ELSEVIER

Chemometrics and Intelligent Laboratory Systems 30 (1995) 239–245

Chemometrics and intelligent laboratory systems

Prediction error in partial least squares regression: a critique on the deviation used in The Unscrambler

S. De Vries [1], Cajo J.F. Ter Braak *

Agricultural Mathematics Group (GLW-DLO), P.O. Box 100, 6700 AC Wageningen, Netherlands

Received 29 August 1994; accepted 14 March 1995

Abstract

Partial least squares (PLS) regression is commonly used for multivariate calibration of instruments. Because of the need to know the quality of the prediction in a specific unknown sample and the lack of theory, an 'empirically found formula' to express the uncertainty is utilized in The Unscrambler II software, the de-facto standard in computer software for PLS. In this critique the formula is examined theoretically and by simulation. It is concluded that this formula underestimates the root mean squared error of prediction in most practical applications of PLS. A change of the formula is planned in the next version of The Unscrambler. In the mean time users of The Unscrambler ver 5.5 or lower should multiply the reported deviation by a factor of at least $\sqrt{2(1-(A+1)/n)}$, to get a reasonable estimate of the prediction error.

Keywords: Partial least squares; Calibration; Multivariate calibration

1. Introduction

Partial least squares (PLS) regression [1–3] is a popular method for multivariate calibration of instruments, for example in the field of near infrared (NIR) reflectance spectroscopy. After calibration, the NIR instrument is used to make predictions of the amount of a chemical substance in an object from its spectral reflectances or absorbances.

In PLS, the average performance of the calibration is usually estimated through cross-validation. PLS lacks the theory about how to estimate the quality of each individual prediction. The error is expected to increase with the eccentricity (defined in an appropriate way) of the prediction object with respect to the set of calibration objects. However, the standard theory of ordinary least squares (OLS) regression is not directly applicable because it disregards the variability in the components in PLS. To appreciate this point we need to explain PLS regression in some detail. In PLS regression, the predictor matrix X and the response vector y are modeled as linear combinations of a set of orthogonal components. The components are latent variables which are linear combinations of the predictor variables. These linear combinations are chosen successively in such a way that they have maximum covariance with y [1].

* Corresponding author.

[1] Present address: Event AS, Gaustadalleen 21, N-0317 Oslo, Norway.

SSDI 0169-7439(95)00030-5

After the components have been derived, the predicted y values are obtained by ordinary least squares (OLS) regression of y onto the components. PLS regression of y onto X is thus equivalent to an OLS regression in which the set of regressors is replaced by a (usually smaller) set of components. In the standard theory of OLS, regressors are assumed fixed (measured without error). This theory is however not directly applicable because the components of PLS are random, since they depend on the random variable y. Despite the lack of theory, The Unscrambler II [4], the de-facto standard in computer software for PLS, provides for each prediction a 'deviation', also referred to as 'uncertainty limit'. This deviation is computed on the basis of an 'empirically found formula' developed in the eighties (Schönkopf, personal communication). It is noted that 'the deviation is not a standard formula that can be found in PLS or PCR theory, but an empirically found relationship that has given satisfactory indications on the uncertainty in predictions for a large range of applications' [4], page 342. The manual remains vague about the precise statistical interpretation of 'deviation'. In this critique we attempt an interpretation in terms of root mean squared error [1].

We used The Unscrambler II ver 3, but the formula is unchanged up to and including The Unscrambler II ver 5.5 (May 1994). Camo, the supplier of The Unscrambler II, appreciates the critique and plans to improve the uncertainty measure in a later version.

This critique consists of a theoretical part and a simulation part. In the theoretical part the special case is examined in which the number of PLS components is equal to the number of predictors. In this case PLS is equivalent with OLS regression. Here we show that the Unscrambler formula for deviation (from now on referred to as U-deviation) systematically underestimates the root mean squared error of prediction if the ratio of number of components over number of objects is small, but overestimates it if this ratio is large. In the simulation part we illustrate the theory and demonstrate that the U-deviation underestimates the root mean squared error of prediction in most practical applications, namely when the number of PLS components (i.e., number of latent variables to retain in the PLS model) is much smaller than the number of objects.

2. Theory

The linear model

$$y = \mathrm{X}\beta + \epsilon \tag{1}$$

is assumed where $y = (y_1,..., y_n)^T$ is the $(n \times 1)$ vector of responses, $\mathrm{X} = (x_1,..., x_n)^T$ is a $(n \times p)$ matrix of a ones vector and $p-1$ predictors, $x_i = (1, x_{i1},..., x_{ip-1})^T$, $i = 1,...,n$, is the $(p \times 1)$ vector of predictors for object i, $\beta = (\beta_0,..., \beta_{p-1})^T$ is a $(p \times 1)$ vector of unknown parameters and $\epsilon = (\epsilon_1,..., \epsilon_n)^T$ is a $(n \times 1)$ vector of random errors, which are identically and independently distributed with mean 0 and variance σ^2. In this notation, β_0 is the intercept. Correspondingly we assume a PLS model in which the model centre is the mean of the variables. The ith diagonal element of the hat matrix $H = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$ is called the leverage of object i and is given by $h_{ii} = x_i^T(\mathrm{X}^T\mathrm{X})^{-1}x_i$.

In the manual of The Unscrambler [4], page 342 'the empirically found formula for the computation of the deviation of Y-variable j in prediction object i' (shortly, U-deviation) is given as

$$\mathrm{Dev}(i,j) = \sqrt{\frac{\mathrm{Vy\_Val}}{2} \cdot \left(\frac{\mathrm{Vxi,pr}}{\mathrm{Vx\_Tot,val}} + \mathrm{Hi} + \frac{1}{\mathrm{Ical}} \right)} \tag{2}$$

where Vy_Val is the y-residual variance in the validation set, Vxi,pr is the x-residual variance in the prediction object, Vx_Tot,val is the average x-residual variance in the validation objects, Hi is the leverage of the prediction object with respect to the A PLS components (this leverage does thus not include the contribution of the intercept) and Ical is the number of calibration objects (n).

The mathematical definition of each of the terms in the U-deviation (2) is given in the Appendix. For the moment it suffices to remark that PLS is suitable for multivariate y (whence the index j in (2)), whereas we consider a single response variable only. Also, The Unscrambler uses the term calibration object for an object of the training set from which the PLS model must be estimated. For a single response variable y as in Eq. (1) and cross-validation by

leave-one-out the squared U-deviation for prediction object o can be rewritten without the index j as

$$\mathrm{Dev}^2(o) = \frac{1}{2n} \cdot \mathrm{PRESS_A} \cdot \left(\frac{\mathrm{Vxi,pr}}{\mathrm{Vx\_Tot,val}} + h_{oo,A} \right) \tag{3}$$

where PRESS_A is the prediction error sum of squares (see [5]) for the PLS regression model with A components and $h_{oo,A}$ is the leverage of prediction object o using a model with A components, because $\mathrm{PRESS}_A = n \cdot \mathrm{Vy\_Val}$ and $h_{oo,A} = \mathrm{Hi} + 1/\mathrm{Ical}$. Intuitively, the terms between brackets make sense. PLS divides the x-space in a part that is modeled by the components and a part that remains unexplained. The prediction error is thus likely to increase with both the amount of unexplained variance in the x-space and the eccentricity of the prediction object with respect to the components. The latter is measured by the leverage, in which the term $1/\mathrm{Ical}$ in Eq. (2) accounts for the contribution of the intercept, that is implicit in centred or autoscaled PLS. In the following our critique is focused on the term before the brackets.

If the number of components (A) is equal to the number of predictors ($p-1$), then PLS regression reduces to ordinary least squares (OLS) regression. The mean squared error of prediction (MSEP) in PLS regression with $A = p-1$ is thus given by the usual OLS formula [1,6]

$$\begin{aligned} \mathrm{MSEP}(\hat{y}_o) &= \sigma^2\left(1 + x_o^T(\mathrm{X}^T\mathrm{X})^{-1}x_o\right) \\ &= \sigma^2(1 + h_{oo,p-1}) \end{aligned} \tag{4}$$

where $\hat{y}_o = x_o^T\hat{\beta}$ is an estimate of the unknown response y_o, and x_o is the predictor vector of object o.

Unless the prediction object is an outlier

$$\frac{\mathrm{Vxi,pr}}{\mathrm{Vx\_Tot,val}} \approx 1$$

so that Unscrambler appears to use $(1/2n) \cdot \mathrm{PRESS}$ as an estimator of σ^2.

However, if $\mathrm{X}^T\mathrm{X}/n$ converges to a positive definite matrix as $n \to \infty$, then

$$E\left(\frac{1}{2n} \cdot \mathrm{PRESS}\right) = \frac{\sigma^2}{2n} \sum_{i=1}^{n} \frac{1}{1-h_{ii}} \to \frac{\sigma^2}{2} \text{ as } n \to \infty \tag{5}$$

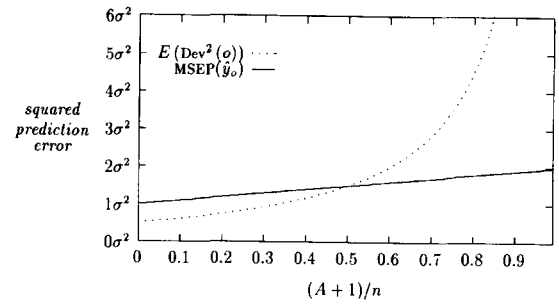

Fig. 1. Expected squared U-deviation in The Unscrambler II ($E(\mathrm{Dev}^2(o))$) and mean squared error of prediction ($\mathrm{MSEP}(\hat{y}_o)$) in relation to $(A+1)/n$ with A the number of components and n the number of training objects in OLS regression (PLS with $A = p-1$ components) if all leverages are equal. For unequal leverages, the two curves cross at $(A+1)/n < 0.5$. The simulations show that the same sort of curves are obtained for PLS with $A < p-1$, even in the case that $p > n$.

because $\max_{1 \le i \le n}(h_{ii}) \to 0$ as $n \to \infty$ [7]. By conclusion, the squared U-deviation asymptotically underestimates the MSEP by a factor of $1/2$ in PLS regression with $p-1$ components ($A = p-1$).

In [8] is it proven that for finite n

$$\frac{\sigma^2}{2}\left(1 + \frac{p}{n}\right) < E\left(\frac{1}{2n} \cdot \mathrm{PRESS}\right) \le \frac{\sigma^2}{2}\left(\frac{1 + p/n}{1 - \max_{1 \le i \le n}(h_{ii}^2)}\right) \tag{6}$$

Depending on the number of objects n and the number of predictors p, the squared U-deviation can under- or overestimate the MSEP. This is illustrated in Fig. 1 for the balanced case in which all leverages are equal (i.e., $h_{ii} = p/n$ for all i). In this case the expectation of $(1/2n) \cdot \mathrm{PRESS}$ reaches the upperbound of Eq. (6). If the number of predictors is small compared to the number of objects, then the squared U-deviation underestimates the MSEP. If the number of predictors almost equals the number of objects, then it overestimates the MSEP. Only in case there are twice as much objects as predictors, the squared U-deviation and the MSEP are equal. However, in practice the leverages h_{ii} are not all equal, so that $\max_{1 \le i \le n}(h_{ii}) > p/n$. To get an impression of what happens for unequal leverages, assume that the maximum leverage is less than k times the average leverage (p/n). Then, from the upper bound in Eq. (6), we derive that the break-even point where the squared

U-deviation is equal to the MSEP is at a value of p/n greater than $((1 + 8k^2)^{1/2} - 1)/4k^2$. For $k = 1$, the expression yields $p/n = 0.5$, as in Fig. 1. For $k = 2$ or 3 (i.e., when the training set does not contain outliers in the sense of high leverage points) the break-even point is at a value of p/n greater than 0.30 and 0.21, respectively. This indicates that, if p/n is small and the prediction object is not an outlier, the squared U-deviation will underestimate the MSEP.

This concludes the theory of PLS regression with $p - 1$ components. In practice, fewer components are used in PLS regression. Now recall that PLS regression is equivalent to an OLS regression with the PLS components as regressors. As a first guess of what happens to the MSEP and the squared U-deviation, we neglect the randomness of the PLS components. This is asymptotically correct ($n \to \infty$ for fixed $A \leq p - 1$) because the PLS components depend on first and second moments only (e.g., [9]), which are all estimated consistently. Now assuming fixed components, the foregoing theory holds true with predictors replaced by components and, consequently, p replaced by $A + 1$. (This is why the abscissa in Fig. 1 is labelled $(A + 1)/n$ instead of p/n.)

With fewer than $p - 1$ components, the leverages decrease, so that also the expectation in Eq. (5) decreases. Consequently, the U-deviation can be expected to underestimate the true mean squared error of prediction. A first guess of the amount of underestimation can be obtained from Fig. 1. Because PLS tends to use few components compared to the number of objects ($(A + 1)/n$ small) the squared U-deviation is conjectured to be only half the true MSEP (cf. asymptotically result, Eq. (5)).

3. Methods

To illustrate the theoretical results and to investigate what happens with the squared U-deviation if fewer than $p - 1$ components are used ($A < p - 1$), we carried out a simulation study. For the simulations three data sets are used. They are

MP42: This data set is taken from Table 4.2 in [6]. There are 25 objects, 2 predictor variables and a constant. The response variable is the delivery time in minutes and the predictor variables are the number of cases and the distance in feet.

SIAM: This data set is a chemical data example from [10]. It contains 15 objects, 8 predictor variables and a constant. The response variable is the biological activity. The predictor variables describe various properties of chemical compounds of the molecule.

GRAS: This data set is a subset of the grass data from the Laboratory for Soil and Crop testing (Oosterbeek, Netherlands). It contains 25 objects (samples), 44 predictor variables and a constant. The response variable is the digestibility coefficient and the predictor variables are from a NIR spectrum after preprocessing, sampled at 32 nm intervals from 1100–2500 nm.

Each data set is first analyzed by OLS (or PLS for the GRAS data) to obtain estimates for β and σ^2. The estimated parameters are taken as the true parameters of model (1). The model is then used to generate 1000 simulation data sets by adding independent normal errors with variance σ^2, where σ^2 is set to the estimated variance in the data set. In addition, responses of prediction objects are simulated. Prediction objects had the same predictor vectors as the calibration objects. Each simulation data set is subjected to OLS or PLS to estimate its model parameters. The responses of the prediction objects were predicted by inserting the newly estimated parameters in Eq. (1). For each prediction object, the prediction error (simulated response-prediction) was calculated and its square averaged across all 1000 simulated data sets to obtain an empirical estimate of the MSEP of the object. In the OLS-case ($A = p - 1$), the true MSEP can be calculated from Eq. (4). Because the empirical and theoretical values for the MSEP differed less than a few percent in our simulations, we present only the theoretical MSEP in the OLS-case.

The OLS-case is carried out for the first two data sets, the PLS-case for the last two data sets.

The simulations were carried out by a special purpose program written in C++. It used the Box-Muller transformation with the pseudo-random number generator ran1 from [11] to generate normal random errors, and the CGLS algorithm [12] to carry out autoscaled PLS regression. In the PLS-case the number of components was chosen such as to minimize the leave-one-out cross-validation MSE in each simulation data set based on the SIAM data, and, for tech-

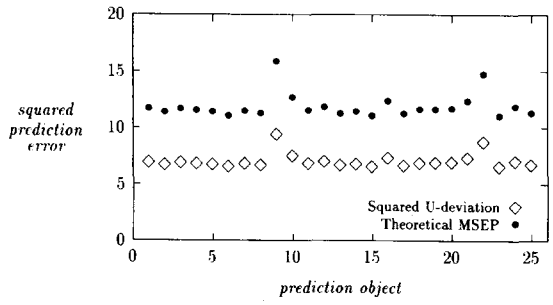

Fig. 2. Squared prediction error for the MP42 data in the OLS-case (PLS with 2 components). Parameters used: $\beta = (2.341, 1.616, 0.01438)^T$, variance used: $\sigma^2 = 10.62$.

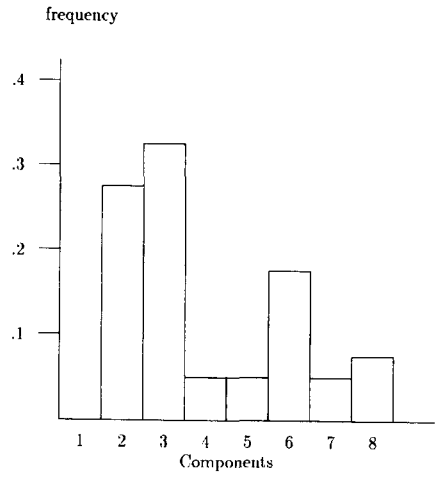

Fig. 4. Histogram of the optimal number of PLS-components A in the 1000 simulation data sets based on the SIAM data (A chosen by leave-one-out cross-validation).

nical reasons, was set to equal to the optimum number of components in the training data set in the simulations based on the GRAS data. The program produced the same numerical results for PLS as The Unscrambler, as was checked on some test data, including the SIAM data set.

4. Results

The OLS simulations (PLS with $A = p - 1$) are reported first. For the MP42 data (Fig. 2) the squared U-deviation underestimates the MSEP for all prediction objects whereas for the SIAM data (Fig. 3) it overestimates the MSEP.

In the PLS simulations ($A < p - 1$) based on the SIAM data set, the optimal number of components, estimated by leave-one-out cross-validation varied among simulated data sets with a mode at 3 components (Fig. 4). The squared U-deviation for the resulting PLS-models is about one-third of the empirical average squared deviation (Fig. 5). The same magnitudes of underestimation were obtained if the simulation data sets were generated using the PLS-estimate of β of the SIAM data or if σ^2 was increased with a factor ten (in which leave-one-out cross-validation judged usage of a single component optimal in 57% of these simulations). In the PLS simulations ($A < p - 1$) based on the GRAS data set,

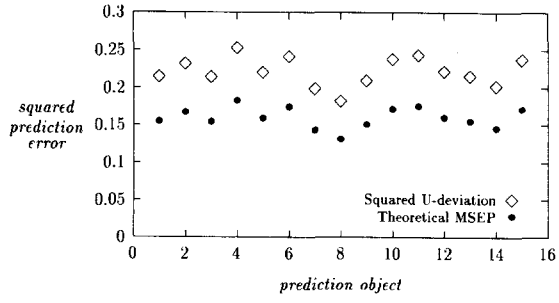

Fig. 3. Squared prediction error for the SIAM data in the OLS-case (PLS with 8 components). Parameters used: $\beta = (0.0638, 0.56, 0.17, 0.18, -0.68, 0.15, 0.97, -0.33, 0.55)^T$, variance used: $\sigma^2 = 0.1$.

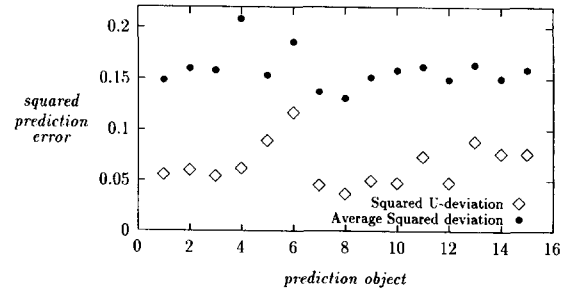

Fig. 5. Squared prediction error for the SIAM data in the PLS-case (A chosen by leave-one-out cross-validation). Parameters used: $\beta = (0.0638, 0.56, 0.17, 0.18, -0.68, 0.15, 0.97, -0.33, 0.55)^T$, variance used: $\sigma^2 = 0.1$.

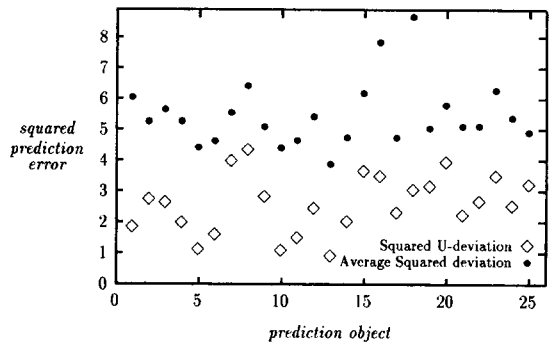

Fig. 6. Squared prediction error for the GRAS data in the PLS-case ($A = 10$). Parameters used: $\beta = (105.1, 4.44, -45.51, ..., -10.12, 0.5253)^T$, variance used: $\sigma^2 = 3.109$.

number of components was set to 10 (the optimal number of components for the GRAS data set). Here the squared U-deviation varied between one-and two-third of the empirical average squared deviation (Fig. 6). In some additional simulations in which the number of components was varied between 5 and 15, we observed that for smaller than 10 components the underestimation by the squared U-deviation was even more severely, whereas for higher than 10 components the underestimation gradually disappeared.

5. Discussion

The U-deviation is not a reliable estimator of the root mean squared error of prediction. The simulations largely confirm the theory. Although the leverages in the test data sets are unequal, the simulation results for the OLS-case ($A = p - 1$) can be explained qualitatively from Fig. 1. The ratio p/n is 0.12 and 0.6 for the MP42 and SIAM data sets, respectively. Fig. 1 shows that for $p/n = 0.12$ and 0.6 the squared U-deviation is 0.56 and 1.25 times the theoretical MSEP (Fig. 1), close to the multiplication factors in Fig. 2 and Fig. 3, respectively. The MP42 data set comes close to the asymptotically result that for large n the squared U-deviation is only half the true MSEP.

In practice, PLS is used with less than $p - 1$ components. We conjectured on the basis of the theory for the OLS-case that the squared U-deviation would be only half the true MSEP. In the simulations we found sometimes an even stronger underestimation. A possible explanation is that the true PRESS_A is underestimated in PLS because the minimum PRESS is selected in the cross-validatory choice of the number of components. Although we studied only small data sets in the simulations, we expect the conclusions to scale up to large data sets for the following reason. PLS components, at least the first few, are stable in large data sets, because they depend on the crossproducts only [9,13]. Therefore the PLS-case converges to the OLS-case with A fixed predictors for A and p fixed and $n \to \infty$.

This study has demonstrated that the U-deviation underestimates the MSEP in the practical application of PLS. A temporary fix is to replace the 2 in the U-deviation by $1/(1 - (A + 1)/n)$. The latter replacement reduces to 1 for $A \ll n$ and also works reasonably as $A \to p$ and $A \to n$. This replacement is suggested by formula (5) by taking equal leverages and thus corrects for the fault of the U-deviation shown in Fig. 1. Users of the Unscrambler II ver 5.5 or lower should multiply the reported deviation by a factor of at least $\sqrt{2(1 - (A + 1)/n)}$, until the formula has been improved in a new version.

Alternative, mathematically rigorous estimators for the MSEP in PLS regression have already been suggested. Phatak et al. [14] used Taylor series expansion and we developed a jackknife estimator of the MSEP [8] which we hope to publish in the near future [15].

Acknowledgements

We thank S. de Jong (Unilever Research Laboratorium, Vlaardingen) for help with unravelling the U-deviation and the suggestion of the temporary fix, J.G. de Gooijer for his supervision and interest in this project and Suzanne Schönkopf (CAMO AS), H. van der Voet and the referees for comments on the manuscript.

Appendix A

Mathematical definitions of the terms in the U-deviation (2) are as follows [4]. The notation of the orthogonalized PLS algorithm [1] frame 3.4 is used.

Vy_Val is the y-residual variance in the validation set,

$$\text{Vy\_Val} = \frac{1}{n_{\text{CV}}} \sum_{i=1}^{n_{\text{CV}}} \hat{f}_{ij,v(A)}^2 = \frac{1}{n_{\text{CV}}} \sum_{i=1}^{n_{CV}} \left(y_{ij} - \bar{y}_j - \sum_{a=1}^{A} \hat{t}_{ia} \hat{q}_{aj} \right)^2$$

with n_{CV} the number of objects used for (cross) validation, $\{\hat{f}_{ij,v(A)}\}$ the y-residuals in the validation set using a model with A components, y_{ij} the value of response variable j for validation object i, $\bar{y}_j$ the average for response variable j, $\hat{t}_{ia}$ the estimated score of validation object i for component a and $\hat{q}_{aj}$ the estimated loading of variable j for component a.

Vxi,pr is the x-residual variance in the prediction object,

$$\text{Vxi,pr} = \frac{1}{K-A} \sum_{k=1}^{K} \hat{E}_{ik,pr(A)}^2 = \frac{1}{K-A} \sum_{k=1}^{K} \left(x_{ik} - \bar{x}_k - \sum_{a=1}^{A} \hat{t}_{ia,pr} \hat{p}_{ak} \right)^2$$

with $\{\hat{E}_{ik,pr(A)}\}$ the x-residuals for the prediction object i using a model with A components, x_{ik} the value of predictor k for object i, $\bar{x}_k$ the average for predictor variable k, $\hat{t}_{ia,pr}$ the estimated score of prediction object i for component a, $\hat{p}_{ak}$ the estimated loading of predictor x_k for component a and K the number of predictor variables (exclusive the intercept, i.e., $K = p - 1$).

Vx_Tot,val is the average x-residual variance in the validation objects,

$$\text{Vx\_Tot,val} = \frac{1}{n_{\text{CV}}(K-A)} \sum_{i=1}^{n_{\text{CV}}} \sum_{k=1}^{K} \hat{E}_{ik,A}^2$$

with $\{\hat{E}_{ik,A}\}$ the x-residuals in the validation set defined analogously to the x-residuals of a prediction object.

Finally, Hi is the leverage of the prediction object with respect to the A PLS components (this leverage does thus not include the contribution of the intercept),

$$\text{Hi} = \sum_{a=1}^{A} \frac{\hat{t}_{ia,pr}^2}{t_a^T t_a}$$

with t_a the n-vector of component estimated scores in the calibration set.

References

[1] H. Martens and T. Næs, Multivariate Calibration, Wiley, Chichester, 1989.

[2] P. Geladi and B.R. Kowalski, Anal. Chim. Acta, 185 (1986) 1–17.

[3] A. Höskuldsson, J Chemom., 2 (1988) 211–228.

[4] The Unscrambler User's Guide, version 5.5, Camo A/S, Trondheim, 1994.

[5] D.M. Allen, Technometrics, 16 (1974) 125–127.

[6] D.C. Montgomery and E.A. Peck, Introduction to Linear Regression Analysis, 2nd edn., Wiley, New York, 1992.

[7] R.G. Miller, Ann. Statist., 2 (1974) 880–891.

[8] S. De Vries, Jackknife Methods and Partial Least Squares Regression, Report LWA-93-12, Agricultural Mathematics Group, Wageningen, 1993, pp. 29–34.

[9] S. De Jong and C.J.F. Ter Braak, J. Chemom., 8 (1994) 169–174.

[10] S. Wold, A. Ruhe, H. Wold and W.J. Dunn III, SIAM J. Stat. Comput., 5 (1984) 735–743.

[11] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.

[12] C.G. Paige and M.A. Saunders, ACM Trans. Math. Software, 8 (1982) 43–71.

[13] F. Lindgren, P. Geladi and S. Wold, J. Chemom., 7 (1993) 45-59.

[14] A. Phatak, P.M. Reilly and A. Penlidis, Anal. Chim. Acta, 277 (1993) 495–501.

[15] S. De Vries, C.J.F. Ter Braak and S. De Jong, in preparation.