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

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

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

J.C. BOULET [1], D. BERTRAND [2], G. MAZEROLLES [1], R. SABATIER [3], J.M. ROGER [4]

[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 usefull information from X

(1) NIPALS

Known properties:

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

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

(2) a new writing of NIPALS

New properties:

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

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

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

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

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

(simplified)

Definitions:

- **T** NxA scores $\{t_1 \ldots t_A\}$
- **W** PxA weights $\{w_1 \ldots w_A\}$
- **P** PxA loadings of **X** $\{p_1 \ldots p_A\}$
- **c** Ax1 loadings of **y**
- **b** Px1 regression vector

New definitions:

- **Σ** PxP $\Sigma = (\mathbf{X}'\mathbf{X})^{+}$ (Moore-Penrose)
- $\mathbf{P}_i$ Pxi loadings of **X** $\{p_1 \ldots p_i\}$
- $\mathbf{Q}_i$ PxP $\mathbf{Q}_i = \mathbf{I}_P - \Sigma\,\mathbf{P}_i\,(\mathbf{P}'_i\,\Sigma\,\mathbf{P}_i)^{-1}\,\mathbf{P}_i$
- **r** Px1 $\mathbf{r} = \mathbf{X}'\,\mathbf{y}$

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

3-1: a new calculation of P:

$\mathbf{p}_1 = (\mathbf{X}'\,\mathbf{X})\,(\mathbf{r})$

loop: $\mathbf{p}_{i+1} = (\mathbf{Q}'_i\,\mathbf{X}'\,\mathbf{X})\,(\mathbf{Q}'_i\,\mathbf{r})$

3-2: choice of r

(1) $\mathbf{r} = \mathbf{X}'\mathbf{y}$ => 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

<u>Spectra:</u> 500-1900nm

$\mathbf{k}, \mathbf{k}_w, \mathbf{k}_G, \mathbf{k}_L$: pure spectra of ethanol, water, glycerol, lactate

$\mathbf{X}_G$ 165 spectra (EtOH=0)

$\mathbf{X}$ 315 spectra (EtOH ≠ 0)

$\mathbf{X}_v$ 1000 spectra (EtOH≠ 0)

<u>Processing:</u>

$\mathbf{X}_G \rightarrow$ PCA $\rightarrow$ $\mathbf{G}$= 1-4 PCs

$\mathbf{R}=[\mathbf{G}\ \mathbf{k}_w\ \mathbf{k}_G\ \mathbf{k}_L]$

$\mathrm{NAS}=(\mathbf{I}-\mathbf{R}\,(\mathbf{R}'\mathbf{R})^{-1}\,\mathbf{R}')\mathbf{k}$

Calibration / Validation: RMSEP

Model	r choice	Notes	LV5	LV6	LV7	LV8	LV9	LV10	LV11	LV12	LV13	LV14
m_1	$\mathbf{1}_P$		2.30	2.94	1.43	1.12	1.09	1.08	0.99	0.96	0.97	0.96
m_2	$\mathbf{X}'\mathbf{1}_N$	Mean of **X** spectra	2.22	2.50	2.23	1.46	**0.94**	**0.93**	1.02	0.97	1.01	1.00
m_3	$\mathbf{X}'\mathbf{y}$	NIPALS	1.26	1.04	1.03	1.34	1.02	1.38	1.19	1.08	1.19	1.18
m_4	$\mathbf{k}$	Pure spectra	1.93	2.42	1.88	1.21	1.02	1.01	1.02	1.03	1.03	1.02
m_5	NAS	Net analyte signal	**0.94**	**0.92**	**0.92**	**0.93**	0.97	0.99	1.02	1.04	1.04	1.01
m_6	$\mathbf{X}'_c\,\mathbf{y}_c$	NIPALS centered	1.05	1.00	**0.95**	1.25	1.02	1.40	1.20	1.11	1.23	1.22

m2 and m5 → 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