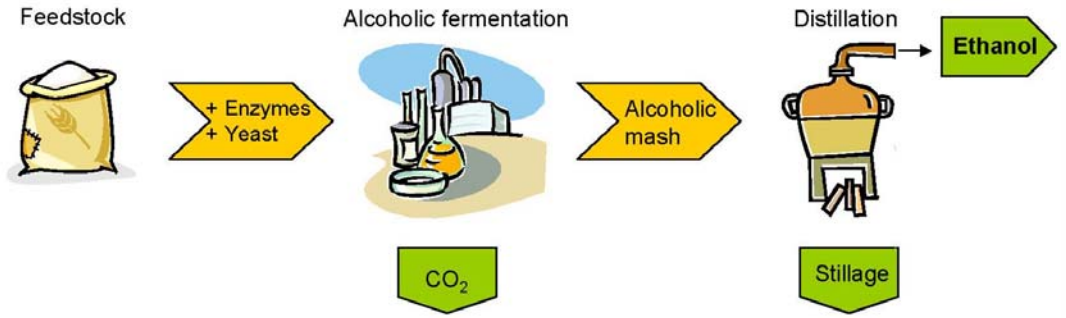
Applicability of near-infrared (NIR) spectroscopy for process monitoring in bioethanol production

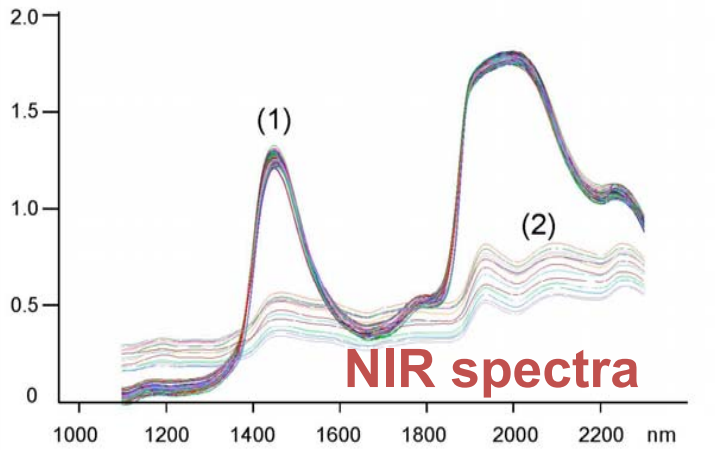
Bettina Liebmann
Institute of Chemical Engineering, TU Vienna

Overview

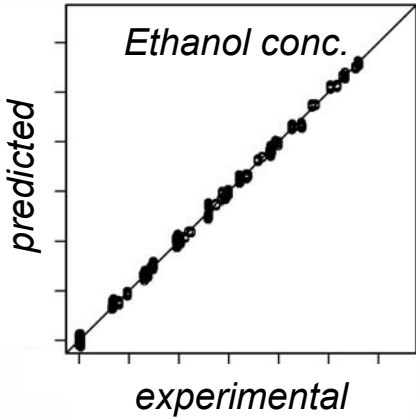
- Bioethanol Production:



- NIR spectroscopy:



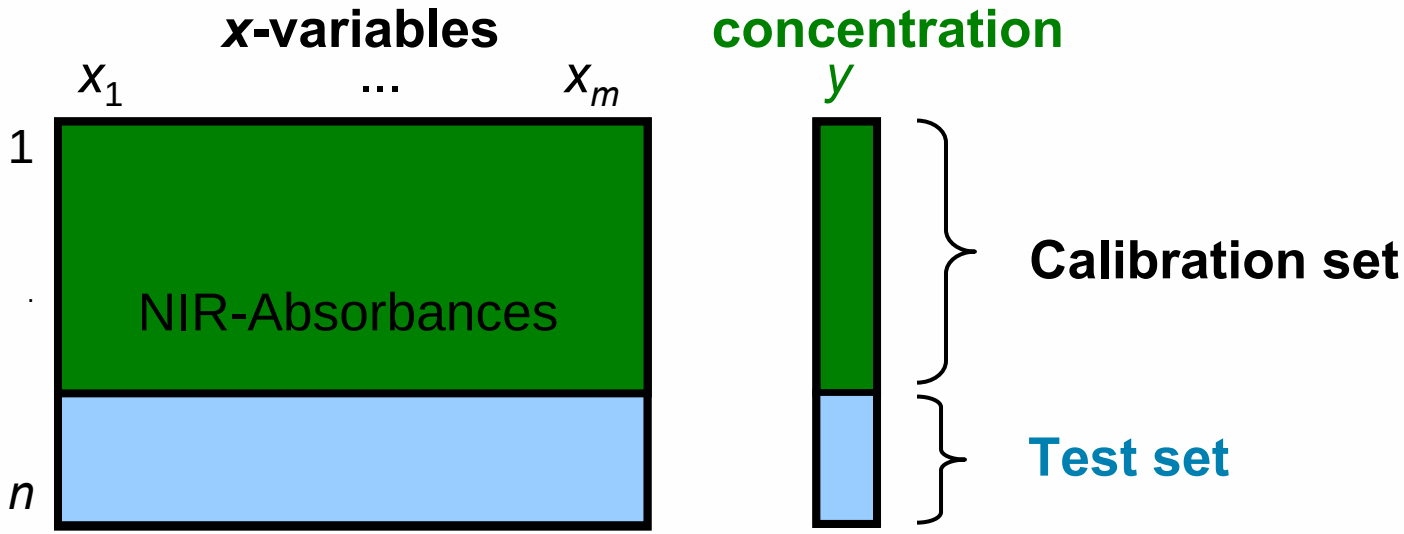
- Chemometrics:



Empirical model

→ **careful validation!**

Analysis of data from NIR spectroscopy



→ **Create** linear PLS **model** from **calibration** data: $y = f(x)$

$y = b_0 + b_1 x_1 + ... + b_m x_m$

→ **Optimize** PLS model's **complexity** within calibration data (CV)

→ **Validate** PLS model with **test** data: $\hat{y}_{TEST} = f(x)$

We want small errors ($\hat{y}_{TEST} - y$)

Repeated double cross validation (rdCV)

... Consists of 3 nested loops

Repetition loop

with different random sequences of the samples

Outer CV loop

Split data into calibration sets and test sets

Create model from calibration set

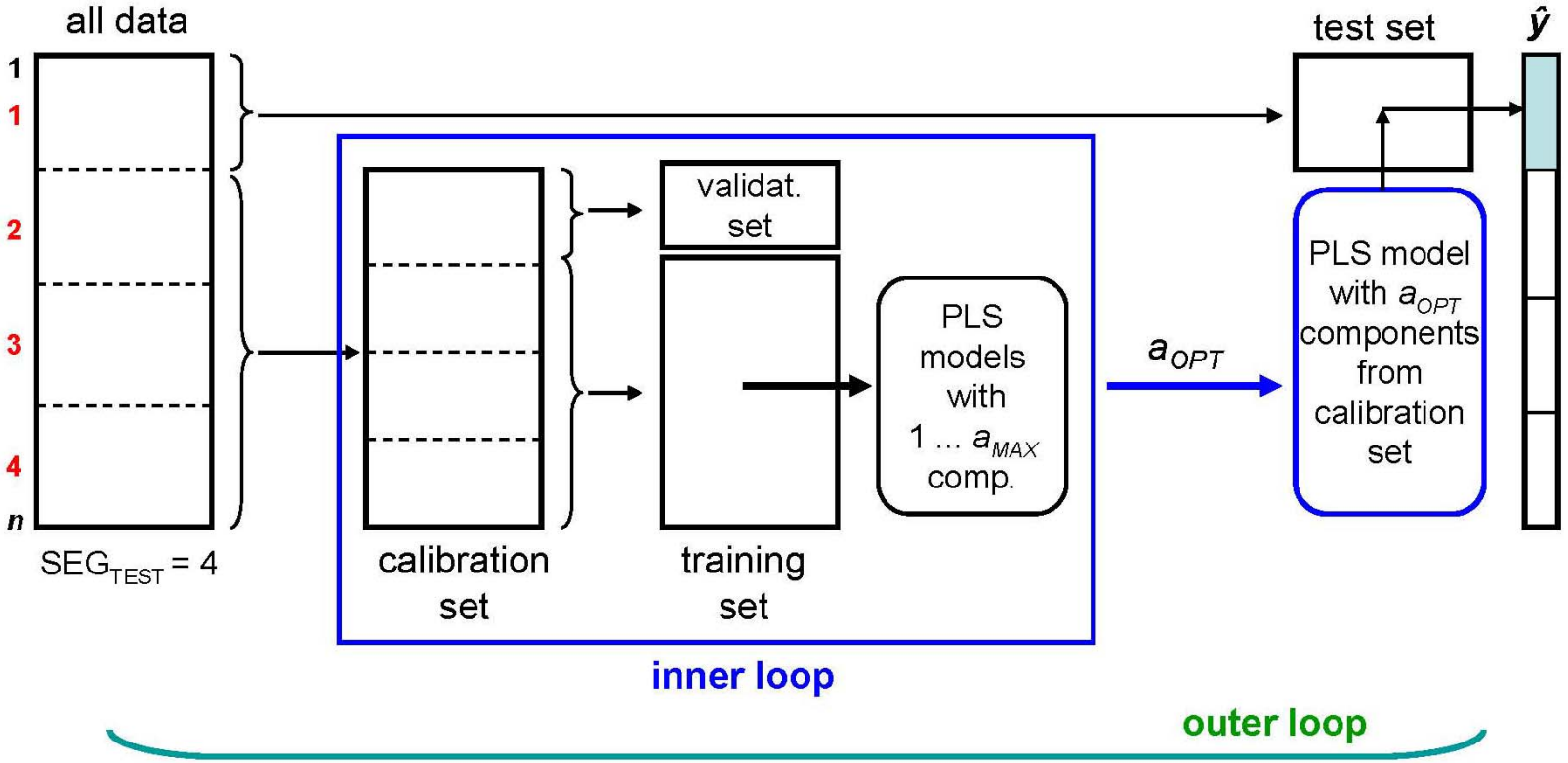
Estimate $\hat{y}$ prediction errors for **test** set

Inner CV loop

Estimate **optimum model complexity**, that is,

Estimate of the optimum number of PLS-components

Outer / inner loop of rdCV, schematically



Filzmoser, Liebmann, Varmuza:
Repeated Double Cross Validation.
Journal of Chemometrics, 23 (2009) 160-171

SOFTWARE for R: www.r-project.org (free)
Package ‚chemometrics' (Filzmoser et al.)
rdCV as function **‚mvr_dcv'**

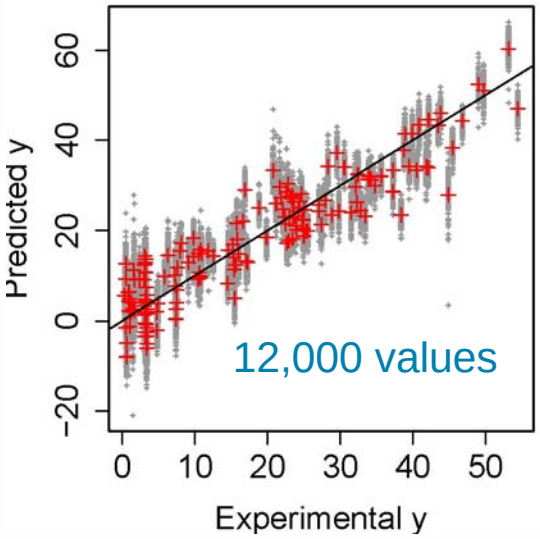
Repeated double cross validation (rdCV)

... for n = 120 samples, rdCV results in ...

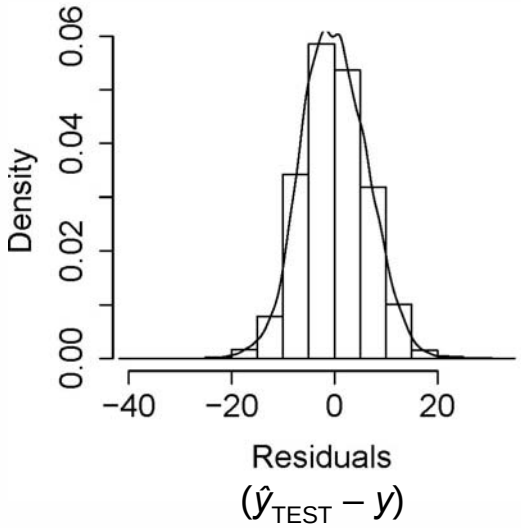
Repetition loop with n_{REP} = 100 repetitions

$n * n_{REP}$ = 120 * 100 = 12,000 predicted values $\hat{y}$ from test set samples

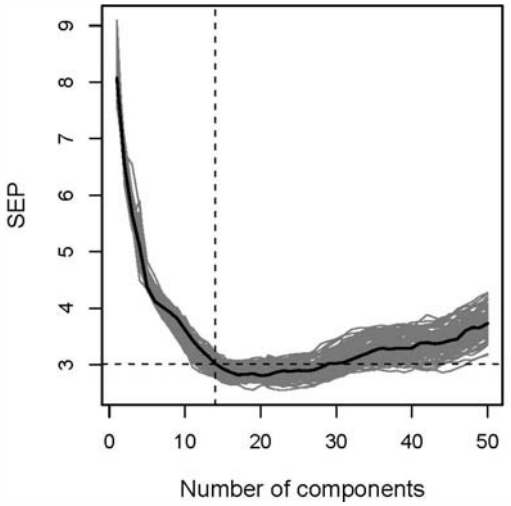
Predicted vs. experimental concentration y



Density distribution of errors



SEP versus model complexity

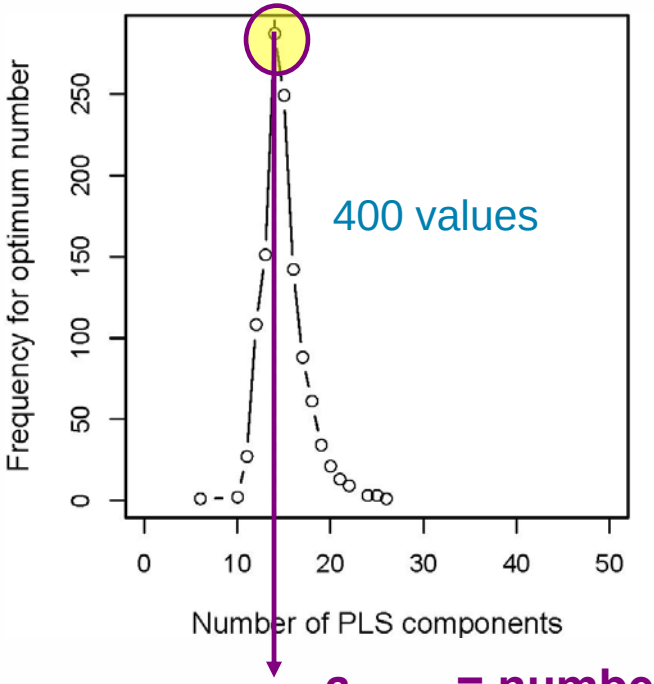


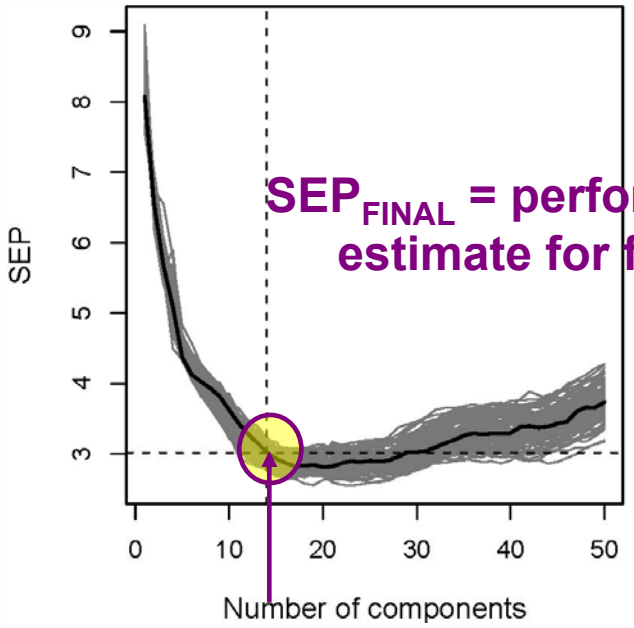
Repeated double cross validation (rdCV)

... for *n* = 120 samples, rdCV results in ...

Outer CV loop with SEG_{TEST} = 4 segments

n_{REP} * SEG_{TEST} = 100 * 4 = 400 values for ‚optimum complexity'
(= number of PLS components)

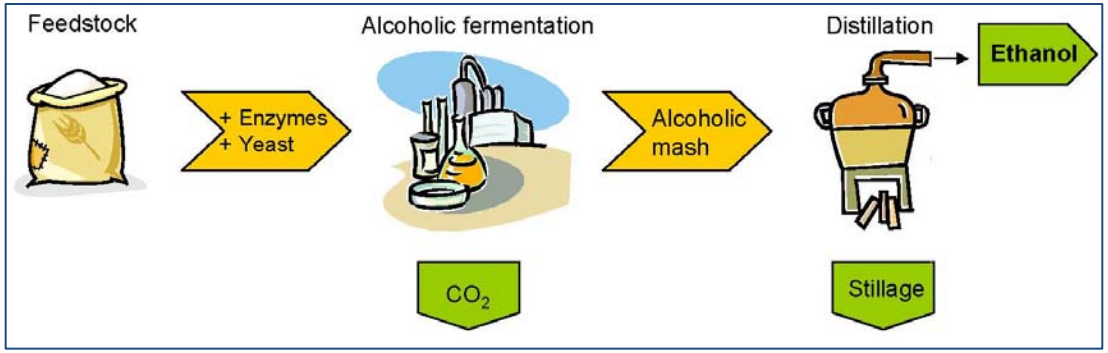




SEP_{FINAL} = performance estimate for future samples

a_{FINAL} = number of PLS components for very final regression model from all data

Selected results



rdCV+GA	95 % error interval g/l	R^2	concentration range in g/l
Mash			
Glucose	± 9.0	0.900	0-54
Ethanol	± 2.4	0.997	22-88
Glycerol	± 1.0	0.988	2-17
Stillage			
Glucose	± 3.4	0.949	0-24
Ethanol	± 1.6	0.998	0-58
Glycerol	± 1.2	0.941	3-14
Acetic Acid	± 0.4	0.461	0-1
Lactic Acid	± 0.2	0.812	0-1
Fructose	± 1.0	0.909	0-6
Maltose	± 0.8	0.938	0-6
Arabinose	± 0.2	0.938	0-1

Liebmann, Friedl, Varmuza: *Analytica Chimica Acta*, 642 (2009), 171-178
Liebmann, Friedl, Varmuza: *Biochemical Engineering Journal*, in prep.

$R^2 < 0.5$ w/o GA

Conclusions

- NIR spectroscopy was successfully applied:
  - Incoming grain analysis
  - Fermentation monitoring
  - Analysis of distillation residue
- Process implementation of NIR allows:
  - Fast analytical results, minimum sample presentation
  - Quantification in multi-constituent solutions
  - Determination of concentrations >> 1g/l
- Multivariate data analysis:
  - Validate NIR models thoroughly (rdCV)
  - ‚Good‘ reference values necessary
  - Incorporate sufficiently different calibration samples