Data treatment in near infrared spectroscopy

P. Chaminade, A. Baillet and D. Ferrier

Laboratoire de Chimie Analytique, Faculté de Pharmacie, 1 rue Jean-Baptiste Clément, 92296 Châtenay-Malabry Cedex, France

Since the last ten years, the use of near-infrared spectroscopy is increasing due to the improvement of instrumentation and software. The calculation capacity of today computers allows a promising future for this technique. According to the complexity of structural attribution of absorption bands, only a chemometric approach is able to solve qualitative and quantitative applications. Depending on spectrophotometers and analytical context, the signal treatment and the calculation strategy may vary. For an identification task, cluster analysis is built on spectral data from only a few wavelengths using disperse spectrophotometer when principal component analysis is applied on overall spectral range with interferometer based apparatus. The main quantitative calculation modes i.e. MLR, PCR and PLS are discussed.

The near infrared (NIR) takes place between middle infrared and visible region of the spectrum. As quartz is transparent in near infrared region, transmittance measurements of liquids can be done using standards cuvettes and, reflectance measurements of powders can be realised using fiber optics. NIR spectroscopy is thus a method that require few or no sample preparation. Absorption bands observed in NIR spectra are due to overtones of, mainly, hydrogenic stretching vibrations or combination involving stretching and bending modes of vibration. Those bands are thus broader than in middle infrared and spectra are considerably more complex. Due to this complexity, NIR spectroscopy has soon taken advantage of sophisticated calibration techniques and is now using state of the art data treatment.

This article will illustrate, first the spectral transformation and then the two current uses of NIR spectroscopy : qualitative discriminant analysis and quantitative applications.

Experimental

Qualitative discriminant application: mannitol, sorbitol P100T, sorbitol P60, sorbitol P20-60, hydroxyethylcellulose, hydroxypropylcellulose (Roquette, Lestrèmes, France).

Quantitative application: Paracetamol (Sigma, St. Quentin Fallavier, France) with variable amounts of hydroxypropylcellulose (Roquette) and lactose monohydrate (Sigma).

The measurement where done with a BUHLER-NIRVIS spectrometer and a 2 meters optical fibre. Calculations where done using NIRCAL 2.0 software (BUHLER-ANATEC, Uzwil, Switzerland).

Spectrum transformation

Four different goals can be reached by spectra transformation:

Article available at http://analusis.edpsciences.org or http://dx.doi.org/10.1051/analusis:199826040033

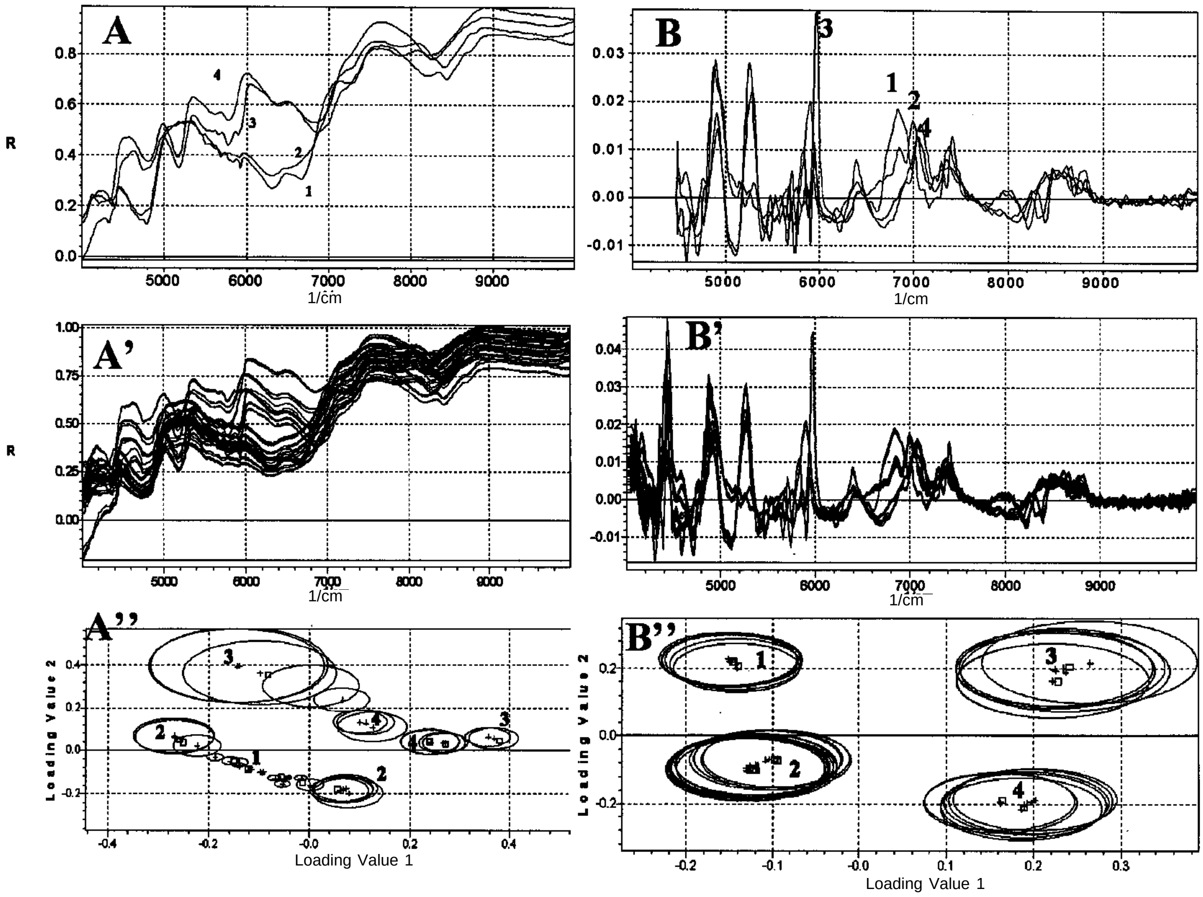

Figure 1. A Typical spectra of mannitol (1), sorbitol (2), hydroxypropylcellulose (3) and hydroxyethylcellulose. A') individual spectra and A") corresponding clusters. B, B' and B") same spectra and clusters after applying derivative.

- correction of baseline offset due to irregular shaped samples,
- reduction of spectral variability for the same class of compound,
- increase of the spectral differences between classes of compounds,
- linearization of the response.

Spectrum transformation is achieved by applying one, or a sequence of, mathematical function. A little library was built using hydroxypropylcellulose, hydroxyethylcellulose, mannitol, and three qualities of sorbitol differing by the size of particles. Typical spectra of those four compounds are presented in figure 1A.

The spectral transformation to be used will depend on the aim of the analysis:

1. does the sorbitol should be simply identified as well as the three other compounds?
2. can we perform a simultaneous identification of the four compounds and the different classes of sorbitol?

In the first case, differences in particle size of sorbitol will provoke difficulties for the following reasons. According to the Kubelka-Munk theory, reflectance (*R*) vary with the concentration (*c*), the absortivity (*a*) and the scattering coefficient (*S*) of the material.

$$\frac{(1-R)^2}{2\,.\,R} = \frac{K}{S} = \frac{\ln(10)\,.\,a.c}{S}\,. \qquad (1)$$

Thus, for a pure compound existing with different particle sizes the corresponding spectra will present different baseline offsets. This phenomenon can be seen in figure 2A, where each spectrum corresponds to the average of spectra from the three classes of sorbitol. Sample packing or the strength used to apply an optical fibre on the sample will provokes similar baseline effects as shown in figure 2B where three spectra where recorded for the same sorbitol batch.

Resulting from this diversity of structure between compounds and their respective qualities, spectra of figure 1A' totally overlap. The result of cluster analysis in figure 1A" shows that none of those compounds could be distinguished from another. The variability is high for each class of compound (as explained previously) but low between classes of compounds due to structural and spectral similarities of on one hand sorbitol and mannitol, and on the other hand, hydroxy ethyl and propyl cellulose.

Working with the first derivative of the previous spectra will overcome those problems. First, working with derivative is of common use to enhance spectral differences between compounds [1,2] (Fig. 1B). Second, as derivative deals with reflectance variation along the spectrum, spectra

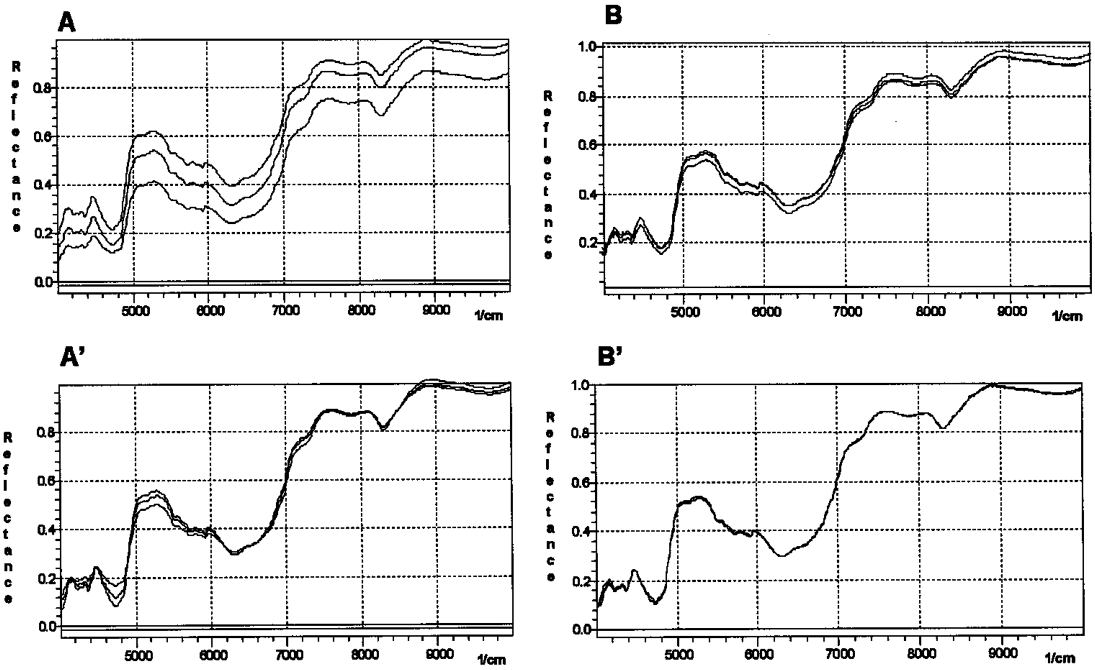

Figure 2. Typical spectra of each sorbitol class without A) and with A') MSC. Representative spectra of one batch of sorbitol without B) and with B') MSC.

differing from only their baseline level will present the same derivative. Figure 1B' shows that the effect of batch and measurements variability is minimised. As a consequence, cluster analysis (Fig. 1B'') now shows four well distinct clusters representing the four compounds. The goal of spectra transformation is now achieved since the intra-class variability is minimised and the inter-class variability is maximised.

The case were the three different classes of sorbitol should be distinguished together with the three other compounds will need a different spectrum transformation. Figure 2A' shows typical spectra of the three sorbitol classes transformed by Multiplicative Scatter Correction [3]. Spectra from the three batches still exhibit differences while the three measurements of the same batch (Fig. 2B') now looks similar. In resulting cluster analysis (Fig. 3): mannitol, hydroxyethylcellulose and hydroxypropylcellulose are correctly separated from the three individualised sorbitol classes. However, the shape of clusters shows that intra-class variability is not as well corrected as in the preceeding case.

As shown is this example, spectrum transformation is achieved by using classical functions such as derivative, or "special" functions devoted to near infrared analysis.

Classical functions include normalisations, derivatives and smoothing.

- Normalisation can be considered with different algorithms, either, at each wavelength, reflectance is divided by the average reflectance, the maximum reflectance, the sum of the total reflectance or each spectra is scaled beween 0 and 1. The classical purpose of normalisation is the reduction of offset deviations in spectra.
- The first derivative is classically computed by replacing the reflectance value at one wavelength by the difference of reflectance at the nearest wavelengths. The second derivative may be computed by submiting the first derivative to the same treatment. More sophisticated algorithm are of common use such as the Savitsky-

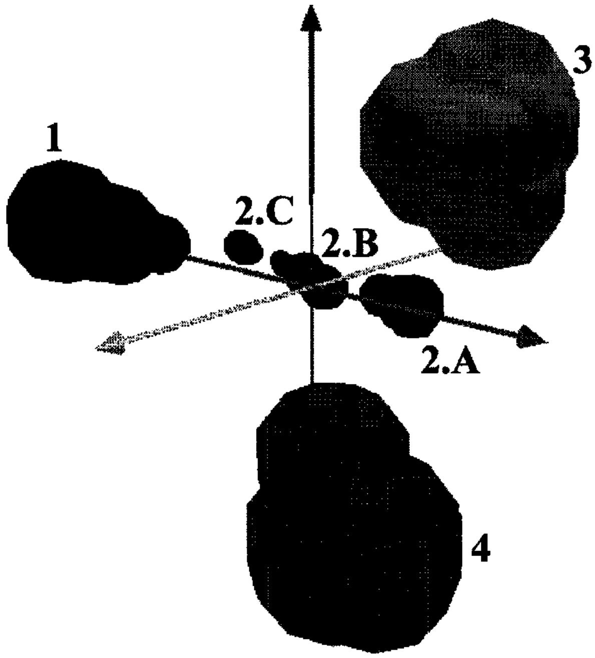

Figure 3. Cluster analysis for the identification of mannitol (1), hydroxypropylcellulose (3), hydroxyethylcellulose (4) and 3 classes of sorbitol (1 A,B,C).

Golay algorithm [4] which uses a polynomial fit of spectra. As shown in the preceeding example, derivative is used to improve spectral differences but may also correct baseline effects.

- Smoothing is used to minimise the effect of spectral noise. Here also, algorithms using a polynomial fit of spectra are classically used.

Special functions are devoted to minimise the effect of light scattering:

- Multiplicative Scatter Correction [4] establish a linear relationship between each spectrum and the mean of all spectra used in the calibration. The spectrum is then corrected by substracting the offset of the regression to the reflectance value and dividing this difference by the slope.
- Standard Normal Variate Transformation (SNV) [5] is an algorithm intended to remove effect of light scattering from spectra. SNV normalise each spectrum by dividing the difference between the transmitance and the average transmitance by the standard deviation of transmitance.

Linearization of response: for quantitative applications, the response at each wavelength can be modified to be proportional to the analyte concentration:

- In the case of measurements using transmitance (T), spectra may be converted to absorbance (A) using the Beer-Lambert law $A = \log (1/T)$.
- In the case of reflectance measurements, by using the Kubelka-Munk equation (Eq. 1).

Qualitative discriminant analysis

All qualitative discriminant analysis depends on comparing spectral data of known samples. Some algorithms match spectra across the overall spectral range (interference and filter based spectrometers), while others use only few selected wavelengths (monochromator based spectrometers).

Discriminant analysis at selected wavelength

The method consists in visually examining spectra to select wavelengths where maximal differences appears for the different compounds to be distinguished. Although this method may look simple the main difficulties are first, the elimination of non informative wavelengths and second, the objective evaluation of the pertinence of the wavelength selection when increasing the size of the library with new compounds.

Selecting two wavelengths of interest for the classification leads to represent the groups in a two dimensional space. However, it cannot be expected to classify a large number of different substances using only two wavelengths. The use of more than three wavelengths stops from proceeding with a visual approach and needs to create a mathematical function for locating data in a multidimensional space.

Discriminant analysis is based on comparison of the Mahalanobis distance [6] (D) between the coordinate of an unknown compound to the centre of each group of known compounds recorded in the library. In a multidimensional space, the distance D from a point X to the centre of a group i (X_i) is given by the matrix product:

$$D^2 = \left(X - \bar{X}_i\right)^{\mathrm{T}} M \left(X - \bar{X}_i\right) \qquad (2)\ [7]$$

Where (M) a matrix determining the distance measures of the multidimensional space and ($^{\mathrm{T}}$) denote the transpose of the ($X - X_i$) matrix. The compound is positively identified if its coordinates are included in the confidence interval of the nearest centre.

Discriminant analysis based on the full spectra

Principal component analysis (PCA) is of current use when spectral data are obtained from interferometer based spectrometers. PCA is a mathematical method widely used to explain causes of variance in data, spectra in the case of near infrared spectrometry.

Rapid description of PCA [8]: A set of n spectra can be expressed as a $n \times p$ data matrix D containing n values of reflectance at each of the p wavelength. The data matrix D is first centered, by replacing each reflectance value by its difference to the average reflectance at the given wavelength, which leads to the matrix X. The general equation for principal components (PCs) calculation is:

$$X = TP^{\mathrm{T}} + E \qquad (3)$$

where T is the score matrix,
P^{T} is the transposed loading matrix,
E is the residual matrix.

The scores are the coordinates of spectra in a coordinate system defined by principal components. The loadings are the link between the original system of coordinate (wavelengths) of the X matrix and the principal component space. The residual represent the amount of spectral information not described by the principal components.

From a practical point of view, the calculation consists in an iterative process: equation (2) is used to extract the first $T_1P_1^{\mathrm{T}}$ term from X. The residual matrix E is submitted to the same calculation to extract $T_2P_2^{\mathrm{T}}$ and leads to a new residual matrix which contain less information. This process is repeated until the residual matrix contain an amount of information comparable to the noise level. In order to select the minimum number of PCs needed to model spectra without any loss of valuable spectral information, it is of common practice to use only one part of the spectra to build the model (calibration set) while the other part (validation set) is used to test the occurrence of an overfitting.

From the calculation principle of PCs, several statements can be done:

- principal components are not selected wavelengths or spectral regions but describe variance in the full spectral range, there is much less PCs than wavelengths,
- principal components are used to model spectra,
- each PC describes one part of the total variance. Two PCs cannot describe the same part of variance,
- the amount of variance described by the first principal component is greater than for the second and so forth.
- any addition of new spectra or compound will modify the computation of PCs.

The principle of identification using PCA is very similar to the identification using wavelength. The only difference is that the score of a spectrum is used to locate it on the PCs based coordinate system. Since the number of PCs (usually less than 10) used to describe the whole library is smaller than the number of wavelengths, the distinction between a large number of substances is still possible. Here also, Mahalanobis distances are used to measure the distance between the coordinate of the spectra to be identified and the groups of spectra continuing the library.

The development of a discriminant analysis using PCA corresponds to the selection of the minimum number of PC needed to model spectra. The objective is to model the variability of each class of compounds and to represent each class as an unique cluster (Fig. 1B'') computed from the calibration set. Spectra from the validation set are used to test the prediction ability of the model. As discussed previously, spectrum transformation is used prior to PCA to minimise intra-class variability and enhance inter-class variability.

Quantitative analysis

NIR quantitative analysis are devoted to be secondary methods calibrated against reference techniques such as Karl Fisher in the case of moisture. The objective of the calibration is to obtain results as accurate and precise as the reference method.

Considering mathematical calculations, the same distinction may be done between methods dealing with only a few selected wavelengths and full spectrum methods.

Multiple Linear Regression (MLR) uses a few selected wavelengths to establish the relationship between quantitative characteristics of samples (often called properties) and spectra. The choice of exploitable wavelengths is here also critical. In its principle [9] MLR tends to establish a linear combination of responses that minimizes the error in recalculating properties. This approach is correct if no interfering phenomena are present and no colinearities between properties exist at the selected wavelengths. MLR is also very sensitive to noise. The selection of an inappropriate wavelength(s) may result in the establishment of an inappropriate model and thus to errors in prediction.

Principal Component Regression (PCR) and Partial Least Squares (PLS) [9] are two full spectrum methods based on PCA. The assumption is that if the main cause of variability between spectra is the variation of analyte(s) concentration in the sample, this variation will be depicted by the principal component. Thus PCR and PLS try to establish a correlation between spectra's scores and their corresponding properties.

The first step of PCR is a PCA and then, a MLR is performed between the score matrix and the property values. In the case of PLS, PCs are calculated on both the spectra and the property values. The underlying idea is to find PCs from spectra that better define the properties.

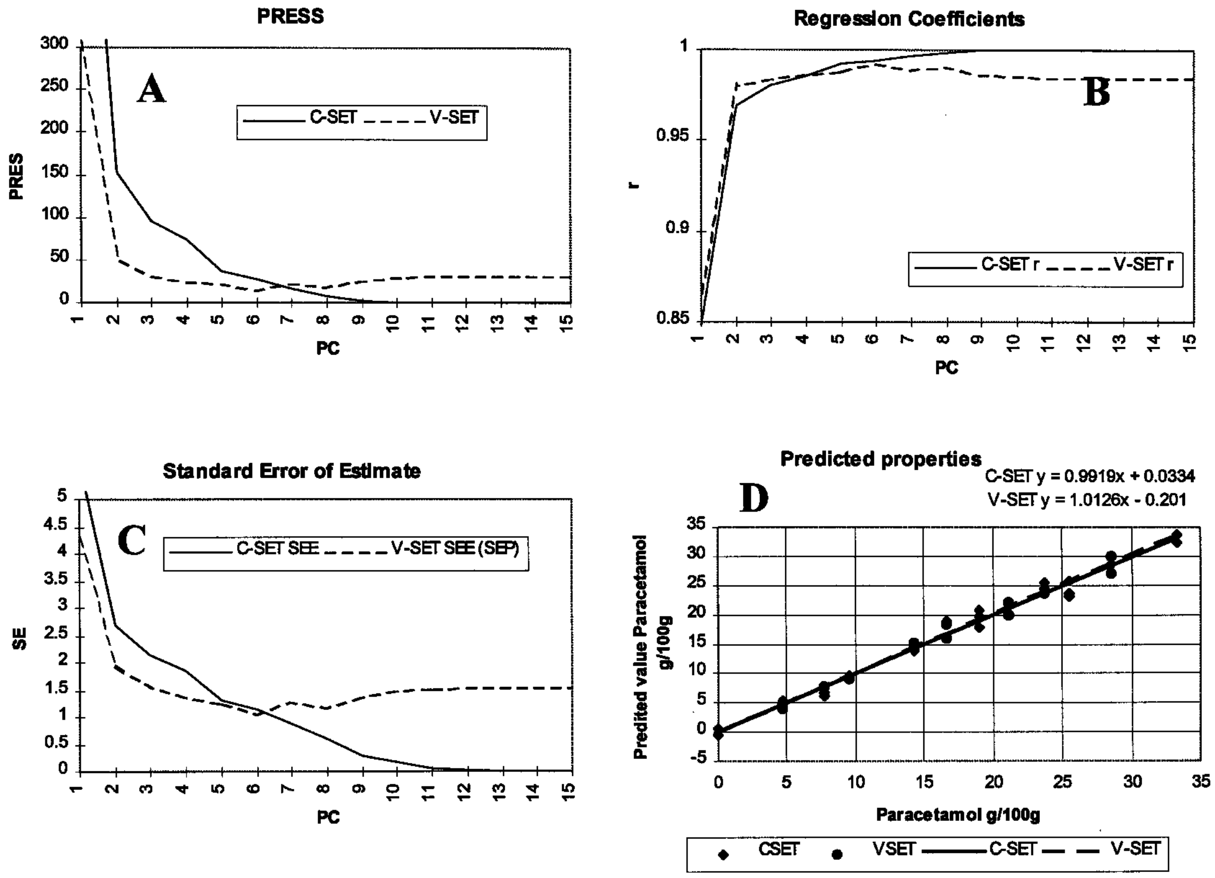

Figure 4. Developpement of a quantitative evaluation of paracetamol (PLS).

Both PCR and PLS present an advantage upon MLR considering noise since only significant variations are selected to build the model.

Here also, the development of the PCR or PLS calibration consists in selecting the minimum number of PCs needed to ensure a reliable calibration. As shown in figure 4A, for the calibration set, the predictive residual sum of square (PRESS) continuously decreases as more PCs are added. The situation is identical for the validation set until six PCs, beyond this point, the PRESS trends to increase which denotes an overfitting of the calibration set spectra. The number of PCs is also chosen with regard to accuracy expressed, as first guess, as the correlation between predicted and true property values (Fig. 4B) and, the precision expressed by the standard error of estimate (SEE) in figure 4C. SEE, also called standard error of prediction (SEP) in the case of the validation set must be consistent with the standard deviation of the reference method. Finally, the calibration is expressed as a linear relationship between predicted and true property values. The slope should be equal to unit for an unskewed calibration, and the intercept close to zero for an unbiased calibration.

In the specific case of PCR, where the correlation is established after the PCA, some PC can be deselected. For example it is possible to work with PCs 1 to 4 and 6 but not 5 if this one is found irrelevant to model the property. From the calculation principle, this feature is impossible for PLS where PCs describe both spectra and property values.

Acknowledgment

The authors would like to thank Kontron Instruments France to have left to our disposal the Nirvis Spectrometer for this study.

References

1. Giese, A.; French, C. S. “The analysis of overlapping spectral absorption bands by derivative spectrophotometry” *Appl. Spec*. **1955**, *9*, 78-96.
2. Osborne, B. G.; Fearn, T.; Hindle, P. H. in “Practical NIR spectroscopy with applications in food and beverage analysis” 2nd Ed. Longman Scientific & Tecnical.
3. Geladi, P.; MacDougal, D.; Martens, H. “Linearization and scatter correction for near infrared reflectance spectra of meat” *Appl. Spec.* **1985**, *39*, 491-500.
4. Savitsky; Golay *Anal. Chem.* **1964**, *36*, 1627.
5. Barnes, R. J.; Dhanoa, M. S.; Lister, M. S. “Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra” *Appl. Spec.* **1989**, *43*(5) 772.
6. Mahalanobis, P. C. *Proc. Natl. Sci.* **1936**, *2,* 49-55.
7. Burns, D. A.; Ciurczak, E. W. in: “Handobook of Near-Infrared Analysis”, Marcel Dekker, 1992.
8. Esbensen, K. et al. In: “Multivariate Analysis in practice”, CAMO, Trondheim, Norvay, 1994.
9. Beebe, K. R.; Kowalski, B. R. “An introduction to multivariate calibration and analysis” *Anal. Chem.* **1987**, *59*(17), 1007A-1017A.