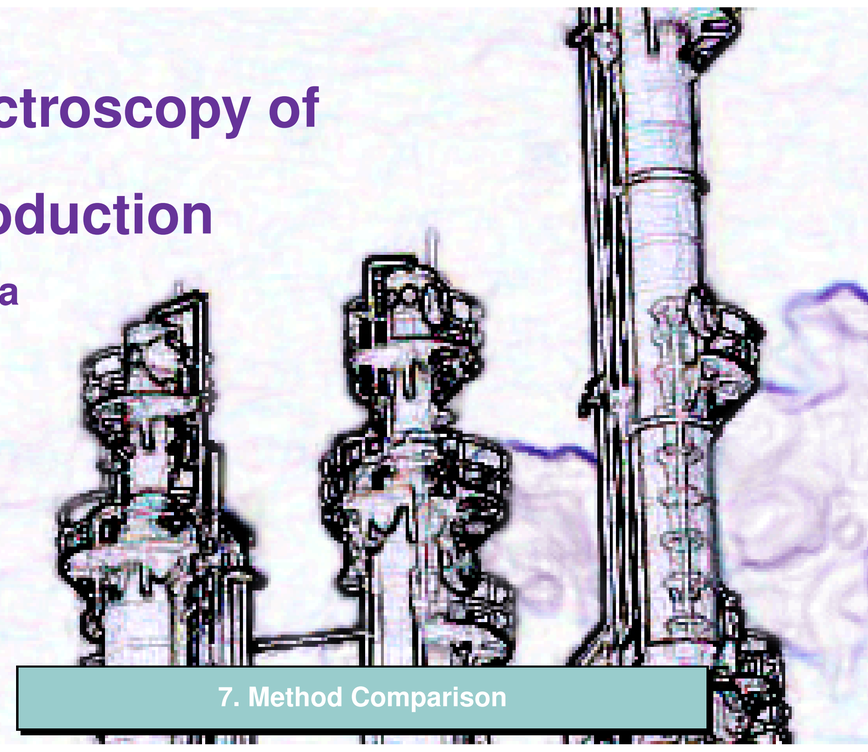
Quantitative Analysis by Near-Infrared Spectroscopy of Compounds Relevant in Bioethanol Production

Bettina Liebmann*, Anton Friedl, Kurt Varmuza

Vienna University of Technology
Institute of Chemical Engineering
Getreidemarkt 9/166-2, A-1060 Vienna, Austria
www.lcm.tuwien.ac.at, www.thvt.at
bettina.liebmann@tuwien.ac.at



1. Introduction

Near-infrared (NIR) spectroscopy was applied to bioethanol fermentations with

- High sample variability from batch to batch due to changes in feedstock and enzymatic pretreatment
- Multi-constituent substrates
- Minimal sample preparation for rapid, nondestructive analysis

Objectives

- Quantify relevant compounds: glucose, ethanol, glycerol, lactic acid, fructose, maltose, arabinose
- Develop PLS regression models based on NIR absorbance data
- Select important variables by a Genetic Algorithm (GA)
- Optimize the PLS models' complexity and estimate its prediction performance for new cases by "rdCV"

2. Experimental

Process Steps

Wheat/rye/corn → enzymatic pretreatment → enzymatic starch degradation → fermentation by yeast → ethanol containing **mash** → separate ethanol by distillation → **stillage** remains as residue

Sample Preparation

- Centrifugation to remove solids
- Stepwise addition of known amounts of the compound under investigation (for calibration)
- Determination of reference concentrations (g/L) by HPLC with refractive index detector

NIR Absorbance Data

1100-2300 nm at 5 nm intervals, AOTF-NIR spectrometer *Brimrose Luminar 5030*, fiber-optic transflectance probe.
1st derivative Savitzky-Golay results in 235 *x*-variables; variable reduction by GA [1,2] to 15 variables (different variables for each compound)

3. Method

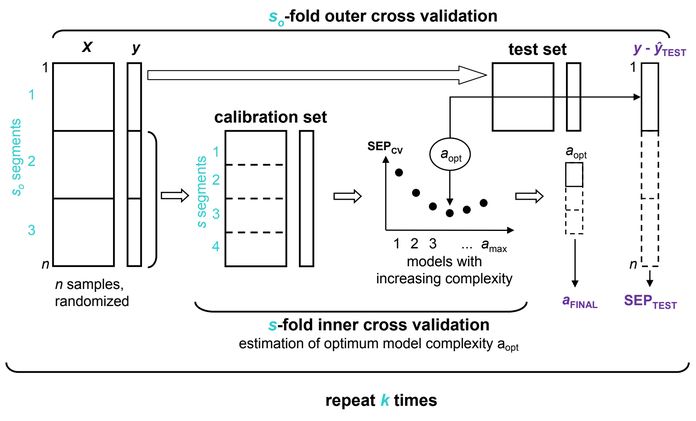
Repeated Double Cross Validation rdCV

- The data set is randomly partitioned into s_o segments: s_o - 1 segments for calibration, 1 segment as test set.
- A PLS model is derived from the calibration set with optimum number of PLS components estimated by *s*-fold inner cross validation.
- Application of PLS model to test set results in n/s_o predicted values $\hat{y}_i$.
- Systematic variation gives a $\hat{y}$ for each object.
- The whole process is repeated *k* (e.g. 100) times.
- Finally, $k \cdot n$ values $\hat{y}$ are available.

Implementation in R

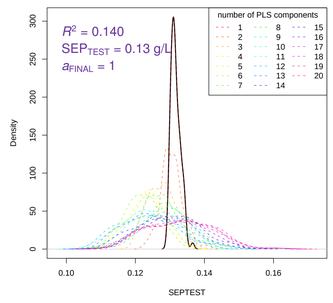
rdCV is available as function `mvr_dcv` in new package *chemometrics* [3,4] developed in R [5,6].

Scheme of repeated double cross validation with s_o = 3 segments in the outer loop and *s* = 4 segments in the inner loop. The process is repeated *k* times.



4. Evaluation

Example: Lactic acid quantification in stillages by NIR, range: 0.06-0.63 g/L
Density distribution of 100 SEP$_{TEST}$ values with increasing model complexity for 100 repetitions

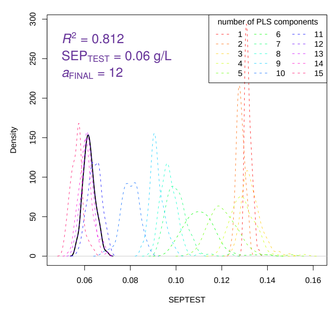


all 235 NIR variables

very low predictive performance

higher complexity → larger errors

broad error distributions



15 GA selected NIR variables

sound predictive performance

higher complexity → reduced errors

narrow error distributions

no overfitting

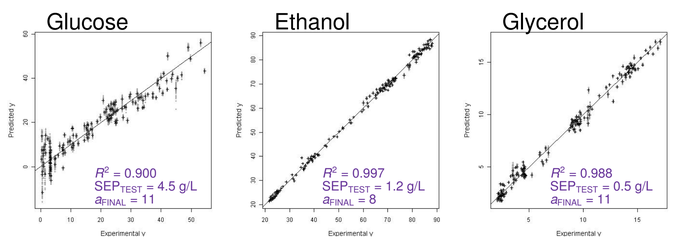
Performance criteria derived from rdCV:

SEP$_{TEST}$	standard deviation of test set predicted errors y-$\hat{y}$ ($k \cdot n$ values $\hat{y}$ available)
a_{FINAL}	final optimum of $s \cdot k$ calculated numbers of PLS components (method: [7])

5. Predicted vs. Experimental

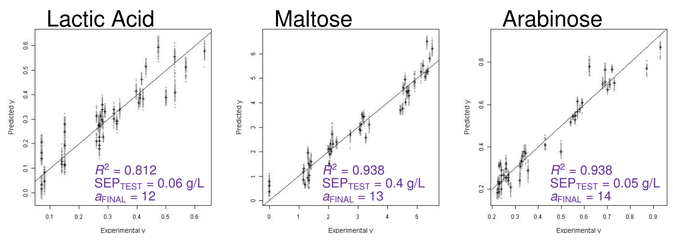
Compounds in Mashes

166 samples, 15 GA selected NIR variables
experimental/predicted *y* in g/L



Compounds in Stillages

50 samples, 15 GA selected NIR variables
experimental/predicted *y* in g/L



6. Prediction Performances by rdCV

Compound	*n*	SEP$_{TEST}$ NIR all	SEP$_{TEST}$ NIR GA	Concentration range in g/L
Mashes				
glucose	166	5.6	4.5	0-54
ethanol	166	1.5	1.2	22-88
glycerol	166	0.7	0.5	2-17
Stillages				
glucose	50	4.0	1.7	0-24
ethanol	50	3.0	0.8	0-58
glycerol	50	1.7	0.6	3-14
lactic acid	50	0.1	0.1	0-1
fructose	50	0.7	0.5	0-6
maltose	50	0.8	0.4	0-6
arabinose	50	0.1	0.1	0-1

n	number of samples
SEP$_{TEST}$	standard deviation of 100·*n* prediction errors (g/L)
NIR all	all 235 NIR absorbance values available
NIR GA	15 GA selected NIR absorbance values

7. Method Comparison

Results of repeated double cross validation (rdCV) are compared with 4-fold cross validation as implemented in software Unscrambler [8]. All data sets with 15 GA selected variables.

		rdCV		CV	
Compound	*n*	SEP$_{TEST}$	a_{FINAL}	SEP$_{CV}$	a_{CV}
Mashes					
glucose	166	4.5	11	5.2	8
ethanol	166	1.2	8	2.7	2
glycerol	166	0.5	11	1.0	4
Stillages					
glucose	50	1.7	13	2.3	5
ethanol	50	0.8	15	2.1	4
glycerol	50	0.6	15	0.8	10
lactic acid	50	0.1	12	0.1	10
fructose	50	0.5	12	0.6	4
maltose	50	0.4	13	0.5	6
arabinose	50	0.1	14	0.1	5

n	number of samples
repeated double cross validation in R:	
SEP$_{TEST}$	standard deviation of 100·*n* test set predicted errors (g/L)
a_{FINAL}	optimum number of PLS components
4-fold random cross validation in Unscrambler:	
SEP$_{CV}$	standard deviation of *n* CV predicted errors (g/L)
a_{CV}	optimum number of PLS components

8. Conclusions

- Easily available **near-infrared spectroscopy** data are **very promising** for the quantification of diverse compounds in **highly variable substrates** of the bioethanol process. Samples included three different feedstock options (wheat, rye, and corn) and six different enzymatic pretreatments.
- **Variable selection** by Genetic Algorithm **improves prediction performance** for all PLS models.
- Repeated double cross validation offers a **sophisticated optimization strategy** for model complexity (number of PLS components). Furthermore, prediction performance can be reasonably estimated.
- In comparison, **4-fold cross validation** yields **higher prediction errors**, as the optimum number of PLS components is chosen more conservatively.
- Evaluation of prediction quality suggests that a higher number of PLS components does not necessarily imply overfitting.
- Implementation of repeated double cross validation **in software R is fast and easy** with typical computation times of 0.5 to 10 minutes.

9. References

1. Software MobyDigs, v 1.0. Talete srl, www.talete.mi.it, Milan, Italy, 2004.
2. Leardi, R.: J. Chromatogr. A 1158 (2007) 226-233.
3. Varmuza, K., Filzmoser, P.: Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton, FL,2009.
4. Filzmoser, P., Liebmann B., Varmuza, K.: J. Chemom. 23 (2009) 160-171.
5. Software R, v 2.8.1. R Development Core Team, www.r-project.org, 2009.
6. Mevik, B. H., Wehrens, R.: J. Statistical Software 18 (2007), issue 2.
7. Hastie, T., Tibshirani, R. J., Friedman, J.: The Elements of Statistical Learning. Springer, New York, USA, 2001.
8. Software The Unscrambler v 9.0. Camo Process AS, www.camo.no, Oslo, Norway, 2004.

We gratefully acknowledge support by the *Austrian Research Promotion Agency (FFG), BRIDGE program*, project no. 812097/11126 and W. Krenn, Vogelbusch GmbH Vienna. We thank P. Filzmoser (Institute of Statistics and Probability Theory, Vienna University of Technology) for collaboration in statistics.

Poster Presentation: 11th Scandinavian Symposium on Chemometrics, SSC11, 8th - 11th June 2009, Loen/Stryn, Norway