p}_{K,2} = \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y}$ and so on; $\mathbf{p}_{K,i} = (\mathbf{X}'\mathbf{X})^i\mathbf{X}'\mathbf{y}$. The loadings obtained by the Krylov sequence have no particular properties, moreover they can be highly (but not completely) colinear, so it is convenient to replace them by vectors $\mathbf{v}_i$ obtained after a Gram-Schmidt orthogonalization of the $\mathbf{p}_{K,i}$ [3]. One difference between standard PLSR and the Krylov sequence is that the $\mathbf{v}_i$ are orthogonal with a Euclidian metric: $\mathbf{v}'_i\mathbf{v}_j = 0$ for $i \neq j$, whereas the $\mathbf{p}_i$ are orthogonal with the $\mathbf{\Sigma}$ metric: $\mathbf{p}'_i\mathbf{\Sigma}\mathbf{p}_j = 0$ for $i \neq j$. The $\mathbf{p}_i$ and the $\mathbf{v}_i$ span the same subspaces, but only the $\mathbf{p}_i$ constitute a $\mathbf{\Sigma}$ orthonormal basis, according to Eq. 10. Thus the proposed algorithm for the loadings calculation is not based on the Krylov sequences.

- The weights $\mathbf{W}$

  The relationships between the vectors $\mathbf{p}_i$ and $\mathbf{w}_i$ can be determined directly from the standard PLSR algorithm or from [13]. However, they remain complex and it has not been possible to identify a relationship between the matrices $\mathbf{P}$ and $\mathbf{W}$ of same dimension ($P \times A$). The apparent proximity between Eqs. 8 and 10 is misleading because ($\mathbf{P}'\mathbf{W}$) is a bidiagonal matrix [14] and ($\mathbf{P}'\mathbf{\Sigma}\mathbf{P}$) is strictly diagonal (see Appendix

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

A, Property 4). Therefore, clearly $\mathbf{W} \neq \mathbf{\Sigma P}$ and $\mathbf{W}$ is neither used nor explained here.

- Properties of $\mathbf{\Sigma}$

  If $\mathbf{\Sigma}$ is the identity or if $\mathbf{\Sigma}$ is built such that $\mathbf{\Sigma} = \mathbf{M'M}$ for any matrix $\mathbf{M}$, the inner or dot product defined by $f(\mathbf{u}, \mathbf{v}) = \mathbf{u'\Sigma v}$ verifies the conditions of a metric or pseudo-metric. Using the singular value decomposition, it applies also to $\mathbf{\Sigma} = (\mathbf{M'M})^+$. So, under our conditions $\mathbf{\Sigma} = (\mathbf{X'X})^+$ has the position and the properties of a metric. Moreover, it is a metric associated with a Mahalanobis distance [15]. Each observation i is represented by a point i into $\mathbb{R}^P$, or a vector $\mathbf{x}_i$ between the origin and the point i. In the particular case where $\mathbf{X}$ has been previously centered, $(\mathbf{X'X})/N$ is the variance-covariance matrix of the P spectral variables and $\mathbf{\Sigma} = (\mathbf{X'X})^+$ represents its pseudo-inverse, the coefficient $1/N$ can be dropped. The origin is located at the center of the cloud of points representing the observations of $\mathbf{X}$. It is important to note that this metric depends on the data, which is not the case for several other regression methods such as Multiple Linear Regression (MLR) or Principal Component Regression (PCR). If spectra are removed or added to $\mathbf{X}$, then the metric and the inner product are changed. As a consequence, the scores and loadings are different. For instance, suppose that two PLSR models are built: the first one with $(\mathbf{X}, \mathbf{y})$ for A latent variables, yielding the loadings $\mathbf{P}_1$; the second one with $\mathbf{X}^U$ defined above and $\mathbf{y}$, yielding $\mathbf{P}_2$. The two metrics are different, $(\mathbf{X'X})^+$ and $(\mathbf{X}^{U\prime}\mathbf{X}^U)^+$ respectively, and so are the two models.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

It is not possible to recover the same models, as reported previously [16], but this is logical and not an artifact. With orthogonal and oblique projections, standard PLSR is based on very simple and basic rules of matrix algebra, and its calculation is correct [16, 17].

The added value of a generalized Euclidian metric such as $\mathbf{\Sigma}$ is to remove or decrease the weights associated with certain directions of the space. Thus, these directions become more weak. The Mahalanobis distance lightens the weights of the variables which are the most variable and also takes into account their relationships with the other variables. This property is relevant for prediction problems, in which spectral variables of low variability may be good predictors. The Mahalanobis metric can accentuate them whereas the usual Euclidian metric, used in PCR for instance, cannot.

Other multivariate methods based on the Mahalanobis distance, such as the Hotelling test [15], are perfectly in accord with standard PLSR.

- The regression vector $\mathbf{b}$

  In studying the geometry of PLSR [4], Phatak concluded that ¨ *the PLS estimator of the vector of b-coefficients...is an oblique projection of the ordinary least square (OLS) estimator* ¨ . If $\mathbf{\Sigma}$ is of full rank, the OLS estimator is $\widehat{\mathbf{b}}_{OLS} = \mathbf{\Sigma X'y}$ [6, 4]. From Eq. 10, it is straightforward that the oblique projector is $\mathbf{\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'}$. Thus, the PLSR regression vector is the oblique $\mathbf{\Sigma}$ projection of $\widehat{\mathbf{b}}_{OLS}$ onto the space spanned by the loadings $\mathbf{p}_i$.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

- The scores $\mathbf{t}_i$ and the prediction of new observations

  In the standard PLSR algorithm, a score $\mathbf{t}_i$ is calculated before the corresponding loading $\mathbf{p}_i$, which suggests that the loadings calculation depends on the scores [16]. In fact, the reciprocal is true: from Eq. A.4, scores also depend on the loadings for their calculation. To conclude, scores and loadings are calculated together and they depend on each other. An example has been shown previously for SIMPLS [3]: standardization of the scores in SIMPLS simultaneously standardizes the loadings. It is also concluded from Eq. 9 that the so-called scores are true scores: for an observation $\mathbf{x}$ and for $i = 1$ to $i = A$, the score value t_i represents the expansion of the loading $\mathbf{p}_i$.

  For a new observation $\mathbf{x}_v$ which does not belong to $\mathbf{X}$, the prediction $\widehat{y}_v$ is deduced from Eq. 10 :

$$\widehat{y}_v = \mathbf{x'}_v \mathbf{\Sigma P} (\mathbf{P'\Sigma P})^{-1} \mathbf{P'\Sigma X'y}$$

$$\widehat{y}_v = \mathbf{x'}^U_v \mathbf{\Sigma X'y} \tag{11}$$

  The prediction can be viewed as two steps:

  - First step: an oblique $\mathbf{\Sigma}$ projection of $\mathbf{x}_v$ onto $\mathbf{P}$, yielding the useful part of $\mathbf{x}_v$ which is $\mathbf{x}^U_v$; the scores of $\mathbf{x}^U_v$ in the basis $\{ \mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_A\}$ are: $\mathbf{t}_v = \mathbf{x}_v \mathbf{\Sigma P} (\mathbf{P'\Sigma P})^{-1}$;

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

– Second step: $\widehat{y}_v$ is the $\mathbf{\Sigma}$ inner product between $\mathbf{x}_v^U$ 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 specific place in PLSR when building the loadings, as seen in Appendix B, which leads us to propose the following method.

2.3. VODKA regressions: 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: $\| \mathbf{w}_i \| = 1$ [18]. Using Eq. A.4 and the normalization of the $\mathbf{t}_i$ in the proposed algorithm, this constraint can be switched from $\mathbb{R}^N$ to $\mathbb{R}^P$ and expressed as: maximizing $\mathbf{p'}_i\mathbf{\Sigma}\mathbf{X}'\mathbf{y}$ under the condition $\mathbf{p'}_i\mathbf{\Sigma}\mathbf{p}_i = 1$. The question is whether $\mathbf{X}'\mathbf{y}$ is truly the best vector? Would another vector $\mathbf{r}$ of the same dimension be more representative of the relevant information from $\mathbf{X}$ which explains $\mathbf{y}$?

This issue has been discussed previously in a context independent from PLSR. The net analyte signal (NAS) [19] is the most condensed spectral information about the compound to be predicted, and it also constitutes the principle of direct calibration [20]. Two definitions of the NAS have been proposed [21]: (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: $\mathbf{s}_{nas} = (\mathbf{I}_P - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}')\mathbf{k}$.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

The regression coefficients obtained from PLSR or other regression models constitute another estimation of the NAS [21]. Moreover, in certain conditions, the regression vector of PLSR can be exactly the NAS [9]. Therefore, if a good approximation of the NAS can be obtained with additional information, this justifies using the NAS for value of $\mathbf{r}$.

When no other information than $(\mathbf{X}, \mathbf{y})$ is available, this vector $\mathbf{r}$ can be built using only $\mathbf{X}$ and $\mathbf{y}$. Moreover, $\mathbf{r}$ is set in the form: $\mathbf{r} = \mathbf{X}'g(\mathbf{y})$, where $g(\mathbf{y})$ is a vector function of $\mathbf{y}$. The vector $\mathbf{X}'\mathbf{y}$ from standard PLSR with $g(\mathbf{y}) = \mathbf{y}$ is collinear to the mean of the observations, weighted by the concentrations of the response variable. The higher the concentrations are, the more the corresponding spectra will contribute to $\mathbf{r}$. When $\mathbf{X}$ or $\mathbf{y}$ are centered, the intermediate values in $\mathbf{y}$ are not well represented in $\mathbf{r}$, which is obtained mainly by subtracting the spectra associated with the lowest values of the response variable from the spectra associated with the highest values of the response variable. A different weighting is possible which is useful for taking into account the nonlinearities between the response variable and the observations, or for simply overexpressing the contribution of the response variable. For example, four different functions $g(\mathbf{y})$ can be developed along with several other possibilities. Let $g_1(\mathbf{y}) = \mathbf{y}^2$, $g_2(\mathbf{y}) = exp(\mathbf{y})$, $g_3(\mathbf{y}) = \sqrt{}(\mathbf{y})$ and $g_4(\mathbf{y}) = log(\mathbf{y})$ be four functions in which each result is a vector obtained after: raising each of the elements of $\mathbf{y}$ to the power of 2 (g_1); taking the exponential of the elements of $\mathbf{y}$ (g_2); the square root of the elements of $\mathbf{y}$ (g_3); or the logarithm of the elements of $\mathbf{y}$ (g_4). If we suppose that all the elements of $\mathbf{y}$ are larger than 1, then using $g_1(\mathbf{y})$ or $g_2(\mathbf{y})$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

will accentuate the weight of the highest concentrations whereas using $g_3(\mathbf{y})$ or $g_4(\mathbf{y})$ will reduce the weight of the highest concentrations.

New regression models can be obtained by switching the vector $\mathbf{X}'\mathbf{y}$ of standard PLSR with another vector $\mathbf{r}$ of the same dimension. The loadings are first calculated with the following algorithm derived from Appendix B:

- Step 1:

$$\begin{aligned} \mathbf{p}_1 &= \mathbf{X}'\mathbf{X}\mathbf{r} \\ \mathbf{p}_1 &= \mathbf{p}_1(\mathbf{p}'_1\boldsymbol{\Sigma}\mathbf{p}_1)^{-0.5} \end{aligned}$$

- Step $i+1$:

$$\begin{aligned} \mathbf{p}_{i+1} &= \mathcal{P}'^{\perp}_{1:i}\mathbf{X}'\mathbf{X}\mathcal{P}'^{\perp}_{1:i}\mathbf{r} \\ \mathbf{p}_{i+1} &= \mathbf{p}_{i+1}(\mathbf{p}'_{i+1}\boldsymbol{\Sigma}\mathbf{p}_{i+1})^{-0.5} \end{aligned}$$

Other parameters, such as scores and regression coefficients are obtained with Eqs. A.6 and 10 respectively. From Figure B.2, it is straightforward that a choice for $\mathbf{r}$ different from $\mathbf{X}'\mathbf{y}$ leads to a new orientation of the $\mathbf{p}_i$ vectors and of the useful space they span. For the models to be relevant, however, the choice of $\mathbf{r}$ cannot be left to chance. It is advisable to first review the available knowledge regarding the data to be processed, and then to choose the most useful elements (e.g. an estimation of the NAS) to build $\mathbf{r}$. This step is summarized in the proposed name for this method: vector orientation decided through knowledge assessment (VODKA) regression.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

3. Applications

We used VODKA in two applications: the quantification of ethanol in fermenting wines and of mannoproteins in polysaccharide extracts. These applications presented several respective differences: the samples were liquid and powder, the spectra were NIR and MIR, and they employed different numbers of samples for calibration. However, the main difference was in the nature of the compounds of interest: ethanol is a single compound, whereas the mannoproteins are a family of similar molecules with slight differences.

3.1. Material and methods

For each of the two applications, spectra $\mathbf{X}$ were centered, but not the reference values $\mathbf{y}$ which contained the raw values. Four models were calculated depending on the choice of the vector $\mathbf{r}$, see Table B.2. For the classical PLSR model m_{plsr}, $g(\mathbf{y}) = \mathbf{y}$, so $\mathbf{r}$ was set to $\mathbf{X}'\mathbf{y}$, which is exactly the same whether $\mathbf{y}$ is centered or not, provided $\mathbf{X}$ is centered. Two other functions were chosen, leading to the model m_{y^2} with $g(\mathbf{y}) = \mathbf{y}^2$ and $\mathbf{r} = \mathbf{X}'\mathbf{y}^2$, and model $m_{exp(y)}$ with $g(\mathbf{y}) = exp(\mathbf{y})$ and $\mathbf{r} = \mathbf{X}'exp(\mathbf{y})$. Finally, for model m_{nas}, $\mathbf{r} = \mathbf{s}_{nas}$ where $\mathbf{s}_{nas}$ is an estimation of the NAS. The root-mean square error of cross-validation ($RMSECV$) was calculated for each of the four models and used to determine the optimal number of latent variables, which were chosen among the lowest values and according to the shape of the $RMSECV$. All the calibration models were applied to the validation dataset $(\mathbf{X}_V, \mathbf{y}_V)$, then characterized by the root-mean square error of prediction ($RMSEP$) and the norm of the regression

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

vectors $\mathbf{b}$. Low $RMSEP$ values and low norms for $\mathbf{b}$ are expected for the best models.

3.1.1. Application 1: quantification of the ethanol concentration in fermenting musts

This application aimed at quantifying ethanol in wines using near-infrared spectroscopy. Spectra were acquired with a Jasco spectrophotometer (optical length 1 mm, wavelength range $500-2500$ nm, acquisition step 2 nm, water reference) at the Skalli-Fortant de France winery (Sète, France). The wavelength range was reduced to $500-1898$ nm such that $P = 700$ spectral variables, for reasons of compatibility with fiber optics. A vertical shift of baselines was performed such that each corrected absorbance at 1170 nm was null. The original data was split into two datasets: (1)$\mathbf{X}$: the N first 480 samples of fermenting musts, for calibration; and (2) $\mathbf{X}_V$: the last 1000 samples of fermenting musts, for validation. A subset of $\mathbf{X}$ was called $\mathbf{X}_m$ and contained 165 musts before fermentation, without ethanol. The reference values were expressed as a percent volume, e.g. 12% vol., and were measured with a WineScan MIR spectrophotometer (Foss) whose standard error of prediction was estimated to be 0.2%. The two datasets formed vectors $\mathbf{y}$ and $\mathbf{y}_V$ of dimensions (480×1) and (1000×1) respectively, and were associated with $\mathbf{X}$ and $\mathbf{X}_V$. These data were completed by the pure spectra of ethanol, water, glycerol and lactic acid $(\mathbf{k}_{etch}, \mathbf{k}_{water}, \mathbf{k}_{glyc}, \mathbf{k}_{lact})$ acquired against air using the same Jasco spectrophotometer.

Matrices $\mathbf{X}$ and $\mathbf{X}_V$ were centered around the row mean of $\mathbf{X}$. The pure

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

spectrum of ethanol yielded $\mathbf{k}$. A matrix $\mathbf{D}$ of dimensions (700×7) gathered all the influences to remove: the first four eigenvectors of a PCA performed onto $\mathbf{X}_m$; and the spectra $\mathbf{k}_{water}, \mathbf{k}_{glyc}$ and $\mathbf{k}_{lact}$. The NAS represented by the vector $\mathbf{s}_{nas}$ was calculated using $\mathbf{k}$ and $\mathbf{D}$, as described above. The $RMSECV$ and the $RMSEP$ were calculated from 1 to 20 latent variables.

3.1.2. Application 2: quantification of mannoproteins in wines

Wines produced from healthy grapes contain three main families of polysaccharides: arabinogalactan-proteins (AGPs), rhamno-galacturonan II (RG-II) and mannoproteins (MPs). We were interested in quantifying the MPs which impact on the physical and sensory properties of wines.

The calibration dataset of 40 samples was built with powder mixtures of four pure fractions: RG-II, neutral MP (MP0), neutral (AGP0) and acidic (AGP4) AGPs, as previously described [22]. The validation dataset consisted of powdered polysaccharide extracts from 65 wines. Spectra acquisition were performed with an Avatar 360 MIR spectrophotometer (Nicolet) equipped with an attenuated total reflectance cell and germanium crystal. The spectra in the range of $950 - 1850\ cm^{-1}$ were standardized to adjust the absorbance to 0 and 1 for the respective wavenumbers of $1850\ cm^{-1}$ and $1035\ cm^{-1}$, with the latter corresponding to the highest glucoside bond absorbance peak. MP reference values were deduced for the calibration dataset from the experimental design, and were obtained for the validation dataset by a chemical method involving hydrolysis and quantification by gas chromatography of the alditol acetates of the neutral sugars. The raw data were $\mathbf{X}$, (40×951), and $\mathbf{X}_V$, (65×951), with

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

the associated response variables $\mathbf{y}$ and $\mathbf{y}_V$. The NAS was estimated with $\mathbf{k}$, the pure spectrum of MP0, and $\mathbf{D}$, which contained the spectra of RG-II and seven different fractions of AGPs.

3.2. Results

3.2.1. Application one

Standard errors of cross validation and of prediction. According to the $RMSECV$ values, the optional numbers of latent variables chosen for models m_{plsr}, m_{nas}, m_{y^2} and $m_{exp(y)}$ were 10, 10, 5 and 10 respectively (Table B.3). The corresponding $RMSEP$ were similar: 0.92, 0.92, 0.92 and 0.90. The chosen number of latent variables was optimal for PLSR because it corresponded to the lower $RMSEP$. However this was not the case for the three other models, which presented their best $RMSEPs$ for several latent variables within a range equivalent to or slightly better than the best one of m_{plsr}. Thus, a small amount of error in the number of latent variables may be less serious for m_{nas}, m_{y^2} and $m_{exp(y)}$ than for m_{plsr}.

Comparison of the regression vectors $\mathbf{b}$. For each of the four models, the regression vectors leading to the lower $RMSEPs$ are presented in Figure B.3. It is noteworthy that the vectors of models m_{plsr} and m_{y^2} are very similar, but not equal; they seem to converge towards the same optimal solution. On the other hand, the b-coefficients vector for $m_{exp(y)}$ presents certain common peaks with the m_{plsr} and the m_{y^2} vectors, but its global shape is very different. The shape of the b-coefficients vector for m_{nas} has nothing in common with the three oth-

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

ers; but it is very close to the NAS (Figure B.3(b)). Thus three different (but equivalent in terms of prediction) solutions were identified by the four models.

The shape of $\mathbf{r}$ provides little information about the shape of the b-coefficients and the quality of prediction. For example, although $\mathbf{r}$ and $\mathbf{b}$ are very different for m_{plsr} and very similar for m_{nas}, these two models are nevertheless equivalent in terms of prediction errors.

The norms of the regression vectors for the four models and different latent variables are compared in Table B.4. In general, the norms increase with the number of latent variables. However, the standard PLSR model is the only model for which the norm increases steadily. For the three other models, the norm can decrease locally, for example m_{nas} and $m_{exp(y)}$ decrease between 10 and 11 latent variables. As a consequence, models with high numbers of latent variables can yield regression vectors of low norm, in contrast to standard PLSR.

3.2.2. Application two

Standard errors of cross validation and of prediction. The results are summarized in Table B.5. The $RMSECV$ values of the four models, calculated for 1 to 30 latent variables, decrease steadily. The choice for A should be close to 30, which is not reasonable for only 40 samples. Thus, the $RMSECV$ cannot be used here for the determination of A. However, $RMSEP$ values shows that good predictions were obtained with a low number of latent variables. In this example, the lowest $RMSEP$ for standard PLSR is 0.058 for seven latent variables. The model m_{nas} is difficult to compare to m_{plsr}, since it presents lower $RMSEP$ values but was obtained with a higher number of latent variables. On

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

the other hand, it is obvious that m_{y^2} and $m_{exp(y)}$ perform better than m_{plsr}: they present equivalent or lower $RMSEP$ values in a range of 12 latent variables, compared to the unique latent variable for m_{plsr} in the same range of $RMSEP$ values.

Comparison of the regression vectors. We compared the vector $\mathbf{X}'\mathbf{y}$ and the NAS to the regression vectors resulting from the four models, Figure B.4. Two spectral regions were identified: (1) $1200 - 950\ cm^{-1}$, in which the same peaks are found in all six examples, that is $\mathbf{X}'\mathbf{y}$, the NAS and the four regression vectors; and (2) $1850 - 1200\ cm^{-1}$, in which $\mathbf{X}'\mathbf{y}$ and the NAS have similar shapes, but the b-coefficients resemble noise near the baseline, with m_{plsr} being the most noisy.

In this application, the regression vectors tended towards a unique optimal solution, in contrast to Application one in which several solutions were identified. We suspect that the spectral information of each chemical compound in the Application two is more precise, specific peaks more localized within the spectrum and thus more easily identified in the MIR spectra than in the NIR spectra of Application one, leading to a unique solution. The relevant information in the $1200 - 950\ cm^{-1}$ region appears to be treated in the same way by the four models. Thus, the differences in their predictive abilities lie in the way noise is minimized in the $1850 - 1200\ cm^{-1}$ region.

We compared the norms of the regression vectors for the four models and the different latent variables (Table B.6) and found the comparisons to be similar to those made for Application one. Moreover, for the selected latent variables,

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

the regression vector of standard PLSR had the highest norm, which is not a good characteristic.

4. Discussion and conclusion

We have written a new algorithm for standard PLSR. We introduced a new parameter, a matrix $\mathbf{\Sigma}$ which has the position of a metric or pseudo-metric. We also dropped two parameters: the weights $\mathbf{W}$ and the loadings $\mathbf{c}$ for $\mathbf{y}$. We showed that the deflation of $\mathbf{X}$ into $\mathbb{R}^N$ can be replaced by a deflation into $\mathbb{R}^P$ by means of the loadings and the metric $\mathbf{\Sigma}$. These results have several consequences. Firstly, the extraction of the useful information from $\mathbf{X}$ is the $\mathbf{\Sigma}$ oblique *projection* of $\mathbf{X}$ onto the *latent structures* represented by the loadings, according to PLSR. The work of Phatak is completed by the identification of the metric and the space. Secondly, from a mathematical point of view, a metric and a basis of a subspace are independent elements of a vectorial space. Either $\mathbf{\Sigma}$ or $\mathbf{P}$ can be replaced into Eqs. A.6 and 10 by another metric or another matrix containing an other basis respectively, and the mathematics remain correct. This point specifically addresses the metric that attempts to weight the variables and to take into account their collinearity. The more comprehensive the observational data are within the space spanned by the spectra, the better the metric will be defined. Suppose that the observations associated with a response variable and represented by $\mathbf{X}$ have been extracted from a much larger database $\mathbf{X}_t$. For example, the whole database of a spectrometer that has been used for years for calibration is $\mathbf{X}_t$. Then, if a new parameter is cali-

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

brated, the response values are only available for a small subset of $\mathbf{X}_t$ which is $\mathbf{X}$. Since a much better representation of the space spanned by the spectra can be obtained from $\mathbf{X}_t$ than from $\mathbf{X}$, $\mathbf{\Sigma}_t = (\mathbf{X'}_t\mathbf{X}_t)^+$ is a better estimation of the metric to be applied than $\mathbf{\Sigma} = (\mathbf{X'X})^+$. Thus the PLSR should be built with the metric $\mathbf{\Sigma}_t$ rather than $\mathbf{\Sigma}$. This applies also to the cross-validation, which consists of extracting observations for prediction and calculating a model with the remaining observations. Instead of calculating a new metric at each loop, it seems mathematically more logical to determine a unique metric which is used independently from the set of observations being processed at any given time. This also would increase the calculation speed.

This new presentation of standard PLSR allows the deflation of $\mathbf{X}$ into $\mathbb{R}^P$, yielding a new algorithm for the building of the loadings. This puts forward an inner parameter, a vector $\mathbf{r}$ whose value is $\mathbf{X'y}$ for standard PLSR. We show that it is also a tool for orienting the calculation of the loadings, that is, the information contained within the spectra used by the model for prediction. The choice of $\mathbf{r} = \mathbf{X'y}$ is very likely to yield good predictions, but in certain cases other choices for $\mathbf{r}$ may be possible and yield VODKA regressions. Like standard PLSR, VODKA models use orthogonal scores and $\mathbf{\Sigma}$-orthogonal loadings. Among the first motivation of the authors of PLSR was to find relevant latent variables (orthogonal scores into $\mathbb{R}^N$, loadings into $\mathbb{R}^P$) that could explain an observation [23]. Our proposed modification aims at doing that also, but different choices for $\mathbf{r}$ can lead to dramatically different solutions, whereas the different PLSR algorithms tend to converge towards the same solution [7]. To

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

avoid confusion, the PLSR acronym has not been linked to VODKA, even if standard PLSR is one of its particular solutions.

Standard PLSR and VODKA both use orthogonal scores into $\mathbb{R}^N$ associated with orthogonal loadings into $\mathbb{R}^P$ with the metric $\mathbf{\Sigma}$ to identify latent variables relevant for prediction. However, these two methods are different in terms of means. When calculating the loadings by the new algorithm, standard PLSR seeks into $(\mathbb{R}^P, \mathbf{\Sigma})$ vectors of norm 1 whose inner product with $\mathbf{X}'_{1:i}\mathbf{y}$ is maximal; and Vodka seeks into $(\mathbb{R}^P, \mathbf{\Sigma})$ vectors of norm 1 whose inner product with a vector $\mathbf{r}_i$ is maximal, with $\mathbf{r}_i = \mathcal{P}'^{\perp}_{1:i}\mathbf{r}$. The results of the two methods can differ. Standard PLSR takes into account all the information within the observations which is correlated to the response variable; that is, direct information from the response variable itself, plus indirect information provided by other compounds correlated to the response variables. This property is a strength because good models are often obtained with the indirect information, and sometimes without any contribution by the direct information. It is also a weakness because unexpected modifications in the correlations among the different compounds will modify the predictions and lead to lower robustness. VODKA can be a solution to these situations. One strength is that it allows the introduction of external and selected knowledge to be introduced in the calculation process of the regression, and this enhances performance. In the two applications examined here, VODKA was able to produce models with prediction errors equivalent to or slightly lower than the errors associated with standard PLSR models. Perhaps the best result of VODKA has been to yield large ranges of the best prediction

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

errors, which implies that these models are more stable near the optimal number of latent variables.

In our first application of VODKA, we obtained three different models with comparable performances. This raises the question of the unicity of a best regression model. According to definition 1 of the NAS, one unique and best solution should exist. However definition 2 allows a wider interpretation and is more in accordance with the functioning of PLSR. If several compounds contribute to the prediction of the compound of interest, increasing the contribution of one will decrease the contribution of the others, so several equivalent solutions may exist. An other possible explanation lies in the nature of the spectra, for example NIR vs MIR, as suggested by the second application. However, we do not have enough elements or applications to identify the most likely hypothesis for these differences.

To conclude, VODKA regression provides an opportunity to take into account all the available information, not just that from the calibration dataset, allowing regression models to be built which can present some advantages over those produced by standard PLSR.

5. Acknowledgements

This work was made possible within the IRVIN program, was carried out with the Skalli winery and was supported by the Languedoc-Roussillon region. The authors thank all the people who contributed to this manuscript, and especially Steven Brown for his relevant advice.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

References

[1] H.Wold, in: Multivariate analysis, Krishnaiah (ed), Academic Press, London, 1966.

[2] S.Wold, H.Martens, H.Wold, The multivariate calibration method in chemistry solved by the pls model, in: Proc. conf. matrix pencils, lecture notes in mathematics, Springler-Velay, Heildelberg, A.Ruhe, B.Kagstrom, 1983, pp. 286–293.

[3] S.DeJong, Simpls: an alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.

[4] A.Phatak, S.DeJong, The geometry of partial least squares, J. Chemom. 11 (1997) 311–338.

[5] H.Martens, T.Naes, in: Near infrared technology in the agricultural and food industries, Williams-Norris (eds), Amer. Assn. of Cereal Chemists, St Paul, 1987.

[6] H.Martens, T.Naes, Multivariate Calibration, Wiley, Chichester, 1989.

[7] M.Andersson, A comparison of nine pls1 algorithms, J. Chemom. 23 (2009) 518–529.

[8] A.Hoskuldsson, The h principle: new ideas, algorithms and methods in applied statistics and physics, Chemom. Intell. Lab. Syst. 23 (1994) 1–28.

[9] B.Nadler, R.R.Coifman, Partial least squares, beer's law and the net analyte signal: statistical modeling and analysis, J.Chemom. 19 (2005) 45–54.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

[10] A.Hoskuldsson, Pls regression methods, J.Chemom. 2 (1988) 211–228.

[11] P.Geladi, B.Kowalski, Partial least squares regression: a tutorial, Anal.Chim.Acta 185 (1986) 1–17.

[12] I.S.Helland, Partial least squares regression and statistical models, Scand. J. of Statist. 17 (1990) 97–114.

[13] R.J.Pell, L.S.Ramos, R.Manne, The model space in partial least squares regression, J. Chemom. 21 (2007) 165–172.

[14] R.Ergon, Re-interpretation of nipals results solves plsr inconsistency problem, J. Chemom. 23(2) (2009) 72–75.

[15] R.DeMaesschlack, D.JouanRimbaud, D.L.Massart, The mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000) 1–18.

[16] R.Manne, R.J.Pell, L.S.Ramos, The pls model space: the inconsistency persists, J.Chemom. 23 (2009) 76–77.

[17] S.Wold, M.Hoy, H.Martens, J.Trygg, F.Westad, J.MacGregor, B.M.Wise, The pls model space revisited, J.Chemom. 23 (2009) 67–68.

[18] J.Trygg, Parsimonious multivariate models, Ph.D. thesis, Umea University, Sweden (2001).

[19] A.Lorber, N.K.M.Faber, B.R.Kowalski, Net analyte signal calculation in multivariate calibration, Anal. Chem. 69(8) (1997) 1620–1626.

[20] R.Marbach, A new method for multivariate calibration., Journal of Near Infrared Spectroscopy 13 (2005) 241–254.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

[21] J.Ferre, N.K.M.Faber, Net analyte signal calculation for multivariate calibration, Chemom. Intell. Lab. Syst. 69 (2003) 123–136.

[22] J.C.Boulet, T.Doco, J.M.Roger, Improvement of calibration models using two successive orthogonal projection methods, application to quantification of wine polysaccharides, Chemom. Intell. Lab. Syst. 87 (2007) 295–302.

[23] S.Wold, Personal memories of the early pls development, Chemom. Intell. Lab. Syst. 58 (2001) 83–84.

[24] M.Tenenhaus, La régression PLS, Technip, Paris, 1998.

[25] B.S.Dayal, J.F.MacGregor, Improved pls algorithms, J. Chemom. 11 (1997) 73–85.

Appendix A. Properties of standard PLSR

Several properties of standard PLSR are recalled or developed. They are logically ordered to demonstrate the seventh and last property, each of them relying on the previous ones.

- Property 1: the projection of the $\mathbf{t}_i$ onto the space spanned by the columns of $\mathbf{X}$

  The $\mathbf{t}_i$ belong to the subspace of $\mathbb{R}^N$ spanned by the columns of $\mathbf{X}$ [10]. Thus their orthogonal projection onto $\mathbf{X}$ is invariant:

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{t}_i = \mathbf{t}_i \tag{A.1}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

- Property 2: the relationship between $(\mathbf{t'}_i\mathbf{t}_i)$ and $(\mathbf{p'}_i\mathbf{\Sigma}\mathbf{p}_i)$

Equation 5 can be simplified [24] leading to:

$$\mathbf{p}_i = \mathbf{X'}\mathbf{t}_i(\mathbf{t}_i'\mathbf{t}_i)^{-1} \tag{A.2}$$

Thus:

$$\mathbf{p'}_i\mathbf{\Sigma}\mathbf{p}_i = (\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{t'}_i\mathbf{X}(\mathbf{X'X})^{+}\mathbf{X'}\mathbf{t}_i(\mathbf{t}_i'\mathbf{t}_i)^{-1}$$

$$\mathbf{p'}_i\mathbf{\Sigma}\mathbf{p}_i = (\mathbf{t'}_i\mathbf{t}_i)^{-1} \tag{A.3}$$

In his presentation of SIMPLS [3], De Jong wrote a similar equation: $\mathbf{P'\Sigma P} = \mathbf{T'T} = \mathbf{I}_A$, which was developed under the hypothesis that the $\mathbf{t}_i$ had been normed and thus, for this particular case, $(\mathbf{t'}_i\mathbf{t}_i)^{-1} = (\mathbf{t'}_i\mathbf{t}_i) = 1$. However, for standard PLSR and current versions of SIMPLS for which the scores are not normed, only Equation A.3 is valid.

- Property 3: expression of $\mathbf{t}_i$ in terms of $\mathbf{X}$, $\mathbf{p}_i$ and $\mathbf{\Sigma}$

Each term of Equation A.2 is multiplied on the left by $\mathbf{X\Sigma}$, that is $\mathbf{X}(\mathbf{X'X})^{+}$; the terms are permuted and simplified according to Equation A.1 and then Equation A.3:

$$\mathbf{t}_i \quad = \quad \mathbf{X\Sigma}\mathbf{p}_i(\mathbf{p'}_i\mathbf{\Sigma}\mathbf{p}_i)^{-1} \tag{A.4}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

- Property 4: $\mathbf{\Sigma}$-orthogonality of the $\mathbf{p}_i$

For $i \neq j$ [10]:

$$\mathbf{p'}_i \mathbf{\Sigma} \mathbf{p}_j = 0 \tag{A.5}$$

Moreover the matrix $\mathbf{P'\Sigma P}$ is diagonal and the nonzero term at the i^{eme} line and i^{eme} column is $\mathbf{p'}_i \mathbf{\Sigma} \mathbf{p}_i$.

- Property 5: expression of $\mathbf{T}$ in terms of $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{\Sigma}$

Because of Property 4, Equation A.4 leads to the Property 5:

$$\mathbf{T} = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1} \tag{A.6}$$

- Property 6: expression of the b-coefficients using $\mathbf{X}$, $\mathbf{y}$, $\mathbf{\Sigma}$ and $\mathbf{P}$

$\widehat{\mathbf{y}}$ is obtained after an orthogonal projection of $\mathbf{y}$ onto $\mathbf{T}$ [4], then the value of $\mathbf{T}$ is replaced by its value from Equation A.6 and the expression simplified:

$$\widehat{\mathbf{y}} = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'\Sigma X'y}$$

Equation 10 is straightforward.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

- Property 7: The relationship between $\mathbf{X}\mathcal{P}_{1:i}^{\perp}$ and $\mathcal{T}_{1:i}^{\perp}\mathbf{X}$

Let $\mathcal{P}_{1:i}^{\perp}$ be the oblique $\mathbf{\Sigma}$ anti-projector to $\mathbf{P}_{1:i}$, and $\mathcal{T}_{1:i}^{\perp}$ the orthogonal anti-projector to $\mathbf{T}_{1:i}$. Due to Property 4:

$$\mathcal{P}_{1:i}^{\perp} = \mathbf{I}_P - \sum_{k=1}^{k=i} \mathbf{\Sigma}\mathbf{p}_k (\mathbf{p'}_k \mathbf{\Sigma}\mathbf{p}_k)^{-1} \mathbf{p'}_k$$

The matrix $\mathbf{X}_{1:i}$ can be written by means of two expressions. From Equation 6, it is obvious that $\mathbf{X}_{1:i} = \mathcal{T}_{1:i}^{\perp}\mathbf{X}$. Using the values of $\mathbf{t}_i$ from Equation A.4, it is possible to substitute the $\mathbf{t}_i$ into Equation 6. Thus, a new expression of $\mathbf{X}_{1:i}$ is deduced:

$$\begin{aligned}
\mathbf{X}_{1:i} &= \mathbf{X}_{1:i-1} - \mathbf{X}\mathbf{\Sigma}\mathbf{p}_i (\mathbf{p'}_i \mathbf{\Sigma}\mathbf{p}_i)^{-1} \mathbf{p'}_i \\
&= \mathbf{X} - \sum_{k=1}^{k=i} \mathbf{X}\mathbf{\Sigma}\mathbf{p}_k (\mathbf{p'}_k \mathbf{\Sigma}\mathbf{p}_k)^{-1} \mathbf{p'}_k \\
&= \mathbf{X}\mathcal{P}_{1:i}^{\perp}
\end{aligned}$$

Finally:

$$\mathbf{X}_{1:i} = \mathcal{T}_{1:i}^{\perp}\mathbf{X} = \mathbf{X}\mathcal{P}_{1:i}^{\perp} \tag{A.7}$$

The anti-projection of $\mathbf{X}$ orthogonally to $\mathbf{T}$ into $\mathbb{R}^N$ gives the same result as its oblique $\mathbf{\Sigma}$ anti-projection to $\mathbf{P}$ into $\mathbb{R}^P$.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Appendix B. New calculation of the $\mathbf{p}_i$ into $\mathbb{R}^P$ from standard PLSR

The deflation of $\mathbf{y}$ is not necessary when $\mathbf{X}$ is deflated [25], so equation 1 can first be simplified and then written using $\mathcal{T}_{1:i}^{\perp}$:

$$\begin{aligned} \mathbf{w}_{i+1} &= \mathbf{X}'_{1:i}\mathbf{y} \\ &= \mathbf{X}'\mathcal{T}_{1:i}^{\perp}\mathbf{y} \end{aligned}$$

The combination of Equations 1, 3 and 5 from standard PLSR plus Equation A.7 leads to:

$$\mathbf{p}_{i+1} = \alpha_{i+1}\mathbf{X}'\mathcal{T}_{1:i}^{\perp}\mathcal{T}_{1:i}^{\perp}\mathbf{X}\mathbf{X}'\mathcal{T}_{1:i}^{\perp}\mathbf{y} \quad \text{(B.1)}$$

$$= \alpha_{i+1}\mathbf{X}'\mathcal{T}_{1:i}^{\perp}\mathbf{X}\mathbf{X}'\mathcal{T}_{1:i}^{\perp}\mathbf{y} \quad \text{(B.2)}$$

$$= \alpha_{i+1}\mathcal{P}'^{\perp}_{1:i}\mathbf{X}'\mathbf{X}\mathcal{P}'^{\perp}_{1:i}\mathbf{X}'\mathbf{y} \quad \text{(B.3)}$$

with α_{i+1} a nonzero scalar associated to $\mathbf{p}_{i+1}$. The value of α_{i+1} is not important because it is simplified into Equation 10, but vectors $\mathbf{p}_i$ should have small norms to improve the stability of the calculation. For this reason, the $\mathbf{p}_i$ are $\mathbf{\Sigma}$-normalized to 1 such that they form a $\mathbf{\Sigma}$-orthonormal basis of the subspace of $\mathbb{R}^P$ containing the useful information. The new algorithm for the calculation of the $\mathbf{p}_i$ is thus written:

- Step 1:

$$\begin{aligned} \mathbf{p}_1 &= \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} \\ \mathbf{p}_1 &= \mathbf{p}_1(\mathbf{p}'_1\mathbf{\Sigma}\mathbf{p}_1)^{-0.5} \end{aligned}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

- Step $i+1$:

$$\begin{aligned} \mathbf{p}_{i+1} &= {\mathcal{P}'}^{\perp}_{1:i}\mathbf{X}'\mathbf{X}{\mathcal{P}'}^{\perp}_{1:i}\mathbf{X}'\mathbf{y} \\ \mathbf{p}_{i+1} &= \mathbf{p}_{i+1}(\mathbf{p'}_{i+1}\mathbf{\Sigma}\mathbf{p}_{i+1})^{-0.5} \end{aligned}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

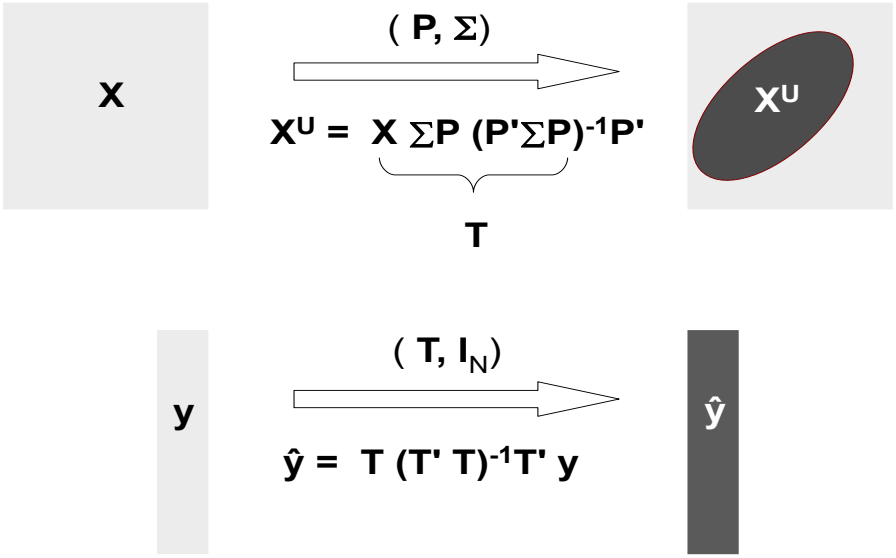

Figure B.1: Standard PLSR as a double projection. The upper panel shows an oblique $\mathbf{\Sigma}$-projection of $\mathbf{X}$ onto $\mathbf{P}$ yielding $\mathbf{X}^U$ and its scores $\mathbf{T}$. The lower panel shows an orthogonal projection of $\mathbf{y}$ onto $\mathbf{T}$ yielding $\widehat{\mathbf{y}}$.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

$\mathbf{X}$	matrix $(N \times P)$, N observations and P spectral variables
$\mathbf{y}$	vector $(N \times 1)$, the response variable
$\mathbf{X}_{1:i}$	anti-projection of $\mathbf{X}$ orthogonally to $\{ \mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_i\}$
$\mathbf{y}_{1:i}$	anti-projection of $\mathbf{y}$ orthogonally to $\{ \mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_i\}$
$\mathbf{T}$	matrix $(N \times A)$, scores for $\mathbf{X}$
$\mathbf{P}$	matrix $(P \times A)$, loadings for $\mathbf{X}$
$\mathbf{T}_{1:i}$, $\mathbf{P}_{1:i}$	matrices containing the i first columns of $\mathbf{T}$ and $\mathbf{P}$
$\mathbf{W}$	matrix $(P \times A)$, weights for $\mathbf{X}$
$\mathbf{c}$	vector $(A \times 1)$, loadings $c_1, c_2, ...c_A$ for $\mathbf{y}$
$\mathbf{\Sigma}$	Moore-Penrose pseudo-inverse of $(\mathbf{X}'\mathbf{X})$
$\mathbf{I}_N$, $\mathbf{I}_P$	identity matrices for $\mathbb{R}^N$ and $\mathbb{R}^P$ spaces
$\mathbf{r}$	vector $(P \times 1)$
$\mathcal{T}_{1:i}^{\perp}$	$(N \times N)$ orthogonal anti-projector to $\{ \mathbf{t}_1, \mathbf{t}_2, , ...\mathbf{t}_i\}$
$\mathcal{P}_{1:i}^{\perp}$	$(P \times P)$ oblique anti-projector to $\{ \mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_i\}$ with the metric $\mathbf{\Sigma}$
$\mathbf{t}_i$	score vector at step i of standard PLSR; also i^{th} column vector of $\mathbf{T}$
$\mathbf{p}_i$	loading vector at step i of standard PLSR; also i^{th} column vector of $\mathbf{P}$
$\mathbf{w}_i$	weight vector at step i of standard PLSR; also i^{th} column vector of $\mathbf{W}$
$\mathbf{b}$	regression vector, or vector of b-coefficients, for A latent variables
A	number of latent variables; also dimension of the PLSR model and rank of $\mathbf{T}$,$\mathbf{W}$ and $\mathbf{P}$

Table B.1: Main notations

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

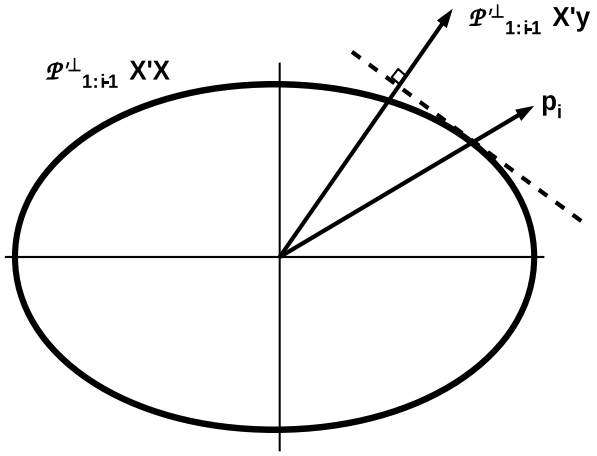

Figure B.2: Geometric building of the loadings $\mathbf{p}_i$ using standard PLSR (from Figure 13 of Phatak [4])

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Model	Data	Reference values	$\mathbf{r}$
m_{plsr}	$\mathbf{X}$	$\mathbf{y}$	$\mathbf{X}'\mathbf{y}$
m_{nas}	$\mathbf{X}$	$\mathbf{y}$	$\mathbf{s}_{nas}$
m_{y^2}	$\mathbf{X}$	$\mathbf{y}$	$\mathbf{X}'\mathbf{y}^2$
$m_{exp(y)}$	$\mathbf{X}$	$\mathbf{y}$	$\mathbf{X}'exp(\mathbf{y})$

Table B.2: The four models obtained using VODKA regression with different values for $\mathbf{r}$.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Latent	$RMSECV$				$RMSEP$			
variables	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$
4	0.96	1.07	0.68	1.36	1.76	1.99	1.04	2.59
5	0.89	0.99	0.61	1.25	1.50	1.90	**0.92**	2.31
6	0.73	0.90	0.62	1.22	1.20	1.76	**0.91**	2.17
7	0.68	0.72	0.61	1.12	1.06	**0.87**	**0.90**	2.22
8	0.67	0.68	0.58	1.02	0.99	**0.89**	0.96	2.19
9	0.64	0.64	0.57	0.72	0.93	**0.92**	0.96	1.11
10	0.56	0.56	0.56	0.63	**0.92**	**0.92**	0.94	**0.90**
11	0.56	0.56	0.56	0.61	0.95	0.94	0.94	**0.89**
12	0.56	0.56	0.55	0.57	1.02	0.99	0.96	0.93

Table B.3: Application 1: The standard errors of cross-validation ($RMSECV$) and prediction ($RMSEP$) for the four models and a range including the lower latent variables. $RMSEP$ values less than or equal to 0.92 are represented in bold.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Latent variables	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$
4	52.8	137.7	51.6	50.7
5	56.3	114.3	54.2	50.1
6	63.7	105.7	57.9	50.9
7	68.8	**117.8**	**81.9**	51.3
8	70.8	190.2	94.3	64.7
9	76.7	189.4	100.1	144.5
10	**98.4**	204.9	98.8	99.3
11	116.0	147.9	110.8	**95.2**
12	152.2	206.8	142.4	139.9

Table B.4: Application 1: The norms of the regression vectors $\mathbf{b}$ for models m_{plsr}, m_{nas}, m_{y2} and $m_{exp(y)}$ of Application 1. Values corresponding to the selected number of latent variables are shown in bold

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

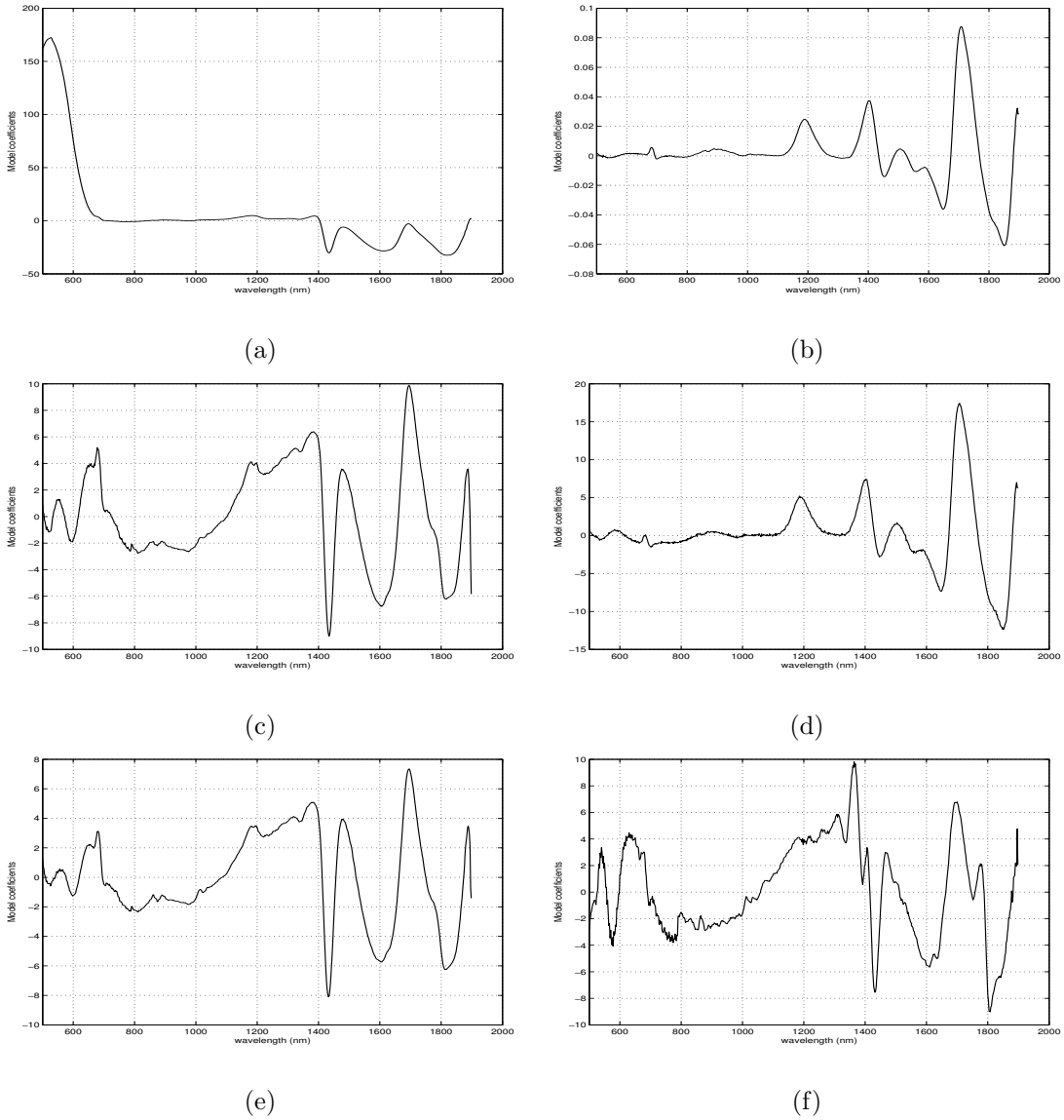

Figure B.3: Spectra and b-coefficient vectors for Application 1. Spectra of $\mathbf{X}'\mathbf{y}$ (a) and $\mathbf{s}_{nas}$ (b). The coefficient vectors for models m_{plsr} (c), m_{nas} (d), m_{y2} (e) and $m_{exp(y)}$ (f) were calculated with 10, 7, 7 and 11 latent variables, respectively.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Latent	$RMSECV$				$RMSEP$			
variables	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$
6	0.076	0.059	0.056	0.069	0.073	0.15	0.17	0.069
7	0.083	0.060	0.048	0.051	**0.058**	0.19	**0.058**	**0.050**
8	0.053	0.059	0.051	0.062	0.076	0.15	**0.058**	**0.050**
9	0.049	0.049	0.048	0.041	0.079	0.12	**0.051**	**0.049**
10	0.042	0.047	0.047	0.046	0.090	0.089	0.059	**0.047**
11	0.039	0.040	0.041	0.038	0.088	0.084	0.059	**0.051**
12	0.039	0.042	0.042	0.043	0.096	0.071	**0.054**	**0.050**
13	0.037	0.039	0.042	0.043	0.095	0.061	**0.053**	**0.050**
14	0.035	0.039	0.035	0.036	0.096	0.060	**0.053**	**0.045**
15	0.029	0.041	0.035	0.033	0.095	0.059	**0.053**	**0.046**
16	0.029	0.033	0.031	0.030	0.094	0.059	**0.050**	**0.049**
17	0.029	0.036	0.029	0.029	0.094	**0.056**	**0.052**	**0.056**
18	0.027	0.028	0.029	0.028	0.094	**0.054**	**0.054**	0.059
19	0.027	0.030	0.028	0.027	0.094	**0.057**	0.065	**0.058**

Table B.5: The standard errors of cross-validation ($RMSECV$) and prediction ($RMSEP$) for the four models and a range including the lower latent variables of Application 2. $RMSEP$ values less than or equal to 0.058 are represented in bold.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

Latent variables	m_{plsr}	m_{nas}	m_{y^2}	$m_{exp(y)}$
6	1.44	1.31	1.16	1.15
7	**1.52**	1.14	1.17	1.15
8	1.61	1.21	1.22	1.19
9	1.71	1.23	**1.25**	1.20
10	1.77	1.21	1.24	**1.25**
11	1.8	1.21	1.28	1.24
12	1.85	1.25	1.30	1.24
13	1.86	1.28	1.31	1.32
14	1.86	1.34	1.30	1.36
15	1.86	1.46	1.31	1.34
16	1.86	1.42	1.39	1.50
17	1.86	1.42	1.41	1.52
18	1.86	**1.42**	1.42	1.52
19	1.86	1.44	1.41	1.53

Table B.6: The norms of the regression vectors $\mathbf{b}$ for models m_{plsr}, m_{nas}, m_{y2} and $m_{exp(y)}$ of Application 2. Values corresponding to the selected number of latent variables are shown in bold.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2013, 120, 116-125.
The original publication is available at http://www.sciencedirect.com
DOI : 10.1016/j.chemolab.2012.11.002

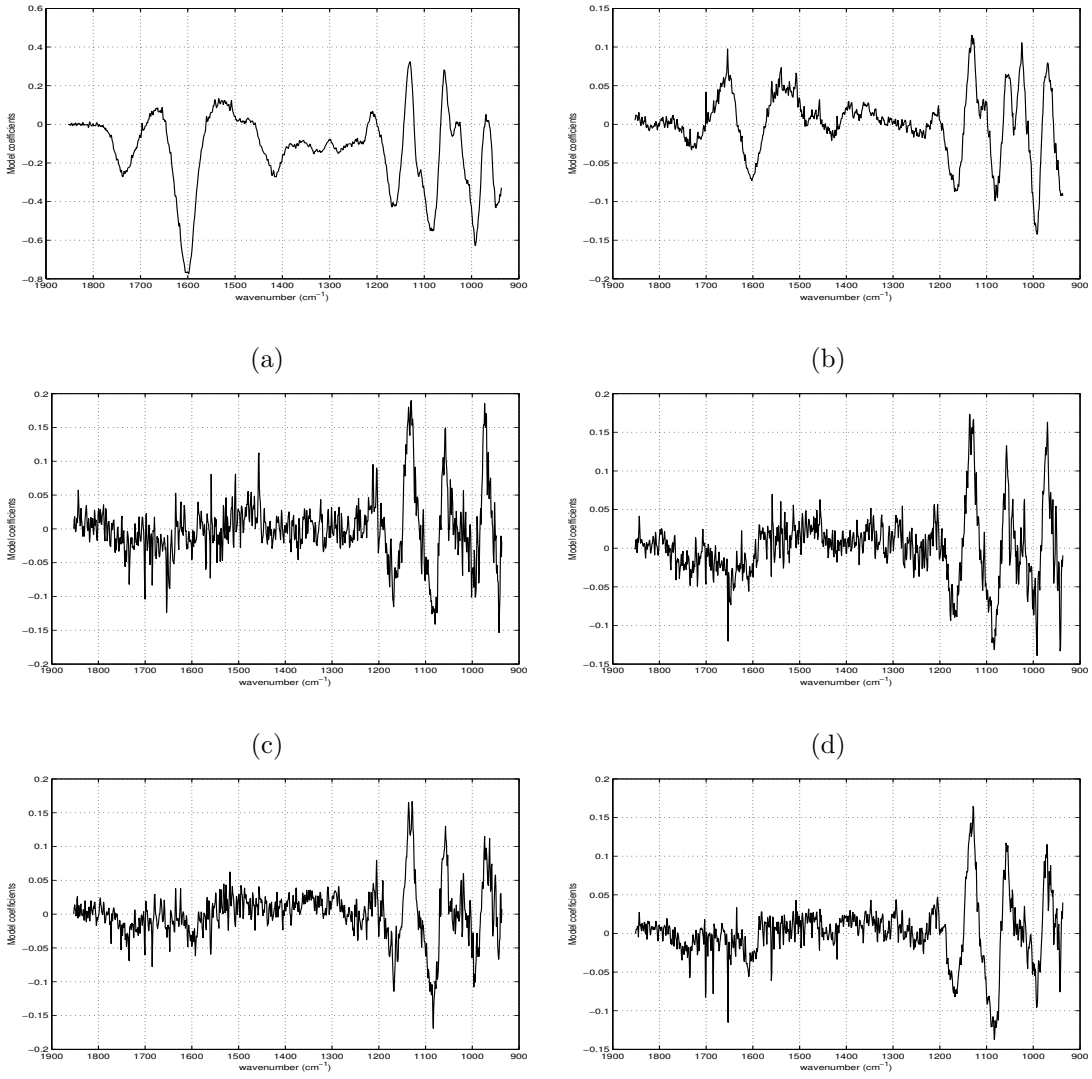

Figure B.4: Spectra and b-coefficient vectors for Application 2. Spectra of $\mathbf{X}'\mathbf{y}$ (a) and $\mathbf{s}_{nas}$ (b). The b-coefficient vectors for models m_{plsr} (c), m_{nas} (d), m_{y2} (e) and $m_{exp(y)}$ (f) were calculated with 7, 18, 9 and 10 latent variables, respectively.