Presenting: Group 11, Group 12, Date: 23.06.2021

Exercise 1:

Consider the following data generating process with $n = 100$ observations and p covariates:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$$

Initially set the number of predictors $p = 2$ and assume $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\Sigma}$ is the covariance matrix with the variances on the diagonal and small values on the off-diagonal (all values chosen by you) and $\boldsymbol{\mu} = \begin{pmatrix} 0 & 10 \end{pmatrix}$. The true coefficients $\boldsymbol{\beta}$ range from 0.1 to 2 (you can sample values from that range or use equispaced values on that interval, for example, for $p = 2$ you will have $\boldsymbol{\beta} = (0.1 \quad 2)^\top$) and the errors are drawn from a normal distribution $\varepsilon \sim \mathcal{N}(0, 1)$.

The aim of this exercise: compare OLS, ridge regression, lasso and PCR.

a) Write a function to calculate the principal component scores as shown in the lecture. Visualize the principal components along with the original observations.

b) Perform PCR using one and two principal components. Comment on your results.

c) Calculate the prediction error (test MSE) for simple OLS ridge regression, lasso and PCR using one and two principal components.

Exercise 2 (Simulation Study)**:**

a) Evaluate the difference in prediction performance of the four methods for $p = 10$ in a simulation study, choose the number of principal components using k-fold cross validation with $k = 5$ or $k = 10$.

b) Propose a manipulation of the dgp that would make the lasso and ridge regression perform worse than PCR. Try to investigate the reason for the inferior performance by evaluating model selection for the lasso.