A Backward Variable Selection method for PLS regression (BVSPLS)

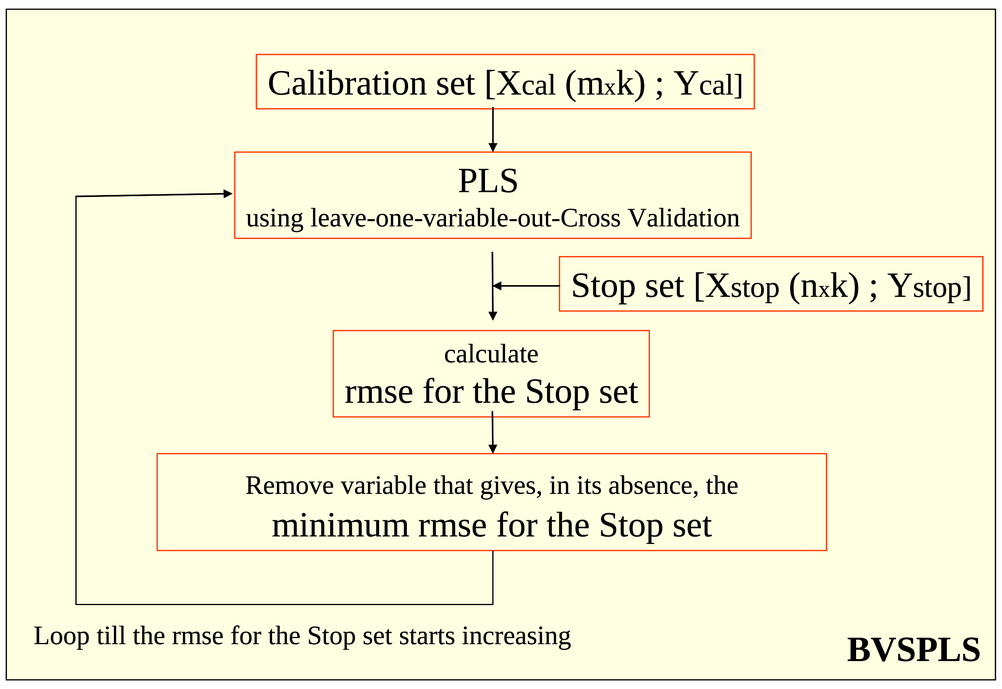
Juan Antonio Fernández Pierna, Vincent Baeten & Pierre Dardenne

Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department
Chaussée de Namur, 24, 5030 Gembloux, Belgium
Tel : + 32 (0) 81 62 03 52 / fernandez@cra.wallonie.be

Introduction

In spectroscopy, thanks to the modern techniques of analysis, objects described by a large number of variables are easily measured in a very short time. Then, powerful 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 we have 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 parts 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.



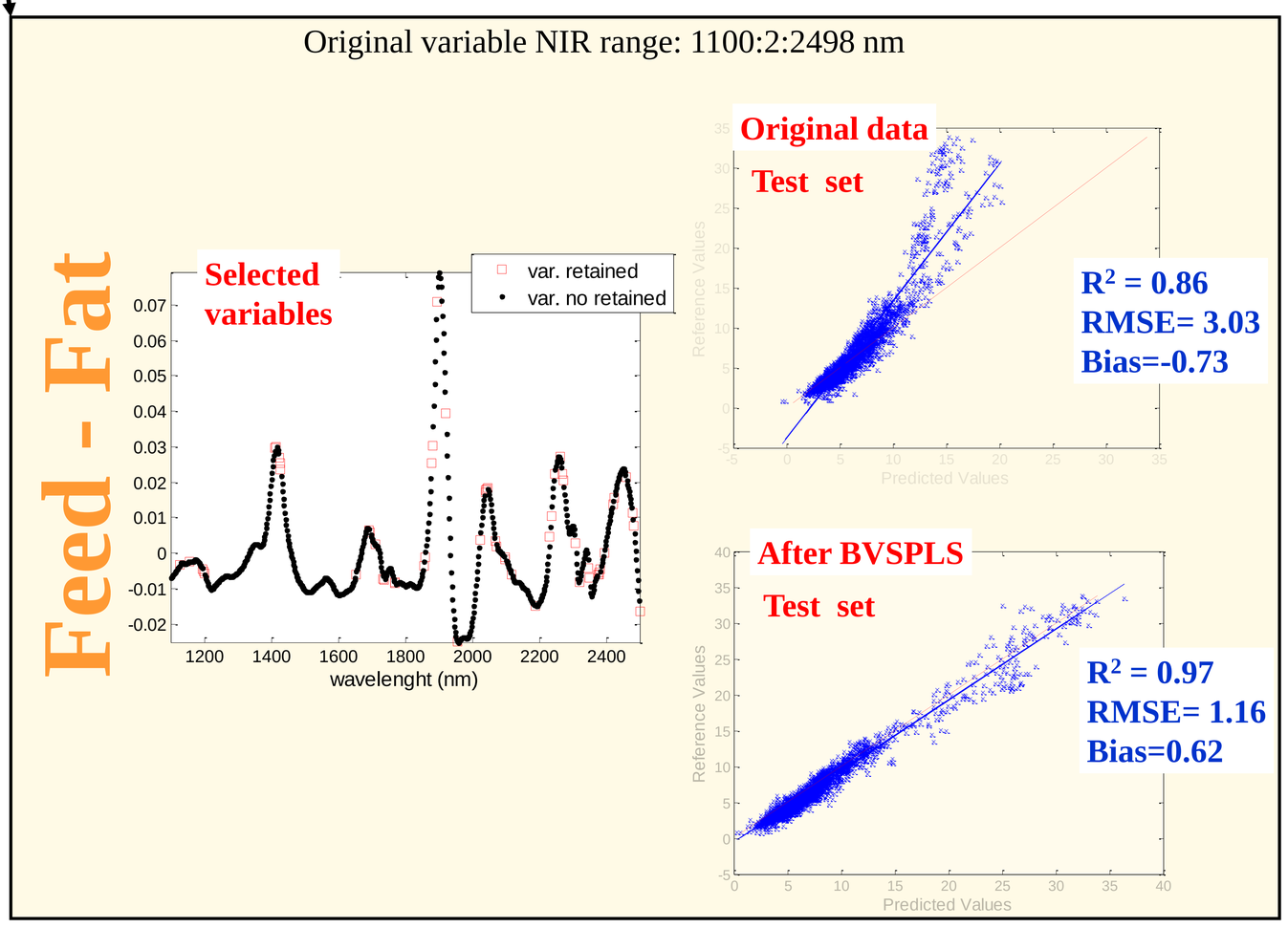
The goal of variable selection is to obtain the smallest set of variables that gives the generalization ability comparable with the original set of variables [Nagatani and Abe 2007]. The main benefits of variable selection are the improvement of data visualization and data understanding, the reduction of measurement requirements and training and utilization times by defying the curse of dimensionality to improve prediction performance (better robustness) [Guyon and Elissee 2003].

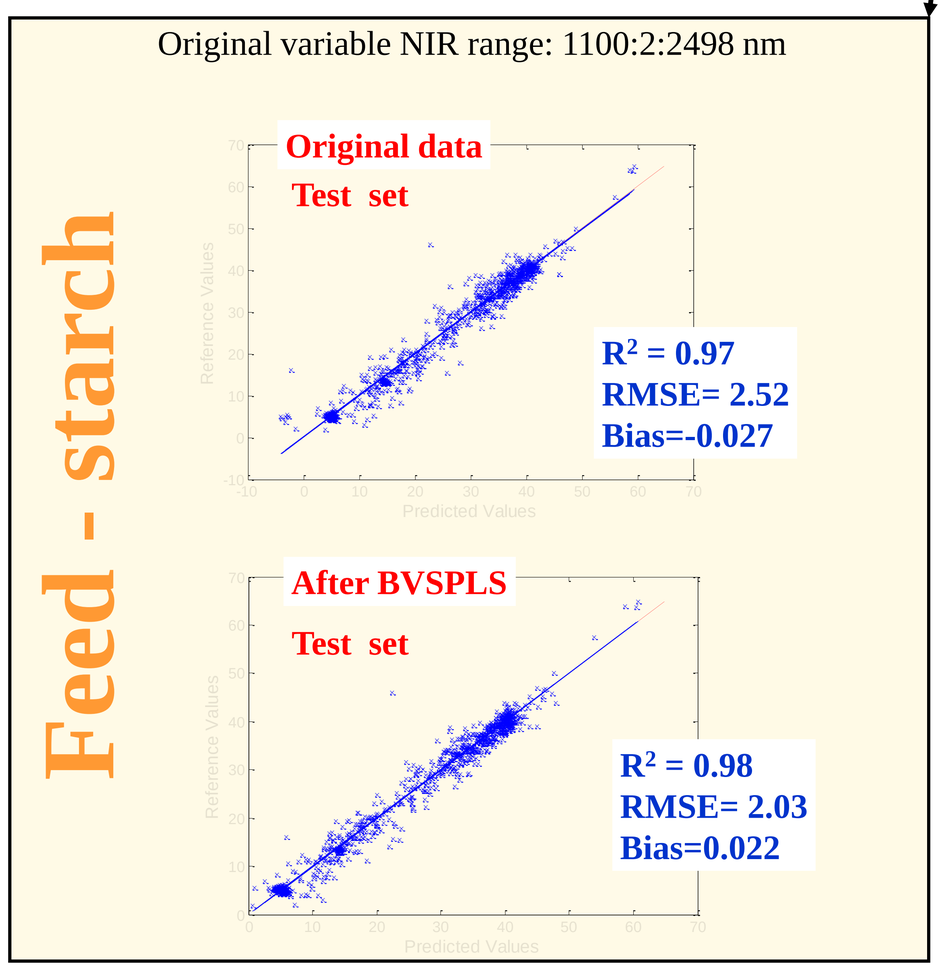
Here, we present a variable selection method that directly evaluates the selected subset of input variables and estimates the generalization ability using the regression equation (wrapper method) having, then, into account the particular biases of the algorithm. We apply a backward iterative step-by-step method for the selection of spectral variables using PLS (the 'Backward Variable Selection method for PLS regression' or BVSPLS). We use the root-mean-square criterion to quantify the improvement obtained using the selected range of variables.

It is an iterative method. In the first step, the method starts by fitting a model with all the variables and then a leave-one-variable-out is performed. 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 root mean square error (RMSEP). The minimum RMSEP is selected indicating that that 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).

Results

Samples	Property	Number of samples cal	stop	test	RMSEP test samples before bvspls	after bvspls	number of original variables	number of variables retained	% variables retained
Complete feed	*fat*	*362*	*518*	*3043*	3.03	1.16	700	58	8
Complete feed	*protein*	*336*	*873*	*6182*	1.68	1.57	700	62	9
Complete feed	*moisture*	*337*	*1010*	*5920*	0.84	0.70	700	224	32
Complete feed	*starch*	156	200	1000	2.52	2.03	700	226	32
Maize	*protein*	*272*	*338*	*3869*	0.63	0.54	700	260	37
Maize	*starch*	*165*	*331*	*4558*	2.84	2.41	700	122	17
Maize	*cel*	*175*	*355*	*3526*	1.37	1.35	700	110	16
Maize	*ndf*	*264*	*279*	*3354*	2.44	2.99	700	95	14





Conclusion

The objective of this technique was mainly to construct and select subsets of features that are useful to build a good predictor, i.e. to construct a suitable model with only few variables. However, in most of the cases studied, the RMSEP we obtain decreases drastically with up to 79% of the variables removed in average.
We can improve or at least keep constant the prediction performances by providing faster predictors as well as an improvement in the stability of the PLS models and a clear 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.

In a further study, this method will be apply 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

Guyon I & Elisseeff A, 'An introduction to variable and feature selection', Journal of machine learning research 3 (2003),pp 1157-1182.
Nagatani T & Abe S, 'Backward variable selection of support vector regressors by block delection'; International Joint Conference on Neural Networks (IJCNN) 2007, pp: 2117-2122 Orlando, FL, USA.