Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

A Backward Variable Selection method for PLS regression (BVSPLS)

Juan Antonio Fernández Pierna, Ouissam Abbas, Vincent Baeten, Pierre Dardenne*

Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department, Chaussée de Namur n°24, 5030 Gembloux, Belgium

ARTICLE INFO

Article history:
Received 29 July 2008
Received in revised form 2 December 2008
Accepted 2 December 2008
Available online xxx

Keywords:
Variable selection
Q2 ~~PLS~~
Regression
~~NIR~~

ABSTRACT

Variable selection has been discussed in many papers and it became an important topic in areas as chemometrics and science in general. Here a backward iterative step-by-step wrapper method is proposed using PLS. The root-mean-square error of prediction (RMSEP) for an independent test set is used as selection criterion to quantify the gain obtained using the selected range of variables. The method has been applied to different data sets and the results obtained revealed that one can improve or at least keep constant the prediction performances of the PLS models compared to the full-spectrum models. Moreover with the advantage that the number of variables is reduced driving to an easier interpretation of the relationship between model and sample composition and/or properties. The aim is not to compare to other variable selection methods but to show that a simple one can improve or at least keep constant the prediction performances of the PLS models by using only a limited number of variables.

1. Introduction

In spectroscopy, thanks to the modern techniques of analysis, objects described by a large number of variables (i.e. absorbance at defined wavelengths or wavenumbers) can be easily measured in a short time. Then, multivariate chemometric techniques become necessary in order to extract the most relevant information from such data. PLS is probably the most common multivariate technique used for this extraction.

Often, PLS is trained with the full-spectrum region including variables that are unrelated to the variation of the response (e.g. concentration); in this case there are risks of obtaining an overfitted model, i.e. a model with a small residual error but poor prediction ability. However training PLS with a range of selected spectral variables should allow the informative part of the spectrum to be extracted and modelled rapidly and to discard the other spectral variables that are redundant or not correlated to the response. In this case when too few variables are kept, one can reach the case of an underfitted model, i.e. an overestimation of the error variance.

The goal of variable selection is to obtain a small set of variables that gives the generalization ability better or at least equivalent to the original set of variables [1]. The main benefits of variable selection are the improvement of data visualization and data understanding, the reduction of measurement requirements and training as well as utilization times by defying the curse of dimensionality to improve prediction performance (better robustness) [2].

Variable selection in PLS consists in obtaining latent variables based on a reduced subset of variables in such a way that they retain as much as information as the latent variables obtained with the whole spectra (i.e. with all the variables) with a minimum error in prediction. Variable selection methods are usually grouped into two categories: filter (or univariate) methods and wrapper (or multivariate) methods. Nagatani and Abe [1] define the filter methods as those that estimate the generalization ability for the selected subset of input variables by some indirect estimator that relies solely on properties of the data (for instance the correlation coefficient) and the wrapper methods as those that directly estimate the generalization ability using a learning algorithm, i.e. the regression equation.

As explained by Talavera [3], wrapper methods are a good alternative in supervised learning problems, since by employing the learning algorithm to evaluate the selected subset of input variables they have into account the particular biases of the algorithm. However, this kind of methods requires a high computational cost.

Classical wrapper methods include also sequential techniques, i.e. the use of forward selection and its counterpart backward selection or a combination of both. Forward selection methods start with one variable and then building up a model by adding variables whereas backward selection methods start with all available variables, and then removing the unnecessary variables step-by-step. Both methods have their drawbacks. When working with forward selection, weaker subsets are obtained because the importance of a certain variable is not assessed in the context of other variables that are not included yet. When working with backward selection, sometimes the variables that are removed would be significant when added to the final reduced models.

Q1 * Corresponding author. Tel.: +32 81 62 03 50; fax: +32 81 62 03 88.
E-mail address: dardenne@cra.wallonie.be (P. Dardenne).

0003-2670/$ – see front matter © 2008 Published by Elsevier B.V.
doi:10.1016/j.aca.2008.12.002

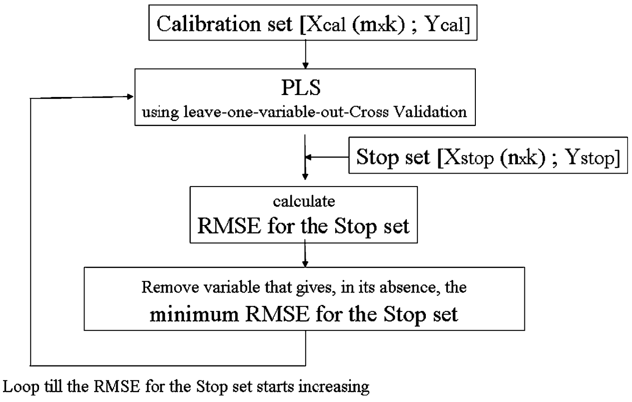
Calibration set [Xcal (mxk) ; Ycal]
PLS
using leave-one-variable-out-Cross Validation
Stop set [Xstop (nxk) ; Ystop]
calculate
RMSE for the Stop set
Remove variable that gives, in its absence, the
minimum RMSE for the Stop set
Loop till the RMSE for the Stop set starts increasing

Fig. 1. BVSPLS algorithm.

Several wrapper selection methods can be found in the literature [4,5]. In this paper, we apply a backward iterative step-by-step PLS-oriented method for the selection of spectral variables (the 'Backward Variable Selection method for PLS regression' or BVSPLS). A similar algorithm has been proposed by Abe [6] using Support Vector Machines [7] and applied to some benchmark data sets. In his paper, Abe showed that the number of variables can be drastically reduced without deteriorating the generalization ability of the constructed models. Here a similar algorithm is applied but working with PLS as regression technique. The objective is to build a correct model with only few variables and to show that we can get similar results or even improve its prediction performance as well as an improvement in the stability of the models and a clear interpretation of the relationship between model and sample composition and/or properties.

2. Theory

The idea of the BVSPLS method is simple and the algorithm is defined in Fig. 1. Here it is explained for NIR spectroscopic data but it could be generalized to other kind of data.

It is a wrapper method that relies on PLS as regression algorithm and works in an iterative way based in backward selection. We use the root-mean-square error of prediction (RMSEP) as goodness measure, i.e. as selection criterion to quantify the gain obtained using the selected range of variables. In the first step the method starts by fitting a model with all the initial set of variables and then a leave-one-variable-out (deleting one variable at a time) is performed. In such a way and in order to avoid overfitting, each time a variable is removed, a new model is constructed using a training set (cal) and applied to an independent dataset (stop set) calculating its RMSEP, used as selection criterion. The minimum RMSEP is selected indicating that the particular variable, when it is not included in the construction of the model, drives to an improvement of the model performance. Then the selected variable is considered as not significant and then removed. The procedure is repeated again *X* (number of variables) times by successively re-fitting reduced models. All the RMSEP are kept and a graphic showing each RMSEP versus the number of iterations can be constructed. Based on this plot, the minimum RMSEP is selected. And the final model is constructed with the variables which remain at the stage when the RMSEP starts to increase. In order to test the ability of the constructed models to generalize, a set of independent data is needed (test set).

3. Experimental

3.1. Data

Different samples have been used in this study. In all the cases the datasets consist on NIR spectra of different agronomical products in order to model certain properties. The first data set consists on complete feed spectra in order to model the fat, protein, moisture and starch content. The second set of data consists on maize samples that have been cut and dried (not fermented) to model the protein, starch, cellulose and ndf (neutral detergent fibre) content. In all the cases the data have been pre-processed by SNV (Standard Normal Variate) [8], Detrend [9] and Savitsky–Golay (SG) first derivative (9:2:1) [10]. SNV corrects spectra for spectral noise and background effects which cause baseline shifting and tilting; Detrend removes linear trends and the SG first derivative removes baseline offsets.

3.2. Software

All computations, chemometric analyses, and graphics were executed with programs developed in Matlab v7.4 (The Mathworks, Inc., Natick, MA, USA). PLS calibrations were derived using the SIMPLS algorithm [11] included in the PLS Toolbox (Eigenvector Research, Inc., Manson, WA, USA).

4. Results

The method has been applied to the different data sets in order to compare the prediction ability before and after variable selection for the models constructed for each of them. One of the simplest mean of avoiding overfitting is to split the data into three subsets: a calibration set, used to build the model, a stop set, used to optimize the model (minimize the RMSEP useful for the selection of variables) and a test set, used to validate in a completely independent way the constructed model. This test set is needed because the stop set is used to decide when to stop training, and thus the constructed model is no longer entirely independent of the stop set. The split of the data has been performed using the duplex method [12], which is a modification of the Kennard–Stone method. In the first step, the two points which are furthest away from each other are selected for a first set. From the remaining points, the two objects which are furthest away from each other are included in a second set. In the

Table 1
Q3 Results obtained for all the datasets with and without application of BVSPLS.

Samples	Property	Number of samples			Model dimensionality	RMSEP test samples		Number of original variables	Number of retained variables
		Cal	Stop	Test		Before BVSPLS	After BVSPLS		
Complete feed	Fat	362	518	3043	8	3.03	1.16	700	58 (8%)
Complete feed	Protein	336	873	6182	3	1.68	1.57	700	62 (9%)
Complete feed	Moisture	337	1010	5920	4	0.84	0.70	700	224 (32%)
Complete feed	Starch	156	200	1000	9	2.52	2.03	700	226 (32%)
Maize	Protein	272	338	3869	3	0.63	0.54	700	260 (37%)
Maize	Starch	165	331	4558	3	2.84	2.41	700	122 (17%)
Maize	Cellulose	175	355	3526	9	1.37	1.35	700	110 (16%)
Maize	Neutral detergent fibre	264	279	3354	6	2.44	2.99	700	95 (14%)

Please cite this article in press as: ~~T.A.~~ Fernández Pierna, et al., A Backward Variable Selection method for PLS regression (BVSPLS), Anal. Chim. Acta (2008), doi:10.1016/j.aca.2008.12.002

Table 2

Samples	Property	Range property (min–max)	Skill score test samples		*R*-square	
			Before BVSPLS	After BVSPLS	Before BVSPLS	After BVSPLS
Complete feed	Fat	0.66–33.87	0.447	0.7885	0.8616	0.9684
Complete feed	Protein	8.04–61.74	0.7832	0.7978	0.954	0.9673
Complete feed	Moisture	1.65–17.94	0.5782	0.6487	0.8395	0.8811
Complete feed	Starch	0.0009–64.76	0.5902	0.6714	0.9242	0.9473
Maize	Protein	4.02–13.67	0.4797	0.5562	0.7364	0.809
Maize	Starch	0.001–53.17	0.7121	0.756	0.936	0.9438
Maize	Cellulose	9–38.56	0.6559	0.6602	0.883	0.886
Maize	Neutral detergent fibre	24.25–67.30	0.6231	0.5374	0.866	0.837

third step, the remaining point which is furthest away from the two previously selected for the first set is included in that set. Following the same procedure, points are added alternately to each set [13]. This splitting has been performed after removing of all the extreme samples that have been included by default in the test set.

In all cases, two models, before and after variable selection, have been constructed and applied to the independent test set. Table 1 shows a summary of the results obtained for all the datasets. The first and second columns represent the kind of samples used and the property to model, the three next columns the number of samples used as calibration, stop and test sets respectively for each data, the next column represents model dimensionality and then, the next two columns show the RMSEP for the independent test set before and after variable selection respectively. Last columns show the number of original variables as well as the retained number of variables including the percentage. In order to have more information concerning model prediction, another parameter has been used, the skill score. The skill score is defined as one minus the RMSEP divided by standard deviation of the observed data. This function interprets model predictability using residual error and observed variability in the data. A skill score of 1 means a perfect fit and a skill score less than 0 means that the model error is larger than the variability of the data. The results are shown in Table 2.

The first example is based on a NIR data set of complete feed spectra to model the fat content. The data set contains 3923 samples measured at 700 wavelengths (1100:2:2498 nm). As previously explained the data set has been split into a calibration (362 samples), stop (518 samples) and test set (3043 samples). The classical PLS model on the raw data led to a RMSEP for the test set of 3.03. The application of the Backward Variable Selection for PLS regression method drove to the selection of 58 variables (instead of 700, i.e. only around 8% of the variables are retained) as being important for the determination of the fat content and a RMSEP of 1.16. Fig. 2 shows the evolution of the RMSEP for all the subsets versus the number of iterations, the process is stopped when the minimum RMSEP is obtained for the stop set as indicated in the figure. A large bias can be observed in the figure mainly between the test and the other data sets. This can be easily explained by the fact that all the extreme samples have been included in the test set previous the duplex selection.

Fig. 3 shows the variables selected after BVSPLS. This method allows obtaining dispersed variables rather than regions, but as it can be observed in the figure, these dispersed variables could be grouped into spectroscopically logical regions, which are easier to interpret [14]. The broad band between 1100 and 1200 nm presents intensities at 1134, 1158, and 1176 nm associated to the second overtones of C–H while the peak at 1418 nm and the shoulder at 1352 nm correspond to C–H combinations. Smaller peak observed at 1564 nm can be attributed to N–H stretch first overtones. The region at 1686 nm, showing a shoulder at 1712 nm and very weak peaks at 1752 and 1782 nm, is generally associated to the C–H stretch first overtones [15]. An intense peak at 1900 nm is observed. It can be associated to the combination of O–H and C–O stretch or C=O stretch second overtones.

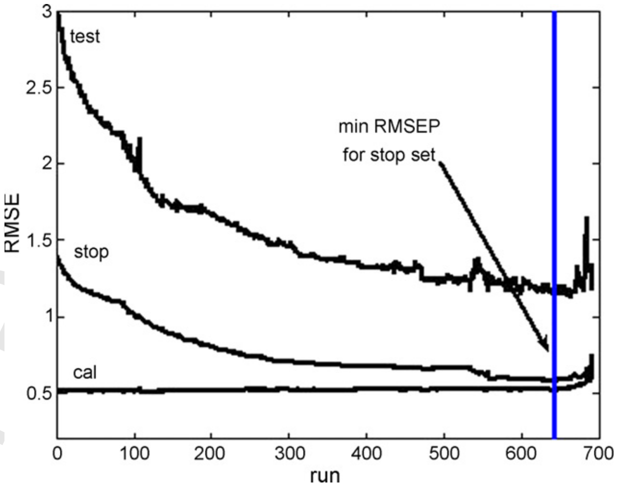

Fig. 2. Evolution of the RMSE for all the subsets for the complete feed when measuring the fat content.

It is important to note that the combination band of OH associated to water usually observed in near infrared spectra of feed shows intensity at 1940 nm, and it was indicated by the programme as a retained variable.

The region above 2000 nm is characterized by combination bands of different groups. Band at 2044 nm represents a combination of N–H stretch and amide bond while the band pointed at 2258 nm indicates the combination band of O–H characteristic of starch [15]. In addition, intensities at 2300, 2340, and 2448 nm can be affected to C–H combination bands [16].

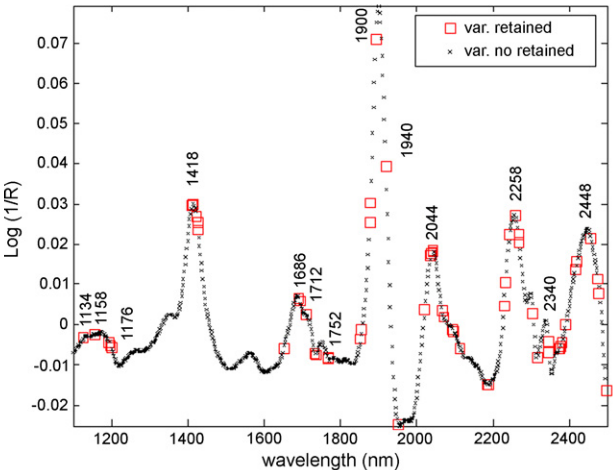

Fig. 3. Selected variables by backward selection for the compound feed data set for the fat content.

Please cite this article in press as: ~~T.A.~~ Fernández Pierna, et al., A Backward Variable Selection method for PLS regression (BVSPLS), Anal. Chim. Acta (2008), doi:10.1016/j.aca.2008.12.002

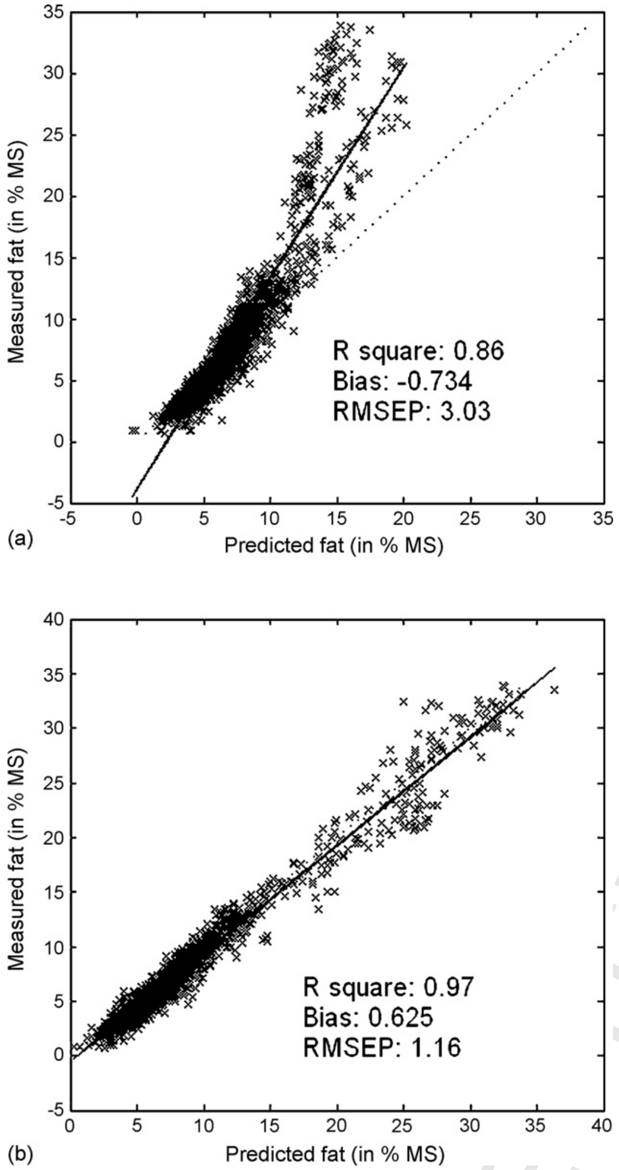

Fig. 4. Prediction of fat content for the compound feed data set. Results for the test set (a) before variable selection and (b) after BVSPLS.

A model constructed for the fat content in the complete feed using only the retained variables, indicated in Fig. 3, decreases the RMSEP to 1.16 for the test set. Fig. 4 shows, for the same data set, the prediction results for the test set before (4a) and after (4b) variable selection. The same conclusions can be obtained when looking at the skill score values. The application of an *F*-test [17] (not shown) at the 95% confidence level to the RMSEP and the skill scores before and after variable selection have shown that the results after variable selection are significantly better than before selection for this property.

For the rest of the properties for the complete feed also an improvement in the error is obtained as it can be observed in Tables 1 and 2. In these cases, an *F*-test has shown that there are no significant differences in both the RMSEP and the skill score at the 95% confidence level. Then, in all cases similar results are obtained with a 20% of number of retained variables in average.

The second example is based on a NIR data on maize samples. As in the previous data set an average of 21% of number of retained variables is obtained. These dispersed variables grouped into regions, drive to similar spectroscopically logical conclusions as for the complete feed. For all the studied properties, except for the ndf, also an improvement in the error, expressed as RMSEP or skill score, is obtained. An *F*-test applied has shown no significant differences before and after variable selection in terms of error. But, in all cases and in general, a mean reduction in the number of variables of 79% is obtained, which is useful for an easy interpretation of the calibration models.

5. Conclusion

The objective of this technique was mainly to construct and select subsets of variables that are useful to build a good predictor, i.e. to construct a suitable model with only few variables. In all the cases studied, except for the ndf in maize, the RMSEP and the skill score we obtain improved. In most of the cases; there are no significant differences in the RMSEP or the skill values at the 95% confidence level, but in all the cases a mean reduction of variables about 79% is obtained.

The splitting of the data sets into calibration, stop and test sets was done to emphasize the effect of the variable selection on the model performance. The variable selection is all the more important that the number of samples is low.

With this study we prove that one can improve or at least keep constant the prediction performances by providing correct predictors as well as an improvement in the stability of the PLS models and an easier interpretation of the relationship between model and sample composition and/or properties. This shows the possibility of constructing fast spectrometers based on a reduced number of variables aiming to a reduction of training and utilization times to be used, for instance, in a conveyer belt. Moreover, the fact of grouping the different selected variables in regions allows an easy chemical interpretation of the spectra. The main advantage is a better accuracy in prediction mode (routine analyses) especially when the calibration set is limited (few samples).

The aim of this paper was not to compare different feature selection methods. This kind of studies has been done intensively at the literature. With this paper, our objective was to demonstrate that the predictive ability of the models obtained with the wavelengths selected by the algorithm is in most of the cases better or at least similar than the predictive ability of the full spectrum. In a further study this method could be compared to other variable selection techniques as genetic algorithms and applied to different data sets. The next step is the application in the framework of the SAFEED-PAP EU project (FOOD-CT-2006-036221) aiming the species-specific determination in feed using NIR microscopic data.

References

[1] T. Nagatani, S. Abe, Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN 2007), Orlando, FL, USA, 2007, pp. 1540–1545.
[2] I. Guyon, A. Elisseeff, Journal of Machine Learning Research 3 (2003) 1157–1182.
[3] L. Talavera, in: A. Fazel Famili, Joost N. Kok, José María Peña, Arno Siebes, A. J. Feelders (Eds.), Advances in Intelligent Data Analysis VI, 6th International Symposium on Intelligent Data Analysis, IDA 2005, Madrid, Spain, September 8–10, 2005, Proceedings. Lecture Notes in Computer Science 3646 Springer 2005, ISBN 3-540-28795-7, pp. 440–451.
[4] G.H. John, R. Kohavi, K. Pfleger, in: M. Kaufmann (Ed.), Machine Learning: Proceedings of the Eleventh International Conference, New Brunswick, NJ, USA, 1994, pp. 121–129.
[5] D.J. Stracuzzi, P.E. Utgoff, Randomized variable selection, Journal of Machine Learning Research 5 (2004) 1331–1362.
[6] S. Abe, ESANN'2005 Proceedings—European Symposium on Artificial Neural Networks, Bruges (Belgium), April 27–29, 2005.
[7] J.A. Fernández Pierna, V. Baeten, A. Michotte Renier, R.P. Cogdill, P. Dardenne, Journal of Chemometrics 18 (2004) 341–349.
[8] A.M.C. Davies, T. Fearn, Spectroscopy Europe 19 (4) (2007) 24–28.
[9] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Applied Spectroscopy 43 (1989) 772–777.
[10] A.M.C. Davies, Spectroscopy Europe 19 (2) (2007) 32–33.
[11] S. De Jong, Chemometrics and Intelligent Laboratory Systems 18 (1993) 251–263.
[12] R.D. Snee, Technometrics 19 (1977) 415–428.

Please cite this article in press as: ~~T.A.~~ Fernández Pierna, et al., A Backward Variable Selection method for PLS regression (BVSPLS), Anal. Chim. Acta (2008), doi:10.1016/j.aca.2008.12.002

[13] R. De Maesschalck, F. Estienne, J. Verdú-Andrés, A. Candolfi, V. Centner, F. Despagne, D. Jouan-Rimbaud, B. Walczak, D.L. Massart, S. de Jong, O.E. de Noord, C. Puel, B.M.G. Vandeginste, Internet Journal of Chemistry 2 (1999) 19.

[14] R. Leardi, A. Lupiáñez González, Chemometrics and Intelligent Laboratory Systems 41 (1998) 195–207.

[15] B.G. Osborne, T. Fearn, Near Infrared Spectroscopy In Food Analysis, Longman Scientific & Technical, Harlow, UK, 1986.

[16] S.J. Lister, M.S. Dhanoa, Journal of Near Infrared Spectroscopy 5 (1997) 99–111.

[17] C. Mello, R.J. Poppi, J.C. de Andrade, H. Cantarella, Analyst 124 (1999) 1669–1674.

Please cite this article in press as: ~~T.A.~~ Fernández Pierna, et al., A Backward Variable Selection method for PLS regression (BVSPLS), Anal. Chim. Acta (2008), doi:10.1016/j.aca.2008.12.002