Validation of Multivariate Empirical Models

Kurt Varmuza

Vienna University of Technology

Institute of Chemical Engineering

Laboratory for ChemoMetrics

www.lcm.tuwien.ac.at

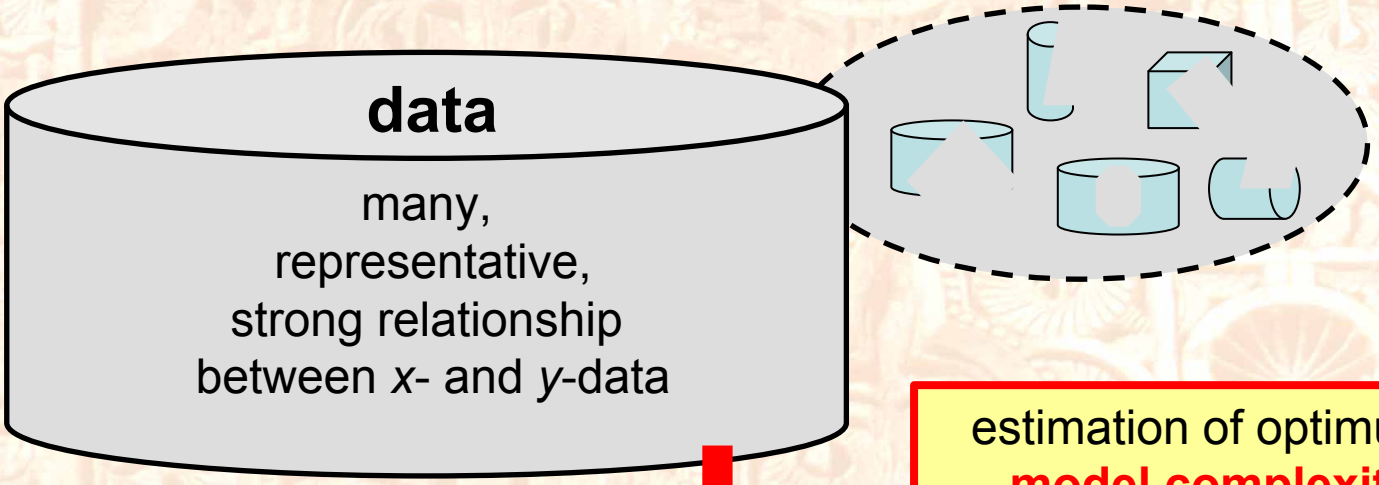
Vienne (Autriche)

First African-European Conference on Chemometrics

Rabat, Morocco, 20-24 September 2010 (Lecture 21 September 2010)

CONTENTS

The prominent task in chemometrics ...



only a few, and often not so good data

estimation of optimum **model complexity**

estimation of prediction errors of *y* for new cases **model performance**

several *x*-variables → **empirical model** $\hat{y} = f(\boldsymbol{x})$ → predicted $\hat{y}$

Final aim: low prediction errors of *y* for new cases

Life is rather easy with

many and "friendly" data

no outliers,

good (linear) relationship between the *x*'s and *y*

> good models,
>
> small uncertainties of estimated
>
- optimum model complexity,
- model performance

Strategy (split of data set) for model generation and validation not so crucial.

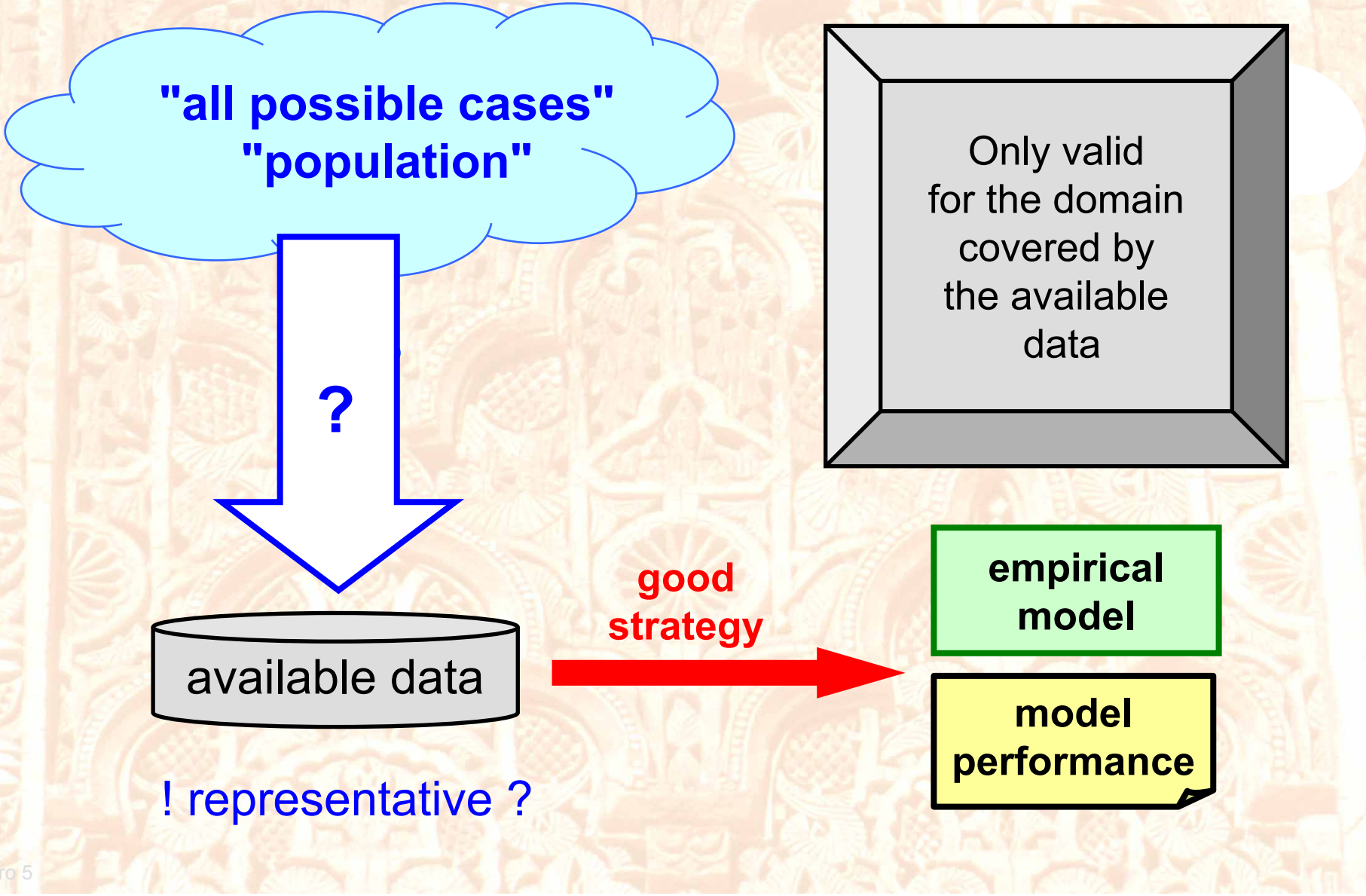
Life is troublesome and necessarily uncertain with

- small data sets (often not representative) ,
- "unfriendly" data

> not so good models,
>
> large uncertainties of estimated
>
- optimum model complexity,
- model performance

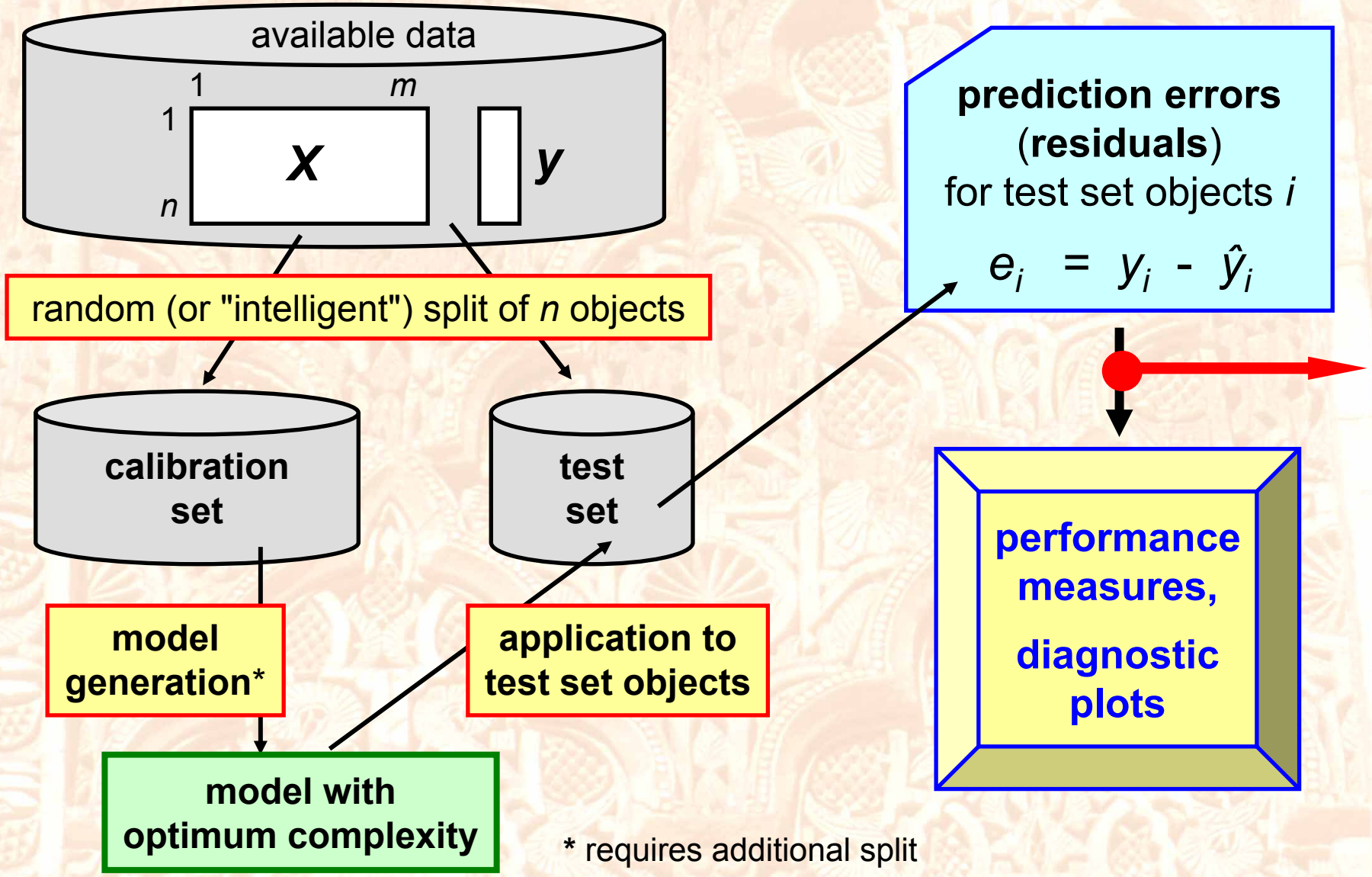
Chemometrics has to live with this troublesome situation.
On the other hand: more interesting, controversy, open-end, ...

Another fundamental problem with empirical models



CONTENTS

Basic strategy for model validation



Data set PAC

Polycyclic Aromatic Compounds (QSPR)

n = **209 chemical structures from polycyclic aromatic compounds**, approx. 3D, all H-atoms; *Corina* [1]

y **gas-chromatographic retention index**; *Lee* et al. [2]

X_1 m_1 = **467 molecular descriptors**; *Dragon* [3]

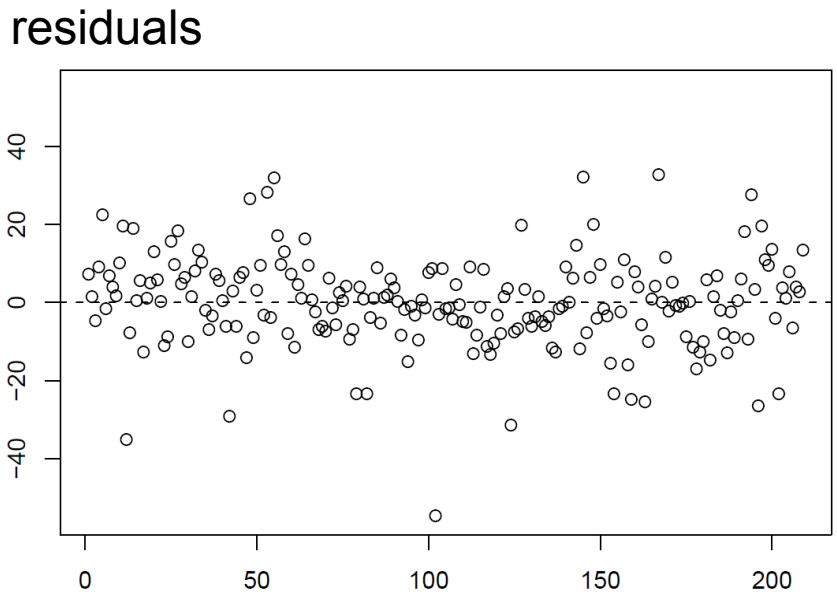
X_2 m_2 = **13 molecular descriptors selected**; genetic algorithm; from all data; *MobyDigs* [4]

[1] Corina software, Molecular Networks GmbH Computerchemie, www.mol-net.de, Erlangen, Germany (2004).
[2] Lee M.L., et al., Anal. Chem. 51 (1979) 768-773.
[3] Dragon software, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004).
[4] MobyDigs software, 1.0. Talete srl, www.talete.mi.it, Milan, Italy (2004)

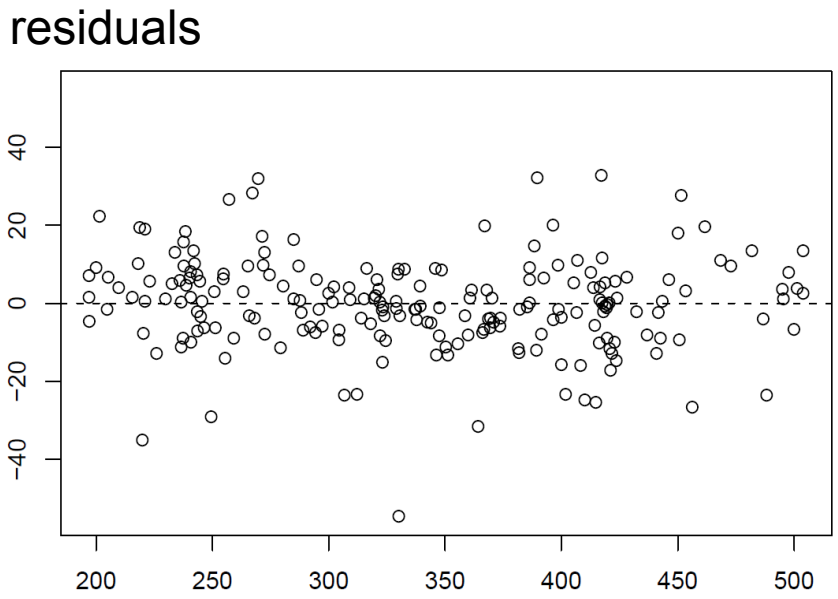
Use of prediction errors (residuals), e_i

"properly generated (for test set objects), and many"

residuals versus object no.



residuals versus *y*

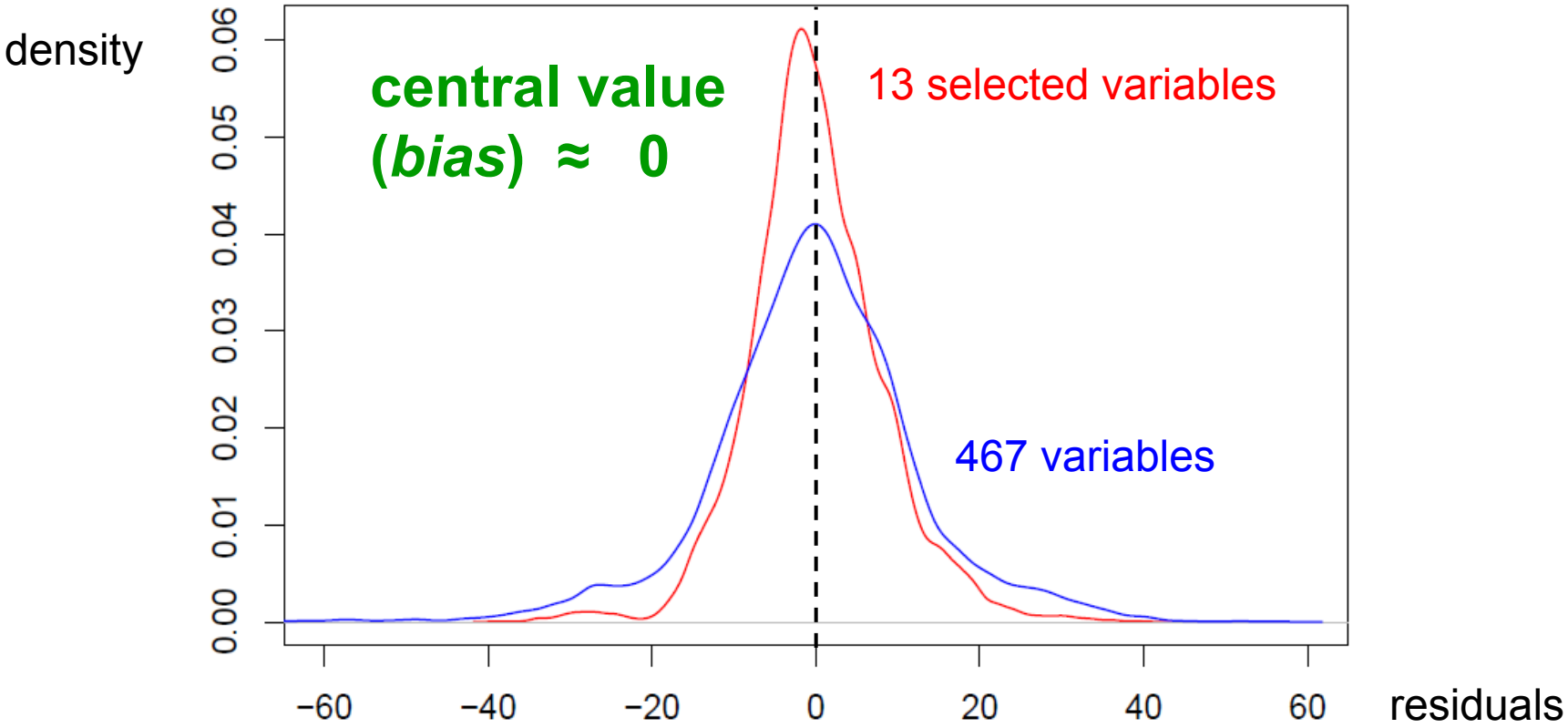


PAC data, n = 208, m = 467 variables, mean of 100 repetitions (rdCV)

Use of prediction errors (residuals), e_i

"properly generated (for test set objects), and many"

distribution of prediction errors

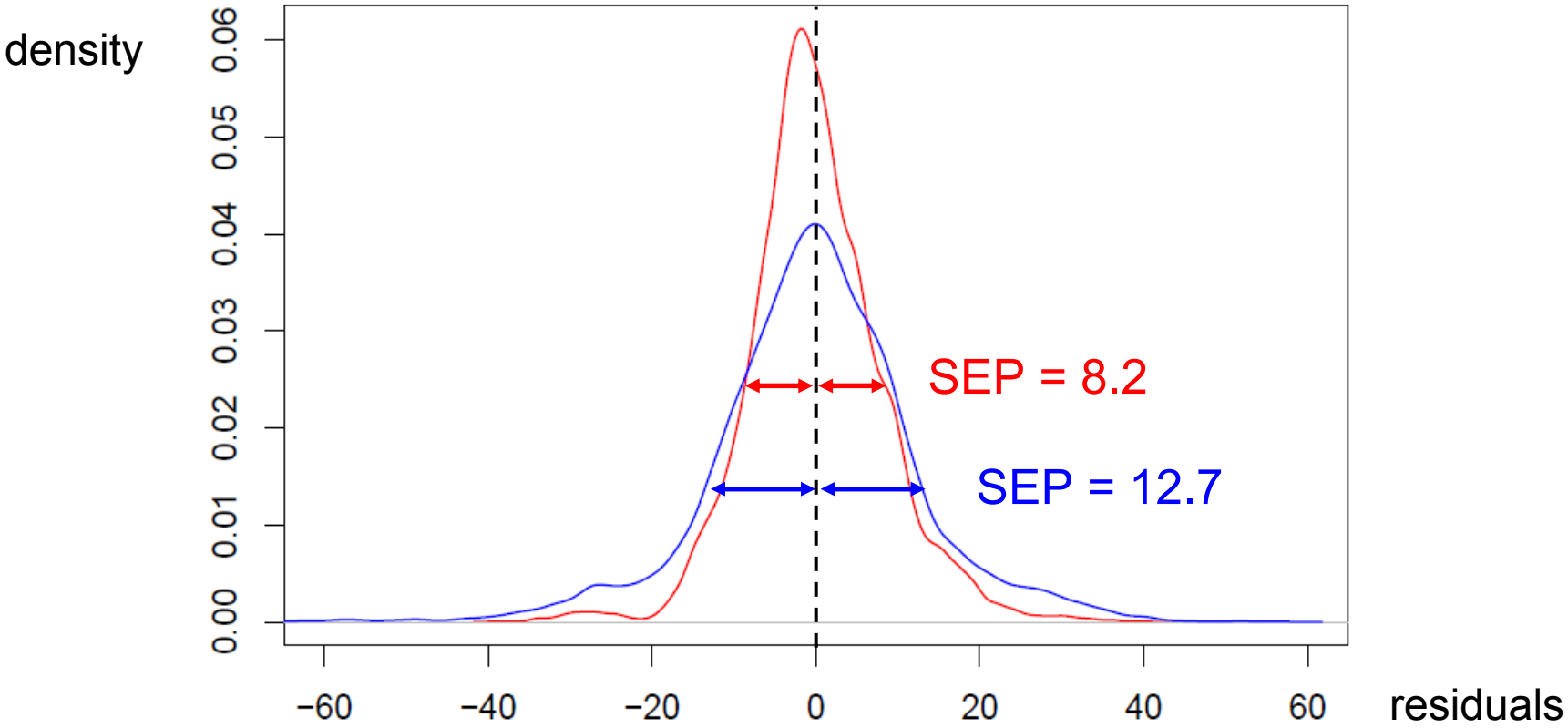


PAC data, n = 208, m = 467 or 13 variables, 100 repetitions (rdCV),
20,800 test set residual for each distribution

Use of prediction errors (residuals), e_i

"properly generated (for test set objects), and many"

standard deviation of residuals
(SEP, *standard error of prediction*)



PAC data, n = 208, m = 467 or 13 variables, 100 repetitions (rdCV),
20,800 test set residual for each distribution

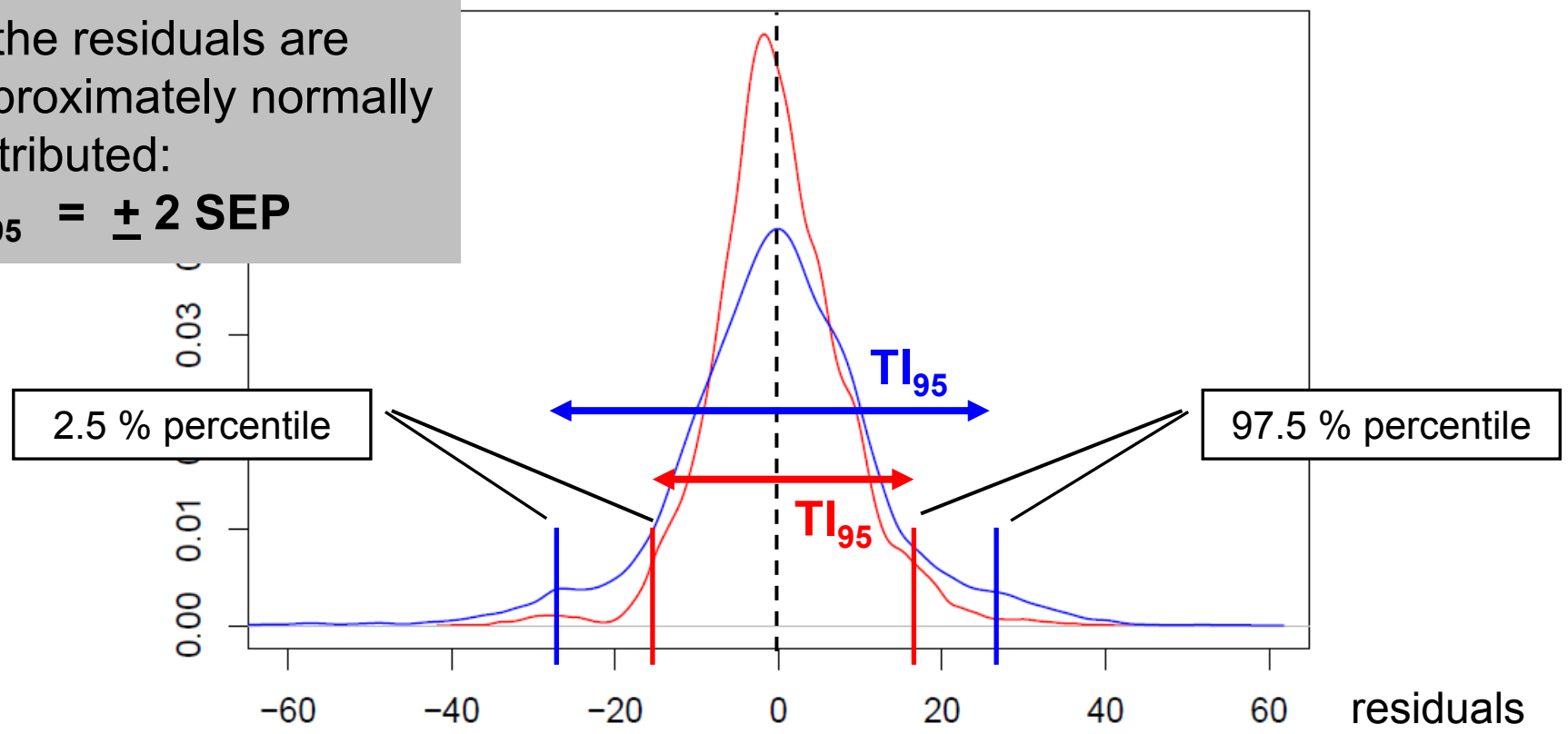
Use of prediction errors (residuals), e_i

"properly generated (for test set objects), and many"

95%-tolerance interval (TI_{95}) for prediction errors

IF the residuals are approximately normally distributed:

TI_{95} = ± 2 SEP

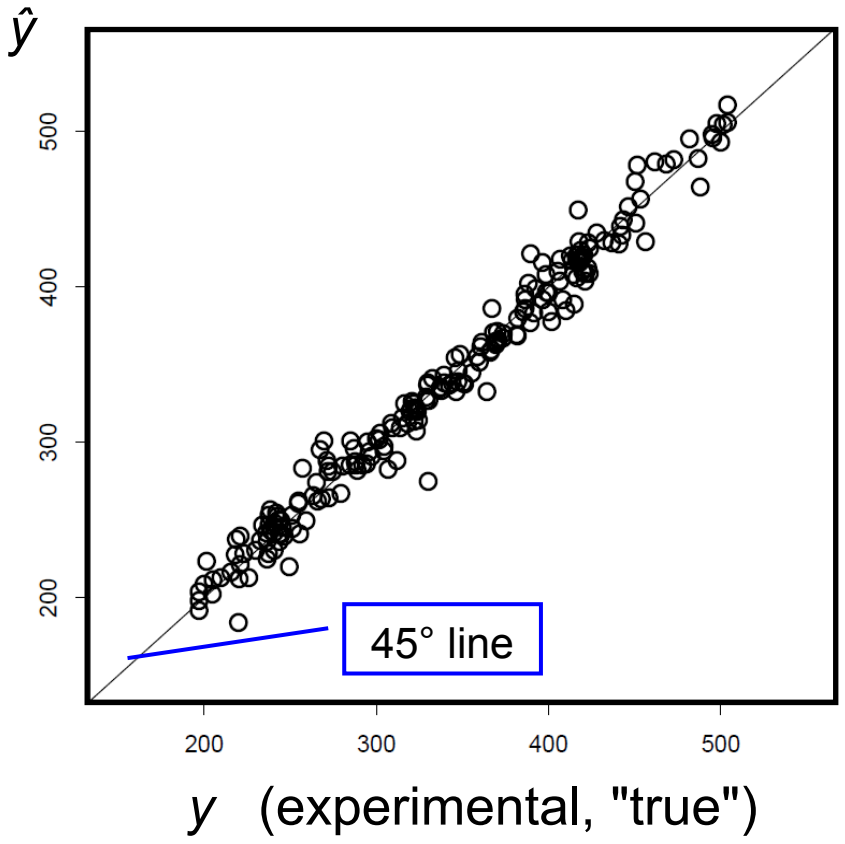


PAC data, *n* = 208, *m* = 467 or 13 variables, 100 repetitions (rdCV),
20,800 test set residual for each distribution

Use of predicted *y* 's ($\hat{y}$)

"properly generated (for test set objects), and many"

$\hat{y}$ versus *y*



- **Pearson correlation coefficient, squared,** $R^2 = 0.9785$
- **robust measures for correlation**
- **adjusted correlation coefficient,**

  $_{ADJ}R^2 = 1 - (n\text{-}1)(1\text{-}R^2)/(n\text{-}m\text{-}1)$

  considers the number of variables (*m*)

  others: AIC, BIC, Cp

PAC data, *n* = 208, *m* = 467 variables, mean of 100 repetitions (rdCV)

Comparison of performance criteria

💣 *"Trivial, but often ignored"* 💣

A comparison of models (e.g. by SEP)
is only reasonable if also
the variability of the criterion is estimated !

☹ A single split of the objects into a calibration set and a test set, or simple CV methods give only a single number, e.g. for SEP.

☺ **THERFORE: Repeat the basic validation procedure many times, with different splits into calibration set and test set, to obtain a sufficient large number of test set prediction.**

Comparison of performance criteria

"Trivial, but often ignored"

WANTED

- **Very many residuals,** prediction errors from test set objects
- **Many estimations of the performance criterion**

For data sets with small number of objects:

appropriate techniques are necessary (**resampling**)

Resampling techniques

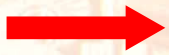
Bootstrap

Cross Validation

various strategies

repeated double Cross Validation (rdCV)

Rather small number of objects (*n*)

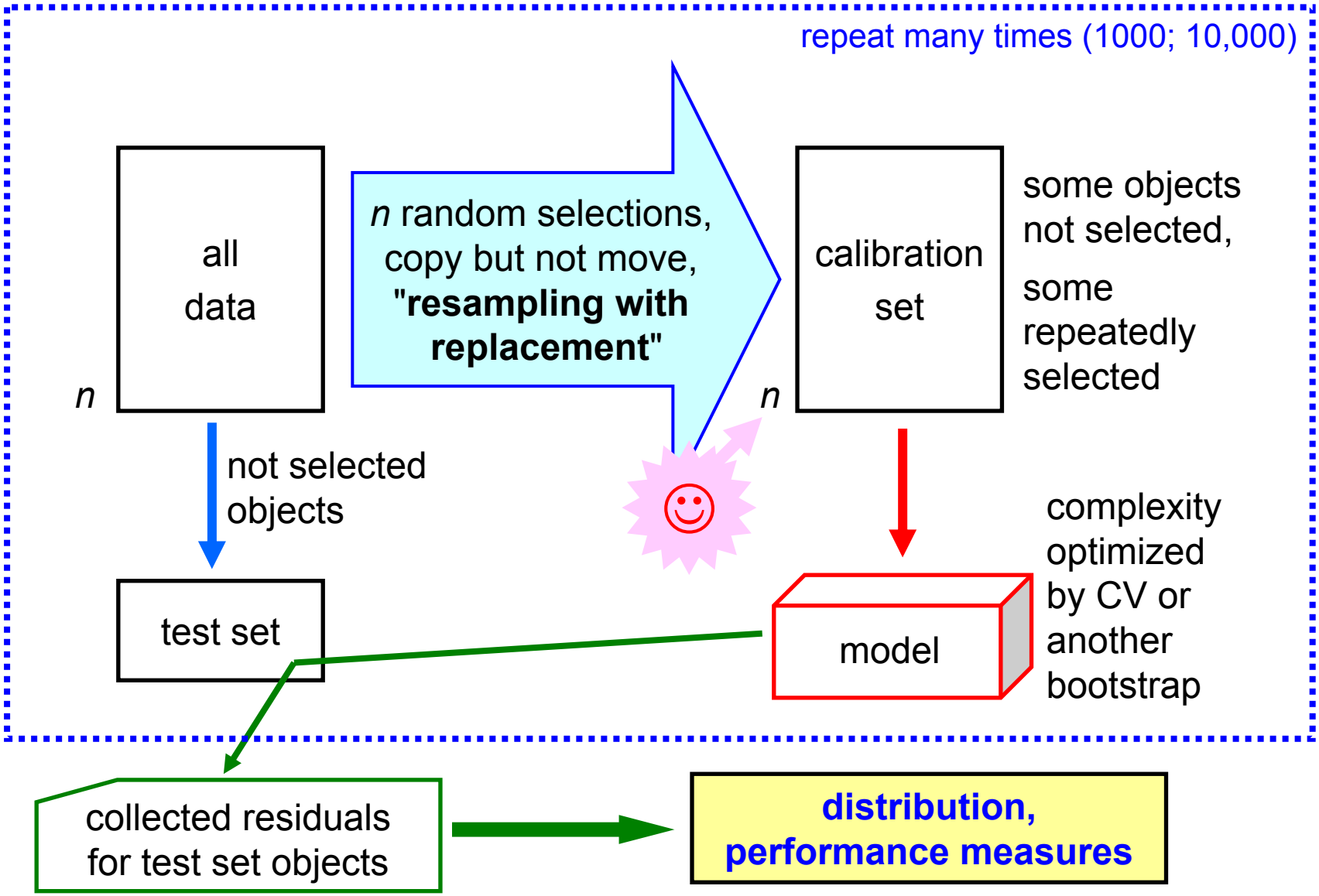


Realistic estimation of
- optimum model complexity,
- prediction errors for new cases, including their variability

CONTENTS

- Introduction
- Basic Strategy
- **Bootstrap**
- Cross Validation
- repeated double Cross Validation (rdCV)
- Realization of rdCV and Examples
- Optimum number of PLS components
- Classification

Boot strap (for model evaluation, short)



Boot strap (for model evaluation, short)

Advantages

+ simple,

+ always the maximum number, *n*, of objects in the calibration set

Disadvantages

- not all objects are considered equally,
- not a fixed but a varying number of prediction errors,
- calibration set contains a varying number of identical copies of objects,
  (on the average 63% of the objects are in the calibration set,
  t.m. 37% are copies)
- optimization of model complexity (for the calibration set) requires
  another bootstrap (or CV) with an even increased no. of
  copies in the training sets

CONTENTS

- Introduction
- Basic Strategy
- Bootstrap
- **Cross Validation**
- repeated double Cross Validation (rdCV)
- Realization of rdCV and Examples
- Optimum number of PLS components
- Classification

Cross Validation (CV)

☞ Most used resampling in chemometrics (statistics)

☞ For optimization of

* model complexity (no. of PLS or PCA components, ...),
* model parameter (no. of neighbors in KNN classification),
* estimation of model performance

if only a rather small number of objects is available

"random + some strategy"

If not properly applied: too optimistic or even wrong

"Some do not like CV"

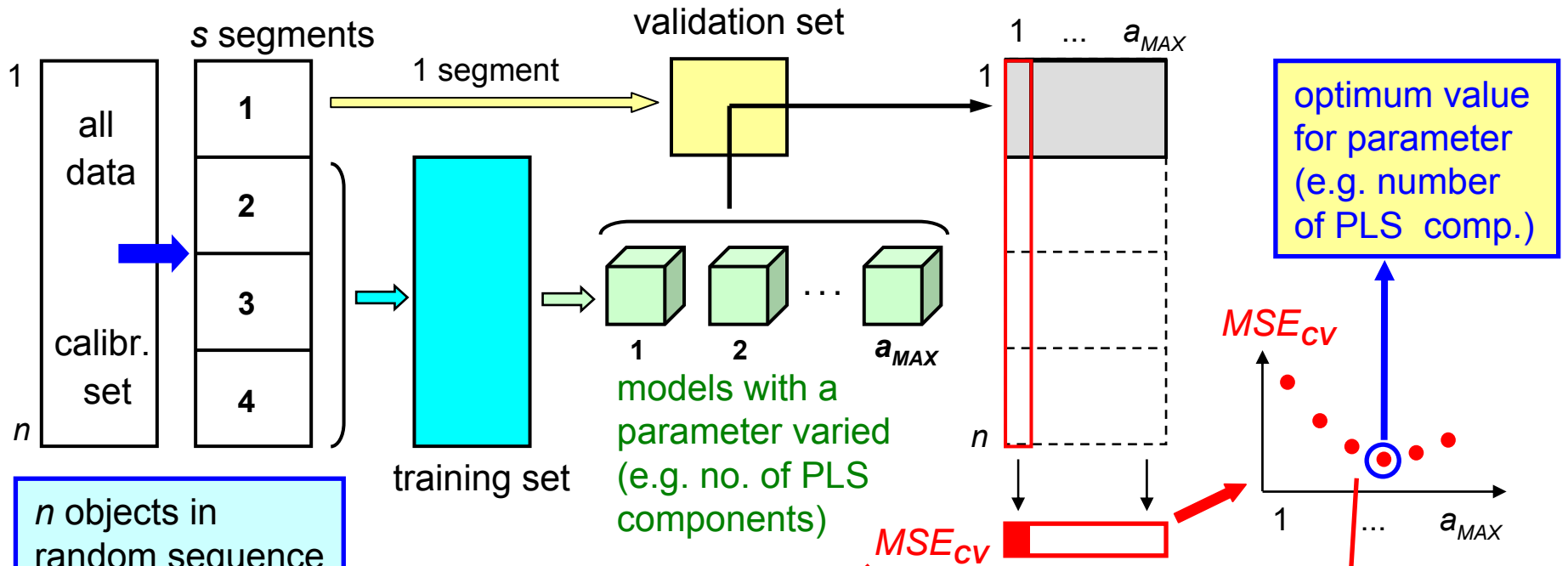
Esbensen K.H., Geladi P.: *J. Chemometrics*, 24, 168 (2010)

What else ??!

Cross Validation (CV) for optimization

Split into *s* segments with appr. equal no. of objects;
s = 2, 3, ... , *n* (often 4 - 7)
s = 4 leave-a-quarter-out
s = *n* leave-one-out (LOO, full CV)

residuals from CV
$e = y - \hat{y}_{CV}$



n objects in random sequence or split "112233", "123123", ...

mean squared error
$= (1/n)\Sigma e_i^2$

NOT appropriate as a measure of the model performance for new cases!

CONTENTS

- Introduction
- Basic Strategy
- Bootstrap
- Cross Validation
- **repeated double Cross Validation (rdCV)**
- Realization of rdCV and Examples
- Optimum number of PLS components
- Classification

double Cross Validation (dCV)

Outer CV (outer loop)

- Split all data into calibration sets and corresponding test sets
- Make optimized model from calibration set (see inner CV) and apply it to the test set
- Results in "properly generated" residuals, suitable for estimation of a **final performance**

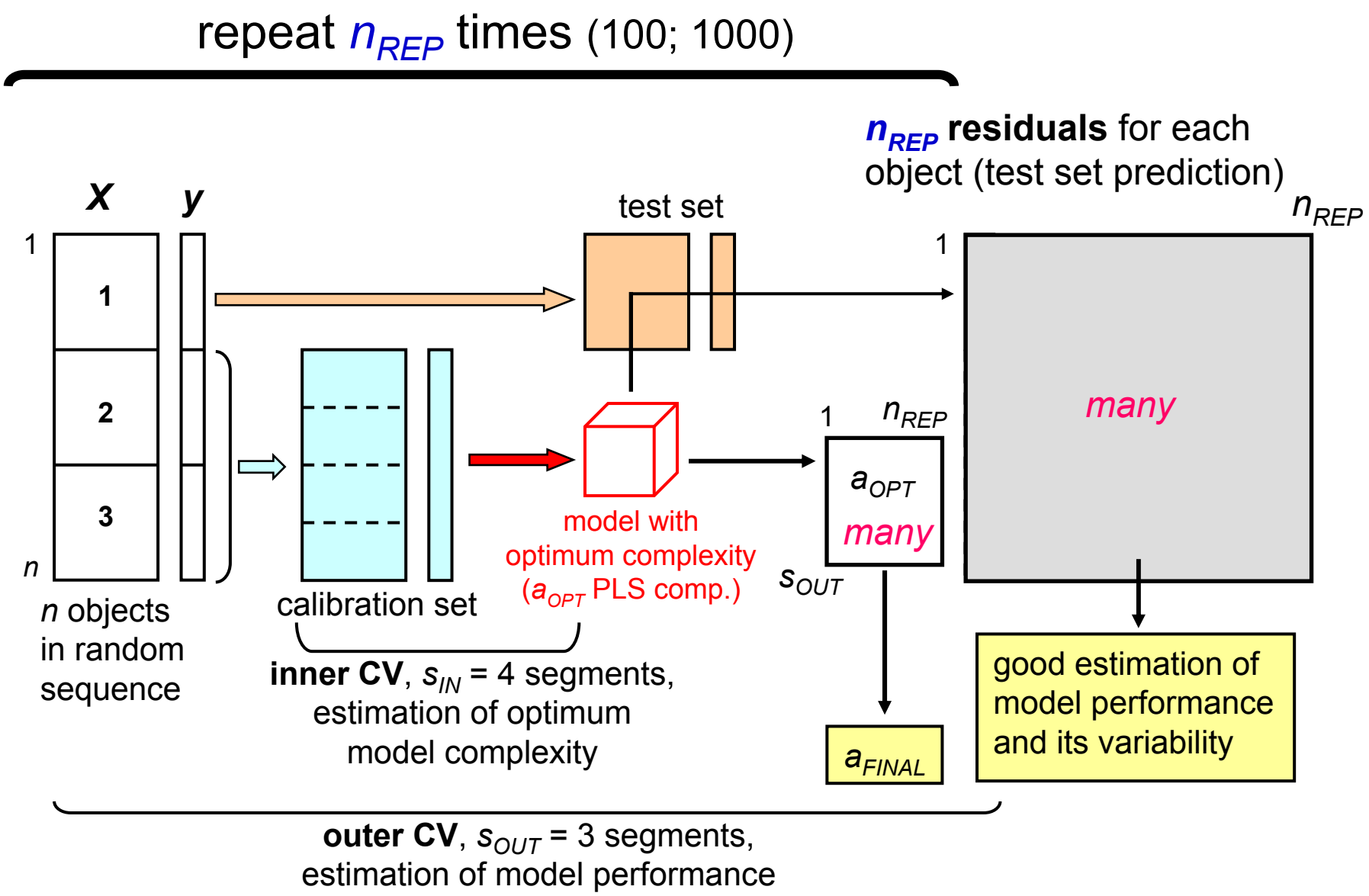
Inner CV (inner loop)

- Only for estimation of **optimum complexity** of model (no. of PLS components) but not for estimation of model performance.
- Split the calibration set into training sets and corresponding validation sets for this optimization.

No final performance measure from optimization !

No model optimization with results from test sets !

repeated double Cross Validation (rdCV)



repeated double Cross Validation (rdCV)

- Separates the estimation of the optimum model complexity (e.g. number of PLS-components) from the estimation of the model performance (for new cases).

- Yields many values for the optimum model complexity (e.g. number of PLS-components), and thereby allows a reasonable estimation of the final optimum model performance (a_{FINAL}) and its variability.

- Yields a large number of residuals from test set prediction, and thereby allows a reasonable estimation of the model performance (e.g. SEP) and its variability.

CONTENTS

- Introduction
- Basic Strategy
- Bootstrap
- Cross Validation
- repeated double Cross Validation (rdCV)
- **Realization of rdCV and Examples**
- Optimum number of PLS components
- Classification

Realisation (R)

www.lcm.tuwien.ac.at [click at **R**]

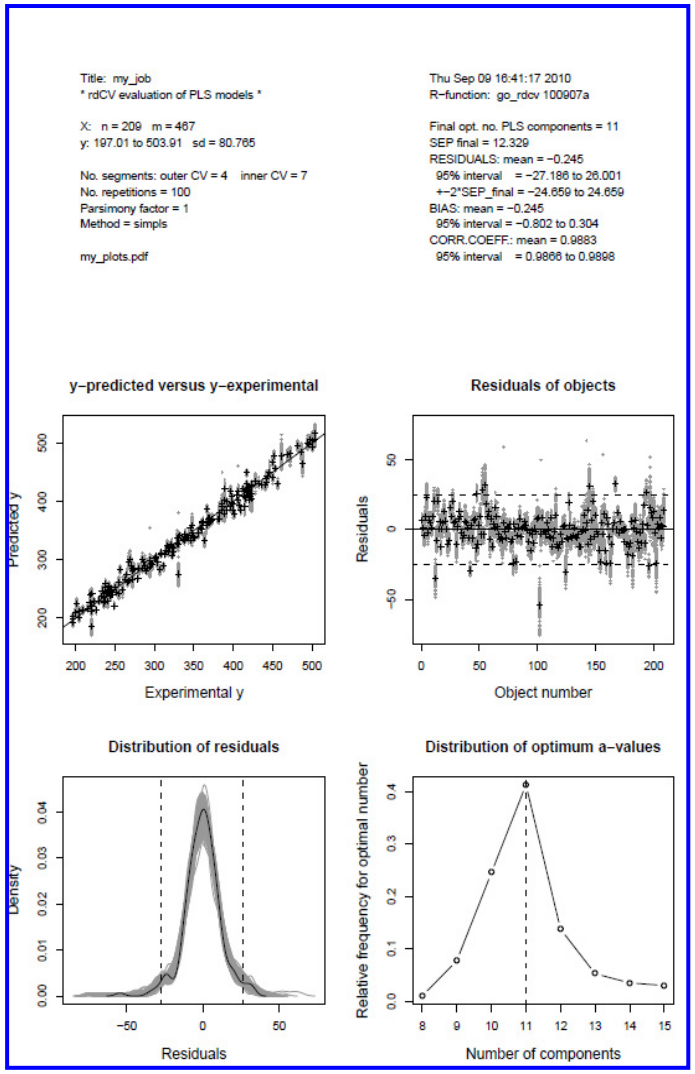
R-software (**go_rdcv.R**), short documentation, data

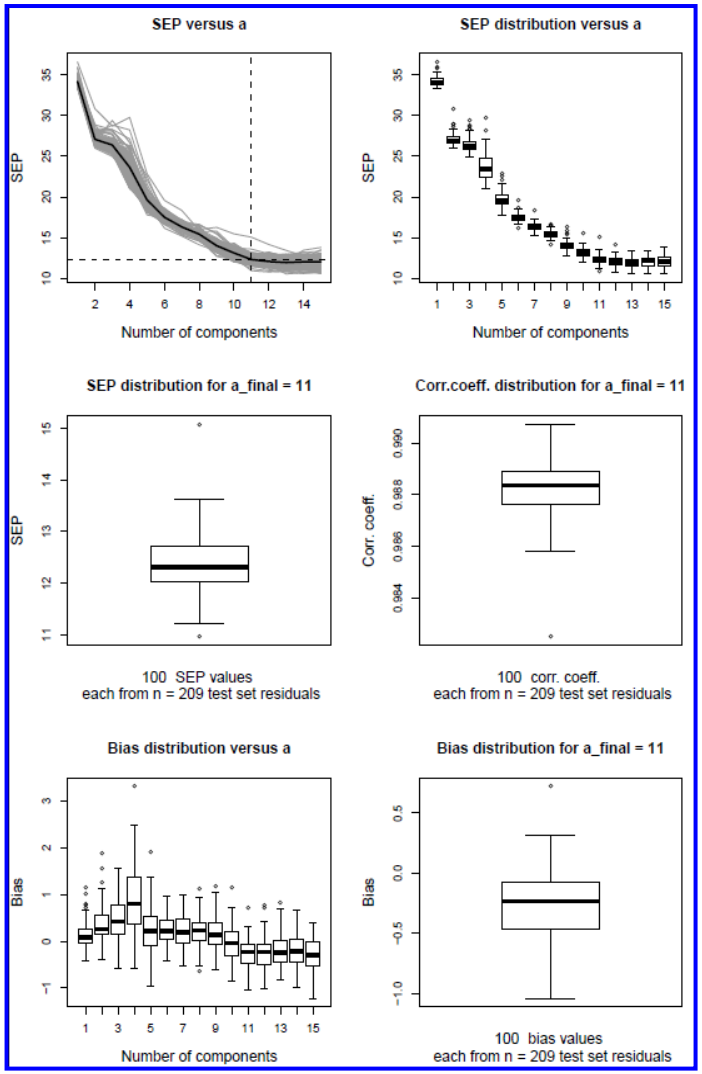
"download rdcv.zip and unzip"

```
>  source("go_rdcv.R")
>  # X   load/import X-data as matrix
>  # y   load/import y-data as vector
>  my_result = go_rdcv("my_job", X, y,
                 PDFfile = "my_plots.pdf")
```

Realisation (R)

www.lcm.tuwien.ac.at [click at **R**]





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

Research Article

Journal of CHEMOMETRICS

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

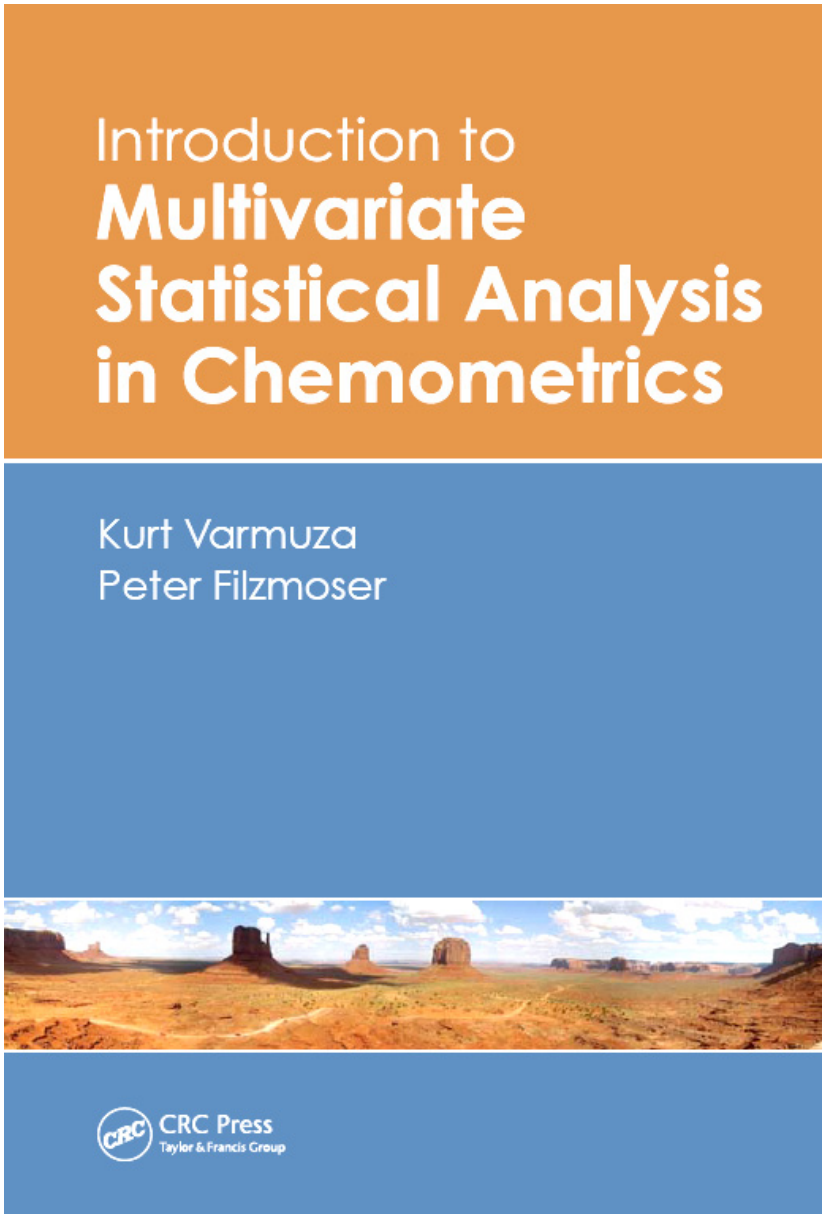
(www.interscience.wiley.com) DOI: 10.1002/cem.1225

Repeated double cross validation

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

Repeated double cross validation (rdCV) is a strategy for (a) optimizing the complexity of regression models and (b) for a realistic estimation of prediction errors when the model is applied to new cases (that are within the population of the data used). This strategy is suited for small data sets and is a complementary method to bootstrap methods. rdCV is a formal, partly new combination of known procedures and methods, and has been implemented in a function for the programming environment R, providing several types of plots for model evaluation. The current version of the software is dedicated to regression models obtained by partial least-squares (PLS). The applied methods for repeated splits of the data into test sets and calibration sets, as well as for estimation of the optimum number of PLS components, are described. The relevance of some parameters (number of segments in CV, number of repetitions) is investigated. rdCV is applied to two data sets from chemistry: (1) determination of glucose concentrations from near infrared (NIR) data in mash samples from bioethanol production; (2) modeling the gas chromatographic retention indices of polycyclic aromatic compounds from molecular descriptors. Models using all original variables and models using a small subset of the variables, selected by a genetic algorithm (GA), are compared by rdCV.

Keywords: prediction performance; optimum complexity of linear PLS models; cross validation; bootstrap; R



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

Ca 320 pages, appr. Euro 85

Book info
www.lcm.tuwien.ac.at

Book includes many R-codes

R-package "*chemometrics*"

Data set GLC

Glucose concentration from NIR

n = **120 alcoholic fermentation mashes** [1]

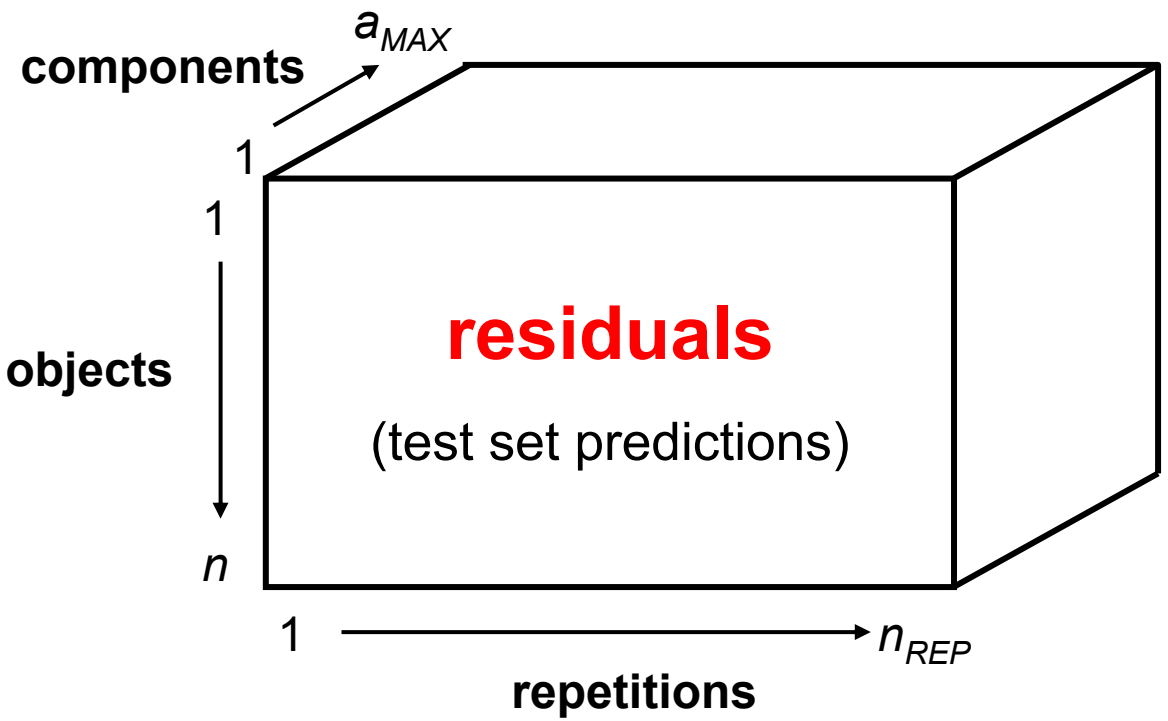
$\boldsymbol{y}$ **glucose concentration (HPLC)**; 0.1 to 55.3 g/L

$\boldsymbol{X_1}$ m_1 = **235 NIR absorptions**;1115-2285 nm,
first derivative (Savitzky-Golay, 7 points, 2nd order)

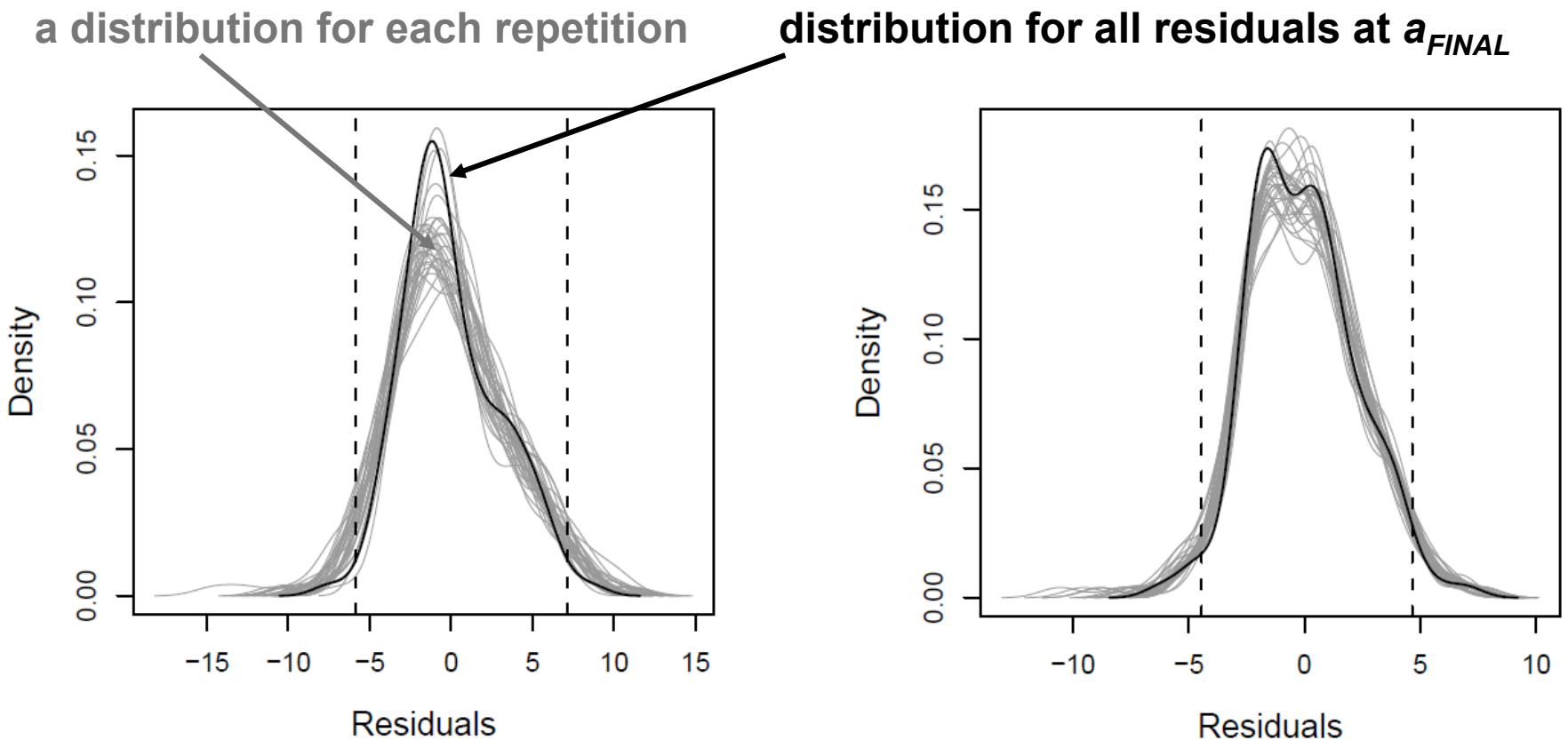
$\boldsymbol{X_2}$ m_2 = **15 NIR absorptions selected**;
genetic algorithm; from all data; *MobyDigs* [2]

[1] Liebmann B., Friedl A., Varmuza K.: Anal. Chim. Acta, **642**, 171-178 (2009)
[2] MobyDigs software, 1.0. Talete srl, www.talete.mi.it, Milan, Italy (2004)

rdCV Results: Residuals



rdCV Results: Distribution of Residuals



m = 235

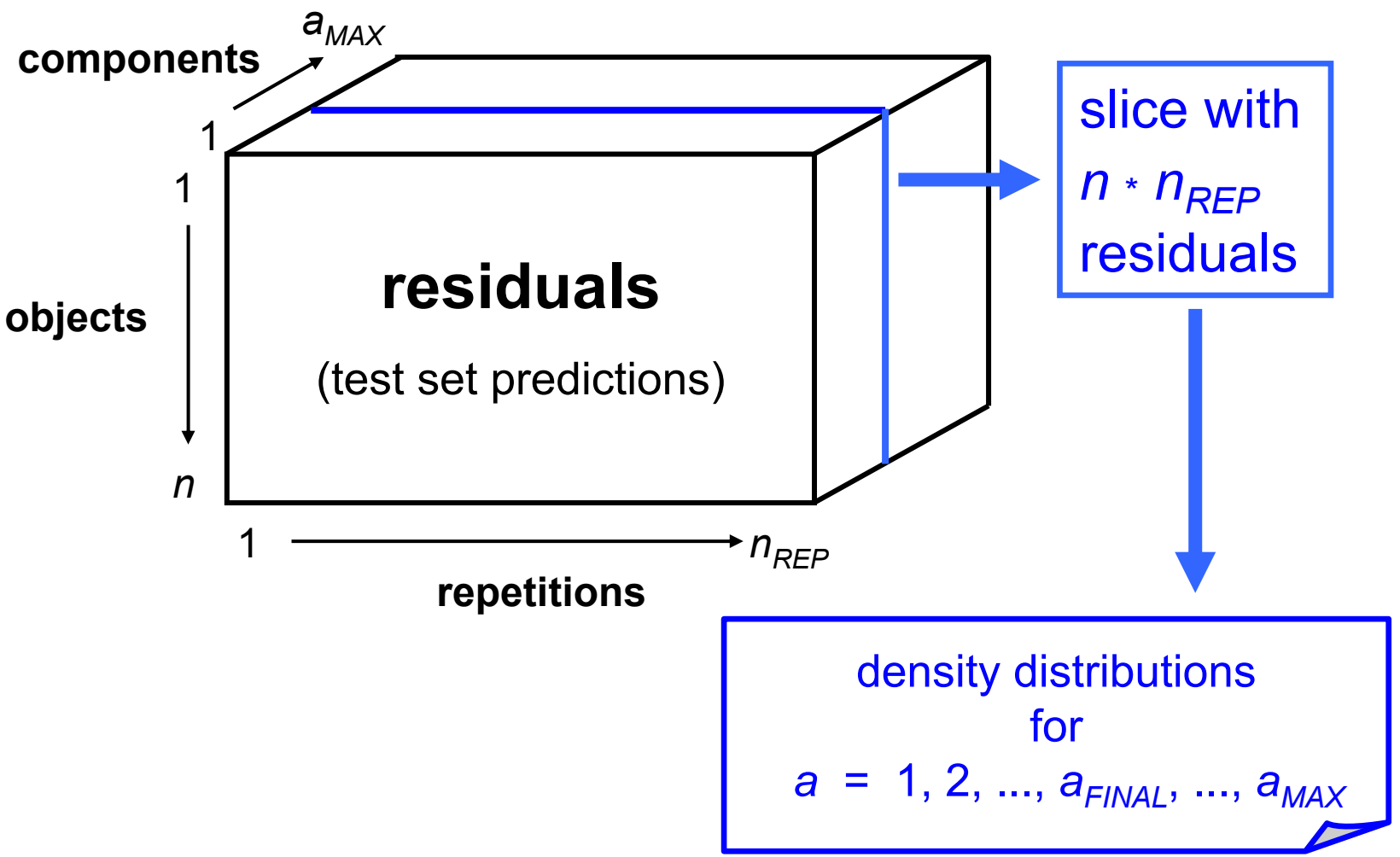
a_{FINAL} = 13, SEP_{FINAL} = 3.4

m = 15

a_{FINAL} = 8, SEP_{FINAL} = 2.3

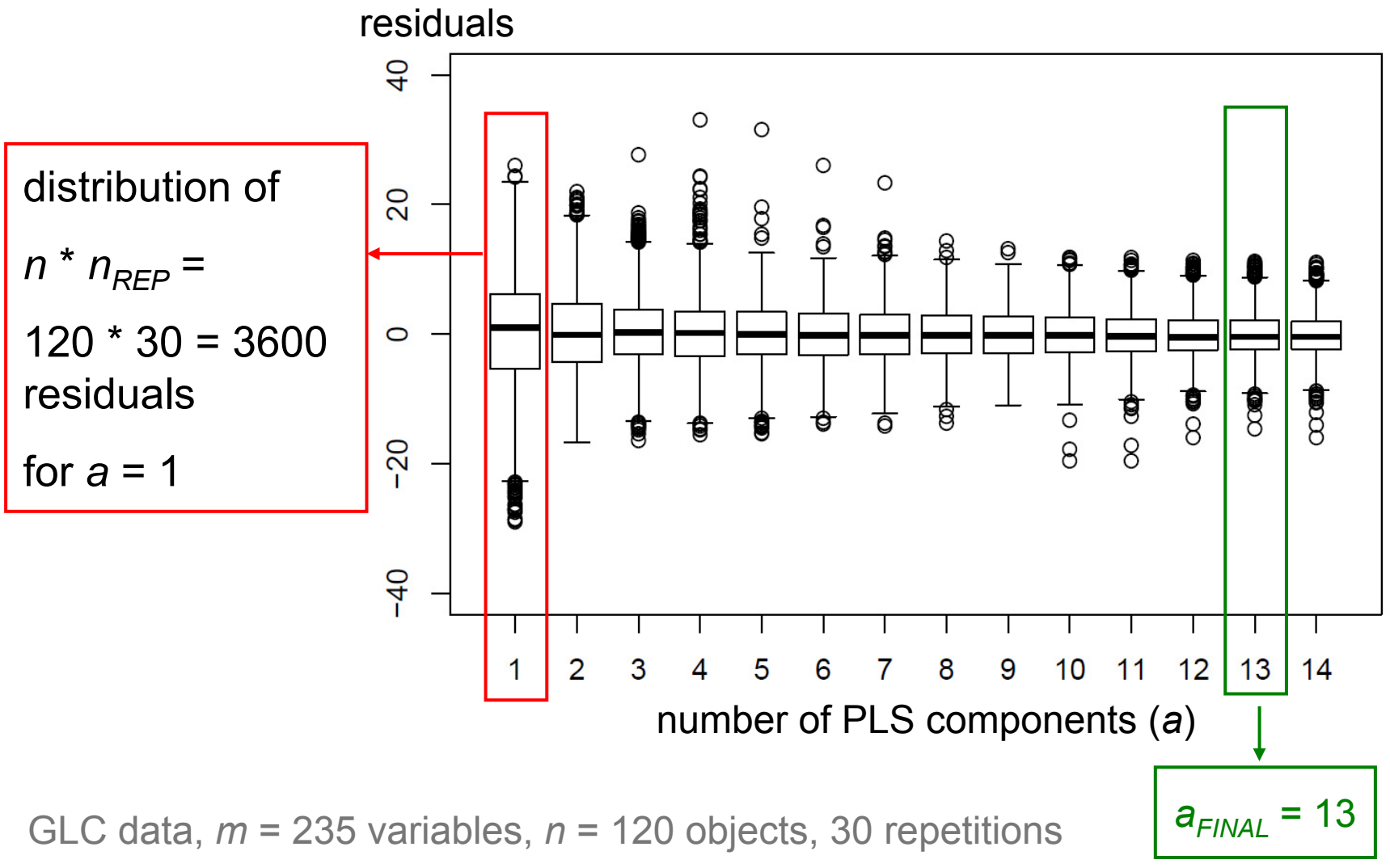
GLC data, m = 235 and 15 variables, n = 120 objects, 30 repetitions

rdCV Results: Distribution of Residuals



rdCV Results: Distribution of Residuals

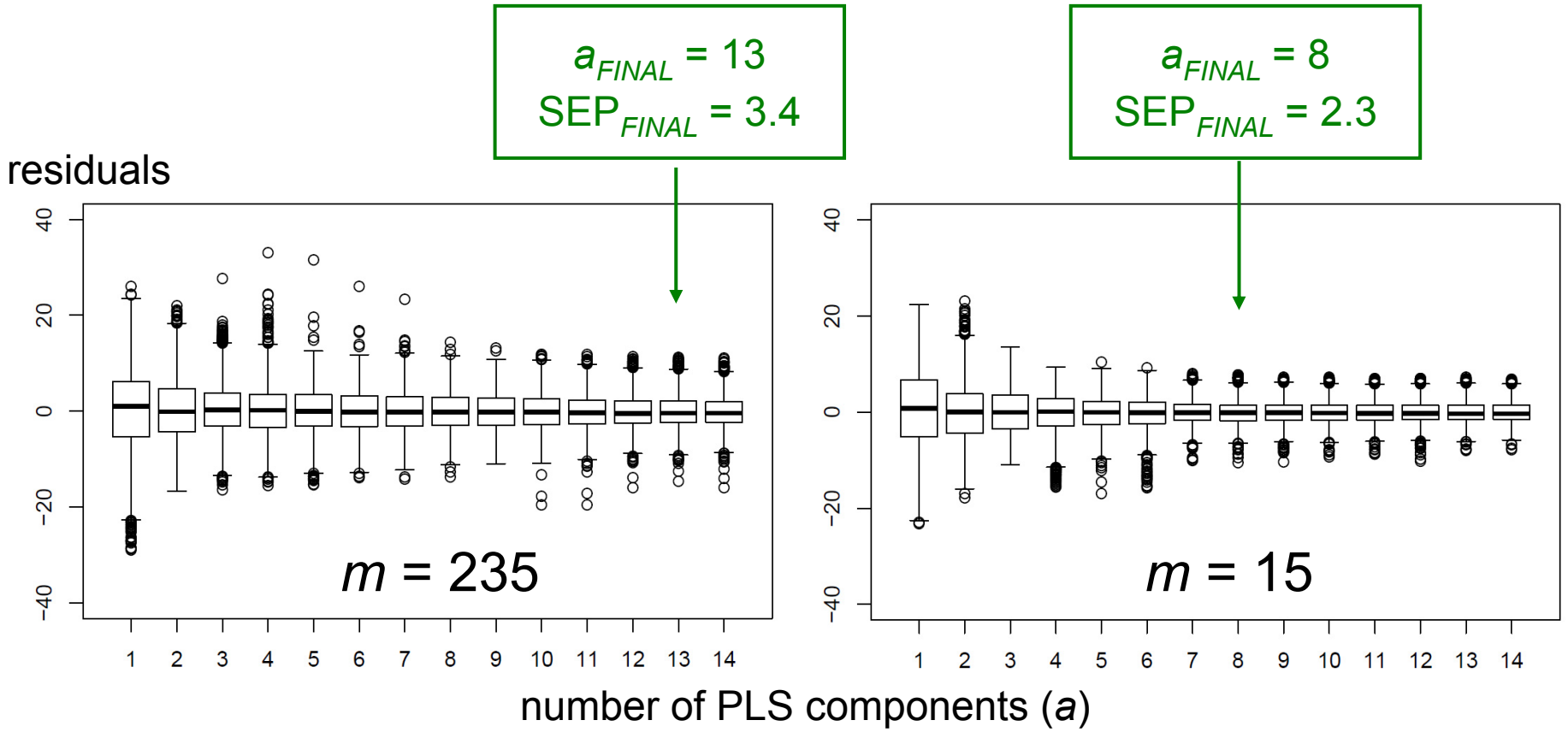
Distribution of residuals for a = 1, 2, ..., a_{FINAL}, ..., a_{MAX}



GLC data, m = 235 variables, n = 120 objects, 30 repetitions

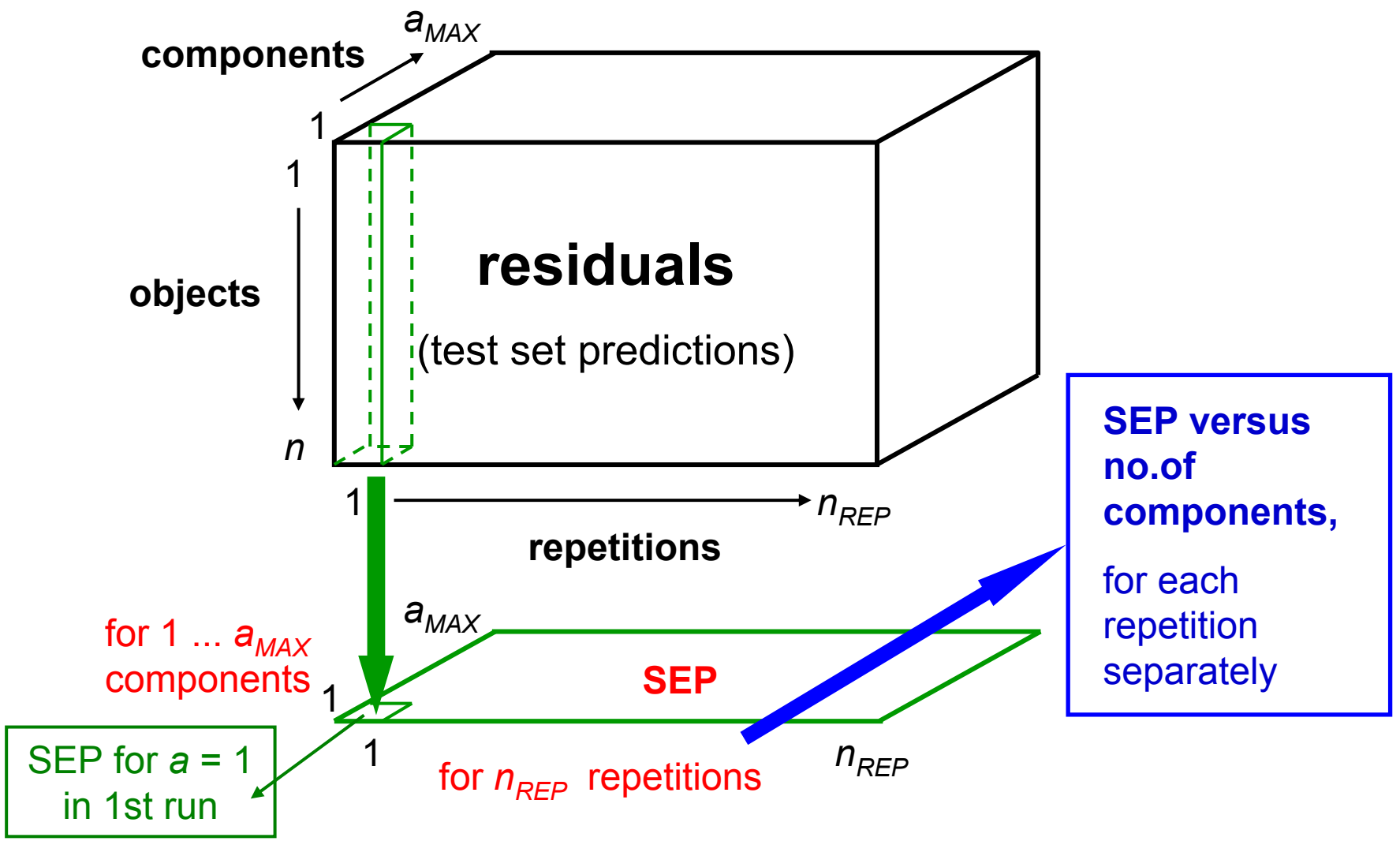
rdCV Results: Distribution of Residuals

Distribution of residuals for $a = 1, 2, ..., a_{FINAL}, ..., a_{MAX}$



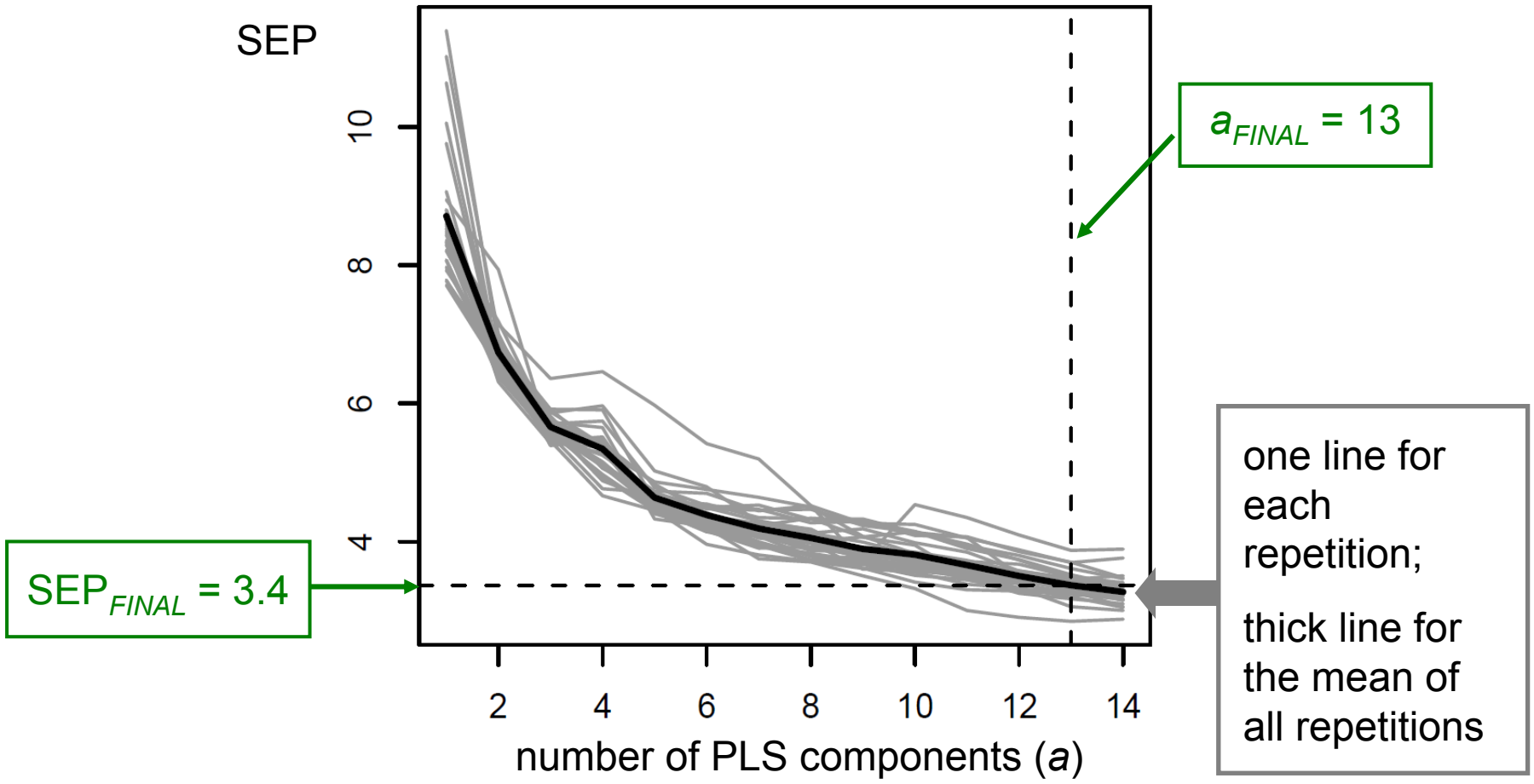
GLC data, m = 235 and 15 variables, n = 120 objects, 30 repetitions

rdCV Results: Residuals (SEP)



rdCV Results: Residuals (SEP)

SEP versus number of PLS components



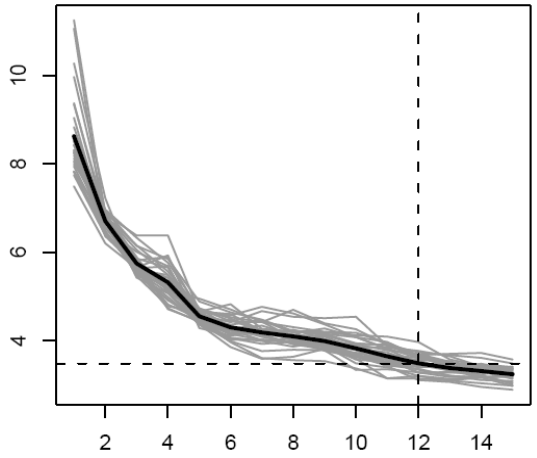
GLC data, *m* = 235 variables, *n* = 120 objects, 30 repetitions

rdCV Results: Residuals (SEP)

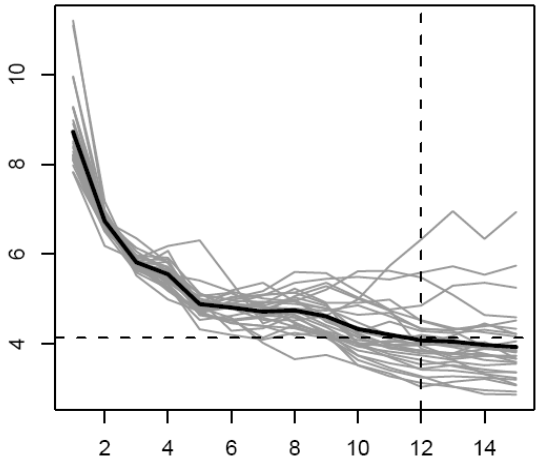
With artificial outliers (3 randomly selected variables "times 2")

SEP versus number of PLS components

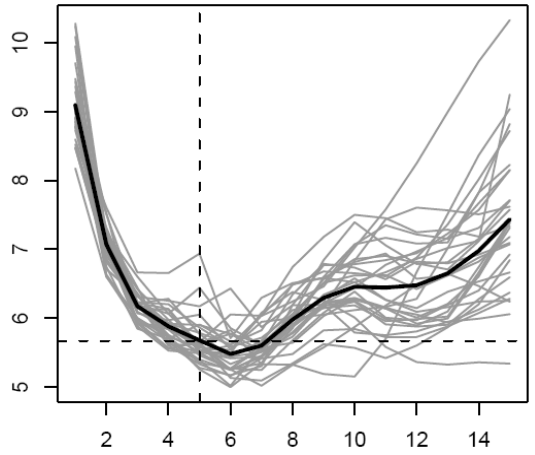
no outlier objects



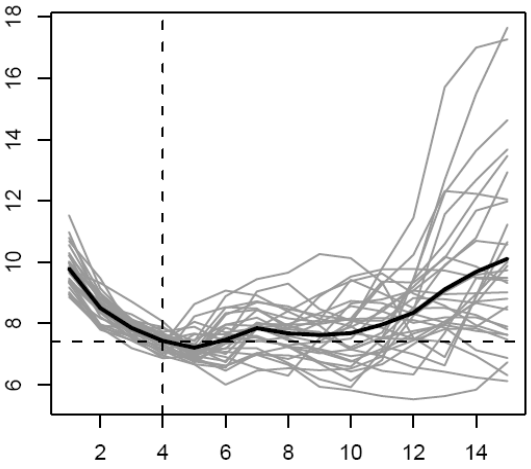
1 outlier object



5 outlier objects



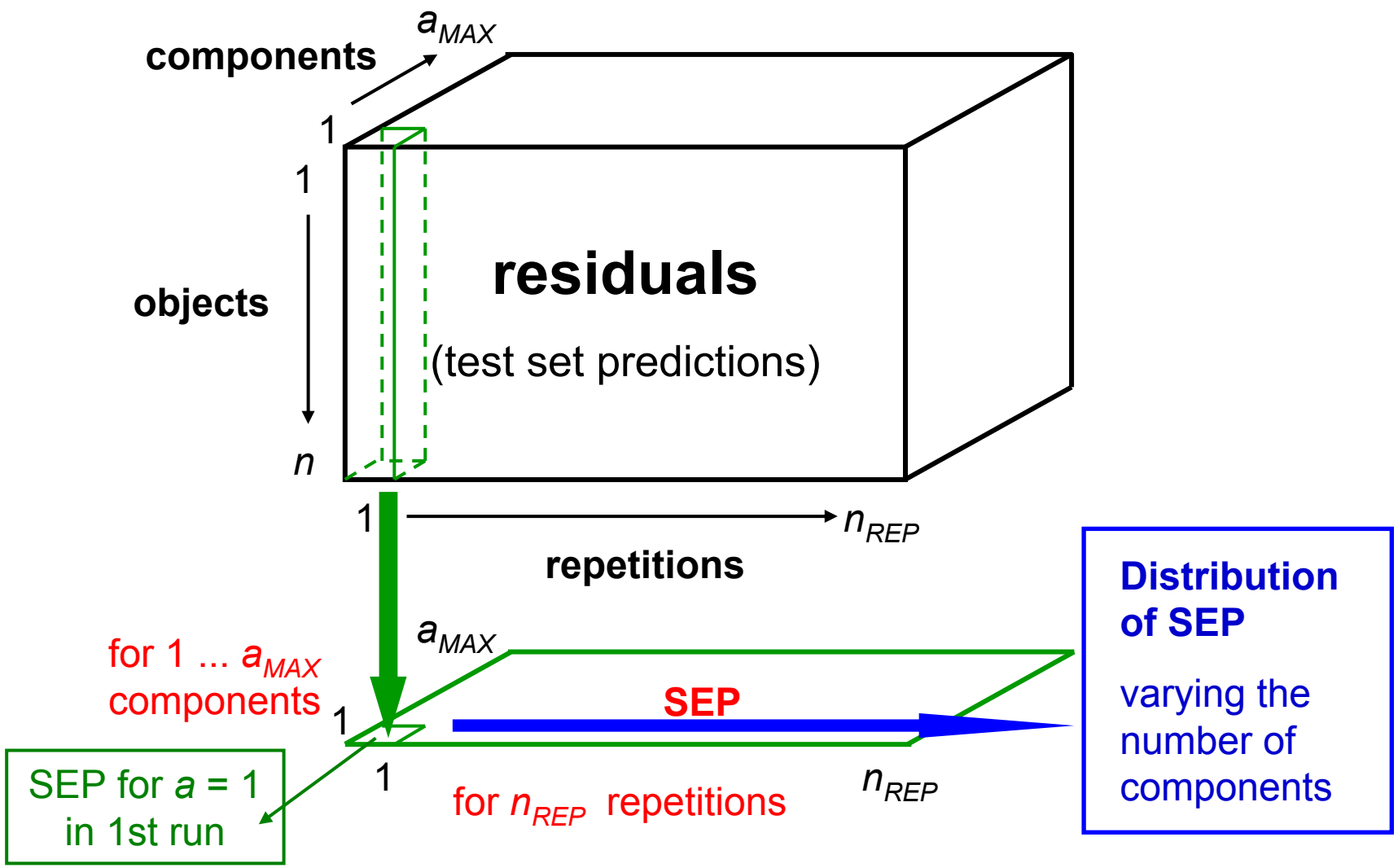
10 outlier objects



GLC data, *m* = 235 variables, *n* = 120 objects, 30 repetitions

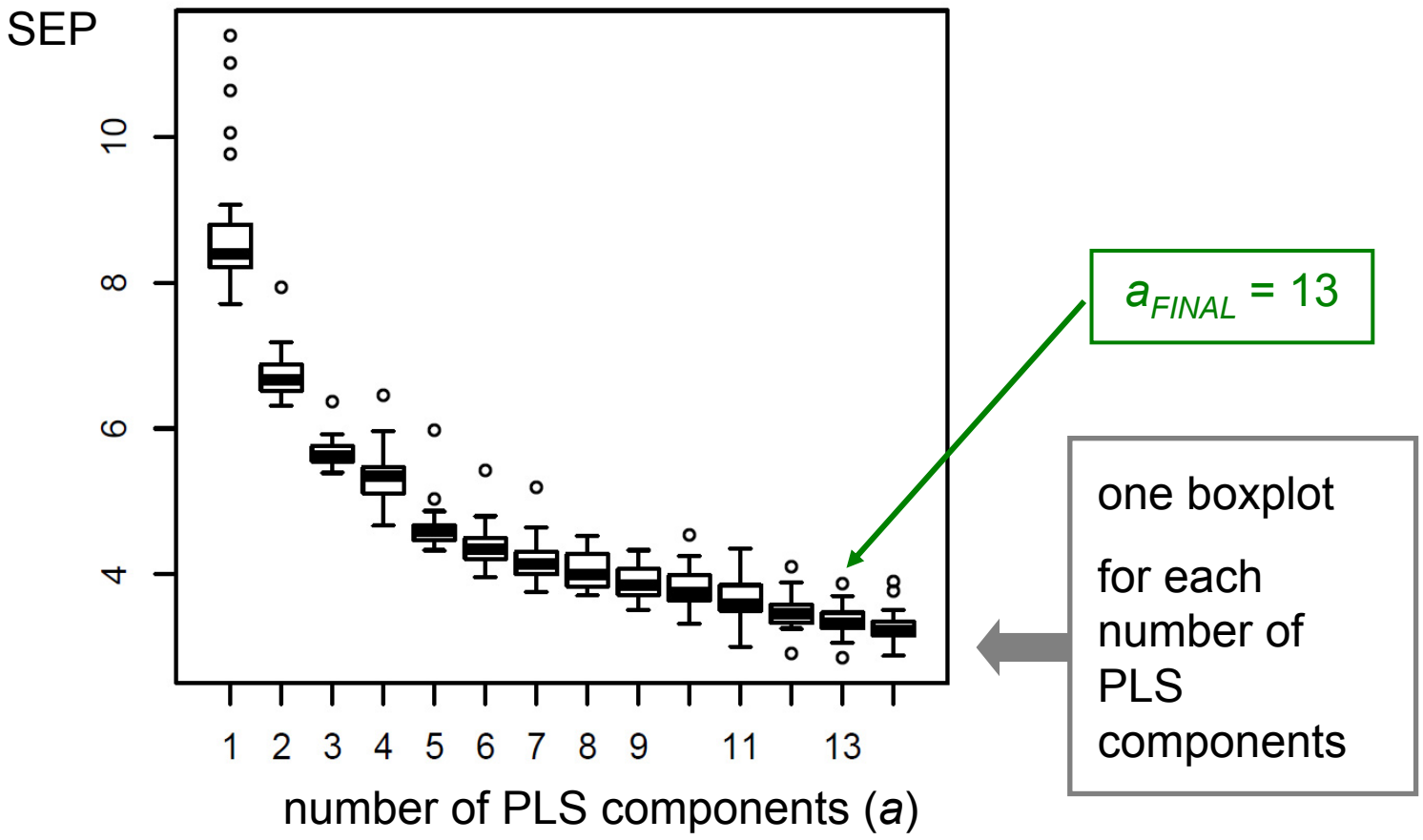
SEP 1 outlier

rdCV Results: Residuals (SEP)



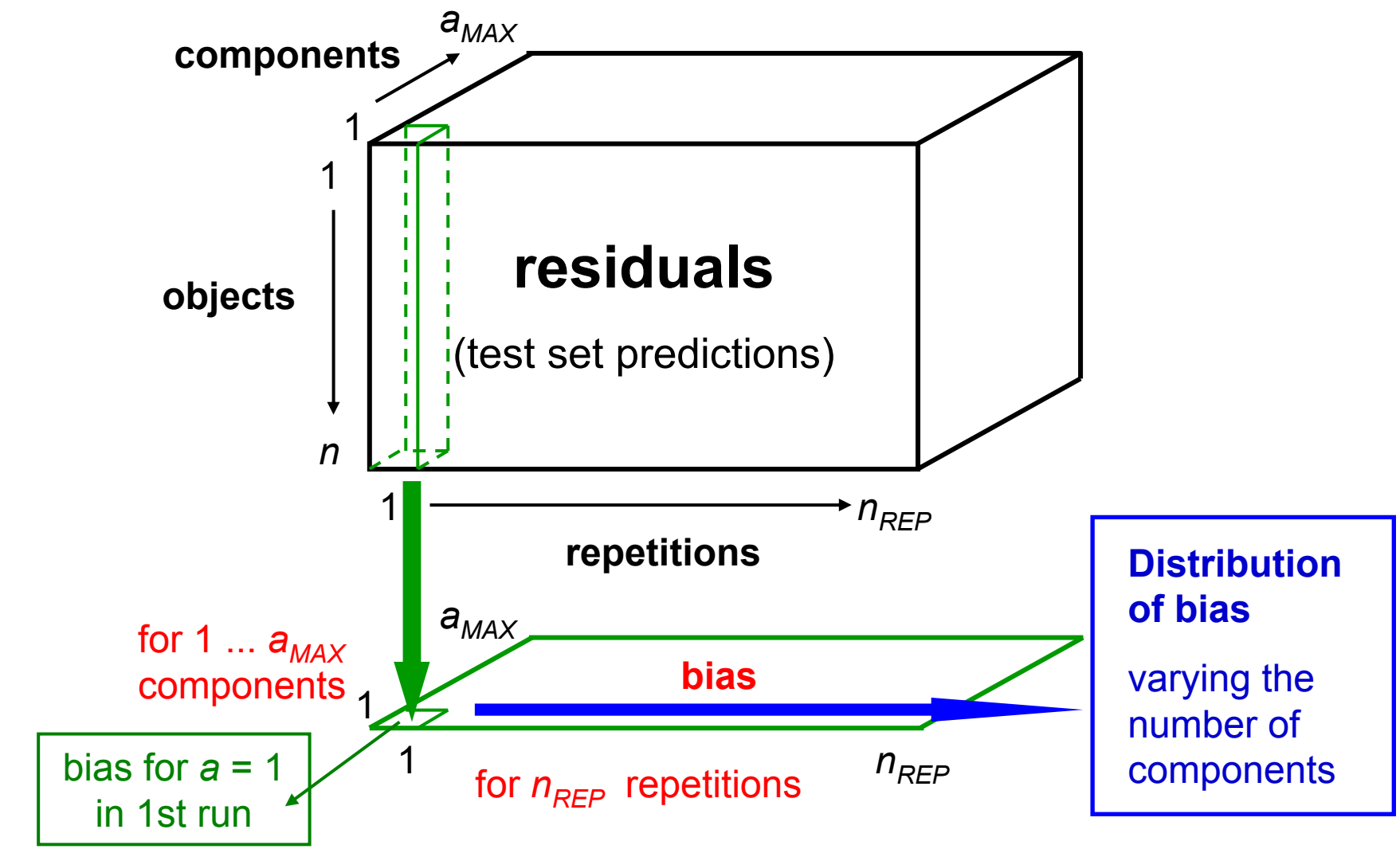
rdCV Results: Residuals (SEP)

Distribution of SEP versus number of PLS components



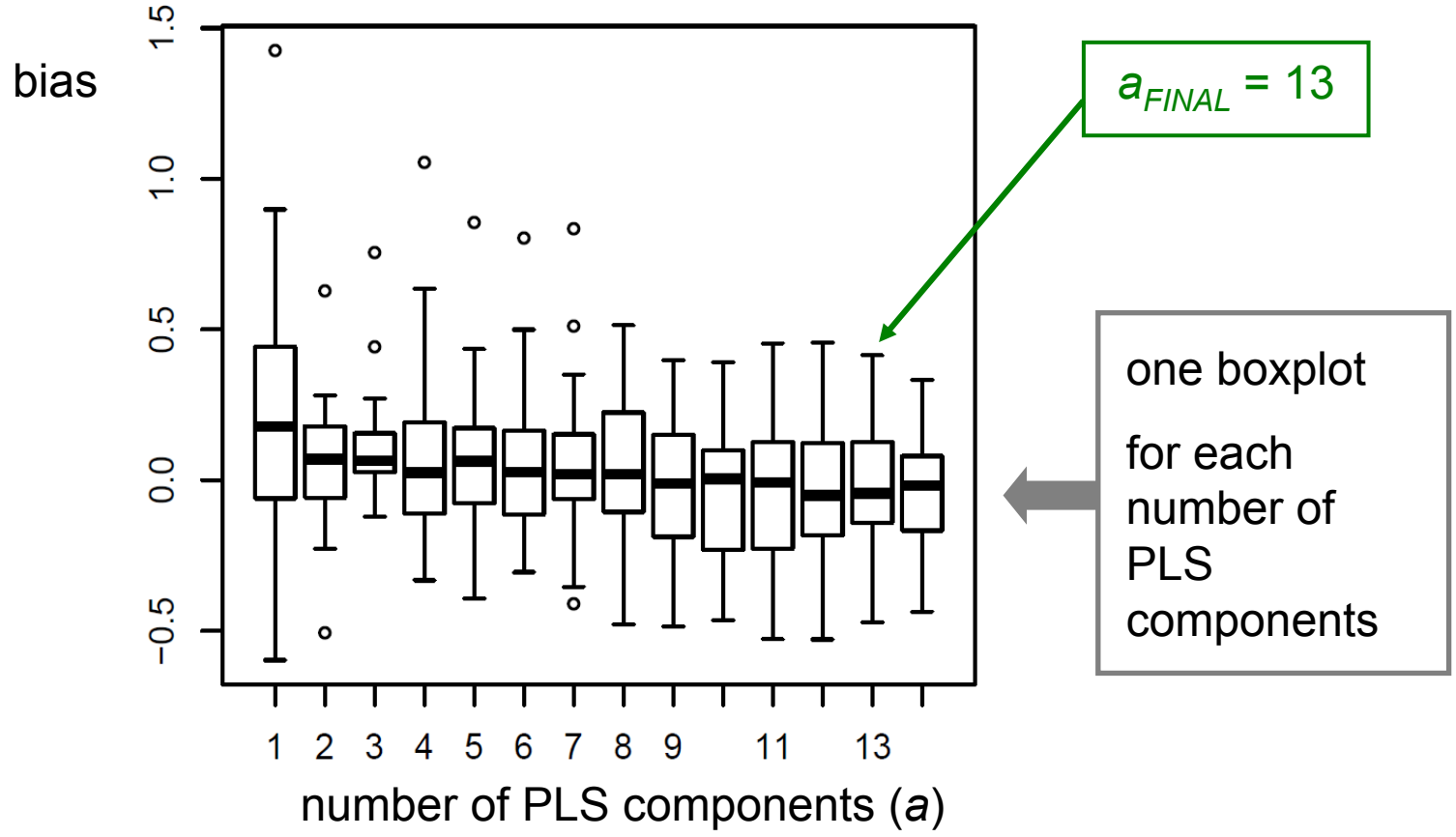
GLC data, *m* = 235 variables, *n* = 120 objects, 30 repetitions

rdCV Results: Residuals (Bias)



rdCV Results: Residuals (Bias)

Distribution of bias versus number of PLS components



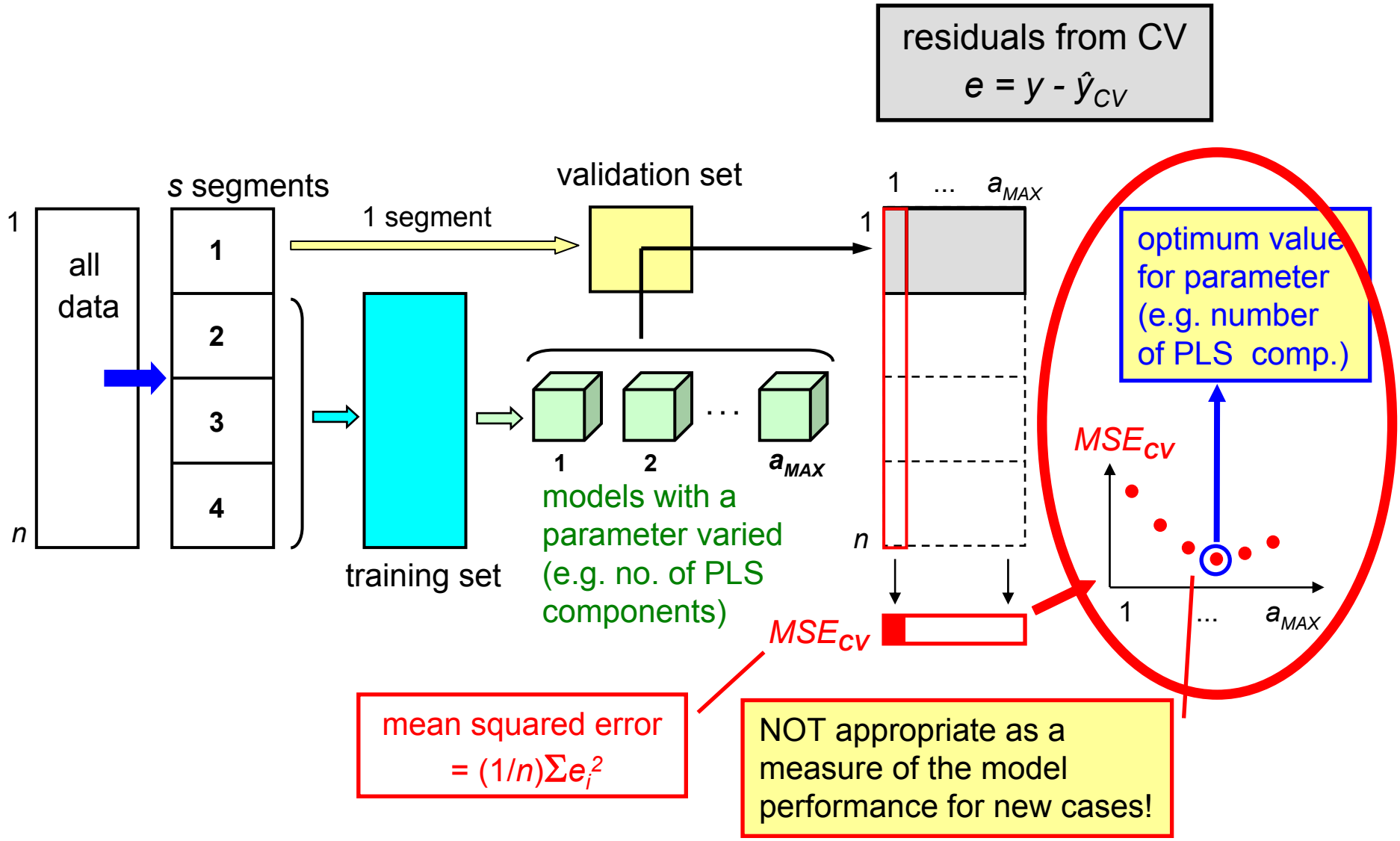
GLC data, *m* = 235 variables, *n* = 120 objects, 30 repetitions

bias plots

CONTENTS

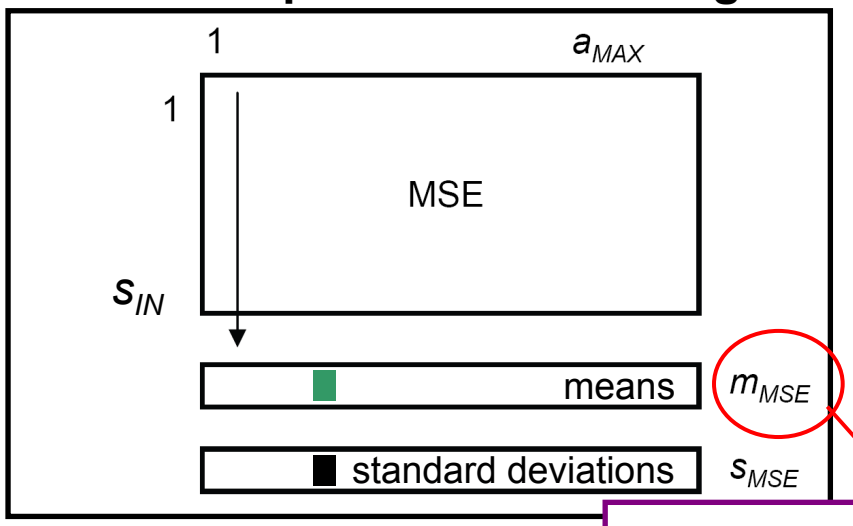
- Introduction
- Basic Strategy
- Bootstrap
- Cross Validation
- repeated double Cross Validation (rdCV)
- Realization of rdCV and Examples
- **Optimum number of PLS components**
- Classification

Estimation of optimum no. of PLS components



Estimation of optimum no. of PLS components

Inner CV loop: MSE for each segment



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

π parsimony factor

$\pi = 0$ global minimum

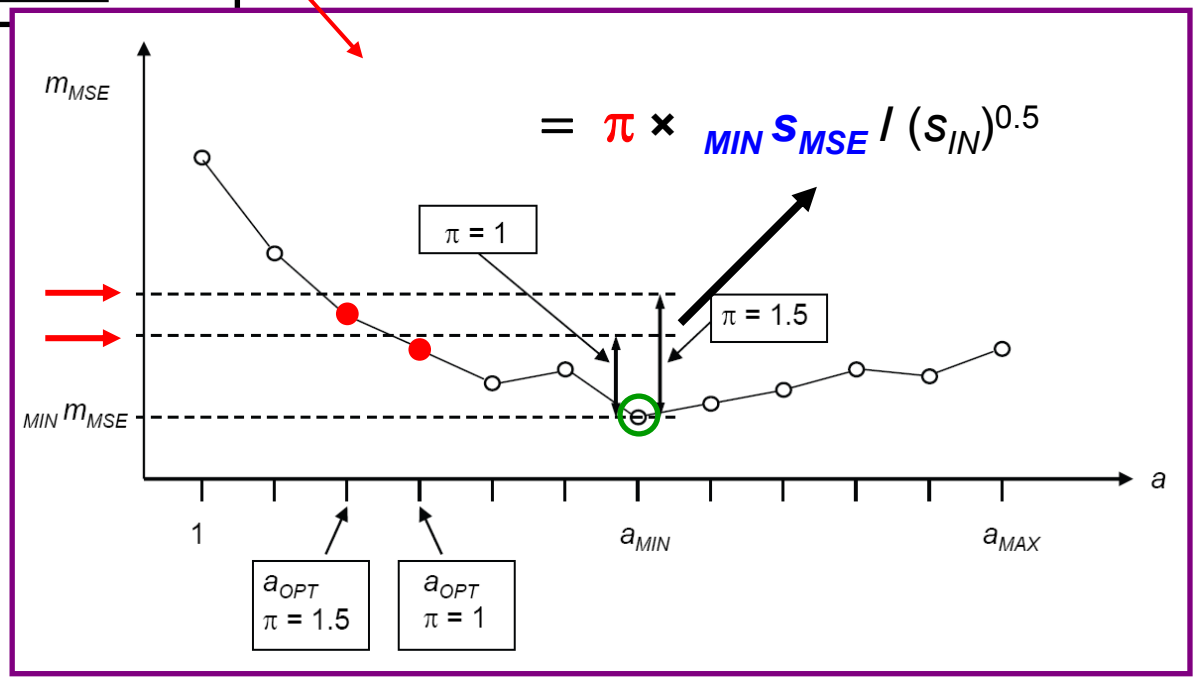
$\pi = 2$ 95% confidence interval

global minimum

$MIN\ m_{MSE}$

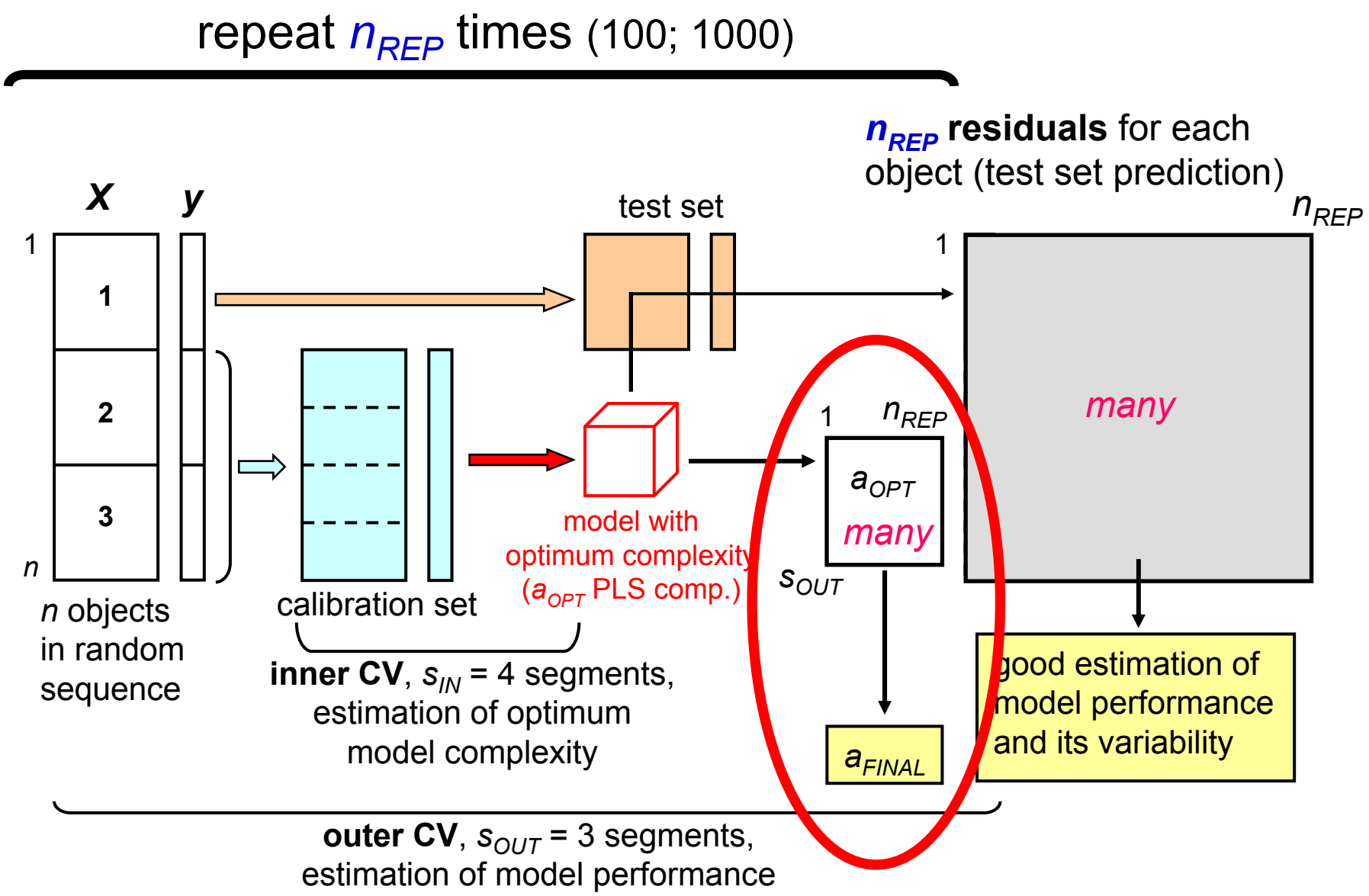
with

$MIN\ s_{MSE}$

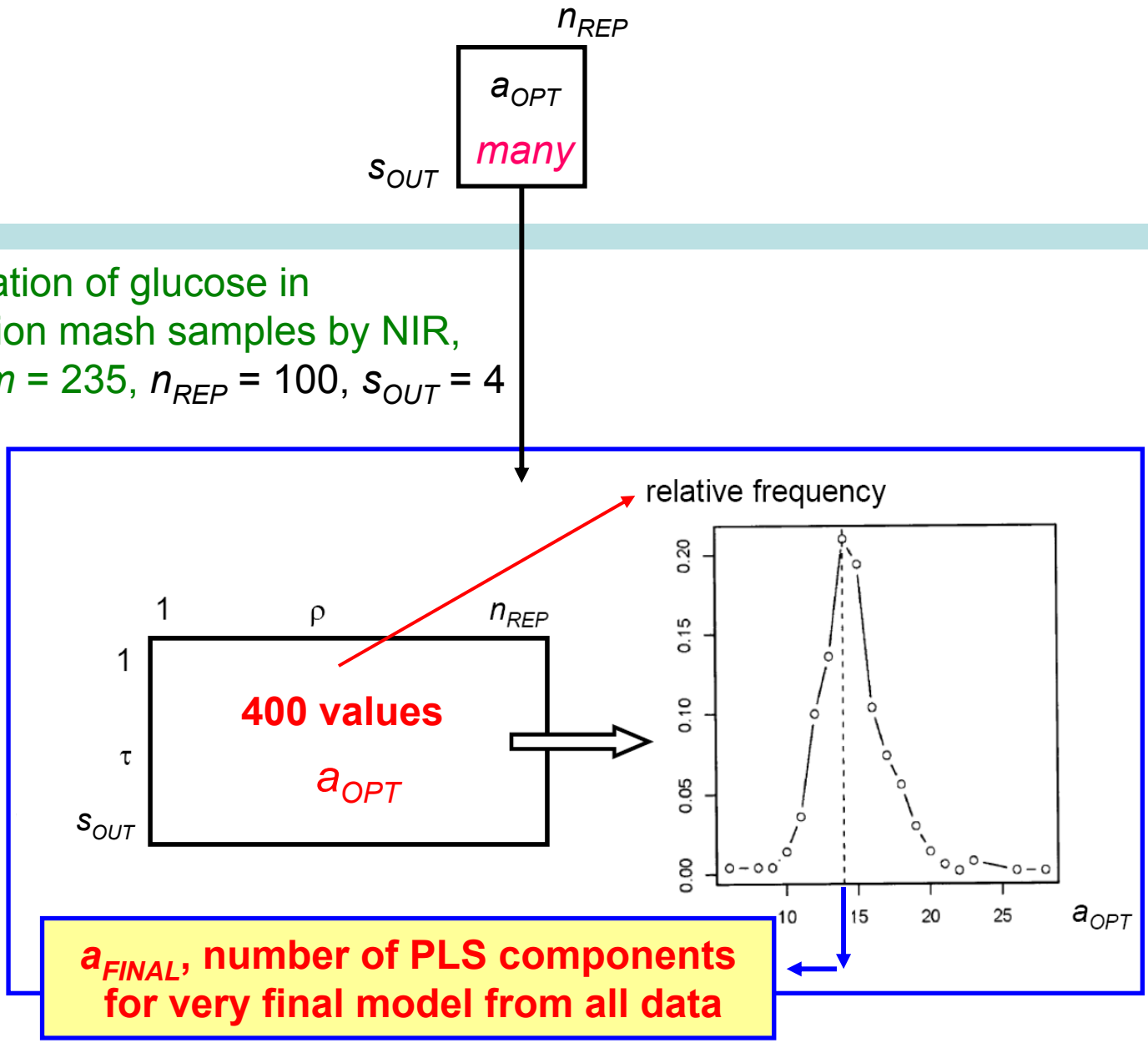


stand err meth

Estimation of FINAL optimum no. of PLS comp.



Estimation of FINAL optimum no. of PLS comp.



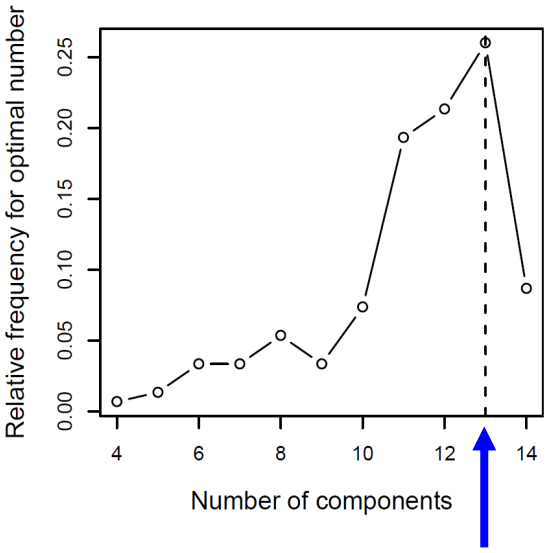
aOPT

rdCV Results: Optimum no. PLS comp.

Variation of parsimony factor π

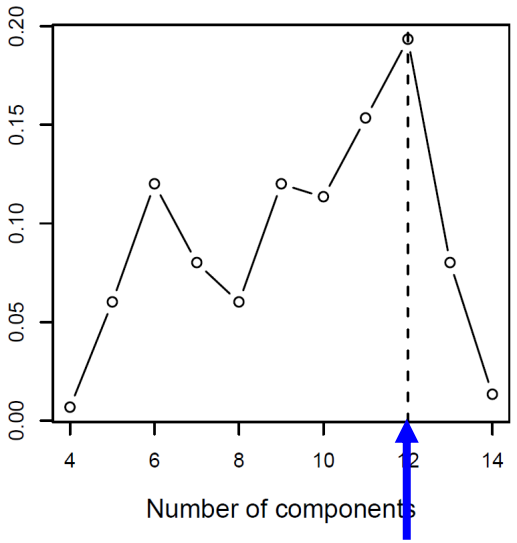
Distribution of a_{OPT} values and a_{FINAL}

$\pi = 1$



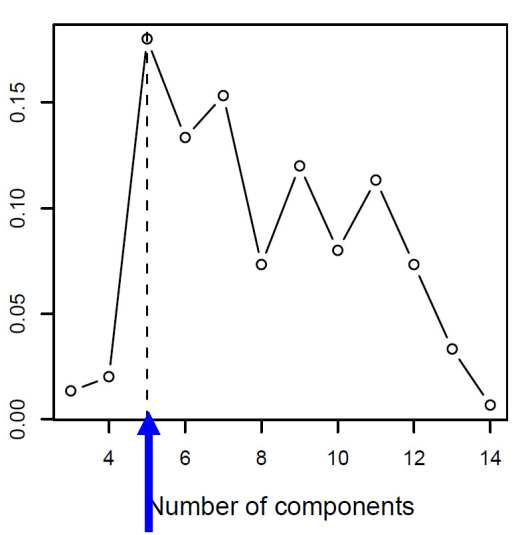
$a_{FINAL} = 13$

$\pi = 2$



$a_{FINAL} = 12$

$\pi = 3$



$a_{FINAL} = 7$

GLC data, *m* = 235 variables, *n* = 120 objects, 50 repetitions

aopt distr

CONTENTS

- Introduction
- Basic Strategy
- Bootstrap
- Cross Validation
- repeated double Cross Validation (rdCV)
- Realization of rdCV and Examples
- Optimum number of PLS components
- **Classification**

Classification

Binary classification by regression methods (e.g. D-PLS)

Target value $y = y_1$ for class 1 e.g. $y_1 = 0$

$y = y_2$ for class 2 e.g. $y_2 = 1$

Discrimination value $y_D = (y_1 + y_2) / 2$ e.g. $y_D = 0.5$

Classification if $\hat{y} \leq y_D$ assign to class 1

if $\hat{y} > y_D$ assign to class 2

Classification

Binary classification by regression methods (e.g. D-PLS)

n test objects		assigned class 1	assigned class 2	sum
true class	1	n_{11}	n_{12}	n_1
	2	n_{21}	n_{22}	n_2
	sum	$n_{\rightarrow 1}$	$n_{\rightarrow 2}$	n

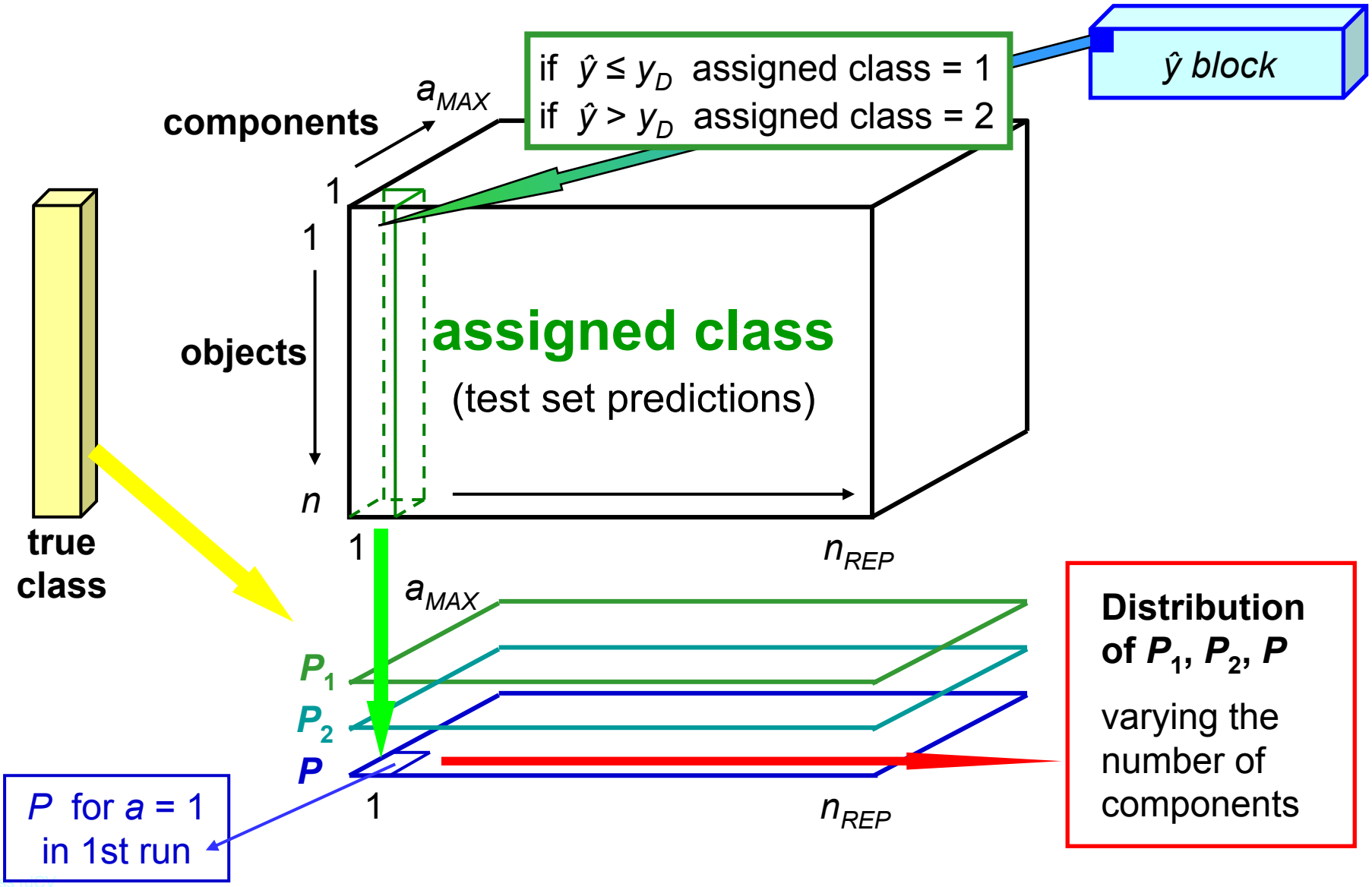
Predictive ability

class 1	$P_1 = n_{11} / n_1$
class 2	$P_2 = n_{22} / n_2$
Average predictive ability	$P = (P_1 + P_2) / 2$

No: Overall predictive ability $(n_{11} + n_{22}) / n$

Classification - rdCV

Binary classification by regression methods (e.g. D-PLS)



Data set AMES

QSAR: mutagenicity from AMES test

n = **6458 chemical structures from organic compound** [1], approx. 3D, all H-atoms; *Corina* [2]

y **AMES mutagenicity**

n_1 = 3488 (mutagenic), n_2 = 2970 (not mutagenic);

binary classification

X m = **1604 molecular descriptors**; *Dragon* [3]

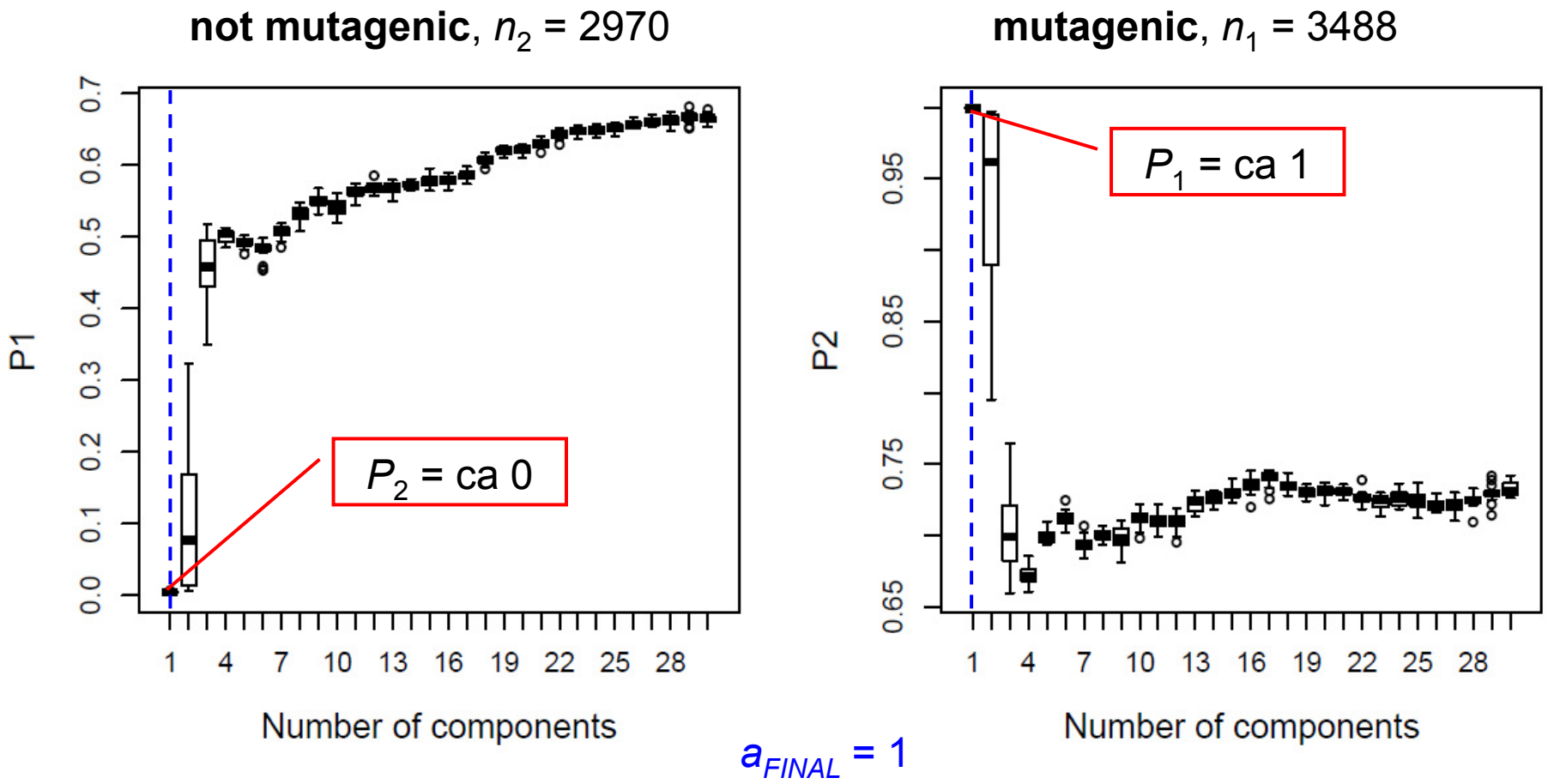
[1] K. Hansen, Technical University Berlin, Germany
http://ml.cs.tu-berlin.de/toxbenchmark/index.htm

[2] Corina software, Molecular Networks GmbH Computerchemie, www.mol-net.de, Erlangen, Germany (2004).

[3] Dragon software, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004).

Classification (erroneous AMES data)

Distribution of predictive ability

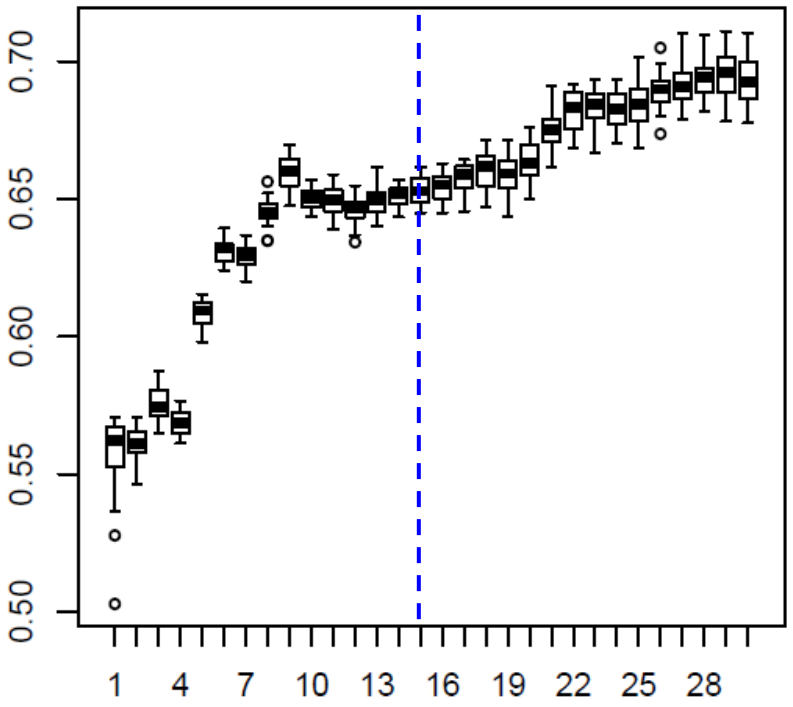


n = 6458, m = 1604 descriptors (variables), 20 repetitions (for each box plot)

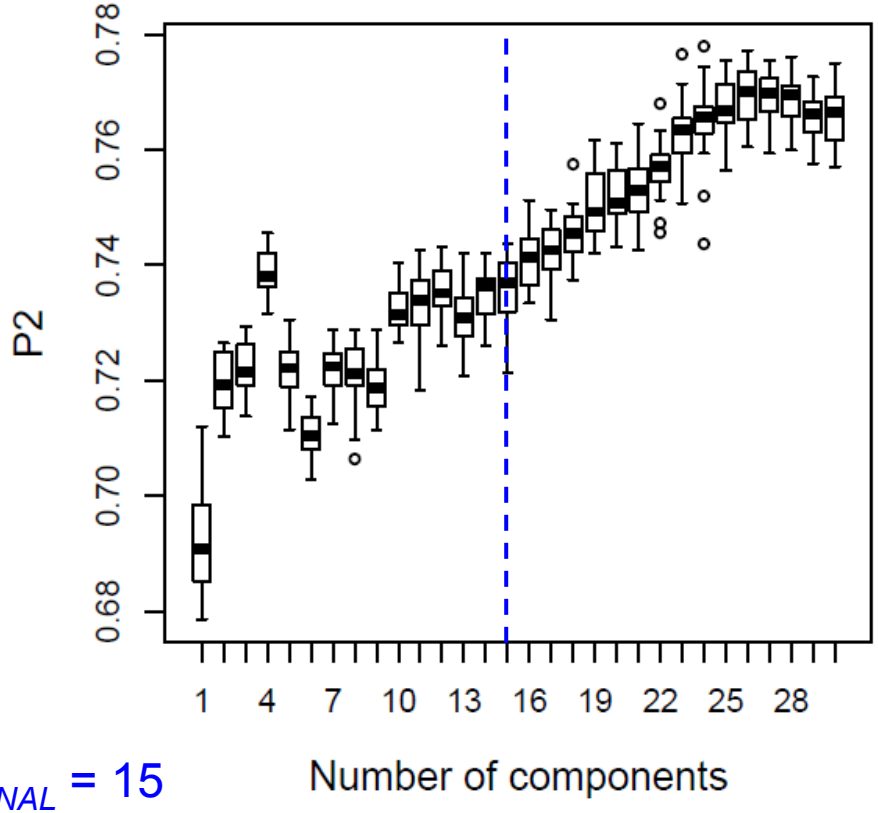
Classification (AMES data)

Distribution of predictive ability

not mutagenic, n_2 = 2970



mutagenic, n_1 = 3488



a_{FINAL} = 15

n = 6458, m = 1440 descriptors (variables), 20 repetitions (for each box plot)

Take time for validation

Consider variability

Accept variability/uncertainty

Like diversity

Thank You