Light Scattering and light absorbance separated by Extended Multiplicative Signal Correction (EMSC). Application to NIT analyses of powder mixtures

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2003

Impact Factor: 5.64 · DOI: 10.1021/ac020194w · Source: PubMed

3 AUTHORS, INCLUDING:

Harald Martens
Norwegian University of Science and Techno…
33 PUBLICATIONS **2,534** CITATIONS
SEE PROFILE

Søren Balling Engelsen
University of Copenhagen
260 PUBLICATIONS **5,889** CITATIONS
SEE PROFILE

Anal. Chem. **2003,** *75,* 394−404

Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures

Harald Martens,* Jesper Pram Nielsen, and Søren Balling Engelsen

Centre for Advanced Food Studies, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30 DK-1958, Frederiksberg C, Denmark

The extended multiplicative signal correction (EMSC) preprocessing method allows a separation of physical light-scattering effects from chemical (vibrational) light absorbance effects in spectra from, for example, powders or turbid solutions. It is here applied to diffuse near infrared transmission (NIT) spectra of mixtures of wheat gluten (protein) and starch (carbohydrate) powders, linearized by conventional log(1/*T*). Without any correction for uncontrolled light scattering variation between the powder samples, these absorbance spectra could give reasonable predictions of the analyte (gluten), but only when using multivariate calibration with a much more complex model than expected. Standard MSC preprocessing did not work for these data at all; it removed too much analyte information. However, the EMSC preprocessing yielded powder spectra that obeyed Beer's Law more or less as if they had been obtained from transparent liquid solutions, apparently by isolating the chemical light absorption from additive, multiplicative, and wavelength-dependent effects of uncontrolled light-scattering variations. The model-based EMSC and its converse, the extended inverted signal correction (EISC), gave rather complete descriptions of the diffuse absorbance spectra and virtually indistinguishable performance in the calibration set and the test set of samples.

Preprocessing for Multivariate Calibration. Multivariable electromagnetic spectrophotometry in the near- or mid-infrared region offers great practical and economical advantages for analysis of large sample series, as demonstrated by diffuse NIR reflectance or transmittance spectroscopy in areas such as agriculture, food technology, pharmaceutics, and petrochemistry. Today, such high-speed instruments are routinely designed to yield precise quantitative determination for a variety of chemical and physical properties, using multivariate calibration to solve the selectivity problems caused by the lack of sample preparation and for automatic detection of outliers.[1] Preprocessing of the spectral measurements is used for optimizing the subsequent multivariate calibration. An example of this is the common linearization of transmittance T into absorbance $\log(1/T)$, which under ideal conditions is linearly related to chemical composition according to Beer's law.

When analyzing more or less intact complex samples by diffuse reflectance or transmittance spectroscopy, uncontrolled variations in light scattering is often a dominating artifact that complicates subsequent quantitative chemical analysis. This undesired scattering variation is due to uncontrolled physical variations in the measured samples: particle size and shape, sample packing, sample surface, etc. If the light scattering could be modeled and corrected for mathematically in a more elaborate preprocessing stage, these problems could be reduced or eliminated. The cost of NIR analysis could then be reduced, because the need for controlled sample preparation could be further reduced, the number of calibration samples could be reduced, and the statistical calibration modeling process could be simplified. Moreover, a pragmatic but reasonably accurate model-based light scatter correction may shed new light on the light scattering processes themselves. If successful, such a method for light-scatter correction might also be used for other types of instruments, for example, for reducing the need to remove uncontrolled turbidity prior to UV or VIS spectroscopy in general. The problem is how to describe light scattering mathematically in practice.

Additive and Multiplicative Models for Light Scattering. In some simple systems, a purely multiplicative effect of light scattering can be observed. In transmittance spectroscopy of, for example, transparent solutions, a change in the optical path length (e.g., cuvette width) scales the whole absorbance spectrum by a given factor, according to the Beer−Lambert law. In diffuse reflectance of powders under ideal conditions, a variation in the overall light-scatter coefficient between the samples likewise scales their chemical light absorption spectra according to the Kubelka−Munk theory.[2] In either case, if the scaling factor for each sample is known or can be measured, the multiplicative interference effect is easy to correct for by a simple rescaling.

In other systems, a purely additive effect of light scattering can under certain rare conditions be observed. In visible-range "transflection" spectroscopy in turbid aqueous solutions, a variation in turbidity level sometimes causes a simple baseline shift,[3]

* Corresponding author. E-mail: ham@kvl.dk.

(1) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, 1993.

(2) Kubelka, P.; Munk, F. *Z. Technol. Phys.* **1931**, *12*, 593−604.

10.1021/ac020194w CCC: $25.00
Published on Web 01/01/2003

which can be corrected for by a baseline subtraction according to Beer's law.

However, in practical diffuse spectroscopy of complex samples under realistic measurement conditions, such a fixed, purely multiplicative or purely additive modeling of light scattering appears to be an oversimplification. A more extensive modeling of light scattering variations is generally required. This modeling can be done as an implicit part of the multivariate calibration process or in an explicit preprocessing stage.

Implicit Scatter Correction during Multivariate Calibration Regression. Pragmatic multivariate calibration techniques[1] can to some extent implicitly compensate for unknown scatter variations. Multivariate techniques based on additive regression models, such as bilinear regression using partial least squares regression,[4] can automatically pick up and account for various unknown types of "scatter" by introducing extra regression components. The price for this implicit scatter compensation is a higher complexity of the calibration models, which then become prone to noise and difficult to interpret and require more calibration samples.[5] Bilinear additive modeling can be regarded as an implicit Taylor expansion of the underlying, unknown relationships in the data,[6] and for multiplicatively scatter-affected relationships the first-order (additive) approximation of bilinear models such as PCA or PLSR is simply not good enough.[1]

Explicit data analytical preprocessing prior to calibration can sometimes eliminate nonrelevant systematic sources of covariance and may then lead to more simple and robust regression models. Several additive preprocessing methods exist for removing irrelevant spectral contributions using covariance-based linear (additive) methods to subtract or down-weigh spectral components expected to represent interfering agents, such as spectral interference subtraction,[7] direct orthogonalization,[8] orthogonal signal correction (OSC)[8,9] and GLS preprocessing.[10] However, light-scatter effects tend to give more or less multiplicative contributions to the spectral data, for which these purely additive preprocessing methods[1] cannot be expected to work very well.

Explicit Methods for Scatter Correction in "Dirty" Systems. A common preprocessing alternative is to reduce the scattering problems by replacing the input spectra with their first or second derivatives, which will remove between-sample variations in baseline offset and linear baseline trends in the spectra. The tradeoff is usually noisier spectra as a result of the numerical calculation of the derivatives, and the derivative spectra may be more difficult to interpret.

Explicit multiplicative *and* additive preprocessing methods have also been developed, for example, multiplicative signal correction (MSC),[11,12] piecewise multiplicative signal correction (PMSC),[13] standard normal variate (SNV),[14] and the path length correction method PLC-MS.[15] Particularly the spectral derivative and MSC methods are now being used successfully in many applications, possibly because they have been built into several standard software systems for multivariate calibration. The mentioned data transformations have in common that they are relatively simple and that they can be applied without a priori knowledge about the samples and their spectra.

Using Prior Knowledge in the Preprocessing. If spectral regions are present where the target analyte or certain chemical interfering agents exhibit strong absorption, then the MSC parameter estimation may confuse chemical absorption and physical light-scattering effects, with dramatically bad results. An example of this will be demonstrated in the present paper. However, if a priori theory or quantitative data exist on the kind of samples and spectra involved, this can enhance the performance of the preprocessing. The simplest way is to use prior spectroscopic knowledge about the constituents' spectra, to ignore (down-weigh) spectral regions where dominating chemical constituents absorb very strongly, when determining how to scatter-correct each spectrum in the MSC.

A more ambitious way to use prior knowledge is to extend the MSC model to include new parameters to account for the physical and chemical phenomena that affect the measured absorbance spectra. One such approach is the extended multiplicative signal correction (EMSC). The basic version of EMSC was originally published by Martens and Stark.[7] In the present study, the EMSC is modified and applied to a different kind of spectral data from powder mixtures.

The inverse of the MSC model was first described by Helland et al. who named it inverted scatter correction (ISC).[16] In preparation for the present study, the method was extended in analogy to the EMSC and is here referred to as *extended inverted signal correction* (EISC). In a parallel application study,[17] a simpler version of the EISC (ISC with wavelength-dependent scattering correction) was then found to improve the calibration to protein content in wheat kernels from single seed NIT spectra.

The aim of this study is to first describe the EMSC method and compare it theoretically to the MSC, then to test the performance of MSC and EMSC on a set of diffuse transmittance spectra of powder mixtures and to compare the EMSC results to those from EISC.

THEORY

Modified Beer−Lambert's Law and the EMSC Model. For transparent solutions of a set of J absorbing chemical constituents, when Beer's law is obeyed, the theoretical chemical absorbance spectrum for sample i over a certain range of wavelength channels $\lambda = 1, 2, ..., \Lambda$, row vector $\mathbf{z}_{i,\mathrm{chem}} = [z_{i,\mathrm{chem},\lambda}, \lambda = 1, 2, ..., \Lambda]$ may

(3) Martens, H.; Martens, M. *Multivariate Analysis of Quality. An Introduction*; John Wiley and Sons Ltd.: Chichester, U.K., 2001.
(4) Wold, S.; Martens, H.; Wold, H. *Lect. Notes Math.* **1983**, *973*, 286−293.
(5) Martens, H.; Næs, T. *Near-Infrared Technology in the Agriculturel and Food Industries*; Williams, P. C., Norris, K. H., Eds., 2nd ed.; American Association of Cereal Chemists: St. Paul, 2001; Chapter 4.
(6) Wold, S. *Technometrics* **1978**, *20*, 397−405.
(7) Martens, H.; Stark, E. *J. Pharm. Biomed. Anal.* **1991**, *9*, 625−635.
(8) Andersson, C. A. *Chemom. Intell. Lab. Syst.* **1999**, *47*, 51−63.
(9) Wold, S.; Antii, H.; Lindgren, F.; Öhman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 229−244.
(10) Høy, M.; Ph.D. Thesis, NTNU, Trondheim, Norway, 2002; ISBN 82-471-5512-5.
(11) Martens, H.; Jensen, S. A.; Geladi, P. *Nordic Symposium on Applied Statistics*, Skagenkaien 12, 1983, 6-12-1983; Stokkand Forlag Publishing: CITY, 1983, pp 208−34.
(12) Geladi, P.; McDougall, D.; Martens, H. *Appl. Spectrosc.* **1985**, *39*, 491−500.
(13) Isaksson, T.; Kowalski, B. R. *Appl. Spectrosc.* **1993**, *47*, 702−709.
(14) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(15) Miller, C. E.; Næs, T. *Appl. Spectrosc.* **1990**, *44*, 895−898.
(16) Helland, I. S.; Næs, T.; Isaksson, T. *Chemom. Intell. Lab. Syst.* **1995**, *29*, 233−241.
(17) Pedersen, D. K.; Martens, H.; Nielsen, J. P.; Engelsen, S. B. *Appl. Spectrosc.* **2002**, *56* (9), 1206−1214.

be assumed to be a linear combination of the absorbance contributions of the J constituents,

$$\mathbf{z}_{i,\mathrm{chem}} = c_{i,1}\mathbf{k}_1' + \ldots + c_{i,j}\mathbf{k}_j' + \ldots + c_{i,J}\mathbf{k}_J' \qquad (1)$$

where $c_{i,j}$ is the concentration, and column vector $\mathbf{k}_j = [k_{\lambda,j}, \lambda = 1, 2, \ldots, \Lambda]'$, the absorptivity spectrum of the jth constituent. Under near-ideal conditions, with fixed optical path length, the measured absorbance spectrum for sample i, row vector $\mathbf{z}_i = [z_{i,\lambda}, \lambda = 1, 2, \ldots, \Lambda]$, is $\mathbf{z}_i \approx \mathbf{z}_{i,\mathrm{chem}}$.

If the constituent spectra $\mathbf{k}_j$, $j = 1, 2, \ldots, J$, are sufficiently distinct to be linearly independent of each other, eq 1 can be used for quantitative analysis in multicomponent systems based on multivariate calibration[1], assuming that the data have an additive chemical information structure.

To approximate physical effects related to light scatter variations, the measured absorbance spectra $\mathbf{z}_i$ of each sample i can be modeled as a scaled version of the ideal spectrum $\mathbf{z}_{i,\mathrm{chem}}$ according to Lambert's law, or some other linear transformation. Moreover, the light-scattering effect depends on the wavelength $\boldsymbol{\lambda}$, for which reason a smooth, polynomial wavelength dependency should also be taken into account.

The EMSC model can be written

$$\mathbf{z}_i \approx a_i + b_i\mathbf{z}_{i,\mathrm{chem}} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 \qquad (2)$$

where the coefficients a_i and b_i represent the baseline offset a_i and the path length b_i, relative to the baseline offset and path length in a reference spectrum. Coefficients d_i and e_i allow for unknown, smoothly wavelength-dependent spectral variations from sample to sample.

If the coefficients in eq 2 had been known theoretically, or estimated perfectly, then the EMSC correction

$$\mathbf{z}_{i,\mathrm{corrected}} = (\mathbf{z}_i - a_i - d_i\boldsymbol{\lambda} - e_i\boldsymbol{\lambda}^2)/b_i \qquad (3)$$

would remove the baseline and path-length variations as well as the wavelength-dependent spectral effects, yielding corrected spectra with only chemical absorbance information left: $\mathbf{z}_{i,\mathrm{corrected}} \approx \mathbf{z}_{i,\mathrm{chem}}$. Ideally, it would then be advantageous to replace the measured spectra $\mathbf{z}_i$ with $\mathbf{z}_{i,\mathrm{corrected}}$ in subsequent multivariate calibration, since the latter have a simpler and more linear relationship to the analyte concentration. Unfortunately, the parameters are usually unknown and have to be estimated from the available spectrum $\mathbf{z}_i$.

EMSC Parameter Estimation. The success of EMSC requires that good statistical estimates of the model parameters a_i, b_i, d_i and e_i in eq 2 are found from the measured spectrum $\mathbf{z}_i$ in such a way that they are insensitive to variations in the unknown chemical constituents' concentrations $c_{i,1}, c_{i,2}, \ldots, c_{i,J}$. This is done by including a quantitative description of the constituents' spectra $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_J$ in the EMSC model.

If the model for a sample's ideal chemical absorbance spectrum $\mathbf{z}_{i,\mathrm{chem}}$ (eq 1) is inserted directly into the physical model (eq 2), then the model obtains nonlinear parameter products $b_ic_{i,1}, b_ic_{i,2}, \ldots, b_ic_{i,J}$. This complicates the parameter estimation, because the important path length parameter b_i cannot be observed and estimated independently of the unknown chemical compositions $c_{i,1}, c_{i,2}, \ldots, c_{i,J}$. To have only a sum of linear parameters, Beer's Law expression in eq 1 can instead be rewritten in terms of variations around a chosen reference spectrum, which will be termed row vector $\mathbf{m}$,

$$\mathbf{z}_{i,\mathrm{chem}} = \mathbf{m} + \Delta c_{i,1}\mathbf{k}_1' + \Delta c_{i,2}\mathbf{k}_2' + \ldots + \Delta c_{i,J}\mathbf{k}_J' \qquad (4a)$$

where $\mathbf{m}$ is some reference spectrum, for example, measured in a "typical" sample or computed as the mean of a set of spectra, and $\Delta c_{i,j}$ represents the difference in constituent number j's concentration between the sample i and reference $\mathbf{m}$. Equation 4a can be inserted into the physical model (eq 2), which now gets a purely additive term $b_i\mathbf{m}$. This construction removes the problem of parameter products $b_ic_{i,1}, b_ic_{i,2}, \ldots, b_ic_{i,J}$ because b_i is obtained without other unknown terms. However this generates another, more statistical problem of "collinearity": If the reference spectrum $\mathbf{m}$ in eq 4a is more or less equal to a linear combination of the constituent spectra $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_J$, then it will be impossible for the parameter estimation to distinguish clearly between the contributions b_i from the reference $\mathbf{m}$ and the combined contributions $\Delta c_{i,j}$ from all the individual constituents, even if the contituent spectra themselves, $\mathbf{k}_j$, $j = 1, 2, \ldots, J$, are linearly independent of each other.

This collinearity problem between reference $\mathbf{m}$ and the constituent spectra $[\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_J]$ can be overcome in different ways,[7] for example, by replacing the J constituent spectra with $J - 1$ (or fewer) linear combinations of them. This new set of "chemical constituent spectra" to be used in the EMSC model can be obtained by singular value decomposition of the centered constituent matrix $\mathbf{G} = [(\mathbf{k}_1 - \mathbf{m}), (\mathbf{k}_2 - \mathbf{m}), \ldots, (\mathbf{k}_J - \mathbf{m})]$ using only the loadings/eigenvectors that correspond to clearly nonzero singular values.

The Binary Case. Assume, for example, that the samples are expected to contain only two main chemical constituents with different, linearly independent spectra, $\mathbf{k}_1$ and $\mathbf{k}_2$. Then, if $\mathbf{m}$ represents their average spectrum, there is only one nonzero singular value in $\mathbf{G} = [(\mathbf{k}_1 - \mathbf{m}), (\mathbf{k}_2 - \mathbf{m})]$. Its loading is proportional to $\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2$. Thus, eq 1 can be rewritten for this two-constituent model as

$$\mathbf{z}_{i,\mathrm{chem}} = \mathbf{m} + \Delta c_i\mathbf{k} \qquad (4b)$$

In the example of gluten/starch mixtures studied in this paper, the pure constituent spectra were defined from two calibration samples known to represent the pure constituents, namely a spectrum of pure gluten (sample 3) and a sample of pure starch (sample 93). With $\mathbf{k}_1 = \mathbf{k}_{\mathrm{gluten}} = \mathbf{z}_3$ and $\mathbf{k}_2 = \mathbf{k}_{\mathrm{starch}} = \mathbf{z}_{93}$, the reference spectrum $\mathbf{m}$ was defined by their average, $\mathbf{m} = (\mathbf{k}_{\mathrm{gluten}}' + \mathbf{k}_{\mathrm{starch}}')/2$, and the chemical variation spectrum, $\mathbf{k}$, by their difference, $\mathbf{k} = \mathbf{k}_{\mathrm{gluten}}' - \mathbf{k}_{\mathrm{starch}}'$.

A quantitative understanding of the resulting EMSC model can in this example be obtained from the fact that the constituent concentrations here add up to 1; $c_{i,\mathrm{gluten}} + c_{i,\mathrm{starch}} = 1$. With eq 1 rewritten as $\mathbf{z}_{i,\mathrm{chem}} = c_{i,\mathrm{gluten}}\mathbf{k}_{\mathrm{gluten}}' + c_{i,\mathrm{starch}}\mathbf{k}_{\mathrm{starch}}'$, then eq 4b can be rewritten more explicitly as

$$\mathbf{z}_{i,\mathrm{chem}} = \mathbf{m} + (c_{i,\mathrm{gluten}} - 0.5)\mathbf{k}' \qquad (4c)$$

This binary mixture model in eq 4b or 4c can now be inserted into the physical model in eq 2, yielding a linear statistical model with only additive terms, even for b_i,

$$\mathbf{z}_i = a_i\mathbf{1} + b_i\mathbf{m} + h_i\mathbf{k}' + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\epsilon}_i \qquad (5)$$

where vector $\mathbf{1} = [1, 1, 1, ..., 1]$ is introduced for matrix formality. Vector $\boldsymbol{\epsilon}_i$ is added to represent the residual spectrum of sample i containing random measurement noise and possible unmodeled spectral structures. Note that $h_i = b_i \cdot \Delta c_i = b_i \cdot (c_{i,\text{gluten}} - 0.5)$.

Vectors $\mathbf{m}$ and $\mathbf{k}'$ were already assumed to be sufficiently different from each other. Ideally, all five of the row "spectra" or "model vectors", $\mathbf{1}$, $\mathbf{m}$, $\mathbf{k}'$, $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}^2$, in eq 5 should be clearly linearly independent of each other. Then the EMSC parameters in vector $\mathbf{p}_i = [a_i, b_i, h_i, d_i,$ and $e_i]$ (eq 5) can be estimated by least squares regression of each input spectrum $\mathbf{z}_i$ to the model regressor matrix $\mathbf{M} = [\mathbf{1}; \mathbf{m}; \mathbf{k}'; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$ according to the regression model $\mathbf{z}_i = \mathbf{p}_i\mathbf{M} + \boldsymbol{\epsilon}_i$.

A versatile solution for the EMSC parameters in sample i is the conventional weighted least squares estimator,

$$\mathbf{p}_i = \mathbf{z}_i\mathbf{V}\mathbf{M}'(\mathbf{M}\mathbf{V}\mathbf{M}')^{-1} \qquad (6)$$

where the diagonal matrix $\mathbf{V}$ allows different weights for different wavelengths. Subsequently, the fit of the individual spectra to the model can be assessed by summarizing the estimated residual spectrum $\boldsymbol{\epsilon}_i = \mathbf{z}_i - \mathbf{p}_i\mathbf{M}$ for each sample $i = 1, 2, ...$ The weights can be defined on the basis of prior knowledge or by default set to 1.0.

After the unknown physical and chemical parameters [a_i, b_i, h_i, d_i, e_i] have been estimated for every sample in calibration and test sets, their corrected spectra, $\mathbf{z}_{i,\text{corrected}}$, $i = 1, 2, ...,$ can be obtained by eq 3. These corrected spectra can then be used as $\mathbf{X}$ in subsequent multivariate calibration for $\mathbf{y} =$ [gluten].

If there are more than two chemical constituents, an expression similar to eq 5 can be obtained by inserting eq 4a into eq 2. The EMSC model will then have more than one chemical difference spectrum, and more than one chemical parameter $h_{i,1}$, $h_{i,2}$, ..., but eq 6 can still be used for the parameter estimation, and eq 3, for the scattering correction.

Causality and Approximation. If the EMSC model in eq 5 and EMSC parameter estimation in eq 6 were perfect, the corrected spectra would be linearly related to the concentration of the chemical analyte, in this case $\mathbf{z}_{i,\text{corrected}} = \mathbf{m} + (c_{i,\text{gluten}} - 0.5)\mathbf{k}' + \boldsymbol{\epsilon}_i/b_i$, and this would be optimal for subsequent linear multivariate calibration modeling. In practice, the EMSC model will usually not be causally perfect because of, for example, unmodeled constituent interactions, stray light, and more intricate wavelength dependencies. Neither will the parameter estimates be perfect because of, for example, measurement noise in $\mathbf{m}$, $\mathbf{k}$, or $\mathbf{z}_i$. However, although the modeling inside the EMSC can be rather ambitious, the EMSC correction itself is rather conservative, since eq 2 also passes unmodeled information $\boldsymbol{\epsilon}_i$ from $\mathbf{z}_i$ to $\mathbf{z}_{i,\text{corrected}}$ (although scaled by factor b_i). As a consequence, the EMSC correction (eq 3) can be seen as just an approximation tool analogous to the subsequent multivariate calibration modeling itself: If the EMSC is found to simplify the structure without losing too much valuable information, then it has had a net positive effect on the calibration process as a whole.

The EMSC is intended primarily as a spectral preprocessing method to simplify the subsequent multivariate calibration, in which $\mathbf{z}_{i,\text{corrected}}$ in a set of samples $i = 1, 2, ..., N$ is used as $\mathbf{X}$ in multivariate calibration for some chemical constituent $\mathbf{y}$. However, the EMSC parameters themselves may also provide valuable information. First, if the EMSC model gives a sufficiently complete description of the absorbance spectra, a reasonably good estimate of the analyte concentration can be obtained already during the EMSC preprocessing before any multivariate calibration. In the present case, a good EMSC modeling would yield good gluten estimates from $c_{i,\text{gluten}} = h_i/b_i + 0.5$. Second, if the desired information in the absorbance measurements $\mathbf{z}_i$ are of a physical, rather than chemical nature, then the parameter estimates a_i, b_i, d_i and e_i can be used for extracting various types of physical variations. For instance, the matrix of EMSC parameters from eq 6, [$\mathbf{p}_i$; $i = 1, 2, ..., N$], could be used as $\mathbf{X}$ in the calibration step to describe a physical property, $\mathbf{y}$, such as particle size, sample packing, cuvette thickness etc.

Conventional MSC. The conventional MSC can be seen as a simplification of the EMSC. It assumes the same structure as the EMSC in a reduced version of eq 2.

$$\mathbf{z}_i \approx a_i + b_i\mathbf{z}_{i,\text{chem}} \qquad (7)$$

However, the MSC uses a much simpler model of $\mathbf{z}_{i,\text{chem}}$ than eqs 4a−c, because it assumes $\mathbf{z}_{i,\text{chem}} \approx \mathbf{m} + \boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i$ represents "unknown and irrelevant types of variation", including the spectral contributions due to chemical constituents. For estimating the MSC parameters a_i and b_i, $\boldsymbol{\delta}_i$ is simply ignored, and the basic MSC model is therefore written

$$\mathbf{z}_i = a_i\mathbf{1} + b_i\mathbf{m} + \boldsymbol{\epsilon}_i \qquad (8)$$

With the simplified model regressor matrix $\mathbf{M} = [\mathbf{1}; \mathbf{m}]$ with the only two rows presumably linearly independent, eq 6 yields estimates of the physical a_i and b_i parameters. The MSC correction, of course, is correspondingly simpler, and eq 3 reduces to $\mathbf{z}_{i,\text{corrected}} = (\mathbf{z}_i - a_i)/b_i$. The MSC may in many cases give useful light-scattering correction, because the unknown chemical contributions, $\boldsymbol{\delta}_i$, are often relatively uncorrelated with vectors $\mathbf{1}$ and $\mathbf{m}$ and, hence, do not affect the estimation of parameters a_i and b_i very much. However, major spectral effects due to changes in the chemical composition of the samples may render MSC inappropriate.

The results from the inverse version of EMSC, the EISC, will be briefly summarized at the end of the Results and Discussion section.

MATERIAL AND METHODS

Spectral Measurement. Transmittance spectra (T) in the range of 850−1050 nm were collected using an Infratec 1255 Food and Feed Analyzer (FossTecator, Höganäs, Sweden) fitted with a standard sample holder for five cylindrical cuvettes. A tungsten lamp (50 W) and a diffraction grating were used to create monochromatic light. The light passed through the powders and reached the silicon detector, and transmittance was recorded as $T = I/I_0$.

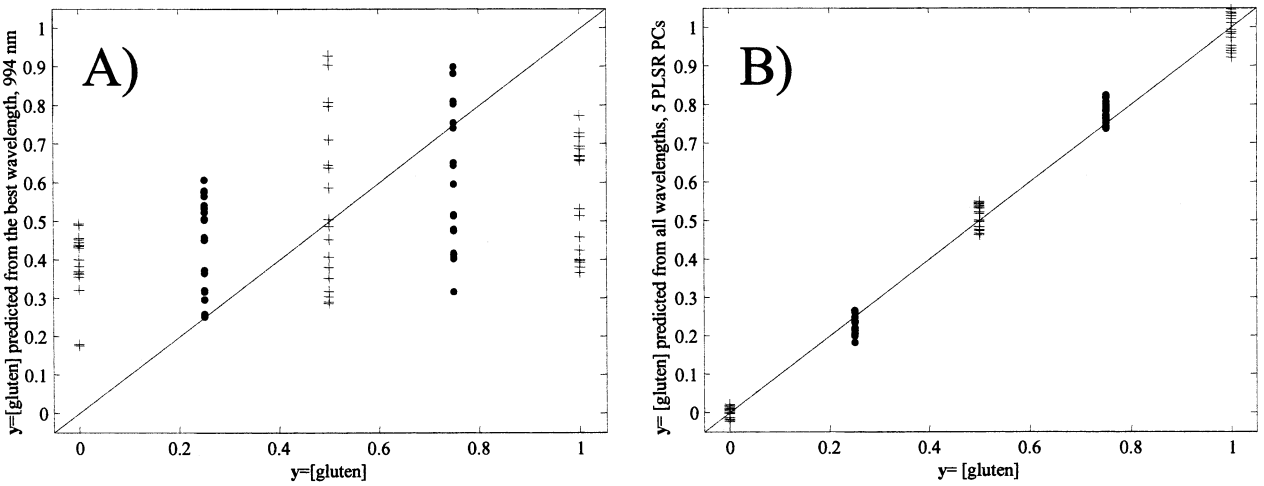

Figure 1. Calibrations without preprocessing of the log(1/*T*) measurements: (A) univariate calibration using the best wavelength channel, 994 nm ($\mathbf{x}_{73}$), and (B) multivariate calibration, combining all wavelength channels between 850 and 1050 nm ($\mathbf{x}_1 - \mathbf{x}_{100}$) via a PLSR model with 5 PCs. Abscissa: Analyte concentration **y** ([gluten] in powders), Ordinate: Prediction at the optimal model rank, $\hat{\mathbf{y}}_{A=5}$. Diagonal line: Theoretically correct predictions, + = calibration samples, ● = independent test samples.

Table 1. Experimental Design for the $N = 100$ Measured Spectra

samples	mixture	[gluten]/[starch]	packing	reps 1,2	use
1−10	1	1/0	loose	1−5, 6−10	calibration
11−20	1	1/0	firm	11−15, 16−20	calibration
21−30	2	0.75/0.25	loose	21−25, 26−30	test
31−40	2	0.75/0.25	firm	31−35, 36−40	test
41−50	3	0.5/0.5	loose	41−45, 46−50	calibration
51−60	3	0.5/0.5	firm	51−55, 56−60	calibration
61−70	4	0.25/0.75	loose	61−65, 66−70	test
71−80	4	0.25/0.75	firm	71−75, 76−80	test
81−90	5	0/1	loose	81−85, 86−90	calibration
91−100	5	0/1	firm	91−95, 96−100	calibration

Samples. Industrial-grade wheat gluten (∼80% protein) and pure wheat starch (Merck Eurolab 73502-250) were used for the binary mixture design. Five mixtures with different ratios of gluten/starch (by weight, Table 1) were prepared by weighing on a precision balance and were mixed thoroughly. For each of these mixtures, ∼2 g was taken out randomly in five sampling replicates and filled loosely into five different glass cuvettes (horizontal diameter, 25 mm) to a vertical sample thickness of ∼8 mm. The NIT spectra of the five sampling replicates were measured vertically in two consecutive spectral replicates. Then the powder in each of the five cuvettes was packed more firmly (compressed by hand) to a sample thickness of ∼5−6 mm, and their NIT spectra were measured again in two consecutive spectral replicates. In total, the five mixtures × two powder packings × two spectral replicates × five sample holders/powder sampling replicates amounted to a factorial design with $N = 100$ samples ($5 \times 2 \times 2 \times 5 = 100$ spectra).

Data Analysis. Each transmittance spectrum T was first changed into absorbance $A = \log(1/T)$. These absorbance spectra were then subjected to various preprocessing methods. The corrected spectra from eq 3 were submitted to multivariate calibration and prediction[1]. The 100 wavelength channels between 850 and 1050 nm were used in regressor matrix $\mathbf{X} = [\mathbf{z}_{i,\text{corrected}}, i = 1, 2, ..., N]$, and the concentration of the analyte, [gluten], was used as regressand $\mathbf{y} = [c_{i,\text{gluten}}, i = 1, 2, ..., N]$. Bilinear modeling by partial least squares regression[4] (PLSR) was used as a low-rank calibration method. To ensure the validity of the obtained results, only the spectra from three of the five mixtures (mixtures 1, 3, and 5; in total, 60 spectra) were used as calibration samples for developing the calibration model ($\mathbf{y} \approx f(\mathbf{X})$). The remaining mixtures (mixtures 2 and 4; in total, 40 spectra) were used as an "independent" test set, for predictions $\hat{\mathbf{y}} = f(\mathbf{X})$. The predictive validity for models with 0, 1, 2, ... latent variables (PCs) in the PLSR model was then estimated in two different ways, as described in Martens and Martens[3]: (1) $\text{RMSEP}_{\text{test}}$ = the prediction error in the independent test set, and (2) RMSEP_{cv} = the corresponding cross-validated root mean square error of prediction in the calibration set using a three-segment version of cross-validation, keeping all replicates for one of the three calibration mixtures out at a time for independent prediction testing. The preprocessing, multivariate calibration and graphics were performed using software written in MatLab version 6.1 (The Matworks, Inc., Natick, MA).

RESULTS AND DISCUSSION

No Correction for Light Scattering. Figure 1 shows the performance or the spectral measurements without any preprocessing, that is, log(1/*T*). Figure 1A illustrates traditional univariate calibration: the wavelength with the best correlation to the analyte, **y** = [gluten], 994 nm ($\mathbf{x}_{73}$), was used as regressor for **y** in the calibration samples, and the resulting model was applied to the 40 samples in the independent test set. The figure shows how selectivity problems render such traditional calibration useless in diffuse spectroscopy of light-scattering samples, even at the "best" wavelength, which also agrees with the knowledge gained through experience with NIR spectroscopy over the past decades.

In contrast, Figure 1B shows that multivariate calibration—in this case by PLSR—can overcome most of the selectivity problems if there are enough calibration samples and if enough PCs are used in the model. The cross-validation within the calibration sample set (+) showed that five PCs gave the best compromise between predictive error and model complexity. The figure also

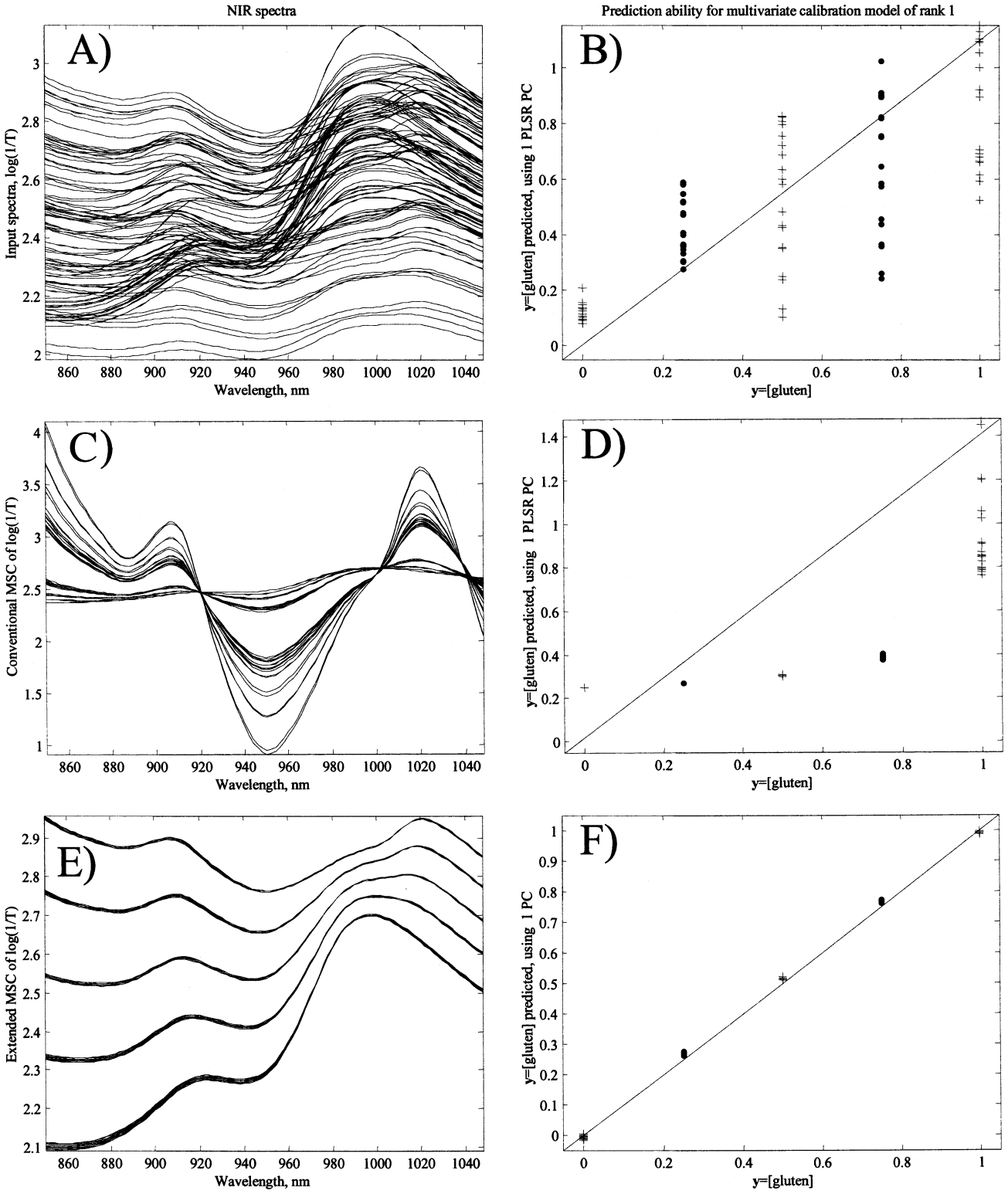

Figure 2. Different preprocessing methods: (left) NIT spectra of five gluten/starch mixtures in 20 replicates (i.e., all 100 measured spectra) and (right) analyte concentration predicted from spectra at the expected model rank, $\hat{\mathbf{y}}_{A=1PC}$ vs input $\mathbf{y}$. (A,B) No preprocessing, log(1/*T*). (C,D) Simple multiplicative scatter correction (MSC) of the log(1/*T*) spectra. (E,F) Extended multiplicative scatter correction (EMSC) of the log(1/*T*) spectra. Diagonal line: theoretically correct predictions, + = calibration samples, ● = independent test samples.

shows that this model gives fairly good predictive ability also for the test samples (●).

Considering the simplicity of the chemical composition of these gluten/starch mixtures, the complexity of the input spectra shown in Figure 2A is surprising. Yet, a five-PC model (Figure 1B) gave relatively good predictive ability for $\mathbf{y}$ = [gluten] from these spectra. However, since the samples represent binary mixtures in which the analyte concentrations add up to 1, only one single PC should ideally suffice in the calibration model if the spectral response is truly bilinear. Figure 2B clearly illustrates that a single-

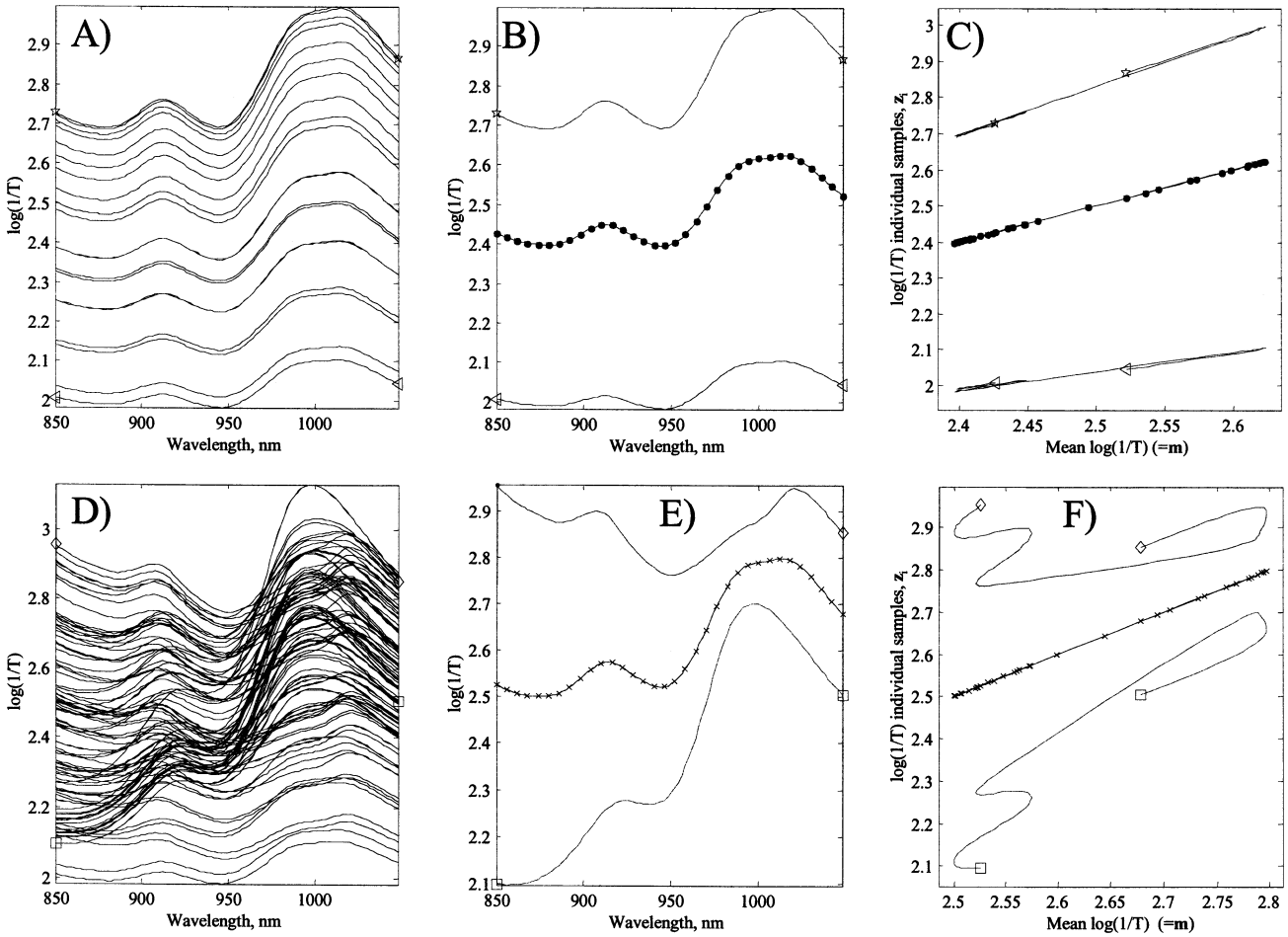

Figure 3. MSC and its problem when the chemical differences are large. Upper row: constant chemistry, different physics. (A) NIT spectra of 20 samples of identical chemical composition (50/50 gluten/starch mixture, samples 41−60). (B) NIT spectra of two of the samples, $\mathbf{z}_{41}$ (☆) and $\mathbf{z}_{58}$ (△), and the mean spectrum $\mathbf{m}_{50/50}$ of 20 samples (●). (C) $\mathbf{z}_{41}$ (☆), $\mathbf{z}_{58}$ (△) and $\mathbf{m}_{50/50}$ (●) plotted against $\mathbf{m}_{50/50}$ (abscissa). Lower row: different chemistry, different physics. (D) NIT spectra of 20 replicate spectra of each of the five mixtures of gluten and starch (samples 1−100). (E) Two of the NIT spectra, $\mathbf{z}_3$ (◊) and $\mathbf{z}_{93}$ (□), and the mean spectrum $\mathbf{m}$ of all of the spectra (×). (F) $\mathbf{z}_3$ (◊), $\mathbf{z}_{93}$ (□), and $\mathbf{m}$ (×) plotted against $\mathbf{m}$ (abscissa).

PC model of the absorbance spectra does not give a satisfactory selectivity for the analyte.

Conventional Multiplicative Signal Correction (MSC). It is expected that most of the selectivity problems in this study are due to physical variations caused by uncontrolled variations in light scattering due to sample packing, sample surface topology, particle size, and possible variations in amount of sample in the cuvettes. To separate such physical light-scattering variations from the chemical absorbance variations, MSC was first applied to the spectra in Figure 2A. Assuming the simplified chemical model that ignores the systematic chemical variations, each calibration and test set spectrum $\mathbf{z}_i$, $i = 1, 2, ..., 100$ was modeled by eq 8, with parameter estimation according to eq 6. As usual in MSC, the reference spectrum $\mathbf{m}$ was defined as the average of the calibration samples' spectra. Changing the weights $\mathbf{V}$ in eq 6 did not appreciably improve the present results; hence, for simplicity, equal weights of 1 were used for all wavelength channels. Figure 2C shows the corrected $\log(1/T)$ spectra $\mathbf{z}_{i,\text{corrected}}$ after MSC preprocessing. Although it is clear that the absorbance spectra now look a lot less complex, the variation in some of the spectra (the pure gluten samples) seems to have become exaggerated, and the others have become virtually indistinguishable. This is confirmed in Figure 2D, which shows the predictive performance for $\mathbf{y}$ from MSC corrected absorbance spectra $\mathbf{X}$, using the multivariate calibration model with the expected model complexity, one PC. The 20 replicates of the pure gluten samples give very different predictions, but the remaining samples give almost identical and quite erroneous results, both within the calibration set and in the test set. Apparently, because the simple MSC ignored the large, systematic chemical variations caused by the gluten and starch differences, $\boldsymbol{\delta}_i$, these differences contaminated the estimates of a_i and b_i in eq 6 whereby much of the chemical information in the spectra was erroneously removed by the MSC correction.

Extended Multiplicative Signal Correction (EMSC). Each of the 100 NIT absorbance input spectra $\log(1/T)$ in Figure 2A were instead modeled according to eq 5, submitted to EMSC parameter estimation by eq 6, and corrected according to eq 3. The lowest part of Figure 2 shows the effect of the EMSC pretreatment.

The preprocessed spectra are shown in Figure 2E, where the five mixtures are seen as five distinct spectral patterns. The 20 replicate samples for each mixture, so clearly different in the input spectra (Figure 2A), now appear more or less superimposed and indistinguishable.

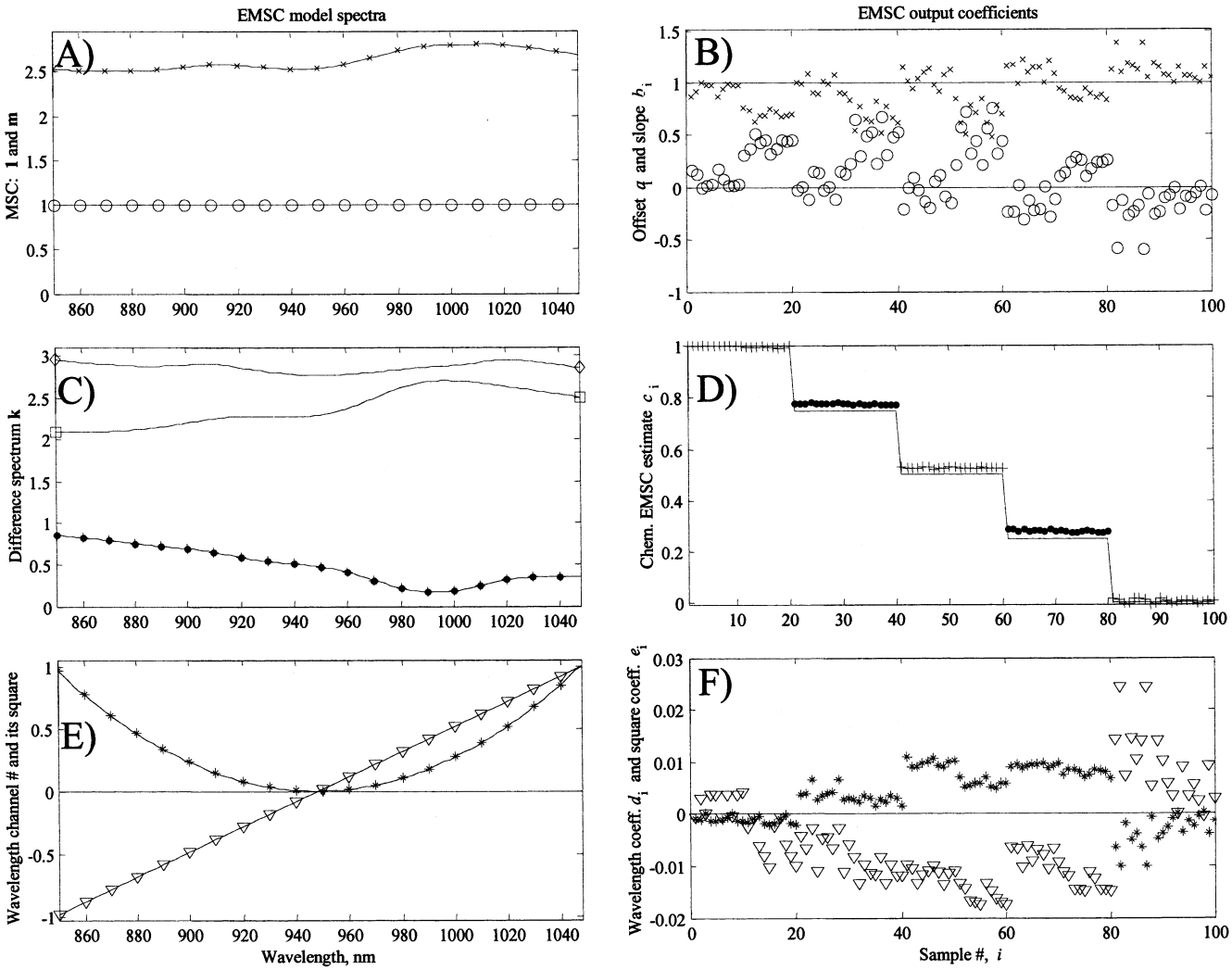

Figure 4. EMSC model and its estimated parameters: left, EMSC model spectra (eq 5) as functions of wavelength; right, estimated EMSC model parameters from eq 6 shown as functions of sample no. (A) MSC part of the model: vector **1** (○) and reference spectrum **m** (×). (B) Additive parameter a_i (○) and multiplicative parameter b_i (×). (C) Chemical model extension: spectra of two samples, $\mathbf{z}_3$ (◇) and $\mathbf{z}_{93}$ (□), and the EMSC model vector $\mathbf{k} = \mathbf{z}_3 - \mathbf{z}_{93}$ (●). (D) Chemical EMSC parameter: analyte concentration estimated already during the EMSC, c_i, for the calibration samples (+) and the independent test samples (●). The true analyte fractions are given by the step line. (E) Physical model extensions: wavelength index λ (▽, between −1 and 1) and its square λ^2(*). (F) Physical EMSC parameters: wavelength coefficients d_i (▽) and wavelength-squared coefficient e_i (*).

Figure 2F shows the predictive performance of the EMSC corrected spectra (Figure 2E) using the expected model complexity (one PC) in the PLSR calibration model. The five mixtures give gluten concentration predictions close to the theoretical expectation (dotted line). Hence, almost all of the selectivity and linearity problems evident in Figure 2B,D have been eliminated by the EMSC preprocessing.

The EMSC pretreatment in this data set provided a good reduction of the light-scattering problems, and this simplified the subsequent calibration for the analyte, $\mathbf{y}$ = [gluten]. The results are seen to be equally good for the three mixtures used for calibration (+) and the two intermediate mixtures used for the independent test set (●). A small response curvature can be observed, indicating room for possible improvement of the EMSC preprocessing or the subsequent calibration.

The Problem of the MSC: Mixed Chemical and Physical Variation. Figure 3 illustrates the reason for the poor performance of the simple MSC. The upper half of Figure 3 shows how the spectra behave when all the input spectra resemble each other chemically, while the lower half demonstrates how the spectra behave when chemical variations are dominant. Figure 3A shows the 20 replicate input spectra of mixture no. 3 (containing equal amounts of gluten and starch). The measured samples differ only in physical properties, such as light scattering. Since the mean $\mathbf{m}_{50/50}$ of the spectra in Figure 3A also represents this 50/50 mixture, the "unknown chemical variability", $\boldsymbol{\delta}_i$, is actually equal to 0, and every spectrum in Figure (3a) should therefore follow the MSC model (eq 8), $\mathbf{z}_i \approx a_i\mathbf{1} + b_i\mathbf{m}_{50/50}$. Therefore, when one of the spectra $\mathbf{z}_i$ in Figure 3A is plotted against this mean spectrum, $\mathbf{m}_{50/50}$, the consecutive wavelength channels should form a series of spots that generate a straight line with a certain offset, a_i, and a certain slope, b_i. Figure 3B shows two spectra selected from Figure 3A (nos. 41 and 58) together with this mean spectrum $\mathbf{m}_{50/50}$. When the individual spectra are plottet against the mean spectrum (Figure 3C), they are indeed seen to form nice straight lines from which their different offsets, a_i, and slopes, b_i, can be unambiguously estimated.

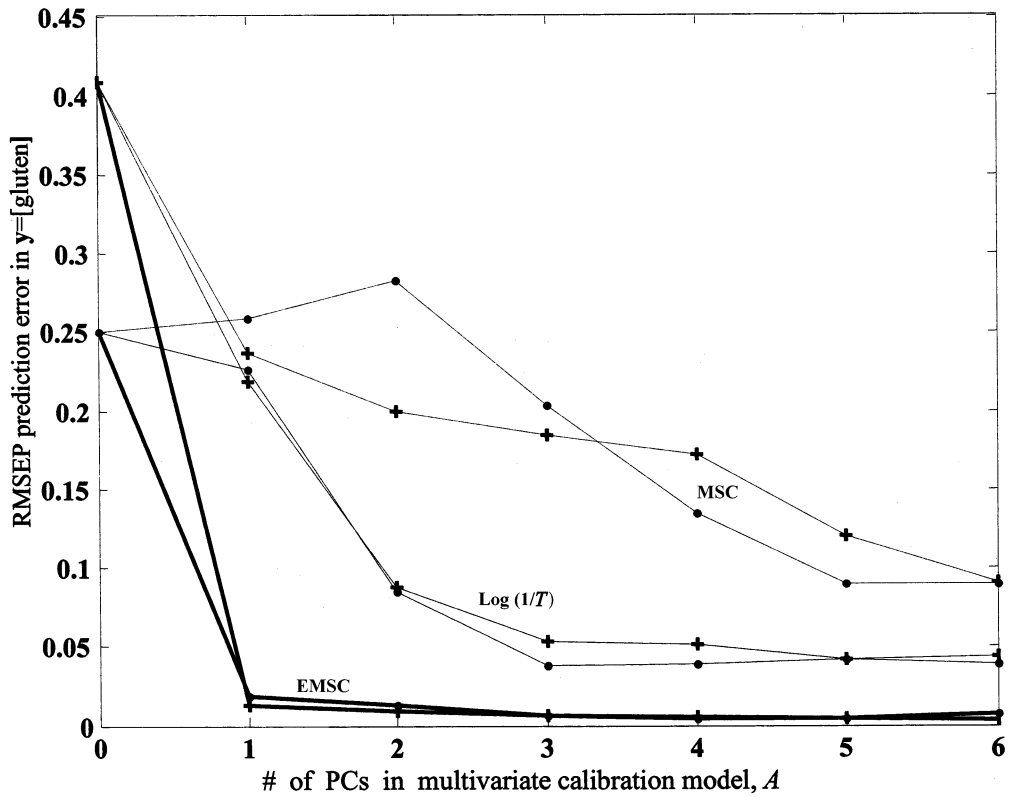

Figure 5. Comparison of preprocessings. Prediction error (RMSEP) for analyte **y** = [gluten] vs no. of PCs in the multivariate calibration. Calibration set (+) (mixtures 1, 3, and 5); cross-validation between mixtures. Independent test set (●) (mixtures 2 and 4).

In contrast, the lower part of the figure illustrates what happens when the input spectra exhibit clear chemical variations $\boldsymbol{\delta}_i$ (Figure 3D). In this case, two arbitrarily chosen spectra (nos. 3 and 93) (Figure 3E) now display strange wiggles when plotted against the mean spectrum of all of the calibration samples, **m** (Figure 3F). Clearly, in this case, it is difficult to estimate and correct for by spectral slopes b_i and offsets a_i without removing chemical information. The chemical and physical effects are seriously entangled in these spectra, and it is difficult to isolate wavelength ranges where they can be disentangled. The purpose of the EMSC is to reduce this problem.

The EMSC Model and its Parameter Estimates. Although the EMSC can be used in a software module for "black box" preprocessing, it is interesting to study the reasons for the good EMSC modeling reported in Figure 2E and F. An overview of the EMSC process is given by Figure 4, which shows the spectra (ACE) in the EMSC model (eq 5) and the corresponding parameter estimates for the 100 samples obtained by eq 6 (BDF). The first row in Figure 4 illustrates the offset and slope (the "MSC part") of the EMSC model. Figure 4A shows vectors **1** and **m** ($= (\mathbf{k}_{\text{gluten}}' + \mathbf{k}_{\text{starch}}')/2$), and Figure 4B, the corresponding additive and multiplicative parameter estimates a_i and b_i. These parameter estimates are seen to vary in a complex pattern around their expected values (0 and 1, respectively). For example, it is observed that among the first 20 samples (pure gluten), the first 10 (loosely packed) differ markedly from the next 10 (densely packed), illustrating that the EMSC parameters can be used in their own right for quantifying physical properties, such as sample packing. This pattern is also seen clearly in the next samples 21−40 and 41−60 representing 75 and 50% gluten, but less clearly for samples 61−80 and 81−100, that is, 25 and 0% gluten. Parameter estimates a_i and b_i are seen to be strongly negatively correlated; $r(a_i, b_i) = -0.998$ over all 100 samples. The reason for this correlation is presently unclear, but it could reflect instrument geometry. Work is in progress to check this.

The second row in Figure 4 illustrates the chemical part of the EMSC model. The two thin curves in Figure 4C represent the two input spectra $\mathbf{z}_3$ ($\mathbf{k}_{\text{gluten}}'$, top) and $\mathbf{z}_{93}$ ($\mathbf{k}_{\text{starch}}'$) used for defining **m** and the chemical difference spectrum $\mathbf{k} = \mathbf{k}_{\text{Gluten}}' - k_{\text{Starch}}'$ (●). Figure 4D shows the resulting estimate of the analyte concentration c_i, obtained already during the EMSC preprocessing. For comparison, the true analyte fractions, elements y_i in vector **y**, are shown as step lines. The figure reveals that good concentration estimates c_i in this case were obtained already during the EMSC preprocessing, both for the calibration samples (+) and for the test samples (●).

The third row in Figure 4 shows the wavelength correction part of the EMSC model. Figure 4E shows wavelength vector $\boldsymbol{\lambda}$ as a linear function of the number of nanometers, ranging from −1 to +1, as well as its square. Figure 4F shows their coefficient estimates d_i and e_i. Both effects are seen to be centered around 0, and their variations are rather small compared to that of offset a_i (Figure 4B). A negative correlation between d_i and e_i is observed: $r(d_i,e_i) = -0.80$ over all 100 samples, but the reason for this is not yet clear.

How the Different Preprocessing Methods Perform for Multivariate Calibration. Figure 5 summarizes the three ways to preprocess the NIT absorbance spectra in terms of their ability to predict **y** = [gluten] when used as **X** in the subsequent multivariate linear calibration modeling by PLSR. Each curve represents a root mean square error of prediction (RMSEP), that is, an "average" of the prediction error in **y**, plotted against

calibration model complexity (0, 1, 2, ..., 6 PCs in the PLSR model). The three line pairs were obtained using the three different preprocessing methods reported above: the middle pair of curves represent the untreated NIT absorbance spectra shown in Figure 2A; the upper pair of curves represent the MSC pretreated spectra (Figure 2C); and the lower pair of curves, the new EMSC pretreated spectra (Figure 2E). Curves marked with the symbol "+" ($RMSEP_{cv}$) were estimated by the cross-validation within the calibration set, and the curves marked with symbol "●" were obtained for the test set ($RMSEP_{test}$). At one PC, the curves summarize the predictions already shown in Figure 2B,D,F, respectively.

As usual, the cross-validation and the test-set methods gave rather similar RMSEP error estimates when enough PCs were used to avoid underfitting. An example of this is seen at the "optimal" five-PC solution for the EMSC preprocessed spectra, where $RMSEP_{cv} = RMSEP_{test} = 0.005$. Since all RMSEP errors are given in **y**'s concentration scale 0−1, this latter corresponds to an average prediction uncertainty of ±0.5% on a scale of 0−100%, which is quite low, considering the complexity of the original input spectra (Figure 2A).

Figure 5 confirms that the conventional MSC had a detrimental effect in the present case: the subsequent bilinear modeling behaved much worse than the untreated $\log(1/T)$ spectra and required more PCs. With the ideal model complexity (one PC), both of these simple methods displayed poor performance. In contrast, already after one PC, the EMSC solution yielded very low prediction errors ($RMSEP_{cv} = 0.013$ and $RMSEP_{test} = 0.019$). In fact, this solution was better than any of the other solutions using raw $\log(1/T)$ or MSC pretreated spectra. The EMSC model (eq 5) fitted the 100 spectra quite well. Averaged over all $N = 100$ samples and all $K = 100$ wavelength channels, the variability of the spectra was reduced from a total initial standard deviation of 0.2 absorbance units in the spectra $\mathbf{z}_i$ (Figure 2A) to a total standard deviation of only 0.0005 absorbance units in the EMSC residuals ϵ_i (eq 5).

The choice of samples 3 and 93 to represent the pure constituents in the EMSC model was somewhat arbitrary. Work is in progress to study the effect of using other pairs of samples and to simplify and optimize the EMSC modeling.

Extended Inverse Scatter Correction: EISC. As mentioned previously, the EISC is an extension of the ISC method, which switches the roles of $\mathbf{z}_i$ and **m** compared to MSC. The ISC model was recently extended with smooth wavelength-dependent terms, $d_i\boldsymbol{\lambda}$ and $e_i\boldsymbol{\lambda}^2$, for successful preprocessing of NIT spectra of intact wheat kernels.[17] Presently, it was further extended with the chemical term $h_i\mathbf{k}$, to yield the Extended ISC (EISC) model with residual vector $\boldsymbol{\gamma}_i$.

$$\mathbf{m} = a_i\mathbf{1} + b_i\mathbf{z}_i + h_i\mathbf{k} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\gamma}_i \qquad (9)$$

After having estimated the parameters a_i, b_i, h_i, d_i, and e_i by regressing **m** on $\mathbf{M} = [\mathbf{1}\ \mathbf{z}_i\ \mathbf{k}'\ \boldsymbol{\lambda}\ \boldsymbol{\lambda}^2]$ in analogy to eq 6, the spectra were corrected by $\mathbf{z}_{i,\text{corrected}} = a_i\mathbf{1} + b_i\mathbf{z}_i + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2$.

The pretreated NIT spectra $\mathbf{z}_{i,\text{corrected}}$ resulting from the EISC were visually indistinguishable from those of EMSC (Figure 4A), and so was their predictive performance in subsequent multivariate calibration by PLSR (Figures 2B and 5); hence, they are not reported here. The similarity between the present EMSC and EISC modeling results is interesting, since the difference between the two methods is analogous to the difference between reverse ("classical") and forward ("inverse") calibration methods, except that EMSC and EISC employ regression over wavelengths, whereas calibration employs regression over samples. It is well-known that forward and reverse calibration[1] give almost identical results when the models fit the data very well. Hence, the small residuals ϵ_i from EMSC and $\boldsymbol{\gamma}_i$ from EISC explain the similarity in the results from the EMSC and EISC preprocessing.

In the present study, the EMSC (and EISC) method was applied to diffuse transmittance spectra of binary powder mixtures. More work is needed to test the method on more complex data. But this type of preprocessing is expected to be useful for diffuse reflectance or transmittance spectra obtained in other spectral ranges, for example, in UV, vis, or IR and from more complex types of mixtures and for other light-scattering materials. Preliminary experience has also shown benefits from applying EMSC or EISC to very different types of data, for example, in chromatography to correct uncontrolled variations in baseline offset and total sample concentration and in descriptive sensory analysis to correct for uncontrolled variations in how individual assessors use the sensory scale (work is in progress in this respect).

In many cases, the measured data cannot be expected to be completely modeled already at the preprocessing stage. However, preprocessing could still be quite useful by simplifying the subsequent multivariate calibration regression as well as by revealing something about the nature of the selectivity problems. The ambitious, theory-driven EMSC and EISC preprocessing can then describe and correct for interference phenomena that are more or less expected and understood, whereas the subsequent data-driven multivariate calibration regression can reveal and correct for unexpected or poorly understood phenomena. In summary, this could be seen as a flexible and powerful combination of deductive and inductive traditions in analytical chemistry.

CONCLUSION

EMSC preprocessing simplified a set of diffuse NIT absorbance spectra measured in the lower NIR range by transmittance through 5−8 mm of highly light-scattering mixtures of gluten and starch powders. The success is presumably due to the ability of spectral modeling to separate chemical light-absorbance and physical light-scatter effects. Using prior knowledge about the absorbance spectra of the major constituents and assumptions about smooth wavelength-dependency of the light scattering variation, the corrected spectra became insensitive to light scattering variations and responded linearly to the analyte concentration. Thus, the subsequent multivariate calibration regression model became much simpler and had better predictive performance. In fact, the preprocessing proved so effective in the present study that the multivariate calibration regression became superfluous, since the analyte fraction estimated from the EMSC modeling itself provided a direct measure of the desired analyte fraction. Extended inverted signal correction (EISC) yielded corrected spectra and calibration models that were almost indistinguishable from those of the corresponding EMSC.

ACKNOWLEDGMENT

The authors thank Danish Slaughteries and the Ministry of Food, Agriculture, and Fisheries for financial support through the project "Early post mortem measurement of WHC (water-holding capacity) and drip loss in fresh pork". The Danish Cereal Network (B4) and the Advanced Quality Monitoring program (FELFO) are also acknowledged for financial support. The Matlab programming was supported by Consensus Analysis AS. Gilda Kischinovsky is acknowledged for assistance with the manuscript. The relationship of EMSC and EISC to the following patents is acknowledged: E. Stark and H. Martens, "Multiplicative Signal Correction Method and Apparatus", U.S. Patent 5,568,400; and E. Stark and H. Martens, " Improved Multiplicative Signal Correction Method and Apparatus", European patent 0415401.

Received for review March 25, 2002. Accepted September 11, 2002.

AC020194W