



**HAL**  
open science

## VODKA-PLSR, une famille de modèles PLSR dérivés de l'algorithme NIPALS

Jean Claude J. C. Boulet, Dominique Bertrand, Gerard Mazerolles,  
Jean-Michel Roger

► **To cite this version:**

Jean Claude J. C. Boulet, Dominique Bertrand, Gerard Mazerolles, Jean-Michel Roger. VODKA-PLSR, une famille de modèles PLSR dérivés de l'algorithme NIPALS. Chimiométrie 2010, Dec 2010, Paris, France. pp.13-15. hal-02750965

**HAL Id: hal-02750965**

**<https://hal.inrae.fr/hal-02750965>**

Submitted on 3 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# VODKA-PLSR: une famille de modèles PLSR dérivés de l'algorithme NIPALS

J.C. Boulet<sup>1</sup>, D.Bertrand<sup>2</sup>, G.Mazerolles<sup>1</sup>, R.Sabatier<sup>3</sup>, J.M.Roger<sup>4</sup>

<sup>1</sup> INRA, UMR1083, 2 place Viala, F-34070 Montpellier, bouletjc@supagro.inra.fr

<sup>2</sup> INRA, Bioinformatique, rue de la Géraudière, F-44136 Nantes, Dominique.Bertrand@nantes.inra.fr

<sup>3</sup> UM1, EA2415, 15 av. C.Flahault, F-34093 Montpellier, sabatier@univ-montp1.fr

<sup>4</sup> CEMAGREF, UMR1201, 361 rue J.F.Breton, F-34093 Montpellier, jean-michel.roger@cemagref.fr

**Mots clés :** regression, PLSR, NIPALS

Un nouveau modèle de régression basée sur NIPALS est proposé. Il s'appuie sur trois entrées: un jeu d'étalonnage constitué de N spectres sur P variables donnant  $\mathbf{X}$ , les valeurs de référence pour une grandeur d'intérêt donnant  $\mathbf{y}$  de dimension  $N \times 1$ , enfin un vecteur  $\mathbf{r}$  de dimensions  $P \times 1$  choisi arbitrairement. Ces entrées sont utilisées pour le calcul des deux paramètres  $\Sigma$  et  $\mathbf{P}$ , avec  $\Sigma = (\mathbf{X}'\mathbf{X})^+$  la pseudo-inverse de  $(\mathbf{X}'\mathbf{X})$  au sens de Moore-Penrose.

## 1- Théorie

### 1.1 Nouvelle écriture de NIPALS

NIPALS utilise  $(\mathbf{X}, \mathbf{y})$  pour calculer itérativement les A colonnes de trois matrices contenant des scores  $\mathbf{T}$ , des vecteurs  $\mathbf{P}$  et des poids  $\mathbf{W}$  [3]. Par exemple, soit i un indice entre 1 et A, si nous notons  $\mathbf{P}_{1:i-1}^+$  le projecteur de  $\mathbb{R}^N$  orthogonal à  $\{\mathbf{t}_1 \dots \mathbf{t}_{i-1}\}$ , à chaque boucle les vecteurs  $\mathbf{p}_i$  sont calculés ainsi:

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

Le projecteur  $\mathbf{P}_{1:i-1}^+$  est inutile puisque  $\mathbf{t}_i$  est déjà orthogonal à  $\{\mathbf{t}_1 \dots \mathbf{t}_{i-1}\}$ , d'où:

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

Multiplions de chaque coté à gauche par  $\mathbf{X}\Sigma$  soit  $\mathbf{X}(\mathbf{X}'\mathbf{X})^+$ . Le terme  $\mathbf{X}(\mathbf{X}'\mathbf{X})^+\mathbf{X}'$  est le projecteur orthogonal sur  $\mathbf{X}$  [2]. Or  $\mathbf{t}_i$  est obtenu par combinaison linéaire des colonnes de  $\mathbf{X}$ ,  $\mathbf{t}_i$  appartient obligatoirement à l'espace décrit par les colonnes de  $\mathbf{X}$ , donc sa projection sur  $\mathbf{X}$  donne  $\mathbf{t}_i$ . Ainsi après arrangement des termes:

$$\mathbf{t}_i = \mathbf{X}\Sigma \mathbf{p}_i (\mathbf{t}_i' \mathbf{t}_i) \quad (2)$$

L'expression de  $\mathbf{t}_i' \mathbf{t}_i$  selon  $\mathbf{p}_i$  et  $\Sigma$  s'obtient en calculant  $\mathbf{p}_i' \Sigma \mathbf{p}_i$  à partir de l'équation (1). Ce résultat est introduit dans (2):

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

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

On démontre aussi que les  $\mathbf{p}_i$  sont strictement orthogonaux entre eux au sens de  $\Sigma$ , soit pour  $i \neq j$ ,  $\mathbf{p}_i' \Sigma \mathbf{p}_j = 0$ . Donc  $\mathbf{P}' \Sigma \mathbf{P}$  est une matrice diagonale dont le terme de la  $i^{\circ}$  ligne et de la  $i^{\circ}$  colonne est  $\mathbf{p}_i' \Sigma \mathbf{p}_i$ . En conséquence l'équation (4) conduit à:

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

$\mathbf{T}$  est la matrice des scores de la projection  $\Sigma$ -orthogonale de  $\mathbf{X}$  sur  $\mathbf{P}$ . Les b-coefficients sont déduits après une régression aux moindres carrés de  $\mathbf{y}$  sur  $\mathbf{T}$ .

## 1.2. Calcul de P par NIPALS-P.

NIPALS-P est un calcul plus direct des  $\mathbf{p}_i$  puisqu'il reste dans  $\mathbb{R}^p$ . Transposons chaque terme de l'équation (1) et multiplions à gauche par  $\mathbf{t}_i$ , puis remplaçons le  $\mathbf{t}_i$  à gauche par sa valeur donnée par l'équation (4):

$$\mathbf{t}_i \mathbf{p}_i' = \mathbf{t}_i (\mathbf{t}_i' \mathbf{t}_i)^{-1} \mathbf{t}_i' \mathbf{X} \quad (6)$$

$$\mathbf{X} \Sigma \mathbf{p}_i (\mathbf{p}_i' \Sigma \mathbf{p}_i)^{-1} \mathbf{p}_i' = \mathbf{t}_i (\mathbf{t}_i' \mathbf{t}_i)^{-1} \mathbf{t}_i' \mathbf{X} \quad (7)$$

Soit  $\mathbf{Q}_{1:i}^+ = \mathbf{I}_p - \Sigma \mathbf{P}_{1:i} (\mathbf{P}_{1:i}' \Sigma \mathbf{P}_{1:i})^{-1} \mathbf{P}_{1:i}'$  le projecteur  $\Sigma$ -orthogonal à  $\{\mathbf{p}_1 \dots \mathbf{p}_i\}$ . On démontre à partir de l'équation (7) que:

$$\mathbf{P}_{1:i}^+ \mathbf{X} = \mathbf{X} \mathbf{Q}_{1:i}^+ \quad (8)$$

D'autre part, à partir de NIPALS, la substitution de  $\mathbf{w}_{i+1}$  dans l'expression de  $\mathbf{t}_{i+1}$  puis celle de  $\mathbf{t}_{i+1}$  dans l'expression de  $\mathbf{p}_{i+1}$  donne:

$$\begin{aligned} \mathbf{p}_{i+1} &= \alpha_{i+1} \mathbf{X}' \mathbf{P}_{1:i}^+ \mathbf{X} \mathbf{X}' \mathbf{P}_{1:i}^+ \mathbf{y} \\ \mathbf{p}_{i+1} &= \alpha_{i+1} (\mathbf{P}_{1:i}^+ \mathbf{X})' \mathbf{X} (\mathbf{P}_{1:i}^+ \mathbf{X})' \mathbf{y} \end{aligned} \quad (9)$$

avec  $\alpha_{i+1}$  un scalaire. Reprenons la formule (9) en remplaçant  $\mathbf{P}_{1:i}^+ \mathbf{X}$  par  $\mathbf{X} \mathbf{Q}_{1:i}^+$ :

$$\mathbf{p}_{i+1} = \alpha_{i+1} \mathbf{Q}_{1:i}^{+'} \mathbf{X}' \mathbf{X} \mathbf{Q}_{1:i}^+ \mathbf{X}' \mathbf{y} \quad (10)$$

Chaque nouveau vecteur  $\mathbf{p}_{i+1}$  est à un coefficient près le produit de  $\mathbf{X}' \mathbf{X}$  par  $\mathbf{X}' \mathbf{y}$  après  $\Sigma$ -orthogonalisation par rapports aux vecteurs  $\{\mathbf{p}_1 \dots \mathbf{p}_i\}$  obtenus précédemment.

## 1.3. Le modèle VODKA-PLSR.

VODKA-PLSR propose de remplacer  $\mathbf{X}' \mathbf{y}$  par tout autre vecteur  $\mathbf{r}$  de même dimension, et de normer les  $\mathbf{p}_i$  selon  $\Sigma$ . L'algorithme est le suivant.

- A l'étape 1:

$$\mathbf{p}_1 = \mathbf{X}' \mathbf{X} \mathbf{r}$$

$$\mathbf{p}_1 \leftarrow \mathbf{p}_1 (\mathbf{p}_1' \Sigma \mathbf{p}_1)^{-0.5}$$

- A l'étape  $i+1$ :

$$\mathbf{p}_{i+1} = \mathbf{Q}_{1:i}^{+'} \mathbf{X}' \mathbf{X} \mathbf{Q}_{1:i}^+ \mathbf{r}$$

$$\mathbf{p}_{i+1} \leftarrow \mathbf{p}_{i+1} (\mathbf{p}_{i+1}' \Sigma \mathbf{p}_{i+1})^{-0.5}$$

Le vecteur  $\mathbf{r}$  permet d'intégrer des informations supplémentaires dans le modèle de régression. D'où le nom: Vector Orientation Decided through Knowledge Assessment-Partial Least Square Regression (VODKA-PLSR). Nous retrouvons NIPALS-P lorsque  $\mathbf{r} = \mathbf{X}' \mathbf{y}$ . Les b-coefficients sont obtenus par une régression aux moindres carrés de  $\mathbf{y}$  sur  $\mathbf{T}$ . Après simplification:

$$\mathbf{b} = \Sigma \mathbf{P} (\mathbf{P}' \Sigma \mathbf{P})^{-1} \mathbf{P}' \Sigma \mathbf{X}' \mathbf{y} \quad (11)$$

## 2- Matériels et méthodes

L'objectif était la quantification de l'éthanol en cours de fermentation par spectrométrie proche infra-rouge entre 500 et 1900 nm. Les spectres disponibles étaient:

- $\mathbf{X}_G$ : 165moûts, ne contenant pas d'éthanol;
- $\mathbf{X}$ : 315moûts en fermentation ou vins, pour l'étalonnage;
- $\mathbf{X}_V$ : 1000moûts en fermentation ou vins pour la validation;
- $\mathbf{k}$ ,  $\mathbf{k}_W$ ,  $\mathbf{k}_G$ ,  $\mathbf{k}_L$ : spectres purs de l'éthanol, de l'eau, du glycérol et de l'acide lactique.

Les valeurs de référence  $\mathbf{y}$  et  $\mathbf{y}_V$  des individus de  $\mathbf{X}$  et  $\mathbf{X}_V$  étaient connues. Les 4 premiers vecteurs propres d'une ACP sur  $\mathbf{X}_G$  donnent  $\mathbf{Q}$ . Une matrice  $\mathbf{R}$  est obtenue en concaténant  $\mathbf{Q}$ ,  $\mathbf{k}_W$ ,  $\mathbf{k}_G$ , et  $\mathbf{k}_L$ . Le NAS-Net Analyte Signal [1] est la projection de  $\mathbf{k}$  orthogonalement à  $\mathbf{R}$ . Différents choix ont été faits pour  $\mathbf{r}$ :  $m1$  vecteur composé de la valeur 1;  $m2$   $\mathbf{X}' \mathbf{1}_N$  colinéaire au spectre moyen;  $m3$   $\mathbf{X}' \mathbf{y}$  NIPALS;  $m4$   $\mathbf{k}$  spectre pur de l'éthanol;  $m5$  NAS;  $m6$  NIPALS sur données centrées. Les 6 modèles ont été construits sur  $(\mathbf{X}, \mathbf{y})$  avec 1 à 20 variables latentes, leurs prédictions sur  $(\mathbf{X}_V, \mathbf{y}_V)$  comparées selon l'erreur standard de prédiction RMSEP.

### 3- Résultats

Les 6 modèles sont tous différents. Leurs valeurs de *RMSEP* sont reportées dans le tableau 1. Deux modèles: *m2* et *m5* ont de meilleures performances que les modèles NIPALS-P centrés (*m6*) ou non (*m3*). Le gain de *RMSEP* est certes modeste, environ 3 p.cent, l'atout majeur des modèles *m2* et *m5* est de présenter une large plage de variables latentes dans laquelle la qualité de prédiction est stable. Au contraire, pour NIPALS-P, la plage optimum est réduite à une seule variable latente.

Modèle	LV4	LV5	LV6	LV7	LV8	LV9	LV10	LV11	LV12	LV13	LV14	LV15
<i>m1</i>	2.09	2.30	2.94	1.43	1.12	1.09	1.08	<b>0.99</b>	<b>0.96</b>	<b>0.97</b>	<b>0.96</b>	1.22
<i>m2</i>	3.16	2.22	2.50	2.23	1.46	<b>0.94</b>	<b>0.93</b>	<b>1.02</b>	<b>0.97</b>	<b>1.01</b>	<b>1.00</b>	1.11
<i>m3</i>	1.81	1.26	1.04	1.03	1.34	<b>1.02</b>	1.38	1.19	1.08	1.19	1.18	1.16
<i>m4</i>	2.26	1.93	2.42	1.88	1.21	<b>1.02</b>	<b>1.01</b>	<b>1.02</b>	1.03	1.03	1.02	1.17
<i>m5</i>	1.04	<b>0.94</b>	<b>0.92</b>	<b>0.92</b>	<b>0.93</b>	<b>0.97</b>	<b>0.99</b>	<b>1.02</b>	1.04	1.04	1.01	1.28
<i>m6</i>	1.23	1.05	<b>1.00</b>	<b>0.95</b>	1.25	<b>1.02</b>	1.40	1.20	1.11	1.23	1.22	1.21

Tableau 1 : Erreurs standard de prédiction (*RMSEP*)

### DISCUSSION ET CONCLUSION

Les modèles d'étalonnage inverse utilisent uniquement de l'information expérimentale. Le modèle VODKA-PLSR offre la possibilité de compléter cette information expérimentale en introduisant de l'information experte comme le spectre pur de la grandeur d'intérêt ou le NAS via le paramètre  $r$ . Le calcul de  $P$  est orienté par  $r$  de manière à identifier au mieux l'espace utile de  $X$ . Plusieurs options pour  $r$  se sont révélées intéressantes, mais beaucoup d'autres choix restent encore possibles. Ainsi VODKA-PLSR ouvre une infinité de modèles de régression de type PLSR issus de NIPALS. Le problème est maintenant d'identifier le vecteur  $r$  optimum. L'utilisation du NAS nous paraît une première piste intéressante puisqu'elle utilise la complémentarité entre informations expérimentales et expertes. Mais d'autres pistes sont envisageables.

### Références

- [1] A.Lorber, K.Faber, B.R.Kowalski. Net analyte signal calculation in multivariate calibration, *Analytical Chemistry*, 69 (8):1620-1626, 1997.
- [2] J.F.Durand. Eléments de calcul matriciel et d'analyse factorielle de données. Université Montpellier II, 2002.
- [3] J.Trygg. Parsimonious multivariate models. PhD thesis, Umea University, Sweden, 2001.