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

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

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

J.C. Boulet[1], D.Bertrand[2], G.Mazerolles[1], R.Sabatier[3], J.M.Roger[4]

[1] INRA, UMR1083, 2 place Viala, F-34070 Montpellier, bouletjc@supagro.inra.fr
[2] INRA, Bioinformatique, rue de la Geraudière, F-44136 Nantes, Dominique.Bertrand@nantes.inra.fr
[3] UM1, EA2415, 15 av. C.Flahault, F-34093 Montpellier, sabatier@univ-montp1.fr
[4] 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 **X**, les valeurs de référence pour une grandeur d'intérêt donnant y de dimension Nx1, enfin un vecteur **r** de dimensions Px1 choisi arbitrairement. Ces entrées sont utilisées pour le calcul des deux paramètres **Σ** et **P**, avec **Σ**=(X'X) + la pseudo-inverse de (**X'X**) au sens de Moore-Penrose.

1- Théorie

1.1 Nouvelle écriture de NIPALS

NIPALS utilise (**X,y**) pour calculer itérativement les A colonnes de trois matrices contenant des scores **T**, des vecteurs **P** et des poids **W** [3]. Par exemple, soit i un indice entre 1et A, si nous notons $P^{\perp}_{1:i}$ le projecteur de $\mathbb{R}^N$ orthogonal à $\{t_1...t_i\}$, à chaque boucle les vecteurs pi sont calculés ainsi:

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

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

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

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

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

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

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

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

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

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

T est la matrice des scores de la projection **Σ**-orthogonale de **X** sur **P.** Les b-coefficients sont déduits après une régression aux moindres carrés de **y** sur **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):

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

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

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

$$\mathbf{P}^{\perp}_{1:i}\,\mathbf{X} = \mathbf{X}\,\mathbf{Q}^{\perp}_{1:i} \tag{8}$$

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:

$$\mathbf{p}_{i+1} = \alpha_{i+1}\,\mathbf{X}'\,\mathbf{P}^{\perp}_{1:i}\,\mathbf{X}\mathbf{X}'\mathbf{P}^{\perp}_{1:i}\,\mathbf{y}$$

$$\mathbf{p}_{i+1} = \alpha_{i+1}\,(\mathbf{P}^{\perp}_{1:i}\,\mathbf{X})'\mathbf{X}(\mathbf{P}^{\perp}_{1:i}\,\mathbf{X})'\mathbf{y} \tag{9}$$

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

$$\mathbf{p}_{i+1} = \alpha_{i+1}\,\mathbf{Q}^{\perp\prime}_{1:i}\,\mathbf{X}'\,\mathbf{X}\,\mathbf{Q}^{\perp\prime}_{1:i}\,\mathbf{X}'\mathbf{y} \tag{10}$$

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 Σ-orthogonalisation par rapports aux vecteurs $\{\mathbf{p}_1 \ldots \mathbf{p}_i\}$ 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 Σ. 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:

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

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

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:

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

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 6modèles 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 meilleurs 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*
m1	2.09	2.30	2.94	1.43	1.12	1.09	1.08	**0.99**	**0.96**	**0.97**	**0.96**	1.22
m2	3.16	2.22	2.50	2.23	1.46	**0.94**	**0.93**	**1.02**	**0.97**	**1.01**	**1.00**	1.11
m3	1.81	1.26	1.04	1.03	1.34	**1.02**	1.38	1.19	1.08	1.19	1.18	1.16
m4	2.26	1.93	2.42	1.88	1.21	**1.02**	**1.01**	**1.02**	1.03	1.03	1.02	1.17
m5	1.04	**0.94**	**0.92**	**0.92**	**0.93**	**0.97**	**0.99**	**1.02**	1.04	1.04	1.01	1.28
m6	1.23	1.05	**1.00**	**0.95**	1.25	**1.02**	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 parait 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.