EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits

J.M. Roger, F. Chauchard, Véronique Bellon Maurel

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V. 2003. Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

EPO - PLS
External Parameter Orthogonalisation of PLS
Application to temperature-independent measurement of sugar content of intact fruits

Jean-Michel ROGER(1), Fabien CHAUCHARD(2), Véronique BELLON-MAUREL(3) *1-Information and Technologies for Agro-processes,*

Cemagref BP 5095, 34033 Montpellier Cedex 1, France
2- Ondalys, www.ondalys.com

Abstract

NIR spectrometry would present a high potential for online measurement if the robustness of multivariate calibration was improved. The lack of robustness notably appears when an external parameter varies - e.g. the product temperature. This paper presents a preprocessing method which aims at removing from the **X** space the part mostly influenced by the external parameter variations. This method estimates this parasitic subspace by computing a PCA on a small set of spectra measured on the same objects, while the external parameter is varying. An application to the influence of the fruit temperature on the sugar content measurement of intact apples is presented. Without any preprocessing, the bias in the sugar content prediction was about 8^o Brix for a temperature variation of 20^oC. After EPO preprocessing the bias is not more than 0.3^o Brix, for the same temperature range. The parasitic subspace is studied by analysing the b-coefficient of a PLS between the temperature and the influence spectra. Further work will be achieved to apply this method to the case of multiple external parameters and to the calibration transfer issue.

1 Introduction

Near Infrared spectrometry (NIR) is a powerful analytical tool widely used in routine laboratory in applications as various as food processing, pharmaceutical products ([1]), chemical industry ([2]). Although on-line applications are much fewer, NIR spectroscopy offers numerous advantages for on-line implementation ([3]) : there is no sampling and no sample preparation, NIR radiations can penetrate through thick samples ; ultrafast measurements are possible thanks to multichannel detectors and last, optical components and

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

instrumentation are low-cost. However, NIR spectra are made up of harmonics and combinations of fundamental absorption bands of the Mid Infrared (MIR) range and therefore, are very encumbered with numerous overlapping bands. As a consequence, multivariate data processing is often the only way for building a calibration : a model exhibits relations between spectra ($\mathbf{X}$ matrix of predicting values) and the quantities to be predicted, also called "reference values", which are in general concentrations ($\mathbf{Y}$ matrix of responses) [4]. The multivariate calibration technique most used for processing NIR data is undoubtedly Partial Least Square Regression (PLS) [5], because it handles full spectra or continuous parts of spectra.

1.1 Robustness of NIR models

However, the models generated by PLS generally suffer from a lack of robustness with regard to "influence quantities", which hinder their use in industrial conditions. Influence quantities also called "external parameters" are quantities different from the measurand but which affect the result of a measurement ([6]). Whereas external parameters are well controlled in routine laboratories, they can vary greatly in industrial conditions and alter the measured spectra. From a mathematical point of view, if the model is linear and if the external parameter level is stable during the test, a bias appears in the predicted values. A systematic alteration of the spectrum, when processed by the model, causes a systematic value to be added to the prediction. If the external parameter is correlated to the response, the trend line joining the actual values to the predicted values presents a slope different from the unity. If the influence parameter is not stable, the parasitic information could appear like a noise and results in higher variance of the prediction error (pure scattering).

Product temperature, spectrometer temperature, stray light, wavelength shifts (in particular when model must be transferred from one instrument to another) are most common disturbing external parameters. Other influencing phenomena are embedded into the product as for instance, fruit variety, oil origin, crop season, etc. Product temperature is certainly the most studied parameter in NIR and MIR calibrations ([7], [8], [9], [10], [11], [12]).

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

1.2 Correcting strategies

Correcting strategies - i.e. strategies to reduce the effect of external parameters - vary depending on the availability - or not - of the external parameter value. If it is known or measured, two options are possible : (i) the spectrum is corrected as a function of the parameter level, as proposed by [13] or the model is chosen according to the external parameter level ; this is called *a priori* correction, or (ii) the value of the predicted response is corrected on the basis of the external parameter level, and therefore, this method is called *a posteriori* correction. If the external parameter value is unknown, which is the most common case, the only solution is to build a so-called "robust model". In that case again, two major ways are available to make the model robust : (i) the optimisation of the calibration sample basis and (ii) the preprocessing techniques.

The first way, i.e. optimisation of the calibration sample basis, is also the most straightforward one. "Robust calibration" is based on the use of an exhaustive calibration set covering all the variations both of the response of interest but also of the external parameters, influence of which must be eliminated. During the calibration phase, the model is built automatically to be as insensitive as possible to the external parameter influence ([8]). This smart building of the calibration set can be improved by experimental designs ([14]) or even by more refined methods based on a ruggenedness test such as in [15]. Wulfert et al ([12]) also tried to improve this method by simultaneously predicting the response value and the external parameter value, based on the PLS2 algorithm ([5]). However, results were disappointing because external parameter values and response values were not correlated, therefore reducing the performance of the PLS2 algorithm.

The second way, i.e. preprocessing methods, includes very classical methods, such as the derivatives (for instance using the Savitsky Golay algorithm as in [16]) or geometric transformations, as SNV ([17] et [18]) and Multiplicative Scattering Correction ([4]) developed to address the problem of scattering and its multiplicative effects ; data can also be filtered as in variable selection ([19], [20], [21]) ; whatever the method, variable selection aims at making the model more parcimonious ([22]). Reducing the number of variables has a positive effect on error propagation in the model ([23]). More generally, reducing the dimensionality of $\mathbf{X}$ matrix before any calibration is a major pre-processing step to make the model more robust.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

1.3 EPO : A method to reduce the space dimensionality with regard to external parameters

The method proposed here, External Parameter Orthogonalisation (EPO), deals with the case where the external parameter can not be measured on-line. Spectra are pre-processed by projection onto the orthogonal to the space in which alterations induced by the external parameter variations occur. As seen above, the most straightforward way is to select variables, but a recent trend in chemometrics is devoted to reducing the $\mathbf{X}$ dimensionality by orthogonal projection. The general principle, which is the theoretical basis, is that the column space of $\mathbf{X}$ is made up of the sum of two subspaces, among which only one contains information useful to the model. By an adequate projection of $\mathbf{X}$, the model is therefore created using the useful subspace only. By the way, selecting variables is a particular case of subspace projection, in the canonical basis. The parasitic subspace can be estimated in two ways, either by finding the space which is orthogonal to $\mathbf{Y}$, or by finding the space in which the influence of external factors occurs.

Various papers describe the first way. Several projection methods have been recently proposed for improving the PLS performance. In [24], the orthogonal signal correction (OSC) is described as a filtering method. At each step of the NIPALS algorithm, the score vector $\mathbf{t}$ is corrected from its part orthogonal to $\mathbf{Y}$, giving $\mathbf{t}^*$. This processing is said to be very efficient as soon as the spectrum contains systematic variations. In [21], the OSC algorithm is altered in order to build $\mathbf{t}^*$ into the column space of $\mathbf{X}$. Orthogonal projection to Latent Structures (O-PLS) [25] uses the NIPALS algorithm to built the subspace orthogonal to $\mathbf{Y}$. This subspace is then split up by a PCA. Principal components are then removed one by one from the $\mathbf{X}$ matrix.

The second way is less studied. Hansen ([26]) built a data set using a very large number of spectra having the following specificity : all the spectra came from samples having the same level for the response of interest but showing very large variations of the external parameters. The parasitic subspace is determined by PCA on such sample sets. The pre-processing technique deals with projecting the spectra onto the orthogonal of this subspace.

The EPO method also belongs to this "second way". The parasitic subspace is estimated, not by orthogonalising with regard to $\mathbf{Y}$, but by taking into account the effects of the major external parameter. Although being close to Hansen's method, this method only needs a small sample set and does not require the response of interest to be constant. It is particularly

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

fitted to improving the robustness of an existing calibration - i.e. to take advantage of an existing data base - with regard to a particular external parameter. In addition to finding the adequate subspace, this method makes it possible to interpret the influence of the external factor on the spectrum. In this paper, the method is first introduced by a theoretical part. Second, it is applied to robustness enhancement in a real case : reducing the temperature influence on sugar content prediction in fruits. Results are presented and discussed and tracks are explored for tuning and refining this method.

2 Theory and notations

Capital bold characters will be used for matrices, e.g. $\mathbf{X}$; small bold characters for column vectors, e.g. $\mathbf{x}_i$ will denote the i^{th} column of $\mathbf{X}$; row vectors will be denoted by the transpose notation, e.g. $\mathbf{x}_j^T$ will denote the j^{th} row of $\mathbf{X}$; non bold characters will be used for scalars, e.g. matrix elements x_{ij} or indices i.

Let C be the property to predict (e.g. a chemical concentration) ; G an external parameter, the effect of which has to be eliminated (e.g. the temperature of the product). Let $\vec{S}$ be the p-dimensional space of the measured spectra. The space $\vec{S}$ can be split up into three subspaces : $\vec{C}$ spanned by the chimical spectral responses independent from G ; $\vec{G}$ generated by the perturbations caused by G and independent from C ; and $\vec{R}$ containing the rest of the spectral information. The dimensions of these subspaces are resp. c, g and r and $p = c+g+r$. If C and G are independent, $\vec{R}$ contains independent residuals (measurement noise). In practice, $\vec{R}$ contains the data correlated to both C and G (the co-variant part). The EPO proposes to remove a part of the perturbations caused by G by projecting the spectra onto the subspace orthogonal to $\vec{G}$; i.e. onto $\vec{C} \oplus \vec{R}$. If the influence of G on the spectra was perfectly known (e.g. if a temperature increase caused an absorbance shift), one should be able to build the space $\vec{G}$. This knowledge is rarely available, and the subspace decomposition must be calculated by a calibration process.

A set of spectra $\mathbf{X}$ can be written as :

$$\mathbf{X} = \mathbf{XP} + \mathbf{XQ} + \mathbf{R} \tag{1}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the matrices of the projection operators onto $\vec{C}$ and $\vec{G}$, and $\mathbf{R}$ is a residual matrix. Normally, when G does not vary, an inverse

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

calibration model (e.g. a linear regression) is invoked to separate the matrices $\mathbf{P}$ and $\mathbf{R}$. The EPO processing aims at separating $\mathbf{XP}$ from $\mathbf{XQ}$. In other words, $\mathbf{X}$ will be split up into a useful part $\mathbf{X}^* = \mathbf{XP}$ and a parasitic part $\mathbf{X}^\dagger = \mathbf{XQ}$.

If it is possible to calculate an estimation $\widehat{\mathbf{G}}(g \times p)$ of a basis of the space $\vec{G}$, an estimation of $\mathbf{Q}$ will be given by :

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{G}}\left(\widehat{\mathbf{G}}^T\widehat{\mathbf{G}}\right)^{-1}\widehat{\mathbf{G}}^T$$

The EPO preprocessing will then transform $\mathbf{X}$ into $\mathbf{X}^*$ by :

$$\mathbf{X}^* = \mathbf{X}\left(\mathbf{I} - \widehat{\mathbf{Q}}\right)$$

Let $(\mathbf{X}^0, \mathbf{Y}^0)$ be a set of n^0 samples, aquired while G was remaining constant. This set is used to calibrate a PLS model which has to be robustified with respect to G. Let $\{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^k\}$ be k matrices $(n \times p)$ of n spectra by p wavelengths acquired on another set of n samples for k values of G $\{g^1, g^2, \cdots, g^k\}$. These values do not need to be known, but they must be representative of the $\vec{G}$ space. Let $\mathbf{M}$ be the matrix $(k \times p)$ of the k spectra averaged on $\{\mathbf{X}^i\}_{i=1\cdots k}$:

$$\mathbf{m}_i^T = \frac{1}{n}\sum_{j=1}^{j=n}\mathbf{x}_j^{iT}$$

Or, by using the equation 1 splitting :

$$\mathbf{m}_i^T = \mathbf{m}_i^T\mathbf{P} + \mathbf{m}_i^T\mathbf{Q} + \frac{1}{n}\sum_{j=1}^{j=n}\mathbf{r}_j^{iT}$$

Let $\mathbf{D}$ be the matrix $(k \times p)$ of the k influence spectra, defined by $\mathbf{d}_i^T = \mathbf{m}_i^T - \mathbf{m}_1^T$ $(\mathbf{d}_1^T = \mathbf{0})$. From the dimensionnality point of view, this operation is equivalent to a mean centering, but it allows us to clearly view the effect of G if g^i are ordered. $\mathbf{D}$ is calculated as :

$$\mathbf{d}_i^T = (\mathbf{m}_i^T - \mathbf{m}_1^T)\mathbf{P} + (\mathbf{m}_i^T - \mathbf{m}_1^T)\mathbf{Q} + \frac{1}{n}\sum_{j=1}^{j=n}\left(\mathbf{r}_j^{iT} - \mathbf{r}_j^{1T}\right)$$

Since all $\mathbf{m}_i^T$ are the mean spectra of the same chemicals, $\mathbf{m}_i^T\mathbf{P} = \mathbf{m}_j^T\mathbf{P}$, and consequently $(\mathbf{m}_i^T - \mathbf{m}_1^T)\mathbf{P} = \mathbf{0}$. Then :

$$\mathbf{d}_i^T = (\mathbf{m}_i^T - \mathbf{m}_1^T)\mathbf{Q} + \frac{1}{n}\sum_{j=1}^{j=n}\left(\mathbf{r}_j^{iT} - \mathbf{r}_j^{1T}\right)$$

Finally, in a matrix form :

$$\mathbf{D} = \mathbf{AQ} + \mathbf{R}'$$

Since $\mathbf{R}'$ is the mean value of residuals, we propose to estimate $\widehat{\mathbf{G}}$ by computing the PCA of $\mathbf{D}$, thus writting :

$$\mathbf{D} = \sum_{i=1}^{i=g} \mathbf{t}_i \widehat{\mathbf{g}}_i^T + \mathbf{R}''$$

Each column $\mathbf{g}_i$ is a vector of the basis $\widehat{\mathbf{G}}$. Since the $\mathbf{g}_i$ vectors are the principal components of a PCA, they are orthogonal and of unitary length. Then, $\mathbf{g}_i^T\mathbf{g}_j = 0$ for $i \neq j$ and $\mathbf{g}_i^T\mathbf{g}_i = 1$. Thus $\widehat{\mathbf{G}}^T\widehat{\mathbf{G}} = \mathbf{I}$ and $\widehat{\mathbf{Q}} = \widehat{\mathbf{G}}\widehat{\mathbf{G}}^T$. Finally, $\mathbf{X}^{0*}$ is calculated as :

$$\mathbf{X}^{0*} = \mathbf{X}^0\left(\mathbf{I} - \widehat{\mathbf{G}}\widehat{\mathbf{G}}^T\right)$$

Once g is chosen and $\widehat{\mathbf{G}}$ is identified, a calibration is computed between $\mathbf{X}^{0*}$ and $\mathbf{Y}^0$. A new example $\mathbf{x}_{new}$ will be preprocessed like $\mathbf{X^0}$, i.e. by :

$$\mathbf{x}_{new}^{*T} = \mathbf{x}_{new}^T\left(\mathbf{I} - \widehat{\mathbf{G}}\widehat{\mathbf{G}}^T\right)$$

From $\mathbf{D}$, it is possible to calculate at the most $k-1$ principal components. Then, the EPO preprocessing can be applied with a number of components g varying from 1 up to $k-1$. Determining the optimal value of g may be donne using several methods. Two of them are proposed below.

A cross validation on $\{\mathbf{X}^i\}_{i=1\cdots k}$, with k splits defined according to G values, produces an error as a function of EPO component number and PLS latent variable number : $SECV(g, n_{LV})$. It should be noted that this method obviously requires knowing the responses corresponding to $\{\mathbf{X}^i\}_{i=1\cdots k}$. This condition is not at all needed by the EPO preprocessing itself.

Another way, which does not require to know the values of the responses, adresses the effect of the EPO preprocessing on the similarity between all the spectra of the same sample. Without preprocessing, the model fails because

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

the spectra of the same sample measured at two different values of G can mismatch more than the spectra of two different samples measured at the same value of G. From a classification point of view, the n clusters of k spectra of the same chemical sample do not well separate. This cluster separation can be measured by the Wilks' Λ which expresses the ratio between inter-group variance and total variance. Wilks' Λ can be calculated with several formula. If the number of individuals is less than the number of variables (which is often the case in spectrometry), one has to chose a calculation which is resistant to the rank deficience. For example : $\Lambda = trace(\mathbf{B})/trace(\mathbf{T})$, where $\mathbf{B}$ is the inter-group variance-covariance matrix (each group is replaced by its centre of mass weighted by its size) and $\mathbf{T}$ is the total variance-covariance matrix. A value of 1 for Λ expresses a perfect separation (each group exactly matches its center of mass). A value of 0 for Λ reveals a null separation (all the centers of mass are confounded). We thus propose to observe the evolution of the Wilks' Λ as a function of g, to select the optimal g value.

3 Material and methods

3.1 Building the data basis

The fruits used in this study were apples, of Golden Delicious variety. They all came from the same cultivar and the same harvest. NIR spectra were collected at a precise location of apple surface using a NIR spectrometer (MMS1, ZEISS, Germany). These spectra covered the range from 300 to 1100 nm, using $p = 256$ equally spaced wavelengths. At each fruit surface location, a sample of the juice was collected and its sugar content level was measured by refractometry (Refractometer EUROMEX RD 645, precise at ± 0.2 Brix). In order not to interfere with the method presented here, no pre-processing has been carried out on the spectra. For instance, the whole wavelength range has been used. No optical reference was used in order to avoid any other external influence. Therefore, the study was carried out on intensity spectra.

Three data sets S^0, S^1, S^2 were used in this study.

S^0 was made up of $n^0 = 80$ fruits measured at ambient temperature (25^oC), providing a $(n^0 \times p)$ matrix $\mathbf{X}^0$ of spectra, and a $(n^0 \times 1)$ vector $\mathbf{y}^0$, containing sugar content values.

S^1 and S^2 were made up of $n = 10$ fruits measured at $k = 8$ different

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

temperatures: $\mathbf{t} = \{$ 5, 10, 15, 20, 25, 30, 35, 40 $\}^o$C. Initially kept at 4^oC, the fruits were taken to these temperatures as follows : A water bath was set at $t^1 = 5^o$C and the fruits were immersed in it for 30 minutes. Then, the ten fruits were measured one by one by the spectrometer, as quickly as possible, in order to avoid any temperature variation. They were then plunged back into the water bath, temperature of which was increased by 5^oC-steps; the operation described above was repeated for each temperature step. After this series of measurements, the fruits was let for 30 minutes and their sugar content levels were measured. Therefore, S^1 and S^2 both contained k $(n \times p)$ matrices, respectively $\{\mathbf{X}^{1,i}\}_{i=1\cdots k}$ and $\{\mathbf{X}^{2,i}\}_{i=1\cdots k}$ and both a $(n \times 1)$ vector, resp. $\mathbf{y}^1$ and $\mathbf{y}^2$, of sugar content levels. Let $\mathbf{z}^2$ be the $(nk \times 1)$ vector in which the $\mathbf{y}^2$ vector is concatenated k times. S^1 was used to carry out the orthogonalisation, using the EPO algorithm and S^2 for the test.

S^0, S^1 and S^2 sets were acquired on separate days.

3.2 Mathematical processing

First, a sugar PLS model was calibrated using S^0, without any pre-processing. The number of latent variables, (n_{LV0}), was determined by leave-one out cross-validation. This model was applied onto S^2, in order to provide a (rough) non-corrected estimate $\widehat{\mathbf{z}^2}$.

Then EPO was applied. To determine the appropriate number of components, g and n_{LV}, the two methods described in section 2, i.e. cross-validation and Wilks' Λ have been applied to S^1. The number g has been used for the S^0 and S^2 spectra preprocessing and n_{LV} has been used during the prediction of the EPO-corrected estimate $\widehat{\mathbf{z}^{2*}}$. The RMSEP and the biases have been calculated for the raw and the corrected prediction, as follows :

$$RMSEP_{raw} = \sqrt{\frac{1}{nk}\left(\widehat{\mathbf{z}^2} - \mathbf{z}^2\right)^T\left(\widehat{\mathbf{z}^2} - \mathbf{z}^2\right)}$$

$$Bias_{raw}(t_i) = \frac{1}{n}\sum_{j=(i-1)n+1}^{j=in}\left(\widehat{\mathbf{z}_j^2} - \mathbf{z}_j^2\right)$$

$$RMSEP_{corr} = \sqrt{\frac{1}{nk}\left(\widehat{\mathbf{z}^{2*}} - \mathbf{z}^2\right)^T\left(\widehat{\mathbf{z}^{2*}} - \mathbf{z}^2\right)}$$

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

$$Bias_{corr}(t_i) = \frac{1}{n} \sum_{j=(i-1)n+1}^{j=in} \left(\widehat{\mathbf{z}_j^{2*}} - \mathbf{z}_j^2 \right)$$

4 Results and discussion

4.1 Results without any correction

In order to show the typical peaks of a fruit in a NIR spectrum, the second derivative of the average spectrum of $\mathbf{X}^0$ has been computed (figure 1) using the Savitsky Golay algorithm, with a 21 point-width window and a polynomial of degree 3. The chlorophyll peak clearly appears at 680 nm, water peaks are visible at 760 and 960 nm, and at 838 nm appears a peak typical for water and sugar and related to combinations of O-H bond vibration modes. The evolution of the SECV computed on S^0 is reported in figure 2. As the SECV really drops when 10 latent variables are used, n_{LV0} was set to 10.

The $RMSEP_{raw}$ computed with the non corrected spectra is as high as 4.68^oBrix. This huge error is mainly due to bias depending on the temperature as shown in figure 3. For each temperature, the prediction values are well aligned in parallel to the first bisecting line but they are shifted by a bias value. This bias value linearly varies with the temperature level (figure 4). It can be as high as 8^o Brix and the Root Mean Square of the bias is 4.63^o Brix. The lowest bias ($-0,14^o$ Brix) was obtained for an experimental test temperature close to the calibration temperature.

Influence spectra ($\mathbf{D}$ matrix) show the effect of temperature onto the spectrum (figure 5). There is no direct relationship between the temperature t_i and the height of the influence spectrum $\mathbf{d}_i^T$. In order to analyse the influence of the temperature on the spectra - and since we knew the values of $\{t_i\}$ - a PLS regression was carried out between the influence spectra $\mathbf{D}$ and the temperature $\mathbf{t}$. The b-coefficient spectrum (figure 6) outlines the influence of the temperature on the main bands of the influence spectrum : the major b-coefficient feature is a pseudo-sinusoid centred at 760 nm and covering the $[700, 820]$ region. In spectrometry, pseudo-sinusoid are very typical b-coefficient features : when convoluted to the spectrum, they enhance the effect of the horizontal translation of the spectrum bands (figure 7). Therefore, this pseudo-sinusoid clearly shows that the effect of the temperature on fruit NIR spectra is a translation of the band at 760nm, i.e. of the water band. Such a translation of the water peak, due to the alteration of vibration

energy of the water molecular bonds, has already been described by Osborne and Fearn ([27]). The same feature occurs in the $860-1000$ nm range, and is centred on 915 nm. This absorption band, that may be related to CH bonds, is typical of the sugar and more generally of carbohydrates ([28]) ; it is not clearly seen in the fruit spectra. Last, a negative peak is centred at 838 nm in the b-coefficient spectrum. This peak indicates a relationship between the temperature and the area of the negative peaks centred at 838 nm (typical for OH combination bands) in the influence spectra (figure 5).

4.2 Results of the EPO preprocessing

The SECVs obtained using S^1 with various g, n_{LV} combinations are plotted in figure 8. A low-SECV basin, with SECV level inferior to 1^o Brix, is described by (g, n_{LV}) couples roughly following equation : $n_{LV}+g \geq 16$. This relationship between (g, n_{LV}) combinations and the SECV level is described more in detail in Table 1. The optimal SECV, close to 0.5^o Brix is seen for $g = 2$ and $n_{LV} = 16$.

When these values are applied, the prediction obtained on S^2 gives a $RMSEP_{corr}$ equal to 0.65^o Brix, with a major reduction of the bias of each temperature data set (figure 9a). Moreover, bias are no longer correlated to the temperature (figure 9b). They oscillate between -0.35 and $+0.39^o$ Brix, which is much better than the bias level obtained without correction. The Root Mean Square of bias is now only 0.25^o Brix instead of 4.63^o Brix.

The other way to find the best (g, n_{LV}) couple is to study the Wilks Λ evolution as a function of g, using the EPO-corrected S^{1*} data set. The higher the Wilks Λ, the tighter the grouping of identical samples measured at different temperatures. The optimum is found at $g = 4$, because Λ reaches a maximal plateau for this number (Table 2). The effect of the EPO correction using $g = 4$ is demonstrated in figure 10, plotting the first factorial map of a PCA calculated on the spectra of S^1. Without any correction (figure 10a), identical samples measured at various temperature are spread along straight lines, whereas after EPO correction (figure 10b), the samples are more tightly grouped together. Applying a cross-validation PLS process on EPO-corrected S^{0*} enabled us to find an optimal $n_{LV} = 12$. Prediction carried out using EPO-corrected S^{2*} (with $g = 4$ and $n_{LV} = 12$) are shown in figure 11. The $RMSEP_{corr}(4, 12)$ is very low (0.52^o Brix) and the bias are less important than with the first method. The Root Mean Square of bias is now only 0.19^o Brix.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

5 Conclusion

This paper is dedicated to presenting a new method, named External Parameter Orthogonalisation (EPO), for defining the useful space in which prediction models must be found, when external parameters alter the spectrum. An orthogonalisation is carried out with regard to the influence of an external parameter onto the spectrum. The power of such a method is proven through an application, i.e. the reduction of the temperature influence on NIR spectra used to predict sugar content in fruits. EPO only requires to measure a small set of appropriate samples measured at different levels of the external parameter. The projection matrix which is generated from EPO can then be applied to any existing calibration basis. Orthogonalising with regard to external parameter is an original alternative to the method developed by Hansen ([26]), which needs from the beginning a very large data set including all the expected external parameter variations. This point is clearly a major advantage of the EPO method.

Next step will be to apply this method to other external factors. EPO could also be applied to the issue of calibration transfer, with the apparatus change taken as the external parameter. The combination of different external parameter effects will also be considered.

References

[1] C. W. Huck, R. Maurer, G. K. Bonn, Quality control of liquid plant extracts in the phytopharmaceutical industry in near infrared spectroscopy, in: A. M. C. Davis, R. Giangiacomo (Eds.), Near Infrared Spectroscopy : Proceedings of the 9th International Conference, NIR Publications, 2000, pp. 487–491.

[2] G. Lachenal, Structural investigations and monitoring of polymerisation by nir spectroscopy, Journal of Near Infrared Spectroscopy 1-4 (1998) 299–306.

[3] H. W. Siesler, Quality control and process monitoring by mid infrared, near infrared and raman spectroscopy, in: A. M. C. Davis, R. Giangiacomo (Eds.), Near Infrared Spectroscopy : Proceedings of the 9th International Conference, NIR Publications, 2000, pp. 331–337.

[4] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[5] P. Geladi, B. R. Kowalski, Partial least squares regression : A tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[6] AFNOR, Nf x 07 - 001 : Fundamental standards - international vocabulary of basic and general terms in metrology, AFNOR (Paris) 07-001.

[7] V. Bellon, C. Vallat, D. Goffinet, Quantitative analysis of individual sugars during starch hydrolysis by ftir/ atr spectrometry. part ii: Influence of external factors and wavelengths parameters, Applied Spectroscopy 49:5 (1995) 563–568.

[8] S. Kawano, H. Abe, M. Iwamoto, Development of a calibration equation with temperature compensation for determining the brix value in intact peaches, J. Near Infrared Spectrosc. 3 (1995) 211–218.

[9] K. DeBraekeleer, F. C. Sánchez, P. A. Hailey, D. C. A. Sharp, A. J. Pettman, D. L. Massart, Influence and correction of temperature perturbations on nir spectra during the monitoring of a polymorph conversion process prior to self-modelling mixture analysis, J. Pharm. Biomed. Anal. 17 (1998) 141–152.

[10] H. Abe, C. Iyo, S. Kawano, A study on the universality of a calibration with sample temperature compensation, J. Near Infrared Spectrosc. 8 (2000) 209–213.

[11] W. G. Hansen, S. C. C. Wiedemann, M. Snieder, V. A. L. Wortel, Tolerance of near infrared calibrations to temperature variations ; a practical evaluation, in: A. M. C. Davis, R. Giangiacomo (Eds.), Near Infrared Spectroscopy : Proceedings of the 9th International Conference, NIR Publications, 2000, pp. 307–311.

[12] F. Wülfert, W. T. Kok, O. E. de Noord, A. K. Smilde, Linear techniques to correct for temperature-induced spectral variation in multivariate calibration, Chemom. Intell. Lab. Syst. 51 (2000) 189–200.

[13] F. Wülfert, W. T. Kok, O. E. de Noord, A. K. Smilde, Correction of temperature-induced spectral variation by continuous piecewise direct standardisation, Anal. Chem. 72 (2000) 1639–1644.

[14] E. V. Thomas, Development of robust multivariate calibration models, Technometrics 42-2 (2000) 168–177.

[15] H. Swierenga, A. P. de Weijer, R. J. van Wijk, L. M. C. Buydens, Strategy for constructing robust multivariate calibration models, Chemom. Intell. Lab. Syst. 49 (1999) 1–17.

[16] P. A. Gorry, General least-squares smoothing and differentiation by the convolution (savitzky-golay) method, Anal. Chem. 62 (1990) 570–573.

[17] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43-5 (1989) 772–777.

[18] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Correction to the description of standard normal variate (snv) and de-trend transformations in practical spectroscopy with applications in food and beverage analysis – 2nd edition, J. Near Infrared Spectrosc. 1 (1993) 185–186.

[19] H. Swierenga, P. J. de Groot, A. P. de Weijer, M. W. J. Derksen, Improvement of pls model transferability by robust wavelength selection, Chemom. Intell. Lab. Syst. 41 (1998) 237–248.

[20] J. M. Roger, V. Bellon-Maurel, Using genetic algorithms to select wavelengths in near-infrared spectra: application to sugar content prediction in cherries, Applied Spectroscopy 54-9 (2000) 1313–1320.

[21] A. Höskuldsson, Variable and subset selection in pls regression, Chemom. Intell. Lab. Syst. 55 (2001) 23–38.

[22] M. B. Seasholtz, B. Kowalski, The parsimony principle applied to multivariate calibration, Analytica Chimica Acta 277-2 (1993) 165–177.

[23] K. Faber, B. R. Kowalski, Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares, J. Chemometrics 11 (1997) 181–238.

[24] S. Wold, H. Antti, F. Lindgren, J. Ohman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.

[25] J. Trygg, S. Wold, Orthogonal projections to latent structures (o-pls), J. Chemometrics 16 (2002) 119–128.

[26] P. W. Hansen, Pre-processing method minimizing the need for reference analysis, J. Chemometrics 15 (2001) 123–131.

[27] B. G. Osborne, T. Fearn, Near infrared spectroscopy in food analysis, John Wiley and Sons, N.Y., 1986.

[28] P. C. Williams, K. Norris, Near infrared technology in the agricultural and food industries, American association of cereal chemistry Inc., St Paul - Minnesota, 1987.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

RMSECV ^oBrix		n_{LV}												
		1	$\cdots$	8	9	10	11	12	13	14	15	16	17	18
g	1	1.43	$\cdots$	1.30	2.22	2.52	2.23	1.73	1.58	1.39	1.21	0.88	0.83	0.97
	2	1.43	$\cdots$	0.97	1.13	0.93	0.99	0.88	0.74	0.68	0.62	0.56	0.57	0.82
	3	1.24	$\cdots$	0.97	1.11	1.28	0.96	0.87	0.81	0.75	0.71	0.69	0.73	0.90
	4	1.31	$\cdots$	0.91	1.19	0.92	0.82	0.78	0.69	0.65	0.66	0.74	0.73	0.85
	5	1.22	$\cdots$	1.01	1.10	0.82	0.80	0.73	0.65	0.64	0.62	0.65	0.70	0.83
	6	1.26	$\cdots$	0.97	1.02	0.79	0.76	0.69	0.65	0.62	0.61	0.62	0.64	0.82

Table 1: Evolution of the Cross-Validation Error (RMSECV) computed using S^1, with regard to the EPO component number, g, and with the number of latent variables of the PLS, n_{LV}.

g	Λ
0	0.625
1	0.935
2	0.958
3	0.969
4	0.981
5	0.980
6	0.982
7	0.983

Table 2: Evolution of the Wilks' *Lambda* as a function of g; $g = 0$ corresponds to no EPO preprocessing.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

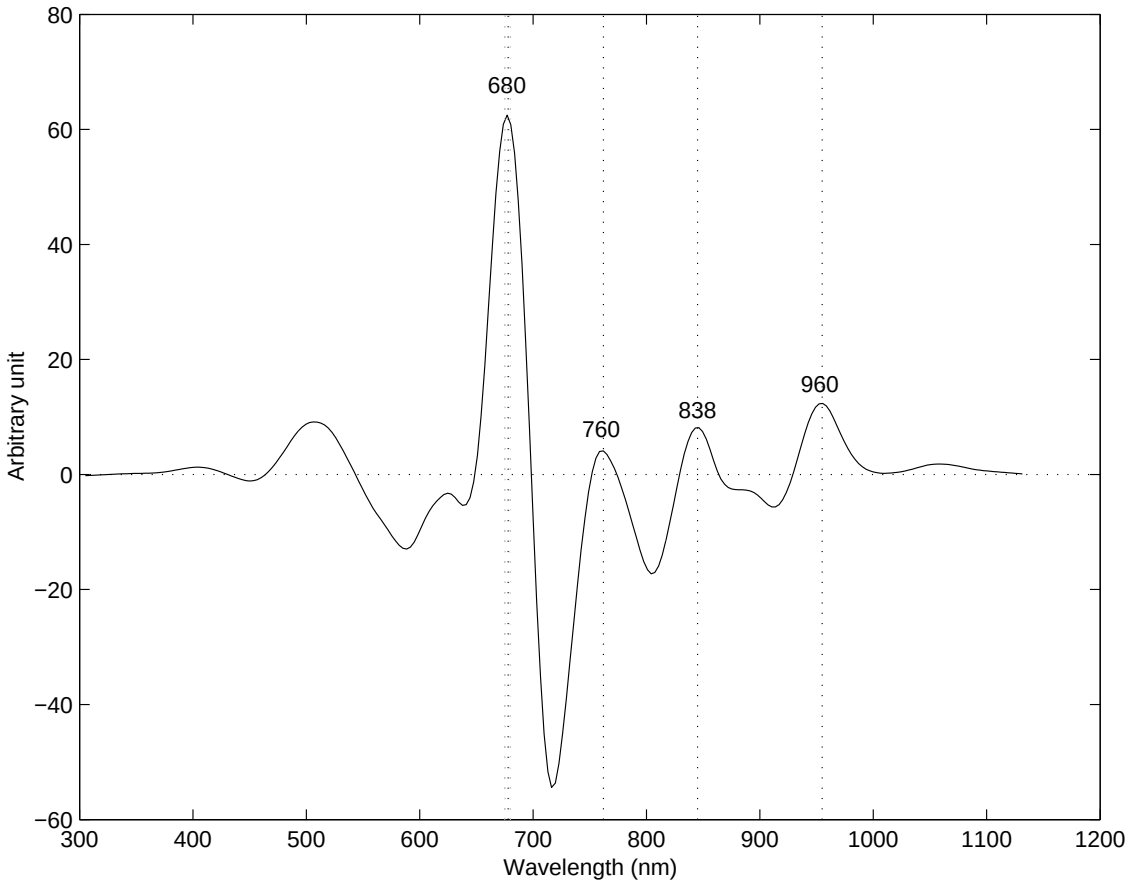

Figure 1: Second derivative of the mean spectrum of S^0 data set.

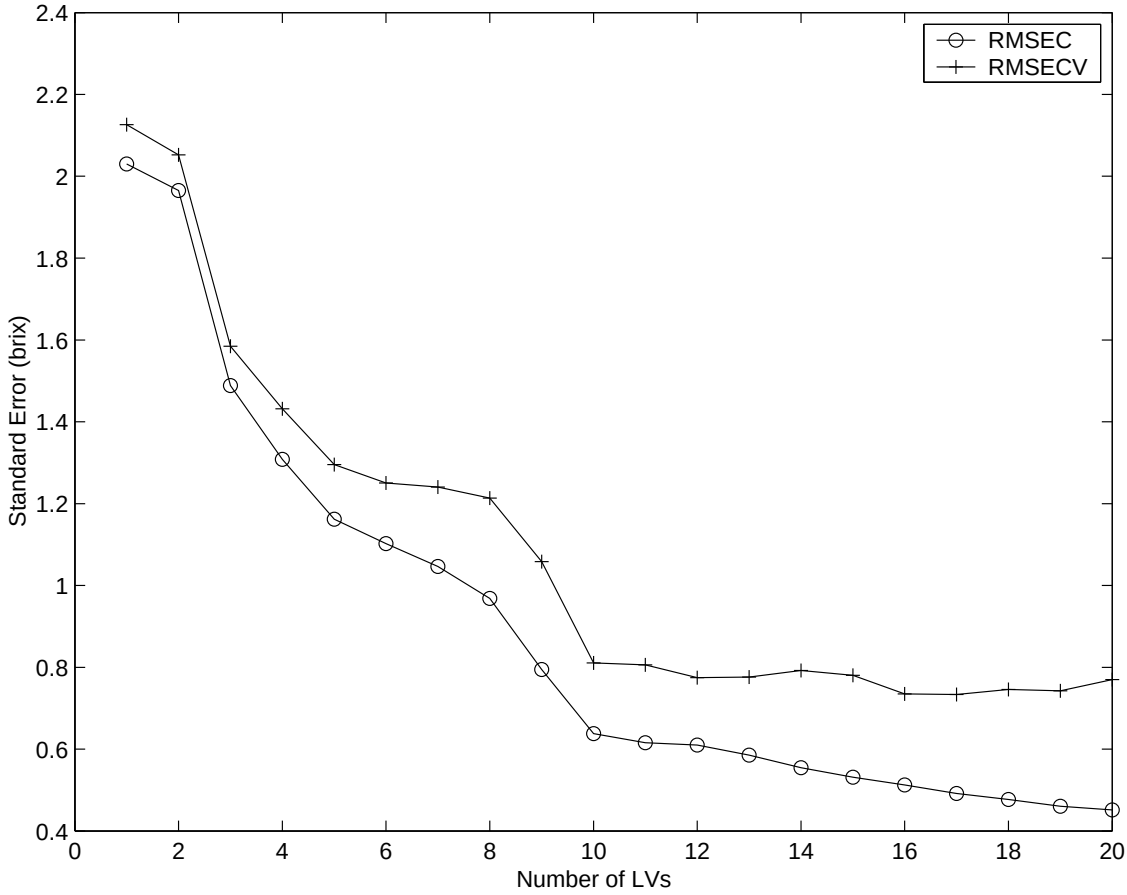

Figure 2: Evolution of the calibration error (RMSEC) and of the cross-validation error (RMSECV) calculated on the data set S^0, as a function of the number of latent variables n_{LV0}.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

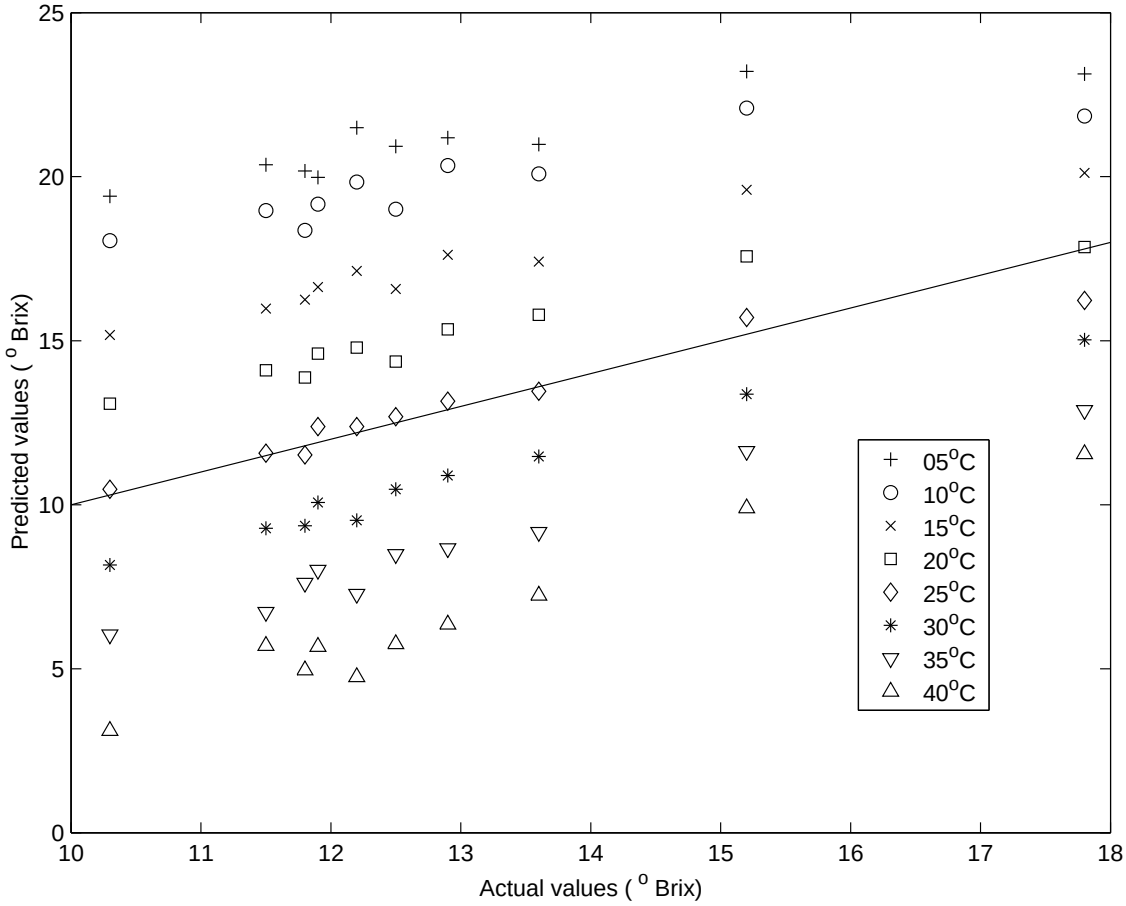

Figure 3: Predicted sugar contents ($\widehat{\mathbf{z}^2}$ as a function of actual sugar contents $\mathbf{z}^2$). The same sample set was measured at different temperatures, and the prediction model has been calibrated at one temperature (25^o C).

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

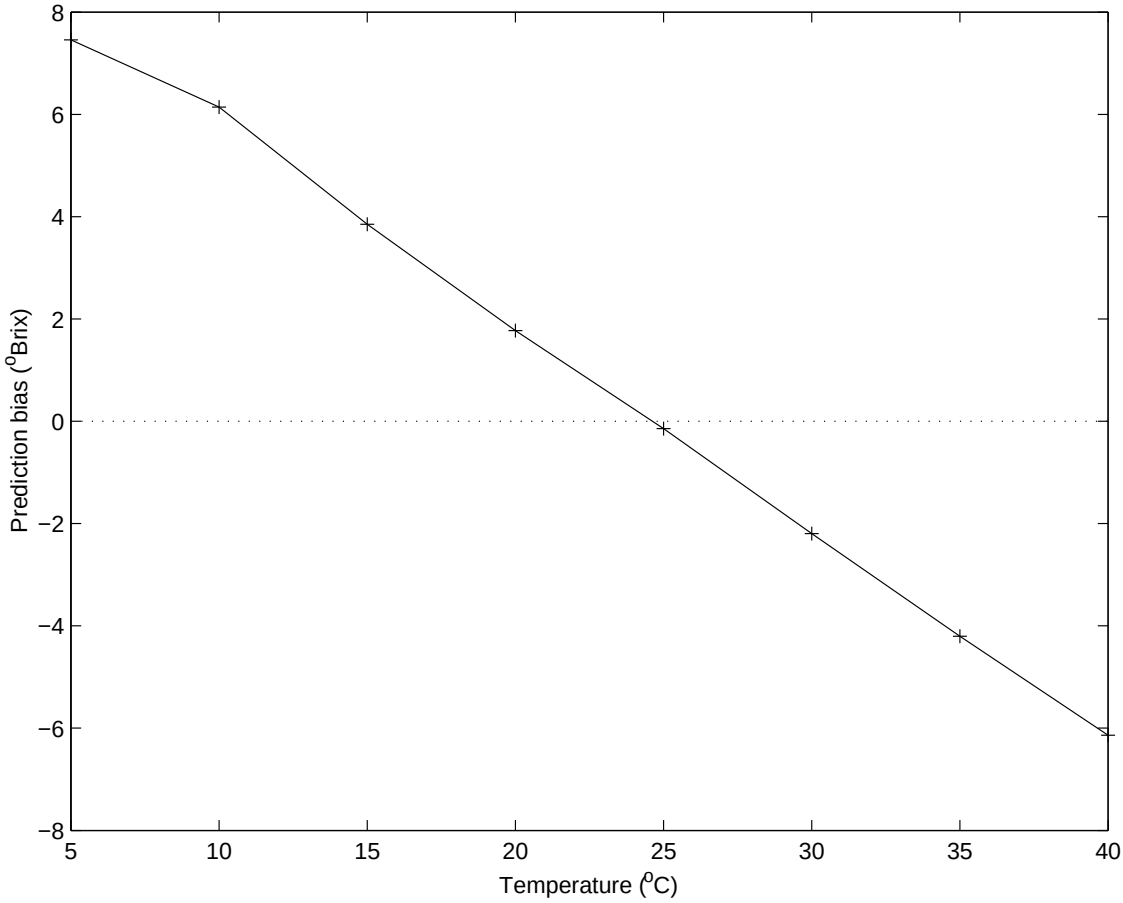

Figure 4: Evolution of the bias as a function of the temperature for the test set S^2, using a prediction model calibrated at 25^o C ($Bias_{raw}(t_i)$).

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

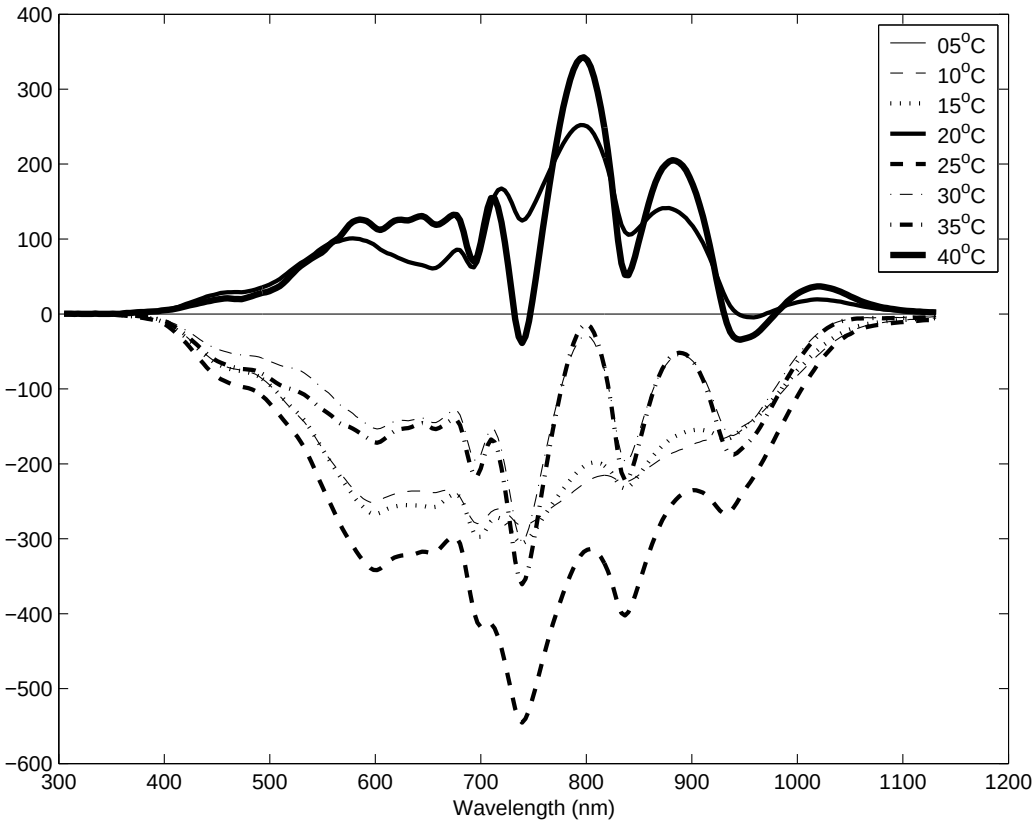

Figure 5: Temperature influence spectra calculated with regard to the reference spectrum measured at 5^oC (**D** matrix).

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

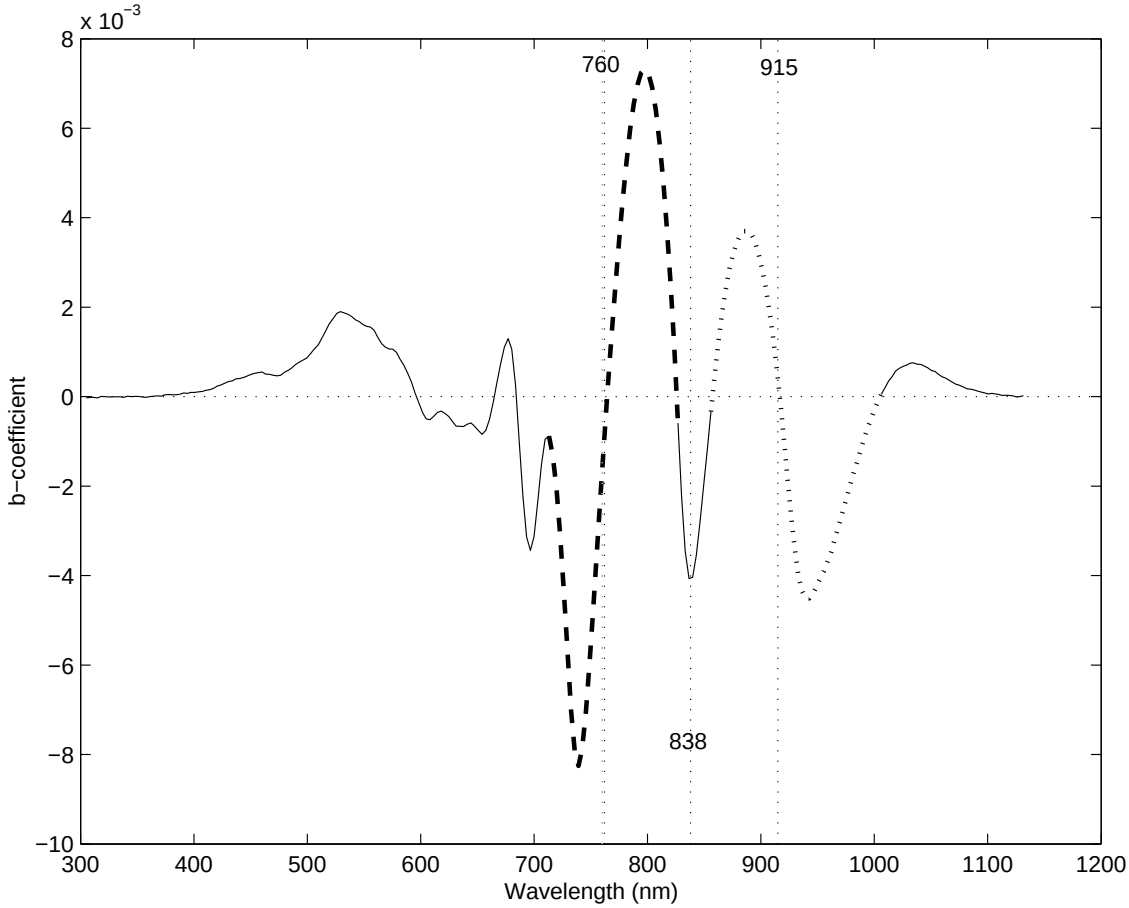

Figure 6: b-coefficient spectra of a PLS regression carried out between the temperature influence spectra and the temperature.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

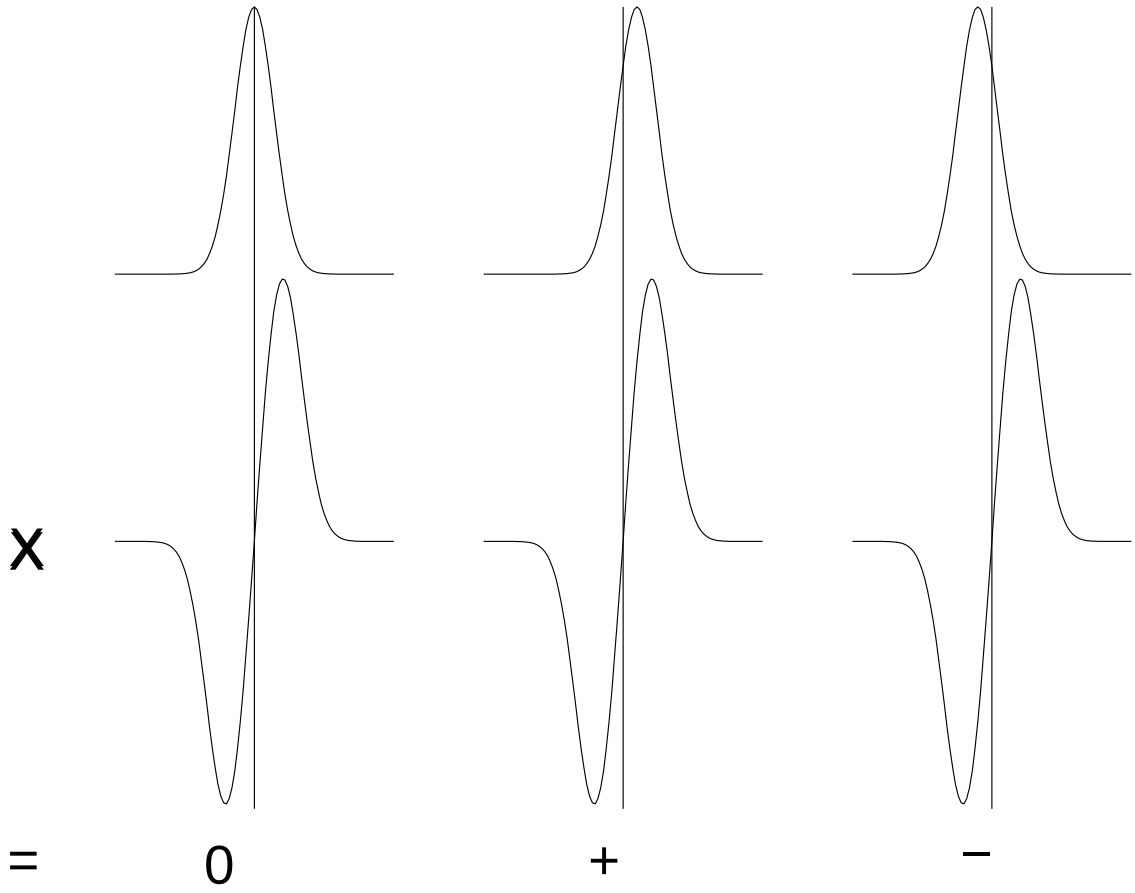

Figure 7: Example of convolutions between a peak and a sinusodal profile.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

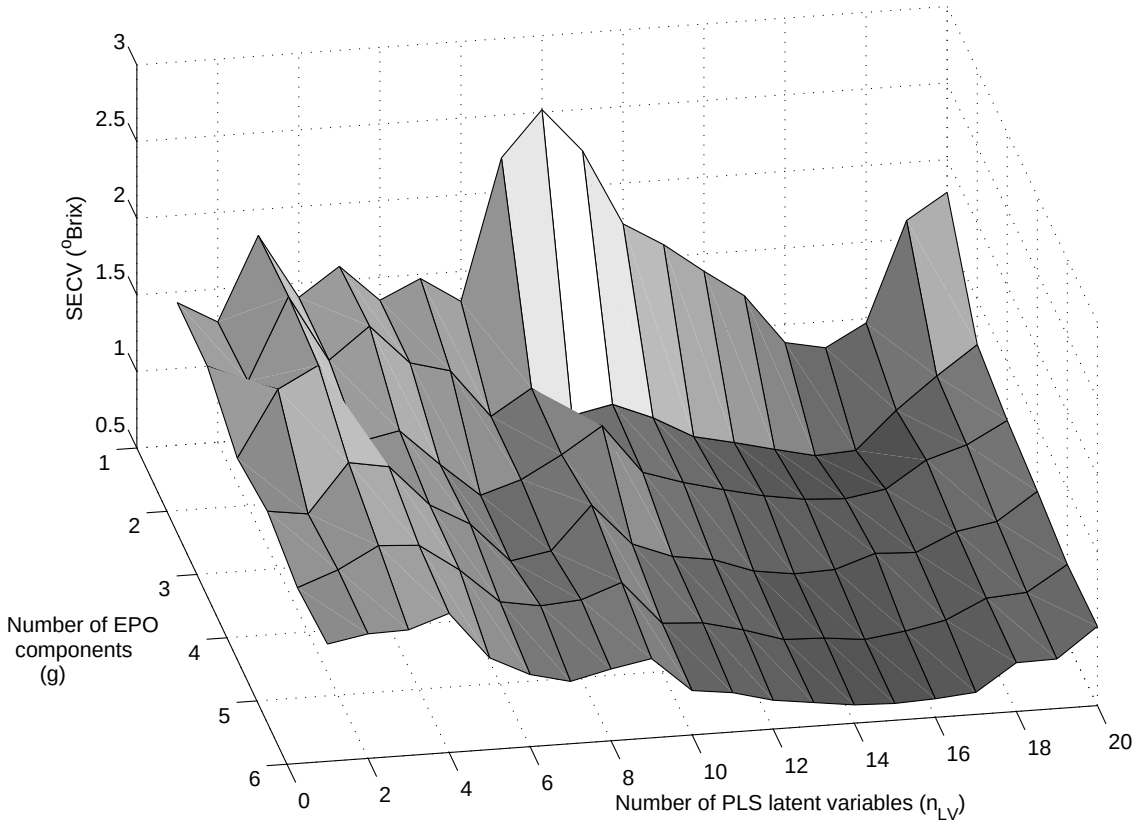

Figure 8: Cross-validation error on S^1, as a function of the number of EPO components (g) and of the number of latent variables used in the PLS regression (n_{LV}).

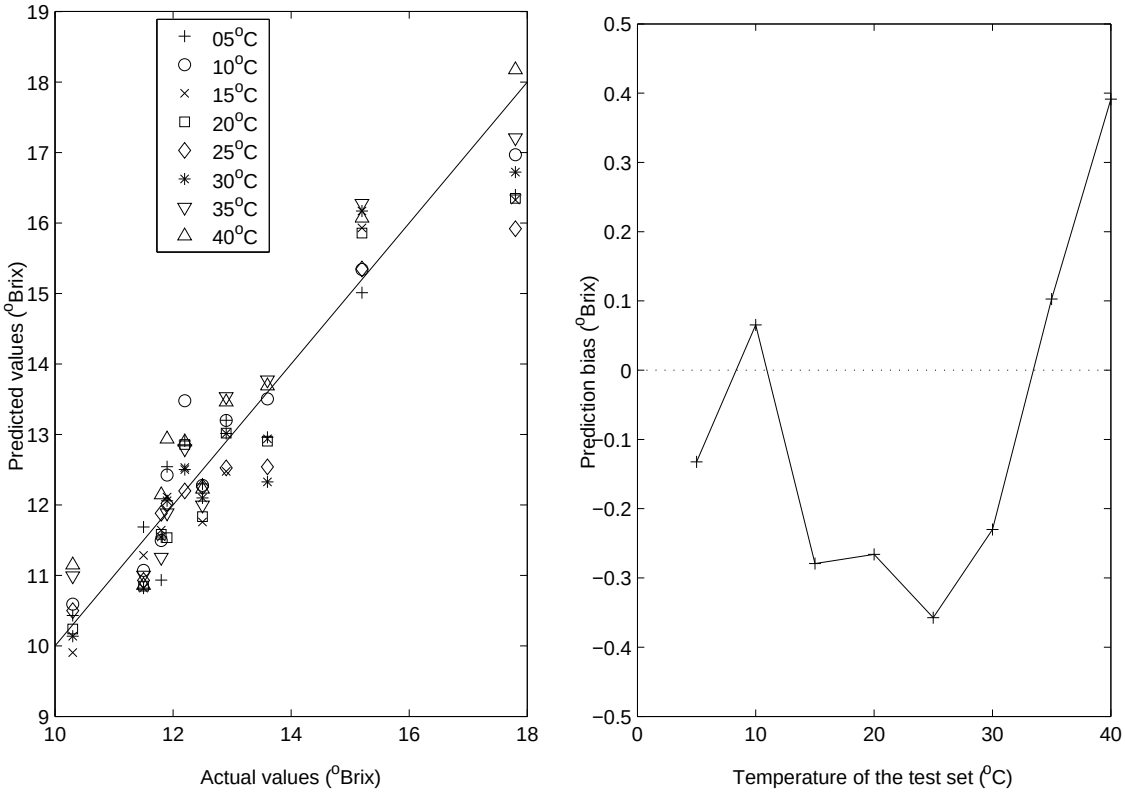

Figure 9: Results of the test on S^2, with EPO preprocessing ($g = 2$ components) and PLS regression ($n_{LV} = 16$ latent variables). Left (a) : $\widehat{\mathbf{z}^{2*}}$ as a fonction of $\mathbf{z}^2$). Right (b) : $Bias_{corr}$ as a function of t_i.

Roger, J.M. ; Chauchard, F. ; Bellon Maurel, V., Chemometrics and intelligent laboratory system, vol. 66, n° 2, p. 191-204

author-produced version of the final draft post-refeering
the original publication is available at www.elsevier.com

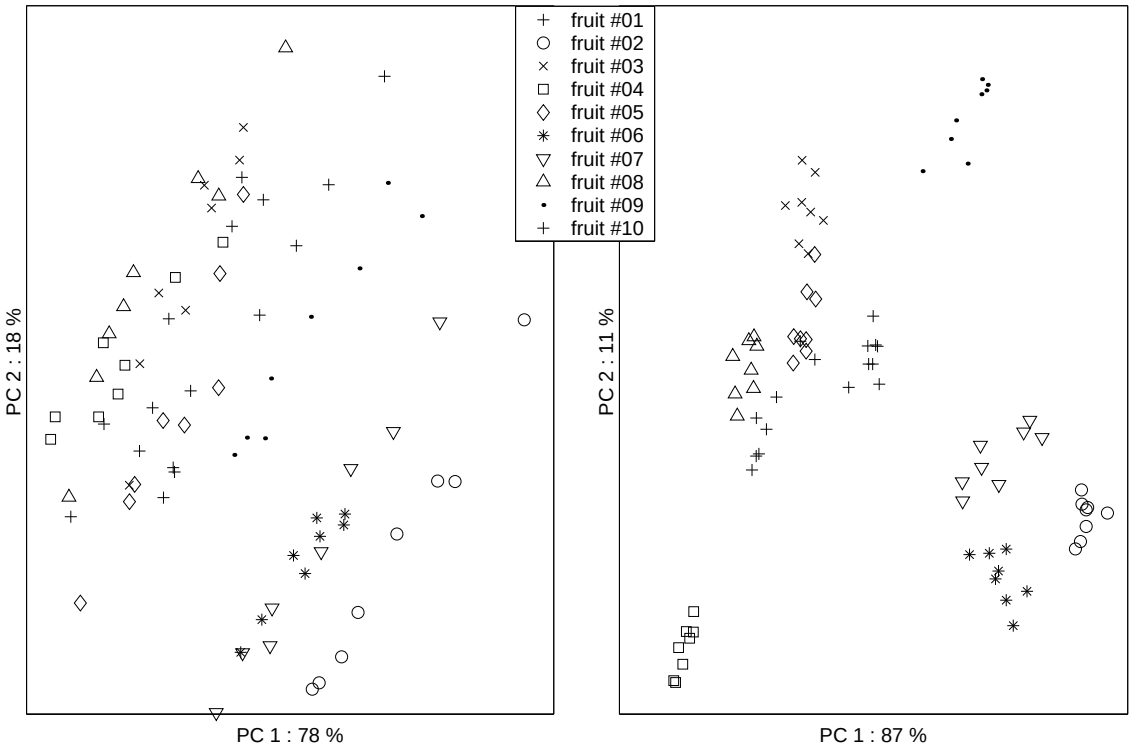

Figure 10: Two first component scores of a PCA calculated on the spectra of S^2 ; Left (a) : without preprocessing. Right (b) : with EPO preprocessing at $g = 4$ components.

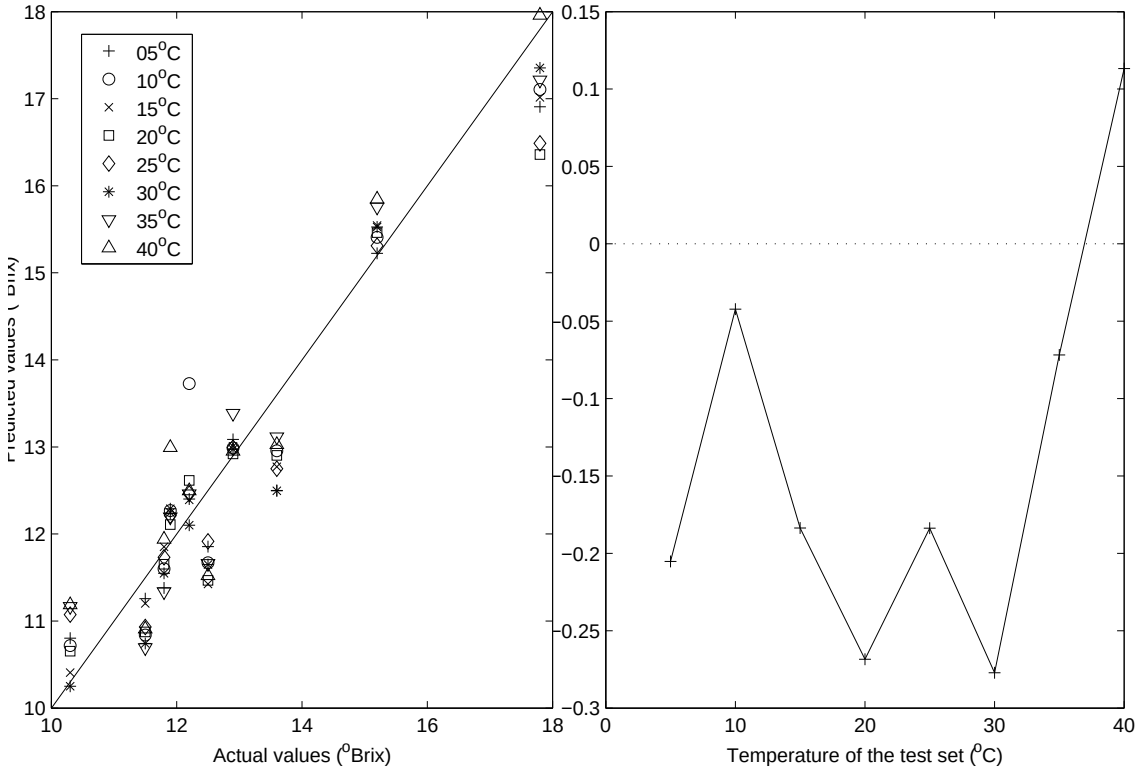

Figure 11: Results of the test on S^2, with EPO preprocessing ($g = 4$ components) and PLS regression ($n_{LV} = 12$ latent variables). Left : $\widehat{\mathbf{z}^{2*}}$ as a function of $\mathbf{z}^2$). Right : $Biass_{corr}$ as a function of t_i.