Toulouse School of Economics

Worksheet 1 – Principal Components Analysis

High-Dimensional Data Analysis and Machine Learning

Camille Mondon

January 5, 2026

1 PCA on US Crime Data

Exercise 1. In this exercise, we explore the **US crime data** with Principal Component Analysis (PCA). The goal is to understand data preprocessing, standardization, and interpretation of PCA results.

1. Load the data `crime.csv`[1] into a data frame `crimeus` using `read.csv`, with row names given by the first column.
2. Plot `Population` vs `Index` as a scatter plot.
   - Use `plot(..., type = "n")` first and then `text(...)` to label the points.
   - What patterns or clusters do you notice?
3. Perform a PCA on columns 5 to 12 (the crime variables) using `PCA`.
4. Examine the eigenvalues: why might some of them be exactly 0?
5. Comment on why the results may not be very informative.
6. What does this bit of code do?

```r
crimerate <- crimeus
crimerate[2:12] <- round(crimeus[2:12] * 100000 / crimeus$Population)
```

7. Perform PCA on `Population` and the `crimerate` data (columns 1 and 5 to 12).
   - Why is it necessary to scale the data before PCA?
   - Plot a scree plot and the individuals on the first two components.
8. Perform PCA on the `crimerate` data excluding `Inmates` (columns 5 to 11).
   - Plot a scree plot and the individuals on the first two components.
     - What differences do you observe with the previous question?
9. Examine the contributions of variables to the first two components. Which variables drive the main directions of variability?
10. Write an interpretation answering the following questions:
    - Why does PCA on raw counts fail to give meaningful results?
    - Why do we transform the data to **rates per population** before PCA?
    - How does scaling (standardization) affect PCA results?
    - Which regions are the most extreme according to the first principal component?

[1] You can download it here.