Evaluation of empirical chemometric models for calibration and classification

Kurt VARMUZA

Vienna University of Technology

Institute of Chemical Engineering

Laboratory for **C**hemo**M**etrics

www.lcm.tuwien.ac.at, kvarmuza@email.tuwien.ac.at

Autumn School of Chemoinformatics, 15 - 16 November 2011, 16 Nov. 2011

The University of Tokyo

Contents

Common situation in science

available data

$x_1, x_2, \ldots, x_m$

= vector $\boldsymbol{x}^T$

Measured or calculated

desired data

y *(e.g., property)*

Cannot be determined directly or only with high cost

model

$$y = f(x_1, x_2, \ldots, x_m) = f(\boldsymbol{x}^T)$$

f: mathematical equation or algorithm,

derived from data (**empirical model**) or from knowledge (theoretical model)

Common situation in science 1/3

available data

$x_1, x_2, \ldots, x_m$

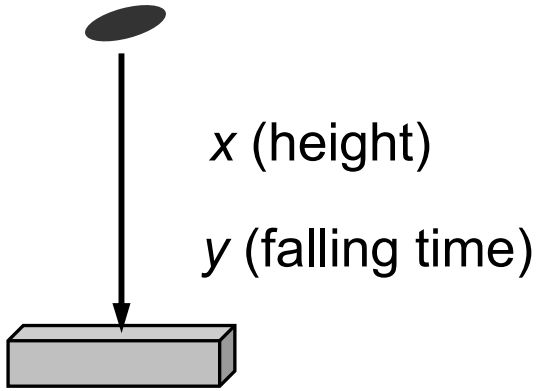
= vector $\boldsymbol{x}^T$

Measured or calculated

desired data

y *(e.g., property)*

Cannot be determined directly or only with high cost



Fundamental (scientific) law,
first principle

$y = (2x / g)^{0.5}$

model parameter g: gravity constant

Common situation in science 2/3

available data

$x_1, x_2, \ldots, x_m$

= vector $\boldsymbol{x}^T$

Measured or calculated

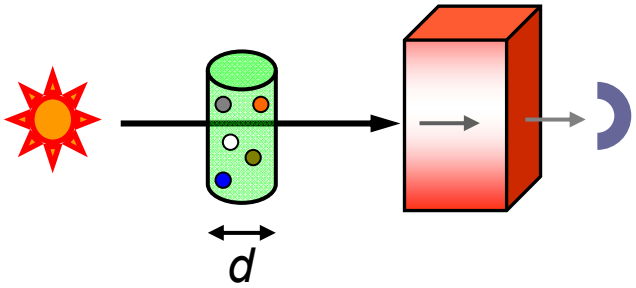
desired data

y *(e.g., property)*

Cannot be determined directly or only with high cost

$x_1, x_2, \ldots, x_m$ (N)IR absorbances

y concentration of a **compound**



Lambert-Beer's law,

reasonable relationship between x and y (parameter unknown)

$$y = \log (I_0 / I) / (\alpha d)$$

$$y = \sum_{\nu} {}_{\nu}\beta \log({}_{\nu}I_0 / {}_{\nu}I) + \beta_0$$

Common situation in science 3/3

available data

$x_1, x_2, \ldots, x_m$

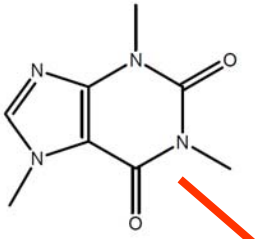
= vector $\boldsymbol{x}^T$

Measured or calculated

desired data

y *(e.g., property)*

Cannot be determined directly or only with high cost



set of numbers, molecular descriptors, $\boldsymbol{x}^T$

property y

Only an assumption:

y (property) is simply related with x (variables)

= "very empirical" ("dangerous")

$$y = \boldsymbol{x}^T \boldsymbol{b} + b_0 \quad \text{(linear model)}$$

Common situation in chemistry

available data

$x_1, x_2, \ldots, x_m$

= vector $\boldsymbol{x}^T$

Measured or calculated

desired data

y *(e.g., property)*

Cannot be determined directly or only with high cost

(N)IR absorbances	>>>	concentration of a substance in a complex mixture
molecular descriptors	>>>	property, activity (QSPR/QSAR)
spectral data	>>>	origin of sample
spectral data	>>>	presence/absence of chem. substructure

Contents

Empirical (linear) models

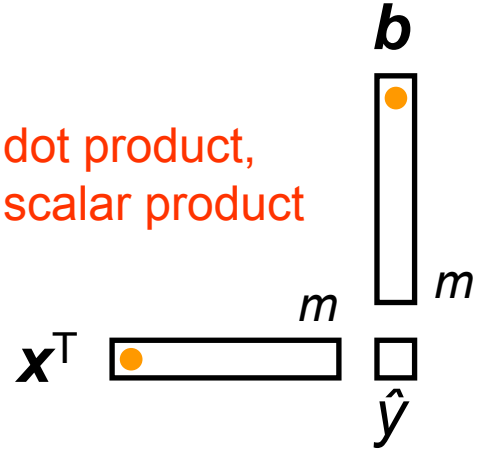
$y = f(x_1, x_2, ..., x_m)$

y	● continuous	multivariate calibration
	● discrete, categorical	multivariate classification pattern recognition

Linear model

$$\hat{y} = \boldsymbol{x}^T \boldsymbol{b} + b_0 = b_1 x_1 + b_2 x_2 + ... + b_m x_m + b_0$$

- $\hat{y}$: calculated (predicted) property ["*y-hat*"]
- $\boldsymbol{x}^T$: vector with variables (features, descriptors)
- $\boldsymbol{b}$: vector with regression coefficients
- b_0: intercept



Empirical (linear) models

- errors
- tolerance intervals,
- no unique solutions

Creation of a model:
estimation of the
model parameters ($\boldsymbol{b}$, b_0)
from given data, $\boldsymbol{X}$ and $\boldsymbol{y}$
(calibration set)

Linear model

$$\hat{y} = \boldsymbol{x}^T \boldsymbol{b} + b_0 = b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + b_0$$

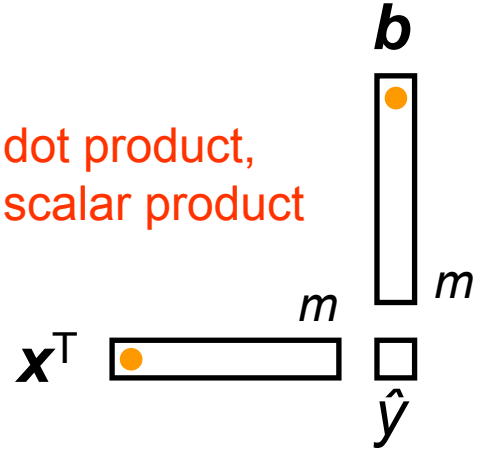
intercept

vector with regression coefficients

vector with variables (features, descriptors)

calculated (predicted) property ["*y-hat*"]

dot product,
scalar product



Empirical (linear) models

Creation of a model:
estimation of the
model parameters ($\boldsymbol{b}$, b_0)
from given data, $\boldsymbol{X}$ and $\boldsymbol{y}$
(calibration set)

Guiding principle

NOT best fit of the calibration data is important,

BUT **optimum prediction for new cases**
(**test set** data, never used in model creation)

Empirical (linear) models

1

Optimum
model complexity
estimated from
calibration data

2

Performance
of model
estimated from
test set data

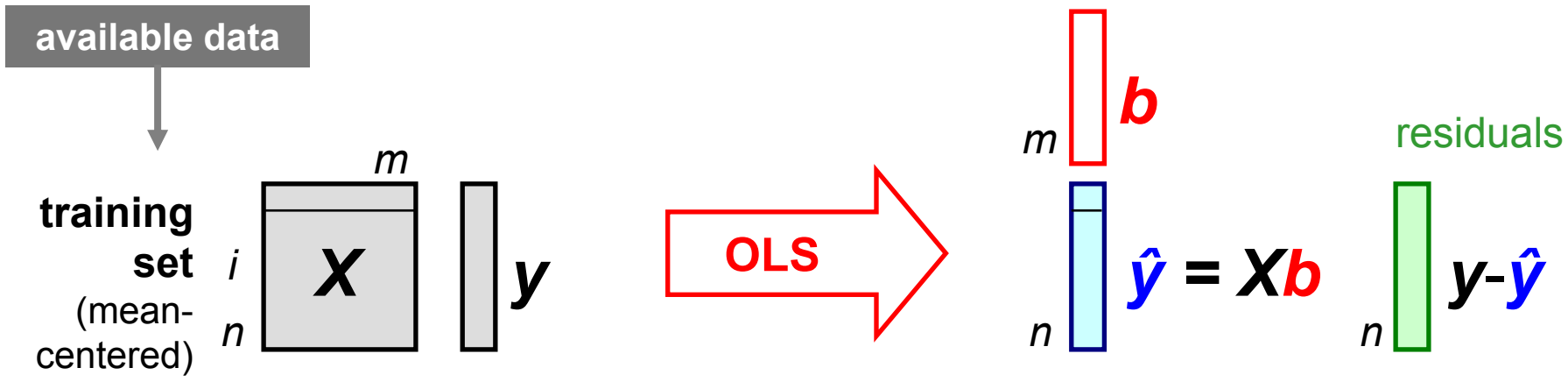
Guiding principle

NOT best fit of the calibration data is important,

BUT **optimum prediction for new cases**
(**test set data**, never used in model creation)

Creation of linear regression models

only a few selected topics



OLS Ordinary Least-Squares Regression

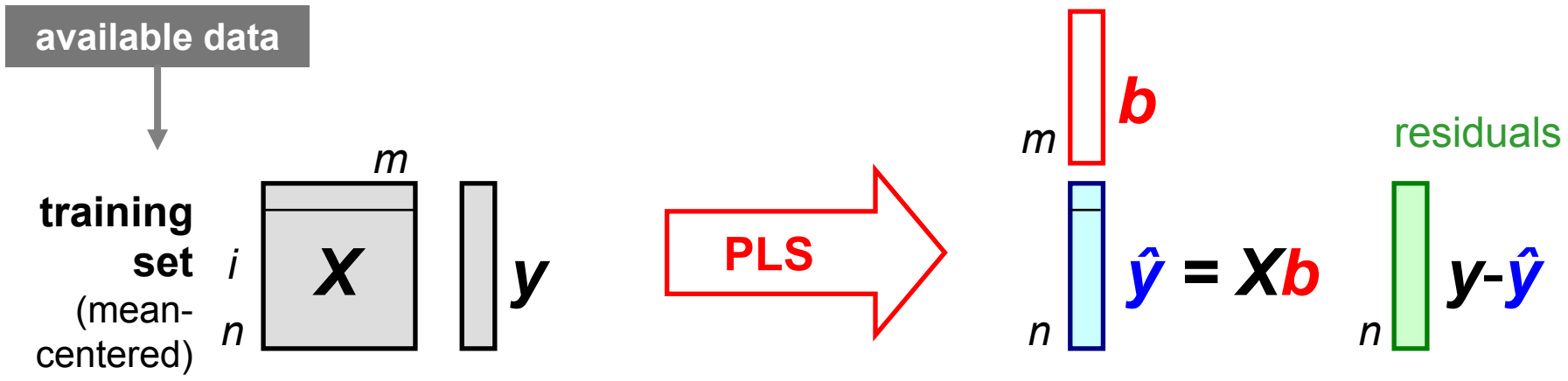
$\sum_i (y_i - \hat{y}_i)^2 \rightarrow \min$ $\qquad \boldsymbol{b}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Requirements
- $m < n$
- no highly correlating *x*-variables (columns)

☹ No optimization of model complexity (possibly a variable selection); rarely applicable in chemistry

Creation of linear regression models

only a few selected topics



PLS Partial Least-Squares Regression

simplified

(1) $U_{PLS} = X B_{PLS}$

Intermediate, linear (latent) variables (**components**):
- ☐ maximum covariance with *y*,
- ☐ uncorrelated or orthogonal directions in *x*-space,
- ☐ number of PLS-components is optimized

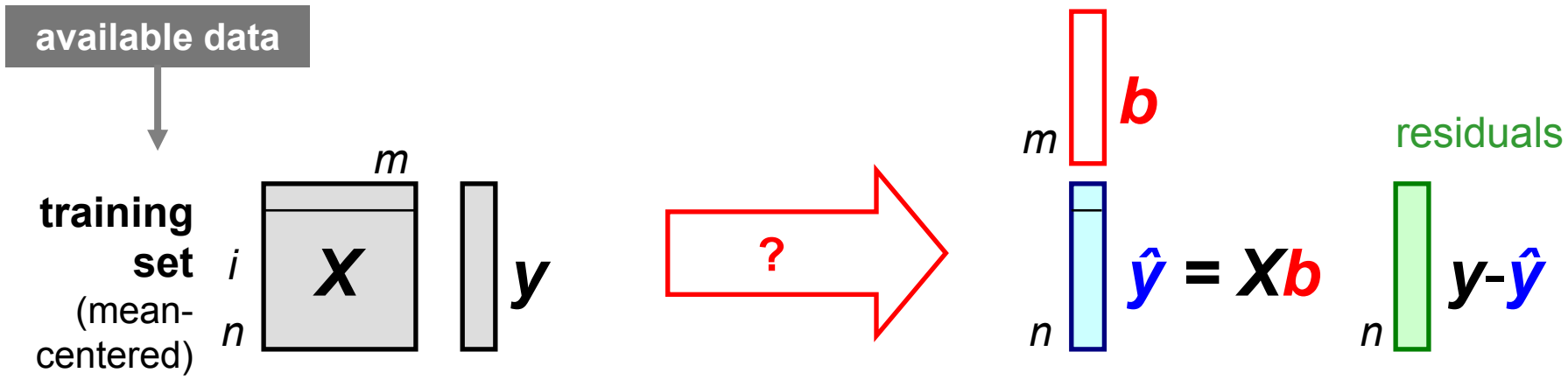
(2) OLS with U_{PLS}

☺
- ☐ applicable if $m > n$,
- ☐ applicable for highly correlating variables,
- ☐ optimization of model complexity !!!

☹ Various different approaches and algorithms

Creation of (linear) regression models

only a few selected topics



Some other regression methods in chemometrics

PCR Principal Component Regression (similar to PLS)

Lasso includes variable selection

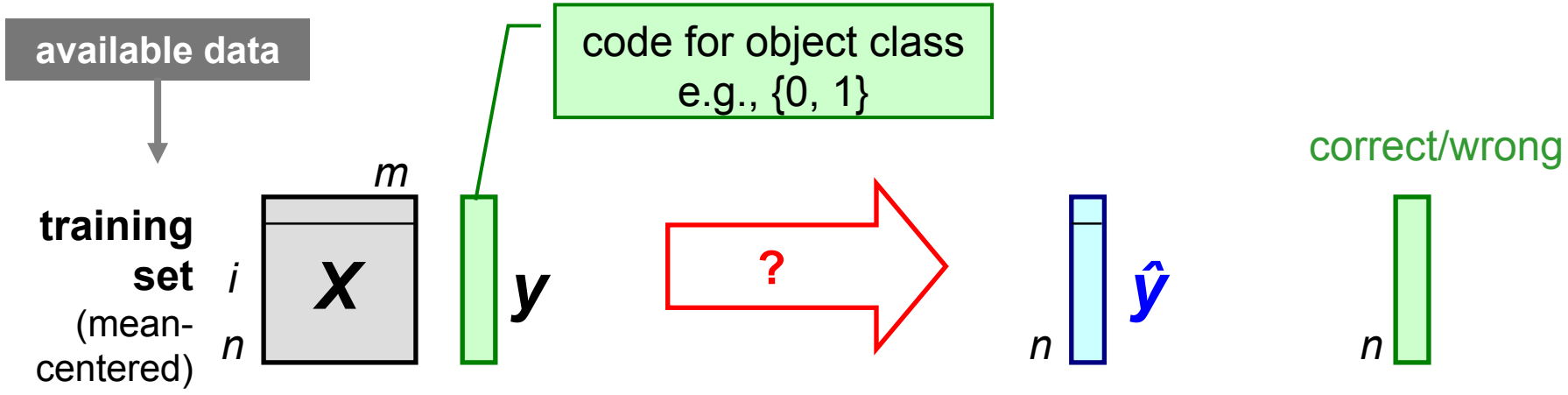
Ridge similar to PCR (weighting of all PCA scores)

ANN Artificial Neural Networks (nonlinear)

PLS2 PLS for more than one *y*-variable

Multivariate classification

only a few selected topics



D-PLS Discriminant PLS *

Binary classification (2 classes): y = -1 and +1 for class 1 and 2, resp.

PLS applied as a regression method (resulting in a discriminant vector $\boldsymbol{b}$).
Optimization of model complexity: number of PLS-components.

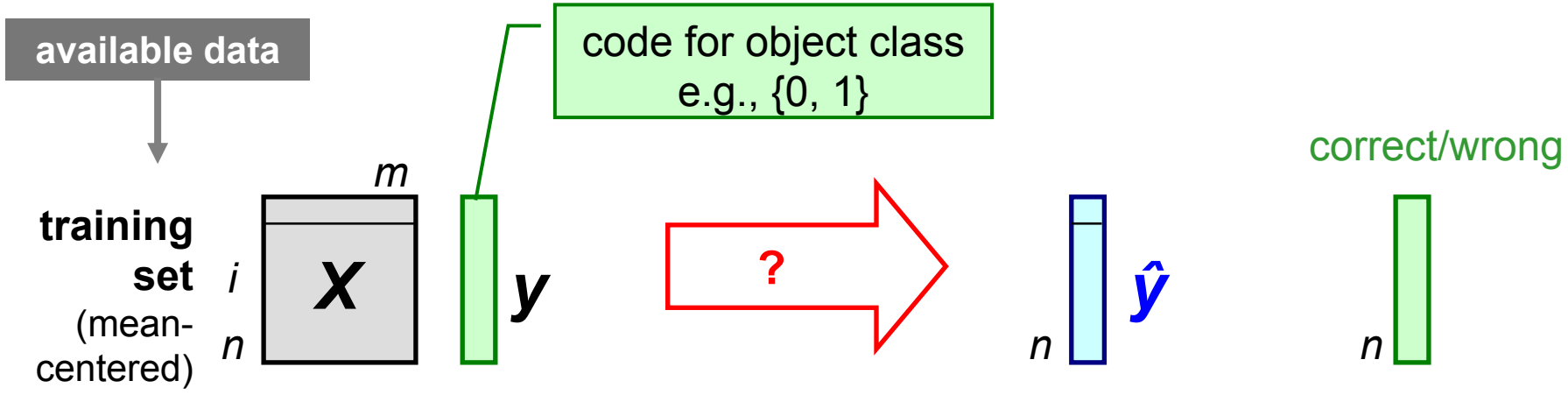
Class assignment: $\hat{y} = \boldsymbol{x}^T\boldsymbol{b}$; if $\hat{y} < 0$ assign to class 1, else to class 2.

Often used instead of LDA (linear discriminant analysis) because of advantages of PLS

* D-PLS is not recommended for >2 classes

Multivariate classification

only a few selected topics



KNN (*k*-nearest neighbor) classification

An algorithm; nonlinear; no discriminant vector.

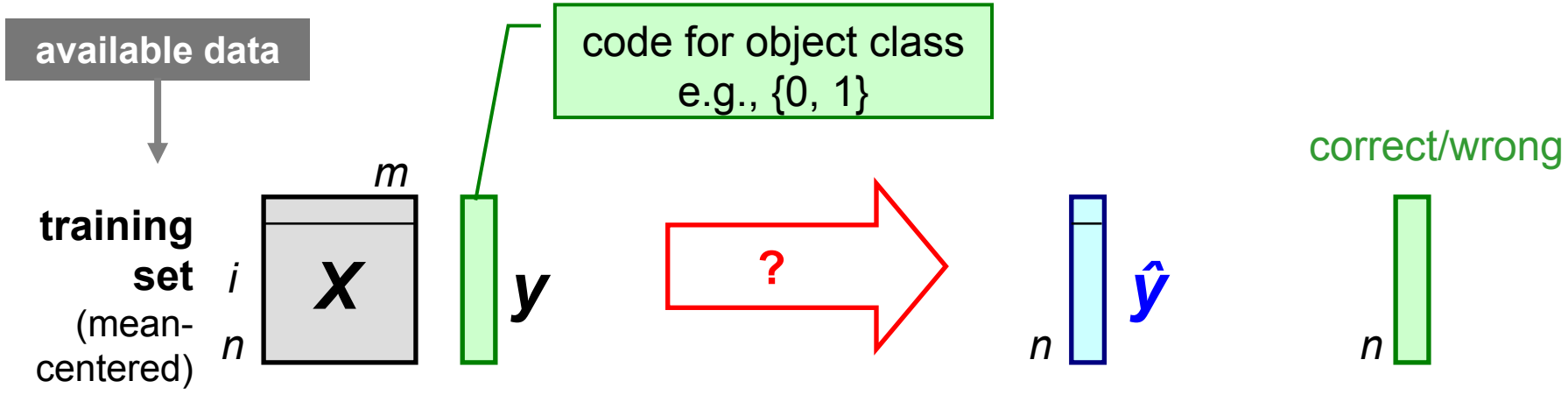
Usually the **Euclidean distance** between objects (in *x*-space) is used to find the nearest neighbors (objects with known class membership) to a query object.

A majority voting among the neighbors determines the class of the query object.

Optimization of model complexity: *k*, number of neighbors.

Multivariate classification

only a few selected topics



Some other classification methods in chemometrics

SVM Support Vector Machine (nonlinear)

CART Classification tree (nonlinear, evident)

SIMCA PCA models for each class (nonlinear, outlier detection)

ANN Artificial Neural Networks (nonlinear)

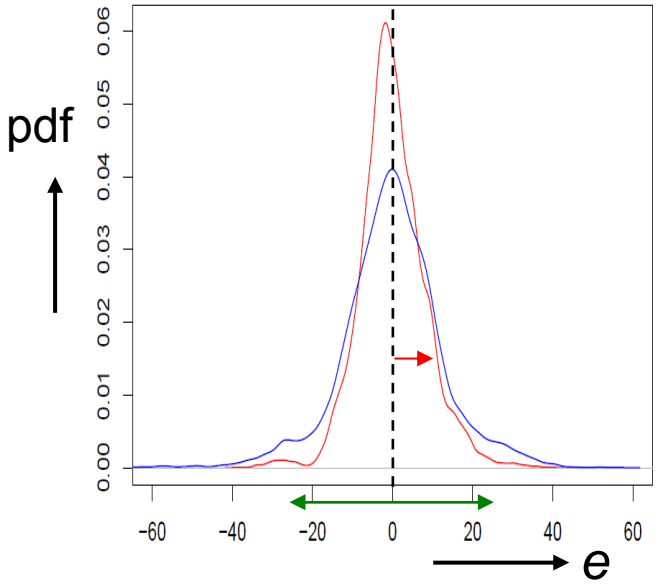
Contents

Performance measures (calibration)

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object *I* (residual)
$i = 1 \ldots z$	*z* is the number of objects used

Specify: ☞ which data set (calibration set, test set)
☞ which strategy (cross validation, ...)

Distribution of prediction errors



bias = mean of prediction errors e_i

SEP = standard deviation of prediction errors e_i

= **S**tandard **E**rror of **P**rediction

SEC = **S**tandard **E**rror of **C**alibration

CI = confidence interval, $CI_{95\%} \approx \pm 2*SEP$

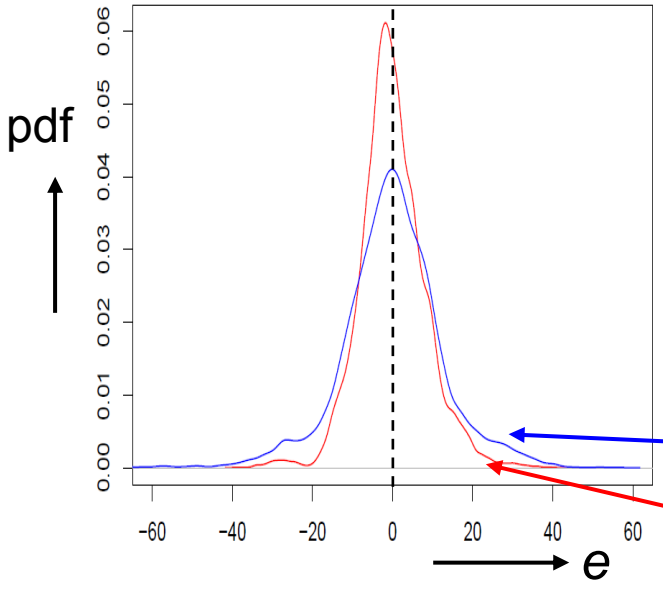
All in units of *y* ! Result: $\hat{y} \pm 2*SEP$

Performance measures (calibration)

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object *i*
$i = 1 \ldots z$	*z* is the number of objects used

Specify: ☞ which data set (calibration set, test set)
☞ which strategy (cross validation, ...)

Distribution of prediction errors



Modeling the GC retention index (*y*) for *n* = 208 PAC by *m* = 467 molecular descriptors (*Dragon* software).

Repeated double cross validation (rdCV) with 100 repetitions (*z* = 20 800)

m = 467; SEP = 12.7

m = 13; SEP = 8.2

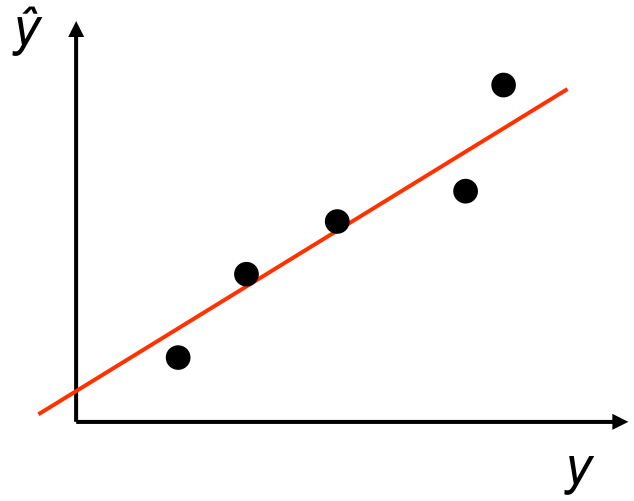
test set objects

Performance measures (calibration)

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object *i*
$i = 1 \ldots z$	*z* is the number of objects used

Specify: ☞ which data set (calibration set, test set)
☞ which strategy (cross validation, ...)

Predicted versus reference *y*'s



R^2 = **squared (Pearson) correlation coefficient**

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

squared adjusted correlation coefficient

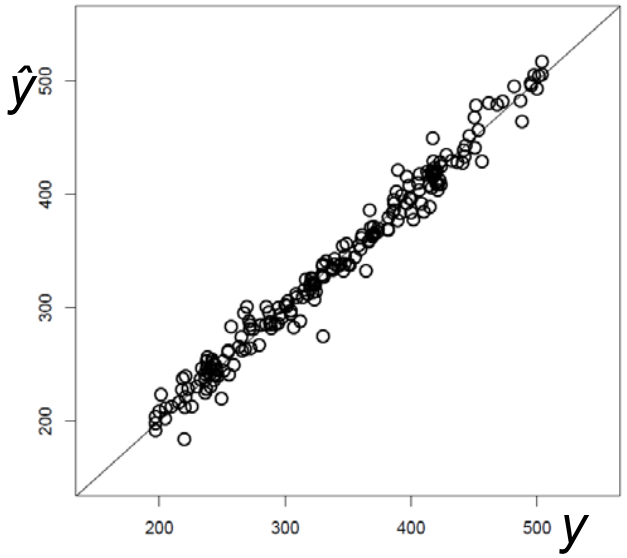
Penalizes models with a higher number of variables (*m*)

Performance measures (calibration)

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object *i*
$i = 1 \ldots z$	*z* is the number of objects used

Specify: ☞ which data set (calibration set, test set)
☞ which strategy (cross validation, ...)

Predicted versus reference *y*'s



Modeling the GC retention index (*y*) for *n* = 208 PAC by *m* = 467 molecular descriptors (*Dragon* software).

Repeated double cross validation (rdCV)
$\hat{y}$ are means of 100 repetitions

$R^2 = 0.979$ (test set objects)

Various other diagnostic plots.

Performance measures (calibration)

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object *i*
$i = 1 \ldots z$	*z* is the number of objects used
	Specify: ☞ which data set (calibration set, test set)
	☞ which strategy (cross validation, ...)

Some other measures

MSE	mean squared error	mean of prediction errors e_i
RMSE	root mean squared error	square root of MSE
PRESS	predicted residual error sum of squares	sum of squared errors e_i
AIC	Akaike's information criterion	consider *m*
BIC	Bayes information criterion	
C_p	Mallow's C_p	

Performance measures (classification)

Class assignment table (binary classification)

no. of objects

	assigned class 1	2	sum
true class 1	n_{11}	n_{12}	n_1
true class 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 $P = (P_1 + P_2)/2$

! Avoid: Overall predictive ability $= (n_{11} + n_{22})/n$

Performance measures (classification)

Example (warning)

$n = 100$; $n_1 = 95$; $n_2 = 5$

E. g.: All objects from class 1 are correctly classified;
all objects from class 2 are wrong classified.
Result: $P_1 = 1$; $P_2 = 0$; $P = 0.5$ (a bad classifier, OK)

However, $P_{OVERALL} = 0.95$ (although a bad classifier)

Predictive ability **class 1** $P_1 = n_{11}/n_1$

class 2 $P_2 = n_{22}/n_2$

Average predictive ability $P = (P_1 + P_2)/2$

! **Avoid: Overall predictive ability** $= (n_{11} + n_{22})/n$

Performance measures (classification)

Another notation: only for binary (positive/negative, bad/good) classifications

E. g.: medical test results ⟶		assigned class 1 *positive*	assigned class 2 negative	sum
positive,sick	1	n_{11} *true positive*	n_{12} *false negative*	n_1
negative,healthy	2	n_{21} *false positive*	n_{22} *true negative*	n_2
	sum	$n_{\rightarrow 1}$	$n_{\rightarrow 2}$	n

sensitivity = n_{11} (true positive) / n_1 (all sick) = P_1

"correct from all sick" (desired: near 1)

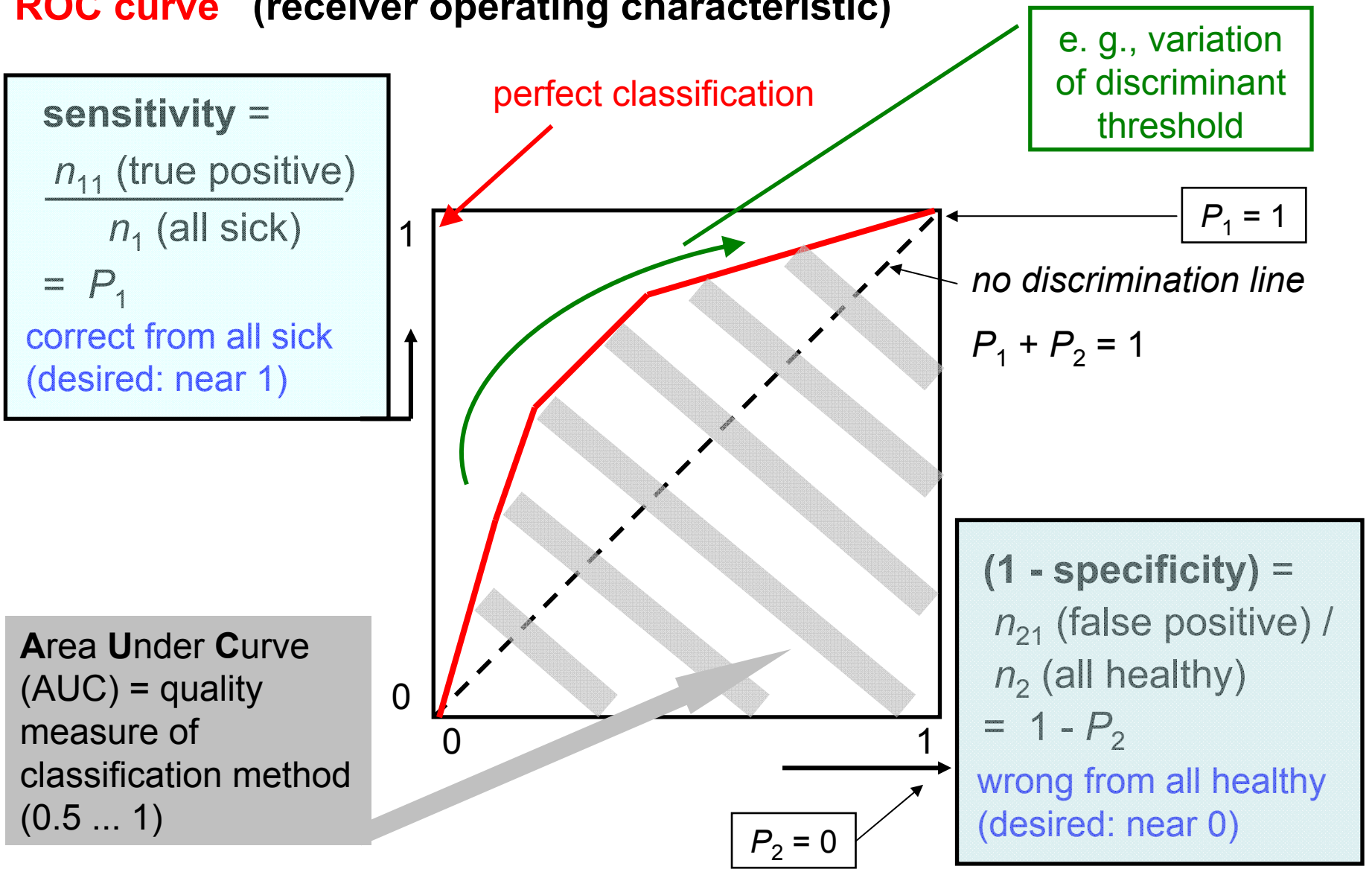
specificity = n_{22} (true negative) / n_2 (all healthy) = P_2

(1 - specificity) = n_{21} (false positive) / n_2 (all healthy) = 1 - P_2

"wrong from all healthy" (desired: near 0)

Performance measures (classification)

ROC curve **(receiver operating characteristic)**

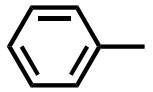


Performance measures (classification)

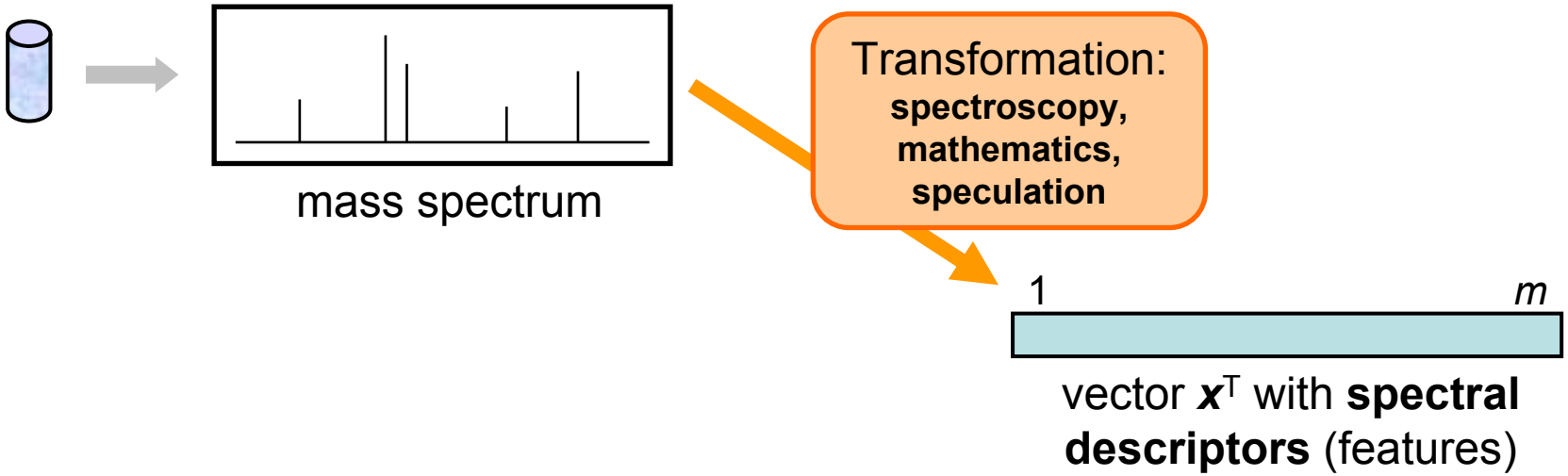
Other measures for classification performance

misclassification rate	for each class separately and summarized
risk of wrong classification	different risks for wrong classification of the different classes can be defined
rejection rate	if no assignment to any class is allowed (dead zone)
confidence of answers	ratio of correct answers (assignment to a specific class 1, 2, 3, ...); depends on relative group sizes like overall predictive ability

Performance measures (classification)



Example: Spectra-structure relationship (KNN, DPLS, SVM)



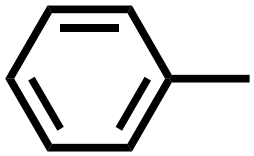
Binary classification

Chemical substructure present / not present (class 1 / class 2)

n = 600 (class 1: 300; class 2: 300), *m* = 658

Dataset '*phenyl*' in R-package '*chemometrics*'

C_6H_5 -



Werther W., Demuth W., Krueger F.R., Kissel J., Schmid E.R., Varmuza K.: *J. Chemom.*, **16**, 99 (2002)

Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, FL, USA (2009)

Performance measures (classification)

Example: Spectra-structure relationship (KNN, DPLS, SVM)

rdCV

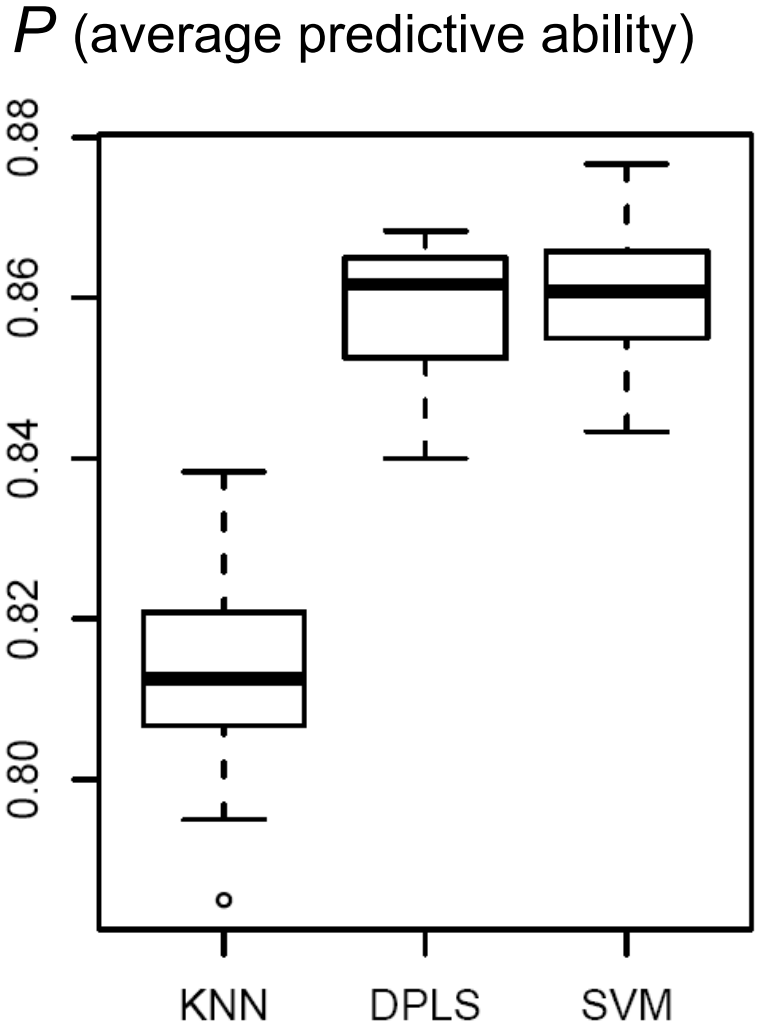
20 repetitions;
$s_{OUT} = 2$; $s_{IN} = 6$

Optimized parameter

KNN: $k_{FINAL} = 3$
DPLS: $a_{FINAL} = 2$
SVM: $\gamma_{FINAL} = 0.0002$

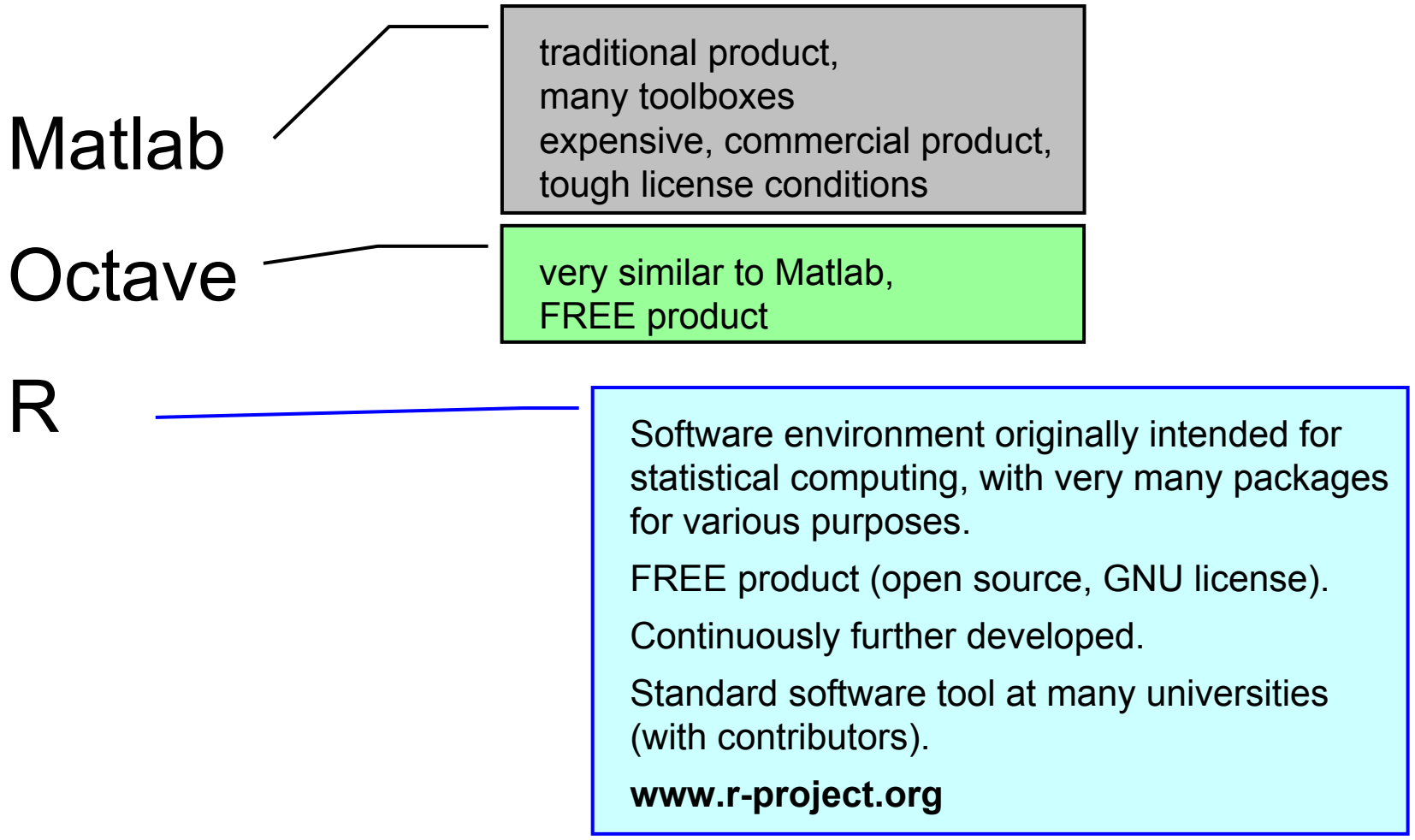
Computation time

KNN	550 s
DPLS	42 s
SVM	940 s



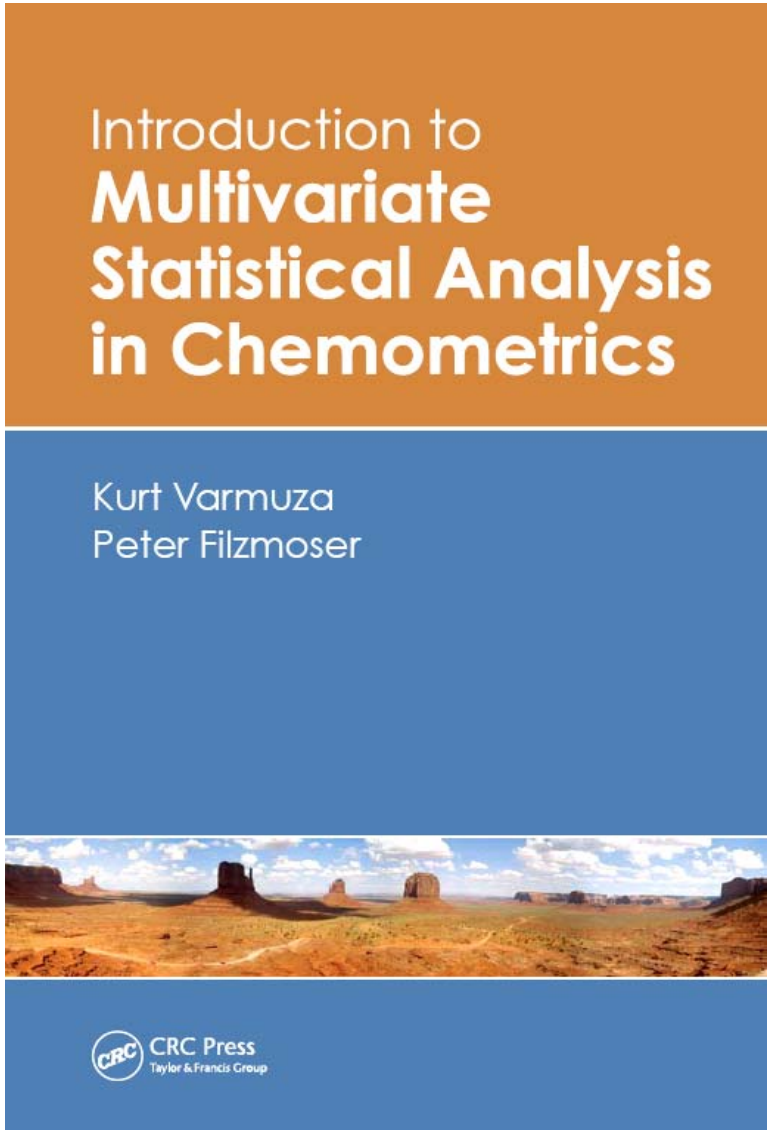
Contents

R (software system)



various software products
(Unscrambler, SIMCA, Sirius, ...)

Book including examples and data sets for R



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

Ca 320 pages,
appr. € 100

Includes many R-codes (examples)

However, description of methods without R

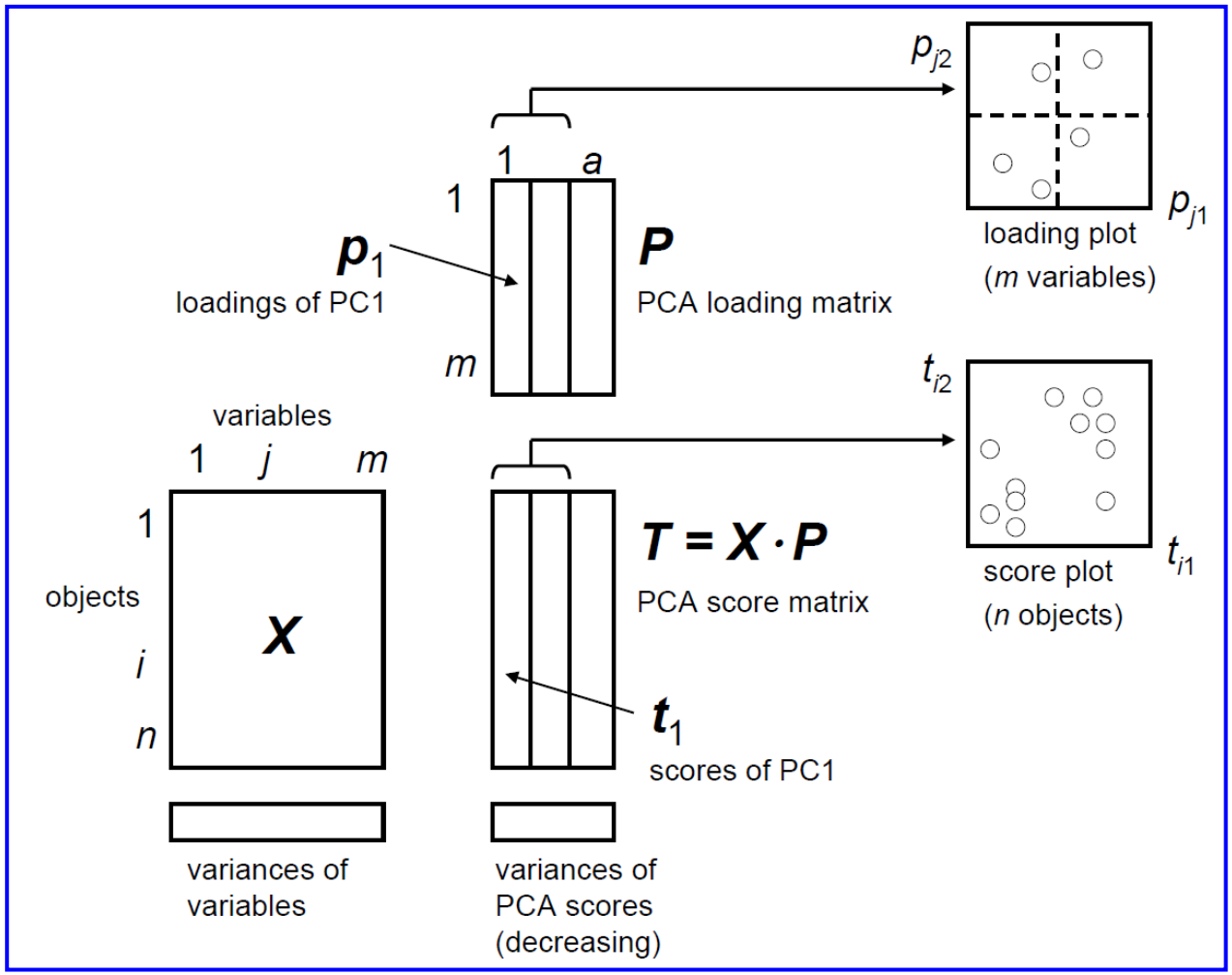
Info: www.lcm.tuwien.ac.at

Book including examples and data sets for R

R-package *chemometrics:* Sources, data sets, examples, tutorial (vignette), by P. Filzmoser and K. Varmuza

Example from book:

Principal **C**omponent **A**nalysis

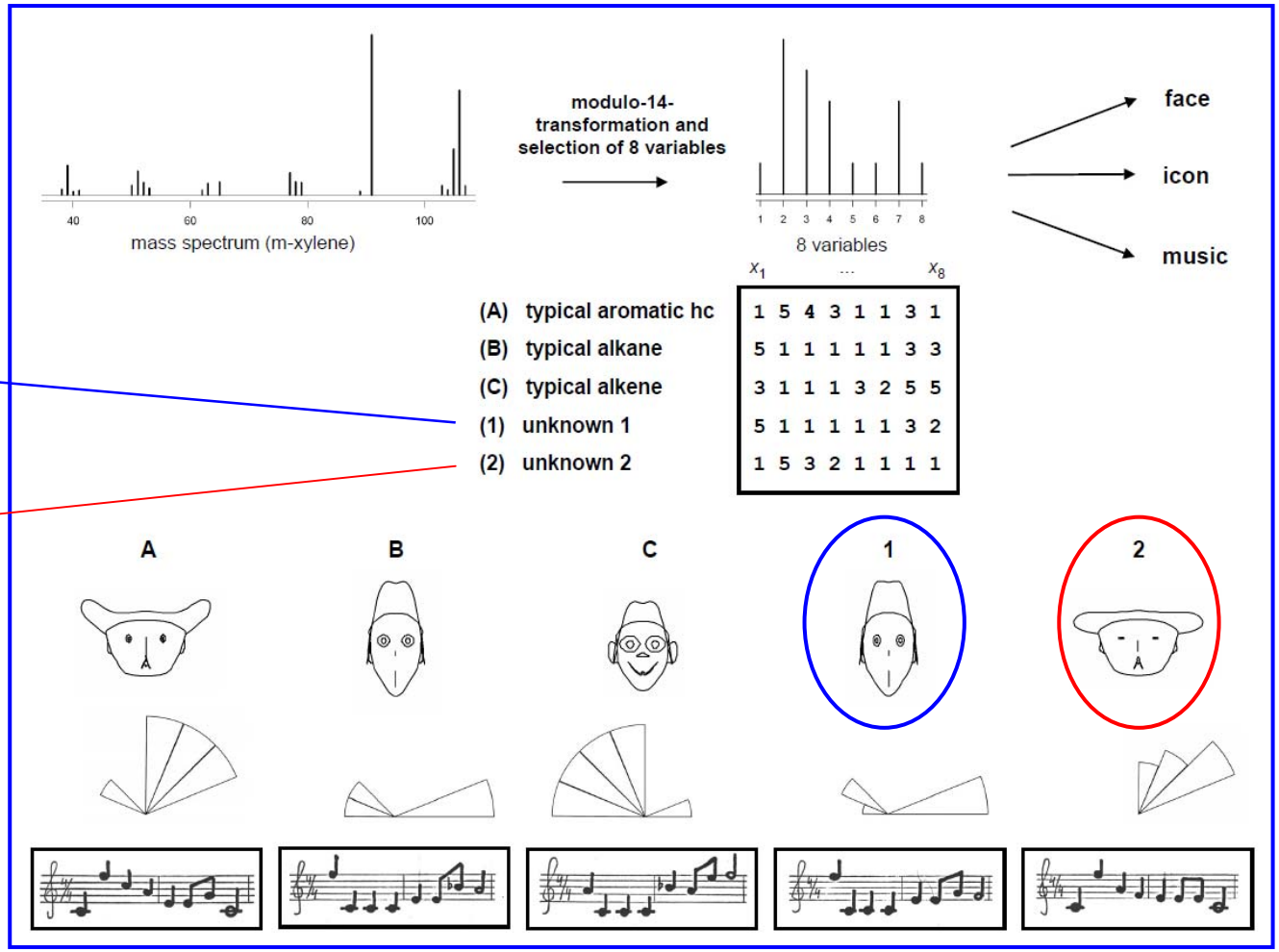


Book including examples and data sets for R

R-package *chemometrics:* Sources, data sets, examples, tutorial (vignette), by P. Filzmoser and K. Varmuza

2-methyl-heptane

meta-xylene



Example from book:

Visual
Cluster analysis

R (software system)

www.r-project.org

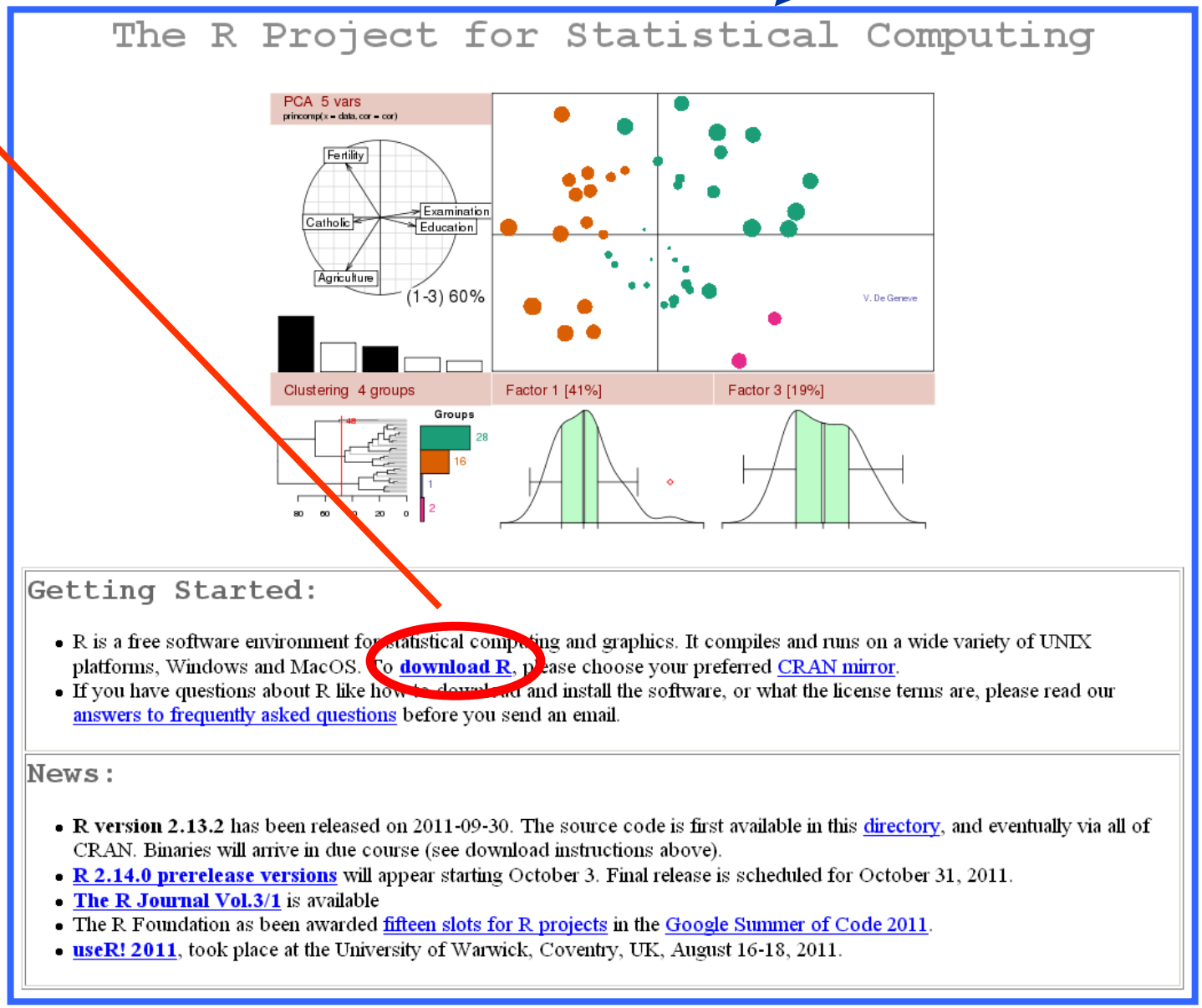
Select a "mirror", e.g., cran.md.tsukuba.ac.jp/

Download "R for Windows" (or Linux or Mac)

Select "base" and then "Download R 2.14.0 for Windows" (R-2.14.0-win.exe, 46 MB)

Perhaps download FAQs and Info's.

Run R-2.14.0-win.exe for easy installation.



Contents

Strategies for *optimum* models

Calibration — small SEP, large R^2

Classification — high P_1, P_2, ...

for NEW CASES

model with **optimum complexity** — 1

separated

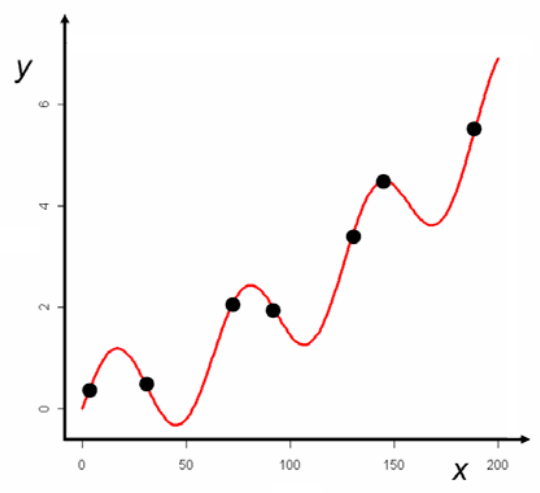
2 — performance measures from **test set** objects

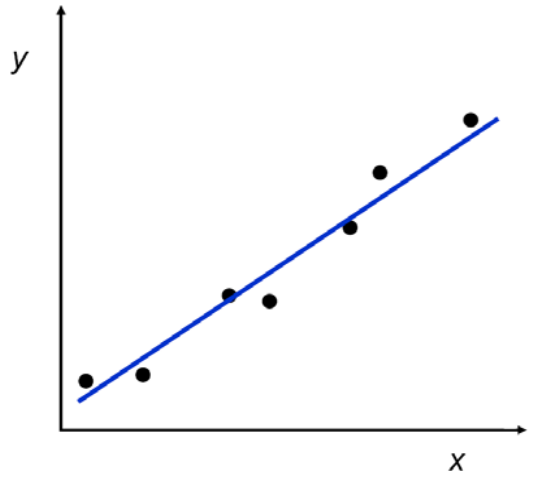
Best fit for calibration data is not the primary aim (overfitting),

but high performance for new objects.

Estimations !

Strategies (1) Optimum model complexity





fit error = 0

too high complexity of model

overfitted

error for new cases probably large

fit error > 0

perhaps better (optimum) complexity of model

perhaps optimal fitted

error for new cases probably smaller than for overfitted model

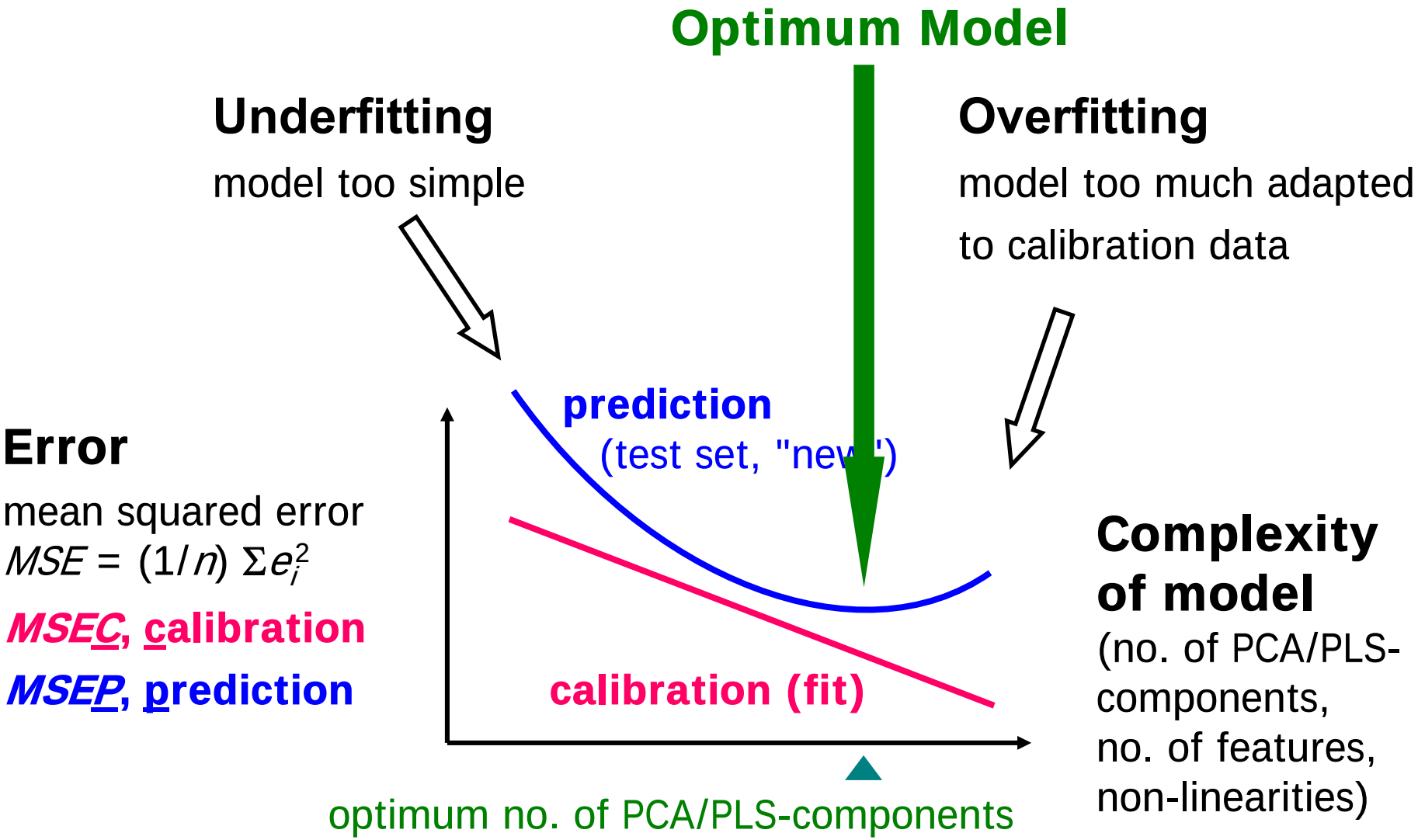
Strategies (1) Optimum model complexity

Optimum complexity of model has to estimated by trial and error.
Usually not a unique solution.

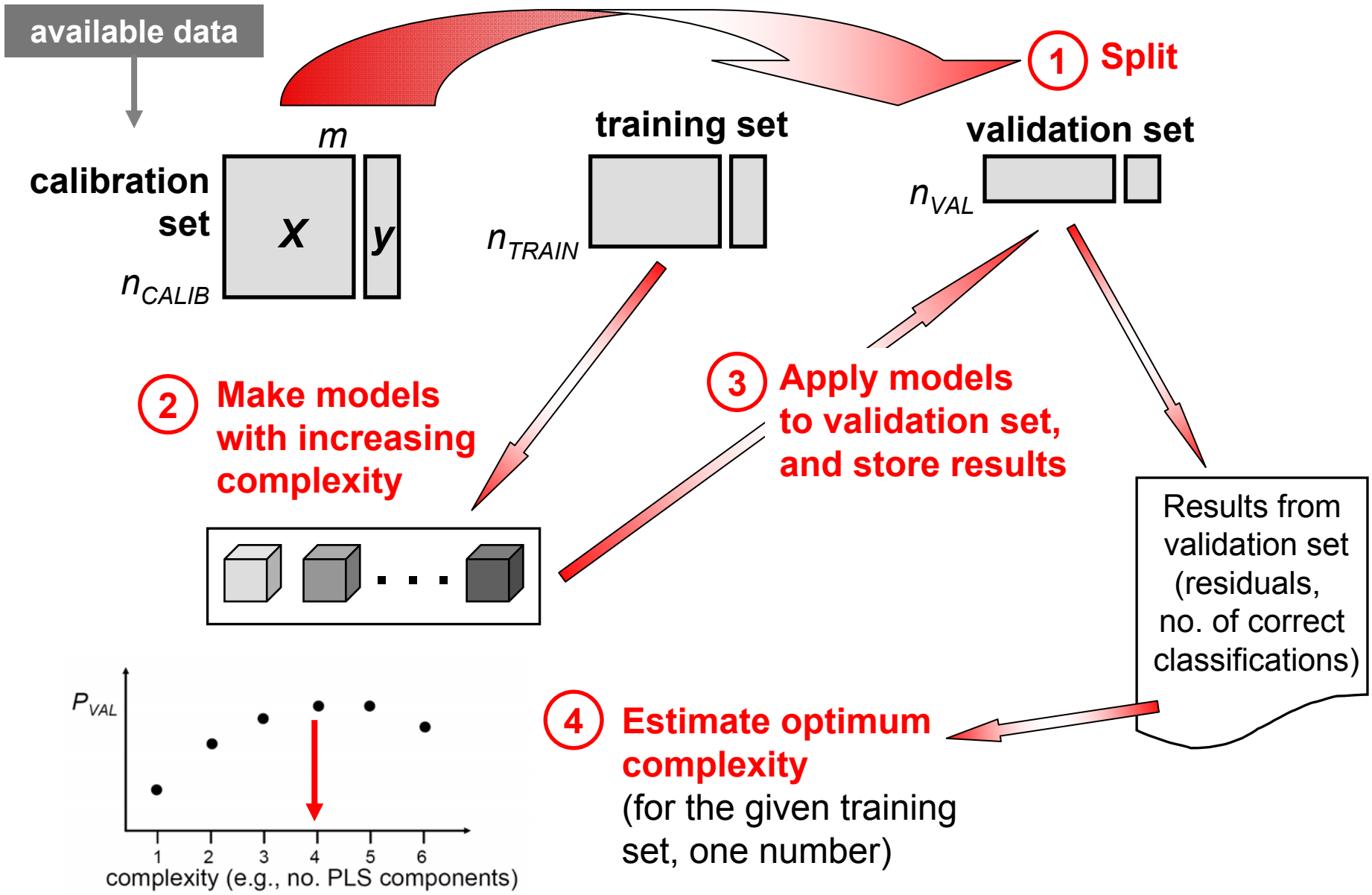
Optimum complexity: parameter of the method for model generation

Calibration		Classification	
PLS	no. of PLS components	**DPLS**	no. of PLS components
PCR	no. of PCA components	**PCA + LDA**	no. of PCA components
Ridge	complexity parameter λ_R	**KNN**	no. of neighbors
Lasso	complexity parameter λ_L	**SVM**	gamma
ANN	no. of hidden neurons	**SIMCA**	no.s of PCA components
OLS	(no. of variables)	**CART**	tree size
		ANN	no. of hidden neurons

Strategies (1) Optimum model complexity



Strategies (1) Optimum model complexity



Strategies (1) Optimum model complexity

Optimum model complexity: estimation, statistics

- ❑ more data are better,
- ❑ more estimations are better

However, usual data sets (in chemistry) are small
(number of objects, n = 20 ... 200)

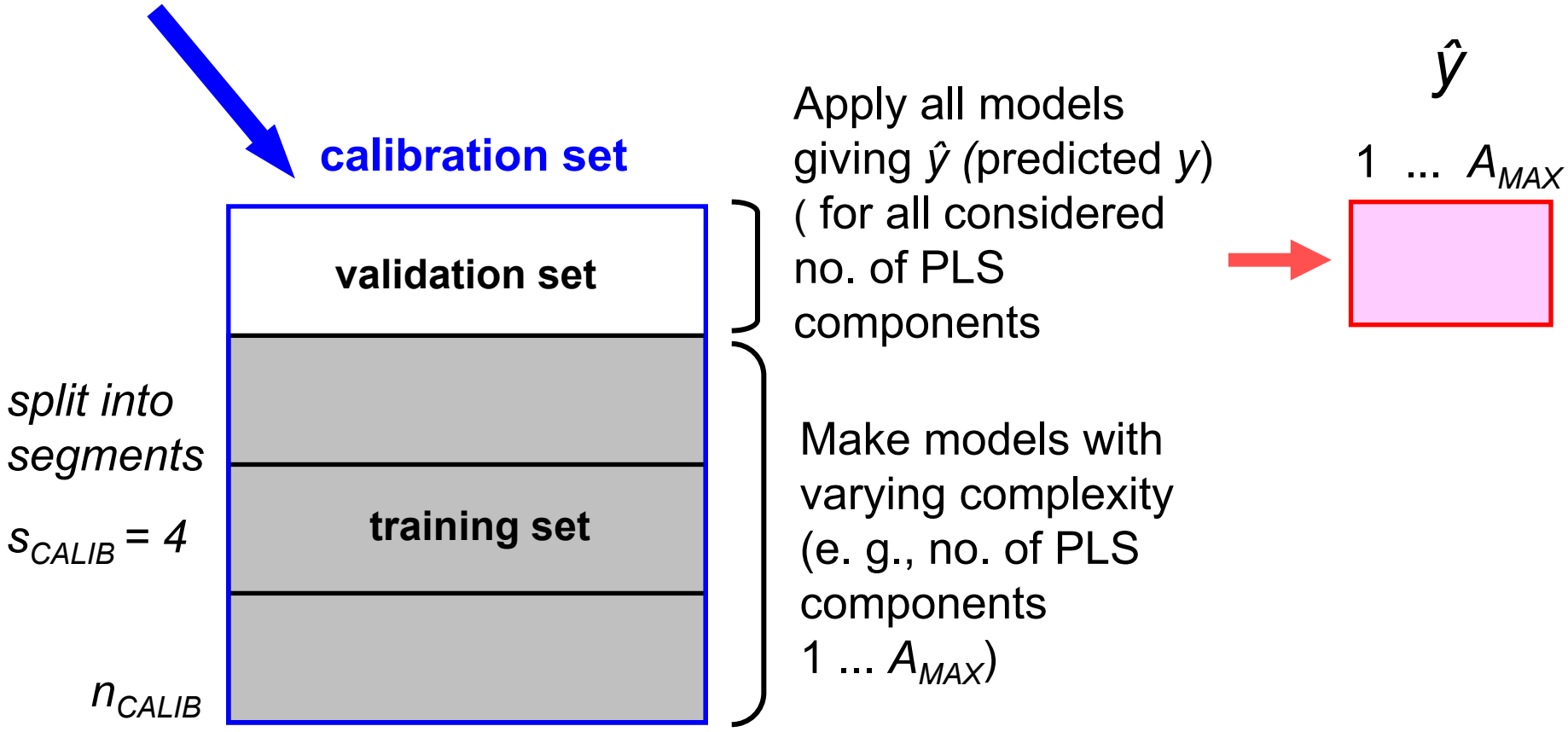
Resampling strategies

- ◆ **cross validation (CV)**
- ◆ bootstrap

Strategies for split of data set into subsets of objects:
calibration set (training set, validation set), test set

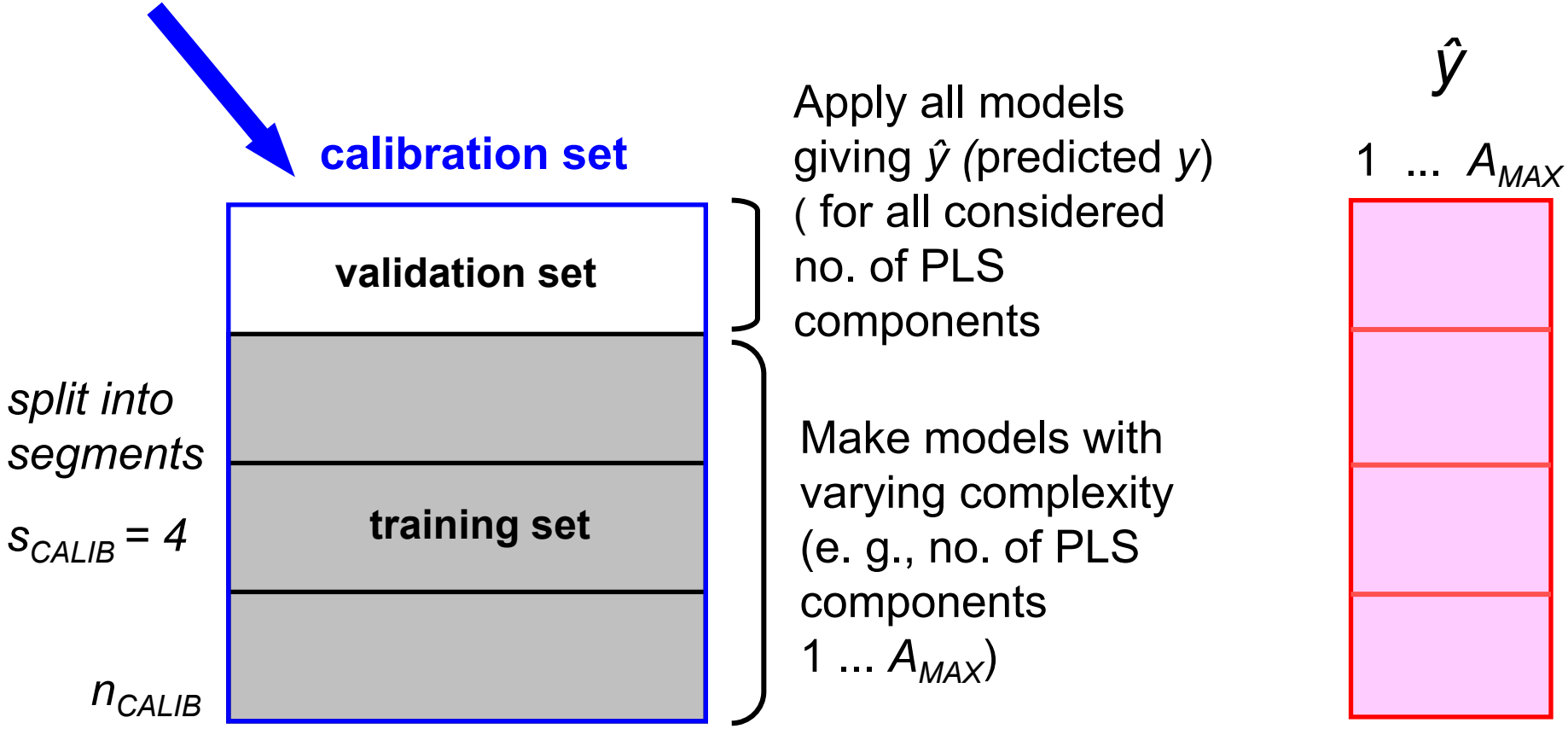
Strategies (1) Optimum model complexity **by CV**

available data



Strategies (1) Optimum model complexity **by CV**

available data

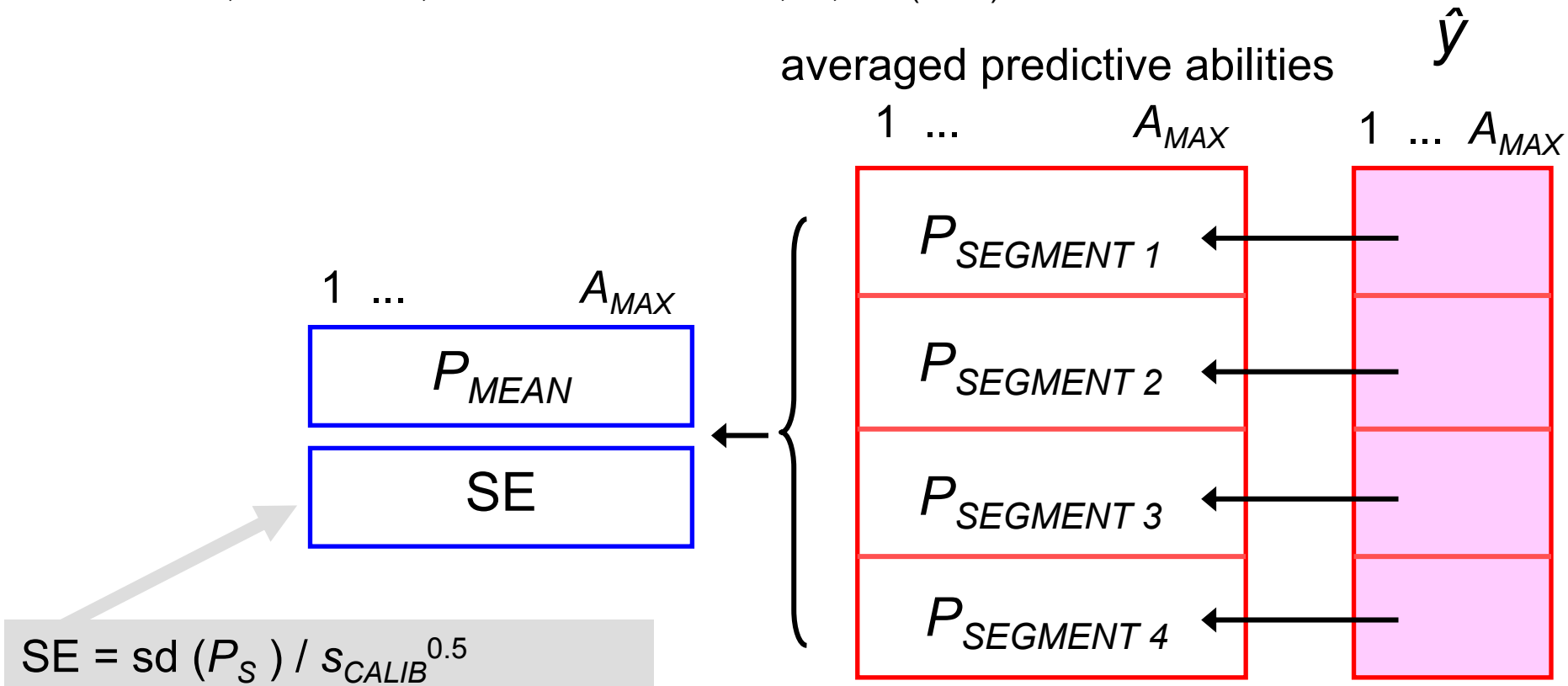


CV loop: each segment as validation set

Strategies (1) Optimum model complexity **by CV**

Evaluation of CV results based on One Standard Error Method

Hastie T.,Tibshirani R.J., Friedman J.: The Elements of Statistical Learning, Springer, New York (2001)
Filzmoser P., Liebmann B.,Varmuza K.: *J. Chemom.*, **23**, 160 (2009)



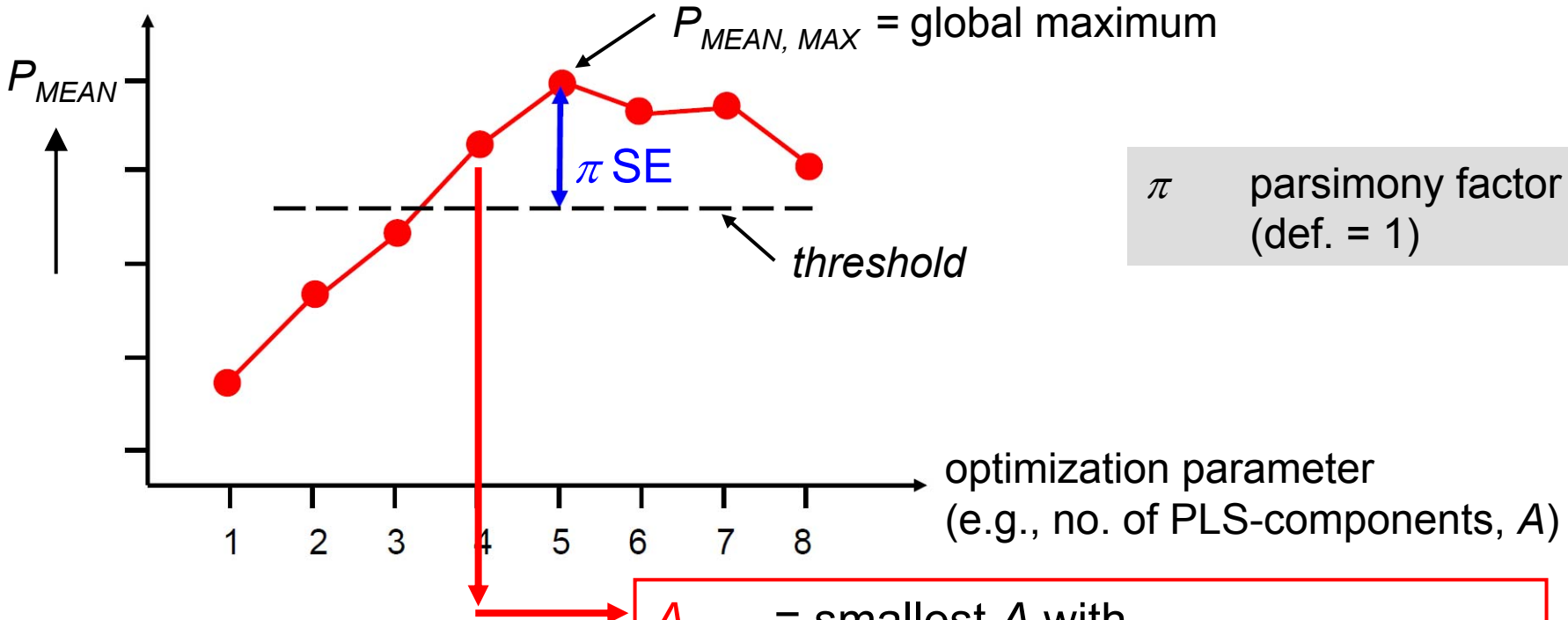
SE = sd (P_S) / $s_{CALIB}^{0.5}$
= standard error of mean

s_{CALIB} = no. of mean values
= no. of segments

Strategies (1) Optimum model complexity **by CV**

Evaluation of CV results based on One Standard Error Method

Hastie T.,Tibshirani R.J., Friedman J.: The Elements of Statistical Learning, Springer, New York (2001)
Filzmoser P., Liebmann B.,Varmuza K.: *J. Chemom.*, **23**, 160 (2009)



π parsimony factor (def. = 1)

A_{OPT} = smallest A with

$P_{MEAN} \geq P_{MEAN, MAX} - \pi SE$

(one number for given calibration set)

Strategies (1) Optimum model complexity **by CV**

Evaluation of CV results based on One Standard Error Method

Hastie T.,Tibshirani R.J., Friedman J.: The Elements of Statistical Learning, Springer, New York (2001)
Filzmoser P., Liebmann B.,Varmuza K.: *J. Chemom.*, **23**, 160 (2009)

☞ s_{CALIB} *number of segments in CV loop*

≥ 4 for a reasonable estimation of SE,
each segment with ≥ 5 objects

☞ π parsimony factor

$\pi = 0$ global maximum

$\pi = 1$ one standard error (default)

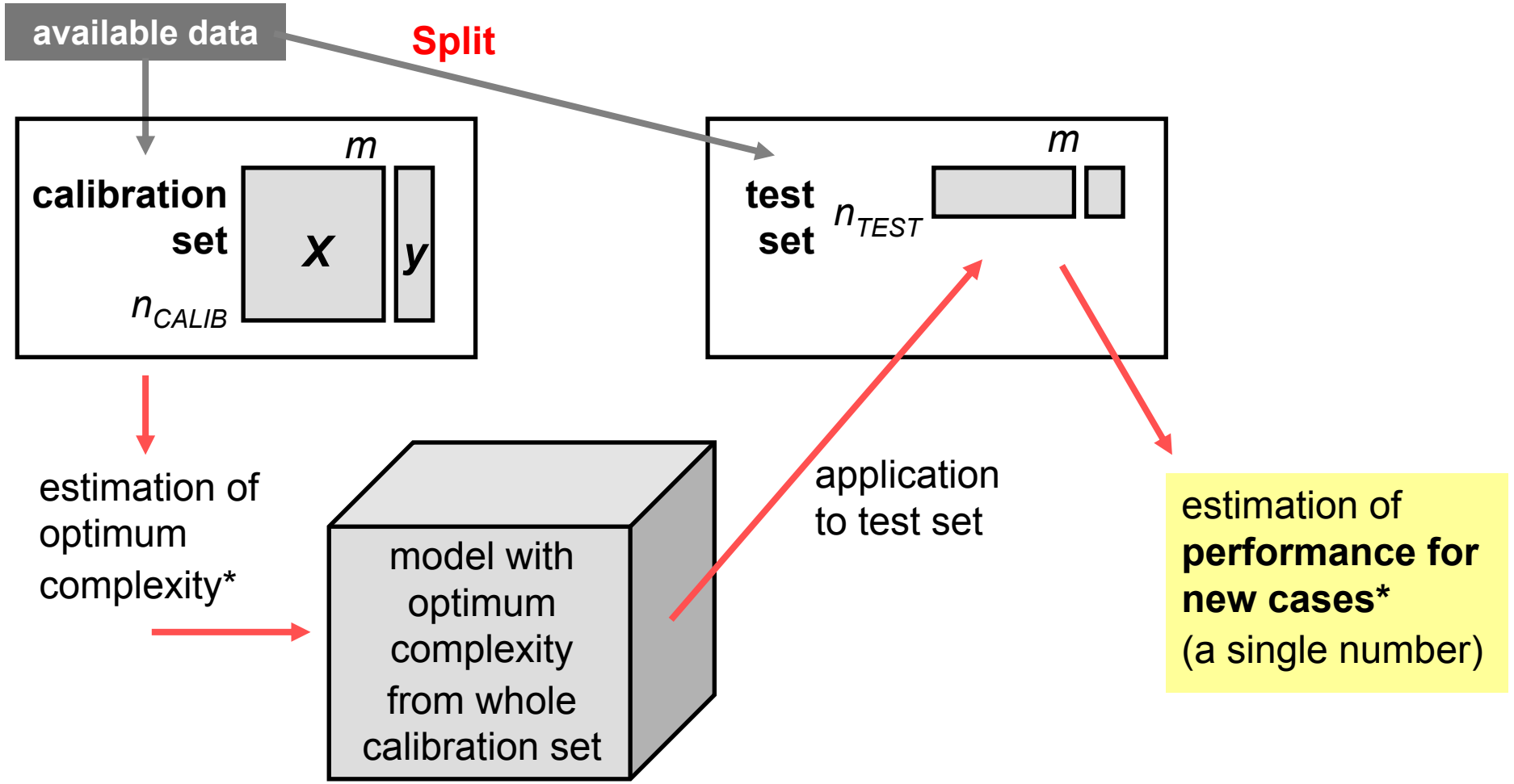
$\pi = 2$ 95% confidence interval

☞ Measure used for **calibration**: MSE (minimization)

Measure used for **classification**: *P* (maximization)

Also for regression methods applied to classification,
such as DPLS or SVM.

Strategies (2) Performance for new cases



* However, results depend on (random) split into calibration and test set

Strategies (3) repeated double Cross Validation (rdCV)

For calibration

Filzmoser P., Liebmann B., Varmuza K.: *J. Chemom.*, **23**, 160 (2009).
Repeated double cross validation.

Similar (*cross model validation and permutation*)

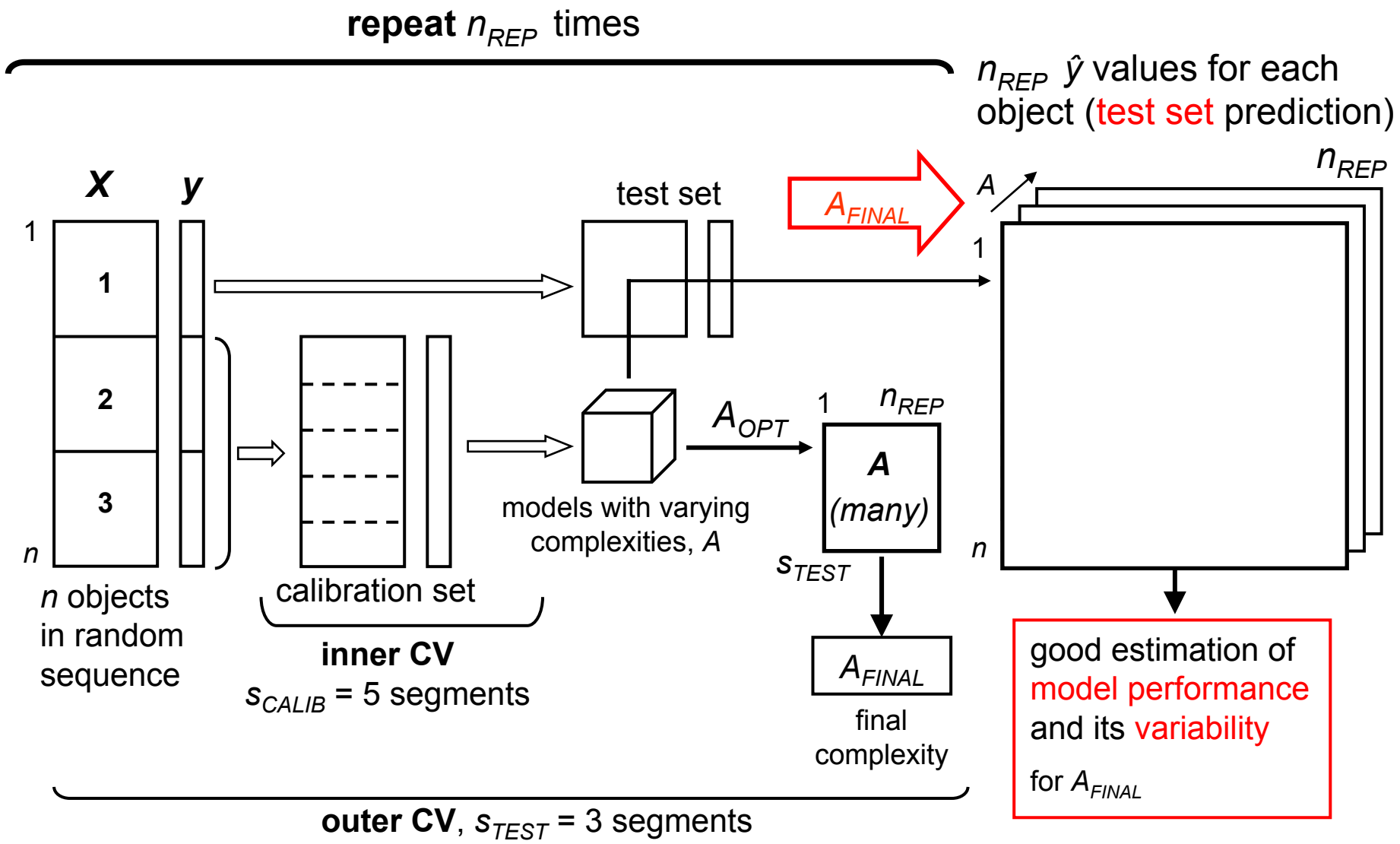
Westerhuis J.A. et al.: *Metabolomics*, **4**, 81 (2008).
Assessment of PLSDA cross validation.

Applications of rdCV

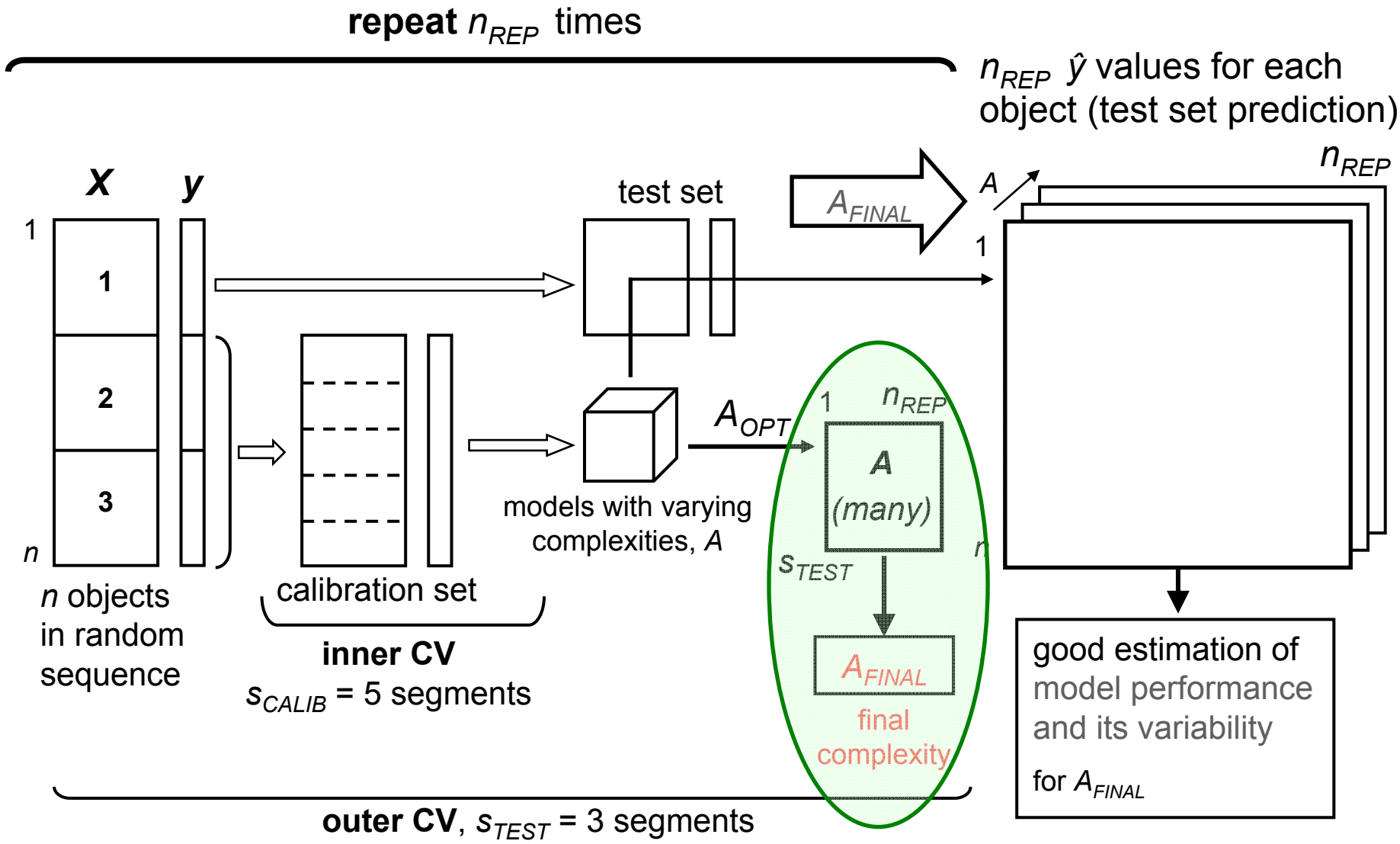
- Liebmann B., Friedl A., Varmuza K.: *Anal. Chim. Acta*, **642**, 171 (2009).
Determination of **glucose and ethanol in bioethanol** production by near infrared spectroscopy and chemometrics.
- Felkel Y., Dörr N., Glatz F., Varmuza K.: *Chemom. Intell. Lab. Syst.*, **101**, 14 (2010).
Determination of the **total acid number** (TAN) of used gas **engine oils** by **IR** and chemometrics applying a combined strategy for variable selection.
- Liebmann B., Filzmoser P., Varmuza K.: *J. Chemom.* **24**, 111 (2010). **Robust and classical PLS** regression compared.

R-package *chemometrics*; see also **www.lcm.tuwien.ac.at/R**

Strategies (3) repeated double Cross Validation (rdCV)

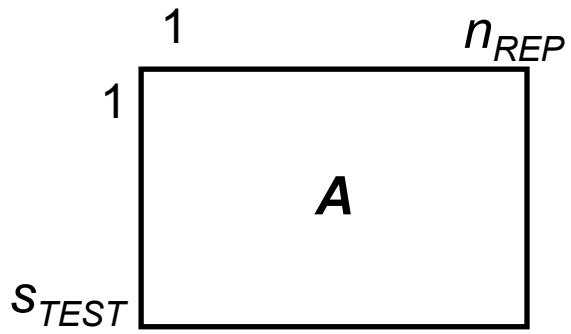


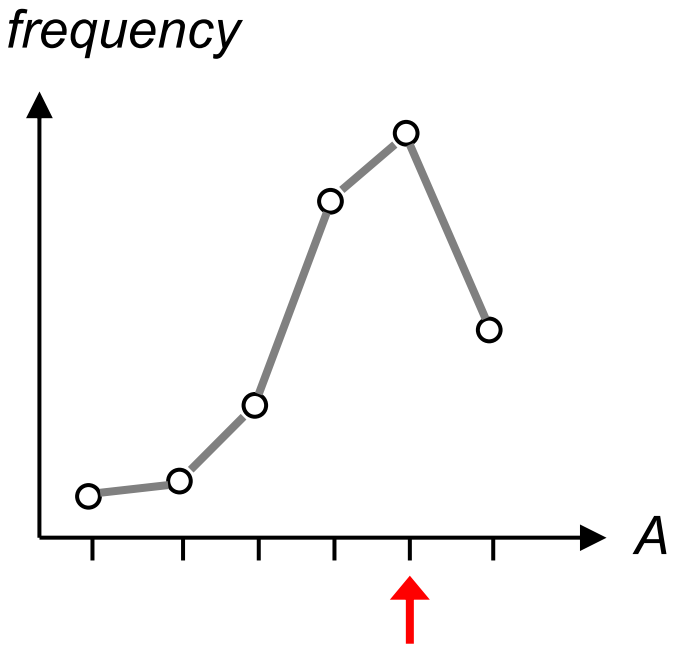
Strategies (4) Final optimum complexity (rdCV)



Strategies (4) Final optimum complexity (rdCV)

s_{TEST} * n_{REP} values for optimization parameter, *A*



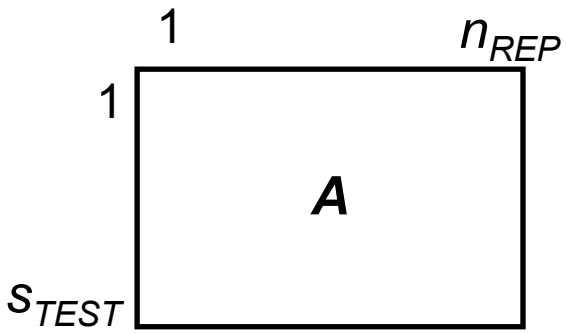


Typical, e. g.,
s_{TEST} = 3
n_{REP} = 100
give 300 estimations for the optimum complexity

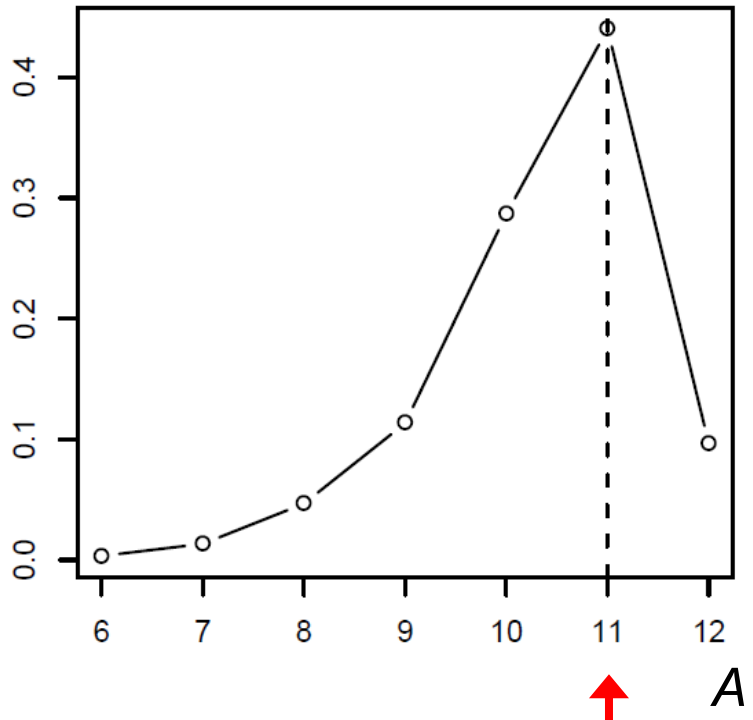
- Most frequent value of *A* = A_{FINAL}
- Or other heuristics, or a set of values for A_{FINAL} (consensus models)

Strategies (4) Final optimum complexity (rdCV)

s_{TEST} * n_{REP} values for optimization parameter, *A*



frequency (300 values)



optimum no. of PLS components; "final" value = 11

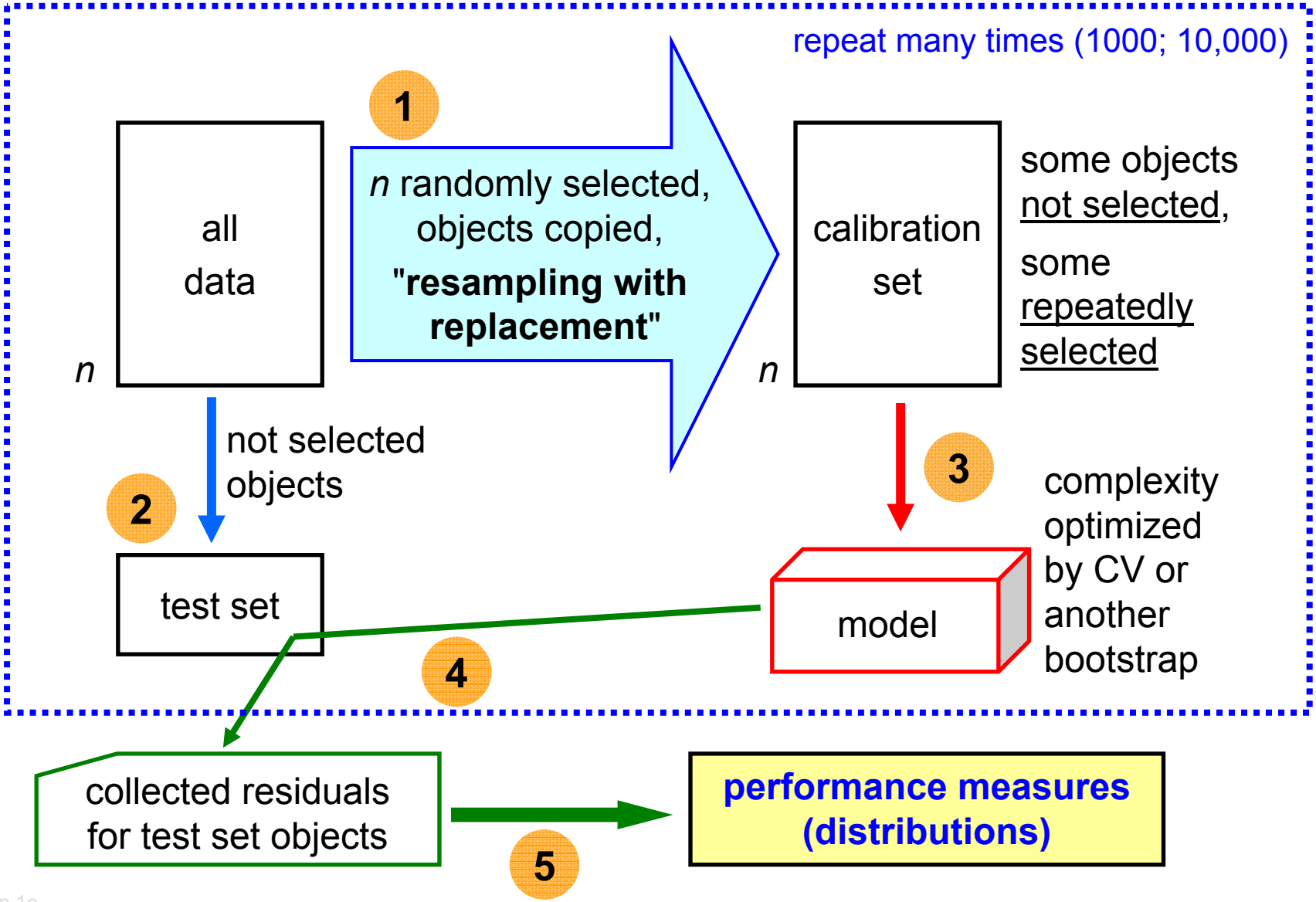
Modeling the GC retention index (*y*) for *n* = 208 PAC by *m* = 467 molecular descriptors (*Dragon* software).

rdCV with s_{TEST} = 3 segments in outer loop, and n_{REP} = 100 repetitions

Strategies (5) Summary of rdCV

- A resampling method combining some systematics and randomness.
- For calibration and classification.
- For data sets with ca ≥ 25 objects.
- Optimization of model complexity (model parameter) is separated from the estimation of model performance.
- Provides estimations of the variability of model complexity and of performance.
- Easily applicable and fast
  - R-package "*chemometrics*"
  - www.lcm.tuwien.ac.at/R

Strategies (6) Bootstrap



Strategies (6) Bootstrap

Advantages

+ simple,

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

Disadvantages

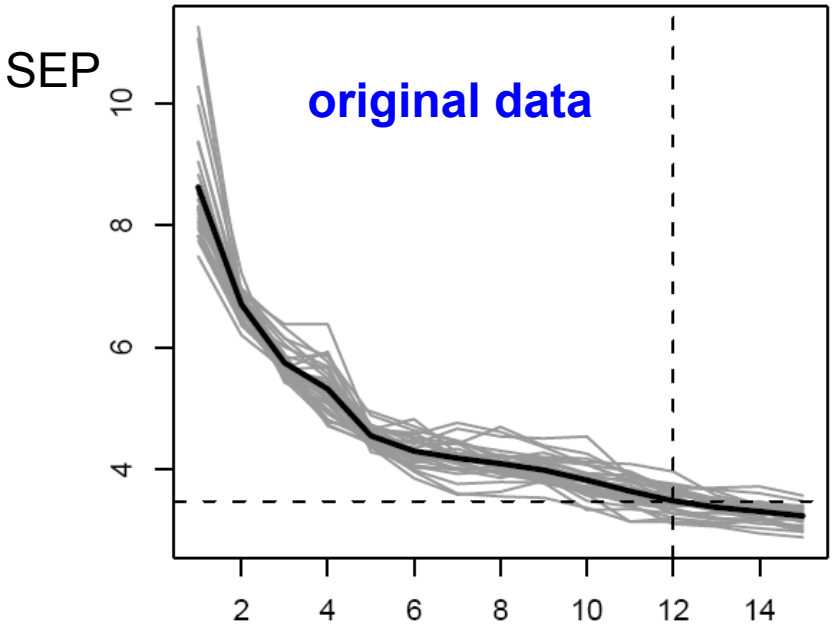
- not all objects are considered equally,
- not a fixed but a varying number of prediction errors per object,
- calibration set contains a varying number of identical copies of objects,
  (on the average 63% of the objects are in the calibration set;
  t. m., 37% are copies)
- optimization of model complexity (for the calibration set) requires another bootstrap (or CV) with an even increased number of copies in the training sets

Contents

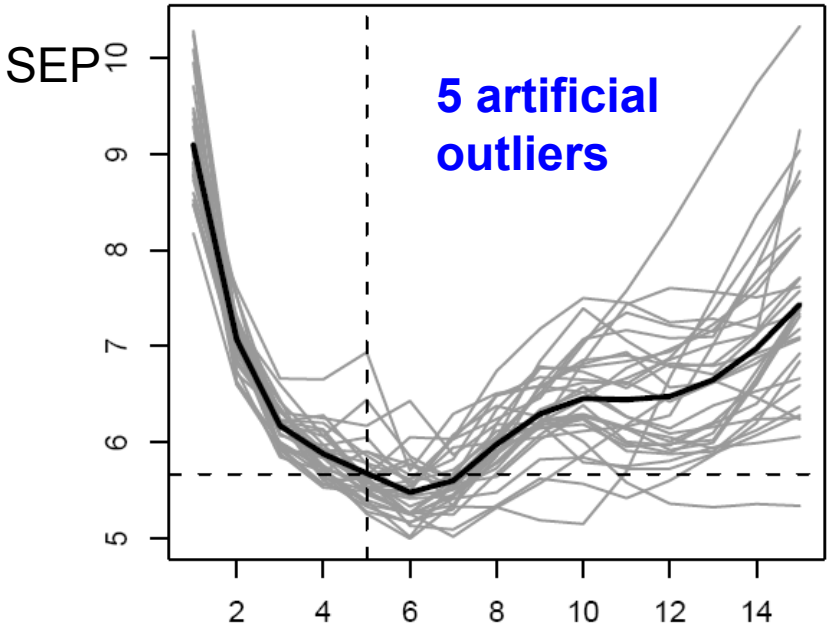
rdCV for calibration - example "concentration"

***X**: n* = 120 alcoholic **fermentation mashes**; *m* = 235 NIR absorptions (1st deriv.);
***y**:* glucose concentration (HPLC, 0.1 - 55 g/L)

rdCV: segments for test sets: 3; segments for CV within calibration sets: 5;
30 repetitions



no. of PLS components;
"final" value = 12



no. of PLS components;
"final" value = 5

Liebmann B., Friedl A., Varmuza K.: Anal. Chim. Acta, **642**, 171-178 (2009)

rdCV for classification - example "iris"

Data "iris"

n = 150; *m* = 4 (length measurements on blossom)

no. of classes = 3

Anderson E. (1935, collected); Fisher R.A. (1936)

rdCV

s_{TEST} = 4; s_{CALIB} = 6; no. of repetitions = 100

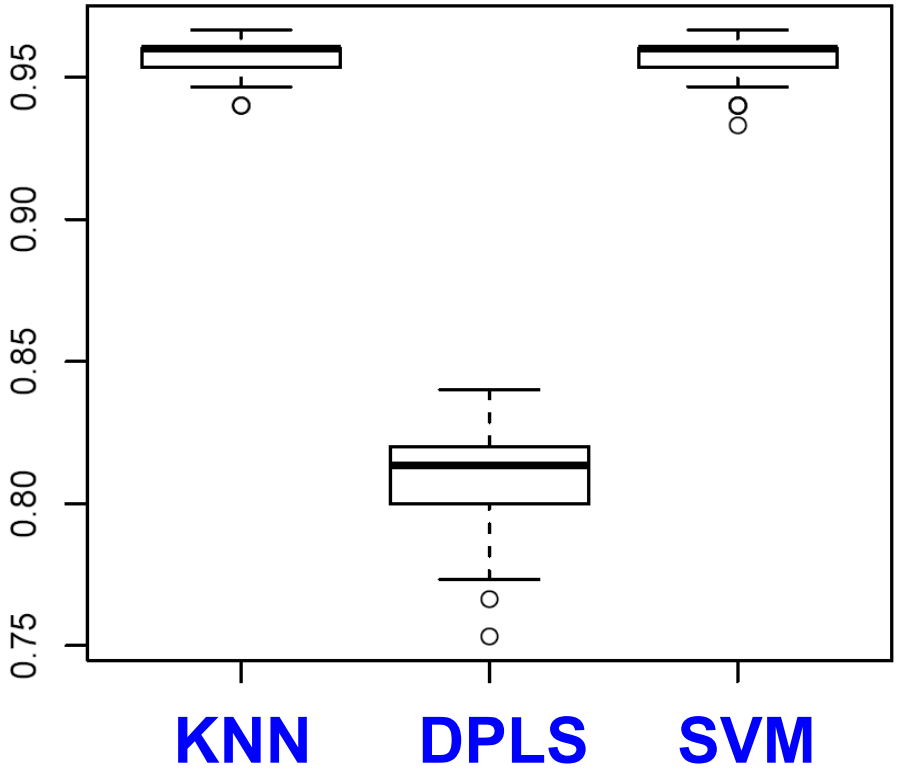
Comparison of KNN, DPLS, SVM

rdCV for classification - example "iris"

P

average predictive ability for 3 classes

100 repetitions



rdCV: $s_{TEST} = 4$; $s_{CALIB} = 6$

$k_{opt} = 1$, $A_{PLS} = 2$

DPLS not recommended for more than 2 classes

rdCV for classification - QSAR example "AMES"

Data "AMES"

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 = 1440 molecular descriptors; *Dragon* [3]

Classification method: rdCV with DPLS (1 - 30 components)

---

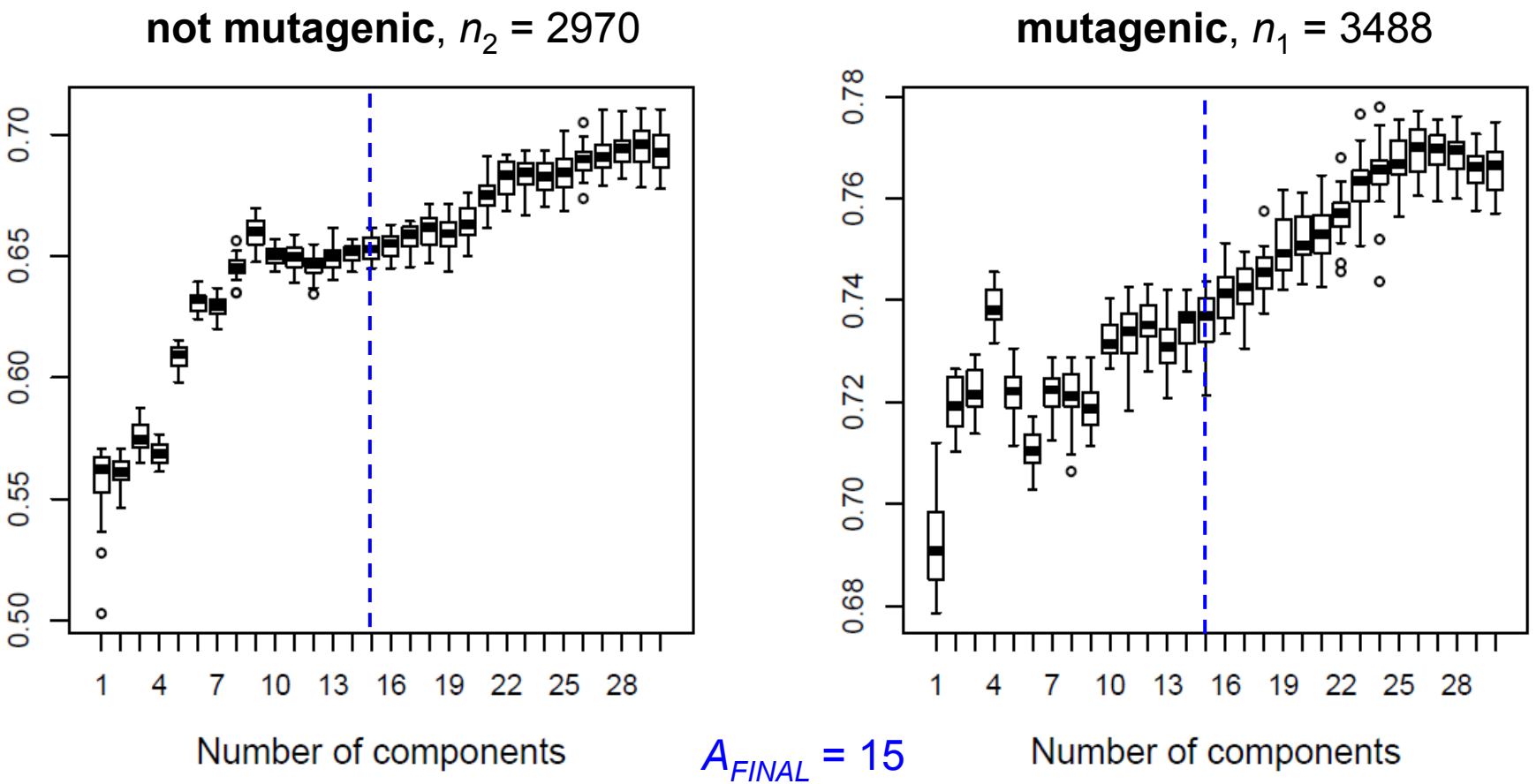
[1] K. Hansen, Technical University Berlin, Germany
http://ml.cs.tu-berlin.de/toxbenchmark/index.htm

[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).

rdCV for classification - QSAR example "AMES"

Distribution of predictive abilities for the 2 classes



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

Conclusions

A good theory (appropriate !)
is most practical.

However, in chemistry often
empirical models are necessary -
and (sometimes) useful.

A necessary (but not sufficient)
prerequisite for good empirical
models is a careful statistical
evaluation (this contribution).

However, a statistical evaluation
cannot fully avoid spurious
(accidental) correlation.

Model interpretation is desired.

Aim at parsimonious models.

Take time for validation

Consider variability

Accept variability and uncertainty

END.