# PYU44C01 Linear Algebra Assignment 2 2021

You are provided with four data sets for a set of 29 EU and EFTA countries from https://ec.europa.eu/eurostat which are: hours worked per week (hours), employment rates (%), Labour mobility (000's of workers working outside their country of citizenship by citizenship) and working population (000's of workers).

(1) Prove that $\mathbf{A}_1 = \sigma_1 \mathbf{u_1} \mathbf{v_1}^\mathbf{T}$ is a rank 1 approximation to **A** where $\mathbf{u_1}$ and $\mathbf{v_1}$ are columns of the $\mathbf{Q_1}$ and $\mathbf{Q_2}$ matrices in the SVD of **A**.

(2) Write a Python script which reads the four data sets provided into a 4 x 29 array **A** and plots them *in the order given in the hours worked data set*. Note that order varies by data set. You will find the arrays in ***script.py*** useful for this.

(3) Modify the data so that each row of **A** has zero mean.

(4) Find a diagonal matrix **D** which scales the rows of **A** so that each row has unit magnitude.

(5) Form the correlation matrix $\mathbf{C} = (\mathbf{D} \mathbin{.} \mathbf{A}) \mathbin{.} (\mathbf{D} \mathbin{.} \mathbf{A})^\mathbf{T}$

(6) Perform a principal component analysis of the data sets using the correlation matrix, i.e. project the data points in **D . A** onto the two most important principal axes of the correlation matrix.

(7) Put on your economist hat and discuss the meaning of the principal component analysis. You may find the *Nature* article on Blackboard on genetic variation throughout Europe as well as material covered in the lecture on principal component analysis useful for this.