Exploration of the variability of variable selection based on distances between bootstrap sample results

Christian Hennig[1] Willi Sauerbrei[2]

[1]University College London

[2]Universität Freiburg

1 The problem

Regression variable selection can be very unstable.
Different models may yield very similar fits;
an ambiguous dataset
may allow multiple quite different fits.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i,$$

$i = i, \ldots, n,\ e_i \sim \mathcal{N}(0, \sigma^2)$ iid.
Variable selection: choose $V \subseteq \{1, \ldots, p\}$:
$j \notin V \Leftrightarrow \beta_j = 0.$

Variable selection can be useful,
but it can also be problematic
and is easily misinterpreted.

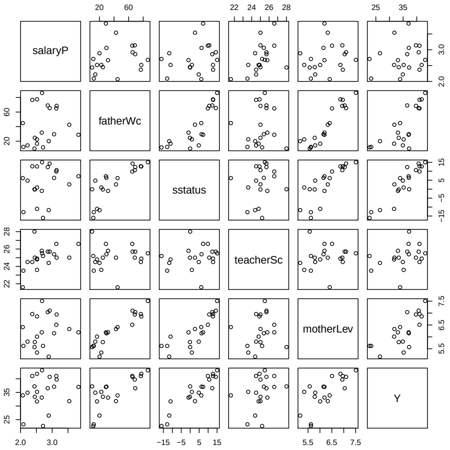
Exploring its stability and a variety of models
gives a more comprehensive picture of
how the variables "collaborate".

Here: Use LS linear regression,
backward selection with AIC or BIC stopping criterion.

Here: Use LS linear regression,
backward selection with AIC or BIC stopping criterion.

But our techniques are much more general,
could use with GLMs, robust regression, Lasso, forward selection, trees and forests…

Dataset 1: (Coleman et al. 1966)
Data on $n = 20$ schools,
y: verbal mean test score,
x_1: staff salary per pupil,
x_2: percentage of white collar fathers,
x_3: socioeconomic status composition indicator,
x_4: mean teacher's verbal test score,
x_5: mean mother's educational level.


salaryP
fatherWc
sstatus
teacherSc
motherLev
Y

Dataset 2
Study on ozone effects on school childrens lung growth,
$n = 496$ children, $p = 24$.
Ihorst et al. (2004), Buchholz et al. (2008).

Sauerbrei et al. (2015) investigate
stability of variable selection
using nonparametric bootstrap.

Response: FFVC - forced vital capacity (l) in autumn 1997
Explanatory variables:
ALTER age (years) at 1996-03-01
ADHEU allergic rhinitis diagnosed by physician
SEX 0male, 1female
HOCHOZON patient lives in a village with high ozone values
AMATOP maternal atopy (asthma, allergic rhinitis, eczema)
AVATOP paternal atopy (asthma, allergic rhinitis, eczema)
ADEKZ eczema diagnosed by physician
ARAUCH Tobacco smoke exposure at home (no/yes)
AGEBGEW weight (g) at birth
FSNIGHT cough at night or in the morning
FLGROSS height (cm) at pulmonary function testing
FMILB sensitization to dust mite allergens
FNOH24 maximal NO2 value of last 24h before pulmonary function testing (μg/m3)
FTIER sensitization to animal (dog and cat) danders
FPOLL sensitization to pollens (hazel, birch, grass)
FLTOTMED total number of medications at pulmonary function testing
FO3H24 max. O3 value of last 24h before pulmonary function testing (μg/m3)
FSPT sensitization to any of pollens, dog and cat danders or dust mites
FTEH24 max. temperature of last 24h before pulmonary function testing (Cel.)
FSATEM shortness of breath
FSAUGE itchy or watery eyes
FLGEW weight (kg) at pulmonary function testing
FSPFEI wheezing or whistling in the chest
FSHLAUF cough following exercise

2 Ingredients

Analysis uses B bootstrap models (selected variables) $V_1, \ldots, V_B$.

Schools data: $B = 500$ finds 17 models (backward/AIC).
Ozone data: $B = 500$ each backward/AIC and backward/BIC finds 798 models.

2.1 Distances

Use distance-based methods:
Multidimensional scaling, cluster analysis.

Distances between models

(a) Variable-based distance (Kulczynski 1927)

$$d_V(V_1, V_2) = 1 - \left(\frac{|V_1 \cap V_2|}{2|V_1|} + \frac{|V_1 \cap V_2|}{2|V_2|} \right)$$

Can also apply as distance between variables according to presence in models.

Distances between models

(a) Variable-based distance (Kulczynski 1927)

$$d_V(V_1, V_2) = 1 - \left(\frac{|V_1 \cap V_2|}{2|V_1|} + \frac{|V_1 \cap V_2|}{2|V_2|} \right)$$

Can also apply as distance between variables according to presence in models.
...but more relevant how models treat points.

Distances between models

(a) Variable-based distance (Kulczynski 1927)

$$d_V(V_1, V_2) = 1 - \left(\frac{|V_1 \cap V_2|}{2|V_1|} + \frac{|V_1 \cap V_2|}{2|V_2|} \right)$$

Can also apply as distance between variables according to presence in models.
…but more relevant how models treat points.

(b) Fit-based distance

$$d_F(V_1, V_2) = \sum_{i=1}^{n} |f_{V_1}(\mathbf{x}_i) - f_{V_2}(\mathbf{x}_i)|$$

(Manhattan-distance gives every fit difference same weight.)

2.2 Multidimensional Scaling

Kruskal's (1964) nonmetric MDS maps distances on Euclidean space with distances $\hat{d}$, optimising

$$\text{Stress} = \sqrt{\frac{\sum_{i,j}[f(d(z_i, z_j)) - \hat{d}_{ij}]}{\sum_{i,j} \hat{d}_{ij}}},$$

f monotonic transformation.

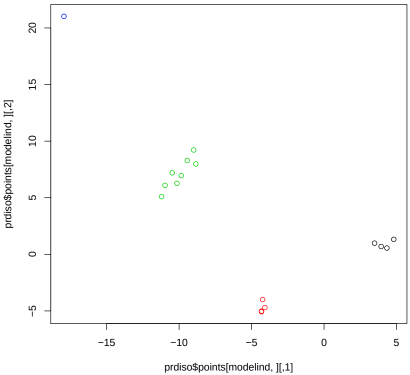
2.3 Suitable clustering methods

E.g., hierarchical (single, complete, average linkage).
Use average linkage (AL) here.
SL allows large within-cluster distances too easily,
CL too often divides what is not separated.

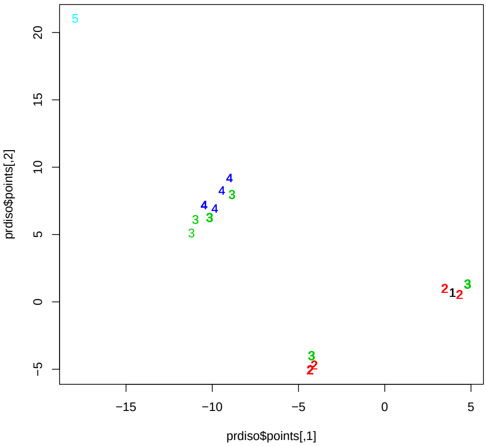
3 Data analysis

3.1 Schools data

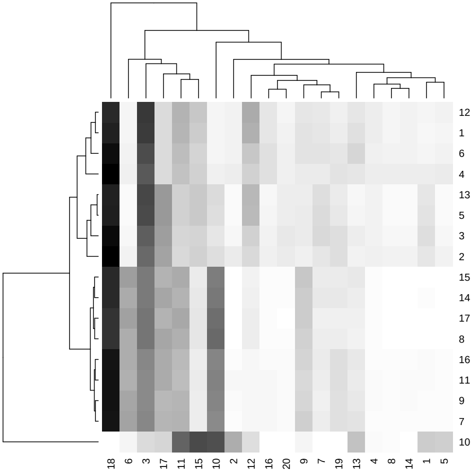
MDS on fit distance (with clustering)



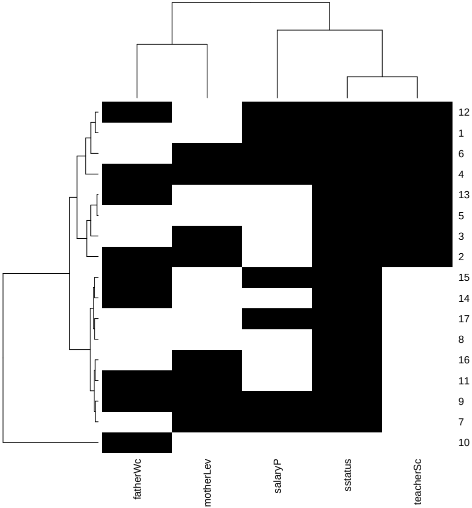
Where are best models?



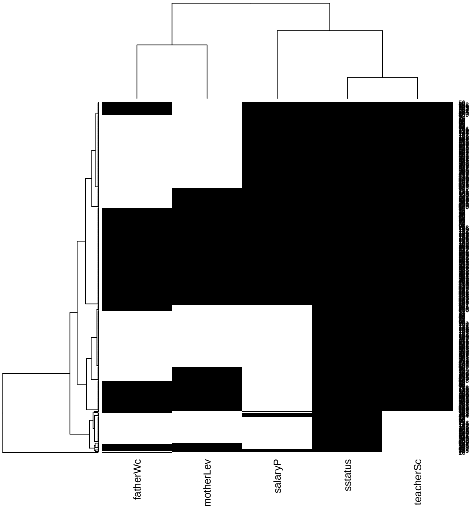
Models and squared residuals



Models and variables

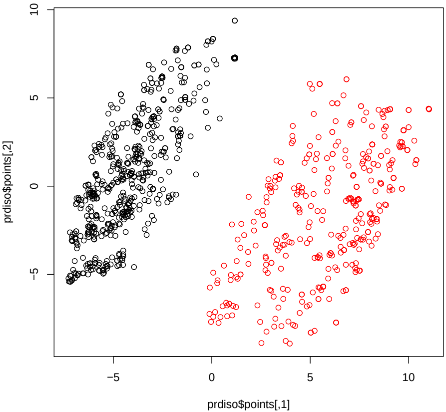


Models and variables (by bootstrap run)

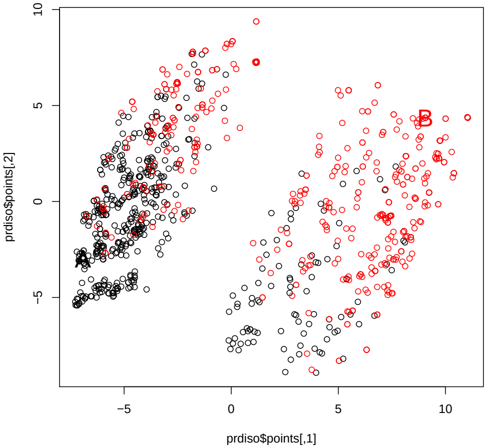


3.2 Ozone data

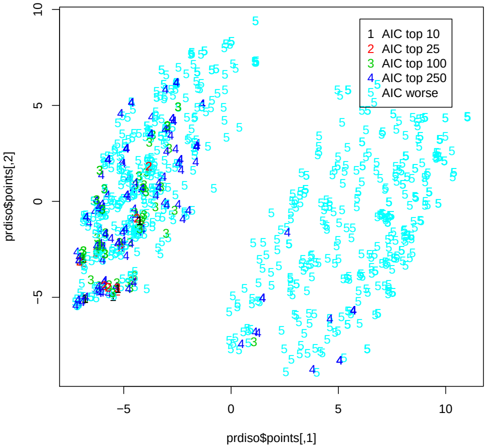
MDS on fit distance with clusters



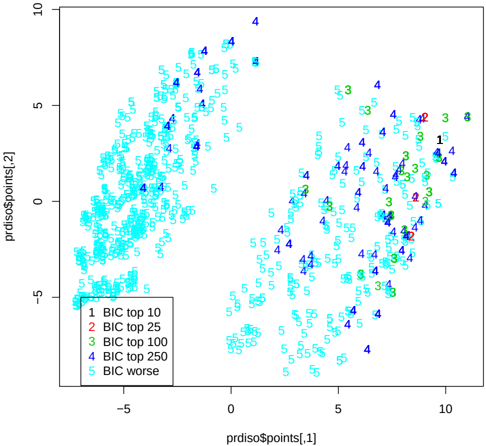
Models found by AIC, BIC



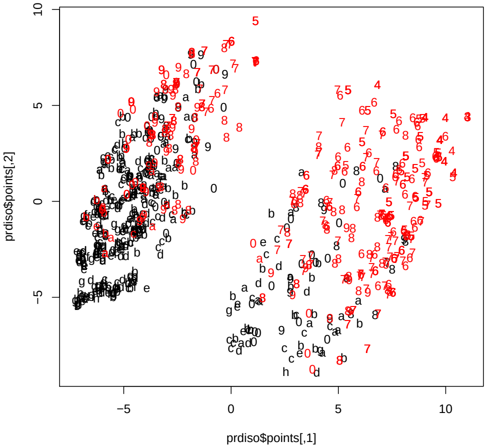
Best AIC


1 AIC top 10
2 AIC top 25
3 AIC top 100
4 AIC top 250
5 AIC worse
prdiso$points[,1]
prdiso$points[,2]
−5
0
5
10

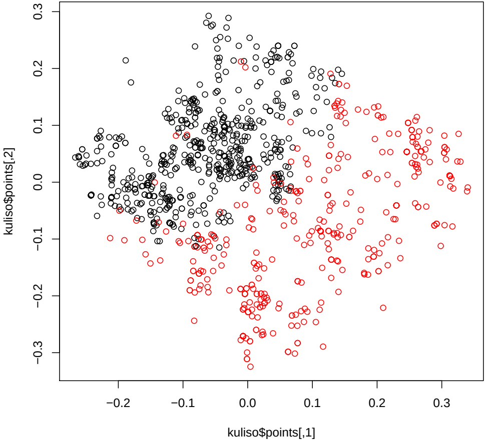
Best BIC


1 BIC top 10
2 BIC top 25
3 BIC top 100
4 BIC top 250
5 BIC worse
prdiso$points[,1]
prdiso$points[,2]
−5
0
5
10
−5
0
5
10

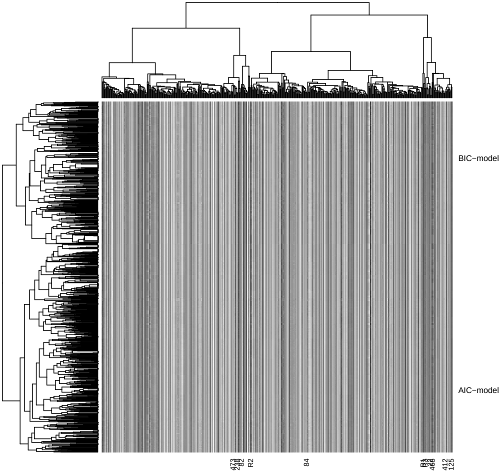
Size of models



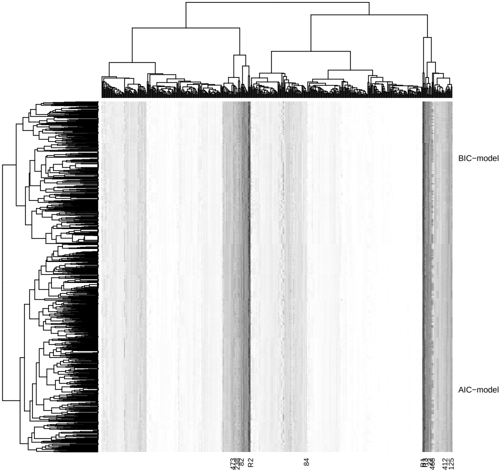
Variable-based distance (with clusters from before)



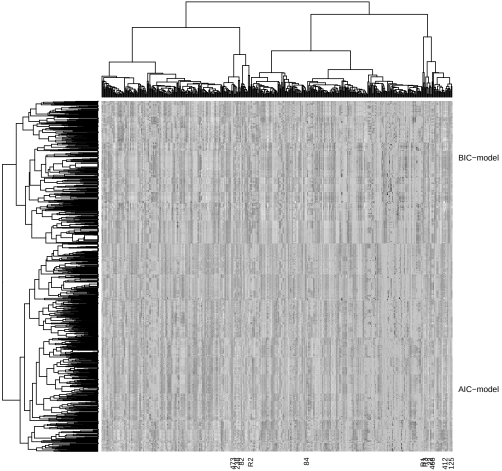
Models and fits



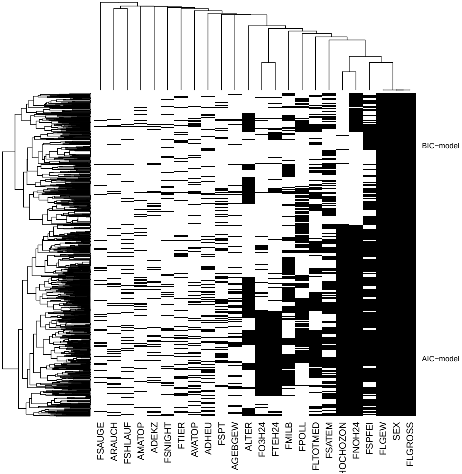
Models and squared residuals



Models and squared residuals (column standardised)



Models and variables



Variables that make a difference can clearly be seen.

4 What we learnt

Large variability in models for both datasets.

Schools data:

- Four clusters of models deliver quite different fits.
- Some models fit some ($\sim$ half) points very well, disregarding others.
- Better AIC achieved by “compromise fits” (including TeacherSc variable).

Ozone data

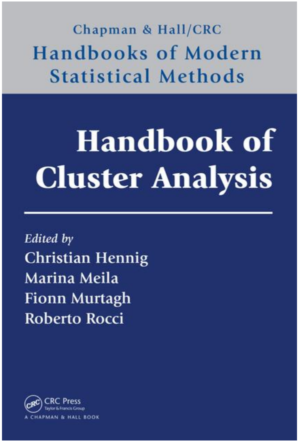
- Two clusters of model fits,
  not aligned with BIC/AIC-models,
  rather connected to vars HOCHOZON, FNOH24 and FN3H24.
- BIC- and AIC-selected models are quite different.
- Little variation between model fits and residuals,
  choice between them somewhat arbitrary.

Ozone data

- Two clusters of model fits,
  not aligned with BIC/AIC-models,
  rather connected to vars HOCHOZON, FNOH24 and FN3H24.
- BIC- and AIC-selected models are quite different.
- Little variation between model fits and residuals,
  choice between them somewhat arbitrary.

Not shown: atypicality of models,
and observations supporting atypical models.

A bit of marketing:



This work is supported by EPSRC Grant EP/K033972/1.