Pure Appl. Chem., Vol. 76, No. 6, pp. 1215–1225, 2004.
© 2004 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

ANALYTICAL CHEMISTRY DIVISION*

GUIDELINES FOR CALIBRATION IN ANALYTICAL CHEMISTRY PART 2. MULTISPECIES CALIBRATION

(IUPAC Technical Report)

Prepared for publication by
KLAUS DANZER[1,‡], MATTHIAS OTTO[2], AND LLOYD A. CURRIE[3]

[1]*Department of Inorganic and Analytical Chemistry, Friedrich Schiller University of Jena, Lessingstrasse 8, D-07743 Jena, Germany;* [2]*Department of Analytical Chemistry, Technical University Bergakademie Freiberg. Leipziger Strasse 29, D-09599 Freiberg, Germany;* [3]*National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*

*Membership of the Analytical Chemistry Division during the final preparation of this report (2002–2003) was as follows:

President**:** D. Moore (USA); ***Titular Members*****:** F. Ingman (Sweden); K. J. Powell (New Zealand); R. Lobinski (France); G. G. Gauglitz (Germany); V. P. Kolotov (Russia); K. Matsumoto (Japan); R. M. Smith (UK); Y. Umezawa (Japan); Y. Vlasov (Russia); ***Associate Members*****:** A. Fajgelj (Slovenia); H. Gamsjäger (Austria); D. B. Hibbert (Australia); W. Kutner (Poland); K. Wang (China); ***National Representatives*****:** E. A. G. Zagatto (Brazil); M.-L. Riekkola (Finland); H. Kim (Korea); A. Sanz-Medel (Spain); T. Ast (Yugoslavia).

[‡]Corresponding author

Republication or reproduction of this report or its storage and/or dissemination by electronic means is permitted without the need for formal IUPAC permission on condition that an acknowledgment, with full reference to the source, along with use of the copyright symbol ©, the name IUPAC, and the year of publication, are prominently visible. Publication of a translation into another language is subject to the additional condition of prior approval from the relevant IUPAC National Adhering Organization.

Guidelines for calibration in analytical chemistry Part 2. Multispecies calibration

(IUPAC Technical Report)

Abstract: Calibration in analytical chemistry refers to the relation between sample domain and measurement domain (signal domain) expressed by an analytical function $x = f_s(\mathrm{Q})$ representing a pattern of chemical species Q and their amounts or concentrations x in a given test sample on the one hand and a measured function $y = f(z)$ that may be a spectrum, chromatogram, etc.

Simultaneous multispecies analyses are carried out mainly by spectroscopic and chromatographic methods in a more or less selective way. For the determination of n species Q_i ($i = 1, 2\ldots n$), at least n signals must be measured which should be well separated in the ideal case. In analytical practice, the situation can be different.

CONTENTS

1. INTRODUCTION
2. CLASSICAL MULTIVARIATE CALIBRATION
3. UNCERTAINTY IN MULTIVARIATE CALIBRATION
4. INVERSE CALIBRATION
5. ERROR DIAGNOSIS AND VALIDATION
6. REFERENCES

LIST OF SYMBOLS

N	Number of species
M	Number of signal values (absorbances) used for calibration
P	Number of calibration standards (mixtures used for calibration)
Q_i	Chemical species (elements, ions, compounds), $i = 1 \ldots n$
x_i	Amount (concentration) of species Q_i; $i = 1 \ldots n$
a_{ij}	Calibration coefficients (regression coefficients, sensitivity coefficients), $i = 1, \ldots, n$, $j = 1, \ldots, m$
z_j	Signal positions, e.g., wavelengths at which signal absorbances y_j are measured, $j = 1, \ldots, m$
y_j	Signal absorbances at the signal positions z_j, $j = 1, \ldots, m$
e_j	Uncertainties (errors) of the actual y-measurement, $j = 1, \ldots, m$

Matrices are written in bold italic type, and in capital (upper-case) letters, except for column matrices (column vectors), which are printed in small (lower-case) letters, and scalars, which are written in italic type. Transposed matrices are indicated by superscript T; the transpose of a column matrix is a row matrix. Estimates of matrices, vectors (row or column matrices), or scalars are characterized by ^; e.g., $\hat{\boldsymbol{x}}$ is the estimate of vector $\boldsymbol{x}$.

1. INTRODUCTION

Calibration in analytical chemistry refers to the relation between sample domain and measurement domain (signal domain) expressed by an analytical function $x = f_s(\mathrm{Q})$ representing a pattern of chemical species Q and their amounts or concentrations x in a given test sample on the one hand (see Fig. 1, left-hand side) and a measured function $y = f(z)$ that may be a spectrum, chromatogram, etc. (Fig. 1, right-hand side).

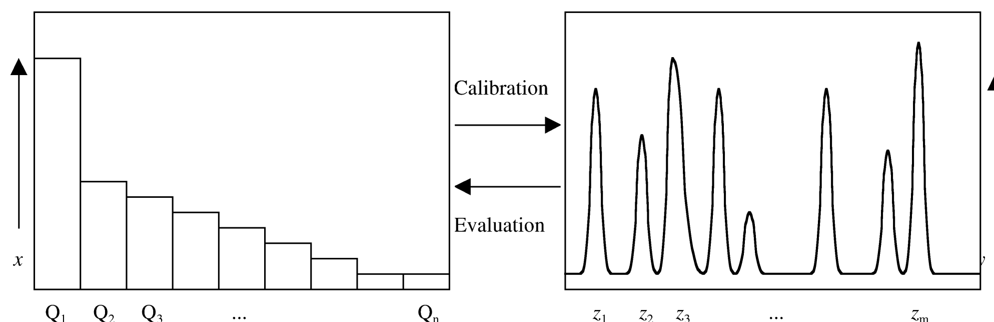

Fig. 1 Relationship between sample domain and signal domain in the case of elemental analysis.

Simultaneous multispecies analyses are carried out mainly by spectroscopic and chromatographic methods in a more or less selective way. For the determination of n species Q_i (i = 1, 2…n), at least n signals must be measured, which should be well separated in the ideal case. In analytical practice, the situation can be different as shown in Fig. 2 (b) and (c) where an illustration is given only for a detail out of the signal domain of Fig. 1, right.

In case (a), each species can be calibrated and evaluated independently from the other. In that fully selective case, the following equation system (1) corresponds to the matrix $\boldsymbol{A}$ in eq. 3:

$$\begin{aligned} y_1 &= a_{10} + a_{11}x_1 + e_1 \\ y_2 &= a_{20} + a_{22}x_2 + e_2 \\ &\vdots \\ y_m &= a_{m0} + a_{mn}x_n + e_m \end{aligned} \tag{1}$$

where the number of measured signal absorbances usually is equal to the number of species, $m = n$. On the other hand, case (b), in which neighboring signals overlap to a certain degree, can be handled by means of multiple linear calibration when the absorbances y_i are additive and related signal maxima can be measured for each species:

$$\begin{aligned} y_1 &= a_{10} + a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n + e_1 \\ y_2 &= a_{20} + a_{21}x_1 + a_{22}x_2 + \cdots a_{2n}x_n + e_2 \\ &\vdots \\ y_m &= a_{m0} + a_{m1}x_1 + a_{m2}x_2 + \cdots a_{mn}x_n + e_m \end{aligned} \tag{2a}$$

or, in matrix form,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \tag{2b}$$

When the preconditions mentioned above are valid, the multispecies system can be calibrated with a high degree of confidence. The uncertainties e_i include both deviations from the model and random errors (noise).

In the case of strongly overlapped signals, see (c) in Fig. 2, multiple linear calibration cannot be used with validity for the following reasons.

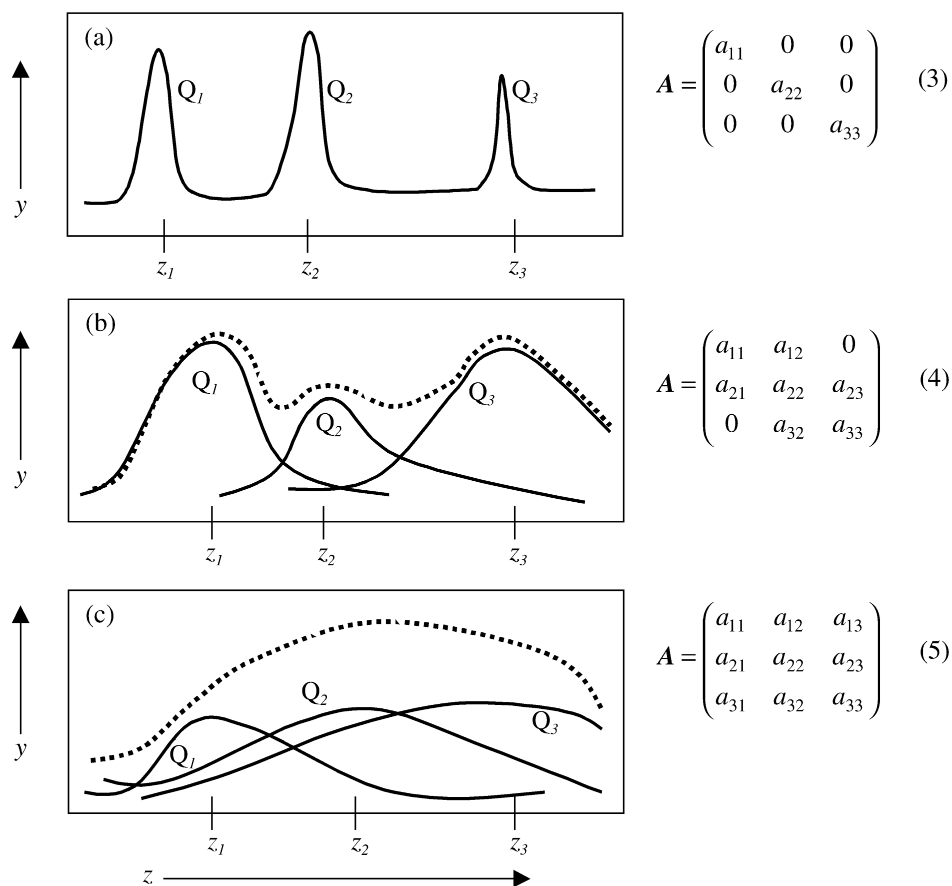

Fig. 2 Evaluation of multispecies analysis in different cases of signal relations [1]: well-separated (a), moderately overlapped (b), and strongly overlapped (c) in form of spectra (left) and relevant matrices 3–5 (right); Q_1, Q_2, Q_3 different species (analytes), z_1, z_2, z_3 wavelengths at which the intensities y_1, y_2, y_3 are measured.

(i) In real analytical systems, not all the sample species are known. In such cases, an alternative possibility may be the inverse calibration model of eq. 2b

$$\boldsymbol{x} = \boldsymbol{Ya} + \boldsymbol{e}_x \tag{6}$$

where $\boldsymbol{a}$ may be a spectral matrix with m given wavelengths for n species mixtures in which other variations like baseline effects must also be included. The vector of sensitivity coefficients can be estimated by

$$\hat{\boldsymbol{a}} = (\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{x} \tag{7}$$

(ii) In principle, for spectra like (c), multicollinearities have to be expected. That means that overlapping signal curves and consequently the resulting sum curve are correlated and the measured absorbances at the respective wavelengths are not independent from each other. Therefore, eq. 7 becomes unstable and other methods of estimation of $\boldsymbol{a}$-values must be used. These methods use overdetermined equation systems as their basis, such as

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \end{pmatrix} \tag{8}$$

instead of eq. 5. The number of wavelengths (sensors, detecting channels), m, is usually much higher than the number of species n (here, $n = 3$). Estimation of the coefficients is then carried out by multivariate calibration.

Depending on whether the spectra $\boldsymbol{Y}$ are calibrated as dependent on concentrations $\boldsymbol{X}$ or conversely, on different methods of multispecies calibration, see Fig. 3, can be used.

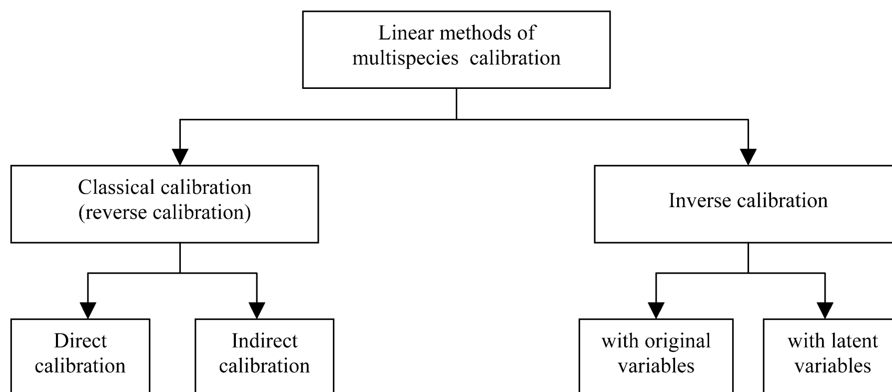

Fig. 3 Multivariate calibration methods.

2. CLASSICAL MULTIVARIATE CALIBRATION

Classical multivariate calibration represents the transition of common single species analysis from one dependent variable (measured value, or measurand) to m dependent variables, e.g., wavelengths or sensors which can be simultaneously included in the calibration model. It is possible to determine $n \geq 1$ species in the analytical system. The classical linear calibration [2] is therefore represented by the generalized matrix relation

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{A} \tag{9}$$

where $\boldsymbol{Y}$ is the $(p \times m)$-matrix of dependent variables (e.g., absorbances at m wavelengths or responses at m sensors), $\boldsymbol{X}$ is the $(p \times n)$-matrix of independent variables (e.g., concentrations of n species), and $\boldsymbol{A}$ is the $(n \times m)$-matrix of the calibration coefficients, often called "sensitivity matrix" [3–5]; p is the number of calibration standards (mixtures), which is identical with the number of spectra or similar measurements. The rows of the matrix correspond to the spectra of the pure species, which can be directly measured or indirectly estimated.

Direct calibration can be applied when the calibration coefficients are known, otherwise—in the case of *indirect calibration*—the calibration coefficients are computed by means of experimentally estimated spectra-concentrations relations.

Classical calibration procedure can only be applied when all the species that contribute to the shape of the spectra are known and can be included into the calibration. Additionally, there is the constraint that no interactions between the analytes and other species (e.g., solvent) or effects (e.g., of temperature) should occur.

The analytical values (concentrations) are estimated by

$$\hat{\boldsymbol{X}} = \boldsymbol{Y}\boldsymbol{A}^{+} \tag{10}$$

where $\boldsymbol{A}^+$ is the so-called Moore–Penrose generalized pseudo inverse

$$\boldsymbol{A}^+ = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} \tag{11}$$

with the same dimensions ($m \times p$) as the transposed matrix.

In case of baseline shift, the sensitivity matrix in eqs. 9 and 10 must be complemented by a vector $\boldsymbol{1}$:

$$\boldsymbol{A} = (\boldsymbol{1}\,\boldsymbol{A}) = \begin{pmatrix} 1 & a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{p1} & \cdots & a_{pm} \end{pmatrix} \tag{12}$$

Instead of the addition of the $\boldsymbol{1}$-vector the calibration data may be centered ($y_i - \bar{y}$ and $x_i - \bar{x}$, respectively). Even if the spectra of the pure species cannot be measured directly then the $\boldsymbol{A}$-matrix can be estimated indirectly from the spectra *provided that all species of the analytical system are known*

$$\hat{\boldsymbol{A}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{13}$$

Note: Instead of the symbol $\boldsymbol{A}$ and the term sensitivity matrix also the symbol $\boldsymbol{K}$ (matrix of calibration coefficients, matrix of linear response constants, and so on) is used. Because of the direct metrological and analytical meaning of the sensitivities a_{ij} in the $\boldsymbol{A}$-matrix, the term *sensitivity matrix* is preferred.

For inversion of the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, it is necessary that a sufficient number of spectra for different concentration steps have been measured. The concentration vectors must vary independently from each other. For this reason, experimental design [8] should be used. In the case that the preparation of samples of defined composition is impossible, then the samples should be selected as representative and as uncorrelated as possible (natural design [3]).

The prediction of analytical values $\boldsymbol{X}$ according to the classical indirect calibration model follows eq. 10

$$\hat{\boldsymbol{X}} = \boldsymbol{Y}\hat{\boldsymbol{A}} \tag{14}$$

The desired independence between the variables of the different analytical signals corresponds directly with the ***selectivity*** of the analytical system [6,7]. In the case of multivariate calibration, the selectivity is characterized by means of the condition number

$$\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|\cdot\|\boldsymbol{A}^{-1}\| \tag{15}$$

where $\|\boldsymbol{A}\|$ is the matrix norm of $\boldsymbol{A}$ and $\|\boldsymbol{A}^{-1}\|$ the norm of the inverse matrix. The matrix norm of $\boldsymbol{A}$ is calculated from $\sqrt{\lambda_{max}}$, the square root of the largest eigenvalue λ_{max}, and the norm of $\boldsymbol{A}^{-1}$ from the reciprocal square root of the lowest eigenvalue λ_{min}:

$$\mathrm{cond}(\boldsymbol{A}) = \sqrt{\lambda_{\mathrm{max}}}\;\frac{1}{\sqrt{\lambda_{\mathrm{min}}}} \tag{16}$$

Equation 14 is valid for exactly determined systems ($m = n$). In the case of overdetermined systems, $m > n$, the condition number is given by

$$\mathrm{cond}(\boldsymbol{A}) = \sqrt{\mathrm{cond}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})} \tag{17}$$

If systems are well conditioned, the selectivity is expressed by condition numbers close to 1.

3. UNCERTAINTY IN MULTIVARIATE CALIBRATION

The evaluation is carried out according to eqs. 10 and 14. The prediction of a row vector $\boldsymbol{x}$ of dimension n from a row vector $\boldsymbol{y}$ of dimension m results from

$$\boldsymbol{x} = \boldsymbol{y}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1} \tag{18}$$

The relative uncertainty for the prediction of the x-values can be estimated by

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \mathrm{cond}\,\boldsymbol{A}\left(\frac{\|\delta\boldsymbol{y}\|}{\|\boldsymbol{y}\|} + \frac{\|\delta\boldsymbol{A}\|}{\|\boldsymbol{A}\|}\right) \tag{19}$$

where $\|\delta\boldsymbol{y}\|/\|\boldsymbol{y}\|$ is the relative uncertainty of the y-values (error of measurement) and $\|\delta\boldsymbol{A}\|/\|\boldsymbol{A}\|$ the relative uncertainty of the estimation of (modeling error). The condition number is calculated from eqs. 15–17.

4. INVERSE CALIBRATION

The classical direct or indirect calibration is carried out by least squares minimization according to Gauss. Error-free analytical values x are assumed or at least that the errors in x are very small compared with those of the y-values [2]. Additionally, all the species in the analytical system must be known and included in the calibration. If these preconditions are not fulfilled the inverse calibration must be applied.

The inverse calibration regresses the analytical values (concentrations), x, on the measured values, y. Although with it a *prerequisite of the Gaussian least squares minimization is violated* because the y-values are not error-free, it has been proved that predictions with inverse calibration are more precise than those with the classical calibration [6]. This holds particularly for multivariate inverse calibration.

In chemometrics, the inverse calibration model is also denoted as the P-matrix model (the dimension of $\boldsymbol{P}$ is $m \times n$)

$$\boldsymbol{X} = \boldsymbol{Y}\,\boldsymbol{P} \tag{20}$$

The calibration coefficients are elements of the matrix $\boldsymbol{P}$, which can be estimated by

$$\hat{\boldsymbol{P}} = \boldsymbol{Y}^{+} \cdot \boldsymbol{X} = (\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X} \tag{21}$$

The analysis of an unknown sample is carried out by multiplication of the measured spectrum $\boldsymbol{y}$ by the $\boldsymbol{P}$-matrix

$$\hat{\boldsymbol{x}} = \boldsymbol{y}\hat{\boldsymbol{P}} \tag{22}$$

In the case that the original variables, the measured values $\boldsymbol{y}$, are used for inverse calibration, there are no significant advantages of the procedure apart from the fact that no second matrix inversion has to be carried out in the analysis step (see eq. 22). On the contrary, it is disadvantageous that the calibration coefficients (elements of the $\boldsymbol{P}$-matrix) do not have any physical meaning because they do not reflect the spectra of the single species. In addition, multicollinearities may appear which can make inversion of the $\boldsymbol{Y}$-matrix difficult (see eq. 21).

On the other hand, when latent variables instead of the original variables are used in inverse calibration then powerful methods of multivariate calibration arise which are frequently used in multispecies analysis and single species analysis in multispecies systems. These so-called *"soft modeling methods"* are based, like the $\boldsymbol{P}$-matrix, on the inverse calibration model by which the analytical values are regressed on the spectral data:

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{B} \tag{23}$$

Where $\boldsymbol{B}$ is the $(m \times n)$-matrix of calibration coefficients, in concrete terms the matrix of $\boldsymbol{B}$-coefficients. In contrast to the $\boldsymbol{P}$-matrix, *not all the dimensions of the spectra (the* $\boldsymbol{Y}$*-matrix) are used*, but only those that are significant are realized by certain principal components. Therefore, the estimation of the matrix of $\boldsymbol{B}$-coefficients can be carried out by PCR (principal component regression) or PLS (partial least squares) regression.

Both PCR and PLS form latent variables $\boldsymbol{T}$ (principal components, factors) from the original variables, viz., from the matrix of measured values according to:

$$\boldsymbol{Y} = \boldsymbol{T}\,\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{E}_{Y} \tag{24}$$

where $\boldsymbol{T}$ is the factor (score) matrix and $\boldsymbol{L}$ the loading matrix with the dimension $m \times n$; $\boldsymbol{E}_{Y}$ is the matrix of nonsignificant factors, which is regarded as an error matrix. Additionally, in PLS the matrix of analytical values (e.g., concentrations) is decomposed in the same way:

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{E}_{X} \tag{25}$$

PCR and PLS have in common the following steps:

i. Estimation of a weight matrix (eigenvalues) $\boldsymbol{V}$ (from $\boldsymbol{Y}$ in PCR and from $\boldsymbol{Y}$ and $\boldsymbol{X}$ in PLS)
ii. Calculation of the factor matrix $\boldsymbol{T} = \boldsymbol{ZV}$ by means of the standardized variables $\boldsymbol{Z}$
iii. Calculation of the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ according to

$$\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{T}^{+}\boldsymbol{Y} \tag{26}$$

$$\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{T}^{+}\boldsymbol{X} \tag{27}$$

In PCR, the calibration coefficients (B-matrix) are estimated column by column according to

$$\hat{\boldsymbol{b}} = \boldsymbol{V}\,\boldsymbol{Q}^{\mathrm{T}} \tag{28a}$$

and

$$\hat{\boldsymbol{b}}_0 = \bar{\boldsymbol{c}} - \bar{\boldsymbol{Y}}\,\hat{\boldsymbol{b}} \tag{28b}$$

The prediction then is carried out by

$$\hat{\boldsymbol{c}} = \boldsymbol{Y}\,\hat{\boldsymbol{b}} + \hat{\boldsymbol{b}}_0 \tag{29}$$

The significance and nonsignificance of principal components are decided on the basis of the variance that is explained by each of them. Normally, in analytical methods the main variance is caused by the analyte concentration. But sometimes properties of the sample, such as moisture or surface roughness, or effects of the measuring procedure such as spectral baselines or scattered light, can exceed the effect of analyte concentration. Therefore, additional tests should be made as to what degree the principal components postulated to be nonsignificant by the software are correlated with the analytical values. Principal components that are highly correlated with the variable of interest (e.g., concentration) should be included in the calibration procedure notwithstanding their share in the variance.

In PLS, both the matrices of measured values $\boldsymbol{Y}$ and analytical values $\boldsymbol{X}$ are decomposed according to eqs. 24 and 25: $\boldsymbol{Y} = \boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}_{Y}$ and $\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{E}_{X}$ and thus relations between spectra and concentrations are considered from the outset. The $\boldsymbol{B}$-matrix of calibration coefficients is estimated by

$$\hat{\boldsymbol{B}} = \boldsymbol{V}\,(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{V})^{-1}\boldsymbol{Q}^{\mathrm{T}} \tag{30}$$

Because the $\boldsymbol{Y}$-matrix and $\boldsymbol{X}$-matrix are interdependently decomposed, the $\boldsymbol{B}$-matrix fits better and is more robust than the calibration using PCR. The evaluation is carried out by eq. 23 according to $\hat{\boldsymbol{X}} = \boldsymbol{Y}\hat{\boldsymbol{B}}$. The application of PLS to only one y-variable is denoted as PLS 1. When several y-variables are considered in the form of a-matrix, the procedure is denoted PLS 2 [12].

5. ERROR DIAGNOSIS AND VALIDATION

The reliability of multispecies analysis has to be validated according to the usual criteria: selectivity, accuracy (trueness and precision), confidence and prediction intervals and, calculated from these, multivariate critical values and limits of detection. In multivariate calibration, collinearities of variables caused by correlated concentrations in calibration samples should be avoided. Therefore, the composition of the calibration mixtures should not be varied randomly, but by principles of experimental design [8].

5.1 Selectivity

In general, selectivity of analytical multispecies systems [9] can be expressed qualitatively and estimated quantitatively according to a statement of Kaiser [4] and advanced models [5]. In multivariate calibration, selectivity is mostly quantified by the condition number; see eqs. 15–17. Unfortunately, the condition number does not consider the concentrations of the species and gives therefore only an aid to orientation of maximum expectable analytical errors. Inclusion of the concentrations of calibration standards into selectivity models makes it possible to derive multivariate limits of detection [10–12].

5.2 Precision

The uncertainty of calibration and prediction of unknown concentrations are expressed by the standard error of calibration (SEC), defined as:

$$\hat{s}_{\text{cal}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{(\text{calc})} - y_i^{(\text{true, cs})})^2}{n}} \tag{31}$$

and the standard error of prediction (SEP), defined as:

$$\hat{s}_{\text{pred}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{(\text{calc})} - y_i^{(\text{true, ts})})^2}{n}} \tag{32}$$

where $y_i^{(\text{true,cs})}$ are the true values of the calibration samples (standards), $y_i^{(\text{true,ts})}$ the true values of test samples with which the prediction power independently is estimated, and $y_i^{(\text{calc})}$ are the respective y-values calculated by the model.

Another measure for the precision of multivariate calibration is the so-called PRESS value (predictive residual sum of squares) [13], defined as:

$$s_{\text{res}}^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i^{(\text{calc})} - y_i^{(\text{true})})^2 \tag{33}$$

It can be calculated as usual for SEP (see eq. 32) by use of test samples. It is also possible to estimate the PRESS value on the basis of standard samples only applying cross-validation by means of the so-called hat matrix $\boldsymbol{H}$ [12,13]:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}} \tag{34}$$

The $n \times n$ hat matrix transforms the vector of the measured y-values to the vector of the estimated $\hat{\boldsymbol{y}}$-values. An element h_{ij} of the hat matrix is calculated by

$$h_{ij} = \boldsymbol{x}_i^{\text{T}} (\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_j \tag{35}$$

From the elements of the hat matrix, some important relations can be derived, e.g., the rank of the $\boldsymbol{X}$-matrix from the sum of the significant diagonal elements of the hat matrix:

$$\text{rank}\,\boldsymbol{X} = \sum_{i=1}^{n} h_{ii} \tag{36}$$

(the rank of the hat matrix is equal to its trace) and the residuals

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{y} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{y} \tag{37}$$

The residuals can be calculated from a given set of calibration samples in a different way. Cross-validation is an important procedure to estimate a realistic prediction error like PRESS. The data for k samples are removed from the data matrix and then predicted by the model. The residual errors of prediction of cross-validation in this case are given by

$$e_{(\mathrm{cv})} = \frac{\hat{e}_{(k)}}{1 - h_{kk}} \tag{38}$$

The PRESS value of cross-validation is given by the sum of all the k variations:

$$s^2_{(\mathrm{cv})\mathrm{res}} = \sum \hat{e}^2_{(\mathrm{cv})} \tag{39}$$

Prediction limits for the estimation of an unknown concentration x_i can be calculated. The calculation depends on the specific multivariate calibration model:

$$\Delta\bar{x}_{i,\mathrm{pred}} = \hat{\bar{x}}_i \pm \Delta\hat{x} = \hat{\bar{x}}_i \pm s_x\, t_{\alpha,f} \tag{40}$$

where s_x is the standard deviation of prediction estimated from

$$s_x^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{m - p}\boldsymbol{\Lambda} \tag{41}$$

$t_{\alpha,f}$ is the Student-t statistic for f degrees of freedom at the α confidence level. The variance s_x^2 depends on the number of sensors or wavelengths, m, the number of species, n, the number of parameters p and a factor Λ, which takes the form:

$$\boldsymbol{\Lambda} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1} \tag{42a}$$

in the case of classical multivariate calibration. For inverse calibration,

$$\boldsymbol{\Lambda} = 1 + \boldsymbol{y}_{\mathrm{o}}^{\mathrm{T}}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\,\boldsymbol{y}_{\mathrm{o}} \tag{42b}$$

and for cross-validation when the leverage values are applied in calibration,

$$\Lambda = 1 + h_{kk} \tag{42c}$$

5.3 Trueness

Absence of systematic errors can be tested traditionally by means of recovery functions [14]. For this reason, the concentration estimated by the model, $\hat{c}$, is compared with the true concentration value, c, by a regression model:

$$\hat{c} = \alpha + \beta c \tag{43}$$

where c can be the known values of an independent set of test samples or reference values estimated on the same samples by means of an independent method, which yields true values as is well known. The regression coefficients have to be $\alpha = 0$ and $\beta = 1$, where values outside of the confidence interval $\pm\Delta\alpha$ indicate additive (constant) systematic errors and values exceeding the confidence interval $1 \pm \Delta\beta$ upwards or downwards show proportional systematic errors. By means of recovery studies, both accuracy can be tested and precision can be estimated.

6. REFERENCES

1. K. Eckschlager and K. Danzer. *Information Theory in Analytical Chemistry*, John Wiley, New York (1994).
2. K. Danzer and L. A. Currie. *Pure Appl. Chem*. **70**, 993 (1998).
3. M. Martens and T. Naes. *Multivariate Calibration*, Wiley, New York (1989).
4. H. Kaiser. *Fresenius' J. Anal. Chem*. **260**, 252 (1972).
5. K. Danzer. *Fresenius' J. Anal. Chem*. **369**, 397 (2001).
6. V. Centner, D. L. Massart, S. de Jong. *Fresenius' J. Anal. Chem*. **361**, 2 (1998).
7. M. Otto. *Chemometrics, Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH, Weinheim (1998).
8. S. N. Deming and S. L. Morgan. *Experimental Design: A Chemometric Approach*, 2nd ed., Elsevier, Amsterdam (1993).
9. J. Vessman, R. L. Stefan, J. F. van Staden, K. Danzer, W. Lindner, D. Thorburn Burns, A. Fajgelj, H. Müller. *Pure Appl. Chem*. **73**, 1382 (2001).
10. G. Bauer, W. Wegscheider, H. M. Ortner. *Fresenius' J. Anal. Chem*. **340**, 135 (1991).
11. K. Faber and B. R. Kowalski. *J. Chemometrics* **11**, 181 (1997).
12. K. Faber and B. R. Kowalski. *Appl. Spectrosc.* **51**, 660 (1997).
13. I. E. Frank and R. Todeschini. *The Data Analysis Handbook*, Elsevier, Amsterdam (1994).
14. D. Thorburn Burns, K. Danzer, A. Townshend. *Pure Appl. Chem*. **74**, 2201 (2002).