



HAL
open science

VODKA-PLSR, a family of PLS models based on the NIPALS algorithm

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

► **To cite this version:**

Jean Claude J. C. Boulet, Dominique Bertrand, Gerard Mazerolles, Jean-Michel Roger, Robert Sabatier. VODKA-PLSR, a family of PLS models based on the NIPALS algorithm. CAC2010, Oct 2010, Anvers, Belgium. pp.1, 2010. hal-02750739

HAL Id: hal-02750739

<https://hal.inrae.fr/hal-02750739>

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, a family of PLS models based on the NIPALS algorithm



J.C. BOULET¹, D. BERTRAND², G. MAZEROLLES¹, R. SABATIER³, J.M. ROGER⁴

¹ INRA, UMR1083, F-Montpellier
² INRA, Bioinformatique, F-Nantes

³ UM1, EA2415, F-Montpellier
⁴ CEMAGREF, UMR1201, F-Montpellier



Theory

Including expert information into regression models using NIPALS

A re-writing of NIPALS puts forwards a new parameter, a vector r chosen by the operator. This vector allows the extraction of useful information from X

(1) NIPALS

Known properties:

$$T = X W (P' W)^{-1}$$

$$b = W (P' W)^{-1} c$$

(2) a new writing of NIPALS

New properties:

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

$$c' = y' T (T' T)^{-1}$$

$$\hat{y} = T (T' T)^{-1} T' y$$

$$b = \Sigma P (P' \Sigma P)^{-1} (T' T)^{-1} T' y$$

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

(simplified)

Definitions:

T $N \times A$ scores $\{t_1 \dots t_A\}$

W $P \times A$ weights $\{w_1 \dots w_A\}$

P $P \times A$ loadings of X $\{p_1 \dots p_A\}$

c $A \times 1$ loadings of y

b $P \times 1$ regression vector

New definitions:

Σ $P \times P$ $\Sigma = (X'X)^+$ (Moore-Penrose)

P_i $P \times 1$ loadings of X $\{p_1 \dots p_i\}$

Q_i $P \times P$ $Q_i = I_P - \Sigma P_i (P_i' \Sigma P_i)^{-1} P_i$

r $P \times 1$ $r = X' y$

(3) Vector Orientation Decided through Knowledge Assessment: VODKA- PLSR

3-1: a new calculation of P:

$$p_1 = (X' X)^{-1} (r)$$

loop: $p_{i+1} = (Q_i' X' X)^{-1} (Q_i' r)$

3-2: choice of r

(1) $r = X'y$ => NIPALS (postulate)

(2) $r =$ any vector of dimension P

Expert knowledge can be used for the choice of r

Application

Ethanol quantification in wines and musts

Calibration

Validation: RMSEP

Spectra: 500-1900nm
 k, k_w, k_G, k_L : pure spectra of ethanol, water, glycerol, lactate
 X_G 165 spectra (EtOH=0)
 X 315 spectra (EtOH \neq 0)
 X_v 1000 spectra (EtOH \neq 0)

Processing:

$X_G \rightarrow$ PCA $\rightarrow G = 1-4$ PCs

$$R = [G \ k_w \ k_G \ k_L]$$

$$NAS = (I - R (R'R)^{-1} R')k$$

Model	r choice	Notes
m_1	1_p	
m_2	$X'1_N$	Mean of X spectra
m_3	$X'y$	NIPALS
m_4	k	Pure spectra
m_5	NAS	Net analyte signal
m_6	$X'_c y_c$	NIPALS centered

LV5	LV6	LV7	LV8	LV9	LV10	LV11	LV12	LV13	LV14
2.30	2.94	1.43	1.12	1.09	1.08	0.99	0.96	0.97	0.96
2.22	2.50	2.23	1.46	0.94	0.93	1.02	0.97	1.01	1.00
1.26	1.04	1.03	1.34	1.02	1.38	1.19	1.08	1.19	1.18
1.93	2.42	1.88	1.21	1.02	1.01	1.02	1.03	1.03	1.02
0.94	0.92	0.92	0.93	0.97	0.99	1.02	1.04	1.04	1.01
1.05	1.00	0.95	1.25	1.02	1.40	1.20	1.11	1.23	1.22

m_2 and $m_5 \rightarrow$ better predictions than NIPALS

Discussion and conclusion

Practical aspects

- An infinity of regression models based on NIPALS
- Expert information (e.g. NAS) can be directly introduced into regression models through r
- NIPALS ($r = X'y$) isn't always the best choice

Theoretical aspects

- A more general model depending on the choices of P and Σ

VODKA-PLSR synopsis

