ELSEVIER

Chemometrics and Intelligent Laboratory Systems 61 (2002) 133–149

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Standard error of prediction for multiway PLS 1. Background and a simulation study

Nicolaas (Klaas) M. Faber [a,*], Rasmus Bro [b]

[a] *Department Production and Control Systems, ATO, P.O. Box 17, 6700 AA Wageningen, The Netherlands*
[b] *Food Technology, Royal Veterinary and Agricultural University, Rolighedsvej 30, 1958 Frederiksberg C, Denmark*

Received 27 February 2001; received in revised form 25 September 2001; accepted 18 October 2001

Abstract

While a multitude of expressions has been proposed for calculating sample-specific standard errors of prediction when using partial least squares (PLS) regression for the calibration of first-order data, potential generalisations to multiway data are lacking to date. We have examined the adequacy of two approximate expressions when using unfold- or tri-PLS for the calibration of second-order data. The first expression is derived under the assumption that the errors in the predictor variables are homoscedastic, i.e., of constant variance. In contrast, the second expression is designed to also work in the heteroscedastic case. The adequacy of the approximations is tested using extensive Monte Carlo simulations while the practical utility is demonstrated in Part 2 of this series. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Multiway calibration; Unfold-PLS; Multilinear PLS; Standard error of prediction

1. Introduction

Applications of multiway data analysis are rapidly increasing. This development has prompted researchers to develop methods for exploiting the full information content of these intrinsically rich data structures. In a seminal paper, Wold et al. [1] generalised standard (i.e., linear) partial least squares (PLS) to the calibration of multiway data. Their method is known as unfold-PLS, because it amounts to unfolding or matricising the $I \times J_1 \times J_2 \times \ldots \times J_N$ stack of predictor arrays, $\underline{\mathbf{X}}$, to obtain an $I \times J_1 J_2 \ldots J_N$ matrix $\mathbf{X}$. This matrix is subsequently modelled using standard PLS in terms of $I \times 1$ score vectors $\mathbf{t}_a$ ($a = 1,\ldots,A$) and $J_1 J_2 \ldots J_N \times 1$ weight vectors $\mathbf{w}_a$ ($a = 1,\ldots,A$). Bro [2] introduced multilinear PLS as a powerful alternative to unfold-PLS. Multilinear PLS derives its name from the fact that it models $\underline{\mathbf{X}}$ in terms of $I \times 1$ score vectors $\mathbf{t}_a$ ($a = 1,\ldots,A$) and $J_n \times 1$ ($n = 1,\ldots,N$) weight vectors $\mathbf{w}_a^n$ ($a = 1,\ldots,A$). Stated differently, multilinear PLS is a genuine Nth-order approach, because weights are calculated for each individual mode ($n = 1,\ldots,N$), whereas unfold-PLS amounts to a pseudo first-order one. Depending on the particular number of modes of $\underline{\mathbf{X}}$, multilinear PLS is called tri-PLS ($N+1=3$), quadri-PLS ($N+1=4$), etc. For a variety of applications, an improvement over the unfold-PLS results could be reported [2–4], while Smilde [5] and De Jong [6] detailed theoretical advances with respect to multilinear PLS.

From a practical point of view, it is desirable to have a quantitative measure for the uncertainty in

* Corresponding author.
E-mail address: n.m.faber@ato.wag-ur.nl (N.M. Faber).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(01)00204-0

predictions obtained when applying a multiway calibration model. The subject of sample-specific standard errors of prediction has attracted considerable attention in the first-order case and a multitude of expressions has been proposed for standard PLS [7–20]. In contrast, potential generalisations to multiway data have not been reported to the best of our knowledge. Here it is proposed to build on an expression that has been derived for standard PLS [12]. This particular expression, which results from applying an additional approximation to the so-called local linearisation of the PLS regression vector, is attractive in the sense that it accounts for all sources of (random) error variance in the predictor and predictand variables. Its adequacy has been thoroughly tested using Monte Carlo simulations [19]. It is reasonable to expect this expression to work to the same extent when using unfold-PLS, because the intrinsic structure of the multiway array is lost owing to the unfolding operation. A drawback of this approach is that the original expression is derived under the assumption that the errors in the predictor variables have constant variance (the homoscedastic case). Its application is further restricted because estimates are required of the error variances for both predictor and predictand variables. To overcome part of these limitations, a second expression is derived. This expression is designed to also work in the heteroscedastic case and its application only requires an estimate of the variance of the measurement error in the predictand. We conjecture that both expressions are valid using multilinear PLS.

The adequacy of the proposed expressions is verified using simulated $\underline{\mathbf{X}}$'s that follow the trilinear model (i.e., PARAFAC structure) while in Part 2 of this series [21] we will test the practical utility on experimental excitation emission matrix (EEM) data. The reason for focussing on the calibration of second-order predictor arrays is that data of higher order still form the exception. The particular choice for trilinear $\underline{\mathbf{X}}$ is motivated by the fact that many instruments generate data that approximately follow this model (see Ref. [22] and Table 2 in Ref. [23]). Moreover, making this assumption about the true structure allows one to discuss the PLS results with respect to the calibration framework recently developed by Linder and Sundberg [24,25]. Characteristic for this framework is that it explicitly takes account of the bilinear structure of the individual predictor arrays. Finally, the trilinear assumption enables one to interpret the quality of the prediction results in terms of analytical figures of merit [26–29]. The expressions underlying these figures of merit reflect the trilinear structure of the data too.

2. Theory

2.1. Preliminaries

Calibration can be performed in two modes, namely according to the *classical* or the *inverse* model formulation, respectively. In applied work, the inverse model is often preferred over the classical model, because it is more flexible. Thus, in the remainder of this paper focus will be on the inverse calibration model. For convenience, and also because many applications follow this description, it is assumed that the goal of calibration is to replace a reference method by predicting the analyte concentration in an unknown (chemical) sample from instrument responses. Each analyte is modelled separately. There are two reasons for considering this scenario. First, modelling each analyte separately (i.e., PLS1) is often favoured over constructing a single model for all analytes (i.e., PLS2). Second, the proposed expressions for sample-specific standard error of prediction are based on an expression that has been derived specifically for PLS1 [12]. In the remainder of this paper, the acronym PLS stands for PLS1. Finally, because focus is entirely on second-order arrays, the general notation for the numbers of variables, i.e., J_n ($n=1,\ldots,2$), is dropped in favour of J and K. The two corresponding modes are conveniently denoted as J-mode and K-mode, respectively.

2.2. Model, prediction and prediction error

The inverse multiway calibration model for the unknown sample is written as:

$$y = \mathbf{x}^{\mathrm{T}}\beta + e \tag{1}$$

where y is the true analyte concentration, $\mathbf{x}$ and β are the unfolded $JK \times 1$ arrays with true instrument

responses and regression coefficients, e is the residual, and the superscripted "T" denotes transposition. It is stressed that the model equation relates the true quantities rather than the measured, estimated or predicted ones.

It is important to note that the term residual is often used to denote the difference between the measured and fitted or predicted data. The current terminology is borrowed from the errors-in-variables literature (see Van Huffel and Vandewalle [30] for an excellent discussion of these models). Eq. (1) constitutes a so-called general regression model with errors in the variables, when $\mathbf{x}$ and y are observed with measurement errors. This model corresponds to a non-zero residual problem, because the true values for $\mathbf{x}$ and β do not reproduce the true y. The residual e could, for example, summarise a deviation from Beer–Lambert's law owing to turbidity of the sample solution; it can be treated as a zero-mean random variable if a suitable background correction has been applied. (Beer–Lambert's law states that an exact relationship holds between the true concentration and absorbance.) The residual is often confused with a measurement error. A major difference between the two is that the residual can only be made smaller by including informative x-variables, whereas the measurement error is only reduced by improving the measurement itself (see Van Huffel and Vandewalle [30], p. 229). The residual is also different from a model error. An example of the latter arises when fitting a straight line where a parabola would better describe the data. Model errors can only be countered by changing to a more appropriate model. Eq. (1) is general in the sense that it covers numbers (univariate calibration) and vectors (multivariate calibration) as special cases.

A prediction for the unknown sample is obtained as:

$$\hat{y} = \tilde{\mathbf{x}}^{\mathrm{T}}\hat{\beta} \tag{2}$$

where the "hat" (^) signifies prediction of a random variable (e.g., y) or estimation of a parameter (e.g., β), and the "tilde" (~) indicates that the associated quantity is measured (here the instrument responses for the unknown sample).

The prediction error (PE) is defined as:

$$\mathrm{PE} \equiv \hat{y} - y. \tag{3}$$

In practice, the prediction error is unknowable since y is unknown. However, using certain assumptions, one may derive an expression for determining its expected size. The relevant statistics are mean squared error of prediction (MSEP), prediction error variance and prediction bias. MSEP is defined as:

$$\begin{aligned}\mathrm{MSEP} &\equiv E[\mathrm{PE}^2] \\ &= E[(\hat{y} - E[\hat{y}])^2] + (E[\hat{y}] - y)^2 \\ &= \text{variance} + (\text{bias})^2 \end{aligned} \tag{4}$$

where $E[\cdot]$ denotes expected value.

MSEP contains a variance and a bias contribution. In the next sections we will focus on deriving approximate expressions for the standard error of prediction, which is defined as the square root of the prediction error variance. The handling of prediction bias is integrated in these derivations.

2.3. General expression for standard error of prediction

A general expression for standard error of prediction is derived as follows. The PE is a non-linear function of the input data, which are assumed to be unbiased (all errors are considered to be zero mean random variables). Every differentiable function can be approximated using a Taylor series expansion. Truncating this expansion after the linear term is known as local linearisation (in statistics) or error propagation (in chemistry). Such a first-order approximation is useful, because it is relatively easy to derive a variance expression for linear functions. This approach does not, however, account for bias. The subject of prediction bias is taken up in Section 2.4.1.

Using local linearisation and neglecting the products of error terms, the prediction error is approximated as:

$$\begin{aligned}\mathrm{PE} &\approx (\mathbf{x} + \Delta\mathbf{x})^{\mathrm{T}}(\beta + \Delta\beta) - \mathbf{x}^{\mathrm{T}}\beta - e \\ &\approx \mathbf{x}^{\mathrm{T}}\Delta\beta + (\Delta\mathbf{x}^{\mathrm{T}})\beta - e \end{aligned} \tag{5}$$

where the prefix "Δ" signifies the error in the associated quantity. Taking expectation of the squared linear approximation of PE yields the approximate prediction error variance as:

$$V_{\mathrm{PE}} \approx \mathbf{x}^{\mathrm{T}}\mathbf{V}_{\Delta\beta}\mathbf{x} + \beta^{\mathrm{T}}\mathbf{V}_{\Delta x}\beta + V_e \tag{6}$$

where $\mathbf{V}_{\Delta\beta}=E[\Delta\beta\Delta\beta^{\mathrm{T}}]$ is the covariance matrix for the regression coefficient estimates, $\mathbf{V}_{\Delta x}=E[\Delta\mathbf{x}\Delta\mathbf{x}^{\mathrm{T}}]$ is the covariance matrix of the measurement errors in the instrument responses, and $V_e=E[e^2]$ is the variance of the residual. The general expression for standard error of prediction follows as:

$$\sigma_{\mathrm{PE}} \equiv (V_{\mathrm{PE}})^{1/2} \approx (\mathbf{x}^{\mathrm{T}}\mathbf{V}_{\Delta\beta}\mathbf{x} + \beta^{T}\mathbf{V}_{\Delta x}\beta + V_e)^{1/2}. \tag{7}$$

It is seen that the (approximate) standard error of prediction has two distinct contributions, namely the model contribution from the calibration step, i.e., $\mathbf{x}^{\mathrm{T}}\boldsymbol{V}_{\Delta\beta}\mathbf{x}$, and the unknown sample contribution from the prediction step, i.e., $\beta^{\mathrm{T}}\mathbf{V}_{\Delta x}\beta+V_e$. The first term depends explicitly on the estimation method, whereas the second term is, in principle, method-independent. When it is evaluated, however, the true values for β have to be replaced by their respective estimates so that the *practical* value is method-dependent.

Because local linearisation amounts to a first-order approximation of an uncertainty, it yields promising results only if this uncertainty is small. In other words, the parameter estimates and the measurements should not be too noisy, otherwise the method breaks down and higher-order approximations must be considered. The latter is prohibitive for complicated estimation methods such as PLS. The level of bias, e.g., of the regression coefficients, on the other hand, is not essential for the quality of the linearisation because it is based on variances in which the bias does not appear.

2.4. Specific expressions for standard error of prediction

Eq. (7) is general in the sense that no assumption has been made about the estimation method (or number of analytes modelled simultaneously). We now proceed by finding suitable expressions for $\mathbf{x}^{\mathrm{T}}\mathbf{V}_{\Delta\beta}\mathbf{x}$.

2.4.1. Homoscedastic errors in the predictors

For pseudo first-order calibration using unfold-PLS one has several expressions at one's disposal that have been proposed for standard PLS [7–20]. Of special distinction is the work of Phatak et al. [10] and Denham [13] who performed local linearisation of the PLS model to obtain expressions for $\mathbf{V}_{\Delta\beta}$ in the case of negligible errors in the predictor variables. This work has been extended by Faber and Kowalski [14] to include errors in the predictor variables and to principal component regression (PCR). Generally, these expressions, which accommodate for heteroscedastic and correlated errors, are difficult to interpret. Simply ignoring the most complicated terms and assuming that all errors are independently and identically distributed (iid) yields the first specific expression for standard error of prediction as:

$$\sigma_{\mathrm{PE}} \approx [h(\|\beta\|^2 V_{\Delta x} + V_e + V_{\Delta y}) + \|\beta\|^2 V_{\Delta x} + V_e]^{1/2} \tag{8}$$

where the scalar h is the unknown sample leverage with respect to the origin, $\|\cdot\|$ denotes the Euclidean norm, $V_{\Delta y}=E[(\Delta y)^2]$ is the variance of the measurement error in the reference method (uncertainty in y-values for the calibration set), and $V_{\Delta x}$ is the iid simplification of $\mathbf{V}_{\Delta x}$. For zero-intercept models, the leverage is calculated as $h=\mathbf{t}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{t}$, where $\mathbf{t}$ is the $A\times 1$ unknown sample score vector, likewise the rows of $\mathbf{T}$ $(I\times A)$ for the calibration set. Mean centring is accounted for by adding $1/I$. The different terms in Eq. (8) correspond to similar parts in Eq. (7).

In the case of negligible predictor noise, ignoring the most complicated terms has been labelled the "naive" approach by Denham [13]. To the best of our knowledge, it was first proposed by Höskuldsson [8]. A simple interpretation of the "naive" approach is as follows. PLS regression can be thought of in two stages. The first stage constructs a set of scores as a linear transformation of the predictor variables. The second stage relates the predictand to the scores using ordinary least squares (OLS). The "naive" approach essentially ignores the uncertainty in the scores. In other words, the resulting expression for standard error of prediction is identical to the well-known OLS formula applied to the scores (see Eq. (9) in Ref. [13]). The reasoning extends in straightforward fashion to the case of non-negligible predictor noise [14]. It is important to note that OLS assumes the predictor matrix to be of full column rank. However, for rank-deficient predictor matrices all results carry through for the minimum length least squares (MLLS) solution.

It is reiterated that Eq. (8) is obtained by ignoring the uncertainty in the scores. This implies that it can be used in connection with any method that amounts to regression onto a subspace generated by scores, unfold-PCR being another example of the family. (The only difference arises in the numerical values inserted for h and β.) Covering a broad range of methods can be seen as an intrinsic advantage of the "naive" approach. Formally, it amounts to a zeroth-order approximation, because first-order terms are discarded. It is important to note that these terms tend to vanish if the model approaches the OLS or MLLS limit [14]. Practically, this situation is indicated when the model explains most of the systematic variance of **X**, which is often the case when the predictor variables constitute spectroscopic measurements. Thus, the practical significance of the additional approximation may be quite small in many applications. To distinguish the expressions derived in Refs. [10,13,14] from Eq. (8), they are referred to as the "full linearisation" results in the remainder of this paper.

The model contribution from the calibration step, i.e., $h(\|\beta\|^2 V_{\Delta x} + V_e + V_{\Delta y})$, has a very simple interpretation. The leverage quantifies the distance of the unknown sample to the calibration samples in A-dimensional space (A is the number of PLS factors). A small leverage corresponds to positions sufficiently close to the average calibration sample. Owing to this closeness, the model is precisely estimated at these positions and, consequently, the model contribution to Eq. (8) is small. The converse holds for high-leverage points. Extensive Monte Carlo simulations have shown that Eq. (8) performs better than expressions implemented in certain commercial software systems [19]. Unfortunately, the practical utility of Eq. (8) is limited because the iid assumption may not be realistic for the predictor variables. (The iid assumption is believed to be realistic for the data studied by McCue and Malinowski [31], but they designed an instrument with specific properties.) Another drawback is that estimates are required of the error variances for both predictor and predictand variables [16].

Eq. (8) only accounts for variance, not for bias, cf. Eq. (4). However, methods such as PLS derive much of their popularity from the ability to trade off variance against bias. By selecting fewer factors (A) than the number of independently contributing analytes (M), a prediction bias is introduced. This bias–variance trade-off is profitable if the increase of squared prediction bias is more than offset by the reduction of prediction variance. The importance of incorporating non-negligible prediction bias in prediction intervals has been recently discussed [15,18]. We have accounted for prediction bias as follows. Höskuldsson presents a formula for estimating prediction bias (see Eq. (8.24) on p. 248 in Ref. [32]). The bias is simply calculated by summing the contribution of the factors that are left out in the regression. This is done by estimating the full PLS model, where all components are included. The estimated bias is then the sum of the scores not used for prediction weighted by their corresponding regression coefficients. However, including the contribution of the "noise factors" leads to a bias estimate with relatively high uncertainty. Thus, we have modified this formula by restricting the summation to the factors that give a systematic contribution to the predictor array. The motivation is that the pseudo-rank of the predictor array (a single number) is an upper bound for the number of factors required for prediction (possibly different for individual analytes). The resulting bias estimate is added to the result of Eq. (8). This modification is generally applicable to score-based methods (for PCR, the relevant prediction bias formula is given by Næs and Martens [33]). Finally, it is noted that this approach provides a sample-specific root mean squared error of prediction rather than a standard error of prediction. However, to simplify the presentation, this distinction is not made in the rest of the paper.

2.4.2. Heteroscedastic errors in the predictors

An expression that is designed to work in the heteroscedastic predictor error case as well is derived as follows. First, it is observed that, in the absence of bias, the mean squared error of calibration (MSEC) estimates $\|\beta\|^2 V_{\Delta x} + V_e + V_{\Delta y}$ [34], so that Eq. (8) can be rewritten as:

$$\sigma_{\mathrm{PE}} \approx [(1+h)\mathrm{MSEC} - V_{\Delta y}]^{1/2}. \tag{9}$$

For zero-intercept models, the MSEC is obtained in the usual way from the squared fit errors as:

$$\mathrm{MSEC} = \frac{\sum_{i=1}^{I} (\hat{y}_i - \tilde{y}_i)^2}{I - A} \tag{10}$$

where $\hat{y}_i$ and $\tilde{y}_i$ denote the fitted and measured reference value for the ith calibration sample, respectively. A division by $I-A$, rather than I, is required to account for loss of degrees of freedom (one for each factor). For mean-centred models, the denominator in Eq. (10) is $I-A-1$ (the additional one is for the model centre).

It is reiterated that MSEC estimates $\|\beta\|^2 V_{\Delta x} + V_e + V_{\Delta y}$ in the absence of bias. However, if the calibration and prediction samples are exchangeable, Eq. (9) *implicitly* accounts for prediction bias through the incorporation of fit bias in MSEC. For a rigorous underpinning of this property, see Eqs. (6) and (7) in Denham [17]. (Although Denham assumes errorless predictor variables, the argument holds more generally.) The study of Denham is concerned with selecting the number of PLS factors without resorting to an independent test set. The latter is often referred to as external validation. To achieve this goal twelve estimators of mean squared error of prediction (MSEP) are compared. Some afterthought shows that Eq. (9) is intimately related to the estimator denoted as $\mathrm{MSEP}_{\mathrm{rss}_1}$. For both examples considered by Denham, the behaviour of $\mathrm{MSEP}_{\mathrm{rss}_1}$ did not reveal the correct number of factors (see Tables 1 and 3 in Ref. [17]). Consequently, $\mathrm{MSEP}_{\mathrm{rss}_1}$ should not be used to select the optimum model dimensionality. However, $\mathrm{MSEP}_{\mathrm{rss}_1}$ was close to the true MSEP for the correct number of factors. These results lend credibility to using Eq. (9) for calculating sample-specific standard errors of prediction when the model dimensionality has been selected using a dependable method. Denham concluded that cross-validation and bootstrapping methods are the best alternatives to external validation.

It is seen that by substituting the expression for MSEC, one effectively eliminates the error variances that are associated with *both* model and unknown sample term in Eq. (8). The measurement error in the reference values only contributes to the model term. As a result, $V_{\Delta y}$ is the only measurement variance present in Eq. (9). Although Eq. (9) is derived under the iid assumption, we conjecture that it applies to most types of heteroscedasticity (e.g., in spectroscopy). This conjecture is believed to be reasonable if the measurement noise is the same during the calibration and prediction stage: it is merely the *total effect* of the measurement noise that counts, rather than the *specific property* of the individual measurement errors. In the case that predictor noise increases with signal amplitude, Eq. (9) should work best if the variation in signal amplitude is limited. (In the current work, the validity of Eq. (9) is tested using proportional noise.) A rigorous approach to deal with heteroscedastic errors in the predictors is the full linearisation, as done in Ref. [14] for standard PLS. Obviously, the extreme simplicity of the current approach comes at a certain price.

Monte Carlo simulations, as conducted in this paper, are ideally suited for testing the validity of conjectures that are hard to verify theoretically. Eq. (9) is intended to be more generally applicable than Eq. (8), but it has a distinct disadvantage. Subtracting $V_{\Delta y}$ may yield negative variance estimates if $V_{\Delta y}$ is relatively large. Moreover, even if this subtraction leads to an admissible value, constructing a prediction interval on the basis of Eq. (9) is complicated, because the prediction variance need not be approximately distributed proportional to a simple χ^2. However, these problems are not unique to the currently proposed approach: they are generally persistent when estimating variance components [35].

2.5. Selection of optimum rank of PLS models

Faber [19] has found the performance of Eq. (8) to rely heavily on the ability to correctly estimate the optimum model dimensionality. The recent study of Denham [17] puts the same demand on the use of Eq. (9). We propose to verify the adequacy of the approximations leading to Eqs. (8) and (9) using Monte Carlo simulations. Since these simulations are performed unsupervised, i.e., without intervention of a human operator, the factor selection problem is not a trivial one. A common procedure for selecting the number of PLS factors is to monitor the average prediction error for an independent test set. Many researchers consider external validation to be wasteful because samples are set aside that could be used for model building [17]. Cross-validation or internal validation is a popular alternative, but we did not consider it here because it is very time-consuming. External validation yields an estimate of root mean squared error of prediction (RMSEP). The selected optimum number of factors is the one for which RMSEP is either a minimum or reaches a plateau.

Chance effects owing to the uncertainty in the estimated RMSEP values are likely to influence the decision, especially when focussing on a minimum. This reasoning suggests that, to avoid overfitting, PLS factors should be added to the model only if RMSEP decreases more than the associated uncertainty. Faber [36] has found that, when extreme outliers are excluded from the prediction set, the MSEP estimate is approximately distributed proportional to a χ^2 with degrees of freedom equal to the number of validation samples (I_{val}). Applying a linear approximation to standard expressions for mean and variance of a χ^2-variable yields that:

$$\frac{\sigma_{\text{RMSEP}}}{\text{RMSEP}} \approx \frac{1}{\sqrt{2I_{\text{val}}}}. \tag{11}$$

For certain Monte Carlo trials we found *two* adjacent RMSEP values differing less than their relative uncertainty, as calculated from Eq. (11), after which the RMSEP significantly dropped. In other words, a plateau had not yet been reached. After some trial and error we settled for the current Monte Carlo simulations to demand that *three* successive RMSEP values should differ less than their relative uncertainty. It is conceivable that this selection rule depends on the specific structure of the simulated data. However, this is of little practical importance, because selecting the number of factors is better not carried out unsupervised in practice.

2.6. Criterion for assessing the adequacy of the approximations

In Ref. [19], it is argued that a statistically sound criterion for assessing the quality of an approximate standard error of prediction is, that it should enable the construction of prediction intervals. It was found that, when applying Eq. (8) to standard PLS, the random variable

$$t = \frac{\text{PE}}{\hat{\sigma}_{\text{PE}}} \tag{12}$$

is approximately distributed as Student's t with an appropriate number of degrees of freedom. A quotient is distributed as Student's t if the numerator is normally distributed and the squared denominator is distributed as χ^2, independent of the numerator. While the normal assumption is usually tenable, the χ^2 assumption may be reasonable only when using Eq. (8). (See Ref. [15] for more details, such as degrees of freedom.) When using Eq. (9), the χ^2 assumption is too crude, unless $V_{\Delta y}$ is sufficiently small. Consequently, setting up generally applicable prediction intervals on the basis of Eq. (9) is, at the time, too ambitious.

Fortunately, Eq. (12) remains useful for validation purposes, even if prediction intervals are out of reach. This can be understood as follows: Clearly, one may consider the approximations to be adequate if the true prediction error is predicted correctly on average. This much softer requirement leads to a suitable criterion, because subtracting the measurement variance $V_{\Delta y}$ in Eq. (9) affects the shape of the distribution, rather than the associated standard deviation. The standard deviation of a t-value with f degrees of freedom is $\sqrt{f/(f-2)}$. Consequently, demanding that the true prediction error be predicted correctly on average amounts to demanding the expected standard deviation to approach $\sqrt{f/(f-2)}$. This expected value is easily obtained by averaging the standard deviations obtained for a series of independent repetitions.

3. Description of the simulations

Four-component systems ($M=4$) are simulated by multiplying Gaussian elution profiles by experimentally obtained ultraviolet (UV) spectra for adenine (A), cytidine (C), guanine (G) and uracyl (U) [37] (see Fig. 1). This automatically leads to zero-intercept models. Not testing the influence of mean centring is believed to be reasonable, because the contribution of

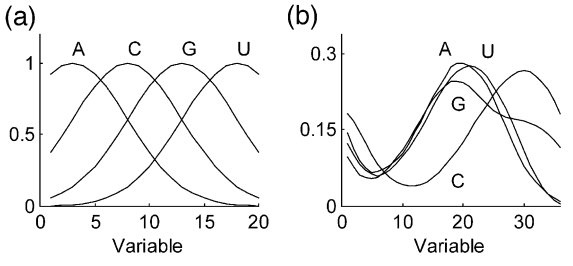

Fig. 1. (a) Simulated elution profiles and (b) experimental UV spectra for adenine (A), cytidine (C), guanine (G) and uracyl (U).

the intercept is rather trivial. (The intercept is accounted for by adding $1/I$ to the leverage.) Normally distributed noise is added at different levels (see Tables 1 and 2). Various noise settings are considered to thoroughly test the adequacy of the approximations leading to Eqs. (8) and (9). The initialisation of the pseudo-random number generator is identical for each noise setting. In practice, one or more potential sources of error variance may be negligible. For example, "clean" bilinear data correspond to a zero-residual problem (i.e., $\sigma_e = 0$). Consistent with this condition, the terms associated with σ_e and $\sigma_{\Delta y}$ are either zero or negligible in Refs. [38,39]. In contrast, σ_e dominates for the data analysed in Refs. [13,17,40]. Finally, Ref. [10] reports on an application where only $\sigma_{\Delta y}$ is considered. All these examples deal with first-order data, but the order of the data is irrelevant for the relative importance of sources of error variance. For each noise setting, 100 repetitions of three independent data sets are generated, namely a calibration set for model estimation, a validation set for factor selection and a prediction set for prediction error estimation (given the model). A fairly large number of independent repetitions (100) is conducted to reliably estimate the average standard deviation of the "*t*-values". Both the sizes of the calibration and validation set take values that can be considered intermediate (30) and fairly large (50). The number of prediction samples (2000) is chosen to be large enough to allow examining distributions. It follows that for each single case the evaluation of Eqs. (8) and (9) is based on $100 \times 2000 = 2 \times 10^5$ predictions. The range of the predictand is chosen so that 20% of the samples will fall slightly outside the calibrated space. In this way, the adequacy to cope with mild extrapolation is tested. While the predictand is corrupted exclusively by additive noise, the noise in the predictors is additive (cases 1–12) as well as proportional (cases 13–20). Proportional noise is added to test the validity of Eq. (9) when the predictor noise depends on signal magnitude, which is often the practical condition. The criterion for selecting the level of the noise in the predictors is that it should give an appreciable contribution to prediction error. For additive noise, it follows from Eq. (8) that $\|\beta\|^2 V_{\Delta x}$ should be of the same order as V_e and/or $V_{\Delta y}$. Table 3 lists the values for $\|\beta\|$ that have led to the selected variances. It is emphasised that different regression vectors are obtained in the absence of noise and using $A = M = 4$ factors. Elsewhere [41], it is shown that under these circumstances unfold-PLS yields the same regression vector as the bilinear least squares (BLLS) method proposed by Linder and Sundberg [24,25]. This observation is believed to be of interest because the BLLS regression vector has the

Table 1
Simulation parameters

Number of constituents (M)	4
Number of calibration samples (I)	30, 50
Number of validation samples	30, 50
Number of prediction samples	2000
Range of predictand for calibration and validation set	1–9
Range of predictand for prediction set	0–10
Number of J-mode variables (J)	20
Position of Gaussian peaks	3, 8, 13, 18
Standard deviation of Gaussian peaks	5
Number of K-mode variables (K)	36
Identity of spectra	A, C, G, U
Standard deviation of residual (σ_e)	0, 0.1
Standard deviation of noise in predictands ($\sigma_{\Delta y}$)	0, 0.1
Standard deviation of noise in predictors ($\sigma_{\Delta x}$)	0, 0.01, 0.05, 0.1, 3%, 5%

Table 2
Noise settings for different cases. The symbols are explained in the text

Case	σ_e	$\sigma_{\Delta y}$	$\sigma_{\Delta x}$
1	0	0	0.01
2	0	0	0.05
3	0	0	0.1
4	0	0.1	0
5	0	0.1	0.05
6	0	0.1	0.1
7	0.1	0	0
8	0.1	0	0.05
9	0.1	0	0.1
10	0.1	0.1	0
11	0.1	0.1	0.05
12	0.1	0.1	0.1
13	0	0	3%
14	0	0	5%
15	0	0.1	3%
16	0	0.1	5%
17	0.1	0	3%
18	0.1	0	5%
19	0.1	0.1	3%
20	0.1	0.1	5%

Table 3
Size of regression coefficient vectors ($\|\beta\|$) obtained with noiseless data and $A=M=4$ factors

Analyte	Unfold-PLS	Tri-PLS
Adenine	0.477	0.485
Cytidine	0.554	0.586
Guanine	0.803	0.904
Uracyl	0.720	0.827

smallest norm. The proper noise level in the heteroscedastic case is determined as follows. For the data sets generated in 100 repetitions, the average standard deviation of 1% proportional noise is 0.0178. Thus, to achieve a comparable effect as homoscedastic noise in the predictors, where $\sigma_{\Delta x}$ is 0.05 and 0.1, respectively, 3% and 5% relative noise is added, yielding average standard deviations of $3 \times 0.0178 \approx 0.05$ and $5 \times 0.0178 \approx 0.09$, respectively. All calculations are performed in Matlab (Mathworks, Inc.) and copies of the programs are available from the authors.

4. Results and discussion

Detailed results are presented only for the simulations where the number of calibration and validation samples is 30, because increasing the set sizes to 50 generally leads to similar results.

4.1. Selection of optimum rank of PLS models

Since the simulations are carried out unsupervised, one must critically examine the performance of the factor selection rule. The optimum selected ranks for unfold-PLS are summarised in Table 4. In all cases the selected ranks range from 3 to 5. With few exceptions, the selected rank for cytidine and uracyl is 4, which equals the number of analytes. The same behaviour is observed for all analytes when using tri-PLS (not shown). In contrast, the calibration of adenine and guanine using unfold-PLS often requires only three factors. A typical example of the behaviour of RMSEP is displayed in Fig. 2. For unfold-PLS, the RMSEP values are 2.18, 1.33, 0.167, 0.151, 0.137, 0.138 and 0.138, while for tri-PLS one obtains 2.18, 1.69, 0.832, 0.088, 0.092, 0.095 and 0.095. A three-dimensional model is selected for unfold-PLS because adding the fourth factor yields a 9.6% decrease in RMSEP and this value is smaller than the threshold dictated by Eq. (11), namely $1/\sqrt{2 \cdot 30} = 12.9\%$. Adding the fifth factor yields a further decrease in RMSEP of 9.3%, which is not considered to be significant either. It is noted that the same model

Table 4
Optimum selected rank (A) for unfold-PLS (100 runs). The number of calibration and validation samples is 30

Case	Adenine			Cytidine			Guanine			Uracyl		
	3	4	5	3	4	5	3	4	5	3	4	5
1	0	100	0	0	100	0	0	100	0	0	100	0
2	0	100	0	0	100	0	1	99	0	0	100	0
3	0	100	0	0	100	0	88	12	0	0	100	0
4	18	81	1	0	100	0	49	51	0	0	100	0
5	19	81	0	0	100	0	55	44	1	0	100	0
6	22	78	0	0	100	0	78	22	0	0	100	0
7	17	83	0	0	100	0	47	53	0	0	100	0
8	16	84	0	0	100	0	55	45	0	0	100	0
9	19	81	0	0	100	0	78	22	0	0	100	0
10	46	54	0	0	100	0	63	37	0	0	100	0
11	50	50	0	0	100	0	67	33	0	0	100	0
12	44	56	0	0	100	0	80	19	1	0	100	0
13	0	100	0	0	95	5	3	89	8	0	94	6
14	0	100	0	0	94	6	66	17	17	0	87	13
15	20	80	0	0	99	1	52	48	0	0	98	2
16	24	76	0	0	99	1	61	32	7	0	95	5
17	18	82	0	0	100	0	58	40	2	0	100	0
18	23	77	0	0	100	0	67	26	7	0	100	0
19	49	51	0	0	100	0	65	33	2	0	100	0
20	51	49	0	0	100	0	71	24	5	0	97	3

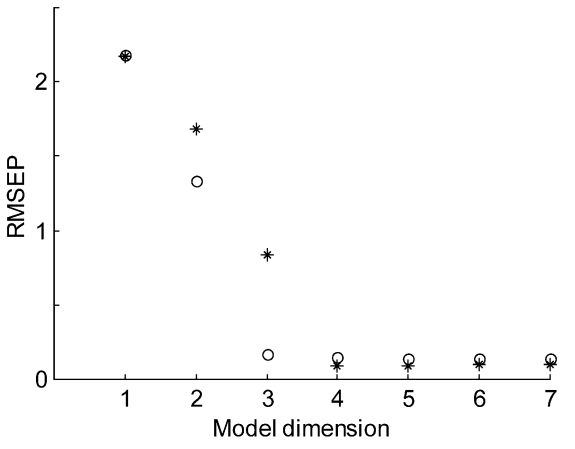

Fig. 2. RMSEP as a function of model dimension for guanine for a single run of case 3 where the number of calibration and validation samples is 30: unfold-PLS (○) and tri-PLS (*).

dimension would have been arrived at when using the threshold value for 50 validation samples, namely $1/\sqrt{2 \cdot 50} = 10\%$. One might argue that, without prior knowledge about the true underlying dimension, a human operator would have selected 5 factors, because this leads to an apparently stable minimum in RMSEP. Thus, the currently deployed selection rule seems to be somewhat conservative. This property could, however, be an advantage for the current evaluation study, because it leads to additional cases where the incorporation of bias in Eqs. (8) and (9) can be tested.

Despite the conservative character of the factor selection rule, the true underlying dimension is overestimated in some cases (see Table 4). This phenomenon is most often observed for the cases where proportional noise is added to the predictor variables (cases 13–20). A plausible explanation is that structured noise "displaces" genuine predictive information to the higher-numbered factors. For tri-PLS one finds $9+5+16+5=35$ of these cases against $1+13+50+29=93$ for unfold-PLS. Thus, unfold-PLS seems to be more prone to this effect than tri-PLS, which is in agreement with the earlier observation that unfold-PLS is more likely to overfit than tri-PLS [2]. This is especially true for the cases where the effect of the noise in $\mathbf{X}$ is not diluted by the effect of the noise in $\mathbf{y}$ (cases 13 and 14).

In summary, being able to find an explanation for the optimum selected ranks suggests that these numbers are reasonable.

4.2. Predictive ability of unfold- and tri-PLS

Although we are not engaged in a method comparison study, it is of practical interest to compare the predictive ability of unfold- and tri-PLS. It turns out that tri-PLS performs consistently better with the sole exception of case 1, which is characterised by a low level of the noise (Table 5). These results are in agreement with the earlier observation that tri-PLS is less sensitive to noise [2].

It is noted that the RMSEP values for cases 4–6, 10–12, 15, 16, 19, 20 are based on comparing the predictions with "measured" reference values ($\sigma_{\Delta y}=0.1$), rather than the true ones. DiFoggio [42] has pointed out that this procedure leads to a so-called *apparent* RMSEP. The apparent RMSEP systemati-

Table 5
Average RMSEP (100 runs). The number of calibration and validation samples is 30. The numbers printed bold mark the instances where tri-PLS performs worse than unfold-PLS

Case	Unfold-PLS				Tri-PLS			
	A	C	G	U	A	C	G	U
1	0.0053	0.0061	0.0088	0.0079	**0.0054**	**0.0064**	**0.0097**	**0.0089**
2	0.028	0.035	0.056	0.047	0.027	0.032	0.049	0.045
3	0.061	0.090	0.168	0.130	0.054	0.064	0.099	0.090
4	0.107	0.107	0.110	0.105	0.107	0.107	0.109	0.105
5	0.110	0.113	0.124	0.115	0.109	0.112	0.121	0.114
6	0.122	0.141	0.196	0.166	0.118	0.125	0.149	0.139
7	0.108	0.105	0.107	0.107	0.108	0.105	0.106	0.107
8	0.110	0.111	0.121	0.117	**0.111**	0.111	0.117	0.117
9	0.122	0.139	0.195	0.167	0.119	0.124	0.145	0.141
10	0.152	0.149	0.156	0.150	0.149	0.149	0.154	0.150
11	0.154	0.153	0.166	0.157	0.150	0.153	0.162	0.157
12	0.160	0.175	0.224	0.196	0.157	0.163	0.184	0.175
13	0.032	0.043	0.067	0.055	0.032	0.037	0.059	0.053
14	0.056	0.087	0.151	0.115	0.053	0.063	0.100	0.089
15	0.111	0.116	0.129	0.118	**0.112**	0.113	0.124	0.117
16	0.119	0.140	0.184	0.155	0.118	0.124	0.148	0.137
17	0.112	0.114	0.127	0.121	0.112	0.112	0.121	0.120
18	0.121	0.137	0.183	0.157	0.119	0.123	0.145	0.140
19	0.154	0.156	0.170	0.159	0.151	0.154	0.164	0.159
20	0.159	0.174	0.214	0.188	0.156	0.162	0.183	0.175

cally overestimates the *actual* RMSEP, because the measured reference values contain a spurious random component that cannot be predicted. This random component (measurement noise) is, unlike the residual (e), not associated with the true value and its contribution to the RMSEP estimate is therefore misleading. Case 4, where this measurement error is the only source of error variance, represents the most extreme example. The apparent RMSEP is bounded below by $\sigma_{\Delta y}=0.1$, whereas the (unknown) actual RMSEP is much smaller. DiFoggio has already pointed out that model predictions can be much more precise than the reference method. A simple correction procedure has been proposed by Faber and Kowalski [43]. For a recent critical review of the subject, see DiFoggio [44].

4.3. Adequacy of approximations leading to Eq. (8)

The criterion for testing the adequacy of the approximations leading to Eq. (8) is that the average standard deviation of the "t-values" calculated using Eq. (12) should be close to $\sqrt{f/(f-2)}$. Since we have evaluated Eq. (8) using the input values for the error variances, the degrees of freedom are taken to be infinite. This is not exact, because the uncertainty in the estimates for h and $\|\beta\|$ is ignored. Exact degrees of freedom are only possible if the predictor variables are error-free ($\sigma_{\Delta x}=0$). In practice, f would be estimated as a function of the degrees of freedom of the individual error variance estimates [15]. With the postulated value for f, the average standard deviation should approach unity. Three general observations can be made from the results presented in Table 6: First, the results are clearly best for tri-PLS. Second, incorporating bias often leads to standard deviations closer to the target value, which is expected. Third, the procedure is slightly conservative for tri-PLS. The latter can be understood as follows: The standard deviation of the "t-values" is systematically underestimated, which must be due to systematically overestimating the standard error of prediction when using Eq. (8). However, the amount of overestimation does not seem to be of practical importance.

For unfold-PLS, the standard deviations are too large for cases 2, 3, 6, 9 and 12. This means that the standard error of prediction is systematically *underestimated* when using Eq. (8). The explanation for this behaviour is found by inspecting the regression coefficient estimates. For the worst case encountered in Table 6, namely case 3, these estimates are extremely noisy for unfold-PLS while they are rather smooth for tri-PLS (see Fig. 3). For linear methods such as PCR, variance in the parameter estimates strictly increases with increasing number of factors (complexity of the model). This general rule for linear methods corresponds to a strong tendency for non-linear methods such as PLS (see Ref. [45] for an exception). Consequently, the apparently high variance in the unfold-PLS coefficient estimates is quite unexpected, because unfold-PLS uses less factors than tri-PLS (see Fig. 2). The excellent results obtained for tri-PLS suggest that

Table 6
Average standard deviation of "t-values" calculated using Eqs. (8) and (12) (100 runs). The number of calibration and validation samples is 30. The numbers in parentheses are calculated without taking account of bias

Case	Unfold-PLS				Tri-PLS			
	A	C	G	U	A	C	G	U
1	1.00	0.99	1.01	1.00	0.99	0.99	0.99	0.98
2	1.06	1.16	1.29 (1.30)	1.23	0.99	0.99	1.00	0.99
3	1.24	1.55	1.76 (2.12)	1.76	1.00	0.99	1.01	1.00
4	0.79 (1.22)	0.85	0.88 (1.28)	0.81	0.79	0.85	0.84	0.81
5	0.84 (1.13)	0.96	1.06 (1.28)	0.98	0.84	0.91	0.96	0.92
6	1.00 (1.20)	1.31	1.63 (1.90)	1.50	0.90	0.95	1.00	0.97
7	0.96 (1.03)	1.00	0.97 (1.02)	0.98	0.99	0.99	0.97	0.98
8	0.97 (1.03)	1.01	1.02 (1.08)	1.00	0.98	0.99	0.98	0.97
9	1.02 (1.08)	1.15	1.37 (1.50)	1.28	0.98	0.99	0.99	0.98
10	0.95 (1.10)	0.97	0.97 (1.05)	0.96	0.96	0.97	0.95	0.96
11	0.95 (1.11)	0.98	1.01 (1.09)	0.98	0.96	0.96	0.96	0.96
12	0.97 (1.11)	1.11	1.32 (1.46)	1.21	0.96	0.96	0.98	0.97

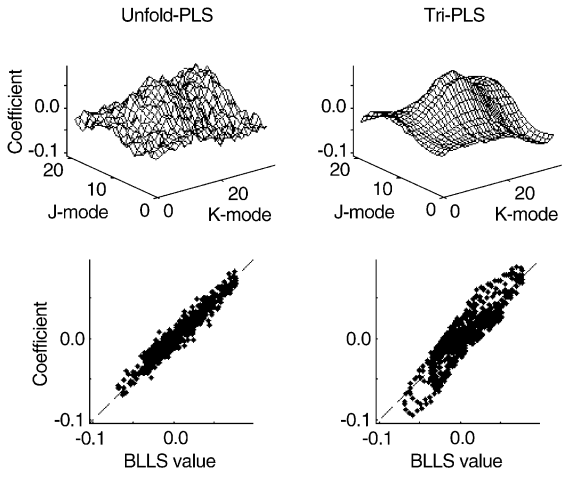

Fig. 3. Top: regression coefficient estimates obtained for guanine for a single run of case 3 where the number of calibration and validation samples is 30. Bottom: same, but plotted against the "true" BLLS values. The unfold- and tri-PLS models are constructed using $A=3$ and $A=4$ factors, respectively. The BLLS values are based on $A=M=4$ factors.

the additional approximation leading to Eq. (8) is not the basis of the problem. Apparently, the local linearisation of the model is not adequate for unfold-PLS at the current level of the noise. (Recall that local linearisation works best if the errors are small.) As a result, the standard error in the parameter estimates increases *non-linearly* with increasing standard deviation of the noise, likewise the standard error of prediction. The results are slightly better for unfold-PLS when increasing the number of calibration samples to 50 (not shown). The preceding discussion is subjective in the sense that it is not clear how one should assess the coefficient estimates to be smooth enough for Eq. (8) to yield trustworthy results. Thus, an important subject for future research is to develop diagnostics that indicate the onset of *non-linear* behaviour with respect to predictor noise. An important requirement for these diagnostics is that they be reliably estimable from the data. Stewart [46] has derived such a diagnostic for OLS with errors in the predictor variables. This diagnostic is similar to a signal-to-noise ratio; a basic ingredient is the smallest singular value of the predictor matrix (full column rank). It is consistent with a criterion that governs asymptotic results for the singular value decomposition (SVD) (see Theorem 4 in Ref. [47]). While it should be straightforward to generalise this diagnostic to unfold-PLS (possibly rank-deficient **X**), a difficulty arises in the case of tri-PLS when $N>2$. In that specific case, one needs a diagnostic that considers the N-way structure of the predictors. Hopke et al. [48] have recently generalised the matrix condition number (ratio of largest and smallest singular value) to $N>2$.

Unfold-PLS and Linder and Sundberg's BLLS method yield identical parameters in the absence of noise and using $A=M=4$ factors [41]. Consistent with this result, it is observed in Fig. 3 that the unfold-PLS parameters randomly scatter around the "true" BLLS values. In contrast, the tri-PLS parameters are either shrunk or expanded. This observation suggests that some prediction bias may be present when using tri-PLS with $A=M=4$ factors. It is emphasised that this additional source of bias has been ignored in the currently proposed approach, because Höskuldsson's procedure [32] only accounts for underfactoring bias ($A<M$). However, simulations without noise demonstrated it to be negligible for the current structure of the data (not shown). Investigating this phenomenon theoretically as well as experimentally could be of considerable interest, but is outside the scope of this paper. It is worth pointing out that De Jong [49] has shown the length of the regression vector estimated by standard PLS to be *shrunk* in comparison with the OLS estimator (or with PLS estimators based on a larger number of dimensions) (see Ref. [50] for more details). De Jong's result should be contrasted with the current observation that the norm of the tri-PLS estimator *exceeds* the norm of the BLLS estimator. This observation is believed to be relevant, because the BLLS estimator is best (minimum variance) linear unbiased (BLUE) when calibration is error-free and the noise in the unknown sample predictor variables is iid. In other words, the BLLS estimator has the status of "golden standard" for second-order bilinear calibration, similar to the OLS estimator for zeroth- and first-order calibration.

The distributions of the standard errors of prediction and "*t*-values" are shown in Fig. 4 for the models characterised in Figs. 2 and 3. The norms of the regression vector estimates are 0.770 and 0.895 for unfold- and tri-PLS, respectively (for "true" values, see Table 3). As a result, 0.0770 and 0.0895 are hard lower bounds for the results of Eq. (8). Clearly, Eq. (8)

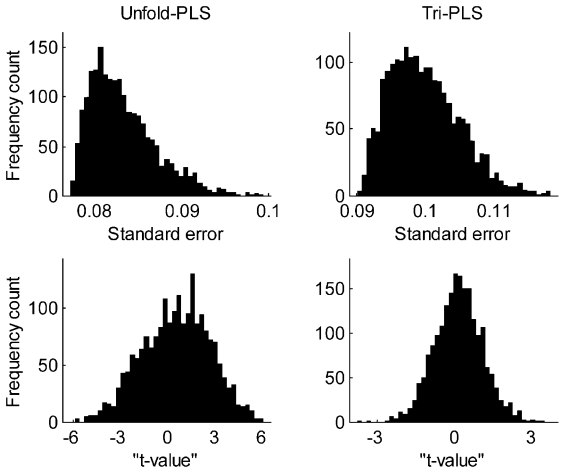

Fig. 4. Distribution of standard errors of prediction calculated using Eq. (8) and "*t*-values" calculated using Eq. (12) for guanine for a single run of case 3 where the number of calibration and validation samples is 30. The unfold- and tri-PLS models are constructed using $A=3$ and $A=4$ factors, respectively.

predicts unfold-PLS to perform better than tri-PLS, which is misleading because the unfold-PLS parameter estimates do not respond linearly to predictor noise. This is particularly obvious from the distribution of the "*t*-values", which is far too wide for unfold-PLS.

Cases 4 and 5 require special attention because some of the averages are much smaller than unity for both unfold- and tri-PLS. This result implies that the model contribution to Eq. (8) is seriously overestimated. In other words, Eq. (8) is overly conservative. We will focus on case 4, for which this behaviour is most pronounced. It is observed that for case 4 the model term is entirely determined by the measurement error in the reference values. Since the prediction sample term is absent, it is inferred that the leverage captures between 79% and 88% of the true model term. On first sight this may seem disappointing, because developing expressions for the model term has been the focus of this paper. However, it is actually a promising result, because in practice the model term is usually (much) smaller than the unknown sample term. The reason for this is that, unless the unknown sample is an extreme outlier, the leverage, which largely determines the size of the unknown sample term, is smaller than one. This can be derived as follows. A rule of thumb for labelling a sample as an outlier is that its leverage be larger than three times the average leverage of the calibration set. For zero-intercept models, the average leverage is A/I; if mean centring is applied, A is replaced by $A+1$. For example, the current simulations (zero-intercept models) lead to a maximum value for A/I of $5/30 \approx 0.17 \ll 1$. It is noted that for case 5, where noise is added to the predictors, the conservative character of Eq. (8) is best observed for tri-PLS, because for unfold-PLS the opposite tendency due to noisy regression coefficients is at work (see above). For case 6, the latter effect even dominates which leads to results that are difficult to interpret. Since tri-PLS yields rather smooth parameter estimates for all noise settings, the conservative character of Eq. (8) is most pronounced for the cases where $\sigma_{\Delta y}=0.1$, i.e., cases 4–6 and 10–12.

For cases 1–3 homoscedastic noise is added to the predictors only so that Eq. (8) predicts RMSEP to scale up linearly with the standard deviation of the noise, the norm of the regression vector estimate being the scale factor. This behaviour is correctly predicted for tri-PLS, but not for unfold-PLS (see Table 5). The norm of the regression vector estimate is an analytical figure of merit, because it determines the effect of homoscedastic predictor noise on prediction error. Because of the homoscedasticity requirement, it is known as the index of random error [51]. It has also been termed the "inverse sensitivity", since it is the reciprocal of the "sensitivity" encountered in the classical model [29]. It is important to note that currently two sets of definitions exist for analytical figures of merit. This has led to some controversy in the literature [29]. It is discussed in detail elsewhere [41] that the definition of Ho et al. [27] should be used in combination with the generalized rank annihilation method (GRAM). By contrast, a calibration model constructed using Linder and Sundberg's BLLS method, unfold-PLS or unfold-PCR is correctly characterised using the figures of merit derived by Messick et al. [28]. The preceding discussion illustrates that the "inverse sensitivity" is only informative if the pertinent model is sufficiently smooth to yield a linear response to predictor noise. Consequently, only for case 1, the smaller "inverse sensitivity" or larger "sensitivity" of unfold-PLS (see Table 3) correctly predicts the smaller RMSEP in Table 5. Bro has given an example where the coefficient estimates are larger

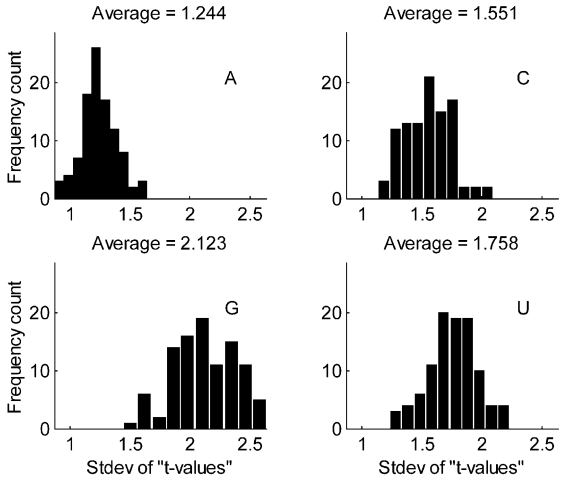

Fig. 5. Distribution of standard deviation of "*t*-values" calculated using Eqs. (8) and (12) when using unfold-PLS for case 3 where the number of calibration and validation samples is 30 (100 runs): adenine (A), cytidine (C), guanine (G) and uracyl (U).

for multilinear PLS by approximately a factor of two (see Fig. 7 in Ref. [2]). However, the coefficient estimates are much noisier for unfold-PLS. The nonlinear response to predictor noise explains why for that particular example the prediction results are better for multilinear PLS.

The tri-PLS RMSEP values obtained for cases 13 and 14 show that the "inverse sensitivity" is a semiquantitative measure if the predictor noise is heteroscedastic. This is best understood by observing that the RMSEP values for cases 3 and 14 are almost identical. However, for case 14 the average standard deviation of the proportional noise is $5 \times 0.0178 \approx 0.09$ (see Section 3). In other words, it is approximately 10% smaller than the standard deviation of the homoscedastic noise added in case 3 ($\sigma_{\Delta X} = 0.1$). The interpretation of this result is that, for the current simulations, the effect of predictor noise is approximately 10% larger if it is proportional. (A similar result has been reported for the calibration of experimental UV data using GRAM [52].) Consequently, even in the heteroscedastic predictor noise case practitioners may benefit from the enhanced interpretability of Eq. (8) if the analytical determination is to be customised to specific needs.

Finally, the distributions of the standard deviations of the "*t*-values" (Figs. 5 and 6) illustrate why series of independent repetitions must be performed to assess the potential utility of Eq. (8). While the distributions seem to be arbitrary for unfold-PLS, they are more or less reproducible for tri-PLS. The latter distributions are positively skewed so that the mode is smaller than unity. In practice, one would use a single model for many predictions. Depending on the particular quality of this (single) model, the use of Eq. (8) would lead to systematically under- or overestimating the prediction error. This problem is, however, hard to avoid. Consider, for example, the current practice of validating each individual prediction by the RMSEP, calculated from the prediction errors for a validation set. Since the RMSEP is a random variable, one will continuously under- or overestimate the prediction error too. In fact, the current approach is preferable in this respect, because Denham [17] has shown that it leads to RMSEP estimates with relatively low variability.

4.4. Adequacy of approximations leading to Eq. (9)

Eq. (9) is evaluated using the input value for $V_{\Delta y}$. As a result, f equals the degrees of freedom of MSEC and the standard deviation of the "*t*-

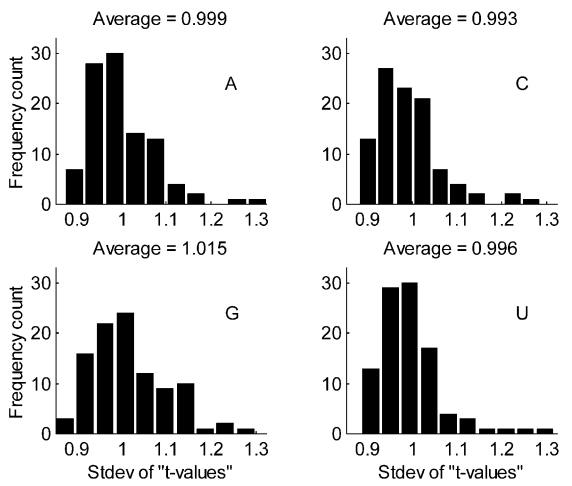

Fig. 6. Distribution of standard deviation of "*t*-values" calculated using Eqs. (8) and (12) when using tri-PLS for case 3 where the number of calibration and validation samples is 30 (100 runs): adenine (A), cytidine (C), guanine (G) and uracyl (U).

values'' should be compared with $\sqrt{f/(f-2)} = \sqrt{(I-A)/(I-A-2)} \approx \sqrt{(I-M)/(I-M-2)} \approx \sqrt{26/(24)} \approx 1.04$. Although directly following from Eq. (8) by a simple substitution, Eq. (9) differs substantially in that it may lead to inadmissible values for the standard error of prediction (square root of a negative number) when the measurement noise in the reference values is sizeable. We have discarded models where such a situation arises (see Table 7). The explanation of the results obtained when using unfold-PLS is found in Section 4.3 for cases 1–12, while for cases 13–20 it is obtained by identifying these cases with cases 2, 3, 5, 6, 8, 9, 11 and 12, respectively. As for Eq. (8), the results are consistently better for tri-PLS. However, contrary to Eq. (8), Eq. (9) does not seem to be conservative: the results are often quite close to the target value (1.04).

A tacit assumption underlying Eq. (9) is that the predictor noise should have a similar effect for calibration and prediction samples. The promising results obtained for tri-PLS suggest that violation of this assumption need not have serious consequences. This is remarkable, because the size of the predictor noise varies with signal amplitude, which, owing to the large range of analyte concentrations, is highly sample-dependent for the current simulations. Clearly, a more rigorous approach than the currently proposed one is to carry out the ''full linearisation'' for tri-PLS. Unfortunately, this is a cumbersome task. The reason for this is that tri-PLS uses an SVD to calculate the weight factors if $N=2$ and local linearisation of SVD is rather complex [14,47]. In addition, estimating a PARAFAC model forms the heart of the method if $N>2$. Linearisation results are just emerging for PARAFAC models, see Refs. [53,54] for $N=3$ and Ref. [54] for $N=4$. To the best of our knowledge, results are lacking for $N>4$. An attractive feature of the ''naïve approach'' is that the order of the predictor arrays is irrelevant. What counts is whether the uncertainty in the scores can be ignored.

In addition, two difficulties are envisioned when applying the ''full linearisation'' expressions in practice. First, the noise in the predictors must be fully characterised. Performing replicate measurements is the best way to do this, but this is obviously time-

Table 7
Average standard deviation of ''*t*-values'' calculated using Eqs. (9) and (12) (100 runs). The number of calibration and validation samples is 30. The numbers in parentheses denote the number of models (out of 100) that have been discarded because negative values were encountered for prediction error variance

Case	Unfold-PLS				Tri-PLS			
	A	C	G	U	A	C	G	U
1	1.02	1.03	1.05	1.03	1.03	1.02	1.06	1.03
2	1.12	1.21	1.36	1.28	1.04	1.03	1.07	1.03
3	1.38	1.70	1.83	1.90	1.05	1.03	1.09	1.05
4	0.78 (37)	0.60 (52)	0.75 (33)	0.55 (55)	0.57 (49)	0.72 (51)	0.58 (42)	0.55 (55)
5	0.89 (42)	0.85 (46)	1.11 (25)	0.91 (45)	0.65 (44)	0.76 (41)	0.84 (29)	0.85 (38)
6	1.16 (39)	1.50 (42)	2.04 (2)	1.77 (25)	0.87 (23)	0.98 (18)	1.10 (1)	1.11 (7)
7	1.02	1.03	1.02	1.03	1.03	1.03	1.00	1.03
8	1.04	1.08	1.09	1.07	1.02	1.03	1.01	1.02
9	1.16	1.32	1.51	1.44	1.03	1.05	1.03	1.03
10	1.11	1.03	1.05 (1)	1.05 (2)	1.09 (1)	1.03	1.01 (2)	1.05 (2)
11	1.12 (1)	1.11	1.13 (1)	1.10 (2)	1.08	1.04	1.02 (1)	1.06 (2)
12	1.26 (1)	1.44	1.59 (1)	1.51 (1)	1.06	1.06	1.04 (1)	1.09
13	1.06	1.30	1.47	1.38	1.00	1.01	1.03	1.01
14	1.17	1.66	2.02	1.87	1.01	1.02	1.05	1.03
15	0.86 (36)	0.88 (46)	1.16 (19)	0.93 (46)	0.68 (38)	0.84 (36)	0.92 (17)	0.90 (31)
16	1.05 (28)	1.36 (32)	1.74 (12)	1.64 (23)	0.83 (21)	0.94 (21)	1.04 (4)	1.09 (10)
17	1.04	1.09	1.14	1.08	1.03	1.03	1.01	1.02
18	1.10	1.26	1.50	1.30	1.03	1.03	1.02	1.02
19	1.11 (1)	1.12	1.16 (2)	1.15 (1)	1.07	1.04	1.01	1.07 (1)
20	1.19	1.35	1.46 (5)	1.38 (5)	1.05 (1)	1.05	1.02	1.07

consuming. Alternatively, Wang and Hopke [55] have proposed a method for estimating heteroscedastic noise from the data. The assumption is that the data array is large enough for the noise to be constant in a small region. It works by averaging squared residuals from an SVD in the neighbourhood of the target element. The method can be easily adapted to estimate correlations by averaging products of suitably neighbouring residuals. The second difficulty concerns storage and manipulation of the noise information. If the noise is correlated, a covariance matrix must be built that is dimensioned $IJK \times IJK$. For the current simulations, where the number of variables is rather small, this already leads to a $30 \cdot 20 \cdot 36 \times 30 \cdot 20 \cdot 36 = 21\,600 \times 21\,600$ matrix. Depending on the sparsity of this covariance matrix, storage problems can occur. (Storing the full symmetric matrix in double precision requires 1900 MB!) In addition, one would have to evaluate the Jacobian matrices that contain the partial derivatives of the parameter estimates with respect to the predictand and predictor variables. The Jacobian associated with the predictor variables is the largest of the two: it is dimensioned $JK \times IJK$. It is an open question whether the increased accuracy of the results outweighs the extreme "user-friendliness" of Eqs. (8) and (9). Finally, it is noted that similar problems often impede estimation of multiway models using the maximum likelihood method [56].

5. Conclusions and outlook

Two approaches to estimating standard error of prediction for multiway PLS have been investigated. The first approach, which is derived for homoscedastic predictor noise, requires estimates of all error variances. In contrast, the second approach, which is intended to work in the heteroscedastic case as well, only requires an estimate of the standard deviation of the measurement error in the reference values. Each approach has its strengths and weaknesses. Interestingly, both approaches are disfavoured by a relatively large measurement error in the reference values. While the first approach tends to be somewhat conservative, the second approach yields standard errors of prediction that in extreme cases cannot be used to construct prediction intervals in the usual way. For the current simulations, tri-PLS performed better than unfold-PLS. This cannot be expected to be a general rule. However, the current research may lead to insight that is required for carrying out a thorough method comparison. Eventually, this may lead to identification of regions where one PLS algorithm is to be preferred over the other.

From the current work, the following directions for future research can be given:

1. Demonstrate the practical utility of Eqs. (8) and (9) on real data. Work is in progress to apply the current error analysis to experimental EEM data [21].
2. Set up prediction intervals on the basis of Eq. (9). This problem is general (see [35]).
3. Develop diagnostics that indicate the breakdown of local linearisation. Potentially suitable precursors have been developed in connection with the SVD.
4. Quantify bias in the tri-PLS regression vector when $A = M$. This bias amounts to an identifiability problem that does not seem to have an analogue in the lower-order domain.
5. Carry out the full linearisation for tri-PLS. This amounts to incorporating the results of Paatero [53] and Liu and Sidiropoulos [54].

Acknowledgements

Constructive criticism by the reviewers is appreciated by the authors.

References

[1] S. Wold, P. Geladi, K. Esbensen, J. Øhman, J. Chemom. 1 (1987) 41.
[2] R. Bro, J. Chemom. 10 (1996) 47.
[3] R. Bro, H. Heimdal, Chemom. Intell. Lab. Syst. 34 (1996) 85.
[4] J. Nilsson, S. De Jong, A.K. Smilde, J. Chemom. 11 (1997) 511.
[5] A.K. Smilde, J. Chemom. 11 (1997) 367.
[6] S. De Jong, J. Chemom. 12 (1998) 77.
[7] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 93.
[8] A. Höskuldsson, J. Chemom. 2 (1988) 211.
[9] T.V. Karstang, J. Toft, O.M. Kvalheim, J. Chemom. 6 (1992) 177.
[10] A. Phatak, P.M. Reilly, A. Penlidis, Anal. Chim. Acta 277 (1993) 495.

[11] S. De Vries, C.J.F. Ter Braak, Chemom. Intell. Lab. Syst. 30 (1995) 239.
[12] K. Faber, B.R. Kowalski, Chemom. Intell. Lab. Syst. 34 (1996) 283.
[13] M.C. Denham, J. Chemom. 11 (1997) 39.
[14] K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 181.
[15] T. Morsing, C. Ekman, J. Chemom. 12 (1998) 295.
[16] M. Høy, K. Steen, H. Martens, Chemom. Intell. Lab. Syst. 44 (1998) 123.
[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] R. Bro, N.M. Faber, Chemom. Intell. Lab. Syst., in preparation.
[22] S. Leurgans, R.T. Ross, Stat. Sci. 7 (1992) 289.
[23] M. Kubista, Chemom. Intell. Lab. Syst. 7 (1990) 273.
[24] M. Linder, R. Sundberg, Chemom. Intell. Lab. Syst. 42 (1998) 159.
[25] M. Linder, R. Sundberg, J. Chemom., in press.
[26] A. Lorber, Anal. Chem. 58 (1986) 1167.
[27] C.-N. Ho, G.D. Christian, E.R. Davidson, Anal. Chem. 52 (1980) 1071.
[28] N.J. Messick, J.H. Kalivas, P.M. Lang, Anal. Chem. 68 (1996) 1572.
[29] K. Faber, A. Lorber, B.R. Kowalski, J. Chemom. 11 (1997) 419.
[30] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem, SIAM, Philadelphia, 1991.
[31] M. McCue, E.R. Malinowski, J. Chromatogr. Sci. 21 (1983) 229.
[32] A. Höskuldsson, Prediction Methods in Science and Technology, vol. 1, Basic Theory, Thor Publishing, Denmark, 1996.
[33] T. Næs, H. Martens, J. Chemom. 2 (1988) 155.
[34] S.D. Hodges, P.G. Moore, Appl. Stat. 21 (1972) 185.
[35] S.R. Searle, G. Casella, C.E. McCullogh, Variance Components, Wiley, New York, 1992.
[36] N.M. Faber, Chemom. Intell. Lab. Syst. 49 (1999) 79.
[37] F.P. Zscheile, H.C. Murray, G.A. Baker, R.G. Peddicord, Anal. Chem. 34 (1962) 1776.
[38] A.J. Berger, M.S. Feld, Appl. Spectrosc. 51 (1997) 725.
[39] J.F. Garcia, A. Izquierdo-Ridorsa, M. Toribio, G. Rauret, Anal. Chim. Acta 331 (1996) 33.
[40] N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler, Anal. Chem. 70 (1998) 2972.
[41] N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, Chemom. Intell. Lab. Syst., submitted for publication.
[42] R. DiFoggio, Appl. Spectrosc. 49 (1995) 67.
[43] K. Faber, B.R. Kowalski, Appl. Spectrosc. 51 (1997) 660.
[44] R. DiFoggio, Appl. Spectrosc. 54 (2000) 94A.
[45] N.M. Faber, J. Chemom. 13 (1999) 185.
[46] G.W. Stewart, Contemp. Math. 112 (1990) 171.
[47] L.A. Goodman, S.J. Haberman, JASA 85 (1990) 139–145.
[48] P.K. Hopke, P. Paatero, J. Hong, R.T. Ross, R.A. Harshman, Chemom. Intell. Lab. Syst. 43 (1998) 25.
[49] S. De Jong, J. Chemom. 9 (1995) 323.
[50] O.C. Lingjærde, N. Christophersen, Scand. J. Stat. 27 (2000) 459.
[51] H. Mark, Principles and Practices of Spectroscopic Calibration, Wiley, New York, 1990.
[52] N.M. Faber, J. Chemom. 15 (2001) 169.
[53] P. Paatero, Chemom. Intell. Lab. Syst. 38 (1997) 223.
[54] X. Liu, N.D. Sidiropoulos, IEEE Trans. Signal Process. 49 (2001) 2074.
[55] J.-H. Wang, P.K. Hopke, Anal. Chim. Acta 412 (2000) 177.
[56] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 339.