Comparison of some Linear Regression Methods – Available in R – for a QSPR Problem

Kurt Varmuza [1*] and Peter Filzmoser [2]

Vienna University of Technology

[1] **Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology**
Getreidemarkt 9/166,
A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at,
www.lcm.tuwien.ac.at

[2] **Institute of Statistics and Probability Theory,
Vienna University of Technology**
Wiedner Hauptstrasse 8-10,
A-1040 Vienna, Austria
P.Filzmoser@tuwien.ac.at,
www.statistik.tuwien.ac.at/public/filz

Poster Presentation:

4th German Conference on Chemoinformatics / 22nd CIC-Workshop

Goslar, November 9-11, 2008

Introduction

A chemical/physical/biological property y of chemical compounds can be modeled by a set of molecular descriptors x_j derived from the chemical structures.

In a **linear regression model** we estimate y by

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

using m regressor variables.

The regression coefficients $b_1, \ldots, b_m$ and intercept b_0 are estimated using a data set $\boldsymbol{X}(n \times m)$ and $\boldsymbol{y}(n \times 1)$.

For highly correlating x-variables and/or $m > n$ the traditional OLS (ordinary least-squares) regression method cannot be used. Alternatives are for example

- PLS (partial least-squares) regression
- robust PLS regression
- PCR (principal component regression)
- Ridge regression
- Lasso regression

All these methods are available in the free software system $\mathcal{R}$ [1] by the package **"chemometrics"** [2].

This package includes the function "mvr_dcv" [3] for **repeated double cross validation (RDCV)**, comprising

- selection of an optimal model complexity of PLS models [4], and
- careful evaluation of the prediction performance.

PLS and robust PLS regression

Replace $\boldsymbol{X}$ in the original model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

by **latent variables** $\boldsymbol{T}$ of lower dimension, such that

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathsf{T}} + \boldsymbol{E}$$

Consider the regression model for $\boldsymbol{y}$ on $\boldsymbol{T}$,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} = (\boldsymbol{T}\boldsymbol{P}^{\mathsf{T}})\boldsymbol{b} + \boldsymbol{e}_T = \boldsymbol{T}(\boldsymbol{P}^{\mathsf{T}}\boldsymbol{b}) + \boldsymbol{e}_T = \boldsymbol{T}\boldsymbol{g} + \boldsymbol{e}_T$$

and estimate the coefficients $\boldsymbol{g}$.

$\boldsymbol{t}_1, \ldots, \boldsymbol{t}_a$ are the columns of $\boldsymbol{T}$, and they are obtained sequentially by

$\mathrm{cov}(\boldsymbol{X}\boldsymbol{w}_j, \mathbf{y}) \rightarrow \max$ under $||\boldsymbol{t}|| = ||\boldsymbol{X}\boldsymbol{w}_j|| = 1$ and orthogonality constraints.

Using for "cov" a robust estimator like the M-estimator [5] results in **robust PLS**, see [6].

PCR

Like for PCR a latent variable model is used,

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{g} + \boldsymbol{e}_T$$

with $a < m$ regressor variables $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_a$. These are taken as the first a principal components (PCs) of $\boldsymbol{X}$. Using robust PCs results in **robust PCR** [2].

Ridge and Lasso regression

Minimize the sum of squared residuals,

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \rightarrow \min$$

under

$b_1^2 + \ldots + b_m^2$	< const	Ridge regression
$\lvert b_1\rvert + \ldots + \lvert b_m\rvert$	< const	Lasso regression

Ridge regression gives an explicit solution for the regression coefficients, $\boldsymbol{b}_{RIDGE} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda_R \boldsymbol{I})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.

Lasso regression has to be solved by an optimization routine. Depending on the size of "const", some of the regression coefficients are exactly zero. Thus, Lasso regression acts like a **variable selection method.**

Usage within $\mathcal{R}$

PLS:	`plsr`	in `library(pls)`
rob. PLS:	`prm`	in `library(chemometrics)`
PCR:	`pcr`	in `library(chemometrics)`
Ridge:	`lm.ridge`	in `library(MASS)`
Lasso:	`lars`	in `library(lars)`

Further, and more sophisticated evaluation schemes are in the library **"chemometrics"**, see the help file [2].

Application (1)

QSPR example

n = **209 polycyclic aromatic compounds**, 3D, all H-atoms; *Corina* [7]

$\boldsymbol{y}$ **gas-chromatographic retention indices**, *Lee* indices [8]

$\boldsymbol{X}$ m_1 = 467 molecular descriptors; *Dragon* [9]

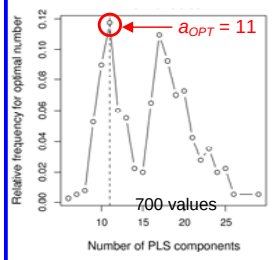
m_2 = 13 descriptors selected by a genetic algorithm; *MobyDigs* [10]

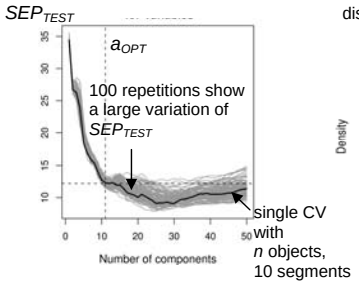
$\mathcal{R}$: `data(PAC)  # load data from library chemometrics`

PLS

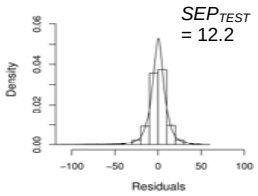
Evaluation: **RDCV** with 7 and 4 segments in outer and inner loop, resp.; 100 repetitions

m_1 = 467 descriptors

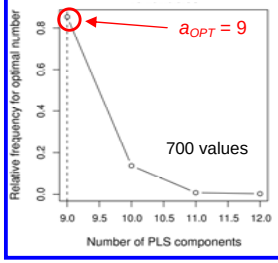


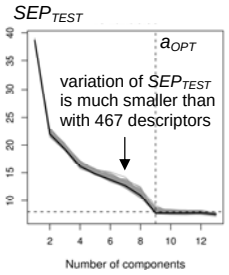


distribution of 20,900 residuals

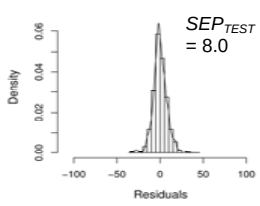


m_2 = 13 descriptors





distribution of 20,900 residuals



A single cross validation can give misleading results.
Repeated double cross validation (or bootstrap) is recommended.

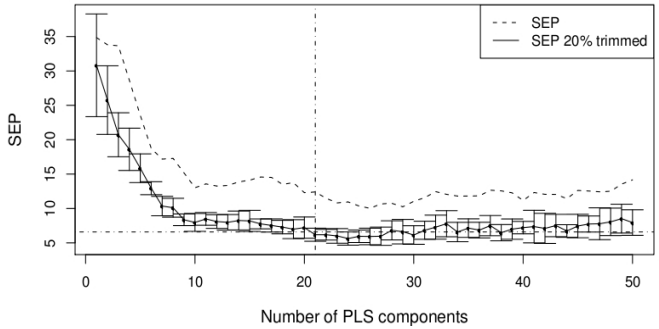
```
R: res_pls <- mvr_dcv(y~X,ncomp=50,data=PAC,method="simpls")
   plotSEPmvr(res_pls,res$optcom,PAC$y,PAC$X)
```


Application (2)

Robust PLS

Evaluation: 10-fold CV

Result: optimal number of PLS components is 21 (trimmed SEP)

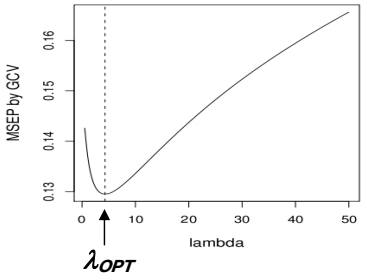


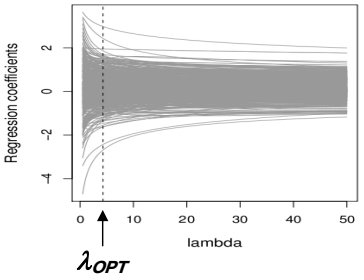

```
R: rpls <- prm_cv(PAC$X,PAC$y,a=50,trim =0.2,plot.opt="TRUE")
```

Ridge regression

Evaluation: generalized cross validation (GCV, an approx. leave-1-out)

Result: optimal Ridge parameter (λ) is 4.3, see x-axis in plots



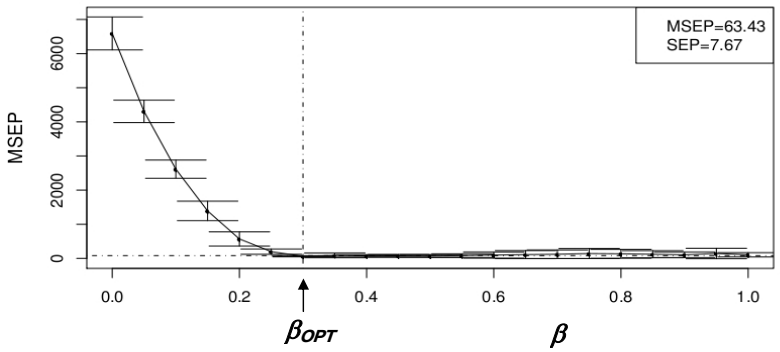



```
R: res_rid <- plotRidge(y~X,data=PAC,lambda=seq(0.5,50,by=0.05))
```

Lasso regression

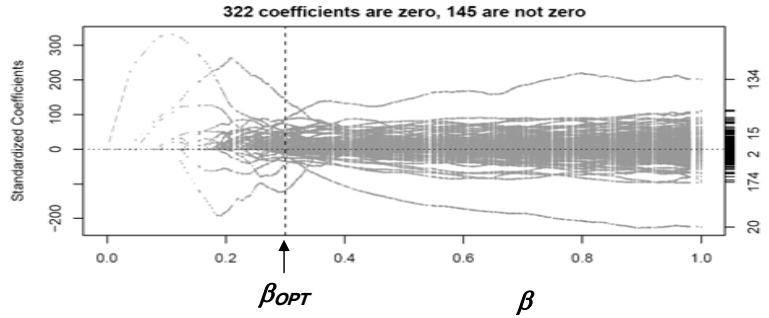
Evaluation: 10-fold CV

Result: optimal Lasso parameter β (horizontal axis) is 0.3.




```
R: res_lasso <- lassoCV(y~X,data=PAC,K=10,fraction=
                seq(0,1,by=0.05))   # K: number of CV segments
```

Resulting model: Plot shows the regression coefficients depending on the size constraint β (horizontal axis); for β_{OPT}, 332 coefficients are exactly zero.




```
R: res_coef <- lassocoef(y~X,data=PAC,sopt=res_lasso$sopt)
```


Comparison of results

Method	*m**	*a*	SEP_{TEST}	$SEP^{0.2}$
PLS	467	11	12.2	5.7
PLS	13	9	8.0	4.7
Robust PLS	467	21	-	6.2
PCR	467	21	14.2	7.9
Ridge regression	467	-	-	4.0
Lasso regression	145	-	-	5.0

*m**	number of variables in the final model
a	number of PLS/PCR components
SEP_{TEST}	SEP from repeated double cross validation
$SEP^{0.2}$	SEP with 20% trimming of largest absolute residuals

A fair comparison with robust methods is only possible with the **trimmed SEP $^{0.2}$** which excludes potential outliers.

For this data set, **Ridge regression** results in the best prediction model with a $SEP^{0.2}$ of 4.0. PLS with 13 GA-selected variables and Lasso regression with 145 variables have a similar performance with a $SEP^{0.2}$ of 4.7 and 5.0, respectively.

References

[1] R: software, a language and environment for statistical computing. R Development Core Team, Foundation for Statistical Computing, www.r-project.org, Vienna, Austria, 2008.
[2] Varmuza K., Filzmoser P.: Introduction to ultivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA, in print (2009).
[3] Filzmoser B., Liebmann B., Varmuza K.: submitted (2008).
[4] Our R function "mvr_dcv" uses a PLS package, described by Mevik B.H. and Wehrens R., J. Stat. Software 18 (2007) issue 2, 1-24.
[5] Maronna R., Martin D., Yohai V.: Robust statistics: Theory and methods. Wiley, Toronto, ON, Canada (2006).
[6] Serneels S., Croux C., Filzmoser P., Van Espen P. J.: Chemom. Intell. Lab. Syst. 79 (2005) 55-64.
[7] Corina software, Molecular Networks GmbH Computerchemie, www.mol-net.de, Erlangen, Germany (2004).
[8] Lee M.L., et al., Anal. Chem. 51 (1979) 768-773.
[9] Dragon software, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004).
[10] MobyDigs software, 1.0. Talete srl, www.talete.mi.it, Milan, Italy (2004).

Acknowledgment. This work was partly founded by the Austrian Research Promotion Agency (FFG), BRIDGE program, project no. 812097/11126.