

Introduction

The LASSO as variable selection method is applied to 4 regression data sets from QSPR/QSAR and analytical chemistry. Subsequently, the selected subsets are subjected to repeated double cross validation (rdCV) for PLS models, and their prediction performance is compared with using all variables as well as variables selected by a Genetic Algorithm.

LASSO

The LASSO method is a **variable selection method** proposed by R. Tibshirani [1].

"LASSO" = Least Absolute Shrinkage and Selection Operator.

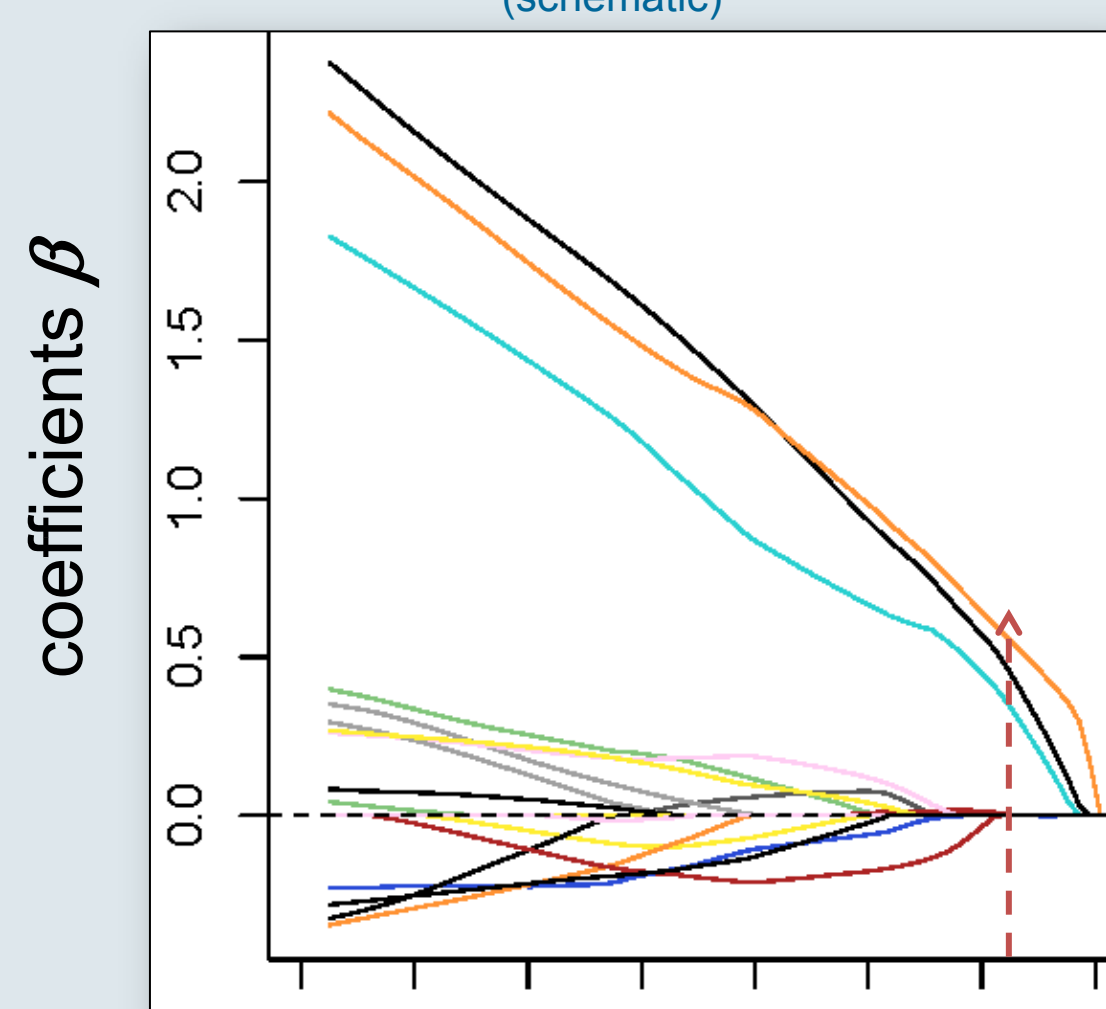
It is a constrained version of ordinary least squares (OLS) regression, and typically used for regression of a single response variable, y , on a predictor matrix X .

$$\begin{aligned} \text{OLS} \quad & \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^m \beta_j x_{ij} \right)^2 \Rightarrow \min \\ \text{LASSO} \quad & \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^m \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^m |\beta_j| \Rightarrow \min \end{aligned}$$

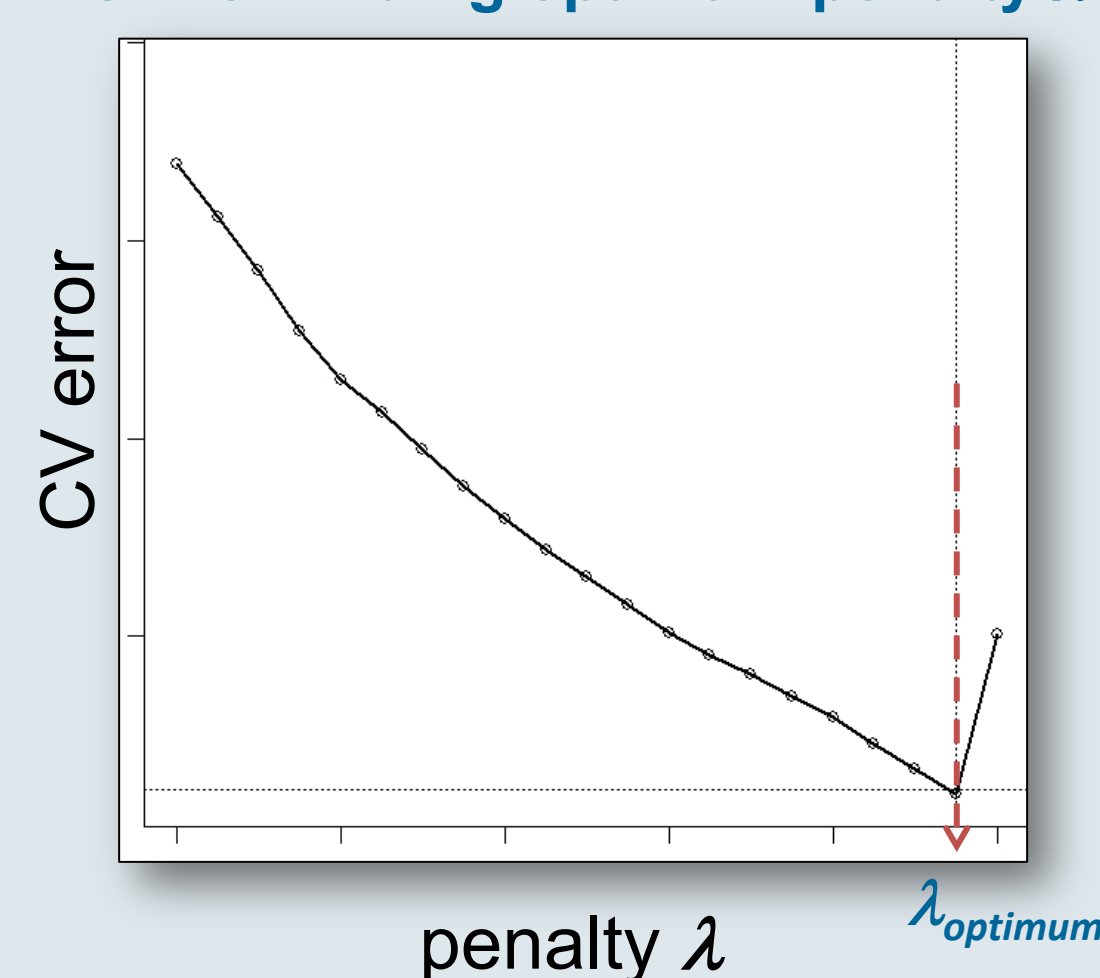
estimate regression coefficients β by minimizing an error function based on squared residuals

- The absolute size of the regression coefficients β is constrained.
- The higher the penalty λ , the more regression coefficients are shrunk towards zero \Rightarrow variable selection! [2]
- Penalty $\lambda = 0$ gives the OLS solution if $m < n$; if $m > n$, the maximum possible number of selected variables is the number of samples.
- Determine optimum λ by cross-validation or related methods [3].

Path of regression coefficients [3]
(schematic)



CV for finding optimum penalty λ



Data


PAC. 209 x 467 ($n \times m$), GC retention index of polyaromatic compounds by molecular descriptors [4,5]

PCB. 209 x 2106, octanol-water partition coefficient (log P) of polychlorinated biphenyls by molecular descriptors [6, 7]

Selwood. 53 x 31, in-vitro antifilarial activity of antifilarial antimycin A1 analogues by physicochemical descriptors [8, 9]

NIR. 166 x 235, glucose concentration of fermentation mash by NIR absorbance spectra [4, 10]

Methods and Software

LASSO. LASSO based on Least Angle Regression [3,11] with  [12] package "lars" [13]; Optimum LASSO penalty determined by 10-fold cross-validation with R-package "chemometrics" [4], function lassoCV; Extract variables with coefficients $\beta \neq 0$, i.e. "LASSO selected variables".

Genetic Algorithm GA. [14] Extracts "GA selected variables".

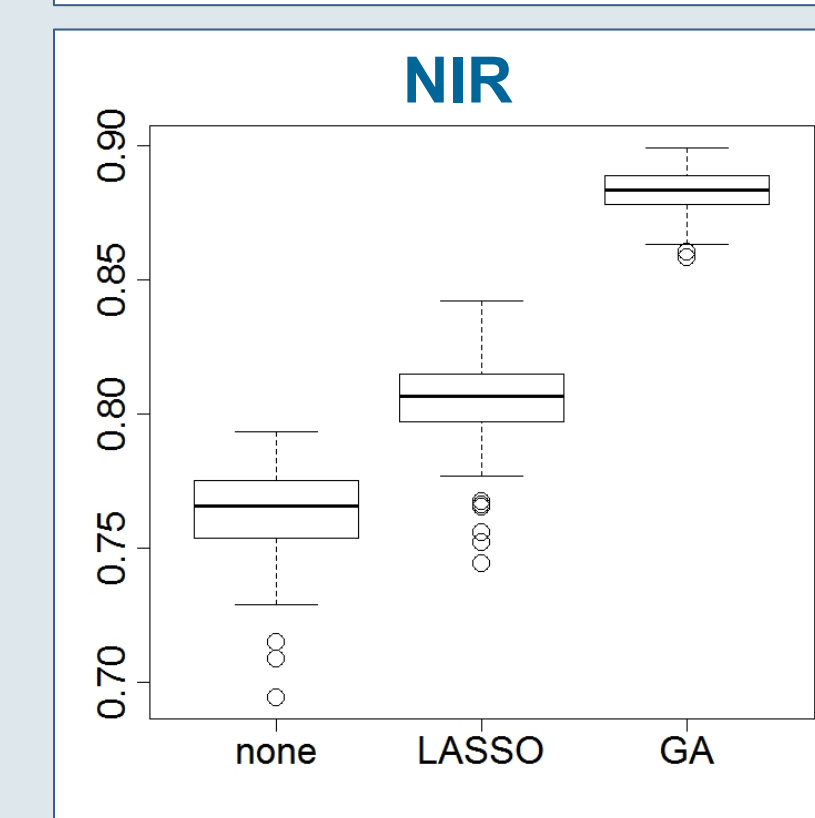
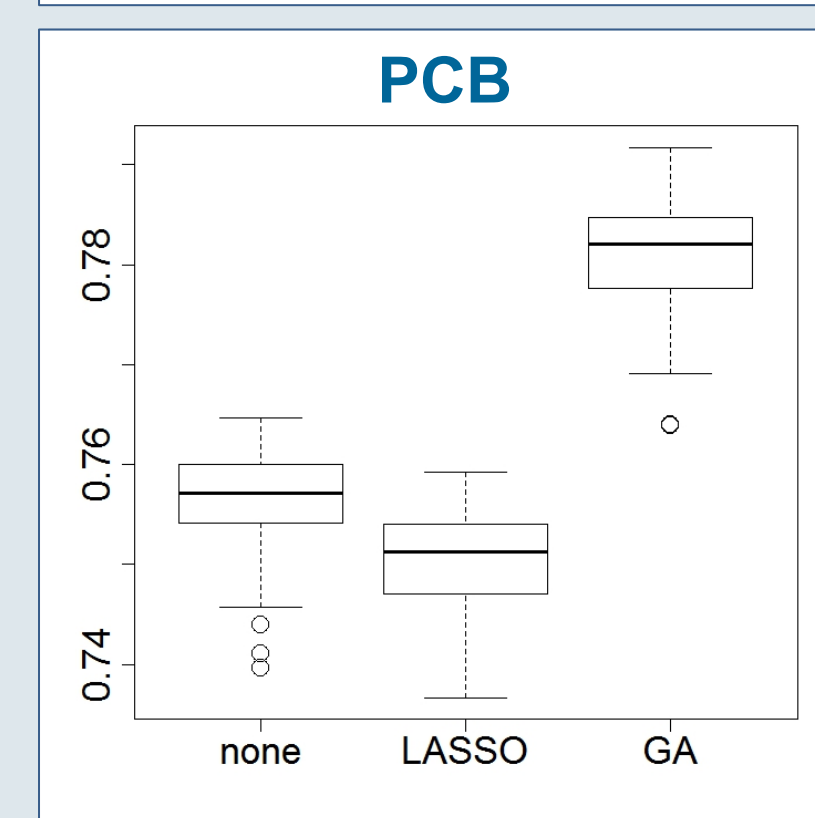
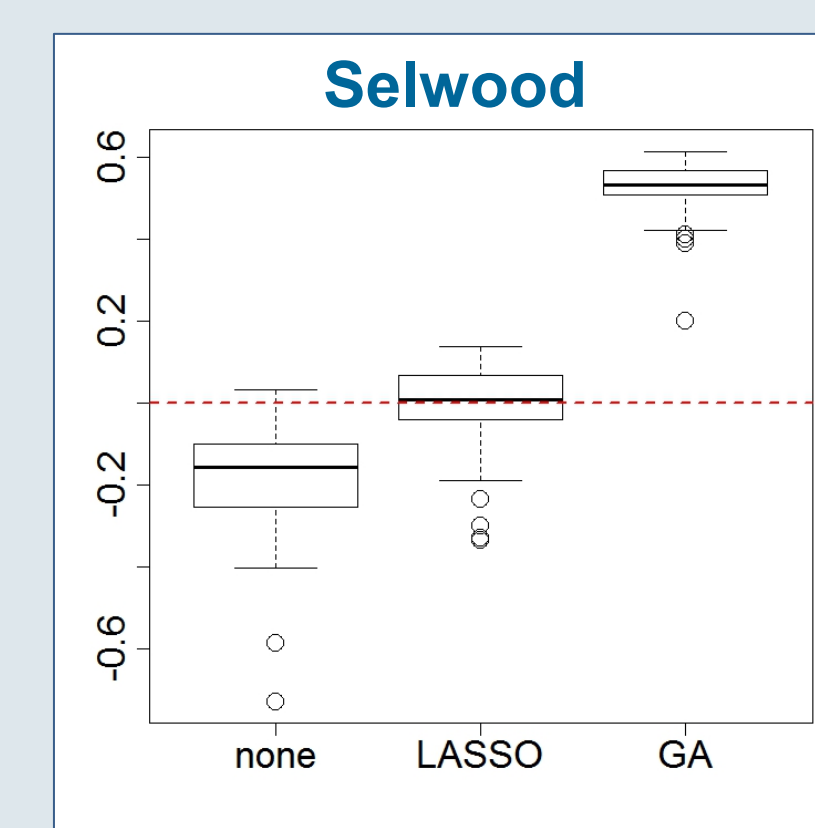
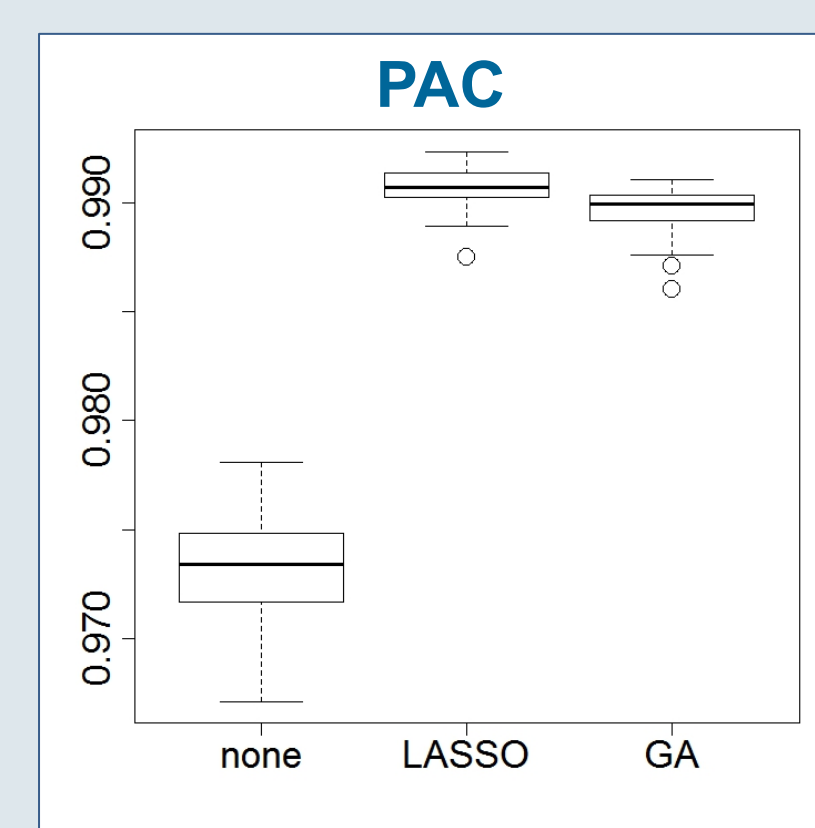
Repeated double Cross-Validation. rdCV [4,5] for estimating model performance of **PLS models with (a) all variables, (b) LASSO selected variables, (c) GA selected variables**. Data is split into 4 outer random segments for test set validation, and 10 inner random segments for CV to determine optimum number of PLS components. The random data split is repeated 100 times, giving different sample order. Each repetition yields n test set predicted y values, which are summarized by the performance measure $Q^2_{rdCV} = 1 - (\text{PRESS}/n_{\text{EXTERNAL}})/(\text{TSS}/n_{\text{TRAIN}})$ based on [15]. In the case of rdCV, $n_{\text{EXTERNAL}} = n_{\text{TRAIN}} = n$.

Summary of PLS model performance

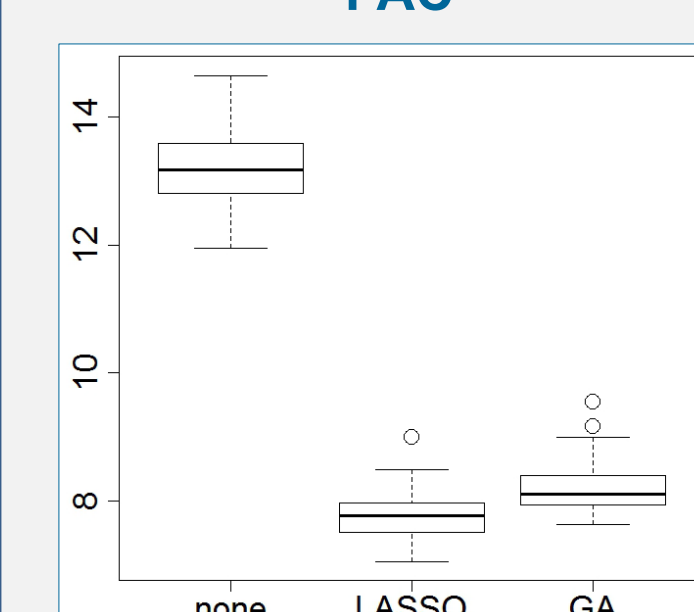
Data set	Var. Sel.	n	m	mean Q^2_{rdCV}	mean SEP _{rdCV}	a_{rdCV}	$y_{\min} - y_{\max}$ (y_{mean})
PAC	-	209	467	0.973	13.2	10	197 - 504 (338)
	LASSO		74	0.991	7.8	14	
	GA		13	0.990	8.2	9	
PCB	-	209	2106	0.757	0.36	1	4.5 - 8.2 (6.4)
	LASSO		3	0.750	0.37	1	
	GA		3	0.781	0.35	2	
Selwood	-	31	53	-0.184	0.89	1	-1.00 - 1.84 (0.46)
	LASSO		7	-0.002	0.81	1	
	GA [8]		4	0.529	0.56	3	
NIR	-	166	235	0.763	6.9	9	0.32 - 54.4 (17.5)
	LASSO		116	0.804	6.3	12	
	GA		15	0.883	4.8	10	

- n , number of samples m , number of variables $y_{\min}/y_{\max}/y_{\text{mean}}$, min./max./mean value of y -variable
- Q^2_{rdCV} , predictive squared correlation measure based on test data (rdCV), cf. **boxplots** below
- mean Q^2_{rdCV} , average of Q^2_{rdCV} from 100 random splits into test data and training data
- mean SEP_{rdCV}, standard deviation of prediction errors from test sets at a_{rdCV} PLS components, mean from 100 values
- a_{rdCV} , optimum number of PLS components (rdCV)

Variation of Q^2_{rdCV} for 100 different test sets



Variation of SEP_{rdCV} for 100 different test sets: PAC



The alternative performance criterion SEP shows similar (inverse) behaviour as Q^2 , and has the advantage of being in the same units as y .

Conclusions

- In general, variable selection improves the model performance. Only for PAC data LASSO is (slightly) better than GA. For many variables ($m = 2106$, PCB), LASSO does not improve the model performance.
- However, LASSO is substantially faster to compute than GA.
- LASSO functions are (freely) available for R, Matlab, Python, Java etc.
- LASSO may select NO variable at all for poor data, but often selects more variables than GA.
- LASSO suffers problems with highly correlated variables (cf. NIR).

General open problem: Different variable subsets are obtained from different sample subsets; how to combine these variable subsets reasonably?

References

- [1] R. Tibshirani, Journal of the Royal Statistical Society B, 58, 267 (1996).
- [2] K. Varmuza, P. Filzmoser: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009).
- [3] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd ed., Springer, NY, USA (2009).
- [4] P. Filzmoser, K. Varmuza: chemometrics: Multivariate Statistical Analysis in Chemometrics, R-package v. 1.3.8, 16-02-2012. <http://cran.r-project.org/web/packages/chemometrics/index.html>
- [5] P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometrics, 23, 160 (2009).
- [6] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, J. Chem. Inf. Comput. Sci, 42, 693 (2002).
- [7] P. Gramatica, N. Navas, R. Todeschini, Chemom. Intell. Lab. Syst 40, 53 (1998).
- [8] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Anal. Chim. Acta, 515, 199 (2004).
- [9] D. L. Selwood, D. J. Livingstone, J. C. W. Comley, A. B. O'Dowd, A. T. Hudson, P. Jackson, K. S. Jandu, V. S. Rose, J. N. Stables, J. Med. Chem., 33, 136 (1990).
- [10] B. Liebmann, A. Friedl, K. Varmuza, Biochem. Eng. J., 52, 187 (2010).
- [11] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, The Annals of Statistics, 32, 407 (2004).
- [12] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2011. www.r-project.org
- [13] T. Hastie, B. Efron: lars: Least Angle Regression, Lasso and Forward Stagewise, R-package v. 0.9-8, 07-03-2011. <http://cran.r-project.org/web/packages/lars/index.html>
- [14] MobyDigs, Software, version 1.0. Talete srl, Milano, Italy, 2004. www.talete.mi.it
- [15] V. Consonni, D. Ballabio, R. Todeschini, J. Chemom., 24, 194 (2010).

Acknowledgements. Anton Friedl (Institute of Chemical Engineering, TU Vienna, Austria) is warmly thanked for providing financial support for the presenting author.