# Modified jack-knifing in multivariate regression for variable selection

**Frank Westad and Michael Byström**
*CAMO ASA, Nedre Vollgate 8, N-0158 Oslo, Norway.*

**Harald Martens**
*The Technical University of Denmark, Institute of Biotechnology, Denmark Technical University, DK-2800 Lyngby, Denmark and Institute of Physical Chemistry, Norwegian University of Science and Technology, N-7034 Trondheim, Norway.*

## Introduction

One of the criticisms of PLS regression has been the lack of significance testsof the model parameters. Up to now "rules of thumb" exist, mainly based on previous experience when selecting significant wavelengths. We present a novel method based on jack-knifing to estimate variances of model parameters in PCR and PLS regression for cross-validation. The purpose of this work is to develop a model that predicts ethanol concentrations in mixtures of methanol, ethanol and propanol from near infrared (NIR) absorbance data with different techniques.

## Theory

### Programs

The calculations were performed with Matlab version 5.1.0 (MathWorks, Inc., Natick, MA, USA) and The Unscrambler® version 7.5 (CAMO A/S, Norway).[1]

### Regression theory

Let a linear model be described in matrix notation as model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ where $\beta$ contains the regression coefficients estimated with a desired loss function, $\mathbf{X}$ is a matrix of spectra and $\mathbf{y}$ is some response-variable to be predicted for future sampled spectra. For methods like PCR and PLS, one of the important aspects is to find the optimal number of components (Aopt), preferably from a suited validation method like cross-validation (CV) or independent test set validation.[2]

When cross-validation is applied in regression, Aopt is determined based on prediction of kept-out objects (samples) from the individual models. The root mean square error (*RMSE*) is an error measure for how well the model performs and is given by the expression

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (y - \hat{y})^2}{N}}$$

The notation $RMSEP_{CV}$ is used to indicate cross-validation whilst $RMSEP_{T}$ indicates test set validation. *RMSEC* is the fit from the calibration. Normally, one would chose Aopt from the lowest *RMSE* value, but this can lead to overfitting in the validation. If the decrease in *RMSE* is small when adding

one more component, one might want to be more conservative (we will focus on automatic ways to find the "best" model, and not use supervised visual inspection of loadings, regression coefficients, residual variance curves etc.). Below is pseudo-code for a conservative assessment of Aopt using a "punish factor" to avoid overfitting.

```
PunishFactor = 0.03% 3 percent
ReduceAopt = 1;
while ReduceAopt
   if RMSECV(Aopt)*1 + PunishFactor > RMSECV(Aopt -1)
     A_opt = Aopt -1
   else
     ReduceAopt = 0
   end
end
```

This procedure was used for finding Aopt in the full cross-validated PLS model.

## Uncertainty estimates and variable selection

The approximate uncertainty variance of the PCR and PLS regression coefficients **b** can be estimated by jack-knifing[3,4]

$$s^2\mathbf{b} = \left( \sum_{m=1}^{M} (b - b_m)^2 \right) \left( N - 1/N \right)$$

where
$N$ = the number of samples
$s^2\mathbf{b}$ = estimated uncertainty variance of **b**
**b** = the regression coefficient at the cross-validated Aopt components using all the $N$ samples
$\mathbf{b}_m$ = the regression coefficient at the rank A using all objects except the object(s) left out in cross-validation segment $m$
On the basis of such jack-knife estimates of the uncertainty of the model parameters, useless or unreliable variables may be eliminated automatically, in order to simplify the final model and making it more reliable. This is done by significance tests, where t-tests are performed for each element in **b** relative to the square root of its estimated uncertainty variance $s^2\mathbf{b}$, giving the significance level for each parameter. It is not to be expected that the significance test for a data set with many variables will converge to a stable set of variables in the first round. This is due to the fact that there is a high degree of redundancy (and collinearity) in NIR spectra and relevant X-information extracted for latter components is often near the noise level. The significance for each variable in the regression coefficient can be used as a variable selection method:

```
perform PLS on full spectrum with cross-validation
pLimit = 0.05
X = empty
sort variables after pValues
while decrease in RMSECV for Aopt
  X = X + most significant variable
  perform PLS with the new variable set
  compute RMSECV (and RMSEP for test samples)
```

```
end
```

## Experimental

The data set contained 101 wavelengths each 5 nm, from 1100 to 1600 nm. The spectra were transformed to absorbance units. No single wavelength can be used alone because of strongly overlapping spectra. Sixteen calibration samples were mixed in a such a way that they spanned the different variations in alcohol concentration. A mixture triangle was used. All the mixture samples were carefully made up in the laboratory. The mixing proportions were used directly as *Y*-values. The inaccuracy in the reference method, therefore, only consists of the variation in sample preparation, volume measurement etc. Ten new samples were also prepared from scratch to be used as test samples.

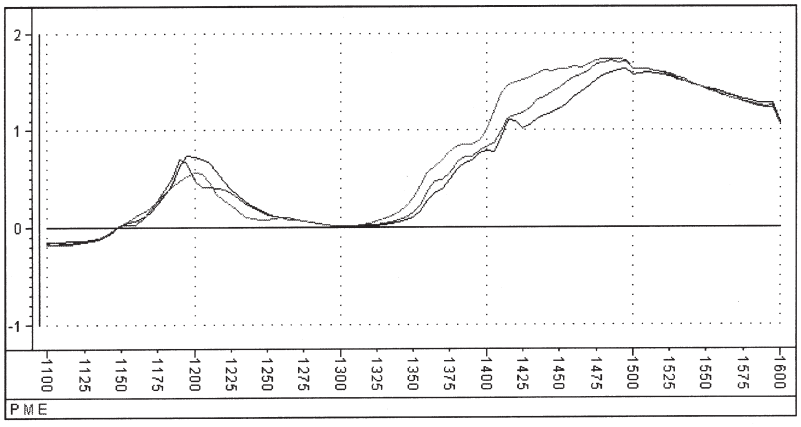


**Figure 1. Line plot of the pure spectra of methanol, ethanol and propanol.**

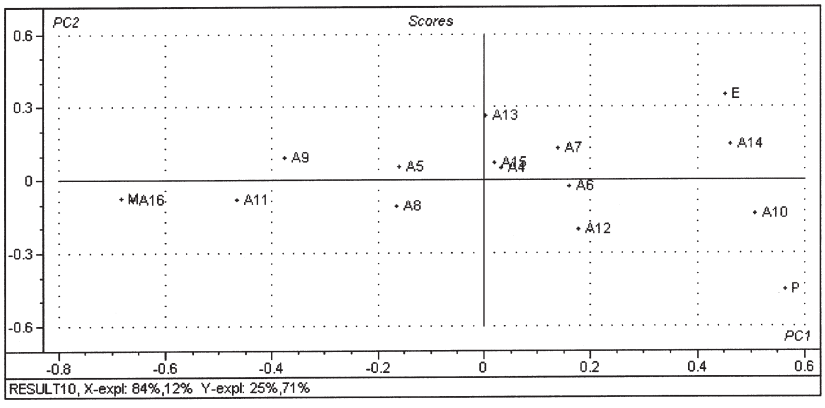


**Figure 2. Score plot of the first model, no wavelength selection.**

Spectra of mixtures often exhibit scatter effects due to interference. This causes shifts in the spectra. Therefore, Multiplicative Scatter Correction was used to correct for base line shifts and multiplicative effects.

Pure spectra of methanol, ethanol and propanol are shown in Figure 1.

## Results and discussion

A PLS model with all variables was calculated on 16 calibration samples. The test set consisted of ten samples. The mixture structure of the three alcohols is clearly visible in the score plot in Figure 2. Propanol and methanol are negatively correlated in PLS component 1 and methanol is negatively correlated to the combination of the other two in PLS component 2. This means that PLS component 1 describes the variation of the proportions of propanol and methanol, while PLS component 2 describes

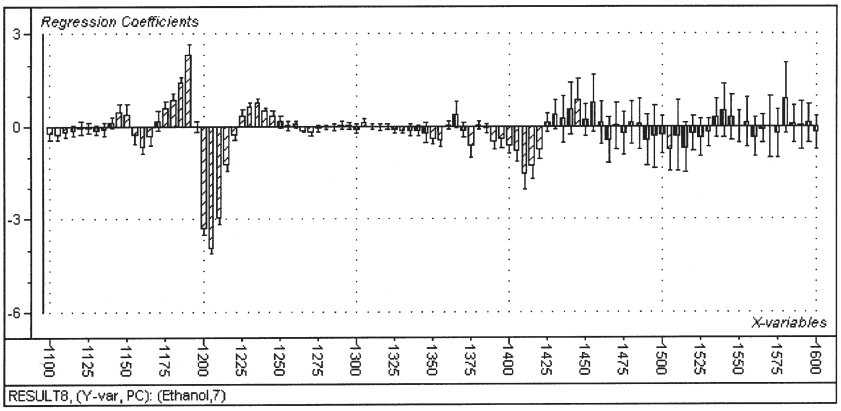


**Figure 3. Plot of regression coefficients with uncertainty limits ±2 × sdev. Significant variables are marked with striped bars.**

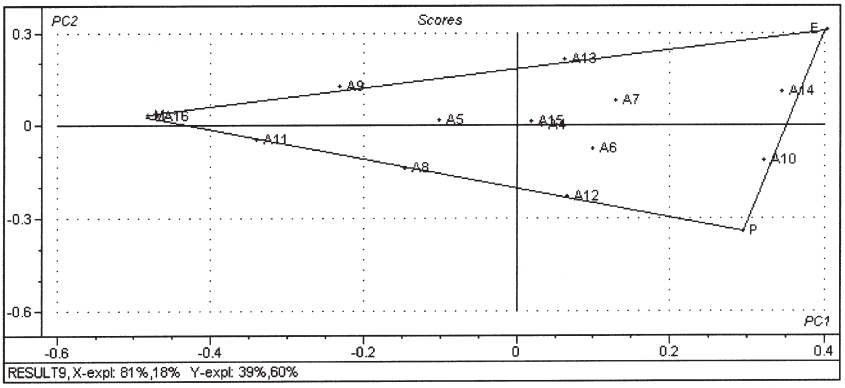


**Figure 4. Score plot after wavelength selection.**

**Table1. Summary of results for the investigated regression models.**

| Method | # Comp. | # Variables | *RMSEC* | $RMSEP_{CV}$ | $RMSEP_{T}$ |
|---|---|---|---|---|---|
| Full spectrum PLSR | 7 | 101 | 0.03 | 1.13 | 1.11 |
| Jack-knife PLSR | 7 | 39 | 0.26 | 0.94 | 1.13 |
| Jack-knife forward selection | 4 | 4 (1190, 1200, 1205 and1210 nm) | 0.39 | 0.64 | 0.66 |
| MLR Best Combination search | — | 2 (1190 and 1205 nm) | 0.63 | 0.81 | 0.93 |

the variation of the three alcohols together. There is also a curvature present in the data along the line of mixtures of methanol and propanol.

Using wavelength selection based on jack-knifing, 39 variables were identified as significant. In the regression coefficient plot, in Figure 3, the selected variables are marked with striped bars. The uncertainty limit, ± 2 × estimated standard deviation, is also displayed. After the wavelength selection based jack-knifing, the known curvature was smoothed out and the mixture triangle became even clearer. This is well illustrated in the score plot in Figure 4.

$RMSEP_{CV}$ was the chosen error measure in the variable selection for all tested methods, i.e. the set of variables and Aopt chosen for a given method is the set that gives the lowest $RMSEP_{CV}$ among the combinations tested with that method. The $RMSEP_{CV}$ and other information regarding the models are listed in Table1. When applying jack-knife estimates for forward selection, the four most significant wavelengths in the spectra are; 1190, 1200, 1205 and 1210. $RMSEP_{T}$ is an estimate of the prediction error based on ten samples and its value also reveals if there are problems with overfit for some of the methods. The model based on four jack-knife selected variables predicted the test samples better than the MLR best combination search. We have, therefore, shown that the variable selection method from significance tests based on jack-knifing for these data is an alternative to best combination search MLR. The method also has the interpretational advantages of PLS regression and can be applied to other types of data.

## Acknowledgement

We thank Howard Mark for running experiments on the data with his best combination search program.

## References

1. The Unscrambler® 7.5. CAMO ASA, Oslo, Norway (1999).
2. H. Martens and T. Næs, *Multivariate Calibration.* J. Wiley & Sons Ltd, Chichester, UK (1989).
3. B. Efron, *Society for Industrial and Applied Mathematics*.Philadelphia, Pennsylvania, USA (1982).
4. H. Martens and M. Martens, *Food quality and preference*, in press (1999).