Variable Selection for Multivariate Models - Myth and Reality

Kurt Varmuza

Vienna University of Technology, Dept. of Statistics and Probability Theory, A-1040, Austria
kvarmuza@email.tuwien.ac.at; www.lcm.tuwien.ac.at/vk/

Autumn School of Chemoinformatics, Nara, Japan, 28 November 2013

Abstract for lecture

1. Introduction

For estimating a property y, often empirical models are used in chemometrics that relate a set of x-variables with y by a function or algorithm $\hat{y} = f(x_1, x_2, ..., x_m)$, with $\hat{y}$ for the predicted value of the modeled property, and m for the number of x-variables. Typical successful applications of such multivariate models are in the areas: (1) QSA(P)R with up to some thousand molecular descriptors used as x-variables for modeling chemical, physical or biological properties; (2) quantitative analytical chemistry of complex mixtures with, e. g., some hundred NIR absorbance values used as x-variables for modeling the concentrations of relevant compounds in mixtures. For data in chemistry the number of variables is often larger than the number of samples (objects, chemical structures), n; and furthermore, the variables are often highly correlated. Under these conditions, linear PLS regression is an appropriate and widely used method. Thus, the number of variables is usually big, and simply determined by the available (spectroscopic) data or software for calculation of molecular descriptors. Some relationship between the used x-variables and the desired y is usually evident from the scientific background; however, existing theoretical concepts often do not allow formulating a knowledge-based equation for function $f(\boldsymbol{x})$. A closely related aspect is an efficient selection of relevant variables that give a powerful model. Basically, PLS regression does not require variable selection for mathematical reasons, and models with many (often all available) variables may possess satisfying performances.

Nevertheless, there are arguments for variable selection as follows: (1) Use of many variables gives a good fit of the model for the calibration data; however, the user is interested in a high prediction performance for new data not used during model development. A reduction of the number of variables can avoid overfitting and can lead to an improved prediction performance. (2) A model with many variables is practically impossible to interpret. Interpretation of model parameters seems feasible only if not more than about a dozen variables are used. (3) Elimination of noise variables may yield more stable models.

Variable selection (feature selection) is an important and crucial topic in multivariate data analysis because for most practical problems only suboptimal methods can be applied - that means these methods will not necessarily find the best variable subset. "Best" means highest prediction performance of a model for new cases (test set data) but not best fit to the calibration data [1].

Essential limiting facts for variable selection are as follows.

(1) The number of possible variable subsets is 2^m-1 (including the set with all m variables); m = 20 is a low number in chemometrics but would require to develop and test models for > 10^6 variable sets; for m = 300 (typical for NIR data) a "more than astronomical" number 2.10^{90} results.

(2) No approach is known for finding the best variable subset (in the sense defined above) without testing all possible subsets. Actually, the method "best subset regression" finds the variable

subset(s) which have best fit (but not best prediction performance) by applying a sophisticated and fast strategy with testing only a small part of all possible subsets. This method is applicable up to ca 35 variables. An exhaustive test of all variable subsets (necessary for finding the subset with best prediction performance for test data) may be practicable up to ca 15 variables (32,767 subsets).

(3) Merging two "good" variable subsets (or two single "good" variables) will not necessarily result in an improved performance - it may even become worse than the performance of each subset. There is no way out of this situation - except testing all subsets.

(4) A great amount of effort and creativity has gone into the development of variable selection methods. Some are simple and fast, others are complicated, dependent on random effects, and computer time consuming. Sometimes, a "myth" appears which is stimulated by the simplicity of some (traditional) methods, or by fascinating names of other (new) methods. In reality, variable selection methods cannot guarantee to find the best variable subset in most practical situations. It is surprising that the "success" of variable selection methods (in terms of improving the prediction performance) is often not sufficiently documented or is estimated by inadequate measures. This contribution tries to set up a scheme for variable selection in practice based on a strict evaluation of the results.

Considering the not completely solvable problems of variables selection on one hand and the need for variable selection on the other hand, the following strategy is claimed here (Fig. 1) [2]. Variable selection is performed with all objects thus exploiting all information present in the available data. Several different approaches for variable selection are applied in parallel and the controlling parameters of the methods are varied, and also randomness may be included. Results are typically 5 to 50 "good" variable subsets of different size and content. No performance measures obtained during variable selection are considered for the performance of final models. The resulting variable subsets are only considered as suggestions - hopefully good one's. Consequently, these variable subsets (together with the complete set) have to be carefully tested for their capability to produce models with a high performance for test set objects. The applied strategy rdCV (repeated double cross validation) [1,3] typically gives for each variable set a series of estimations (say 20 to 100) for an appropriate performance measure (here SEP, standard error of prediction). The power of the different variable sets can be easily compared from boxplots that represent the distributions of the SEP values.

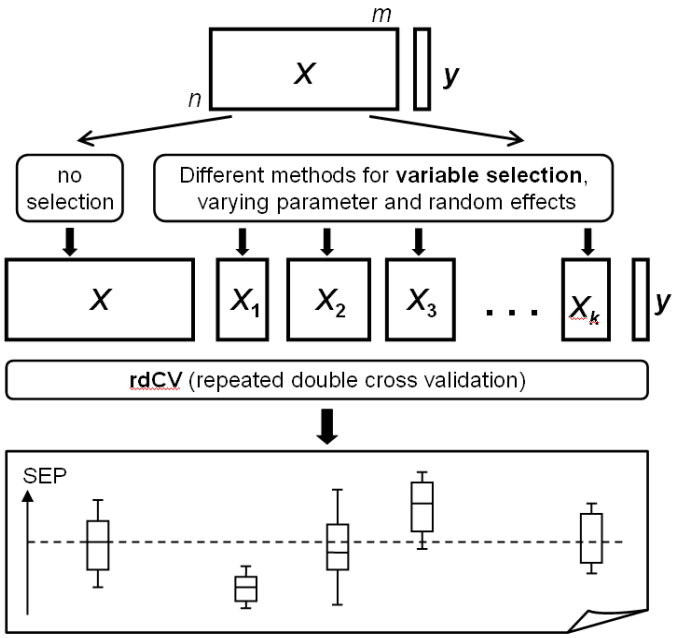

Figure 1. Strategy for variable selection. From the original data $\boldsymbol{X}$ and $\boldsymbol{y}$ (n objects, m variables), several subsets $\boldsymbol{X}_1$, ... $\boldsymbol{X}_k$, of the x-variables are created, and then tested with repeated double cross validation (rdCV). Each variable set gives, e. g., 30 estimated SEP values (standard error of prediction) and the distribution of them is presented by a boxplot. Performances of the variable subsets and the original variable set can be easily compared from the boxplots.

2. Methods for variable selection used

Among the many methods described in literature the following have been compared; for details see [1,4]. Correlation coefficient, R, is always Pearson.

(1) **corr xy**: Select x-variables with highest correlation to y.

(2) **corr xx**: Delete x-variables with a high correlation to another x-variable (e. g., $R^2 > 0.99$).

(3) **reg coeff**: Select variables with high absolute standardized regression coefficients in a PLS regression model made from all variables.

(4) **step**: Apply stepwise variable selection (usually forward/backward strategy) using BIC (Bayes Information Criterion [1,4]) as fit criterion; applicable for ca $m \leq 3000$.

(5) **repl**: Apply the sequential replacement method [5,6], resulting in variable subsets of given size with best fit (BIC); applicable for ca $m \leq 1500$.

(6) **ga**: Apply a genetic algorithm, resulting in "good" variable subsets of given size; applicable for ca $m \leq 500$.

(7) **best subset**: Apply best subset regression, an exhaustive search for the variable subsets with best fit (BIC); applicable for $m \leq 35$.

(8) **all subsets**: Check all 2^m-1 possible variable subsets with rdCV (see below), resulting in the subsets with best prediction performance (SEP, test set data); applicable for $m \leq 15$.

3. Optimum PLS models and evaluation

The rdCV (repeated double cross validation) strategy [1,3] was applied for a strict evaluation of variable sets. The CV procedure for a random split of the n objects into a calibration set (75%) and a test set (25%) was repeated 30 times, giving 30 estimations of SEP for a variable subset. SEP is the standard deviation of the prediction errors (only for test set objects) and is a user-oriented measure for the model performance. Because the prediction errors usually show a normal distribution a 95% confidence interval for predictions can be defined as $\hat{y} \pm 2\text{SEP}$. Within each calibration set an inner CV loop was used to estimate the optimum number of PLS components (A, model complexity). This approach also demonstrates the variability of A; for simplicity the most frequent value of A was used for a final model - an alternative would be a consensus machine. All software is in R [7].

4. Data

The data sets used for a comparison of variable selection methods comprise:

PAC-QSPR: n = 209 chemical structures from polycyclic aromatic compounds (3D and all H-atoms, software *Corina*); m = 2661 molecular descriptors (software *Dragon 6.0*); y is a GC retention index [8] with range 197.0 - 503.9.

GLU-NIR: n = 166 fermentation samples (centrifuged, [9]); m = 232 NIR absorbances (1115 - 2285 nm); y is the glucose content (reference method HPLC) with range 0.32 - 54.44 g/L.

HEAT-ELE: n = 122 biomass samples from plants [10]; m = 13 elemental composition of C, H, and N, and 10 derived variables (ratios, cross products, log); y is the higher heating value (HHV, calorimetric reference values with estimated experimental errors of $\pm$ 60 kJ/kg) with range 15,719 - 25,948 kJ/kg.

5. Results

Fig. 2 compares several variable selection methods for the data set **GLU-NIR.** The original data with m =232 ("all") gave SEP values with a median of 7.0 (g/L glucose). Variable selection by the (widely used) methods **corr xy** or **corr xx** (limit for R^2 0.99 and 0.999, respectively) was not successful. Considerable improvement showed **step**wise variable selection (median of SEP values 4.8) and especially the methods **ga** and **repl** (median 4.2). For data set **HEAT-ELE** with only 13 variables the exhaustive method **all subsets** could be applied; results showed that the methods **best subset** and **ga** gave near optimum variable subsets (3rd best subset, and 20th best subset, respectively). In conclusion, variable selection usually improves the model performance, however, mostly not dramatically - and the correct effect can be seen only by a strict evaluation.

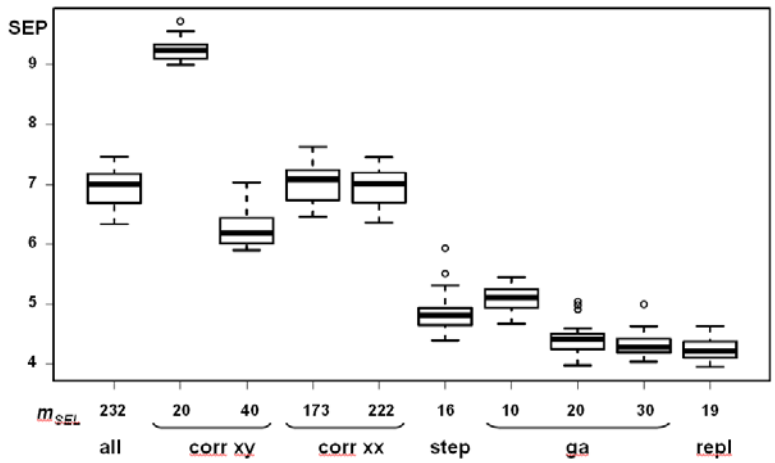

Figure 2. Application of rdCV to data set **GLU-NIR** and nine variable subsets selected by different methods.

Acknowledgement. Thanks to Peter Filzmoser (TU Vienna) for support in statistics and R-programming.

References

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

[2] K. Varmuza, P. Filzmoser, M. Dehmer, Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. Open access: http://dx.doi.org/10.5936/csbj.201302007; data and software: www.lcm.tuwien.ac.at/R/, Computational and Structural Biotechnology Journal 5 (2013) e201302007.

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

[4] T. Hastie, R.J. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, Springer, New York, NY, 2001.

[5] C. Mercader, E.A. Castro, Partial-order ranking and linear modeling: Their use in predictive QSAR/QSPR studies, in: M. Dehmer, K. Varmuza, D. Bonchev (Eds.), Statistical modelling of molecular descriptors in QSAR/QSPR, Wiley-VCH, Weinheim, Germany, 2012, pp. 149-174.

[6] A. Miller, Subset selection in regression, Chapman & Hall (CRC), Boca Raton, FL, USA, 2002.

[7] R, A language and environment for statistical computing, R Development Core Team, Foundation for Statistical Computing, www.r-project.org, Vienna, Austria, 2013.

[8] M.L. Lee, D.L. Vassilaros, C.M. White, M. Novotny, Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons, Anal. Chem. 51 (1979) 768-773.

[9] B. Liebmann, A. Friedl, K. Varmuza, Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics, Anal. Chim. Acta 642 (2009) 171-178.

[10] A. Friedl, E. Padouvas, H. Rotter, K. Varmuza, Prediction of heating values of biomass fuel from elemental composition, Anal. Chim. Acta 544 (2005) 191-198.