Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 68 (2003) 41–51

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

An evaluation of the PoLiSh smoothed regression and the Monte Carlo Cross-Validation for the determination of the complexity of a PLS model

S. Gourvénec[a], J.A. Fernández Pierna[a], D.L. Massart[a,*], D.N. Rutledge[b]

[a] ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, Brussels B-1090, Belgium
[b] Laboratoire de Chimie Analytique, Institut National Agronomique 16, rue Claude Bernard, Paris 75005, France

Abstract

A crucial point of the PLS algorithm is the selection of the right number of factors or components (i.e., the determination of the optimal complexity of the system to avoid overfitting). The leave-one-out cross-validation is usually used to determine the optimal complexity of a PLS model, but in practice, it is found that often too many components are retained with this method. In this study, the Monte Carlo Cross-Validation (MCCV) and the PoLiSh smoothed regression are used and compared with the better known adjusted Wold's R criterion.

Keywords: PLS; Complexity; Monte Carlo Cross-Validation; Smoothing; Durbin–Watson criterion; Adjusted Wold's R criterion

1. Introduction

PLS regression [1–3] is widely used in spectroscopy for calibration and prediction and can be applied to several kinds of signals. The determination of the number of components or factors to retain is one of the most important steps in the building of a model. This determination of the correct number of latent variables (LVs) is crucial in order to avoid overfitting and, therefore, to obtain robust predictive models. Leave-one-out cross-validation is usually applied to establish the number of components needed. One can evaluate the predictive power of the model leaving out one sample at a time by calculating a statistic for lack of prediction accuracy, called Predicted Residual Error Sum of Squares (PRESS). This value is plotted against the number of components, and the number of components that gives a minimum PRESS is considered to be the optimal number of components to achieve the best prediction. In practice, it is found that often too many components are retained in this way. To avoid this, the Monte Carlo Cross-Validation (MCCV) and the PoLiSh smoothed regression have been proposed and are compared with the adjusted Wold's R criterion in this study.

The MCCV [4–6] is an asymptotically consistent method, but is rarely used in chemometrics. It is based on the same principle as the leave-one-out cross-validation, but instead of leaving out only one point ($n_v = 1$), subsets of different sizes ($n_v \geq 2$) are left out during the calibration. These n_v samples are then used for the validation. MCCV is considered to

[*] Corresponding author. Tel.: +32-2-477-47-34; fax: +32-2-477-47-35.

E-mail address: massart@vub.vub.ac.be (D.L. Massart).

0169-7439/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0169-7439(03)00086-8

avoid overfitting if n_v is more than 50% of the dataset [4].

The PoLiSh procedure [7] is an alternative way to decide the optimal complexity of a model. It is based on two main ideas.

The first one is to use Savitsky–Golay smoothing of the loadings weights vectors (***w***) obtained at each iteration step of the NIPALS procedure in order to progressively "displace" the random or quasi-random variations from earlier (most important) to later (less important) PLS latent variables. The second idea is to measure the structure of the regression PLS vectors (loadings and weights, *b*-coefficients, etc.) in order to determine the number of components required and to evaluate the optimal PLS model dimensionality. Two criteria are used: the Durbin–Watson [7–9] method and the comparison of the rotation angle between adjacent *b*-coefficient vectors [7]. The adjusted Wold's *R* criterion [10] is a way of interpreting leave-one-out cross-validation. With this criterion, one does not look for the minimum of PRESS but considers and compares two different values of PRESS corresponding to two different numbers of successive latent variables.

2. Theory

2.1. Leave-one-out cross-validation

Using a set of *n* calibration spectra, the PLS algorithm is performed on $(n-1)$ calibration spectra and, with this calibration, the quantitative variable (concentration) of the sample left out during calibration is predicted. This procedure is repeated *n* times until each sample has been left out once. The prediction for each sample is then compared with the known value of the reference sample. The sum of the squared variable prediction errors for all calibration samples is a measure of how well a particular PLS model fits the quantitative variable:

$$\text{PRESS} = \sum (y_{\text{predicted}} - y_{\text{real}})^2 \quad (1)$$

PRESS is calculated in the same way each time a new factor is added to the PLS model. The optimal order (number of factors) of the PLS model is the one that yields the minimum PRESS or Root Mean Square Error of Cross-Validation (RMSECV):

$$\text{RMSECV} = \sqrt{\frac{\text{PRESS}}{n}} \quad (2)$$

2.2. The PoLiSh procedure [7]

The PoLiSh procedure is a combination of three independent procedures: the Savitsky–Golay smoothing, the comparison of Durbin–Watson criteria, and the comparison of the angles between *b*-coefficient vectors.

The smoothing of the regression vectors within the PLS calculations is useful in reducing the noise level of the first latent variables and in transferring the noise into the later latent variables. After the smoothing step, the algorithm continues as in classical PLS and the new PRESS is calculated. The difference between the PRESS from the PoLiSh PLS and the PRESS from the classical PLS is calculated and plotted as a function of the number of latent variables.

2.2.1. The Savitsky–Golay smoothing [11–13]

In order to eliminate uninformative local spectral variations, it may be useful to smooth the data. The simplest way to smooth the data is by a moving average. In this method, a finite-size window is selected and the average of all points inside it is calculated. This average replaces the central point in the window:

$$x_{\text{smooth},i} = \frac{1}{(2p+1)} \sum_{j=i-p}^{j=i+p} x_j \quad (3)$$

where *i* is the index of the data points, *p* is the number of variables on each side of the actual variable, and $(2p+1)$ is the size of the window.

In the PoLiSh procedure, the smoothing using the Savitsky–Golay filters (also called digital smoothing polynomial filters or least squares smoothing filters) is used. It is also a moving window averaging method but now the central point in the window is replaced by the value of a polynomial that fits the data inside the window. In such a way, Savitsky–Golay filters are optimal in the sense that they minimize the least squares error in fitting a polynomial to each window of noisy data.

2.2.2. The Durbin–Watson criterion [8,9]

The second aspect of the PoLiSh procedure is the use of the Durbin–Watson criterion [8,9]. The Durbin–Watson test is usually used to investigate the (non) randomness (i.e., autocorrelation) of regression residuals. It examines the null hypothesis (H_0) that there is no correlation between the successive residuals, versus the alternative hypothesis that the correlation exists. The following statistic d is computed:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (4)$$

where e_i is the residual corresponding to the object i, and e_{i-1} is the residual of the preceding object (the objects being ranked according to time).

The d value is compared to two (lower and upper) critical values, d_L and d_U [8]. When $d < d_L$, the null hypothesis is rejected; when $d > d_U$, it is considered that there is no correlation between residuals. If $d_L < d < d_U$, the test is inconclusive.

In the PoLiSh procedure, the Durbin–Watson test is not applied to residuals but to the loadings (**p**), weights (**w**), and vectors of b-coefficients (**b**). The Durbin–Watson values are then plotted against each number of latent variables and it is easy to detect the inclusion of "unstructured" information (increasing Durbin–Watson values).

2.2.3. Angles between b-coefficient vectors

The study of the rotation angles between adjacent b-coefficient vectors is the third test used in the PoLiSh procedure. If the angle between two consecutive **b**-vectors does not change significantly, this means that they are strongly correlated and, therefore, that the second one contains no new information. On the plot of angles between **b**-vectors as a function of number of latent variables, the first angle that is low is considered as the optimal complexity of the model.

2.3. The Monte Carlo Cross-Validation [4–6]

The cross-validation method consists of splitting the data (n points) that are available for building the model into two parts. The first part (n_c points) is used to fit the model (calibration). The second part (n_v points) is left out and kept to validate the model and to assess its predictive ability. Cross-validation selects the model with the best average predictive ability, calculated and based on different values of n_c and n_v.

The cross-validation criterion is the RMSECV:

$$\mathrm{RMSECV}n_v = \sqrt{\frac{\mathrm{PRESS}}{n}} \qquad (5)$$

$$\text{with}\ \ \mathrm{PRESS} = \sum (y_{\text{predicted}} - y_{\text{real}})^2 \qquad (6)$$

This RMSECV is calculated for every kth component added to the model. The optimal complexity (i.e., the optimal number of components that should be included in the model) is determined by k, which gives the minimum RMSECV. The simplest way to carry out cross-validation (with $n_v = 1$ or leave-one-out cross-validation) is shown to be asymptotically incorrect (inconsistent) [4,5]. It tends to include an excessive number of components in the model and consequently brings overfitting. The Monte Carlo Cross-Validation keeps a large number of points (n_v) for the validation. It has been proven by Shao [5] that, under these conditions and when $n \to \infty$ and $n_v/n \to 1$, the probability for cross-validation to choose the correct model tends to 1. The MCCV method consists of repeating the procedure of CV n_v N times (in general, $N = n^2$ is enough in order to make MCCV n_v perform as well as CV n_v [4]).

The MCCV criterion is then:

$$\mathrm{MCCV}n_v(k) = \frac{1}{N}\mathrm{RMSECV}_{n_v} \qquad (7)$$

2.4. The adjusted Wold's R criterion [10]

This criterion compares two successive values of PRESS obtained with the leave-one-out cross-validation:

$$R = \frac{\mathrm{PRESS}_{h+1}}{\mathrm{PRESS}_h} \qquad (8)$$

where PRESS_h is calculated from the leave-one-out cross-validation (see Section 2.1) for the latent variable h.

When this ratio is in excess of unity, it is considered that the optimal number of latent variables is *h*, and this is called the Wold's *R* criterion. This criterion leads to the minimum of PRESS, which has been shown to have poor statistical properties. Instead of comparing this ratio to unity, it was then proposed to include the sampling variability in the limit, and to fix it at 0.90 or 0.95. This is called "the adjusted Wold's criterion." To decide on the optimal complexity of the model using this criterion, it is considered that if *R* is larger than 0.90 (or 0.95) for the following latent vectors, the latent variable $h+1$ does not bring significantly new information to the model and should not be included. This criterion will be denoted *R* in the following sections and the limit will be fixed at 0.90.

3. Data and pretreatment

Dataset 1 consists of 305 samples of forages and 174 wavelengths measured in the range 1108–2492 nm [14]. Two properties are modelled: the humidity at 103 °C and the raw protein. The dataset was split into two subsets: the first one (subset 1), consisting of 205 samples, was used to build the model; the second one (subset 2), consisting of 100 samples, was used to validate it. The data were mean-centered in the two cases.

Dataset 2 is based on a dataset first published by Kalivas [15]. It can be obtained from the database of the *Chemometrics and Intelligent Laboratory Systems*. It consists of two clusters and, in this work, only the first cluster is considered. The dataset is then reduced to 40 wheat samples and 701 variables (between 1100 and 2500 nm, each 2 nm) for the determination of moisture.

4. Results

4.1. Dataset 1 a: response variable: humidity

The standard PLS model was first built using subset 1. Leave-one-out cross-validation was used to compute RMSECV values to assess model performance. The optimal model, with the minimum RMSECV, requires 17 latent variables (Fig. 1).

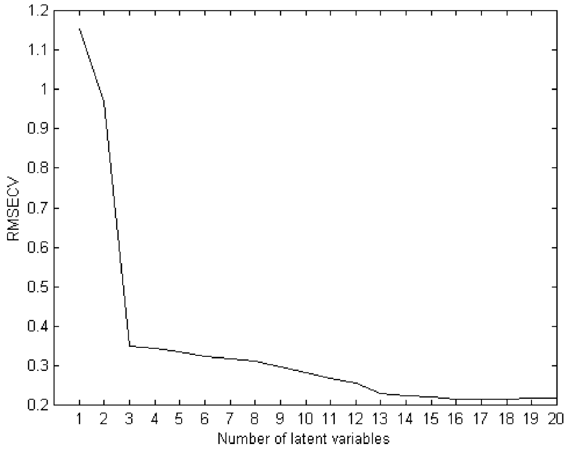

Fig. 1. RMSECV of the PLS model for dataset 1a (humidity).

A randomisation test [16] is used to compare the quality of prediction of two different complexities. In this case, it was concluded that the complexity can be reduced to 13 latent variables. The PoLiSh procedure was then applied to this set of data. On Fig. 2, one can observe the difference between the PoLiSh PRESS values and the PLS PRESS values (using internal cross-validation leave-one-out procedure).

In this figure, it is possible to see that for three latent variables, the difference of PRESS values is negative, meaning that the PRESS of PoLiSh is lower than the PRESS of PLS. This is explained by the fact that the smoothing in the PoLiSh procedure is useful in reducing the noise level in the first latent variables and in transferring the noise into the later latent variables. Despite the fact that the PRESS values do not decrease significantly for this number of latent variables (when compared to PLS), this comparison can highlight some model features. The important characteristic of this plot is the low PRESS values for PoLiSh for the third LV and the increase (i.e., the transition) between the third and the fourth LVs. Fig. 3 shows the Durbin–Watson profiles of the regression vectors (**p**, **w**, and **b**).

As can be seen from the dwP (Durbin–Watson values for the vector **p**) and dwW (Durbin–Watson values for the vector **w**) profiles, the vector shows structure up to the third LV (low dw values) and then these values start to increase, indicating the inclusion of "unstructured" information. This behaviour is not well reflected in the dwB (Durbin–Watson values for

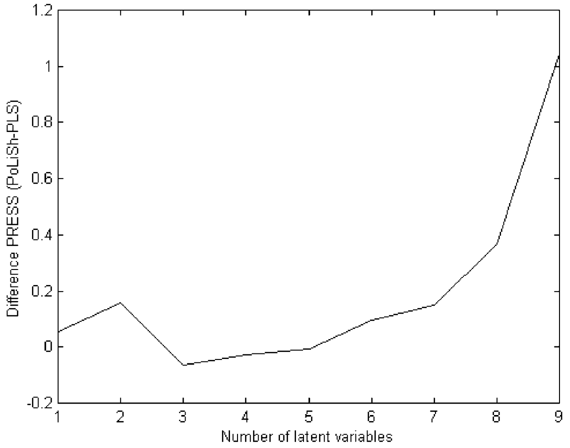

Fig. 2. Difference of PRESS values between PoLiSh and PLS models for dataset 1a (humidity).

the vector of b-coefficients) profiles, but using the two previous criteria (mainly dwP), one can have a better insight into the signal structure. From the previous discussion, one can assume that only the first three latent variables are important to build a stable PLS calibration model. These results give an indication of the model dimensionality [e.g., three or four latent variables—compared to 17 (or 13) latent variables given by leave-one-out cross-validation (or randomisation test)].

In Fig. 4, the plot of the angle between adjacent b-coefficient vectors in the PLS regression is given.

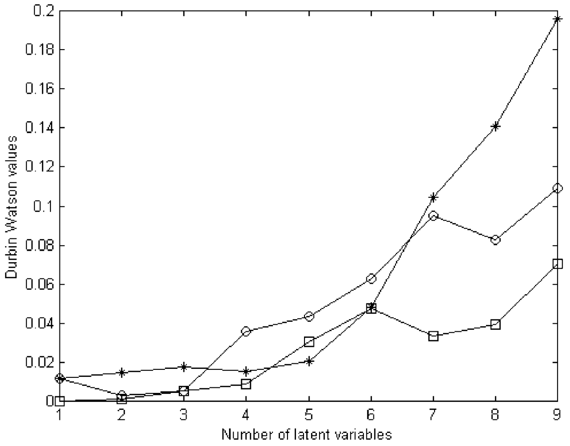

Fig. 3. Durbin–Watson profiles for the loadings **p** (–○–), weights **w** (–□–), and b-coefficient vectors (–∗–) of the PLS model for dataset 1a (humidity).

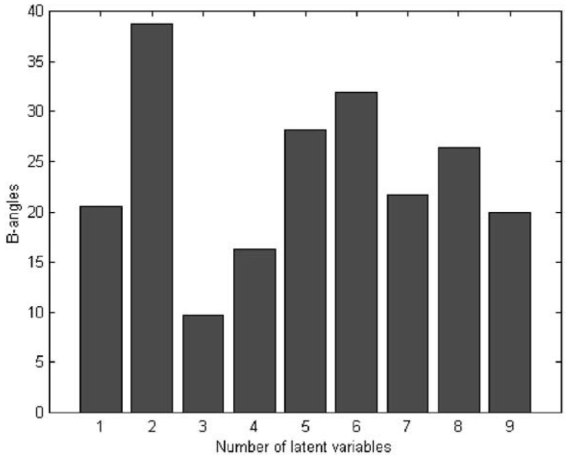

Fig. 4. Angles between adjacent b-coefficient vectors for dataset 1a (humidity).

A first minimum is obtained in the transition between the third and the fourth LV, which means that they point almost in the same direction. The angle between the fourth and the fifth LV clearly increases, suggesting that the model starts to introduce noise. From all these results, it can be concluded that the optimal complexity of the PLS model is 3.

In the Monte Carlo Cross-Validation, different values of n_v are tested to visualise the effect of the size of n_v on the cross-validation. For this dataset, n_v is successively equal to 1 (leave-one-out cross-validation), 150, and 175.

The results are shown in Fig. 5.

For $n_v = 1$ and 150, the minimum of MCCV is 17 latent variables, even if a local minimum is reached for three latent variables. For $n_v = 175$, the minimum is 3.

The adjusted R criterion gives also the same number of latent variables to build the model. Table 1 shows the values of R for the first latent variables.

R of step 3 (ratio of PRESS for four latent variables and PRESS for three latent variables) is larger than the limit 0.90 and leads to the conclusion that only three latent variables are enough to build the model. One can notice that this result is also valid with the limit of 0.95.

This criterion is in agreement with the results obtained with the PoLiSh procedure and MCCV, and shows that the model should be built with three components.

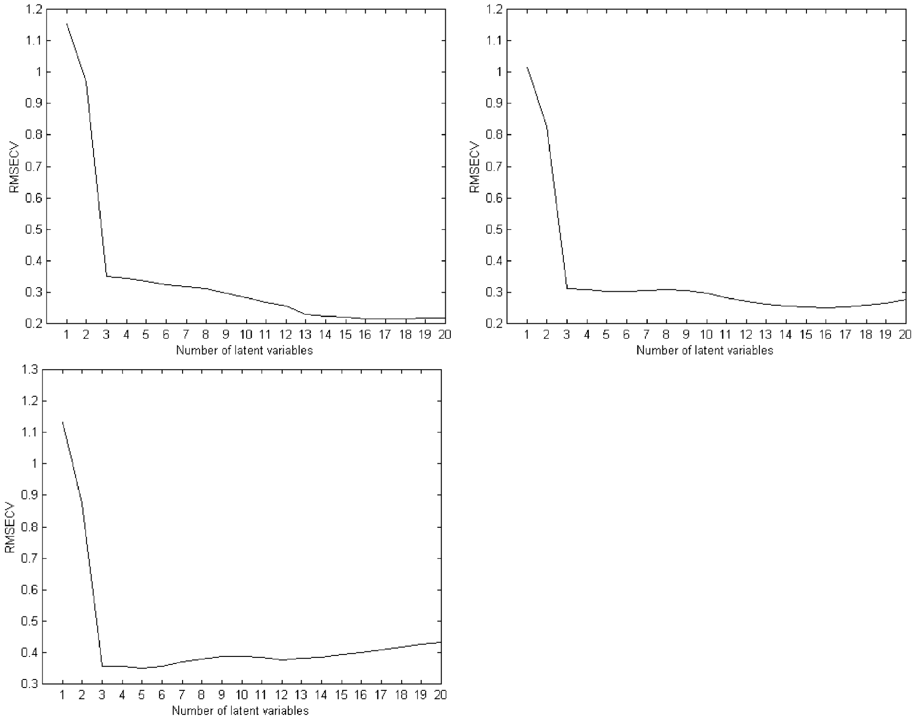

Fig. 5. MCCV plots: leave-n_v-out for dataset 1a (humidity).

This could be confirmed by validation using subset 2. The Root Mean Square Error of Prediction (RMSEP) for a model with 17 components is 0.65, with 13 components is 0.64, and with 3 components as suggested with the PoLiSh procedure or the MCCV is 0.80. These values are not so different, especially according to the range of y used for validation (between 3.12 and 145.1), and show that more robust models can be found with the methods presented here without affecting the quality of prediction.

4.2. Dataset 1 b: response variable: raw protein

For subset 1, the standard PLS regression (with leave-one-out cross-validation) shows a minimum of RMSECV for 18 latent variables (Fig. 6).

Table 1
Values of adjusted Wold's R criterion for the dataset 1a

Number of latent variables	Adjusted Wold's R criterion
1	0.704
2	0.132
3	0.954

By applying the randomisation test [16], it is possible to reduce the number of components to 16 latent variables.

The PoLiSh procedure gives interesting results, in the sense that there is a minimum of the difference of PRESS values for PoLiSh and for classical PLS for

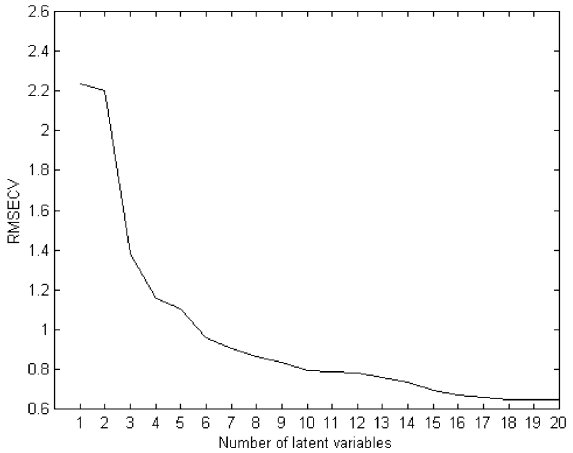

Fig. 6. RMSECV of the PLS model for dataset 1b (raw protein).

five latent variables (Fig. 7). It is then possible to reduce the complexity of the model from 18 (or 16) to 5 latent variables.

The Durbin–Watson profiles for the **p** vectors, **w** vectors, and *b*-coefficient vectors increase from the fifth latent variable (Fig. 8).

The calculation of the rotation angles between *b*-coefficient vectors shows that the transition between the fourth and the fifth latent variable produces a low angle. It means that these vectors point almost in the same direction and that only four components are enough to build the model (Fig. 9).

The PoLiSh procedure therefore indicates that it is possible to reduce the number of latent variables to five or even four latent variables but shows that there is no big difference between four and five.

The MCCV confirms these results. Once again, different sizes of n_v are tested ($n_v = 1$, 100, 175, 180, 185, and 190). Results are shown in Fig. 10.

These plots indicate clearly (especially for $n_v = 190$) that only five latent variables are necessary to build the PLS model. From $n_v = 1$ to $n_v = 190$, one can see the minimum of MCCV moving from 18 to 5.

Values of *R* for this dataset are shown in Table 2.

With the limit that was predefined (0.90), the number of latent variables that should be kept to build the model is 4. It is important to notice that the choice of the number of latent variables is given by the fact that the *R* criterion should be larger than 0.90 for the following latent variables, and leads here to 4 instead of one. A second remark concerns the comparison with previous methods (PoLiSh and MCCV) because the result is not exactly the same.

In this case, the PoLiSh method, the MCCV, and the *R* show that the PLS regression computed with the leave-one-out cross-validation includes too many components. It is preferable to reduce this number of components from 18 (or 16) variables to five or four variables to build a more robust model.

Validation using subset 2 shows that the RMSEP increases from 0.73 (for a model with 16 or 18 components) to 1.3 (for a model with four or five

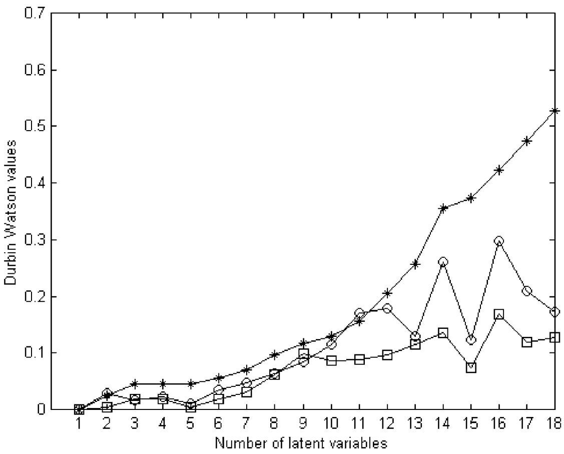

Fig. 8. Durbin–Watson profiles for the loadings **p** (–○–), weights **w** (–□–), and *b*-coefficient vectors (–∗–) of the PLS model for dataset 1b (raw protein).

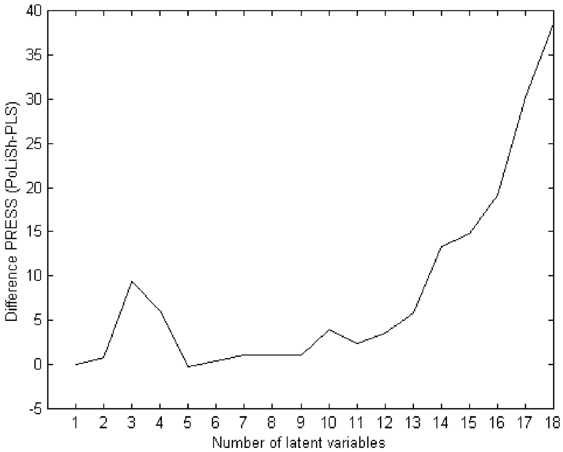

Fig. 7. Difference of PRESS values between PoLiSh and PLS models for dataset 1b (raw protein).

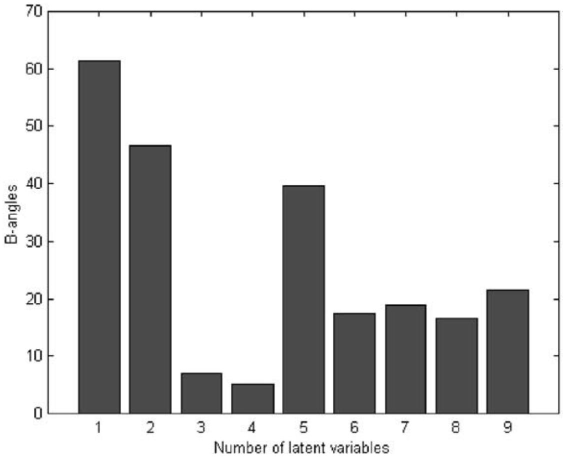

Fig. 9. Angles between adjacent *b*-coefficient vectors for dataset 1b (raw protein).

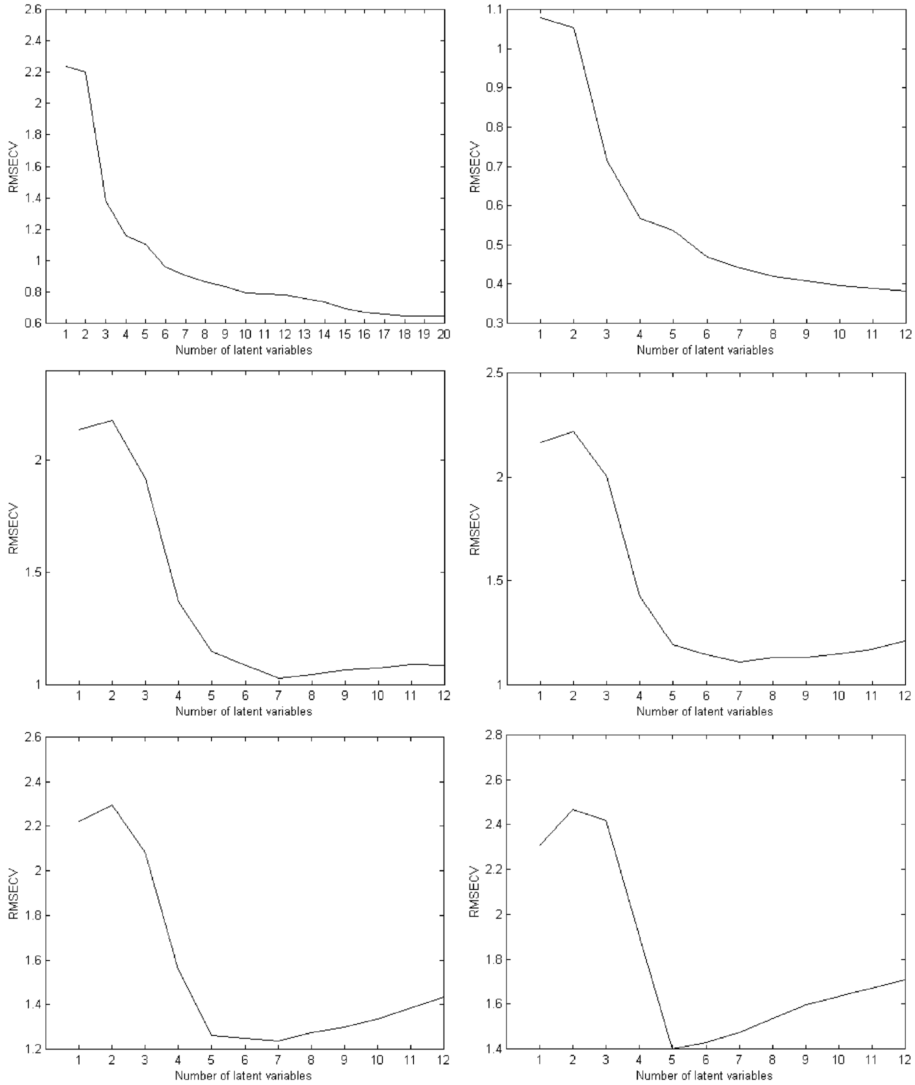

Fig. 10. MCCV plots: leave-n_v-out for dataset 1b (raw protein).

components). These values show that the prediction is not so much affected (even if the difference is more remarkable than in the case of the humidity) when decreasing considerably the number of components of the PLS model. The low number of components included in the model contributes to have a robust model.

Table 2
Values of adjusted Wold's *R* criterion for the dataset 1b

Number of latent variables	Adjusted Wold's *R* criterion
1	0.969
2	0.393
3	0.702
4	0.916

4.3. Dataset 2

The spectra used in this case are corrected with an offset correction in order to remove the baseline drift. The standard PLS with a leave-one-out cross-validation was first performed and the minimum of

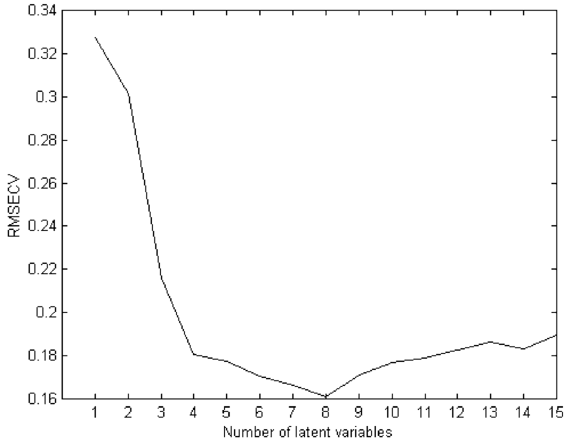

Fig. 11. RMSECV of the PLS model for dataset 2 (Kalivas data).

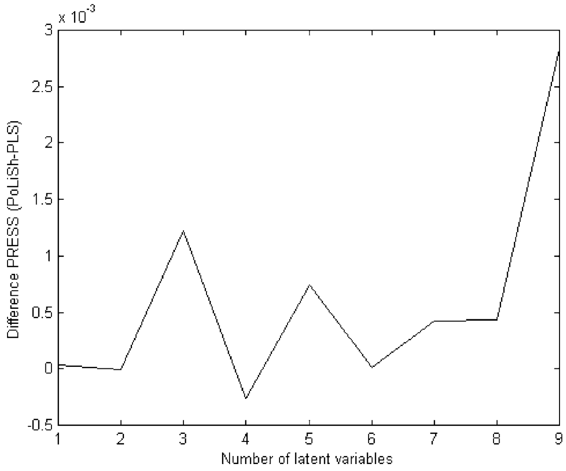

Fig. 13. Durbin–Watson profiles for the loadings **p** (–○–), weights **w** (–□–), and *b*-coefficient vectors (–∗–) of the PLS model for dataset 2 (Kalivas data).

RMSECV is obtained for eight latent variables (Fig. 11).

A randomisation test [16] was applied and shows that including four latent variables in the model gives the same performance as the model with eight latent variables.

Fig. 12 shows clearly that the Durbin–Watson values start to increase after the fourth value, indicating the inclusion of unstructured information with the fifth latent variable.

The plot of the difference of PRESS between PoLiSh and PLS is not as clear as the previous ones (Fig. 13). There is nevertheless a minimum for four latent variables, even if for two or six variables, there are also local minima.

The angle between three and four components is low and this means that three or four latent variables are enough to build the model (Fig. 14).

The MCCV gives again the same results as the PoLiSh PLS regression. Once again, different sizes of n_v are tested ($n_v = 1$, 5, 10, 15, and 20). Results are shown in Fig. 15.

These plots indicate clearly that only four latent variables are requested to build the PLS model. Again,

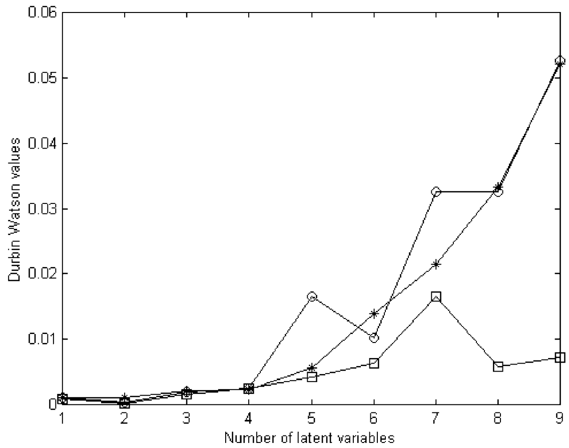

Fig. 12. Difference of PRESS values between PoLiSh and PLS models for dataset 2 (Kalivas data).

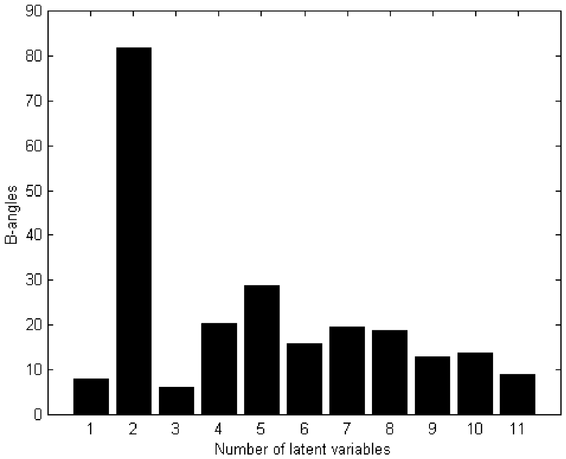

Fig. 14. Angles between adjacent *b*-coefficient vectors for dataset 2 (Kalivas data).

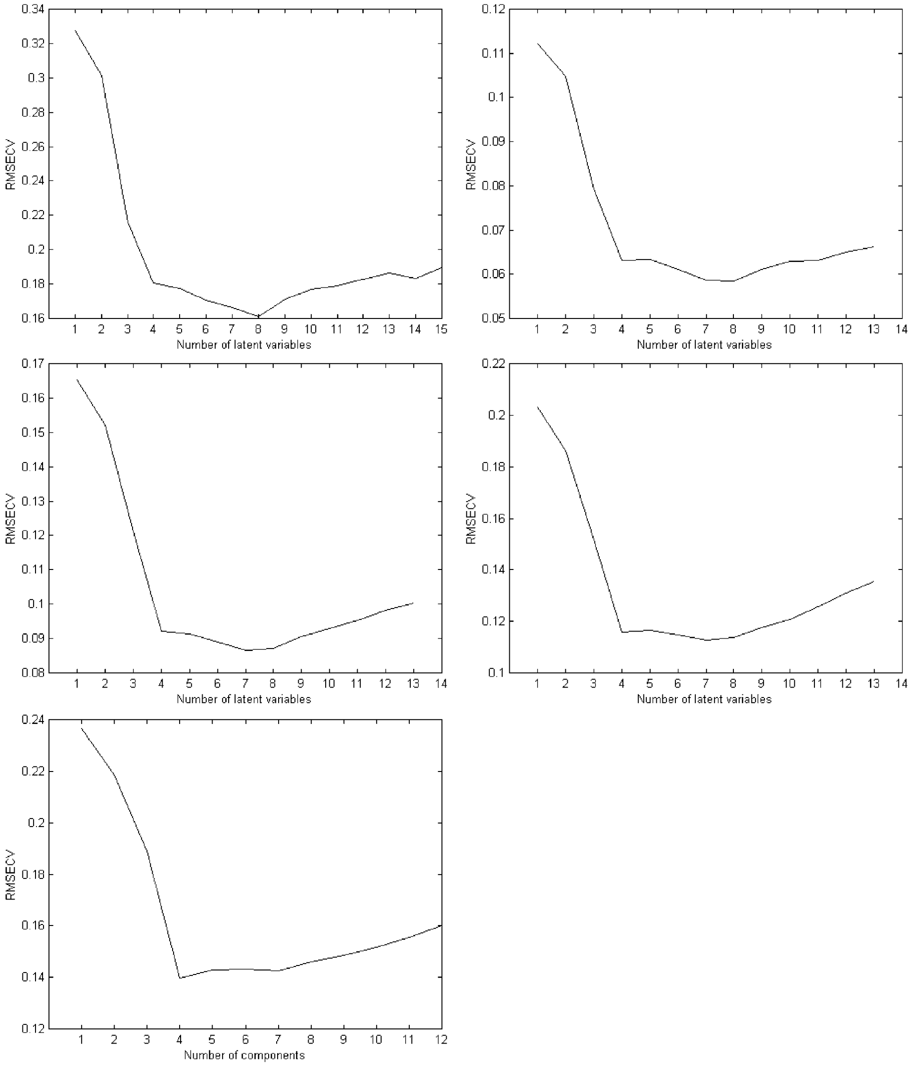

Fig. 15. MCCV plots: leave-n_v-out for dataset 2 (Kalivas data).

it should be added that it is may not be necessary to leave out 25 samples and that, for 20, a minimum is already visible.

Table 3 shows the values of R for the first latent variables.

R of step 4 (ratio of PRESS for five latent variables and PRESS for four latent variables) is larger than the limit 0.90 and leads to the conclusion that only three latent variables are enough to build the model. It can be noticed that this result is also valid with the limit of 0.95.

This criterion confirms again the results obtained with the PoLiSh procedure and MCCV, and shows that the optimal complexity is four.

Table 3
Values of adjusted Wold's R criterion for the dataset 2

Number of latent variables	Adjusted Wold's R criterion
1	0.846
2	0.514
3	0.699
4	0.963

External validation was not performed in this case since there were no subsets available. Due to the particular structure of the original dataset (two clusters), the calibration dataset constituted of samples coming from one cluster and was not divided in two parts.

5. Discussion

For the three datasets that are studied in this paper, the PoLiSh PLS regression, the Monte Carlo Cross-Validation, and the adjusted Wold's *R* criterion are shown to be useful tools to determine the complexity of a PLS model. The three different ways of choosing the complexity of a PLS model were compared to the leave-one-out cross-validation method, which is very common and widely used. It can be seen that the number of PLS components to be used is considerably reduced with the PoLiSh smoothed regression, the Monte Carlo Cross-Validation, and the adjusted Wold's *R* criterion. This reduction of the number of variables may avoid overfitting and lead to a more robust model. It is, however, not evident that the increase in robustness will lead to better precision in estimation. To ascertain this, it would really require more complete studies in which validation would be performed with a completely independent prediction set for several cases and with different structures and qualities of data.

Another point of discussion concerns the Monte Carlo Cross-Validation. It is explained in the Results section that the best results of the cross-validation are obtained by leaving out a large number of samples. In the studied cases, 85.4%, 92.7%, and 50% of the whole calibration set, respectively, are left out during the calibration. One might wonder if making a calibration model with so few samples is representative of the structure of the data and enough to build a correct model. However, one can see that the results of MCCV are comparable to the results of the PoLiSh regression and the adjusted Wold's *R* criterion, and that the same complexity is obtained. Nevertheless, the maximal ratio of objects in the calibration set to objects in the validation set is a point that should still be clarified and studied to better understand and interpret results of MCCV. A possibility might be an intermediate solution with 50% in calibration and validation set and application of the adjusted Wold's *R* criterion to the result.

References

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

[2] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.

[3] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1202–1208.

[4] Q. Xu, Y. Liang, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.

[5] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.

[6] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.

[7] D.N. Rutledge, A. Barros, I. Delgadillo, Anal. Chim. Acta 446 (2001) 281–296.

[8] N.R. Draper, H. Smith, Applied Regression Analysis, Wiley, New York, 1981.

[9] V. Centner, O.E. de Noord, D.L. Massart, Anal. Chim. Acta 376 (1998) 153–168.

[10] S. Wold, Technometrics 24 (1978) 397–405.

[11] A. Savitsky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627–1639.

[12] J. Steiner, Y. Termonia, J. Deltour, Anal. Chem. 44 (1972) 1906–1909.

[13] D.L. Massart, B.G.M. Vandeginste, et al., Chemometrics: A Textbook, vol. 2, Elsevier, Amsterdam, 1988.

[14] I Ruisánchez, F.X. Rius, S. Maspoch, J. Coello, T. Azzouz, R. Tauler, L. Sarabia, M.C. Ortiz, J.A. Fernández, D. Massart, A. Puigdomènech, C. García, Chemom. Intell. Lab. Syst. 63 (2) (2002) 93–105.

[15] J.H. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.

[16] H. van der Voet, Chemom. Intell. Lab. Syst. 25/28 (1994/1995) 313–323, 315.