Journal of Pharmaceutical & Biomedical Analysis
Vol. 9, No. 8, pp. 625–635, 1991
Printed in Great Britain

0731–7085/91 $3.00 + 0.00

Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy*

HARALD MARTENS†‡ and EDWARD STARK§

‡*Consensus Analysis AS, Ski Business Park, N-1400 Ski, Norway*
§*KES Analysis Inc, 3M, 160 Westend Avenue, New York, NY 10012, USA*

Abstract: Near infrared (NIR) spectroscopy spectra can be converted mathematically to precise quantitative information of chemical and physical nature by multivariate calibration. This makes NIR analysis useful for a variety of "difficult" sample types (powders, slurries), more or less without any sample preparation.

The paper emphasizes the importance of using prior knowledge for spectral preprocessing of spectral data prior to the linear multivariate calibration modelling. Two new preprocessing methods are presented: extended multiplicative signal correction (EMSC) for elimination of uncontrollable path length or scattering effects, and spectral interference subtraction (SIS) for elimination of known spectral interferences.

Determination of toluene in mixtures with benzene and xylene from NIR spectra with gross simulated light scattering effects is used for illustration.

Keywords: *Near infrared (NIR) spectroscopy; multivariate calibration; chemometrics; spectral preprocessing; extended multiplicative signal correction (EMSC); spectral interference subtraction (SIS).*

Introduction

Knowledge-driven versus data-driven modelling in analytical chemistry

Results from chemical analysis must be selective. Traditionally, selectivity problems in terms of chemical or physical interferences had to be removed by filtering or standardization of the samples prior to measurement. With multichannel instruments and multivariate calibration software the selectivity enhancement can instead be done mathematically.

For effective multivariate calibration modelling it is important to combine *a priori* assumptions ("prior knowledge") and empirical data in a balanced way.

Classical analytical chemistry has focused very much on prior assumed knowledge. "Hard modelling" (mathematical modelling based on explicit causal and statistical assumptions) is used for extracting information from data. An example of this is the modelling of spectroscopic measurements from chemical mixtures. These are often modelled as linear combinations of known pure constituents. Such modelling can only be used in systems where this assumed knowledge is adequate, e.g. where there are no unidentified constituents, constituent interactions, temperature effects, light scattering variations, etc. This knowledge-driven modelling has led to a tendency of academic over-simplification (simple transparent solutions of a few well known constituents), leaving many practical analytical problems unsolved in the fields of food and agriculture, biology, process industry and pharmacy.

On the other hand, in analytical chemistry based on quantitative chemometric modelling, empirical measurements are used instead of causal assumptions. "Soft modelling" (mathematical modelling with as few statistical and causal assumptions as possible) is used for information extraction. Background knowledge is only used mentally, for design of experiments to obtain the empirical data, as well as in the graphical interpretation of the results. But knowledge about, for example, pure constituents is not used as an integral part of the mathematical modelling of the mixture data obtained. This data-driven modelling allows

*Presented at the Symposium on "Chemometrics in Pharmaceutical and Biomedical Analysis", November 1990, Stockholm, Sweden.

†Author to whom correspondence should be addressed.

reliable quantitative analysis in systems where classical analysis has failed. But it makes the calibration process unnecessarily costly, by requiring the empirical data to carry all the information required to make the necessary calibration models.

The present paper assumes a flexible intermediate between hard modelling and soft modelling: *a priori* knowledge is applied in "hard" modelling during preprocessing, to simplify the structure in the spectral data. "Soft" modelling is then used for cleaning up empirically what the causal modelling could not explain. The goal is to obtain maximal understanding and maximal predictive reliability and relevance at minimum experimental and data analytic costs.

The paper presents two new spectral preprocessing methods for improving the multivariate calibration of multichannel analytical instruments based on spectroscopic background knowledge: extended multiplicative signal correction (EMSC) is designed to improve the separation of light scattering and light absorbance, and spectral interference subtraction (SIS) for elimination of interferences with known spectral effects. Conventional projection on latent structures regression (PLSR) is then used for the subsequent empirical "soft modelling" calibration. Near infrared (NIR) data are used for illustrating their application.

Near infrared spectroscopy

NIR spectroscopy [1], operating within the wavelength range 900–2600 nm, is a relatively new analytical technique with a high potential for pharmaceutical and biomedical analysis. With little or no sample preparation it can provide precise chemical and physical characterization of a variety of "difficult" sample types — powders and intact tablets, intact biological tissue, slurries, suspensions and emulsions as well as turbid or clear solutions — and even gases. The method even has interesting potentials for *in-vivo* applications.

NIR spectroscopy can also be used effectively for high-speed qualitative control purposes, to check that raw materials, processes or products have spectra within systematic ranges corresponding to set quality specifications (multivariate control charting).

Most organic molecules as well as water and many inorganic compounds display useful NIR absorbance patterns. These absorbances are overtones and combination bands from fundamental molecular vibration bands in the IR region. The IR absorbances themselves are often too strong to allow simple, representative analysis of complex samples. But the NIR bands are sufficiently weakened to allow the light to penetrate anywhere from a few millimetres to a couple of centimetres through the samples. The NIR measurements can be taken as reflectance or transmittance, depending on what is most practical.

The development of the highly successful NIR technology has been application driven [1], rather than theory driven. NIR analysis relies on empirical statistical estimation of the mathematical transfer functions required to convert spectral measurements into chemical information. Thus it violates the conventional, misleading (and often subconscious?) academic desire for one-to-one correspondence between data and fundamental information.

To wring highly selective and precise results from highly non-selective and confusing measurement may for traditional analytical chemists seem like unreliable "black magic" — almost like cheating. To make things worse, the prime fields of NIR application till now, food and feeds analysis, may have had a rather low standing on the academic status ladder. Finally, since NIR instruments require both multivariate calibration chemometrics and spectroscopic insight, this multidisciplinary technique may fall between the traditionally specialized academic chairs.

However, the NIR technique is now well understood theoretically, and has proven to work well in many practical applications where other analytical methods like IR or UV spectroscopy fail. While it is very popular in industry, it is still largely ignored in many universities.

The purpose of the present paper is to illustrate that there is solid spectroscopic rationale behind the multivariate calibration techniques in NIR spectroscopy. This will be done by demonstrating that apparently confusing and wildly non-selective data can be made selective by either applying background knowledge through preprocessing, or by PLSR, or (preferably) by a combination of these.

A somewhat extreme example is presently used, in order to illustrate two aspects: (1) the power of multichannel NIR and quantitative chemometrics, compared with traditional univariate calibration; and (2) the importance of spectral preprocessing.

The example concerns calibration for one constituent in mixtures of two other constituents with very similar NIR spectra. The NIR spectra display gross uncontrolled path length and baseline variations, like the ones found when the turbidity or path length changes dramatically. Toluene, whose NIR spectrum is similar to a mixture of benzene and xylene, is considered the analyte to be determined from NIR spectra, in future unknown samples containing unknown levels of the two interferents benzene and xylene. In this case we know the spectra of the three pure solvents. So under ideal conditions one could resolve each future mixture's absorbance spectrum into these three constituent spectra. But the constituents do not have quite the same spectra in mixture as in pure form; various interactions may be expected. Likewise, the present data set has been modified to include major baseline and path length effects, to simulate major uncontrolled light scattering variations and/or path length variations. The knowledge about the three constituents' pure spectra will be used for preprocessing.

Calibration theory

The calibration theory for NIR instruments and other multichannel non-selective chemometric sensors is described in detail, e.g. by Martens and Naes [2].

The goal of the calibration of NIR instruments is to find the transfer function $f()$ that allows us to convert a multichannel input spectrum $\mathbf{z}_i = \{z_{ik}, k = 1,2,\ldots,K\}$ (say, K = 100 NIR wavelengths of a pharmaceutical sample i), into sample quality y_i (say, content of a certain chemical constituent):

$$y_i = f(\mathbf{z}_i). \qquad (1)$$

NIR instruments may be calibrated by statistical regression based on purely empirical data from a representative training set of samples. PLSR [3] is one popular method for this purpose. These calibration data consist of reasonably precisely known data for both chemometric sensor $\mathbf{z}_i$ and reference method y_i from a representative set of training samples (objects) $(y_i, \mathbf{z}_i), i = 1,2,\ldots,N$.

Preprocessing

If background knowledge about the nature of the relationship between spectral data $\mathbf{z}_i$ and chemical data y_i is available, it is advisable to apply this during preprocessing, for instance for converting the measurements $\mathbf{z}_i$ into corrected spectra $\mathbf{x}_i$:

$$\mathbf{x}_i = g(\mathbf{z}_i). \qquad (2)$$

The purpose is to simplify the subsequent statistical regression calibration to estimate the predictor function:

$$y_i = b(\mathbf{x}_i) = b[g(\mathbf{z}_i)] = f(\mathbf{z}_i). \qquad (3)$$

The goal of the preprocessing is to reduce the need for calibration data, to improve the statistical precision of the predictor function $b()$, and to simplify the spectroscopic interpretation of this function and its underlying calibration model.

The preprocessing function $g()$ can involve a number of different stages and types of transformations. Rather than "black box" approaches such as neural net or optimal scaling, techniques with distinct spectroscopic interpretation are here employed.

Response linearization

If general knowledge exists about the mathematical shape of the relationship between measurements $\mathbf{z}_i$ and qualities y_i, one may apply this knowledge in the preprocessing stage in order to simplify the final modelling. Changing the measured transmittances T to optical density, OD = $\log(1/T)$, is here used as an example of a useful (although not perfect) instrument response linearization.

Extended multiplicative signal correction

While many types of spectroscopic selectivity problems are of an additive nature (e.g. absorbance effects of chemical interferents), others have a strong multiplicative component. Examples of the latter are light scattering variation or optical path length variations. If these vary uncontrollably from sample to sample, it is advantageous to reduce their effect in the preprocessing stage. Otherwise their multiplicative nature may otherwise destroy the subsequent additive PLSR calibration modelling.

If the physical and chemical effects in the spectra are sufficiently different, they may be separated by multivariate statistical modelling. One method is the MSC technique [4–6]. This is now termed "multiplicative signal correction" [2, chap. 7], as a generalization of the original term "multiplicative scatter correction" [4], since it is also applicable to other types of data, e.g. correcting for varying amounts of sample applied to a chromatography column. MSC seeks to correct the baseline and amplification effects to the same "average" level in every spectrum.

As outlined in ref. 7 (p. 350) Stark and Martens in 1989 developed MSC into the extended multiplicative signal correction (EMSC) in order to attain a more effective separation of chemical and physical effects in light spectroscopy. The EMSC method employs knowledge about the spectra of the analytes and interference effects to improve the path length estimation. The method is described in detail under *Mathematical method description*.

Spectral interference subtraction

When spectral information about the analyte and interferents is available, it is possible to reduce or eliminate the spectral effects of the interferents in the preprocessing stage. This further reduces the need for empirical calibration data. The technique presented here is called spectral interference subtraction (SIS) and was developed by the authors in conjunction with the development of the EMSC. Its purpose is to filter out the effects of known constituents from the spectral data, with as little modification as possible of the effects the unknown constituents and phenomena. The method is described in detail under *Mathematical method description.*

Mathematical method description

Extended multiplicative signal correction. Let $\mathbf{Z} = \{\mathbf{z}_i = 1,2,\ldots,N\}$ be the measured spectra of a set of N samples. These data are to be corrected into spectra $\mathbf{X} = \{\mathbf{x}_i, i = 1,2,\ldots,N\}$ by EMSC. (All vectors are assumed to be column vector.)

The spectral model used here is:

$$\mathbf{z}_i = \mathbf{x}_i b_i + \mathbf{1}a_i + \mathbf{e}_i, \tag{4}$$

where $\mathbf{1}$ is vector $(1,1,1,\ldots,1)'$, $\mathbf{e}_i$ is the residual in the model, and where a_i represents an unknown additive effect (e.g. baseline offset) and b_i represents an unknown multiplicative effect (e.g. optical path length or light scattering level).

This basic EMSC signal model may be extended in different ways, to include, for example, wavelength dependencies for a_i or b_i. But the basic EMSC model is presently used, as a local simplified approximation to various more complicated nonlinear models. The unknown parameters a_i and b_i are now to be estimated from the data.

Under ideal conditions (Beer's model), the absorbance data $\mathbf{x}_i$ can be seen as a sum of the contributions from the different chemical constituents (analyte and interferents) with spectra $\mathbf{K} = \{\mathbf{k}_j, j = 1,2,\ldots,J\}$ and concentrations $\mathbf{c}_i = \{c_{ij}, j = 1,2,\ldots,J\}'$:

$$\mathbf{x}_i = \mathbf{k}_1 c_{i1} + \mathbf{k}_2 c_{i2} + \ldots + \mathbf{k}_J c_{iJ} = \mathbf{K}\mathbf{c}_i. \tag{5}$$

Spectrum (6) $\mathbf{x}_i$ may be rewritten as a deviation from a reference spectrum — $\mathbf{z}_0$, which could be, for example, the average of a set of empirical spectra:

$$\mathbf{x}_i = \mathbf{z}_0 + \mathbf{K}\mathbf{d}_i. \tag{6}$$

This deviation $\mathbf{d}_i$ represents, in the case of absorbance data, the deviations in the analyte and interference concentration compared with that of the reference sample: $\mathbf{d}_i = \mathbf{c}_i - \mathbf{1}c_0$.

This yields

$$\mathbf{z}_i = (\mathbf{z}_0 + \mathbf{K}\mathbf{d}_i)b_i + \mathbf{1}a_i + \mathbf{e}_i, \tag{7}$$

which can be rewritten

$$\mathbf{z}_i = \mathbf{z}_0 b_i + \mathbf{K}\mathbf{d}_i b_i + \mathbf{1}a_i + \mathbf{e}_i. \tag{8}$$

Reference spectrum $\mathbf{z}_0$ is usually chosen as the average of a set of spectra $\mathbf{z}_i, i = 1,2,\ldots,N$. Some method (MSC or EMSC) is used for estimating a_i and b_i.

Once estimated, the actual signal correction for converting the measured spectra $\mathbf{z}_i$ into corrected estimates of the unknown desired spectra $\mathbf{x}_i$ [equation (2)] is:

$$\mathbf{x}_i = (\mathbf{z}_i - \mathbf{1}a_i)/b_i. \tag{9}$$

Conventional multiplicative signal correction. In conventional MSC the estimation of a_i and b_i is done by simply ignoring the term $\mathbf{K}\mathbf{d}_i b_i$ in equation (8):

$$\mathbf{z}_i = \mathbf{z}_0 b_i + \mathbf{1}a_i + \mathbf{e}_i. \tag{10}$$

The least-squares solution is then:

$$[b_i,a_i] = ([\mathbf{z}_0\ \mathbf{1}]'[\mathbf{z}_0\ \mathbf{1}])^{-1}[\mathbf{z}_0\ \mathbf{1}]'\mathbf{z}_i. \quad (11)$$

To avoid wavelength regions where the chemical absorbance variations $\mathbf{d}_i$ might influence the estimation of a_i and b_i, weighted least-squares (WLS) estimation is used in practice:

$$[b_i,a_i] = ([\mathbf{z}_0\ \mathbf{1}]'\mathbf{V}[\mathbf{z}_0\ \mathbf{1}])^{-1}[\mathbf{z}_0\ \mathbf{1}]'\mathbf{V}\mathbf{z}_i, \quad (12)$$

where $\mathbf{V}$ is a diagonal matrix with a weight factor for each wavelength. For instance, $\mathbf{V}$ could have elements $v_{kk} = 1$ for wavelengths to be used, and $v_{kk} = 0$ for wavelengths not used.

This MSC pre-treatment can greatly simplify the subsequent calibration modelling. However, if different chemical constituents in the samples have very different absorbance levels, additive variations in the spectra due to composition variations are mistakenly treated as if they were due to multiplicative effects. In the present example, the interferent benzene has considerably higher absorbance than analyte toluene and interferent xylene in the relevant wavelength range. Variations in benzene–(toluene + xylene) concentration ratios will in MSC be taken as path length variations and removed by division. This will introduce errors in the subsequent calibration modelling.

Extended multiplicative signal correction. EMSC is designed to allow explicit compensation for the chemical variabilities by including information about the major analyte and interferent spectra in the estimation of a_i and b_i. In equation (8) the term $\mathbf{d}_ib_i$ may be termed $\mathbf{h}_i$. The spectral model equation (8) rewritten as

$$\mathbf{z}_i = \mathbf{z}_0b_i + \mathbf{K}\mathbf{h}_i + \mathbf{1}a_i + \mathbf{e}_i. \quad (13)$$

Ideally, the WLS solution should then be attained by including $\mathbf{K}$ into equation (12):

$$[b_i,\mathbf{h}_i,a_i] = ([\mathbf{z}_0\ \mathbf{K}\ \mathbf{1}]'\mathbf{V}[\mathbf{z}_0\ \mathbf{K}\ \mathbf{1}])^{-1}[\mathbf{z}_0\ \mathbf{K}\ \mathbf{1}]'\mathbf{V}\mathbf{z}_i. \quad (14)$$

However, if the mixture modelling is reasonably complete (data on all major constituents' spectra are sufficiently reliable and included in the model), $\mathbf{z}_0$ will be more or less linearly dependent on $\mathbf{K}$ and $\mathbf{1}$. The matrix inversion in equation (14) will then be more or less unstable. Therefore it is necessary to replace the spectral model (13) by an expression that spans the same variability. Expression $\mathbf{K}\mathbf{h}_i$ in equations (13) and (14) is replaced by some other expression $\mathbf{P}\mathbf{t}_i$:

$$\mathbf{z}_i = \mathbf{z}_0b_i + \mathbf{P}\mathbf{t}_i + \mathbf{1}a_i + \mathbf{e}_i \quad (15)$$

$$[b_i,\mathbf{t}_i,a_i] = ([\mathbf{z}_0\ \mathbf{P}\ \mathbf{1}]'\mathbf{V}[\mathbf{z}_0\ \mathbf{P}\ \mathbf{1}])^{-1}[\mathbf{z}_0\ \mathbf{P}\ \mathbf{1}]'\mathbf{V}\mathbf{z}_i. \quad (16)$$

One such approach is to let $\mathbf{P}$ be the $J - 1$ main eigenvectors obtained by singular value decomposition of the J columns in matrix $(\mathbf{K} - \mathbf{1}\mathbf{z}_0)$. The EMSC could then be seen as an extension of the path length correction presented by Miller and Naes [7].

Another solution is used presently: instead of using $\mathbf{K} = [\mathbf{k}_{\text{benzene}}, \mathbf{k}_{\text{toluene}}$ and $\mathbf{k}_{\text{xylene}}]$, we use $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2]$, where $\mathbf{p}_1 = \mathbf{k}_{\text{benzene}} - \mathbf{k}_{\text{xylene}}$ and $\mathbf{p}_2 = \mathbf{k}_{\text{toluene}} - \mathbf{k}_{\text{xylene}}$. Thereby the inversion in equation (16) becomes easy.

Equation (16) is thus used for EMSC estimation of unknown additive baseline offset a_i and unknown multiplicative path length coefficient b_i. Equation (9) is finally used for the actual conversion of each EMSC input spectrum $\mathbf{z}_i$ into EMSC output spectrum $\mathbf{x}_i$.

The $J - 1$ estimated elements in score vector $\mathbf{t}_i$ may now be divided by b_i and used as information about the concentration variations of the modelled constituents. However, the mathematical steps required to solve the inversion problems above make this a little complicated. Such explicit modelling of the constituent concentrations is more readily done when reformulated as a separate preprocessing step termed SIS.

Spectral interference subtraction. Each obtained (EMSC-corrected) spectrum is assumed to contain contributions from various analytes and interferents, as described in equation (5). Normally, the subsequent multivariate calibration regression modelling (e.g. PLSR) would pick up and correct for the different interference effects, provided the calibration data set spanned each of them independently. However, if we can estimate and subtract some if these interference effects already at the preprocessing stage, that reduces the cost and improves the interpretability of the subsequent PLSR modelling.

Now let $\mathbf{z}_i$ represent a spectrum input to the SIS preprocessing [equation (2)]. Note that in the present case SIS input $\mathbf{z}_i$ is the EMSC output. In the SIS process, these SIS input spectra $\mathbf{z}_i$ are to be converted into SIS output spectra $\mathbf{x}_i$ where certain interferences have been filtered out.

If we know the approximate spectra $\mathbf{k}_j$ of some of the major interferences $j = 1,2,\ldots,J$, one may in principle filter their effects out from the mixture spectra $\mathbf{x}_i$, for instance by projecting $\mathbf{z}_i$ on these interference spectra. In practice this is not usually advisable, since parts of the (unknown) analyte spectra will also be subtracted in the process. This makes the subsequent regression modelling rather difficult to interpret.

However, when we also know the approximate spectra of the analyte itself, then the SIS technique allows filtering of these main interferents in $\mathbf{x}_i$ without removing the analyte contributions.

Assume that $\mathbf{K}$ consists of both analyte spectrum (in this case $\mathbf{k}_{\text{toluene}}$) and the major interference spectra ($\mathbf{k}_{\text{benzene}}$ and $\mathbf{k}_{\text{xylene}}$).

The ideal additive mixture model in equation (5) is now expanded to include a residual spectrum $\mathbf{e}_i$ (noting that input $\mathbf{z}_i$ now represents output $\mathbf{x}_i$ from the previous preprocessing operation):

$$\mathbf{z}_i = \mathbf{K}\mathbf{c}_i + \mathbf{e}_i, \qquad (17)$$

where residual $\mathbf{e}_i$ reflects unmodelled constituent effects in $\mathbf{z}_i$ as well as non-linearities and measurement noise in $\mathbf{z}_i$ and effect of errors in $\mathbf{K}$.

The SIS modelling consists of WLS solution of equation (17), estimating concentrations $\mathbf{c}_i$ from spectrum $\mathbf{z}_i$, assuming constituent spectra $\mathbf{K}$ and statistical wavelengths weights diag($\mathbf{V}$):

$$\mathbf{c}_i = [\mathbf{K}'\mathbf{V}\mathbf{K}]^{-1}\mathbf{K}'\mathbf{V}\mathbf{z}_i, \qquad (18)$$

with

$$\mathbf{e}_i = \mathbf{z}_i - \mathbf{K}\mathbf{c}_i. \qquad (19)$$

Per definition, equation (17) can be written out explicitly as a sum of the different constituents, in this case

$$\mathbf{z}_i = \mathbf{k}_{\text{toluene}}c_{i,\text{toluene}} + \mathbf{k}_{\text{benzene}}c_{i,\text{benzene}} + \mathbf{k}_{\text{xylene}}c_{i,\text{xylene}} + \mathbf{e}_i. \qquad (20)$$

The subsequent SIS correction consist in reconstructing the mixture spectrum with the interferents weighted to zero:

$$\mathbf{x}_i = \mathbf{k}_{\text{toluene}}c_{i,\text{toluene}} + \mathbf{e}_i = \mathbf{k}_{\text{toluene}}{}^{*}1^{*}c_{i,\text{toluene}} + \mathbf{k}_{\text{benzene}}{}^{*}0^{*}c_{i,\text{benzene}} + \mathbf{k}_{\text{xylene}}{}^{*}0^{*}c_{i,\text{xylene}} + \mathbf{e}_i. \qquad (21)$$

Expressed in general terms, the SIS correction based on equation (17) is:

$$\mathbf{x}_i = \mathbf{K}\mathbf{W}\mathbf{c}_i + \mathbf{e}_i, \qquad (22)$$

where $\mathbf{W}$ is a diagonal matrix with one diagonal element for each modelled constituent $j = 1,2,\ldots,J$ (here 3), and with $w_{jj} = 1$ for the analyte(s) and $w_{jj} = 0$ for the modelled interferences. In the present example we have

$$\mathbf{W} = \begin{matrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$

Other weights may also be chosen. One such possibility is to use $w_{jj} > 1$ for the analyte, in order to enhance its contribution in output $\mathbf{x}_i$ relative to the residuals $\mathbf{e}_i$. This will probably have a strong graphic effect; whether or not it improves subsequent multivariate calibration further is unclear. Off-diagonal non-zero elements may also be used, if the concentrations of certain constituents are known to be inter-correlated.

Calibration Regression

PLSR [3] is a linear (additive) calibration method. That means that the final predictor function $f()$ in equation (3) can be summarized by the linear formulation

$$y_i = b_0 + \mathbf{x}_i\mathbf{b}. \qquad (23)$$

In order to obtain the calibration model (b_0,$\mathbf{b}$), the spectral data $\mathbf{x}_i$ are modelled as a sum of a few "factors" (mathematically defined difference spectra). The composition data y_i are then modelled as another sum of these same factors. This is described extensively in the literature, see, for example, ref. 2.

For linear instruments one expects to find a number of factors $a = 1,2,\ldots,A$ that corresponds to A_{expected}, the number of chemical or physical phenomena affecting the X-data. However, the optimal number of factors may be lower than A_{expected} if the initial calibration data set is small and noisy, and it may be higher than A_{expected} if there are unexpected interferents, curvatures, etc.

Explicit validation methods are therefore used to decide the optimal number of factors, A. Full cross validation, a conservative statistical validation method, is employed here: each sample is in turn kept as "secret", while the others are used for calibration; the "secret" sample is then used to test the predictive performance of that calibration model.

Experimental

Input data

Forty-seven known mixtures of the organic solvents benzene, toluene and xylene in various ratios were prepared. The NIR transmission spectra were measured for these mixtures, as well as for the three pure solvents, in a Guided Wave Model 200-45 process spectrophotometer, using a 2-m single-strand optical fibre and a transmission probe configuration. The transmittance data T were converted to optical densities.

The optical density spectra of the mixtures were then modified by adding random "baseline" offset and multiplying random "path length" scale factors, in order to simulate light scattering problems. These modified OD data are here regarded as the "raw input spectra".

Preprocessing

Light scattering correction. Traditional multiplicative signal correction (MSC) and EMSC were performed on the raw input OD spectra in order to separate physical and chemical effects in the spectra. In the EMSC the physical "baseline" offset and the "path length" scale factors were estimated in such a way that chemical OD differences between the three solvents were not mistakenly counted as physical effects. The analysis was done using the EMSC module (version 1.0) in the quantitative inference engine toolbox (QUIET) from Consensus Analysis AS (Ski Business Park, N-1400 Ski, Norway). The QUIET package is a set of independent batch-oriented C-programs for off- and on-line chemometrics and qualimetrics, running under PC DOS, Windows 3.0, Unix and VAX VMS.

Correction for known interferents. The EMS treated OD spectra were further simplified by SIS. In this analysis the approximate spectral contributions due to the interferents benzene and xylene were subtracted in such a way as not to modify the contributions of the analyte toluene and the contributions from unknown effects (chemical interactions, residual light scattering variations, etc). The analysis was done using the SIS module (Version 1.0) in QUIET (Consensus Analysis AS).

Calibration regressions

The PLSR calibration regressions for toluene from the NIR data were performed in the UNSCRAMBLER program, Version 3.01, from CAMO AS (Jarleveien 4, Lademoen, N-7041 Trondheim, Norway). The computations were carried out on an IBM PC.

Results and Discussion

Input data: fibre-optic process near infrared spectra

Figure 1(a) shows the OD spectrum of the analyte, $\mathbf{k}_{\text{toluene}}$, together with the two interferents benzene and xylene. The figure shows that the spectra are strongly overlapping. Wavelength channel 39 (1676 nm) seems to be the single most typical wavelength for the analyte.

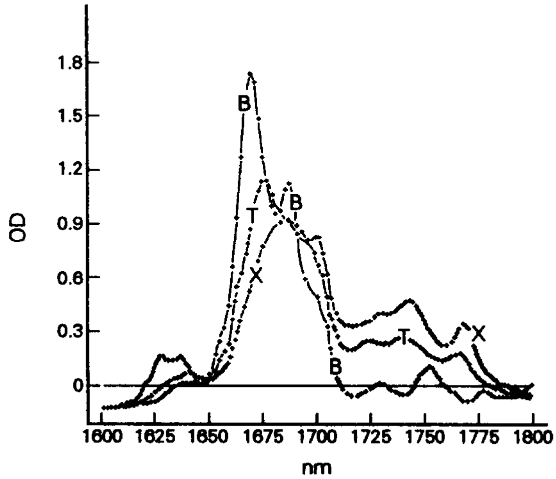

Figure 1
Pure constituent spectra. NIR OD spectrum of the analyte toluene (T) and the two interferences, benzene (B) and xylene (X).

Figure 2 shows the NIR OD spectra of some representative samples after different preprocessing, and Fig. 3 shows the corresponding univariate and multivariate predictive performance when calibrating for the analyte toluene.

Raw input spectra

Figure 2(a) shows the raw input OD spectra

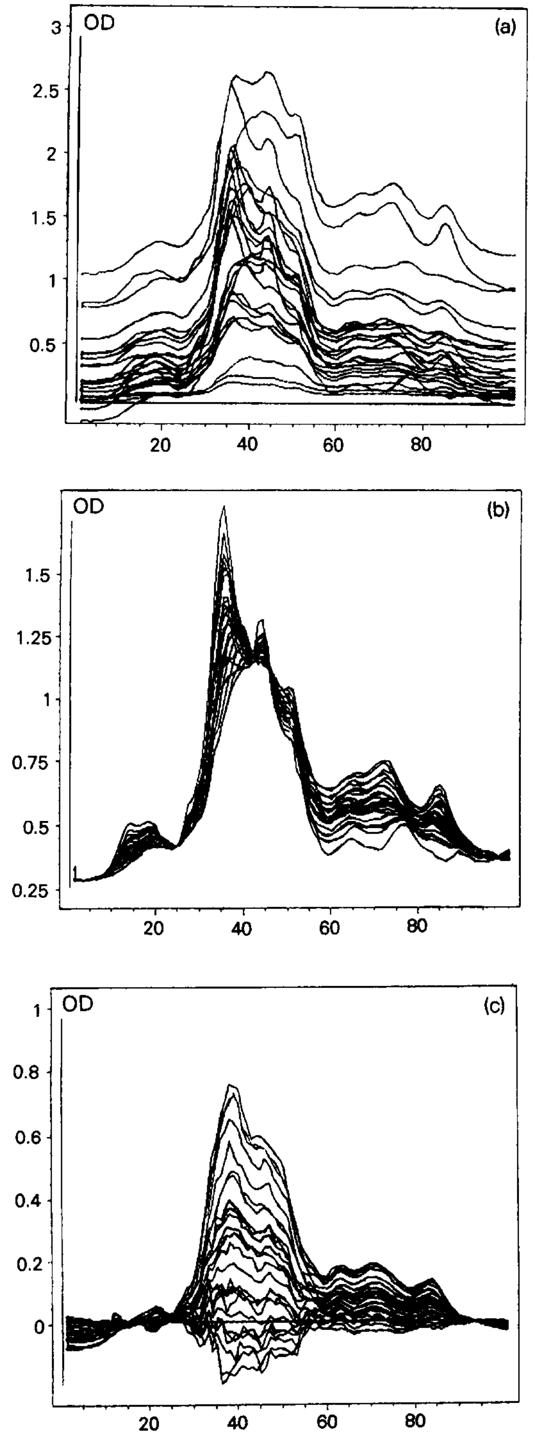

Figure 2
Spectra for solvent mixtures. (a) "Input spectra", illustrating uncontrolled random path length and baseline variations; (b) same spectra after EMSC; (c) same spectra after EMSC and SIS.

prior to preprocessing. A large degree of variation is evident. Figure 3(a) shows virtually zero correlation between the "best" single wavelength (channel 39) and toluene concentration for these data. Figure 3(b) shows that with multivariate PLSR calibration these 101 wavelengths together yielded a clearly improved predictive ability, but the relationship is not satisfactory.

Extended multiplicative signal correction

Figure 2(b) shows the same spectra after EMSC preprocessing. This normalizes all the spectra to an average estimated baseline level and an average estimated path length ("light scattering") level. The variability in the spectra is now much smaller. Figure 3(c) shows that satisfactory predictions can still not be attained using only one single wavelength, due to the spectral overlap between the analyte (toluene) and the interferences (benzene, xylene) and to "unknown interferences". However, when using all 101 wavelengths in multivariate PLSR calibration [Fig. 3(d)] an excellent predictive ability is attained.

Spectral interference subtraction

Figure 2(c) again shows the same EMSC treated spectra, after SIS preprocessing steps to remove additive effects of the known interferents benzene and xylene. The variability in the spectra is now quite systematic and represents mainly increasing levels of the analyte toluene (cf. Fig. 1). However, some low-toluene samples now show negative OD (as opposed to the expected level close to zero). This is probably due to non-representativity in the SIS component spectra, or to non-additivity, for example, caused by constituent interactions of some kind. Such unexpected spectral phenomena have to be "cleaned up" in the subsequent multivariate calibration.

Figure 3(e) shows that good predictions can now be attained with a single wavelength. But the single-wavelength calibration is still not optimal, due to the spectral overlap between the analyte (toluene) and "unknown interferences" in the spectra. With all 101 wavelengths combined in multivariate PLSR calibration [Fig. 3(f)] an excellent predictive ability is attained.

Figure 4(a) shows the prediction variance, as estimated by full cross-validation, for the raw input OD spectra, the MSC treated spectra, the EMSC treated spectra and the EMSC plus SIS treated spectra, as functions of the number of PLSR factors [i.e. the complexity of the calibration model $y_i = b(\mathbf{x}_i)$]. Figure 4(b) is an

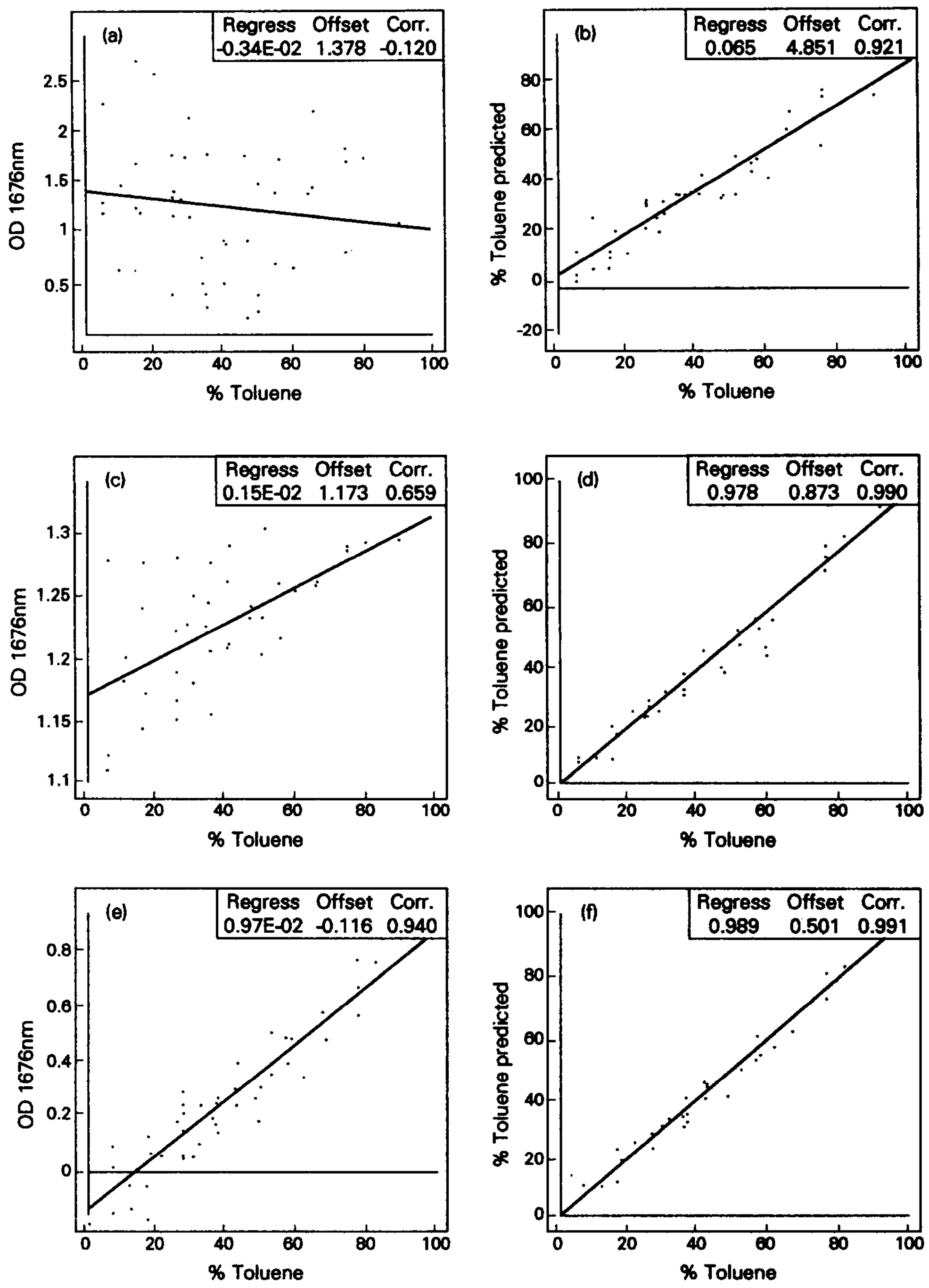

Figure 3
Calibration performances for toluene (abscissa). Left side: traditional univariate calibration, using the "best" single wavelength; ordinate, OD at 1676 nm, channel 39. Right side: multivariate calibration, using a combination of four PLSR factors; ordinate, predicted toluene concentration as obtained in full cross-validation. Parts (a) and (b) are "Input spectra". (b) Same spectra after EMSC. (c) Same spectra after EMSC and SIS.

expansion of Fig. 4(a), giving detailed comparison of the three preprocessing methods.

Figure 4(b) shows that the unpreprocessed spectra (upper curve) contained interference problems that the PLSR calibration could not handle well. A four-factor model was required in order to obtain reasonable predictive ability, but this yielded rather high predictive error (root-mean-square error of prediction RMSEP = ±8.9% toluene) [cf. Fig. 3(b)]. The MSC treated spectra had drastically improved predictive ability and needed two

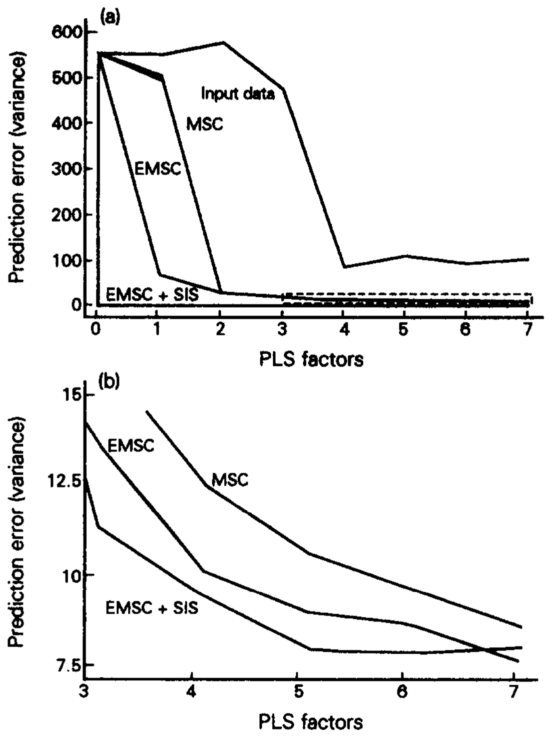

Figure 4
Predictive performance of calibration models for different spectral preprocessing. Abscissa: number of PLSR factors (calibration model complexity). Ordinate: average prediction error for toluene concentration, given as error variance in full cross-validation. Upper line: raw input OD spectra. Middle curves: OD after MSC and EMSC. Lower curve: OD after EMSC and SIS. (b) Expansion of data in (a) above.

factors to give good predictive ability, as expected for a three-constituent mixture system where the sum of the constituents is constant. The EMSC treated spectra gave a further improvement. The PLSR loadings for MSC and EMSC treated spectra (not shown here) were somewhat difficult to interpret, since they represent two difference spectra between the three solvents. A couple of minor factors gave slightly improved predictive ability [RMSEP = ±3.2% toluene after four factors, cf. Fig. 3(d)]. Other studies have revealed that these effects represent inter-constituent interactions of optical or chemical kind, deviations from fully linear instrument response, etc.

The SIS treated spectra (lower curve) gave good predictive ability already after one factor. The PLSR loading of this factor (not shown here) was virtually identical to the spectrum of the analyte itself. Very good predictive ability was again attained after, for example, four factors [RMSEP = ±3.1% toluene, cf. Fig. 3(f)].

Conclusion

This paper has shown that complicated NIR spectra which traditionally would be regarded as useless, can yield very good predictive ability in multivariate calibration by, for example, PLSR. It has also illustrated that data preprocessing can be very important, particularly for data with mixed multiplicative (path length) and additive (baseline, spectrally overlapping interferants) effects.

The MSC and EMSC preprocessing effectively removed most of the path length and baseline effects, allowing the subsequent additive PLSR to work well. EMSC gave a small, but clear improvement over the traditional MSC treatment. This difference is expected to be greater in systems where the absorbance spectra of the constituents differ more widely than in the present case (e.g. for mixtures containing water and displaying water temperature effects).

The SIS preprocessing allowed the removal of most of the known additive interference, making the resulting PLSR model easier to interpret. This preprocessing can be advantageous in situations where it is difficult to generate real calibration samples to span the full variability of the expected future sample qualities, e.g. in on-line process control applications. Every mixture spectrum, both in the calibration set and in future unknown samples, can instead be made "immune" against certain expected future interference types by SIS preprocessing.

However, it should be noted that for every interference corrected for, be it in PLSR or in preprocessing, there is a certain price to be paid in terms of reduced precision. The magnitude of this increased noise sensitivity depends on the degree of overlap between the analyte spectrum and the interference spectra corrected for.

NIR spectroscopy has a high potential in pharmaceutical and biomedical analysis. It is the first analytical technique to benefit fully from multivariate calibration. But this type of selectivity and reliability enhancement is equally applicable to other types of multichannel pharmaceutical or biomedical measurements — other types of spectroscopy (UV, vis, IR), chromatography, electrophoresis, image analysers — scanners, etc.

References

[1] P.C. Williams and K.H. Norris (eds), *Near-Infrared Technology in Agricultural and Food Industries*. Am. Assoc. Cereal Chem., St Paul, MN (1987).

[2] H. Martens and T. Naes, *Multivariate Calibration*. Wiley, Chichester (1989).

[3] S. Wold, H. Martens and H. Wold, *The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. Proc. Conf. Matrix Pencils* (A. Ruhe and B. Kågström, Eds), *Lecture Notes in Mathematics*, pp. 286–293. Springer, Heidelberg (1983).

[4] H. Martens, S.A. Jensen and P. Geladi, *Multivariate Linearity Transformation for Near-Infrared Reflectance Spectrometry. Proc. Nordic Symp. on Applied Statistics*, pp. 235–268. Stokkand Forlag, Stavanger (1983). Reprinted in ref. 6.

[5] P. Geladi, D. McDougall and H. Martens, *Appl. Spectrosc.* **39**, 491–500 (1985).

[6] H. Martens, Dr.techn. thesis, University of Trondheim, Norway (1985).

[7] C.E. Miller and T.A. Naes, *Appl. Spectrosc.* **44**, 895–898 (1990).

[*Received for review 7 January 1991;*
revised manuscript received 25 March 1991]