Received: 4 October 2011 ■ *Revised: 15 April 2012* ■ *Accepted: 15 May 2012* ■ *Publication: 24 May 2012*

JNIRS JOURNAL OF NEAR INFRARED SPECTROSCOPY

Prediction error improvements using variable selection on small calibration sets—a comparison of some recent methods

Stein Ivar Øvergaard,[a,b,*] Juan Antonio Fernández Pierna,[c] Vincent Baeten,[c] Pierre Dardenne[c] and Tomas Isaksson[d]

[a]The Norwegian Institute for Agricultural and Environmental Research, 2849 Kapp, Norway. E-mail: stein.ivar.overgaard@bioforsk.no

[b]The Norwegian University of Life Sciences, Department of Mathematical Sciences and Technology, 1430 Ås, Norway

[c]Walloon Agricultural Research Centre, Valorisation of Agricultural Products Department, Food and Feed Quality Unit, Gembloux, Belgium

[d]The Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, 1430 Ås, Norway

Variable selection provides useful information about the most important predictors in the dataset, information which is not always available at the beginning of an analysis. Two recent variable selection methods, backward variable selection for partial least squares (BVSPLS) and powered partial least squares (PPLS), were compared against each other and against forward stepwise selection (FSS) and full spectrum partial least squares (PLS) in terms of their ability to produce accurate prediction models in NIR spectroscopy data. All four regression methods were studied using three different NIR datasets. PPLS and BVSPLS gave good prediction results in all three datasets, even with a very limited number of calibration samples available (<40). All methods gave similar prediction results when the number of calibration samples was higher (>150). PPLS gave the best predictive performance of all methods and also gave the selections of variables that were most easily assigned to specific chemical bonds. Hence, the PPLS models were more easily interpretable than the other models. This study quantifies differences between the two recent variable selection methods as well as the differences between recent methods and more established methods. Moreover, if the number of calibration samples can be reduced through variable selection, the labour and cost associated with wet chemistry reference methods can be reduced accordingly.

Keywords: BVSPLS, FSS, NIR, PLS, PPLS, variable selection

Introduction

Regression is probably the most widely studied and applied statistical analysis method in the chemometric literature. The aim is to develop models which can be used to predict properties of interest based on measurements of the chemical system, such as spectroscopic data. Multivariate calibration techniques such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares regression (PLS)[1] can then be used to compute a mathematical model. It correlates the multivariate measurement (spectrum) to the concentration of the analyte of interest and such a model can be used to predict the concentrations of new samples.

When the number of measured predictor variables is large and it is not known beforehand which specific predictors are

ISSN: 0967-0335
doi: 10.1255/jnirs.966

most influential on the responses, selection of variables could be feasible. Variable selection tries to identify a sub-set of variables that still possess sufficient features to build a robust regression model. Moreover, due to a number of practical and statistical reasons (for example, to avoid collinearity, improve predictive ability and improve interpretability), a large set of variables should be reduced to a smaller, more manageable set. The main goal of any variable selection technique is to obtain a small sub-set of variables that gives a model with the prediction and generalisation abilities better than, or at least equivalent to, a model based on the original set of variables. Variable selection in regression is a difficult part of model building because the number of sub-sets to be considered grows exponentially with the number of candidate variables. The advantages of variable selection are the exclusion of irrelevant and redundant variables leading to better signal-to-noise ratio, better data visualisation and model interpretability, reduction of measurement requirements as well as increased prediction accuracy and precision. Subsequently, these properties could induce the development of cheaper instruments, cheaper analysis as well as faster prediction models. Moreover, an increase in the model robustness can be achieved by the application of a variable selection technique. One possible drawback of variable selection is that certain outlier detection methods may be more difficult to undertake. Numerous methods have been developed for variable selection such as varieties of sub-set selection methods,[2,3] stepwise regression, jack-knife or bootstrapping algorithms,[4–7] evolutionary algorithms,[8] genetic algorithms[9–12] and thresholding algorithms.[13] Recently, a backward variable selection method for PLS regression (BVSPLS) has been proposed.[14] Another relatively recent method is the powered partial least squares,[15] which is a generalisation of the traditional non-linear iterative partial least squares (NIPALS) algorithm. PPLS can also be used for variable selection purposes. The development of new variable selection methods is constantly evolving and the need for comparative studies is raised. There are several studies aimed at comparing various methods,[16] but due to the ongoing development of new methods, comparative studies will always be needed.

The objective of this paper is to compare BVSPLS and PPLS and with some widely used methods. The most established, simplest and most pragmatic method for variable selection is forward stepwise selection of variables (FSS). There are several examples in the literature where varieties of FSS have been used as a reference method.[2,16] Moreover, the FSS algorithm is implemented in a multitude of data analysis software and is, hence, widely used. For this reason, FSS was our choice for a reference method for this comparative study. Moreover, the traditional PLS solution, without any variable selection, should also be included as a reference method in order to address the question whether variable selection itself has a positive effect for the predictive ability of the models. Assuming that prediction error is the main concern, why would analysts bother with variable selection in the first place? This question was our main rationale for comparing models against full spectrum PLS. In order to validate the feasibility of the methods, we chose to compute prediction models based on small datasets. Decreasing the number of calibration samples while still retaining good model performance can potentially reduce reference sampling costs.

Materials and methods

Datasets, pre-processing and sample selection

Three large datasets used in Fernández Pierna *et al.*[14] were also used in this study: fat/feed (Feed), fibre/maize (Maize I) and protein/maize (Maize II). All datasets had a spectral range from 100 nm to 2498 nm, with every second wavelength removed, thus containing 700 variables each. See Table 1 and Fernández Pierna *et al.*[14] for more details. There were some duplicate samples in the datasets and, in order to achieve a proper validation, duplicate response variable values and their corresponding spectra were removed from the datasets. Hence, the number of samples in each dataset (N) was reduced to 2721 for fat/feed, 2488 for Maize I and 1349 for Maize II. All datasets were pre-processed with the standard normal variate (SNV[17]) procedure. Splitting of the data into calibration and validation sets was done with the DUPLEX algorithm.[18] This algorithm splits a dataset (i.e. the spectra) into two parts by means of a Euclidean distance measure. The algorithm goal is to create two datasets with homogeneous statistical properties for calibration and validation purposes. A sub-set of 200 samples from each dataset were selected with DUPLEX and reserved for calibration purposes. The remaining samples were allocated as the validation set. Since we chose to work with smaller calibration sets, the 200 selected samples were further decimated to 20 samples in 19 steps with the DUPLEX

Table 1. Overview of the datasets (calibration and validation set together).

Product	Constituent	Range[a] [w/w–%]	*STD*[a] [w/w–%]	No. of validation samples	S_{REF}[b]
Feed	Fat	0.660–33.9	5.07	2521	0.20
Maize I	Fiber	24.3–67.3	6.82	2288	0.60
Maize II	Protein	4.02–13.7	1.60	1149	0.20

[a]Standard deviation for calibration and validation set together
[b]Standard error of reference method

algorithm. The first step selected 190 samples out of the original 200 samples and the second step selected 180 samples. For each successive step, the number of selected samples was decreased by 10. Thus, 19 calibration sets (200,190,...,20 samples) and one validation set (all samples except the 200 calibration samples) were calculated from each main dataset.

Software

All analyses were performed using MATLAB (version R2007b, MathWorks Inc., USA, www.mathworks.com) with PLS Toolbox (version 4.2, Eigenvector Research Inc., USA, www.eigenvector.com). The BVSPLS and DUPLEX algorithms were implemented in MATLAB code by the authors. The PPLS algorithm was implemented in MATLAB by the authors based on the original code written by Ulf Indahl.[15]

Variable selection methods

Forward stepwise selection (FSS)

The FSS algorithm is a simple and widely used procedure for variable selection. Three different basic varieties of stepwise regression are commonly used: forward selection, backward elimination and stepwise method. Forward selection sequentially introduces new predictors into the model one at a time while the backward loop eliminates predictors one at a time from the current variable set. The stepwise method is a hybrid between forward selection and backward elimination. It starts as forward selection, but for each selection step it runs an elimination step to compute the need for deleting predictors. The algorithm uses a Fisher *F*-statistic in order to decide when variables should be removed or included. To construct the final prediction model, we used a PLS algorithm on the retained variables. We chose to set the inclusion and removal values so that the FSS algorithm selected 40–80 variables. The *p*-values for inclusion and removal were both set to 0.11 in order to let the FSS select approximately 40–80 variables in the used datasets. Selecting that many variables will almost certainly introduce some multicollinearity between the variables, but the PLS algorithm will handle this more robustly than, for instance, the least squares method in the MLR algorithm.

Backwards variable selection for PLS (BVSPLS)

The BVSPLS is a recently proposed method and is a backward elimination method. Unlike the other algorithms in this study, the BVSPLS needs three datasets. In addition to the usual calibration and validation set, a dedicated dataset (the stop-set) for decision on which variables to retain is needed. The first algorithm step is to compute a PLS model based on the full calibration dataset with all variables included. Consequently, one variable is removed each time the algorithm loop executes. For each loop execution, the root mean square error of prediction (*RMSEP*) of the stop-set, is computed. When all variables have been discarded, a plot of the *RMSEP* from the stop-set against the number of variables can be presented. The algorithm then chooses the number of variables corresponding to the minimum *RMSEP*. This sub-set of variables is then retained for the final model which is a traditional PLS algorithm. For further details, see Fernández Pierna *et al.*[14] In order to provide a stop to BVSPLS, we chose to let 10% of the calibration samples form the stop-set. Hence, for a 100 samples calibration set, 10 of the samples (selected with DUPLEX) were put in the stop-set.

Powered partial lest squares (PPLS)

The PPLS is a generalisation of the traditional NIPALS algorithm. Rather than optimising the covariance between the predictors and the response, the PPLS splits the covariance expression in the weight vector optimization criterion into a variance part and a correlation part. The user can then choose the weighting between the variance component and correlation component through an additional control parameter, gamma (γ). The algorithm can be used both for modelling and variable selection through the choice of γ. A γ value of 0.5 makes the PPLS solution equivalent to the traditional PLS solution, whereas values close to 0 or 1 makes the algorithm select variables based on predictor variance and correlation with the response, respectively. It is also possible to let the algorithm optimise the γ value within a predefined numerical range using an optimisation procedure that maximises the correlation between PPLS scores and the response. In this study, we chose to let the PPLS work with γ values optimised from the interval [0.99,1]. The algorithm hence focuses almost exclusively on the variables with strong correlation to the response and also possibly strong predictive ability. As suggested by Indahl,[15] the variables that had loading weights less than the relative numerical resolution of MATLAB (2.2204×10^{-17}) were discarded. See Indahl[15] for further details.

Selection of optimal number of PLS/PPLS factors

All methods tested in this study have the feature of latent variables. Hence, model complexity has to be selected by the user. To make the resulting models more comparable, we chose to perform the selection of latent variables just once for each combination of method and dataset. For each dataset, the model complexity was determined on the basis of the complete calibration set of 200 samples and the number of factors was held constant throughout the whole range of calibration sets. Selection of the number of PLS/PPLS components was carried out as a conservative chi-square test. The main idea is to consider the minimum mean square error of cross-validation (*MSECV*) as a realisation of the true model error variance, σ_0^2. Using the chi-square power function, an acceptance region for *MSECV* can be computed. The model with the fewest number of components that also have an *MSECV* inside the acceptance region is then selected as the final model. See Indahl[15] for further mathematical details. However, several numbers of factors for each model were computed and compared, but the differences between models were only modestly affected by the choice of number of factors.

Validation procedure

Each of the 19 calibration sets was used to construct a prediction model, which was used to predict the validation set, which was the same for all 19 models and for all four regression methods. We chose to use a dedicated validation set instead of cross-validation for comparison of the predictive ability of each model because several studies have pointed out that cross-validation can lead to severe over-fitting and over-optimistic estimation of the models diagnostic measures.[2,19,20] The test procedure was performed in the following way:

The number of PLS factors was determined by computing a model with the full calibration set of 200 samples. The same number of factors was used for every variable selection method.

Each variable selection algorithm was executed once on the 19 smaller calibration sets (200,190,...,20 samples) and predictions for the validation set samples were computed each time.

Based on these validation set predictions, the coefficients of determination (r^2) between measured and predicted constituents for each method were computed and reported.

Steps 2 to 3 were repeated for each main dataset (fat/feed, Maize I and Maize II).

Hence, r^2 for 20 to 200 calibration samples were obtained in a comparable way with 10 samples increments. Paired *t*-tests on the absolute value of the residuals were performed in the sense of Cederkvist *et al.*[21] and *p*-values from these tests are indicated in the results section.

Results

Figures 1-3 (upper plots) show the coefficients of determination between measured and predicted constituents from the validation-set predictions as a function of the number of calibration samples for the Feed, Maize I and Maize II samples, respectively. For all datasets, variable selection gave little or no prediction error improvement over the PLS algorithm for calibration sets larger than 60 samples (Feed, Figure 1), 120 samples (Maize I, Figure 2) and 160 samples (Maize II, Figure 3). For calibration datasets smaller than this, all variable selection techniques gave better predictive performance compared to full spectrum PLS ($p<0.05$). Especially for low sample numbers, the BVSPLS and PPLS gave better performance than both PLS and FSS.

In the Feed dataset (Figure 1), all models were stable at a high r^2 above 50 samples calibration sets. From 50 to 20 calibration samples, BVSPLS and PLS started to show a decrease in performance ($p<0.05$). The PPLS, in particular, had an advantage over the other methods for calibration sets smaller than 50 samples ($p<0.05$).

In the case of Maize I (Figure 2), the situation was similar in the sense that all solutions were stable at a relatively high level of explained variance for calibration sets larger than 120 samples. Here, the PPLS gave the best predictions for all datasets smaller than 120 samples with the BVSPLS slightly lower prediction ability. FSS had performance between PLS and PPLS/BVSPLS for calibration sets smaller than 120 samples.

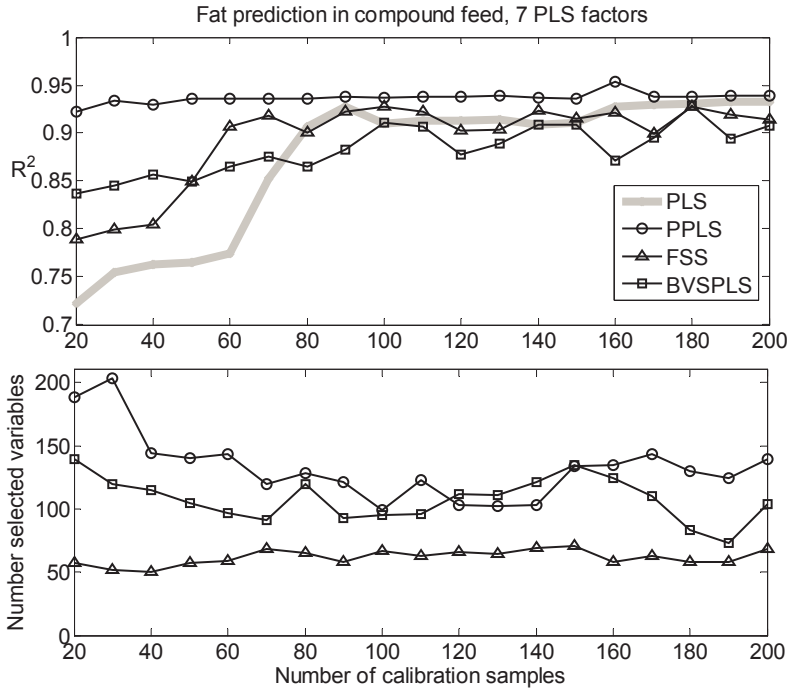

Figure 1. Upper plot: coefficient of determination (r^2) between measured and predicted constituents from the validation-set predictions of the Feed dataset as a function of the number of calibration samples for the fat content in feed mixture. The regression models used was full spectrum partial least squares (PLS), partial powered least squares (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Lower plot: number of selected variables for the three variable selection methods PPLS, FSS and BVSPLS.

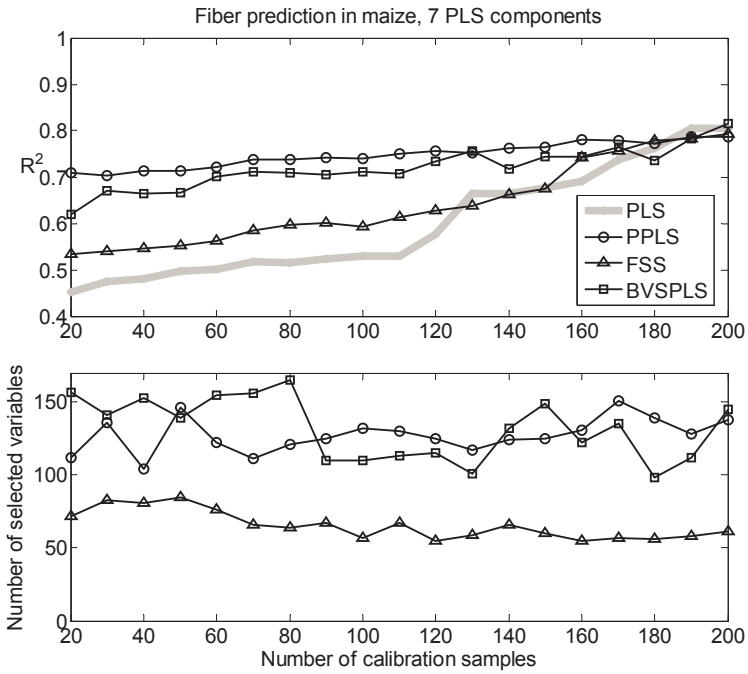

Figure 2. Upper plot: coefficient of determination (r^2) between measured and predicted constituents from the validation-set predictions of the Maize I dataset as a function of the number of calibration samples for the fat content in feed mixture. The regression models used was full spectrum partial least squares (PLS), partial powered least squares (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Lower plot: number of selected variables for the three variable selection methods PPLS, FSS and BVSPLS.

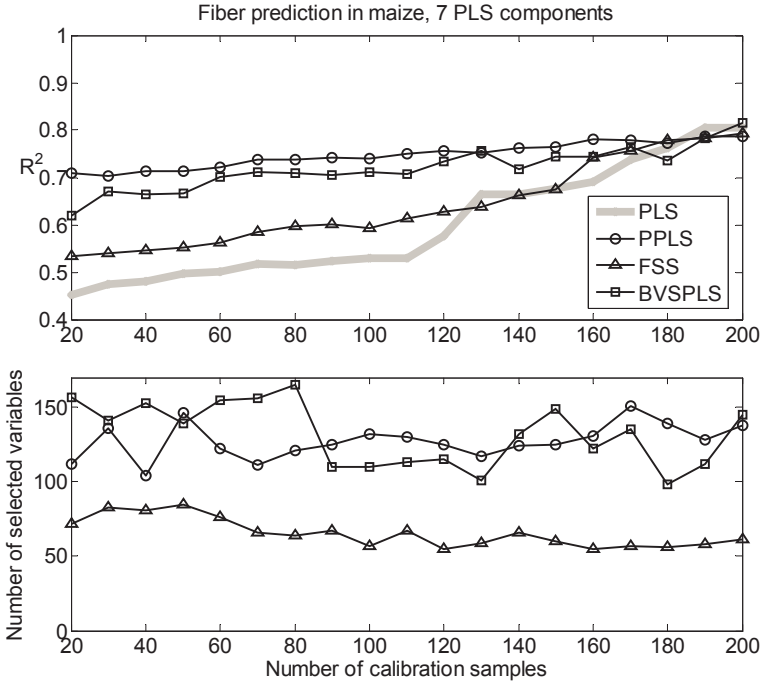

Figure 3. Upper plot: coefficient of determination (r^2) between measured and predicted constituents from the validation-set predictions of the Maize II dataset as a function of the number of calibration samples for the fat content in feed mixture. The regression models used was full spectrum partial least squares (PLS), partial powered least squares (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Lower plot: number of selected variables for the three variable selection methods PPLS, FSS and BVSPLS.

In the last dataset, the Maize II (Figure 3), all methods performed less well than in the first two datasets. Full spectrum PLS gave the lowest prediction ability for datasets smaller than 170 samples. For datasets larger than 170 samples, there were just small differences between the methods. For smaller datasets than 170 samples, the PPLS gave slightly better performance than FSS and BVSPLS.

In the fat dataset, PPLS models had an almost constant *RMSEP* of 1.8, whereas some of the PLS models had a *RMSEP* as high as 3.2. In the Maize I dataset, the *RMSEP* were approximately 2.7 for all methods at 200 samples calibration sets. As the calibration set became smaller, *RMSEP* values increased to 3.0 for PPLS and 4.2 for PLS. This behaviour was also present in the Maize II dataset. All methods had *RMSEP* of 0.7 at 200 calibration samples, but the *RMSEP* rapidly increased as the number of calibration samples decreased. The PPLS increased to 1.6 whereas PLS increased to 5.0

For all datasets, the PPLS and BVSPLS selected more variables for inclusion in the prediction models than the FSS. All variable selection methods differed markedly with regard to frequency of the selected variables. The PPLS and FSS selected some variables in the spectrum more often than other variables and especially PPLS gave a quite clear and structured image of which variables contribute positively to the prediction models. BVSPLS, however, selected variables more evenly spread throughout the spectrum (Figures 4–6).

To illustrate the improvement of variables selection (Figure 7), an example for 60 calibration samples of the Feed dataset is illustrated (Figure 1). We chose the PPLS and the PLS models for this case and plotted the predicted fat content data against the measured fat content from each method.

Discussion

Two recent methods, BVSPLS and PPLS, were compared against each other and against FSS and PLS methods. Several studies on variable selection are found in the literature, for example References 22–24, but we could not find comparative studies on variable selection methods similar to BVSPLS and PPLS.

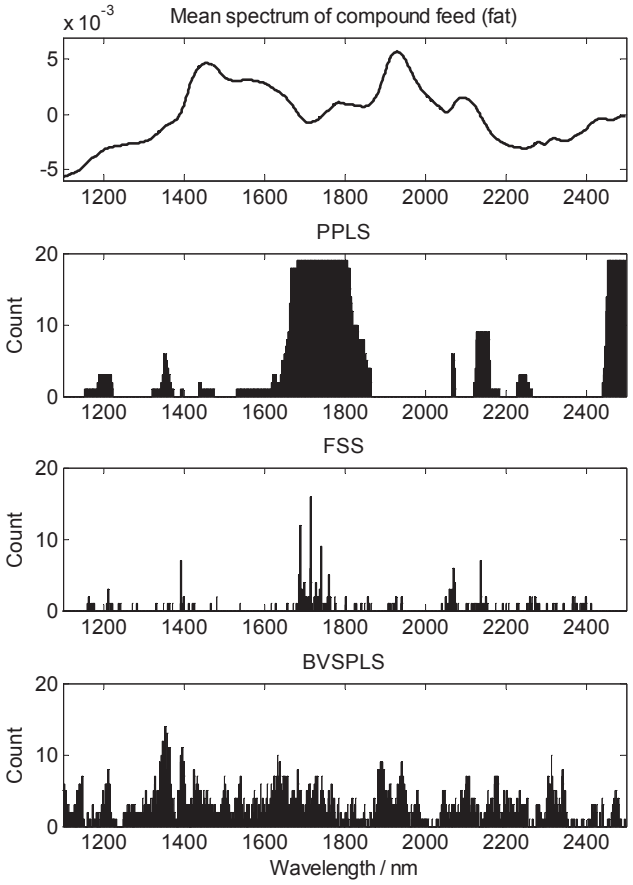

Figure 4. Frequencies of the selected variables for powered partial least square (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Histogram of the selected variables in each method (i.e. the number of times each variable was selected out of the total of 19 models) vs the wavelength for fat content in the Feed dataset.

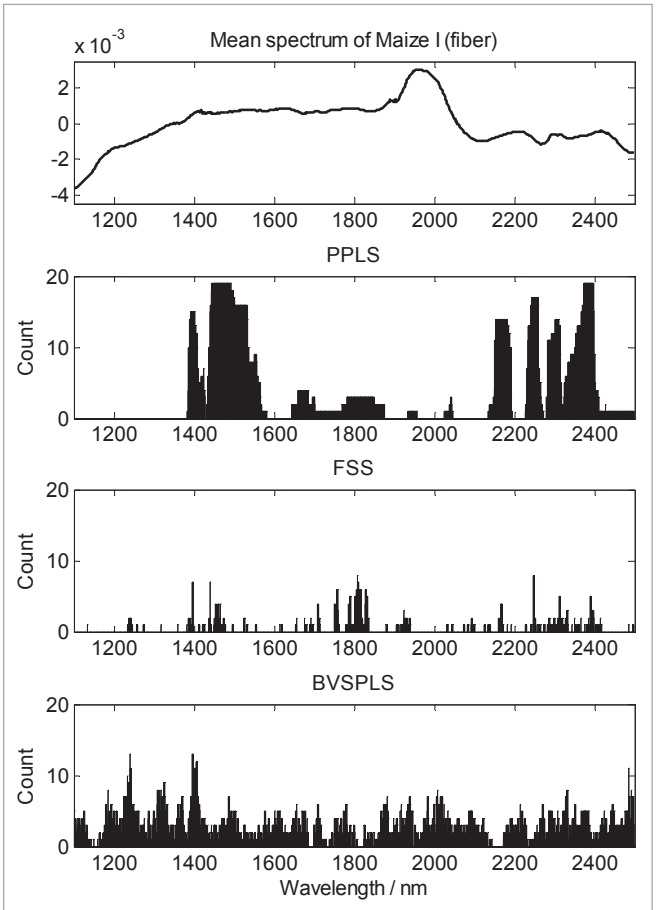

Figure 5. Frequencies of the selected variables for powered partial least square (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Histogram of the selected variables in each method (i.e. the number of times each variable was selected out of the total of 19 models) versus the wavelength for the Maize I dataset.

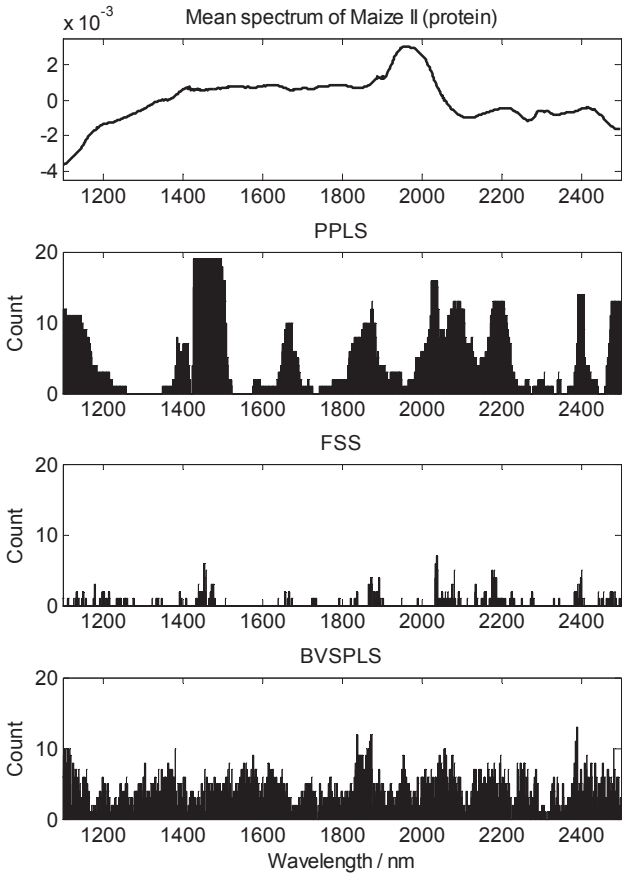

Figure 6. Frequencies of the selected variables for powered partial least square (PPLS), forward stepwise selection (FSS) and backwards variable selection for PLS (BVSPLS). Histogram of the selected variables in each method (i.e. the number of times each variable was selected out of the total of 19 models) versus the wavelength for the Maize II dataset.

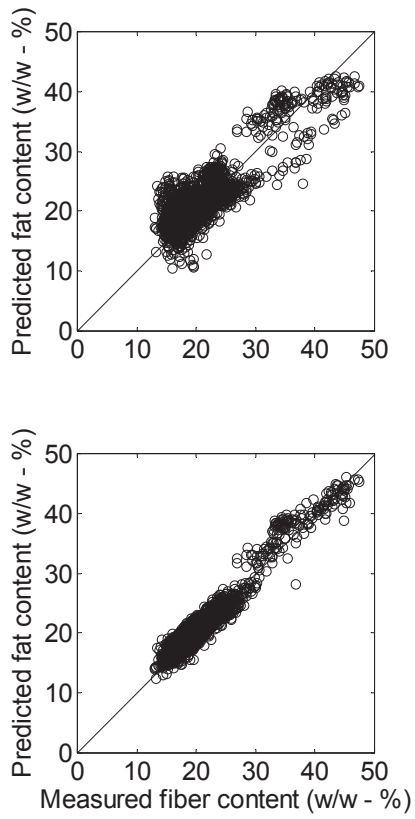

Figure 7. Predicted fat content from partial least squares (PLS, upper plot) and partial least squares (PPLS, lower plot) in feed mix plotted against the measured fat content. The PLS gave a coefficient of determination (r^2) between measured and predicted constituent of 0.78, whereas the PPLS model gave $r^2 = 0.93$. Both models were calibrated using a calibration set with 60 samples and validated on the full validation set ($N = 2521$, Figure 1).

All variable selection techniques showed improvements over PLS in some cases and the improvements were more pronounced at smaller calibration sets. The numbers of PLS factors were determined using the full 200 samples calibration sets and the numbers of components found were held constant for all other calibration sets. The actual numbers of components were determined with a conservative chi-square-test. We tried, however, several other model dimensionalities, but the general results and improvements in prediction ability were only slightly affected by this.

To explain the differences in predictive ability, we have pointed out three reasons. 1. Some predictor variables have only remote relevance to the response Variable. 2. The signal-to-noise ratio (S/N) in some predictor variables may be so low that the elimination of those variables improves the model. 3. Some predictor variables may have a nonlinear relationship to the response. Thus, elimination of these variables may give more parsimonious and linear prediction models and, hence, improve the prediction abilities. All, or a sub-set of, these reasons could explain the prediction error improvements that we have presented, but further research is needed to exploit the details in the mechanisms behind this phenomena.

FSS, and especially PPLS, selected very interpretable sets of variables (Figure 4–6). For the Feed dataset, PPLS and to a certain degree FSS, emphasised the C–H stretch bands in the 1700 nm range, whereas BVSPLS selected apparently random variables (Figure 4). This behaviour was also present in the Maize I dataset where the PPLS and FSS selected many variables in the O–H stretch band at 1450 nm and the C=O stretch band from 2000 nm to 2200 nm (Figure 5). An even stronger interpretability was found in the Maize II dataset, where the FSS and PPLS focused very strongly on the N–H stretch bands in the 1800 nm, 2000 nm and 2400 nm regions (Figure 6). For the BVSPLS, the picture was more difficult to interpret because the algorithm selected variables almost evenly spread out in the measured spectrum (Figures 4–6) but still with better prediction results than those obtained with PLS and similar to those of PPLS.

Both FSS and BVSPLS are related in the sense of being stepwise methods selecting or discarding predictors at certain

schemes. Their selection patterns are quite similar (Figures 4–6) and have a more random appearance than variables selected by PPLS. PPLS works in a fundamentally different way than the other methods because of the specially designed optimisation step that heavily weights the predictors with high correlation with the response.

Conclusion

In this paper, several validation procedures were conducted on the recent variable selection methods BVSPLS and PPLS and compare these to FSS and full spectrum PLS. The comparisons were carried out on three different NIR spectroscopic datasets predicting fat in compound feed (Feed), fibre in maize(Maize I) and protein in maize (Maize II). We have drawn three conclusions from this study.

1. Variable selection gave a positive effect on the prediction ability of small calibration sets. Since calibration samples are often costly to collect, this may be an important finding in order to make the best regression models out of few calibration samples.
2. The results clearly showed a consistent and well-structured selection of variables. FSS and BVSPLS gave less consistent variable selections as PPLS.
3. Both PPLS and BVSPLS showed high ability to compute good prediction models on small datasets which were better than FSS. This shows that variable selection techniques are evolving and require continued comparisons with existing algorithms.

Acknowledgement

In 2009, the first author spent four weeks working at the Walloon Agricultural Research Centre (CRA-W) in Gembloux, Belgium. Thanks are due to all the staff at the CRA-W for their generosity and many fruitful discussions. Also thanks to Mr Audun Korsaeth at The Norwegian Institute for Agricultural and Environmental Research (Bioforsk) for the funding that made this project possible.

References

1. S. Wold, H. Martens and H. Wold, "The multivariate calibration-problem in chemistry solved by the PLS method", *Lecture Notes in Mathematics* **973,** 286 (1983). doi: 10.1007/BFb0062108

2. L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, "Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy", *Appl. Spectrosc*. **54,** 413 (2000). doi: 10.1366/0003702001949500

3. A. Hoskuldsson, "Variable and subset selection in PLS regression", *Chemometr. Intell. Lab. Syst*. **55,** 23 (2001). doi: 10.1016/S0169-7439(00)00113-1

4. F. Westad and H. Martens, "Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression", *J. Near Infrared Spectrosc*. **8,** 117 (2000). doi: 10.1255/jnirs.271

5. H. Martens and M. Martens, "Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)", *Food Qual. Prefer.* **11,** 5 (2000). doi: 10.1016/S0950-3293(99)00039-7

6. N.M. Faber, "Uncertainty estimation for multivariate regression coefficients", *Chemometr. Intell. Lab. Syst*. **64,** 169 (2002). doi: 10.1016/S0169-7439(02)00102-8

7. R. Wehrens, H. Putter and L.M.C. Buydens, "The bootstrap: a tutorial", *Chemometr. Intell. Lab. Syst*. **54,** 35 (2000). doi: 10.1016/S0169-7439(00)00102-7

8. M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad and M. Akhond, "Ant colony optimisation: a powerful tool for wavelength selection", *J. Chemometr.* **20,** 146 (2006). doi: 10.1002/cem.1002

9. R. Leardi, M.B. Seasholtz and R.J. Pell, "Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data", *Anal. Chim. Acta* **461,** 189 (2002). doi: 10.1016/S0003-2670(02)00272-6

10. R. Leardi, "Application of genetic algorithm-PLS for feature selection in spectral data sets", *J. Chemometr.* **14,** 643 (2000). doi: 10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E

11. D. Broadhurst, R. Goodacre, A. Jones, J.J Rowland and D.B. Kell, "Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry", *Anal. Chim. Acta* **348,** 71 (1997). doi: 10.1016/S0003-2670(97)00065-2

12. D. Whitley, "An overview of evolutionary algorithms: practical issues and common pitfalls", *Inform. Software Technol*. **43,** 817 (2001). doi: 10.1016/S0950-5849(01)00188-4

13. S. Saebo, Almøy, J. Aarøe and A.H. Aastveit,"ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS", *J. Chemometr.* **22,** 423 (2008). doi: 10.1002/cem.1157

14. J.A.F. Pierna, O. Abbas, V. Baeten and P. Dardenne, "A backward variable selection method for PLS regression (BVSPLS)", *Anal. Chim. Acta* **642,** 89 (2009). doi: 10.1016/j.aca.2008.12.002

15. U. Indahl, "A twist to partial least squares regression", *J. Chemometr.* **19,** 32 (2005). doi: 10.1002/cem.904

16. I.G. Chong and C.H. Jun, "Performance of some variable selection methods when multicollinearity is present", *Chemometr. Intell. Lab. Syst*. **78,** 103 (2005). doi: 10.1016/j.chemolab.2004.12.011

17. R.J. Barnes, M.S. Dhanoa and S.J. Lister, "Standard normal variate transformation and de-trending of

near-infrared diffuse reflectance spectra", *Appl. Spectrosc*. **43,** 772 (1989). doi: 0.1366/0003702894202201

18. R.D. Snee, "Validation of regression-models-methods and examples", *Technometrics* **19,** 415 (1977). doi: 10.2307/1267881

19. E. Anderssen, K. Dyrstad, F. Westad and H. Martens, "Reducing over-optimism in variable selection by cross-model validation", *Chemometr. Intell. Lab. Syst*. **84,** 69 (2006). doi: 10.1016/j.chemolab.2006.04.021

20. C.M. Andersen and R. Bro, "Variable selection in regression-a tutorial", *J. Chemometr.* **24,** 728 (2010). doi: 10.1002/cem.1360

21. H.R. Cederkvist, A.H. Aastveit and T. Naes, "A comparison of methods for testing differences in predictive ability", *J. Chemometr.* **19,** 500 (2005). doi: 10.1002/cem.956

22. A.F.C. Pereira, M.J. Coelho Pontes, F.F. Gambarra Neto, S.R. Bezerra Santos, R.K. Harrop Galvão and M.C. Ugulino Araújo, "NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection", *Food Res. Int*. **41,** 341 (2008). doi: 10.1016/j.foodres.2007.12.013

23. F. Liu, Y.H. Jiang and Y. He, "Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer", *Anal. Chim. Acta* **635,** 45 (2009). doi: 10.1016/j.aca.2009.01.017

24. N. Shetty, R. Gislum, A.M. Dahl Jensen and B. Boelt, "Development of NIR calibration models to assess year-to-year variation in total non-structural carbohydrates in grasses using PLSR", *Chemometr. Intell. Lab. Syst.* **111,** 34 (2011). doi: 10.1016/j.chemolab.2011.11.004