Nonconvex penalties: Case studies

Patrick Breheny

March 23

Introduction

- In our lecture for today, we will revisit our two high-dimensional studies from the previous chapter / topic and analyze them with the reduced-bias approaches of this topic
- First, we consider an adaptive lasso model for the BRCA1 gene expression data
- As our initial estimator, let's use lasso estimates with λ chosen according to BIC:

```
fit <- ncvreg(X, y, penalty='lasso')
b <- coef(fit, which=which.min(BIC(fit)))[-1]
```

(using `ncvreg` for fitting due to its compatibility with `BIC`)

- Cross-validation would of course be a reasonable alternative

Adaptive lasso fit

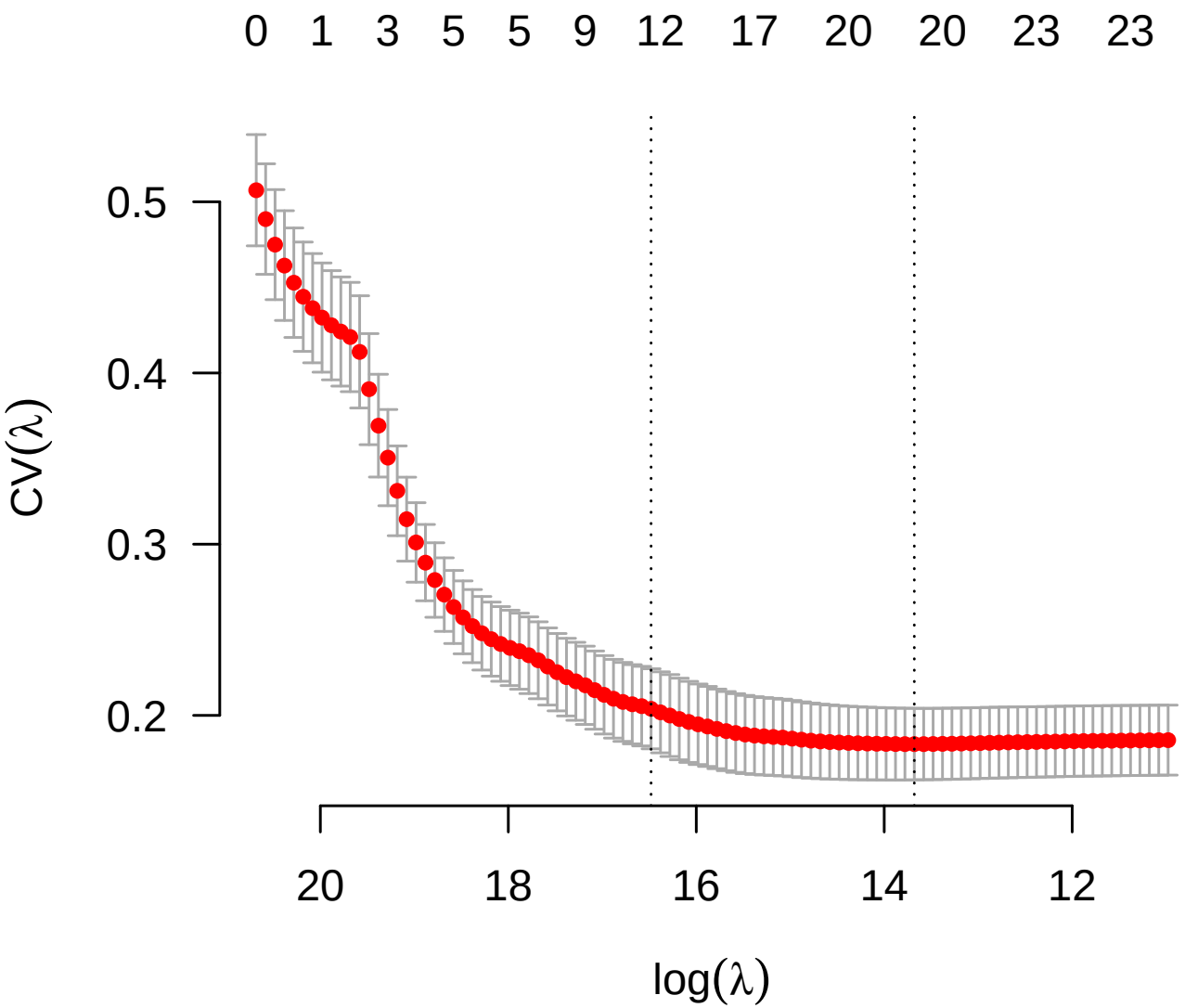
Once we have the initial estimator, we can fit an adaptive lasso model as follows:

```
w <- abs(b)^(-1)          # Calculate weights
w <- pmin(w, 1e10)        # cv.glmnet does not allow
                          # infinite weights
cvfit <- cv.glmnet(X, y, penalty.factor=w)
```

and plot the results as usual:

```
plot(cvfit)
```

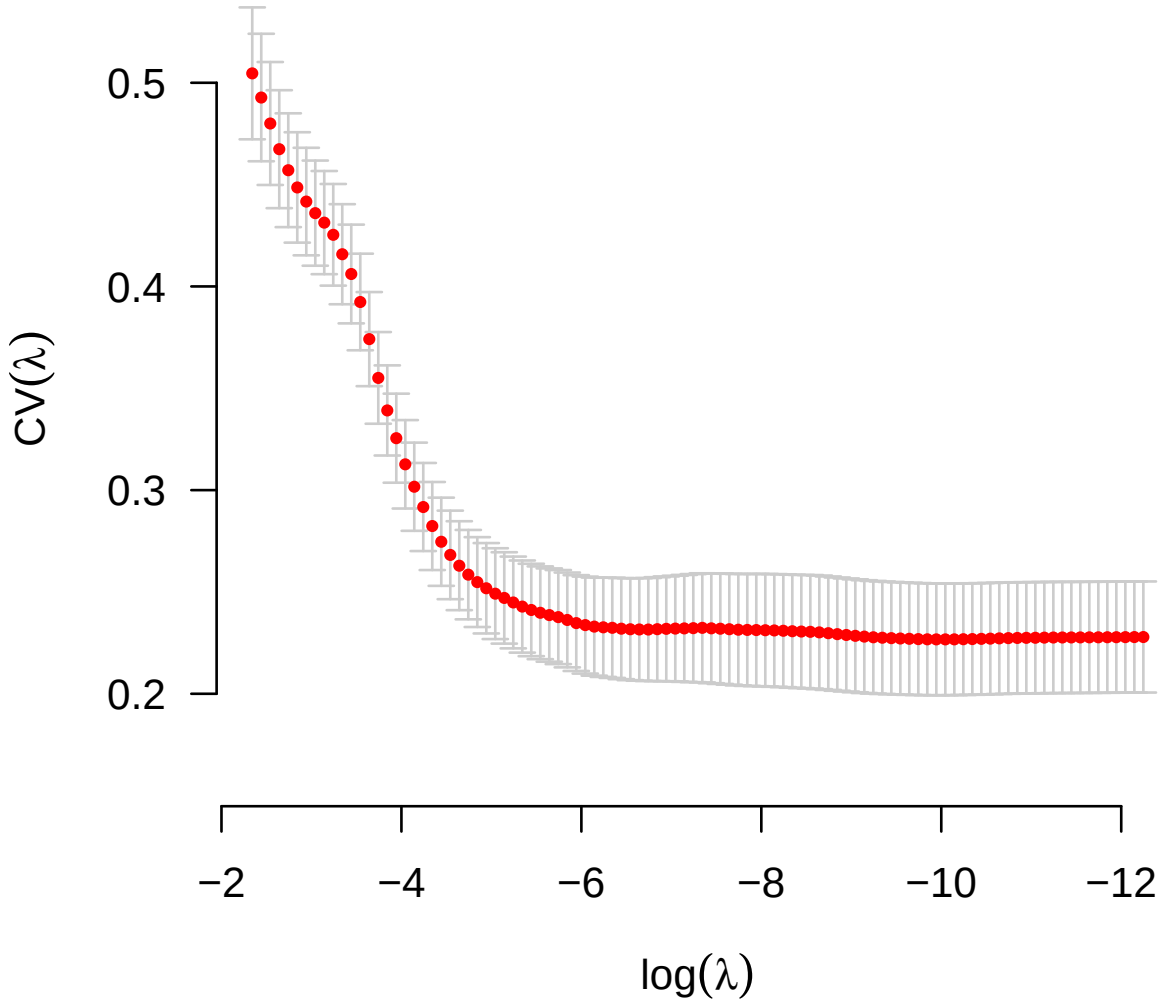
Adaptive lasso: Cross-validation (biased)



Source of bias

- In the figure, the CV error is not estimated in an unbiased manner
- The reason is that the left-out fold is not truly external to the fitting procedure, as it was used to obtain an initial estimator
- As a result, prediction error is underestimated
- To obtain an (approximately) unbiased estimate of CV error, one must cross-validate the entire procedure, including the initial estimate

Adaptive lasso: Cross-validation (unbiased)



Remarks

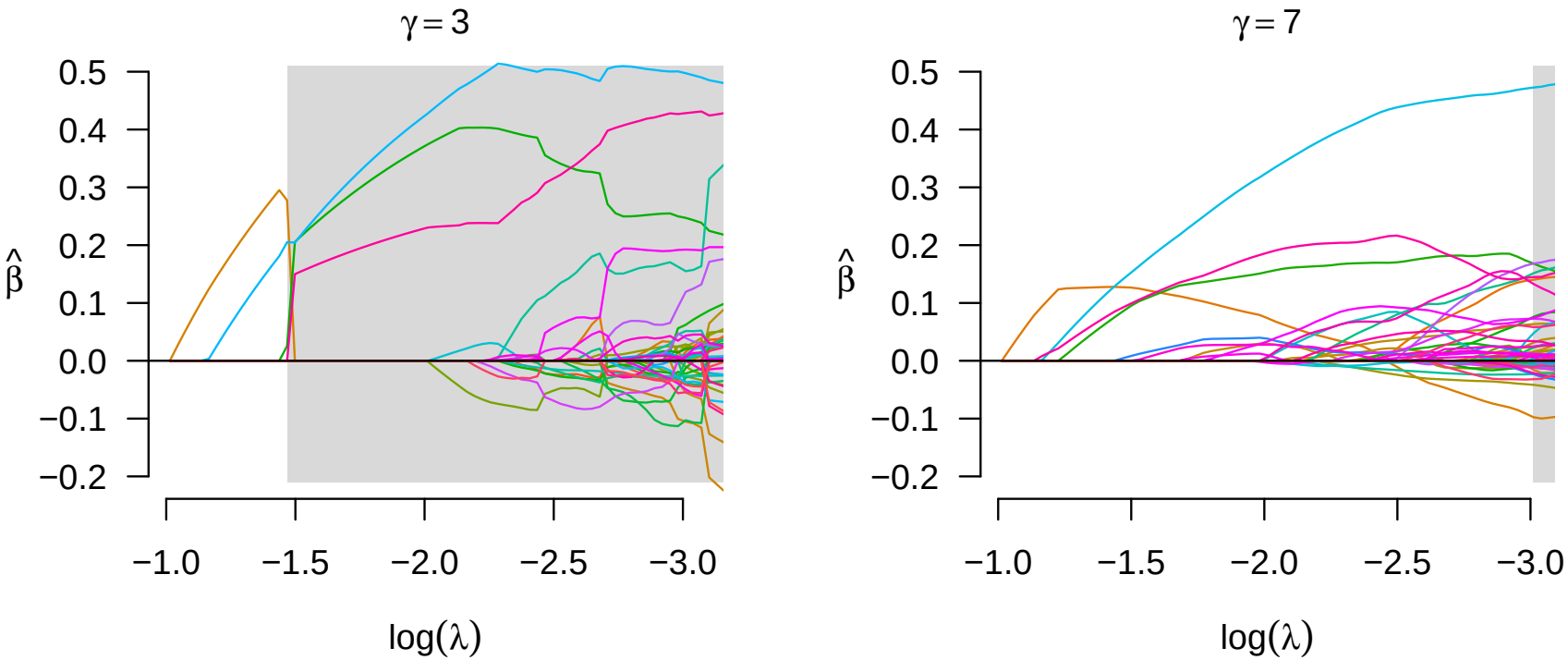
- This is an important cautionary example to keep in mind for the adaptive lasso: flexible, two-stage methods have certain advantages in terms of simplicity, but are also easy to make mistakes with
- Unfortunately, while existing `R` packages can be used to fit adaptive lasso models, there are not currently any comprehensive software packages for the adaptive lasso (that I am aware of) that carry out full cross-validation

MCP analysis

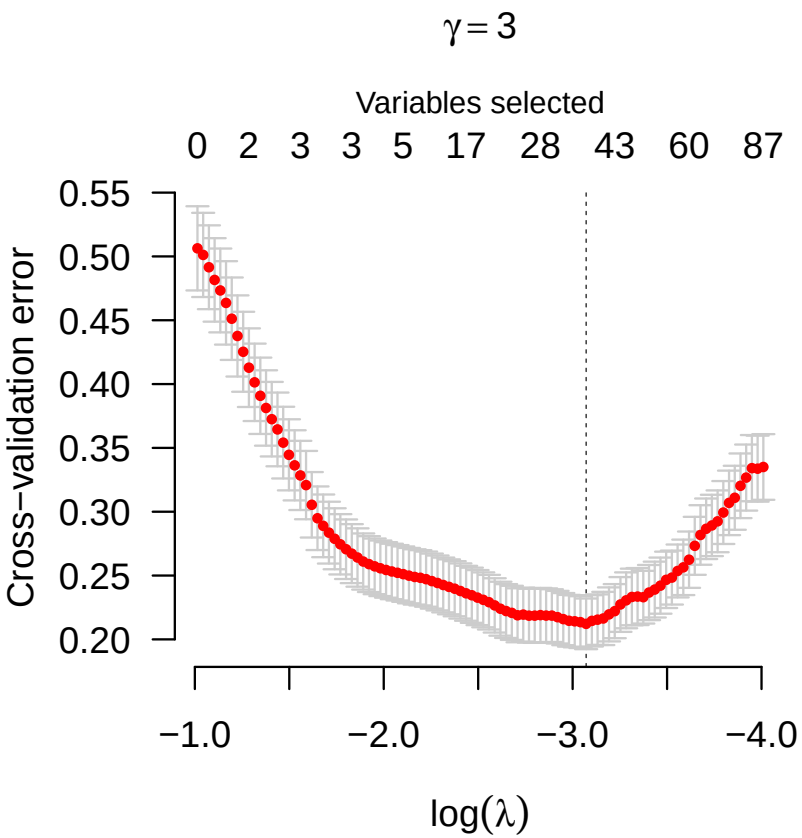
- MCP and SCAD achieve the adaptive lasso's goal of reducing the bias associated with the lasso, but do so in a single step and thus prove a bit more amenable to carrying out inference concerning predictive accuracy using cross-validation
- Let's fit two penalized regression models to the BRCA1 data, one with $\gamma = 3$ and the other with $\gamma = 7$:

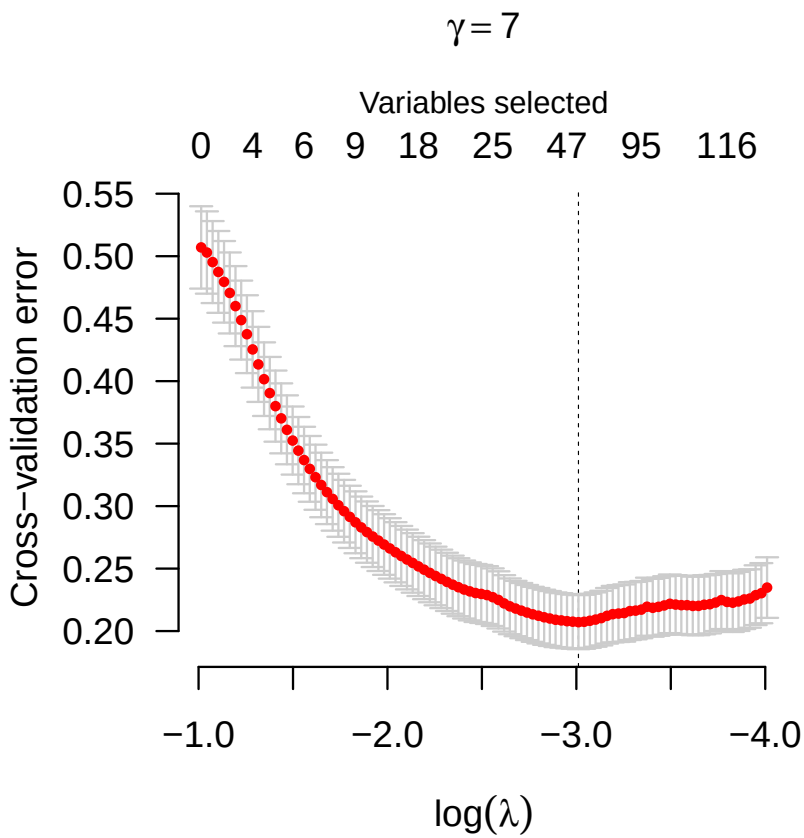
```
cvfit3 <- cv.ncvreg(X, y)           # gam=3 is default
cvfit7 <- cv.ncvreg(X, y, gamma=7)
```

Results: MCP



CV Results: MCP





summary

ncvreg provides a useful summary function for fitted CV objects:

```
> summary(cvfit3)
MCP-penalized linear regression with n=536, p=17322
At minimum cross-validation error (lambda=0.0464):
-------------------------------------------------
  Nonzero coefficients: 38
  Cross-validation error (deviance): 0.21
  R-squared: 0.58
  Signal-to-noise ratio: 1.39
  Scale estimate (sigma): 0.461
```

summary

And the equivalent summary for $\gamma = 7$:

```
> summary(cvfit7)
MCP-penalized linear regression with n=536, p=17322
At minimum cross-validation error (lambda=0.0492):
-----------------------------------------------
  Nonzero coefficients: 52
  Cross-validation error (deviance): 0.21
  R-squared: 0.59
  Signal-to-noise ratio: 1.45
  Scale estimate (sigma): 0.455
```

Remarks

- For both models, the minimum error is $\mathrm{CV} = 0.21$; very close to, although slightly larger than the $CV = 0.20$ achieved by the lasso
- However, the two models select very different numbers of variables, both compared to each other and compared to the lasso, which selected 96 nonzero coefficients
- The most striking difference between the two solution paths is that for MCP with $\gamma = 3$, the the optimal solution occurs in the region that is not locally convex
- As this is real data, we cannot know which estimates are more accurate, but personally, I would prefer the $\gamma = 7$ solution

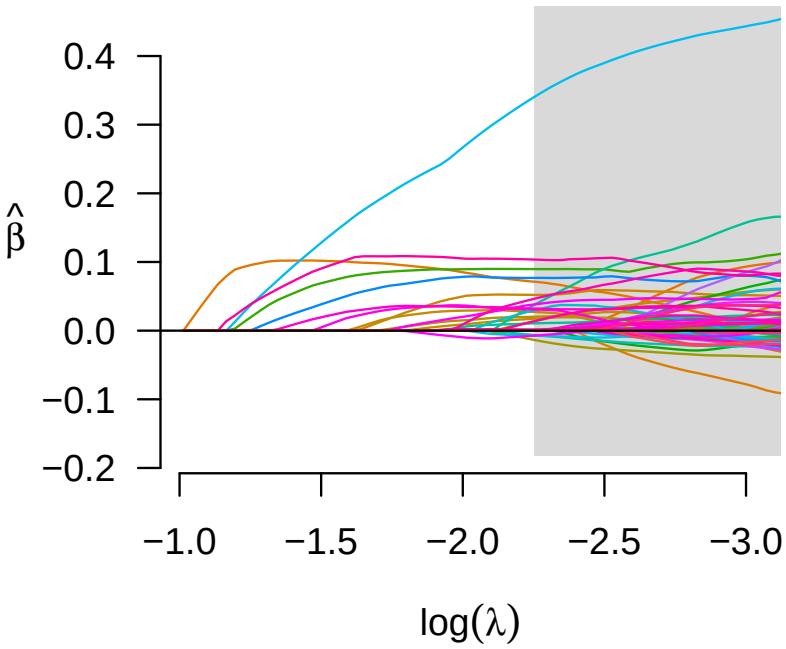
SCAD

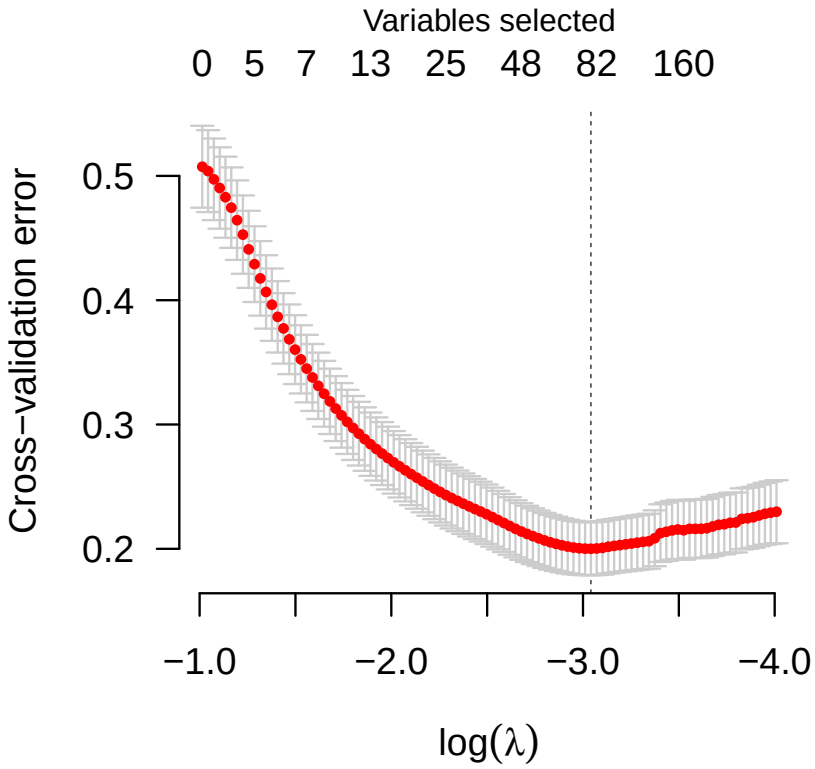
Finally, let us fit a SCAD-penalized regression model to this data; similar to the MCP case, we set $\gamma = 8$ here to increase the stability of the solution path:

```
> cvfit <- cv.ncvreg(X, y, gamma=8, penalty='SCAD')
> summary(cvfit)
SCAD-penalized linear regression with n=536, p=17322
At minimum cross-validation error (lambda=0.0478):
-------------------------------------------------
  Nonzero coefficients: 79
  Cross-validation error (deviance): 0.20
  R-squared: 0.61
  Signal-to-noise ratio: 1.53
  Scale estimate (sigma): 0.447
```

Results: SCAD ($\gamma = 8$)

The SCAD results are more lasso-like than that of the MCP models, as one would expect from the fact that the SCAD and lasso penalties are more similar





Remarks

- This is just one example, but these results seen are fairly representative, in my experience
- The prediction performance (as estimated by cross-validation) is typically similar between MCP/SCAD/lasso, but there can be substantial differences in terms of the estimates themselves
- The main advantage in practice of MCP (or SCAD) is the ability to achieve that prediction performance using fewer features
- Finally, the results of SCAD are almost always in between those of MCP and lasso

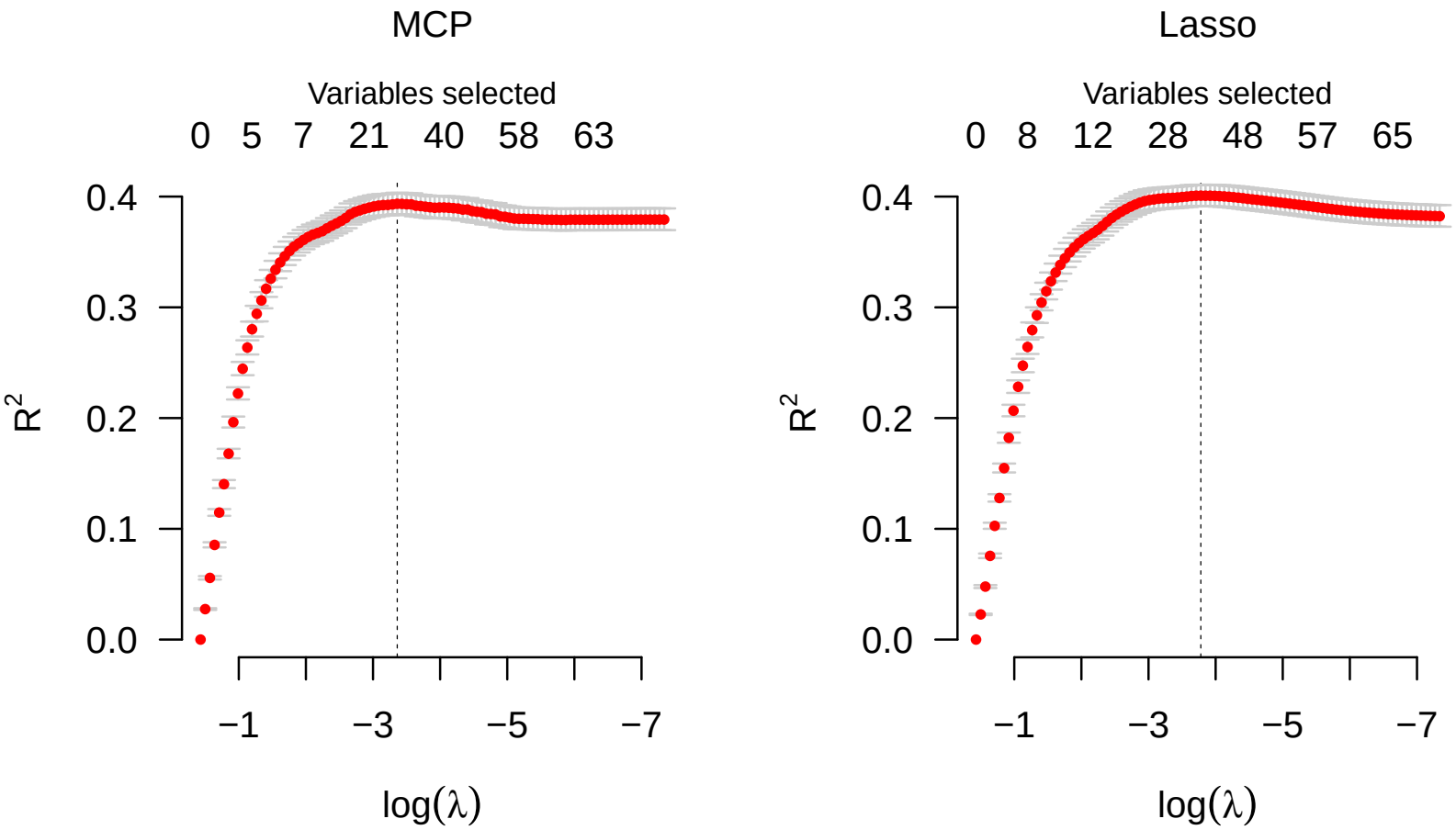
WHO-ARI: MCP

- Let us also revisit the WHO study of acute respiratory illness, which you have looked at a few times in your homework assignments
- Let us fit an MCP-penalized regression model to this data using $\gamma = 6$ and compare it to the fit of the lasso:

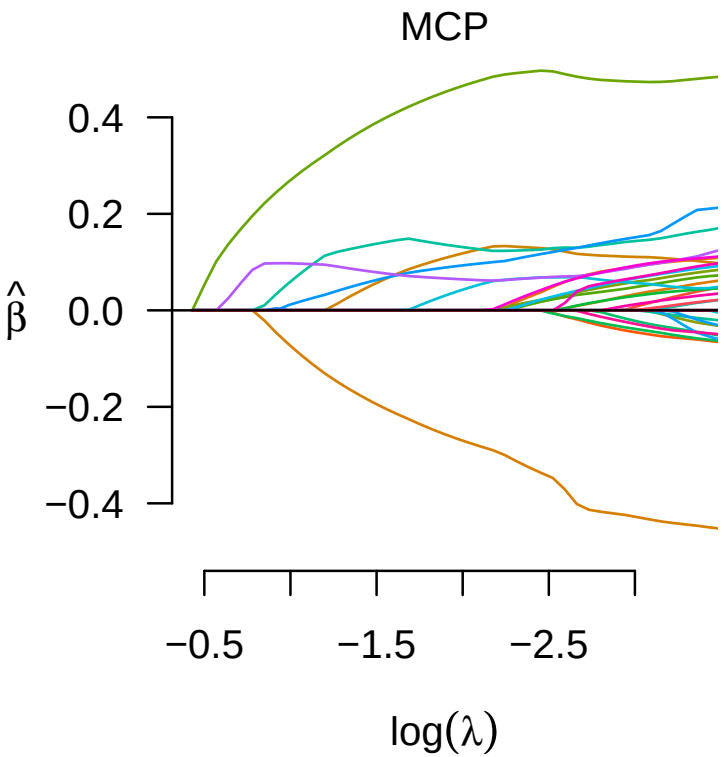
```
set.seed(2)
cvfit.mcp <- cv.ncvreg(XX, y, gam=6)
set.seed(2)
cvfit.las <- cv.ncvreg(XX, y, penalty="lasso")
```

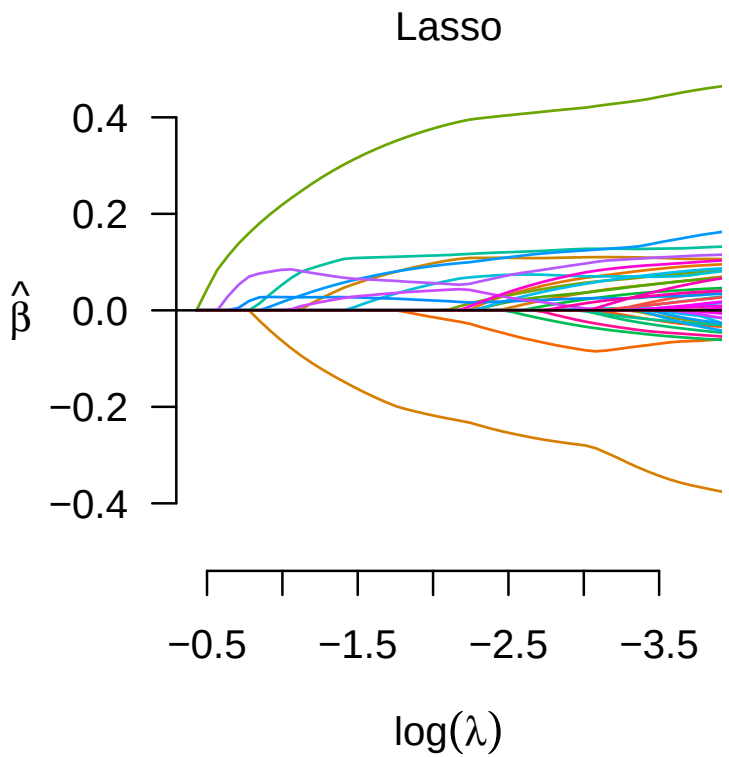
- In making these kinds of comparisons, it is important to keep the CV fold assignments the same across methods, otherwise you may mistake the effect of different folds for the effect of the penalty

Results: CV


MCP
Variables selected
0 5 7 21 40 58 63
0.4
0.3
0.2
0.1
0.0
R^2
−1 −3 −5 −7
log(λ)
Lasso
Variables selected
0 8 12 28 48 57 65
0.4
0.3
0.2
0.1
0.0
R^2
−1 −3 −5 −7
log(λ)

Results: Coefficient path





Summary: MCP

```
> summary(cvfit.mcp)
MCP-penalized linear regression with n=816, p=67
At minimum cross-validation error (lambda=0.0347):
-----------------------------------------------
  Nonzero coefficients: 27
  Cross-validation error (deviance): 1.23
  R-squared: 0.39
  Signal-to-noise ratio: 0.65
  Scale estimate (sigma): 1.111
```

Summary: Lasso

```
> summary(cvfit.las)
lasso-penalized linear regression with n=816, p=67
At minimum cross-validation error (lambda=0.0228):
-------------------------------------------------
  Nonzero coefficients: 39
  Cross-validation error (deviance): 1.22
  R-squared: 0.40
  Signal-to-noise ratio: 0.67
  Scale estimate (sigma): 1.104
```