Improvement of Direct Calibration in spectroscopy

J.C. Boulet, J.M. Roger

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

Improvement of Direct Calibration in spectroscopy

Jean-Claude Boulet, Jean-Michel Roger

Abstract

Several linear calibration methods have been proposed for predicting the concentration of a particular compound from a spectrum. Some methods are based on experimental data, such as Partial Least Square Regression. Other methods are based on expert data, e.g. Direct Calibration. This article proposes a new method, called Improved Direct Calibration, which uses expert and experimental information. It performs a projection onto the pure interest spectrum, after correcting it from influence factors. No calibration dataset is necessary to build this model. This method has been successfully applied to the quantification of ethanol in musts during fermentation, using near infra-red spectrometry.

1 Introduction

Tests applied to many agrifood, pharmaceutical or chemical processes involve quantifying a factor of interest: for example, alcoholic fermentation is tested by measuring the ethanol produced by yeasts. Infrared spectroscopy tools are increasingly used for these applications. However, a spectrum does not only provide information specific to the factor of interest, it also contains contributions from influence factors such as the concentration of other chemical elements, temperature or turbidity. Nevertheless the value of the factor of interest can be extracted using chemometric tools. These build a calibration based on expert and/or experimental knowledge.

Expert knowledge is fundamental data regardless of the experiment. A pure spectrum, a chemical composition or a molecular weight are all examples of expert knowledge linked to a chemical factor. Some calibration methods use pure spectra, for example Direct Calibration (DC) [1], to produce a direct model. Extended Multiplicative Scattering Correction (EMSC) [2] uses baseline models to correct for the effect of a physical influence factor, the scattering.

Experimental knowledge is represented by data collected from samples: spectra and associated quantitative and qualitative values. A calibration database, consisting of a set of spectra associated with the values of the

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

factor of interest is thus an example of experimental knowledge. A matrix of spectra acquired according to an experimental design in which the influence factors vary is another example. The most conventional regression methods such as Principal Component Regression (PCR) [1] and Partial Least Square Regression (PLSR) [3] use the calibration database to produce a forward-inverse model. Other methods use experimental knowledge complementary to the calibration database to correct for the effect of influence factors. For example, Independent Interference Reduction (IIR) [4] uses spectra for which the factor of interest has zero value, to identify and remove useless space. External Parameter Orthogonalisation [5] performs the same correction but with spectra acquired according to an experimental design in which only the influence factors vary.

The most commonly used chemometric calibration methods are based only on experimental knowledge. They are constrained by the calibration database management, which must represent the expected variability in future samples, for both the factor of interest and the influence factors. Thus, a robust model requires an often tedious and costly experimental phase. However, using a mixture of experimental and expert data should reduce this constraint. For instance, Science-Based Calibration (SBC) [6] produces a projection using the pure spectrum of the factor of interest, weighted by the effects of influence factors identified in an experimental database. However, as proposed by Marbach, this method does not specifically take into account the pure spectra of chemical influence factors. The present article proposes a new method, mixing experimental and expert knowledge while taking these two types of informations into account. The first part provides a theoretical description of the method. The second part presents an application of monitoring alcohol fermentation using infrared spectrometry, and finally the third part presents and discusses the results.

2 Theory

Generally speaking, matrices are in bold upper case, vectors in bold lower case and scalars in normal characters. Vectors are listed in columns. The layout used for matrix elements depends on their nature: elements which are the same length as the spectra are in lines, those which are the same length as the number of individuals are in columns.

Let $\mathbf{x}(P,1)$ be a spectrum acquired for a sample and y the value of an interest factor associated with this sample; $\mathbf{k}(P,1)$ the pure spectrum of the factor of interest; $\mathbf{K}$ the pure spectra matrix (Q,P) of the Q constituents other than the factor of interest and $\mathbf{t}_\chi$, of dimensions $(Q,1)$, their concentrations. Let $\mathbf{X}_G$ be a matrix of spectra acquired while the influence factors vary and the interest factor does not. Three approaches have been proposed to characterize $\mathbf{X}_G$. Hansen [4] uses a set of samples in which the factor

of interest is naught. Marbach [6] uses a set of samples measured for the same value of the factor of interest, then centered. Roger [5] uses a set of samples measured then centered at different levels of the factor of interest. Centering in the last two cases removes the spectral influence of the interest factor.

The factor of interest contribution to $\mathbf{x}$ is $y\mathbf{k}$. Chemical influences are the spectral results of concentrations of all the chemical compounds other than the compound of interest in the sample analysed. If all the chemical compounds obey the Beer-Lambert law, each contributes a profile of its pure spectrum, weighted by its concentration, to the final spectrum. Their contribution to the spectrum is therefore given by: $\mathbf{K}'\mathbf{t}_\chi$. The physical influence factors include all the disturbances involved in measuring the spectrum, such as temperature and granulometry for example. Even if they cannot be represented by pure spectra, their influences evolve in a subspace of $\mathbb{R}^P$. Let $\{\mathbf{p}_1 ... \mathbf{p}_A\}$ be an orthogonal basis in this subspace and $\mathbf{P}$ the matrix (A, P) containing them. Let $\mathbf{t}_\phi$ be the score vector representing the physical influences on this basis. The contribution to the spectrum of physical values is therefore given by $\mathbf{P}'\mathbf{t}_\phi$. $\mathbf{x}$ can therefore be written:

$$\mathbf{x} = y\mathbf{k} + \mathbf{K}'\mathbf{t}_\chi + \mathbf{P}'\mathbf{t}_\phi + \varepsilon$$

where ε is a noise vector, each component of which is independent of the others and identically distributed. Assuming that this noise is weak enough to be negligible, the previous expression leads to:

$$\mathbf{x} = y\mathbf{k} + \mathbf{K}'\mathbf{t}_\chi + \mathbf{P}'\mathbf{t}_\phi \tag{1}$$

Calibrating a linear model consists in determining the b-coefficient vector $\mathbf{b}$ and the intercept b_0 such that $\widehat{y}$ defined by:

$$\widehat{y} = \mathbf{x}'\mathbf{b} + b_0 \tag{2}$$

is the best estimate of y that minimises $|y - \widehat{y}|$ under certain constraints. Several direct calibration strategies have been proposed, which give different results in the light of equation (1). They differ in the way in which influence factor information is handled.

DC proposes to project $\mathbf{x}$ onto $\mathbf{k}$ orthogonally to $\mathbf{K}$. Two conditions are assumed to be fulfilled: (c1) the pure spectra of all the chemical values are known and linearly independent so that $(\mathbf{K}\mathbf{K}')$ is invertible; (c2) the effect on the spectra of physical factors is assumed to be negligible. Let $\mathbf{\Sigma}_{DC}$ of dimension (P, P) be the $\mathbf{K}$-orthogonal projector:

$$\mathbf{\Sigma}_{DC} = (\mathbf{I} - \mathbf{K}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{K}) \tag{3}$$

Transposing and right multiplying by $\mathbf{\Sigma}_{DC}$ the equation (1) yields:

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

$$\mathbf{x}'\mathbf{\Sigma}_{DC} = y\mathbf{k}'\mathbf{\Sigma}_{DC} + \mathbf{t}'_{\chi}\mathbf{K}\mathbf{\Sigma}_{DC} + \mathbf{t}'_{\phi}\mathbf{P}\mathbf{\Sigma}_{DC}$$

The two terms on the right of the equation are null, the first by construction, the second in application of the previous hypothesis (c2), and finally:

$$\mathbf{x}'\mathbf{\Sigma}_{DC} = y\mathbf{k}'\mathbf{\Sigma}_{DC}$$

Right multiplying by $\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1}$ produces the DC formula if a single factor of interest is predicted [6]:

$$\widehat{y} = \mathbf{x}'\mathbf{\Sigma}_{DC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1} \tag{4}$$

leading to:

$$\mathbf{b}_{DC} = \mathbf{\Sigma}_{DC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1} \tag{5}$$

Very attractive in principle, this approach is very difficult to apply because hypotheses (c1) and (c2) are rarely fulfilled.

SBC proposes to project $\mathbf{x}$ onto $\mathbf{k}$ using noise-reduction metrics. The SBC b-coefficients are given by:

$$\mathbf{b}_{SBC} = \mathbf{\Sigma}_{SBC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{SBC}\mathbf{k})^{-1} \tag{6}$$

where $\mathbf{\Sigma}_{SBC}$ is given by:

$$\mathbf{\Sigma}_{SBC} = (\mathbf{X}'_G\mathbf{X}_G)^{-1} \tag{7}$$

Each variable of $\mathbf{x}$ is weighted. The weight of each variable depends on its variability on $\mathbf{X}_G$ plus those of other correlated variables. Highest variabilities in $\mathbf{X}_G$ lead to the lowest scores. This approach gives relatively more weight to noise-free variables of $\mathbf{x}$ which contain information about the interest factor. Therefore, SBC increases the signal-to-noise ratio. Problems can occur when $\mathbf{X}_G$ is not invertible. Pseudo-inverse calculation or matrix dimension reduction using PCA are alternative solutions.

The new method presented in this paper (IDC) is performed by completing the expert information in $\mathbf{K}$ with the experimental information in $\mathbf{X}_G$. For this, a vector basis $\{\mathbf{p}_1 ... \mathbf{p}_A\}$, representing the space spanned by $\mathbf{X}_G$, forming matrix $\mathbf{P}$, is identified and added to $\mathbf{K}$, giving a matrix $\mathbf{R}$. Let $\mathbf{\Sigma}_{IDC}$ be the $\mathbf{R}$ orthogonal projector:

$$\mathbf{\Sigma}_{IDC} = (\mathbf{I} - \mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}\mathbf{R}) \tag{8}$$

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

By transposing, then right multiplying by $\mathbf{\Sigma}_{IDC}$ the equation 1, the effect of physical and chemical influence factors becomes mathematically null, which gives:

$$\mathbf{x}'\mathbf{\Sigma}_{IDC} = \widehat{y}\mathbf{k}'\mathbf{\Sigma}_{IDC}$$

Right multiplying by $\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$ provides:

$$\widehat{y} = \mathbf{x}'\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1} \tag{9}$$

which gives:

$$\mathbf{b}_{IDC} = \mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k}')^{-1}$$

The easiest way of identifying $\mathbf{P}$ is to perform an SVD on $\mathbf{X}_G$, and to retain the first A eigenvectors. The choice of A is an important stage in this method. On the one hand, if A is too high, residual information on the interest factor that could be present in $\mathbf{X}_G$ can be captured by $\mathbf{P}$, leading to a poor prediction. On the other hand, if A is too small, not all the influence factors are corrected, leading also to a poor prediction. To choose A, calculating the prediction error as a function of A onto a prediction set is possible. However, this approach loses the major advantage of direct calibration, which is precisely not to need a calibration database. Another approach consists of applying the IDC model to $\mathbf{X}_G$ (for which the interest factor is constant and usually known) for different values of A and examining the evolution of the prediction error.

Although matrix $\mathbf{K}$ must theoretically contain the pure spectra of all the compounds in the sample, in practice this approach comes up against a certain number of difficulties. For instance, in very complex samples, some pure products cannot be extracted and stabilised in large enough quantities for spectra measurement to be possible. In order to capture all the information concerning influence factors in either $\mathbf{X}_G$ or $\mathbf{K}$, a simple rule should consist in considering that $\mathbf{K}$ must contain the pure spectra of products whose concentration does not vary in $\mathbf{X}_G$, or that the influence factors not found in $\mathbf{K}$ must be varied in $\mathbf{X}_G$.

3 Materials and methods

The application concerned ethanol quantification in musts and wines using near infrared spectroscopy. Spectra were acquired for all samples analyzed in the Skalli laboratory (Sète, France) during Septembre 2005, i.e. 80p.cent were fermenting musts, and 20p.cent were wines from previous vintages and sometimes other wineries, controled during the buying, ageing or bottling

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

processes. Due to changes in the chemical composition of musts during vinification, all samples will be considered as different. The experimental database consisted of 1480 spectra, acquired using a Jasco spectrophotometer (optical path 1 mm, range $500 - 2500$ nm, step 2nm, with water as optical reference). The ethanol reference values of these same samples were measured by mid-infrared spectrophotometry (Foss). These data yielded an $\mathbf{X}$ matrix $(1480, 1001)$ and a $\mathbf{y}$ vector $(1480, 1)$. The pure spectra were acquired using the same spectrophotometer onto pure samples of ethanol, glycerol, water and lactic acid, with air as optical reference.

3.1 Data processing

The data were processed using Scilab software *(www.scilab.org)* completed with PCA and PLSR functions using the Saisir toolbox *(www.chimiometrie.fr)*. All spectra baselines were previously adjusted to value 0 for wavelength 1100nm, chosen in a region where experimental and expert spectra both had the lowest variabilities. The interest factor was ethanol, its pure spectrum $\mathbf{k}$ was divided by 100 for the predicted values to be directly expressed in percent volume. Experimental information was represented by $(\mathbf{X}, \mathbf{y})$ and split into three sets:

- $\mathbf{X}_1$, containing non fermented musts, i.e. 165 sample spectra with zero ethanol concentration;
- $(\mathbf{X}_2, \mathbf{y}_2)$, containing the first 315 samples with non-zero ethanol content, in their chronological order of acquisition;
- $(\mathbf{X}_3, \mathbf{y}_3)$ containing the last 1000 samples with non-zero ethanol content, in their chronological order of acquisition.

This chronological splitting was chosen to ensure maximum independence between the sets of $(\mathbf{X}_2, \mathbf{y}_2)$ and $(\mathbf{X}_3, \mathbf{y}_3)$. It was verified that the histograms of $\mathbf{y}_2$ and $\mathbf{y}_3$ were comparable. The $\mathbf{P}$ matrix was obtained from the A first eigenvectors of an SVD of $\mathbf{X}_1$ taken here as the $\mathbf{X}_G$ matrix. Expert information about influence factors was represented by glycerol, lactic acid and water spectra, yielding to the $\mathbf{K}$ matrix $(1001, 3)$.

Seven calibration models were calculated then tested on $(\mathbf{X}_3, \mathbf{y}_3)$. The first three models were designed to explain how IDC works. The first model (m1) was a simple projection onto $\mathbf{k}$. The second model (m2) used IDC with only $\mathbf{k}$ and $\mathbf{K}$, corresponding to DC with very few pure spectra. The third model (m3) used IDC with only $\mathbf{k}$ and $\mathbf{P}$. The fourth model (m4) used complete IDC with $\mathbf{k}$, $\mathbf{K}$ and $\mathbf{P}$. The fifth model (m5) used PLSR, calibrated on $(\mathbf{X}_2, \mathbf{y}_2)$ by cross-validation of the NIPALS algorithm. The number of latent variables was chosen to minimise the RMSECV. The sixth model (m6) was an IDC with $\mathbf{k}$, $\mathbf{K}$ and $\mathbf{P}$ with A distinctly higher than the optimal value chosen in models (m3) and (m4). Finally, the seventh model

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

(m7) repeated model (m4) after the water spectrum had been removed from $\mathbf{K}$.

3.2 Model comparison

The models were first evaluated visually according to their general predictive aptitude, i.e. by aligning the predicted values relative to the reference values along the line ($\widehat{y} = y$). RMSEP, bias and RMSEPc corrected for bias were calculated for each model. The b-coefficient peaks were interpreted by comparison with the pure spectra of ethanol and water.

4 Results

4.1 Choice of y unit

The Beer-Lambert law is limited to low concentrations, which is not the case here. However, absorbance is due to electron excitement by photons, so y should reflect electronic properties of the molecules. Mark [7] recently raised the question of the unit of predicted values using DC. He suggested that y represent the percentage of hydrogen atoms contributed by the compound of interest. The number of H-moles contributed by 1mL of ethanol and water is almost the same, respectively 0.104 and 0.111, so in first approximation y would represent the percent volume of ethanol, that is the common unit used in enology.

The unit of y is given by the unit chosen for $\mathbf{k}$. Here, $\mathbf{k}$ was the pure spectrum of ethanol, so $\mathbf{k}/100$ represents the spectrum of 1 p.cent vol. of ethanol, and the corresponding y values obtained using the IDC model are also expressed as percent volume of ethanol.

4.2 Construction of the K matrix

It is unusual, in spectroscopy, to use spectra not acquired under the same conditions. However, this is the case here, because the experimental data represented by $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ and the expert data represented by $\mathbf{k}$ and $\mathbf{K}$ used water and air respectively as optical references. The explanation is that experimental spectra are usually acquired with a water reference for practical reasons, whereas expert data are acquired with an air reference to meet the requirements of Beer-Lambert's law relative to mixtures. For any spectrum, let note $\mathbf{x}^w$ its value with a water reference and $\mathbf{x}^a$ its value with an air reference. If $\mathbf{k}^a_{water}$ is the water spectrum with an air reference, then:

$$\mathbf{x}^w = \mathbf{x}^a - \mathbf{k}^a_{water}$$

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

Therefore, the difference between the different $\mathbf{x}$ spectra measured with air and water references is simply the water spectra. Once this same spectrum is inserted in $\mathbf{K}$, because $\mathbf{\Sigma}_{IDC}$ is an orthogonal projection onto $\mathbf{K}$, the product $\mathbf{\Sigma}_{IDC}\mathbf{k}^a_{water}$ is null. In conclusion, incorporating the water spectrum $\mathbf{k}^a_{water}$ in $\mathbf{K}$ means that the experimental data acquired with water reference can be used directly instead of the experimental data with air reference.

The main natural constituents of musts and wines, excluding ethanol, are: water, glucose, fructose, glycerol, tartaric, malic and lactic acids. However, the musts used for $\mathbf{X}_G$, named $\mathbf{X}_1$ in our application, contain variable quantities of glucose and fructose as well as tartaric and malic acids. The pure spectra of these compounds were therefore not included in $\mathbf{K}$. On the other hand, glycerol and lactic acid are not found in musts, so they were not present in $\mathbf{X}_G$. This is why their pure spectra were included in $\mathbf{K}$. Finally, $\mathbf{K}$ contains spectra for water, glycerol and lactic acid.

4.3 Parameter determination

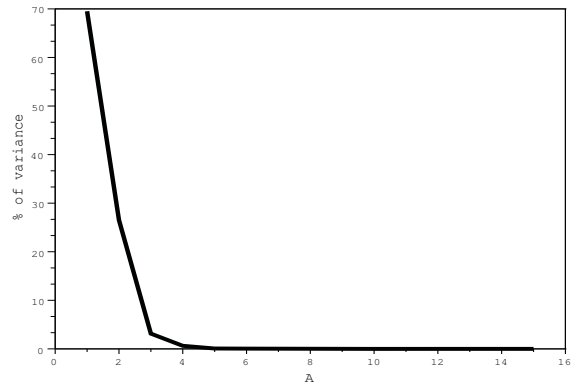
Fig.1 is used to choose the dimensions of models (m3), (m4), (m6) and (m7). Fig.1a shows the evolution of the percentage of inertia of $\mathbf{X}_G$ captured by $\mathbf{P}$ vectors. Fig.1b shows the evolution of the standard error of prediction for model (m4) applied to $\mathbf{X}_G$ as a function of A. The value $A = 4$ was used because it allowed almost all $\mathbf{X}_G$ information to be captured while giving a minimal prediction error. This optimal value of $A = 4$ was also applied to (m2) and (m7). The prediction error increases for $A > 10$. That confirms the risk that excessively high values of A can cause $\mathbf{P}$ to capture residual information about the factor of interest, as mentioned in the theory part. To verify this, model (m6) with $A = 12$ was also constructed. Fig.2 was used to tune the dimension of the PLS model. The standard cross-validation error of the PLSR (RMSECV) was stabilised for 5 and minimum for 8 latent variables, so the PLSR model was built using 8 latent variables.

4.4 Analysis of models m1 to m7

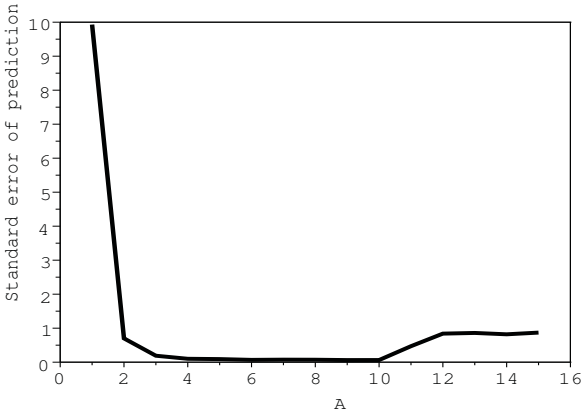
Fig.3-m2 shows the prediction obtained by model (m2). The correction induced by the pure spectra led to a prediction which certainly includes too much noise to be used as a prediction model, but which is sensitive to $\mathbf{y}$: the square correlation coefficient between predicted values and reference values is 0.87.

Table 1 shows the correlations obtained by different models, derived from (m2) by eliminating 0 to 2 spectra from the $\mathbf{K}$ matrix. Highest correlations are clearly obtained by the presence of the pure spectrum of water and at least one of the other two pure spectra: glycerol or lactic acid. In all other cases, correlation does not exceed 0.20 and can be considered to be null. The need of the water spectrum is explained by the different references

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010



(a)



(b)

Figure 1: (a) Evolution of the percentage of inertia of $\mathbf{X_G}$ captured by A for the first 15 vectors of $\mathbf{P}$; (b) Standard error of prediction for the IDC model applied to $\mathbf{X_G}$ for $A = 1$ to 15

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

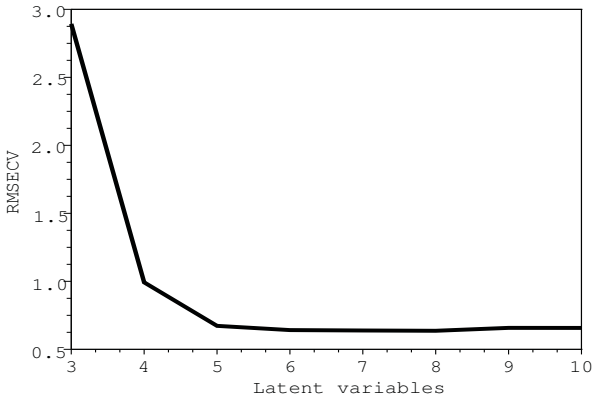


Figure 2: PLSR RMSECV for latent variables 3 to 10.

used for data acquitition, air or water. The cosines of glycerol and lactic acid spectra with the ethanol spectrum are 0.93 and 0.92 respectively. This strong colinearity removes from the spectra a huge amount of useless information close to the ethanol signal. The remaining information is more specific to ethanol, signal to noise ratio is increased, leading to more accurate models.

Fig.3-m3 shows the prediction obtained by the model (m3). This model does not use any of the three spectra mentioned previously: water, lactic acid, glycerol. It is therefore normal for the correlation between predicted and reference values to be close to 0. However, the introduction of $\mathbf{X}_G$ eliminates all prediction variability and thus noise. Finally, this model predicts a value close to 0 for all the samples.

Fig.3-m1 shows the prediction obtained with no correction. The absence of $\mathbf{K}$ leads to a prediction which is insensitive to ethanol, and the absence of $\mathbf{P}$ gives a very noisy prediction. Logically, the prediction for (m1) is very noisy and centered on 0.

The complete IDC model (m4) gives very satisfactory predictions, comparable with those of the PLSR (m5), see Fig.3-m4 and m5, Tables 2 and 3. The IDC RMSEP is very close to that of the PLSR for ethanol concentrations of less than 10p.cent vol., but higher for ethanol concentrations over 10p.cent vol. In general, models (m2) to (m4) display greater error in the high ethanol concentration zone, corresponding to wines at the end of fermentation or finished wines. The problem is certainly due to the appearance of new influence factors, such as an effect of physical and chemical stabilisation of finished wines. The PLSR was able to handle these new influence factors because the calibration set contained finished wines, thus improving robustness for PLSR in this situation. These new influence factors haven't been taken into account by IDC calibration because they were represented neither in $\mathbf{K}$ nor in $\mathbf{X}_G$. Another set of data from finished wines could have

been be taken to improve IDC calibration. The problem is then to obtain samples which all have exactly the same ethanol content so that simply centering eliminates the effect of ethanol on these spectra. A solution can be an experimental design based on samples before and after stabilisation, under such conditions that no loses of ethanol are possible.

Model (m6), Fig.3-m6 had poor prediction, much worse than that of model (m4). The difference between these models is that the optimal value $A = 4$ was chosen for (m4), whereas a value deliberately chosen to be much higher, $A = 12$, was chosen for (m6). Thus, the experimental design leading to $\mathbf{X}_G$ was intended to maximize the expression of the influence factors, which are modelised by the first loadings of $\mathbf{P}$. But native grapes can also contain low amounts of ethanol, because of anaerobic metabolism known as carbonic maceration, or yeast activity on wounded berries. This ethanol can lead to a weak spectral signal. If the dimension of $\mathbf{P}$ is too large, more or less information about the interest factor is captured by $\mathbf{X}_G$. Thus, $\mathbf{\Sigma k}$ tends to the null vector. This was verified in our example, an IDC model with $A = 40$ led to predictions centered around 0, results not shown.

The difference between models (m4) and (m7) lies only in the presence of the water spectrum for the first one, and its absence for the second one. Their differences of accuracy demonstrates the importance of the water spectrum into $\mathbf{K}$.

Fig. 4 represents the IDC b-coefficients compared to pure spectra of ethanol and water, completed by a wine spectrum. The 4 main peaks of the ethanol spectrum ($1580, 1710, 2085$ and 2295 nm) are found in the IDC b-coefficients (Fig.4). The ethanol peak at 2085 nm is attenuated in the b-coefficients. One explanation is that sugars also have a strong absorbance peak at this wavelength [8]. In addition to the ethanol peaks, two other peaks are visible in the b-coefficients. The negative peak at 1450 nm as well as the positive peak at 1940 nm may be linked to the water absorbance in this zone. For ethanol prediction using IDC, the opposite contributions of these two peaks cancelled out the influence of water, that is precisely the interest of the orthogonal projection.

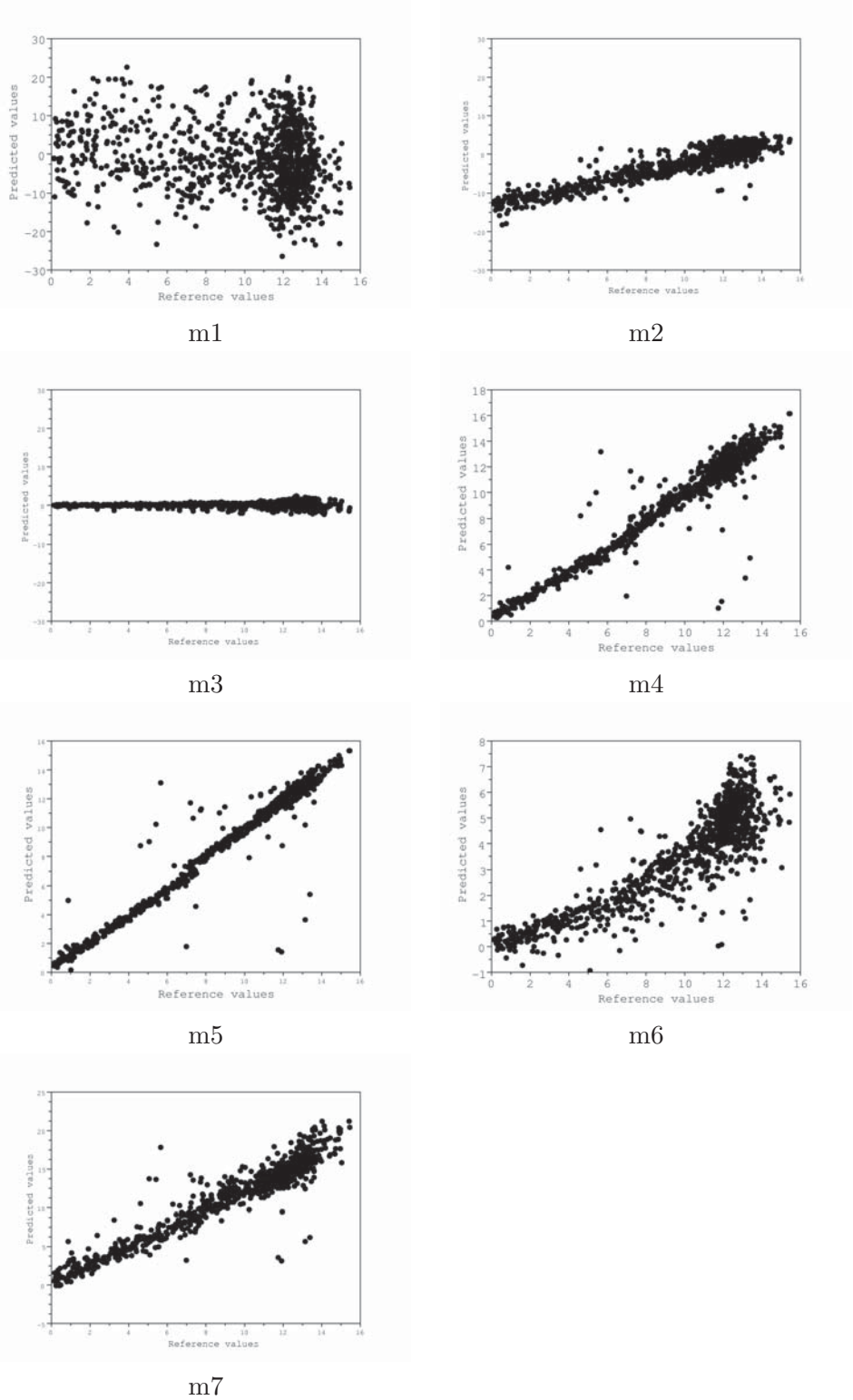
The comparison between PLSR b-coefficients and IDC shows that they are different: their cosine is 0.47, so the angle between these two vectors is nearly 60 degrees. This example illustrates the non-uniqueness of the models: equivalent predictions are obtained from very different models.

4.5 Bias and slope management

From a theoretical point of view, a well-built IDC model do not have bias and slope different from 0 and 1 respectively. Otherwise that means that some influence factors haven't been taken into account.

An IDC model can be built in certain conditions then applied under conditions where an unexpected influence factor hasn't been taken into account

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010



m1 m2 m3 m4 m5 m6 m7

Figure 3: Model tests: (m1) simple projection without correction; (m2) IDC using only $\mathbf{K}$; (m3) IDC using only $\mathbf{X_G}$; (m4) complete IDC with $A = 4$; (m5) PLSR using 8 latent variables; (m6) complete IDC with $A = 12$; (m7) model (m4) after withdraw of water spectrum

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

Spectra in $\mathbf{K}$	R^2
Water (W)	0.20
Lactate (L)	0.00
Glycerol (G)	0.06
L + G	0.03
W + L	0.74
W + G	0.85
W + L + G	0.87

Table 1: Coefficients of correlation R^2 between predicted and reference values, for models obtained from (m2) after removing 0, 1 or 2 spectra

Model	Slope	Bias	$RMSEP_c$	$RMSEP$	R^2
m1 ($\mathbf{\Sigma} = \mathbf{I}$)	-0.34	11.3	9.98	15.1	0.02
m2 ($\mathbf{\Sigma} = \mathbf{K}$)	1.17	11.9	1.87	12.1	0.87
m3 ($\mathbf{\Sigma} = \mathbf{P}, A = 4$)	-0.16	11.6	4.55	12.4	0.29
m4-IDC ($\mathbf{\Sigma} = [\mathbf{KP}], A = 4$)	0.99	0.03	0.96	0.96	0.94
m5-PLSR (8LV)	0.97	0.09	0.85	0.85	0.95
m6 ($\mathbf{\Sigma} = [\mathbf{KP}], A = 12$)	0.43	6.32	2.41	6.76	0.74
m7 (m4 without water spectrum)	1.10	-1.51	1.51	2.13	0.89

Table 2: Figures of merit of the models

Model	Ethanol < 10 %	Ethanol ≥ 10 %
IDC (m4)	0.87	1.01
PLS (m5)	0.85	0.86

Table 3: RMSEP of IDC and PLSR according to the concentration of ethanol in the sample

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

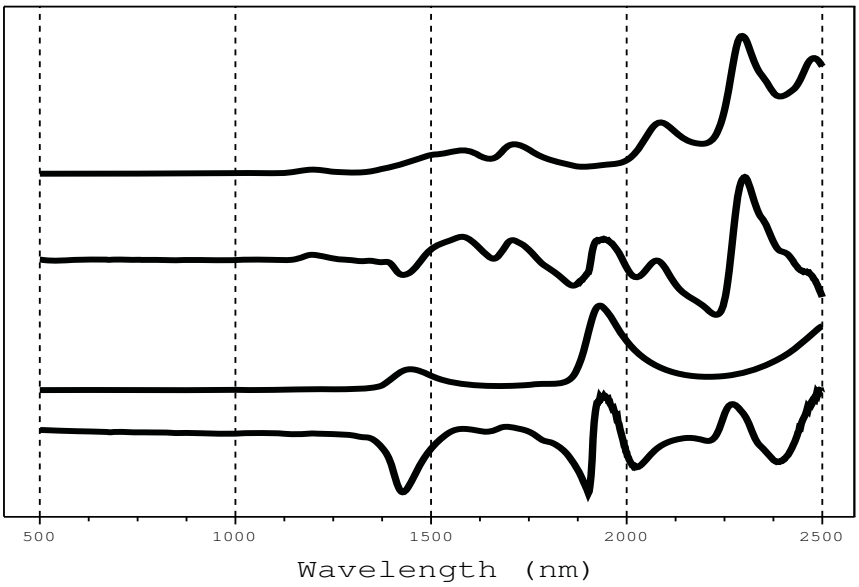


Figure 4: From top to bottom, shapes of pure ethanol spectrum, ICD b-coefficients, pure water spectrum and example of wine spectrum

by the model. Let us assume that this new influence factor is described by only one vector $\mathbf{e}$. For a sample i, let $\mathbf{x}_i$ and $\widehat{y}_i$ be respectively the spectrum and the IDC estimation that would be obtained without the new influence factor, $\mathbf{x}_i^*$ and $\widehat{y}_i^*$ the measured spectrum and the IDC estimation obtained with the new influence factor. Then there exists a scalar λ_i that verifies:

$$\mathbf{x}_i^* = \mathbf{x}_i + \lambda_i \mathbf{e} \tag{10}$$

and after multiplying each side by $\mathbf{b}'_{IDC}$:

$$\widehat{y}_i^* = \widehat{y}_i + \lambda_i \mathbf{b}'_{IDC}\mathbf{e} \tag{11}$$

This explains why some IDC models present a non-null biais, a slope different from 1, or both. When the $\{\lambda_i\}$ coefficients take a constant value λ for all i, predicted values have a constant biais equal to $\lambda \mathbf{b}'_{IDC}\mathbf{e}$. When the $\{\lambda_i\}$ coefficients are correlated to the y_i, the resulting predictions have a slope different from 1. In the other cases, the prediction variance is increased.

The bias problem can be due to the way $\mathbf{X}_G$ is built. A first approach consists in acquiring spectra with a constant value of y and variation of influence factors as proposed by Roger [5] and Marbach [6]. The resulting matrix $\mathbf{X}_G$ must be centered to withdraw all information about the interest factor. This operation also eliminates all constant information, so that the

correction of any constant influence factor $\mathbf{e}$ is not performed. A second approach proposed by Hansen [4] uses a set of samples where the interest factor is null whereas influence factors vary, as used in this paper. No centering is necessary, so the effect of $\mathbf{e}$ is taken into account.

However if $\mathbf{e}$ is known, it is always better to put it in $\mathbf{K}$ in order to take it into account explicitely. This can be illustrated by a simple example. Let $\mathbf{x}_G$ be a vector of $\mathbf{X}_G$. The orthogonal projection to the vector $(\mathbf{x}_G+\mathbf{e})$ removes a space of dimension 1, whereas the orthogonal projection to the matrix containing $\mathbf{x}_G$ and $\mathbf{e}$ removes a space of dimension 2. All the information about $\mathbf{x}_G$ and $\mathbf{e}$ is withdrawn in the second case, but not in the first. In our application, $\mathbf{e}$ is the water spectrum used to get spectra with water reference. Despite its presence in all spectra of $\mathbf{X}_G$, it must be put into $\mathbf{K}$ to obtain an accurate IDC model.

4.6 Links with the Net Analyte Signal

The term $(\mathbf{\Sigma}_{DC}\mathbf{k})$ is written $(\mathbf{I}-\mathbf{K}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{K})\mathbf{k}$; it represents the projection of the pure spectrum of the factor of interest orthogonally to the matrix of pure spectra of chemical influence factors. Similarly, $(\mathbf{\Sigma}_{IDC}\mathbf{k})$ is written $(\mathbf{I}-\mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}\mathbf{R})\mathbf{k}$. Therefore the $\mathbf{k}$ spectrum of the factor of interest is projected orthogonally to the space defining the chemical and physical influence factors. These two cases are in accordance with the definition of the Net Analyte Signal (NAS) given by Lorber [9]:"*the net analyte signal may be computed as the part of its spectrum orthogonal to the contribution of other coexisting constituents*", with the difference that with IDC this definition is extended to physical influence factors: IDC improves the definition of NAS. Let the scalar $\alpha = (\mathbf{k}\mathbf{\Sigma}_{IDC}\mathbf{k}')^{-1}$. The prediction of the interest factor is also written:

$$\widehat{\mathbf{y}}_{IDC} = \alpha\mathbf{X}\widehat{\mathbf{NAS}}_{IDC}$$

Therefore, prediction by IDC is proportional to the inner product of $\mathbf{X}$ spectra and the estimate of NAS calculated by IDC using Euclidean metrics. Coefficient α adjusts the notation scale of the interest factor which is arbitrary: for example, mg/L or g/L. IDC b-coefficients tend towards the NAS, which is not the case for PLSR b-coefficients.

5 Conclusion

This study confirms the fact that DC cannot be applied in cases where certain chemical and physical influence values are not taken into account. However, these expert data can be judiciously completed with experimental information acquired in the form of a spectral matrix which is used to

Author-produced version of the article published in Analytica Chimica Acta, 2010, 668, 2, 130-136.
The original publication is available at http://www.sciencedirect.com/
doi:10.1016/j.aca.2010.04.010

characterise the influence factors missing in the expert data. Simultaneous use of these expert and experimental data leads to the IDC. IDC does not require a calibration set, but results may need bias and/or slope correction before use. It is also shown that IDC is a predictive method based on the NAS, i.e. the predicted value is proportional to the inner product between the sample spectrum and the NAS.

IDC is much more efficient than DC. Sometimes it could perform as well as PLSR. It should be noted that IDC and PLSR b-coefficients are not similar, even when their respective predictions are close to each other. The applications of IDC are mainly those in which PLSR is not applicable or can only be used with difficulty, i.e. when construction of the calibration database is problematic or impossible. For instance IDC should be a powerful tool when applied to hyperspectral image analysis. In these situations, IDC provides high analytical added value.

Indirectly, this study showed that three methods of direct calibration: DC, SCB and IDC are built on the same formula and based on two matrices: 1) the pure spectrum $\mathbf{k}$ of the factor of interest, and 2) a $\mathbf{\Sigma}$ matrix characterising the influence factors, defined differently by DC, SBC or IDC. In a future article, we shall show that this same writing provides links between direct calibration and inverse calibration methods such as PLSR or PCR.

A Acknowledgements

This work was made possible within the IRVIN program, carried out with the Skalli winery thanks to a financial support from the Languedoc-Roussillon Region.

B References

References

[1] H.Martens, T.Naes, Multivariate Calibration, Wiley, 1989.

[2] H.Martens, J.P.Nielsen, S.B.Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction, application to near infra-red transmission analysis of powder mixtures, Analytical Chemistry 75(3) (2003) 394–404.

[3] S.Wold, A.Ruhe, H.Wold, W. D. III, The collinearity problem in linear regression, the partial least square (pls) approch to generalized inverses, Journal of Science and Statistical Computations 5 (1984) 735–743.

[4] P.W.Hansen, Pre-processing method minimizing the need for reference analyses, Journal of Chemometrics 15 (2001) 123–131.

[5] J.M.Roger, F.Chauchard, V.Bellon-Maurel, Epo-pls external parameter orthogonalisation of pls, application to temperature-independant measurement of sugar contents in fruits, Chemometrics ant Intelligent Laboratory Systems 66 (2003) 191–204.

[6] R.Marbach, A new method for multivariate calibration, Journal of Near Infrared Spectroscopy 13 (2005) 241–254.

[7] H.Mark, R.Rubinovitz, Chemometric calibration without matrices (almost), in: Pittcon-Chicago, 2009.

[8] B.G.Osborne, T.Fearn, Near infrared spectroscopy in food analysis, Wiley, N.Y., 1986.

[9] A.Lorber, K.Faber, B.R.Kowalski, Net analyte signal calculation in multivariate calibration, Analytical Chemistry 69(8) (1997) 1620–1626.