Calibration of wine mannoproteins using two successive orthogonal projections

Jean Claude J. C. Boulet, Thierry Doco, Jean-Michel Roger

Calibration of wine mannoproteins using two successive orthogonal projection methods

J.C. Boulet[1], T. Doco[2], J.M. Roger[3]
[1] INRA-UMR1083, 2 Place Viala, F-34060 Montpellier, bouletjc@supagro.inra.fr
[2] INRA-UMR1083, 2 Place Viala, F-34060 Montpellier, doco@supagro.inra.fr
[3] CEMAGREF-UMR ITAP, 361 Rue JF Breton, F-34196 Montpellier, jean-michel.roger@montpellier.cemagref.fr

Keywords: OSC, EPO, SOP, robustness, FTIR, wine, mannoproteins, quantification.

1 Introduction.

Spectroscopic methods require the development of models on an ($\mathbf{X_0}$, $\mathbf{y_0}$) calibration base before application to unknown samples ($\mathbf{X_1}$, $\mathbf{y_1}$). $\mathbf{X}$ is the matrix of the spectra and $\mathbf{y}$ is the vector of the reference values for the analytical parameter Y to be quantified. The quality of a calibration model depend on influence factors G_k that may reduce both the accuracy and robustness of models. The present study addresses chemical influence factors, intrinsic to the sample.

Orthogonal projection methods were designed to overcome the problem of influence factors. We concentrated on two of them: Orthogonal Signal Correction (OSC) [4] and External Parameter Orthogonalisation [3]. We wish to integrate known influence factors in orthogonal projections and this can be performed naturally by EPO or OSC with an experimental design dedicated to this application. We also wish to continue to benefit from the capacity of OSC to minimise influence factors that are not identified but that are present in the calibration set ($\mathbf{X_0}$, $\mathbf{y_0}$). We have proposed new methodology based on two successive orthogonal projection (SOP) levels in order to combine these two objectives [1]. The first level consists of an EPO or an OSC in which we integrate known information on the influence factors G_k. The second level is a classic OSC applied to the resultant of ($\mathbf{X_0}$, $\mathbf{y_0}$) after the previous projection. This approach is validated using an example based on the MIR quantification of mannoproteins (MPs), a polysaccharide family present in wine, taking into account two influence factors of chemical origin belonging to the two other wine polysaccharide families: arabinogalactans-proteins (AGPs) and rhamnogalacturonan II (RG-II).

2 Theory.

2.1. Application of EPO and OSC for characterised chemical influence factors.

It is assumed that we know the pure spectra of the compound of interest and the interference compounds. We propose that EPO and OSC are applied as follows:

a) The case of EPO performed on a single factor of influence. Assuming that the spectral effect of the factor of influence is proportional to its level—as is often the case with chemical influence factors and with application of the Beer-Lambert law—the spectra of $\mathbf{D}$ are perfectly collinear in theory and $\mathbf{D}$ can be reduced to a single one of them. We now take the case of an EPO performed with k influence factors G_1 to G_k. $\mathbf{D}$ is constructed with the k pure spectra of these compounds, and a PCA is performed onto $\mathbf{D}$ giving $\mathbf{P}$. EPO will thus accomplish a projection that is orthogonal to the space containing all the influence factors.

b) A matrix $\mathbf{X_{OSC}}$ comprising the spectrum of Y —the chemical parameter analysed—and the k spectra of the influence factors G_1 to G_k is constructed for OSC to eliminate the influence factors $G_1 \ldots G_k$. Opposite these spectra, a vector $\mathbf{y_{OSC}}$ is constructed containing 1 opposite the spectrum of the factor of interest and 0 elsewhere. Each of the values in $\mathbf{y_{OSC}}$ corresponds to the concentration of Y in the sample.

S_{G1}
.
.
S_{Gi}

S_Y	1
S_{G1}	0
.	.
.	.
S_{Gi}	0

Table 1- Structure of the X_{EPO} (left) and X_{OSC} (right) data used for the first orthogonal projection.

2.2. Successive orthogonal projection (SOP).

The first EPO or OSC described before are applied to the calibration and validation data sets. Onto these new data sets, we apply successively a classic OSC and a PLS.

3 Material and methods.

3.1. Calibration data

The set of calibration data (X_0, y_0) was obtained by mixtures of varied concentrations of 4 purified fractions chosen to represent all wine polysaccharides, consisting of dimer of RG-II, neutral mannoprotein MP0, neutral arabinogalactan-protein AGP0 and acidic arabinogalactan-protein AGP4, prepared using a full three-component mix or Scheffé plan.

3.2. External validation data

The external validation data set (X_1, y_1) was obtained from 20 wines chosen from the INRA-UE Pech Rouge wine collection in such a way as to express maximum variability between samples.

3.3. Spectrum acquisition

Spectral acquisition was performed with a MIR spectrophotometer with an ATR single reflection cell. Only the section of spectrum between 950 and 1850 cm^{-1} was used. The spectra were subjected to standardisation to adjust absorbance values to 0 and 1 for respective wave numbers of 1850 cm^{-1} where there was no absorbance and 1035 cm^{-1} at the highest glucoside bond absorbance peak. The acquisition method induced variability resulting from the quality of the contact between the sample and the germanium crystal. The standardisation performed allowed simple correction of this physical factor of influence.

4 Results and discussion.

In agreement with [2], we checked in our data that PLS without the use of OSC does not give a satisfactory calibration model in spite of the optimum conditions for calibration set development with a full experimental plan. This shows that PLS alone is not capable of finding orthogonal space at influence factors. This rules out the exhaustive calibration approach in the present case.

4.1. Calibration using OSC-PLS.

For purposes of comparison, classic OSC-PLS calibration was performed as described by Coimbra *et al*, [7]. After OSC 2 factors, all the spectra for samples containing mannoproteins very soon acquired typical mannoprotein form. Furthermore, the spectra are grouped in 4 classes corresponding to the MP fractions in

the sample : 0, 0.33, 0.66 and 1. The 4 classes are more homogeneous visually when the number of OSC factors increases.

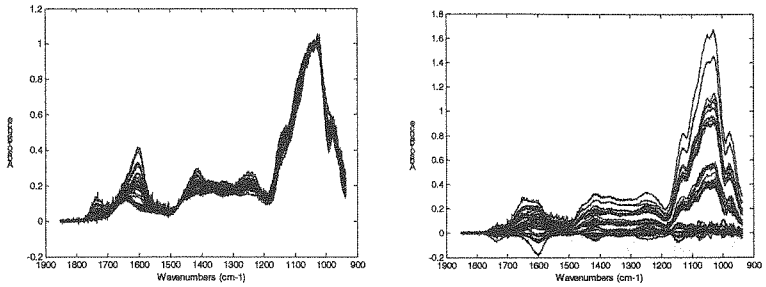

Figure 1- Spectra of $\mathbf{X_0}$ before and after classical OSC.

The model chosen was that with 2-factor OSC and PLS with 1 latent variable. The result of external validation is shown in Figure 4. RMSEP is 0.13, of the same scale of magnitude as the 0.14 obtained [2]. Substantial bias estimated at 0.12 can be seen, resulting in remainders that are always positive.

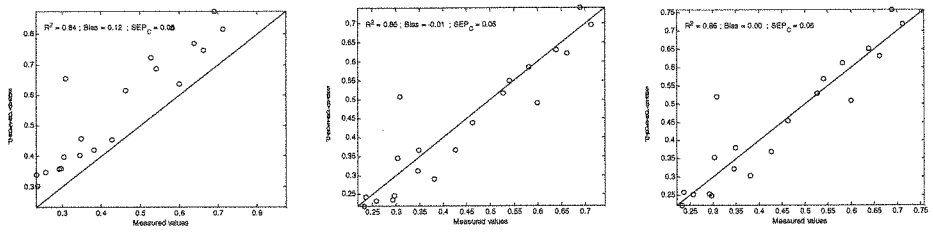

Figure2 - External validation of OSC-PLS, EPO-OSC-PLS and OSC-OSC-PLS models.

4.2. Calibration by EPO-OSC-PLS.

The first orthogonal projection is EPO. As **D** matrix contains 2 lines, we chose 2 factors for the performance of EPO, that is to say the number of factors expected to represent two non-collinear spectra. The spectra from the $\mathbf{X_{0\_EPO}}$ matrix obtained after 2-factor EPO on $\mathbf{X_0}$ have lost the initial typical polysaccharide form. However, they display well-marked peaks in a zone where it is possible to differentiate the three polysaccharide families.

The second orthogonal projection is OSC on the $\mathbf{X_{0\_EPO}}$ matrix. We performed successively EPO with 2 factors, OSC with 1 factor and, finally, PLS with 1 latent variable. The result of external validation is shown in Figure 6. RMSEP at 0.06 is significantly lower than comparable values found with the preceding OSC-PLS combination, that is to say RMSEP 0.13 and bias-corrected RMSEPc 0.08 . The EPO-OSC-PLS combination thus results in a distinct improvement of the calibration model in comparison with the OSC-PLS combination. The classification of spectra in $\mathbf{X_{0\_EPO}}$ reveals 4 MP concentration levels. In contrast, substantial noise also appears, decreasing after OSC (Figure 3), whence the advantage of performing this second orthogonal projection.

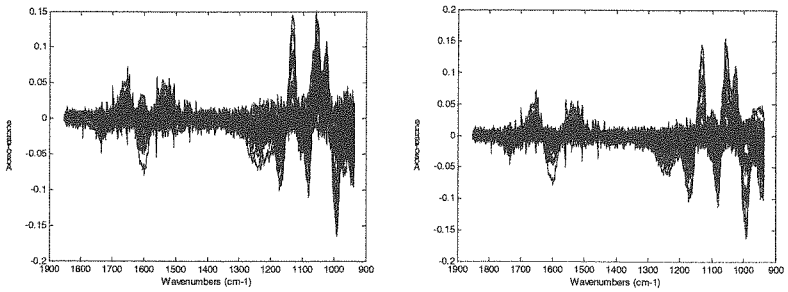

Figure 3 - Calibration data after one EPO ($\mathbf{X}_{0\_EPO}$) and EPO-OSC ($\mathbf{X}_{0\_EPO\_OSC}$)

4.3. Calibration by OSC-OSC-PLS.

The result of external validation is shown in Figure 2. A RMSEP of 0.06 is obtained, the same as in the EPO-OSC-PLS projection. It is also substantially better than that obtained with OSC-PLS reported in paragraph 4.1.

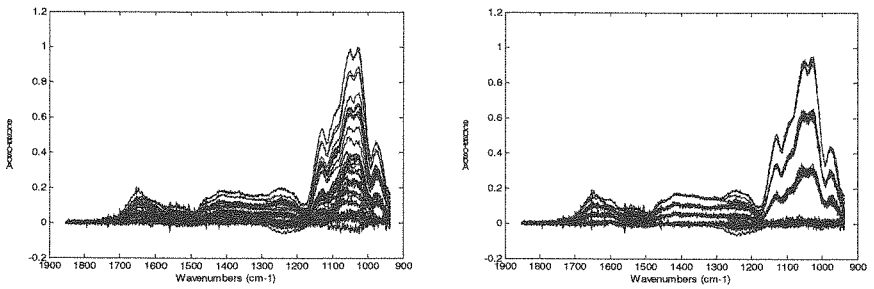

Figure 4 - Calibration data after one OSC ($\mathbf{X}_{0\_OSC}$) and two OSCs ($\mathbf{X}_{0\_OSC\_OSC}$)

Two successive OSCs thus give a better model than a single OSC. Nevertheless, in this particular case, the two OSCs have the same origin. We extracted only 3 spectra, those of pure compounds, from the 40 spectra of RGII, AGPs and MPs in order to perform the first OSC. We also checked that when used alone, this OSC based on 3 spectra performs less well than OSC based on the 40 spectra. When they are compared with a single projection level in the form of OSC-PLS prediction is much better with 40-individual OSC than with 3-individual OSC. This is visible in the quality of the $\mathbf{X}_0$ spectra after each of these two projections (Figure 4). The use of 3-individual OSC as the first stage of projection thus makes it possible to guide projection more clearly in relation to the influence factors, thus allowing a significant gain in RMSEP. The evolution of the $\mathbf{X}_0$ spectra obtained after one classical OSC (Figure 1) and after two OSCs (Figure 4) shows a decrease in noise that is particularly marked in the 1600 cm^{-1} region. This corresponds precisely with an area in which other polysaccharides such as RG-II display a distinctly higher absorbance peak than MPs. Thus, a single projection level per OSC left a residue of information related to influence factors in $\mathbf{X}_{0\text{-}OSC}$. However, two projection levels made it possible to minimise or eliminate from $\mathbf{X}_{0\text{-}OSC\_OSC}$ the residual information linked with AGP and RG-II influence factors.

4.4. Similarities between the first EPO and the first OSC.

We observed that EPO-OSC-PLS and OSC-OSC-PLS gave us quite exactly the same external validation. As the difference between these two models depends of the first orthogonalisation step, we can conclude that we have the same information about the interest factor *Y* after the first EPO or the first OSC. In the first case the

spectra contain only the information relevant to Y, corresponding to the Net Analyte Signal, in the second one the spectra shapes were kept too. It can be explained by the structure of the boolean reference values used in the first OSC. During the OSC calculation, the **T** matrix orthogonal to **y** will have 0 values in regard to the spectra of Y, and then all the variability in **X** will be extracted from the influence factors G_i. That's why the first OSC and the first EPO have the same prediction ability.

EPO is based on a single **P** matrix, as OSC is based on two matrix: **W** et **P**. In this particular case, we observed that the **P** matrix of both EPO and OSC are exactly the same.

5 Conclusion.

The study shows that EPO or OSC can be used to take into account several known and quantified chemical influence factors. In chronological order, at first step, EPO or OSC projectors are computed on the pure spectra of all the influence factors –for OSC, the **y** vector encodes, as a boolean, the Y concentrations-. This projection is applied on the calibration database. It allows solely for known influence factors in order to minimize their effect. EPO or OSC projections are defined using small sets of data specially designed for this purpose. Then, a second and classical OSC projection acts for minimizing influence factors that are unknown but present in the calibration data. Finally, quantification is performed with PLS. The first projection is performed in a subspace optimised with regard to the influence factors and so the second projection OSC and PLS have more relevant information to process. When OSC is used at the first step, the spectra obtained resemble the spectrum of the factor of interest Y. When EPO is used at the first step, the spectra obtained are the part of the calibration spectra that differ from the influence factors. The spectra obtained are noisier and more difficult to interpret. However, the models obtained require only a small number of dimensions in both cases. This has two practical advantages: (i) it becomes much easier to set the parameters of the model, and (ii) the models with few factors or latent variables are considered to be more robust than those with many. As a result, two successive orthogonal projections (EPO-OSC or OSC-OSC) can give much more satisfactory models than a single orthogonal projection.

6 Références.

[1] J.C. Boulet, T.Doco, J.M. Roger, Improvement of calibration models using two successive orthogonal projections methos. Application to quantification of wine mannoproteins, Chemometrics and Intelligent Laboratory Systems, 87:275-302, 2007.

[2] M.A. Coimbra, A.S. Barros, E.Coelho, F.Gonçalves, S.M. Rocha, I.Delgadillo, Quantification of polymeric mannose in wine extracts by FT-IR spectroscopy and OSC-PLS1 regression, Carbohydrate Polymers, 61:434-440, 2005.

[3] J.M. Roger, F.Chauchard, V. Bellon-Maurel, EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits, Chemometrics and Intelligent Laboratory Systems 66 :191-204, 2003.

[4] S. Wold, H.Antti, F.Lindgren, J. Ohman, Orthogonal signal correction of near-infrared spectra, Chemometrics and Intelligent Laboratory Systems 44 :175-185, 1998.