Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Chemometrics and Intelligent Laboratory Systems 65 (2003) 281–291
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemometrics

Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error

J.A. Fernández Pierna[a], L. Jin[a,b], F. Wahl[c], N.M. Faber[d], D.L. Massart[a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Institute of Applied Chemistry, Nanchang University, 330047 Nanchang, PR China*
[c] *Institut Français du Pétrole, BP3, 69390 Vernaison, France*
[d] *Department Production and Control Systems, ATO, P.O. Box 17, 6700 AA Wageningen, The Netherlands*

Received 25 June 2002; received in revised form 11 October 2002; accepted 1 November 2002

Abstract

The prediction uncertainty is studied when using a multivariate partial least squares regression (PLSR) model constructed with reference values that contain a sizeable measurement error. Several approximate expressions for calculating a sample-specific standard error of prediction have been proposed in the literature. In addition, Monte Carlo simulation methods such as the bootstrap and the noise addition method can give an estimate of this uncertainty. In this paper, two approximate expressions are compared with the simulation methods for three near-infrared data sets.

Keywords: Multivariate calibration; Partial least squares regression; Uncertainty estimation; Standard error of prediction; Monte Carlo simulation; Bootstrap; Noise addition; Near-infrared spectroscopy

1. Introduction

The primary goal of using a partial least squares regression (PLSR) model in multivariate calibration is to predict the value of a property of interest, the so-called predictand, and its uncertainty [1,2]. The uncertainty of a calculated value is defined as a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand [3]. The analysis of the uncertainty consists in the study of the 'output' uncertainty, i.e. the uncertainty present in the outputs of the model. When the model is constructed, the uncertainty of a predicted value for unknown samples depends on this output uncertainty. In most of the cases, this uncertainty is calculated as a function of the different sources of uncertainty present in the model.

In the univariate context, prediction uncertainty is quantified by a sample-specific standard error of prediction. For the calculation of these univariate quantities, one can rely on standard expressions taken from basic statistics. Unfortunately, multivariate models are inherently much more complex than their univariate analogues. As a result, theoretical advances with respect to the corresponding error analysis are relatively slow. Developing approximate expressions

* Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(02)00139-9

for sample-specific standard error of prediction when applying a PLSR model has received considerable attention in the Chemometrics-related literature [4–23]. We found only few examples dealing with alternative methods such as principal component regression (PCR) [13,24,25] and artificial neural networks (ANNs) [26]. In addition, Monte Carlo simulation methods such as the bootstrap [27] and the noise addition method [28,29] can give an estimate of this uncertainty.

The latest contributions with respect to approximate expressions converge on two proposals; hence, these will constitute the focus of this paper. The first proposal is the correction made by De Vries and Ter Braak [8] on the expression derived by Martens [15] and used in the Unscrambler® software package (CAMO) [30]. The second proposal is the simplification of Faber and Bro [21] of an expression derived earlier under the errors-in-variables (EIV) model [9,13]; its validity was verified using extensive Monte Carlo simulations. Promising results have been reported for practical application of the 'old' EIV expression [14,31]. However, these examples treated models that where constructed using accurately known references values. It is well known that accurately known references values are the exception in typical Chemometrics work such as the prediction of octane number or quality parameters of agricultural products. The purpose of this study is to compare the two proposals (Unscrambler and 'new' EIV) on near-infrared (NIR) data sets for which the reference values carry a sizeable measurement error. Their performance is assessed from the results obtained using bootstrapping and noise addition.

2. Theory

2.1. Notation and conventions

Standard notation is used to denote scalars (uppercase and lowercase italic), vectors (lowercase bold) and matrices (uppercase bold). The symbols $\mathbf{X}$ and $\mathbf{y}$ are used to represent the true predictor matrix and predictand, respectively. The predictors constitute, for example, NIR spectra while the predictand is the property of interest, for example, analyte concentration. The true quantities are unobservable. The measured and therefore observed values for the predictor matrix and predictand vector are denoted by $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$, respectively. A 'hat' added to the symbol of the true value indicates that a quantity is estimated, fitted or predicted.

The term sample is used in two distinct meanings, namely to denote a *chemical* sample or a *statistical* one. A chemical sample is an object for which a property of interest is to be determined, whereas a statistical sample is a random draw from a population. Since the difference is substantial, the meaning should be clear from the context.

2.2. Approximate expressions for sample-specific standard error of prediction

The different steps that one must follow in order to determine the uncertainty are presented in Refs. [32,33]. Initially, the relationship between all the input quantities has to be defined. This step enables one to list the potential sources of uncertainty. The next step is to quantify every uncertainty component associated with the listed sources. Finally, the various contributions to the overall uncertainty have to be combined. Two proposed combinations are studied here, namely the Unscrambler expression [8,15] and the 'new' EIV expression [21]. Since the derivations leading to these expressions are detailed elsewhere, only the main results are given here.

2.2.1. The Unscrambler expression

Implemented in The Unscrambler® software package (CAMO) is an expression proposed by Martens [15] and subsequently improved by De Vries and Ter Braak [8]. This expression considers prediction via the scores:

$$\hat{y} = \bar{\tilde{\mathbf{y}}} + \hat{\mathbf{t}}^T \hat{\mathbf{q}} \tag{1}$$

where $\hat{y}$ is the prediction, $\bar{\tilde{\mathbf{y}}}$ is the average (measured) predictand for the training set (I samples and K variables), $\hat{\mathbf{t}}$ is the estimated score vector of the prediction sample, the superscripted 'T' symbolizes transposition and $\hat{\mathbf{q}}$ is the estimated y-loading vector for the training set (F factors). Each term of this equation has an associated uncertainty. The error on the mean can be estimated using the residual variance in a validation data set. The error in the product of scores and loadings is more difficult to determine

because both scores and loadings contain error and depend on each other. It can be approximated by making two assumptions: the first one is to assume that the scores are without error but the loadings are not, and the second one is to assume that the loadings are without error and the scores are not. This approach leads to two expressions that can be averaged to yield the expression proposed by Martens and modified by De Vries and Ter Braak:

$$\sigma(\mathrm{PE}) \approx \left[\left(1 - \frac{F+1}{I}\right) \left(\underbrace{hR(\mathbf{y}_{\mathrm{val}})}_{A} + \underbrace{\frac{R(\mathbf{x})R(\mathbf{y}_{\mathrm{val}})}{R(\mathbf{X}_{\mathrm{val}})}}_{B} + \underbrace{\frac{2R(\mathbf{y}_{\mathrm{val}})}{I}}_{C} \right) \right]^{1/2} \quad (2)$$

where $\sigma(\cdot)$ is the standard deviation of the associated quantity (square root of the variance), $\mathrm{PE} = \hat{y} - y$ is the prediction error, $R(\cdot)$ is the mean squared residual (MSR) of the associated quantity, $\mathbf{y}_{\mathrm{val}}$ contains the predictands of the validation set, h is the leverage of the prediction sample, which can be seen as the distance of that sample to the mean of the training set data [2], $\mathbf{x}$ contains the predictors of the prediction sample, and $\mathbf{X}_{\mathrm{val}}$ contains the predictors of the validation set. The symbol $R(\cdot)$ is used on the right-hand side of Eq. (2) to emphasize that these quantities are estimated directly from residuals. The term denoted as A corresponds to the part of the equation when scores are considered without error, B corresponds to the part when loadings are considered without error and C represents the error in the mean of Eq. (1). The main characteristic of the Unscrambler expression is that no independent noise estimates are required; all ingredients are estimated directly from the data.

The Unscrambler expression attempts to account for an unexpected interference by the term B, which can be understood as follows. An unexpected interference may lead to a poor model fit of the predictors. A poor model fit is recognized as large x-residuals (in the numerator of B), in comparison with the validation set (denominator of B). Hence, relatively large x-residuals lead to a relatively large contribution of B to the estimated prediction uncertainty. Unfortunately, accounting for an unexpected interference is problematic when having no additional information. Clearly, the part of the unexpected interference that deteriorates prediction should be spanned by the factors that are included in the model. In fact, this part should be parallel to the regression vector, which is a special (one-dimensional) direction within the factor space. In the 'worst' case, the unexpected interference is aligned with the regression vector, leading to normal x-residuals. In the 'best' case, the unexpected interference is spanned by the factors excluded from the model, leading to large x-residuals. In the first case, the Unscrambler expression leads to a normal estimate for prediction uncertainty, which is too optimistic. Conversely, an overly pessimistic estimate is obtained in the last case. It follows that the size of the residuals does not relate to the detrimental effect of an unexpected interference in an obvious way. It is the orientation with respect to the regression vector that counts but this information is typically not available.

It is seen that in its original formulation, Eq. (2) requires a separate validation set to be monitored for calculating residuals. (At the same time, the validation set can be used to optimize the model.) This implies that the available data should be split into three subsets, namely for training, validation and testing. However, to save data for training and testing, we have used cross-validation, i.e. internal validation, instead of setting aside a validation set, see Ref. [34] for more details.

2.2.2. The EIV expression

This approach starts from the so-called EIV regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (3)$$

$$\tilde{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y} \quad (4)$$

$$\tilde{\mathbf{X}} = \mathbf{X} + \Delta\mathbf{X} \quad (5)$$

where $\mathbf{y}$ is the predictand vector, $\mathbf{X}$ is the predictor matrix, $\mathbf{b}$ is the regression vector, $\mathbf{e}$ is the difference between $\mathbf{y}$ and its expectation, and $\Delta\mathbf{y}$ and $\Delta\mathbf{X}$ are the unobservable measurement errors in $\mathbf{y}$ and $\mathbf{X}$, respectively. Eq. (3) describes the relationship between the true predictor and predictand variables, while Eqs. (4) and (5) summarize that the true predictor and predictand variables are unobservable owing to measure-

ment errors. For this model, Faber and Kowalski [13] derived formulas for PLSR and PCR that account for heteroscedastic as well as correlated noise.

A first simplification is obtained by neglecting terms that tend to vanish when the models explains a substantial part of the variance of **X**. Further assuming similar independently and identically distributed (iid) noise for training and prediction data [9] yields

$$\sigma(\mathrm{PE}) \approx \left[\underbrace{(h + 1/I)\Big(V(\mathbf{e}) + V(\Delta\mathbf{y})}_{A} + \|\mathbf{b}\|^2 V(\Delta\mathbf{X})\Big) + \underbrace{V(\mathbf{e}) + \|\mathbf{b}\|^2 V(\Delta\mathbf{X})}_{B} \right]^{1/2} \quad (6)$$

where $\|\cdot\|$ symbolizes the Euclidean vector norm and $V(\cdot)$ represents the variance of the associated quantity. The first term (A) corresponds to the model contribution from the calibration step, while the second term (B) accounts for the unknown sample contribution from the prediction step. The first term (A) depends explicitly on the estimation method, whereas the second term (B) is, in principle, method-independent. When it is evaluated, however, the true values for **b** have to be replaced by their respective estimates so that the practical value of B is method-dependent. Faber and Bro [21] have further simplified Eq. (6) to:

$$\sigma(\mathrm{PE}) \approx [(1 + h + 1/I)\mathrm{MSEC} - V(\Delta\mathbf{y})]^{1/2} \quad (7)$$

where MSEC denotes the mean squared error of calibration estimated as

$$\mathrm{MSEC} = \frac{\sum_{i=1}^{I} (\tilde{y}_i - \hat{y}_i)^2}{I - df} \quad (8)$$

in which $\tilde{y}_i$ is the observed predictand for the i-th training sample, $\hat{y}_i$ is the associated fit and df denotes the degrees of freedom consumed by the model parameters.

The following comments seem to be in order:

(1) The model contribution in Eq. (6) depends on the leverage (h). This makes sense because the leverage determines the distance of the prediction sample from the center of the training set data, where the model is relatively precise.

(2) In Eqs. (6) and (7), there is no provision for prediction error due to an unexpected interference. An unexpected interference may lead to large x-residuals (see discussion following the Unscrambler expression). Hence, these expressions should not be used for unknown samples with abnormally large x-residuals.

(3) Eq. (8) requires inserting an appropriate number of degrees of freedom. For ordinary least squares (OLS), each parameter takes away a degree of freedom from the data, likewise a potential intercept. Similarly, PCR consumes a single degree of freedom for each factor because these factors are entirely determined by the original predictor variables. By contrast, the appropriate number of degrees of freedom for PLSR is not such a trivial matter, since the construction of the factors includes the predictand vector. The rigorous study of Van der Voet [35] has clearly established that the conventional number, i.e. a single degree of freedom for each factor, is not correct. A sound alternative can be calculated using the results of leave-one-out cross-validation, see Eq. (26) in Ref. [35].

(4) Eq. (7) is very similar to the early proposal of Höskuldsson [5]. The difference lies in the subtraction of $V(\Delta\mathbf{y})$, which will always lead to smaller values. Interestingly, Höskuldsson's expression is identical to the one discussed by Næs and Martens for PCR [24]. It has been adopted by the American Society for Testing and Materials (ASTM) [36] and implemented in certain commercial software [37]. Successful generalization to nonlinear variations of PLSR has been recently reported [22].

(5) Inserting a pessimistic estimate for $V(\Delta\mathbf{y})$ may lead to overoptimistic estimates for the standard error of prediction. In absence of a dependable estimate for $V(\Delta\mathbf{y})$, it is safe to assume $V(\Delta\mathbf{y})$ equal to zero and Eq. (7) simply reduces to Höskuldsson's expression. $V(\Delta\mathbf{y})$ is considered as a constant value; if this value is concentration-dependent, i.e. $V(\Delta\mathbf{y})$, Eq. (7) should be changed accordingly by inserting $V(\Delta\mathbf{y}) = f(\hat{y})$ where $\hat{y}$ is the associated prediction.

(6) MSEC is assumed to adequately account for the measurement noise in **X**. The validity of this assumption has been confirmed by Monte Carlo

simulations [21]. For this reason, Eq. (7) requires only a single independent noise estimate, namely the error variance of the reference method, $V(\Delta\mathbf{y})$. Note that the Unscrambler expression requires no noise estimates at all.

(7) MSEC may contain a bias term, see Denham [17] for more details. The reason for this is that the number of factors is selected as a compromise between bias (too few factors) and variance (too many factors). Consequently, Eq. (7) results in a sample-specific root mean squared error of prediction (RMSEP), rather than a standard error of prediction. The importance of prediction bias should, however, not be overrated, since a successful bias-variance trade-off implies bias to be relatively unimportant.

(8) As argued by Faber and Bro [21], Eq. (7) applies to all calibration methods that amount to the two-step procedure of constructing scores, followed by the OLS regression of these scores onto the predictand. Consequently, Eq. (7) should be valid for, among others, PLSR and PCR.

2.3. Bootstrapping residuals

The bootstrap is a computer simulation procedure in which resampling data replaces experimental replication [27]. It can be inferred from the tutorial by Wehrens et al. [38] that bootstrapping samples is the preferred mode in Chemometrics. Here we decided to use bootstrapping residuals because it allows us to directly work on the noise. A recent comparison of resampling methods has shown this method to work better for estimation of the uncertainty in multivariate regression coefficients [39]. Fig. 1 presents a flow chart for the bootstrapping algorithm that is applied in this study. After centering the training data, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$, a PLSR model with F factors is built. Then, the fit residuals $\mathbf{e}$ ($I \times 1$) between $\hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}$ are calculated as:

$$\mathbf{e} = \hat{\mathbf{y}} - \tilde{\mathbf{y}} \tag{9}$$

The next step is to generate B bootstrap samples $\mathbf{e}_1^*$, $\mathbf{e}_2^*, \ldots, \mathbf{e}_B^*$. Each bootstrap sample $\mathbf{e}_b^* = (e_{b1}^*, e_{b2}^*, \ldots, e_{bI}^*)$ is obtained by randomly sampling with replacement I times from the original fit residuals $e_1, e_2, \ldots, e_I$. Sampling with replacement implies that the same residual may be included several times in a particular sample $\mathbf{e}_b^*$. For instance, with $I = 6$ one might obtain

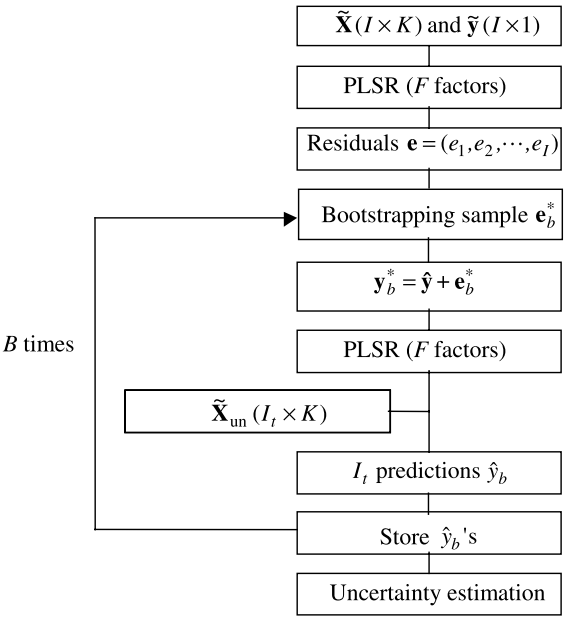

Fig. 1. Flow chart of the bootstrapping method. The symbols are explained in the text.

$\mathbf{e}_1^* = (e_4, e_5, e_1, e_1, e_3, e_4)$, $\mathbf{e}_2^* = (e_2, e_2, e_5, e_1, e_2, e_3)$, etc. Adding a bootstrap sample $\mathbf{e}_b^*$ to the fitted $\hat{\mathbf{y}}$ generates a new predictand vector, $\mathbf{y}_b^*$:

$$\mathbf{y}_b^* = \hat{\mathbf{y}} + \mathbf{e}_b^* \tag{10}$$

Then, a PLSR model is constructed using the centered $\tilde{\mathbf{X}}$ and $\mathbf{y}_b^*$ to obtain the regression coefficients, $\hat{\mathbf{b}}_b$. Finally, the b-th prediction for a single sample is calculated as:

$$\hat{y}_b = \tilde{\mathbf{x}}^{\mathrm{T}} \hat{\mathbf{b}}_b \tag{11}$$

Having repeated the calculation B times, an estimate of sample-specific standard error of prediction is obtained as the standard deviation of the B predictions. The number of repetitions, B, should be large enough in order to ensure the accuracy of the estimation of the uncertainty value for the entire data set.

2.4. Noise addition

The procedure for the noise addition method is very similar to the bootstrapping method (see Fig. 2):

(1) A PLSR model is constructed for the centered training data,

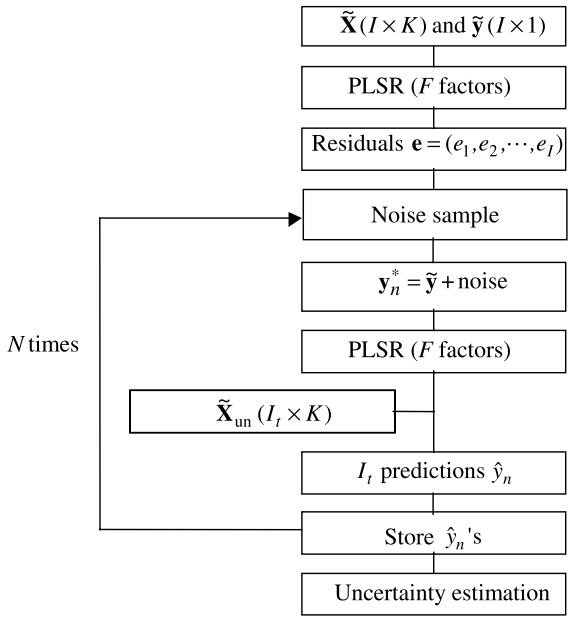

Fig. 2. Flow chart of the noise addition method. The symbols are explained in the text.

(2) the variance of the noise is estimated from the residual vector **e**,
(3) normally distributed noise is added to the predictands of the training set ($\tilde{\mathbf{y}}$) to obtain a new predictand vector ($\mathbf{y}_n^*$),
(4) PLSR is performed using $\mathbf{y}_n^*$ in order to determine the regression coefficients, $\hat{\mathbf{b}}_n$, and
(5) the predictions ($\hat{y}_n$) for new samples are obtained using these coefficients.

Steps 2–5 are repeated N times and the sample-specific standard error of prediction is estimated as the standard deviation of the N predictions.

3. Experimental

3.1. Data sets

Three NIR data sets, which are characterized by a rather imprecise reference method, are used in this study. The first one consists of 99 samples of green tea [40] measured between 1100–2500 nm (each 2 nm). The property of interest is caffeine content determined by RP-HPLC. For this data set the reference method has an estimated measurement standard deviation (reproducibility) $\sigma(\Delta\mathbf{y})=V(\Delta\mathbf{y})^{1/2}=0.15$ g/100 g dry leaves. The data were split into two subsets by using the duplex algorithm [41]. This method starts by selecting the two points furthest from each other and puts them both in a first set (training). Then the next two points furthest from each other are put in a second set (testing), and the procedure is continued by alternatively placing pairs of points in the first or second set. As a result, 57 samples were used for training and 42 samples for testing. The second data set [42] consists of 84 polyether polyol samples measured between 1100–2158 nm (each 2 nm). The property of interest is the hydroxyl number measured on a Pacific Scientific 6250 scanning spectrometer (NIR-System, Silver Spring, MD). The reference method has an estimated measurement standard deviation $\sigma(\Delta\mathbf{y})=V(\Delta\mathbf{y})^{1/2}=0.7$ mg KOH/g. As for the previous data set, the samples were divided using the duplex algorithm into 60 samples for training and 24 samples for testing. The third data set [43] consists of 239 gas oil samples measured between 4900–9000 cm^{-1} (each 2 cm^{-1}). The property of interest is the percent of hydrogen determined by RMN. For this data set, the estimated measurement standard deviation of the reference method is $\sigma(\Delta\mathbf{y})=V(\Delta\mathbf{y})^{1/2}=0.025$ g/100 g. With the duplex method the data set was split into 84 samples for training and 155 samples for testing.

3.2. Software

All computations were executed using Matlab 5.2.0 (1998) from the Mathworks, as support. Bootstrapping and noise addition use 1000 repetitions each.

4. Results and discussion

Various data pretreatments are considered, i.e. first and second derivatives and the standard normal variate (SNV) transformation. The optimum complexity of the model is determined by two methods, namely leave-one-out cross-validation to calculate the minimum root mean squared error of cross-validation (RMSECV) and the randomization test proposed by

Van der Voet [44]. Faber and Bro [21] have claimed Eq. (7) to be valid for PCR; this conjecture is tested here too. For considerations of space, only a selection of the results is presented.

4.1. Green tea data

Recall that the 'old' EIV expression [9,13] has already given promising results for NIR data sets for which the reference values are relatively precise [14,31]. Thus, the purpose of this study was to compare the Unscrambler and the 'new' or simplified EIV expression on NIR data sets for which the reference values are not very precise. The relative importance of this source of uncertainty is determined as follows. The variance of the reference method is $V(\Delta\mathbf{y}) = 0.15^2$ g^2/(100 g dry leaves)^2. Adding excessive noise to the spectra, as suggested in [14], has little effect on the predictive ability of the resulting PLSR and PCR models, measured as RMSEP. This implies that the effect of the original spectral noise is negligible. With negligible spectral noise, MSEC = $V(\mathbf{e}) + V(\Delta\mathbf{y})$, see, e.g. Ref. [9]. In this way, one obtains the value $V(\mathbf{e}) = 0.12^2$ g^2/(100 g dry leaves)^2 for the first model in Table 1, which shows that the uncertainty of the reference method is sizable indeed ($V(\Delta\mathbf{y}) > V(\mathbf{e})$).

The best models are obtained without pretreating the spectra or after applying the SNV transformation. For all models summarized in Table 1, factor selection is based on the randomization test. The values for root mean squared error of calibration (RMSEC) and RMSECV are very similar. The reason for this similarity is that RMSEC is calculated using Van der Voet's degrees of freedom [35], which are directly related to cross-validation, see third comment following Eq. (8). The RMSEP values exceed the RMSECV values in all cases, which is in agreement with the often-reported observation that cross-validation is slightly optimistic. Bootstrapping and noise addition lead to almost identical estimates for sample-specific standard error of prediction. For example, for the first model in Table 1, an average ratio of 0.97 is obtained between bootstrapping and noise addition results (correlation coefficient is 0.99). This excellent agreement is expected because of the close relationship between them: they both directly work with the noise in the data. In the remainder of this paper, we will restrict ourselves to the noise addition method to avoid duplication of material. For the first model in Table 1, the Unscrambler results often differ considerably from the noise addition results: the deviations range between underestimation and overestimation by a factor of two, see Fig. 3. In contrast, the simplified EIV expression seems to work very well for this data set. The largest difference between the EIV and noise addition results is observed for test sample 14: 0.16 g/100 g dry leaves for noise addition versus 0.13 g/100 g dry leaves for the EIV expression. Considering that an uncertainty estimate is reported in at most two decimal digits leads to the impression that the discrepancies are of little practical significance. The PCR and PLSR results for the EIV approach are equally satisfactory. This observation corroborates the conjecture of Faber and Bro [21] that Eq. (7) should be valid for both methods. It is important to note that the performance of the EIV approach does not depend on the cur-

Table 1
Results for the green tea data

Method	Pretreatment	Factors[a]	RMSEC (g/100 g dry leaves)	RMSECV (g/100 g dry leaves)	RMSEP (g/100 g dry leaves)	Unscrambler expression[b]		EIV expression[b]	
						Ratio	Correlation	Ratio	Correlation
PLSR	No	4	0.19	0.20	0.28	0.89	0.81	0.98	0.81
	SNV	5	0.16	0.17	0.23	1.61	0.78	0.93	0.79
	SNV	4	0.18	0.19	0.23	1.05	0.45	0.98	0.82
PCR	No	5	0.19	0.20	0.27	–	–	1.05	0.57
	SNV	5	0.16	0.17	0.25	–	–	0.95	0.81

[a] Determined using the randomization test.
[b] Compared with the result of 1000 noise additions.

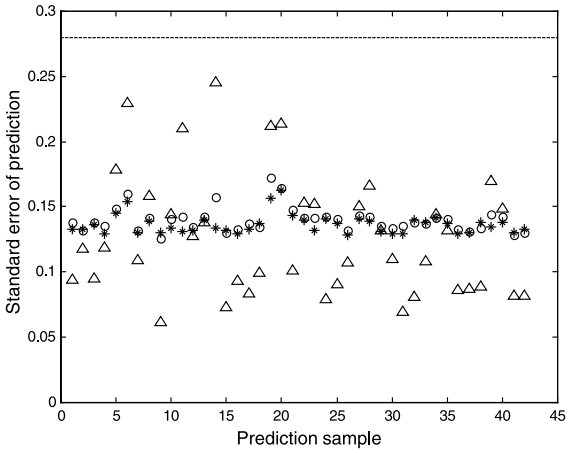

Fig. 3. Estimated sample-specific standard error of prediction (g/100 g dry leaves) for the four-dimensional PLSR model of the green tea data (no data pretreatment): noise addition (○), EIV (*) and Unscrambler (△). The apparent RMSEP (---) is added as guide to the eye.

rently tested spectral pretreatment methods. If, however, the pretreatment amounts to removing a subspace from the predictor matrix, then the leverage in Eq. (7) should be modified as explained by Olivieri [23].

With imprecisely measured reference values, predictive measures such as RMSEP are biased high because this measurement error is confounded with the true prediction error. This has been very well explained by DiFoggio [45] who coined the terms apparent and actual RMSEP. The apparent RMSEP is calculated using imprecise reference values whereas the actual RMSEP is based on the true values. DiFoggio proposed the following correction:

$$\text{corrected RMSEP} = [\text{apparent MSEP} - V(\Delta\mathbf{y})]^{1/2} \quad (12)$$

where MSEP stands for mean squared error of prediction. Eq. (12) has been successfully applied by Sørensen [46] to a number of NIR applications. Comparing Eqs. (12) and (7) reveals that the term $(1+h+1/I)$ MSEC has the interpretation of an apparent sample-specific MSEP. In other words, Eq. (7) performs the correction for each individual sample, while Eq. (12) operates on the set level. Consequently, sample-specific standard errors of prediction obtained using Eq. (7) tend to be smaller than the apparent RMSEP. This holds for the entire test set, see Fig. 3. Finally, the preceding discussion implies that the apparent RMSEP is consistent with Höskuldsson's early proposal [5], see fourth comment following Eq. (8).

4.2. Polyether polyol data

Using the procedure explained in the previous section, one obtains the variance estimate $V(\mathbf{e}) = 1.3^2$ (mg KOH)^2/g^2 for the first model in Table 2. Comparing this value with $V(\Delta\mathbf{y}) = 0.7^2$ (mg KOH)^2/g^2 shows that the relative importance of the measurement noise in the reference values is smaller than for the previous case ($V(\Delta\mathbf{y}) < V(\mathbf{e})$). Spectral pretreatment does not improve the predictive ability. The minimum RMSECV leads to different model complexities

Table 2
Results for the polyether polyol data

Method	Pretreatment	Factors	RMSEC (mg KOH/g)	RMSECV (mg KOH/g)	RMSEP (mg KOH/g)	Unscrambler expression[a]		EIV expression[a]	
						Ratio	Correlation	Ratio	Correlation
PLSR	No	7[b]	1.44	1.76	1.75	0.46	0.77	0.98	0.81
	No	5[c]	2.23	2.55	2.47	0.32	0.94	0.98	0.79
PCR	No	7[b]	2.01	1.76	2.77	–	–	1.03	0.57
	No	5[c]	2.88	2.55	4.85	–	–	0.97	0.81

[a] Compared with the result of 1000 noise additions.
[b] Determined using the minimum RMSECV.
[c] Determined using the randomization test.

than the randomization test. Bootstrap and noise addition yield very similar results: for the first model in Table 2, the average ratio is 0.99 (correlation coefficient is 0.99). For this model, the Unscrambler results are persistently too small, sometimes by more than a factor of three (Fig. 4). A factor of three is even the average underestimation for the second model in Table 2. By contrast, the EIV results closely follow the noise addition results. The largest difference is observed for test sample 4: when using noise addition, one would report an error value of 2.3 instead of 2.2 mg KOH/g when relying on the EIV expression. Finally, it is observed that the EIV expression maintains its excellent performance for models with relatively high RMSEP. Stated differently, the validity of this approach is not limited to the 'best' model.

One might argue that the poor results obtained using the Unscrambler expression are caused by inserting severely optimistic results from cross-validation in Eq. (2), instead of using external validation as suggested in the original work. However, it can be inferred from the ratio between RMSECV and RMSEP (larger than unity!) that the often-observed optimism is absent for the first two models in Table 2.

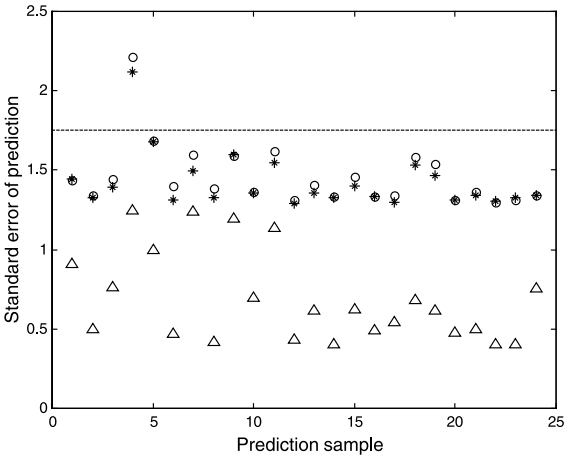

Fig. 4. Estimated sample-specific standard error of prediction (mg KOH/g) for the seven-dimensional PLSR model of the polyether polyol data (no data pretreatment): noise addition (○), EIV (*) and Unscrambler (△). The apparent RMSEP (---) is added as guide to the eye.

4.3. Gas oil data

Results are discussed only for the PLSR and PCR models for which the number of factors is determined using the randomization test (no spectral pretreatment). The optimum selected PLSR model is based on five factors, while PCR requires six factors. For these models, one obtains the variance estimates $V(\mathbf{e}) = 0.052^2$ g^2/(100 g)^2 and $V(\mathbf{e}) = 0.058^2$ g^2/(100 g)^2, respectively. As for the previous data sets, the EIV results are in excellent agreement with the noise addition results (ratio is 0.99 and 1.17 for PLSR and PCR models respectively), whereas the Unscrambler results are unsatisfactory (not shown).

In comparison with the previous data sets, the current one has a large number of test samples (155). Such a large test set enables one to further scrutinize the adequacy of the approximate expressions (2) and (7) by monitoring studentized prediction residuals,

$$z = \frac{\hat{y} - \tilde{y}}{\sigma(\hat{y} - \tilde{y})} \tag{13}$$

with obvious notation. The reference values for the test set contain a measurement error with variance $V(\mathbf{\Delta y})$; hence, the denominator of Eq. (13) follows as

$$\sigma(\hat{y} - \tilde{y}) = \sqrt{\sigma(\mathrm{PE})^2 + V(\mathbf{\Delta y})} \tag{14}$$

where σ(PE) is the result of Eq. (2) or Eq. (7). Eq. (14) has the simple interpretation that it is easier to predict an error-free reference value than one that contains a measurement error. The distributions of the studentized prediction residuals are presented in Fig. 5. The number of training samples is large (84); hence, the degrees of freedom are large and, ideally, the distribution of the studentized residuals should approach the normal distribution with standard deviation unity. The Unscrambler expression yields a standard deviation that is too large (1.41), which must be caused by underestimating the standard error of prediction—the denominator of Eq. (13). In addition, the distribution is heavily skewed, which violates the requirement that it can be characterized using a single width parameter—the standard error of prediction. By contrast, the standard deviation of the studentized residuals is too small when the EIV

expression is used (0.757). This means that the standard error of prediction is overestimated. The reason for this becomes clear when investigating the relevant summary statistics RMSEC, RMSECV and RMSEP (Fig. 6). It turns out that the average fit error (RMSEC), which is a major ingredient of Eq. (7), is larger than the average prediction error (RMSEP) by 19%. This is quite unusual, but can be explained from the particular splitting of samples in a training and test set: on average, the training samples are further away from the model center than the test samples. Splitting the samples in a random fashion, rather than using the duplex algorithm, leads to an RMSEP that is slightly larger than RMSEC and a standard deviation of the studentized residuals of 0.994, which is excellent. The distribution remains symmetric around zero.

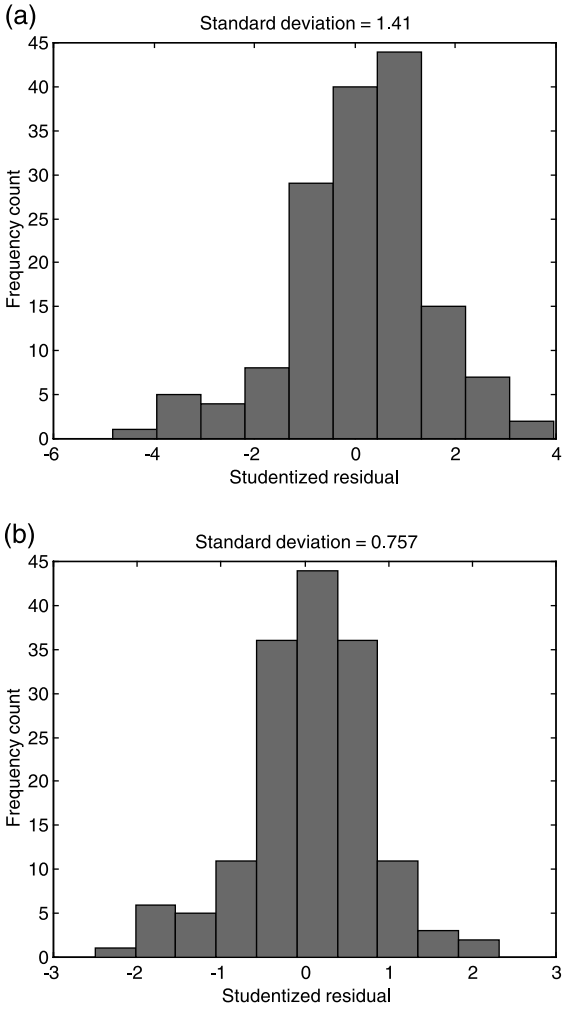

Fig. 5. Studentized residuals for gas oil data set: (a) Unscrambler expression and (b) EIV expression.

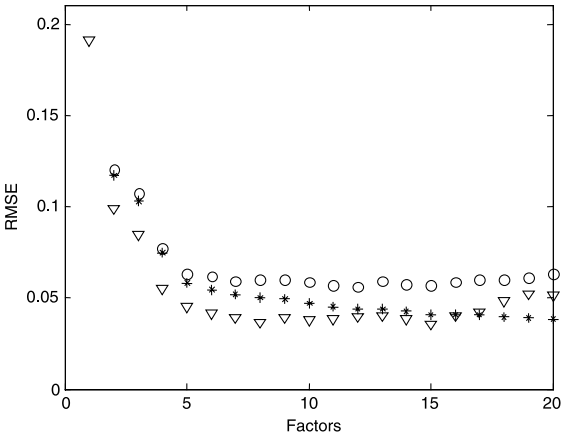

Fig. 6. Root mean squared errors for gas oil data set: RMSEC (*), RMSECV (○) and RMSEP (▽).

5. Conclusions

The Unscrambler expression [8,15] and the 'new' EIV expression [21] have been compared with Monte Carlo simulation-based methods on NIR data sets with imprecise reference values. A major plus of the Unscrambler expression is that no independent noise estimates are required; all values are obtained directly from the data. Unfortunately, the extreme user-friendliness of the Unscrambler expression has limited value owing to the relatively poor results. In contrast, promising results are obtained using the 'new' EIV expression. An important aspect of the resulting sample-specific uncertainty estimates is that they can be significantly smaller than the apparent RMSEP when the reference values are relatively imprecise. This can be seen as an added bonus, which, however, is lost if the reference error variance is unknown. In that case, the EIV expression reduces to Höskuldsson's early proposal [5], which has been adopted by the ASTM [36].

Acknowledgements

Frank Schreutelkamp is thanked for bringing the commercial implementation of the ASTM guideline to the authors' attention.

References

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.
[2] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[3] ISO Guide to the Expression of Uncertainty in Measurement (GUM) and the International Vocabulary of Basic and General Terms in Metrology (VIM).
[4] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 93.
[5] A. Höskuldsson, J. Chemom. 2 (1988) 211.
[6] T.V. Karstang, J. Toft, O.M. Kvalheim, J. Chemom. 6 (1992) 177.
[7] A. Phatak, P.M. Reilly, A. Penlidis, Anal. Chim. Acta 277 (1993) 495.
[8] S. De Vries, C.J.F. Ter Braak, Chemom. Intell. Lab. Syst. 30 (1995) 239.
[9] K. Faber, B.R. Kowalski, Chemom. Intell. Lab. Syst. 34 (1996) 283.
[10] R.E. Kleinknecht, J. Chemom. 10 (1996) 687.
[11] A.J. Berger, M.S. Feld, Appl. Spectrosc. 51 (1997) 725.
[12] M.C. Denham, J. Chemom. 11 (1997) 39.
[13] K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 181.
[14] N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler, Anal. Chem. 70 (1998) 2972.
[15] M. Høy, K. Steen, H. Martens, Chemom. Intell. Lab. Syst. 44 (1998) 123.
[16] T. Morsing, C. Ekman, J. Chemom. 12 (1998) 295.
[17] M.C. Denham, J. Chemom. 14 (2000) 351.
[18] N.M. Faber, J. Chemom. 14 (2000) 363.
[19] N.M. Faber, Chemom. Intell. Lab. Syst. 52 (2000) 123.
[20] X.-H. Song, N.M. Faber, P.K. Hopke, D.T. Suess, K.A. Prather, J.J. Schauer, G.R. Cass, Anal. Chim. Acta 446 (2001) 329.
[21] N.M. Faber, R. Bro, Chemom. Intell. Lab. Syst. 61 (2002) 133.
[22] G. Baffi, E. Martin, J. Morris, Chemom. Intell. Lab. Syst. 61 (2002) 151.
[23] A.C. Olivieri, J. Chemom. 16 (2002) 207.
[24] T. Næs, H. Martens, J. Chemom. 2 (1988) 155.
[25] W.J. Egan, W.E. Brewer, S.L. Morgan, Appl. Spectrosc. 53 (1999) 218.
[26] G. Chryssolouris, M. Lee, A. Ramsey, IEEE Trans. Neural Netw. 7 (1996) 229.
[27] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
[28] W.H. Press, B.P. Flannery, S.A. Teukolski, W.T. Vetterling, Numerical recipes, The Art of Scientific Computing, Cambridge Univ. Press, Cambridge, 1988, §14.5.
[29] R.J. Carrol, D. Ruppert, L.A. Stefanski, Measurement Error in Nonlinear Models, Chapman & Hall, London, 1995, Chapter 4.
[30] Available at: http://www.camo.com/Software/Overview/unscrambler.htm.
[31] R. Boqué, M.S. Larrechi, F.X. Rius, Chemom. Intell. Lab. Syst. 45 (1999) 397.
[32] L. Cuadros, M.E. Hernandez, E. Almansa, F.J. Egea, F.J. Arrebola, J.L. Martinez, Anal. Chim. Acta 454 (2002) 297.
[33] EURACHEM/CITAC Guide: Quantifying Uncertainty in Analytical Measurement, 2nd ed., EMPA, St. Gallen, 2000.
[34] J.A. Fernández Pierna, F. Wahl, O.E. de Noord, D.L. Massart, Chemom. Intell. Lab. Syst. 63 (2002) 27.
[35] H. Van der Voet, J. Chemom. 13 (1999) 195.
[36] Annual Book of ASTM Standards, vol 03.06, E1655, Standard Practices for Infrared, Multivariate, Quantitative Analysis, ASTM International, West Conshohocken, Pennsylvania, USA, 1998.
[37] Quant + software, Version 4.51.02, PerkinElmer, Wellesly, Maryland, USA.
[38] R. Wehrens, H. Putter, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 54 (2000) 35.
[39] N.M. Faber, Chemom. Intell. Lab. Syst. (accepted).
[40] J. Luypaert, M. Zhang, D. L. Massart, Anal. Chim. Acta (accepted).
[41] R.D. Snee, Technometrics 19 (1977) 415.
[42] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 388 (1999) 283.
[43] F. Wahl, personal communication.
[44] (a) H. Van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313;
(b) H. Van der Voet, Chemom. Intell. Lab. Syst. 28 (1995) 315.
[45] R. DiFoggio, Appl. Spectrosc. 49 (1995) 67.
[46] L.K. Sørensen, J. Near Infrared Spectrosc. 10 (2002) 15.