Übung: Stability Selection

Fabian Scheipl

Schauen Sie sich auf jeden Fall zunächst das Video zu "Stability Selection" an. Benutzen Sie als Ausgangsbasis für die folgenden Aufgaben den Code in `get_stability_paths.R`.

Die `refit`-Funktion können Sie hier zunächst mal als "black box" betrachten.

Aufgabe 1: ***Bootstrap-Sampling***

Schreiben Sie die fehlenden Funktionen

```
sample_without_replacement <- function(nrows, strata = NULL, fraction = 0.5) {
  # ??
}
get_selected <- function(new_model) {
  # ??
}
make_paths <- function(selected) {
  # ??
}
```

`get_selected` sollte für ein gegebenes Modell eine Matrix mit (max. Subsetgröße+1)×(Anz. Kovariablen) zurückgeben, `make_paths` sollte für eine Liste solcher Matrizen eine Matrix die die *stability paths* enthält zurückgeben. Die erste Zeile der Matrizen sollte (Selektionshäufigkeiten für) ein Model ohne Kovariablen repräsentieren.

Überprüfen Sie Ihren Code mit folgenden Test:

```
library(ElemStatLearn)
library(MASS)
library(leaps)

data(prostate)
data <- prostate

max_formula <- lpsa ~ (.  - train)
model <- regsubsets(max_formula, data=data, nbest = 1, nvmax = 8,
  really.big = TRUE)

set.seed(20141020)
stability_paths <- get_stability_paths(model, data, reps=10)
stability_paths
```

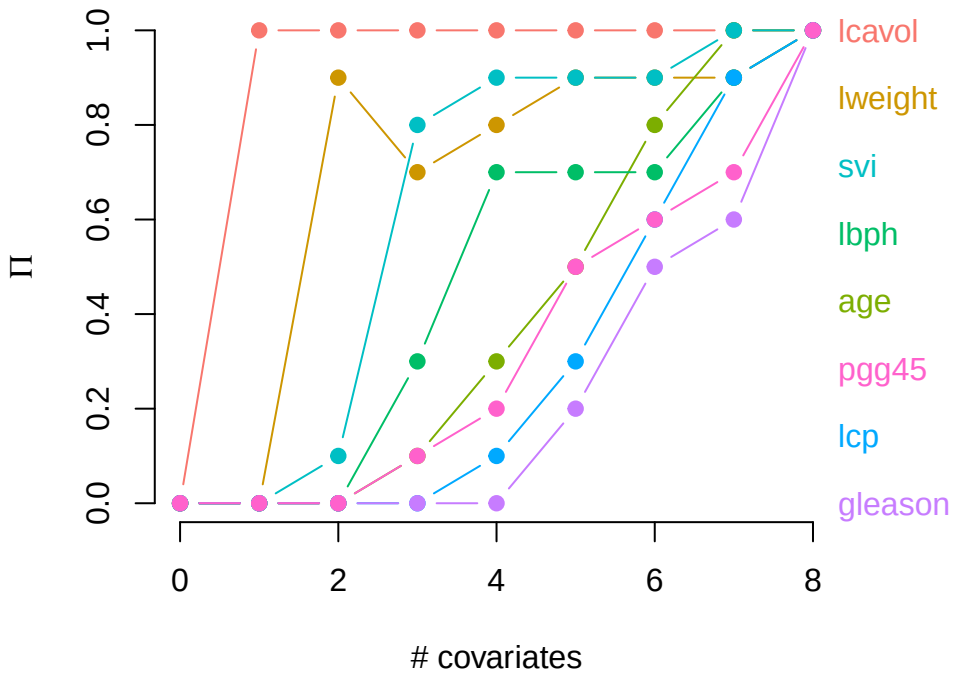
```
##   lcavol lweight age lbph svi lcp gleason pgg45
##        0     0.0 0.0  0.0 0.0 0.0     0.0   0.0
## 1      1     0.0 0.0  0.0 0.0 0.0     0.0   0.0
## 2      1     0.9 0.0  0.0 0.1 0.0     0.0   0.0
## 3      1     0.7 0.1  0.3 0.8 0.0     0.0   0.1
```

```
## 4      1     0.8 0.3  0.7 0.9 0.1     0.0   0.2
## 5      1     0.9 0.5  0.7 0.9 0.3     0.2   0.5
## 6      1     0.9 0.8  0.7 0.9 0.6     0.5   0.6
## 7      1     0.9 1.0  0.9 1.0 0.9     0.6   0.7
## 8      1     1.0 1.0  1.0 1.0 1.0     1.0   1.0
```

Aufgabe 2: ***Grafik***

Schreiben Sie eine Funktion `plot_stability_paths`, die in etwa so etwas wie die untenstehende Grafik erzeugt.

```
plot_stability_paths(stability_paths)
```



Aufgabe 3: ***Fancy Grafik***

Modifizieren Sie ihre Plotfunktion so, dass Sie Grafik-Parameter sowohl für die Label am rechten Rand als auch für die Darstellung der Pfade übergeben können. Testen Sie ihre Funktion mit dem folgenden Beispielcode:

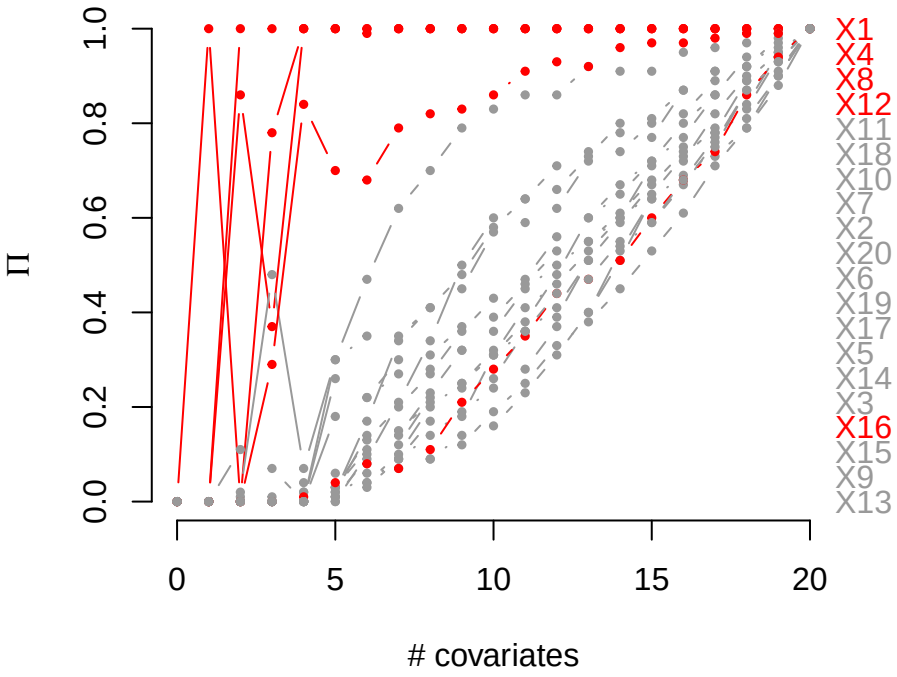
```
## generate a n x (p+1) dataset with p strongly collinear covariates,
## p_signal non-zero effects that are spread out evenly among the covariates and
## that decrease in size from p_signal to 1.
## covariates are generated with corr(x_i, x_j) = corr_x^|i-j|.
## response y = x %*% coef + error; error is N(0, sd_error)
simulate_collinear <- function(n = 200, p = n / 10, p_signal = 3, sd_error = 2,
```

```
    corr_x = .95) {
  library(mvtnorm)

  cov_x <- corr_x ^ abs(outer(1:p, 1:p, "-"))
  x <- rmvnorm(n, sigma = cov_x)

  true_coef <- rep(0, p)
  true_coef[seq(1, p, l = p_signal + 1)] <- p_signal : 0

  y <- x %*% true_coef + rnorm(n, sd = sd_error)
  structure(data.frame(y, x), true_coef=true_coef)
}

set.seed(1991788)
data <- simulate_collinear(p_signal = 5)
```

```
max_formula <- y ~ .
model <- regsubsets(max_formula, data=data, nbest = 1, nvmax = 20,
  really.big = TRUE)
set.seed(110101)
stability_paths <- get_stability_paths(model, data, reps=100)
head(stability_paths)
```

```
##   X1   X2   X3   X4   X5 X6   X7   X8   X9  X10  X11  X12  X13  X14  X15
##    0 0.00 0.00 0.00 0.00  0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
## 1  0 0.00 0.00 1.00 0.00  0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
## 2  1 0.00 0.00 0.00 0.00  0 0.11 0.86 0.01 0.00 0.02 0.00 0.00 0.00 0.00
## 3  1 0.00 0.00 0.78 0.00  0 0.07 0.37 0.00 0.01 0.48 0.29 0.00 0.00 0.00
## 4  1 0.00 0.00 1.00 0.00  0 0.00 1.00 0.00 0.00 0.07 0.84 0.02 0.04 0.02
## 5  1 0.03 0.03 1.00 0.02  0 0.01 1.00 0.00 0.02 0.30 0.70 0.01 0.06 0.03
##    X16  X17 X18  X19  X20
##   0.00 0.00 0.0 0.00 0.00
## 1 0.00 0.00 0.0 0.00 0.00
## 2 0.00 0.00 0.0 0.00 0.00
## 3 0.00 0.00 0.0 0.00 0.00
## 4 0.01 0.00 0.0 0.00 0.00
## 5 0.04 0.01 0.3 0.26 0.18
```

Plotten Sie die Pfade der Kovariablen deren Koeffizienten tatsächlich $\neq 0$ sind in rot, die anderen in grau, Ihre Grafik sollte in etwa so aussehen:


Π
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
covariates
X1
X4
X8
X12
X11
X18
X10
X7
X2
X20
X6
X19
X17
X5
X14
X3
X16
X15
X9
X13