UvA-DARE (Digital Academic Repository)

A New Indicator for Optimal Preprocessing and Wavelengths Selection of Near-Infrared Spectra

Skibsted, E.; Boelens, H.F.M.; Westerhuis, J.A.; Witte, D.T.; Smilde, A.K.

DOI

[10.1366/000370204322886591](https://doi.org/10.1366/000370204322886591)

Publication date

2004

Published in

Applied Spectroscopy

[Link to publication](#)

Citation for published version (APA):

Skibsted, E., Boelens, H. F. M., Westerhuis, J. A., Witte, D. T., & Smilde, A. K. (2004). A New Indicator for Optimal Preprocessing and Wavelengths Selection of Near-Infrared Spectra. *Applied Spectroscopy*, 58, 264-271. <https://doi.org/10.1366/000370204322886591>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

New Indicator for Optimal Preprocessing and Wavelength Selection of Near-Infrared Spectra

E. T. S. SKIBSTED, H. F. M. BOELEN, J. A. WESTERHUIS,* D. T. WITTE, and A. K. SMILDE

Process Analysis and Chemometrics, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands (E.T.S.S., H.F.M.B., J.A.W., A.K.S.); Pharmaceutical Site Maaloev, Novo Nordisk A/S, Novo Nordisk Park, 2760 Maaloev, Denmark (E.T.S.S.); and QA/ALC-A, N.V. Organon, Molenstraat 10, P.O. Box 20, 5340 BH Oss, The Netherlands (D.T.W.)

Preprocessing of near-infrared spectra to remove unwanted, i.e., non-related spectral variation and selection of informative wavelengths is considered to be a crucial step prior to the construction of a quantitative calibration model. The standard methodology when comparing various preprocessing techniques and selecting different wavelengths is to compare prediction statistics computed with an independent set of data not used to make the actual calibration model. When the errors of reference value are large, no such values are available at all, or only a limited number of samples are available, other methods exist to evaluate the preprocessing method and wavelength selection. In this work we present a new indicator (SE) that only requires blank sample spectra, i.e., spectra of samples that are mixtures of the interfering constituents (everything except the analyte), a pure analyte spectrum, or alternatively, a sample spectrum where the analyte is present. The indicator is based on computing the net analyte signal of the analyte and the total error, i.e., instrumental noise and bias. By comparing the indicator values when different preprocessing techniques and wavelength selections are applied to the spectra, the optimal preprocessing technique and the optimal wavelength selection can be determined without knowledge of reference values, i.e., it minimizes the non-related spectral variation. The SE indicator is compared to two other indicators that also use net analyte signal computations. To demonstrate the feasibility of the SE indicator, two near-infrared spectral data sets from the pharmaceutical industry were used, i.e., diffuse reflectance spectra of powder samples and transmission spectra of tablets. Especially in pharmaceutical spectroscopic applications, it is expected beforehand that the non-related spectral variation is rather large and it is important to remove it. The indicator gave excellent results with respect to wavelength selection and optimal preprocessing. The SE indicator performs better than the two other indicators, and it is also applicable to other situations where the Beer–Lambert law is valid.

Index Headings: Spectral preprocessing; Wavelength selection; Near-infrared spectroscopy; Error indicator; Net analyte signal; Signal-to-noise ratio; Pharmaceutical powders and tablets.

INTRODUCTION

Near-infrared (NIR) spectroscopy is gaining popularity as a quantitative analytical method in the pharmaceutical industry.^{1–3} Quality control of incoming raw materials and quantitative analysis of intermediate^{4,5} and finalized products³ are examples of that. Spectra can be recorded quickly and in a non-invasive manner and they can be combined with a multivariate calibration technique, e.g., principal component regression⁶ (PCR) and partial least-squares regression⁷ (PLS), whereby quantitative measures

can easily be obtained. One known problem in near-infrared spectroscopy is spectral variations that are not related to the property of interest.⁸ This non-related variation is especially important in pharmaceutical applications of NIR. In the pharmaceutical industry, spectra are often recorded in reflectance mode. Varying particle sizes and varying compression of, e.g., powders cause non-related spectral variation. To correct for this variation various spectral preprocessing techniques are used prior to calibration, e.g., multiplicative scatter correction⁹ (MSC), offset correction, or Savitzky–Golay¹⁰ derivatives. Another problem is that if a large part of the recorded spectrum does not contain any information about the analyte, wavelength selection becomes very important. Several methods have been proposed for wavelength selection.^{11,12} Until recently, it was believed that full spectrum methods, e.g., PLS, would automatically overcome the problem of wavelength selection by setting the regression coefficients for non-informative wavelengths to zero or near zero. However, this is not the case and PLS-based calibrations can in many cases be improved by a proper selection of wavelengths.¹³

The most common way of judging whether a preprocessing method is beneficial for the analytical performance is to compute the prediction uncertainty for an independent test set, i.e., the root mean square error of prediction (RMSEP) or root mean square error of prediction cross-validation (RMSECV) if only a smaller dataset is available, and then select the preprocessing method that gives the lowest RMSEP/RMSECV. Some pitfalls with this method are that it requires a fairly large number of samples, i.e., both calibration and test set data. Secondly, if the uncertainty of the reference values is high then judgments are based on reference values with errors. Finally, when using PCR or PLS the RMSEP/RMSECV values are influenced by the model dimensionality. If the model dimensionality is not estimated correctly with some kind of validation technique, then the RMSEP/RMSECV values will be misleading and therefore, judgments of preprocessing method selection or wavelength selection may also be incorrect.

Other methods exist to help choose the optimal preprocessing method, i.e., methods using the net analyte signal (NAS) concept. Net analyte signal is defined as the part of a signal that is unique for the analyte of interest.¹⁴ Lorber¹⁴ demonstrated how figures of merit, e.g., multivariate sensitivity, signal-to-noise ratio, selectivity, and limit of detection could be computed from the net analyte signal of the analyte. These figures of merit can be used

Received 19 June 2003; accepted 30 October 2003.

* Author to whom correspondence should be sent. E-mail: westerhuis@science.uva.nl.

to judge whether a preprocessing method is beneficial for the analytical performance, and they can also be used for wavelength selection. Faber¹⁵ used the inverse multivariate sensitivity of the analyte to judge whether a certain preprocessing method, e.g., derivative, would improve the predictive ability of the calibration model or not. Xu and Schechter¹⁶ developed an error indicator for wavelength selection. Boelens et al. have also demonstrated the usability of NAS for improving the detection limit for a spectroscopic process analysis by tuning Savitzky–Golay filters.¹⁷ All these methodologies use the net analyte signal of the analyte of interest.

In this work we introduce a new error indicator called the signal-to-error indicator (SE). A signal-to-error (SE) value is computed for the analyte when various preprocessing methods and wavelength selections are applied to the spectra. The highest SE value indicates the optimal preprocessing and wavelength interval.

We will demonstrate the performance of the inverse sensitivity indicator, the error indicator, and the signal-to-error indicator with two NIR data sets from different stages in a pharmaceutical tablet production process. The indicators are compared to the standard PLS methodology and the RMSECV. For the applications presented in this paper the PLS method is used as a standard to which the other indicators can be compared. This is possible since the reference method is known to be accurate. The first set contains spectra of powder samples after mixing the tablet constituents. In the second data set, finalized tablets using the powder composition from the first data set are measured. In both cases the analyte is the active pharmaceutical ingredient (API) and the optimal preprocessing and wavelength selection is sought.

First some theory about net analyte signal and the method to compute figures of merit will be presented. Secondly, the different error indicators will be described and compared. Then in the Experimental section the instrumentation and different data sets used are described in detail, and finally, in the Results and Discussion section the different error indicators are compared and the results are commented on.

THEORY

Notation. Boldface capital characters denote matrices, boldface lower-case characters denote vectors, and lower-case italic characters denote scalars. $\|\mathbf{r}\|$ is the Euclidean norm of the vector $\mathbf{r}$, superscript T denotes the transposed matrix or vector, and the superscript + denotes the Moore–Penrose generalized inverse of a matrix. The matrix $\mathbf{I}_J$ is the $J \times J$ identity matrix.

Net Analyte Signal. The net analyte signal is defined as the part of a spectrum that is orthogonal to a subspace spanned by the spectra of all constituents except the analyte, i.e., all interfering constituents.¹⁴ So the net analyte signal of analyte k can be found by the following orthogonal projection:

$$\mathbf{r}_k^* = (\mathbf{I}_J - \mathbf{S}_{-k} \mathbf{S}_{-k}^+) \mathbf{r}_k \quad (1)$$

$$\mathbf{s}_k^* = (\mathbf{I}_J - \mathbf{S}_{-k} \mathbf{S}_{-k}^+) \mathbf{s}_k \quad (2)$$

where $\mathbf{r}_k$ is a $J \times 1$ vector containing the spectral response for a sample including the analyte k measured at J wavenumbers. The pure analyte spectrum $\mathbf{s}_k$ is a $J \times 1$

vector. $\mathbf{S}_{-k}$ is a $J \times L$ matrix with L spectra of blank samples. In some publications¹⁴ pure spectra of the interfering constituents are used to construct the $\mathbf{S}_{-k}$ matrix. In our experience this is not the best method, e.g., pure spectra are not always available and the pure constituent spectrum might differ slightly in shape from the spectral contribution in a mixture of interfering constituents. Practically, the $\mathbf{S}_{-k}$ matrix is most easily constructed by measuring mixtures of the interfering constituents. $\mathbf{S}_{-k}^+$ is its Moore–Penrose inverse, an $L \times J$ matrix. $\mathbf{r}_k^*$ and $\mathbf{s}_k^*$ are $J \times 1$ vectors called the net analyte signal vector of the k th constituent. The net analyte signal for constituent k in any sample can now be computed with Eq. 1.

Inverse Sensitivity. Various figures of merit¹⁴ can be computed using the net analyte signal concept, e.g., analyte sensitivity. Faber¹⁵ evaluated the effect of various preprocessing methods of near-infrared spectra with an error indicator based on computation of the inverse of the analyte sensitivity (α^{-1} ; from here we denote this as invSEN) using the net analyte signal concept. Faber used the assumption that the length of the net analyte signal vector is proportional to the concentration of the analyte. Faber converted the net analyte signal vector into a scalar value by taking the Euclidean norm¹⁵ of the net analyte signal vector and plotted the value against the analyte concentration of the sample, thereby constructing a univariate calibration plot. The analyte sensitivity can then be computed with:

$$\alpha = \frac{\|\mathbf{r}_{k,c}^*\|}{c_{k,c}} \quad (3)$$

$$\text{invSEN} = \alpha^{-1} = \frac{c_{k,c}}{\|\mathbf{r}_{k,c}^*\|} \quad (4)$$

where $\|\mathbf{r}_{k,c}^*\|$ is the norm of the net analyte signal of a calibration sample with concentration $c_{k,c}$ and the slope of the calibration line α is the sensitivity of analyte k .

Faber concluded that a preprocessing method is beneficial for the final predictive ability if the inverse sensitivity is decreasing with that particular pretreatment. The effect on the inverse sensitivity when doing first and second derivatives compared to multiplicative scatter corrected (MSC) spectra was evaluated. This indicator needs a collection of spectra to span the interference space and spectra containing the analyte and their respective reference concentrations of the analyte to compute analyte sensitivity.

Error Indicator. Xu and Schechter¹⁶ developed an error indicator (EI) for wavelength selection. The assumption for their EI is that the prediction error in multivariate analysis is determined by the quality of the corresponding net analyte signal. By minimizing the relative error in the norm of the NAS, the analytical conditions are optimized and lower prediction errors are achieved. The EI was defined as follows:

$$\text{EI} = \frac{\text{var}(\|\mathbf{r}_k^*\| - \|\mathbf{r}_{k,\text{true}}^*\|)^{1/2}}{\|\mathbf{r}_{k,\text{true}}^*\|} \quad (5)$$

Due to non-related variations (interferents or baseline offsets) the norm of the NAS may be affected. The numerator of the EI describes the variance in the norm of the NAS caused by noise in the spectra due to non-related

variations. Xu and Schechter assume that the noise in the spectra due to non-related variations is homoscedastic, i.e., each wavenumber has the same variance, and that the noise is not correlated for neighboring wavenumbers. In that case, the variance in the norm of the NAS due to non-related variation can be written as follows:¹⁸

$$\text{var}(\|\mathbf{r}_k^*\| - \|\mathbf{r}_{k,true}^*\|) = \frac{[(2\|\mathbf{r}_{k,true}^*\|^2 s^2 + (Js^2)^2)]}{(\|\mathbf{r}_k^*\| + \|\mathbf{r}_{k,true}^*\|)^2} \quad (6)$$

Here J is the number of wavenumbers in the spectra used. The standard deviation of the spectral noise described above is represented by s . Since $\|\mathbf{r}_{k,true}^*\|$ cannot be known, Xu and Schechter propose to replace it with $\|\mathbf{r}_k^*\|$, which leads, according to Ferre and Rius,¹⁸ to the following expression for the error indicator:

$$\text{EI} = \frac{\left[s^2 \left(1 + \frac{J^2 s^2}{4\|\mathbf{r}_k^*\|^2} \right) \right]^{1/2}}{\|\mathbf{r}_k^*\|} \quad (7)$$

The standard deviation of the spectral noise, s , is found from the net analyte signal regression plot (NASRP). First take the NAS vector of the pure analyte spectrum $\mathbf{s}_k^*$ and the NAS vector of a sample spectrum containing the analyte $\mathbf{r}_k^*$. Then the absorbance at each wavelength j in $\mathbf{s}_k^*$ is plotted against the absorbance in $\mathbf{r}_k^*$ at the same wavelength, for all $j = 1, \dots, J$ wavelengths in the vectors. This results in the NASRP plot. In the ideal case with no non-related variation, both NAS vectors will point in the same direction and the points in the NASRP plot will form a perfectly straight line passing (0, 0). The assumption made by Xu and Schechter¹⁶ is that at each wavelength the error is normally distributed with the same standard deviation, i.e., white noise. A straight line is fitted through the points in the NASRP plot in a least-square sense and by computing the residual vector, i.e., deviation of each of the points from the line, s can be computed:¹⁸

$$s = \sqrt{\frac{\mathbf{e}_{k,\text{res}}^T \cdot \mathbf{e}_{k,\text{res}}}{J - 1}} \quad (8)$$

where $\mathbf{e}_{k,\text{res}}$ is a $J \times 1$ vector containing the residuals. The residuals are computed in the following manner:

$$\mathbf{e}_{k,\text{res}} = \mathbf{r}_k^* - \mathbf{s}_k^* c_k \quad (9)$$

$$c_k = \mathbf{r}_k^{*T} \cdot \mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2 \quad (10)$$

The error indicator needs a collection of blank spectra to span the interference space, the pure analyte spectrum, and a sample spectrum containing the analyte.

Signal-to-Error Indicator. In this work we present a new indicator based on the computations of the signal-to-error (SE). We assume that the error in the spectra is made of two contributions, i.e., noise and bias. If a certain preprocessing method or wavelength selection is not removing unwanted interference, then extra blank samples may have a small contribution orthogonal to $\mathbf{S}_k$ when they are projected onto the interference space. We compute this contribution as the projection ($PROJ_{\text{blank}}$) of some extra blank spectra ($\mathbf{r}_{\text{blank}}$) on the normed $\mathbf{s}_k^*$ vector, i.e., normed to unit length. We call the normed $\mathbf{s}_k^*$ vector for the net analyte signal regression vector $\mathbf{nas}_{\text{reg}}$.

FIG. 1. Geometrical display of the interference space and the net analyte signal vectors.

$$PROJ_{\text{blank}} = \mathbf{r}_{\text{blank}}^T \cdot \frac{\mathbf{s}_k^*}{\mathbf{s}_k^{*T} \mathbf{s}_k^*} \quad (11)$$

$$PROJ_{\text{blank}} = \mathbf{r}_{\text{blank}}^T \cdot \mathbf{nas}_{\text{reg}} \quad (12)$$

The error taking into account both bias and noise is computed by:

$$\begin{aligned} \text{error} &= \sqrt{\frac{\sum_{i=1}^I (PROJ_{\text{blank},i} - 0)^2}{I}} \\ &= \sqrt{\frac{\sum_{i=1}^I (PROJ_{\text{blank},i})^2}{I}} \end{aligned} \quad (13)$$

In the nominator we use I and not $I - 1$ because no mean is subtracted so the degrees of freedom are preserved.

The signal is then computed by projecting the analyte spectrum on the NAS regression vector and the SE can be computed as the ratio between the signal and the error:

$$\text{Signal} = \mathbf{s}_k^T \cdot \mathbf{nas}_{\text{reg}} \quad (14)$$

$$\text{SE} = \frac{\text{Signal}}{\text{error}} \quad (15)$$

This error indicator needs a collection of blank spectra to span the interference space and to quantify the error part plus the pure analyte spectrum. If the pure analyte spectrum is not available, a sample spectrum containing the analyte can be used.

Although the error indicator and the signal-to-error indicator seem to be comparable, there are some important differences. The EI minimizes the difference between the length of two vectors, $\mathbf{r}_{k,true}^*$ and $\mathbf{r}_k^*$. However, these vectors will not necessarily point in the same direction. Therefore, the difference in lengths is not directly related to errors in concentration. The SE indicator focuses on errors in the direction of the NAS regression vector, i.e., the same direction. The projections on the NAS regression vector are used (can also be negative) and not only the lengths of the projected vector. These projections are directly related to the concentrations (cf. Fig. 1).

Toolboxes for net analyte signal calibrations are available for free download at <http://www-its.chem.uva.nl/research/pac/index.html>.

EXPERIMENTAL

The powder samples were measured with a BOMEM MB 160 FT-NIR spectrometer equipped with a SpinningVialTM accessory for measuring powder samples, the samples were measured with diffuse reflectance, and an InGa detector was used. The SpinningVialTM accessory measured through the sample vials through the side of the glass vials (where the glass walls are assumed to be the most homogeneous). The wave number range from 3800 cm^{-1} to $12\,000\text{ cm}^{-1}$ was recorded and the spectral resolution was set to 8 cm^{-1} . For each spectrum a total of 32 scans were averaged (the scanning time for 32 scans measured with a spectral resolution of 8 cm^{-1} is the same time as the SpinningVialTM accessory uses to spin the sample vial one revolution). The tablet samples were measured with a BOMEM MB 160 FT-NIR spectrometer equipped with a TabletSamplIRTM accessory. The tablets were measured with a transmission measurement and an InGaAs detector was used. The wave number range from 4000 cm^{-1} to $12\,000\text{ cm}^{-1}$ was recorded and the spectral resolution was set to 16 cm^{-1} . When measuring transmission spectra of tablets, normally only broad peaks in the first and second overtone region are useful for quantification and 16 cm^{-1} is a reasonable resolution. For each spectrum a total of 32 scans were averaged. In both cases the data were collected with GRAMS32 (ThermoGalactic.com, GRAMS/32, 1998) software and imported into Matlab (MathWorks Inc., Matlab ver. 12.1., 2001) with in-house written software. Computations were performed in Matlab with in-house written routines plus the PLS-toolbox (Eigenvector Research, Inc. PLS-Toolbox. Version 2.1., 1998).

Dataset 1: Powder Samples. The samples were made according to a triangular mixture design. The samples contained five constituents, i.e., the active pharmaceutical ingredient (API), two filler binders (A and B), and two glidants (C and D). Three doses are normally produced, i.e., 0.64, 1.27, and 2.57 API w/w % (low, medium, and high strength). To have samples that resemble the heterogeneous nature of powder mixtures, samples with over- and under-dose of API, filler binder A, and filler binder B were produced according to a triangular mixture design (Fig. 2). Samples with $\pm 10\%$ of target dose of API and ± 10 and ± 20 w/w % of target dose of filler binder A and filler binder B, respectively, were made while the added amounts of glidant C and glidant D were kept constant. Initial experiments (not shown here) indicated that homogeneity of filler binder A and filler binder B could be difficult to obtain in a large-scale mixing process. It was therefore assumed that the span, i.e., $\pm 20\%$ from target concentration, of those constituents would resemble the heterogeneity that could be expected in the interference matrix, while glidant C and glidant D are assumed to be less important and for practical reasons the added amount was kept constant. Blank samples without API were also prepared (marked with squares in Fig. 2). The samples were prepared in 25 mL glass vials that fitted into the SpinningVialTM accessory. The total sample size was 8.0 g and the samples were prepared in the following manner. First the filler binder A was weighed with an electronic precision weight and transferred into the vial. Then API was weighed and transferred into the vial.

FIG. 2. Triangular mixture design for powder samples.

The constituents were mixed manually with a metal spatula. Then filler binder B, glidant C, and finally the glidant D were added, each time manually mixing with a metal spatula. Each sample was measured eight times in the SpinningVialTM accessory and between each measurement the sample was removed and shaken vigorously. The mean of the eight spectra was then used to represent the sample. The powder samples are generally problematic to measure because of the heterogeneous distribution of the sample constituents, but other studies (not shown) have shown that the SpinningVialTM accessory and the use of the mean spectrum is a valid methodology, and the methodology has also been reported elsewhere.¹⁹ As a reference method, the weighed amount was used (gravimetric) and the uncertainty of this value was believed to be low, i.e., $\pm 10^{-4}\text{ g}$.

Dataset 2: Tablet Samples. No specific experimental design was used for the tablet samples, but a small data set based on a stratified sampling scheme was used. Tablets were taken from nine different production batches (pilot scale batches): three batches with placebo tablets, i.e., blank samples without API, and six batches with API in three different levels. From each batch two tablets were used, for a total of 18 tablets. Because it is not possible to measure a transmission spectrum of the pure API (s_k), we used a spectrum of a tablet from a batch with high concentration of API as a replacement for the pure analyte spectrum. One tablet spectrum from each of the placebo batches, i.e., three spectra, were used to span the interference space and the three remaining spectra were used as blank samples to quantify the error.

RESULTS AND DISCUSSION

A selection of different preprocessing methods (Table I) that are normally^{20,21} applied when doing preprocessing of NIR spectra obtained from diffuse reflectance measurements of powders and transmission spectra of tablets were compared. For both data sets we compared the same preprocessing methods. The wavelength selection can be conducted in many different ways. In this article we used the prior knowledge we have about the analyte, i.e., lo-

TABLE I. Preprocessing methods.

No.	Method	Note
1	No preprocessing	
2	MSC	
3	Offset	Using 9990–10 000 cm^{-1} as offset point
4	First derivative	Using 11 spectral points
5	First derivative	Using 25 spectral points
6	Second derivative	Using 11 spectral points
7	Second derivative	Using 25 spectral points

cation of main analyte peaks. The search for the optimal wavelength interval was conducted by choosing a starting point, i.e., a wavenumber where an analyte peak is present, and then computing the various indicator values and RMSECV for a wavelength interval defined around this starting point. Then the interval was extended in both directions and new indicator values and RMSECV were computed. This was done a proper number of times using an increasing interval width until a large part of the wavelength axis was examined. The selection of wavelength intervals to examine can be made in numerous ways either using prior knowledge about major peak locations or more automatic routines, e.g., moving windows. In any case, the indicator values can be computed and therefore applied to existing wavelength selection methods.

For the powder samples the starting point was 6000 cm^{-1} , i.e., an analyte peak is found there (Fig. 3) with an interval width of 160 cm^{-1} , i.e., from 5920 to 6080 cm^{-1} . Then the interval was extended 160 cm^{-1} to 5840–6160 cm^{-1} . This was repeated until 20 intervals were examined, the last covering 4400–7600 cm^{-1} . For the tablet samples the starting point was 8800 cm^{-1} , i.e., an analyte peak is found there with an interval width of 120 cm^{-1} , i.e., from 8740 to 8860 cm^{-1} . Then the interval was extended 120 cm^{-1} to 8680–8920 cm^{-1} , and this was repeated until 15 intervals were examined, the last covering 7900–9700 cm^{-1} .

Comparing the Indicator Values and RMSECV Using GAIN Values. To compare the indicator values and the RMSECV we computed the GAIN for each value. The GAIN is computed as the ratio between an indicator or RMSECV value to a reference value. The reference value for the indicators or RMSECV is the value when using spectra without any preprocessing applied and using the whole wavelength range:

$$SE_{\text{gain}} = \frac{SE_{\text{pre}}}{SE_{\text{ref}}} \quad (16)$$

$$\text{invSEN}_{\text{gain}} = \frac{\text{invSEN}_{\text{ref}}}{\text{invSEN}_{\text{pre}}} \quad (17)$$

$$\text{EI}_{\text{gain}} = \frac{\text{EI}_{\text{ref}}}{\text{EI}_{\text{pre}}} \quad (18)$$

$$\text{RMSECV}_{\text{gain}} = \frac{\text{RMSECV}_{\text{ref}}}{\text{RMSECV}_{\text{pre}}} \quad (19)$$

where the subscript “ref” means that the indicator and RMSECV value are computed using non-preprocessed spectra and the whole wavelength range, i.e., 4000 to 10 000 cm^{-1} for the powder samples and 7300 to 10 000 cm^{-1} for the tablet samples. The subscript “pre” means

Fig. 3. NIR spectra of blank powder sample, powder sample with 2.57 w/w % API, and analyte spectrum.

that a preprocessing method or preprocessing method and wavelength interval selection have been applied to the spectra. Note that SE_{ref} is the denominator in Eq. 17. This is because the optimal preprocessing and wavelength selection is equal to the highest SE value opposite the other indicators and RMSECV where the lowest value equals the optimal preprocessing and wavelength selection. If the gain value is bigger than one, then the preprocessing or preprocessing and wavelength selection will improve the final calibration model, while if the gain value is equal to or lower than one then the preprocessing or preprocessing and wavelength selection are not improving or worsen the final calibration model.

Results for Powder Samples. In Fig. 3 the pure analyte spectrum, a spectrum of a blank sample, and a spectrum of a sample containing 2.57 w/w % API are depicted. The difference between the blank spectrum and the spectrum containing 2.57 w/w % API is mainly caused by scattering phenomena seen as offset differences from 7000 to 12 000 cm^{-1} . In the API spectrum main peaks are identified in the combinational band region, i.e., 4650 cm^{-1} and 4940 cm^{-1} , and in the first overtone region we find a peak at 6000 cm^{-1} , and in the second overtone region a peak at 8800 cm^{-1} is apparent.

Choosing the Optimal Preprocessing Method for Powder Samples. To span the interference space for the invSEN, SE, and EI indicator we used five blank sample spectra, symbolized with open squares in Fig. 2. To compute the invSEN we used two sample spectra containing the analyte, i.e., samples marked with grey circles in Fig. 2. To compute the SE we used two analyte spectra to compute the signal and an additional twenty-five blank sample spectra to compute the error. To compute the EI, two sample spectra, i.e., samples marked with grey color in Fig. 2, and two analyte spectra were used. The RMSECV values were calculated using the 32 samples depicted in Fig. 2. When computing the RMSECV values the 32 samples were divided into 11 blocks, i.e., 10 blocks with three samples each and one block with two samples, and then cross-validation was performed leaving out one block each time. Based on the cross-validation

FIG. 4. Optimal preprocessing method. Gain values for indicators and RMSECV for powder samples.

results, five PLS components were selected for the PLS model of the whole wavelength range.

The indicator values and the RMSECV were calculated using the 4000–10000 cm^{-1} wavelength region and by applying the preprocessing methods listed in Table I. In Fig. 4 the gain values are depicted for the indicators and RMSECV. The RMSECV shows that the best preprocessing method is first derivatives using 25 spectral points with a gain value of 2.9. The SE indicator has the highest gain for first derivatives, while the EI indicator has the highest gain for second derivatives. The invSEN indicator has the highest gain for MSC, which is clearly wrong compared to the PLS results.

Wavelength Selection for Powder Samples. Indicator and RMSECV values were computed for twenty wavelength intervals around 6000 cm^{-1} . For all intervals, four PLS components were used to calculate the RMSECV values. Again the number of PLS components is based on cross-validation results. This was done for all seven preprocessing methods and the highest gain value for the RMSECV was then found to be 5.95 when preprocessing method 5 was used with the wavelength interval from 5840–6160 cm^{-1} (Fig. 5). This matched perfectly the SE indicator that had the highest gain value for the same preprocessing method and wavelength interval as the PLS method. Also, the EI indicator had the highest gain value for preprocessing method 5, but using the wavelength interval from 5760–6240 cm^{-1} . The shape of the RMSECV gain curve corresponded well with the shape of the SE gain curve, and also the gain values were all above one for the RMSECV and the SE. The gain values for the EI when applying preprocessing method 5 were only above one for three intervals, i.e., $I - 2$, $I - 3$, and $I - 4$, while the remaining intervals were less than one, indicating that no preprocessing and using the whole wavelength region was better for those intervals (Fig. 5). The invSEN indicator was not useful for wavelength selection using any of the preprocessing methods. The highest gain value for the invSEN was 11.8 using MSC as the preprocessing method and the wavelength interval from 4000 to 10000 cm^{-1} , and when using all other preprocessing methods the gain for the invSEN was always below one, with the lowest value for the smallest wave-

FIG. 5. Optimal preprocessing/wavelength selection. Gain values for indicators and RMSECV for powder samples.

length interval, i.e., $I - 1$ and increasing with increasing interval width (see insert in Fig. 5).

It is important to notice that the selection of preprocessing method using the whole wavenumber range is not representative of the results when only a small wavelength region is used. Therefore, combining preprocessing and wavelength selection, as is done here, seems to be necessary.

Results for Tablet Samples. To span the interference space for the invSEN, SE, and EI indicators we used three blank sample spectra. To compute the invSEN we used two samples with a high concentration of API. To compute the SE we used two sample spectra, i.e., using two samples with a high concentration of API as substitution for pure analyte tablet spectra that were not available to compute the signal, and an additional three blank sample spectra to compute the error. To compute the EI four sample spectra with a high API concentration were used. Two of the sample spectra were used to compute the average $\mathbf{r}_k^*$ and the two other sample spectra were used to compute the average $\mathbf{s}_k^*$ (Eqs. 2 and 3) because no pure analyte tablet samples are available. The RMSECV values were calculated using all 18 samples. When the RMSECV value was computed the leave-one-out principle was used because of the limited size of the dataset.

Choosing the Optimal Preprocessing Method for Tablet Samples. Also for the tablet samples, comparison of the preprocessing methods using a broad spectral range was not a feasible method, i.e., preprocessing combined with wavelength selection was necessary.

Wavelength Selection for Tablet Samples. Indicator and RMSECV values were computed for fifteen wavelength intervals around 8800 cm^{-1} with all the preprocessing methods described in Table I. All PLS models were calculated using four PLS components. The highest gain value for the RMSECV was 3.6 when using preprocessing method 5, i.e., first derivatives with 25 spectral points and the wavelength interval 8620–8980 cm^{-1} (Fig. 6). Also, the SE had the maximum gain value of 3.8 using preprocessing method 5 and the interval 8620–8980 cm^{-1} (Fig. 6). The shape of the RMSECV and the SE gain curves were fairly similar. As for the powder samples,

FIG. 6. Optimal preprocessing/wavelength selection. Gain values for indicators and RMSECV for tablet samples.

the invSEN was not useful for wavelength selection and the gain values were less than one except for the MSC method. The EI had a maximum gain at 1.29 when MSC was used for preprocessing and the wavelength interval was 8320–9280 cm^{-1} (not depicted) and was in general not useful for wavelength selection of the tablet samples.

The problem with the invSEN indicator is that when the spectra are preprocessed using first and second derivatives the Euclidean length of the spectra and subsequently the net analyte signal vectors are lowered. This decreases the analyte sensitivity as computed in Eq. 4 without regard to the analytical performance of a calibration model using derivative spectra. In the original publication, Faber assumed that only white noise is present, which is a huge simplification of real spectroscopic systems in pharmaceutical applications. This might also explain why the method fails with our examples.

The EI indicator performed reasonably well but with failures. Wavelength selection of the tablet samples was not possible. The reason for the failure with the tablet samples might be that no “pure analyte tablet” was available. In the EI, the net analyte signal vector of a sample and analyte spectra are compared. But as pure analyte spectra are not always available and generally not for tablet samples the EI is not usable for this sample type.

The validation of the SE method is only performed on the zero concentration level. Therefore, it can be expected that the method will work better for low concentrations. During the work we discovered that a good selection of blank samples is the “key” to the SE indicator. For the powder samples we had measured each of the five blank samples eight times, giving forty blank spectra. Among these spectra we picked a few spectra to span the interference space and a larger portion to quantify the error. We recommend that as many blank samples as possible be measured using repeated measurements, and in that manner, instrumental noise and baseline drift are included. This is easy to do in most industrial applications, but might be more difficult for environmental products. Also reposition the samples and for powder samples, shake the samples. In that manner, heterogeneous samples are best measured.

A problematic issue for all NAS methods is that it is

unclear how interactions between the analyte and the interferences are dealt with. This is a general problem of the NAS approach, but even for more commonly used inverse calibration methods such as PLS or PCR this is not clear.

CONCLUSION

We have demonstrated a new indicator for choosing the optimal preprocessing method and conducting wavelength selection of NIR spectra. The indicator was compared to existing indicators also using net analyte signal computations and the standard methodology using cross-validation results from a PLS regression model. The indicator performed better than the two reference methods using net analyte signal methodology. The invSEN failed generally to find the optimal preprocessing method and was also not useful for wavelength selection. The EI indicator was developed for wavelength selection but we tried to use it for selection of optimal preprocessing method without success for both the powder and tablet samples. For wavelength selection the EI indicator performed reasonably for the powder samples and identified a few wavelength intervals that improved the calibration model, but not the optimal selection (Fig. 5). The indicator could not be used for wavelength selection of the tablet samples. The SE indicator identified the right preprocessing method and also the optimal wavelength selection both for the powder and the tablet samples. For the tablet samples the right preprocessing method was not obvious and was identified only after subsequent wavelength selection was performed (Fig. 6). Thus, in cases where only a few samples are available, reference values are determined with a high error, or are not available, we recommend this new indicator.

In this study the proposed method is only demonstrated for reflectance spectra of powder samples and transmittance spectra of whole tablets. More and different spectroscopic applications are necessary to corroborate the obtained results and to understand the limitations of this method. It might be the case that for different applications, the proposed indicator will not always be the best choice for selection of the optimal preprocessing and wavelength points.

ACKNOWLEDGMENT

Novo Nordisk, Corporate Research Affairs (CORA) sponsored this work as a part of E.T.S. Skibsted's Ph.D. project.

1. M. Blanco, H. Iturriaga, S. Maspocho, and C. Pezuela, *Analyst* (Cambridge, U.K.) **123**, 135 (1998).
2. M. Blanco, J. Coello, A. Eustaquio, H. Iturriaga, and S. Maspocho, *Anal. Chim. Acta* **392**, 237 (1999).
3. M. Blanco, J. Coello, H. Iturriaga, S. Maspocho, and D. Serrano, *Analyst* (Cambridge, U.K.) **123**, 2307 (1998).
4. R. D. Maesschalck, F. C. Sánchez, D. L. Massart, P. Doherty, and P. Hailey, *Appl. Spectrosc.* **52**, 725 (1998).
5. F. C. Sánchez, J. Toft, B. Bogaert, S. S. Dive, and P. Hailey, *Fresenius' J. Anal. Chem.* **352**, 771 (1995).
6. H. Martens and T. Næs, *Trends Anal. Chem.* **3**, 204 (1984).
7. H. Martens and T. Næs, *Multivariate Calibration* (John Wiley and Sons, Chichester, 1989).
8. H. Swierenga, A. P. de Weijer, R. J. van Wijk, and L. M. C. Buydens, *Chemom. Intell. Lab. Syst.* **49**, 1 (1999).
9. P. Geladi, D. McDougall, and H. Martens, *Appl. Spectrosc.* **39**, 491 (1985).

10. A. Savitzky and M. J. E. Golay, *Anal. Chem.* **36**, 1627 (1964).
11. Q. Ding and G. W. Small, *Anal. Chem.* **70**, 4472 (1998).
12. L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, *Appl. Spectrosc.* **54**, 413 (2000).
13. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, and G. L. Coté, *Anal. Chem.* **70**, 35 (1998).
14. A. Lorber, *Anal. Chem.* **58**, 1167 (1986).
15. N. M. Faber, *Anal. Chem.* **71**, 557 (1999).
16. L. Xu and I. Schechter, *Anal. Chem.* **68**, 1842 (1996).
17. H. F. Boelens, W. T. Kok, O. E. de Noord, and A. K. Smilde, internal report, available upon request.
18. J. Ferré and F. X. Rius, *Anal. Chem.* **70**, 1999 (1998).
19. O. Berntsson, L. G. Danielsson, M. O. Johansson, and S. Folestad, *Anal. Chim. Acta* **419**, 45 (2000).
20. O. Berntsson, Ph.D. Thesis, "Characterization and Application of Near Infrared Reflection Spectroscopy for Quantitative Process Analysis of Powder Mixtures", Kungliga Tekniska Högskolan (KTH), Stockholm, Sweden (2001).
21. H. Swierenga, Ph.D. Thesis, "Robust Multivariate Calibration Models in Vibrational Spectroscopic Applications", Katholieke Universiteit Nijmegen, The Netherlands (2000).