A new formulation to estimate the variance of model prediction: application to near infrared spectroscopy calibration

E. Fernandez-Ahumada, J.M. Roger, B. Palagos

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

A new formulation to estimate the variance of model prediction. Application to near infrared spectroscopy calibration

E. FERNANDEZ-AHUMADA*⋆ J.M. ROGER*, B. PALAGOS*

**Cemagref BP 5095 - 34033 Montpellier Cedex1 France*

Abstract

Evaluation of uncertainty affecting predictions is a major trend in analytical chemistry and chemometrics. Several approximate expressions and resampling methods have been proposed for the estimation of prediction uncertainty when using multivariate calibration. This article proposes a new expression for the variance of prediction, adapted to near infrared spectroscopy specificities and particularly to the spectral error structure, induced by the high colinearity of the variables. The proposed analytical expression enables a detailed evaluation of the different contributions and components of uncertainty affecting the model. An application to real data of feedstuff near infrared spectra related to protein content has shown its advantages.

Key words: prediction uncertainty, multivariate calibration, NIR spectroscopy

⋆ corresponding author g82feahe@uco.es

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

1 Introduction

Many decisions in different human activities (industrial, commercial, scientific, healthcare, safety, environmental) are based on analytical results. These results must be reported with some indication of their quality, so that assessment of their reliability is enabled [1]. In metrology, there are different terms to characterize the quality of a method or an instrument, such as reproducibility, repeatability, accuracy, precision, trueness and uncertainty. Definitions of all these terms can be found elsewhere [2]. This paper focuses on the uncertainty whose formal definition in the Guide to the Expression of Uncertainty in Measurement is as follows: "parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand" [3].

For the case of near infrared (NIR) spectroscopy where multivariate calibration is commonly used to construct a predictive model on the basis of multiple predictor variables, the predicted value is incomplete without a statement about its uncertainty. In the chemometrics-oriented literature, considerable attention has been paid to prediction uncertainty estimation. According to recent reviews [4], there are two basic ways of estimating prediction uncertainty, namely, error propagation and resampling strategies. Error propagation leads to closed-form expressions where different assumptions are considered but which provide a platform for evaluating the different sources of uncertainty. Resampling is essentially a black box approach which, however, is often more accurate because fewer approximations are taken into account.

The theory of error propagation has provided the framework from which many

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

authors have developed multiple expressions. These expressions are the result of evaluating each source of uncertainty associated to model inputs and considering its contribution. Most of the existing approximate expressions have been developed with a Partial Least Squares (PLS) regression model but there exist some works which used other methods such as Principal Component Regression (PCR) [5], [6] or Artificial Neural Networks (ANNs) [7]. First expressions proposed by Hoskuldsson [8], Phatak et al.[9] and Denham [10] assumed the hypothesis of negligible errors in the predictor variables. The expression of Holkuldsson was then adopted by the American Society of Testing and Materials (ASTM), considering that data were not mean-centred. Those works were expanded by Faber and Kowalski [11] who included errors in the predictor variables under the general errors-in-variables (EIV) model. A drawback of their approach is that the original expression is derived under the assumption that the errors in the predictor variables have constant variance (the homoscedastic case). Later on Faber and Bro [12] proposed a new expression which accommodated for heteroscedastic and correlated errors. But in fact, the expression was derived under the assumption that the errors in predictor variables are identically and independently distributed (i.i.d.) and the authors conjectured that it applied to most types of heteroscedasticity.

In spectroscopy, error measurements in predictors are unlikely to be uncorrelated and with constant variance. That is the reason why the purpose of this study is the proposal of a new prediction uncertainty expression for linear calibration models, where these issues are considered.

The idea of the new proposal is to build an expression as general as possible where the minimum of hypotheses are assumed. The procedure, as for other approximate expressions, is based on the classical EIV model, aiming at con-

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

sidering all sources of uncertainty affecting predictors, the dependent variable and model coefficients. The expression may be used for evaluating the model. It is an analytical expression which enables to characterize the different sources of uncertainty affecting the model. For the case of near infrared spectra, these sources might be the instrument (repeatability, reproducibility), the sampling, deviation of Lambert Beer's law, etc. The aim of this new expression is to give the basis to calculate the overall uncertainty depending on different types of error sources.

The paper is organised as follows: first, a theoretical section reminds the most complete expression found in literature, considering the assumptions taken into account. Then the new expression is proposed. A discussion and interpretation of both expressions is provided. Material and methods section gives details on the real data set used for comparison of both expressions when different types of errors are evaluated and the procedure followed to estimate each term of expressions. Next section shows the main results obtained from the comparison of both expressions. Their performance is also assessed from estimates obtained using a resampling method. A discussion of all these results is provided. Last section contains the most important conclusions achieved.

2 Theory

2.1 Notations and theoretical recalls

Capital bold characters will be used for matrices, e.g. $\mathbf{A}$; small bold characters for column vectors, e.g. $\mathbf{a}_i$ will denote the i^{th} column of $\mathbf{A}$; row vectors will be denoted by the transpose notation, e.g. $\mathbf{a}_j^{\mathrm{T}}$ will denote the j^{th} row of $\mathbf{A}$. Non

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

bold characters will be used for scalars, e.g. matrix elements a_{ij} or indices i. If needed, matrix dimensions will be indicated as indexes, e.g $\mathbf{A}_{(N\times P)}$. The trace of a square matrix $\mathbf{A}$ will be noted $\mathrm{tr}(\mathbf{A})$.

Any measured or estimated quantity (scalar a or vector $\mathbf{a}$) is assumed to be a random variable and will be noted as $\widehat{a}$ or $\widehat{\mathbf{a}}$. The random part of these entities, which carries the errors, will be noted as δa or $\delta\mathbf{a}$ and the true (unknown) part as $\check{a}$ or $\check{\mathbf{a}}$, so that :

$$\widehat{a} = \check{a} + \delta a \quad \text{or} \quad \widehat{\mathbf{a}} = \check{\mathbf{a}} + \delta\mathbf{a} \tag{1}$$

Operator $\mathbf{Var}()$ will denote the variance-covariance matrix of a random vector and Var() will denote the variance of a random scalar variable. From equation (1), $\mathbf{Var}(\widehat{\mathbf{a}}) = \mathbf{Var}(\delta\mathbf{a})$ and $Var(\widehat{a}) = Var(\delta a)$. For clarity reasons, $\mathbf{Var}(\delta\mathbf{a})$ could be noted as $\mathbf{\Sigma}_a$ and $\mathrm{Var}(\delta a)$ could be noted as σ_a^2.

Let $\mathbf{k}$ be a fixed (not random) vector, $\mathbf{u}$ and $\mathbf{v}$ two random vectors. The following formulae are recalled:

$$\mathrm{Var}(\mathbf{k}^{\mathrm{T}}\mathbf{u}) = \mathbf{k}^{\mathrm{T}}\mathbf{\Sigma}_u\mathbf{k} \tag{2}$$

$$\mathbf{Var}(\mathbf{u} \pm \mathbf{v}) = \mathbf{\Sigma}_u + \mathbf{\Sigma}_v \pm 2\mathbf{Cov}(\mathbf{u}, \mathbf{v}) \tag{3}$$

Where $\mathbf{Cov}(\mathbf{u}, \mathbf{v})$ is the matrix containing the covariances between all the components of $\mathbf{u}$ and those of $\mathbf{v}$.

If $\delta\mathbf{u}$ and $\delta\mathbf{v}$ are independent, the following relation is verified:

$$\mathrm{Var}(\delta\mathbf{u}^{\mathrm{T}}\delta\mathbf{v}) = \mathrm{tr}(\mathbf{\Sigma}_u\mathbf{\Sigma}_v) \tag{4}$$

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

The vector $\mathbf{b}_{(P\times 1)}$ is the regression vector between the P predictors of $\mathbf{x}$ and the response y, calibrated on N samples, so that:

$$\widehat{y} = (\widehat{\mathbf{x}} - \widehat{\mathbf{x}}_c)^{\mathrm{T}}\widehat{\mathbf{b}} + \widehat{y}_c = \widehat{\mathbf{z}}^{\mathrm{T}}\widehat{\mathbf{b}} + \widehat{y}_c \tag{5}$$

where $\mathbf{x}_c$ and y_c correspond to the mean sample of the learning set and $\mathbf{z} = \mathbf{x} - \mathbf{x_c}$.

The scenario is covered under the so-called classical errors-in-variables (EIV) model ([13]) :

$$\widehat{y} = \check{y} + \delta y \tag{6}$$
$$\widehat{\mathbf{z}} = \check{\mathbf{z}} + \delta\mathbf{z} \tag{7}$$

Substituting eq. (6) and eq. (7) into eq. (5) results in the following equation:

$$\widehat{y} = \check{\mathbf{z}}^{\mathrm{T}}\check{\mathbf{b}} + \delta\mathbf{z}^{\mathrm{T}}\check{\mathbf{b}} + \check{\mathbf{z}}^{\mathrm{T}}\delta\mathbf{b} + \delta\mathbf{z}^{\mathrm{T}}\delta\mathbf{b} + \check{y}_c + \delta y_c \tag{8}$$

2.2 Classical approach

In the literature ([11], [14]) the following assumptions are considered:

C1: Measurement errors are independently and identically distributed (i.i.d.) with zero mean and constant variance

C2: Error characteristics of the prediction objects are the same as the ones of training objects

C3: The product of errors (random parts of an entity) is neglected

C4: $\widehat{\mathbf{b}}$ is independent of $\widehat{\mathbf{x}}_c$ and $\widehat{y}_c$

Hypothesis (C1) yields $\mathbf{Var}(\delta\mathbf{x}) = \sigma_x^2\mathbf{I}$, where $\mathbf{I}$ is an appropriate dimensioned identity matrix. By applying all the other hypotheses on eq. (8), the following

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

expression stands for the variance of the estimation $\widehat{y}$:

$$\mathrm{Var}(\widehat{y}) = 0 + \left(1 + \frac{1}{N}\right) \check{\mathbf{b}}^2 \sigma_x^2 + \check{\mathbf{z}}^{\mathrm{T}} \mathbf{Var}(\delta\mathbf{b}) \check{\mathbf{z}} + 0 + 0 + \frac{\sigma_{lab}^2}{N} \tag{9}$$

where σ_{lab}^2 is the laboratory variance, affecting the y reference. Equation (9) yields:

$$\mathrm{Var}(\widehat{y}) = \left(1 + \frac{1}{N}\right) \check{\mathbf{b}}^2 \sigma_x^2 + \check{\mathbf{z}}^{\mathrm{T}} \mathbf{\Sigma}_b \check{\mathbf{z}} + \frac{\sigma_{lab}^2}{N}$$

In practice, to evaluate this expression, true values are replaced by measured or estimated values, yielding:

$$\mathrm{Var}(\widehat{y}) = \left(1 + \frac{1}{N}\right) \widehat{\mathbf{b}}^2 \sigma_x^2 + \widehat{\mathbf{z}}^{\mathrm{T}} \mathbf{\Sigma}_b \widehat{\mathbf{z}} + \frac{\sigma_{lab}^2}{N} \tag{10}$$

The above expression will be referred to in the following as the *classical expression.*

2.3 New proposal

In this study, the following hypotheses are supposed to be fulfilled:

N1: Measurement errors have zero mean

N2=C2: Error characteristics of the prediction objects are the same as the ones of training objects

N3: $\delta\mathbf{z}$ and $\delta\mathbf{b}$ are independent

N4=C4: $\widehat{\mathbf{b}}$ is independent of $\widehat{\mathbf{x}}_c$ and $\widehat{y}_c$.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Considering the hypotheses (N3) and (N4), each term of the sum of equation (8) is independent of the others. Thus, all the covariance terms are null and the variance of $\widehat{y}$ is given by:

$$\mathrm{Var}(\widehat{y}) = 0 + \check{\mathbf{b}}^{\mathrm{T}}\mathbf{Var}(\delta\mathbf{z})\check{\mathbf{b}} + \check{\mathbf{z}}^{\mathrm{T}}\mathbf{Var}(\delta\mathbf{b})\check{\mathbf{z}} + \mathrm{Var}(\delta\mathbf{z}^{\mathrm{T}}\delta\mathbf{b}) + 0 + \frac{\sigma_{lab}^2}{N} \tag{11}$$

But

$$\begin{aligned}\mathbf{Var}(\delta\mathbf{z}) &= \mathbf{Var}(\delta\mathbf{x} - \delta\mathbf{x}_c) &(12)\\ &= \mathbf{Var}(\delta\mathbf{x}) + \mathbf{Var}(\delta\mathbf{x}_c) - 2\mathbf{Cov}(\delta\mathbf{x}, \delta\mathbf{x}_c) &(13)\end{aligned}$$

Since it is difficult to establish the independence of $\delta\mathbf{x}$ and $\delta\mathbf{x}_c$, it is considered that if they are dependent, they both vary in the same sense. Thus, their covariance is either null or positive. By the way, neglecting the covariance terms will at the best be ineffective and at the worst give an overestimation of the variance. The expression (13) becomes:

$$\begin{aligned}\mathbf{Var}(\delta\mathbf{z}) &= \mathbf{Var}(\delta\mathbf{x}) + \mathbf{Var}(\delta\mathbf{x}_c) &(14)\\ \mathbf{Var}(\delta\mathbf{z}) &= \mathbf{Var}(\delta\mathbf{x}) + \frac{1}{N}\mathbf{Var}(\delta\mathbf{x}) &(15)\\ \mathbf{Var}(\delta\mathbf{z}) &= \left(1 + \frac{1}{N}\right)\mathbf{Var}(\delta\mathbf{x}) &(16)\end{aligned}$$

Equation (11) then becomes:

$$\mathrm{Var}(\widehat{y}) = \left(1 + \frac{1}{N}\right)\check{\mathbf{b}}^{\mathrm{T}}\mathbf{Var}(\delta\mathbf{x})\check{\mathbf{b}} + \check{\mathbf{z}}^{\mathrm{T}}\mathbf{Var}(\widehat{\mathbf{b}})\check{\mathbf{z}} + \mathrm{Var}(\delta\mathbf{z}^{\mathrm{T}}\delta\mathbf{b}) + \frac{\sigma_{lab}^2}{N} \tag{17}$$

Using the hypothesis (N3) on $\delta\mathbf{z}$ and $\delta\mathbf{b}$ and combining eq. (4), eq. (16), and eq. (17) yields:

$$\mathrm{Var}(\widehat{y}) = \left(1 + \frac{1}{N}\right)\check{\mathbf{b}}^{\mathrm{T}}\boldsymbol{\Sigma}_x\check{\mathbf{b}} + \check{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\Sigma}_b\check{\mathbf{z}} + \left(1 + \frac{1}{N}\right)\mathrm{tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b) + \frac{\sigma_{lab}^2}{N} \tag{18}$$

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Like for the classical approach, true values are replaced by measured or estimated values. This yields the following expression, reordered to match the classical one:

$$\mathrm{Var}(\widehat{y}) = \left(1 + \frac{1}{N}\right) \widehat{\mathbf{b}}^{\mathrm{T}} \boldsymbol{\Sigma}_x \widehat{\mathbf{b}} + \widehat{\mathbf{z}}^{\mathrm{T}} \boldsymbol{\Sigma}_b \widehat{\mathbf{z}} + \frac{\sigma_{lab}^2}{N} + \left(1 + \frac{1}{N}\right) \mathrm{tr}(\boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_b) \qquad (19)$$

The above expression will be referred to in the following as the *new expression*.

2.4 Theoretical discussion

2.4.1 Discussion of the hypotheses

Hypothesis (N1), assuming that the measurement errors have zero mean, is a minimal assumption. On the contrary, hypothesis (C1) is very restrictive and rarely fulfilled in spectrometry, especially when baselines occur. With this type of noise, the errors affecting each variable of $\mathbf{x}$ are very dependent of the others and do not present constant variance.

Hypotheses (C2) and (N2) assume that the measurement conditions are the same when calibrating and when predicting. That means that this study mainly addresses the problem at the calibration stage and does not cover the case of robustness against unknown influence factors.

The hypothesis (C3) is classically stated in error propagation schemes; it assumes that product of errors can be neglected on the basis that errors are small. In the new formulation, this hypothesis has been replaced by the hypothesis (N3), which enabled the calculation of a new term: $\mathrm{tr}(\boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_b)$.

Hypotheses (C4) and (N4) assume that the model coefficients are independent

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

from the centre. In other words, that means that the slope and the offset of the model are independent.

2.4.2 Term by term discussion of the two expressions

The first term of both expressions expresses the propagation of the error part of $\mathbf{x}$ through the model $\mathbf{b}$. The classical expression may misestimate this term for two reasons : (i) it does not take into account the relationship between the space spanned by $\mathbf{\Sigma}_x$ and the model; (ii) it does not take into account the correlations between the variables of $\mathbf{x}$, which are particularly intense in spectrometry. In the new expression, the vectorial structure of this term is respected, instead of the classical expression, which considers only the norm of the vectors. It is then likely that the new expression will better manage the structured errors, like the baselines.

The second term is identical in the two expressions. It reveals how the modelling error $\delta\mathbf{b}$ is amplified by the vector $\mathbf{x}$. In the new expression, it has a similar form to the first term, contrary to the classical expression.

The third term is identical in the two expressions. It expresses the explicit contribution of the error of the reference value y, through the calculation of the model centre y_c. It should be noticed that this laboratory error also affects the second term implicitly, through $\mathbf{\Sigma}_b$.

The fourth term of the new expression has no correspondence in the classical one, which neglects the products of errors. The matrix $\mathbf{\Sigma}_x\mathbf{\Sigma}_b$ represents the intersection of the two spaces spanned by $\delta\mathbf{x}$ and $\delta\mathbf{b}$. The quantity $\mathrm{tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_b)$ measures the common part of the instances of $\delta\mathbf{x}$ and $\delta\mathbf{b}$.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

3 Material and methods

3.1 Description of the data

Data set consisted of $N = 385$ samples of compound feedstuffs with protein content as reference data. Measurements were performed at $P = 656$ wavelengths in the visible and near infrared region (380 to 1690 nm) with the instrument located on a conveyor belt. For each sample, 10 spectra were acquired, which gave N blocks of 10 repetitions. The reference protein content was determined by Kjeldahl method, with a standard error of laboratory $\sigma_{lab} = 0.2\%$ ($\sigma^2_{lab} = 0.04\%^2$). Two matrices of spectra were considered for calculations: $\mathbf{X_{rep}}(10N \times P)$ which contained all the spectra and $\mathbf{X}(N \times P)$ which contained the spectra averaged over blocks of repetitions. The vector $\mathbf{y}(N \times 1)$ contained the protein content values.

The following preprocessings were applied to the data:

- RAW: no preprocessing
- DTR: the linear trend was estimated by means of a linear regression on each spectrum and then removed
- SNV: the spectra were centred and normalized by their standard deviation
- D2: the spectra were replaced by their second derivative, calculated by Savitsky and Golay algorithm, with a width of 42 nm and a polynomial order of 3.
- D2SNV: the spectra were processed by D2 and then by SNV
- SNVD2: the spectra were processed by SNV and then by D2

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

3.2 Description of calculations

All calculations were performed and programmed with Matlab version 7.1.0 (The Mathworks, Inc.).

3.2.1 Estimation of terms

Several parameters needed to be estimated in order to obtain an approximation of each term. These parameters were $\widehat{\mathbf{b}}$, $\mathbf{\Sigma}_x$ (new proposal), σ_x^2 (classical expression) and $\mathbf{\Sigma}_b$. Depending on the type of error evaluated, some of these parameters were estimated differently. Errors studied in this paper were: sample specific error and repeatability. What we have called "sample specific error" is the error intrinsic of each sample which causes the fitting error. It is like a bias of each sample which cannot be reduced by replicating the measurements. It is an error depending only on the sample and not on the measurement. This error can come from $\mathbf{x}$ and / or y, but in this paper only influence from $\mathbf{x}$ has been considered. The other kind of error evaluated is the repeatability of the spectral measurement regarding sample presentation.

The model $\widehat{\mathbf{b}}$ between the spectra of $\mathbf{X}$ and the protein contents of $\mathbf{y}$ was estimated by means of a PLS regression (SIMPLS algorithm). Cross validation was used for the choice of the optimal number of latent variables.

For the evaluation of the sample specific error ($\delta\mathbf{x}_{ss}$):

- For each spectrum $\mathbf{x}_i$ of $\mathbf{X}$, an ideal spectrum $\widetilde{\mathbf{x}}_i$ was calculated by means of a kernel centred on its y_i value and applied on the other samples, like explained in [15]. Then, $\delta\mathbf{x}_{ss}$ was estimated as the difference between $\widetilde{\mathbf{x}}_i$ and

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

$\mathbf{x}_i$. A matrix containing all these differences was built and referred to as $\mathbf{S}$. The matrix $\mathbf{\Sigma}_x$ was calculated as the covariance matrix of $\mathbf{S}$. The variance σ_x^2 was calculated as the mean of the diagonal elements of $\mathbf{\Sigma}_x$.

- $\mathbf{\Sigma}_b$ was extracted from the bootstrap calculations detailed in (3.2.2).
- σ_{lab}^2 was considered to be estimated by replicates and provided by the laboratory which performed the analysis.

For the evaluation of the repeatability error ($\delta\mathbf{x}_{rep}$):

- Matrix $\mathbf{X_{rep}}$ was centred by blocks of 10 repetitions and the matrix obtained was referred to as $\mathbf{R}$. To obtain an estimation of $\mathbf{\Sigma}_x$, the covariance matrix of $\mathbf{R}$ was calculated. For the estimation of σ_x^2, the mean of the diagonal elements of $\mathbf{\Sigma}_x$ was considered.
- $\mathbf{\Sigma}_b$ was extracted from the bootstrap calculations detailed in 3.2.2.
- σ_{lab}^2 was considered to be estimated by replicates and provided by the laboratory which performed the analysis.

Once all these parameters were estimated, their values were injected in equations 10 and 19 to calculate the terms of classical and new expressions. As term 2 depends on the individual, its median value was retained.

3.2.2 Variance estimation by resampling

Bootstrap was used as resampling method to obtain a general variance value, where no assumptions were considered. In the following it will be noted as BS variance. The bootstrap procedure used performed n drawing with replacement of n samples.

For the sample specific errors, the procedure was similar to a cross-validation:

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

- all blocks were averaged to suppress the repeatability error
- each block was successively kept out from the others
- the other blocks were bootstrapped and used for developing a model
- the model was applied on the block kept out

For the repeatability errors, the following process was used:

- a bootstrap on the N initial blocks came up with a set of N new blocks $\{I_1, I_2, ..., I_N\}$
- a bootstrap was performed inside each block
- a random error with distribution $N(0; \sigma^2_{lab})$ was added to each y value
- the $\{I_1, I_2, ..., I_N\}$ blocks were used for calibration, and the others for the test. Two versions of model were calculated. The first one was calculated on the N averaged spectra and the second one on the $10N$ individual spectra.

At the end of 1000 iterations, the variance of predictions was calculated.

4 Results and discussion

Both expressions presented here are intended to give an estimation of $Var(\widehat{y})$ for each prediction. Nonetheless, only the second term of both expressions (which is identical in classical and new forms) depends on the sample. All the other terms are related to global characteristics of data and model. Thus, the new expression would not present a very high added value for individual uncertainty. Then, as explained in 3.2.1, results will show the median value of uncertainty calculated on the whole calibration set, enabling the discussion to be focused on the estimation of the global uncertainty.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Table 1 presents the variances calculated for sample specific errors by both classical and new expressions, according to the different preprocessings. Results obtained by resampling and the mean squared error of calibration (MSEC) are also presented.

First, the row containing results from resampling method (BS variance) is studied. These values can be considered as representing the true variance since this method uses no approximation. These BS variance values show that most of pretreatments do not produce a variance much lower than the raw model. One can notice that low variances are observed for linear pretreatments as D2, DTR or even RAW. The lowest value is obtained for D2SNV. Thus, the hypothesis could be put forward that the sample specific spectral error is made up mainly of an additive effect like a baseline, and secondly of a slight multiplicative effect. It should also be noticed that the highest values of variance are obtained when SNV is used as the single pretreatment or in first position before D2. That confirms the prevalence of the additive effect in the sample specific error.

Regarding the BS variance and MSEC, the same trend is observed. The lowest MSEC are obtained when D2 is applied as unique pretreatment or in first position followed by SNV. Model performance with DTR is similar to RAW, which could be explained by assuming that DTR is a light pretreatment. As found for the BS variance, application of SNV clearly degrades the model. The accordance observed between the BS variance and MSEC validates the way this variance is estimated.

Results with the new expression show a good agreement with those of the BS variance, values of both methods are very close to each other, independently of

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

the pretreatment applied. However values obtained with the classical expression are always overestimated, very far from the BS variance and with high sensitivity to the applied pretreatment. The explanation of these results lies in the estimation of first term and especially because of the C1 hypothesis. With this hypothesis, the norm of spectral measurement errors is considered. This implies that only the magnitude of the noise affects the estimation. By contrast, the new expression takes into account also the structure of the noise through $\mathbf{\Sigma}_x$. The fact of taking into account this structure enables the model to orthogonalize itself, possibly, against the noise. An example of this behaviour is illustrated in Figure (1). Under C1 hypothesis, the noise is represented as a sphere which cannot be orthogonal to the model and the dependence between the model and the noise does not affect the first term. On the contrary, if $\mathbf{\Sigma}_x$ presents a low dimension, as usually in spectrometry, under N1 hypothesis the model can be orthogonal to this noise and the first term can be reduced.

Also in terms of comparison of the two analytical expressions, it is worth noting that the fourth term, only present in the new expression, has an important weight in the final result (always more than 10 % of the whole variance). It is also noticeable that, in the present application, the third term, related to the direct effect of the standard laboratory error, is completely negligible.

Analyzing in depth each term of the new expression for each pretreatment, several comments can be stated:

For all pretreatments, first term is the most important one, representing between 78 and 85 % of the sum, whilst the importance of the second term varies between 4.7 and 6.3 % and the fourth one between 10.3 and 15.3 % of the whole variance.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

The highest values of first term were always observed when SNV was involved. The explanation of this observation can be found in the correction performed by SNV, which is especially employed to reduce multiplicative effects. However the effects present in the data seem to be mostly additive ones and with reduced dimension (e.g. a simple baseline). Thus, the SNV spoils the result because the normalization it performs converts the existing linear noise into a non-linear one, which cannot be handled properly by the model. The lowest value of first term was obtained for D2. With this pretreatment, the spectral error becomes less important in magnitude, decreasing the classical approach terms; and also in structure, decreasing the new expression terms. For RAW and DTR, average values of first term were observed. From these results, it can be concluded that the model itself without pretreatment is able to partially manage the sample specific error. Actually, the vector $\mathbf{b}$ is partially orthogonal to the space spanned by this noise.

Second term is considered as a median value of the distance between the spectrum and the model centre, weighted by the model noise. It depends mainly on three factors: (i) the length of the centred spectra $\mathbf{z}$, which is related to the classical concept of leverage, (ii) the norm of $\mathbf{\Sigma}_b$ and (iii) the colinearity between $\mathbf{z}$ and $\mathbf{\Sigma}_b$. Fourth term represents the dependency between two noise spaces: spectral noise and model noise. Many different parameters can affect these terms. Among others, one may think that if there is model overfitting and/or an important part of the spectral noise is used by the model, second term and fourth term will show high values. This statement can be illustrated with examples shown in Table 1:

- The regression with D2 managed to obtain a model quite insensitive to noise. The pretreatment was effective, but it probably removed some information

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

forcing the model to overfit the data. Consequently, its second and fourth terms are among the highest values.

- For the DTR model, second and fourth terms presented the lowest values. This model is the simplest, with a good performance in terms of variance and probably not overfitted.
- The SNVD2 model can be considered the most overfitted according to second and fourth terms and it also presents the highest variance. In this case, it seems that the pretreatment created a non-linear noise which prevented a good model adjustment.

Analyzing results concerning repeatability errors (Table 2 and 3), the first observation is that BS variance is less important than that of sample specific error. The order of magnitude of BS variance for the repeatability error is less than 1%, against more than 2% for sample specific error. This result implies, for the case studied here, that errors dealing with sample presentation are less important than those concerning model lack of fit, such as the presence of unknown compounds, deviations of Lambert Beer's law, etc.

As for sample specific error, the model which used SNVD2 shows the highest BS variance value. For the case of models on averaged spectra (Table 2), the BS variance presents an extremely high value (1.83), comparable to variances obtained for sample specific errors. Thus, applying SNV in first position makes second derivative harmful.

It is also observed, contrary to what was found for sample specific errors, that uncertainties obtained with D2 and D2SNV pretreatments were not better that those of simpler models (RAW and DTR). The hypothesis may be put forward that the repeatability error is more complex than the sample specific

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

one.

It is important to note that uncertainties of models performed on averaged spectra are higher than uncertainties of model developed keeping repetitions. This result shows that regression, if repetitions are kept, is able to develop a model partially independent of repeatability error. By contrast, this is not possible when using averaged spectra because a large part of the repeatability error is removed.

It is also worth noting that the more complex is the model, the bigger is the difference between uncertainties of models on averaged spectra and uncertainties of models with repetitions. The ratio is 1.1 for RAW, 1.15 for DTR and 3.0 when SNVD2 pretreatment is applied.

For the case of models performed on averaged spectra (Table 2), estimations obtained with the new expression are less sensitive to pretreatments than those of the classical approach. Compared to the BS variance, the uncertainty of the new expression is not always close to it and most of times, underestimated. The explanation of this phenomenon may be found in the violation of hypothesis N2: objects used in calibration (averaged spectra) are not comparable to objects used in test (individual spectra). Thus, in the following, all comments will be referred to the case of calibration and test performed on individual spectra (Table 3).

Regarding Table 3, estimations obtained with the new expression show very close agreement with results of BS variance. On the other hand, estimations obtained with the classical approach are overestimated except for the cases of complex models (SNVD2 and D2SNV) where the underestimation can be explained by the missing fourth term.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Concerning the analysis of new expression terms:

First term for repeatability errors (Table 3) is not systematically higher than the others, as observed for sample specific errors (Table 1). That implies that uncertainty due to propagation of repeatability error through the model is comparable to the effect of modelling errors. Since second term is of the same order than for sample specific error, the explanation may be found in the reduction of first term.

Fourth term is not negligible in any case (more than 10% of the total result for all pretreatments). For cases of SNVD2 and D2, fourth term is even the highest. The same explanations found for sample specific errors can be addressed here. For the model with SNVD2, the presence of non-linear noise makes the model very complex; and for the model with D2, the pretreatment has removed important information which provokes model overfitting.

Conclusion

The uncertainty that affects the model predictions is of major importance in analytical chemistry. However, no clear expression of this uncertainty is fully adapted to the case of NIR spectrometry. This article proposes a new expression for the variance of the prediction adapted to any linear calibration models, like e.g. PLS. This formulation respects the specificities of spectrometry and particularly the spectral error structure which is induced by the high colinearity of the variables. Four terms appear in this expression: (i) the amplification of the spectral error by the model; (ii) the amplification of the model error by the spectrum; (iii) the direct impact of the laboratory error; (iv) the depen-

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

dency between spectral error and model error. An application to real data of feedstuff NIR spectra related to protein content has shown the ability of this new expression to manage different types of errors. The estimated values of the uncertainty was in total accordance with those yielded by a resampling method. The analysis of the four terms showed that they provide complementary information on the model behaviour. Further applications of this new expression to other problems should be carried out to exploit the knowledge it yields. New figure of merits based on the four terms could be created to better qualify the calibration models.

Acknowledgements

First author wish to thank the CEMAGREF for the funds that enabled the post-doctoral contract, during which this work was developed.

References

[1] Evaluation of measurement data - guide to the expression of uncertainty in measurement .

[2] M. Zeaiter, J. M. Roger, V. Bellon-Maurel, D. N. Rutledge, TrAC Trends in Analytical Chemistry 23 (2) (2004) 157 – 170.

[3] Eurachem/citac guide, quantifying uncertainty in analytical measurement .

[4] P. Geladi, Some recent trends in the calibration literature, Chemometrics and Intelligent Laboratory Systems 60 (1-2) (2002) 211 – 224.

[5] T. Isaksson, T. Naes, Effect of multiplicative scatter correction (msc) and

linearity improvement in nir spectroscopy, Applied Spectroscopy 42 (7) (1988) 1273–1284.

[6] W. J. Egan, W. E. Brewer, S. L. Morgan, Measurement of carboxyhemoglobin in forensic blood samples using uv-visible spectrometry and improved principal component regression, Applied Spectroscopy 53 (2) (1999) 218–225.

[7] G. Chryssolouris, M. Lee, A. Ramsey, Confidence interval prediction for neural network models, IEEE Transactions on Neural Networks 7 (1) (1996) 229–232.

[8] A. Höskuldsson, Pls regression methods, Journal of Chemometrics 2 (3) (1988) 211–228.

[9] A. Phatak, P. Reilly, A. Penlidis, An approach to interval estimation in partial least squares regression, Analytica Chimica Acta 277 (2) (1993) 495–501.

[10] M. Denham, Prediction intervals in partial least squares, Journal of Chemometrics 11 (1) (1997) 39–52.

[11] K. Faber, B. Kowalski, Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares, Journal of Chemometrics 11 (3) (1997) 181–238.

[12] N. K. M. Faber, R. Bro, Standard error of prediction for multiway pls: 1. background and a simulation study, Chemometrics and Intelligent Laboratory Systems 61 (1-2) (2002) 133 – 149.

[13] A. Kapteyn, T. Wansbeek, Identification in the linear errors in variables model, Econometrica 51 (6) (1983) pp. 1847–1849.

[14] L. Zhang, S. Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using partial least squares (pls): A practitioner's perspective, Chemometrics and Intelligent Laboratory Systems 97 (2) (2009) 152–158.

[15] M. Zeaiter, J. M. Roger, V. Bellon-Maurel, Dynamic orthogonal projection. a new method to maintain the on-line robustness of multivariate calibrations. application to nir-based monitoring of wine fermentations, Chemometrics and Intelligent Laboratory Systems 80 (2) (2006) 227 – 235.

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Table 1

Estimations of the prediction variance for the sample specific error, according to different pretreatement and different methods of estimation.

			RAW	DTR	SNV	D2	D2SNV	SNVD2
	Pretreatment		RAW	DTR	SNV	D2	D2SNV	SNVD2
	Number of LV		14	13	15	15	12	11
Classical	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^2\sigma_x^2$	763.36	41.31	200.21	27.36	8.64	8.99
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$	0.13	0.11	0.14	0.14	0.12	0.15
	T3	$\frac{\sigma_{lab}^2}{N}$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Sum		763.49	41.43	200.35	27.50	8.77	9.14
New	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{b}$	1.99	1.97	2.39	1.74	2.00	2.64
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$	0.13	0.11	0.14	0.14	0.12	0.15
	T3	$\frac{\sigma_{lab}^2}{N}$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	T4	$\left(1+\frac{1}{N}\right)\mathrm{tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_b)$	0.32	0.24	0.30	0.34	0.31	0.39
	Sum		2.44	2.32	2.84	2.22	2.43	3.18
BS variance			2.42	2.32	2.71	2.42	2.18	3.16
MSEC			1.85	1.84	2.05	1.55	1.57	2.14

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Table 2

Estimations of the prediction variance for the repeatability error, according to different pretreatement and different methods of estimation. The models were calculated on averaged samples.

			Pretreatment	RAW	DTR	SNV	D2	D2SNV	SNVD2
			Number of LV	14	13	15	15	12	11
Classical	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^2\sigma_x^2$		11.04	0.57	1.31	0.65	0.28	0.22
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$		0.14	0.11	0.14	0.12	0.13	0.16
	T3	$\frac{\sigma_{lab}^2}{N}$		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Sum			11.18	0.68	1.45	0.77	0.41	0.38
New	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{b}$		0.21	0.25	0.31	0.21	0.39	0.53
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$		0.14	0.11	0.14	0.12	0.13	0.16
	T3	$\frac{\sigma_{lab}^2}{N}$		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	T4	$\left(1+\frac{1}{N}\right)\mathrm{tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_b)$		0.03	0.03	0.07	0.14	0.06	0.26
	Sum			0.39	0.39	0.53	0.48	0.58	0.95
BS variance				0.43	0.44	0.65	0.84	0.71	1.83

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

Table 3

Estimations of the prediction variance for the repeatability error, according to different pretreatement and different methods of estimation. The models were calculated on individual samples.

	Pretreatment		RAW	DTR	SNV	D2	D2SNV	SNVD2
	Number of LV		14	13	15	15	12	11
Classical	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^2\sigma_x^2$	9.67	0.61	1.21	0.62	0.28	0.22
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$	0.18	0.14	0.18	0.17	0.16	0.18
	T3	$\frac{\sigma_{lab}^2}{N}$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Sum		9.85	0.74	1.38	0.79	0.43	0.41
New	T1	$\left(1+\frac{1}{N}\right)\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{b}$	0.15	0.18	0.19	0.11	0.26	0.22
	T2	$\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_b\mathbf{z}$	0.18	0.14	0.18	0.17	0.16	0.18
	T3	$\frac{\sigma_{lab}^2}{N}$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	T4	$\left(1+\frac{1}{N}\right)\mathrm{tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_b)$	0.05	0.04	0.07	0.15	0.06	0.22
	Sum		0.37	0.36	0.44	0.43	0.48	0.62
BS variance			0.39	0.38	0.46	0.37	0.49	0.54

Author-produced version of the article published in Analytica Chimica Acta, 2012, 721, 28-34.
The original publication is available at http://www.sciencedirect.com/
Doi: 10.1016/j.aca.2012.01.044

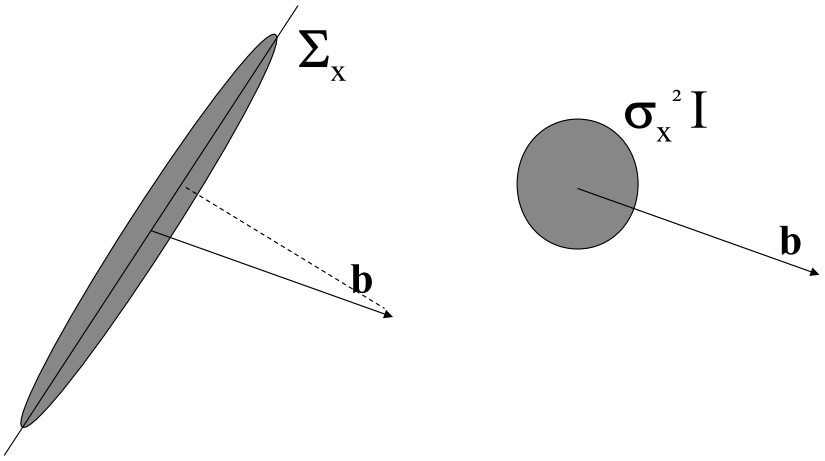

Fig. 1. Schematic example of difference between N1 and C1 hypothesis. On the left, with N1 hypothesis, the structure of the noise is taken into account, and the model can be almost orthogonal to the error space. On the right, with C1 hypothesis, the noise space is assumed to be spherical, then the model cannot be orthogonal to the error space.