ANEXO V

Orthogonal signal correction in near infrared calibration

M. Blanco, J. Coello, I. Montoliu y M. A. Romero

Analytica Chimica Acta, 434, 2001, 125-132

ELSEVIER

Analytica Chimica Acta 434 (2001) 125–132

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Orthogonal signal correction in near infrared calibration

M. Blanco*, J. Coello, I. Montoliu, M.A. Romero

Departament de Química. Unitat de Química Analítica. Facultat de Ciències. Universitat Autònoma de Barcelona. E-08193 Bellaterra. Spain

Received 16 June 2000; received in revised form 29 December 2000; accepted 10 January 2001

Abstract

The different physical characteristics and treatments of the solid samples are responsible for the spectral variability that takes place in near infrared (NIR) measures. These changes, not related to the analyte concentration, may yield in complex and not very robust calibration models. Mathematical treatments are usually applied for the correction of this variability, being the most common derivation, standard normal variate (SNV) and multiplicative scatter correction (MSC). Orthogonal signal correction (OSC) is a new mathematical treatment designed to minimize, in a set of spectral data, the variability not related with the concentration of the analyte. In this work the application of this new treatment to minimize the spectral differences of two types of samples: production samples and laboratory samples, is evaluated. A method is developed for the determination of the content of the active component in a pharmaceutical preparation by means of PLS calibration. Results obtained by OSC are compared with those obtained with the original data and with those corrected by derivation, SNV and MSC. OSC treatment leads to PLS calibration models with good prediction ability and simpler than those obtained using other pretreatments.

Keywords: Orthogonal signal correction; Near infrared spectroscopy; Pharmaceutical analysis; Data pretreatment

1. Introduction

Pharmaceutical analysis is a field in which the use of near infrared spectroscopy (NIRS) has grown dramatically over the past two decades. The characteristics of the NIR spectrum make possible to analyze solid samples directly, minimizing or eliminating its manipulation, which results in smaller time and cost of analysis. Its speed and low cost of analysis transform NIR spectroscopy into a suitable technique for quality control of pharmaceutical products. In NIR spectroscopy, the assignment of bands to a certain analyte or functional group is really difficult due to their own characteristics (wide and overlapped bands) so, it is necessary the use of multivariate calibration techniques to extract analytical information from NIR spectra, either qualitative or quantitative.

NIR spectrum does not only depend on the chemical composition of the sample. Physical characteristics as size, form and distribution of particles and degree of compression of the sample, affect in a significant way the obtained signal. This is useful in some cases, since it allows the determination of physical parameters of the sample. Nevertheless, the physical differences can derive in multiplicative effects in the spectrum, which together with other additive effects, as baseline shift or chemical absorptions, will complicate the calibration models and will reduce the quality of the results when carrying out a quantitative analysis.

These effects can be modeled introducing in the calibration set all the possible variability of the samples, or they can be reduced applying appropriate

* Corresponding author. Tel.: +34-93-581-1367; fax: +34-93-581-2379.
E-mail address: iqan8@blues.uab.es (M. Blanco).

0003-2670/01/$ – see front matter
PII: S0003-2670(01)00820-0

mathematical treatments [1]. The introduction of all possible sources of variability in the calibration set is not always possible, and if it is possible, it involves to use a very high number of samples in the calibration set, obtaining complex models with a low prediction ability. It is very common the use of mathematical treatments to minimize the spectral variations due to physical changes, being the most used derivation, standard normal variate (SNV) and multiplicative scatter correction (MSC).

Derivation is one of the most used treatments in NIR since it allows the correction of spectral shifts [2,3]. MSC, developed by Geladi et al. [4], is a treatment based on the separation of multiplicative and additive effects of the scatter in NIR measurements, minimizing in this way spectral variations that are not due to the analyte concentration. SNV, developed by Barnes et al. [5] is based on an autoscaling of each spectrum separately. Each spectrum is mean centered and then divided by its standard deviation. Then the new spectra are centered in 0 and its standard deviations are 1. The objective of this treatment is to obtain a common scale for all the spectra. This way, particle size effects are minimized. It has been demonstrated that MSC and SNV are linearly related, so they should provide similar results [6].

Different production batches of a same pharmaceutical preparation have very similar concentrations of the active compound and excipients, which are close to nominal values. So it is really difficult to have a set of production samples which covered the concentration range needed to establish a calibration model. To solve this problem there are several procedures [1], but the simplest one to expand the concentration range consists on adding to the production samples present in the calibration set, laboratory-made samples, prepared by mixing the pure components. This procedure has some advantages: it is less laborious and less expensive than other procedures and it is very simple to obtain the required concentration for the calibration design. Nevertheless, one of the main problems of this approach is that laboratory-made and production samples may present different physical characteristics, because the preparation processes of both types of samples are quite different. This procedure has demonstrated its effectiveness in several works carried out in our investigation group, with the purpose of determining the content of active principle in some pharmaceutical preparations [7,8]. However, in some cases the spectral differences between laboratory-made and production samples are so high that the resulting models are very complex, diminishing this way the prediction ability.

In this work, we study an orthogonal signal correction (OSC) method, developed in our research group, as an application for the pretreatment of spectral data, which will be subjected later to multivariate calibration. The design of this calibration includes a set of samples prepared in the laboratory, with different physical characteristics from those of the production samples. These physical differences are the main reason of the spectral differences between both sample sets. By means of this treatment, we try to minimize or suppress these spectral differences, diminishing the complexity of the calibration model. The obtained results using OSC will be compared with those obtained using other pretreatments.

1.1. Orthogonal signal correction

Orthogonal signal correction is a spectral data treatment technique developed by Wold et al. [9], whose goal is to correct the $\boldsymbol{X}$ data matrix removing the information that is orthogonal to the concentration matrix $\boldsymbol{Y}$. This treatment is applied jointly to all the spectra in the calibration set. Later, the correction on the $\boldsymbol{X}$ matrix can be applied to an external prediction set to evaluate the prediction ability of the calibration model built with the treated data.

The algorithm used in this type of correction is similar to the NIPALS algorithm, commonly used in PCA and PLS, and it is described in the original work of Wold et al. In each step of the algorithm, the weight vector ($\boldsymbol{w}$) is modified, imposing the condition that $\boldsymbol{t} = \boldsymbol{X} \cdot \boldsymbol{w}$ is orthogonal to the $\boldsymbol{Y}$ matrix, and where $\boldsymbol{t}$ is the corresponding score vector. In PLS the condition that weights would be calculated to maximize the covariance among $\boldsymbol{X}$ and $\boldsymbol{Y}$ is imposed, but in OSC just the opposite is attempted, to minimize this covariance, making $\boldsymbol{t}$ as close as possible to the orthogonality with $\boldsymbol{Y}$. The result of this calculation are scores and loadings matrices that contain the information not related to the concentration. Each internal latent variable (score by loading product) removes a part of the $\boldsymbol{X}$ matrix variance. As we removed more variance, the number of internal latent variables to use in the OSC model will be higher. These internal latent variables are similar to

factors in a PLS calibration. Once the information not correlated with the concentration has been modeled, it is removed from the spectral data, subtracting from the $\boldsymbol{X}$ matrix, the product of the scores orthogonal to the concentration ($\boldsymbol{T}$) and the loadings matrices ($\boldsymbol{P}'$):

$$\boldsymbol{X}_{OSC} = \boldsymbol{X} - \sum_{i=1}^{n} T_i P_i'$$

where n is the number of times that the treatment is applied (OSC factors). Only one factor is commonly used in OSC correction, since a second treatment on the corrected $\boldsymbol{X}$ data can remove useful information, reducing the predictive ability. This type of mathematical treatment has already been applied with success to NIR spectral data in calibration transfer [10].

2. Experimental

2.1. Apparatus

NIR spectra were recorded on a near infrared spectrophotometer NIRSystems 6500 (Foss NIRSystems, Raamsdonksveer, The Netherlands) equipped with a fiber-optic probe model AP6641ANO4P. The instrument is controlled by means of a compatible PC, using Vision 2.22 (Foss NIRSystems, Raamsdonksveer, The Netherlands) for the acquisition of data. For the homogeneity of the laboratory samples a Turbula Mixer Type T2C (WAB, Basel, Switzerland) was used.

2.2. Software

OSC and SNV on the spectra were carried out in MATLAB 5.3 (MathWorks, Natick, USA). The OSC routine used was developed in our investigation group. First derivative, MSC and PLS calibrations were carried out with Unscrambler 7.5 (CAMO, Trondheim, Norway).

2.3. Samples

The pharmaceutical preparation used was Nimesulene (Laboratorios Menarini, Badalona, Spain), containing a nominal 50 mg/g of the active compound, nimesulide, in granules, and several excipients. In all, 50 samples have been used; 29 of them were production samples, all corresponding to different batches. Other 21 samples were prepared in the laboratory, covering a concentration range of 25–75 mg/g (±50% of the nominal content) for the active compound, and were obtained by mixture of the different powdered components, properly homogenized in the shaker mixer. These samples have the texture of a quite fine powder so their physical characteristics clearly differ from those of the production samples.

Both types of sample were used to record three spectra over the wavelength range 1100–2500 nm with a fiber-optic probe. Samples were turned over between successive recordings. The three spectra obtained for each sample were averaged, and the average spectra used for the calculations.

Samples were split in two sets, viz. a calibration set consisting of 16 laboratory samples, which cover the concentration range and 11 production samples, covering the variability of production process; and a prediction set comprising 5 laboratory samples and 18 production samples.

2.4. Data processing

The wavelength range 2200–2500 nm, where the analyte presents a weak absorption and the fiber-optic probe provides a higher spectral noise, was removed in all cases. First derivative was applied using a Savitzky–Golay filter with a window size of three points and a second order polynomial. For MSC, wavelength ranges from 1100 to 1500 nm and from 1750 to 2000 nm, where the active compound nimesulide does not present absorption, were used to correct the signal. Partial least squares (PLS) was used for calibration in the wavelength range 1100–2200 nm. Data were centered before use, and the models were constructed by cross-validation. The quality of the results obtained using the different pretreatments was compared calculating the relative standard error of prediction (RSEP, %) [11].

3. Results and discussion

Due to the physical differences between production and laboratory samples, the scatter effect is very pronounced and two blocks of spectra clearly appear. After applying the different mathematical treatments, the spectral differences are attenuated as it can be seen in

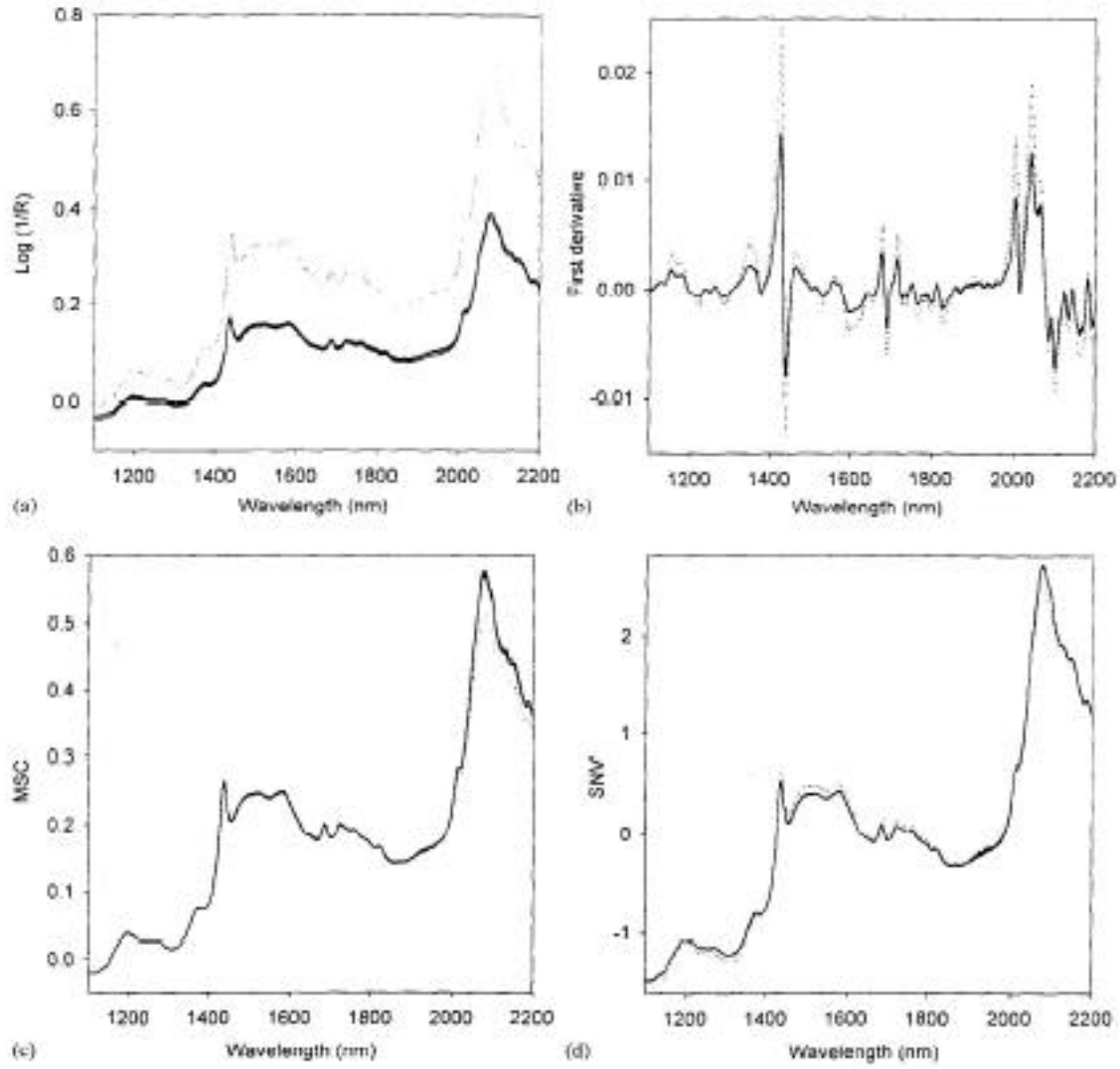

Fig. 1. Production samples (···) and laboratory samples (—) spectra: (a) original spectra; (b) first derivative spectra; (c) MSC spectra; (d) SNV spectra.

the scores plot of the two first components of a principal component analysis (PCA). PLS models were built after correcting the spectra with the different pretreatments; the same calibration and validation sets have been used, as well as the same wavelength range. The standard error of prediction for calibration and validation sets and the complexity of the calibration model, number of PLS components, were compared. The obtained results for the different treatments are discussed below.

3.1. Original spectra

Original spectra are shown in Fig. 1a. The difference in scatter produces a large displacement between the spectra of both types of samples. This is also observed in the scores plot after a PCA of the spectral data (Fig. 2a), where two clusters corresponding to both types of samples appear. A high number of PLS components (seven) is necessary to build a PLS model which will be able to handle the spectral

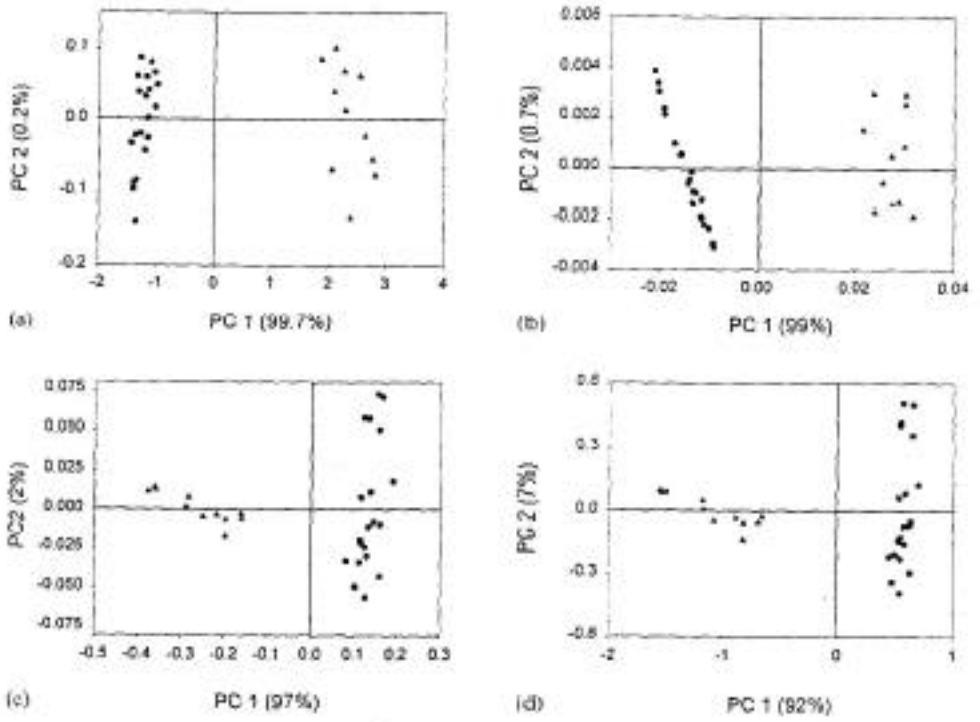

Fig. 2. Principal component analysis (PCA) of production samples (▲) and laboratory samples (●). The explained variance for each PC is shown in brackets: (a) PCA from original spectra; (b) PCA from first derivative spectra; (c) PCA from MSC spectra; (d) PCA from SNV spectra.

differences, but with poor results for the prediction set (see Table 1).

3.2. Derivative spectra

First derivative does not completely remove the differences between laboratory samples and production samples spectra. It only diminishes them (Fig. 1b). PCA still shows both clusters (Fig. 2b). With first derivative spectra, optimal PLS model had three components and, although the complexity of the model decreased, the results produced calibration and prediction errors very similar to those obtained with the original data (Table 1). Models with second derivative were also tested, but they did not improve the results.

Table 1
Calibration (RSEC) and prediction (RSEP) errors obtained with every data pretreatment

	PLS components	RSEC (%)	RSEP (%)
Original spectra	7	2.7	5.6
First derivative	3	4.1	5.3
SNV	5	2.4	2.8
MSC	5	2.4	3.4
OSC	1	3.4	2.9

3.3. MSC spectra

MSC corrected spectra (Fig. 1c) show a wavelength region (towards 2100 nm) where spectra of laboratory-made and production samples are still different. First PC still separates both groups of samples (Fig. 2c). After MSC correction, the spectra were used to build a PLS model, with five PLS components, which provides lower calibration and prediction errors than those obtained with the original data and first derivative spectra (Table 1).

3.4. SNV spectra

After standardization of the spectra using SNV almost all the spectral differences are minimized, although small differences are still observed in some regions of the spectra (Fig. 1d). However, PCA still detects the origin of the samples (Fig. 2d). A PLS model with five components allows to obtain good results, not only for calibration but also for prediction, very similar to those obtained after applying MSC (Table 1).

3.5. OSC spectra

In order to obtain the best quantitation results, the effect of some OSC parameters (number of production samples to introduce in the calibration set, number of latent variables and number of OSC factors) has been studied in order to test their influence on the quantitation of the analyte.

OSC models the information present in the spectra that is not related to the concentration, that means that samples in the calibration set should be similar to those in the prediction set, otherwise the same correction could not be applied to both types of samples. In the case studied, production and laboratory-made samples differ in their particle size and consequently in the scatter properties, so in the calibration set both types of samples should be present. The minimum number of production samples that should be in the calibration set has been studied. With a number of laboratory-made samples fixed at 16, the number of production samples was varied introducing in the calibration set 4, 6, 8, 10 and 11 samples. In Table 2, the obtained results are shown, in all cases the same prediction set, already described in Section 3.3, was used.

Table 2
Results obtained applying OSC varying the number of production samples introduced in the calibration set, fixing at 16 the number of laboratory samples

Number of production samples	Latent variables	PLS components	RSEC (%)	RSEP (%)
4	4	5	6.1	12.9
6	6	1	2.4	6.1
8	6	1	3.1	3.1
10	6	1	3.4	3.4
11	6	1	3.4	2.9

It can be observed that when few production samples were included in the calibration set, OSC could not model and reduce the differences between both types of samples. The number of PLS components was five, and the prediction error very high. When more production samples were added to the calibration set, the obtained models with corrected spectra require only one PLS component, diminishing also the prediction error. The minimum prediction error has been obtained with 11 production samples, so this number of production samples was chosen for the comparison with the other treatments, and that is the reason for having defined this way the calibration and prediction sets. Including more production samples in the calibration set did not reduce the prediction error. This number of production samples may vary in each case and it has to be studied to find the best results, as we have concluded from results not shown in this paper. Moreover, OSC can lead to an overfitted solution, with a very good calibration but with reduced predictive capability. The parameters that allow controlling this overfitting are the number of internal latent variables (LV) and the number of OSC factors. Increasing the number of LV implies to remove more information that is not correlated with the analyte concentration. The number of OSC factors is the number of times that OSC is applied to a set of spectra. This means to make a OSC treatment and to obtain the corrected spectra. Next, the OSC is again applied to the corrected spectra and so on. One OSC factor is usually enough, since a second factor leads to a great overfitting.

In our case, to determine when this overfitting takes place, both parameters were evaluated. In Table 3 the effect of LV is shown, using the calibration set described in Section 2. In the table, the calibration and

Table 3
Results obtained applying the OSC pretreatment with different number of latent variables

Latent variables	PLS components	RSEC (%)	RSEP (%)
1	6	5.1	17.2
2	6	3.8	15
3	6	2.8	6.4
4	1	6.6	14.8
5	1	4.8	11.7
6	1	3.5	2.9
7	1	2.7	4.9
8	1	2.2	5.1

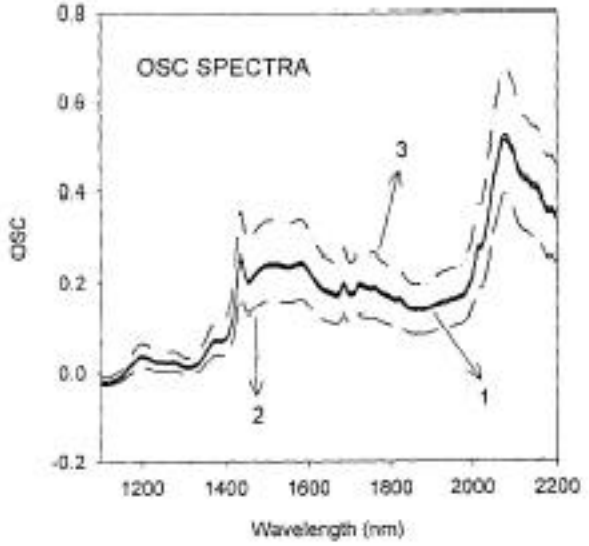

Fig. 3. OSC spectra of laboratory samples and production samples overlapped (1) compared with the original spectra of laboratory samples (2) and production samples (3).

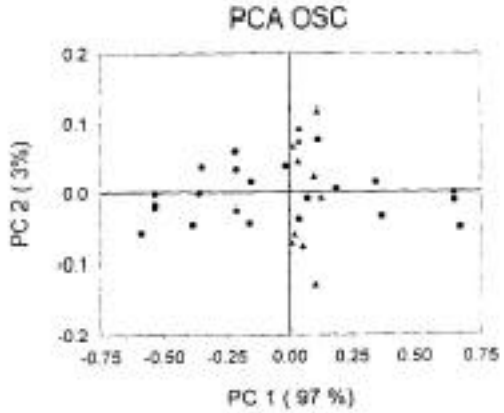

Fig. 4. Principal component analysis (PCA) of OSC corrected spectra from production samples (▲) and laboratory samples (●).

prediction errors for each number of LV are shown. In each step of the calculation, spectra were corrected and the PLS regression calculated, predicting the external group and calculating its error. This way, up to three LV, the number of PLS components is six. From four latent variables, all the PLS models are obtained with one PLS component only. The best prediction results were obtained using six LV, since with more than six, the calibration error diminishes but that of prediction increases yielding an overfitted solution.

Although it is described that one OSC factor is usually enough to correct some data, we have applied OSC to our data twice (two OSC factors). As expected, the second factor led to a very high overfitting in anyone of its LV, for what we decided to work with one OSC factor.

Using OSC, the spectral differences between the two types of samples have been removed. The corrected spectra, from both laboratory and production samples spectra, appear overlapped and located between the original spectra of production and laboratory samples (Fig. 3). The scores plot of OSC corrected spectra shows that the variability between the two types of samples that appeared with the other treatments has been removed (Fig. 4). The samples are now in a single cluster and ordered along the first component in function of their concentration. Production samples appear in the middle of laboratory samples, since they have a content of active compound approximately in the middle of the concentration range.

The results of PLS after OSC correction are shown in Table 1, together with the results provided by the other pretreatments studied. With this treatment, RSEC (%) and RSEP (%) are very similar to those obtained using SNV, but with a lower number of PLS components; they have decreased from five with SNV model to only one with OSC model.

4. Conclusions

The most commonly applied mathematical treatments (derivation, SNV, MSC) of the signal before performing a PLS calibration reduce the spectral differences due to light scatter and reduce the complexity of the calibration model from that obtained with the original spectra. Anyhow, the optimal number of PLS components is still high, and none of them is able to completely eliminate the spectral differences that clearly appears in the scores plot after a PCA of the treated spectra. Orthogonal signal correction succeeds in removing the differences between the spectra even in a case like the studied, where the physical characteristics of production and laboratory-made

samples were so different and with important spectral differences, as it is shown in the plot of the PCA scores. OSC corrected data provide a simpler calibration model, although the pretreatment is much more complex than other pretreatments commonly used, requiring only one PLS component for modeling the data and with a predictive ability comparable to that obtained with other techniques, like SNV, with more complex models.

Acknowledgements

The authors are grateful to Spain's DGICyT (Project PB96-1180) for funding this work. M.A. Romero also acknowledges additional funding from Spain's Ministry of Education and Culture in the form of an FPI grant.

References

[1] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, C. de la Pezuela, Analyst 123 (1998) 135R.
[2] W.F. McClure, NIR News 4 (6) (1993) 12.
[3] W.F. McClure, NIR News 5 (1) (1994) 12.
[4] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (3) (1985) 491.
[5] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (5) (1989) 772.
[6] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, J. Near Infrared Spectrosc. 2 (1994) 43.
[7] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, C. de la Pezuela, Analyst 122 (1997) 761.
[8] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, D. Serrano, Analyst 123 (1998) 2307.
[9] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175.
[10] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, Chemom. Intell. Lab. Syst. 44 (1998) 229.
[11] M. Otto, W. Wegscheider, Anal. Chem. 57 (1985) 63.