ELSEVIER

Chemometrics and Intelligent Laboratory Systems 45 (1999) 397–408

Chemometrics and intelligent laboratory systems

Multivariate detection limits with fixed probabilities of error

R. Boqué *, M.S. Larrechi, F.X. Rius

Departament de Química Analítica i Química Orgànica, Universitat Rovira i Virgili, Pça. Imperial Tàrraco 1, Tarragona, Catalunya E-43005, Spain

Abstract

In this paper, a new approach to calculate multivariate detection limits (MDL) for the commonly used inverse calibration model is discussed. The derived estimator follows the latest recommendations of the International Union of Pure and Applied Chemistry (IUPAC) concerning the detection capabilities of analytical methods. Consequently, the new approach: (a) is based on the theory of hypothesis testing and takes into account the probabilities of false positive and false negative decisions, and (b) takes into account all the different sources of error, both in calibration and prediction steps, which affect the final result. The MDL is affected by the presence of other analytes in the sample to be analysed; therefore, it has a different value for each sample to be tested and so the proposed approach attempts to find whether the concentration derived from a given response can be detected or not at the fixed probabilities of error. The estimator has been validated with and applied to real samples analysed by NIR spectroscopy.

Keywords: Multivariate detection limits; Inverse calibration model; NIR spectroscopy

Contents

1. Introduction 398

2. Theory 399
   2.1. Calibration model 399
   2.2. Variance of the predicted concentration 399
   2.3. Strategy for calculating multivariate detection limits 400

3. Experimental section 403
   3.1. Samples and software 403
   3.2. Instrumental 403
   3.3. Software 403

4. Results and discussion 403
   4.1. Model building and validation 403
   4.2. Validation of the MDL estimator 405
   4.3. Calculation of the MDL 406

* Corresponding author. Tel.: +34-77-558155; Fax: +34-77-559563; E-mail: boque@quimica.urv.es

0169-7439/99/$ - see front matter
PII: S0169-7439(98)00195-6

5. Conclusions 407

Acknowledgements 407

Appendix A. Mathematical derivation of the detection response 407

References 408

1. Introduction

As instrumental techniques get more and more sophisticated and capable of generating multivariate data, the study of different figures of merit, among which are detection limits, has become very important, not only to characterise the different associated analytical methods, but also to serve as a guide for the design of new instruments. The study of detection limits in multivariate regression only dates back a few years and has been reviewed recently [1]. The approaches described and the derived estimators are very varied; Lorber [2] was one of the first to calculate a multivariate detection limit (MDL), starting from the definition of *net analyte signal*. Subsequently, Lorber and Kowalski [3] defined this estimator as a function of the confidence intervals associated with the predicted concentration. Bauer et al. [4] obtained an estimator that is a function of the error in the predicted concentration with the theory of error propagation.

For multivariate inverse calibration models, little has been done. The first approach, by Lorber and Kowalski [3], presented the problem of detection as checking if the predicted concentration and its confidence intervals for a given test sample included or not the zero concentration value. However, the main drawback of their work was the reduced mathematical and statistical support for the derived expression of the variance of the predicted concentration. Subsequently, Faber and Kowalski [5] dealt with the problem of detection in a similar way, but using expressions for the variance of prediction based on the so-called errors-in-variables (EIV) models. Finally, Lorber et al. [6] developed some figures of merit for inverse models based on the concept of *net analyte signal*. In this latter paper, the authors developed an expression for the MDL on the signal domain, following the basis set up in a work by Boqué and Rius [7] on classical models. However, the fact that the estimator by Lorber et al. [6] takes only into account the uncertainty in the signal measurements makes its real application to be rather limited. In analytical methods that use inverse calibration models, it is emphasised that one of the main sources of uncertainty is given by the concentrations of the reference method, from which the former method has been calibrated. Moreover, as it has also been pointed out by the IUPAC, ''detection limits cannot be specified in the absence of a fully defined measurement process'' [8]. Consequently, every source of uncertainty from the whole method contributing to the final predicted value should be included in the derivation of MDLs.

Finally, protection against both false positive (probability of type I error, α) and false negative (probability of type II error, β) decisions has to be achieved when establishing whether the analyte in study can be detected or not, as it has been suggested in the definition of limit of detection proposed for individual signals (zero-order calibration) in Clayton et al. [9] and the latest IUPAC recommendations [8]. It seems therefore that a more complete approach to MDL calculation is needed. Boqué and Rius [7] have applied very recently the principle of testing both hypotheses for the classical multivariate calibration model. Similarly, Faber and Kowalski [10] have derived a simplified expression for classical calibration, which is a modification of the original expression from Bauer et al. [4].

In this paper, we propose a new method for calculating MDL in the concentration domain for the commonly used inverse calibration models, $\boldsymbol{c} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The approach presented is based on the work on classical univariate regression developed by Hubaux and Vos [11]. The method has the advantage that the calculation of the confidence intervals is associated with the predicted concentrations of a multivariate model and takes into account the probabilities of false

positive and false negative decisions. A statistical modification of the above approach, based on the work of Clayton et al. [9] in univariate calibration is also presented and discussed. In multivariate calibration, the detection limit does not only depend on the mathematical model and the calibration sets used but also on the presence of other analytes in the sample to be analysed. Therefore, each analyte in every single sample has a different MDL and, as a consequence, it is important to assess whether the concentration derived from a given response is detectable or not at the chosen probabilities of error.

The estimator has been validated with real data by comparing both the theoretical and the experimental percentages of committing a type II error, from different fixed values of the type I error and at different levels of concentration. Additionally, the procedure described is applied to the calculation of the detection limits of a set of real data, consisting in the determination of the aromatic content of gasolines by near-infrared spectroscopy.

2. Theory

2.1. Calibration model

The general inverse calibration model can be expressed as $\boldsymbol{c} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the concentrations are a linear function of the responses (sensors). Matrix $\mathbf{R}$ contains the responses at J different sensors for the I calibration samples. $\boldsymbol{\varepsilon}$ $(I \times 1)$ is the vector of residual error (unmodelled variation in $\boldsymbol{c}$) and its elements are assumed to be independently and identically distributed (i.i.d.). Assuming that the data are mean centred, the vector $\boldsymbol{\beta}$ of regression coefficients at J wavelengths for an analyte k can be estimated in the calibration step using $\hat{\boldsymbol{b}}_k = \hat{\mathbf{R}}^+ \boldsymbol{c}_k$, where $\hat{\mathbf{R}}^+$ is the pseudoinverse of matrix $\mathbf{R}$ $(I \times J)$ and $\boldsymbol{c}_k$ is the vector $(I \times 1)$ of analyte concentrations in the calibration samples. In general, $\hat{\boldsymbol{b}}_k$ is sought in such a way that the value $\|\boldsymbol{c}_k - \mathbf{R}\hat{\boldsymbol{b}}_k\|^2$ is minimised, with $\|\cdot\|$ as the Euclidean norm of the given vector.

The difference among the existing multivariate regression methods simply lies in the various ways of inverting matrix $\mathbf{R}$ [12]. For inverse least-squares (ILS) regression, matrix $\mathbf{R}^+$ is equivalent to $(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}$, the well-known least-squares solution. For principal components regression (PCR) and partial least-squares regression (PLS), however, the calculation of the pseudoinverse matrix is a three-step procedure: decomposition of matrix $\mathbf{R}$, determination of the number of optimal factors (pseudo or chemical rank) of the matrix and subsequent calculation of its pseudoinverse. One of the most used algorithms for calculating $\mathbf{R}^+$ is the *singular value decomposition* (SVD) [13].

Once the model has been validated (e.g., by cross-validation), if the response of an unknown sample is measured as a row vector, $\boldsymbol{r}_{\mathrm{un}}^{\mathrm{T}}$, the expression which predicts the concentration of the analyte k is given by the scalar (Eq. (1)):

$$\hat{c}_{\mathrm{un},k} = \boldsymbol{r}_{\mathrm{un}}^{\mathrm{T}}\hat{\mathbf{R}}^+\boldsymbol{c}_k = \boldsymbol{r}_{\mathrm{un}}^{\mathrm{T}}\hat{\boldsymbol{b}}_k. \quad (1)$$

2.2. Variance of the predicted concentration

Several expressions for the variance of the predicted concentration in inverse models have been developed [3,14,15]. Traditionally, in the literature on statistics, the calculation of this uncertainty relies on the model $\boldsymbol{c} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, which assumes that there are no errors in the measurement of the concentrations and instrumental responses. This is a severe restriction in many spectroscopic analytical methods, in which the concentration of the analyte in the calibration samples, usually determined by a reference method, has a non-negligible measurement error. As it has been pointed out by Faber and Kowalski [16], the EIV model, where the errors in the measurement variables are taken into account, is an alternative for this kind of situations. A general expression for the variance of the predicted concentration can be derived for the EIV model:

$$\mathrm{var}(\hat{c}_{\mathrm{un},k}) \approx \left(I^{-1} + \hat{h}_{\mathrm{un}}\right)\left[\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 + \hat{\sigma}_{\mathbf{c}}^2 + \|\hat{\boldsymbol{b}}_k\|^2\hat{\sigma}_{\mathbf{R}}^2\right] + \hat{\sigma}_{\boldsymbol{\varepsilon}_{\mathrm{un}}}^2 + \|\hat{\boldsymbol{b}}_k\|^2\hat{\sigma}_{\boldsymbol{r}_{\mathrm{un}}}^2 \quad (2)$$

where $\hat{\sigma}_{\mathbf{c}}^2$, $\hat{\sigma}_{\mathbf{R}}^2$ and $\hat{\sigma}_{\boldsymbol{r}_{\mathrm{un}}}^2$ are variances of the measurement errors for the concentrations and responses of the calibration set and for the response of the test sample, respectively. $\hat{\sigma}_{\boldsymbol{\varepsilon}}^2$ and $\hat{\sigma}_{\boldsymbol{\varepsilon}_{\mathrm{un}}}^2$ are the variances of the residuals for the calibration model and for the test sample, respectively. Finally, $\hat{h}_{\mathrm{un}}$ is the leverage of the unknown sample (which describes the position

of the sample with respect to the centre of the model) and is given by:

$$\hat{h}_{\mathrm{un}} = \boldsymbol{r}_{\mathrm{un}}^{\mathrm{T}} \hat{\mathbf{R}}^{+} \left(\hat{\mathbf{R}}^{+}\right)^{\mathrm{T}} \boldsymbol{r}_{\mathrm{un}}. \quad (3)$$

As the estimated pseudoinverse is used, both $\hat{\boldsymbol{b}}_{\mathrm{k}}$ and $\hat{h}_{\mathrm{un}}$ may be estimated from ILS, PCR or PLS regression without loss of generality.

Eq. (2) is a general expression, which can be simplified. If, for example, signal measurement errors are neglected and it is assumed that $\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 = \hat{\sigma}_{\boldsymbol{\varepsilon}_{\mathrm{un}}}^2$, then Eq. (2) becomes:

$$\begin{aligned} \mathrm{var}\left(\hat{c}_{\mathrm{un},k}\right) &\approx \left(I^{-1} + \hat{h}_{\mathrm{un}}\right)\left[\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 + \hat{\sigma}_{\mathbf{c}}^2\right] + \hat{\sigma}_{\boldsymbol{\varepsilon}_{\mathrm{un}}}^2 \\ &\approx \left(I^{-1} + \hat{h}_{\mathrm{un}} + 1\right)\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 + \left(I^{-1} + \hat{h}_{\mathrm{un}}\right)\hat{\sigma}_{\mathbf{c}}^2. \end{aligned} \quad (4)$$

If the measurement error in the reference concentration values is also neglected, then the expression of the variance for the classical regression model is obtained:

$$\mathrm{var}\left(\hat{c}_{\mathrm{un},k}\right) \approx \hat{\sigma}_{\boldsymbol{\varepsilon}}^2\left(I^{-1} + h_{\mathrm{un}} + 1\right). \quad (5)$$

The first term of Eq. (5) takes into account the error in the data centring (for a centred model), the second is the squared Mahalanobis distance and the third is the error due to the unmodelled part of the true value $c_{\mathrm{un},k}$. The value of $\hat{\sigma}_{\boldsymbol{\varepsilon}}^2$ can be calculated from the residuals of the calibration model and it has to be corrected for its degrees of freedom, $\nu = I - r - 1$ (for a centred model), being r the number of optimal factors of the model. On the other side, if $\hat{\sigma}_{\mathbf{c}}^2 \gg \hat{\sigma}_{\boldsymbol{\varepsilon}}^2$ and the signal measurement errors are neglected, Eq. (2) becomes:

$$\mathrm{var}\left(\hat{c}_{\mathrm{un},k}\right) \approx \hat{\sigma}_{\mathbf{c}}^2\left(I^{-1} + \hat{h}_{\mathrm{un}}\right). \quad (6)$$

Eq. (2) is important because, as well as serving as a base for deriving the detection limit, it gives information about the various error sources that are involved in the model and so directly influence the parameters that affect it, both in the calibration and prediction steps. However, two main problems are to be mentioned concerning Eq. (2). First, bias is not taken into account, so the calibration model should be tested for bias before Eq. (2) is used. Otherwise, the uncertainty in predictions would be incorrect. This point has been dealt with by Faber and Kowalski [5], so the reader is referred there for more information. The second problem is related to the calculation of the different error terms in Eq. (2). The best approach to calculate these uncertainties is by experimental replication, both in $\mathbf{R}$ and $\boldsymbol{\varepsilon}$ matrices.

Two-sided confidence intervals associated with the predicted concentration, $\hat{c}_{\mathrm{un},k}$, can also be derived from Eq. (2), provided that the predicted concentration follows a normal distribution:

$$\hat{c}_{\mathrm{un},k} - \hat{\gamma} \le c_{\mathrm{un},k} < \hat{c}_{\mathrm{un},k} + \hat{\gamma} \quad (7)$$

where

$$\hat{\gamma} = t_{1-\alpha/2,\nu}\mathrm{var}\left(\hat{c}_{\mathrm{un},k}\right)^{1/2} \quad (8)$$

The assumption about the distribution of the predicted concentrations can be safely made in multivariate calibration, because the number of observations is generally large and consequently suitable formulation of the central limit theorem can be invoked. Problems arise with Eq. (8) when the signal measurement, $\hat{\sigma}_{\mathbf{R}}^2$ and $\hat{\sigma}_{\boldsymbol{r}_{\mathrm{un}}}^2$, are considered [5]. First, as $\mathrm{var}(\hat{c}_{\mathrm{un},k})$ is a complex variance estimate, the number of degrees of freedom, ν, for the calculation of the statistic t has to be obtained by applying Satterthwaite's rule [17]. Secondly, $\hat{c}_{\mathrm{un},k}$ and $\mathrm{var}(\hat{c}_{\mathrm{un},k})$ are no longer independent because the error in the spectrum of the test sample is used to estimate both quantities. So, the degree of correlation has to be measured in order to safely apply the *t*-test and derive the confidence intervals.

2.3. Strategy for calculating multivariate detection limits

The predicted concentration and its confidence intervals can be calculated in the prediction step from the response of the problem sample. In order to assess whether the predicted concentration is above the limit of detection, a new approach has been developed in the multivariate field that takes into account the probabilities of both type I and type II errors. In this new scenario, the limit of detection corresponds to the concentration (calculated with a β probability of error) derived from the instrumental response at zero concentration level (calculated with an α probability of error). This approach uses criteria discussed by Hubaux and Vos [11] and subsequently developed by Garner and Robertson [18] in the field of classical univariate regression and is based on the confidence intervals of the regression model. A sta-

tistical improvement of this approach is also presented, which is based on the work of Clayton et al. [9], who uses a non-central t parameter in order to account for the probability of a type II error.

The strategy developed for calculating detection limits in multivariate calibration consists of the following steps:

1. Modelling and validation of the multivariate regression model.
2. Measuring the multivariate response for the unknown sample, $\boldsymbol{r}_{un}$.
3. Calculation of the predicted concentration and its variance (by using either Eqs. (2), (4), (5) or (6)).
4. Calculating the detection response, $\hat{r}_D$ (i.e., the instrumental response corresponding to the unknown sample but with an analyte concentration of zero), with an α probability of committing a type I error (Fig. 1).
5. Calculating the concentration, $(MDL)_k$, for the detection response, $\hat{r}_D$, with a β probability of committing a type II error.
6. Calculating $\hat{c}_{un,k}$, from $\boldsymbol{r}_{un}$ and the model developed in 1.
7. Comparing $\hat{c}_{un,k}$ with $(MDL)_k$ and deducing whether the measured response, $\boldsymbol{r}_{un}$, is for a sample with an analyte concentration which is higher than the detection limit with α and β fixed probabilities of error.

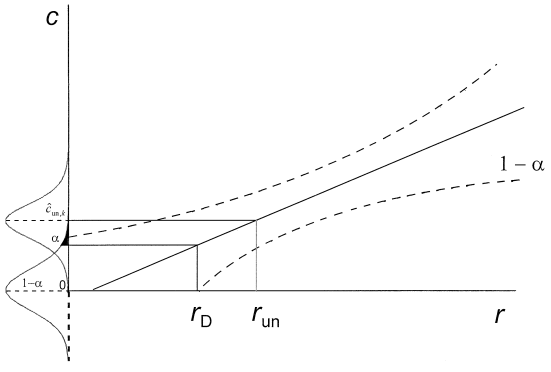

Fig. 1. Projection of a multivariate inverse calibration model on the plane formed by the axis of response variables (concentrations) and predictor variables (instrumental responses). The connection with multivariate models can be seen by realising that instead of *one* slope, there are multiple slopes for a given analyte in the multivariate situation, all of them collected in the regression vector $\boldsymbol{b}_k$. The probability of committing a type I error is shown in black, i.e., the probability that a response corresponding to an unknown sample, $\boldsymbol{r}_{un}$, can be given a concentration higher than zero, when the analyte is not even present.

In step 3, the expression of variance of the predicted concentration must be chosen according to the main sources of variability present in the data. Adapting the methodology of Hubaux and Vos [11] to the multivariate domain, the detection response in step 4, $\hat{r}_D$, is the intersection of the lower confidence interval of the multivariate model with the plane of responses for an analyte concentration of zero. Mathematically, this can be obtained from Eq. (9a):

$$\hat{c}_k - \hat{\gamma} = 0 \tag{9a}$$

which, after introducing Eqs. (1) and (8) and rearranging, becomes:

$$\sum_{j=1}^{J} r_j \hat{b}_j = t_{1-\alpha,\nu} \mathrm{var}(\hat{c}_k)^{1/2} \tag{9b}$$

where the t statistic is one-sided in this particular case.

In multivariate regression, there are multiple detection responses, $\hat{r}_D$, which are the result of the intersection of the lower confidence interval of the calibration model with the hyperplane in dimension J of the instrumental responses. That is to say, different samples, all of which have the same concentration of the analyte in study but have a different matrix composition, and so, different spectra. All of these responses are solutions of Eq. (9b), so to unequivocally define $\hat{r}_D$, the approach adopted in this paper consists of defining a straight line joining $\boldsymbol{r}_{un}$ and the origin of the coordinates (see Appendix A). The intersection of this straight line with the lower confidence interval gives the detection response, $\hat{r}_D$, that can be seen in Fig. 2. Chemically, this means that $\hat{r}_D$ is a response which would correspond to a sample with a matrix composition proportional to that of the unknown sample, but without the presence of the analyte under study (with an α probability of error).

2.3.1. Detection limit calculation

Furthermore, for the same instrumental response, $\boldsymbol{r}_{un}$, there is a β probability of committing a type II error. So, the concentration corresponding to the detection limit (taking into account both probabilities of error) has to be calculated.

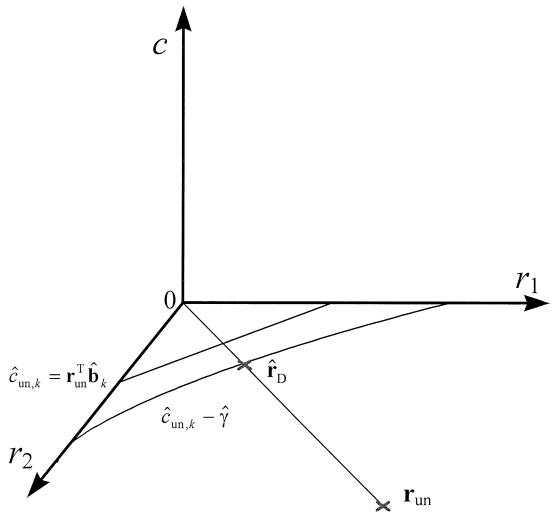

Fig. 2. Plot of the geometrical estimation of the detection response, $\hat{\mathbf{r}}_{\mathrm{D}}$. r_1, r_2 and **c** represent orthogonal *x*-, *y*- and *z*-axes, respectively.

In order to well define both probabilities of error, the detection response, $\hat{r}_{\mathrm{D}}$, is projected onto the top edge of the hyperbola which defines the confidence interval of the multivariate model and which is constructed in this case with a significance level $(1-\beta)$, so obtaining the concentration, $(\mathrm{MDL})_k$, which corresponds to a single detection limit for each unknown sample (Fig. 3):

$$(\mathrm{MDL})_k = \sum_{j=1}^{J} r_{\mathrm{D},j}\hat{b}_j + t_{1-\beta,\nu}\mathrm{var}(\hat{c}_{\mathrm{D},k})^{1/2} \qquad (10a)$$

or analogously, by introducing Eq. (9b) (see also Appendix A):

$$(\mathrm{MDL})_k = (t_{1-\alpha,\nu} + t_{1-\beta,\nu})\mathrm{var}(\hat{c}_{\mathrm{D},k})^{1/2} \qquad (10b)$$

where $\mathrm{var}(\hat{c}_{\mathrm{D},k})$ is the variance of the concentration derived from the detection response, $\hat{r}_{\mathrm{D}}$ and can be calculated from either Eqs. (2), (4), (5) or (6) by substituting $\hat{r}_{\mathrm{D}}$ for $\boldsymbol{r}_{\mathrm{un}}$. The term $t_{1-\alpha,\nu}\mathrm{var}(\hat{c}_{\mathrm{D},k})^{1/2}$ can be easily recognised as the critical level [8], which is the minimum significant value of the estimated concentration that can be taken as different from zero, with an α probability of error. The limit of detection, $(\mathrm{MDL})_k$, can be seen then as the concentration for which there is a probability β that the estimated value (when the analyte is present) does not exceed the critical value. With this definition, the probability of having a false negative can be reduced at the expense of a higher limit of detection. Comparing $(\mathrm{MDL})_k$ with the concentration value predicted by the model will enable us to decide if the analyte in the sample is detectable or not, with pre-set α and β probabilities of error.

A statistical improvement was provided in the univariate field by Clayton et al. [9]. Their approach establishes that when the parameters of the calibration model are not known and have to be estimated, then the probability of a type II error (false negative) follows a non-central *t* distribution, with non-centrality parameter $\Delta(\alpha,\beta)$. The critical value, i.e., the concentration from which the analyte is said to be detected with an α probability of error is defined as in the previous approach. On the other hand, the concentration value for which the false negative is β, for a given value of the critical value, is:

$$(\mathrm{MDL})_k = \Delta(\alpha,\beta)\mathrm{var}(\hat{c}_{\mathrm{D},k})^{1/2} \qquad (11)$$

where $\Delta(\alpha,\beta)$ is the non-centrality parameter of a non-central *t*-distribution. This parameter can be estimated from statistical tables [9,19] or computer algorithms [20]. Results obtained by Eqs. (10b) and (11) tend to be very similar when the number of degrees of freedom is high, which is often the case in multivariate calibration.

It has to be pointed out that Eqs. (10b) and (11) are valid only for the homoscedastic case, that is,

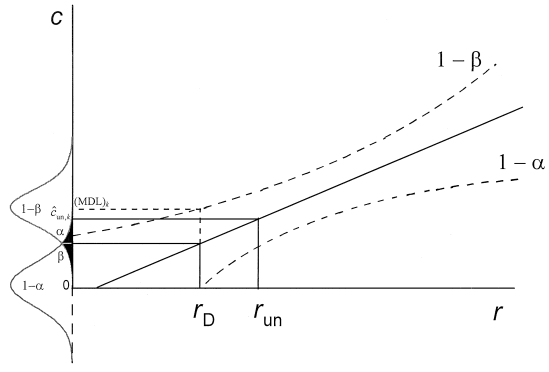

Fig. 3. Comparison of the concentration, $\hat{c}_{\mathrm{un},k}$, derived from a response, $\boldsymbol{r}_{\mathrm{un}}$, with the detection limit, $(\mathrm{MDL})_k$, derived from the detection response, $\hat{r}_{\mathrm{D}}$. In this case, as $\hat{c}_{\mathrm{un},k} < (\mathrm{MDL})_k$, the analyte would not be detectable, with an α probability of committing a type I error and a β probability of committing a type II error. It can be seen that in terms of the magnitude of the confidence intervals (constructed with significance levels of $1-\alpha$ and $1-\beta$, respectively), the analyte may actually be detectable.

when the variance in the concentration domain is constant between the null and the alternative hypothesis, which is, in fact, a reasonable assumption in most occasions. In the heteroscedastic case, however, the concentration at the level of the limit of detection is governed by the variance at that level, and not by the variance at zero concentration level. In fact, a non-central t parameter should be used, but statistical tables are not available in the heteroscedastic case. Although different approximations in the univariate case have been developed by Currie [21], a rigorous expression to deal with this problem is still lacking in the multivariate situation.

According to the IUPAC, the $(\mathrm{MDL})_k$ is the concentration for which the probability that the estimated concentration value, $\hat{c}_{\mathrm{un},k}$, does not exceed the critical value is β. So, for a given test sample, the concentration of analyte will be detected, with fixed α and β probabilities, as long as $\hat{c}_{\mathrm{un},k} \geq (\mathrm{MDL})_k$. So it may happen that for an instrumental response, $\boldsymbol{r}_{\mathrm{un}}$, the analyte concentration found, $\hat{c}_{\mathrm{un},k}$, is detectable or not depending on the α and β probabilities of error chosen, as can be seen in Fig. 3.

The major advantage of both methods is that, as they are based on linear relationships, the normality of distributions for the predicted concentrations is preserved. However, it should be emphasised that MDLs are sample-dependent and also model-dependent. So, it is only possible, for a given sample, to state that the concentration of the analyte is detectable or not (only for that sample) with assumed risks of taking false positive and false negative decisions.

3. Experimental section

3.1. Samples and software

The calibration data set consisted of the NIR spectra of gasoline samples, supplied by Repsol. Twenty-nine samples were used in the modelling step. The aromatic content was determined by a reference gas-chromatographic method of analysis, and ranged between 3% and 12.5% of the total composition of the sample. Furthermore, for the validation step, 30 replicates of three different samples (2.19%, 3.24% and 4.60% of aromatic content) were analysed by NIR in different days and by different analysts (this gives an estimation of the time-intermediate precision of the method). In all cases, samples were stored at 5°C in aluminium bottles, to protect them from UV radiation.

The uncertainty associated with the reference method was calculated from the standard deviation of 10 replicate samples, with a concentration level of aromatics of 11%. A value of $\hat{\sigma}_{\mathrm{c}} = 0.17$ was obtained.

3.2. Instrumental

Spectra were collected of samples at room temperature by a Unicam-Masso Galaxy spectrophotometer, equipped with a tungsten source and a SePb detector. Spectra were taken in the interval from 869 to 1613 nm. Absorbance responses were obtained at 2749 wavelengths; however, it was found out that using fewer data does not affect the prediction ability of the model, so finally, responses at only 687 wavelengths were used. Every spectrum, including the background spectrum, was the average result of 16 scans. Some 1-cm pathlength cells were used. They were made of special quartz for NIR to ensure a small absorption in the working interval.

3.3. Software

The regression method used was PCR, from the SVD algorithm incorporated into the MATLAB software for Windows, version 4.0. (The Mathworks, MA, USA).

4. Results and discussion

4.1. Model building and validation

Prior to the model building, the spectra of the 29 samples were baseline corrected. This was done by subtracting from each spectrum the absorbance corresponding to a non-informative wavelength, in this case at 1100 nm. This procedure has been shown to give good results with this type of data [22]. Fig. 4 shows the spectra before and after performing the baseline correction. Both matrix $\mathbf{R}$ of responses and vector $\boldsymbol{c}$ of concentrations were also mean centred. A PCR model was built and the optimal number of components was chosen by a leave-one-out cross-

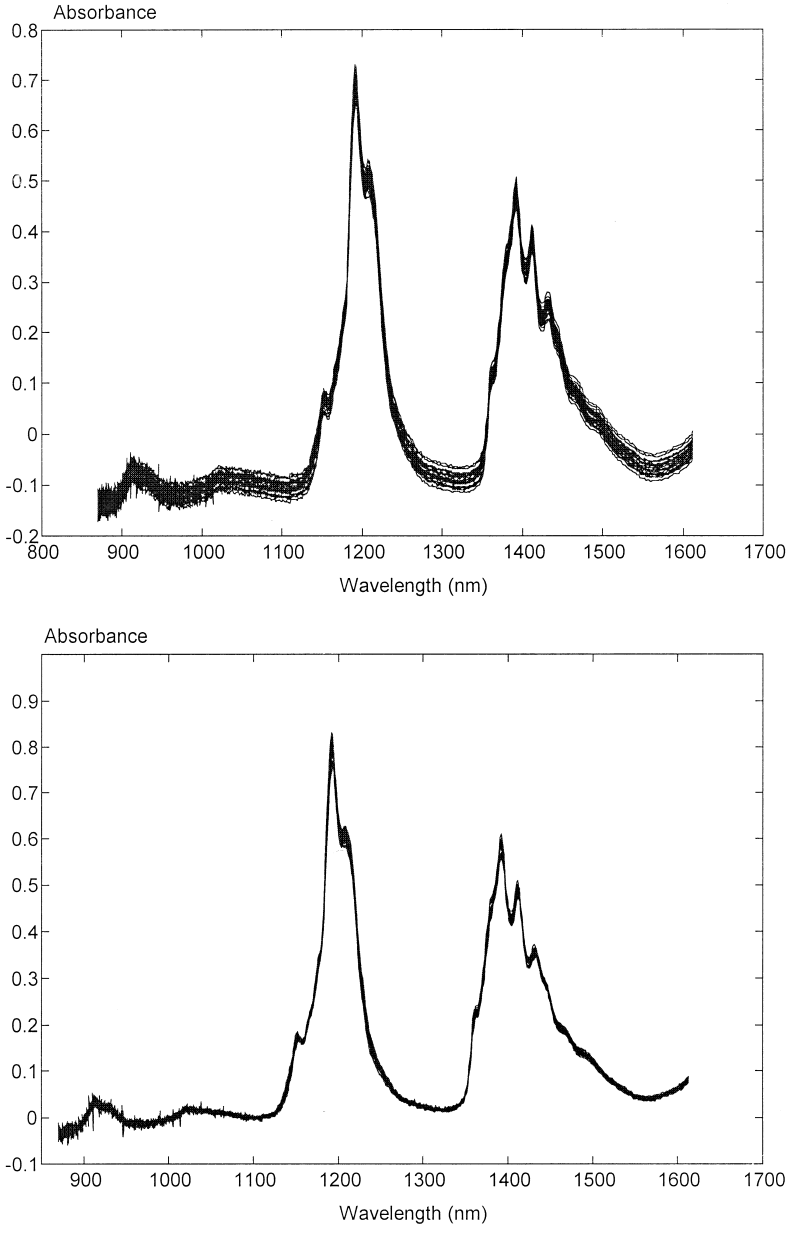

Fig. 4. NIR spectra of the 29 calibration samples. (a) Before baseline correction and (b) after baseline correction.

validation procedure. Fig. 5 shows the predictive residual error sum of squares (PRESS) as a function of the complexity of the model.

A model with five principal components was found to be optimal in terms of prediction. For this model, a root-mean-square error of cross-validation (RMSECV) of 0.87 was obtained. The root-mean-square error of calibration was 0.64 and finally, $\hat{\sigma}_{\varepsilon}^{2}$, the fit of the model to the calibration data (corrected for $\nu = 29 - 5 - 1 = 23$ degrees of freedom) was 0.72.

It is noted that the estimated RMSECV is based on measured reference values and not on true values. So, what is being measured is an *apparent* RMSECV, as

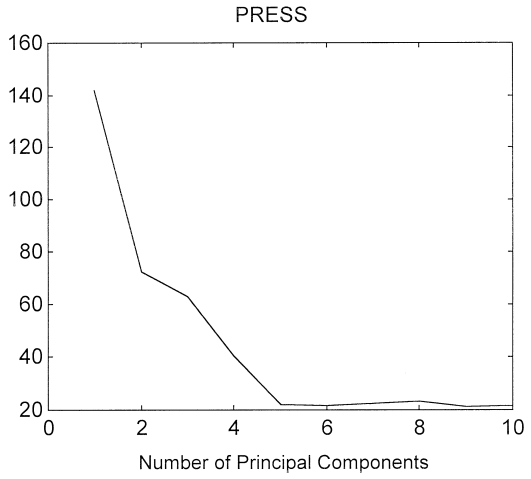

Fig. 5. Cumulative PRESS vs. the number of principal components in the PCR model.

it has been pointed out by DiFoggio [23]. The differences between measured and predicted values depend on both the errors in the reference method and the inherent error of the model (lack of fit). Even if the multivariate model is performing better than the reference method, one will never be able to prove it by using validation samples whose concentrations are known only to the same accuracy as the calibration samples. This is one reason why RMSECV values are always higher than RMSEC (root-mean-square error of calibration) values in common applications.

There is, then, a random component in the measured reference values that contributes to the RMSECV, giving rise to a biased upwards estimator. This random component is precisely the error associated with the reference method, $\hat{\sigma}_{\mathbf{c}}$. In this particular case, $\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 \gg \hat{\sigma}_{\mathbf{c}}^2$ and therefore, the RMSECV is not very influenced by $\hat{\sigma}_{\mathbf{c}}$. Furthermore, in this work the measurement error in the spectra was also neglected, due to the high repeatability of the NIR instrumental measures. Taking into account all these considerations, it seems that application of Eq. (5) is an adequate way of calculating the variance of the predicted concentrations.

The model was tested for bias by using the approach developed by Riu and Rius [24], which consists of a regression of the predicted vs. the measured concentration values for each of the cross-validation samples and a subsequent joint F-test of the slope and the intercept, taking into account the errors in both axes. Fig. 6 shows the joint confidence region of the F-test at a significance level of 95%. The point in the centre of the ellipse represents the experimental values (intercept = 0.84, slope = 0.88). As the theoretical point (intercept = 0, slope = 1) lies within the boundaries of the ellipse, it can be concluded that the PCR model with five factors is not significantly biased at this probability level. So, this model can be safely used for the calculation of both confidence intervals for the predicted concentrations and detection limits.

4.2. Validation of the MDL estimator

The MDL estimator was validated with three samples of different concentrations (2.19, 3.24 and 4.60%), close to the inferior limit of application of the model. Only the results obtained by the approach of Clayton et al. are presented, which are very similar to the ones obtained from the Hubaux and Vos methodology. From Eq. (11), the non-centrality parameter for each concentration level was calculated as:

$$\hat{\Delta}(\alpha,\beta) = \frac{c_k}{\hat{\mathrm{var}}(c_k)^{1/2}} \qquad (12)$$

where c_k is the concentration level of the analyte and $\hat{\mathrm{var}}(c_k)^{1/2}$ is the standard deviation of the replicate

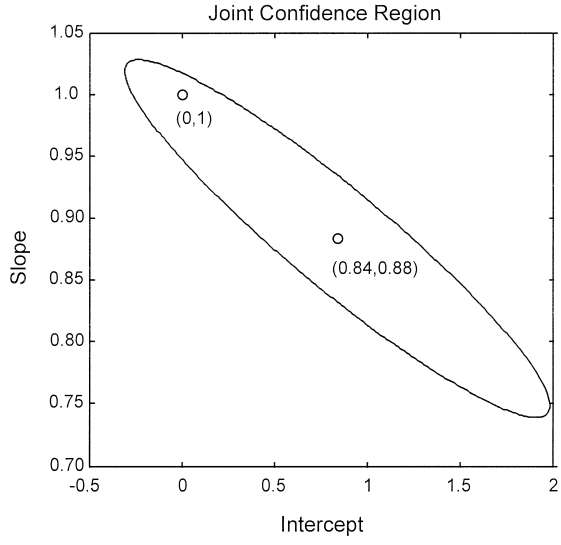

Fig. 6. Joint F-test for bias in the PCR model. Confidence interval for the slope and the intercept at a significance level $(1-\alpha) = 0.95$.

Table 1
Standard deviation of 26 replicates, non-centrality parameter and theoretical and experimental β values (at different α probabilities) for each of the three validation samples

Sample content (%)	STD ($\hat{c}_k$)	$\Delta(\alpha,\beta)$	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.20$	
			β_{the}	β_{exp}	β_{the}	β_{exp}	β_{the}	β_{exp}	β_{the}	β_{exp}
2.19	0.74	2.94	0.32	0.23 (6/26)	0.11	0.12 (3/26)	0.05	0.04 (1/26)	0.02	0.04 (1/26)
3.24	1.26	2.57	0.46	0.19 (5/26)	0.20	0.12 (3/26)	0.11	0.12 (3/26)	0.04	0.08 (2/26)
4.60	1.28	3.60	0.14	0.08 (2/26)	0.03	0.08 (2/26)	0.01	0.04 (1/26)	0.003	0.00 (0/26)

The numbers in brackets are the number of times that the predicted concentration of the replicate was found to be lower than the critical level.

measurements for that sample, considering that the variance is constant along the concentration values. Subsequently, the probability of error β can be calculated, for a given number of degrees of freedom, by fixing the probability of error α. The results are shown in Table 1 (denoted by β_{the}) for the three validation samples and different α probabilities of error.

Four of the thirty replicate measurements in each of the samples were found to be outliers and consequently were skipped in the calculations. Then, using Eq. (1), the aromatic content in every validation sample was predicted. These 26 predicted concentrations, $\hat{c}_{un,k}$, were compared with the values of the critical level, $t_{1-\alpha,\nu}\text{var}(\hat{c}_{D,k})^{1/2}$, for each replicate, being $\text{var}(\hat{c}_{D,k})^{1/2}$ calculated in this case from Eq. (5). The analyte was regarded as being detectable if the predicted concentration is greater than the critical level. So, the percentage of times that $\hat{c}_{un,k}$ is lower than $t_{1-\alpha,\nu}\text{var}(\hat{c}_{D,k})^{1/2}$ can be considered as an estimation of the rate of false negatives. This percentage was compared with the one calculated theoretically (β_{the} in Table 1), given α and the corresponding degrees of freedom. Results (denoted as β_{exp}) are also shown in Table 1. A fairly good agreement between the experimental and theoretical probabilities can be observed. However, it must be borne in mind that relatively few replications were carried out. So, it would be expected to obtain better results for a high number of replications.

4.3. Calculation of the MDL

Two of the previous validation samples were used to test the approaches presented. One with an aromatic content of 2.19% (slightly below the inferior limit of application) and the second with a concentration above this limit (3.24%). Limits of detection were calculated for these two samples by applying the two approaches at different levels of probabilities α and β. Differences between the two approaches were about 1% and even less when the number of degrees of freedom increases and so, only the results from Clayton et al. approach are presented. Eq. (5) was used in all the calculations. Results are shown in Table 2.

Table 2
Multivariate detection limits (expressed in units of % aromatic content) for the two samples in study at different levels of probability (from the approach of Clayton et al.)

	Sample 1 (2.19%)				Sample 2 (3.24%)			
	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$\beta = 0.50$	0.75	1.15	1.48	2.14	0.78	1.19	1.53	2.22
$\beta = 0.20$	1.51	1.90	2.24	2.92	1.55	1.97	2.32	3.03
$\beta = 0.10$	1.90	2.30	2.64	3.33	1.96	2.37	2.73	3.45
$\beta = 0.05$	2.22	2.63	2.97	3.66	2.29	2.71	3.07	3.80
$\beta = 0.01$	2.83	3.24	3.59	4.30	2.92	3.34	3.71	4.45

The values of the limit of detection vs. the β probability of error for different values of α, can also be plotted, giving rise to the so-called *characteristic curves of detection*. Fig. 7 shows this curve for the first sample. From these curves, it can be seen that the concentration level of the analyte in study can be detected or not, depending on the chosen risk of committing a false positive or a false negative. For these particular samples, the predicted concentrations by the model were 2.99 and 3.48%, respectively. By looking at Table 2, both samples would fail to be detected when both α and β have very low values, which is, in fact, a very severe requirement, not chosen in practice.

Any further improvement of the ability of the method to detect lower levels of analytes in samples, can only be achieved by improving the model performance in terms of prediction or by reducing the magnitude of the different sources of error that contribute to the variance of prediction. It is just necessary to have a closer look at Eq. (2) or Eq. (5) in this case, to confirm it. So, as it has been pointed out above, validation of the calibration model, which includes the test for bias, is a fundamental step in the calculation of multivariate limits of detection. Further research is being carried out to study the effect on MDLs of different parameters related to the calibration models (as the type of regression, number of factors or number of samples, among others).

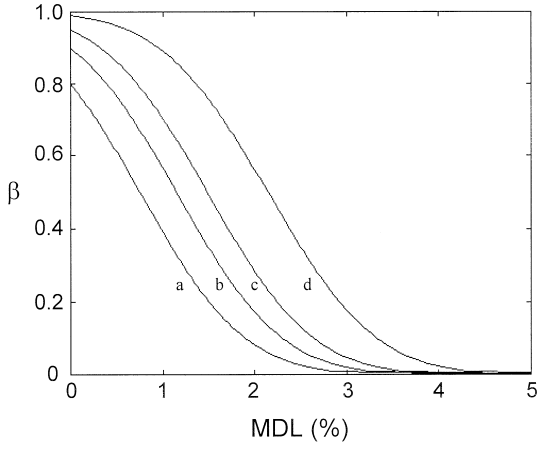

Fig. 7. Characteristic curves of detection at different levels of probability α for a sample with an aromatic content of 2.19%. (a) 0.2, (b) 0.1, (c) 0.05 and (d) 0.01.

5. Conclusions

We have discussed a new different approach for calculating MDL for inverse models which is able to quantify the probabilities of committing type I and type II errors, but further work has still to be done. Several components are involved in the deduction of MDL which can be well differentiated in the process of calculation (such as the individual terms of the variance of the test sample or the calculation of the detection response, $\boldsymbol{r}_D$). Future work has to focus on this two main issues. Development of a general expression for the variance of the predicted concentration, valid also in case of heteroscedastic measurement errors in the response variables, is required. Related to this, a general expression for the MDL when the variance in the concentration axis is non-constant is still not available, which has to be with the no-existence of statistical tables for the non-centrality parameter in the heteroscedastic case. Moreover, calculation of the degrees of freedom (to be used in the statistical tests of hypothesis and confidence intervals) associated to a complex variance estimate is a problem not totally solved. Finally, further research is being developed for the calculation of a non-ambiguous detection response. The idea is to calculate the net analyte signal of the test sample, i.e., the part of the spectrum which is unique for the analyte in study, and statistically compare it with the net signal in the null hypothesis, i.e., when the analyte is not present in the sample. This last net signal might be obtained by performing a regression model based on the net analyte signal of the calibration samples.

Acknowledgements

Financial support from the Spanish Ministry of Education and Science (DGICyT project PB96-1008) is gratefully acknowledged.

Appendix A. Mathematical derivation of the detection response

Assuming that the *n*-dimensional space is formed of *n* orthogonal axes $X_1, X_2, \ldots, X_n$ and that the values $x_1, x_2, \ldots, x_n$ are the coordinates for a point

on each of the axes $X_1, X_2, \ldots, X_n$, respectively, a straight line that passes through two points, $\mathbf{P}_1$ (x_{11}, $x_{21}, \ldots, x_{n_1}$) and $\mathbf{P}_2$ ($x_{12}, x_{22}, \ldots, x_{n_2}$) in the n-dimensional space, is expressed according to Eq. (A.1):

$$\frac{x_1 - x_{11}}{x_{11} - x_{12}} = \frac{x_2 - x_{21}}{x_{21} - x_{22}} = \cdots = \frac{x_n - x_{n_1}}{x_{n_1} - x_{n_2}}. \quad (A.1)$$

Continuing this development, an equation must be deduced for the straight line which passes through the instrumental response of the unknown sample, $\boldsymbol{r}_{un}$ ($r_{un,1}, r_{un,2}, \ldots, r_{un,J}$), and the origin of coordinates of the model, in the space which is now J-dimensional and formed by the variables (or sensors).

For a centred model, defining $\boldsymbol{r}_m$ ($r_{m,1}, r_{m,2}, \ldots, r_{m,J}$) as the vector of means of the responses for each sensor, the origin of coordinates is given by the point $\mathbf{P}_1$ ($-r_{m,1}, -r_{m,2}, \ldots, -r_{m,J}$) and the response of the unknown sample by point $\mathbf{P}_2$ ($r_{un,1} - r_{m,1}, r_{un,2} - r_{m,2}, \ldots, r_{un,J} - r_{m,J}$). The equation of the straight line which passes through these two points is given by Eq. (A.2):

$$\frac{r_1 - (-r_{m,1})}{-r_{m,1} - (r_{un,1} - r_{m,1})} = \frac{r_2 - (-r_{m,2})}{-r_{m,2} - (r_{un,2} - r_{m,2})} = \cdots = \frac{r_J - (-r_{m,J})}{-r_{m,J} - (r_{un,J} - r_{m,J})} \quad (A.2)$$

where $r_1, r_2, \ldots, r_J$ are the coordinates of the response in the J-dimensional space. Reordering and simplifying Eq. (A.3) is finally obtained:

$$\frac{r_1 + r_{m,1}}{r_{un,1}} = \frac{r_2 + r_{m,2}}{r_{un,2}} = \cdots = \frac{r_J + r_{m,J}}{r_{un,J}}. \quad (A.3)$$

The intersection of this straight line with the lower confidence interval at zero concentration, defined by the equation $\hat{c}_k - \hat{\gamma} = 0$ is given by the point which belongs to the previous straight line and which complies with Eq. (A.4):

$$\sum_{j=1}^{J} r_j \hat{b}_j = t_{1-\alpha,\nu} \mathrm{var}(\hat{c}_k)^{1/2}. \quad (A.4)$$

Expressing all the r_j coordinates (r_2 to r_J) as a function of r_1 using Eq. (A.3) and introducing these values into Eq. (A.4), we get a second-degree equation in r_1. Once solved, substituting the numerical value of r_1 into Eq. (A.3) gives $\hat{\nu}_D$ ($\hat{r}_{D,1}, \hat{r}_{D,2}, \ldots, \hat{r}_{D,J}$), the detection response at the $(1-\alpha)$ significance level chosen.

References

[1] R. Boqué, F.X. Rius, Chemometr. Intell. Lab. Syst. 32 (1996) 11–23.
[2] A. Lorber, Anal. Chem. 58 (1986) 1167–1172.
[3] A. Lorber, B.R. Kowalski, J. Chemometr. 2 (1988) 93–109.
[4] G. Bauer, W. Wegscheider, H.M. Ortner, Fres. J. Anal. Chem. 340 (1991) 135–139.
[5] K. Faber, B.R. Kowalski, J. Chemometr. 11 (1997) 181–238.
[6] A. Lorber, K. Faber, B.R. Kowalski, Anal. Chem. 69 (1997) 1620–1626.
[7] R. Boqué, F.X. Rius, Anal. Chim. Acta, submitted.
[8] L.A. Currie, Pure Appl. Chem. 67 (1995) 1699–1723.
[9] C.A. Clayton, J.W. Hines, P.D. Elkins, Anal. Chem. 59 (1987) 2506–2514.
[10] K. Faber, B.R. Kowalski, Fres. J. Anal. Chem. 357 (1997) 789–795.
[11] A. Hubaux, G. Vos, Anal. Chem. 42 (1970) 849–855.
[12] E. Sánchez, B.R. Kowalski, J. Chemometr. 2 (1988) 247–263.
[13] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes: The Art of Scientific Computing, Cambridge Univ. Press, New York, 1989, pp. 52–64.
[14] T. Naes, H. Martens, J. Chemometr. 2 (1988) 155–167.
[15] T.V. Karstang, J. Toft, O.M. Kvalheim, J. Chemometr. 6 (1992) 177–188.
[16] K. Faber, B.R. Kowalski, Chemometr. Intell. Lab. Syst. 34 (1996) 283–292.
[17] F.E. Satterthwaite, Biom. Bull. 2 (1947) 110.
[18] F.C. Garner, G.L. Robertson, Chemometr. Intell. Lab. Syst. 3 (1988) 53–59.
[19] D.B. Owen, J. Am. Stat. Assoc. 60 (1965) 320.
[20] R. Boqué, F.X. Rius, Trends Anal. Chem., in press.
[21] L.A. Currie, Chemometr. Intell. Lab. Syst. 37 (1997) 151–181.
[22] J.J. Kelly, C.H. Barlow, T.M. Jinguji, J.B. Callis, Anal. Chem. 61 (1989) 313–320.
[23] R. DiFoggio, Appl. Spectrosc. 49 (1995) 67–75.
[24] J. Riu, F.X. Rius, Anal. Chem. 68 (1996) 1851–1857.