Performance Optimization of Spectroscopic Process Analyzers

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2004

Impact Factor: 5.64 · DOI: 10.1021/ac0353987 · Source: PubMed

4 AUTHORS, INCLUDING:

Wim Th Kok
University of Amsterdam
173 PUBLICATIONS **4,117** CITATIONS
SEE PROFILE

Onno E. Denoord
Shell Global
45 PUBLICATIONS **1,749** CITATIONS
SEE PROFILE

Anal. Chem. **2004,** *76,* 2656−2663

Performance Optimization of Spectroscopic Process Analyzers

Hans F. M. Boelens,[†] Wim Th. Kok,[†] Onno E. de Noord,[‡] and Age K. Smilde*[,†,§]

Process Analysis and Chemometrics, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands, Shell Research and Technology Centre, Shell International Chemicals B.V., Amsterdam, Badhuisweg 3, 1031 CM Amsterdam, The Netherlands, and TNO Nutrition and Food Research, Utrechtseweg 48, P.O. Box 360, 3700 AJ Zeist, The Netherlands

To increase the power and the robustness of spectroscopic process analyzers, methods are needed that suppress the spectral variation that is not related to the property of interest in the process stream. An approach for the selection of a suitable method is presented. The approach uses the net analyte signal (NAS) to analyze the situation and to select methods to suppress the nonrelevant spectral variation. The empirically determined signal-to-noise of the NAS is used as a figure of merit. The advantages of the approach are (i) that the error of the reference method does not affect method selection and (ii) that only a few spectral measurements are needed. A diagnostic plot is proposed that guides the user in the evaluation of the particular suppression method. As an example, NIR spectroscopic monitoring of a mol−sieve separation process is used.

In the process industry, monitoring the state of a process becomes increasingly important. This holds for both continuous processes as well as batch processes. The latter, for example, are difficult to control because of their inherent nonlinear nature. Good monitoring tools are, therefore, needed to increase the understanding of batch processes, to facilitate optimization, to improve batch-to-batch consistency and for safety reasons. Monitoring continuous processes is important, first, because it allows for checking controller performance and catalyst degradation and, second, to achieve consistent product quality. There is an increasing awareness that monitoring the *physical state* of the process (e.g., by using pressure, temperature, or viscosity measurements) is not sufficient; the *chemical state* of the process should also be measured.

Monitoring the chemical state of a process calls for process analytical measurements.[1] Different types of process analyzers are available for this type of measurements. Traditionally, chromatographic analyzers are used, but increasingly, analyzers based on spectroscopy, such as near-infrared (NIR), Raman, and UV−vis spectroscopy, are becoming popular.[2] The main advantage of spectroscopic process analyzers versus chromatographic process analyzers is their ability to measure in real time (on line or in situ).

To utilize the full power of spectroscopic methods, two problems should be solved. First, in some spectroscopic techniques (NIR and UV−vis), the spectra of chemical compounds in the reaction mixture are severely overlapping. This lack of instrumental selectivity requires the use of multivariate calibration models to extract the property of interest. Second, the acquired spectra will not only contain variation that is directly related to a compositional change of the reaction mixture, they can also contain unwanted variation caused by changing physicochemical properties of the mixture (e.g., viscosity, refractive index, temperature) or by the instruments (noise, baseline drift). Methods are needed to deal with the impact of this unwanted variation.

Different ways of dealing with unwanted spectra variation are possible. One way is to anticipate unwanted spectral variation and include spectra containing such variation in the calibration phase. This means that the multivariate calibration models are "told" that this is irrelevant variation and will recognize this type of variation in future samples. Stated another way, this type of unwanted spectral variation is treated as an unknown interference in the calibration phase, as is known to work well in first-order calibration.[3]

An alternative is preprocessing spectral data before making a multivariate calibration model, for example, by second-derivative Savitzky−Golay filtering.[4] It is well-known that the type of preprocessing affects the prediction errors of calibration models.[5] A practical problem is that many preprocessing methods exist and that the effect of a specific preprocessing method on the performance of the calibration model is a priori unclear. Moreover, some preprocessing methods contain metaparameters (e.g., filter widths in Savitzky−Golay filters) which have to be tuned. For this reason, preprocessing methods and metaparameters are selected by comparing prediction errors of various calibration models on a test set. This exhaustive search is laborious and not particularly insightful. Moreover, the measurement error of the reference

* To whom correspondence should be addressed. Phone: +31 20 525 5062. Fax: +31 20 525 5604. E-mail: asmilde@science.uva.nl.

[†] University of Amsterdam.

[‡] Shell International Chemicals B.V.

[§] TNO Nutrition and Food Research.

(1) Callis, J. B.; Illman, D. L.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 624A−37A.

(2) Chalmers, J. M. *Spectroscopy in Process Analysis*; Sheffield Academic Press Ltd.: Sheffield, 2000.

(3) Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 247−63.

(4) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−39.

(5) De Noord, O. E. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 65−70.

10.1021/ac0353987 CCC: $27.50
Published on Web 04/07/2004

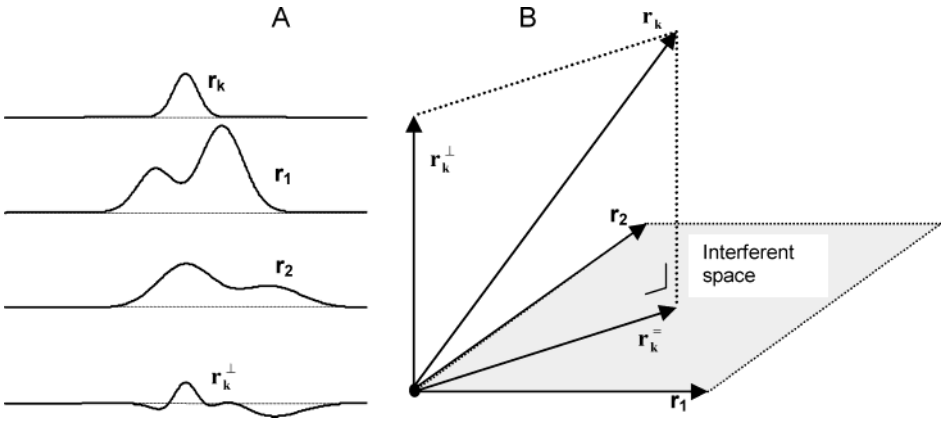


Figure 1. Left panel (A): a considerable band overlap exists between spectrum of analyte ($\mathbf{r}_k$) and spectra of interfering agents ($\mathbf{r}_1$ and $\mathbf{r}_2$). NAS vector ($\mathbf{r}_k^{\perp}$) is different from $\mathbf{r}_k$ in shape and magnitude. Right panel (B): representation in Jth dimensional vector space. NAS vector ($\mathbf{r}_k^{\perp}$) will be different from $\mathbf{r}_k$ in direction and length. The vector $\mathbf{r}_k^{=}$ is the part of $\mathbf{r}_k$ that is in the interference space.

method of the test sample may play a dominant role in the comparison.

The two alternatives are clearly distinct. Including variation as *unknown interference* assumes that this type of variation has the same shape in both the calibration and application phase (e.g., water vapor bands in gas-phase NIR). On the contrary, preprocessing tries to remove the type of unwanted spectral variation which may differ in shape between the calibration phase and application phase (e.g., baseline drift). Of course, combinations of both approaches are possible. In the following, the approach of including interferences will be called option A, and the preprocessing approach will be called option B.

A way of diagnosing multivariate calibration models is by means of the net analyte signal (NAS) concept.[6−9] The NAS is the part of the spectrum that is directly related to the analyte of interest. Some work on the use of NAS in the context of spectral preprocessing has already been described in the literature.[10−14] Faber[10] applied the NAS to get a better understanding of the impact of several preprocessing methods of NIR absorbance spectra. Brown et al.[11,12] used the signal-to-noise ratio of the NAS as a criterion to evaluate the performance of a multivariate calibration when the preprocessing filtering operation is changed. For spectra disturbed by white spectral noise, it was derived that preprocessing with low-pass filters does not improve the performance. Brown et al.[12] also discussed the application and the properties of Savitzky−Golay derivative filters for the reduction of drift noise in spectral data.

In this paper, a unified strategy is provided for optimizing the performance of spectroscopic process analyzers using options A and B and combinations thereof. The strategy does not perform an exhaustive search among calibration models, thereby avoiding the use of reference values of test samples with their intrinsic imprecision. The strategy uses the NAS approach to provide insight into the consequences of each step, which is visualized in the spectral domain. This enables the spectroscopist to check in an easy way the validity of the steps taken.

As an example, the monitoring of the break-through of normal alkanes in a mol−sieve separation process is used.[15] The focus of this application is on the improvement of the detection limit of the spectroscopic analysis. This will lead to earlier detection of the breakthrough and will allow a more efficient control of the mol−sieve process.

THEORY

The NAS Vector. The spectral absorbances that are measured at J wavenumbers are collected in a ($J \times 1$) column vector (r). The matrix $\mathbf{R}$ ($J \times I$) contains the spectra of I samples in its columns. Suppose that the analyte of interest (k) is a compound in a mixture of other spectroscopically active compounds that are called interferening agents. By definition, it is always possible to split up the absorbance spectrum (r_k) of the analyte of interest into two distinct parts ($r_k = \mathbf{r}_k^{=} + r_k^{\perp}$)

$$\mathbf{r}_k = \mathbf{r}_k^{=} + \mathbf{r}_k^{\perp} \tag{1}$$

where $\mathbf{r}_k^{=}$ is the part of the spectrum that could have been generated by a linear combination of the spectra of the interfering agents. The superscript "=" indicates that $\mathbf{r}_k^{=}$ is in the space spanned by the spectra of the interfering agent (Figure 1).

Consequently, $\mathbf{r}_k^{=}$ cannot be unique for the analyte of interest, because a mixture of interfering agents could have produced it. The other part, $\mathbf{r}_k^{\perp}$, is orthogonal to the spectra of the interferences reflecting the part of the absorbance spectrum only depending on the analyte k present in the mixture. This part, called the net analyte signal vector (NAS), can therefore be used for quantification of the analyte k.[6,7,16] The shape of $\mathbf{r}_k^{\perp}$ only depends on the *presence* of interferences in the mixture, not on the their specific concentrations. Only addition or deletion of spectroscopically active components can change $r_k^{\perp}$.

In the following, it as assumed that spectra of samples without the analyte are available. In such a case, the ($J \times J$) orthogonal

(6) Lorber, A. *Anal. Chem.* **1986**, *58*, 1167−72.
(7) Lorber, A.; Faber, K.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 1620−26.
(8) Ferre, J.; Rius, F. X. *Anal. Chem.* **1998**, *70*, 1999−2007.
(9) Faber, N. M. *Chemom. Intell. Lab. Syst.* **2002**, *50*, 107−14.
(10) Faber, N. M. *Anal. Chem.* **1999**, *71*, 557−65.
(11) Brown, C. D.; Wentzell, P. D. *J. Chemom.* **1999**, *13*, 133−52.
(12) Brown, C. D.; Vega-Montoto, L.; Wentzell, P. D. *Appl. Spectros.* **2000**, *54*, 1055−68.
(13) Olivieri, A. C. *J. Chemom.* **2002**, *16*, 207−17.
(14) Goicoechea, H. C.; Olivieri A. C. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 73−81.
(15) Boelens, H. F. M.; Kok, W. T.; De Noord, O. E.; Smilde, A. K. *Appl. Spectrosc.* **2000**, *54*, 406−12.

projection matrix $\mathbf{P}_k^{\perp}$ can be used to determine the NAS vector ($\mathbf{r}_k^{\perp}$) when the spectrum ($\mathbf{r}_k$) of the analyte is known

$$\mathbf{r}_k^{\perp} = \mathbf{P}_k^{\perp}\mathbf{r}_k \tag{2}$$

where $\mathbf{P}_k^{\perp}$ can be calculated using the matrix $\mathbf{R}_{-k}$ that contains the samples without the analyte k in its columns,[6]

$$\mathbf{P}_k^{\perp} = (\mathbf{I} - \mathbf{R}_{-k}\mathbf{R}_{-k}^{+}) \tag{3}$$

and the superscript "+" denotes the Moore−Penrose inverse.

The NAS View on Multivariate Calibration. Using the NAS approach, multivariate calibration can be seen as a two-step procedure. Step 1 is to find the direction in multivariate space that is unique for analyte k in the mixtures (the NAS vector). Step 2 is to calibrate the length of the NAS vector with the known concentration of the analyte in the calibration sample.

Suppose that a spectrum ($\mathbf{r}_{\mathrm{cal}}$) of a calibration sample with known concentration (c_{cal}) and interference spectra (i.e., spectra of samples not containing the analyte) are available. Step 1 then consists of calculating the NAS vector of $\mathbf{r}_{\mathrm{cal}}$, using eq 2.

$$\mathbf{r}_{\mathrm{cal}}^{\perp} = \mathbf{P}_k^{\perp}\mathbf{r}_{\mathrm{cal}} \tag{4}$$

Subsequently, this vector is normalized to length one.

$$\mathbf{r}_k^{\mathrm{nas}} = \frac{\mathbf{r}_{\mathrm{cal}}^{\perp}}{||\mathbf{r}_{\mathrm{cal}}^{\perp}||} \tag{5}$$

In step 2, the length along this NAS direction is calibrated with the known analyte concentration of the calibration sample. In the ideal case of Lambert−Beer conditions and (almost) error free samples of interfering agents, a one-point calibration will suffice. In practice, more analyte-containing samples can (and should) be used. The slope (s) of the resulting calibration line is equal to

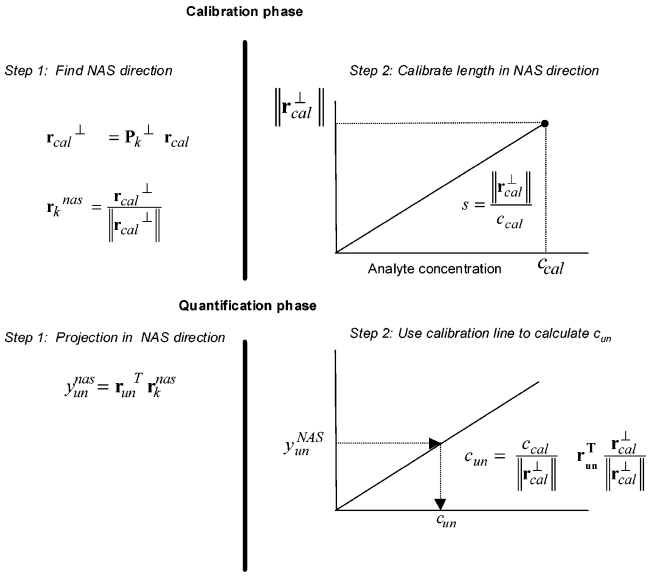
$$s = \frac{||\mathbf{r}_{\mathrm{cal}}^{\perp}||}{c_{\mathrm{cal}}} \tag{6}$$

In the quantification phase, the analyte concentration (c_{un}) is to be determined of an unknown mixture $\mathbf{r}_{\mathrm{un}}$, in which the concentration of the analyte is unknown, but the types of interfering agents that are present in this mixture are assumed to be the same as in the calibration phase. Step 1 now is a projection of $\mathbf{r}_{\mathrm{un}}$ on the NAS direction found in the calibration phase,

$$y_{\mathrm{un}}^{\mathrm{nas}} = \mathbf{r}_{\mathrm{un}}^{\mathrm{T}}\mathbf{r}_{\mathrm{k}}^{\mathrm{nas}} \tag{7}$$

in which $y_{\mathrm{un}}^{\mathrm{nas}}$ is a scalar measure along the NAS direction, and superscript T denotes the transpose operator. (Note that $y_{\mathrm{un}}^{\mathrm{nas}}$ might be negative when the concentration of the analyte is low or 0 (e.g., a blank sample) and $\mathbf{r}_{\mathrm{un}}$ is no longer assumed to be errorless.) Subsequently, the established calibration line is used

Scheme 1. Summary of NAS View on a Multivariate Calibration



to derive the analyte concentration in the unknown mixture.

$$c_{\mathrm{un}} = \frac{1}{s}y_{\mathrm{un}}^{\mathrm{nas}} \tag{8}$$

Summarizing,

$$c_{\mathrm{un}} = \frac{c_{\mathrm{cal}}}{||\mathbf{r}_{\mathrm{cal}}^{\perp}||}\mathbf{r}_{\mathrm{un}}^{\mathrm{T}}\frac{\mathbf{r}_{\mathrm{cal}}^{\perp}}{||\mathbf{r}_{\mathrm{cal}}^{\perp}||} \tag{9}$$

Scheme 1 summarizes the whole procedure.

Description of Spectral Disturbances. For simplicity, let it first be assumed that all calibration spectra are errorless, and a disturbance is present on the measured spectrum of the unknown sample. Furthermore, the spectral disturbance ($\mathbf{e}$) is assumed to be additive to the true spectrum of the unknown ($\mathbf{r}_{\mathrm{un}}$). Each time a new spectrum of the same unknown sample is measured, the spectral disturbance $\mathbf{e}$ will be different in size and possibly in shape. Clearly, the spectral disturbance ($\mathbf{e}$) is an example of irrelevant spectral variation, and the measured spectrum $\tilde{\mathbf{r}}_{\mathrm{un}}$ equals $\mathbf{r}_{\mathrm{un}} + \mathbf{e}$.

Substituting $\tilde{\mathbf{r}}_{\mathrm{un}} = \mathbf{r}_{\mathrm{un}} + \mathbf{e}$ into eq 7 leads to

$$y_{\mathrm{un}}^{\mathrm{nas}} = \tilde{\mathbf{r}}_{\mathrm{un}}^{\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} = \mathbf{r}_{\mathrm{un}}^{\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} + \mathbf{e}^{\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} \tag{10}$$

Both the true spectrum of the unknown sample ($\mathbf{r}_{\mathrm{un}}$) and the spectral disturbance ($\mathbf{e}$) can be split into a part that is orthogonal ($\perp$) to the interference space and a part that lies within ($=$) the interfence space. Moreover, the inner products of $\mathbf{r}_k^{\mathrm{nas}}$ and the parts ($\mathbf{r}_{\mathrm{un}}^{=}$, $\mathbf{e}^{=}$) that lie within the interference space are 0, because $\mathbf{r}_k^{\mathrm{nas}}$ is orthogonal to the interference space. Hence,

$$y_{\mathrm{un}}^{\mathrm{nas}} = \tilde{\mathbf{r}}_{\mathrm{un}}^{\perp\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} = \mathbf{r}_{\mathrm{un}}^{\perp\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} + \mathbf{e}^{\perp\mathrm{T}}\mathbf{r}_k^{\mathrm{nas}} \tag{11}$$

which shows that orthogonal projection of the measured spectrum

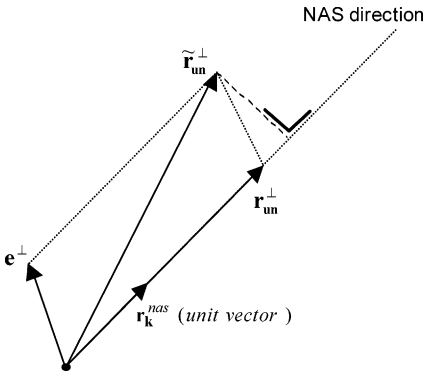


Figure 2. Situation in the space orthogonal to the interference space. The interference space is orthogonal to paper plane.

on the interference space, $\tilde{\mathbf{r}}_{un}^{\perp}$, is not exactly in the NAS direction, but that it is disturbed by the vector $\mathbf{e}^{\perp}$. This situation is depicted in Figure 2.

Using eq 8, the error Δc_{un} in the concentration of the analyte can be expressed as

$$\tilde{c}_{un} = c_{un} + \Delta c_{un} = \frac{1}{S} y_{un}^{nas}$$

$$c_{un} = \frac{1}{S} \mathbf{r}_{un}^{\perp T} \mathbf{r}_k^{nas}$$

$$\Delta c_{un} = \frac{1}{S} \mathbf{e}^{\perp T} \mathbf{r}_k^{nas} \qquad (12)$$

which shows that the concentration error is linearly related to the inner product of $\mathbf{e}^{\perp}$ and $\mathbf{r}_k^{nas}$. The size and shape of $\mathbf{e}^{\perp}$ in relation to a scaled vector $\mathbf{r}_k^{nas}$ can be plotted, and this diagnostic plot will yield the best visual information about the impact a spectral disturbance may have on quantification of the analyte. The reason that the comparison of the shape and size of $\mathbf{e}^{\perp}$ and $\mathbf{r}_k^{nas}$ is different from a comparison of shape and size of e and $\mathbf{r}_k$ in the original spectral domain is that orthogonal projection changes both the size and the shape of the true spectrum and spectral disturbance in a different way. This will be shown in the Results Section.

There are two ways to decrease the concentration error ($\mathbf{e}^{\perp T}\mathbf{r}_k^{nas}$): either $\mathbf{r}_k^{nas}$ is made more orthogonal to $\mathbf{e}^{\perp}$, or the norm of $\mathbf{e}^{\perp}$ is decreased. The first approach can be used when the shape of a specific spectral disturbance is repeatable. In such a case, a measured blank spectrum $\mathbf{e}$ may be used to additionally span the interference space, and a new NAS direction is calculated for the new interference space. The disturbance is, in fact, treated as an additional (instrumentally induced) interference (option A). The new NAS direction will then be more orthogonal to the systematic part of the spectral disturbances. Implicitly, this is done when PLS models are built that contain more factors than the number of chemical compounds in the mixture (assuming linearity and additivity).

Reducing the norm of $\mathbf{e}^{\perp}$ is most effective when the spectral disturbance shape varies from one measurement to another. This could be the case for stochastic spectral disturbances. The vector $\mathbf{e}^{\perp}$ will not have a fixed direction, and the only thing that can be done is to decrease its size by preprocessing the spectra (option B).

Both above-mentioned options have an effect on the direction and the size of the NAS vector. Therefore, the signal-to-noise ratio of the NAS should be used to make decisions about the most appropriate preprocessing method.

Signal-to-Noise Ratio of NAS. The key figure of merit for assessing (combinations of) options A and B is the signal-to-noise ratio of the NAS.[6,11]

$$(S/N)_{nas} = \left(\frac{||\mathbf{r}_{cal}^{\perp}||}{\sigma_0} \right) \qquad (13)$$

where σ_0 is the standard deviation in the signal space of measurements on blank samples projected onto the NAS direction. The spectrum $\mathbf{r}_{cal}$ is a spectrum of a calibration sample that contains the analyte. The $(S/N)_{nas}$ of a compound is closely related to the detection limit of that compound.[6,17,18] It is important to realize that this figure of merit is defined using the signal spectral domain only, and no concentration values enter the formula. Hence, the error of the reference method for establishing the concentration of the analyte in the sample does not play a role here. When comparing different situations, the ratio (ρ_x) of two signal-to-noise ratios can be used.

To use the $(S/N)_{nas}$ as a performance criterion, some spectral measurements are needed. First, spectra of samples that together contain all the interferences are needed to span the interference space. The exact concentrations of the interferences is not important, but (i) the concentration of the interferences should be within the concentration range of interest, and (ii) the analyte of interest should not be present in these mixtures. Second, to establish the NAS direction, a spectrum of a mixture should be measured that contains the analyte within the concentration range of interest. Third, to estimate σ_0, some blank measurements should be available. These measurements can be real instrumental blanks or they can be, for example, repeats of the samples that are used for spanning the interference space. Both types of blanks may cover a different type of variation; for example, the type of spectral variation could depend on the presence of interferences. For FT spectrometers, instrumental blanks are easily acquired when single beam spectra of the backgrounds are stored.

EXPERIMENTAL SECTION

Gas mixtures of low-C-number normal, iso- and cycloalkanes were collected with a FT-NIR spectrometer (BOMEM MB-155, Quebec, Canada) using a homemade gas cell (light path length, 2 cm). Simultaneously, the gas mixture was analyzed by GC as a reference method. The mole fraction error (standard deviation) of the GC method was 0.002% [mole/mole]. All alkane mixtures were prepared from the pure compounds (purity >98%) according to an experimental design. The compounds used are known to be present in the mol−sieve separation process under consideration. A mixture design was used for this. For the "corners" of the mixture design, alkane mixtures are also used. These mixtures are called pseudocomponents. The composition of these pseudocomponents (Table 1) is designed so that several requirements are satisfied: (1) Mixtures with an iso- and cycloalkane fraction that closely resembles the iso- and cycloalkane fraction of the output of the mol−sieve process are present. (2) Mixtures cover a wide range of a normal, iso-, and cycloalkane fraction. Pseudocom-

Table 1. Composition of Pseudo Compounds 1−5

	composition: mole fraction, %
PS1 (iso- and cycloalkanes)	20.8% 2-methyl pentane
	14.2% 3-methyl pentane
	3.7% cyclopentane
	26.8% methyl cyclopentane
	34.5% cyclohexane
PS2 (iso- and cycloalkanes)	91.7% 3-methyl pentane
	9.3% cyclopentane
PS3 (only isoalkanes)	24.4% 2-methyl pentane
	75.6% 2,2,4-trimethyl pentane
PS4 (only cycloalkanes)	3.7% cyclopentane
	4.4% methyl cyclopentane
	4.4% cyclohexane
	87.5% methyl cyclohexane
PS5 (only normal alkanes)	20.0% pentane
	60.0% hexane
	20.0% heptane

ponent 1 contains only isoalkanes and cycloalkanes, and its composition approximately matches the composition of the gas mixture flowing out of the reactor. Pseudocomponent 5 (PS5) contains only the normal alkanes (pentane, hexane, and heptane). Three levels are used for this pseudocomponent, namely, mole fractions of 1, 2, and 5%. At each level, 21 design mixtures are measured with varying amounts of the other four pseudocomponents (total of 63 measurements). Each of these design mixtures is measured only once. The pseudocomponent mixtures are all measured in triplicate (full repeat). More experimental details can be found elsewhere.[15]

The interferograms (30 scans) of the reference (N_2) and the gas mixtures were recorded with the software package WINBE Easy (BOMEM, version 3.01c, 1994). This software also converted the measured interferograms into absorbance spectra (wavenumber range, 4000−10 000 cm^{-1}; resolution, 4 cm^{-1}). The absorbance spectrum of a N_2 blank was calculated by using two single beam spectra of N_2 that were recorded at different points in time. Absorbance spectra and GC results were imported into MATLAB (MathWorks, Natick USA, version 6.1, 2001), in which all further data processing was performed. PLS models were made using the PLS toolbox (Eigenvector Technologies, West Richland, USA, version 2.1, 2000).

RESULTS AND DISCUSSION

Outline. Performance optimization using options A and B and combinations thereof have been systematically tested and treated. To illustrate option A, three interference spaces were defined. The smallest possible interference space, I_1, is spanned by the single samples of PS1−PS4. The interference spaces I_2 and I_3 also contain information regarding unwanted spectral variation. Adding a repeated measurement of PS3 to I_1 defines interference space I_2. Addition of N_2 blank spectra to I_1 yields interference space I_3. To illustrate option B, different customary types of preprocessing were performed, that is, offset correction and Savitzky−Golay derivative filters.

The strategy of using $(S/N)_{nas}$ for performance optimization was validated against the conventional method of comparing RMSEP values of different PLS models on independent test sets. The $(S/N)_{nas}$ values have been calculated using eq 13. Seven N_2 blanks are used to estimate σ_0. The calibration sets used in the PLS models were chosen in such a way that a comparison with the I_1, I_2, and I_3 results (option A) could be made.

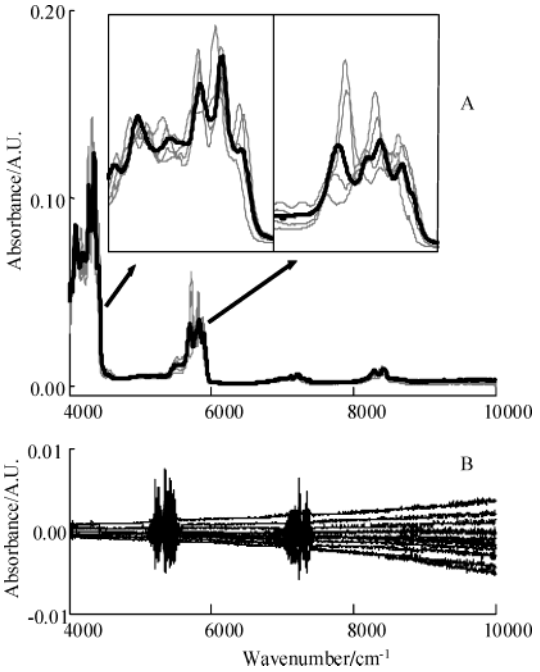


Figure 3. A: NIR absorbance spectrum (bold solid line) of compound PS5 that contains the normal alkanes and the spectra (thin solid lines) of other compounds PS1−PS4 that contain iso- and cycloalkanes. B: N_2 blank spectra.

Spectra and Unwanted Spectral Variation. The analyte of interest is pseudocomponent 5 (PS5) that contains only normal alkanes. The other pseudocomponents (PS1−PS4) that contain cyclo- and isoalkanes were taken as interferences. Figure 3A shows that the NIR absorbance spectrum of PS5 and the spectra of the interferences are severely overlapping (correlation higher than 0.97). In Figure 3B, the NIR absorbance spectrum of 10 N_2 blanks are shown.

These blank spectra (**e**) give an impression of the characteristic shape and intensity of the unwanted spectral variation for the spectrometer, the same equipment and experimental setup being used. The N_2 spectra contain a (nearly) white noise component, an offset component, and a driftlike component. In addition, narrow water vapor bands[19] with varying intensity appear in the wavenumber ranges 5100−5600 and 6900−7450 cm^{-1}. The intensity of the spectral disturbance is at least an order of magnitude smaller than the spectrum of PS5.

Option A, including Unwanted Variation as Interferences. Spectra without preprocessing were considered first. For each interference space, Figure 4 shows the following signals: the signal that corresponds to 20% [mole/mole] normal alkanes ($PS5_{20\%}$) in the respective NAS direction and the orthogonal projections ($\mathbf{e}^{\perp}$) of some N_2 blanks on the interference space.

These diagnostic plots allow a visual comparison of the unwanted spectral variation and the analyte signal. For all three spaces, the $PS5_{20\%}$ signals are nearly the same, whereas at the same time, the $\mathbf{e}^{\perp}$ signals drastically change from one space to

(16) Faber, N. M. *Anal. Chem.* **1998**, *70*, 5108−10.
(17) Faber, K.; Lorber, A.; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 419−61.
(18) Boque, R.; Rius, F. X. *J. Chemom.* **1996**, *11*, 419−61.
(19) Davies, A. M. C. *NIR News* **1992**, *3*, 8−9.

Table 2. NAS Approach and Various PLS Models

interference space	method	preprocessing: none[a]	offset [4830, 4860]	SG1[b]	SG2[c]
I_1 (PS1−PS4)	NAS	S/N = 2	S/N = 4	S/N = 421 w^d = 37	S/N = 784 w^d = 57
	PLS	RMSEP = 2.16 #LV[e] = 8	RMSEP = 1.48 #LV[e] = 7	RMSEP = 0.40 #LV[e] = 6 w^d = 41	RMSEP = 0.33 #LV[e] = 7 w^d = 75
I_2(PS1−PS4+repeat)	NAS	S/N = 13	S/N= 15	S/N = 923 w^d = 37	S/N = 1300 w^d = 61
	PLS	RMSEP = 1.01 #LV[e] = 8	RMSEP = 1.12 #LV[e] = 7	RMSEP = 0.23 #LV[e] = 5 w^d = 41	RMSEP = 0.19 #LV[e] = 5 w^d = 75
I_3 (PS1−PS4+blank)	NAS	S/N = 54	S/N= 132	S/N = 1325 w^d = 37	S/N = 864 w^d = 63
	PLS	RMSEP = 1.01 #LV[e] = 8	RMSEP = 0.98 #LV[e] = 8	RMSEP = 0.31 #LV[e] = 7 w^d = 41	RMSEP = 0.32 #LV[e] = 8 w^d = 63

[a] No preprocessing. [b] First derivative SG filter. [c] Second derivative SG filter. [d] Window width [points] of best SG filter. [e] Number of latent vectors in PLS model (determined by LOO crossvalidation).

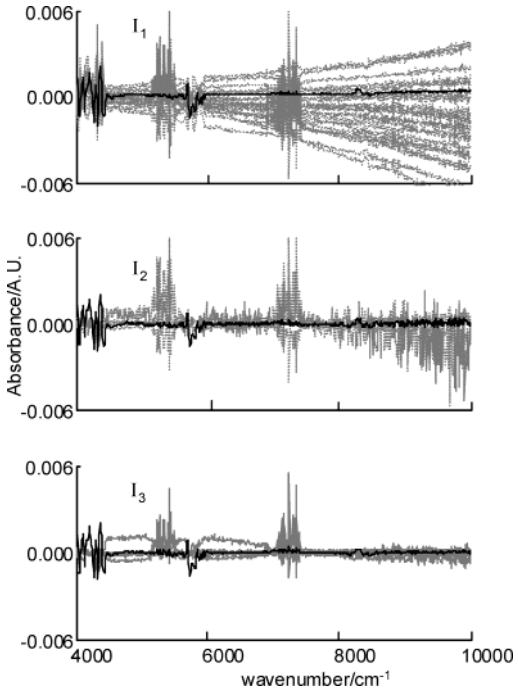


Figure 4. NAS vector of PS5 (bold solid line) corresponding to a 20% mole fraction of normal alkanes and orthogonal projections of N_2 blank spectra (dotted lines) on different interference spaces.

another. The general trend is that from space I_1 to I_3, the size of the $\mathbf{e}^{\perp}$ signals decreases and their shape changes considerably. For space I_1, the $PS5_{20\%}$ signal is completely blurred, especially in the low wavenumber range that is known to contain most information on the normal alkanes.[15] Use of space I_2 clearly results in suppression of the slow, driftlike features of the $\mathbf{e}^{\perp}$ signals. In this case, the $PS5_{20\%}$ signal of I_2 is larger than $\mathbf{e}^{\perp}$ signals in the low wavenumber region. For space I_3, the situation is improved further, and especially the size of the features originating from waterbands has decreased. The $(S/N)_{nas}$ obtained without preprocessing listed in Table 2 confirms this trend. The $(S/N)_{nas}$ for space I_1 is below 3, indicating that the $PS5_{20\%}$ signal is not detectable at all, whereas for spaces I_2 and I_3, the $(S/N)_{nas}$ values are well above 3.

Table 3. $(S/N)_{nas}$ Gain (ρ_x) with Respect to Raw Spectra

interference space	ρ_x for SG1	ρ_x for SG2
$\mathbf{I_1}$ (PS1−PS4)	210	392
$\mathbf{I_2}$ (PS1−PS4 + repeat)	71	100
$\mathbf{I_3}$ (PS1−PS4 + blank)	25	16

Option B, Preprocessing Spectra. The effect of preprocessing the spectra with offset correction, first and second derivative SG filters, is assessed here. The window size of the SG filters was varied between 5 and 81 points. The degree of the interpolation polynomial was selected to be one larger than the order of the derivative. The $(S/N)_{nas}$ was calculated for each window width. The window width (w) having the highest $(S/N)_{nas}$ is reported in Table 2.

When comparing the results with raw spectra (column, "none"), it can be seen that offset correction or preprocessing the spectra with a first derivative SG filter (column "SG_1") leads to higher $(S/N)_{nas}$ values. The beneficial effect of using SG1 filters, however, is much stronger. Preprocessing the spectra with a second derivative filter (column "SG_2") gives only a small additional gain or even a slight decrease (space I_3), as compared to using first derivative filters. The table also shows that the best window width (w) for the SG_2 filter is always higher than the best window width for the SG_1 filter. This can be explained by the fact that calculation of a second derivative leads to a relative increase of (fast) spectral noise. Compensation of this effect can be achieved by increasing the filter width and thereby increasing the smoothing effect of the SG_2 filter. This trend can also easily be seen in diagnostic plots. In Figure 5, the effect of the different preprocessing methods is shown for interference space I_1. It can be seen that offset correction mainly leads to reduction of the amount of unwanted spectral variation at the low wavenumbers, while the wavenumber region around 5800 cm^{-1} is still affected by the spectral variation. SG1 preprocessing reduces the overall amount of variation, but some remaining variation is located at the water band regions. SG2 preprocessing does not help to decrease this particular variation. Using interference spaces I_2 or I_3 helps to further decrease the size of unwanted spectral variation

Table 4. Composition of Calibration and Test Set of Various PLS Models

PLS model	calibration set	test set
similar to I_1	12 mixtures. two compound mixtures: PS1 (99%) and PS5 (1%) PS1 (98%) and PS5 (2%) PS1 (95%) and PS5 (5%) PS2 (99%) and PS5 (1%) PS4 (95%) and PS5 (5%)	remaining 51 mixtures
similar to I_2	same 12 mixtures + single samples of PS1 to PS4	idem
similar to I_3	same 12 mixtures + 4 N_2 blanks	idem

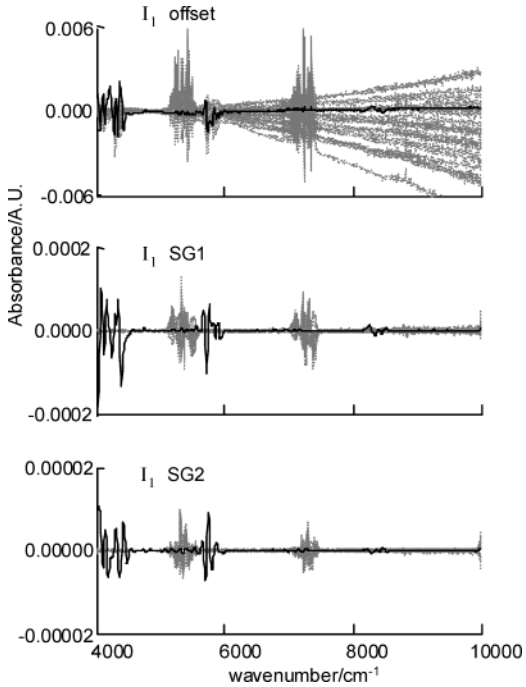


Figure 5. NAS vector of PS5 (bold solid line) corresponding to a 20% mole fraction of normal alkanes and orthogonal projections of N_2 blank spectra (dotted lines) for various preprocessing methods (interference space I_1).

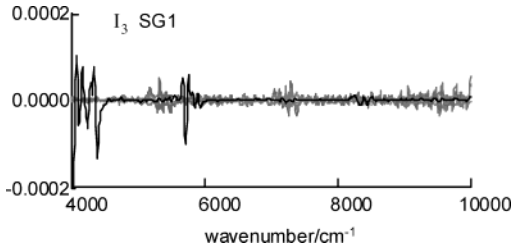


Figure 6. NAS vector of PS5 (bold solid line) corresponding to a 20% mole fraction of normal alkanes and orthogonal projections of N_2 blank spectra (dotted lines) for SG1 preprocessing and interference space 3.

(Figure 6).Table 3 summarizes the $(S/N)_{nas}$ gains (ρ_x) obtained by the SG_1 and SG_2 methods with respect to using no preprocessing. The gain decreases when going from space I_1 to I_3. This is explained by the fact that interference spaces I_2 and I_3 model some of the systematic features of the unwanted spectral variation. With preprocessing, these driftlike features are also suppressed. This illustrates the difference between implicitly modeling these features (option A) in the interference space and removing them before making a calibration model (option B). Option B proved to be more effective in this particular case (see Table 2 and compare Figures 4 and 5).

Comparison with PLS Models. PLS models have been made to allow for a comparison with the NAS approach. The full wavenumber range of the NIR spectra was used. The calibration and test sets of the three PLS models resemble the cases I_1, I_2, and I_3 (see Table 4).

For each possible preprocessing method, a PLS model was made. For the SG preprocessing methods, this means that for each window width, a model was made. The number of latent factors needed in each model was determined using a leave-one-out crossvalidation (RMSECV). For SG preprocessing, the window width of the best PLS model (looking at RMSECV) is reported.

The results of the NAS approach and PLS models are in this case expected to correspond, because the error of the reference method is relatively small (0.002 mol/mol). The prediction error (RMSEP) and the number of latent factors (#LV) of these models are listed in Table 2. A lower RMSEP in general corresponds with a higher $(S/N)_{nas}$. For example, preprocessing of the spectra by offset subtraction or by the use of a SG_1 filter leads to lower RMSEP and to higher $(S/N)_{nas}$ values. The $(S/N)_{nas}$ values show that the SG_2 preprocessing has only a small advantage, as compared to SG_1. This is also concluded from the RMSEP values: only for the I_1 and I_2 models is a slightly smaller RMSEP found. Considering the selection of the best preprocessing method in detail, it is shown that the best window width setting of the filters are similar for the NAS and PLS approach. For SG_1 preprocessing, the width is ~40 points, and for SG_2 preprocessing, it is between ~60 and 75 points. The lowest RMSEP is found for the combination of PLS model I_2 and spectra preprocessed with a SG_2 filter ($w = 75$). The same combination also ranks high using the $(S/N)_{nas}$ approach.

Incorporating more knowledge about unwanted spectral information in the calibration set in general leads to an improvement by a factor 1.2−2 in RMSEP (compare, e.g., model I_2 with model I_1). In the same situation, the $(S/N)_{nas}$ shows more improvement (factor 1.6−7). Comparing results for models I_2 and I_3, it appears that NAS is somewhat more optimistic than PLS about the improvement that can be achieved. These differences between NAS and PLS can be explained by the fact that the PLS models all have a large number of spectra in their calibration set. In this way, some unwanted spectral variation will already be modeled, and the differences between models will become smaller.

The tradeoff between modeling and preprocessing for the PLS models is reflected by the number of factors needed. In general, this number decreases when a better preprocessing method is selected. Without preprocessing, additional factors in the PLS are needed to describe the unwanted spectral variation.

CONCLUSIONS

It is possible to select a suitable spectral preprocessing method using the empirically determined $(S/N)_{nas}$ as criterion. The advantages of this method are that (i) the error of the reference method does not influence the results and (ii) only a few spectral measurements are needed. Furthermore, diagnostic plots of the analyte and error signals orthogonal to the selected interference space supply insight into the effect of preprocessing (option B) and of including unwanted information in the interference space (option A).

For the application at hand, the selection of the preprocessing method is essential for detection of small amounts (<1% mol/mol) of normal alkanes. It was shown that the proposed NAS procedure selects a preprocessing method similar to that of the conventional PLS procedure with considerably less effort.

Received for review November 26, 2003. Accepted March 2, 2004.

AC0353987