MS-C1620 Statistical Inference

Exercise 11

Homework exercise

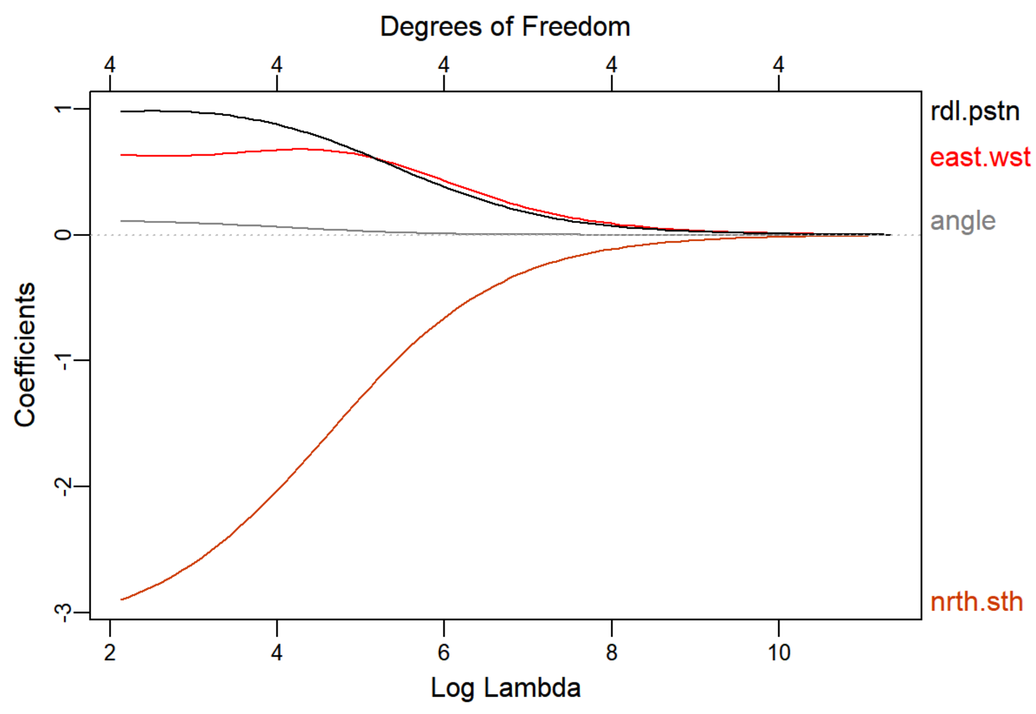
To be solved at home before the exercise session.

---

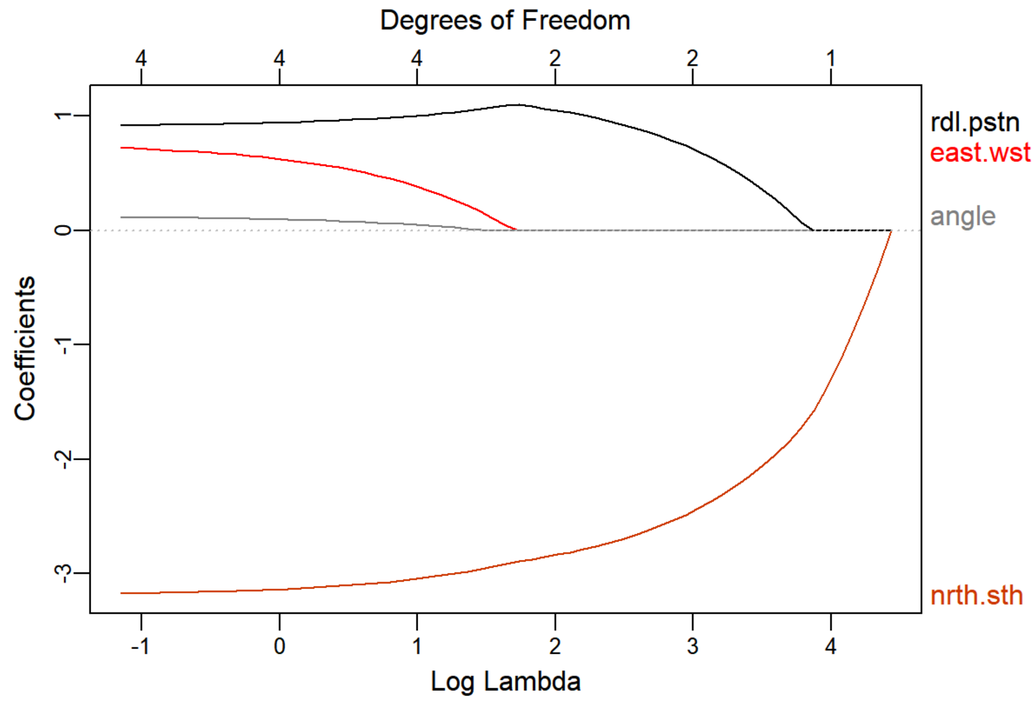
1. a. The data set `galaxy` from the package `ElemStatLearn` contains measurements on the position and radial velocity (the response) of the galaxy NGC7531. The following two plots show the ridge and LASSO coefficient profiles of the four explanatory variables. Compare the two plots and interpret them (for example, deduce which of the explanatory variables are the most ``important''?)

```
library(ElemStatLearn)
library(glmnet)
library(plotmo)

ridge_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 0)
plot_glmnet(ridge_galaxy, xvar = "lambda", label = TRUE)
```




```
lasso_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 1)
plot_glmnet(lasso_galaxy, xvar = "lambda", label = TRUE)
```



b. Consider a regression model where the response `Y` is explained using the covariates `X1`, `X2` and `X3`. The p-values corresponding to the models of all possible combinations of the covariates are listed below. Use them to perform variable selection with both backward and forward selection with the p-value cutoff $\alpha_0 = 0.05$.

```
##     X1
## 0.0071
```

```
##     X2
## 0.4221
```

```
##     X3
## 0.0014
```

```
##     X1     X2
## 0.0055 0.2809
```

```
##     X1     X3
## 0.0021 0.0004
```

```
##     X2     X3
## 0.1267 0.0006
```

```
##     X1     X2     X3
## 0.0010 0.0516 0.0001
```

Class exercise

To be solved at the exercise session. Note: the R-script of lecture 10 might prove helpful in solving the below problems.

---

1. The data set `barro` in the package `quantreg` contains the annual GDP growth rates of several countries along with several explanatory variables. Our objective is to use variable selection to determine which factors are most helpful in predicting the growth rate of GDP.
   a. Visualize the data set.
   b. Fit a standard multiple regression to the data with `y.net` as the response.
   c. Use the function `step` to perform both backward and forward variable selection using the AIC as the criterion.
   d. Do the results of the backward and forward selections agree?
   e. Which variables would you conclude to be the most important in predicting the growth rate of GDP?

---

2. We continue the analysis of problem 1.
   a. Fit a LASSO model to the `barro` data set.
   b. Plot the LASSO coefficient profiles. Which variable does LASSO hold the most important?
   c. Use 10-fold cross validation to choose a suitable value for the parameter λ. Which variables are included in the corresponding model?

---

3. **(Optional)** Investigate how ridge regression and LASSO perform in the presence of “noise” variables. That is, simulate data where the response depends linearly on a few explanatory variables but include in the model also several explanatory variables (noise) which are independent of the response. Plot then the ridge and LASSO profiles of the variables. Do the profiles of the noise variables perform as expected?
