Repeated double cross validation for classification models

Varmuza Kurt *

Filzmoser Peter ♦, Liebmann Bettina *

Vienna University of Technology
* Institute of Chemical Engineering
Laboratory for ChemoMetrics
♦ Institute of Statistics and Probability Theory

Conferentia Chemometrica, CC 2011, 18-21 September 2011, Sümeg, Hungary
20 September 2011
Version 110919b 110923

Empirical models

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

y
- continuous — multivariate calibration
- discrete, categorical — multivariate classification, pattern recognition

Multivariate classification

f classification model, classifier

often a discriminant function
- linear, e. g., DPLS, LDA
- nonlinear, e. g., SVM

Empirical models

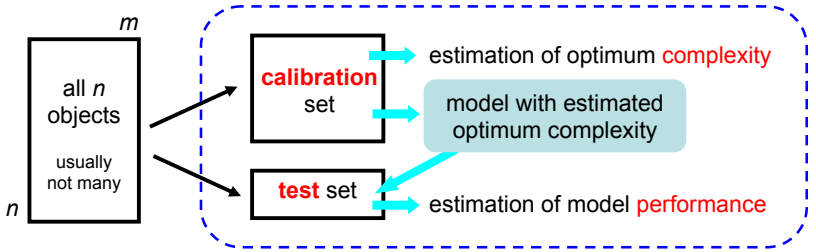
1 Estimation of **optimum model complexity**
optimum model parameter
no underfitting, no overfitting, optimal for new cases,
e. g., number of PLS components, ...

2 Estimation of **model performance**
for new cases, typ.: SEP, R^2, % correctly classified

Estimations
- are **uncertain** (distribution, tolerance interval),
- only **limited** data available (quality, size)

Proposed strategy

☞ Estimation of **optimum model complexity**
☞ Estimation of **model performance**
should be performed **independently**



Repeat !
- **variability** of optimum complexity
- **variability** of performance

Proposed strategy

optimum no. of PLS components,
optimum no. of neighbors (KNN);
SEP, R^2,
% correct

should not be given as single numbers (*)

BUT as

distribution,
confidence/tolerance interval (e. g., 95%),
central value (mean, ...),
spread (standard deviation, ...)

(*) Like any other experimentally determined values

Common methods

for repeated and separate estimations of optimum model complexity and model performance - with a rather low number of objects

- **Double bootstrap**

Calibration sets	sampling with replacement (*outer bootstrap*); *n* objects; however, duplicates, triplicates, ...; *inner bootstrap* (or cross validation) for optimization of model complexity
Predicted *y*	different no. of predictions for the *n* objects

Common methods

for repeated and separate estimations of optimum model complexity and model performance - with a rather low number of objects

- **Double bootstrap**
- **Repeated double cross validation (rdCV)**

Calibration sets	by cross validation scheme (*outer cross validation*), *inner cross validation* for optimization of the model complexity
Predicted *y*	same no. of predictions for each of the *n* objects

THIS CONTRIBUTION

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.

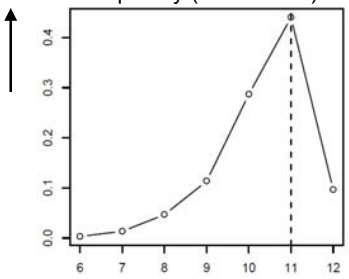
❍ Liebmann B., Filzmoser P., Varmuza K.: *J. Chemom.* **24**, 111 (2010). **Robust and classical PLS** regression compared.

rdCV for calibration - example "QSPR"

***X**: n* = 209 PAC; *m* = 467 molecular descriptors; ***y**:* **GC retention index**;

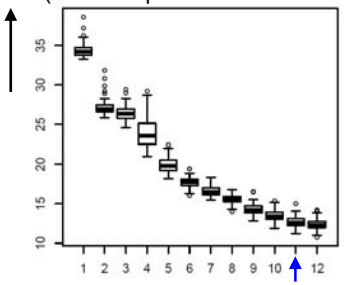
rdCV: segments for test sets: 3; segments for CV within calibration sets: 5;
100 repetitions

relative frequency (300 values)



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

SEP (each boxplot from 100 values)



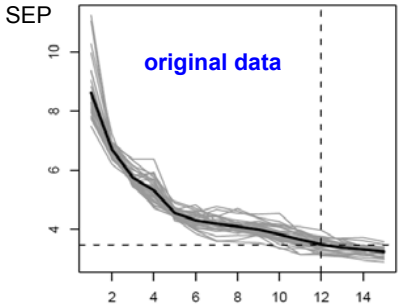
no. of PLS components

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)

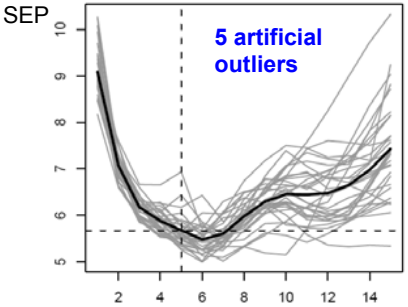
rdCV: segments for test sets: 3; segments for CV within calibration sets: 5;
30 repetitions

SEP



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

SEP

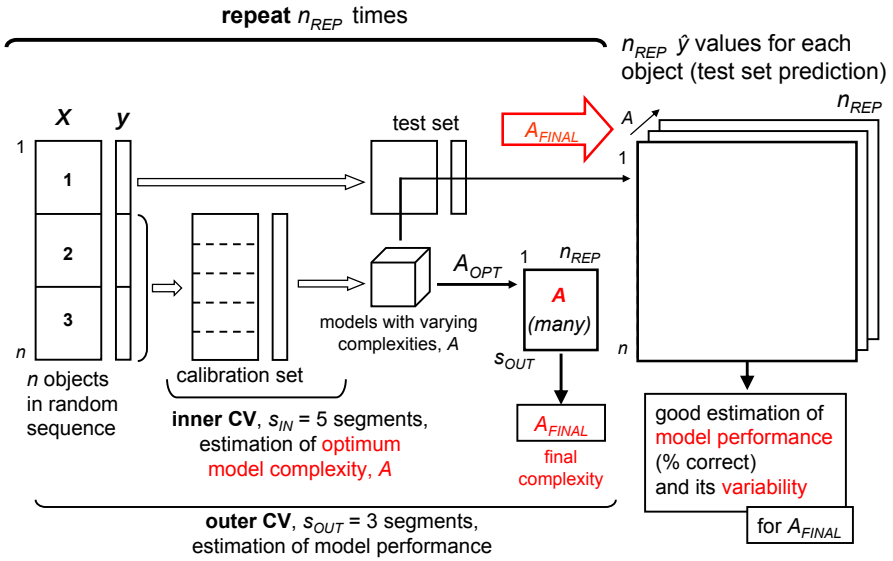


no. of PLS components;
"final" value = 5

rdCV for classification

Method	*Parameter to be optimized*
DPLS	no. of PLS components
PCA + LDA	no. of PCA components
KNN	no. of neighbors
SVM	gamma
SIMCA	no.s of PCA components
CART	tree size
ANN	no. of hidden neurons

rdCV for classification - scheme



Binary classification - performance measure

class assignment table for 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

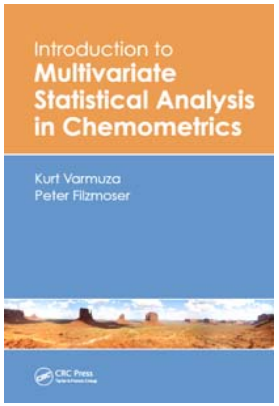
Predictive ability class 1 $P_1 = n_{11}/n_1$

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

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

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

Background - Software



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

Ca 320 pages, appr. € 85
Includes many R-codes

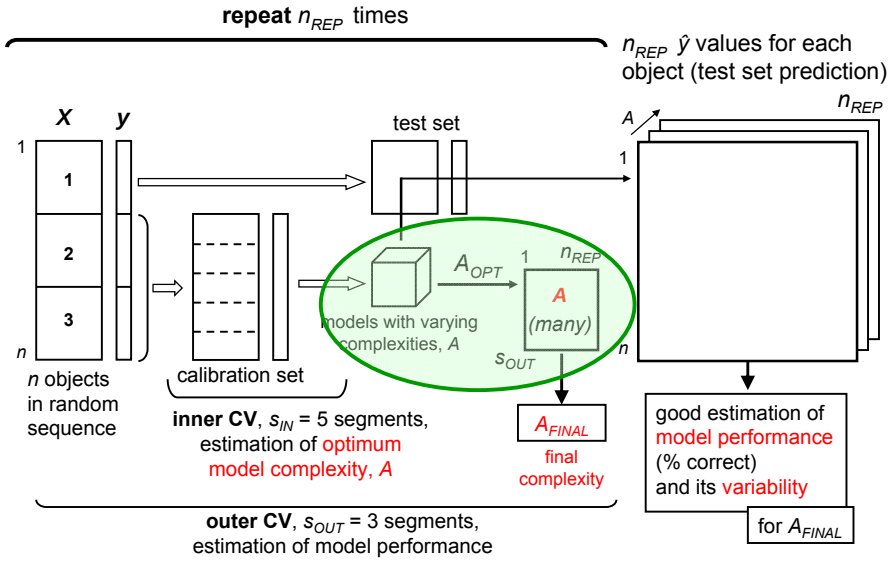
Info: www.lcm.tuwien.ac.at

rdCV software

- R-package "*chemometrics*"
- www.lcm.tuwien.ac.at/R

CM book R (LCM)

Estimation of optimum complexity



Estimation of optimum complexity

Based on **Mean** of predictive abilities, *P*,
and their **standard error**s, SE,
at varying values of the optimization parameter, *A*.

For each tested value of the optimization parameter several results for *P* are necessary (inner CV loop with s_{IN} segments).

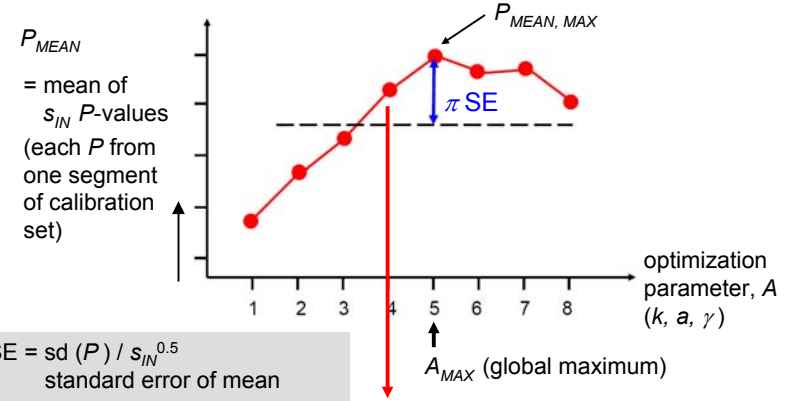
Described as **one standard error method**

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)

Estimation of optimum complexity

Standard error method (schematic example)



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

s_{IN} no. of P-values (= no. of segments in inner CV loop)

π parsimony factor (def. = 1)

A_{OPT} = smallest A with
$P_{MEAN} \geq P_{MEAN, MAX} - \pi$ SE

Estimation of optimum complexity

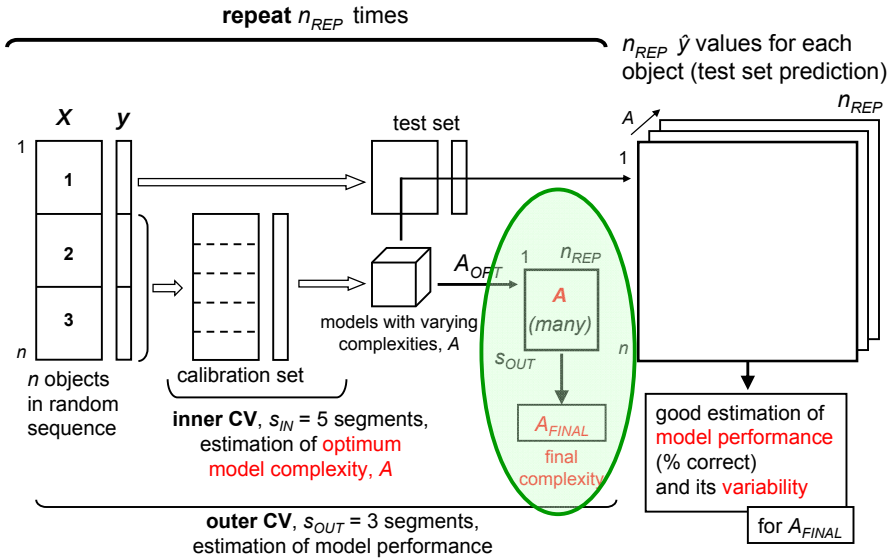
Standard error method

☞ s_{IN} *no. of segments in inner CV loop*
ca ≥ 4 (for a reasonable estimation of SE),
each segment with ca ≥ 5 objects

☞ π parsimony factor
$\pi = 0$ global maximum
$\pi = 1$ one standard error (def.)
$\pi = 2$ 95% confidence interval

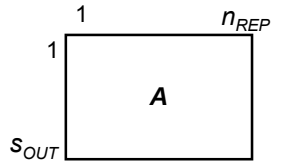
☞ Standard error method with P is also used for DPLS and SVM, because, e. g., MSE is not adequate for classification

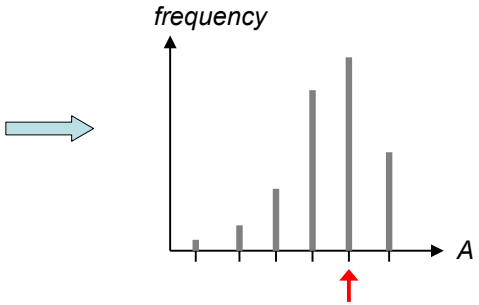
Estimation of "final" optimum complexity



Estimation of "final" optimum complexity

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





Typical, e. g.,
$s_{OUT} = 4$
$n_{REP} = 50$
give 200 estimations for the optimum complexity

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

rdCV / KNN - example "iris"

KNN

autoscaled variables; Euclidean distance; no. of neighbors (k) optimized (1 ... 10)

rdCV

s_{OUT} = 4; s_{IN} = 6; no. of repetitions = 100

Data "iris"

n = 150; m = 4; no. of classes = 3
Anderson E. (1935, collected); Fisher R.A. (1936)

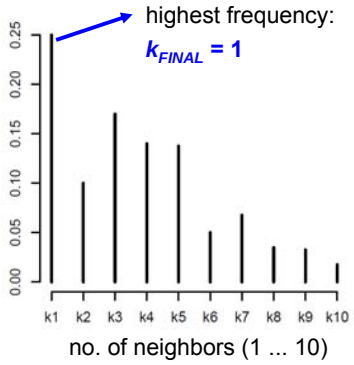
Summarized results

class	no. of objects	predictive ability
1	50	P_1 = 0.995
2	50	P_2 = 0.918
3	50	P_3 = 0.915
all	150	P = 0.943

$A_{FINAL} = k_{FINAL} = 1$
66 s comp. time

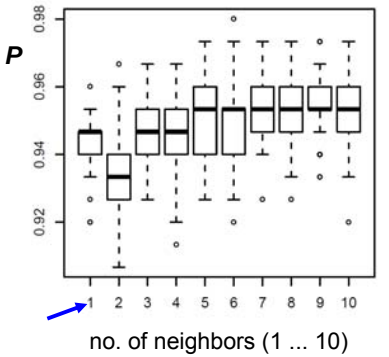
rdCV / KNN - example "iris"

Frequency distribution of found optimum number of neighbors



400 values (s_{OUT} = 4; no. of repet. = 100)

Average predictive ability (P) versus number of neighbors



Test set predictions !
Each box plot from 100 values.

rdCV / DPLS - example "iris"

DPLS

PLS2 with ***Y*** containing binary encoded classes; SIMPLS

rdCV

s_{OUT} = 4; s_{IN} = 6; no. of repetitions = 100

Data "iris"

n = 150; m = 4; no. of classes = 3

Summarized results

class	no. of objects	predictive ability
1	50	P_1 = 0.987
2	50	P_2 = 0.607
3	50	P_3 = 0.837
all	150	P = 0.810

$A_{FINAL} = a_{FINAL} = 2$
39 s comp. time

rdCV / DPLS - example "iris"

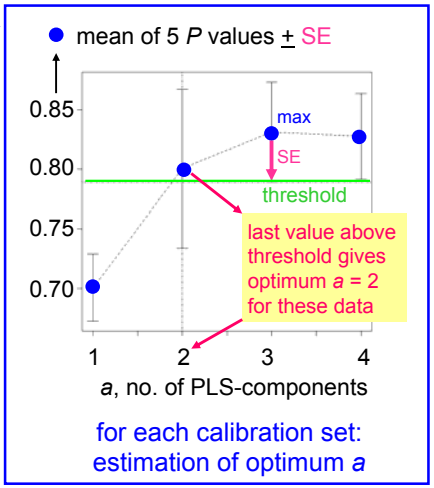
Example for estimation of optimum no. of PLS components with standard error method (inner CV)

n = 150
s_{OUT} = 3 -> 3 calibration sets, each 100 objects
s_{IN} = 5 -> 5 trainings sets, each 80 objects
5 validation sets, each 20 objects

-> 5 values for P from the 5 validation sets, at each tested a

-> mean, SE at each tested a

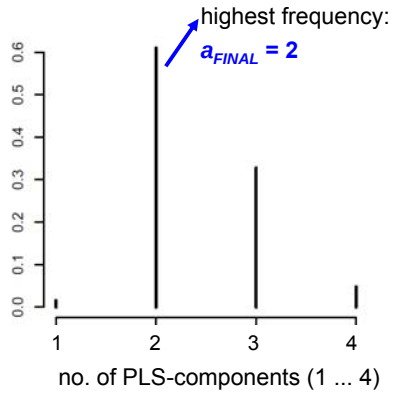
In total s_{OUT} = 3 estimations of a in each repetition



for each calibration set: estimation of optimum a

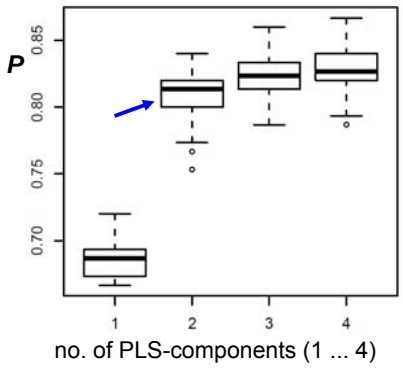
rdCV / DPLS - example "iris"

Frequency distribution of found optimum no. of PLS-components



400 values (s_{OUT} = 4; no. of repet. = 100)

Average predictive ability (*P*) versus no. of PLS-components



Test set predictions !
Each box plot from 100 values.

rdCV / SVM - example "iris"

SVM

Kernel: radial; degree: 3; parameter γ optimized

rdCV

s_{OUT} = 4; s_{IN} = 6; no. of repetitions = 100

Data "iris"

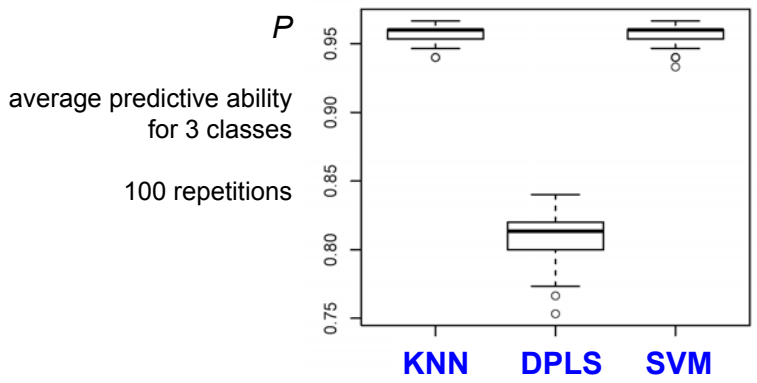
n = 150; *m* = 4; no. of classes = 3

Summarized results

class	no. of objects	predictive ability
1	50	P_1 = 1.000
2	50	P_2 = 0.971
3	50	P_3 = 0.897
all	150	P = 0.956

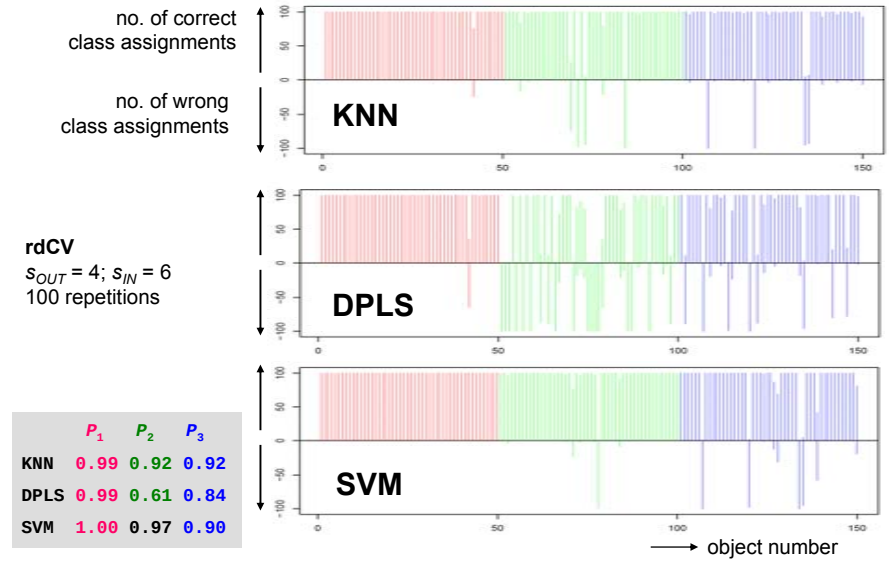
A_{FINAL} = γ_{FINAL} = **0.05**
278 s comp. time

rdCV / comparison KNN-DPLS-SVM - "iris"



rdCV: s_{OUT} = 4; s_{IN} = 6

rdCV / comparison KNN-DPLS-SVM - "iris"

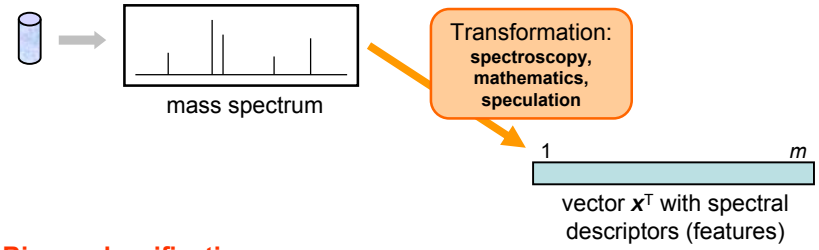


rdCV
s_{OUT} = 4; s_{IN} = 6
100 repetitions

	P_1	P_2	P_3
KNN	0.99	0.92	0.92
DPLS	0.99	0.61	0.84
SVM	1.00	0.97	0.90

rdCV / KNN-DPLS-SVM - example "phenyl"

Spectra-structure relationship



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 '*chemometrcis*'

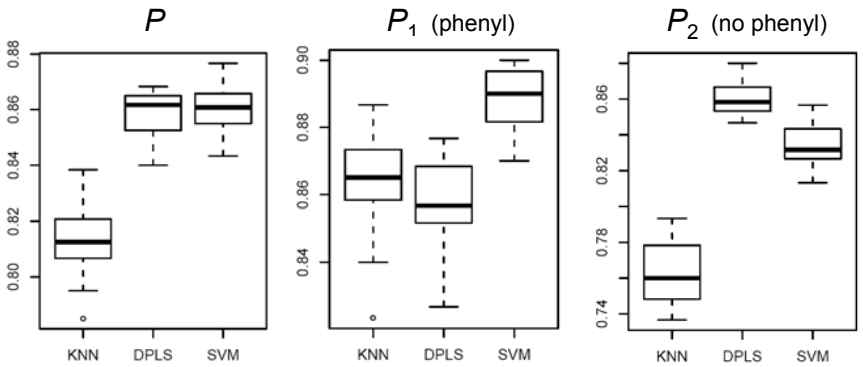
C_6H_5 -

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)

rdCV / KNN-DPLS-SVM - example "phenyl"

rdCV: 20 repetitions; s_{OUT} = 2; s_{IN} = 6



KNN: k_{FINAL} = 3; DPLS: a_{FINAL} = 2; SVM: γ_{FINAL} = 0.0002

Computation time: KNN, 550 s; DPLS, 42 s; SVM, 940 s

rdCV / KNN-DPLS-SVM - example "olive"

Origin of Italian olive oil

n = 572

9 classes (with 25 ...206 samples) from 9 areas in Italy

m = 8 fatty acid concentrations

R-package 'classify', data(olive)

Forina, M. and Armanino C. and Lanteri S. and Tiscornia E.: in Food Research and Data Analysis; ed. Martens, H., Russwurm Jr., H., 189-214. Applied Science Publ. London (1983)

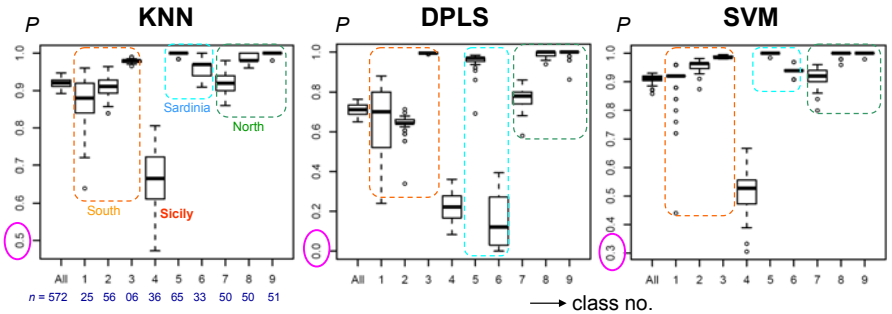
rdCV / KNN-DPLS-SVM - example "olive"

Origin of Italian olive oil

Class no.	Area		*n*	*P* (KNN)	*P* (DPLS)	*P* (SVM)
1	North Apulia		25	0.87	0.65	0.90
2	Calabria	South	56	0.91	0.64	0.95
3	South Apulia		206	0.98	1.00	0.99
4	Sicily		36	0.66	0.23	0.50
5	Sardinia inland		65	1.00	0.96	1.00
6	Sardinia coast		33	0.06	0.16	0.94
7	Liguria west		50	0.92	0.77	0.91
8	Ligura east	North	50	0.99	0.99	1.00
9	Umbria		51	1.00	0.99	1.00
all			572	0.92	0.71	0.91

rdCV / KNN-DPLS-SVM - example "olive"

rdCV: 50 repetitions; s_{OUT} = 2; s_{IN} = 6



P = 0.92 P = 0.71 P = 0.91

KNN: k_{FINAL} = 1; DPLS: a_{FINAL} = 7; SVM: γ_{FINAL} = 0.07

Repeated Double Cross Validation (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 parameter (complexity) and of performance

Easily applicable and fast

Software
- R-package "*chemometrics*"
- www.lcm.tuwien.ac.at/R