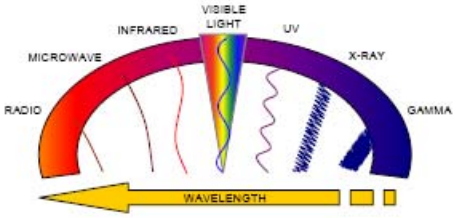
Near-Infrared Spectroscopy and Chemometrics in Bioethanol Production



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

Poster Presentation:

13. Österreichische Chemietage (13th Austrian Chemistry Days)

24 - 27 August 2009, Vienna, Austria

Near-Infrared (NIR) Spectroscopy

... studies the interaction between radiation and matter as a function of wavelength in the near infrared region of the electromagnetic spectrum at approx. 800-2500 nm

Advantages

- easy sampling (long distance fibre optic probing)
- no reagents or waste streams
- non-invasive, non-destructive
- larger penetration depth than IR
- low maintenance costs

Disadvantages

- complex method calibration ($\rightarrow$ chemometrics!)
- no trace component analysis method
- sample temperature control necessary

Chemometrics is ...?

... a data-driven interfacial discipline to ...

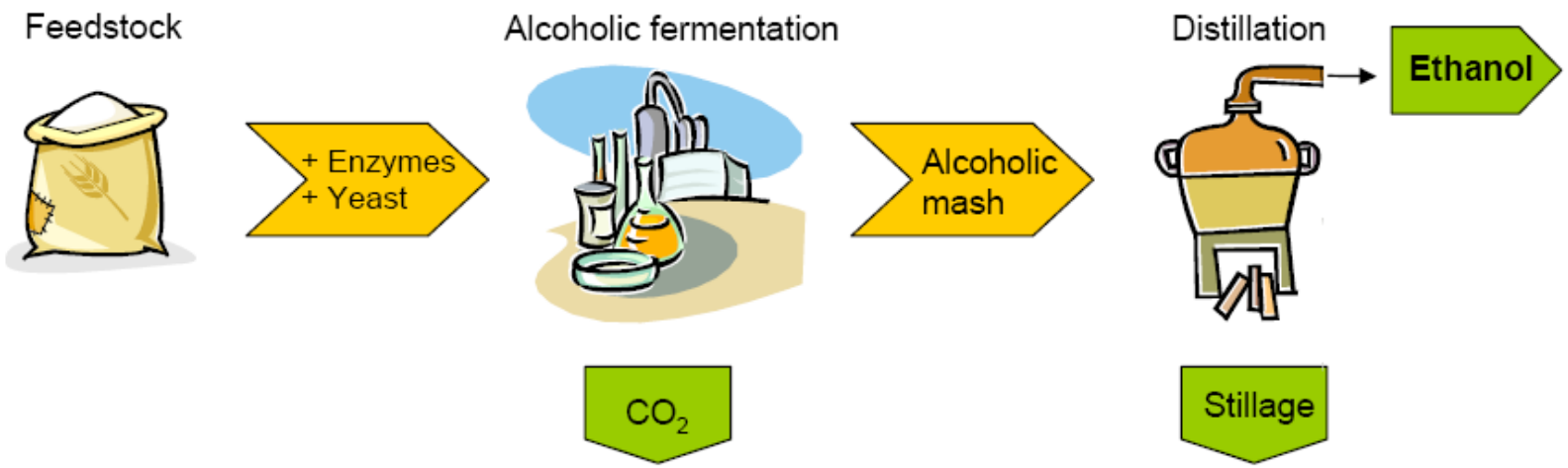
$\rightarrow$ extract information from chemistry-relevant data by mathematical and statistical methods

$\rightarrow$ improve understanding of very large and highly complex datasets in chemistry

$\rightarrow$ reveal underlying relationships in data

$\rightarrow$ correlate data to properties or quality parameters

Bioethanol Production from Starch



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

Application Example

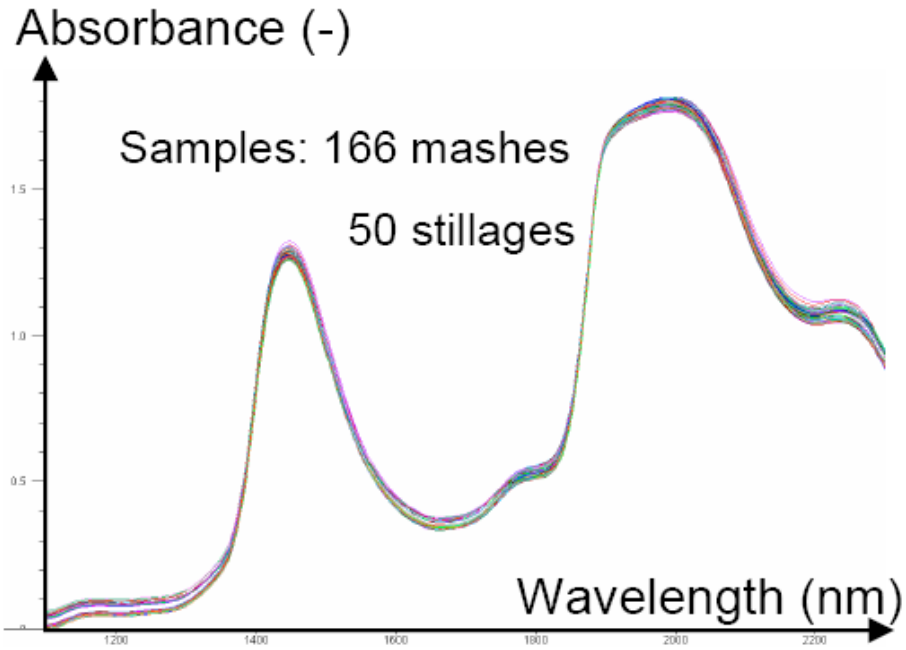
Near-infrared (NIR) spectroscopy was applied to bioethanol fermentations [1] with

- 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

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

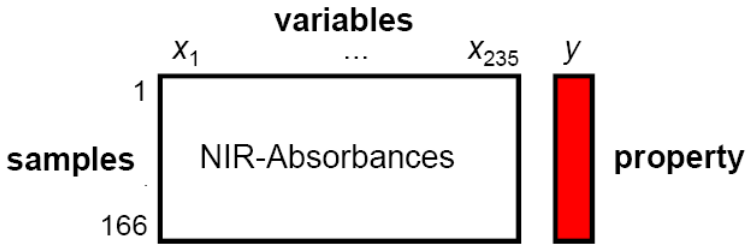


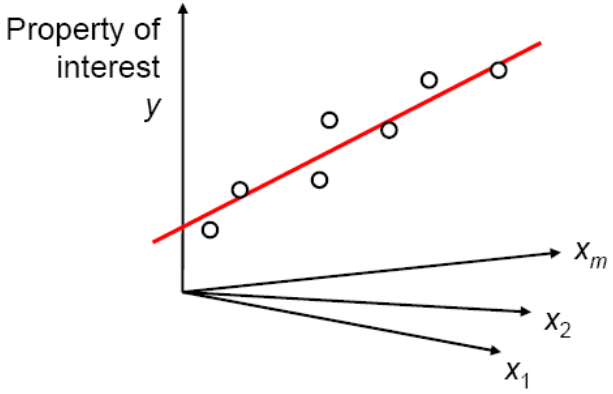
1100-2300 nm at 5 nm intervals, AOTF-NIR spectrometer *Brimrose Luminar 5030*, fiber-optic transflectance probe. 1st derivative spectra contain 235 *x*-variables; variable reduction to 15 (specific variables for each compound) by a Genetic Algorithm (GA) [2,3].

Method

Partial Least Squares Regression PLS

→ Develop PLS calibration models to correlate spectral data **X** with important properties **y** (concentration of glucose, ethanol etc.). [4]





empirical model obtained by PLS regression:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

m experimental variables
(e.g. NIR absorbance data)

Repeated Double Cross Validation rdCV

→ Optimize the models' complexity and test their predictive performances when applied to new samples

- Statistical estimation of the prediction performance of a model is based on a set of prediction errors (y-$\hat{y}$).

- Test set objects, "unknown" to the model under validation, are used.

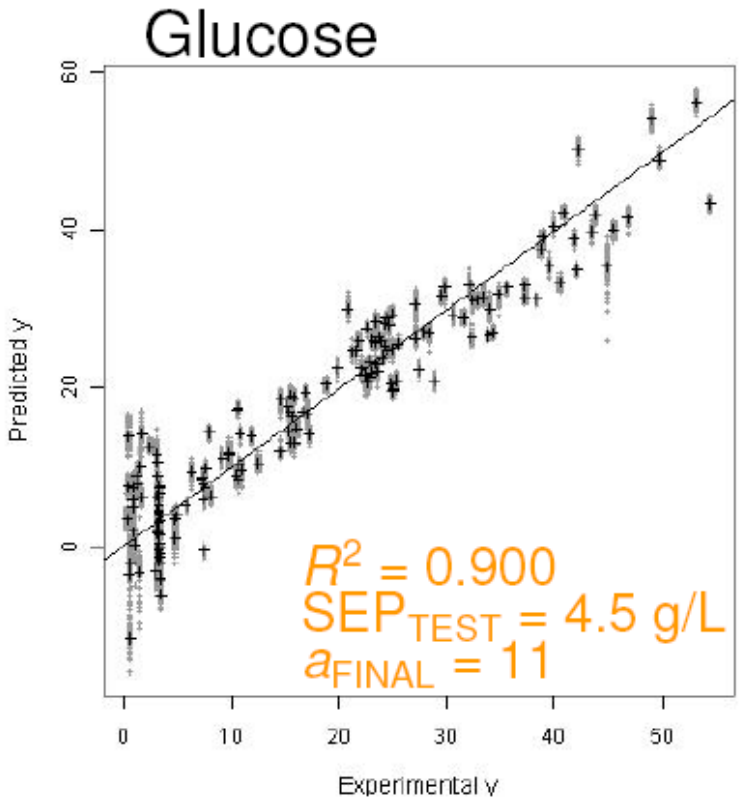
- The rdCV algorithm is described in [5] and is freely available in the package *chemometrics* for software R [6].

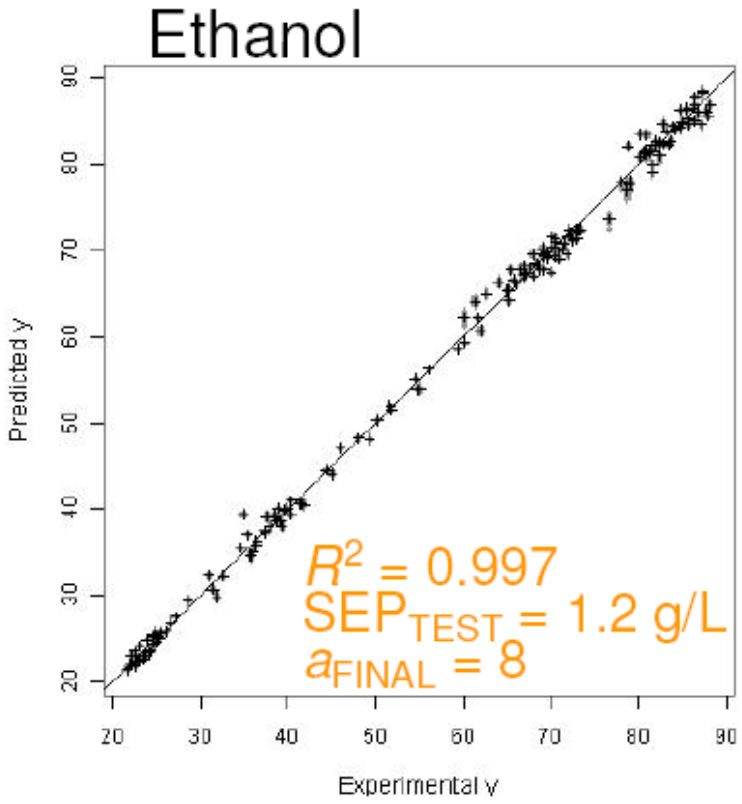
Performance criteria derived from rdCV:

R^2	squared Pearson correlation coefficient between y and $\hat{y}$
SEP_{TEST}	standard deviation of test set predicted errors y-$\hat{y}$ ($100 \cdot n$ values $\hat{y}$ available from rdCV)
$\pm$ 2 SEP	95% tolerance interval for prediction errors (for normally distributed errors)
a_{FINAL}	final model complexity: derived from 400 estimated numbers of PLS components

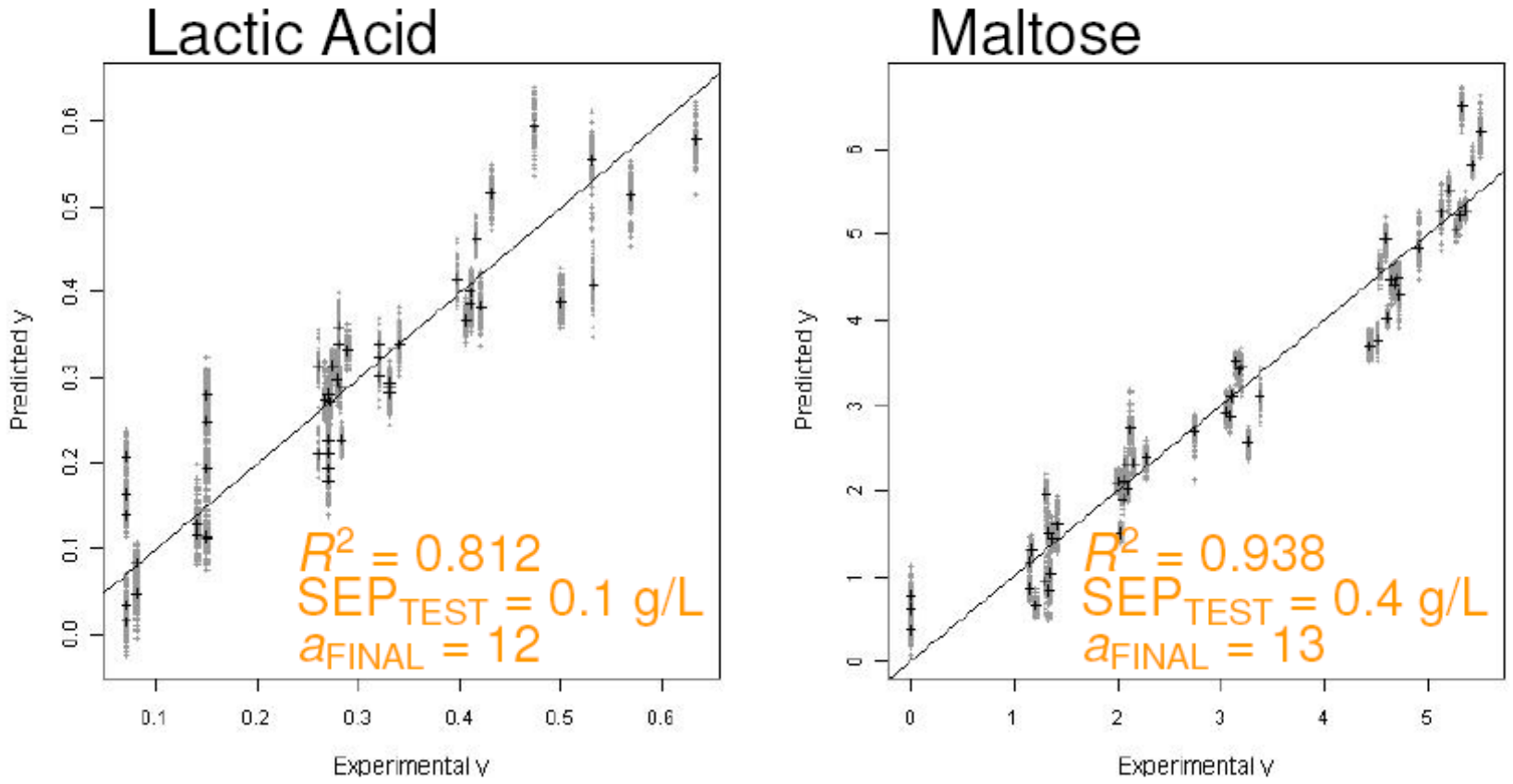
Compounds in Mashes / Stillages

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





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



Summary of Results

Compound	*n*	SEP_{TEST} g/L	Concentration range in g/L
Mashes			
glucose	166	4.5	0-54
ethanol	166	1.2	22-88
glycerol	166	0.5	2-17
Stillages			
glucose	50	1.7	0-24
ethanol	50	0.8	0-58
glycerol	50	0.6	3-14
lactic acid	50	0.1	0-1
fructose	50	0.5	0-6
maltose	50	0.4	0-6
arabinose	50	0.1	0-1

n	number of samples
SEP_{TEST}	standard deviation of prediction errors (g/L) based on test set objects
m=15	15 GA selected NIR absorbance values specific for each model

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 a Genetic Algorithm **improved prediction performance** of 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.
- **Best models for ethanol** quantification in both mash and stillage (range: 0-88 g/L): 95 % of the expected errors are within a tolerance interval of $\pm$ 2 g/L.
- **Poor model performance** for low concentrated components such as lactic acid and arabinose($<$ 1 g/L).
- **Minor sample pretreatment** (removal of solids from liquid samples) is **advisable** for better and more accurate models.

References

1. Liebmann, B., Friedl, A., Varmuza, K.: Anal. Chim. Acta, 642 (2009), 171-178.
2. Software MobyDigs, v 1.0. Talete srl, www.talete.mi.it, Milan, Italy, 2004.
3. Leardi, R.: J. Chromatogr. A 1158 (2007) 226-233.
4. Varmuza, K., Filzmoser, P.: Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton, FL, 2009.
5. Filzmoser, P., Liebmann, B., Varmuza, K.: J. Chemom. 23 (2009) 160-171.
6. Software R, v 2.8.1. R Development Core Team, www.r-project.org, 2009.

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.

Abstract

In large industrial-scale processes such as bioethanol production, chemically undefined multiple substrates are present that can be highly variable from batch to batch. The type of feedstock (e.g. wheat, rye, and corn), the enzymatic pretreatment, as well as yeast fermentation itself vary the complexity of the initial medium.

Near-infrared (NIR) spectroscopy is well suited for rapid, non-destructive, multi-constituent analyses with minimal sample preparation directly in the fermentation broth. Compounds of interest during bioethanol production are glucose, the nutrient for yeast fermentation, its fermentation product ethanol, as well as side-products such as lactic acid, acetic acid, and glycerol.

The objective of this study is the development of PLS regression models for a prediction of concentrations of the above-mentioned compounds in different bioethanol mashes by NIR spectroscopy [1]. We apply a genetic algorithm (GA) for variable selection and evaluate the models' prediction performance by a repeated double cross validation (rdCV). rdCV offers a strategy to estimate the optimum model complexity - that is the number of PLS components. Furthermore, rdCV allows a realistic estimation of the prediction errors and their variations for new cases, based on a large number of test set predicted values [2]. The results are promising for a successful application of NIR spectroscopy and multivariate data evaluation in bioethanol production.

Acknowledgments: FFG - Austrian Research Promotion Agency (BRIDGE program, # 812097/11126)

[1] Liebmann B., Friedl A., Varmuza K.: Anal. Chim. Acta (2009), DOI: 10.1016/j.aca.2008.10.069.
[2] Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics. CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2009.