## Homework 1 (60p)

Instructions:

- submit a single PDF file with your answers, without any code, and indicating all the members in the group
- submit a separate R.script or other software script (or ZIP) file with the code that you used to obtain your answers, making sure that your script can be ran all at once.
- Note that longer answers do not lead to a better grade, but clearer answers do.

# Question 1 (35p) : PCA and factor analysis

The file co2emission.csv contains the annual CO2 emissions per capita of each country for several years. This data is taken from `https://ourworldindata.org/grapher/co-emissions-per-capita` where the data sources are also explained.

**Instructions. For plots, make sure:**

- lines in the plot are thick enough and visible
- legend is present in large enough font
- the x and the y axis values are displayed with a large enough font
- both the x and the y axes are named, and the name is displayed in a large enough font
- the plot includes a title axis that is large enough
- plot is saved as .pdf or .eps, and so can be stretched (do not save it as .png or .jpeg, those get distorted when their size is modified)
- *points will be subtracted if the figure does not look good enough, even if it is correct.*

**a) Data pre-processing:**

- load data in table format
- select the following countries (mostly EU and some other big countries) and drop the rest from the table:
  "Australia","Austria","Belgium","Bulgaria","China","Croatia","Czechia",
  "Denmark", "Estonia", "Finland","France", "Germany", "Greece", "Hungary",
  "India", "Indonesia", "Ireland","Italy","Latvia","Lithuania","Netherlands",
  "Poland", "Portugal","Romania", "Russia", "Slovakia", "Slovenia", "Spain",
  "Sweden", "Switzerland", "Ukraine", "United Kingdom", "United States";
- select years 1907-2022 and drop the rest
- find countries that still have missing observations, and drop them
- you should now have a balanced panel of 29 countries over 116 years.

The goal of the rest of the analysis is to 1) check if CO2 emissions decreased over time; 2) find common trends in CO2 emissions among countries (do they follow a climate goal of reducing pollution?).

b) Plot CO2 emissions over time for all countries, in one plot. What time trend(s) do you see in the data? Do all countries follow the same trend? Which country has a pattern that does not resemble the others? Remove that country from the data and rest of the analysis.

c) Replot the data, and find a real world explanation for the simultaneous dip in the emissions for most countries, followed by a simultaneous pick-up.

d) If we were to perform PCA, what is $n$ (sample size) and what is $p$ (number of variables) in this dataset?

e) Explain intuitively why $n$ and $p$ should both be large to recover a true factor structure, up to a rotation, should that structure describe the true data generating process. Is $n$ and $p$ sufficiently large here?

f) Demean the data, but do not standardize it. Perform PCA and plot the first two loadings for each country on a plot. Do you observe a pattern? Are some countries loading more on the first PC or the second?

*Proceed from now onward with standardized data.*

g) Perform PCA and plot pairwise the first three loadings for each country (three plots, loadings on PC1 and PC2, first plot, loadings of PC2 and PC3 second plot, loadings of PC1 and PC3, third plot). Do you observe a pattern in some or all of the plots? Explain if you do, and possible reasons for it.

*Set the maximum number of PC to 10 from here onward.*

h) Pick the number of PCs with the following two criteria in Bai and Ng (2002, Econometrica), page 201 [paper attached to the Assignment]: $PC_{p1}$, and $IC_{p1}$. Write down the correct mathematical formulae (defining everything).

i) Do a screeplot of the cumulative variance explained by PCA with each new PC, and pick the smallest number of PCs that explain at least 90% of the variation in the dataset. Explain now which of criteria from h) and i) do you trust and why.

j) Remove UK from the analysis and redo h) and i). What do you notice compared to when UK was kept in?

k) Plot the first two PCs over time together with 95% confidence bands based on the normal distribution and the standard errors of the PCs, interpret their evolution over time, and what are possible reasons for this evolution. *(Hint: You may need to Google climate accords for reasons. Note that while the true factors are estimated only up to a rotation matrix, the rotation would be the same for all points in time, so the shape of the plot can be well interpreted, but the quantitative differences between the first and the second PC not.)*

# Question 2 (25p): Clustering

For this question, you will need to verify all the conditions in the following WLLN:

**Theorem 1.** *Let $\{h_i\}_{i=1}^n$ be $p \times 1$ independent random variables with finite means $\mu_i$. If $\sup_i E|h_{ik}|^2 < \infty$ for all $k = 1, \ldots, p$, where $h_{ik}$ are elements of $h_i$, and $\frac{1}{n}\sum_{i=1}^n \mu_i \to \mu$, a finite vector of constants, then*

$$\frac{1}{n}\sum_{i=1}^{n} h_i \xrightarrow{p} \mu$$

Assume:

**Assumption 1.** *The true data generating process is such that $Z_i^{(j)} \sim i.i.d.(\mu_j, \Sigma)$, where $j = 1, 2$ (two clusters, with $\mu_1 \neq \mu_2$ finite means), $i = 1, \ldots, n_j$ are the number of observations in each class, $Z_i^{(j)}$ is $p \times 1$, and $\Sigma$ is a common $p \times p$ variance matrix, which is a positive definite matrix of constants.*

Let $n = n_1 + n_2$.

a) Assume we know to which cluster each observation belongs to. Using Assumption 1, prove that if, for $j = 1, 2$, $n_j \to \infty$, then $\hat{\mu}_j = n_j^{-1}\sum_{i=1}^{n_j} Z_i^{(j)} \xrightarrow{p} \mu_j$. For this, you need to verify each of the conditions stated in Theorem 1 above.

b) If we did not know the cluster membership for any observation, and we assumed there is only one cluster, then we would estimate the mean of that cluster by $\hat{\mu} = n^{-1}\left(\sum_{i=1}^{n_1} Z_i^{(1)} + \sum_{i=1}^{n_2} Z_i^{(2)}\right)$. Using a), show that if $n_j \to \infty$, for $j = 1, 2$ and $\frac{n_1}{n_2} \to 0$, then $\hat{\mu} \xrightarrow{p} \mu_2$.

c) The same holds as in b), but now $n_j \to \infty$, for $j = 1, 2$, and $\frac{n_1}{n} \to p$, where $0 < p < 1$. Show that in this case, $\hat{\mu} \xrightarrow{p} p\mu_1 + (1-p)\mu_2$.

d) Assume you mistakenly classified the first $c_1$ observations in the first cluster to the second cluster, and the first $c_2$ observations in the second cluster to the first cluster. Assume $c_1, c_2$ are constants. As a result, you estimate the mean in the first cluster as $\hat{\mu}_1 = (n_1 - c1 + c_2)^{-1}\left(\sum_{i=c_1+1}^{n_1} Z_i^{(1)} + \sum_{i=1}^{c_2} Z_i^{(2)}\right)$. Prove that $\hat{\mu}_1 \xrightarrow{p} \mu_1$ as $n_1 \to \infty$, that is, these misclassified observations do not affect the consistency of your estimate for the first cluster mean.

*Hint: for showing that the missclassified part of $\hat{\mu}_1$ does not matter asymptotically, call it $h_i$, a $p \times 1$ vector with elements $h_{ik}$, you may use any of the following inequalities. You don't have to use them all, pick the ones which you need.*

- Markov inequality: For each $\eta > 0$, $P[|h_{ik}| > \eta] \leq E|h_{ik}|$
- Triangle inequality: $E|\sum_i h_{ik}| \leq \sum_i E|h_{ik}|$
- Jensen inequality: $E|h_{ik}| \leq [E(h_{ik})^2]^{1/2}$
- Chebyshev inequality: For each $\eta > 0$, $P[|h_{ik}| > \eta] \leq \dfrac{E|h_{ik}|^2}{\eta^2}$.