Exercise 1:

Consider the following data generating process with $n = 100$ observations and p covariates. Initially set the number of predictors $p = 2$ and $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is the covariance matrix with the variance on the diagonal and small values on the off-diagonal (both values chosen by you). $\boldsymbol{\mu} = \begin{pmatrix} 0 & 10 \end{pmatrix}$. The (initially) true coefficients range from $\boldsymbol{\beta} = 0.1 - 0$ (you can sample values from that range or use equispaced values on that interval) and the errors are drawn from a normal distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)$.

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

a) Calculate the principal component scores as shown in the lecture. Visualize the principal components along with the original observations.

b) Perform PCR using and and both principal components.

c) Calculate the prediction error for 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

b) Propose at least two manipulations of the dgp that would make the lasso perform worse than PCR and ridge regression.