ARTICLE IN PRESS

CHEMOM-02660; No of Pages 11

Chemometrics and Intelligent Laboratory Systems xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Assessing feature relevance in NPLS models by VIP

Stefania Favilla [a], Caterina Durante [b], Mario Li Vigni [b], Marina Cocchi [b,*]

[a] Department of Biomedical Sciences, Metabolic and Neuroscience, University of Modena and Reggio Emilia, Italy
[b] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Received 9 January 2013
Received in revised form 17 April 2013
Accepted 26 May 2013
Available online xxxx

Keywords:
VIP
Multi-way data
NPLS
NPLS-DA
Feature selection

ABSTRACT

Multilinear PLS (NPLS) and its discriminant version (NPLS-DA) are very diffuse tools to model multi-way data arrays. Analysis of NPLS weights and NPLS regression coefficients allows data patterns, feature correlation and covariance structure to be depicted. In this study we propose an extension of the Variable Importance in Projection (VIP) parameter to multi-way arrays in order to highlight the most relevant features to predict the studied dependent properties either for interpretative purposes or to operate feature selection. The VIPs are implemented for each mode of the data array and in the case of multivariate dependent responses considering both the cases of expressing VIP with respect to each single y-variable and of taking into account all y-variables altogether.

Three different applications to real data are presented: i) NPLS has been used to model the properties of bread loaves from near infrared spectra of dough, acquired at different leavening times, and corresponding to different flour formulations. VIP values were used to assess the spectral regions mainly involved in determining flour performance; ii) assessing the authenticity of extra virgin olive oils by NPLS-DA elaboration of gas chromatography/mass spectrometry data (GC–MS). VIP values were used to assess both GC and MS discriminant features; iii) NPLS analysis of a fMRI-BOLD experiment based on a pain paradigm of acute prolonged pain in healthy volunteers, in order to reproduce efficiently the corresponding psychophysical pain profiles. VIP values were used to identify the brain regions mainly involved in determining the pain intensity profile.

1. Introduction

Multilinear PLS (NPLS) and its discriminant version (NPLS-DA) are very diffuse tools to model multi-way data arrays. NPLS represents the multi-way extension of two-way partial least squares regression (PLS) for multi-way data and was first developed by Bro in 1996 [1] and successively by Bro, Smilde and De Jong [2–4].

It has been demonstrated that multi-way data analysis tools, taking into account the multi-way structure of data are much more efficient compared to unfolding procedures, that is re-arranging the multi-way data into a two-way matrix structure and then applying bilinear models. Multi-way analysis allows simplifying the interpretation of the results and providing more adequate and robust models using relatively few parameters [5,6]. While this is true in general, it is worth noticing that when dealing with real-time monitoring, e.g. in batch process monitoring, N-way models may not represent a real advantage with respect to adopting a proper unfolding/refolding procedure as by using Multiway-PCA [7].

In particular, the use of NPLS shares all the advantages of latent variable based regression and discrimination methods, from the point of view of data visualization and interpretation [8–10]. In fact, analysis of NPLS weights and NPLS regression coefficients allows data patterns, feature correlation and covariance structure to be depicted.

However, it is often needed to define which are the most relevant features to predict the studied dependent properties either for interpretative purposes, e.g. to provide a better understanding of the underlying process that generated the data, or to operate feature selection in order to reduce the noise generated by irrelevant features or to reduce data redundancy.

Some of the several variable selection methods applied to two-way data matrices in the context of PLS regression [11], such as interval PLS (iPLS) [12] or genetic algorithms [13], can be as well suited for NPLS if the **X**-block multi-way data array has only one spectral dimension, e.g. samples × spectral profiles × time [14]. When the data array has two spectral dimensions, or more generally when a two-dimensional signal map characterizes each sample, as generated by hyphenated analytical techniques, such as emission/excitation fluorescence, chromatography/mass spectrometry, etc., these variable selection methodologies present significant challenges and it is suggested to apply them after unfolding the data array [15]. However, in this way, the multi-way data structure is not taken into account in the variable selection step, thus loosing the multi-way analysis advantage.

Moreover, a general distinction can be made among tools, which accomplish feature selection by deleting a set of features and re-assessing the performance of the reduced models, thus requiring extensive model

* Corresponding author. Tel.: +39 059 2055029; fax: +39 059 373543.
E-mail address: marina.cocchi@unimore.it (M. Cocchi).

0169-7439/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.chemolab.2013.05.013

validation, and those that operate a ranking of the features according to their relevance. Concerning the latter type, congruence loadings [16], VIP (Variable Importance in Projection) [17–20], and selectivity ratio [21] have gained increasing attention as an important measure of each explanatory variable or predictor.

The aim of this work is to extend the VIP method to multi-way arrays and to develop accompanying code. A similar attempt has been reported previously [22] but our formulation is, in our opinion, more straightforward and closer to VIP definition for two-way data. In fact, the VIP definition given in Ref. [22] does not reproduce the two-way VIP formulation, which consists, for each X-variable, of a sum, over latent variables, of its PLS-weight weighted by the percentage of explained Y variance. The reason is due to the fact that the NPLS mode 1 scores for **X**-block (T), which are linked to Y by the NPLS inner-relation, are substituted by a different projection of unfolded-**X** through NPLS weights. Usually, in the case of several dependent responses (multivariate Y) VIP is defined taking into account all y-variables altogether. Here we consider as well, the possibility of expressing VIP with respect to each single y-variable (this is a further difference with the approach presented in Ref. [22] that does not allow this possibility). This offers higher flexibility to the method and can be particularly useful to interpret discriminant NPLS-DA models, since the VIP for single y-variables corresponds, in this case, to the most discriminant feature for each category. However, it is beyond the aim of this paper, to compare the use of VIPs with other feature selection methodologies for multi-way arrays, actually in the applications presented here variable selection is not operated and VIPs are used more on an interpretative ground.

2. Methods

2.1. Multilinear partial least squares (NPLS)

Multilinear PLS (NPLS) represents the extension of two-way partial least squares regression (PLS) to data arrays of any order considering both X and Y-blocks. In the following, the method is described considering the case of a three-way data array, $\underline{\mathbf{X}}$, but the extension to further dimensions can be simply deduced. As for the dependent variables Y-block, we will describe here the case of a two-way matrix, but the method can be easily extended to higher orders in the Y-block [1].

Specifically, PLS regression aims to find a relationship between a set of predictor (independent) data, $\underline{\mathbf{X}}$, and a set of responses (dependent), $\underline{\mathbf{Y}}$. In the more general case, the arrays of independent, $\underline{\mathbf{X}}$ and dependent $\underline{\mathbf{Y}}$ variables are decomposed in such a way that the score vectors from these models have pair-wise maximal covariance [3,4]. Multilinear PLS was firstly developed as a PARAFAC-like model of $\underline{\mathbf{X}}$ and it was shown that the method could be easily extended to any desired order for both $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ arrays. This method was further elaborated and lastly improved with respect to residual analyses by introducing a core array in the model of $\underline{\mathbf{X}}$ [2].

Considering an $\underline{\mathbf{X}}$ array of dimension $I \times J \times K$, the NPLS model is obtained by modeling $\underline{\mathbf{X}}$ as in Tucker3 decomposition:

$$\mathbf{X} = \mathbf{T}\mathbf{G}_{\mathrm{X}}\left(\mathbf{W}^{K} \otimes \mathbf{W}^{J}\right)^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}} \tag{1}$$

where $\mathbf{X}$ is the $\underline{\mathbf{X}}$ array unfolded to an $I \times JK$ matrix, $\mathbf{T}$ holds the first mode scores (sample mode), $\mathbf{W}^{J}$ and $\mathbf{W}^{K}$ are the second and the third mode weights, respectively. The symbol ⊗ denotes the Kronecker product [5].

$\mathbf{G}_{\mathrm{X}}$ is the matricized core array of size $F \times F \times F$ where F is the number of NPLS components (factors) and it is defined by:

$$\mathbf{G}_{\mathrm{X}} = \mathbf{T}^{+}\mathbf{X}\left(\left(\mathbf{W}^{K}\right)^{+} \otimes \left(\mathbf{W}^{J}\right)^{+}\right)^{\mathrm{T}}. \tag{2}$$

Here the superscript '+' means that the Moore–Penrose is pseudo inverse.

In the case of a two-way data matrix, $\mathbf{Y}_{\mathrm{I,M}}$ is defined by:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q} + \mathbf{E}_{Y} \tag{3}$$

where $\mathbf{U}$ holds the $\mathbf{Y}$ scores and $\mathbf{Q}$ is the loading matrix.

$\underline{\mathbf{E}}_{\mathrm{X}}$ and $\mathbf{E}_{\mathrm{Y}}$ hold $\underline{\mathbf{X}}$ and $\mathbf{Y}$ residuals, respectively. In analogy with the two-way PLS algorithm, the weights are determined such that the scores obtained from the $\underline{\mathbf{X}}$ decomposition ($\mathbf{T}$) have maximum covariance with the scores obtained from $\mathbf{Y}$ decomposition (inner relation: $\mathbf{U} = \mathbf{T}\mathbf{B} + \mathbf{E}_{\mathrm{U}}$).

By regressing the data onto their weights vectors, a score vector is found in the $\underline{\mathbf{X}}$-space providing a least squares model of the $\underline{\mathbf{X}}$ data. Furthermore, by choosing the weights such that the covariance between $\underline{\mathbf{X}}$ and $\mathbf{Y}$ is maximized a predictive model is obtained as:

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q} + \mathbf{E}_{\mathrm{Y}}. \tag{4}$$

Regression coefficients that apply directly to $\mathbf{X}(I \times JK)$ may also be derived [4,22]:

$$R = \left[w_1 \left(I - w_1 {w_1}^{T}\right) w_2 \ldots \prod_{f=1}^{F-1} \left(I - w_f {w_f}^{T}\right) w_f\right] \tag{5}$$

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{R}\mathbf{B}\mathbf{1} \tag{6}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}}. \tag{7}$$

The NPLS-DA formulation is the same but the dependent variable block is a matrix $\mathbf{Y}$ holding the class information, i.e. for each category a y-variable is defined as a dummy variable assuming values one/minus one to indicate class membership or not (notation one/zero is also used). As the predicted y-values can assume real values and not only minus one and one, classification of the samples is accomplished by assigning the sample to the category corresponding to the highest value of the predicted response, i.e. if the predicted vector of responses for an unknown sample, is: [−0.5 0.8 0.5] (in the case of three classes problem), it will be assigned to class two.

2.2. VIP calculation

2.2.1. Two-way case

The variable importance in the projection (VIP) [17,19] represents the influence of each variable j of the data matrix $\mathbf{X}_{I,J}$ on the model of the responses matrix $\mathbf{Y}_{I,M}$

$$\mathrm{VIP}^{2}{}_{j} = \mathrm{S}_{f} w^{2}{}_{jf} \cdot \mathrm{SSY}_{f} \cdot J / \left(\mathrm{SSY}_{\mathrm{tot.expl.}} \cdot F\right) \tag{8}$$

where, F is the number of latent variables of the PLS model and J the number of $\mathbf{X}$ variables.

In the case of mono-dimensional $\mathbf{y}^{I \times 1}$ holds:

$$\mathrm{SSY}_{f} = {b_{f,f}}^{2} \mathbf{t}_{f}{}^{\mathrm{T}} \mathbf{t}_{f} \qquad \mathrm{SSY}_{\mathrm{tot.expl.}} = \mathbf{b}^{2}\mathbf{T}^{\mathrm{T}}\mathbf{T} \tag{9}$$

where $\mathbf{T}$ is the $\mathbf{X}$ score matrix and $\mathbf{b}$ the PLS inner relation coefficients.

Thus a VIP value for each variable is computed in order to quantify its importance by using the PLS weight w_{jf} weighted by how much of $\mathbf{y}$ is explained in each model dimension (latent variable).

VIP formulation as originally proposed [17] is intended to be a parameter varying in a fixed range since the sum of squared VIP for all variables is the sum to the number of variables. Thus, the variables with a VIP value larger than 1 (i.e. larger than the average of square VIP values) have an above average influence on the model and are, therefore, considered the most relevant for explaining $\mathbf{Y}$. The choice of the VIP threshold to assess the salient variables is a critical issue, as in any ranking method. The original proposal, that will be adopted here as well, of a threshold of one is acceptable if variable relevance is

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

discussed but feature selection is not accomplished. In the cases of marker identification and variable selection, resampling methods such as bootstrap are more appropriate [19,20] to assess the significance of the VIPs.

2.2.2. Three-way case

In the case of a two-dimensional $\mathbf{Y}(I \times M)$ the previous relation applies to each mode, e.g. in the case of a three-way array $\underline{\mathbf{X}}(I \times J \times K)$:

$$\mathrm{VIP}^2_j = \Sigma_f w^2_{jf} \cdot \mathrm{SSY}_{mf} \cdot J / \left(\mathrm{SSY}_{\mathrm{tot.expl.},m} \cdot F\right) \quad (10)$$

$$\mathrm{VIP}^2_k = \Sigma_f w^2_{kf} \cdot \mathrm{SSY}_{mf} \cdot K / \left(\mathrm{SSY}_{\mathrm{tot.expl.},m} \cdot F\right) \quad (11)$$

where, F is the number of total latent variables, J the number of $\underline{\mathbf{X}}$ variables in Mode2 and K the number of variables in Mode3. For each latent variable f:

$$\mathrm{SSY}_{\mathrm{tot.expl.},m} = \Sigma_i \left(\mathbf{T}_{(I\times F)} \mathbf{B}_{(F\times F)} \mathbf{q}^{\mathrm{T}}{}_{(m,F)}\right)^2 \quad (12)$$

and each y-variable y_m:

$$\mathrm{SSY}_{mf} = \Sigma_i \left(\mathbf{t}_f b_{f,f} q_{m,f}\right)^2 \quad (13)$$

where, I is the number of samples in Mode2, $\mathbf{T}$ is the Mode1 score matrix, $\mathbf{B}$ holds the NPLS inner relation coefficients and $\mathbf{Q}$ the $\mathbf{Y}$ loadings.

For a given model dimension f and each variable j, the VIP value is given by the squared weight w^2_{jf} of that parameter (i.e. the weight w_{jf} indicates the importance of the jth variable in the model dimension f), multiplied by the percent of $\mathbf{Y}$ explained sum of squares by that f dimension.

The variable importance is then normalized so that VIP^2 equals the number of the variables.

While considering all $\mathbf{Y}$ variables together, Eqs. (12) and (13) are reduced to:

$$\mathrm{SSY}_{\mathrm{tot.expl.}} = \Sigma_i \left(\mathbf{T}_{(I\times F)} \mathbf{B}_{(F\times F)} \mathbf{Q}^{\mathrm{T}}{}_{(M,F)}\right)^2 \quad (14)$$

$$\mathrm{SSY}_f = \Sigma_m \Sigma_i \left(\mathbf{t}_f b_{f,f} \mathbf{q}^{\mathrm{T}}{}_{m,f}\right)^2 \quad (15)$$

and Eq. (10) is reduced to:

$$\mathrm{VIP}^2_j = \Sigma_f w^2_{jf} \cdot \mathrm{SSY}_f \cdot J / \left(\mathrm{SSY}_{tot.expl.} \cdot F\right). \quad (16)$$

Extension to the other $\mathbf{Y}$ modes can be easily obtained.

3. Data sets and pretreatment

In this study, we present applications of VIP to different three-way data sets. Two data sets are related to optimization of food processing and authentication issue for products with protected denomination of origin, respectively, and the third one is related to a neuroscience problem. Each data set allows exploring the different situations, predictive and discriminant models, partial and overall VIP contribution with respect to $\mathbf{Y}$ block together with the different aspects of complementary information that VIP can highlight with respect to e.g. NPLS weights or regression coefficients.

3.1. Data set 1: Bread

Li Vigni and Cocchi [24] presented a multi-way study related to the influence of flour formulation on bread quality. Ten different flour mixtures were investigated by means of Near Infrared Spectroscopy (NIRS) to obtain information on flour performance in a critical phase such as dough leavening. For each mixture, a laboratory-scale bread making experiment was carried out according to a standardized recipe and the leavening phase of each dough sample was monitored by means of NIRS at different times. NPLS was applied to model the properties of bread loaves (dimensions, volume, weight, height) from near infrared spectra, acquired at different leavening times, of the dough obtained from different flour formulations.

The data are arranged as follows and schematically shown in Fig. 1:

- *X-block*: a three-way array $\underline{\mathbf{X}}(I \times J \times K)$ (10 flour mixtures × 173 NIR wavelengths × 7 leavening time intervals)

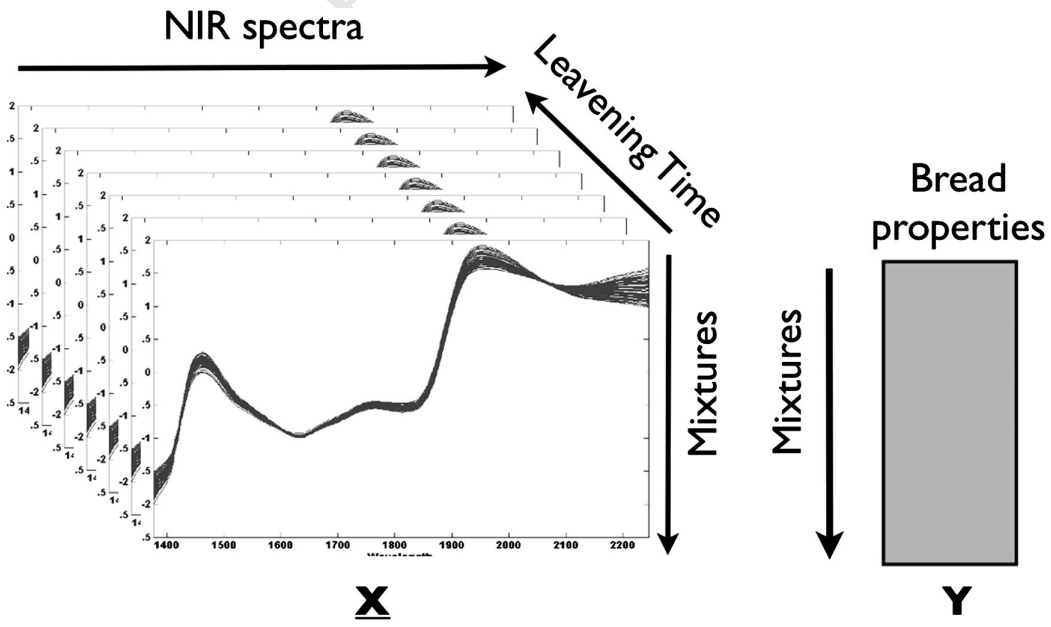

Fig. 1. Data set arrangement for Bread data. (Left) NIR data ($\underline{\mathbf{X}}$): Mode1, samples (the 10 different mixtures); Mode2, NIR spectra (173 wavelengths); Mode3, leavening time points (7). (Right) Bread property data ($\mathbf{Y}$): Mode1, samples (the 10 different mixtures); Mode2, bread properties (4).

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

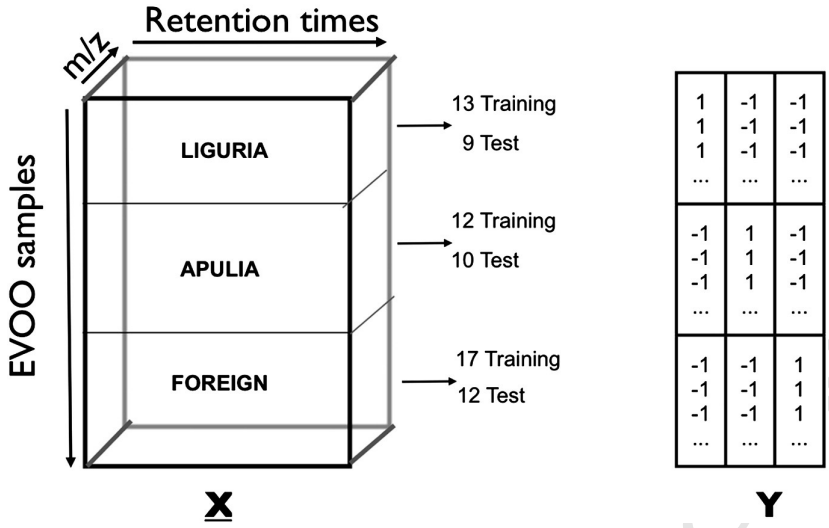

Fig. 2. Data set arrangement for EVOO data. (Left) GC–MS data (**X**): Mode1, samples; Mode2, Retention times (1514 time points); Mode3, 77 selected m/z fragments. (Right) Dummy class variables (**Y**): Mode1, samples; Mode2 classes (3).

- *Y-block*: **Y**($I \times M$) (10 flour mixtures × 4 bread loaf properties: weight, height, diameter and density).

NIR signals were preprocessed by applying Savitsky–Golay Smoothing (15 points window, second order polynomial) coupled to Standard Normal Variate normalization (SNV) to remove the baseline shift.

VIP values are used to assess the spectral regions mainly involved in determining flour performance.

3.2. Data set 2: Extra virgin olive oil (EVOO)

The data set [25] consists of a set of extra virgin olive oil (EVOO), belonging to different olive cultivars and coming from different Mediterranean areas: Liguria (Northern Italy), Apulia (Southern Italy), Greece, Tunisia and Spain. The aim is to assess the authenticity of Ligurian EVOO that has been designed by protected denomination of origin (PDO) certification and represents one of the most highly esteemed EVOOs, of high economic value. The EVOO samples have been characterized by the analysis of aroma (Head Space Solid Phase Micro Extraction coupled with Gas Chromatography–Mass Spectrometry, i.e. HS-SPME/GC–MS), which is well suited for analyzing the volatile fraction that is of relevance for the sensory quality of olive oil. The differentiation among classes has been obtained by NPLS-DA, defining three classes: Liguria, Apulia and Foreign, which includes the EVOO from Turkey, Spain and Greece.

The data set is arranged as follows (Fig. 2):

- *X-block*: a three-way array **X**($I \times J \times K$) (73 EVOO samples × 1514 retention time points × 77 m/z fragments)
- *Y-block*: **Y**($I \times M$) (73 EVOO samples × 3 dummy variables holding class memberships).

For each class the data were randomly split in a training and validation (test) set as shown in Fig. 2. The training set was preprocessed by centering across the first mode, block-scaling within the second mode, by defining four retention time regions in order to allow both major and minor constituents to contribute to the model without up-weighting baseline contribution [25] and scaled by inverse standard deviation within the third mode (selected mass fragments). The pretreatments were applied in the order Mode3, Mode2, and Mode1.

VIP values were used to assess both GC and MS discriminant features.

3.3. Data set 3: Neuroscience data set

This data set derives from a functional magnetic resonance imaging (fMRI) experiment where the psychophysical pain profile, corresponding to subjective responses to acute prolonged noxious stimulation of one hand, was acquired in healthy volunteers. The experiment lasted 20 min (300 time points), the sensory intensity of pain (psychophysical pain profile) and the hemodynamic response (blood-oxygen-level contrast registered by a magnetic resonance pulse sequence, fMRI-BOLD signal) were recorded simultaneously during the experiment. The functional fMRI-BOLD signals (fMRI-BOLD time series) acquired at each brain voxel, as described in Prato et al. [26], were summarized for forty four brain regions of interest (ROIs) by taking the first principal singular vector (1st-SVD) of the data matrix containing the fMRI-BOLD time series for each voxel in that specific ROI.

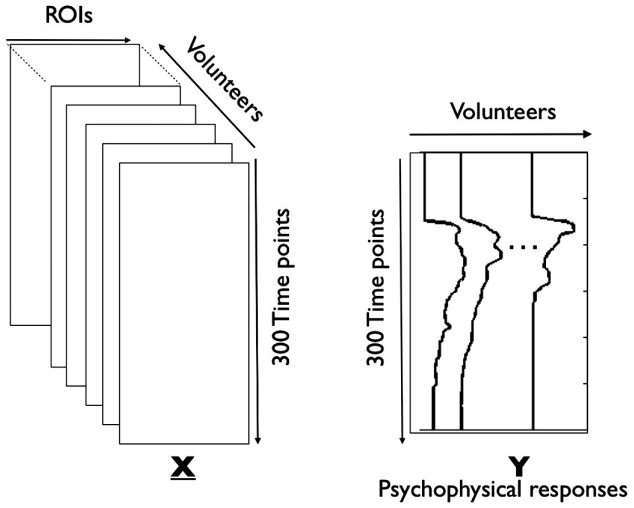

Fig. 3. Data set arrangement for Neuroscience data. (Left) fMRI-BOLD data (**X**): Mode1, Times Points (300); Mode2, fMRI-BOLD intensity for the 44 ROIs; Mode3, volunteers (10). (Right) psychophysical responses (**Y**): Mode1, Time points (300); Mode2, perceived pain intensity by volunteers (10), in this case the actual **Y**($300 \times 1 \times 10$) has been rotated for illustration purposes.

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

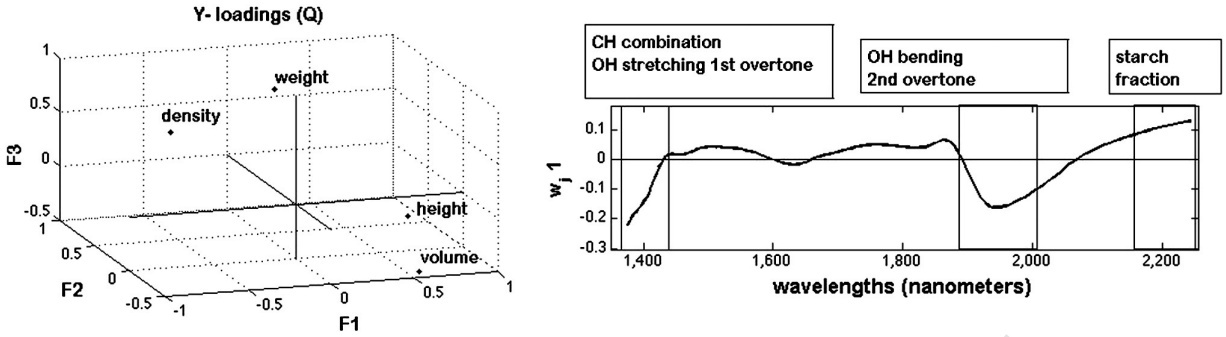

Fig. 4. Data set for Bread. (Left) Y-loadings plot (Q) for the three NPLS components, F1, F2, and F3; (Right) NPLS–weight plot for Mode2 (spectra), F1 vs. wavelengths.

NPLS was applied to build a model that could express the main variation of the ROI time series of different volunteers and obtain a fitted model that could reproduce the corresponding psychophysical pain profile efficiently.

The data arrays have been arranged as (Fig. 3):

- *X-block*: a three-way array $\underline{\mathbf{X}}(I \times J \times K)$ (300 time points × 44 ROIs × 10 volunteers).
- *Y-block*: the $\underline{\mathbf{Y}}(I \times 1 \times K)$ array is actually a matrix $\mathbf{Y}(I \times K)$ comprised of 300 time points (psychophysical pain profile) × 10 volunteers, as shown in Fig. 3, and so computationally handled as such.

The choice of defining time as mode one was motivated by the applicability of the model. In fact, for this approach the main scope was to identify those ROI time series strictly connected (i.e., in terms of covariance) with the psychophysical pain profile of each volunteer (see Fig. 3).

The $\underline{\mathbf{X}}$ and $\mathbf{Y}$ data were not centered or scaled within any mode.

VIP values were used for ranking the ROIs according to their relevance in the NPLS model hence to depict brain region activation profile in response to pain stimulus.

4. Results and discussion

4.1. Bread

Near infrared spectra acquired on dough at subsequent leavening times is an efficient way to characterize the leavening process. In particular, NPLS was used to study the relationship between the modifications recorded by the NIR signal during the leavening time and four properties measured on bread loaves, namely height, weight, volume and density. The dough samples correspond to ten different wheat mixtures (combining four distinct wheat varieties), performed according to a G-optimal design, thus bread performance can be linked to best mixture formulation in terms of wheat varieties.

The dimensionality of NPLS model was chosen on the basis of the best compromise of the minimum values of RMSECV for the four properties, modeled as a single $\mathbf{Y}$ block. Leave One Out cross validation was chosen due to the limited number of samples, ten.

A three factor NPLS model explains 75% of the total $\mathbf{Y}$ variance in fit, with acceptable performances in fit for all responses and generally poor results in cross-validation. As discussed in Ref. [24] these are mainly due to two mixtures showing extreme behavior in property space. However, a qualitative identification of the relationship among NIR signal variability along with the progression of the leavening step and bread properties can be obtained by inspecting NPLS model results.

In particular, $\mathbf{Y}$ loading plot (Fig. 4, left plot) shows that the first NPLS latent variable (F1) mainly models bread loaf height and volume, while latent variables 2 and 3 (F2 and F3) model bread loaf weight. NPLS Mode2 weights indicate which spectral regions mostly influence each factor and hence bread loaf properties. For example, the first factor for Mode2 weights ($\mathbf{w}_{j1}$, Fig. 4, right plot), which is linked to height/volume properties, shows that the most relevant contributions can be assigned to water and its redistribution across the macro-polymeric components of the dough, such as gluten and starch. In fact, the most relevant frequencies are those near 1400, and between 1900 and 1950 nm, for which weights have a negative sign, and above 2100 nm, for which weights have positive sign. These regions correspond, respectively, to absorptions that can be associated to the O–H stretching first overtone, the O–H bending second overtone mode and to overtone and combination mode contributions from the starch, protein and lipid fractions.

Inspection of Mode3 weights (leavening times, Fig. 5) indicates that height and volume are influenced by initial (t0–t10) and last leavening phases (yeast activity, dough strength); in fact these times have relevant weight values on the first NPLS factor (Fig. 5, top plot); weight is mainly influenced by initial (t0) time; hence flour properties are mostly important, as shown by second and third NPLS factor weights (Fig. 5, middle and bottom plots).

The regression coefficient maps are not so easily interpretable, see Fig. 6, while in order to assess the most relevant spectral regions and leavening times VIP values, proved to be very useful (Fig. 7A and B). Fig. 7A reports the VIP values for Mode3. It is interesting to notice that, for bread weight, the spectrum at the beginning of the leavening phase (time 0) has VIP values higher than one while height property exhibits significant alternating variation at times 10, 40 and 60. The other two properties, volume and density, show that the times at which the most significant variations in the NIR spectrum are recorded correspond mainly to the initial conditions (0 and 10 min) and the final one (60 min).

The most influential NIR spectrum regions, as highlighted by Mode2 VIP values (Fig. 7B), are the same for all properties: at about 1425 nm.

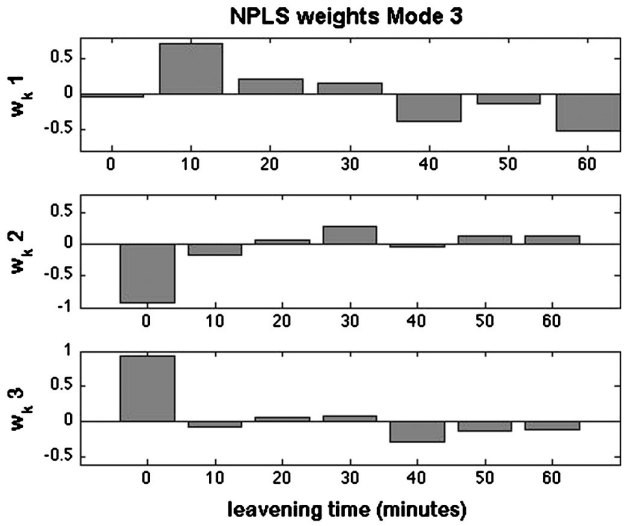

Fig. 5. Data set for Bread. Plot of NPLS weight for Mode3 (leavening times).

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

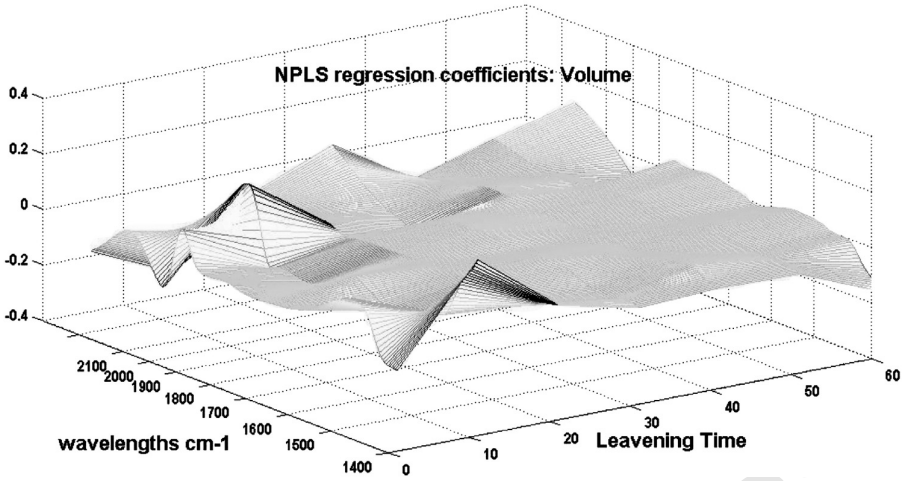

Fig. 6. Data set for Bread. 2D plot of the NPLS regression coefficient map for the y-property Volume.

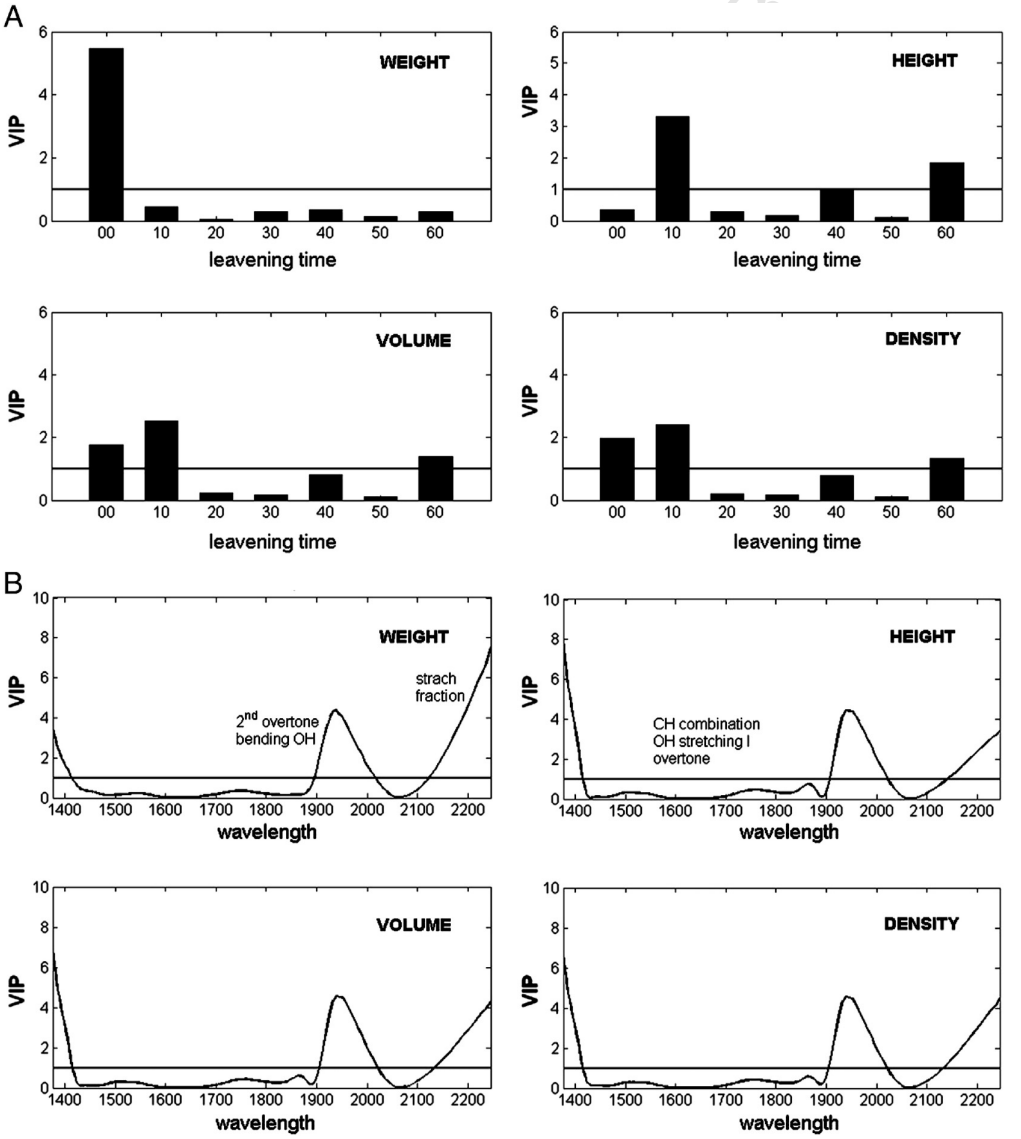

Fig. 7. Data set Bread. (A) Squared VIP values for Mode3 for each bread property; (B) squared VIP values for Mode2 for each bread property.

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

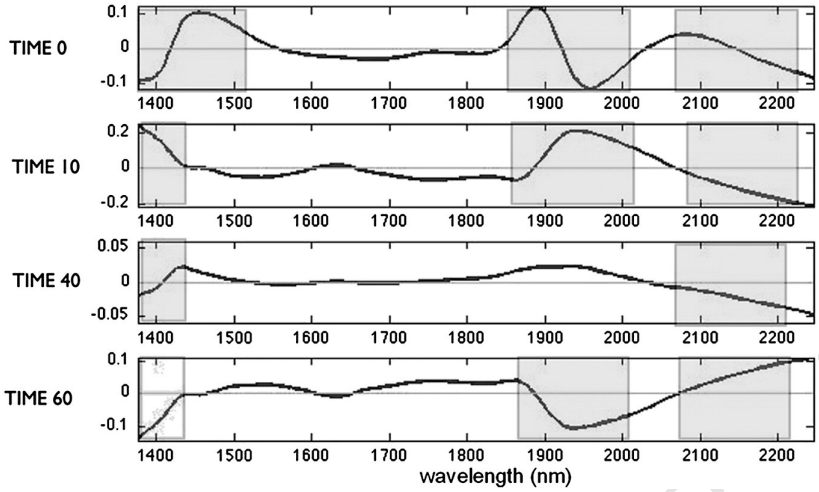

Fig. 8. Data set for Bread. Plot of NPLS regression coefficients at selected leavening times. Gray bar highlights the most relevant NIR spectral regions.

There are also contributions from C–H combination and first O–H stretching overtone, between 1900 and 1950 nm, corresponding to absorptions which can be associated to the second overtone O–H bending mode, and above 2100 nm, where overtone and combination mode contributions from the starch fractions are present.

On the basis of VIP analysis it is now possible to plot for each modeled bread property the NPLS regression coefficients corresponding only to the most influential leavening times, so that the spectral contributions can be discussed in terms of increasing/decreasing of specific absorption bands, taking into account the regression coefficients sign, in an easier way. Fig. 8, as an example, shows the regression coefficient plot for the most relevant leavening times for the bread Volume. The most significant spectral regions discussed above change correlation sign at the different times as it may be expected by the dynamic of the leavening process where several rearrangements of the starch and gluten networks take place [24].

The combination of multi-way methods applied to NIR spectra is here useful to supervise changes of the system according to the leavening time, and can be used as a reference to evaluate the behavior of dough obtained from different wheat flour mixtures, and potentially to identify anomalous leavening situations. Also, it has been shown that VIP values give the same information as the joint discussion of Y loadings and NPLS weights, but with a more direct highlight of the most influential contribution for each property. Moreover, they can be taken into account as a guide to plot in an interpretable manner the NPLS regression coefficients, which may otherwise result of difficult interpretation.

4.2. Extra virgin olive oil (EVOO)

The performance of the NPLS-DA classification models has been evaluated by means of percentage of correct classification in cross-validation (CV), by using venetian blind with six cancelation groups. Three NPLS components gave the best performance with 100% correct classification in fit and CV for Liguria and Foreign classes and 92% (fit and CV) for Apulia. The test set for Liguria and Foreign classes was

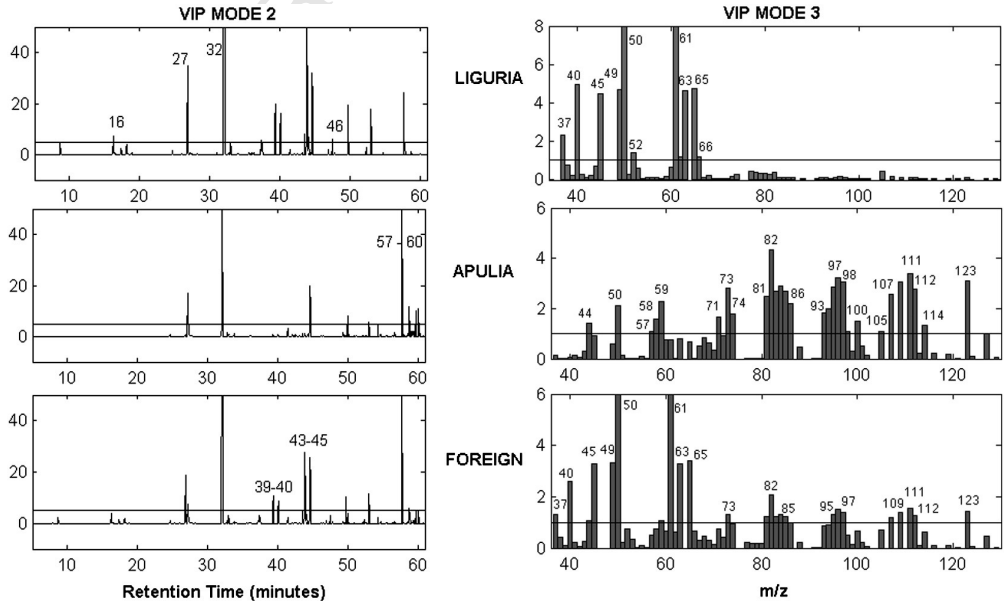

Fig. 9. Data set for EVOO. Plot of the squared VIP. (Left) Mode2 VIPs vs. Retention Times, threshold sets to 5. (Right) Mode3 VIPs vs. m/z fragments, threshold sets to 1.

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

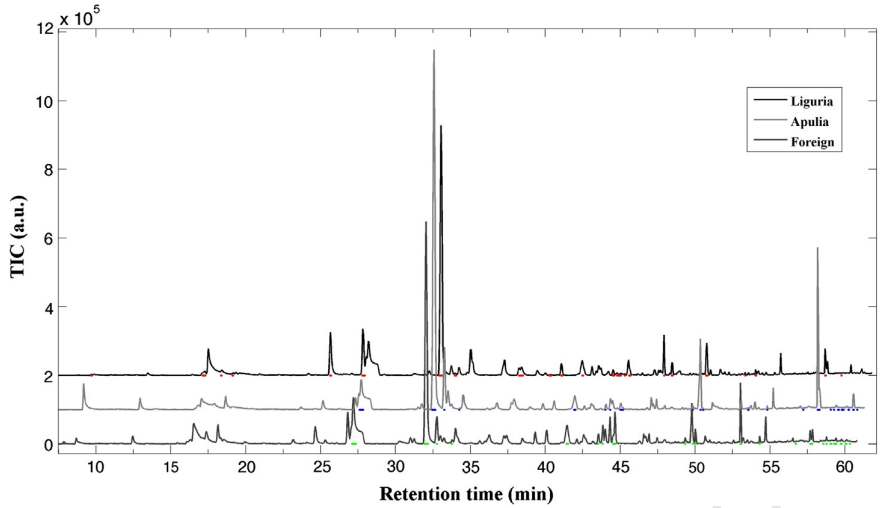

Fig. 10. Data set for EVOO. Average TIC chromatograms for each class, overlaid with a shift of 0.5 on time axis (x-axis) and of 1 on y-axis. Dots indicate retention time with VIP values higher than one.

correctly classified with no error and with one misclassified sample for Apulia.

The VIP values for Mode2 and Mode3 for each y-variable, hence category, are reported in Fig. 9. The VIP values threshold is drawn at 1 for Mode3 (m/z fragment), Fig. 9 on the right; while for Mode2 (Retention time) is drawn at 5, Fig. 9 on the left. This choice is motivated by considering that this is the number of points generally defining a peak, thus a variable to be considered important has to have a contribution at least as one peak in the signal. Taking into account the most important variables of both modes some considerations about the main compositional difference of each EVOO category can be made. The chromatographic peaks, which seem to characterize the volatile fraction of Liguria olive oil, are some low molecular weight compounds (Retention time about 16 min; m/z < 40) and C6 linear unsaturated aldehydes (Retention time regions at about 27 min and 32 min; m/z region 61–70) characteristics of high quality virgin olive oils. The latter compounds are also present in Apulia and the Foreign class but in a lower amount. The Apulia class is mainly characterized by the retention time region 57–60 min and by higher molecular weights compounds (m/z values higher than 80), such as alpha-copaene (Rt = 57.7 min; m/z 81, 93, 105). The high VIP values (mainly for Liguria and Foreign classes) at the retention time regions 39–40 and 43–45 min highlight compounds that are more related to, i.e. specific for, the Foreign class as it can be seen from Fig. 10, where the average total ion count chromatograms (TIC) for each class are shown.

The NPLS weights for Mode2 and Mode3 for each of the three NPLS components (F1, F2 and F3) are reported in Fig. 11. In order to interpret these plots the Y loadings have to be inspected (figure not shown). These indicate that Liguria (y1) has high positive loadings on components 2 and 3 (F2 and F3) and close to zero on component 1 (F1), the opposite holds for Foreign (y3) while Apulia has high positive loadings values on component 1 (F1) and almost zero on the other two components.

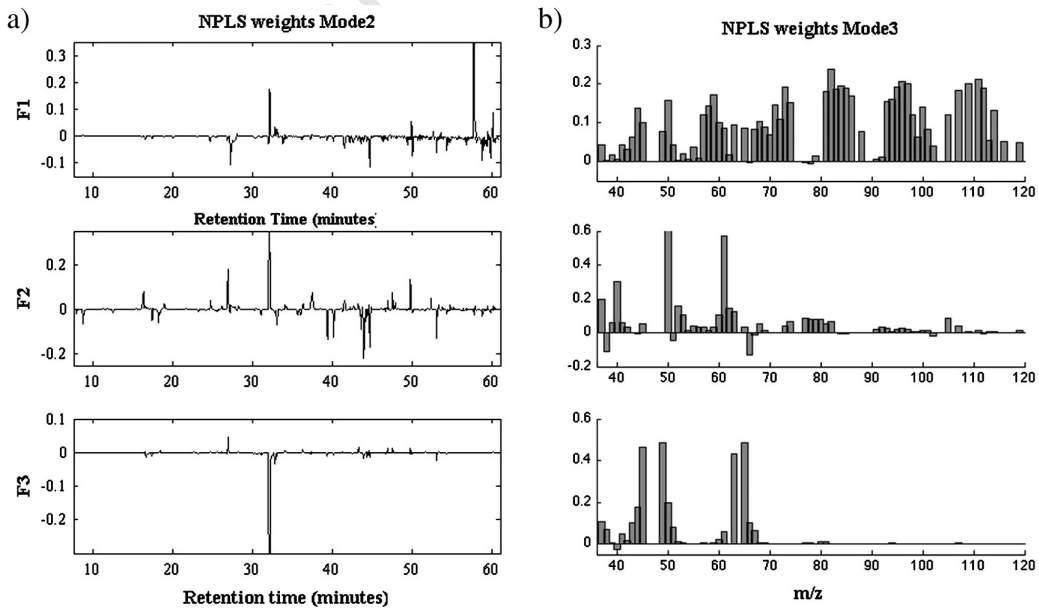

Fig. 11. Data set EVOO. (a) NPLS weights for Mode2 vs. retention times, for NPLS components F1 (top), F2 (middle) and F3 (bottom). (a) NPLS weights for Mode3 vs. m/z fragments, for NPLS components F1 (top), F2 (middle) and F3 (bottom).

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

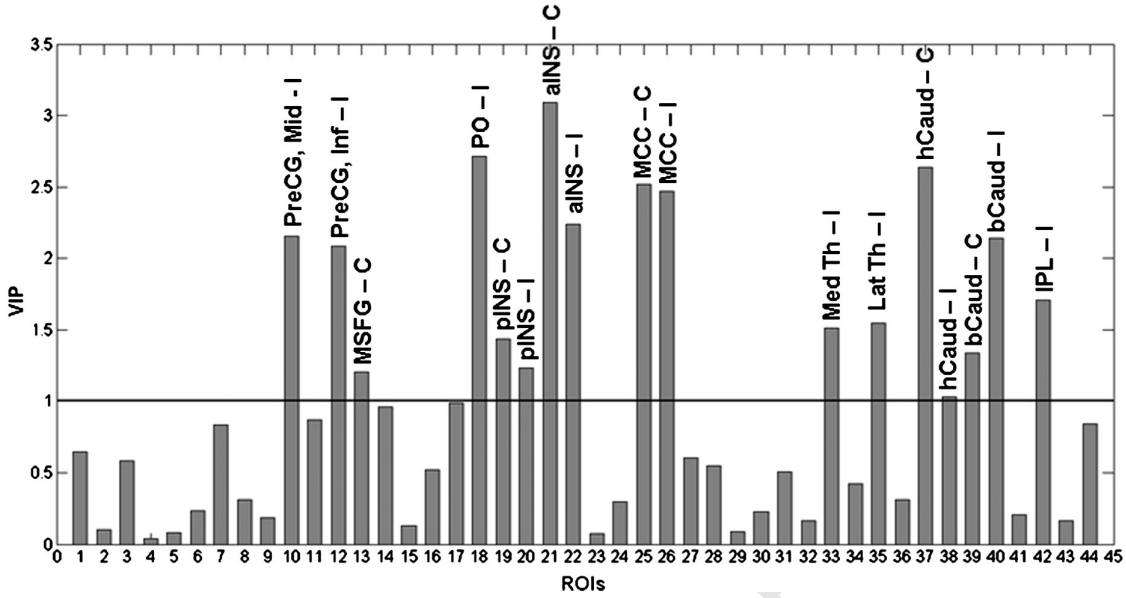

Fig. 12. Data set for Neuroscience. Plot of the squared VIP calculated for all **Y** altogether vs. ROIs, the threshold is set to 1. Labels refer to ROIs: Contralateral (*C*) and Ipsilateral (*I*) to the injected hand, respectively; Pre-Central Gyrus (*PreCG*), Middle (*Mid*); Medial Superior Frontal Gyrus/Paracentral Lobule (*MSFG*); Parietal Operculum (*PO*); Posterior Insula (*pINS*); Anterior Insula (*aIns*); Mid-Cingulate Cortex (*MCC*); Medial Thalamus (*Med Th*); Lateral Thalamus (*Lat Th*); Caudate Nucleus, head (*hCaud*); Caudate Nucleus, body (*bCaud*); Inferior Parietal lobule (*IPL*).

Taking this into account by looking at the Mode2 weights (Fig. 11a) the observations about the characteristics retention time regions for each class, made on the basis of the VIP values, are confirmed, e.g. the negative sign of the weights at retention time 39–40 and 43–45 indicates that this region is characteristic of the Foreign class.

4.3. Neuroscience data set

The goal was to build a model that, starting from the fMRI-BOLD characterization as expressed by the 1st SVD component of each ROIs fMRI-BOLD time series, could predict efficiently the psychophysical pain intensity for each volunteer. A NPLS model with 4 Latent Variables (LVs) was selected according to Cross Validation results, lowest RMSECV value, i.e. 28.12, with explained **Y** variance around 95%.

The VIP values were used for ranking the brain regions (ROIs) mainly involved in pain perception. In this case it is interesting to consider the VIP values calculated for the overall **Y**-responses, to gain a global picture of brain activation common to all subjects. Fig. 12 reports the VIP values versus ROI number as bar plot, highlighting the ROIs showing VIP values greater than one. These brain regions are in agreement with the results reported in Prato et al. [26].

The information gathered by the VIPs is complementary to the one carried by the N-PLS weights. Fig. 13 reports the weights for the first two N-PLS factors for the second mode (ROIs). Extreme positive $\mathbf{w}_j$ values on F1 are those related to ROI with VIP values higher than one. Inspection of the corresponding weights plot for Mode3 (volunteers), i.e. $\mathbf{w}_k$, Fig. 14, shows that the different volunteers are distributed along the first factor that differentiates e.g. volunteer #9 from volunteer #10.

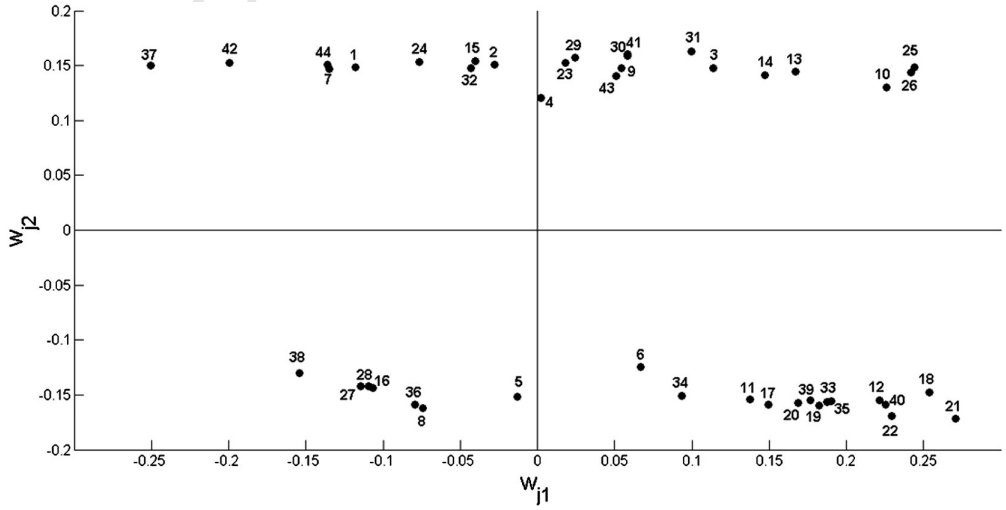

Fig. 13. Data set for Neuroscience. Scatter plot of NPLS weights for Mode2 (ROIs): first vs. second component.

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

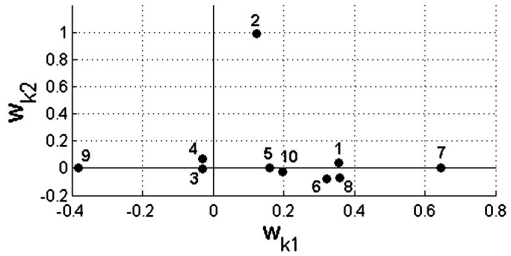

Fig. 14. Data set for Neuroscience. Scatter plot of NPLS weights for Mode3 (volunteers): first vs. second component.

This can be discussed by considering the different behaviors in perceived pain as underlined by the first mode loadings plots (functional fMRI profiles), shown in Fig. 15. The first factor describes the average pain profile, the second seems almost dedicated to a delayed maximum peak and more persistent pain. High positive loading values on factor one for Mode2, Fig. 14 (see for instance volunteer #7) represent an opposite behavior with an anticipated maximum peak in comparison with the average pain profile. The highest negative value (volunteer #9) with the extreme negative position (with respect to the abscissa) identifies a positive shift of the maximum pain perceived in comparison with the mean profile (reference volunteer #3). This may be retrieved by direct observation of the psychophysical responses for these subjects in comparison with those subjects showing a profile close to average as subjects three, see Fig. 16.

The second factor can be considered as a component that takes into account a sort of "prolonged activation due to tonic pain input" (see Fig. 15, in gray dashed) and it is particularly dedicated to describe volunteer #2 with its ample bell-shape of the pain perceived. In Fig. 16 the comparison between the pain profiles of volunteer #2 and the reference volunteer #3 is shown. High weight w_{k2} value (see Fig. 14) for this volunteer seems to be only related to its particular behavior that is also responsible for the separation of the ROI regions in two groups with respect to the weight values on the second mode (Fig. 13).

Thus, discussion of the weights plot is useful to recover the detailed information on specific subject behavior while the overall VIP values point to the most relevant ROIs whose activation is involved in pain perception and that are thus capable of reproducing the Y-psycho responses.

5. Conclusions

The recent developments in feature selection methods for two-way data have addressed the problem of increasing the performance of regression models, such as PLS. Complex filter, wrapper or embedded methods [11,20] improve predictor performance compared to simpler variable ranking methods, but the improvements are not always significant, they are computationally costly and in case of a large number of variables, the risk of over-fitting can be not negligible in the process of variable selection. The extension of variable selection methods to the multi-way data arrays, in multilinear regression context (NPLS, NPLS-DA) without recurring to unfolding, has been rather limited.

In this work, we introduced an extension of the Variable Importance in Projection (VIP) parameter to multi-way arrays. The proposed method has been tested on three different data sets where VIPs were discussed comparatively with respect to NPLS weight and regression coefficients.

VIPs naturally point out the identification of the most relevant variables related to **Y** in a multi-way array **X**. In particular, VIPs can be calculated for each mode of **X** and both considering the single y-responses or all the **Y** altogether. The former can be useful to assess the relevant variables with respect to each modeled properties, especially in the case of discriminant NPLS-DA to highlight the discriminant features, since each y-variable corresponds to a given class. While the latter offer a useful summary in order to operate feature selection.

However, when considering VIP, it is important to remember that, from an interpretative point of view, this metric suffers from the fact that PLS/NPLS components carry with them the unresolved contribution of both Y-related and Y-orthogonal parts of the X-variance. Both contributions are fundamental for a correct prediction of Y and the VIP metric represents a valid support whenever the interest is aimed to rank variables according to their influence to the whole model, while for contexts and purposes where the interest is mainly focused on assessing the X-part covarying with Y only, other methodologies may represent a valuable solution, such as OPLS [28] and Selectivity Ratio metric [21].

In the studied cases, VIPs provided an easier and complementary way to interpret the variable relevance in NPLS models, especially when examination of regression coefficients was not so straightforward due to the unreadable complex patterns associated [27], as in the case of spectral data (as for the Bread data set) and moreover with two signal dimensions as in the case of hyphenated analytical techniques. In the EVOO application, which is an example of chromatography/mass spectrometry data, the joint information from the VIPs on the retention time and m/z directions allows discussion in chemical terms of the most discriminant features.

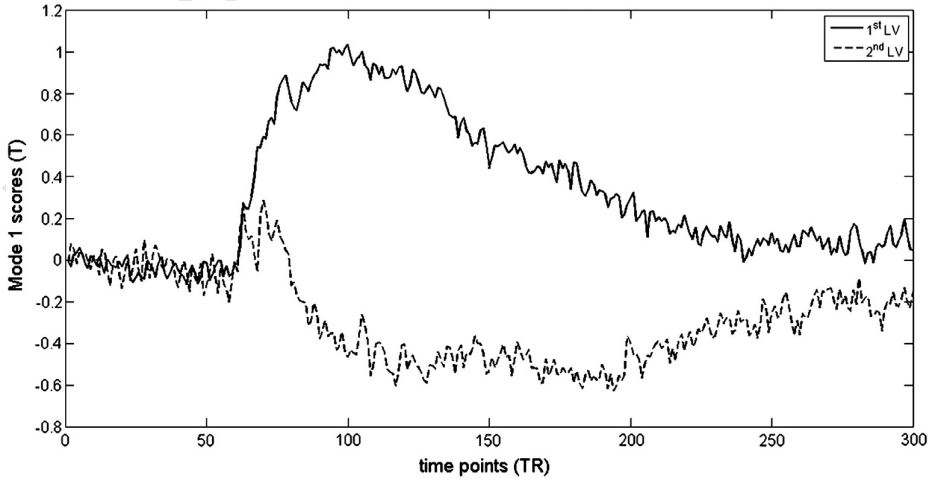

Fig. 15. Data set for Neuroscience. Mode1 loading (score) plot vs. time points (the temporal mode showing the evolution of functional fMRI-BOLD): (black) first component; (gray dashed) second component.

Please cite this article as: S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.05.013

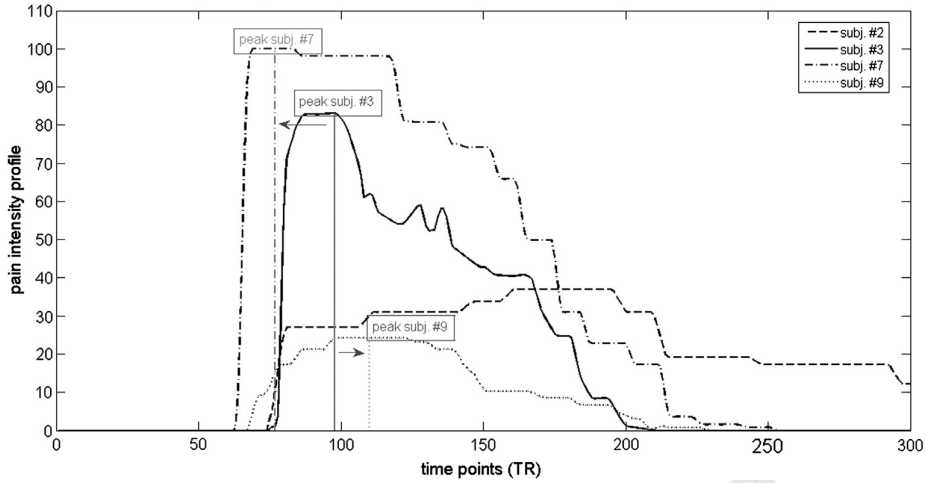

Fig. 16. Data set for Neuroscience. Pain intensity profile (psychophysical response) vs. time points, for some reference volunteers: subject #3 (solid black), subject #2 (dashed), subject #7 (dot-dashed) and subject #9 (dotted).

NPLS weights carry the information about the relevance of variables and sign of correlation with the modeled responses. However, they require to be discussed together with the **Y** loadings and per component. In this respect, the VIPs are complementary pointing to the most influential variables for each property on taking into account all components but of course require inspection of weights to assess the direction of the effect (increasing or decreasing the response values). Finally, the results obtained on the Neuroscience dataset were found to be in line with those published with a completely different method belonging to the machine learning field [26] as far as ranking of the most influential ROIs is concerned, while the use of multi-way models added the possibility to discuss both the common pattern to all volunteers in pain perception as well as the peculiar behavior of specific ones.

6. Uncited reference

[23]

Acknowledgments

The “Functional Neuroimaging” group, of the Department of Biomedical Sciences, Metabolic and Neuroscience, University of Modena and Reggio Emilia are kindly acknowledged for providing the fMRI data and for the support given in the discussion and interpretation of the results on this data set.

References

[1] R. Bro, Multi-way calibration. Multi-linear PLS, Journal of Chemometrics 10 (1996) 47–61.
[2] R. Bro, A.K. Smilde, S. De Jong, On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression, Chemometrics and Intelligent Laboratory Systems 58 (2001) 3–13.
[3] A. Smilde, Comments on multilinear PLS, Journal of Chemometrics 11 (1997) 367–377.
[4] S. De Jong, Regression coefficients in multilinear PLS, Journal of Chemometrics 12 (1998) 77–81.
[5] A. Smilde, R. Bro, P. Geladi, Multi-way analysis with applications, Multi-way Analysis in the Chemical Sciences, Chpt 3.1, John Wiley & Sons, 2004, pp. 35–45.
[6] E. Salvatore, M. Bevilacqua, R. Bro, F. Marini, M. Cocchi, Classification methods for multi-way arrays as a basic tool for food PDO authentication, in: M. De La Guardia, A. Gonzalvez Illueca (Eds.), Food Protected Designation of Origin: Methodologies & Applications, Wilson & Wilson's Comprehensive Analytical Chemistry, vol. 60, Elsevier B.V., 2013
[7] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.
[8] C. Durante, M. Cocchi, M. Grandi, A. Marchetti, R. Bro, Application of N-PLS to gas chromatographic and sensory of traditional balsamic vinegars of Modena, Chemometrics and Intelligent Laboratory Systems 83 (2006) 54–65.
[9] C.M. Rubingh, M.J. van Erk, S. Wopereis, T. van Vliet, E.R. Verheij, N.H.P. Cnubben, B. van Ommen, J. van der Greef, H.F.J. Hendriks, A.K. Smilde, Discovery of subtle effects in a human intervention trial through multilevel modeling, Chemometrics and Intelligent Laboratory Systems 106 (2011) 108–114.
[10] A. Conesa, J.M. Prats-Montalbán, S. Tarazona, M.J. Nueda, A. Ferrer, A multiway approach to data integration in systems biology based on Tucker3 and N-PLS, Chemometrics and Intelligent Laboratory Systems 104 (2010) 101–111.
[11] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemometrics and Intelligent Laboratory Systems 118 (2012) 62–69.
[12] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy 54 (2000) 413–419.
[13] R. Leardi, A.L. Gonzales, Genetic algorithms applied to feature selection in PLS regression: how to use them, Chemometrics and Intelligent Laboratory Systems 41 (1998) 195–207.
[14] L. Stordrange, T. Rajalahtia, F.O. Libnau, Multiway methods to explore and model NIR data from a batch process, Chemometrics and Intelligent Laboratory Systems 70 (2004) 137–145.
[15] P.J. Odman, C. Lindvald, L. Olsson, et al., Sensor combination and chemometric variable selection for online monitoring of *Streptomyces coelicolor* fed-batch cultivations, Applied Microbiology and Biotechnology 86 (2010) 1745–1759.
[16] G. Lorho, F. Westad, R. Bro, Generalized correlation loadings. Extending correlation loadings to congruence and to multi-way models, Chemometrics and Intelligent Laboratory Systems 84 (2006) 119–125.
[17] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: Hugo Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM Science Publishers, Leiden, ISBN: 90-72199-14-6, 1993, pp. 523–550.
[18] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[19] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemometrics and Intelligent Laboratory Systems 78 (2005) 103–112.
[20] R. Gosselin, D. Rodrigue, C. Duchesne, A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemometrics and Intelligent Laboratory Systems 100 (2010) 12–21.
[21] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, Chemometrics and Intelligent Laboratory Systems 95 (2009) 35–48.
[22] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, A.L. Ritaccio, B. Yener, Modeling and detection of epileptic seizures using multi-modal data construction and analysis, Technical Report, 2008, (Computer Science Department at RPI (http://www.cs.rpi.edu/research/tr.html)).
[23] J. Nilsson, S. De Jong, A.K. Smilde, Multiway calibration in 3D QSAR, Journal of Chemometrics 11 (1997) 511.
[24] M. Li Vigni, M. Cocchi, Near infrared spectroscopy and multivariate analysis to evaluate wheat flour doughs leavening and bread properties, Analytica Chimica Acta 764 (2013) 17–23.
[25] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometrics and Intelligent Laboratory Systems 106 (2011) 73–85.
[26] M. Prato, S. Favilla, L. Zanni, C.A. Porro, P. Baraldi, A regularisation algorithm for decoding perceptual temporal profiles from fMRI data, NeuroImage 56-1 (2011) 258–267.
[27] A.J. Burnham, J.F. MacGregor, R. Viveros, Interpretation of regression coefficients under a latent variable regression model, Journal of Chemometrics 15 (2001) 265–284.
[28] J. Trygg, S. Wold, Journal of Chemometrics 17 (2003) 53–64.