Chemistry Central Journal

Poster presentation

Open Access

Comparison of some linear regression methods – available in R – for a QSPR problem

K Varmuza*[1] and P Filzmoser[2]

Address: [1]Vienna University of Technology, Institute of Chemical Engineering, Getreidemarkt 9/166, A-1060 Vienna, Austria and [2]Vienna University of Technology, Institute of Statistics and Probability Theory, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria

* Corresponding author

from 4th German Conference on Chemoinformatics
Goslar, Germany. 9–11 November 2008

Published: 5 June 2009

Chemistry Central Journal 2009, **3**(Suppl 1):P37 doi:10.1186/1752-153X-3-S1-P37

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P37

An important task in science and technology is modeling a property *y* by several variables *x*. In QSPR (quantitative structure-property relationships) the *x*-variables are often numerical molecular descriptors, and the *y*-variable is a chemical or physical property. Several efficient regression methods are available to find appropriate regression coefficients *b*1, *b*2,..., *bm* and the intercept *b*0 for a linear model

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

with $\hat{y}$ for the predicted property and m the number of *x*-variables.

Efficient means that model generation is possible for data with more variables than objects, for data with highly correlating variables, and that the complexity of the model is optimized for best prediction performance (not necessarily for best fit).

The compared methods comprise PLS (partial least-squares) regression, robust PLS, PCR (principal component regression), ridge regression, and lasso regression as implemented in the free software system R [1] by the package "chemometrics" described in [2]. The strategy "repeated double cross validation" [2] has been applied to optimize the model complexity (i.e. to find the optimum number of PLS components), and to estimate the prediction errors for new cases. The QSPR problem used is modeling the gas chromatographic retention indices of 209 polycyclic aromatic compounds characterized by 467 molecular descriptors.

References

1. Software R: *A language and environment for statistical computing, version 2.2.7.* 2008 [http://www.r-project.org]. *Vienna, Austria: R Development Core Team, Foundation for Statistical Computing*
2. Varmuza K, Filzmoser P: *Introduction to multivariate statistical analysis in chemometrics CRC Press, Boca Raton, FL, USA*; 2009 in press.