CHEMOMETRICS

1. Introduction

Chemometrics (1–3) or more general multivariate regression methods (4,5) are applied in many research fields from social science to measurement techniques. There are two competing and equivalent nomenclature systems encountered in the chemometrics literature. The first, derived from the statistical literature, describes instrumentation and data in terms of "ways" that an analysis is performed. Here a "way" is constituted by each independent and nontrivial factor that is manipulated with the data collection system. Multiway techniques (the section Multiway Analysis) have been investigated and applied to hyphenated measurement techniques (6). For example, with excitation–emission matrix fluorescence spectra, three-way data are formed by manipulating the excitation-way, emission-way, and the sample-way. Implicit in this definition is a fully blocked experimental design where the collected data forms a cube with no missing values. Equivalently, a second nomenclature is derived from the mathematical literature where data are often referred to in terms of "orders". In tensor notation (7) a scalar is a zero-order tensor, a vector is first order, a matrix is second order, a cube is third order, etc. Hence, the collection of excitation–emission matrix fluorescence data would form a third-order tensor. However, it should be mentioned that the "way" and "order" based nomenclature are not directly interchangeable. By convention, "order" notation is based on the structure of the data collected from each sample. Analysis of collected excitation-emission fluorescence, forming a second-order tensor of data per sample, is referred to as second-order analysis compared to three-way analysis. In this work the "way" based notation will be adopted.

Although there is a vast area of chemometric applications, analytical chemistry has been chosen to exemplify the principles. Especially optical measurement techniques are usually multivariate and are appropriate for a descriptive

discussion of chemometrics. The first question that arises when introducing chemometrics is What is chemometrics? Simply put, chemometrics is the application of mathematical and statistical methods to the analysis of chemical data. However, it should be stressed that chemometrics is more than a subdiscipline of mathematics or statistics. The key to artfully practicing chemometrics is to extend the limitations of classical mathematics and statistics by understanding, and relying upon, the constraints that chemistry places on possible solutions to a statistically posed question. As Wold noted the impact of chemometrics is in problem solving not data analysis: chemometricians "must remain chemists and adapt statistics to chemistry instead of vice versa (8)." Along these lines Booksh and Kowalski (9) and Brown (10) define chemometrics more as an information science that can be applied to many physical science disciplines. Chemometrics is a truly interdisciplinary science that does draw from mathematics, statistics, and information science; however, the tools from these disciplines cannot be directly applied without sound knowledge and understanding of the chemical system in question. Many statistical tools are useless to chemometricians because the underlying assumptions are violated in the chemical system. Concurrently, a chemometrician could develop very useful 'statistical tools' that cannot be generalized beyond the chemical system in question. Furthermore, the distinction between "statistical significance" and "practical significance" cannot be reliably made without an understanding of both statistics and chemistry. A broad overview of chemometric techniques—without going into depth—has been published recently (11).

From a chemometric standpoint, data and instrumentation can be classified based on the dimensionality of the data set obtained. Instrumentation can be designed to generate a single datum of information per sample analyzed, an ordered vector of data per sample analyzed, or a linked matrix of data per sample analyzed. In general, the higher the dimensionality or number of ways of the data set, the more powerful the instrument. And, consequently, more powerful data analysis methods can be applied to higher directional data set. The different ways of data are presented in Figure 1. A more complete discussion on the interrelationship between data structure and analyzability can be found in references 9 and 12.

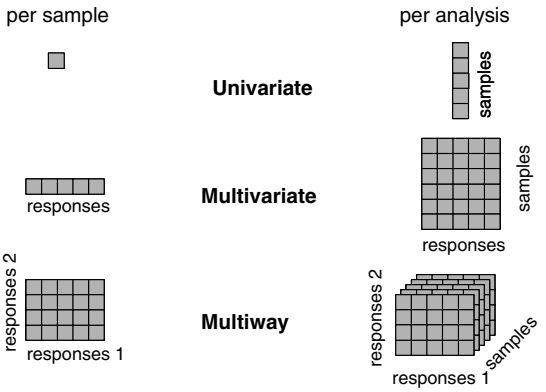

Fig. 1. Matrix representation of data structure for three classes of data.

The most basic type of data is univariate data. Examples of univariate are data collected from a pH meter or single-channel photometer. Univariate data is the lowest dimensionality of data—a univariate instrument returns a zero-dimensional (zero-order) data tensor. Consequently, a collection of data from a univariate sensor forms a vector and is said to be one-way data. That is it varies in only one way: sample-to-sample.

The majority of literature in chemometrics addresses analysis of multivariate data. Examples of multivariate data include chromatograms or spectra. Analysis of a single sample with a multivariate instrument yields a one direction (first-order) data tensor or vector. A collection of samples forms a two directional matrix and is said to be two-way data because it varies from sample to sample and wavelength to wavelength.

Multiway data are formed, eg, by hyphenated instrumentation such as gas chromatography–mass spectrometry (gc–ms) and excitation–emission matrix spectrometers. Analysis of a single sample yields a two-dimensional (2D)(second-order) tensor (matrix) of data. The key to having true multiway data is that one instrument (or order) must modulate the other instrument (or order). For example ultraviolet–visible–infrared (uv–vis–ir) is not multiway data because the uv–vis and ir spectra of a molecule do not modulate each other. A collection of sensor readings during the progression of a batch process form multiway data (sensors by time). There is no upper limit on the number of "ways" that could go into a data set. Conceivably, one could employ an online high performence liquid chrometography (HPLC)–uv–vis spectrometer to monitor a series of batch processes. A four-way data tensor results: wavelength by chromatographic retention time by time in the batch by batch.

2. Linear Regression Analysis

2.1. Notation and Fundamental Mathematical Tools. *Orthogonal matrices* have the following property: $\mathbf{U}^{-1}{}_{(K\times K)} = \mathbf{U}^{\mathrm{T}}$ or $\mathbf{U}\cdot\mathbf{U}^{\mathrm{T}} = \mathbf{1}_{(K\times K)}$. For rectangular $\mathbf{U}_{(N\times K)}$ matrices consisting of orthonormal rows or columns, a similar property holds:

x	Scalar—upper case italics represent fixed values, ie, J samples; lower case italics represent variables, ie, the jth sample.
$\mathbf{x}$	Column vector.
$\|\|\mathbf{x}\|\|_2$	2-Norm of a vector, ie, it's Euclidian length.
$\mathbf{X}_{(N\times K)}$	Matrix with N rows and K columns.
$\mathbf{x}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}$	Transposed vector (row vector), transposed matrix.
$\mathbf{X}^{-1}, \mathbf{X}^{+}$	Inverse matrix (if existent), pseudoinverse or Moore-Penrose pseudoinverse (4,13) for pseudoinverting rank-deficient or rectangular matrices (see below).
$\mathbf{X}$	Third-order tensors—a character with a subscript (or subscripts) is assumed to be the appropriate elements from a higher dimensional data matrix. For example, $\mathbf{X}_k$ is the kth slice of the tensor $\mathbf{X}$.
$\hat{\mathbf{x}}$	Estimate of the true value $\mathbf{x}$.

$$\mathbf{U}\cdot\mathbf{U}^{\mathrm{T}} = \mathbf{1}_{(N\times N)} \quad \text{and} \quad \mathbf{U}^{\mathrm{T}}\cdot\mathbf{U} = \mathbf{1}_{(K\times K)} \tag{1}$$

A *singular value decomposition* (SVD) (13,14) of a matrix

$$\mathbf{X}_{(K\times N)} = \mathbf{U}_{(K\times K)} \cdot \mathbf{S}_{(K\times K)} \mathbf{P}^{\mathbf{T}}_{(K\times N)} \tag{2}$$

factorizes an arbitrary matrix $\mathbf{X}$ into two orthogonal matrices $\mathbf{U}$ and $\mathbf{P}$ as well as a diagonal matrix $\mathbf{S} = \mathrm{diag}(s_1 \ldots s_K)$. Usually, the singular values s_k are order decreasingly; the columns of $\mathbf{U}$ and the rows of $\mathbf{P}^{\mathbf{T}}$ are ordered accordingly. If $\mathbf{X}$ is rank-deficient, ie, the rank is $R < \min(N,K)$, only R singular values are unequal to zero $s_R < \min(N,K) \neq 0$. Hence, the matrices on the right-hand side of (eq. 2) can be a downsized without loss of information to

$$\mathbf{X}_{(K\times N)} = \mathbf{U}_{(K\times R)} \cdot \mathbf{S}_{(R\times R)} \cdot \mathbf{P}^{\mathbf{T}}_{(R\times N)} \tag{3}$$

A *matrix inversion* based on SVD is known to be numerically very stable (14). A full-rank square matrix is inverted utilizing the properties of orthogonal matrices:

$$\mathbf{X}^{-1}_{(K\times K)} = \left[\mathbf{U}_{(K\times K)} \cdot \mathbf{S}_{(K\times K)} \cdot \mathbf{P}^{\mathbf{T}}_{(K\times K)}\right]^{-1} = \mathbf{P} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^{\mathbf{T}} \tag{4}$$

For rectangular or singular $\mathbf{X}$ the *Moore-Penrose pseudoinverse* (4,13) has been defined. It combines the ideas of equations (3 and 4): A SVD of X (eq. 2) is calculated followed by downsizing of the three matrices to R, the number of nonzero singular values (eq. 3). This enables the inversion of S, then equation 4 is formally applied

$$\mathbf{X}^{+}_{(N\times K)} = \left[\mathbf{U}_{(K\times R)} \cdot \mathbf{S}_{(R\times R)} \cdot \mathbf{P}^{\mathbf{T}}_{(R\times K)}\right]^{-1} = \mathbf{P} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^{\mathbf{T}} \tag{5}$$

2.2. Univariate Regression. Two different types of variables are used in regression analysis: predictor or x variables and response or y variables (4). The x variables can be observed but not controlled; values of the y variables are determined by the values of the corresponding x variables. The simplest system is a linear univariate system, which relates the predictor variable x via a proportionality constant a directly to the response variable y, the target figure:

$$y = a \cdot x \tag{6}$$

In spectroscopy, eg, the predictor variable would be the absorption A of a sample at a certain wavelength, the concentration c of a certain chemical in the sample plays the role of the response variable. According to Beer's law, the absorption

$$A = (L \cdot e) \cdot c \tag{7}$$

is directly proportional to the concentration. In this case, the proportionality constant is the inverted product of absorption path length L and the chemical's molar extinction coefficient e at the considered wavelength. Since A is the measur and acquired to determine c the following equation is the analogue to eq. 6:

$$c = \frac{1}{L \cdot e} \cdot A$$

In order to incorporate a constant offset, eg, a constant background absorption A_0, the model (eq. 6) is extended to

$$y = a_0 + a_1 \cdot x \tag{8}$$

or, respectively,

$$c = A_0 + \frac{1}{L \cdot e} \cdot A \tag{9}$$

A model is *linear* in the regression sense, if it is linear in the model parameters a_0 and a_1. A polynomial model

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots + a_q \cdot x^q \tag{10}$$

is also linear in this meaning; $y = a_0 \cdot \exp(-a_1 \cdot x)$, eg, is not linear.

The model parameters a_0 and a_1 (eq. 8) are usually not known and must be determined experimentally by means of a calibration. One would prepare, eg, two samples with known concentrations ${c_1}^{\text{cal}}$ and ${c_2}^{\text{cal}}$ of the target analyte and would measure the absorbances ${A_1}^{\text{cal}}$ and ${A_2}^{\text{cal}}$ of both calibration samples at the chosen wavelength:

$$\begin{aligned} y_1^{\text{cal}} &= a_0 + a_1 \cdot x_1^{\text{cal}} \quad \text{or} \quad c_1^{\text{cal}} = A_0 + \frac{1}{L \cdot e} \cdot A_1^{\text{cal}} \\ y_2^{\text{cal}} &= a_0 + a_1 \cdot x_2^{\text{cal}} \quad \text{or} \quad c_2^{\text{cal}} = A_0 + \frac{1}{L \cdot e} \cdot A_2^{\text{cal}} \end{aligned} \tag{11}$$

These two equations set up an equation system with two unknowns $a_0 = A_0$ and $a_1 = 1/L \cdot e$, which can be solved. Now, unknown samples can be analyzed by measuring their absorption $x_{\text{meas}} = A_{\text{meas}}$. a_0 and a_1 are used then in equations 8 and 9, respectively, for determining concentration $y_{\text{meas}} = c_{\text{meas}}$. Unfortunately, there are always measurement errors ε disturbing the true model (eq. 8). Instead of the undisturbed model (eq. 8) one has to deal with:

$$y = a_0 + a_1 \cdot x + \epsilon \tag{12}$$

Since every measurement is affected by a different and unpredictable error ε including $K > 2$ calibration samples

$$(x_1^{\text{cal}}, y_1^{\text{cal}}), \ldots, (x_K^{\text{cal}}, y_K^{\text{cal}}) \tag{13}$$

does not solve this problem—there are always more unknowns than equations. Hence, a_0 and a_1 cannot be derived from the correct model (eq. 8) by solving a set of calibration equations (eq. 11).

To overcome this problem and to get a workable solution, one has to accept errors in the model parameters and estimate them by means of a least-squares fit. The estimated parameters are denoted by $\hat{a_0}$ and $\hat{a_1}$. As will be discussed

below, $\hat{a}_0$ and $\hat{a}_1$ will be determined from error affected calibration data in such a way, that they are a good compromise explaining the calibration set (eq. 13) as accurate as possible. The conditions for good estimates will be discussed in more detail in the section Statistical Background of Regression Analysis. The idea behind linear least-squares regression is to determine estimators $\hat{a}_0$ und $\hat{a}_1$ from a calibration set (eq. 13) such that the sum of squared errors S is minimized:

$$S = \sum_{i=1}^{K} \epsilon_i^2 = \sum_{i=1}^{K} \left(y_i^{\text{cal}} - [\hat{a}_0 + \hat{a}_1 \cdot x_i^{\text{cal}}]\right)^2 \tag{14}$$

In order to derive the minimum of $S(\hat{a}_0,\hat{a}_1)$, partial derivatives of equation 14 with respect to $\hat{a}_0$ und $\hat{a}_1$ are calculated and set to zero. In theory, this results in an extremum, which could also be a maximum, however, $\hat{a}_0$ und $\hat{a}_1$ can be chosen so out of the way that S can reach basically any value. For all practical purposes a minimum of S is obtained.

$$\begin{aligned} \frac{\partial S}{\partial \hat{a}_0} &= 2 \cdot \sum_{i=1}^{K} (-1)(y_i^{\text{cal}} - [\hat{a}_0 + \hat{a}_1 \cdot x_i^{\text{cal}}]) = 0 \\ \frac{\partial S}{\partial \hat{a}_1} &= 2 \cdot \sum_{i=1}^{K} (-x_i^{\text{cal}}) \cdot (y_i^{\text{cal}} - [\hat{a}_0 + \hat{a}_1 \cdot x_i^{\text{cal}}]) = 0 \end{aligned} \tag{15}$$

Die addends containing the response variable y are transferred to the right side of the equation system (eq. 15). Also the matrix notation is used from here on

$$\begin{pmatrix} \sum_{i=1}^{K} 1 & \sum_{i=1}^{K} x_i^{\text{cal}} \\ \sum_{i=1}^{K} x_i^{\text{cal}} & \sum_{i=1}^{K} x_i^{\text{cal}} \cdot x_i^{\text{cal}} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{K} 1 \cdot y_i^{\text{cal}} \\ \sum_{i=1}^{K} x_i^{\text{cal}} \cdot y_i^{\text{cal}} \end{pmatrix} \tag{16}$$

Equation system 16 is solved resulting is the estimates $\hat{a}_0$ and $\hat{a}_1$ of the true model parameters a_0 and a_1. Now, $\hat{a}_0$ and $\hat{a}_1$ are used in equation 8 instead of a_0 and a_1 for evaluating unknown samples [see paragraph after equation 11]:

$$y_{\text{meas}} = \hat{a}_0 + \hat{a}_1 \cdot x_{\text{meas}} \tag{17}$$

This procedure can easily be extended to handle polynomial models (eq. 10).

2.3. Figures of Merit for Univariate Chemical Analysis. One essential task in chemometrics is comparing quantitatively the performance of multiple types of chemical sensors or of multiple options for calibration models. In order to objectively perform these comparisons, it is useful to have quantifiable criteria on which discussions are based. Examples of such quantifiable criteria include speed, cost, reliability, precision, sensitivity, selectivity, and detection limit of analysis (15). While speed, cost, and reliability weigh heavily in pragmatic decisions of which instrumental technique to employ for a particular application, these figures of merit are not intrinsic to a given instrumental method. For example, the cost and speed of analysis largely depends on the number of

samples to be analyzed; with large quantities of samples economic and time savings can be achieved through bulk purchasing and automation. However, the final four figures of merit are intrinsic to the application of an instrumental method and are directly related to the instrumental response for a particular set of analytes.

The selectivity is the fraction of the instrumental signal that is unique to the analyte. Assuming the instrument is "zeroed" to remove any baseline,

$$\mathrm{SEL} = \frac{r_a}{r} \tag{18}$$

where r_a is the instrumental signal of just the analyte and r is the instrumental signal of the sample. The SEL is a value between 0 and 1 with SEL = 1 being a fully selective sensor. For univariate calibration, an instrument must be fully selective. Otherwise, a bias will be imbedded in the prediction of future samples. There is no way, based only on statistical analysis of collected data, to determine the contribution or existence of nonselective interferents in any given 'unknown' sample.

The sensitivity is the change in instrumental response r with respect to changes in analyte concentration

$$\mathrm{SEN} = \frac{\partial r}{\partial c}$$

For univariate linear calibration, this is the slope of the calibration curve, ie, the constant a_1 in equation 8. The precision of a method is best expressed in the signal to noise ratio (S/N),

$$\mathrm{S/N} = \frac{r_a}{e}$$

where e is a measure of the reproducibility of replicated measurements. In many cases, the measurement reproducibility is not concentration dependant, eg, if thermal noise limited analyses. In this case S/N will not vary with analyte concentration.

The limit of detection (LOD) is defined by the International Union of Pure and Applied Chemists and the American Chemical Society to be the smallest amount of a chemical that can be reasonably detected by a given analytical method (16). Of the many ways to calculate the LOD, determine $r_\mathrm{a} = y_\mathrm{meas} = c_\mathrm{meas}$ for given a signal, x_meas (eq. 17), equal to three standard deviations of replicated instrumental blanks is the most straightforward. Assuming the calibration model is linear, univariate, and free of instrumental offset [$a_0 = 0$ (eq. 8)], the detection limit can be expressed as

$$\mathrm{LOD} = \frac{3 \cdot e}{a_1} = \frac{3 \cdot e}{\mathrm{SEN}} \tag{19}$$

where a_1 (eq. 8) is the slope of the calibration (16).

2.4. Multivariate Linear Regression (MLR). *Calibration.* If the response variable y is linearly dependent on several predictor variables the univariate approach (see section Univariate Regression) has to be extended to the multivariate method named multivariate least-squares regression or multilinear regression (4,17). In the following, only two predictor variables $x_{(1)}$ and $x_{(2)}$ are used to keep the discussion concise, but the concept is easily extended to more x variables. Equation 12 is replaced in a multivariate case by

$$y = a_0 + a_1 \cdot x_{(1)} + a_2 \cdot x_{(2)} + \epsilon \tag{20}$$

In case of two predictor variables, a fit plane is determined by the calibration—in higher dimensional applications a hyperplane is obtained.

As an example, the concentration of a chemical might be additionally temperature dependent, ie, an experimenter has to measure the absorbance $A = x_{(1)}$ at a certain wavelength and the temperature $T = x_{(2)}$ of a sample. From these two reading he/she can calculate the concentration $c = y$. For this purpose, three model parameters have to be estimated: $a_0 = A_0$ the background absorption, $a_1 = 1/L{\cdot}e$ [see discussion after equation 11], and a temperature coefficient $a_2 = 1/\tau$. For this estimation $K \geq 3$ calibration samples $({x^{\text{cal}}}_{(1)1}, {x^{\text{cal}}}_{(2)1}, {y^{\text{cal}}}_1), \ldots,$ $({x^{\text{cal}}}_{(1)K}, {x^{\text{cal}}}_{(2)K}, {y^{\text{cal}}}_K)$ have to be provided. They are obtained in this example from measuring the absorption A and the temperature T of samples with known chemical concentration.

To estimate the three model parameters in

$$y^{\text{cal}} = \hat{a}_0 + \hat{a}_1 \cdot x^{\text{cal}}_{(1)} + \hat{a}_2 \cdot x^{\text{cal}}_{(2)} \tag{21}$$

the sum of squared errors [cf. equation 14]

$$S = \sum_{i=1}^{K} \epsilon_i^2 = \sum_{i=1}^{K} \left(y_i^{\text{cal}} - \left[\hat{a}_0 + \hat{a}_1 \cdot x^{\text{cal}}_{(1)i} + \hat{a}_2 \cdot x^{\text{cal}}_{(2)i} \right] \right)^2$$

is minimized. For this purpose the three partial derivatives are calculated and set to zero equivalent to the univariate case (see section Univariate Regression). This results in the following equation system:

$$\begin{pmatrix} \sum_{i=1}^{K} 1 & \sum_{i=1}^{K} x^{\text{cal}}_{(1)i} & \sum_{i=1}^{K} x^{\text{cal}}_{(2)i} \\ \sum_{i=1}^{K} x^{\text{cal}}_{(1)i} & \sum_{i=1}^{K} x^{\text{cal}}_{(1)i} \cdot x^{\text{cal}}_{(1)i} & \sum_{i=1}^{K} x^{\text{cal}}_{(1)i} \cdot x^{\text{cal}}_{(2)i} \\ \sum_{i=1}^{K} x^{\text{cal}}_{(2)i} & \sum_{i=1}^{K} x^{\text{cal}}_{(2)i} \cdot x^{\text{cal}}_{(1)i} & \sum_{i=1}^{K} x^{\text{cal}}_{(2)i} (2) i^{\text{cal}} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{K} \mathbf{1} \cdot y_i^{\text{cal}} \\ \sum_{i=1}^{K} x^{\text{cal}}_{(1)i} \cdot y_i^{\text{cal}} \\ \sum_{i=1}^{K} x^{\text{cal}}_{(2)i} \cdot y_i^{\text{cal}} \end{pmatrix} \tag{22}$$

This equation can be written in a more compact notation allowing easier operations on the data in the remainder. Therefore, the following vectors and matrices are defined

$$\begin{array}{lll} \mathbf{y}^{\text{cal}} = (y_1^{\text{cal}} \cdots y_K^{\text{cal}})^{\mathbf{T}} & \mathbf{1} = (1 \cdots 1)^{\mathbf{T}} & \mathbf{a} = (a_0 \; a_1 \; a_2)^{\mathbf{T}} \\ \mathbf{x}^{\text{cal}}_{(1)} = \left(x^{\text{cal}}_{(1)t} \cdots x^{\text{cal}}_{(1)K} \right)^{\mathbf{T}} & \mathbf{x}^{\text{cal}}_{(2)} = \left(x^{\text{cal}}_{(2)1} \cdots x^{\text{cal}}_{(2)K} \right)^{\mathbf{T}} & \mathbf{X}^{\text{cal}}_{(K \times 3)} = \left[\mathbf{1} \; \mathbf{x}^{\text{cal}}_{(1)} \; \mathbf{x}^{\text{cal}}_{(2)} \right] \end{array} \tag{23}$$

By using these definitions, equations 20 and 22 can be rewritten in matrix notation

$$\mathbf{y}^{\text{cal}} = \mathbf{X}^{\text{cal}} \cdot \mathbf{a} + \epsilon \tag{24}$$

$$\mathbf{X}^{\mathbf{T}\,\text{cal}} \cdot \mathbf{X}^{\text{cal}} \cdot \hat{\mathbf{a}} = \mathbf{X}^{\mathbf{T}\,\text{cal}} \cdot \mathbf{y}^{\text{cal}} \tag{25}$$

The estimate of the fit parameters $\hat{\mathbf{a}}$ can finally be determined by multiplying the inverse of the covariance matrix from the left:

$$\hat{\mathbf{a}} = (\mathbf{X}^{\mathbf{T}\,\text{cal}} \cdot \mathbf{X}^{\text{cal}})^{-1} \cdot \mathbf{X}^{\mathbf{T}\,\text{cal}} \cdot \mathbf{y}^{\text{cal}} \tag{26}$$

Equation 25 is known as the normal equation of a least-square problem. If the covariance matrix is singular, equation 26 has to employ the Moore-Penrose pseudoinverse [see the section Notation and Fundamental Mathematical Tools, equation 5 and the following Supplementary Topics section]:

$$\hat{\mathbf{a}} = \mathbf{X}^{+\,\text{cal}} \cdot \mathbf{y}^{\text{cal}} \tag{27}$$

Prediction. A response variable $y_{\text{meas}} = c_{\text{meas}}$ of an unknown sample, a concentration value, eg, can be predicted from an unknown data set $\mathbf{x}_{\text{meas}} = (1\ x_{(1)} = A\ x_{(2)} = T)^{\mathrm{T}}{}_{\text{meas}}$ comprising in the given example measured values for the absorbance A and temperature T by

$$y_{\text{meas}} = \hat{\mathbf{a}}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} \tag{28}$$

An offset free model $a_0 = 0$ (eq. 20) can always be obtained by mean centering (see also section Data Pretreatment—Mean Centering and Scaling).

For this purpose, the first row of equation 22 is rewritten

$$k \cdot \hat{a}_0 + k \cdot \hat{a}_1 \cdot \bar{x}_{(1)}^{\text{cal}} + k \cdot \hat{a}_2 \cdot \bar{x}_{(2)}^{\text{cal}} = k \cdot \bar{y}^{\,\text{cal}}$$

The bar on top of the variables indicates mean values. Dividing this equation by k and solving for $\hat{a_0}$ results in

$$\hat{a}_0 = \bar{y}^{\,\text{cal}} - \hat{a}_1 \cdot \bar{x}_{(1)}^{\text{cal}} - \hat{a}_2 \cdot \bar{x}_{(2)}^{\text{cal}}$$

This equation for $\hat{a_0}$ is used in equation 21

$$y^{\text{cal}} - \bar{y}^{\,\text{cal}} = \hat{a}_1 \cdot \left(x_{(1)}^{\text{cal}} - \bar{x}_{(1)}^{\text{cal}}\right) + \hat{a}_2 \cdot \left(x_{(2)}^{\text{cal}} - \bar{x}_{(2)}^{\text{cal}}\right)$$

This mean centered model is reduced by one parameter, and hence one degree of freedom. Now, the whole multivariate least-squares procedure described above is performed on predictor and response variable subtracted by their mean values. If mean centering is applied, equation 28 must also incorporate mean centering

$$y_{\text{meas}} = \hat{\mathbf{a}}^{\mathbf{T}} \cdot (\mathbf{x}_{\text{meas}} - \bar{\mathbf{x}}^{\text{cal}}) + \bar{y}^{\,\text{cal}}$$

For deriving equation 24 only two predictor variables and a bias were assumed. If N predictor variables have to be included into the calibration model, $\mathbf{X}^{\text{cal}}$ must be augmented by additional columns containing the appropriate calibration values, eg, absorbances at several wavelength positions.

If M response variables, concentrations of several chemicals, eg, have to be determined, $\mathbf{y}^{\text{cal}}$ and $\mathbf{a}$ (eq. 24) have to be augmented by one column per response variable:

$$\begin{aligned} \lfloor \mathbf{y}_1^{\text{cal}} \cdots \mathbf{y}_M^{\text{cal}} \rfloor &= \mathbf{X}^{\text{cal}} \cdot [\mathbf{a}_1 \cdots \mathbf{a}_M] + \epsilon \\ \mathbf{Y}^{\text{cal}}_{(K\times M)} &= \mathbf{X}^{\text{cal}}_{(K\times N)} \cdot \mathbf{A}_{(N\times M)} + \epsilon \end{aligned} \tag{29}$$

Instead of estimating a model vector $\mathbf{a}$ by means of equation 26, a model matrix $\mathbf{A}$ is estimated by

$$\hat{\mathbf{A}} = (\mathbf{X}^{\text{cal}^{\mathbf{T}}} \cdot \mathbf{X}^{\text{cal}})^{-1} \cdot \mathbf{X}^{\text{cal}^{\mathbf{T}}} \cdot \mathbf{Y}^{\text{cal}} \tag{30}$$

Unknown predictor variables or measurement vectors $\mathbf{x}^{\text{meas}}$ are evaluated then by

$$\begin{pmatrix} y_1^{\text{meas}} \\ \vdots \\ y_M^{\text{meas}} \end{pmatrix} = \hat{\mathbf{A}}^{\mathbf{T}} \cdot \mathbf{x}^{\text{meas}} \tag{31}$$

Supplementary Topics

1. The concept of pseudoinverses (the section Notation and Fundamental Mathematical Tools) is closely related to MLR applications (eq. 27). For the following short discussion, $\mathbf{X}^{\text{cal}}$ is assumed to have rank $R < \min(K, N)$. Hence, the covariance matrix in equation 26 cannot be inverted. In equation 32, the "inversion" of a rectangular matrix $\mathbf{P}$, which consists of orthonormal columns, is performed in the sense of equation 1. The reader has to keep in mind, that the following is done in a descriptive way without being mathematically thorough.

$$\begin{aligned} (\mathbf{X}^{\mathbf{T}\,\text{cal}} \cdot \mathbf{X}^{\text{cal}})^{-1} \cdot \mathbf{X}^{\mathbf{T}\,\text{cal}} &= \left(\mathbf{P} \cdot \mathbf{S} \cdot \underbrace{\mathbf{U}^{\mathbf{T}} \cdot \mathbf{U}}_{=1} \cdot \mathbf{S} \cdot \mathbf{P}^{\mathbf{T}} \right)^{-1} \mathbf{P} \cdot \mathbf{S} \cdot \mathbf{U}^{\mathbf{T}} \\ &= \mathbf{P} \cdot \mathbf{S}^{-2} \cdot \mathbf{P}^{\mathbf{T}} \cdot \mathbf{P} \cdot \mathbf{S} \cdot \mathbf{U}^{\mathbf{T}} \\ &= \mathbf{P} \cdot \mathbf{S}^{-1} \cdot \mathbf{U}^{\mathbf{T}} \\ &= \mathbf{X}^{+\,\text{cal}} \end{aligned} \tag{32}$$

2. The computation of the pseudoinverse $\mathbf{X}^{+\text{cal}}$ in equation 27 involves a SVD of $\mathbf{X}^{\text{cal}}$ (eq. 5) (section Notation and Fundamental Mathematical Tools), which can be computed in a very reliable way. However, the SVD algorithm takes a lot of computation power: The number of executed floating point

operations (flops), ie, additions and multiplications, of a widely used (14) SVD algorithm is given in Ref. 13 to be

$$\text{flops}\left\{\text{SVD}(\mathbf{X}^{\text{cal}}_{(K\times N)})\right\} = 14\cdot K\cdot N^2 + 8\cdot N^3 \tag{33}$$

According to equation 33, it is evident, that it takes fewer flops to decompose the transposed matrix $\mathbf{X}^{\text{cal}^{\text{T}}}$ than $\mathbf{X}^{\text{cal}}$, whenever $N > K$ since

$$\begin{aligned}\text{flops}\left\{\text{SVD}(\mathbf{X}^{\text{cal T}}_{(N\times K)})\right\} &= 14\cdot N\cdot K^2 + 8\cdot K^3 < 14\cdot K\cdot N^2 + 8\cdot N^3\\ &= \text{flops}\left\{\text{SVD}(\mathbf{X}^{\text{cal}}_{(K\times N)})\right\}\end{aligned} \tag{34}$$

After transposing $\mathbf{X}^{\text{cal}^{\text{T}}}$, one obtains the same matrices from the SVD but in reversed order and transposed. Rearranging and transposing them to get the correct pseudoinverse $\mathbf{X}^{+\text{cal}}$ is in almost all cases much faster than decomposing the original matrix $\mathbf{X}^{\text{cal}}$.

3. Another important property of $\hat{\mathbf{a}}$ is that it is determined in such a way that the vectors $\mathbf{X}\cdot\hat{\mathbf{a}}$ and ε, ie, the residuum, are orthogonal to each other

$$\begin{aligned}(\mathbf{X}\cdot\hat{\mathbf{a}})^{\mathbf{T}}\cdot\epsilon &= (\mathbf{X}\cdot\hat{\mathbf{a}})^{\mathbf{T}}\cdot(\mathbf{y}-\mathbf{X}\cdot\hat{\mathbf{a}})\\ &= \hat{\mathbf{a}}^{\mathbf{T}}\cdot\mathbf{X}^{\mathbf{T}}\cdot(\mathbf{y}-\mathbf{X}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}\cdot\mathbf{y})\\ &= \hat{\mathbf{a}}^{\mathbf{T}}\cdot\left(\underbrace{\mathbf{X}^{\mathbf{T}}-\mathbf{X}^{\mathbf{T}}\cdot\mathbf{X}\cdot(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathbf{T}}}_{=\mathbf{0}}\right)\cdot\mathbf{y}\\ &= 0\end{aligned}$$

This general fact will be of importance for PLS (the section Partial Least Squares).

2.5. Data Pretreatment—Mean Centering and Scaling. The success of multivariate and multiway data analysis often depends on the application of data pretreatment to remove, scale, or standardize the sources of observed variance. The methods described in this section are applicable to univariate, multivariate, and multiway data analysis strategies. Like all tools, the use and power of each preprocessing methods should be understood before it is applied. Pretreating data, if done properly, can bring out desired information. Likewise, pretreating data, if done improperly, can obscure any desired information embedded in the data.

"Mean centering" and "variance scaling" (18) are often performed on multivariate data without much thought to the consequences of these actions. Mean centering removes the average, or mean, response of a given variable or sample. This translates the variance of the data set to be centered about the ordinate axis. Variance scaling normalizes each variable, or sample, such that the data's variance becomes unity. This places the data on a unit sphere. When mean centering and variance scaling are both applied to a collection of data, the data is said to be 'auto scaled.' Auto scaling places the data on a unit sphere centered about the origin of the multivariate space of the data.

There are specific instances when mean centering and variance scaling should and should not be applied to a data set. In general, mean centering

aids in interpretation of factor analysis (FA) models and construction of calibrations. By removing the mean of the data set, often one less factor is required for analysis. An exception may occur when the data is collected under 'closure' (19,20). Closure exists when the sum of the variables or concentrations is constrained to equal a preset value. The most common type of closure is seen in mixture analysis when the sum of percent composition of all detectable species is constrained to equal 100%. Other examples may occur when improper experimental designs are employed. When closure exists, mean centering will not always eliminate a factor. In these instances, the errors introduced by estimating the mean of the data set are not offset by the gains associated with a simpler model.

Mean centering is applied by subtracting the mean spectrum of the data set from every spectrum in the data set. For a data set $\mathbf{R}_{(I\times \mathrm{J})}$ of I samples, each of J predictor variables like discrete digitized wavelengths, the mean centered jth wavelength of the ith sample is defined by

$$^{mc}R_{ij} = R_{ij} - \left(\sum_{j=1}^{J} R_{ij}/J\right) \tag{35}$$

In a multivariate sense, this preprocessing method translates the collection of data to the origin of the multivariate space, where analysis will be performed. The practical consequence of mean centering data is often a more simple and interpretable regression model. In effect, mean centering removes the need for an intercept from the regression model. Consequently, since fewer terms in the regression model may need to be estimated, estimated analyte concentrations may be more precise following mean centering the data. It should be noted that mean centering does not always yield the most precise calibration model. Each calibration method should be tested on mean centered and nonmean centered data.

The effect of mean centering is demonstrated in Figure 2. Figure 2**a** presents raw NIR spectra of the 40 cornflour samples, while Figure 2**b** presents the mean-centered spectra. Although the spectra do not appear to be visually interpretable, none of the variance within the data set has been altered. The major effect of mean centering is removing the broad sloping background from the data collection. The effect of mean centering on principal component based models is shown in the cartoons of Figure 2**c** and **d**. The data cloud in the upper right corner of Figure 2**c** is translated to the origin of the J dimensional space. The arrows of Figure 2**c** and **d** present the direction of greatest variance from the origin. For the nonmean centered data, the direction of greatest variance is the mean of the spectra. With mean centered data, the direction of greatest variance is now the direction of greatest variance *within* the data set. Consequently, more of the information content of a data set can usually be described with a simpler model if the data is mean centered.

When a data set is variance scaled all variables, or samples, are given equal weight in determining the factors of the model. This may be beneficial when variables with small variance have greater predictive variance than variables with larger variance. A prime example is seen in fusing data measurements with

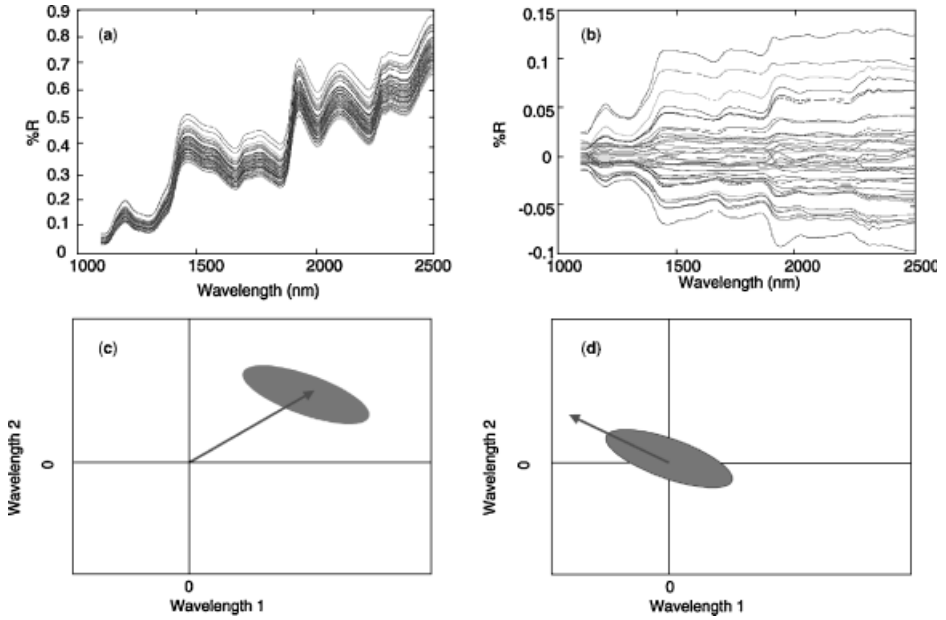

Fig. 2. In this figure the effect of mean centering the data set can be seen. (**a**) The raw data for 40 NIR spectra from corn flour samples. (**b**) The same spectra after mean centering. (**c**) Pictorial representation of a data cloud in two-dimensional space without mean centering. The arrow represents the first PC. (**d**) Pictorial representation of the same data cloud after mean centering; note that the data cloud becomes centered on the ordinate axis. The arrow represents the first PC.

drastically different scales (ie, physical measurements like temperature and pressure with spectroscopic data). However, in most spectroscopic, chromatographic, or electrochemical analyses, the measurement is chosen such to be most sensitive to the analyte of interest. Here, it would not be favorable to give equal weight to background noise in uninformative measurements as is given to measurements with maximum analyte sensitivity.

Variance scaling is applied to the *j*th wavelength of every spectrum by division of the standard deviation of the *j*th wavelength over all spectra in the calibration set. Thus, by variance scaling, the impact each variable has in determining the parameters of the calibration model is equalized. Variance scaling is best employed when the variance of a particular wavelength has no correlation to the useful information content of that particular wavelength. Variance scaled data gives equal weight to all wavelengths, regardless of whether they represent a vibrational overtone, scattering, or just baseline noise. Consequently, variance scaling is seldom beneficial for spectroscopic calibration. However, in instances where the analytically useful signal is very weak compared to other signals, variable scaling can be essential.

The effect of variance scaling and auto scaling are shown in Figure 3 for the cornflour spectra. Variance scaling of the corn flour spectra have little superficial effect seen in Figure 3**a**. However, a close inspection would show that the spread of data at each wavelength is much more uniform across the spectra. This is more

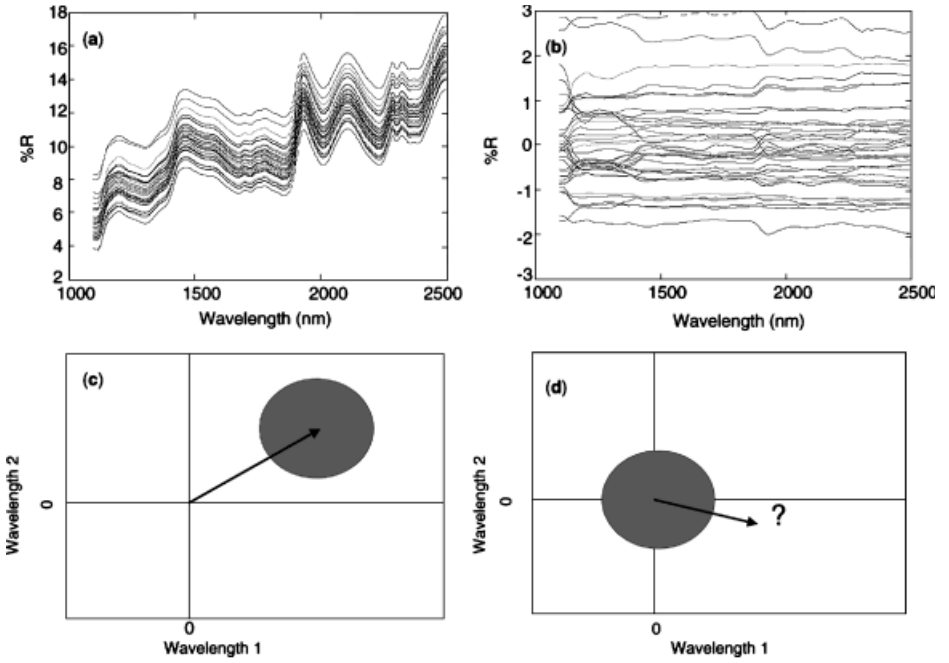

Fig. 3. The same data as in Figure 2 following (**a**) variance scaling and (**b**) auto scaling. Pictorial representations of a data cloud following (**c**) variance scaling and (**d**) auto scaling. The arrow represents the first PC. For variance scaling the data is transformed to lie on a sphere with the same variance in each direction. For auto scaled data, the sphere is translated to the origin of the coordinate axes.

readily seen in Figure 3**b**, where variance scaling is applied to the mean-centered spectra. Thus auto scaling the data. Figure 3**c** shows that variance scaling transforms the oblong data cloud into a unit circle. The direction of greatest variance from the origin is still the mean spectrum. Auto scaling translates the unit circle to the origin of the data space. The direction of greatest variance is now determined by the internal variance of the data with each wavelength having equal weight (regardless of the original magnitude of internal variance).

A third type of scaling often employed is scaling each variable or sample to unit area. This scaling is successfully applied to samples when matrix or sampling effects alter the measurement efficiency of a method. Examples include sample-to-sample variance due to sample thickness in reflectance spectroscopy and effective pathlength in other optical methods. Unit area normalization obscures the absolute concentrations of analytes but preserves the relative concentration of constituents between and among samples. Therefore, absolute calibration cannot be performed unless the calibration set is constrained by closure once the data is normalized.

2.6. Statistical Background of Regression Analysis. As was discussed in the section Univariate Regression, the true model parameters cannot be determined since the experimental calibration set is affected by measurement errors. Instead of the correct model parameters estimates have to be determined

and used for prediction of unknown data sets. The question discussed in this section is: How reliable are these estimates? It has been proven (5), that the expectation values of the least squares estimates are unbiased, ie, $E[\hat{a_i}] = a_i$, if the three Gauss-Markov conditions are fulfilled.

1. ie, the expectation value of all n measurement errors $E[\varepsilon_n] = 0\ \forall n$ (eq. 12) is zero. By means of this condition, it is guarantied that the assumed fit model is appropriate for the measured data. This condition prevents, eg, that a parabola is fitted to a cubic relationship between prediction and response variables.
2. $E[\varepsilon^2{}_n]$ is equal $\forall n$ measurement points ie, the error of the measurement data is independent from the values of predictor variable. This type of error is called homoscedastic. If errors are heteroscedastic, the least-squares fit would be more influenced by large predictor variable values than by small.
3. The errors in different measurements of the predictor variable(s) are uncorrelated $E[\varepsilon_n \cdot \varepsilon_m] = 0\ \forall n \neq m$, this means in the spectroscopic example that the measurement errors of the absorption at different wavelength positions are uncorrelated.

3. Bilinear Chemometric Methods

The reported successes of multivariate chemical analysis are based on three facts. (*1*) Most, if not all chemical processes are multivariate in nature. Consequently, to be able to effectively perform in a multivariate world, multivariate data must be collected and analyzed. (*2*) Even if only a single piece of information is needed from a chemical system it is very difficult to design a sensor that is fully selective to that property of interest. Therefore, to circumvent the lack of fully selective sensors, arrays of partially selective sensors can be constructed that rely on multivariate analysis methods to extract the information of interest. (*3*) There are inherent advantages associated with the redundancy of data when there are many more variables measured per sample than samples collected.

3.1. Classical Least Squares (CLS) versus Inverse Least Squares (ILS). This discussion on CLS versus ILS approaches is based on reference 21. Again a spectroscopic application was used in this discussion since physical meaningful objects help to understand the methods better. The difference between both techniques lies in the approach, which will be explained by means of the Beer's law (eq. 7). The absorbance spectra are written in a matrix $\mathbf{A}_{\text{cal}} = \mathbf{X}^{\text{cal}}$ (eq. 23), the concentrations in a matrix $\mathbf{C}_{\text{cal}} = \mathbf{Y}^{\text{cal}}$ (eq. 29). Replacing both items back to predictor and response variables enables transforming this discussion to applications other than spectroscopy.

The physics oriented CLS approach considers the measured spectra as products of molar extinction coefficients $\mathbf{K}$ (unit spectra) and concentrations $\mathbf{C}_{\text{cal}}$. The spectral errors are contained in $\mathbf{E}_{\text{A}}$:

$$\mathbf{A}_{\text{cal}} = \mathbf{C}_{\text{cal}} \cdot \mathbf{K} + \mathbf{E}_{\mathbf{A}} \tag{36}$$

The CLS calibration step estimates $\mathbf{K}$ by means of a multivariate least-squares procedure equivalent the one presented in the section Multivariate Linear Regression

$$\hat{\mathbf{K}} = (\mathbf{C}_{\text{cal}}^{\mathbf{T}}\mathbf{C}_{\text{cal}})^{-1} \cdot \mathbf{C}_{\text{cal}}^{\mathbf{T}} \cdot \mathbf{A}_{\text{cal}} \tag{37}$$

and evaluates unknown spectra by

$$\hat{\mathbf{c}}_{\text{meas}} = (\hat{\mathbf{K}} \cdot \hat{\mathbf{K}}^{\mathbf{T}})^{-1} \cdot \hat{\mathbf{K}} \cdot \mathbf{a}_{\text{meas}} \tag{38}$$

If $\mathbf{C}_{\text{cal}}$ or $\hat{\mathbf{K}}$ are singular the corresponding pseudoinverse ${\mathbf{C}^{+}}_{\text{cal}}$ or $\hat{\mathbf{K}}^{+}$ (eq. 5) (the section Notation and Fundamental Mathematical Tools) has to be used.

ILS uses a less intuitive calibration routine, which follows the introduction into multivariate least-squares fits as given in the section Multivariate Linear Regression:

$$\mathbf{C}_{\text{cal}} = \mathbf{A}_{\text{cal}} \cdot \mathbf{P}_{(N \times M)} + \mathbf{E}_{\text{C}}$$

In this approach, $\mathbf{P}$ contains calibration coefficients, which relate the spectral intensities to concentration of chemicals. The parameter $\mathbf{E}_{\text{c}}$ contains random concentration errors. This regression matrix $\mathbf{P}$ is purely a mathematical construct and has no physical meaning. A calibration step estimates

$$\hat{\mathbf{P}} = (\mathbf{A}_{\text{cal}}^{\mathbf{T}} \cdot \mathbf{A}_{\text{cal}})^{-1} \cdot \mathbf{A}_{\text{cal}}^{\mathbf{T}} \cdot \mathbf{C}_{\text{cal}} \tag{39}$$

which can then be used for predicting unknown samples:

$$\hat{\mathbf{c}}_{\text{meas}} = \hat{\mathbf{P}}^{\mathbf{T}} \cdot \mathbf{a}_{\text{meas}}$$

Both methods have advantages and drawbacks: CLS minimizes spectral errors—ILS minimizes concentration errors. Usually, the spectroscopic data contain more noise then the calibration concentrations, which can be determined by very precise reference methods. Hence, the CLS calibration (eq. 37) is the more appropriate one compared to the ILS calibration (eq. 39) since CLS calibration is based on the precisely known model. The ILS method, however, uses the less precise calibration based on noisier spectral data. Nonetheless, ILS is supposed to be the superior approach for practical applications since it only needs calibration concentrations of the analytes of interest. In order to make CLS a good predictor calibration concentrations of all analytes must be included that are expected during the measurement process. This restriction is severe, especially in process monitoring where usually a huge number of absorbers are involved. For such applications it is unfeasible or even impossible to determine calibration concentrations of all of them. This is emphasized by means of Figure 4: Second derivative uv spectra obtained from gaseous samples containing different concentrations of NH_3, NO, and SO_2 (22) have been analyzed by CLS and PCR—a ILS based approach (see the section Principal Component Analysis and Principal Component Regression). For demonstration purposes of how CLS fails, only NH_3 and NO had been calibrated although SO_2 was contained in the

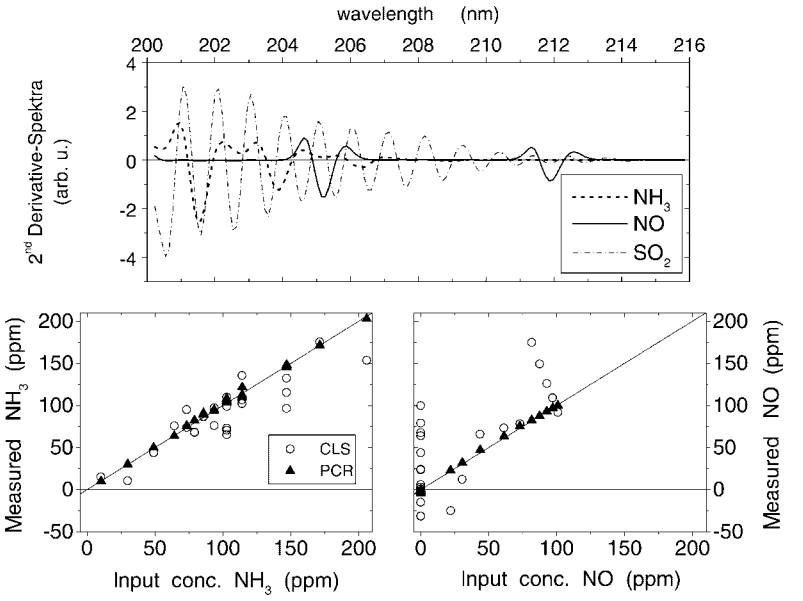

Fig. 4. Comparison of CLS (see section Classical Least Squares) versus PCR (see section Principal Component Analysis (PCA) and Principal Component Regression (PCR)) applied to second derivative spectroscopy in case of incomplete calibration information. Calibration concentrations of gaseous NH_3 and NO have been used—but not of SO_2 (incomplete information), which has also been contained in some calibration samples, though.

calibration samples, too. It is obvious that the CLS calibration cannot handle the SO_2 features strongly overlapping the features of the calibrated analytes. PCR on the other side, is able to calibrate the SO_2 features implicitly and determines correct concentration results. In many applications, eg, spectroscopic estimation of octane number or Reed vapor pressure, no "spectrum" of the extrinsic property would exist. Hence, ILS methods have to been applied.

3.2. Principal Component Analysis (PCA) and Principal Component Regression (PCR). Factor analysis (FA) is employed to aid in visualization of sample (time) dependent trends and measurement (sensor) dependent trends in a multidimensional data space. In general, factor analysis does not give a physically meaningful model—only correlations among samples and measurements are determined. However, FA methods have been modified to apply constraints and assumptions based on previous knowledge of the chemical system being analyzed. These modified FA methods are useful for determining the underlying instrumental and/or sample (time) profiles of the chemical constituents of a process. Perhaps the most commonly applied method of FA is principal component analysis–regression (PCA–PCR) (1–3,23,24)—an ILS based approach. The PCA method only extracts the principal component (PC) or loading vectors by means of which unknown measurement data will be represented. It takes a second step to relate such an abstract data representation to chemical properties, concentrations, eg. Both steps together are PCR and will be discussed in the following.

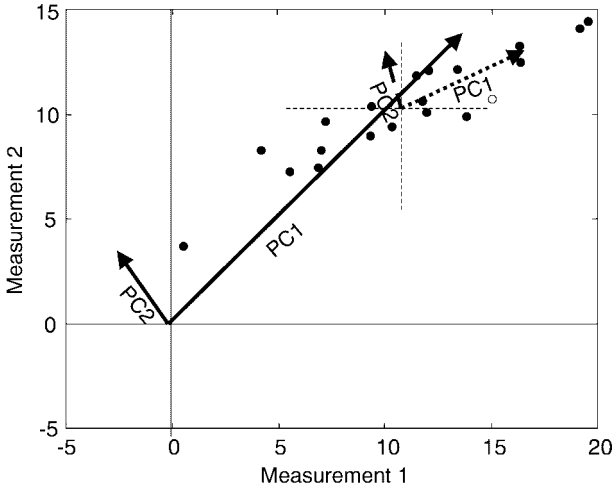

Fig. 5. Graphic representation of principal component analysis. For the data set (red dots) principal components (blue lines) are defined to start at the origin of the coordinate axis system. The first PC describes the main source of variance in the data. For uncentered data, the first PC generally points from the origin of the coordinate axis through the center of the data set; the second PC describes the direction of greatest variance orthogonal to the first PC. For mean centered data, the center of the data set is translated to the origin of the coordinate system; the first PC then describes the direction of greatest variance.

The goal of PCA is to identify the major sources of correlated variance in a collection of data. Once these sources of variance have been identified, they can be exploited to aid in the visualization of the major trends throughout the data collection. The data collection can be reduced from a complicated multidimensional representation to a more easily visualized two- or three dimensional space that describes the majority of the variance (information) in the data collection.

The conceptual idea behind PCA is presented in Figure 5. The largest direction of variance in the data collection is the first PC. The second PC is defined to describe the maximum amount of variance in the data collection while constrained to be orthogonal to the first PC. Consequently, each additional PC is also defined to maximize variance described while constrained to be orthogonal to all preceding PCs. Note that the PCs are defined as vectors originating at the origin of the coordinate space. Therefore, the PCs are dependent on the average value of the data collection; translating the data cloud to a different point in the coordinate space changes the direction of the PCs. For this reason, the data collection is often translated to be centered about the origin of the coordinate space (see the section Data Pretreatment—Mean Centering and Scaling). However, the location of the data collection does not affect the ability of PCA to model the data variance. Only the ease of interpreting the model is affected.

There is a difference between factor analysis and calibration methods: Factor analysis extracts underlying factors or model by means of which the analyzed data can be described—calibration, however, extracts such factors and relates them to chemical or physical properties. A calibration enables a prediction of

chemical or physical properties of unknown future samples, factor analysis analyzes the presented (calibration) data only.

Calibration. This discussion of PCR makes wide use of the multivariate least-squares fit concepts (the section Multivariate Linear Regression) and the same notation is used. To exemplify the discussion for spectroscopy, the corresponding spectroscopic items are mentioned in parenthesis. During the calibration process, K calibration samples (spectra) are acquired. The values of N different predictor variables $x_{(1)} \ldots x_{(N)}$ (absorption at different wavelength positions) are measured for each of these K calibration samples. These N predictor variable values are concluded in N calibration vectors $\mathbf{x}_{(1)}{}^{\mathrm{cal}} \ldots \mathbf{x}_{(N)}{}^{\mathrm{cal}}$ (eq. 23) comprising K values each, one for every calibration sample. These calibration vectors define a calibration matrix $\mathbf{X}_{(K \times N+1)}{}^{\mathrm{cal}} = [1\ \mathbf{x}_{(1)}{}^{\mathrm{cal}} \ldots \mathbf{x}_{(\mathrm{N})}{}^{\mathrm{cal}}]$ (eq. 23). If mean centering was applied, the first column of ones is not needed, ie, $\mathbf{X}_{(K \times N)}{}^{\mathrm{cal}} = [\mathbf{x}_{(1)}{}^{\mathrm{cal}} \ldots \mathbf{x}_{(N)}{}^{\mathrm{cal}}]$.

Calculation of PCs can be accomplished by a singular value decomposition of $\mathbf{X}^{\mathrm{cal}}$ (SVD, eq. 2—see the section Notation and Fundamental Mathematical Tools and Fig. 6):

$$\mathbf{X}^{\mathrm{cal}}_{(K \times N)} = \mathbf{U}_{(K \times K)} \cdot \mathbf{S}_{(K \times K)} \cdot \mathbf{P}^{\mathrm{T}}_{(K \times N)} = \mathbf{T}_{(K \times K)} \cdot \mathbf{P}^{\mathrm{T}}_{(K \times N)} \tag{40}$$

The K orthonormal PCs $\mathbf{p}_k$ each consisting of N loading values are contained in the rows of $\mathbf{P}^{\mathrm{T}} = [\mathbf{p}_1 \ldots \mathbf{p}_K]^{\mathrm{T}}$. The corresponding scores of the calibration spectra are hold in the columns of $\mathbf{T} = \mathbf{U} \cdot \mathbf{S}$. A scores vector, ie, a column of $\mathbf{T}$, contains K weight factors determining how strong which PC contribute to the corresponding calibration sample (calibration spectrum). Often $\mathbf{P}^{\mathrm{T}}$ and $\mathbf{S}$ are multiplied, what is not done here in order to retain orthonormal PCs enabling less costly computations in the following. Calculating the PCs is a PCA, relating the PCs to chemical information extends PCA to PCR.

Due to noise contained in the calibration data $\mathbf{X}^{\mathrm{cal}}$, the "chemical" rank of $\mathbf{X}^{\mathrm{cal}}$ is usually equal to $\min(K,N)$ although there are only $R < \min(K,N)$ chemical meaningful PCs. The number of linear independent influences on the calibration samples determines the number of chemical meaningful PCs.

Especially for process monitoring applications when only incomplete information about the calibration samples is available, the decision about the true

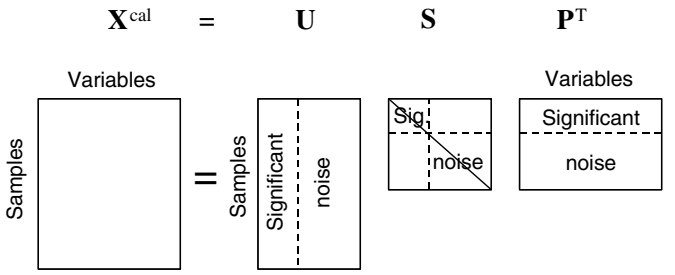

Fig. 6. With PCA the singular value decomposition is often employed to decompose a data matrix $\mathbf{X}^{\mathrm{cal}}$ into three submatrices $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{P}^{\mathrm{T}}$. These three matrices can be partitioned into columns that describe systemic, often chemical, sources of variance and columns that describe only random measurement related noise.

"chemical" dimension R of the calibration model is rather difficult. If all K PCs would be included, the PCA calibration model would usually be overfitted and hence the evaluation results would be downgraded (25). The parameter R is usually determined by methods presented in the section Model Selection. After finding R, $\mathbf{P}_{(R\times N)}{}^{\mathbf{T}}$ and $\mathbf{T}_{(K\times R)}$ are downsized to the number of significant PCs without changing the notation in the following. Evaluating an unknown data set $\mathbf{x}_{\text{meas}}$ is a two-step process: In the first step, $\mathbf{x}_{\text{meas}}$ is projected onto the PCs in order to determine its scores vector $\mathbf{t}_{\text{meas}}$. The second step relates these scores to chemical information (concentration of the calibrated analytes). The scores itself have no chemical meaning, however, the wanted pieces of chemical information are linear combination of the scores. This mapping matrix $\mathbf{B}$ from scores to chemical information has to be extracted from the calibration set, too

$$\mathbf{Y}^{\text{cal}}_{(K\times M)} = \mathbf{T}_{(K\times R)} \cdot \mathbf{B}_{(R\times M)} \Rightarrow \hat{\mathbf{B}} = (\mathbf{T}^{\mathbf{T}}\mathbf{T})^{-1} \cdot \mathbf{T}^{\mathbf{T}} \cdot \mathbf{Y}^{\text{cal}} \tag{41}$$

By comparing equation 41 to eqs. 29 and 30 reveals that $\hat{\mathbf{B}}$ plays the role of $\hat{\mathbf{A}}$ and $\hat{\mathbf{T}}$ the one of $\mathbf{X}^{\text{cal}}$ in this application of multivariate least squares fit. Now the calibration is finalized.

From a different standpoint, these PCs span a R-dimensional subvector space of the N-dimensional vector space of real numbers R^N. The scores are the coordinates of the N-dimensional predictor variable vectors. All features contained in the future unknown predictor vectors will be found in this subvector space. However, the PCs have no physical or chemical meaning—they are linear combinations of all physical or chemical properties present in the calibration samples. The same is true for the score vectors.

Prediction of Unknown Samples. Based on a PCR calibration (see section Calibration) unknown data sets x_{meas} can be evaluated by

$$\begin{aligned} \mathbf{x}_{\text{meas}} &= \mathbf{P} \cdot \mathbf{t}_{\text{meas}} + \epsilon \\ \hat{\mathbf{t}}_{\text{meas}} &= \underbrace{(\mathbf{P}^{\mathbf{T}}\mathbf{P})^{-1}}_{=1} \cdot \mathbf{P}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} = \mathbf{P}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} \\ \mathbf{y}_{\text{meas}} &= \mathbf{B}^{\mathbf{T}} \cdot \hat{\mathbf{t}}_{\text{meas}} \end{aligned} \tag{42}$$

The MLR analogon to the last line of equation 42 is equation 31.

The power of the PCA–PCR approach lies herein: One needs only to know calibration information on the wanted response variables $\mathbf{Y}^{\text{cal}}$ (eq. 29) even if there are plenty of other unknown influences affecting the values of the predictor variables—the algorithm determines by itself an appropriate calibration model, ie, the PCs $\mathbf{P}$ (eq. 40) and a transform matrix $\mathbf{B}$ (eq. 41). However, all influences occurring during the evaluation of unknown data sets $\mathbf{x}_{\text{meas}}$ must also be present during calibration for being implicitly calibrated. In contrast to this, conventional multivariate least squares regression (the section Multivariate Linear Regression) needs the user to include all influences explicitly into an appropriate calibration model $\mathbf{X}^{\text{cal}}$.

An Experimental Example. Second derivative uv spectroscopy has been used in this example of how PCR can implicitly calibrate imperfect measurement

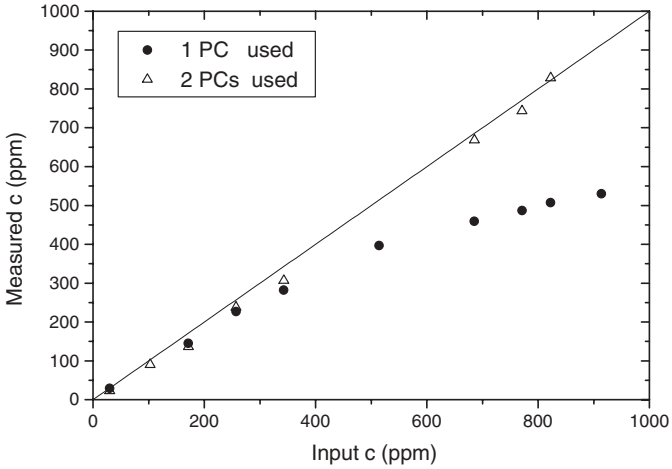

Fig. 7. Concentration errors (solid dots) determined with a one PC model obtained from single compound samples—improved precision (hollow triangles) obtained with a two PC model obtained from the same data set.

data. It was shown in (22) how a linear relation between concentrations and derivative spectra is obtained. As an example gaseous ammonia samples in the concentration range 0–1000 ppm had been prepared and analyzed by means of a PCR. The first calibration set contained two samples (0 and 103 ppm) only which is in theory sufficient for such simple applications. Just one significant PC was found. However, it was found that Beer's law (eq. 7) is not valid over the concentration range aimed at since the measured concentrations are falling short over 300 ppm (Fig. 7).

As is demonstrated by means of Figure 8, this is not a problem of PCR but of the employed measurement technique. Dividing the measured spectra by the concentration values of the sample derives extinction or unit spectra. In Figure 8, three experimentally determined extinction spectra are shown obtained from low, medium, and high concentration samples. In absence of systematic measurement errors, all three would be the same. However, with increasing concentration the peak height is not increasing linearly, furthermore the shape of the extinction spectra is smeared out and the peaks height ratio to each other changes.

In a nutshell, there is more then one influence, ie, concentration, defining the measurement data. Hence, a calibration model including just one predictor variable vector or PC is not appropriate. The CLS (the section Classical Least-Squares versus Inverse Least Squares) would not be able to lift this problem since this algorithm is not flexible enough—it allows only for as many dimensions of the calibration model as analytes have been calibrated. Not so PCR: After increasing the calibration set to three calibration spectra (30, 514, 914 ppm), two significant PCs were found (Fig. 9).

Including both significant PC resulted in clearly improved concentrations (Fig. 7). This second PC enabled an adjustment of the systematic measurement errors. The nonlinear cooperation of the two factors is demonstrated by means of

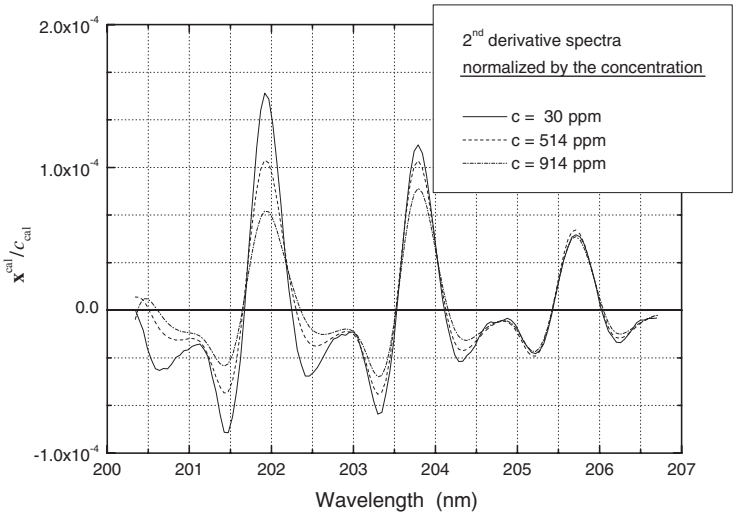

Fig. 8. Comparing three normalized second derivative spectra (extinction spectra) of gaseous NH_3 diluted in N_2 (22), ie, derivative spectra divided by the concentration of the samples. If Beer's law had been applicable, all normalized spectra would be equivalent.

Figure 10, which shows the concentration dependency of the corresponding scores. The scores of the first PC increase with increasing concentration but this increase is slowed with increasing concentration. The second loading is also increasing for concentrations below ~500 ppm due to enhancing the NH_3 peak near 202 nm. Above, its importance is decreasing because of the altering spectrum shape, which is getting more similar to the first factor. At ~800 ppm the second loading is zero, ie, for this concentration the first factor is proportional to the NH_3 derivative spectrum. Above this threshold, the second loading is

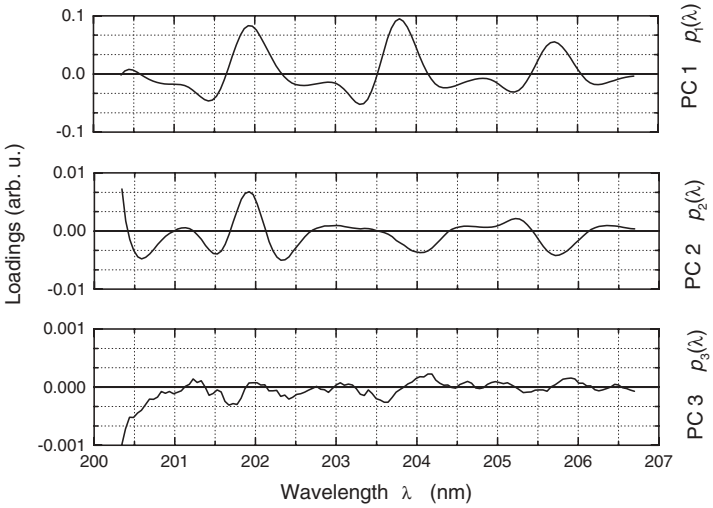

Fig. 9. Two significant PCs and one irrelevant PC obtained from the calibration samples comprising one analyte and systematic measurement errors, ie, nonlinearities (Fig. 8).

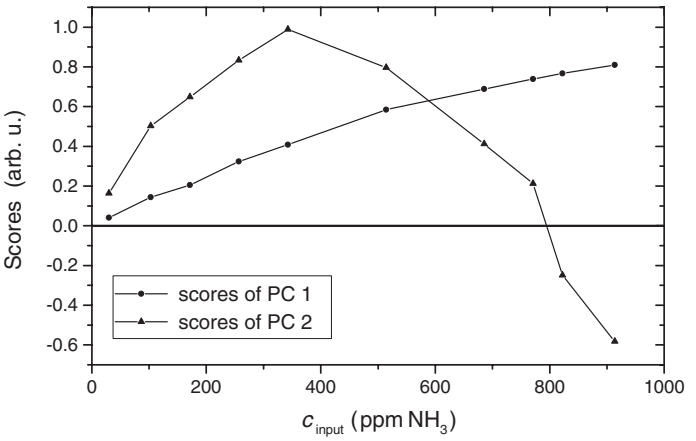

Fig. 10. Nonlinear relationship of the scores of PC 1 and PC 2 (Fig. 9) with increasing concentration.

getting negative, ie, compared to the other two peaks the 202 nm peak is suppressed further by the second factor.

3.3. Partial Least Squares (PLS). Partial least-squares regression (PLS) (21,26–28) has been employed since the early 1980s and is closely related to PCR and MLR. In fact, PLS can be viewed as a compromise midway between PCR (the section Principal Component Analysis and Principal Component Regression) and MLR (29) (the sections Multivariate Linear Regression and Classical Least-Squares versus Inverse Least Squares and 3.1). In determining the decomposition of $\mathbf{X}^{cal}$ and consequently removing unwanted random variance, PCR is not influenced by knowledge of the calibration set's response variables $\mathbf{y}^{cal}$ and $\mathbf{Y}^{cal}$, respectively. Only the variance in $\mathbf{X}^{cal}$ is employed to determine the loading vectors. Conversely, MLR does not factor $\mathbf{X}^{cal}$ prior to regression; all variance correlated to response variables is employed for estimation. PLS determines each loading vector to simultaneously optimize variance described in $\mathbf{X}^{cal}$ and correlation with $\mathbf{y}^{cal}$. The PLS loading vectors are rotations of the PCA PCs for a slightly different optimization criterion. In fact, numerous algorithms exist that are optimized for various sizes of $\mathbf{X}^{cal}$ (30,31).

PLS has two distinct advantages compared to PCR. First, PLS generally provides a more parsimonious model than PCR. The PCR calculates factors in decreasing order of $\mathbf{X}^{cal}$-variance described. Consequently, the first factors calculated, that have the least imbedded errors, are not necessarily most useful for calibration. On the other hand, the first few PLS factors are generally most correlated to concentration. As a result, PLS achieves comparable calibration accuracy with fewer loading vectors in the calibration model. This further results in improved calibration precision because the first factors are less prone to imbedded errors than are lower variance factors.

Second, the PLS algorithm is often faster to implement and optimize for a given application than is the PCR algorithm. The PLS calculates the factors one at a time. Hence, only the loading vectors needed for calibration are determined. The PCR, employing the singular value decomposition, calculates all possible loading vectors for $\mathbf{X}^{cal}$ prior to regression. For large data sets that require relatively few factors for calibration, PLS can be significantly faster than PCR.

PLS extracts iteratively as much variance form $\mathbf{X}^{\text{cal}}$ as it can correlate to the response variable values $\mathbf{y}^{\text{cal}}$. The explained information is subtracted from $\mathbf{X}^{\text{cal}}$ and $\mathbf{y}^{\text{cal}}$—the residuals enter the next iteration step then. Since every iteration operates on the residual of the previous step, the extracted loading vectors are mutually orthogonal to each other (the section Supplementary Topics). The PLS determines a calibration model for every response variable independently—if there are several response variables to be predicted, the algorithm presented in the following has to be run the several times using the $\mathbf{y}^{\text{cal}}$ vectors one after the other.

Calibration. The iterative PLS calibration algorithm is presented first followed by a step-by-step discussion of it.

The following discussion explains how the hth iteration works:

1. Initialization of the algorithm with the mean centered (see section Data Pretreatment—Mean Centering and Scaling) original data $\mathbf{X}^{\text{cal}}$ and $\mathbf{y}^{\text{cal}}$.

	Model	Least-squares estimate
1. initialization:	$\mathbf{E}_0 = \mathbf{X}_{(K\times N)}{}^{\text{cal}}$ and $\mathbf{e}_0 = \mathbf{y}$ (mean centered, see section Data Pretreatment—Mean Centering and Scaling) $h = 1$ iteration counter	
2. determine a weight loading vector $\mathbf{w}_h$ (new predictor variable vector):	$\mathbf{E}_{h-1} = \mathbf{e}_{h-1} \cdot \mathbf{w}_h^{\mathrm{T}} + \epsilon_{\mathbf{E}}$	$\hat{\mathbf{w}}_h = \mathbf{E}_{h-1}^{\mathrm{T}} \cdot \mathbf{e}_{h-1} \cdot (\mathbf{e}_{h-1}^{\mathrm{T}} \cdot \mathbf{e}_{h-1})^{-1}$
3. normalize $\hat{\mathbf{w}}_h$:	$\hat{\mathbf{w}} = \frac{\hat{\mathbf{w}}}{\lVert\hat{\mathbf{w}}\rVert_2}$	
4. determine the corresponding scores vector $\mathbf{t}_h$:	$\mathbf{E}_{h-1} = \mathbf{t}_h \cdot \hat{\mathbf{w}}_h^{\mathrm{T}}$	$\hat{\mathbf{t}}_h = \mathbf{E}_{h-1} \cdot \hat{\mathbf{w}}_h \underbrace{(\hat{\mathbf{w}}_h^{\mathrm{T}} \cdot \hat{\mathbf{w}}_h)^{-1}}_{=1}$ $\hat{\mathbf{t}}_h = \mathbf{E}_{h-1}^{\mathrm{T}} \cdot \hat{\mathbf{w}}_h$
5. relate scores to residues of the response variable $\mathbf{e}_{h-1}$:	$\mathbf{e}_{h-1} = \hat{\mathbf{t}}_h \cdot \upsilon_h + \epsilon_e$	$\hat{\upsilon}_h = (\hat{\mathbf{t}}_h^{\mathrm{T}} \cdot \hat{\mathbf{t}}_h)^{-1} \cdot \hat{\mathbf{t}}_h^{\mathrm{T}} \cdot \mathbf{e}_{h-1}$
6. determine a loading vector $\mathbf{b}_h$	$\mathbf{E}_{h-1} = \hat{\mathbf{t}}_h \cdot \mathbf{b}_h^{\mathrm{T}} + \epsilon_{\mathbf{E}}$	$\hat{\mathbf{b}}_h = (\hat{\mathbf{t}}_h^{\mathrm{T}} \cdot \hat{\mathbf{t}}_h)^{-1} \cdot \hat{\mathbf{t}}_h^{\mathrm{T}} \cdot \mathbf{E}_{h-1}$
7. determine new residues	$\mathbf{E}_h = \mathbf{E}_{h-1} - \hat{t}_h \cdot \hat{\mathbf{b}}_h^{\mathrm{T}}$ $\mathbf{e}_h$ $= \mathbf{e}_{h-1} - \hat{\upsilon}_h \cdot \hat{\mathbf{t}}_h$	
8. if $\lVert\mathbf{e}_h\rVert_2 <$ min, then $R = h$ else $h \rightarrow h+1$	END go to step 2	

2. Determine a loading vector $\hat{\mathbf{w}}_h$, which is used to estimate the scores $\hat{\mathbf{t}}_h$. This is done in such a way that as much variance of $\mathbf{X}^{\text{cal}}$ (for $h = 1$) or the remaining residuals $\mathbf{E}_{h-1}$ (for $h > 1$) is extracted as can be explained by the response variable $\mathbf{y}^{\text{cal}}$ (for $h = 1$) or the remaining residuals $\mathbf{e}_{h-1}$ (for $h > 1$). For $h = 1$, PLS is a CLS [(the section Classical Least Squares versus Inverse Least Squares), especially equation 37 for $M = 1$ with $X^{\text{cal}} = A^{\text{cal}}$

and $\mathbf{y}^{\text{cal}} = \mathbf{C}^{\text{cal}}{}_{(K \times M=1)}]$ and there is actually a physical meaning for $\hat{\mathbf{w}}_h$: PLS estimates the pure component—in spectroscopic applications this would be the pure component or more precisely molar extinction spectrum, ie, $\mathbf{e}(\lambda)$ (eq. 7). This function of wavelength $\mathbf{e}(\lambda)$ should not be confused with the residual vector $\mathbf{e}_h$ used in the PLS algorithm above. This estimation, however, will be poor if more than one analyte is contained in the calibration samples. The following iterations extract correction terms.

3. Normalize $\hat{\mathbf{w}}_h$ to get an orthonormal basis, this simplifies step 4 and gives every vector the same weight during evaluation.
4. Now the scores $\mathbf{t}_h$ for $\hat{\mathbf{w}}_h$ are determined describing how strong $\hat{\mathbf{w}}_h$ are present in $\mathbf{E}_{h-1}$. $\hat{\mathbf{w}}_1$ estimates the pure component spectrum in the aforementioned spectroscopic application; $\mathbf{t}_1$ is a first order approximation of the calibration concentrations.
5. This step the equivalent to equation 41 in PCR: Relate the scores to chemical meaningful items, concentrations, eg—or more precisely the hth contribution to the concentrations.
6. Now the loading vectors $\hat{\mathbf{w}}_h$ have to be "updated" to $\mathbf{b}_h$ incorporating how much of the response variables (concentrations) have actually been explained by the scores. The parameter $\hat{\mathbf{w}}_h$ could not be used in step 4 as final loading vector since the scores $\hat{\mathbf{t}}_h$ had to been estimated first. The difference between $\mathbf{t}_h$ and $\hat{\mathbf{t}}_h$ makes this step and the definition of $\mathbf{b}_h$ necessary.
7. Subtract the extracted and explained information from the residues of the previous step.
8. The algorithm is aborted, if, eg, all information on the calibration's response variable is explained. Otherwise a new iteration is started.

Prediction of Unknown Samples. The prediction algorithm for unknown samples is initialized with the mean-centered predictor and response variables. Then the sample is projected onto the $\hat{\mathbf{w}}_h$ loading vectors in order to estimate the scores value ${t_h}^{\text{meas}}$. By means of the constant $\hat{v}_h$ relating scores and response variable a back transformation is done from the scores representation to physical meaningful objects, concentrations for instance. Step by step the wanted response variable y is updated to the final value. The prediction algorithm has the same number of iterations as the calibration.

1. initialization	$\mathbf{e}_0^{\text{meas}} = \mathbf{x}_{\text{meas}} - \bar{\mathbf{x}}^{\text{cal}}$ and $\mathbf{y}_{\text{meas}} = \bar{\mathbf{y}}^{\text{cal}}$ $h = 1$ iteration counter
2. project the residues of the unknown predictor variable vector onto the new predictor variable vector:	$t_h^{\text{meas}} = \hat{\mathbf{w}}_h^{\text{T}} \cdot \mathbf{e}_{h-1}^{\text{meas}}$
3. update the estimate of the response variable:	$y_h = y_{h-1} + \hat{v}_h \cdot t_h^{\text{meas}}$
4. determine a new residual of the predictor variable vector:	$\mathbf{e}_h = \mathbf{e}_{h-1} - \hat{\mathbf{b}}_h \cdot t_h^{\text{meas}}$
5. $h \rightarrow h + 1$	
6. if $h \leq R$, go to step 2 else END	

3.4. Model Selection. As stated in the sections Principal Component Analysis and Principal Component Regression and Partial Least Squares, the most difficult part of PCR and PLS is to determine the dimension R of the calibration model. The singular values s_K (eq. 2) arranged in decreasing order could be used at least for PCR. As was stated in the section Notation and Fundamental Mathematical Tools, there are R singular values unequal to zero. However, due to the noise no singular value is exactly zero except if mean centering was applied since a degree of freedom was lost. In that case the last one is zero. In most applications the singular values drop to very small values belonging to nonsignificant PCs with a more or less pronounced step. Plotting these singular values can help to get a first guess at least. But this works only in limited cases and often a more sophisticated method is needed.

Usually cross-validation (1) is applied that excludes one of the calibration samples for determining the number R of significant PCs. This is done by including iteratively more PCs and estimating the values of the response variables. Since the true value(s) y^{cal} or $\mathbf{y}^{cal}$ are known for this excluded calibration samples, the estimated value(s) $\hat{y}^{cal}$ or $\hat{\mathbf{y}}^{cal}$ can be compared to the true values. The procedure of excluding a calibration sample and estimate its response variable(s) is done for all calibration samples successively. The number of PCs achieving the closest estimates is chosen for future evaluation.

If a sufficient quantity of calibration samples is available, the best method for selecting and validating a model is to divide the calibration set into three subsets. One set is employed to construct all of the models to be considered. The second set is employed to choose the best model in terms of accuracy and precision. The third set is employed to estimate the performance of the chosen model on future data. There are three statistics often employed for comparing the performances of multivariate calibration models: root-mean-squared error of calibration (RMSEC), root-mean-squared error of cross-validation (RMSECV), and root-mean-squared error of prediction (RMSEP). All three method are based on the calculated root mean squared error (RMSE)

$$\mathrm{RMSE} = \left(\sum_{k=1}^{K} (y_k^{\mathrm{cal}} - \hat{y}_k^{\mathrm{cal}})^2 / K \right)^{1/2} \tag{43}$$

where RMSEC, RMSECV, and RMSEP differ in the determination of $\hat{y}^{cal}$. The best estimate of future performance of a calibration model is the RMSEP. Estimates $\hat{y}^{cal}$ in the RMSEP are determined by applying the calibration model to a subset of data that was not employed in determining the model parameters. The RMSEP may be calculated for a 'validation set' in order to determine the optimal number of factors in a model or to a 'test set' in order to test the performance of the optimal model on future data. If an external subset of data is not available to optimize the calibration model, the RMSEP can be estimated by the RMSECV. The concentration estimates of equation 43 are determined in cross-validation. RMSEC is a measure of how well the calibration model fits the calibration set. This is potentially the least informative of the three statistics. The RMSEC is an extremely optimistic estimation of the model performance. In the limit, if every factor were included in the calibration model, the RMSEC would be zero.

Hence, RMSEC is always decreasing with number of factors. As more factors are included in the calibration model, the model begins to fit the random errors imbedded in the spectra and concentrations. Therefore, the RMSEC will always decrease as more factors are added. However, new samples not included in the calibration set will have a different realization of random errors. Therefore, the calibration model will not fit these errors to the same degree as the errors in the calibration set. When extra factors that mostly describe random errors are included in the calibration model, these factors will introduce the errors in future samples and the RMSECV and RMSEP may increase. Therefore, RMSECV is a better estimate of future performance of model prediction than is RMSEC. This so-called overfitting is well described for PCR in reference 25—similar facts apply to PLS.

The performances of the three statistics are evident in Figure 11**a**–**d** presenting the RMSEC, RMSECV, and RMSEP versus number of factors for PLS calibration of moisture, oil, protein, and starch, respectively. All spectra were preprocessed by MSC and mean centered. The optimal number of factors may be estimated by statistical tests applied to the RMSE, choosing the first minima in the plot, or choosing the global minimum in the plot (3).

Unfortunately, it is often difficult to obtain reliable calibration samples, which are hence too valuable for testing the calibration model only. Furthermore, the dimension of a calibration model defined cross-validation is fixed and cannot be adjust to certain data. Hence, for spectroscopic applications a fine-tuning

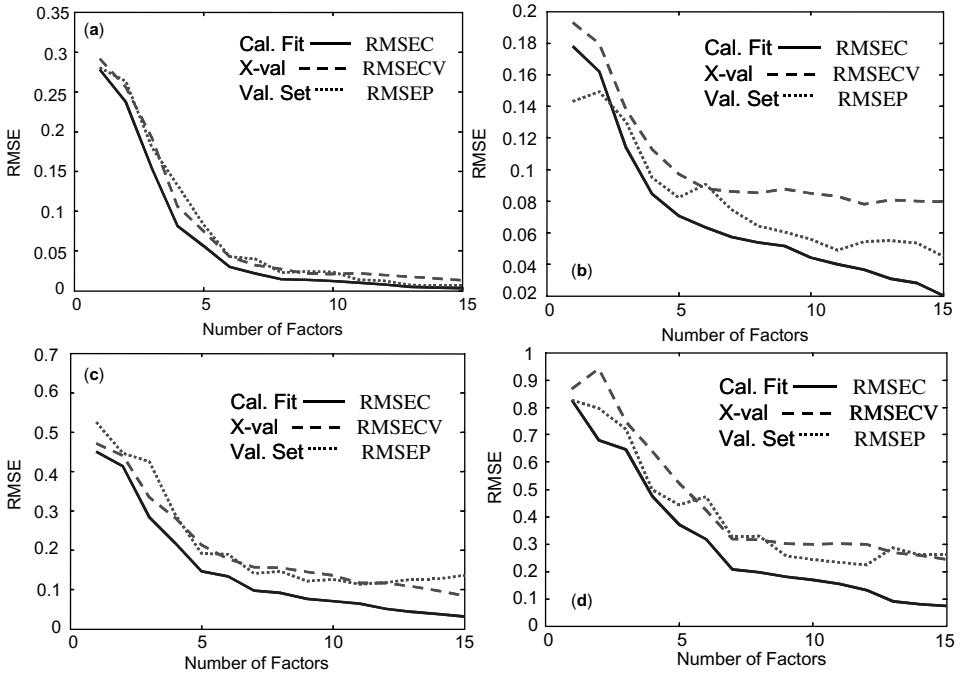

Fig. 11. Plots of RMSEC, RMSECV and RMSEP for prediction of (**a**) moisture, (**b**) oil, (**c**) protein, and (**d**) starch from NIR spectra of cornflour samples.

approach was proposed (32) adjusting the number of used PCs to every measurement spectrum individually. This method can be extended to other multivariate data sets, too. By means of this method no valuable calibration data must be excluded for testing purposes and the calibration model is flexible. In general, there are three types of PCs: Primary PCs are the most important ones modeling the major and true spectroscopic features in a spectrum. The secondary PCs are needed for correcting imperfect measurements like drifts (33) or washed out features (22). Tertiary PCs are due to noise and should not be included into the calibration model to prevent overfitting (25). Starting with one PC the number of PCs is increased stepwise by one. This defines a reduced model. The variance of the residual spectrum obtained at every step of this iteration is F-tested against the variance of the residual spectrum obtained with full model including all PCs. The number of PCs is increased until the F-test cannot find significant differences between the restricted and full model. At that point both methods have the same predictability. By means of this the algorithm reduces overfitting and still extracts all relevant information. By means of synthetic data it was proven that the algorithm selects the correct number of PCs, if noise level is reasonable and if sufficient calibration samples are provided. This fine-tuning of the calibration model could also be applied advantageously to different experimental spectra sets.

3.5. Target Factor Analysis. As mentioned in the section Calibration, the principal components (PCs) derived from PCA do not necessarily describe single, physically meaningful, effects. That is, while a set of data may consist of the NIR spectra of hydrocarbon mixtures, the PCs of the data set are not constrained to be NIR spectra of the constituent hydrocarbons. However, the multivariate space defined by the principal components is the same as the multivariate space defined by the pure (true) spectra of the chemical constituents of the data set plus any other forms of systematic variance. The difference is the basis (see last paragraph in the section Calibration) used for representation: The PCs are rotated versions of the pure component spectra. Target factor analysis (TFA) is a method of testing whether the spectrum of a hypothesized chemical constituent, as defined by an assumed or recorded spectrum, lies in the PC space of the model. If the hypothesized constituent does lie in the PC space, the associated spectrum $\mathbf{x}_{\text{meas}}$ can be expressed as a linear combination of the PCs [rows of $\mathbf{P}^{\mathbf{T}}{}_{(R\times N)}$ (eq. 40)]:

$$\mathbf{x}_{\text{meas}} = \mathbf{P} \cdot \mathbf{t} \tag{44}$$

The coefficients of the linear combination are the scores $\mathbf{t}$ of $\mathbf{x}_{\text{meas}}$. $\mathbf{t}$ can be calculated by regressing, i.e. projecting, the target spectrum, $\mathbf{x}_{\text{meas}}$, onto the orthonormal PCs $\mathbf{P}^{\mathbf{T}}$:

$$\begin{aligned} \hat{\mathbf{t}} &= (\mathbf{P}^{\mathbf{T}} \cdot \mathbf{P})^{-1} \cdot \mathbf{P}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} = \mathbf{P}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} \\ \hat{\mathbf{x}}_{\text{meas}} &= \mathbf{P} \cdot \mathbf{P}^{\mathbf{T}} \cdot \mathbf{x}_{\text{meas}} \end{aligned} \tag{45}$$

Whether $\mathbf{x}_{\text{meas}}$ lies in the vector space spanned by the PCs is tested by comparing $\mathbf{x}_{\text{meas}}$ with $\hat{\mathbf{x}}_{\text{meas}}$. If $\mathbf{x}_{\text{meas}}$ and $\hat{\mathbf{x}}_{\text{meas}}$ are determined to be sufficiently similar by a

statistical or empirical test (3), $\mathbf{x}_{meas}$ is an element of the PCs vector space. If $\mathbf{x}_{meas}$ is not a member of this vector space, the regression [first line in equation 45] estimating $\mathbf{t}$ determines a wrong vector $\hat{\mathbf{t}}$. In return, $\hat{\mathbf{x}}_{meas}$ will be significantly different from $\mathbf{x}_{meas}$. Analogous equations can be constructed that project sample targets onto the scores $\mathbf{T}$ (eq. 40) of $\mathbf{X}^{cal}$. Recently, methods had been developed which extracts the parts of $\mathbf{x}_{meas}$ causing it not being an element of the PC vector space (34,35). Such algorithms can be applied to analyze unknown spectra qualitatively for detecting and correcting of uncalibrated interferences.

3.6. Locally Weighted Regression. The global linear models calculated by PCR or PLS are not always the best strategy for calibration. Global models span the variance of all the samples in the calibration set. If the data are nonlinear, then the linear PCR and PLS methods do not efficiently model the data. This happens for instance, if a linear Beer's law (eq. 7) type relationship between predictor variables (spectra) and response variables (predicted chemical properties) does not hold. One option is to use nonlinear calibration methods employing a global, nonlinear model (the section Nonlinear Methods). The second option is to employ linear calibration methods on small subsets of the data. The locally weighted regression (LWR) philosophy assumes that the data can be efficiently modeled over a short span with linear methods (36–40). The first step in LWR is to determine the Q calibration samples that are most similar with the unknown sample to be analyzed. Similarity can be defined by distance between samples in the spectral space (38), by projections into the principal component space (39), and by employing estimates of the property of interest (40). Once the Q nearest standards are determined, either PLS or PCR is used to calculate the calibration model. The LWR has the advantage of often employing a much simpler and more accurate model for estimation of a particular sample. However, there are three disadvantages associated with LWR. First, two parameters must be optimized for LWR, number of local samples and number of factors, compared to only the latter one for PLS and PCR. Second, a new calibration model must be determined for every new sample analyzed. Third, LWR often requires more samples than PCR or PLS in order to build meaningful, local calibration models.

3.7. Nonlinear Methods. There are numerous nonlinear, multivariate calibration methods described in the chemometric literature. These methods can be divided into two classes. Alternating Conditional Expectations (ACE) (41,42) and Projection Pursuit (PP) (43) seek to transform the nonlinear data such that a linear calibration model is appropriate. Similarly, Global Linearizing Transformations (GLT) is employed to optimally linearize data prior to factor analysis by PCA (44,45). On the other hand, nonlinear-PLS (NPLS) (46,47). Multivariate Adaptive Regression Splines (MARS) (48,49) and Artificial Neural Networks (ANN) (50) determine nonlinear global models that span the entire range of samples. While impossible to provide sufficient detail for each method in this section, some general comments regarding the application of these nonlinear methods are warranted.

Specific nonlinear methods have been compared and contrasted over a wide variety of linear and nonlinear calibration applications (51–53). No single method has demonstrated systematic superiority to the other methods. The safe conclusion is that calibration method superiority is application dependent (54,55) When the underlying type of nonlinearity implicit in the calibration

method matches the latent nonlinearity in the data, the method will optimally model the data. This assertion has been supported by the improvement in calibration performance when theoretical instrument response functions replace the sigmoid transfer function in ANN calibration (56).

Nonlinear methods are much more prone to "over-fitting" the calibration model than linear approaches. Overfitting occurs when the calibration begins to employ random variance (instrumental errors) for determining calibration parameters. The flexibility of the nonlinear models and the relatively large number of parameters that need to be estimated are the primary cause for this phenomenon. Consequently, the more complicated the model, the more prone the method is to overfitting (ie, ANN vs. PCR). A decision tree based on Occums Razor has been proposed to aid chemists in choosing among the nonlinear methods) (54) Linear and nonlinear calibrations were linked in a hierarchical web. The hierarchy is based on nested models and degrees of freedom required to calculate the model. Simple, linear models are at the top of the hierarchy; complex, nonlinear methods are at the bottom. It is recommended that to guard against overfitting and spurious modeling of the data, the method nearest the top of the hierarchy that provides sufficient calibration reliability for the application be employed. That is, use the simplest model that works.

3.8. Multivariate Curve Resolution (MCR). Where TFA (the section Target Factor Analysis) allows the analysis to test for the presence of a hypothesized constituent, TFA is limited in the ability to estimate the spectral profile of any constituents in the data set. This is due to the fact that TFA requires that the spectral profile of the target is available for target testing. If the profile is unavailable, TFA cannot be performed. On the other hand, multivariate curve resolution (MCR) methods allow for the estimation for both the hypothesized and unknown constituents in the data matrix, ie, spectral or chromatographic profile of the separated constituent as well as concentration profiles. Usually MCR techniques are applied in spectroscopy and chromatography. The rotational ambiguity of the decomposition in (eq. 40) is circumvented by making assumptions regarding the nature of the true constituent spectral profiles and sample profiles. These assumptions are translated into constraints applied to the iterative factorization of $\mathbf{X}^{cal}$. Once additional constraints are applied to the factors of $\mathbf{X}^{cal}$, the factors are not true principal components. These factors are properly described as intrinsic factors, but not PCs.

Numerous constraints have been applied to the iterative factorization of $\mathbf{X}^{cal}$ in order to enhance the probability that the determined factors will be physically meaningful. Perhaps, the most common constraint is nonnegativity of estimated spectral and sample profiles (57–68). This constraint is based on the common sense notion that the factorization of $\mathbf{A} = \mathbf{X}^{cal}$ (eq. 36) (see section Classical Least Squares versus Inverse Least Squares) should lead to positive estimates of extinction coefficients or unit spectra (rows of $\mathbf{K}$) and concentrations $\mathbf{C}_{cal}$. In neither case would the true profile likely contain negative values. Another common spectral constraint employs assumptions regarding the content of $\mathbf{X}^{cal}$. If the spectral profile of one or more of the assumed chemical constituents is known, the factorization of $\mathbf{A} = \mathbf{X}^{cal}$ can be constrained to contain the assumed spectral profiles in the solution. It is also possible to employ assumptions regarding the interrelationship among the samples.

For resolving overlapping chromatographic peaks, Gaussian or unimodal elution profiles are assumed for the rows of $\mathbf{X}^{\text{cal}}$ (60,62,63) ie, there is for sure only one maximum in the chromatogram. Concurrently, the presence of samples that contain only one compound may be successfully postulated for chromatographic or kinetic data (62,64). This is referred to as the "uniqueness" constraint. If the concentration of one or more compound is known in any of the particular sample, the resolved profiles can be constrained to reflect this information. For kinetic data, the sample profiles can be constrained to fit a class of differential equations that reflect the postulated reaction pathway (65,66). The validity of the assumed reaction can be tested based of the ability of the data to fit this model.

Of course, application of other constraints and the combination of multiple constraints are possible. The constraints resulting from these assumptions are particularly powerful when well-ordered data, such as kinetics or chromatographic data, are analyzed. Constraining does not ensure that physically meaningful profiles will be determined. In general, application of constraints only reduces the range of feasible solutions where, ideally, the true profiles will lie within this range. The more constraints properly applied to the decomposition, the tighter the estimated range of profiles will resemble the true profiles. However, if a constraint is improperly imposed the estimated profiles will yield erroneous profiles, for instance if nonnegativity when in fact the profile should have negative values.

Practically, iterative MCR methods are capable of resolving spectra and concentrations from complicated, multianalyte mixtures without a priori information aside from constrains. Iterative MCR methods employ the bilinear factorization model

$$\mathbf{A}_{(K\times N)} = \mathbf{C}_{(K\times M)} \cdot \mathbf{P}^{\mathbf{T}}_{(M\times N)} + \mathbf{E}_{(K\times N)} \tag{46}$$

where the K mixture spectra or chromatograms are written in the rows of $\mathbf{A}$. The rows of C contain the analyte concentrations for the samples, rows of $\mathbf{P}^{\mathbf{T}}$ hold the spectral or chromatographic profiles of the pure single analytes at unit concentration, $\mathbf{E}$ is the residuals matrix.

$\mathbf{C}$ and $\mathbf{P}^{\mathbf{T}}$ are estimated by an alternating least squares (ALS) (67,68) algorithm: This algorithm starts with an estimate of either $\mathbf{C}$ or $\mathbf{P}^{\mathbf{T}}$. Assuming $\hat{\mathbf{C}}$ is employed for initialization, estimates of the spectral profiles are calculated based on MLR (the section Multivariate Linear Regression). Either a constrained least-squares fit is employed or the constraints are imposed after $\hat{\mathbf{P}}^{\mathbf{T}}$ is calculated directly by least squares:

$$\hat{\mathbf{P}}^{\mathbf{T}} = (\hat{\mathbf{C}}^{\mathbf{T}} \cdot \hat{\mathbf{C}})^{-1} \cdot \hat{\mathbf{C}}^{\mathbf{T}} \cdot \mathbf{A} \tag{47}$$

or utilizing the pseudoinverse (see section Notation and Fundamental Mathematical Tools and Supplementary Topics) when necessary

$$\hat{\mathbf{P}}^{\mathbf{T}} = \hat{\mathbf{C}}^{+} \cdot \mathbf{A}$$

If, eg, nonnegativity is applied, all negative entries of $\hat{\mathbf{P}}^{\mathbf{T}}$ are set to be zero. Once the constrained estimate of $\hat{\mathbf{P}}^{\mathbf{T}}$ is calculated, $\hat{\mathbf{P}}^{\mathbf{T}}$ is employed to update the

estimate of $\hat{\mathbf{C}}^{\mathrm{T}}$. As with calculating $\hat{\mathbf{P}}^{\mathrm{T}}$, constraints can be imposed during the calculation of $\hat{\mathbf{C}}^{\mathrm{T}}$ by a constrained least squares method, or after the estimation of $\hat{\mathbf{C}}$ by ordinary least squares

$$\hat{\mathbf{C}} = \mathbf{A} \cdot \hat{\mathbf{P}} \cdot (\hat{\mathbf{P}}^{\mathrm{T}} \cdot \hat{\mathbf{P}})^{-1} \qquad \text{or} \qquad \hat{\mathbf{C}} = \mathbf{A} \cdot \hat{\mathbf{P}}^{+} \tag{48}$$

The method iterates by alternating calculating constrained updates of $\hat{\mathbf{P}}^{\mathrm{T}}$ (eq. 47) and $\hat{\mathbf{C}}$ (eq. 48) back and forth until further refinement does not significantly change the model.

3.9. Outlier Detection. Two important statistics for identifying outliers in the calibration set containing K samples are the "sample leverage"and the "studentized residuals". A plot of leverage versus studentized residuals makes a powerful tool for identifying outliers and assigning probable cause. The sample leverage is a measure of the influence, or weight, each sample has in determining the parameters of the calibration model. Samples near the center of the calibration set (average samples) will have a relatively low leverage compared to samples at the extreme edges of the experimental design and outliers. The sample leverage is determined by

$$h_k = 1/K + \mathbf{u}_k^{\mathrm{T}} \cdot \mathbf{u}_k \tag{49}$$

where $\mathbf{u}_k$ is the row of associated matrix $\mathbf{U}$ (eq. 40) with the R significant principal components for the kth sample. Consequently, the sample leverage ranges from 0 for a sample in the center of an infinitely large calibration set to 1 for an extreme sample in a small data set.

The studentized residual is an indication of how well the calibration model estimates the analyte property in each sample. The studentized residual is similar to the Student's t-statistic; the estimation error of each sample is converted to a distance in standard deviations away from zero. An additional term is often added to the calculation to correct for the weight each sample has in determining the calibration model. The studentized residual is increased for samples with a large leverage; this is known as the studentized leverage corrected residuals. The studentized leverage corrected residuals are calculated by

$$t_k = \frac{|c_k - \hat{c}_k|}{\sigma\sqrt{1 - h_k}} \tag{50}$$

where

$$\sigma = \sqrt{\frac{\sum_{k=1}^{K} (c_k - \hat{c}_k)^2}{K - R - 1}} \tag{51}$$

with R being the number of PCs in the calibration model.

The plot of studentized leverage corrected residuals versus sample leverage provides insights into the quality of each calibration sample (Fig. 12). Samples with low leverages and low studentized residuals are typical samples in the

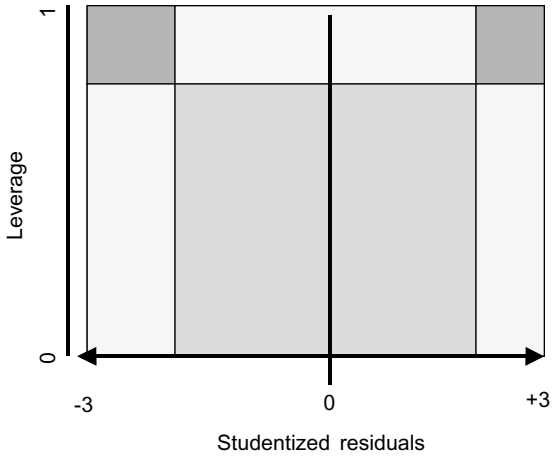

Fig. 12. Leverage-studentized residuals plots can be used to determine suspect data points.

calibration set. Data in the green region are generally "good" data and there is little statistically valid reason to remove any of these data. Data in the far-left and -right yellow regions have large fit errors and small leverages. These points are suspect in that they may have concentration errors or be mislabeled. Data in the upper yellow region are suspect due to spectral anomalies, but might have a high leverage just because they have an extreme concentration. Data in the red regions are most likely to be bad and should probably be removed.

4. Multiway Analysis

4.1. Introduction. Multiway analysis became popular in the late 1970s in the psychometric literature. Psychologists employed the multiway models primarily for factor analysis in order to determine intrinsic factors in large, complex data sets. However, as chemical instrumentation advanced with automated data collection, chemists began to acquire large, multiway data sets. In 1980, Hirschfeld listed 66 instruments capable of generating multiway data (69). Geladi cataloged the manners in which multiway data can be collected in chemical applications (12). Since a different notation has been defined in literature for multiway analysis compared to bilinear chemometrics the standard multivariate notation will be used in this section.

There are six classes of three-way data and four of these classes can be appropriately modeled with the basic trilinear, or PARAFAC (PARAllel FACtor analysis) (70–72) model. The PARAFAC decomposes the data cube (Figs. 1, 13) into N sets of triads, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ (see right part of Fig. 14). The elements of a trilinear $I \times J \times K$ data cube $\mathbf{R}$ can be presented as

$$R_{i,j,k} = \sum_{n=1}^{N} X_{i,n} \cdot Y_{j,n} \cdot Z_{k,n} + E_{i,j,k} \tag{52}$$

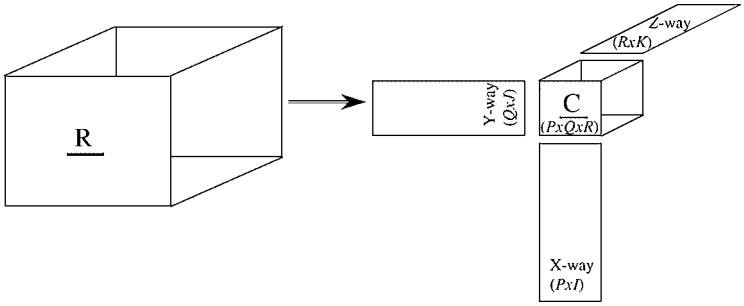

Fig. 13. The Tucker3 model (the section Tucker3 Models) is a generalization of the PARAFAC model (the section Multiway Curve Resolution—PARAFAC/CANDECOMP). The Tucker3 model decomposes the data cube into three sets of spectral and concentration profiles, like the PARAFAC model. However, the Tucker3 model additionally employs a core cube C that governs the mixing between the three spectral and concentration profiles. If the core matrix is all zeros except for having ones along the superdiagonal, the Tucker3 model reduces to the PARAFAC model.

Here, N refers to the rank of the model, ie, number of factors employed by the model. The parameter N must be determined (73) by the user before the algorithm is started. The residual data cube E contains the errors, which cannot be modeled.

To give an example: PARAFAC was used in combination with excitation–emission matrix (EEM) spectroscopy (74–76). The EEM spectroscopy uses a broadband light source usually in the uv–vis range to excite naturally fluorescent analytes, eg, aqueous solution of pesticides and polycyclic aromatic hydrocarbons. By means of two perpendicular spectrographs the excitation and emission spectra are projected in a perpendicular fashion onto a focal plane array (FPA). In other words, emission spectra measured after excitation at different wavelength are measured by the rows the FPA. The 2D $I \times J$ EEM spectra

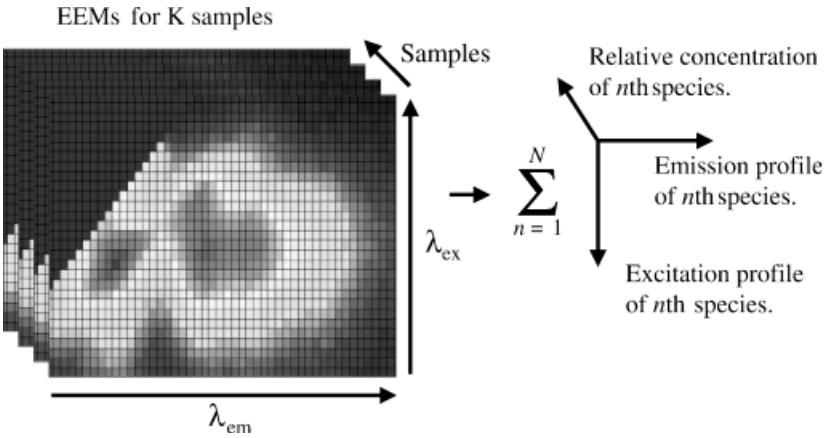

Fig. 14. Pictorial representation of collection of data as it can be factored into pure spectral and concentration profiles by PARAFAC.

obtained from K different samples are stacked to form a three-way data set $\mathbf{R}$ (Fig. 14). The PARAFAC is applied then in order to extract the excitation spectra $\mathbf{X}$, the emission spectra $\mathbf{Y}$, and the concentration profiles $\mathbf{Z}$ of the analytes from $\mathbf{R}$. Since the results are unique to a scaling factor (see discussion below), the excitation and emission spectra of the analytes are normalized to area one; the concentrations or multiplied then with the inverse of the spectra scaling factors. Another example would be obtaining a data cube from a LC-uv/vis device. Each $\hat{\mathbf{x}}_n$ would correspond to one of the true N chromatographic profiles, each $\hat{\mathbf{y}}_n$ to one of the true spectroscopic profiles, and each $\hat{\mathbf{z}}_n$ to the relative concentrations in the K samples.

In general, the number and form of factors are not constrained to be representative of any physical reality. With two-way factor analysis, PCA, this is often referred to as the rotational ambiguity of the factors; there is a continuum of factors that satisfy the PCA model and equivalently describe the data. This is different for three-way analysis. If the following four conditions are given, the factors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ of a chemical component are accurate and unique estimates of the true underling factors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ except for a scaling constant:

1. The true underlying factor in each of the three modes is independent from the state of the other two modes.
2. The true underlying factor in any of the three modes cannot be expressed by linear combinations of the true underlying factors of other components in the same mode.
3. Linear additivity of instrumental responses among the species present is given.
4. The proper number of factors N is chosen for the model.

4.2. Multiway Curve Resolution—PARAFAC/CANDECOMP. PARAFAC is originally based on the work of Kroonenberg (77) and as CANDECOMP (canonical decomposition) on the work of Harshman (78). In either case, the two base algorithms are practically identical. The PARAFAC uses an alternating least squares (ALS) based algorithm for multivariate curve resolution (the section Multivariate Curve Resolution) applied to three-way data sets.

The PARAFAC/CANDECOMP algorithm (79) stores iteratively improved estimates for the X-way, Y-way, and Z-way information in matrices $\mathbf{X}_{(I\times N)}$, $\mathbf{Y}_{(J\times N)}$, and $\mathbf{Z}_{(K\times N)}$. Before the algorithm is presented, six additional matrices have to be defined. Three of which ($\mathbf{R}^{\mathrm{A}}$, $\mathbf{R}^{\mathrm{B}}$, $\mathbf{R}^{\mathrm{C}}$) contain elements of $\mathbf{R}$ and remain unaltered—$\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ will be updated in each iteration step. To keep this discussion concise, these definitions are formally introduced beforehand:

- $\mathbf{R}^{\mathrm{C}}{}_{(I\cdot J\times K)}$ is a matrix constructed by unfolding the K slices of $\mathbf{R}$ in the XY-plane containing the elements

$$R^{C}_{(j-1)I+i,k} = R_{i,j,k}$$

- $\mathbf{C}_{(I\cdot \mathrm{J}\times \mathrm{N})}$ is formed from the N columns of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ with elements

$$C_{(j-1)I+i,n} = \hat{X}_{i,n}\cdot \hat{Y}_{j,n} \tag{53}$$

- $\mathbf{R}^{\mathrm{B}}{}_{(I\cdot K\times J)}$ is a matrix constructed by unfolding the J slices of $\mathbf{R}$ in the XZ plane containing the elements

$$R^{B}_{(k-1)I+i,\,j} = R_{i,j,k}$$

- $\mathbf{B}_{(I\cdot \mathrm{K}\times \mathrm{N})}$ is formed from the N columns of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ with elements

$$B_{(k-1)I+i,n} = \hat{X}_{i,n} \cdot \hat{Y}_{j,n} \tag{54}$$

- $\mathbf{R}^{\mathrm{A}}{}_{(J\cdot K\times I)}$ is a matrix constructed by unfolding the I slices of $\mathbf{R}$ in the YZ plane containing the elements

$$R^{A}_{(k-1)J+j,i} = R_{i,j,k}$$

- $\mathbf{A}_{(J\cdot \mathrm{K}\times \mathrm{N})}$ is formed from the N columns of $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Z}}$ with elements

$$A_{(k-1)J+j,n} = \hat{Y}_{j,n} \cdot \hat{Z}_{k,n} \tag{55}$$

Step 0: Initial guess of the $\mathbf{X}$ and $\mathbf{Y}$ starting profiles—this can be random numbers (80), or an eigenproblem based algorithm like the Direct Trilinear Decomposition (DTLD) (81), or a priori information about the samples. From these two matrices the $\mathbf{Z}_{(K\times \mathrm{N})}$ profiles will be calculated in Step 1 of the algorithm.

Step 1: Employing equation 53, updated estimates of the Z-way data are determined by solving:

$$\mathbf{R}_{\mathrm{C}} = \mathbf{C} \cdot \mathbf{Z}^{\mathbf{T}}$$

This is done by a multivariate least-squares fit (eq. 30) (the section Multivariate Linear Regression) $\hat{\mathbf{Z}}^{\mathrm{T}} = (\mathbf{C}^{\mathrm{T}} \cdot \mathbf{C})^{-1} \cdot \mathbf{C}^{\mathrm{T}} \cdot \mathbf{R}_{\mathrm{C}}$ or by using the pseudoinverse (eq. 5) (the section Notation and Fundamental Mathematical Tools):

$$\hat{\mathbf{Z}}^{\mathbf{T}} = \mathbf{C}^{+} \cdot \mathbf{R}_{\mathrm{C}} \tag{56}$$

In the remainder only the pseudoinverse will be mentioned.

Step 2: Employing equation 54 updated estimates of the Y-way data are determined by solving:

$$\begin{aligned} \mathbf{R}_{\mathrm{B}} &= \mathbf{B} \cdot \mathbf{Y}^{\mathbf{T}} \\ \hat{\mathbf{Y}}^{\mathbf{T}} &= \mathbf{B}^{+} \cdot \mathbf{R}_{\mathrm{B}} \end{aligned} \tag{57}$$

Step 3: Employing equation 55 updated estimates of the X-way data are determined by solving:

$$\begin{aligned} \mathbf{R}_{\mathrm{A}} &= \mathbf{A} \cdot \mathbf{X}^{\mathbf{T}} \\ \hat{\mathbf{X}}^{\mathbf{T}} &= \mathbf{A}^{+} \cdot \mathbf{R}_{\mathrm{A}} \end{aligned} \tag{58}$$

Step 4: Update $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ by using the new estimates $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, and $\hat{\mathbf{Z}}$ in equations 53–55, respectively.

Step 5+: The algorithm proceeds iteratively, cycling through equations 56–58, until the convergence criterion is satisfied.

Two more topics remain to be discussed: The initialization of $\mathbf{X}$ and $\mathbf{Y}$ (Step 0) as well as the stopping criterion of the iteration (Step 5).

The PARAFAC algorithm is sensitive to the starting guess of the solution for $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$. This results from PARAFAC often becoming trapped in local minima and, hence, not converging to the global optimum least squares solution. Furthermore, the PARAFAC algorithm can become delayed in "swamps" far from the optimum solution (82). Although this markedly increases the analysis time, when employing a random starting value, multiple initial guesses should be considered. The solution for each starting value will be different; however, if all or most of the solutions are similar, it is safe to assume that PARAFAC has converged to near the global optimal solution. The convergence time for PARAFAC can be improved by initializing the algorithm with guesses near the optimal solution. These guesses can come from DTLD or reference spectra of species either known or highly suspected to be in the data set. Care should be employed when utilizing the DTLD solutions since DTLD often yields significant imaginary components in predicting X- and Y-way factors. The problems caused by initializing PARAFAC with imaginary components can be circumvented by employing the real components of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ from DTLD or the absolute values of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ from DTLD.

Two popular convergence criteria for the PARAFAC algorithm are based on changes in the residuals (unmodeled data) between successive iterations and changes in the predicted profiles between successive iterations. In the first case, the algorithm is terminated when the root average of the squared residuals between successive iterations agree to within an absolute or relative tolerance, say 10^{-6}. While such fit based stopping criteria are conceptually easy to visualize, a faster method for determining convergence relies on the correlation between the predicted X-, Y-, and Z-way profiles between successive iterations. When the product of the cosines between successive iterations in the X-, Y-, and Z-modes approach arbitrarily close to 1, say within 10^{-6}, the algorithm is terminated. The cosine in the X way is determined by unfolding the $I \times N$ matrices $\hat{\mathbf{X}}_{\text{old}}$ and $\hat{\mathbf{X}}_{\text{new}}$ into a column vectors $\hat{\mathbf{x}}_{\text{old}}$ and $\hat{\mathbf{x}}_{\text{new}}$. The cos θ_X is defined as

$$\cos\,\theta_X = \frac{\mathbf{x}_{\text{old}}\mathbf{x}_{\text{new}}}{\sqrt{(\mathbf{x}_{\text{old}}\mathbf{x}_{\text{old}})(\mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}})}} \tag{59}$$

The other two terms, cos θ_Y and cos θ_Z are defined equivalently. Convergence in all three modes is implied, if

$$\cos\,\theta_X \cdot \cos\,\theta_Y \cdot \cos\,\theta_Z > 1 - 10^{-6}$$

since at least cos $\theta > 1-10^{-6}$ is obtained for all X, Y, and Z ways.

Mitchell and Burdick sight, besides speed, an additional benefit to correlation based convergence (82). In cases when two factors are highly correlated in

one ore more of the three ways, ALS methods may become mired in "swamps" where the fit of the model changes slightly but the correlation between the predicted X, Y, and Z ways change significantly between successive iterations. After many iterations the ALS algorithm will then emerge from the "swamp" and the residuals and estimated profiles will then both rapidly approach the optimum. Hence, correlation based convergence is more resistant to inflection points in the error response surface when optimizing the model.

4.3. Tucker3 Models. The generalization of the PARAFAC model is the Tucker3 model (83,84). The PARAFAC model is intrinsically linear model and straightforward application thus assumes linear interactions and behavior of the samples. While many of the systems of interest to chemists contain nonlinearities that violate the assumptions of the models, the PARAFAC model forms an excellent starting point from which many subsidiary methods are constructed to incorporate nonlinear behavior into calibration models constructed from three-way data collected with hyphenated methods. The trilinear model is actually a specific case of the Tucker3 model. The Tucker3 model is best understood by viewing a graphical representation such as in Fig. 13. A data cube $\mathbf{R}$ is decomposed into three sets of factors, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}\hat{\mathbf{z}}$, as with PARAFAC. However, the Tucker3 model differs from the PARAFAC model in two key ways. The number of factors N in each way of the Tucker3 model is not constrained to be equal. Also, the Tucker3 model employs a small core cube, $\mathbf{C}$, that governs the interactions among the factors. A non zero element at the pth, qth, rth position of the core $\mathbf{C}$ dictates an interaction between the pth factor in the X way, the qth factor in the Y way, and the rth factor in the Z way. This permits modeling of two or more factors that might have, eg, the same chromatographic profile but different spectral and concentration profiles (85,86). If there are the same number of factors in each way, and $\mathbf{C}$ is constrained to only have nonzero elements on the super diagonal, then the Tucker3 model is equivalent to the PARAFAC model.

One alternating least-squares algorithm for estimating the parameters of the Tucker3 model is Tuckals (for TUCK Alternating Least Squares). This iterative Tuckals algorithms proceeds similarly to the PARAFAC/CANDECOMP algorithm except instead of cycling through three sets of parameters, four sets of parameters must be successively updated, $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, $\hat{\mathbf{Z}}$, and $\mathbf{C}$. However, where with PARAFAC just the number of factors in the model N needs to be preassumed, with Tucker3 the three dimensions of the core array P, Q, and R, need to be assumed.

4.4. Solution Constraints. ALS algorithms are more flexible than rank annihilation based algorithms (87) since constraints (cf. the section Multivariate Curve Resolution) can be placed onto the solutions derived from ALS methods. The ALS algorithms implicitly constrain the estimated profiles to lie in the real space. Rank annihilation methods may fit factors with imaginary components to the data. Ideally, constraints are not needed for ALS to achieve accurate, meaningful concentration and spectral profile estimates. However, the presence of slight nonlinear interactions among the true underlying factors, of highly correlated factors, or of low signal to noise in the data will often result in profile estimates that are visually unsatisfying and large quantitative errors are derived from the model. These effects can often be minimized by employing constraints to the solutions that are based on a priori knowledge or assumptions

of the data structure, eg, prior knowledge of sample concentrations or spectral profile characteristics.

Perhaps the most common constraint consciously placed on the PARAFAC or Tucker3 models is nonnegativity. When one of the modes represents concentrations, chromatographic profiles, or in many cases spectra, constraining the solutions to yield only nonnegative profile estimates often improves the quantitative and qualitative accuracy of the models. Care should be taken when applying nonnegativity constraints since some spectral effects, such as absorbance and quenching in fluorescence, can be manifested, detected, and modeled as negative profiles. Nonnegative estimates of the three-way profiles can be obtained by replacing the least squares update of any given profile with the nonnegative least squares (NNLS) solution that is well defined in the mathematics literature (88). The method described in Ref. 88 is readily available as a Matlab function. The downside of this method is that it is numerically intensive compared to computing the regular least-squares solution for each update.

A second constraint often applied in three-way calibration of chromatographic data is unimodality. This constraint exploits the knowledge that chromatographic profiles have exactly one maximum. Unlike NNLS, there is now method to calculate the true unimodal least-squares update during each iteration. Instead a search algorithm must be implemented that finds the maximum of each profile and assures that from that maximum all values are monotonically decreasing.

The third common constraint is based on a priori knowledge of the three-way profiles. In this case, the known relative concentrations of the standards or the known spectral profiles of one or more components can be fixed as part of the solution. In the Tucker3 model, it is common to restrict some of the potential interactions between factors when they are known not to exist. Care must be employed when applying fixed values to the solutions as the scaling of the factors must still be taken into account.

5. Selected Topics

5.1. Background Spectrum Correction. Background correction methods are often employed in spectroscopic applications to remove broad features from the data set. These features hinder calibration as a large source of variance compared to the analyte or as a seemingly random source of variance that consumes many factors in the model. Examples include fluorescence background in Raman spectroscopy and scattering backgrounds in near-ir reflectance spectroscopy.

Simple efforts at background correction include derivatives, polynomial curve fitting, and Fourier Transform (FT) filtering (89). Derivatives remove the portion of a background that can be modeled by a low order polynomial. Taking the first derivative of a spectrum removes the baseline offset. The second derivative removes the linear approximation of the background (and the analyte signal). However, in spite of digital filters for simultaneously smoothing the data while calculating the derivatives (90), the S/N rapidly declines with each derivatization. Polynomial curve fitting is useful when there are regions of the spectra

that contain only background variance. These regions must be distributed across the entire spectrum such that the background can be modeled. The FT filtering removes both low and high frequency variance across the spectrum. It is assumed that the lowest frequency signal is the background and the highest frequency signal is random instrumental errors. Problems may occur with FT filtering due to poorly chosen apodization functions applied to the signal or insufficient ability to distinguish between the signal and the background. This will lead to distortion of the analyte signal.

Multiplicative scatter correction (MSC) was developed to reduce the effect of scattered light on diffuse reflection and transmission NIR spectra (91,92). This method has also shown utility as a means of removing varying background spectra with nonscattering origins. Consequently, MSC sometimes appears as multiplicative signal correction. The basic application of MSC is presented here. However, a more advanced version of MSC exists that assumes a unique scattering model for different regions of the spectra (93).

Scattering theory states that scattering should have a multiplicative effect on reflection (and transmittance) spectra. That is, the observed spectra will contain a broad, changing background from differential scattering at each wavelength. In Fig. 15**a** it is apparent that the largest source of variance within the NIR reflectance spectra of the 40 cornflour samples is derived from scattering. Assuming a multiplicative model for the scattering, the scattering profile in a spectrum can be deduced from a plot of the spectrum of a standard scatterer versus a given spectrum at each wavelength. And ideal 'standard' would have no NIR absorbance (or transmittance) features; however, the mean spectrum from a collection of similar samples will suffice. Fig. 15**b** presents the plot of the intensity of each wavelength for the mean of 40 calibration spectra versus two of the individual calibration spectra. Note that one is scattering more than the average spectrum and one is scattering less than the average spectrum. The plot for each of these two samples lies about a line with a little variation around the line. The difference between each sample and the best fit line through each sample in Figure 15**b** can be interpreted as the chemical signal and the best fit line gives the spectrum of the scattering in the sample. Consequently, the scattering is determined by regressing each spectrum onto the mean spectrum, where the scattering at the jth wavelength of a sample can be modeled by

$$x_j = a + b\bar{x}_j + \epsilon_j \tag{60}$$

with a and b being constant for all J wavelengths in the sample. The scatter corrected data is determined by the scaled deviations about the regression

$$x_{j,\mathrm{MSC}} = (x_{j,raw} - a)/b \tag{61}$$

The corrected spectra for the 40 calibration NIR cornflour samples are shown in Figure 15c. For correction of future samples, the mean of the calibration set may be employed as the scatter standard. Figure 15d shows the corrected spectra of 20 cornflour spectra that were not included in the calibration set. Evident from these figures is that the spectral features are not distorted by

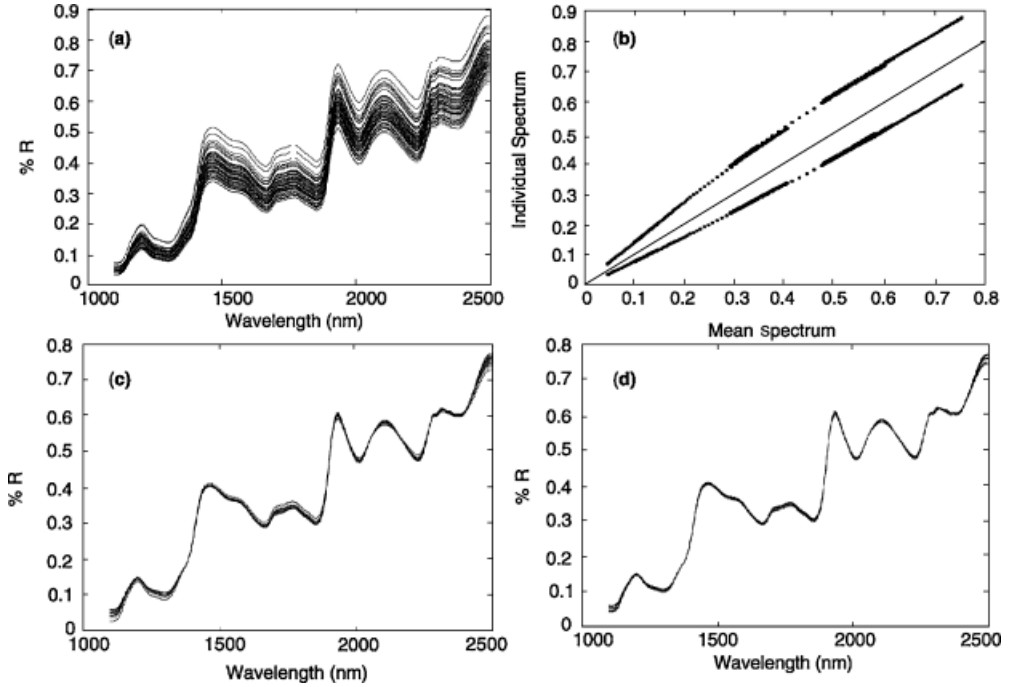

Fig. 15. Demonstration of multiplicative scatter correction. **(a)** The major source of variance for the 40 NIR cornflour spectra is due to scattering affecting the spectral baseline. **(b)** Relationship between two spectra and the mean of the 40 spectra. **(c)** MSC applied to calibration data: corrected spectra are found by the residual of the 40 spectra after regression against the mean of the 40 spectra. **(d)** MSC applied to future data: corrected spectra are found by the residual of the 20 future spectra after regression against the mean of the 40 calibration spectra.

MSC contrasted to scatter correction by calculating the second derivative of each spectrum.

A different approach explicitly including drifts into the calibration was proposed to artificially extend the set of PCs with the so-called pseudo PCs (94). Advantage is taken from the fact that background drifts are usually very broad compared to the more localized absorption features. These pseudo-PCs have been defined to be polynomials up to a user selectable order, however, other linear independent functions could be used, too. It was shown that this combined set of PCs and pseudo PCs is able to determine considerably improved concentration results from highly drift affected spectra compared to a conventional PCR. An example utilizing uv derivative spectroscopy of aqueous samples (95) is given in Figure 16. Several weeks after performing the calibration the zero-point concentrations of three aromatics hydrocarbons have been monitored $>$13 h. Since considerable drifts occurred due to an instable uv light source, the concentration errors without drift correction equal 10% of the measurement range. Most of these concentration errors could be removed based on pseudo-PCs.

An alternative (33) to pseudo-PCs utilizes a similar idea: Polynomials are fitted to the regular PCs and subtracted from them. In this case drift effects,

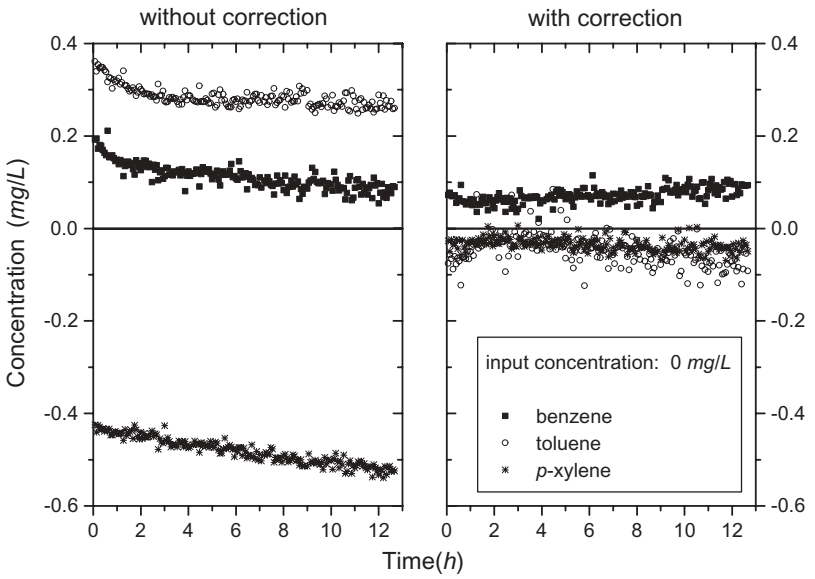

Fig. 16. Comparing the concentration zero points (input 0 mg/L) of three aromatic hydrocarbons dissolved in water (measurement range 0–5 mg/L) obtained from uv derivative spectroscopy (22) without and with drift corrections by means of pseudoprincipal components (94).

which can be modeled by polynomials, are orthogonal to the PCs. This is due to the fact the "corrected" PCs, ie, original PCs minus fit polynomials, are the residuals of these fits. Hence, polynomial like drifts up to the considered order are orthogonal to the corrected PCs (cf. section 3 in Supplementary Topics, chapter Multivariate Linear Regression) and cannot influence the concentration results. The pseudo-PCs method extracts additional information, ie, an estimate of the drift spectrum. The approach fitting polynomials to the PCs is computational less expensive since it is done just once during the calibration. This can be advantageous is computation resources are limited.

5.2. Instrument Standardization. One practical concern with multivariate calibration and prediction is the transport and stability of the calibration models. Ideally, a calibration model can be constructed in the laboratory on a bench-top instrument, then the model can be applied to many similar instruments in the field. Also, once a model is successfully transferred to the field, it will be robust to changes in instrumental sensitivity and alignment. Of course, the goal of a universal transferable and robust instrument–model has not been achieved. Seemingly identical spectrometers have slight wavelength resolution, and sensitivity differences that can prohibit reliable distribution of the calibration model among numerous instruments. Also, time-dependent instrumental drift eventually can render the calibration model obsolete for whichever instrument the model was constructed.

Individual calibration of each instrument is not an acceptable solution to the problem of model distribution. Calibration may be an expensive, time-

consuming task when many calibration samples are needed, the calibration samples are not readily transportable, or the instrument is not easily accessible in the process stream. Concurrently, it is also unacceptable to repeat an entire calibration procedure whenever there are minor changes in the instrumental character.

Instrumental standardization (96–99) strives to solve the problems derived from instrumental differences when constructing one calibration model for multiple instruments. The instrumental standardization philosophy is to construct the best model possible on one instrument then to build a second model that will transform the spectra from other instruments to appear as if they were recorded on the first instrument. Usually, this transfer function can be reliably calculated with less effort.

One standardization method popular in the literature is Piecewise Direct Standardization (PDS) (99–102). With PDS, a set of transfer samples is analyzed on both the original instrument and the instrument to which the calibration model will be transferred. It is best if the transfer samples are a subset of the calibration set; however, other surrogate samples may be employed. A separate transfer function is determined for each wavelength in the spectra by least squares regression using neighboring wavelengths as the independent variables. That is, a local subset of variables measured on the second instrument is employed to build a model that predicts what each measurement would have been if it were measured with the first instrument. This method accounts for shifts and intensity changes over a small spectral window. The drawback of PDS is that success of the standardization is dependent on choice of the transfer samples. The transfer samples must be identical when measured on each instrument and the set of samples must span the space of all encountered spectral changes between the two instruments. Therefore, the choice and number of transfer samples must be optimized by the analyst.

A more useful method of standardization would not require transfer samples to be analyzed. There have been two approaches to this problem. When it can be safely assumed that the only spectral shifts (ie, wavelength or retention time) occur a PCA based method of standardization may be employed (103,104). The spectral (or time) indexes are shifted such that the projection of each sample into the PC space defined by the original instrument is optimized. A more general method based loosely on MSC has also demonstrated success when there are relatively minor performance differences between the original and second instruments (105,106). Here a local selection of wavelengths from each spectrum is regressed against the mean spectrum to build a transfer function. Consequently, the spectra from the second instrument are not transformed to look like the spectra from the first instrument. Instead, spectral responses from both instruments are transformed to lie in a common multidimensional space.

5.3. Optical Computation. Most spectrometer concepts include moving parts like interferometers or scanning gratings. Such moving parts, however, limit the ruggedness of a field analyzer and the time resolution of the concentration runs. The strong point of such spectrometers combined with chemometric software packages is their versatility. For many applications this is not needed, though. In process analytics, eg, a measurement device is usually applied to one very specific task not needing versatility at all. Mechanical stability and good time resolution is of greater importance. In order to overcome both mentioned

drawbacks it was proposed to design so-called multivariate optical elements (MOE) (107–111). MOE are specially designed interference filter in a beam splitter arrangement. The light is emitted from the source, transmitted through the sample and split by the MOE into a transmission part and a reflection part. The idea is to design the transmission spectrum of the interference filter such that it is an imprint of a PC onto a transmission offset. This offset is necessary to enable positive and negative features of the PC. The transmission of light through such an interference filter followed by generating a signal in the detector element resembles the projection of a measured spectrum onto a PC. The transmission through the filter replaces the multiplication of loadings with measurement points of a spectrum; the detector integrating over all wavelengths replaces the summation part of calculating a scalar product. What is left to do is subtracting the transmission offset mentioned above from the results. For this purpose, the interference filter had been placed in a 45° arrangement. Then the transmission and the reflection spectrum can be measured by means of two detectors arranged in perpendicular lines of sight. Calculating the difference signal of both detectors cancels the transmission offset.

5.4. Artificial Neural Networks Combined with Variable Selection. The measurement technique surface plasmon resonance (spr) (112) is sensitive for analyzing refractive indexes of liquids or vapors. Since the matrix, water, eg, and a dissolved analyte have different refractive indices the refractive index of a sample is concentrations dependent. A change of the samples' refractive index is measured by a highly nonlinear wavelength shift of the plasmon absorbance. However, since only one property of a sample, ie, the refractive index, is measured, binary or ternary mixtures cannot be investigated without experimental adjustments. A polymer coating of the spr sensor head was proposed resulting in different, time-dependent enrichment or desorption processes depending on the molecule size. That means different analytes cause a time- and analyte-dependent change of the spr spectra. This idea was applied in references 113 and 114 to measure binary samples of two chlorofluorocarbons and ternary mixtures of alcohols, respectively. Time series of spr spectra monitoring different sorption–desorption behaviors of the analytes were evaluated then by means of a neural network. Inputs into the neural net are the wavelength shifts measured at preselected points of times (variables). Usually, one wants a high time resolution, ie, a large number of variables, in order to capture fast and similar responses and not to lose information. However, there are several disadvantages of using a lot of information like hiding meaningful variables by irrelevant variables or overfitting. Furthermore, danger to change the correlation is increased with the number of variables and many variables mean increased computation time for the neural net training.

Full-connected neural networks employing a large number of variables are prone to overfitting (113). Hence, so-called growing neural nets were applied in Refs. 113 and 114 resulting in sparse, nonuniform structures optimized to a specific problem. The growing of a feedforward back-propagation network is started with one not having hidden neurons or connections. Then one neuron is added at a time, which is connected to one output neuron and two other neurons such that the error decrease regarding the training data is maximized. However, the outcome of this procedure is still dependent on the way the calibration data are split

into training and monitoring data. To overcome this ambiguity two strategies have been proposed: (*1*) To grow neural networks on a rather large number of different training–monitoring sets in parallel. Ranking the variables considering their importance follows this. The number of net growings in which it was selected determines the importance of a variable. The final network is grown in a second step by iteratively adding variables to it in order of decreasing importance until the addition of a variable does not increase the predictability of this final network anymore. (*2*) A certain training–monitoring set is defined and a small number of nets are grown with different initial weights. The best of these is chosen to be the initial topology for the second, different training–monitoring set. Again a number of networks with different initial weights are grown and the best one is selected. This is continued until the topology of the best network does not change anymore. It was found that the procedure (*2*) resulted in better generalizations. Application of such grown networks to binary mixtures resulted in convincing concentration prediction.

As an alternative for variable selection, a genetic algorithm (115–118) has been used in Ref. 114 for selecting the optimum subset for neural networks based on the procedure (*1*) discussed above. A genetic algorithm is applied to a rather large number of different training–monitoring sets in parallel resulting in a set of neural nets. Again, the variables are ranked in decreasing importance and added in the second step one after the other to a final network until the predictability is not improved anymore. The prediction of concentrations could be considerably improved using five selected variables compared to using all 50.

BIBLIOGRAPHY

"Chemometrics" in *ECT* 4th ed., pp. 837–869, by Deborah Illman, University of Washington; "Chemometrics" in *ECT* (online), posting date: December 4, 2000, by Deborah Illman, University of Washington.

1. H. Martens and T. Næs, *Multivariate Calibration*, 2nd ed., John Wiley & Sons, Inc., New York, 1991.
2. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer-Verlag: New York, 2002.
3. E. Malinowski, *Factor Analysis in Chemistry*, 3rd ed., John Wiley & Sons, Inc., New York, 2002.
4. N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., J. Wiley & Sons, Inc., New York, 1998.
5. A. Sen and M. Srivastava, *Regression Analysis—Theory, Methods, Applications*, Springer Verlag, New York, 1990.
6. *Anal. Bio. Chem.* **373**(6), (2002). Special review issue.
7. D. Burdick, *Chemom. Intell. Lab. Syst.* **28**, 229 (1995).
8. S. Wold, *Chemom. Intell. Lab. Syst.* **30**, 109 (1995).
9. K. S. Booksh and B. R. Kowalski, *Anal. Chem.* **66**, 782A (1994).
10. S. D. Brown, *Chemom. Intell. Lab. Syst.* **30**, 49 (1995).
11. P. Hopke, *Anal. Chim. Acta* **500**(1–2), 365 (2003).
12. P. Geladi, *Chemom. Intell. Lab. Syst.* **7**, 11 (1989).

13. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
14. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes in C++*: The Art of Scientific Computing, 2nd ed., Cambridge University Press, Cambridge, 2002.
15. J. K. Taylor, *ChemTech.* **16**, 763 (1986).
16. G. L. Long and J. D. Winefordner, *Anal. Chem.* **55**, 712A (1983).
17. D. L. Massart, B. G. M. Vandegonste, L. M. C. Buydens, S. DeJong, P. J. Lewi, and J. Smeyers, *Handbook of Chemometrics and Qualimetrics*, Elsevier, Amsterdam, The Netherlands, 1997.
18. R. Bro and A. Smilde, *J. Chemom.* **17**, 16 (2003).
19. R. J. Pell, M. B. Seasholtz, and B. R. Kowalski, *J. Chemom.* **6**, 52 (1992).
20. M. B. Seasholtz, B. R. Kowalski, *J. Chemom.* **6**, 103 (1992).
21. D. M. Haaland and E. V. Thomas, *Anal. Chem.* **60**, 1193 (1988).
22. F. Vogt, U. Klocke, K. Rebstock, G. Schmidtke, V. Wander, and M. Tacke, *Appl. Spec.* **53**, 1352 (1999).
23. W. Egan, W. Brewer, and S. Morgan, *Appl. Spectrosc.* **53**, 218 (1999).
24. F. Vogt, M. Karlowatz, M. Jakusch, and B. Mizaikoff, *Analyst* **128**, 397 (2003).
25. J. Mandel, *Am. Stat.* **36**, 15 (1982).
26. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta* **185**, 1 (1986).
27. A. Lorber, L. E. Wangen, and B. R. Kowalski, *J. Chemom.* **1**, 19 (1987).
28. R. Marbach and H. M. Heise, *TRAC* **11**, 270 (1992).
29. M. Stone and R. J. Brooks, *J. R. Stat. Soc. B.* **52**, 337 (1990).
30. S. de Jong, *Chemom. Intell. Lab. Syst.* **18**, 251 (1993).
31. R. Manne, *Chemom. Intell. Lab. Syst.* **2**, 187 (1987).
32. F. Vogt and B. Mizaikoff, *J. Chemom.* **17**, 346 (2003).
33. F. Vogt, H. Steiner, and B. Mizaikoff, *Appl. Spec.* 2003, submitted for publication.
34. F. Vogt and B. Mizaikoff, *J. Chemom.* **17**, 225 (2003).
35. F. Vogt and B. Mizaikoff, *Anal. Chem.* **75**, 3050 (2003).
36. T. Naes and T. Isaksson, *Appl. Spectrosc.* **46**, 34 (1992).
37. K. S. Johnston, S. S. Lee, and K. S. Booksh, *Anal. Chem.* **69**, 1844 (1997).
38. W. S. Cleveland and S. J. Devlin, *J. Am. Stat. Assoc.* **83**, 596 (1988).
39. T. Naes, T. Isaksson, and B. R. Kowalski, *Anal. Chem.* **62**, 664 (1990).
40. Z. Wang, T. Isaksson, and B. R. Kowalski, *Anal. Chem.* **66**, 249 (1994).
41. L. Brieman and J. H. Friedman, *J. Am. Stat. Assoc.* **80**, 580 (1985).
42. I. E. Frank and S. Lanteri, *Chemom. Intel. Lab. Syst.* **3**, 301 (1988).
43. J. H. Friedman and W. Steutzle, *J. Am. Stat. Assoc.* **76**, 817 (1981).
44. M. M. C. Ferreira, W. C. Ferreira, and B. R. Kowalski, *J. Chemom.* **10**, 11 (1996).
45. S. Winsberg and J. O. Ramsey, *Psychometrika* **48**, 575 (1984).
46. I. E. Frank, *Chemom. Intel. Lab. Syst.* **8**, 109 (1990).
47. S. Wold, *Chemom. Intel. Lab. Syst.* **14**, 71 (1992).
48. J. H. Friedman, *Ann. Stat.* **19**, 199 (1991).
49. S. Sekulic and B. K. Kowalski, *J. Chemom.* **6**, 199 (1992).
50. J. Freeman and D. Skapura, *Neural Networks—Algorithms, Applications and Programming Techniques*, Addison-Wesley Publishing Company, New York, 1991.
51. S. Sekulic, M. B. Seasholtz, Z. Wang, B. R. Kowalski, S. E. Lee, and B. R. Holt, *Anal. Chem.* **65**, 835A (1993).
52. I. E. Frank, *Chemom. Intel. Lab. Sys.* **27**, 1 (1995).
53. K. S. Booksh and B. R. Kowalski, *Anal. Chim. Acta* **348**, 1 (1997).
54. M. Gerritsen, J. A. van Leeuwen, B. G. M. Vandeginste, L. Buydens, and G. Kateman, *Chemom. Intel. Lab. Sys.* **15**, 171 (1992).
55. M. B. Seasholtz and B. R. Kowalski, *Anal. Chim. Acta* **277**, 165 (1993).

56. Z. Wang, J. N. Hwang, and B. R. Kowalski, *Anal. Chem.* **67**, 1497 (1995).
57. W. H. Lawton and E. A. Sylvestre, *Technometrics* **13**, 617 (1971).
58. H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbuhler, *Talanta* **32**, 1133 (1985).
59. Y-Z Liang, R. Manne, and O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* **14**, 155 (1992).
60. W. Windig, *Chemom. Intell. Lab. Syst.* **16**, 1 (1992).
61. R. Bro and S. De Jong, *J. Chemom.* **11**, 393 (1997).
62. R. Tauler, I. Marques, and E. Casassas, *J. Chemom.* **12**, 55 (1998).
63. W. C. Bell, K. S. Booksh, and M. L. Myrick, *Anal. Chem.* **70**, 332 (1998).
64. S. P. Gurden, R. G. Bereton, and J. A. Groves, *Chemom. Intell. Lab. Syst.* **23**, 123 (1994).
65. E. A. Sylvestre, W. H. Lawton, M. S. Maggio, *Technometrics* **16**, 353 (1974).
66. R. I. Shrager, *Chemom. Intell. Lab. Syst.* **1**, 59 (1986).
67. J. Saurina, S. Hernandez-Cassou, R. Tauler, and A. Izquierdo-Ridorsa, *J. Chemom.* **12**, 183 (1998).
68. P. Gemperline and E. Cash, *Anal. Chem.* **75**, 4236 (2003).
69. T. Hirschfeld, *Anal. Chem.* **52**, 297A (1980).
70. R. Bro, *Chemom. Intell. Lab. Syst.* **38**, 149 (1997).
71. C. Andersen and R. Bro, *J. Chemom.* **17**, 200 (2003).
72. N. Faber, R. Bro, and P. Hopke, *Chem. Intell. Lab. Syst.* **65**, 119 (2003).
73. R. Bro and H. Kiers, *J. Chemom.* **17**, 274 (2003).
74. A. Muroski, K. Booksh, and M. Myrick, *Anal. Chem.* **68**, 3534 (1996).
75. R. JiJi, G. Cooper, and K. Booksh, *Anal. Chim. Acta* **397**, 61 (1999).
76. R. JiJi, G. Andersson, and K. Booksh, *J. Chemom.* **14**, 171 (2000).
77. P. M. Kroonenberg, *Three-mode Principal Component Analyses. Theory and Applications*, DSWO Press, Leiden, 1983.
78. R. A. Harshman, UCLA Working Paper on Phonetics, Vol. 16, 1970, pp. 1–84.
79. C. Andersson and R. Bro, *Chemo. Intell. Lab. Syst.* **52**, 1 (2000).
80. R. A. Harshman and M. E. Lundy, "The PARAFAC model for Three-Way Factor Analysis and Multidimensional Scaling," in H. G. Law and co-workers, eds., *Research Methods for Multimode Data Analysis*, Praeger, New York, 1984.
81. E. Sanchez and B. R. Kowalski, *J. Chemom.* **4**, 29 (1990).
82. B. C. Mitchell and D. S. Burdick, *J. Chemom.* **6**, 155 (1992).
83. V. Pravdova, F. Estienne, B. Walczak, and D. L. Massart, *Chemom. Intell. Lab. Syst.* **59**, 75 (2001).
84. J. Ten Berge and A. Smilde, *J. Chemom.* **16**, 609 (2002).
85. A. K. Smilde, R. Tauler, J. M. Henshaw, L. W. Burgess, and B. R. Kowalski, *Anal. Chem.* **66**, 3345 (1994).
86. A. K. Smilde, Y. Wang, and B. R. Kowalski, *J. Chemom.* **8**, 21 (1994).
87. C.-H. Ho, G. D. Christian, and E. R. Davidson, *Anal. Chem.* **50**, 1108 (1978).
88. C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood, Cliffs, New York, 1974.
89. Q. Ding, G. W. Small, and M. A. Arnold, *Appl. Spectros.* **53**, 402 (1999).
90. A. Savitzky and M. Golay, *Anal. Chem.* **36**, 1627 (1964).
91. P. Geladi, D. MacDougall, and H. Martens, *Appl. Spectrosc.* **39**, 491 (1985).
92. C. E. Miller, S. A. Svendsen, and T. Naes, *Appl. Spectrosc.* **47**, 346 (1993).
93. T. Isaksson and B. R. Kowalski, *Appl. Spectros.* **47**, 702 (1993).
94. F. Vogt, K. Rebstock, and M. Tacke, *Chemom. Intell. Lab. Syst.* **50**, 175 (2000).
95. F. Vogt, M. Tacke, M. Jakusch, and B. Mizaikoff, *Anal. Chim. Acta* **422**, 187 (2000), Erratum: *Anal. Chim. Acta* **431**, 167 (2001).
96. Y. Wang, M. J. Lysaght, and B. R. Kowalski, *Anal. Chem.* **64**, 562 (1992).
97. C. S. Chen, C. W. Brown, and S. C. Lo, *Appl. Spectros.* **51**, 744 (1997).
98. J. Lin, *Appl. Spectrosc.* **52**, 1591 (1998).

99. P. J. Gemperline, J. H. Cho, P. K. Aldridge, and S. S. Sekulic, *Anal. Chem.* **68**, 2913 (1996).
100. Y. Wang, M. J. Lysaght, and B. R. Kowalski, *Anal. Chem.* **64**, 562 (1992).
101. C. S. Chen, C. W. Brown, and S. C. Lo, *Appl. Spectros.* **51**, 744 (1997).
102. J. Lin, *Appl. Spectrosc.* **52**, 1591 (1998).
103. K. S. Booksh, C. M. Stellman, W. C. Bell, and M. L. Myrick, *Appl. Spectros.* **50**, 139 (1996).
104. B. J. Prazen, C. E. Bruckner, R. E. Synovec, and B. R. Kowalski, *J. Microcol. Sep.* **11**, 97 (1998).
105. T. B. Blank, S. T. Sum, S. D. Brown, and S. L. Monfre, *Anal. Chem.* **68**, 2987 (1996).
106. S. T. Sum and S. D. Brown, *Appl. Spectrosc.* **52**, 869 (1998).
107. M. Nelson, J. Aust, J. Dobrowolski, P. Verly, and M. Myrick, *Anal. Chem.* **70**, 73 (1998).
108. O. Soyemi, D. Eastwood, L. Zhang, H. Li, J. Karunamuni, P. Gemperline, R. Synowicki, and M. Myrick, *Anal. Chem.* **73**, 1069 (2001).
109. M. Myrick, O. Soyemi, H. Li, L. Zhang, and D. Eastwood, *Fresenius J. Anal. Chem.* **369**, 351 (2001).
110. M. Myrick, O. Soyemi, J. Karunamuni, D. Eastwood, H. Li, L. Zhang, A. Greer, and P. Gemperline, *Vibrat. Spec.* **28**, 73 (2002).
111. O. Soyemi, F. Haibach, G. Frederick, P. Gemperline, and M. Myrick, *Appl. Spec.* **56**, 477 (2002).
112. K. Johnston, S. Yee, and K. Booksh, *Anal. Chem.* **69**, 1844 (1997).
113. F. Dieterle, S. Busche, and G. Gauglitz, *Anal. Chim Acta* **490**, 71 (2003).
114. F. Dieterle, B. Kieser, and G. Gauglitz, *Chem Intel. Lab. Syst.* **65**, 67 (2003).
115. C. Lucasius and G. Katerman, *Chem Intel. Lab. Syst.* **19**, 1 (1993).
116. C. Lucasius and G. Katerman, *Chem Intel. Lab. Syst.* **25**, 99 (1994).
117. B. Smith and P. Gemperline, *Anal. Chim. Acta* **423**, 167 (2000).
118. R. Leardi, *J. Chemom.* **15**, 559 (2001).

FRANK VOGT
KARL BOOKSH
Arizona State University