Variable selection and its strict evaluation

Kurt Varmuza, Peter Filzmoser
Vienna University of Technology, Department of Statistics and Probability Theory,
Wiedner Hauptstrasse 7/107, A-1040 Vienna, Austria
E-mails: kvarmuza@email.tuwien.ac.at, P.Filzmoser@tuwien.ac.at

Data sets for multivariate regression models in chemometrics typically have some hundred to some thousand *x*-variables, hence a variable selection appears useful or even necessary [1]. PLS and similar regression methods can use data sets with more variables than objects and also highly correlating variables; nevertheless, there are arguments for variable selection: (1) Use of many variables gives a better fit of the model for the training data; however, an optimum prediction performance of test data is desired; reduction of the variables may avoid overfitting and may result in an improved prediction performance. (2) Models with too many variables are hard to interpret.

An exhaustive search for the best subset of variables is not possible for data sets with the mentioned numbers of variables, and therefore in practice all resulting subsets from variable selection have to be considered as being suboptimal. No general rule is known for suggesting the best variable selection strategy for a given data set. Consequently, several strategies with varying parameters are often applied, resulting in different subsets of the original variables. The performances of models derived from these variable subsets have to be compared - independently from any performance measures obtained during variable selection. Note, that a combination of good variable subsets not necessarily improves or even keeps the model performance.

The evaluation strategy applied here is repeated double cross validation (rdCV) [2] together with PLS regression. This method gives an estimation of the optimum number of PLS components, and an estimation of the model performance for test set objects that have not been used in any step of model creation or optimization. Furthermore, rdCV delivers estimations of the variation of the final optimum number of PLS components, and the final standard error of prediction (SEP), thus supporting a reasonable comparison of different variable sets.

Several variable selection methods have been compared and applied to data sets from analytical chemistry and QSPR [3]. The variable selection methods comprise uni- and bivariate methods (e. g., selection of *x*-variables with maximum correlation coefficients with the response variable y), stepwise selection methods, and multivariate methods (e. g., based on absolute standardized regression coefficients of PLS models with all variables, but also by using random forests and genetic algorithms). A set of user-oriented R-functions makes these variable selection methods and the rdCV-PLS evaluation of the resulting variable subsets easily accessible.

References

[1] Varmuza K, Filzmoser P Introduction to multivariate statistical analysis in chemometrics. Boca Raton, FL, USA: CRC Press. (2009).

[2] Filzmoser P, Liebmann B, Varmuza K Repeated double cross validation. *J. Chemometr.*, **23**, 160-171, (2009).

[3] Varmuza K, Filzmoser P, Dehmer M (2013) Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. Computational and Structural Biotechnology Journal, 5 [6], e201302007, 1-10. Open access: http://journals.sfu.ca/rncsb/index.php/csbj/article/view/csbj.201302007/224.