Variable Selection for Multivariate Models

Myth and Reality

Kurt VARMUZA

Vienna University of Technology

Department of Statistics and Probability Theory

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

www.lcm.tuwien.ac.at/vk/ kvarmuza@email.tuwien.ac.at

Autumn School of Chemoinformatics, 27 - 28 November 2013, Nara, Japan [*28 Nov. 2013]*

Variable Selection
for Empirical Multivariate Models

Contents of Tutorial

Introduction

Performance measure for calibration (rdCV method)

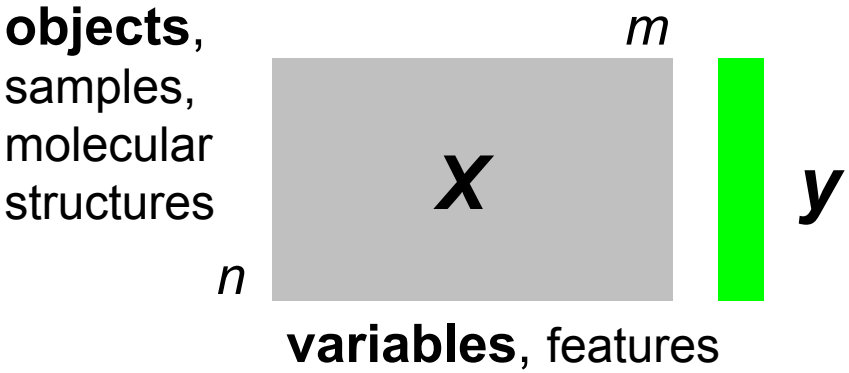
Variable selection strategy

Variable selection methods (basics, applications)

Summary

Introduction

Empirical mathematical models in chemistry



Multivariate Calibration

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

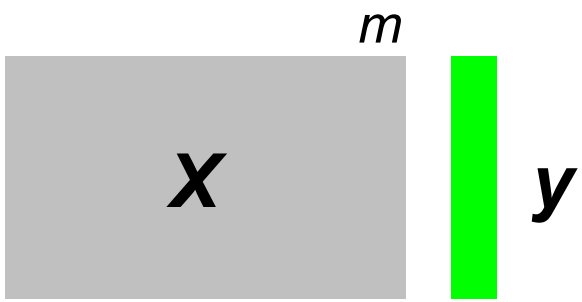
Model parameter — estimated (empirically, data driven),
e. g., by PLS regression, from a calibration set,
including optimization of model complexity

Model performance — carefully estimated for "new cases" (from a test set),
appropriate performance measure,
realistic estimation, including the variability,
requires an appropriate strategy

Introduction

Empirical mathematical models in chemistry

objects, samples, molecular structures



variables, features

desired **property**

Multivariate Calibration

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

Typical m = 10 ... several 1000 no. of variables
("all available" from instruments or software)

Variable selection Better model using only a subset of variables ($m_{SEL} << m$) ?!

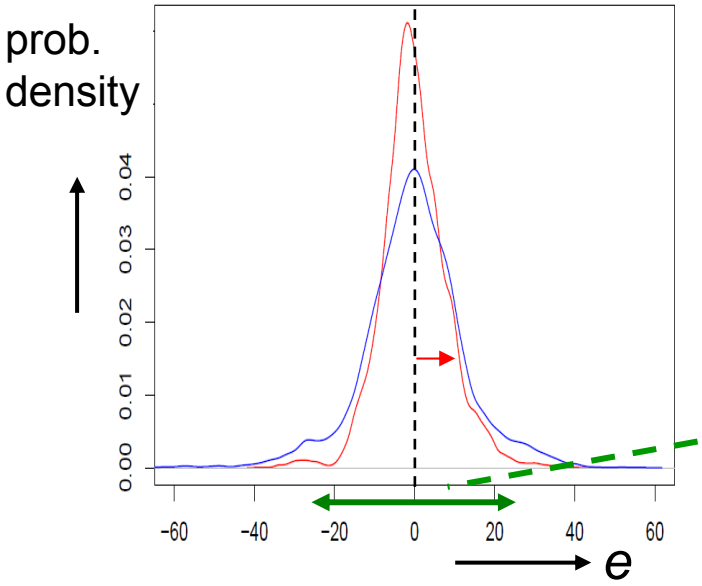
- Methods for variable selection
- Strategy for comparing model performances

Performance measure for 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 (residual)
$i = 1 \ldots z$	*z* is the number of predictions

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, $CI_{95\%} \approx \pm 2*SEP$

User friendly ! All in units of *y* !

Strategy for estimation of SEP

- **SEP has to be estimated from test set objects, not used for creation/optimization of the model**
- **SEP is not a single value but has a distribution, (estimation of the variability of SEP is essential for comparisons of SEP values)**
- **Optimum complexity of models (e. g., no. of PLS components) has to be estimated**

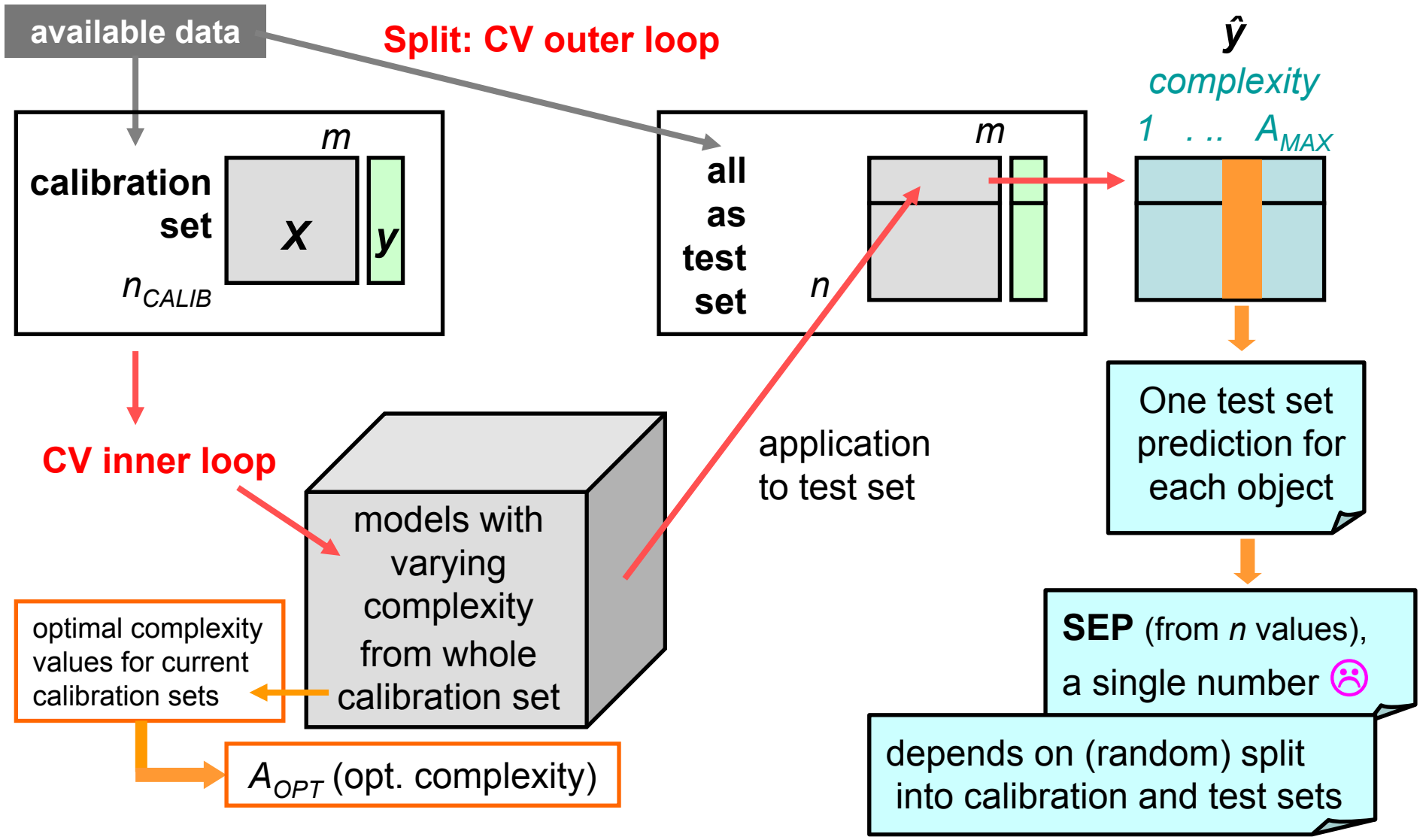
Our strategy is

repeated double Cross Validation (rdCV)

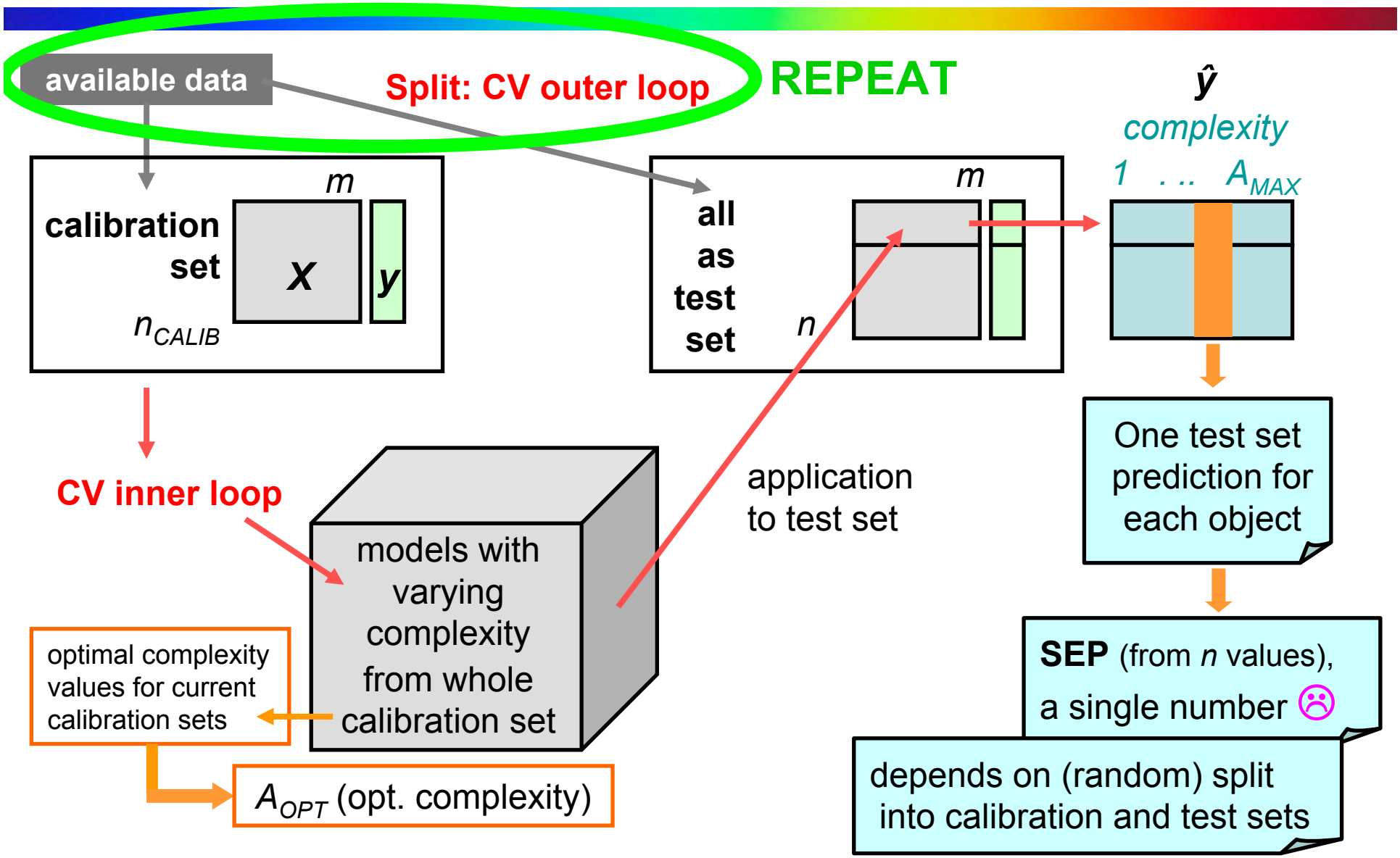
P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometrics 23 (2009) 160-171.
K. Varmuza et al., R software for chemometrics, www.lcm.tuwien.ac.at/R/

Other approaches of ***resampling***: *double bootstrap*, etc.

Strategy for estimation of SEP (rdCV)

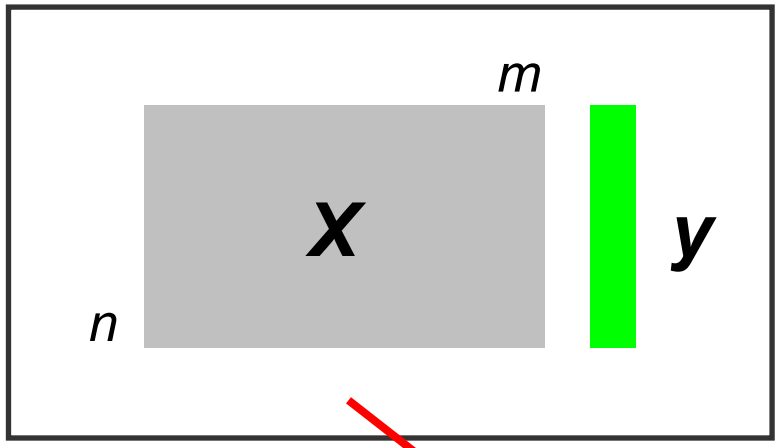


Strategy for estimation of SEP (rdCV)



Strategy for estimation of SEP (rdCV result)

available data

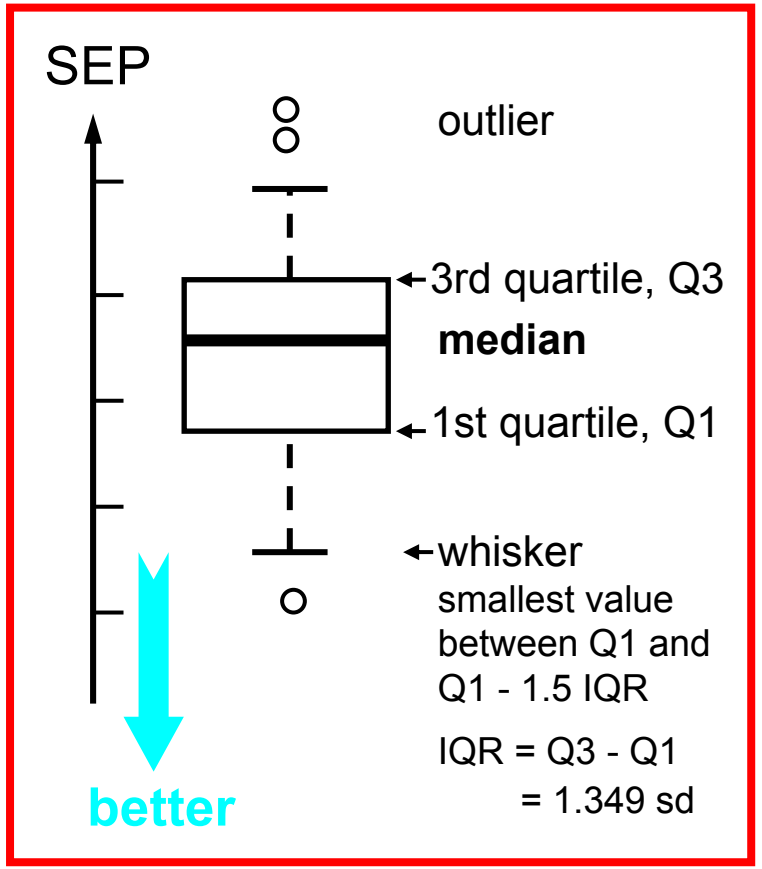


rdCV

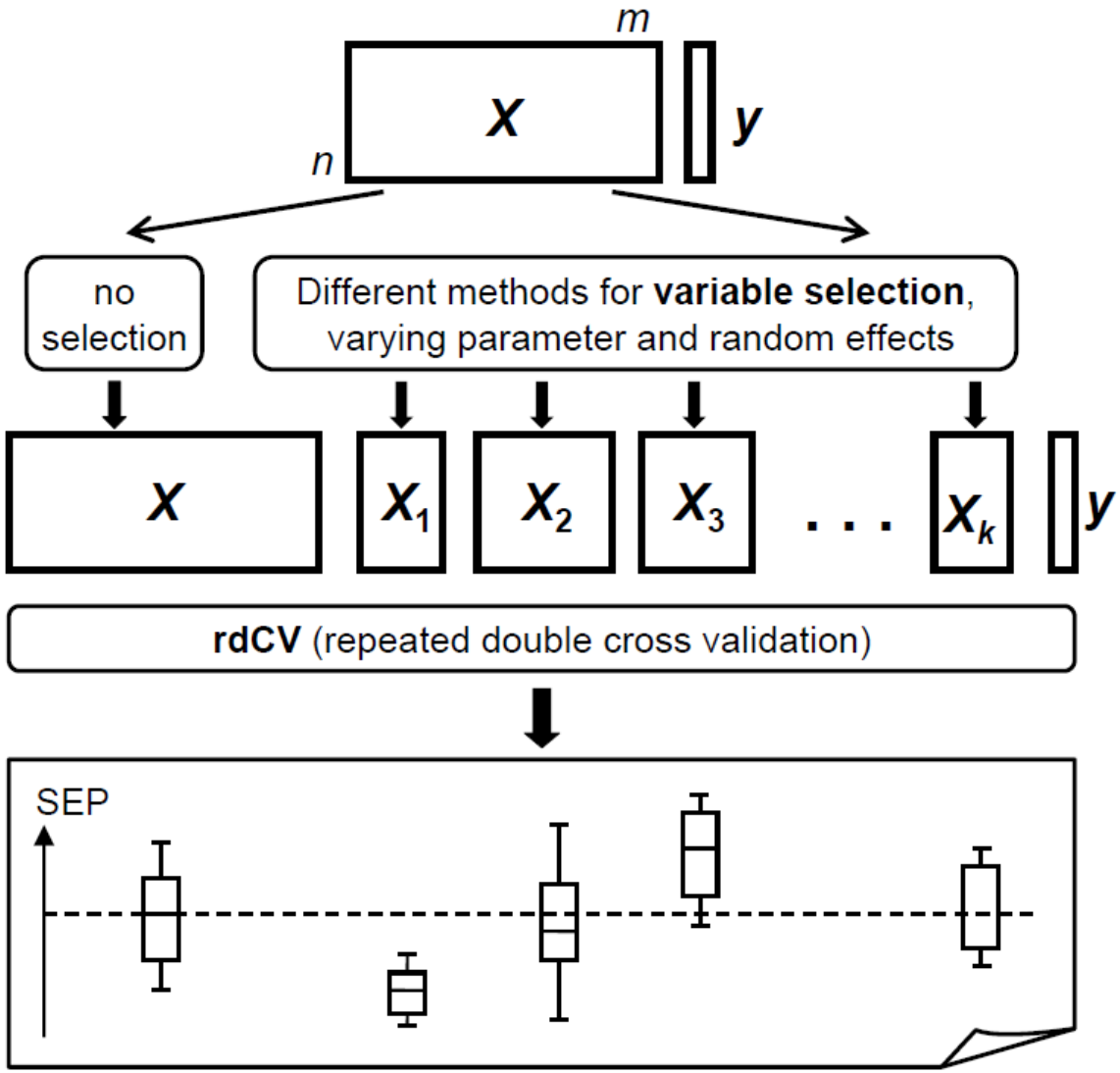
typ.: *rep* = 20 ... 100 repetitions

rep values for SEP

Characterization of the performance of models from a given data set ***X***, ***y*** by a boxplot for SEP



Variable selection / evaluation strategy



Variable selection / evaluation strategy

Performance criteria used during variable selection:

- *prediction performance for test set objects, using optimized models, would be desirable (rdCV-PLS);*

- *however, rarely applicable because too time consuming (e. g., GA may require to test several 100,000 variable sets).*

- *Most often, fit criteria are used (quick and dirty); e. g., SEP from leave-1-out, sum of squared prediction errors (within calibration set), BIC, AIC, Mallows Cp (NOT final estimations of model performance !).*

- *Therefore, evaluate suggested variable subsets carefully (goal-oriented = prediction of new cases), e. g., by rdCV (or double bootstrap).*

Variable Selection / why ?

PLS **OK** with

- $m > n$, $m >> n$
- highly correlating variables
- optimization of model complexity

- better model performance*
- interpretation of model parameter *easier*
- elimination of noise variables, more stable model parameter, *parsimony principle*

* Checked cautiously for test set objects (rdCV, bootstrap, ...).

* Not using performance measures obtained during variable selection as final model performance.

Variable Selection: *The (usually) impossible dream*

No "clever" strategy is possible to find the "best" subset of variables with guarantee.

"clever" **means not checking all possible sets**
"best" **means best test set prediction**

For *m* original ("all") variables, we have

$2^m - 1$ possible variable sets,

containing 1, 2, 3, ..., *m*-1, *m* variables
in all combinations.

?

Exhaustive ("not clever") search for best variable set:

Apply, e. g., rdCV-PLS to all **$2^m - 1$** variable sets !

Variable Selection: *The (usually) impossible dream*

m	$n\_subsets = 2^m - 1$
10	1 023
13	8 191
15	32 767
20	$1.0 \; 10^6$
35	$3.4 \; 10^{10}$
50	$1.1 \; 10^{15}$
100	$1.3 \; 10^{30}$
1000	$1.1 \; 10^{301}$

$m \leq 15$ exhaustive search (rdCV-PLS)

$m \leq 35$ best subset regression (fit !)

Data set (1) HEAT-ELE

n = 122	**biomass samples** from plants [1]
m = 13	**elemental composition** (C, H, N), and derived variables (cross products, logarithms, ratios)
y	higher **heating value**, HHV, calorimetric reference (± 60 kJ/kg) range 15 719 ... 25 948 kJ/kg

Model **heating value = f (element data)**

[1] A. Friedl, E. Padouvas, H. Rotter, K. Varmuza, Prediction of heating values of biomass fuel from elemental composition, *Anal. Chim. Acta* 544 (2005) 191-198.

Data set (2) GLU-NIR

n = 166	**fermentation samples** (cereals), centrifuged [2]
m = 232	**NIR** absorbances, 1115 - 2285 nm
y	**glucose** content, reference method HPLC range 0.32 ... 54.4 g/L

Model **glucose content = f (NIR absorbances)**

[2] B. Liebmann, A. Friedl, K. Varmuza, Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics, *Anal. Chim. Acta* 642 (2009) 171-178.

Data set (3) PAC-QSPR

n = 209	**polycyclic aromatic compounds** chemical structures: 3D and all H-atoms (*Corina*)
m = 2661	**molecular descriptors** (*Dragon 6.0*)
y	**GC retention index** [3, 4], range 197.0 ... 503.9

Model **retention index = f (molecular descriptors)**

[3] M.L. Lee, D.L. Vassilaros, C.M. White, M. Novotny, Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons, *Anal. Chem.* 51 (1979) 768-773.

[4] K. Varmuza, P. Filzmoser, M. Dehmer, Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. *Computational and Structural Biotechnology Journal* 5 (2013) e201302007. Open access: http://dx.doi.org/10.5936/csbj.201302007; data and software: **www.lcm.tuwien.ac.at/R/**

Variable Selection Methods

- **all subsets** exhaustive: all 2^m-1 variable sets, evaluated by rdCV-PLS

$m \leq 15$

- best subset regression
- highest correlation with *y*
- highly correlating *x*-variables
- stepwise selection
- replacement
- genetic algorithm
- many others

Variable Selection: All subsets + rdCV-PLS

Criterion: SEP (rdCV)

Data: **HEAT-ELE** (122 × 13)

Variables		Best subsets (SEP_{FINAL} from rdCV)					
No.	Name	1 st	2 nd	3 rd	4 th	5 th	100 th
1	C						
2	H	■	■	■	■	■	■
3	N	■	■	■	■	■	■
4	C/H						■
5	C*H	■	■	■	■	■	■
6	C*C						
7	H*H						
8	N*N						
9	ln (C)					■	■
10	ln (H)	■	■			■	
11	ln (N)						
12	ln (C/H)						
13	ln (C*H)	■		■		■	■
m_{SEL}		5	4	4	3	6	6
SEP_{FINAL}		352.12	352.13	352.35	352.36	352.90	360.85
a_{FINAL}		3	3	3	3	3	4

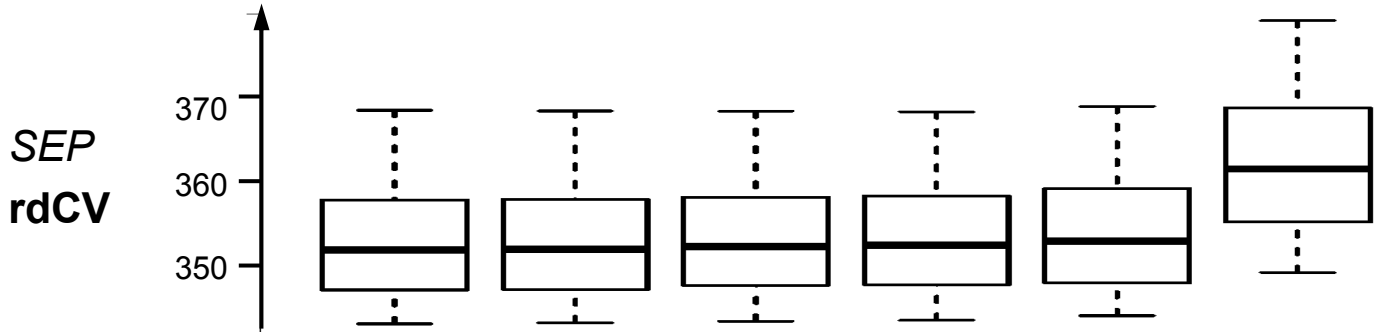
m = 13
no. of all subsets
= 2^m - 1
= 8191

↓

all subsets
tested by rdCV
(*rep* = 30, *amax* = 12
segments: 3 and 4)

↓

subsets ordered
by SEP_{FINAL}



all m = 13
variables:

SEP_{FINAL} = 417

Variable Selection: All subsets + rdCV-PLS

Criterion: SEP (rdCV)

Exhaustive and correctly evaluated.

Recommended method for
data sets with $m \leq 13$ (15) variables.

Variable Selection Methods

- all subsets
- **best subset regression**
- highest correlation with *y*
- highly correlating *x*-variables
- stepwise selection
- replacement
- genetic algorithm
- many others

variable selection optimizes fit

$m \leq 35$

Variable Selection: **Best subset regression**

Criterion: **BIC, fit criterion**

Leaps and bounds algorithm

allows an economic, complete search for subsets with a given number of variables (m_{SEL}) exhibiting best fit (BIC, etc).

Limit: ca $m \leq 35$ (although > 10^{10} subsets)

Best fit does not necessarily provide best prediction!

Therefore: **Resulting 'potentially good' subsets have to be tested for prediction performance**;

e. g., by rdCV (estimating SEP and its variability).

small BIC is fine,
value of BIC is meaningless,
just for comparisons

BIC = $n \log(RSS/n) + m \log(n)$ **B**ayes **I**nformation **C**riterion

RSS = sum of the squared residuals; log with base *e*

Variable Selection: Best subset regression

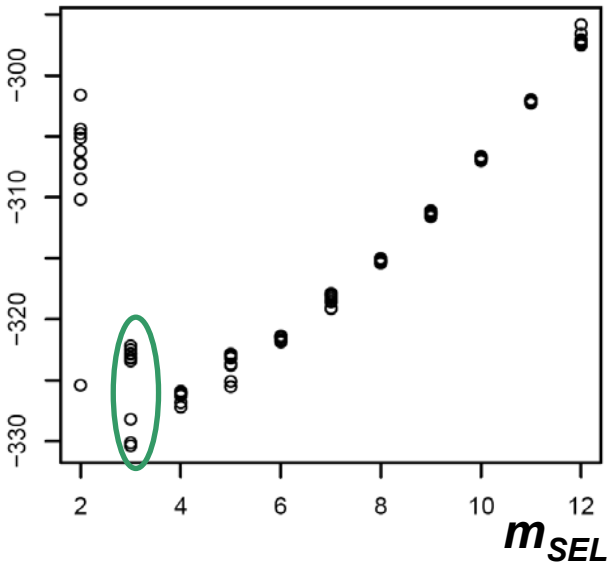
Criterion: BIC, fit criterion [R-library *leaps; regsubsets()*]

Data: **HEAT-ELE** (122 × 13)

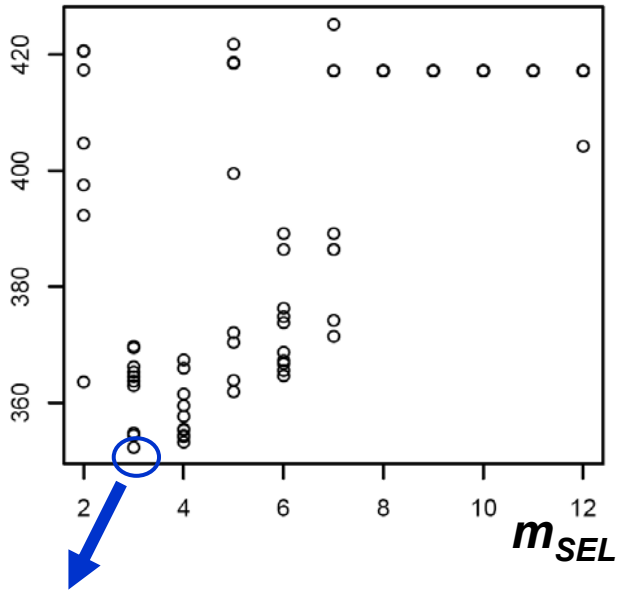
Complete search for the 10 best subsets (minimum BIC); m_{SEL} = **2 ... 12 variables**.

Estimation of prediction performance (SEP_{FINAL}) by rdCV-PLS (**110 subsets**).

BIC



SEP_{FINAL} (rdCV)



Final subset from ***best subset regression + rdCV-PLS***: m_{SEL} = 3, (H, N, C*H)

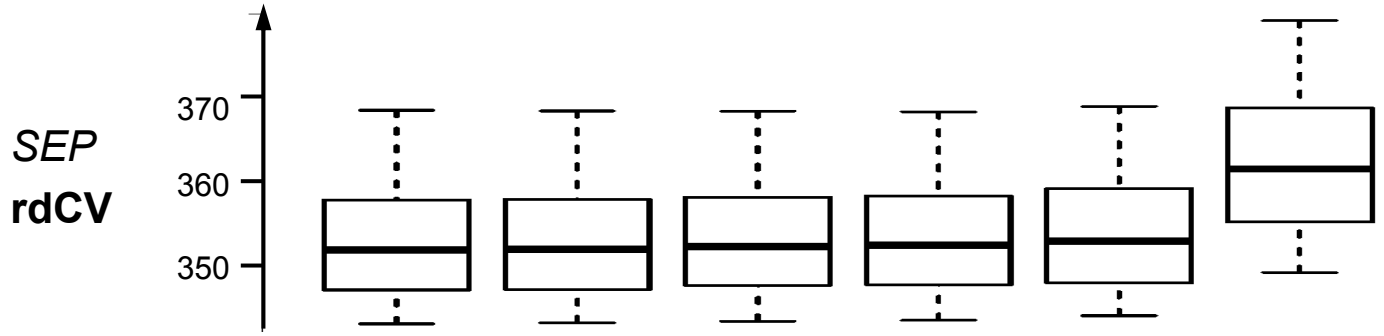
Variable Selection: All subsets + rdCV

Criterion: SEP (rdCV)

Data: **HEAT-ELE** (122 × 13)

Variables		Best subsets (SEP_{FINAL} from rdCV)					
No.	Name	1 st	2 nd	3 rd	4 th	5 th	100 th
1	C						
2	H	■	■	■	■	■	■
3	N	■	■	■	■	■	
4	C/H						
5	C*H	■	■	■	■	■	■
6	C*C						
7	H*H						
8	N*N						
9	ln (C)					■	
10	ln (H)	■	■			■	
11	ln (N)						■
12	ln (C/H)						
13	ln (C*H)	■		■		■	
m_{SEL}		5	4	4	3	6	3
SEP_{FINAL}		352.12	352.13	352.35	352.36	352.90	354.47
a_{FINAL}		3	3	3	3	3	3

Best subset regression



Variable Selection: **Best subset regression**

Criterion: **BIC, fit criterion** [R-library *leaps; regsubsets()*]

Try for datasets with $m \leq 35$ variables.

Consider several suggested variable subsets and test for prediction performance.

May give near optimum results.

Variable Selection Methods

- all subsets
- best subset regression
- **highest correlation with *y***
- **highly correlating *x*-variables**
- stepwise selection
- replacement
- genetic algorithm
- many others

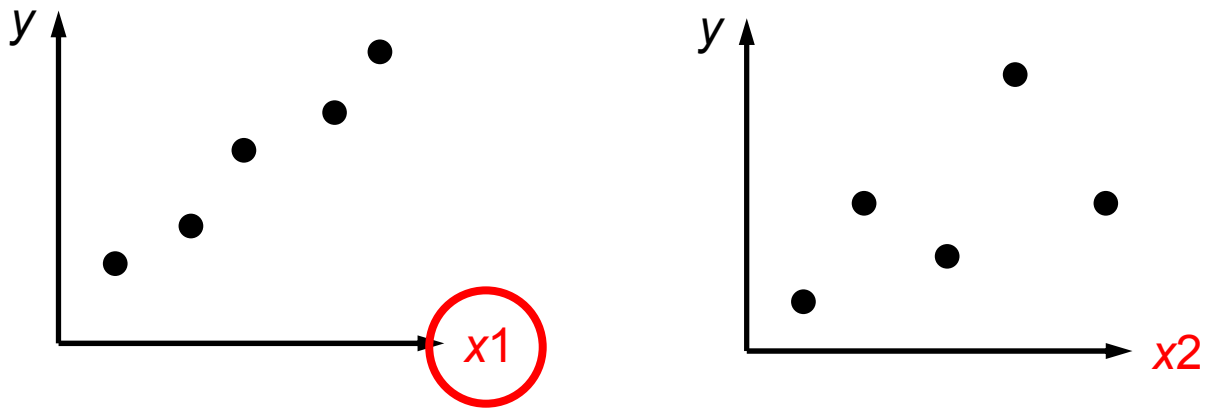
univariate,
fast, simple, evident,
for many variables

Variable Selection: corr xy

Select variables possessing highest correlation with y

Criterion: Squared (Pearson) correlation coefficient [cor(x_j, y)]^2

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



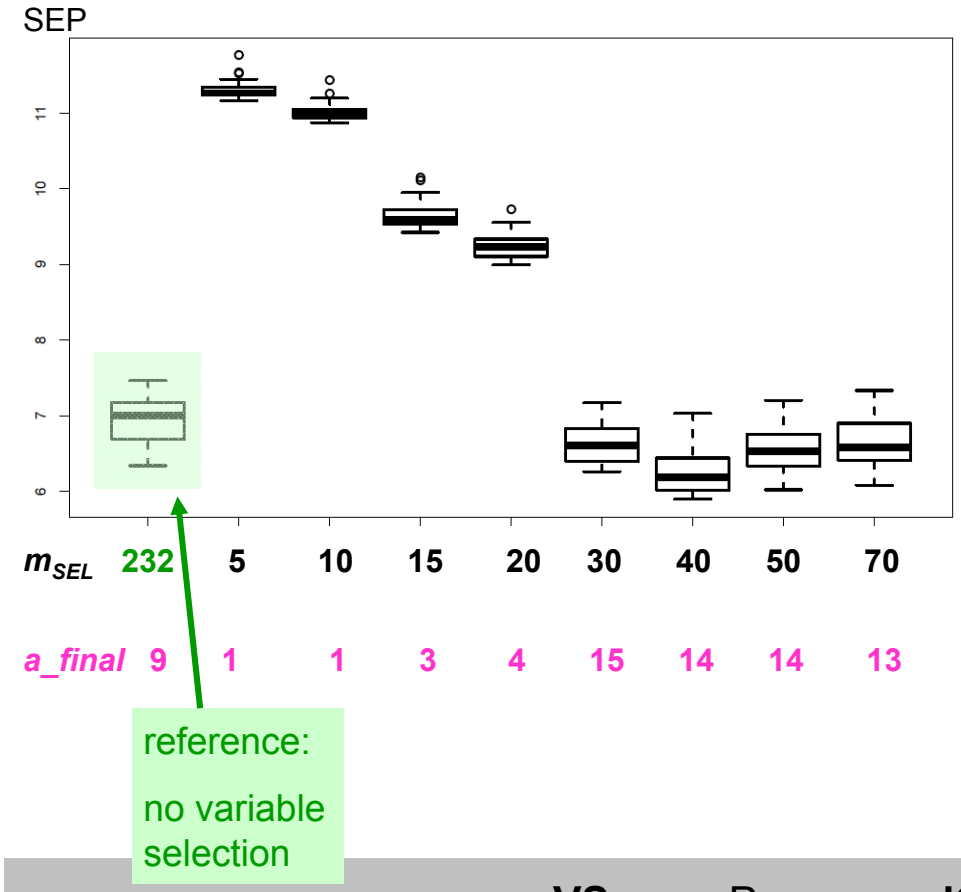
- Select a subset of variables (m_{SEL}) with maximum
  squared Pearson correlation coefficient (etc.), $R^2(x_j, y)$
- Delete variables with very low
  squared Pearson correlation coefficient (etc.), $R^2(x_j, y) <$ **threshold**

Variable Selection: corr xy

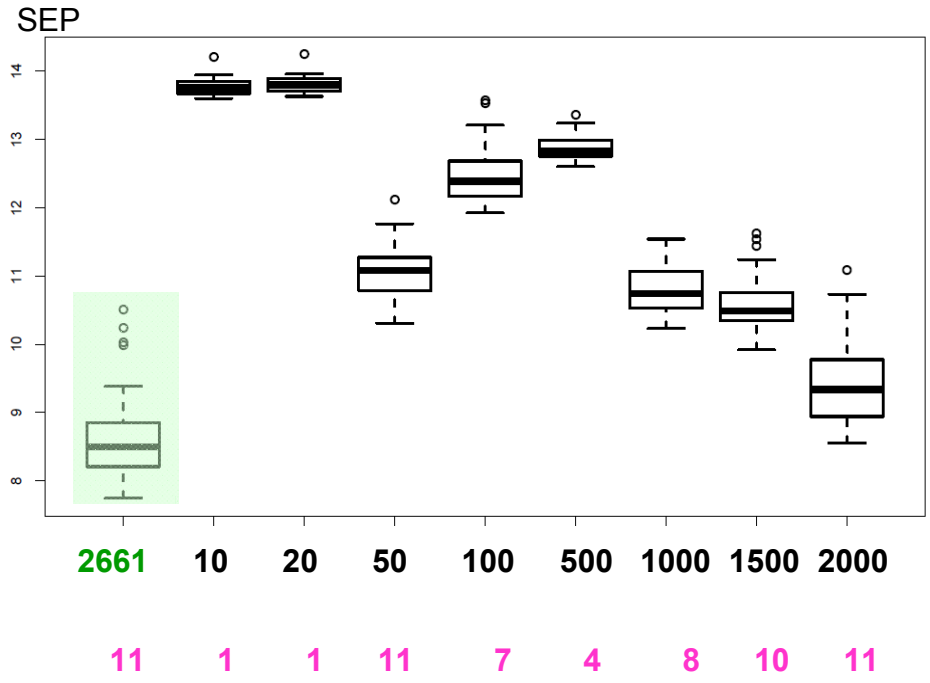
Select variables possessing highest correlation with y

Criterion: Squared (Pearson) correlation coefficient [cor(x_j, y)]^2

Data: **GLU-NIR** (166 × 232)



Data: **PAC-QSPR** (209 × 2661)



VS: squ. Pearson; **rdCV:** rep=30, amax=20

Variable Selection: corr xy

Select variables possessing highest correlation with y

Criterion: Squared (Pearson) correlation coefficient $[\text{cor}(x_j, y)]^2$

No or no significant improvement of prediction performance.

Success depends on data structure.

Variable Selection: corr xx

Eliminate variables possessing very high correlation with another variable

Criterion: Squared (Pearson) correlation coefficient [cor(x_j, x_k)]^2

Eliminate variables that are

- identical to another variable,
- have a large **squared Pearson correlation coefficient** (etc.), $R^2(x_j, x_k)$ to another variable (*), threshold e. g., 0.99, 0.95, ...

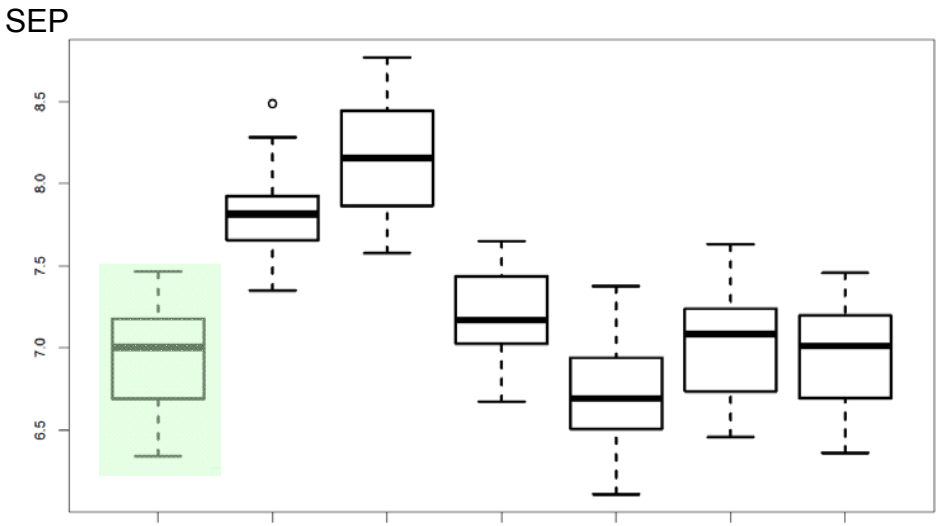
(*) Eliminate the variable with the larger sum of R^2 to all other variables.

Variable Selection: corr xx

Eliminate variables possessing very high correlation with another variable

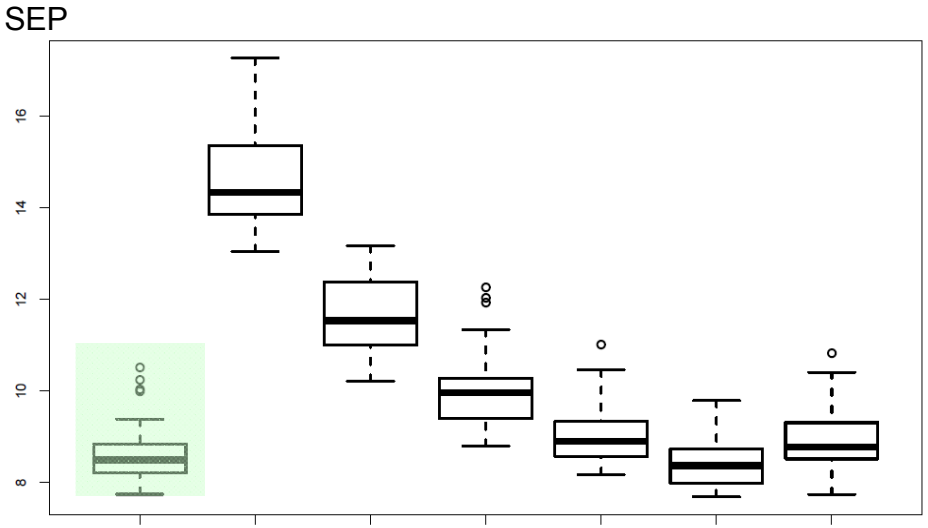
Criterion: Squared (Pearson) correlation coefficient [cor(x_j, x_k)]^2

Data: **GLU-NIR** (166 × 232)



m_{SEL}	232	15	20	42	73	173	222
R^2_{LIMIT}		0.7	0.8	0.9	0.95	0.99	0.999
a_final	9	10	10	12	12	9	9

Data: **PAC-QSPR** (209 × 2661)



m_{SEL}	2661	281	432	685	932	1474	2046
R^2_{LIMIT}		0.7	0.8	0.9	0.95	0.99	0.999
a_final	11	14	10	12	13	12	12

VS: squ. Pearson; **rdCV:** rep=30, amax=20

Variable Selection: corr xx

Eliminate variables possessing very high correlation with another variable

Criterion: Squared (Pearson) correlation coefficient $[\text{cor}(x_j, x_k)]^2$

No improvement of prediction performance.

Elimination of variables with $R^2(x_j, x_k) > 0.99$ does not decrease the prediction performance.

Success depends on data structure.

Variable Selection Methods

- all subsets
- best subset regression
- highest correlation with *y*
- highly correlating *x*-variables
- **stepwise selection**
- replacement
- genetic algorithm
- many others

- start with one variable (or with all variables),
- add/delete single variables ("step"),
- standard method,
- best fit,
- applicable for many variables

Variable Selection: stepwise

Criterion: BIC

"Forward" stepwise selection: starts with an "empty model" (*y* is explained only by the intercept), and adds in each step one variable (the "best") until no further improvement is possible.

"Backward" stepwise selection: starts with the "full model" (all x-variables; not possible if e.g., $m > n$), and removes in each step one variable (the "worst") until no further improvement is possible.

"Both" directions stepwise selection: adding or removing one variable at a time, starting either from the empty (!) or from the full model until no further improvement is possible.

BIC can be used instead of the more traditional F-test

Variable Selection: stepwise

Criterion: BIC

New R function: varsel_stepwise_BIC()

Peter Filzmoser, TU Vienna

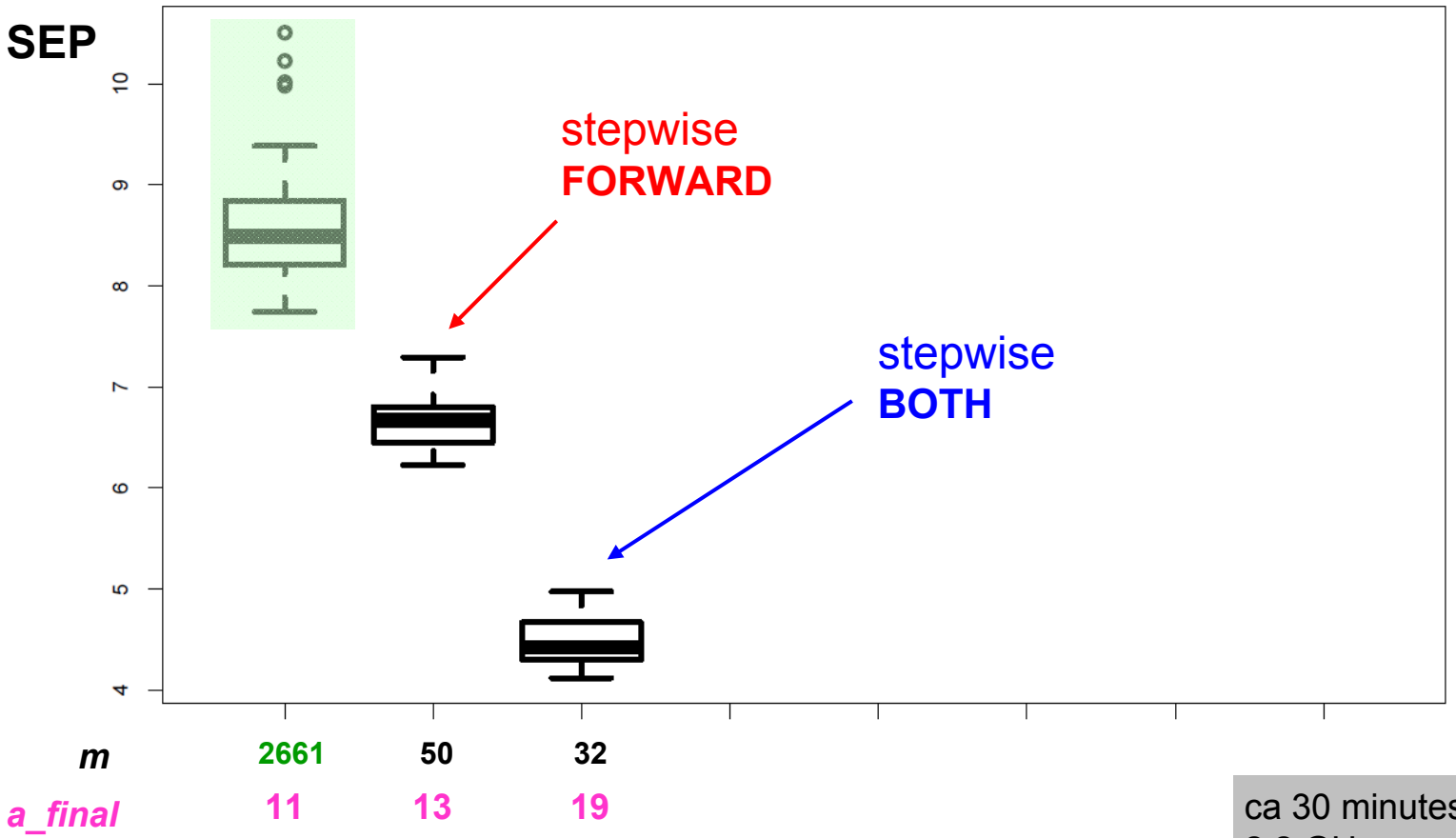
- "Forward" or "Both" strategy.
- Stop until no more improvement (BIC), or until a certain number of steps, or a pre-defined computing time is reached.
- Much faster (for e. g., $m > 100$) than the long existing R function **step()**.
- Still time consuming for $m > 1000$.

Free download: www.lcm.tuwien.ac.at/R/

Variable Selection: stepwise

Criterion: BIC

Data: **PAC-QSPR** (209 × 2661)



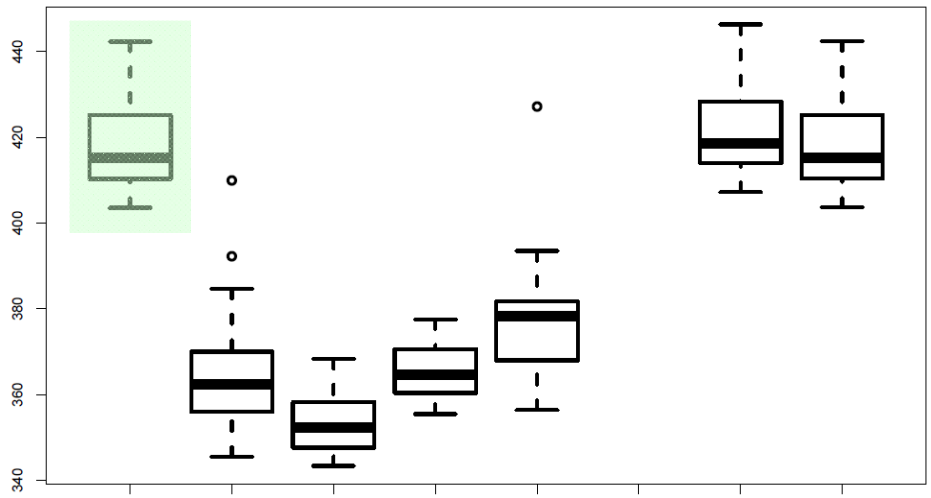
ca 30 minutes
2.2 GHz

VS: BIC, max 50 steps; **rdCV:** rep=30, amax=20

Variable Selection: stepwise

Criterion: BIC

Data: **HEAT-ELE** (122 × 13)



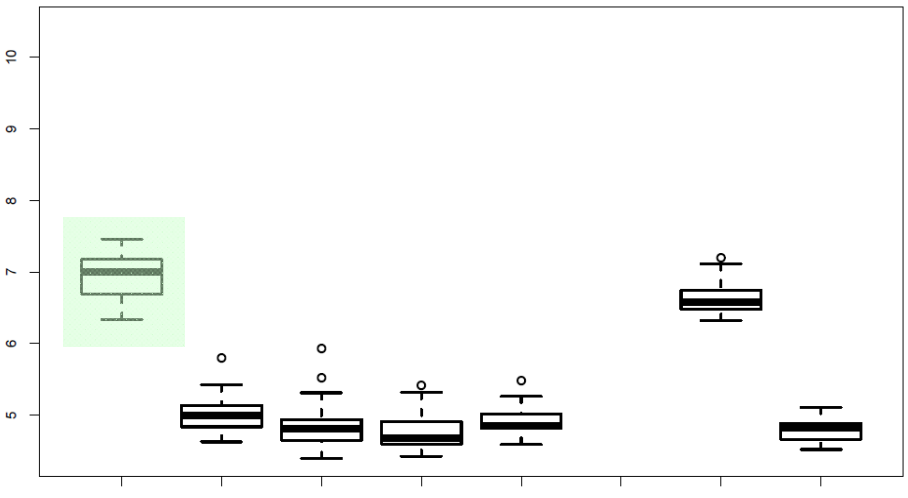
m	13	5	3	2	3	2	4	
a_final	1	4	3	2	3	1	1	

stepwise **FWD | BOTH**

corr xy **(>0.1 | >0.2)** stepwise **BOTH**

corr xx **(<0.95 | <0.99)** stepwise **BOTH**

Data: **GLU-NIR** (166 × 232)



m	232	16	16	15	15	10	15	
a_final	9	13	14	9	13	7	14	

stepwise **FWD | BOTH**

corr xy **(>0.1 | >0.2)** stepwise **BOTH**

corr xx **(<0.95 | <0.99)** stepwise **BOTH**

VS: BIC, squ. Pearson (x, y) or (x, x), max 50 steps; **rdCV:** rep=30, amax=12 or 20

Variable Selection: stepwise

Criterion: BIC

Significant improvement of prediction performance.

Strategy *BOTH* (starting *FORWARD*) is often best.

Computing time	**ca 20 s**	**for $m = 200$**
	ca 20 min	**for $m = 2000$**

Variable Selection Methods

- all subsets
- best subset regression
- highest correlation with *y*
- highly correlating *x*-variables
- stepwise selection
- **replacement**
- genetic algorithm
- many others

- **sequential replacement**
- start with randomly selected m_{SEL} variables
- replace single variables ("step"),
- rarely used,
- best fit

Variable Selection: Sequential replacement

Criterion (fitness): BIC, fit criterion [R-library *leaps; regsubsets()*]

Start with a random set of m_{SEL} variables.

Replace each variable with one of the not selected variables - and keep a new variable if it improves the model performance.

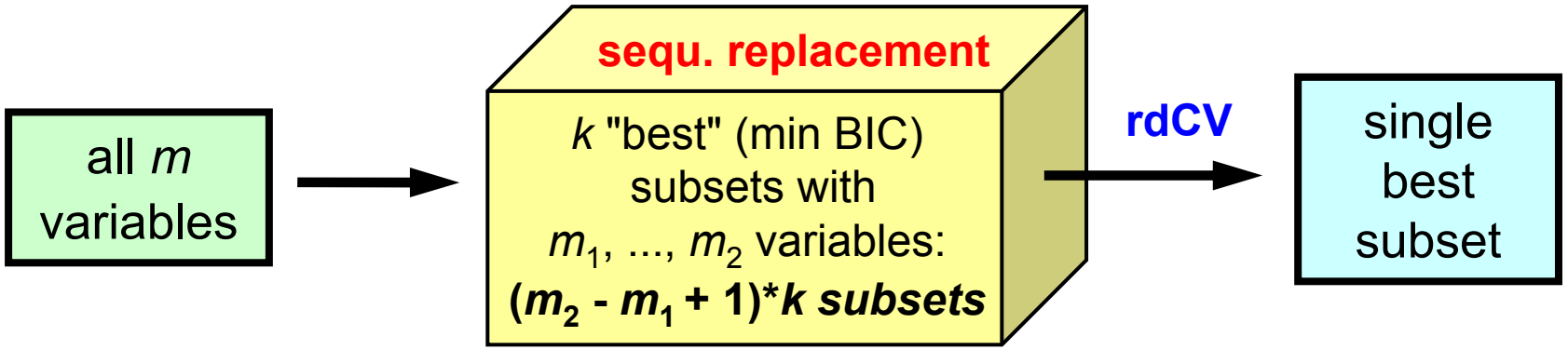
Repeat this procedure until the variable subset is stable or a terminating criterion is reached.

Miller A.: Subset Selection in Regression, 2nd ed., Chapman & Hall - CRC (2002). [1st ed. 1990]

Mercader A.C., Castro E.A.: in Statistical Modelling of Molecular Descriptors in QSAR/QSPR, p. 149, ed. Dehmer M., Varmuza K., Bonchev D., Wiley-VCH, Weinheim (2012).

Todeschini R., et. al.: Reshaped sequential replacement for variable selection (2013).

Our strategy



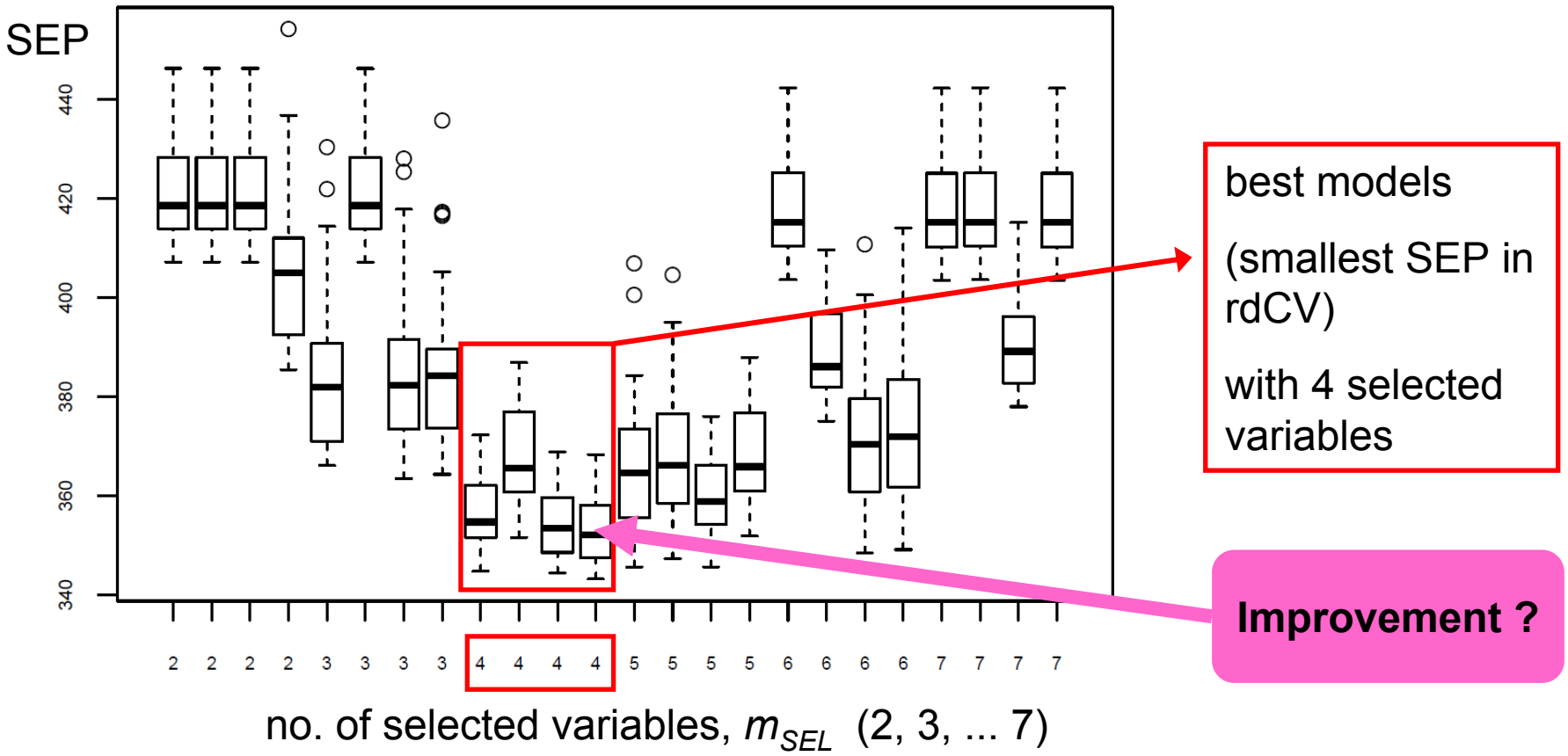
Variable Selection: Sequential replacement

Criterion (fitness): BIC, fit criterion [R-library *leaps; regsubsets()*]

Data: **HEAT-ELE** (122 × 13)

(1) Search **best 4** subsets containing **2 ... 7 variables**: (7-2+1)*4 = **24 subsets**

(2) Evaluate these 24 subsets by rdCV



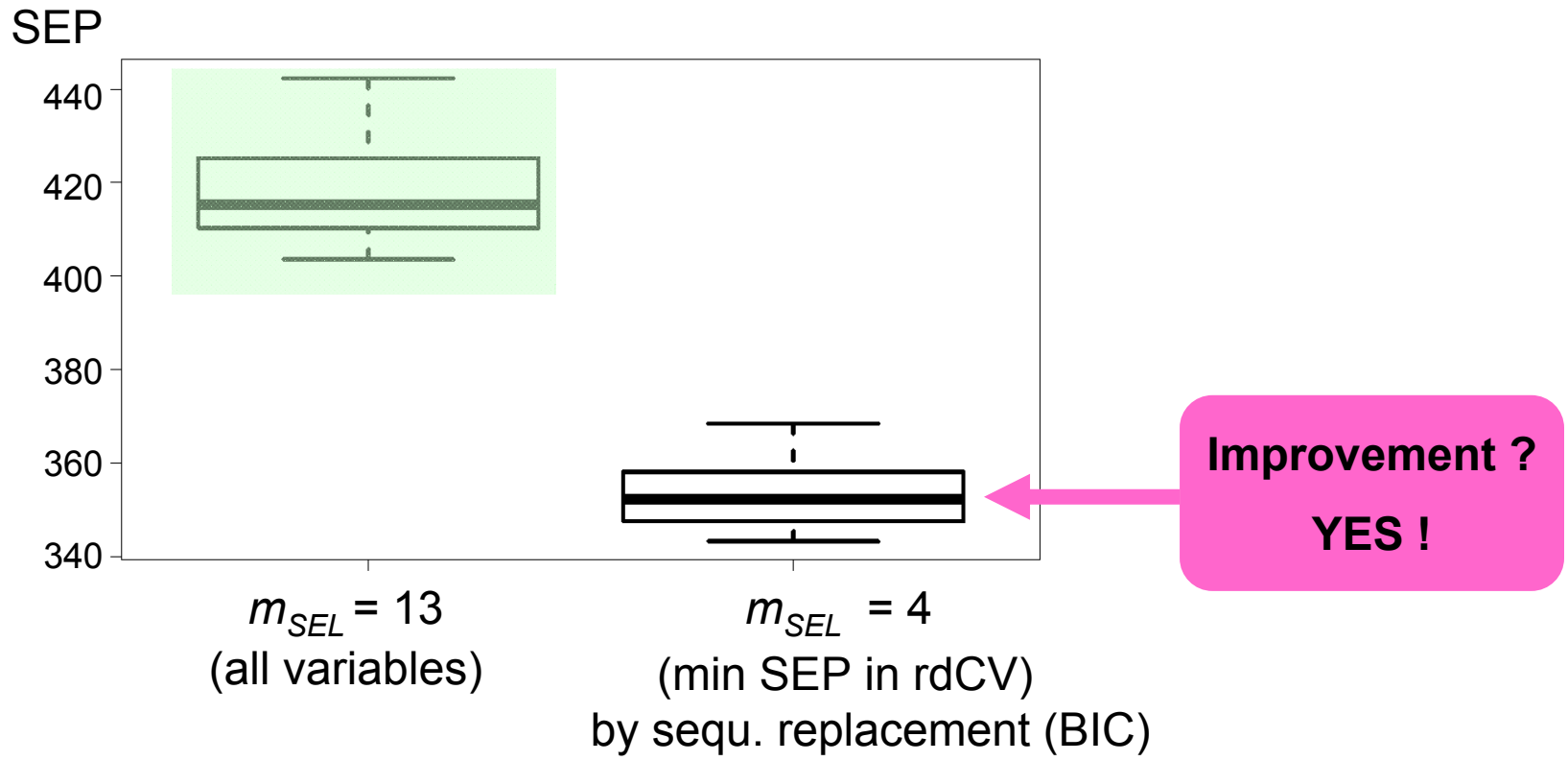
Variable Selection: Sequential replacement

Criterion (fitness): BIC, fit criterion [R-library *leaps; regsubsets()*]

Data: **HEAT-ELE** (122 × 13)

(1) Search **best 4** subsets containing **2 ... 7 variables**: (7-2+1)*4 = **24 subsets**

(2) Evaluate these 24 subsets by rdCV



Variable Selection: All subsets + rdCV

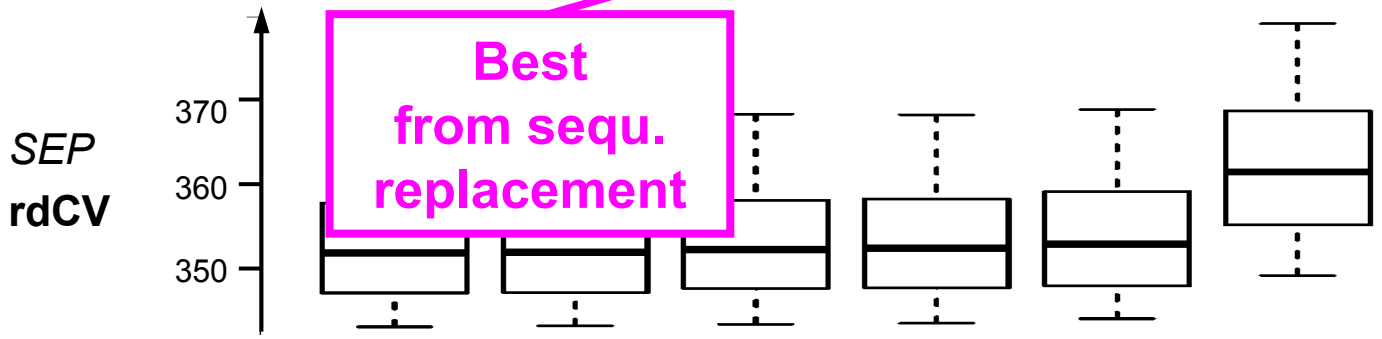
Criterion: SEP (rdCV)

Data: **HEAT-ELE** (122 × 13)

Variables		Best subsets (SEP_{FINAL} from rdCV)					
No.	Name	1 st	2 nd	3 rd	4 th	5 th	**20 th**
1	C						
2	H	■	■	■	■	■	■
3	N	■	■	■	■	■	
4	C/H						
5	C*H	■	■	■	■	■	■
6	C*C						
7	H*H						
8	N*N						
9	ln (C)					■	
10	ln (H)	■	■			■	
11	ln (N)						■
12	ln (C/H)						
13	ln (C*H)	■		■		■	
m_{SEL}		**5**	**4**	**4**	**3**	**6**	**3**
SEP_{FINAL}		**352.12**	**352.13**	**352.35**	**352.36**	**352.90**	**354.47**
a_{FINAL}		3	3	3	3	3	**3**

Best subset regression

Best from sequ. replacement



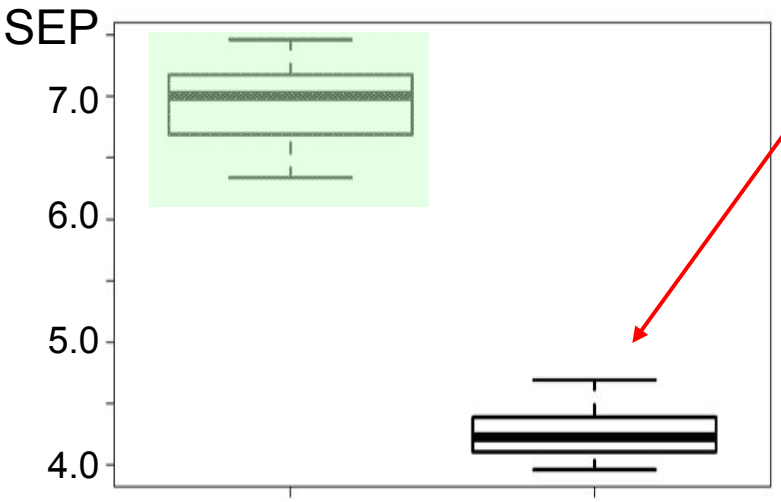
Variable Selection: Sequential replacement

Criterion (fitness): BIC, fit criterion [R-library *leaps; regsubsets()*]

Data: **GLU-NIR** (166 × 232)

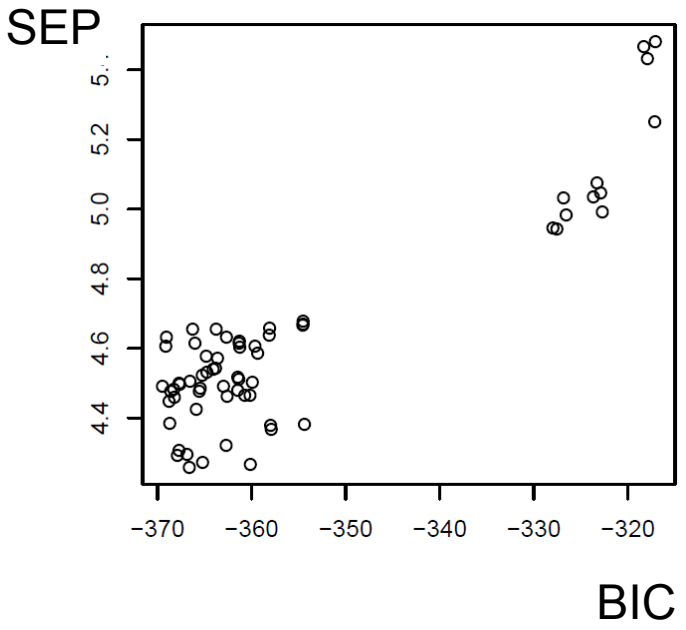
(1) Search **best 4** subsets containing **10 ... 25 variables**: (25-10+1)*4 = **64 subsets**

(2) Evaluate these 64 subsets by rdCV; select best (SEP)



m_{SEL} = 232
(all variables)

m_{SEL} = 20
(min SEP in rdCV)
variables selected by
sequ. replacement (BIC)



Variable Selection: Sequential replacement

Criterion (fitness): BIC, fit criterion [R-library *leaps; regsubsets()*]

Significant improvement of prediction performance.

Time consuming.

Rather not appropriate for *m* >1000.

Variable Selection Methods

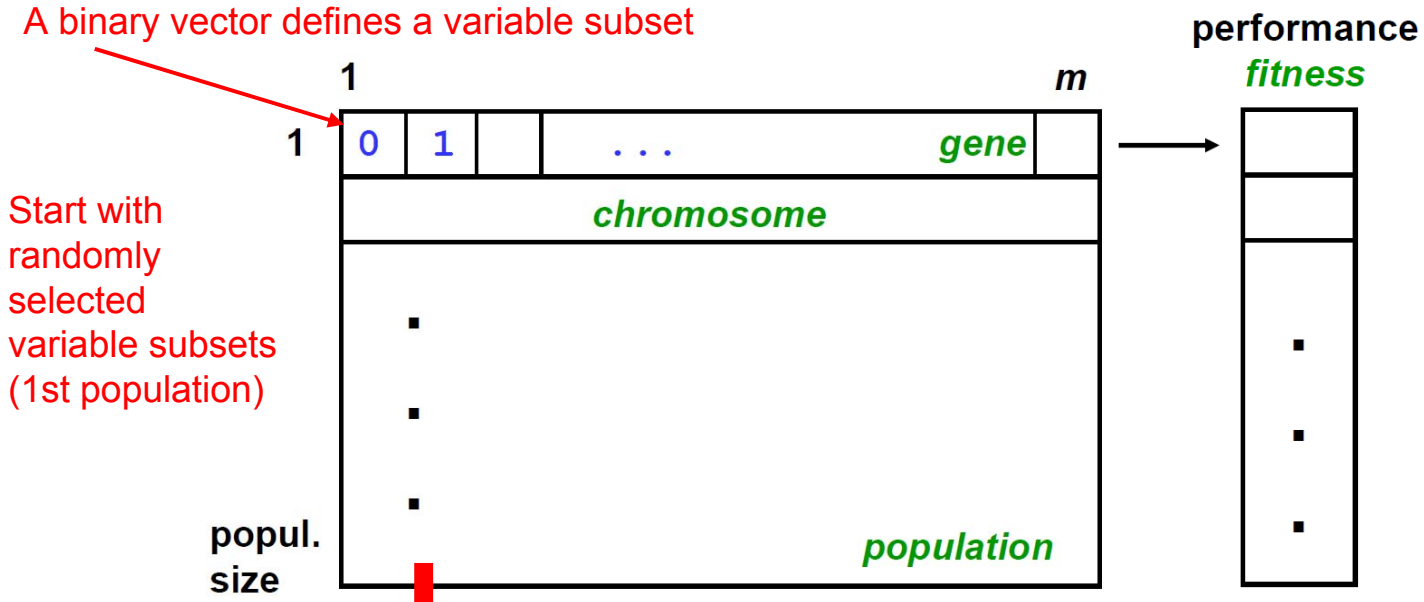
- all subsets
- best subset regression
- highest correlation with *y*
- highly correlating *x*-variables
- stepwise selection
- replacement
- **genetic algorithm**
- many others

- a *nature-inspired* method (terminology),
- search for near-optimum solutions (trying to avoid local optima, some randomness),
- time-consuming,
- rather not for many variables

Variable Selection: Genetic Algorithm (GA)

Criterion (fitness): $R^2(y, \hat{y})$ - adjusted, ...

simplified overview



Modify the population by GA strategies (*evolution, genetics,* ...) with the final aim to improve the fitness of the best variable subsets:

- eliminate worst chromosomes (*fittest survive*);
- combine best chromosomes to a new one (e. g., *cross-over*)
- random modifications of genes (*mutations*)
- consider defined restrictions, ...

Estimate fitness for new variable subsets (new population, a *generation*), until no improvement appears or a termination criterion is reached.

Best variable subset(s) [chromosomes] are the solution for variable selection.

Variable Selection: Genetic Algorithm (GA)

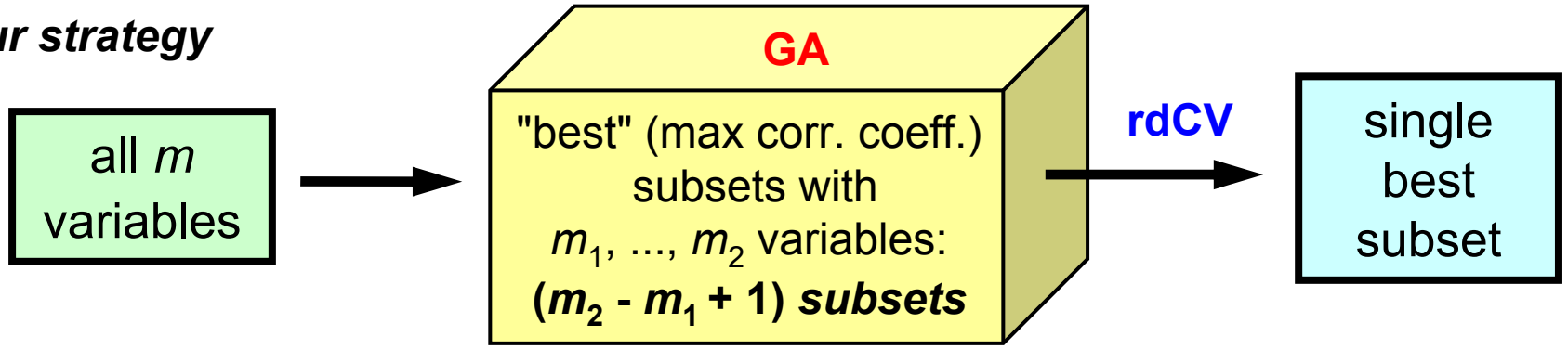
Criterion (fitness): $R^2(y, \hat{y})$, **Pearson (!)** **[R-library *subselect; genetic()*]**

Population size: $2m$ (min 100, max 500)
Stopping rule: no. of generations (typ. 500)

Result of GA: Variable subsets with highest R^2 (Pearson) containing m_{SEL} variables ($m_{SEL} = m_1 \ldots m_2$)

Careful test of the suggested variable subsets necessary (rdCV).

Our strategy

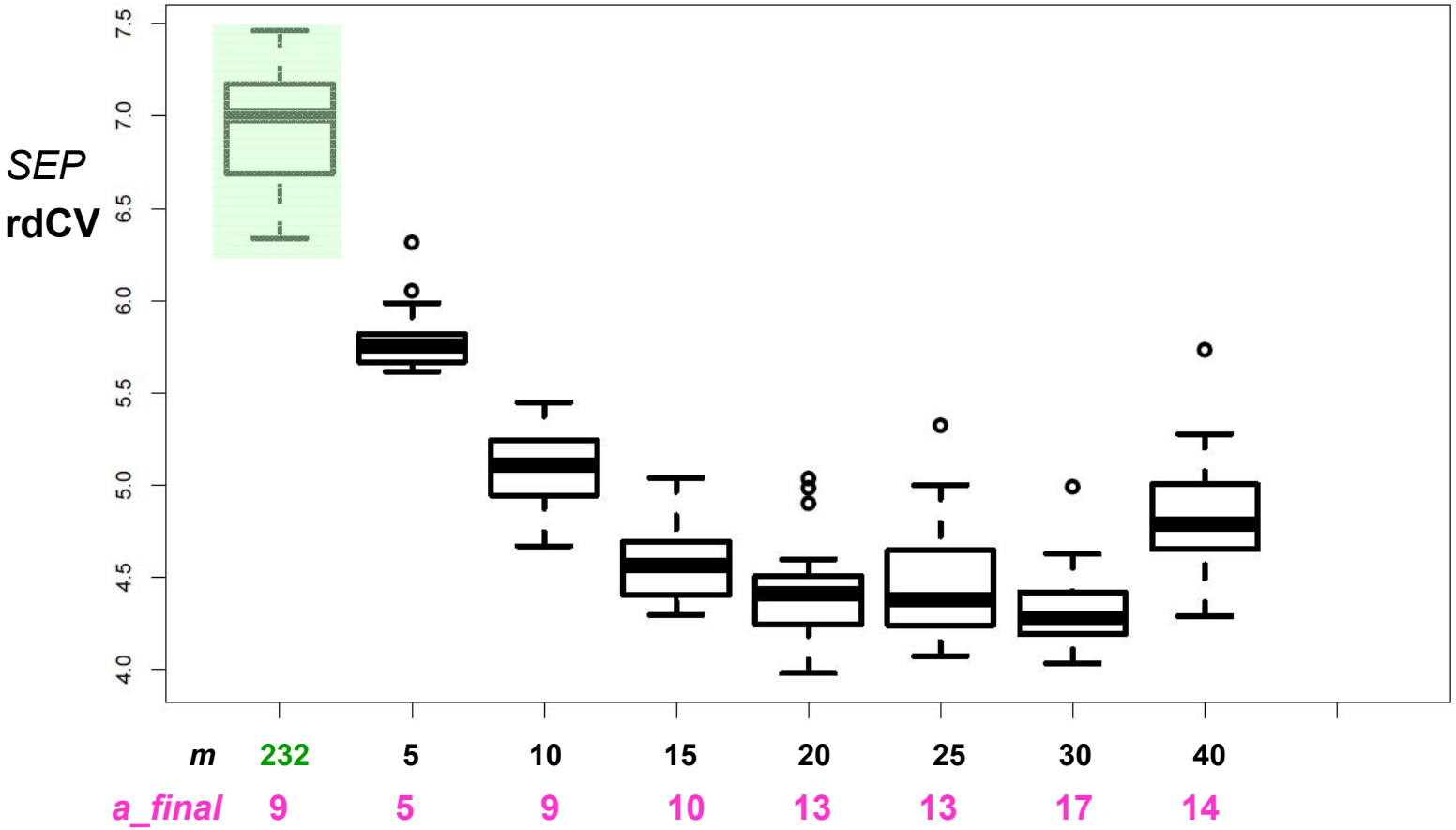


Cadima, J. et al.: Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47, 225-236 (2004).

Variable Selection: Genetic Algorithm (GA)

Criterion (fitness): $R^2(y, \hat{y})$, **Pearson (!)** **[R-library *subselect; genetic()*]**

Data: **GLU-NIR** (166 × 232)



VS: GA, gen=300, mutat.prob.=0.01; **rdCV:** rep=30, amax=20

Variable Selection: All subsets + rdCV

Criterion: SEP (rdCV)

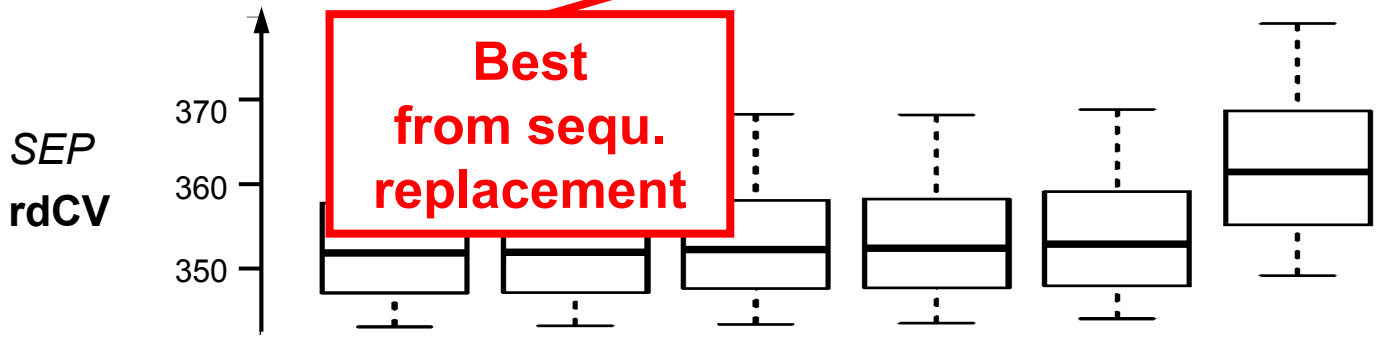
Data: **HEAT-ELE** (122 × 13)

Variables		Best subsets (SEP_{FINAL} from rdCV)					
No.	Name	1 st	2 nd	3 rd	4 th	5 th	**20 th**
1	C						
2	H	■	■	■	■	■	■
3	N	■	■	■	■	■	
4	C/H						
5	C*H	■	■	■	■	■	■
6	C*C						
7	H*H						
8	N*N						
9	ln (C)					■	
10	ln (H)	■	■			■	
11	ln (N)						■
12	ln (C/H)						
13	ln (C*H)	■		■		■	
m_{SEL}		**5**	**4**	**4**	**3**	**6**	**3**
SEP_{FINAL}		**352.12**	**352.13**	**352.35**	**352.36**	**352.90**	**354.47**
a_{FINAL}		3	3	3	3	3	**3**

Best subset regression

Best from sequ. replacement

Best with GA



Variable Selection: Genetic Algorithm (GA)

Criterion (fitness): $R^2(y, \hat{y})$, Pearson (!) [R-library *subselect; genetic()*]

Significant improvement of prediction performance.

Very time consuming.

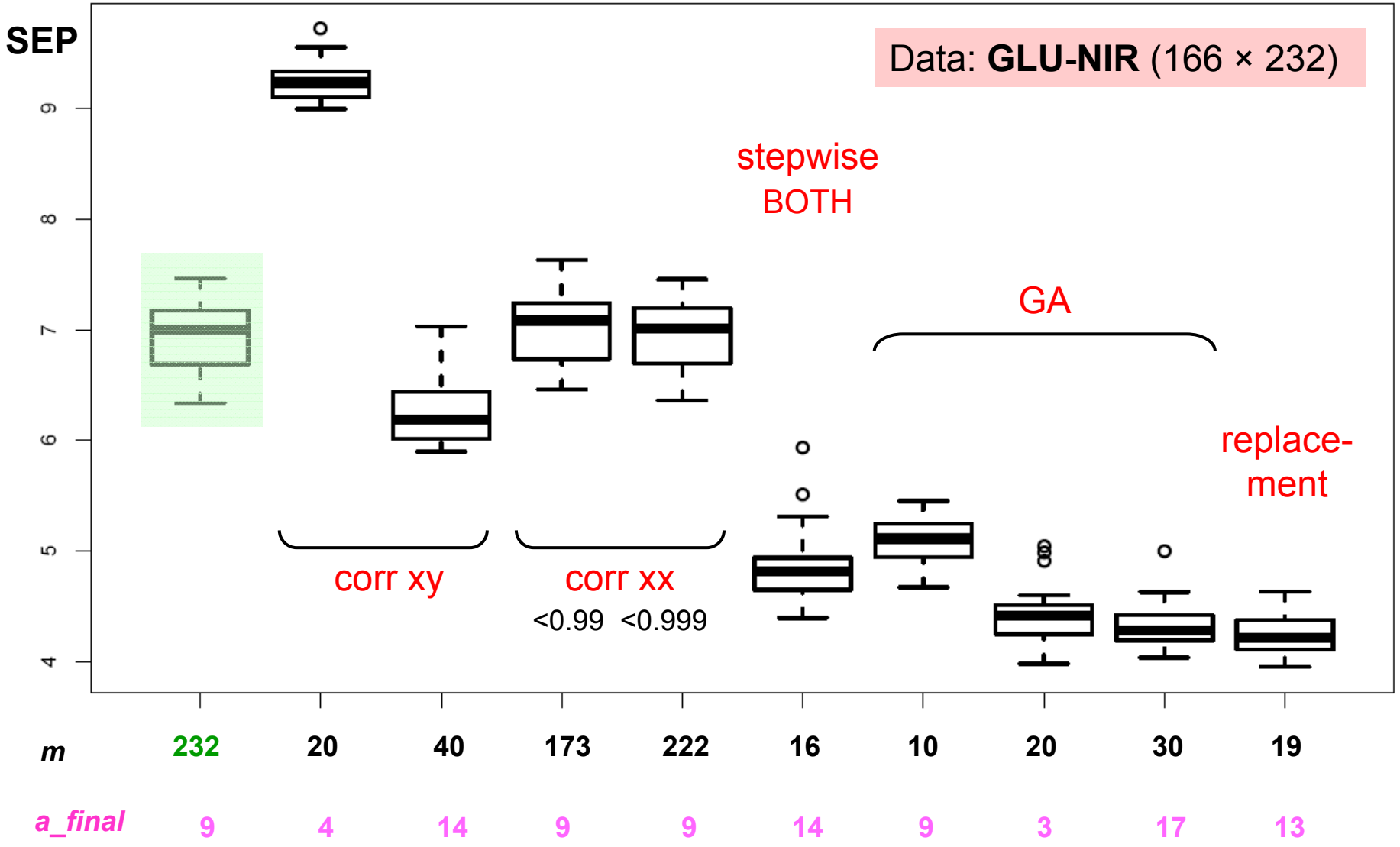
Rather not appropriate for m >500.

Variable Selection Methods

- all subsets
- best subset regression
- highest correlation with *y*
- highly correlating *x*-variables
- stepwise selection
- replacement
- genetic algorithm
- **many others**

- Monte Carlo (purely random),
- lasso regression, elastic net,
- VIP (variable importance in PLS),
- random forest,
- simulated annealing,
- particle swarm optimization,
- ant colony optimization,
- . . .
- ***tempting names (myth?) ...***

Selected best variable selection methods



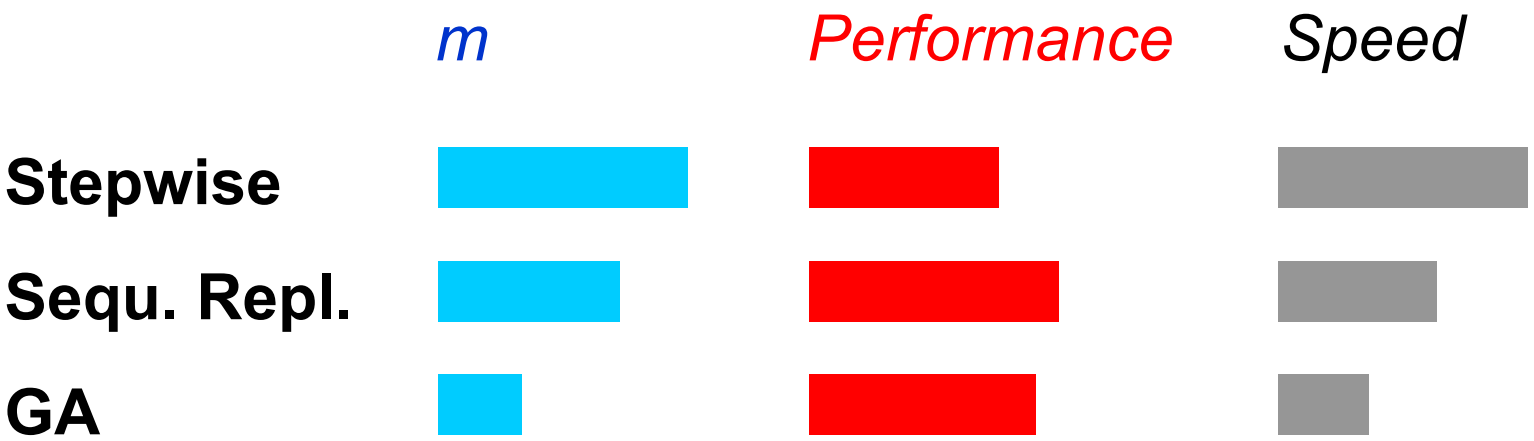
Conclusions

Variable selection often improves the model performance; however, mostly not dramatically.

Usually, no strategy is known for choosing an optimal variable selection method for a given data set.

- **Try several methods and vary parameter(s), resulting in several (many) variable subsets.**
- **Test these suggestions carefully in a similar way as the model will be applied (usually estimate the performance for test set objects - perhaps by rdCV).**

Promising methods for variable selection

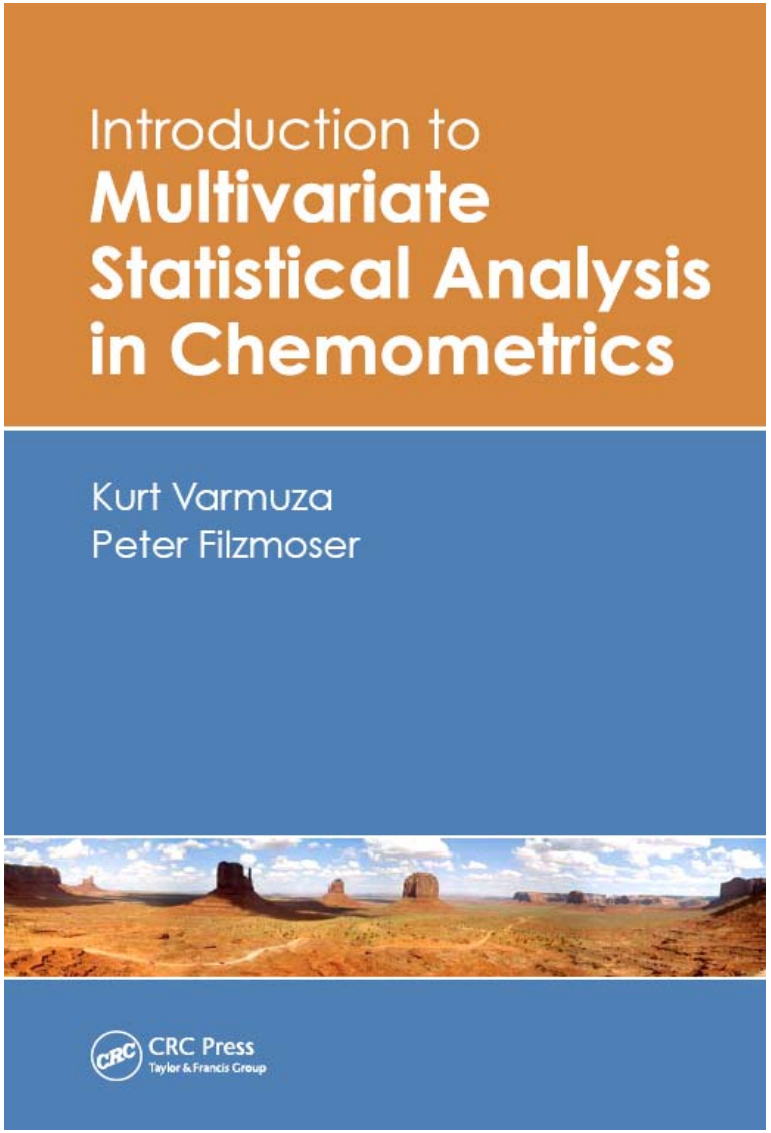


Conclusions may depend on used data sets.

Remember:

Usually, no guarantee for global optimal solution.

Thanks for collaboration: Peter Filzmoser (TU Vienna)



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

Ca 320 pages,
appr. € 100

Includes many R-codes (examples, data)

Methods are explained without using R

Info: www.lcm.tuwien.ac.at/vk/