ELSEVIER

Chemometrics and Intelligent Laboratory Systems 50 (2000) 107–114

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Exact presentation of multivariate calibration model as univariate calibration graph

Nicolaas (Klaas) M. Faber [*]

Dunantsingel 28, 2806 JB Gouda, Netherlands

Received 3 June 1999; accepted 3 September 1999

Abstract

A proof is given that a recently introduced univariate presentation of a multivariate calibration model is exact, i.e., there are no approximations involved. The proof is based on: (1) previously proposed definitions of multivariate net analyte signal and multivariate calibration factors (sensitivity in classical model and inverse sensitivity in inverse model) and (2) the geometrical property of the true regression vector in inverse multivariate calibration that it must be proportional to the true multivariate net analyte signal vector of a particular sample. The extension of the proof to multiway calibration is briefly discussed. A practical example from near-infrared (NIR) spectroscopy is used to illustrate that the proposed univariate presentation may give more meaning to the term 'spectral overlap'. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Multivariate calibration; Net analyte signal; Analytical figures of merit; Calibration factor; Multiway calibration; Near-infrared

1. Introduction

In general, univariate calibration is much better understood than multivariate calibration. One of the main reasons for this situation may well be the fact that univariate data can be presented in an intuitively appealing scatter plot of analyte concentration vs. net analyte signal. For example, the most suitable experimental design for the calibration set is easily explained from such a plot [1]. In addition, the different contributions to prediction intervals (concentration and signal error) can be clearly distinguished this way. In the *classical model*, net analyte signal is expressed as a function of analyte concentration. The slope of the calibration graph is identified as the *sensitivity* of the analytical determination; a high sensitivity is desirable since it leads to a small amount of signal error propagation. In the *inverse model*, analyte concentration is expressed as a function of net analyte signal. Now the slope of the calibration graph is identified as the *inverse sensitivity* and, in contrast, a low inverse sensitivity is desired for the same reason.

It should be clear that the understanding of multivariate calibration methodology can be improved by a suitable generalisation of the concepts that have demonstrated their utility in the univariate context. Lorber [2] has defined the (multivariate) net analyte signal *vector* as the part of the gross signal vector that is useful for calibration. From this definition, it follows that the net analyte signal vector must be or-

[*] Department of Veterinary Anatomy and Physiology, Utrecht University, PO Box 80157, 3508TD Utrecht, Netherlands. E-mail: m.faber@vet.uu.nl

PII: S0169-7439(99)00054-4

thogonal to the signal vectors of the interferences. Lorber [2] has further proposed to calculate the analogue of the univariate net analyte signal, which is a scalar, as the Euclidean norm of the multivariate net analyte signal vector. This procedure leads to the multivariate *scalar* net analyte signal. [1]

Recently, a framework has been developed of analytical figures of merit for multilinear data that is based on Lorber's definition of scalar net analyte signal [3]. This framework incorporates multivariate data as a special case. In particular, it is proposed to present multivariate models (and models obtained for data of higher complexity) as a univariate calibration graph of analyte concentration vs. scalar net analyte signal. Such a presentation for the training and prediction samples should be more informative than the usual plot of known ('lab value') vs. fitted or predicted analyte concentration, since the amount of spectral error propagation is visualised. Spectral error propagation is a key issue when, for example, selecting a proper pre-treatment method. The proposed 'pseudo-univariate' calibration graphs have proved to be insightful in this respect for a near-infrared (NIR) application [4]. However, no rigorous proof is given in Ref. [3] to show that this univariate presentation is exact, i.e., no approximations are involved. Such a proof is relevant, since the validity of this claim is not self-evident. This paper aims to show that only definitions and a fundamental geometrical property are involved, hence the claim is valid. The utility of the proposed univariate presentation is illustrated on the NIR spectroscopic prediction of the oxygenates methyl-*tert*-butyl ether (MTBE) and ethanol in standard reference material gasoline.

2. Theory

In the current presentation, the data follow Beer–Lambert's law for spectroscopy (linear and additive signal). As correctly noted by one of the reviewers, this may help in understanding what is going on, but it is not really needed for the concepts proven in the paper. Indeed, since the development is purely algebraic, the concepts can be applied to regression models in general.

2.1. Univariate calibration

To prove that the proposed univariate presentation of a multivariate model is exact, it is convenient to first consider the univariate zero-intercept inverse calibration model (a complication due to mean centring will be discussed later). Under this model, the analyte concentration, c, and the net analyte signal, r^*, are related for a particular sample as

$$c = br^* + \varepsilon \tag{1}$$

where b denotes the 'inverse sensitivity' (the calibration factor) and ε is a residual. The net analyte signal is obtained by correcting the gross signal, r, for the background contribution, d, as $r^* = r - d$. It is seen that analyte concentration is the predictand while net analyte signal is the predictor. Under the classical model, the roles of net analyte signal and analyte concentration are reversed, which is consistent with the causal relationship between these variables, i.e.,

$$r^* = sc + \varepsilon \tag{2}$$

where s denotes the sensitivity (the calibration factor). The relationship between the two modes of calibration is summarised by the identity

$$b = 1/s \tag{3}$$

It is emphasised that this expression holds for the true quantities, not for specific estimates. Estimates are arbitrary in the sense that they depend on the noise in the data as well as the estimation procedure.

2.2. Multivariate calibration

Under the multivariate zero-intercept inverse calibration model, the analyte concentration is given by

$$c = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r} + \varepsilon \tag{4}$$

where $\boldsymbol{b}$ is the regression vector, $\boldsymbol{r}$ is the instrument response vector and the superscript 'T' symbolises vector transposition.

It is important to note that Eq. (3) is also valid for multivariate models if the corresponding calibration

[1] In the literature confusing terminology can be encountered. Often the term (multivariate) net analyte signal is used to denote the vector as well as its norm.

factors (sensitivity and inverse sensitivity) are consistently defined [5]. Thus, a rigorous proof, which covers both modes of calibration, should demonstrate that Eq. (4) can be brought in the form of Eq. (1) without approximation.

The required proof proceeds as follows. The instrument response vector is expanded in two orthogonal terms as

$$\boldsymbol{r} = \boldsymbol{r}^* + \boldsymbol{r}^{\perp} \tag{5}$$

where $\boldsymbol{r}^*$ denotes the net analyte signal vector and $\boldsymbol{r}^{\perp}$ is the vector orthogonal to the net analyte signal vector (it lies in the space spanned by the spectra of the interferences). The true regression vector should not pick up a signal contribution of the interferences, hence, it should be orthogonal to the space spanned by the interferences' signal vectors. It follows that the regression vector is proportional to the net analyte signal vector; for more details about this proportionality, see Ref. [6].[2] Using this property, Eq. (4) is worked out as

$$c = \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{r}^* + \boldsymbol{r}^{\perp}) + \varepsilon = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}^* + \varepsilon \tag{6}$$

The inner product on the far right-hand side of Eq. (6) is simplified by using the geometrical property that the regression vector is a scalar multiple of the net analyte signal vector, hence

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}^* = \|\boldsymbol{b}\| \cdot \|\boldsymbol{r}^*\| \cos(\boldsymbol{b},\boldsymbol{r}^*) = \|\boldsymbol{b}\| \cdot \|\boldsymbol{r}^*\| \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\cos(\boldsymbol{b},\boldsymbol{r}^*)$ is the cosine of the angle between $\boldsymbol{b}$ and $\boldsymbol{r}^*$. The last step in Eq. (7) follows from the observation that for a zero-intercept model the scalar multiplier is strictly positive, since analyte concentration is strictly positive, hence, the angle must be 0° and $\cos(\boldsymbol{b},\boldsymbol{r}^*) = 1$ (for general vectors, Eq. (7) can be used to derive the well known Schwarz inequality [9]).

The multivariate inverse sensitivity and the scalar net analyte signal are defined as [2,3,5]

$$b = \|\boldsymbol{b}\| \tag{8a}$$

$$r^* = \|\boldsymbol{r}^*\| \tag{8b}$$

and it is easily verified by combining Eqs. (6), (7), (8a) and (8b) that Eq. (4) takes the form of Eq. (1), which completes the proof.

If mean centring is applied, deviations from the mean analyte concentration are modelled, hence the angle between $\boldsymbol{b}$ and $\boldsymbol{r}^*$ must be 0° for analyte concentrations larger than the mean and 180° for analyte concentrations smaller than the mean. In the latter case, the far right-hand side of Eq. (7) obtains a minus sign and it follows that for a mean-centred model, a univariate calibration graph is obtained, which is shifted with respect to a zero-intercept model (the mean is the origin). However, it is customary to report absolute values for analyte concentration rather than deviations from the mean. The common univariate calibration graph with zero as the origin is obtained by simply adding in the means for analyte concentration and net analyte signal. For more details, see Ref. [10].

2.3. Multiway calibration

The framework developed in Ref. [3] includes data types of higher complexity than vectors, i.e., matrices, 'cubes', etc. The proof is easily extended to multiway calibration by recognising that a multiway array can be converted into a vector without loss of information: the special structure of the multiway array is reflected in a suitable structure for the resulting vectors. For example, for multilinear data, the 'overall' regression vector can be expressed as the (multiple) Kronecker product of the regression vectors associated with the individual modes (see Table 2 in Ref. [3]). For general multiway models (i.e., less restrictive than multilinear models), similar results have not yet been reported to the author's best knowledge. The inverse sensitivity follows by applying Eq. (8a) to the 'overall' regression vector, and inserting the result in Eq. (3) yields the sensitivity. These numbers can be used to construct a univariate calibration graph, see Ref. [11] for an example of a univariate presentation of a bilinear model.

[2] As an aside, it is noted that this proportionality implies that the wavelength selection procedures introduced by Xu and Schechter [7] and Spiegelman et al. [8] are strongly related. In both methods, wavelengths are ranked according to the size of wavelength-specific signal-to-noise ratios. While Xu and Schechter calculate these ratios from the net analyte signal vector, Spiegelman et al. use the regression vector instead. Clearly, owing to the constant proportionality, the ranking should be identical if these vectors are consistently estimated.

The characterisation of multiway models in terms of analytical figures of merit such as sensitivity can lead to additional insight. This becomes clear from close examination of the following example taken from the literature. Bro [12] gives plots of regression matrices obtained by tri-PLS and unfold-PLS for fluorescence excitation emission matrix (EEM) data (see Fig. 7(ii) and (iii)). It can be inferred from these plots that the regression matrix obtained for the unfolded data has a larger variance. Bro [12] correctly attributes the increased variance to the larger number of fitted parameters, which is equivalent to a smaller number of degrees of freedom. However, another aspect of unfolding, which is not discussed in Ref. [12], is the decrease of the norm of the regression matrix by approximately a factor of two. (It will be proved elsewhere that, unless analyte and interferences do not overlap in both modes, the norm of the regression matrix decreases upon unfolding.) Applying Eq. (8a) to the regression matrices of Fig. 7(ii) and (iii) leads to the interpretation that the inverse sensitivity has decreased, hence, according to Eq. (3), the sensitivity has increased. It follows that unfolding is a trade-off process where a higher sensitivity (favourable effect on expected prediction error) is obtained at the cost of increased variance in the model parameters (unfavourable effect on expected prediction error). Thus, depending on the specific application at hand, unfolding will either be beneficial or not (for this example, unfolding leads to inferior prediction results, see Ref. [12]).

3. Experimental

Full details on the NIR data are presented elsewhere [13–15]. Calibration and test sets consist of 40 samples each. Calibration models are constructed using partial least squares (PLS). All calculations are based on 391 absorbance values evenly spaced in wavenumber space between 6000 and 9000 cm^{-1}.

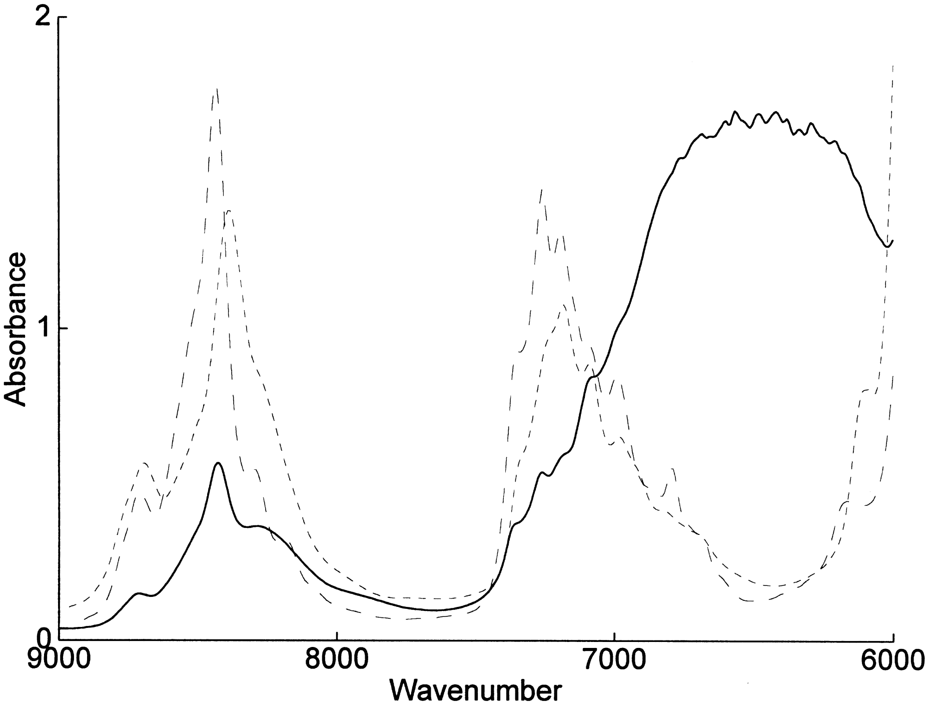

Fig. 1. NIR spectra of ethanol (—), MTBE (— —) and RF-A gasoline (- - -).

4. Results and discussion

In this paper, focus is on quantifying spectral overlap. A detailed discussion of the consequences of spectral overlap for uncertainty in predictions is given elsewhere and will not be repeated (see Ref. [14] and references therein).

The signals of neat MTBE and the industry-average gasoline ('RF-A') used to prepare the mixtures are severely overlapped over the entire spectral region while ethanol exhibits a broad characteristic absorption band between 7000 and 6000 cm^{-1} due to hydrogen-bonded O–H stretch (Fig. 1). Consequently, the hydrogen-bonded O–H stretch gives rise to the main difference between spectra of mixtures of ethanol and RF-A (Fig. 2). Spectral overlap between two components can be conveniently quantified using the inner product or linear correlation coefficient (Table 1). If the spectra are normalised to unit length these quantities range between 0 and 1 (all absorbances are positive, see Fig. 1). As expected, these quantities are rather large for the overlap between MTBE and RFA but only moderate for the overlap between ethanol and MTBE and RFA. However, it is well known that small two-component overlaps do not preclude a large multicomponent overlap, which is unsatisfactory. Eq. (6) shows that prediction is entirely based on the net analyte signal vector. Thus, the net analyte signal vector seems to be a useful basis for the definition of multicomponent overlap. The net analyte signal for ethanol is surprisingly small (Fig. 3). A qualitative interpretation of this vector is difficult owing to the orthogonality constraints. For example, negative regions indicate wavelengths where the contribution of the interferences is large. (Seasholtz and Kowalski [16] have discussed the limited interpretability of the regression vector and their arguments carry through as a result of the proportionality.) Taking the Euclidean norm of the net analyte signal vector leads to the proposed univariate calibration graph (Fig. 4). The slope in Fig. 4 is the multivariate sensitivity (0.048 absorbance units (AU)/% oxygen mass fraction). It is noted that the univariate plot is presented as a *classical* model while

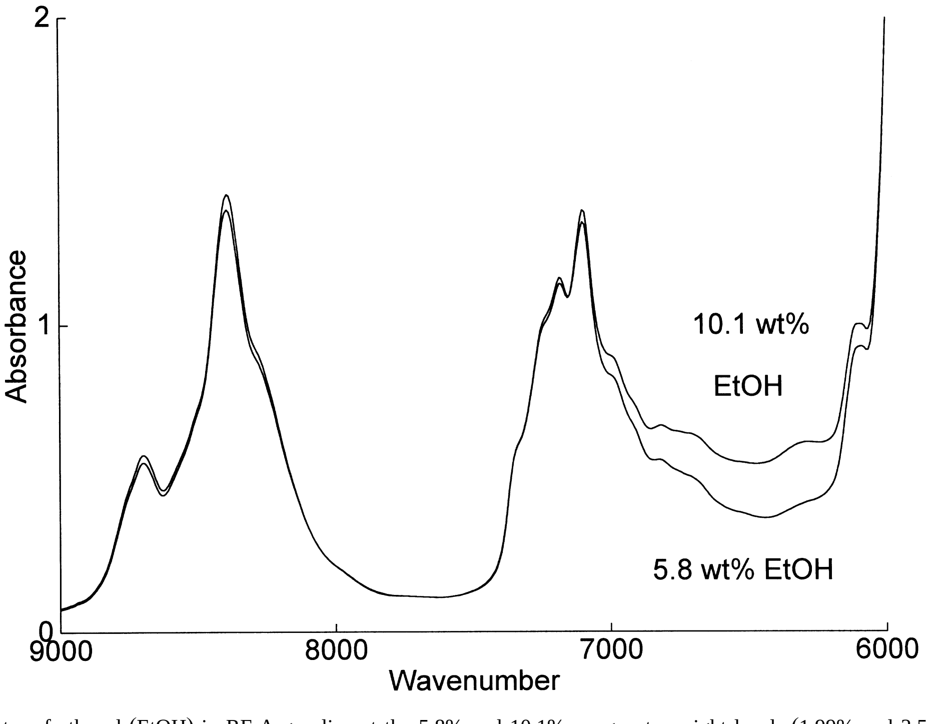

Fig. 2. NIR spectra of ethanol (EtOH) in RF-A gasoline at the 5.8% and 10.1% oxygenate weight levels (1.99% and 3.5% oxygen mass fraction).

Table 1
Inner products (left lower corner) and linear correlation coefficients (right upper corner) for normalised spectra of ethanol, MTBE and RF-A gasoline

Compound	Ethanol	MTBE	RF-A gasoline
Ethanol	1	0.38	0.41
MTBE	0.57	1	0.88
RF-A gasoline	0.60	0.92	1

it is based on a multivariate *inverse* calibration model (eight-dimensional PLS). The reason for doing so is that it facilitates a comparison with Fig. 2. It is seen that the spectra displayed in Fig. 2 differ by approximately 0.2 AU over a range that contains more than 50 absorbance values, while the difference in (scalar) net analyte signal is only approximately 0.1 AU. This finding implies that Fig. 2 is misleading in the sense that it suggests that the region between 7000 and 6000 cm^{-1} is rather unique for ethanol. The uniqueness of a spectrum is conveniently measured in terms of Lorber's selectivity [2],

$$\xi_k = r_k^* / r_k = \|\boldsymbol{r}_k^*\| / \|\boldsymbol{r}_k\| \qquad (9)$$

where k denotes the analyte of interest, $\boldsymbol{r}_k^*$ denotes the net analyte signal vector, and $\boldsymbol{r}_k$ is the analyte contribution to the instrument response vector $\boldsymbol{r}$. It turns out that the selectivity of ethanol is only 0.068, i.e., 93.2% of the spectrum displayed in Fig. 1 is lost due to overlap with the interferences' spectra. Fortunately, the standard deviation of the spectral noise is small (approximately 10^{-4} AU); a detailed analysis of sources of error variance shows that its effect on prediction is (almost) negligible (compare columns four and six in Table 3 in Ref. [14]). From Fig. 2, one might infer that a standard deviation of 10^{-3} AU would still be negligible, which is incorrect. This is easily verified by multiplying the numbers in column six of Table 3 in Ref. [14] by a factor 100: the spectral noise would overwhelm the other sources of error variance. The advantage of a plot such as Fig. 4

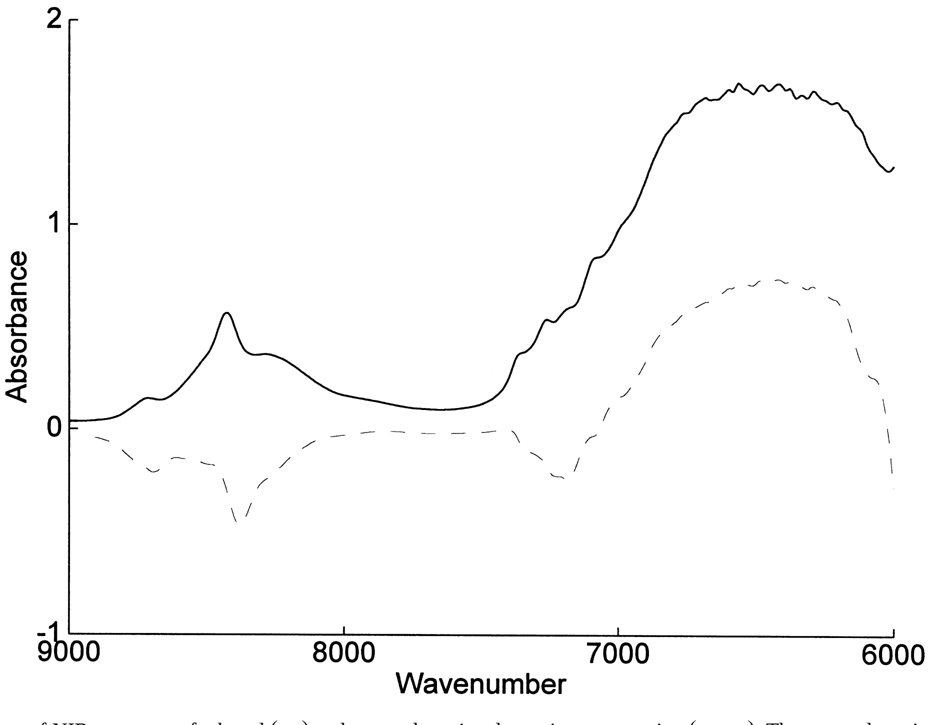

Fig. 3. Comparison of NIR spectrum of ethanol (—) and net analyte signal at unit concentration (— —). The net analyte signal is multiplied by a factor 100 for visual clarity.

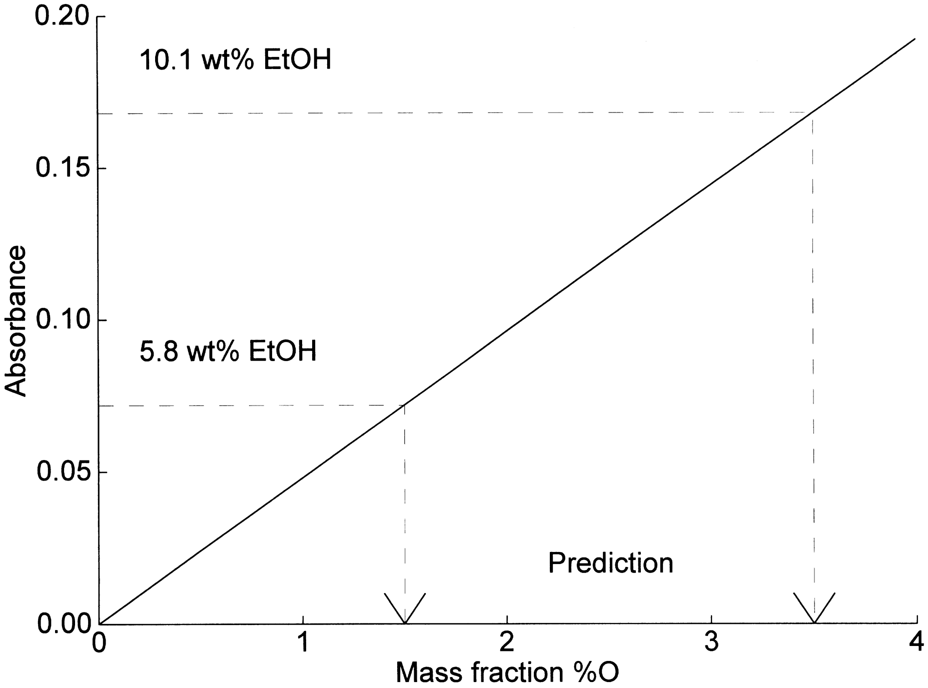

Fig. 4. Univariate presentation of multivariate model for ethanol (EtOH).

is that the contribution of spectral noise to prediction error can be visualised (see Fig. 6 in Ref. [4] for a practical example).

5. Conclusions and outlook

It has been shown that the univariate presentation of a multivariate model may lead to useful insight regarding the amount of the data that effectively enters the model. In this sense, more meaning is given to the term 'spectral overlap'. Otto and Wegscheider [17] have listed practical selectivities based on the condition number for atomic and molecular spectroscopy, and for electroanalytical methods. Lorber [2] has explained that selectivities based on the condition number are of limited utility because a characterisation of individual analytes is not possible. However, continuing the work of Otto and Wegscheider by listing practical selectivities (and other analytical figures of merit) based on Lorber's definitions may, among others, facilitate method selection in the future.

Acknowledgements

Dave Duewer of the National Institute of Standards and Technology (NIST) is thanked for making the NIR spectra available for this work. The author appreciates useful comments made by the reviewers that helped to improve the presentation.

References

[1] D.L. Massart, A. Dijkstra, L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.

[2] A. Lorber, Anal. Chem. 58 (1986) 1167–1172.

[3] K. Faber, A. Lorber, B.R. Kowalski, J. Chemom. 11 (1997) 419–461.

[4] N.M. Faber, Anal. Chem. 71 (1999) 557–565.

[5] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 67–79.

[6] N.M. Faber, Anal. Chem. 70 (1998) 5108–5110.

[7] L. Xu, I. Schechter, Anal. Chem. 68 (1996) 2392–2400.

[8] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35–44.

[9] J. ten Berge, Least Squares Optimization in Multivariate Analysis, DSWO Press, Leiden, 1993.

[10] N.M. Faber, J. Chemom. 12 (1998) 405–409.

[11] N.M. Faber, Anal. Chim. Acta 381 (1999) 103–109.

[12] R. Bro, J. Chemom. 10 (1996) 47–61.

[13] S.J. Choquette, S.N. Chesler, D.L. Duewer, S. Wang, T.C. O'Haver, Anal. Chem. 68 (1996) 3525–3533.

[14] N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler, Anal. Chem. 70 (1998) 2972–2982.

[15] N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler, Anal. Chem. 70 (1998) 4877.

[16] M.B. Seasholtz, B.R. Kowalski, Appl. Spectrosc. 44 (1990) 1337–1348.

[17] M. Otto, W. Wegscheider, Anal. Chim. Acta 180 (1986) 445–456.