Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression

Åsmund Rinnan[1], Martin O. Andersson[2], Carsten Ridder[3] and Søren Balling Engelsen[1]*

[1]Faculty of Life Science, Department of Food Science, Frederiksberg, DK-1958, Denmark
[2]FOSS Japan Ltd., Tokyo Genboku Kaikan 9F, 30-13 Toyo 5-Chome, Koto-Ku, 135-0016 Tokyo, Japan
[3]Lattec I/S, Slangerupgade 69, 3400 Hillerød, Denmark
*Corresponding author: se@life.ku.dk

Introduction

Near infrared (NIR) spectra contain holographic information about the samples being investigated (Figure 1). The same information is repeated again and again and the spectral variables are thus highly correlated (co-linear data structure). For this type of data, multivariate projection methods are very suitable both for supervised and unsupervised exploration of NIR spectral ensembles. The combination of multivariate data analysis and NIR spectroscopy is the working principle in Process Analytical Technology (PAT)[1]. The most important multivariate algorithms are principal component analysis (PCA)[2] and partial least squares (PLS) regression[3].

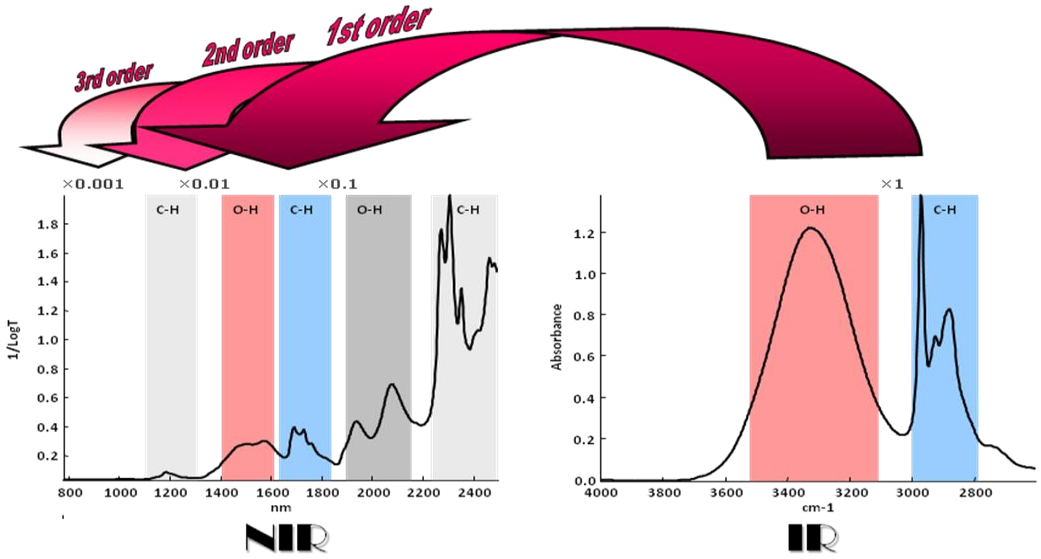

Figure 1. The holographic principle in NIR spectroscopy.

Despite the high redundancy in NIR spectra, the multivariate methods can often be improved by variable selection such as forward selection or by regional selection such as interval PLS (*i*PLS)[4]. The primary reason for the improvements is the reduced number of interferences in the reduced set of variables. Perhaps even more important than improved and parsimonious models is the improved interpretation. The selected variables or region can be assigned by the spectroscopist and referred to a chemical or physical phenomenon.

Variable selection methods generally come in two versions: one that is focussed on finding variables that are good at prediction of the response variable and one that is focussed on uncertainty estimates on the coefficients of the regression vector. If an acceptable global PLS model is obtained, a normal procedure is to inspect the model parameters, for example the regression coefficients. High absolute regression coefficients are considered important and small regression coefficients are considered less important and can be eliminated. This principle is the basis for variable selection by jack-knifing[5,6]. For each cross-validation segment, a new regression vector is calculated and it will thus be possible to estimate the standard deviation of the PLS regression vector. When the distribution of the regression coefficients includes zero, they can be discarded and a new model calculated. Jack-knifing can be implemented iteratively by recalculating the model and eliminating more variables. Jack-knifing differs from other variable selection methods by not looking directly for variables that are good at predictions but just by eliminating variables that possibly have a regression coefficient close to zero and thus do not contribute to the prediction. No matter what value such variables have, they will be multiplied by zero and thus not contribute. To remove variables this way makes

Reference paper as:
Rinnan, Å., Andersson, M.O., Ridder, C. and Engelsen, S.B. (2012). Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 91-95.

the danger of overfitting much smaller than when using variable selection methods that are focussed on finding variables that are good at predicting the response variable.

Recursive weighted PLS (rPLS) is related to jack-knifing but instead of iteratively eliminating variables, rPLS iteratively uses the regression coefficients to magnify important variables and thus down-weight less important variables. Recursive PLS is based on a process of repeated PLS models; the current regression coefficients are used as cumulative weights on **X**:

$$\mathbf{X}_{new} = \mathbf{X} \times \text{diag}(\mathbf{b})$$

where **X** is the previous updated weighted **X**, **b** is the regression coefficient from the last model and $\mathbf{X}_{new}$ is the current **X** used in the PLS model. The rPLS model has the advantage that, under normal conditions, it will converge to a limited number of variables (good for interpretation) but it will exhibit optimal performance before normally including co-linear neighbour variables.

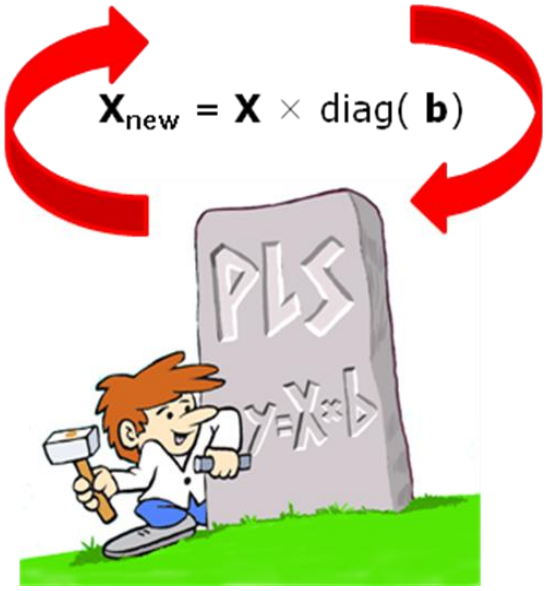

Figure 2. The recursive principle in the rPLS algorithm. © Engelsen & Newlin

Materials and Methods

Samples

We will demonstrate the use and performance of the rPLR algorithm by application to a spectroscopic data set dealing with the determination of the amount of extract from NIR spectra of beers. The data set contains the NIR spectra of 60 beer samples with the related extract content in percentage plato. The percentage extract varies between 4.2-18.8% plato. The extract content is an important quality parameter in the brewing industry as it indicates the yeast potential for fermenting alcohol.

NIR spectroscopy

Spectra were collected using a NIRSystems 6500 spectrophotometer (FOSS NIRSystems, Inc., Silver Spring, Maryland, USA) fitted with a Transport Module (NR-6511). The spectrophotometer uses a split detector system with a silicon (Si) detector between 400 and 1100 nm and a lead sulphide (PbS) detector from 1100 to 2500 nm. The vis-NIR transmission spectra were recorded with a 30 mm quartz cell directly on the undiluted degassed beer and spectral data collected at 2 nm intervals in the range from 400 to 2250 nm were converted to absorbance units.

Chemometrics

Recursively Weighted Regression or rPLS is based on an recursive re-weighting of the independent variable block (**X**) by the regression vector **b** calculated from a PLS regression model between **X** and **y**: $\mathbf{X}_{i+1} = \mathbf{X}_i \times \text{diag}(\mathbf{b}_i)$. The algorithm is started with a standard PLS model between $\mathbf{X}_1$ (equal to **X**) and **y**, giving $\mathbf{b}_1$. The re-weighting is repeated until no further progress in the regression coefficients occurs. The result is a regression vector $\mathbf{b}_{end}$ that contains only ones and zeros. This binary result is a direct output from the rPLS algorithm, i.e. no rescaling of the final regression vector is performed. In a simple case, the number of variables selected (i.e. variables with a corresponding regression coefficient of one) corresponds to the

Reference paper as:
Rinnan, Å., Andersson, M.O., Ridder, C. and Engelsen, S.B. (2012). Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 91-95.

number of latent factors chosen in the original PLS model; this is not the case in more complicated situations. This simple method, which combines multivariate regression and variable selection, is currently being thoroughly tested in a set of different spectroscopic ensembles.

The iPLS and rPLS methods are implemented in MATLAB (The MathWorks, Inc.) and made freely available for non-commercial use at www.models.life.ku.dk. PCA, PLS and spectral plots were made using chemometric software (LatentiX; www.latentix.com).

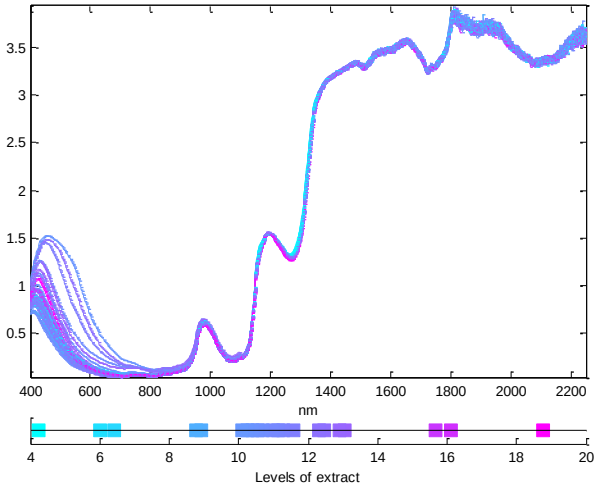

Figure 3. NIR spectra of the beer coloured according to their extract values (LatentiX).

Results and Discussion

Figure 3 shows the NIR spectra of the 60 beer samples coloured according to extract concentration. The NIR spectra contain a rather large noisy part due to absorbance that is too strong, in a region dominated by water. The NIR spectra can roughly be divided into a visible region (400 to 800 nm) which vary systematically but irrelevantly, a centre region (800 to 1400 nm) which seems to have relevant information and a first overtone and combination region (1400 to 2250 nm) which has absorbance values that are too high. This rough picture is exactly the visual outcome of application of iPLS[4] to the data. More importantly, the performance of the PLS model is improved drastically from the global prediction error RMSECV=0.75% plato (blue horizontal line) to the best local model with RMSECV=0.18% plato.

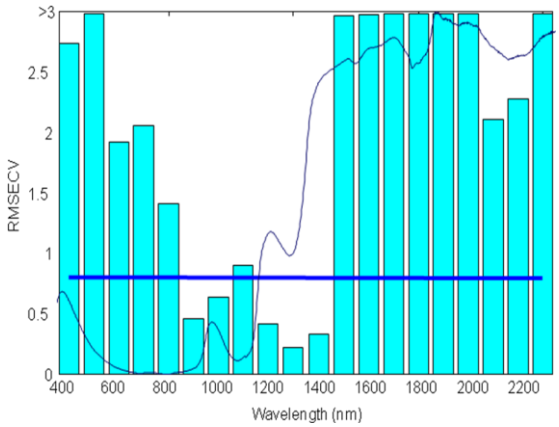

Figure 4. A simple 20 interval PLS on the relationship between NIR spectra and extract values. The global PLS performance is indicated with a blue horizontal line. The performance of local PLS models is shown with cyan bars. The average NIR spectrum is show in the background.

In order to investigate the rPLS performance, it was applied to the same data set (Figure 5). Interestingly the rPLS prediction performance improved significantly during the first 7 iterations to reach a similar performance as the local iPLS model mentioned above after which point the regression performance increased slightly. Only in iteration #26 did the rPLS converge with no further changes in the regression vector being observed.

Reference paper as:
Rinnan, Å., Andersson, M.O., Ridder, C. and Engelsen, S.B. (2012). Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 91-95.

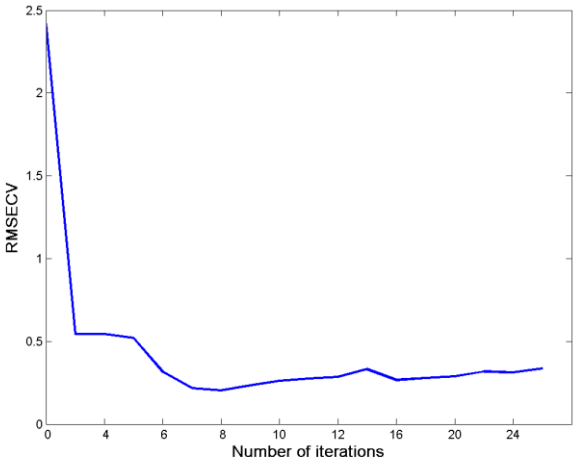

Figure 5. The development in prediction performance (RMSECV) during the iterations of rPLS.

The result of iteration #7 is shown in Figure 6. At this point, 169 of the original 926 variables still have regression coefficients above the numerical uncertainty of the calculation. However, from this point onward there was a significant reduction in variables with regression coefficients above the threshold of $2x10^{-16}$. In the last iteration (Figure 7), only four variables were "retained", three of which had insignificant contributions. In fact, the variable around 1320 nm was the only significant variable after iteration #7 and even before. The many additional rPLS iterations were primarily cleaning up dirty variables in the saturated region. Even in the converged rPLS, a few dirty variables persisted (see Figure 7).

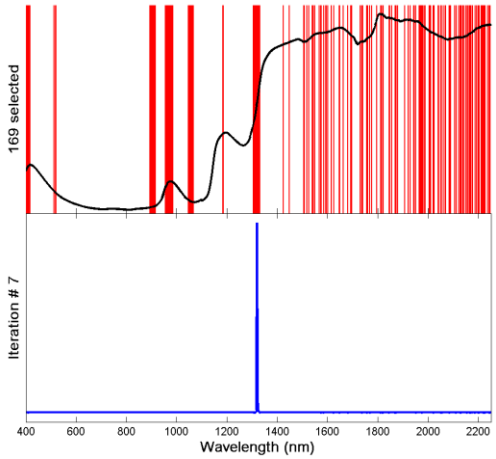

Figure 6. Iteration #7 of the rPLS on beer data. This is the optimal prediction model with 167 variables still having regression coefficients above zero.

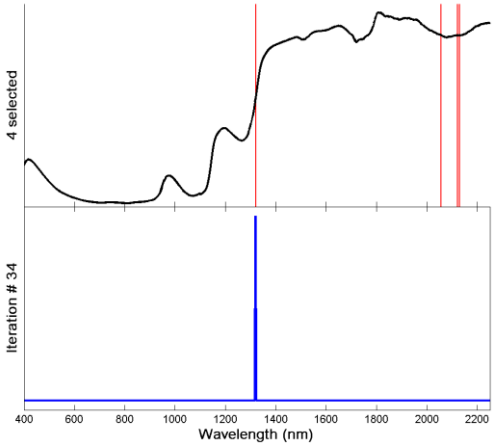

Figure 7. Iteration #34 of the rPLS on beer data. This is the converged rPLS model with 4 variables retained.

Conclusion

In this paper, we have made a preliminary demonstration of recursive weighted PLS (rPLS). The method combines a very promising regression performance (optimal; in this context iteration #7) and simple spectroscopic assignment (end; in this context iteration #34). The method compares favourably to other variable search methods such as principal variable and forward selection due to the inclusion of covariate neighbour channels in the optimal model. The method is superior in finding a very limited set of variables (if not one) in the end-model which is a large benefit for interpretation. We have not discussed validation in this paper which is a most important issue in supervised variable selection in a regression model but the method has been tested in a number of different applications with extensive bootstrapping validation and seems to behave quite well.

The rPLS method is not only applicable to NIR spectroscopy data. We are currently investigating the performance of the method in a metabolomics context both with respect to regression and classification as rPLS-DA. The latter may be especially useful for interpretation of the usually very complex metabolomics data sets.

Reference paper as:
Rinnan, Å., Andersson, M.O., Ridder, C. and Engelsen, S.B. (2012). Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 91-95.

References

1. E. Skibsted and S.B. Engelsen, "Fundamentals of spectroscopy: spectroscopy for process analytical technology (PAT)", in *Encyclopedia of spectroscopy and spectrometry*, Ed by J. Lindon, G. Tranter and D. Koppenaal, 2nd edition, Vol. 3, Elsevier, Oxford, UK, pp. 2651–2661 (2010).
2. H. Hotelling, *J. Educ. Psychol.* **24**, 417-441 (1933).
3. S. Wold, H. Martens and H. Wold, Lect. Notes Math 973, 286-293 (1983).
4. L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, *Appl. Spectrosc.* **54,** 413-419 (2000).
5. B. Efron and G. Gong, *Am. Stat.* **37,** 36-48 (1983).
6. H. Martens and M. Martens, *Food Qual. Prefer.* **11,** 5-16 (2000).

Reference paper as:
Rinnan, Å., Andersson, M.O., Ridder, C. and Engelsen, S.B. (2012). Recursive weighted PLS (rPLS): an efficient and promising multivariate method for spectral variable selection in regression, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 91-95.