Repeated Double Cross Validation

Kurt Varmuza

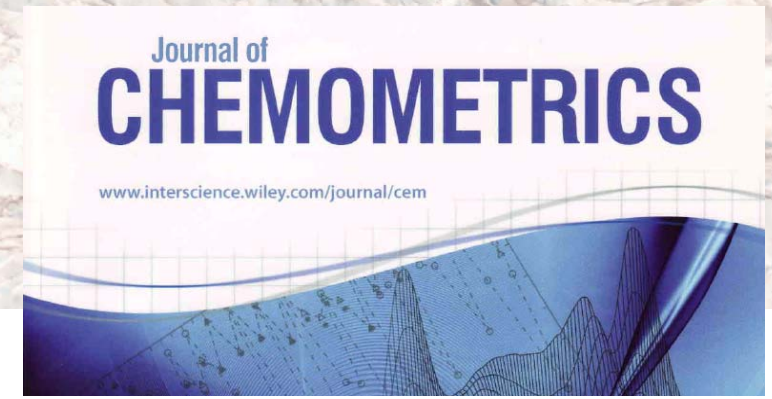
Vienna University of Technology
Institute of Chemical Engineering

Laboratory for ChemoMetrics

www.lcm.tuwien.ac.at

11th Scandinavian Symposium on Chemometrics - SSC 11, 10 June 2009
Loen, Norway, 8-11 June 2009


The tiny boat on front of the huge glacier (at the west coast of Greenland, visit 2007) may be considered as a symbol for the rather simple multivariate methods we apply to complicated chemical problems.
Greenland: Evighedsfjord (July 2007)



Research Article

Received: 7 November 2008, Revised: 22 December 2008, Accepted: 24 December 2008, Published online in Wiley InterScience: 11 February 2009

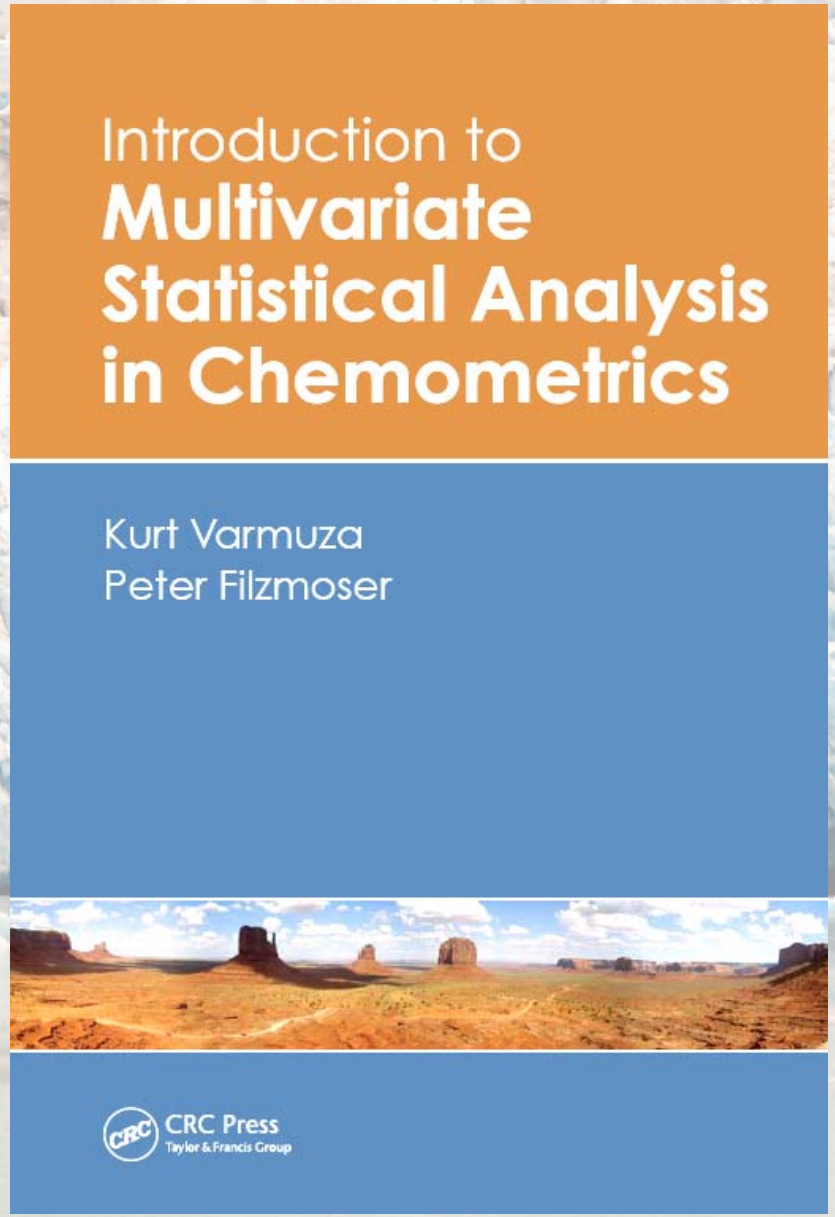
(www.interscience.wiley.com) DOI: 10.1002/cem.1225

Repeated double cross validation

Peter Filzmoser[a], Bettina Liebmann[b] and Kurt Varmuza[b]*

Journal of Chemometrics **23**, 160-167 (2009)

(Sorry, a new book ;-)



Book info:
www.lcm.tuwien.ac.at

R info (download):
www.r-project.org

Includes many R-codes,
R-package *chemometrics*

CRC Press, Taylor & Francis Group,
Boca Raton, FL, USA, 2009
ISBN: 9781420059472

Ca 320 pages, price: appr. US$ 110


1 Entry

Empirical models

Two essential tasks in the generation of empirical models

e.g. PLS regression models

- **optimal complexity of model**
  optimum number of PLS components
- **realistic estimation of the prediction performance**
  based on residuals (prediction errors)

☞ **repeated double Cross Validation (rdCV)**

Empirical models

Cross Validation

repeated double Cross Validation (rdCV)

Bootstrap techniques

Monte Carlo methods

Rather small number of objects (*n*) → Good estimation of prediction errors for new cases

Performance criteria

Statistical estimations of the prediction performance of a model are based on a set of prediction errors

$$e_i = y_i - \hat{y}_i$$

- **properly generated** → **test set objects**
- **many** → **rdCV, bootstrap, ...**
- **well summarized** → **numerical criteria, distribution**

Performance criteria

Standard deviation of the prediction errors
= Standard Error of Prediction (SEP)

$$SEP = \sqrt{[1/(z\text{-}1)] \sum (e_i - bias)^2} \qquad i = 1, \ldots, z$$

$$bias = [1/z] \sum e_i \qquad bias \approx 0$$

z number of available residuals (from test set objects)

IF the residuals are approximately normally distributed:
95% tolerance interval for prediction errors = **± 2 SEP**

ELSE
95% tolerance interval is defined by the
0.025 and 0.975 **quantiles** of the error distribution.


2 rdCV
FRAM
NARVIK
IMO 9370018
a useful engine
a powerful engine
==2: rdCV (Fram)

repeated double Cross Validation - rdCV

3 nested loops

- **Repetition loop** (with different random sequences of the objects).
  - **Outer CV loop**: Split into calibration sets and test sets, model from calibration set, estimation of prediction errors for test set objects.
    - **Inner CV loop**: Optimization of regression model by estimation of the optimum number of PLS-components.

n_{REP} repetitions ➡ $n_{REP} * n$ residuals from test sets

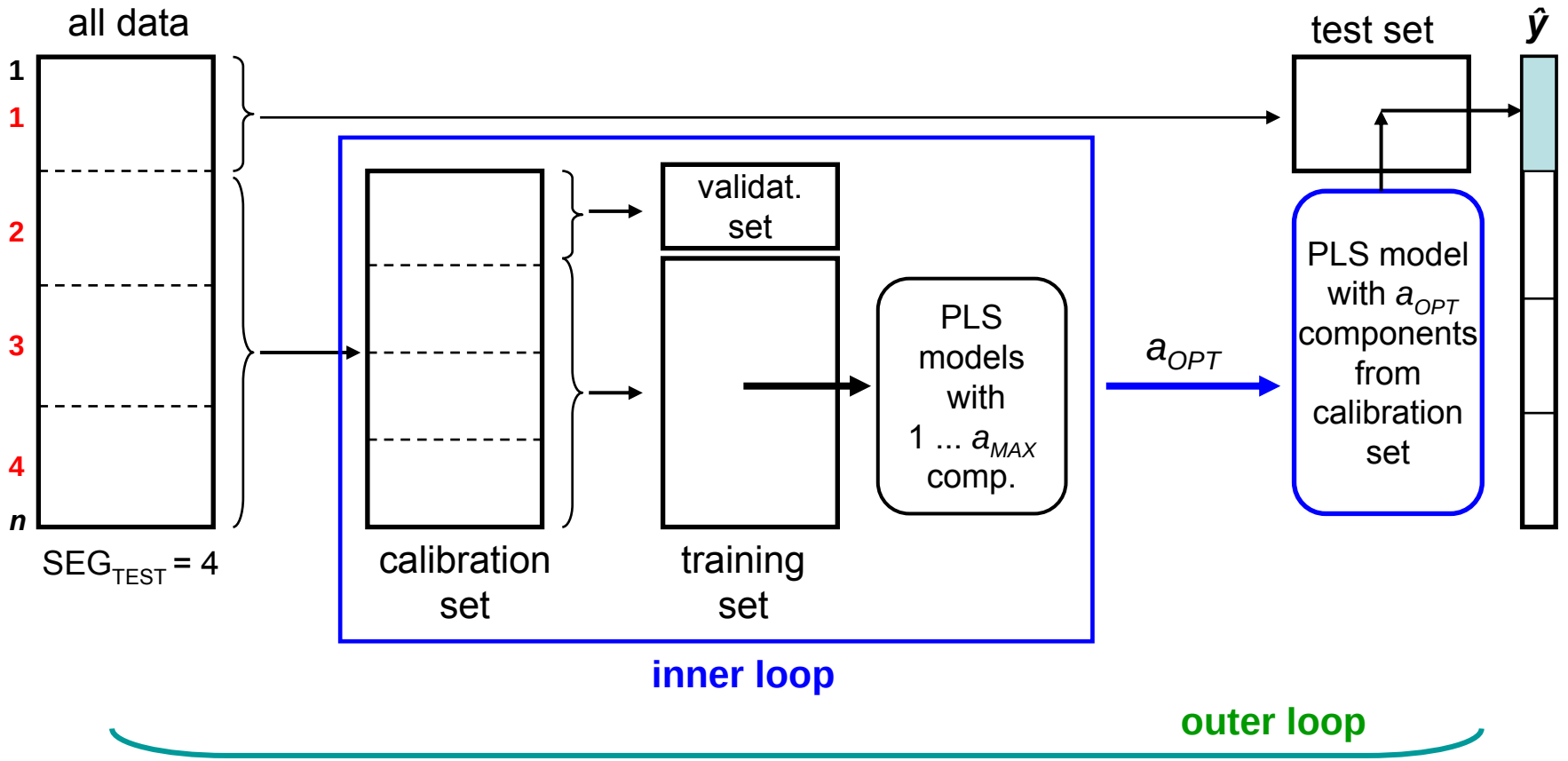
SEG_{TEST} segments in outer CV loop ➡ $n_{REP} * SEG_{TEST}$ values for a_{OPT}

Example:

$n = 70$, $n_{REP} = 100$, $SEG_{TEST} = 4$ ➡ **7000 residuals, 400 values for a_{OPT}**

loops

repeated double Cross Validation - rdCV



rdCV scheme

rdCV in R

R www.r-project.org (free)

R-Package *chemometrics* (Peter Filzmoser et al.)

```
library(chemometrics)
X = ...
y = ...
result = go_rdcv(X,y,PDFfile="myfile.pdf")
```

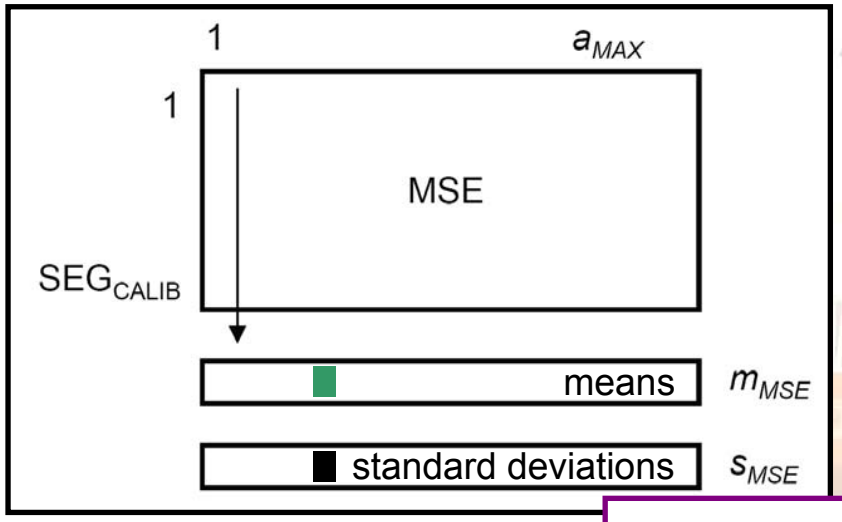
For *data with n*=200 and *m*=500: computation time typ. 4 minutes, including some diagnostic plots;

rdCV function *mvr_dcv* in *chemometrics*

PLS Mevik B.H., Wehrens R., J. Stat. Software 2007, 18(2), 2007

Standard error method (a_{OPT} estimation)

Inner CV loop: MSE for each segment



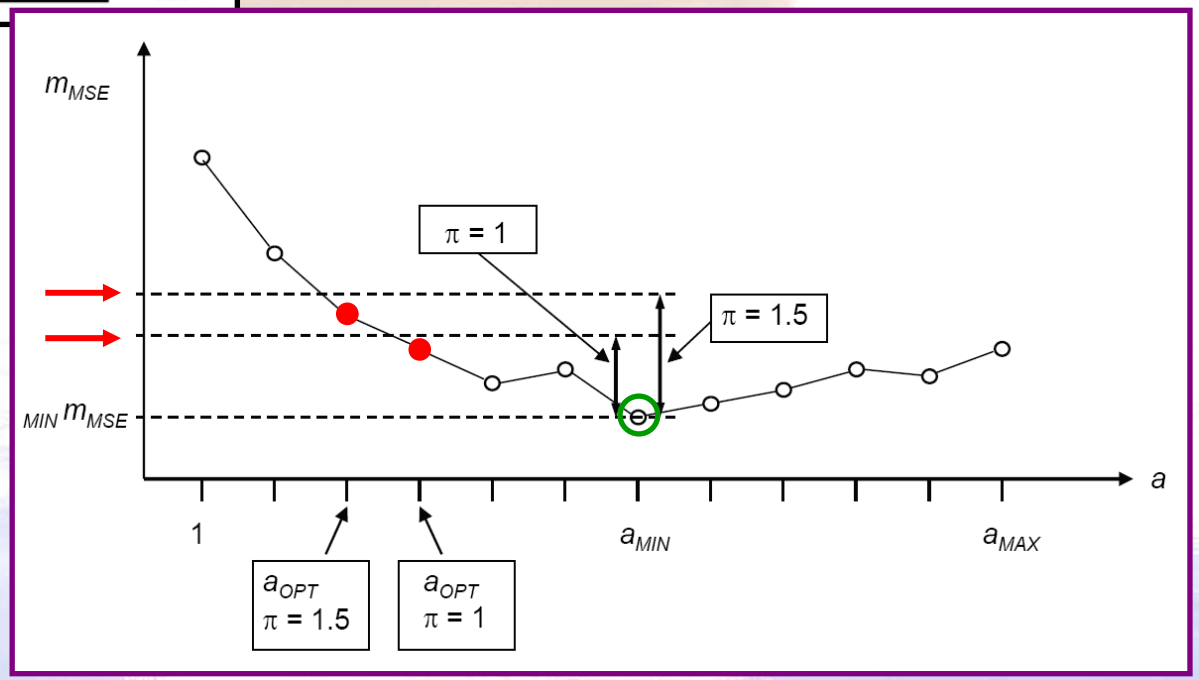
a_{OPT} = minimum value of a with

$$m_{MSE} < {}_{MIN}m_{MSE} + \pi \; {}_{MIN}s_{MSE} / \sqrt{SEG_{CALIB}}$$

π parsimony factor

$\pi = 0$ global minimum

$\pi = 2$ 95% confidence interval



Based on "one standard error method" described in Hastie T.,Tibshirani R.J., Friedman J.: The Elements of Statistical Learning, Springer (2001)

rdCV results evaluation: a_{OPT}

SEG_{TEST} (segments in outer loop)

×

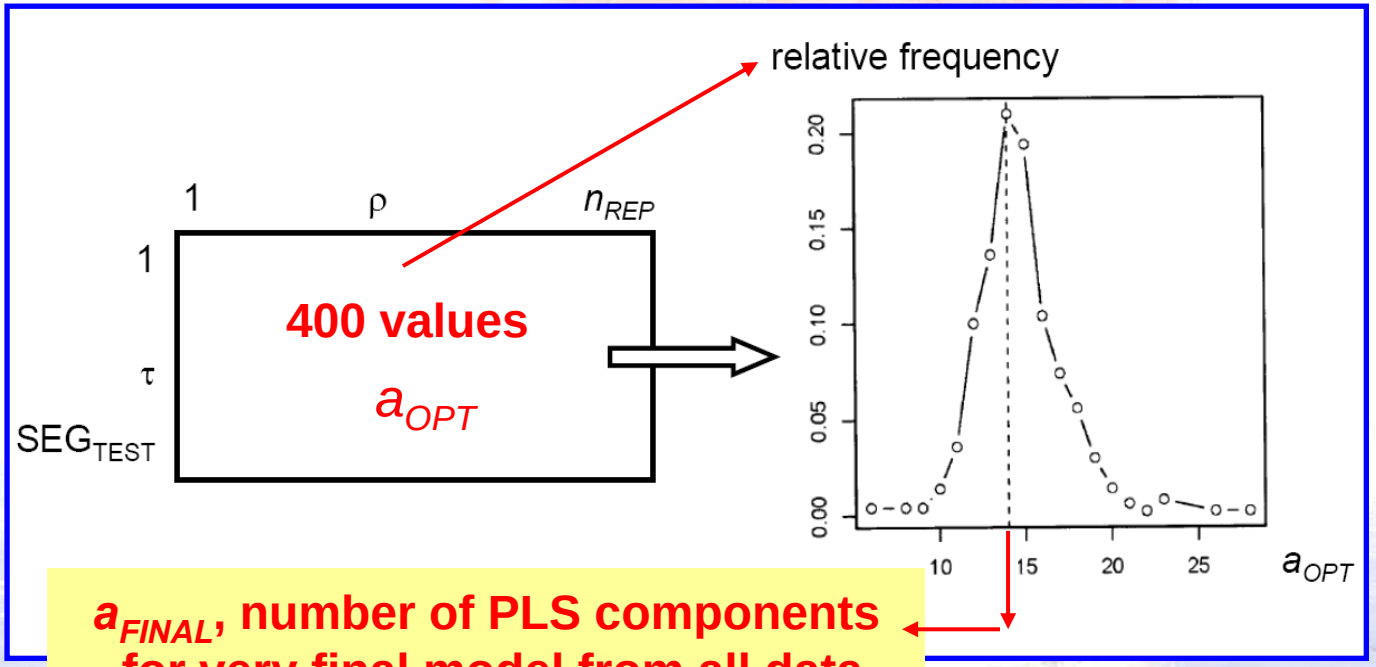
n_{REP} (repetitions)

= number of a_{OPT} values

Determination of glucose in fermentation mash samples by NIR, n = 120, m = 235

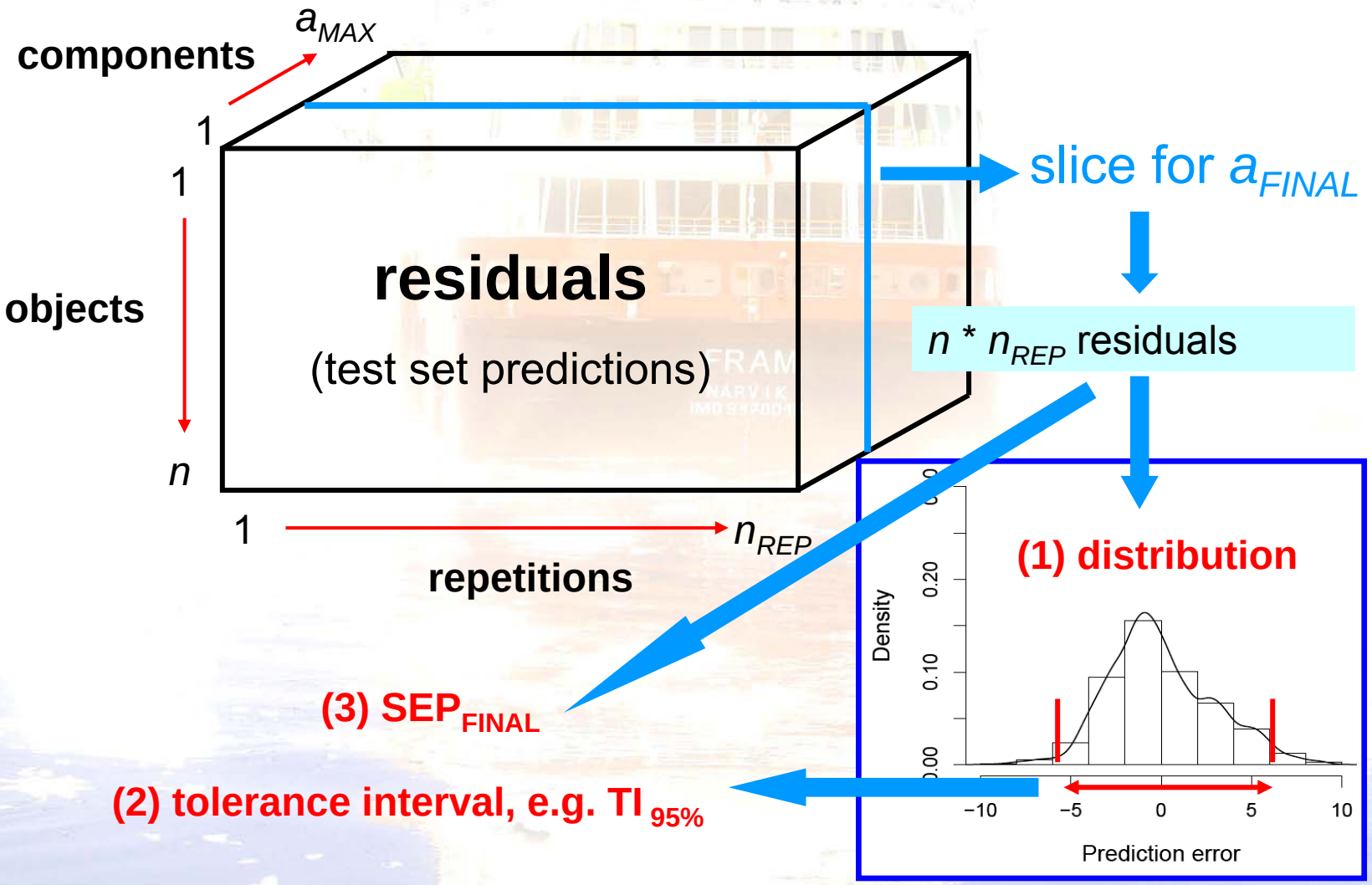
n_{REP} = 100

SEG_{TEST} = 4

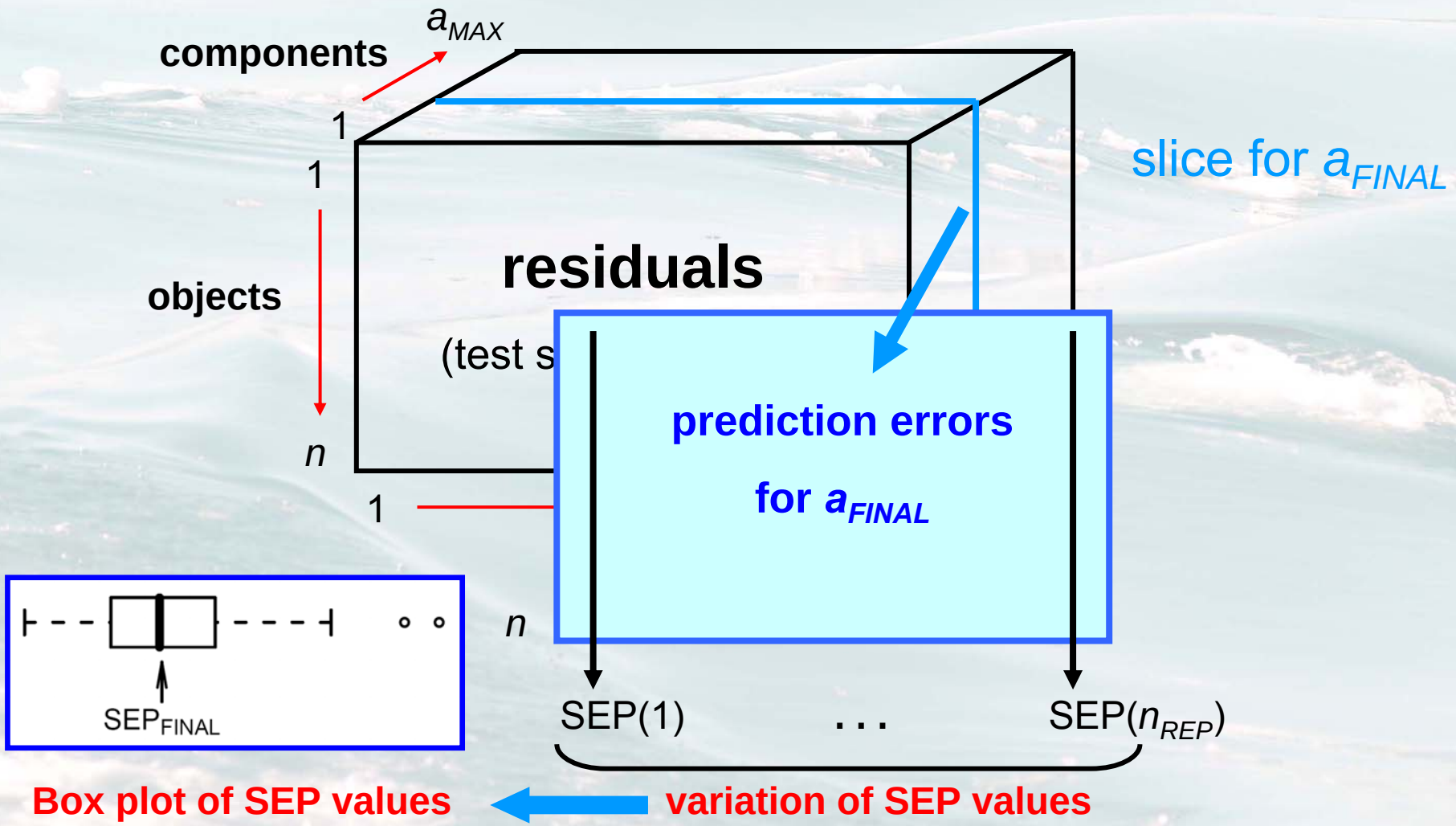


a_{FINAL}, number of PLS components for very final model from all data

rdCV results evaluation: residuals (SEP)

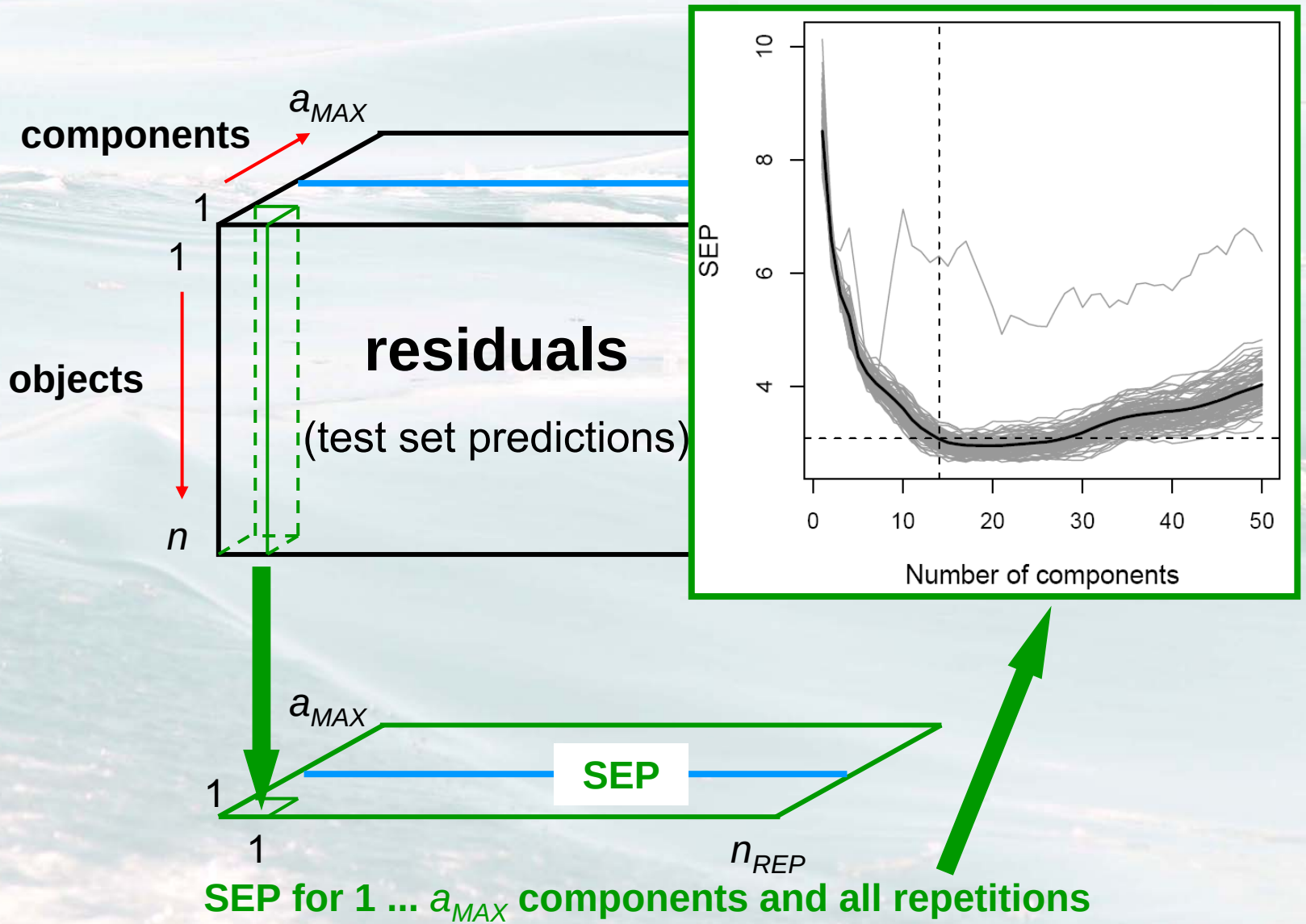


rdCV results evaluation: residuals (SEP)



SEP Boxplot

rdCV results evaluation: residuals (SEP)



3 Applications

Number of segments and repetitions

Four data sets

- *n* = 120, *m* = 235 or *m* = 15 (GA selected)
  Glucose in fermentation mash samples
  (NIR 1100 - 2300 nm)
- *n* = 209, *m* = 467 or *m* = 13 (GA selected)
  GC retention index from molecular descriptors

rdCV

- Outer CV loop (test sets - calibration sets)
  SEG_{TEST} = 4, 5, 7, 10
- Inner CV loop (optimum no. of PLS components)
  SEG_{CALIB} = 4, 5, 7, 10
- Number of repetitions = 5, 20, 100

Number of segments and repetitions

Glucose (NIR)

m = 235

SEG_{TEST}	a_{FINAL}				SEP_{FINAL}			
10	14	15	14	14	3.0	2.9	2.9	3.0
7	14	15	14	14	3.0	2.9	3.0	3.0
5	14	15	14	14	3.1	3.0	3.1	3.1
4	14	14	14	14	3.1	3.1	3.1	3.2

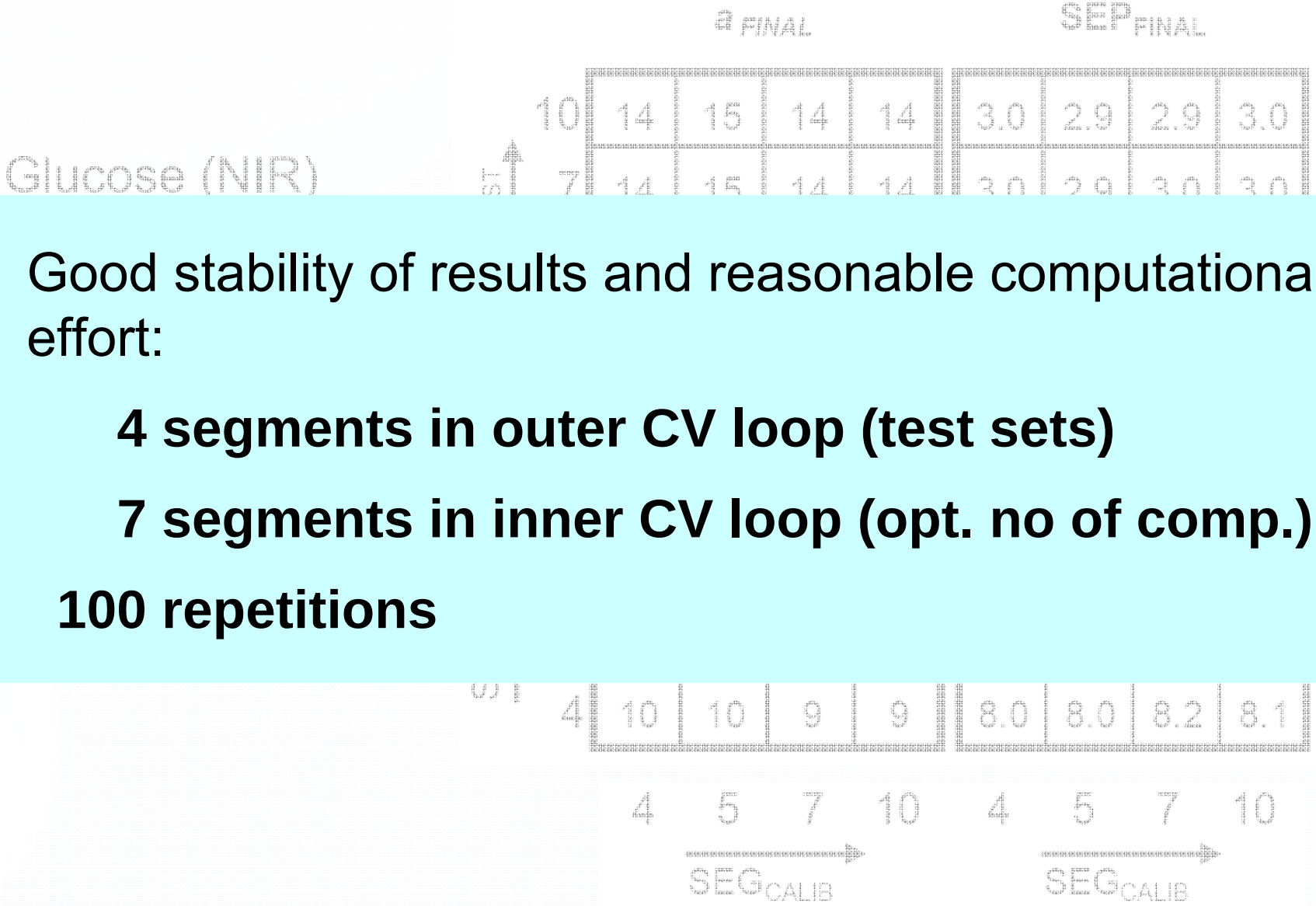
Retention index

m = 13

SEG_{TEST}								
10	9	9	9	9	7.9	7.9	7.9	8.0
7	9	9	9	9	8.0	8.0	8.0	8.0
5	9	10	9	9	8.0	8.0	8.1	8.1
4	10	10	9	9	8.0	8.0	8.2	8.1
	4	5	7	10	4	5	7	10
	SEG_{CALIB}				SEG_{CALIB}			

no segm results

Number of segments and repetitions



Good stability of results and reasonable computational effort:

4 segments in outer CV loop (test sets)

7 segments in inner CV loop (opt. no of comp.)

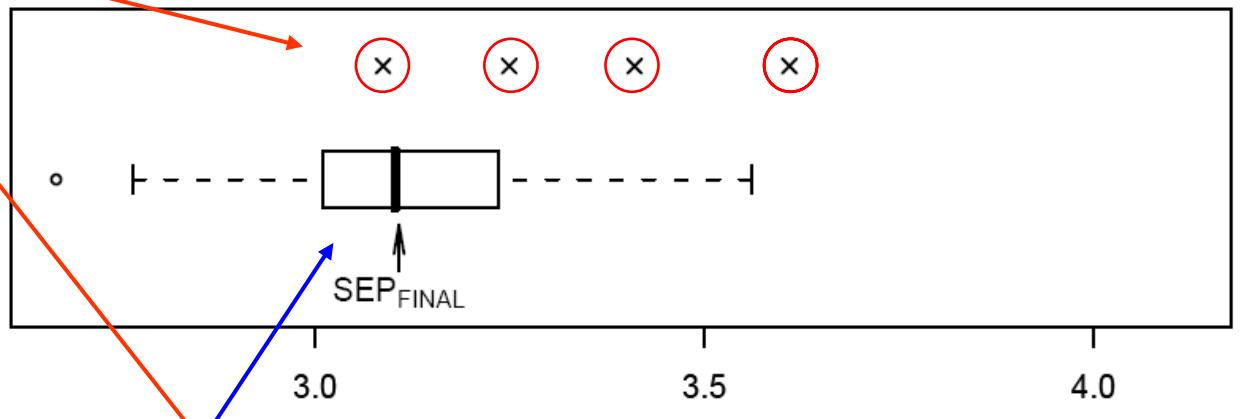
100 repetitions

Variation of SEP

single CV with 4 test sets: SEG_{CALIB} = 7

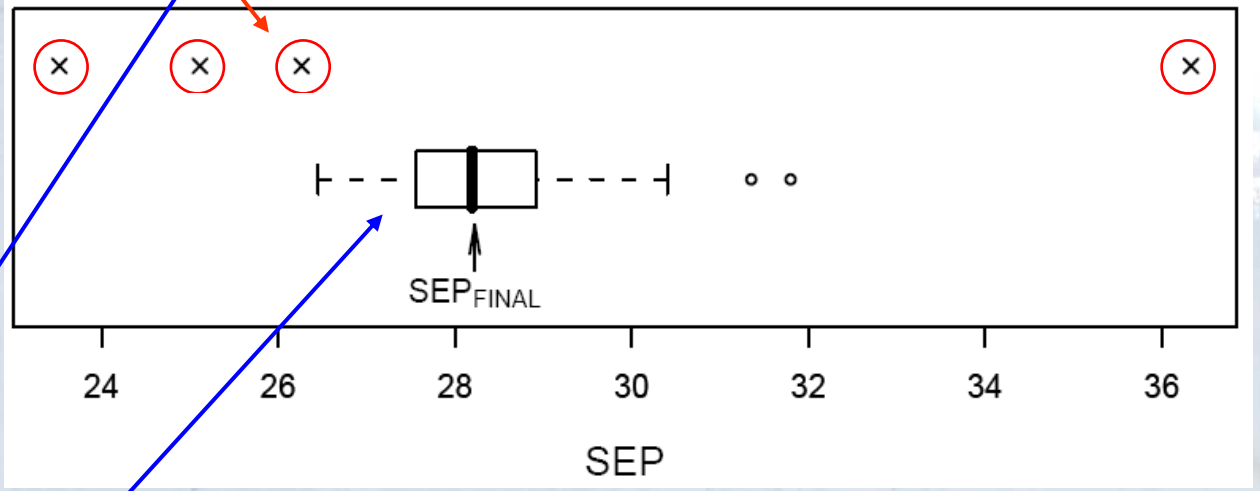
Glucose (NIR)

m = 235



Retention index

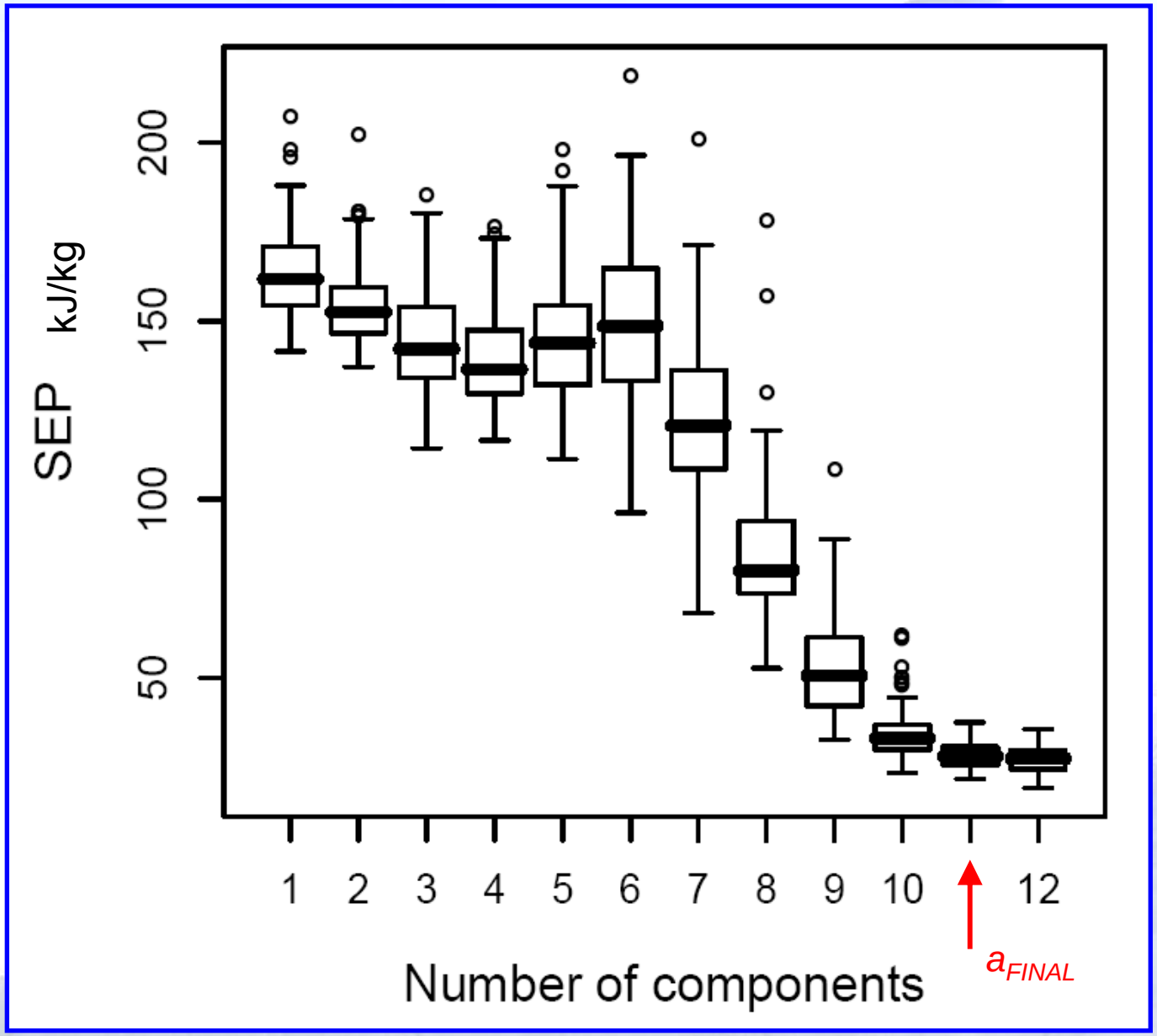
m = 13



rdCV: n_{REP} = 100, SEG_{TEST} = 4, SEG_{CALIB} = 7, box plots from 100 SEP values

no segm results

Variation of SEP with no. of components



Example

n = 35 samples (cereals, wood)

m = 12 NIR abs. (selected from 435 by GA)

y heating value exp. (18.1 - 19.1 MJ/kg)

rdCV

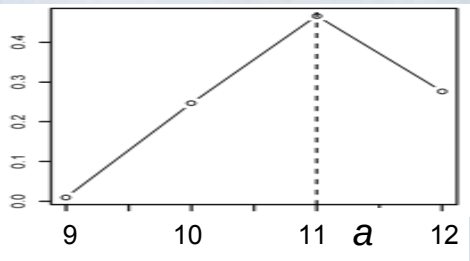
SEG_{TEST} = 3

SEG_{CALIB} = 7

100 repetitions

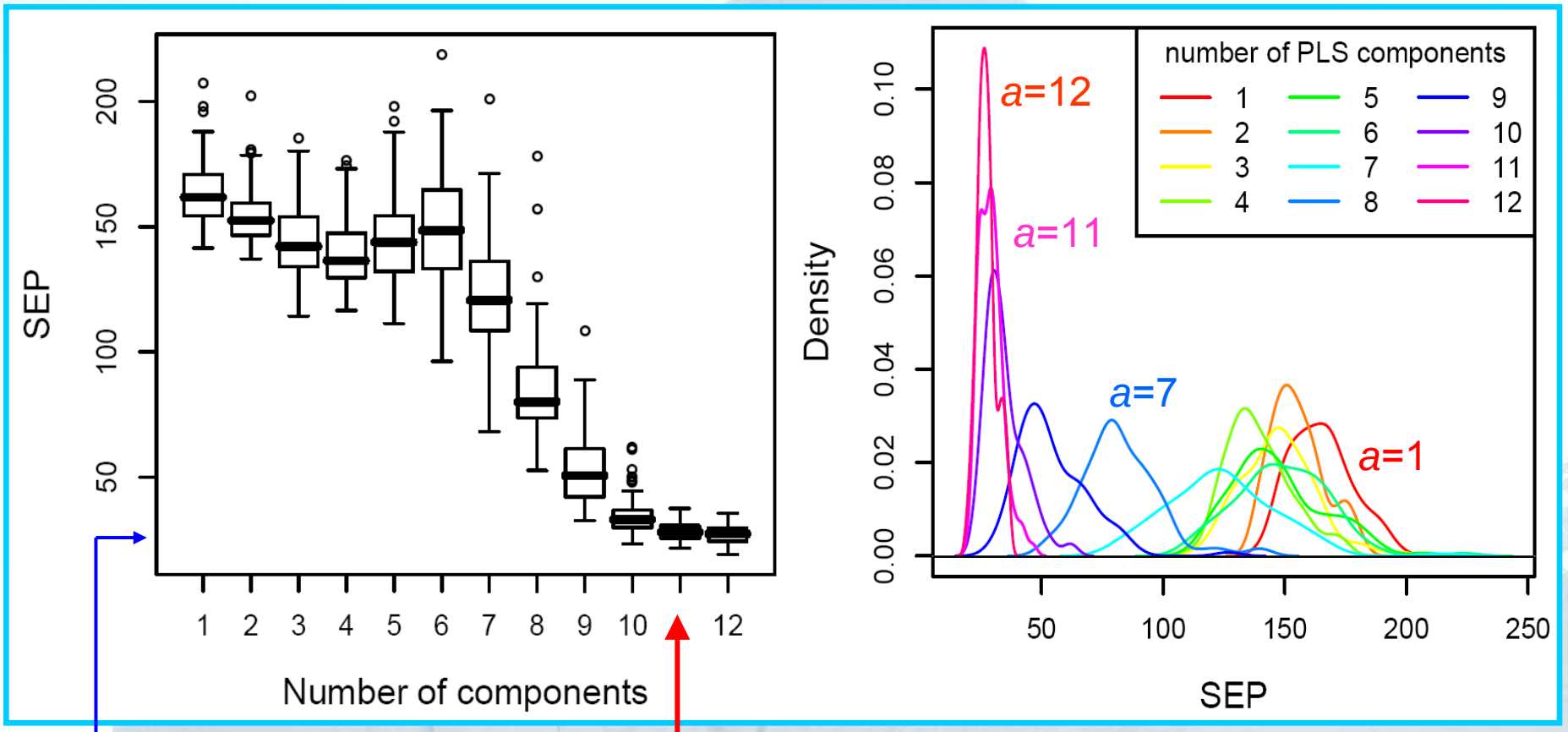
a_{FINAL} = 11

SEP_{FINAL} = 29 kJ/kg



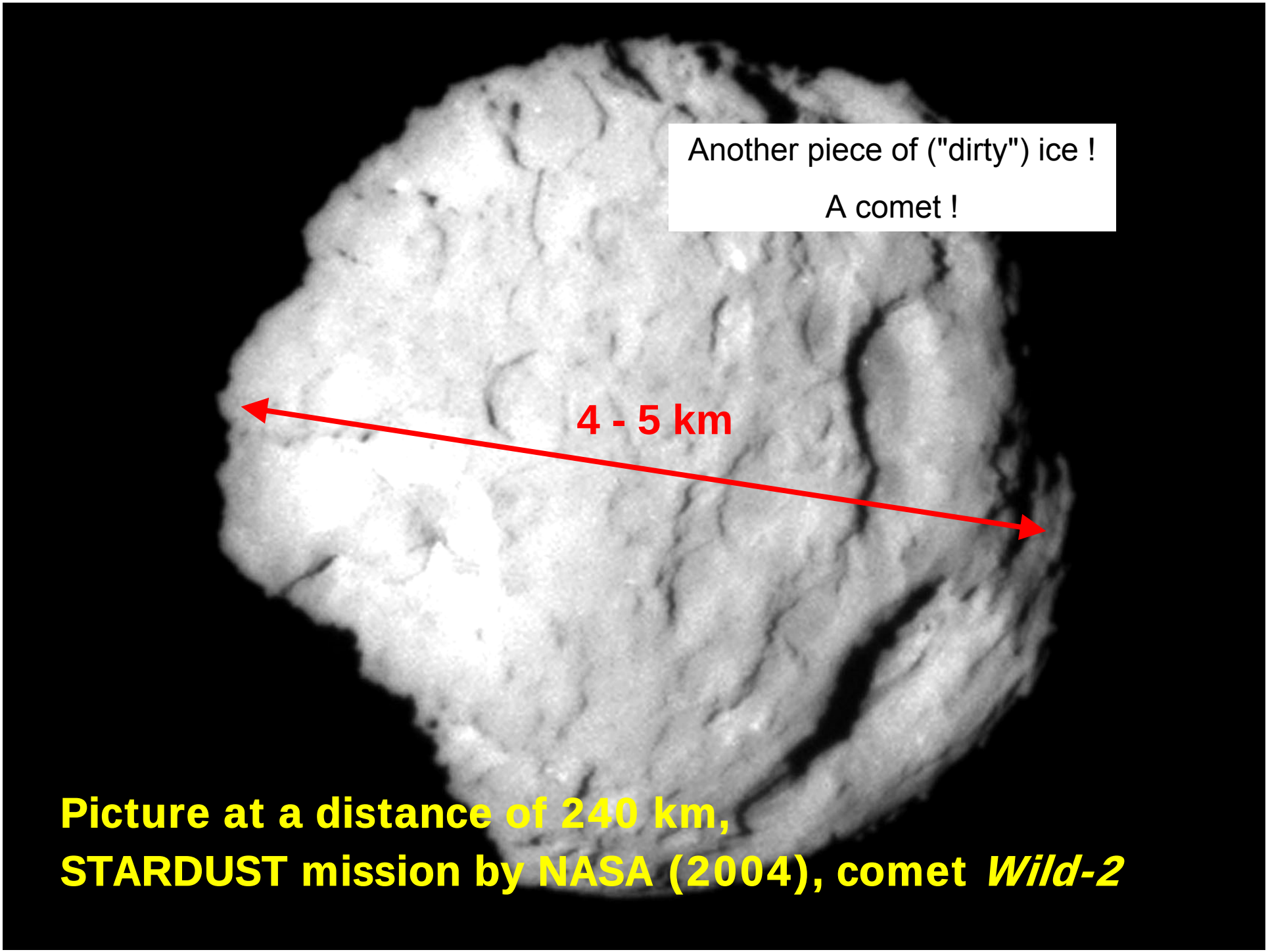
var. SEP boxplots

Variation of SEP with no. of components



a_{FINAL}

SEP_{FINAL} = 29 kJ/kg


Another piece of ("dirty") ice !
A comet !
4 - 5 km
Picture at a distance of 240 km,
STARDUST mission by NASA (2004), comet *Wild-2*

ESA mission Rosetta

European Space Agency project "Rosetta":

First spacecraft to orbit a comet

2.8 x 2.1 x 2.0 m,
two 14 m solar panels,
launch mass 3000 kg,

11 instruments in orbiter*,
9 instruments in lander,

launch 2 March 2004,
arrival at comet May 2014

***Cosima**
TOF-SIMS mass spectrometer,
collection and analysis of dust

Preliminary cosmo chemistry

n = **53 organic reference compounds**

condensed benzene rings, N-aromatic, purines , ...

Mass spectra measured on a similar instrument

Werther, Demuth, Krueger, Kissel, Schmid, Varmuza: *J. Chemom.* **16**, 99 (2002)

y = %N = f (mass spectral data)

This pioneering paper on a similar topic appeared 40 years ago; in the same year the first man was on the moon.

Jurs P.C., Kowalski B.R., Isenhour T.L.: *Anal. Chem.* **41**, 21 (1969)

Computerized learning machines applied to chemical problems.
Molecular formula determination from low resolution mass spectrometry.

Preliminary cosmo chemistry

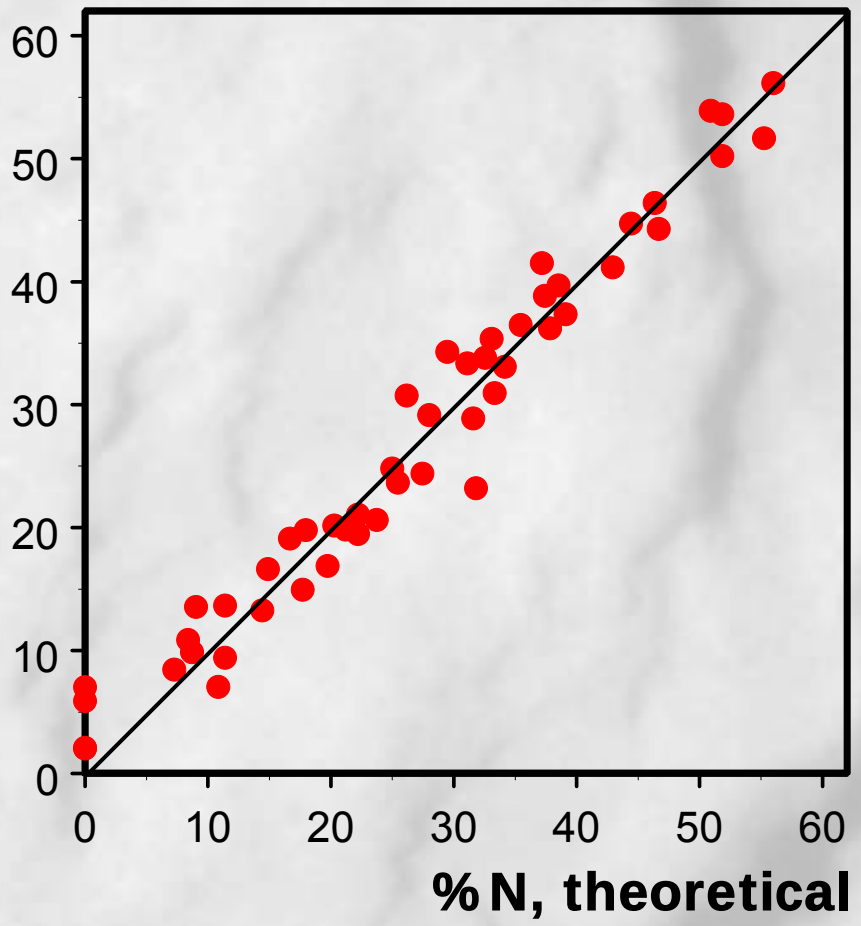
%N = f (mass spectral data)

PLS models with
m = 15 spectral features
selected by GA from 658

rdCV with 10 repetitions

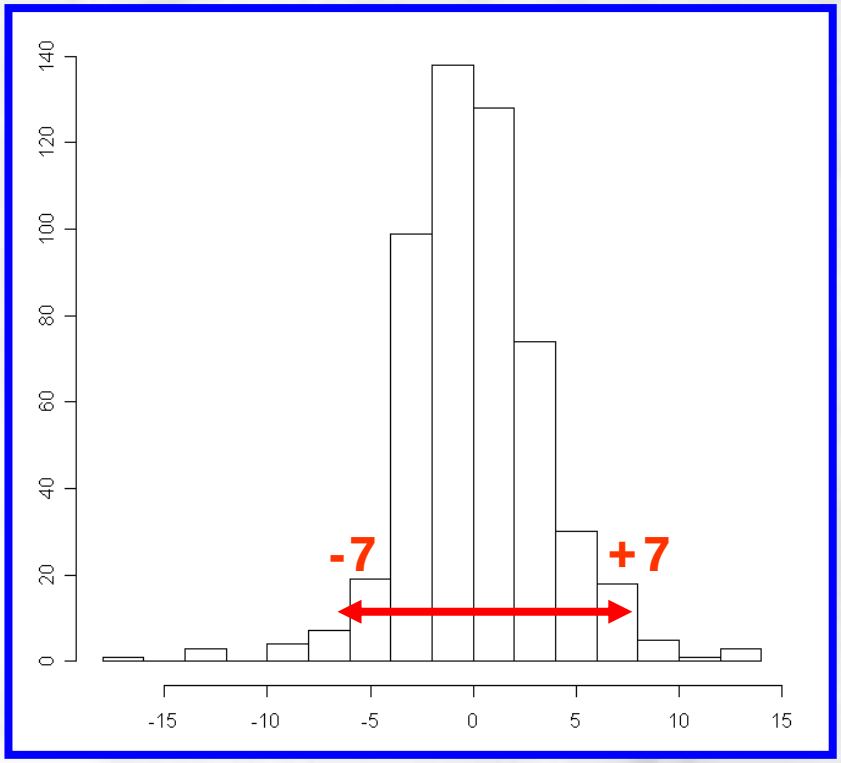
SEP = 3.4 %N

R^2 = 0.967



Preliminary cosmo chemistry

%N = f (mass spectral data)

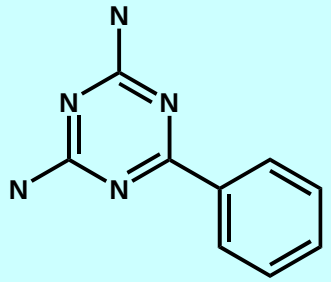


Histogramm of prediction errors
(530 values)
95% tolerance interval: **±7 %N**

tolerance interval ±7 %N

$C_9H_9N_5$ *	**37 %N**
$C_{10}H_{11}N_4$	30 %N
$C_8H_7N_6$	45 %N

* e.g. 2,4-diamino-6-phenyl-1,3,5-triazine (mw 187)




4 Summary
... for a stable arch

Summary

Many residuals (prediction errors)
- from test sets - are better than only a few:

- **distribution of residuals,**
- **tolerance interval for expected prediction errors**

Many values for the optimum number of (PLS) components are better than only a few:

- **variability of this parameter for model complexity**

Summary

Many values for SEP are better that only a few:

- **variability of this parameter for model performance,**
- **spread to be considered when comparing model performances,**
- **distribution of SEP for varying model complexity**

Better knowing
the uncertainties
than relying on
a few single numbers


Thank you