The use of net analyte signal (NAS) in near infrared spectroscopy: Interpretablility and figures of merit

Mafalda Cruz Sarragucua, João Almeida Lopes*

REQUIMTE, Serviço de Química-Física, Faculdade de Farmácia, Universidade do Porto, Rua Aníbal Cunha 164, 4099-030 Porto, Portugal

A R T I C L E   I N F O

Article history:
Received 31 July 2008
Received in revised form 25 September 2008
Accepted 2 October 2008
Available online 14 October 2008

Keywords:
Near infrared spectroscopy
Net analyte signal
Multivariate calibration
Partial least squares
Figures of merit

A B S T R A C T

Near infrared spectroscopy (NIRS) has been extensively used as an analytical method for quality control of solid dosage forms for the pharmaceutical industry. Pharmaceutical formulations can be extremely complex, containing typically one or more active product ingredients (API) and various excipients, yielding very complex near infrared (NIR) spectra. The NIR spectra interpretability can be improved using the concept of net analyte signal (NAS). NAS is defined as the part of the spectrum unique to the analyte of interest. The objective of this work was to compare two different methods to estimate the API’s NAS vector of different pharmaceutical formulations. The main difference between the methods is the knowledge of API free formulations NIR spectra. The comparison between the two methods was assessed in a qualitative and quantitative way. Results showed that both methods produced good results in terms of the similarity between the NAS vector and the pure API spectrum, as well as in the ability to predict the API concentration of unknown samples. Moreover, figures of merit such as sensitivity, selectivity, and limit of detection were estimated in a straightforward manner.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Near infrared spectroscopy (NIRS) has been applied for quality control of pharmaceutical drugs [1] to determine physical or chemical properties by means of multivariate calibration [2]. Still, some drawbacks of the technique, such as the poor selectivity and detection limit, make it difficult to employ it in a more resourceful way [3], especially due to the regulations imposed by pharmaceutical regulating agencies. The NIRS non-selectivity, added to the chemical and physical complexity of pharmaceutical formulations produces a very complex spectrum, which results in difficulties to relate the spectra with samples properties. Pre-processing methods remove some of the variability associated with physical effects, such as base line drifts, particle size differences or unwanted light scattering [4,5]. Besides the associated physical problems, the interferences caused by chemical components present in the mixture are not removed by the methods mentioned above.

To overtake this difficulty some authors have used the net analyte signal theory (NAS) as a pre-processing method to retrieve the information concerning the analyte to be calibrated [5,6]. It is known that using NAS as a pre-processing method prior to a regression (e.g., partial least squares (PLS)) might not improve the prediction results. However, the models obtained with NAS pre-processing are simpler and easier to interpret [6]. The concept of NAS was first defined by Lorber (1986) as the part of the data unique to the analyte of interest, and was used to determine figures of merit, such as the limit of detection, precision, accuracy, sensitivity, and selectivity [7]. Algebraically, the NAS can be defined as the part of the signal orthogonal to the space spanned by the interferences (all components except the analyte of interest). Consequently, this vector can provide important knowledge concerning the main spectral features of the analyte of interest, even when the pure component spectrum is unknown. The concept can be useful to achieve a more deep understanding of the chemical structure and to recognize specific segments of the spectrum related to a given analyte. Two main concepts emerged from the original NAS theory [8]. The most commonly used method is the one in which the interferences space is unknown while the other method requires the prior knowledge of that information. Different applications using both NAS approaches were developed with an assortment of goals: outliers detection [9], variables selection [10], multivariate models represented ordinary two-dimensional plots [11–14], analyte concentrations prediction using multivariate sensitivity analysis [12,14–16] and simply as a pre-processing method [5,6]. Other significant application of NAS concept is the estimation of figures of merit in multivariate calibrations [17]. For multivari-
ate calibrations, the analytical division of IUPAC recommends the use of the NAS theory to calculate figures of merit since its estimation is not easily accomplished by the traditional methods [18]. The NAS vector enables the assessment of the figures of merit in multivariate calibrations in a similar approach to univariate methods [19,20].

In this work, the two different approaches to compute the NAS vector were evaluated and compared in the context of quality control of solid oral dosage forms by NIRS. The evaluation was carried out using both qualitative and quantitative approaches. From a qualitative perspective, the interpretability enhancement of the NIR spectrum was assessed by the similarity between the NAS vector and the spectrum of the pure active pharmaceutical ingredient (API). The ability of each method to predict the API concentration of unknown samples was assessed for the quantitative evaluation. This work was divided in two parts. Initially, a comparison of the two methods was performed using a pharmaceutical formulation based on paracetamol. The calibration and test samples were produced in the laboratory according to an experimental design. Consequently, it was ensured that all mixture components were known, as well as their spectra. In the second part, the method that produced the most satisfactory results, in terms of model interpretability and prediction, was applied to four different pharmaceutical formulations (solid oral dosage forms). The model interpretability and prediction, was applied to four different formulations, acetylsalicylic acid, folic acid and neomycin. For these formulations, the spectrum interpretability, the ability to predict the concentration of unknown samples and the figures of merit were assessed.

2. Theory

2.1. Net analyte signal (NAS)

The net analyte signal concept, which allows to retrieve information concerning the analyte of interest when present in a mixture, mathematically can be defined as the vector orthogonal to the space spanned by the interferences. To calculate the NAS vector a projection matrix is required. In the original theory by Lorber (1986), the interferences pure spectra are used to compute the projection matrix [7]. However, in most applications, the pure spectra of the interferences are unknown. Therefore, alternative methods to compute the projection matrix were developed. Lorber et al. (1997) suggested that the NAS vector for analyte $k$ could be determined using Eq. (1) [9].

$$\mathbf{r}_k^\star = \mathbf{Hr}_k$$

(1)

In Eq. (1) $\mathbf{r}^\star (J \times 1)$ is the NAS vector, $\mathbf{H} (J \times J)$ is orthogonal projection matrix, $\mathbf{r} (J \times 1)$ is the mixture spectrum vector, and $J$ is the number of wavelengths. The projection matrix $\mathbf{H}$ was defined in Eq. (2).

$$\mathbf{H}\!=\!\mathbf{I}_J\!-\!(\mathbf{R}_{A-k}^\dagger \mathbf{R}_{A-k})$$

(2)

In Eq. (2), $\mathbf{I}_J (I \times I)$ is an identity matrix, and $\mathbf{R}_{A-k} (A \times I)$ is a matrix that contains information about all sources of variations of the data, except for the $k$th analyte (exception noted in equation by the $-k$ index). This matrix is determined from the rank annihilation of matrix $\mathbf{R} (J \times I)$, reconstructed with $A$ regression latent variables and $I$ samples. The spectra present in matrix $\mathbf{R} (J \times I)$ contains the information about all the components present in the mixture and is representative of new samples. The superscript $\dagger$ represents the Moore-Penrose inverse of the matrix. The previous approach has some limitations that, accordingly to Ferré et al. (2001), can be overcome by initially calculating the NAS vector (Eq. (1)) using the spectrum projected onto $A$ latent variables [14].

$$\mathbf{r}_k^\star = \mathbf{Hr}_A$$

(3)

These methods involve the evaluation of the full projection matrix $\mathbf{H}$, which due the size of the latter ($J \times J$) is time and memory consuming [13]. To avoid the calculation of the orthogonal projection matrix, Faber proposed a different approach based on the PLS regression vector to calculate the NAS vector $\mathbf{r}^\star$ [8].

$$\mathbf{r}^\star = \mathbf{b}(\mathbf{b}^\top \mathbf{b})^{-1}\mathbf{b}^\top \mathbf{r}$$

(4)

In Eq. (4), $\mathbf{r}$ was defined as above and $\mathbf{b} (J \times 1)$ is the regression vector of the PLS regression. The method in Eq. (4) based on the PLS regression vector, is one of the approaches evaluated in this work and will be henceforth designated by method A.

The second method considered in this work uses the spectra of the interferences to estimate the NAS vector [21]. Any spectrum $\mathbf{r} (J \times 1)$ can be split into three independent parts (Eq. (5)).

$$\mathbf{r} = \mathbf{r}_{\text{INT}}^\star + \mathbf{r}_{\text{RES}}^\star + \mathbf{r}_{\text{INT}}$$

(5)

In Eq. (5), $\mathbf{r}_{\text{INT}} (J \times 1)$ is the vector correspondent to the interferences space, $\mathbf{r}_{\text{RES}} (J \times 1)$ is a residual vector accounting for other sources of variability. Vectors $\mathbf{r}$ and $\mathbf{r}^\star$ were defined above.

To define the space spanned by the interferences, a set of blank samples can be used. The NIR spectra from the blank samples include all sources of variability, both physical and chemical. In this work, all the pure components present in the mixture were known. Consequently, the blank samples were a mixture of the interferences corresponding to the formulation excipients. Afterwards, a principal component analysis (PCA) was performed on the NIR spectra form the blank samples, to separate the systematic from the non-systematic contributions. The optimization of the number of principal components can be performed resorting to cross-validation techniques. Although, in pharmaceutical applications the number of excipients present in the mixture can be a good indicator of the number of components chosen for the PCA [21].

The interference vector $\mathbf{r}_{\text{INT}}$ (Eq. (6)) was determined by projecting the mixture spectrum, $\mathbf{r}$, into the interference space.

$$\mathbf{r}_{\text{INT}} = \mathbf{Pp}^\top \mathbf{r}$$

(6)

In Eq. (6), $\mathbf{P}$ is the loading matrix from the PCA performed on the blank samples spectra. In order to define the NAS vector direction an orthogonal anti-projection is required (Eq. (7)).

$$\mathbf{B}_k = (\mathbf{I}_J - \mathbf{Pp}^\top)\mathbf{R}$$

(7)

$\mathbf{B}_k$ contains $J$ spectra orthogonal to the interference space generating the space spanned by analyte $k$, therefore, indicating the NAS vector direction. To define a unique NAS direction, an average over $\mathbf{B}_k$ is taken, and the NAS regression vector, $\mathbf{b}_k$ is thus defined as in Eq. (8).

$$\mathbf{b}_k = \frac{\sum_{i=1}^{i=I} \mathbf{B}_{ki}}{I}$$

(8)

The NAS vector can subsequently be calculated by projecting each spectrum $\mathbf{r}$ onto the regression vector $\mathbf{b}_k$.

$$\mathbf{r}_k^\star = \mathbf{b}_k(\mathbf{b}_k^\top \mathbf{b}_k)^{-1}\mathbf{b}_k^\top \mathbf{r}$$

(9)

The previous method, which uses the knowledge of the interference space to compute the NAS vector, will be hereafter designated by method B.
2.2. Figures of merit

The NAS concept can be used to determine figures of merit in multivariate calibrations as easily as in univariate systems [18]. In this work, the figures of merit sensitivity, selectivity, and limit of detection, were calculated using the equations proposed by Ferré et al. (2001) [14].

Given the NAS vector of the \( i \)th sample, \( \mathbf{r}_i^* \), the net sensitivity vector, \( \mathbf{s}^* \), can be calculated using the \( i \)th element of \( \mathbf{c}_i (\mathbf{c}_i \) is the predicted concentration using \( A \) latent variables of the PLS regression model), as follows in Eq. (10).

\[
\mathbf{s}^* = \frac{\mathbf{r}_i^*}{\mathbf{c}_i^*} \tag{10}
\]

The sensitivity vector is the same for all samples, even when the concentration values have systematic and/or random errors [14]. Sensitivity, \( \text{SEN} \), is defined as the NAS vector generated by an analyte of unit concentration [18] and can be seen as the slope of the calibration curve. Consequently, and according to Eq. (10), the sensitivity can be defined as the norm of the net analyte signal, Eq. (11).

\[
\text{SEN} = ||\mathbf{s}^*|| \tag{11}
\]

Another important figure of merit used to characterize a calibration curve is the selectivity (\( \text{SEL} \)). Selectivity can be defined as the part of the measured signal unique to the analyte of interest. Applying the NAS theory, the selectivity, \( \text{SEL} \), can be expressed by the ratio between the norm of the NAS vector and the norm of the spectra.

\[
\text{SEL} = \frac{||\mathbf{r}^*||}{||\mathbf{s}||} \tag{12}
\]

Using Eq. (12), a different value for each sample is acquired. To find a unique value to characterize the calibration an average of the \( \text{SEL} \) vector was taken.

The determination of the limit of detection, LOD, in multivariate calibrations is an intricate issue. Several methods to calculate the LOD for a multivariate method were proposed [22]. Using the NAS concept and comparing with the definition of LOD for a univariate model, the limit of detection can be determined by Eq. (13) [17,22].

\[
\text{LOD} = 3 \times \left( \frac{\delta}{||\mathbf{s}^*||} \right) \tag{13}
\]

In Eq. (13), \( \delta \) is a measurement of the instrumental noise, and \( \mathbf{s}^* \) was defined above. The LOD calculated by Eq. (13) is based on the assumption that the spectral measurement gives the dominant contributions to prediction uncertainty. However, that assumption is not always true, in particular for Fourier-transform NIRS applications [15,23]. A reasonable estimation to determine the LOD can be made using the expression in Eq. (14) [15].

\[
\text{LOD} = 3\text{RMSEP} \tag{14}
\]

The limit of detection obtained by the former equation assumes that the prediction uncertainties are approximately constant.

3. Experimental

3.1. Samples

In the initial part of this work, samples were paracetamol based formulations (formulation I) with three excipients (see Table 1 for composition). Two sets of samples were produced: one for calibration and one for testing. Both sets of samples were produced in the laboratory according to an experimental design, by weighing the individual components and thoroughly mixing them in a glass mortar prior to analysis. The total mass of each sample was 3 g with an API concentration range varying between 40% and 99% for the calibration samples, as depicted in Table 2.

In the second part of the work, pharmaceutical formulations with four different API were used (formulations II to V). For each formulation, two types of samples were constructed. For the calibration, synthetic samples constructed in the same way as described above were produced. For testing, the samples were based on commercial samples. Since the commercial samples have a narrow API concentration range, it was necessary to increase the concentration range [24]. To that purpose, tablets were macerated and doped with different API concentrations. To obtain lower API concentrations, the macerated tablets were diluted in a mixture of the formulations excipients. For formulation V, the synthetic samples had one API (neomycin). However, the available commercial samples had another API included. Nevertheless, this formulation was treated as if only neomycin was present. The composition and concentration range of each formulation was shown in Tables 1 and 2, respectively.

3.2. NIR acquisition spectra

NIR spectra were recorded in a Fourier-transform near infrared analyser (FTL2A2000, ABB Bomem, Québec, Canada) equipped with an Indium–Gallium–Arsenide (InGaAs) detector. The analyser was equipped with a diffuse reflectance sampling accessory (ABB ACC101), which provides the means for performing NIR analysis in glass containers. The instrument was controlled by the GRAMS/AI (ThermoGalatic Ma, USA) software. Spectra were recorded with a 2 cm\(^{-1}\) resolution with an average of 64 scans over a wavenumber range between 4000 and 10,000 cm\(^{-1}\). The measurements were performed by placing glass flasks containing the samples in the illuminated area (6 mm diameter) of the diffuse reflectance accessory. The background was measured in the beginning of each day of analysis by placing a flask containing a reflectance reference substance (Teflon) in the measurement area.

All calculations were performed using Matlab version 6.5 (MathWorks, Natick, MA, USA).

4. Results and discussion

4.1. Part I: NAS methods comparison

The two methods described in Section 2.1 were compared following a qualitative and quantitative approach. Both NAS methods were applied to the set of samples of formulation I. To assess the interpretability enhancement of the NIR spectrum, the NAS vector was compared with the spectrum of the analyte of interest, in this case the spectrum of pure paracetamol (API). To remove...
Table 2
Summary of the API, concentrations range and number of samples of the analyzed pharmaceutical formulations.

<table>
<thead>
<tr>
<th>Formulations</th>
<th>Type of samples</th>
<th>Mass fraction</th>
<th>No. of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominal</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>I</td>
<td>Calibration</td>
<td>0.83</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.58</td>
<td>0.92</td>
</tr>
<tr>
<td>II</td>
<td>Calibration</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>III</td>
<td>Calibration</td>
<td>0.83</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>IV</td>
<td>Calibration</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>V</td>
<td>Calibration</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.01</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 3
Comparison between the NAS vector and the pure paracetamol spectrum for NAS methods A and B for varying number of components (A).

<table>
<thead>
<tr>
<th>A</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method A</td>
</tr>
<tr>
<td></td>
<td>Pre-processing method</td>
</tr>
<tr>
<td>None</td>
<td>Savitzky–Golay second derivative (15 points filter)</td>
</tr>
<tr>
<td>2</td>
<td>0.7857</td>
</tr>
<tr>
<td>3</td>
<td>0.2309</td>
</tr>
<tr>
<td>4</td>
<td>0.0467</td>
</tr>
<tr>
<td>5</td>
<td>0.0114</td>
</tr>
<tr>
<td>6</td>
<td>0.0042</td>
</tr>
</tbody>
</table>

the sample-to-sample physical differences, several pre-processing methods were applied, namely, Savitzky–Golay filter with first and second-order derivative, and different filter width, standard normal variate (SNV) and multiplicative scatter correction (MSC). To assess the similarity between the NAS vector and the API spectrum, the correlation coefficient between the two vectors was established. Table 3 compares the results for the correlation coefficient for the two methods with different number of PLS components (method A) and PCA components (method B). For both methods, higher correlations were obtained without pre-processing and with a second derivative Savitzky–Golay filter (15 points filter width). For method A, the best correlation was found for 2 components, with a value of 0.862. Three components were needed to achieve the higher correlation for method B, with a correlation coefficient somewhat higher comparing with method A with a value of 0.975.

In Fig. 1, the NAS vector, the spectrum of pure paracetamol, and the regression vector, for both methods, for two and three components, were depicted. In Fig. 2 the same vectors were represented, but with a second derivative Savitzky–Golay filter applied. Analyzing the vectors profiles in Fig. 1, and as established by the correlation data, the NAS vector for method A was very similar to the pure paracetamol spectra for 2 components. For method B, and

![Fig. 1. Comparison between the pure API spectrum (i), NAS vector (ii) and regression vector (iii) for the prediction of paracetamol in formulation I (no pre-processing): (a) method A, 2 components, (b) method B, 2 components, (c) method A, 3 components, and (d) method B, 3 components.](image-url)
for the same number of components, the correlation was lower because the NAS and the regression vector have some features from the interfering components that were not removed. The sharp peak in the region of 7000 cm$^{-1}$ is a characteristic peak from talc. Even increasing the number of components, this feature could not be fully removed. In Fig. 2, the visual comparison is more intricate, since the spectral noise was amplified when the second derivative was applied. Nevertheless, and for method A, the NAS vector and the pre-processed paracetamol spectrum are still similar, particularly for 2 components, as confirmed by the correlation coefficient value. For method B, the results were improved, compared with the non-pre-processed spectra situation, for 2 and 3 components. The main reason appears to be the removal of the sharp talc peak present in the former situation.

In a quantitative approach, the two methods were compared in terms of the ability to predict the concentration of unknown samples, or test samples. The pre-processing methods were chosen based on the results from the vectors comparison evaluation. The prediction ability was estimated by the root mean square error of prediction (RMSEP) and the correlation coefficient values ($r$). Results were reported on Table 4.

For method A, the lowest RMSEP was found for 3 components without pre-processing with a value of 0.054, and a correlation coefficient of 0.964. Three components were also found to give the lowest RMSEP for the pre-processed data, with a value of 0.094, and a correlation of 0.867. Regarding method B, and for the non-pre-processed data the RMSEP value decreased until 3 components were reached, with a minimum value of 0.0131. For the pre-processed data, the lower RMSEP was found for 2 components with a value of 0.114. The correlation coefficients were 0.842, and 0.826, for the non-pre-processed and pre-processed data, respectively. When comparing both methods in what concerns the prediction results, method A was observed to provide a lower RMSEP value. Pertaining to the similarity between the NAS vector and the spectrum of the pure component, a higher correlation was found for method B.

To choose the best method, the qualitative and the quantitative approaches have to be pondered. While the prediction ability of method A was observed to be better than method B, the NAS vector was more similar to the pure paracetamol spectrum for the latter. Note, that in the current practice, calibrations are typically optimized only in terms of bias. Both cross-validation error (RMSECV) and prediction error (RMSEP) are used. Considering that the NIR spectra interpretability was also possible when using method A, this method was selected for the analyses of the remaining formulations.

![Fig. 2](image.png)

**Fig. 2.** Comparison between the pure API spectrum (i), NAS vector (ii) and regression vector (iii) for the prediction of paracetamol in formulation I (second derivative Savitzky–Golay filter pre-processing): (a) method A, 3 components, (b) method B, 2 components, (c) method A, 3 components, and (d) method B, 3 components.

<table>
<thead>
<tr>
<th>Pre-processing</th>
<th>Method A</th>
<th>Method B</th>
<th>Method B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSEP (% w/w)</td>
<td>$r$</td>
<td>RMSEP (% w/w)</td>
</tr>
<tr>
<td>None</td>
<td>2</td>
<td>15.1</td>
<td>0.827</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5.4</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8.2</td>
<td>0.919</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>14.9</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>13.1</td>
<td>0.894</td>
</tr>
<tr>
<td>Savitzky–Golay second derivative (15 points filter)</td>
<td>2</td>
<td>10.0</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.4</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17.8</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>19.9</td>
<td>0.804</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>21.7</td>
<td>0.754</td>
</tr>
</tbody>
</table>
4.2. Part II

In the previous section it was shown that the NAS theory was capable to retrieve the information regarding the analyte of interest and to predict its concentration in unknown samples. However, the formulation used, had a high API content, and the mixture spectrum was dominated by the API spectrum. In what follows we assessed, the ability of the NAS theory to retrieve the information even when the API has a low concentration. The API concentration in unknown samples and the figures of merit were calculated for four formulations, two with a high API content and two with a low API content (see Tables 1 and 2).

The number of PLS components to be used in method A, was determined based on the lower prediction error RMSEP. For all formulations, except formulation IV, the optimum number of components found was 3. Formulation IV had a lower RMSEP for 4 components. The NAS vector, the pure API spectrum and the regression vector for each formulation were depicted in Fig. 3. For the high API content formulations (II and III), the similarity between the NAS vector and API spectrum was confirmed. However, for formulations IV and V (low API content), the NAS vector does not contain any of the main spectral patterns observed in the pure API spectrum.

The prediction results for the concentration of unknown samples and the figures of merit for each system were shown in Table 5. Since the different formulations have different API concentrations, the RMSEP value cannot be compared. To allow the comparison between the prediction results obtained for the different formulations, the relative error (RMSEP divided by the nominal concentration of the API) was determined. The prediction ability of the method was better for high API content formulations as expected. Conversely, the worst results were found for formulations IV and V with low API contents. Formulation V was undoubtedly, the worst result in prediction ability, with a relative error 10-fold higher, compared with the other formulations. The rationalization of this result can reside in the fact that the test samples, based on commercial products, contain another API other than neomycin (see Section 3.1). The presence of another species in the test samples can explain the higher RMSEP value, since the calibration error (RMSEC) is in the same order or magnitude as the calibration errors from the other three formulations. Comparing the calibration error (RMSEC) with the correspondent prediction error (RMSEP), the former is always lower than the latter, which is attributed to the fact that the test samples have different physical properties from the calibration samples (see Section 3.1).

Sensitivity, selectivity, and limit of detection, were calculated for each formulation, based on the NAS theory, as described in Section 2.2. As mentioned before, the selectivity was determined for each of the calibration samples. However, to have a global perspective, the mean of the selectivity values was taken. The selectivity values were very similar for all formulation with the exception of formulation IV. The latter has a selectivity value almost 10-fold lower than the others. This result can be explained by the higher number of PLS components used (4) against the 3 used for the other formulations. The sensitivity value should be in accordance with the API content. A higher API corresponds to a higher sensitivity. For formulations II, III, and V the relationship was true. However, formulation IV had the lower nominal API content of all, and had

<table>
<thead>
<tr>
<th>Formulations</th>
<th>A</th>
<th>RMSEC (% w/w)</th>
<th>RMSEP (% w/w)</th>
<th>Relative error (%)</th>
<th>r</th>
<th>SEL (mean) (dimensionless)</th>
<th>SEN (dimensionless)</th>
<th>LOD (% w/w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>3</td>
<td>0.9</td>
<td>4.5</td>
<td>4.9</td>
<td>0.960</td>
<td>0.403</td>
<td>17.8</td>
<td>13.5</td>
</tr>
<tr>
<td>III</td>
<td>3</td>
<td>0.9</td>
<td>4.8</td>
<td>5.8</td>
<td>0.651</td>
<td>0.322</td>
<td>16.7</td>
<td>14.4</td>
</tr>
<tr>
<td>IV</td>
<td>4</td>
<td>0.9</td>
<td>0.2</td>
<td>7.0</td>
<td>0.932</td>
<td>0.061</td>
<td>23.0</td>
<td>0.60</td>
</tr>
<tr>
<td>V</td>
<td>3</td>
<td>0.8</td>
<td>38.5</td>
<td>385.0</td>
<td>0.880</td>
<td>0.334</td>
<td>3.75</td>
<td>116</td>
</tr>
</tbody>
</table>
the higher sensitivity value. Again, this value can be explained by the number of PLS components chosen for this formulation.

In multivariate systems, the detection limit does not only depend on the mathematical model and the calibration sets used but also on the presence of others analytes in the sample to be analysed. For that reason, the values of the limit of detection determined for the four systems were different from each others. The LOD values for formulation II and III were very similar, since the formulations had similar content and number of excipients. For formulation IV the LOD value was smaller due to the fact that the nominal content of the respective API was also lower compared with the API from formulations II and III. The high LOD value observed for formulation V can be explained by the high RMSEP value, since the LOD was calculated based on the latter (Eq. (14)).

5. Conclusions

In this paper, the ability of the NAS theory to improve the interpretability of NIRS based calibrations for pharmaceutical solid oral dosage forms was assessed. Two different approaches to compute NAS were evaluated: methods A and B. The former was based on the PLS regression vector while the latter was based on the knowledge of API free formulations NIR spectra (blank samples). A pharmaceutical formulation consisting of synthetic samples based on paracetamol was used to test the methods. Method A was chosen to pursue the work because it proved to have better predictive ability. In the second part of the work, method A was applied to four pharmaceutical formulations with different API. Regression models were able to accurately predict the API concentration in unknown samples. It was also possible to recover the spectrum of the analyte being calibrated, rather like in MCR, although without the need to impose rotational ambiguity constraints. The NAS method was able to estimate the pure API spectrum for the two high API content formulations (API mass fraction above 0.8) but provided poor estimates for the other two formulations (API mass fraction below 0.08). The figures of merit were calculated for the multivariate calibration model in a similar way to univariate calibration.

Acknowledgement

Mafalda Cruz Sarragucã§a would like to acknowledge financial support from the Fundacã§a para a Ciência e Tecnologia (FCT), Portugal (PhD grant, ref: SFRH/BD/32614/2006).

References