Application and properties of O-PLS method

T. Verron,[a] R. Sabatier[b] and R. Joffre [a]

[a] *Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, 1919 route de Mende, F-34293 Montpellier Cedex 5, France. E-mail: thomas.verron@cefe.cnrs-mop.fr*

[b] *Laboratoire de Physique Moleculaire et Structurale, Faculte de Pharmacie, 15 Avenue Charles Flahault, BP 34093 Montpellier Cedex 5, France. E-mail : sabatier@pharma.univ-montp1.fr*

Introduction

The near-infrared spectroscopy (NIRS) is a widely adopted method for performing analytical measurement. One of the principal application of NIRS is to determine a linear relationship between a set of calibration data and a set of concentration component, with the use of multivariate calibration,[1]. However, it's well know that the NIR-spectra contain undesirable variation due to physical properties such as light scattering and particle size, and irrelevant information to the response matrix,[2]. To correct this undesirable variation and to improve multivariate calibration various preprocessing methods, such as orthogonal signal correction (OSC),[3] have been proposed in analytical chemistry literature. Recently, Trygg and Wold have developed a new modification of PLS methodology: O-PLS,[4]. Our main aim is to measure the influence of this preprocessing method on the PLS model when the response is a vector.

Material and method

A long-term field experiment on a soil at Uppsala, Swenden,[5,6] was designed to study the effects of various inorganic fertilisers and manurial treatments on soil structural properties and organic matter changes. We focused our analysis on carbon concentration (response **y**) and near-IR spectra (spectral matrix **X**). To measure the influence of O-PLS method, the root mean square error of prediction (RMSEP), had been used. For a number of I samples, this coefficient RMSEP is defined as

$$\sqrt{\frac{1}{I}\sum_{r=1}^{I}(y_r - \hat{y}_r)^2} \qquad (1)$$

where y_r is the reference value and $\hat{y}_r$ the predicted value for the test sample *r*.

O-PLS method

We name O-PLS method, the research of $\mathbf{w}^{\perp}$ which maximises:

$$\operatorname{cov}(\mathbf{t}, \underbrace{\mathbf{X}\mathbf{w}^{\perp}}_{\mathbf{t}^{\perp}})^2 \quad \text{under the conditions } \left\|\mathbf{w}^{\perp}\right\| = 1 \text{ and } \mathbf{t}^{\perp} \perp \mathbf{y}$$

The difficulties of this method are to choose the number of:

- PLS component to correct,
- O-PLS components to compute for each PLS component.

In order to compare the different possibilities, various combinations of PLS and O-PLS components were tested.

Results

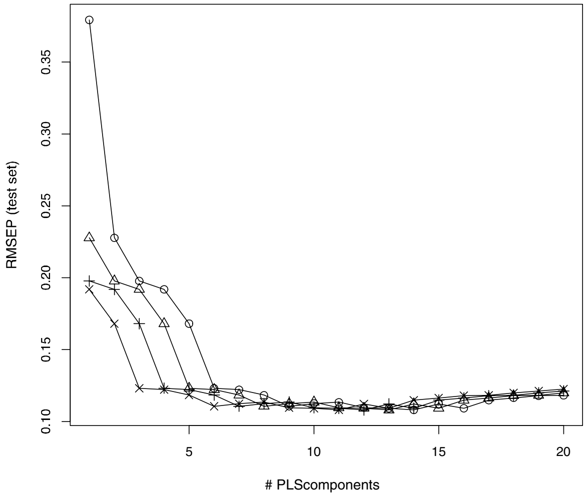


Figure 1. Influence of the number of PLS and O-PLS components on RMSEP of carbon concentration. ○ =PLS(OPLS(1;1),y), Δ =PLS(OPLS(1;2),y), + =PLS(OPLS(1;3),y) and ×=PLS(OPLS(1;4),y).

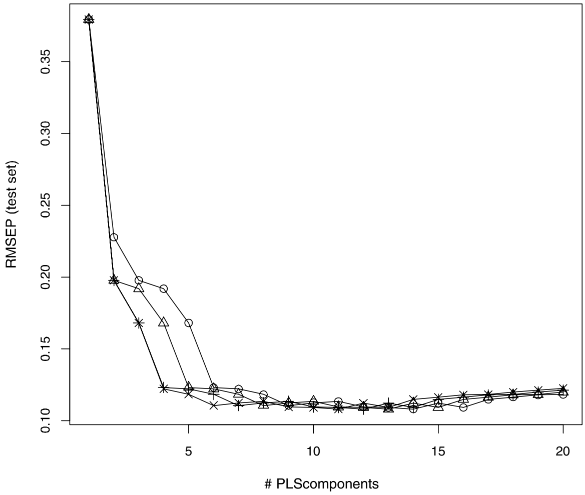


Figure 2. Influence of the number of PLS and O-PLS components on RMSEP of carbon concentration. ○=PLS(OPLS(1;1),y), Δ=PLS(OPLS(2;1,1),y), + =PLS(OPLS(3;1,1,1),y) and ×=PLS(OPLS(4;1,1,1,1),y).

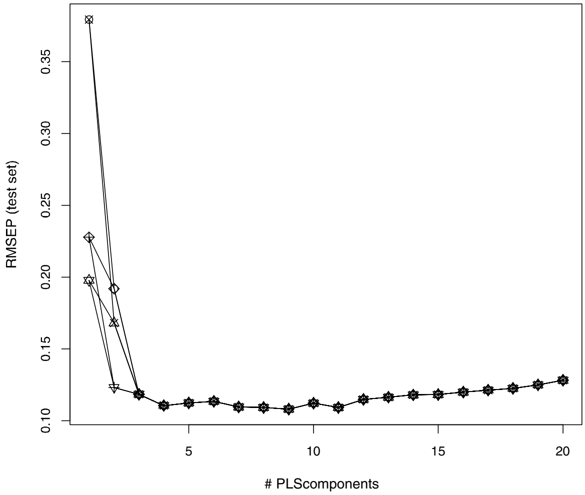


Figure 3. Influence of the number of PLS and O-PLS components on RMSEP of carbon concentration. ○=PLS(OPLS(3;1,2,3),y), Δ=PLS(OPLS(3;3,1,2),y), + =PLS(OPLS(3;2,3,1),y), × =PLS(OPLS(3;1,3,2),y), ◊=PLS(OPLS(3;2,1,3),y) and ∇=PLS(OPLS(3;3,2,1),y).

Discussion

Figure 1, shows the RMSEP when one PLS component was corrected with one, two, three or four O-PLS components. The four curves are clearly similar. In fact, every curve is a horizontal translation of the other.

Figure 2, correspond to model with the first, the first two, the first three or the first four PLS component were corrected with one O-PLS components. The figure 2 is the same as the fig.1 for a number of PLS components superior or equal to four.

We have proved the following property:

$$\mathrm{PLS}(\mathrm{OPLS}(1;k),\mathbf{y}) \xleftrightarrow[k]{} \mathrm{PLS}(\mathrm{OPLS}(k;\underbrace{1,\ldots,1}_{k}),\mathbf{y}) \qquad (2)$$

The notation $\xleftrightarrow[k]{}$ means that the PLS model are the same for a number of PLS components superior or equal to k .

In **Figure 3** six models were compared when the three first PLS components are corrected with a total of six O-PLS components. The six curves are identical for a number of PLS components superior or equal to three.

It is possible to generalize the previous equation by the following property:

$$\mathrm{PLS}(\mathrm{OPLS}(a;k_1,\ldots,k_a),\mathbf{y}) \xleftrightarrow[k]{} \mathrm{PLS}(\mathrm{OPLS}(a;k_1',\ldots,k_a'),\mathbf{y}) \qquad (3)$$

$$\text{if and only if } \sum_{i=1}^{a} k_i = \sum_{i=1}^{a} k_i'$$

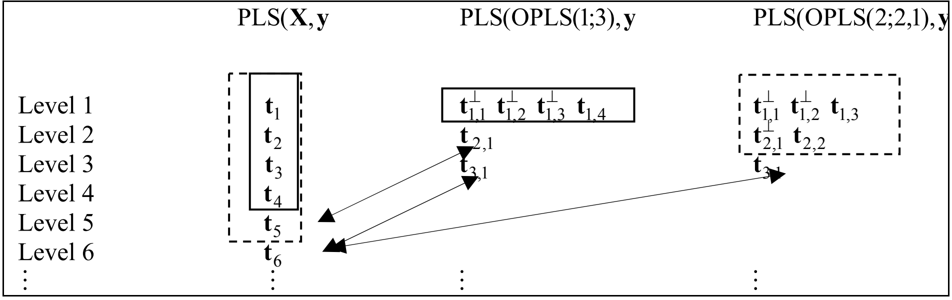
Theory

The main idea of the proof is based on the equality between the space spanned by the PLS components of the initial model and the space spanned by the O-PLS and PLS corrected components. We will denote by

- $\mathbf{t}_j$ the j^{th} PLS component of the initial matrix $\mathbf{X}$,
- $\mathbf{t}_{j,k}$ the j^{th} PLS component corrected by $k-1$ O-PLS components,
- $\mathbf{t}_{j,k}^{\perp}$ the k^{th} O-PLS component computed to correct the j^{th} PLS component.

The following table summarises the principle of the proof :

Table 1. Proof of principle.



	PLS($\mathbf{X}$, $\mathbf{y}$	PLS(OPLS(1;3), $\mathbf{y}$	PLS(OPLS(2;2,1), $\mathbf{y}$
Level 1	$\mathbf{t}_1$	$\mathbf{t}^{\perp}_{1,1}$ $\mathbf{t}^{\perp}_{1,2}$ $\mathbf{t}^{\perp}_{1,3}$ $\mathbf{t}_{1,4}$	$\mathbf{t}^{\perp}_{1,1}$ $\mathbf{t}^{\perp}_{1,2}$ $\mathbf{t}_{1,3}$
Level 2	$\mathbf{t}_2$	$\mathbf{t}_{2,1}$	$\mathbf{t}^{\perp}_{2,1}$ $\mathbf{t}_{2,2}$
Level 3	$\mathbf{t}_3$	$\mathbf{t}_{3,1}$	$\mathbf{t}_{3,1}$
Level 4	$\mathbf{t}_4$		
Level 5	$\mathbf{t}_5$		
Level 6	$\mathbf{t}_6$		
⋮	⋮	⋮	⋮

The vectors in the identical style box define exactly the same space. The vectors links by an arrow are identical. By consequences, the PLS regression model of $\mathbf{y}$ by $\mathbf{X}$, by OPLS(1;3) and by OPLS(2;2,1) respectively with six components, three components and three components are identical.

Conclusion

We have seen that to use O-PLS method, we must to choose the number of PLS component to correct and the number of O-PLS components to compute for each PLS component. In the most of O-PLS papers, only the first PLS component is corrected and empirics methods have proposed to choose the number of O-PLS components to compute. But no mathematical justifications have been proposed to explain these choices.

In this paper, news mathematical properties on O-PLS method are proposed. These properties give a response to these problems in the case where $\mathbf{y}$ is a vector.

First, the property (1) state that it is sufficient to correct only the first PLS component. The correction depends only of the number of O-PLS components to remove.

Second, Each O-PLS components permit to reduce the complexity of the final model of one PLS component. Consequently is possible to choose the number of O-PLS components to compute. This number depends of the number of PLS components that the users want in the final PLS model and the dimension of optimal PLS model given by the cross validation.

Third, we have proved that PLS models prediction quality is preserved when we use the O-PLS method.

So, The O-PLS pretreatment is a method to reduce the number of PLS components but not to increase the prediction quality.

It seems possible to extrapolate these results if $\mathbf{Y}$ is a matrix. Nevertheless, the finals models are not exactly the same, but very similar. To obtain identical models, the PLS components number must be superior to r. This value r depends to the column number of the $\mathbf{Y}$ matrix and the number of O-PLS components used. This is an interesting practical problem but theoretically difficult.

References

1. H. Martens, and T. Naes, *Multivariate Calibration.* (2nd edn), vol. 1. Willey: Chichester, 1989.
2. P.C. Williams and K. Norris, *Near infrared Technology in the Agricultural and Food Industries.* American Cereal Association, St Paul, MN (1987).
3. O. Svensson, T Kourti and JF. Mac Gregor, *J. Chemometrics.* **16**, 176 (2002).
4. J. Trygg, S. and Wold, *J. Chemometrics.* **16**, 119 (2002).
5. H. Kirchmann, J. Persson and K. Carlgren, *The Ultuna long-term soil organic matter experiment, 1956-1991. Departement of Soil Science.* Reports and Dissertation 17, Swedish University of Agricultural Sciences, Uppsala (1994).
6. E. Witter and C. Soil, *Biol Fertil Soils.* **5**, 176 (1996).