# A New Variable Selection Method Based on All Subsets Regression

Andreas Liebminger [1], Leonhard Seyfang [2]
Peter Filzmoser [2], Kurt Varmuza [1]*

[1] Institute of Chemical Engineering
[2] Institute of Statistics and Probability Theory
**Vienna University of Technology, Austria**

* Presenting author
**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

**Poster Presentation:**
**10th SSC 2007, Scandinavian Symposium on Chemometrics**
**11 - 15 June 2007, Lappeenranta, Finland**



## Introduction

An ideal variable selection method for regression models would find one or more subsets of variables which have optimum prediction performance.

Usually,

- not prediction performance is optimized during variable selection;
- no exhaustive test of all possible variable subsets is possible;
- empirical variable selection methods have to be applied that are not optimal.

Consequently, the prediction performance of regression models - obtained from different variable subsets - has to be estimated separately.

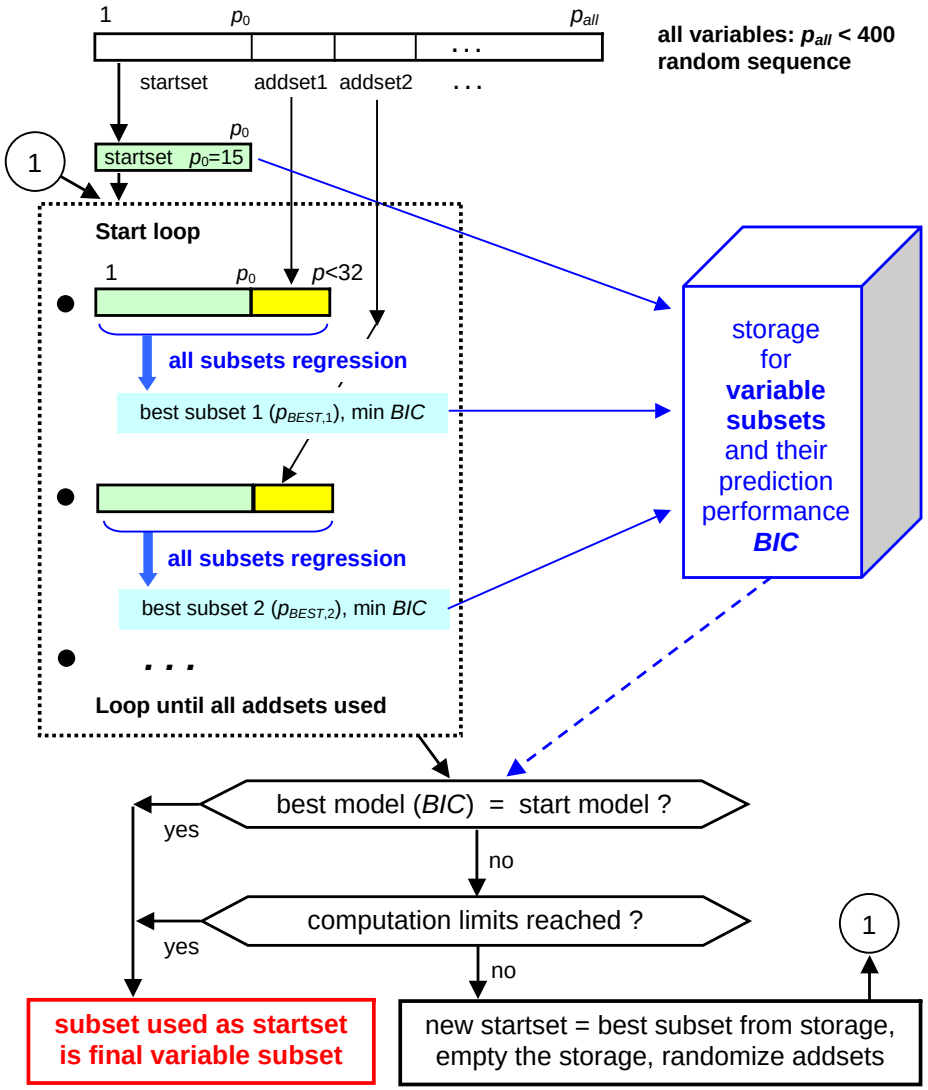
**This study**

- presents the new variable selection method FASS, by combining forward selection with fast all subsets regression;
- compares FASS with other variable selection methods (for instance a genetic algorithm);
- applies a "repeated double cross validation" for estimating the prediction performance of PLS regression models.

# Variable Selection Method FASS

## Forward selection combined with All SubSets Regression

Typical parameters (software in R) [1,2]






# Strategy and Data

## All subsets regression (FASS)

Exhaustive treatment of all variable subsets up to 31 variables. Function "regsubsets" in package "leaps" in R [2]; typ. computation time 2 s per run, 1 - 60 minutes in a FASS application. Regression method OLS; performance criterion *BIC* (Bayesian information criterion, Schwarz criterion, *SIC*), similar to Akaike criterion [3, 11],

$BIC = \ln(RSS/n) + k \ln(n) / n$ for normally distributed residuals

*n*, no. of objects; *p*, no. of variables + 1 (for intercept); *RSS*, sum of squared prediction errors. *BIC* penalizes a large number of variables.

## Genetic algorithm (GA)

Software MobyDigs [4]. Regression method: OLS; performance criterion (fitness): adjusted squared correlation coefficient, $_{ADJ}R^2$, between $y$ and $\hat{y}$ for full cross validation [3]. Maximum number of selected variables is 15, typical computation time 30 - 120 minutes.

## Prediction performance of PLS models

Repeated double cross validation. The data set is randomly partitioned into *s* (typ. 4) segments. A calibration set consists of *s* -1 segments, the remaining segment is a test set. A PLS model is derived from the calibration set (cross validation is used to estimate the optimum number of PLS components), and is applied to the test set, resulting in *n* / *s* predicted values $\hat{y}_i$. Systematic variation gives a $\hat{y}$ for each object. The whole process is repeated *k* times (typ. 10-100). Finally, *k.n* predicted values are available. From the prediction errors several performance criteria are derived, e.g.: $SEP_{TEST}$ (standard deviation), difference of 95% and 5% percentile (confidence interval), density distribution (for visual inspection). Typically, 10 variable subsets (from different selection methods) have been tested by this repeated double cross validation. New software in R [2, 5]; typ. comp. time 2 minutes.

Leave-one-out cross validation. For reference, $SEP_{CV}$ has been determined for full cross validation using all data (Unscrambler [6]).

# Results

## Data sets

**OXY**: $n$ = 180, $p$ = 57. Concentration change of isotope $^{18}O$ in precipitation ($y$) modeled by meteorological and geographical variables [7].

**PAC**: $n$ = 209, $p$ = 467. GC-retention indices ($y$) of polycyclic aromatic compounds [8], modeled by molecular descriptors (Dragon [9]).

**TOX**: $n$ = 846, $p$ = 681. GC-retention indices ($y$) of compounds relevant in forensics [10], modeled by molecular descriptors (Dragon [9]).

| Dataset | $p$ | Variable selection | *SEP TEST* | *SEP CV* |
|---|---|---|---|---|
| **OXY** | 57 | no | **1.01** | 0.90 |
| $y$ = (-16.5)-(-5.5) | 11 | FASS | **0.83** | 0.74 |
| | 15 | GA | **0.84** | 0.79 |
| | 13 | stepwise | **1.09** | 0.77 |
| **PAC** | 467 | no | **11.0** | 7.3 |
| $y$ = 197-504 | 27 | FASS | **5.2** | 5.0 |
| | 15 | GA | **7.2** | 6.7 |
| | 22 | stepwise | **18.7** | 5.3 |
| **TOX** | 681 | no | **108** | 80 |
| $y$ = 1110-3870 | 31 | FASS | **82** | 80 |
| | 15 | GA | **100** | 97 |
| | 33 | stepwise | **205** | 74 |

$n$, number of objects; $p$, number of variables;
*SEP*, standard deviation of prediction errors (standard error of prediction);
*TEST*, test sets in repeated double cross validation (10 repetitions);
*CV*, leave-one-out cross validation



# Results

| Dataset | $n$ | $p$ | **Computing time per job [minutes]**[1] | |
|---|---|---|---|---|
| | | | **FASS**[2] | **GA**[3] |
| OXY | 180 | 57 | 0.2 | 26 |
| PAC | 209 | 467 | 7 | 40 |
| TOX | 846 | 681 | 180 | 120 |

[1] PC processor AMD Athlon 2.2 GHz; [2] Until no further improvement obtained;
[3] Termination after 200000 iterations

- Variable selection by FASS or GA improved prediction performance; stepwise selection was not successful.
- FASS results are similar to GA results or better.
- Advantages of FASS are: less computation time, selection of up to 31 variables (GA in used software allows only 15), more strictly defined algorithm.
- Simple leave-one-out cross validation can be very misleading. A careful estimation of prediction performance is necessary for evaluation of variable selection methods.

**References**

[1] Liebminger, A.: PhD Dissertation. Vienna University of Technology, Austria, 2006.
[2] Software R, 2.2.0. R Development Core Team, www.r-project.org, 2005.
[3] Frank, I. E., Todeschini, R.: The data analysis handbook. Elsevier, Amsterdam, 1994.
[4] Software MobyDigs, 1.0. Talete srl, www.talete.mi.it, Milan, Italy, 2004.
[5] Mevik, B.H., Wehrens, R.: J. Statistical Software 18 (2007) issue 2.
[6] Software The Unscrambler, 9.0. Camo Process AS, www.camo.no, Oslo, Norway, 2004.
[7] Liebminger, A., Papesch, W., Haberhauer, G., Varmuza, K.: Chemom. Intell. Lab. Syst. (2007), doi: 10.1016/j.chemolab.2007.04.005.
[8] Lee, M.L., Vassilaros, D.L., White, C.M., Novotny M.: Anal. Chem. 51 (1979) 768.
[9] Software Dragon, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004).
[10] Garkani-Nejad, Z., Karlovits, M., Demuth, W., Stimpfl, T., Vycudilik, W., Jalali-Heravi, M., Varmuza, K.: J. Chromatogr. A 1028 (2004) 287.
[11] Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time series analysis: Forecasting and control, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 1994.