Advanced Computational Econometrics
Chapter 6: Gaussian mixture models

1 Clustering of countries

The dataset `countries_of_the_world.csv` contains the following information about 227 countries :

1. Region
2. Population
3. Area (sq. mi.)
4. Pop. Density (per sq. mi.)
5. Coastline (coast/area ratio)
6. Net migration
7. Infant mortality (per 1000 births)
8. GDP ($ per capita)
9. Literacy (%)
10. Phones (per 1000)
11. Arable (%)
12. Crops (%)
13. Other (%)
14. Climate
15. Birthrate
16. Deathrate
17. Agriculture
18. Industry
19. Service

In this analysis we consider only the economic and social variables, i.e., variables 6-10 and 15-19.

1. Perform a PCA on this data set. Keep the principal components that explain 90% of the variability.
2. Cluster the data using function `Mclust` in package `mclust`. select the model (inclusing the number of clusters) using the BIC criterion.
3. Try to interpret the clusters.

2 Program effectiveness data

The file `TableF14-1.csv` contains a data set of 32 observations collected to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses (Greene, 2008). Variables in the data set include

— `GPA` : the student's grade point average,
— `PSI` : dummy variable for whether the individual participated in the PSI,
— `TUCE` : the student's score on a pretest in economics.

1. Using function `regmixEM` from package `mixtools`, model the relation between the dependent variable `GPA` and the predictor `PSI` using a mixture of two linear regressions.
2. Explore the relation between the latent class variable Z and variable `TUCE` and explain the results.
3. We now assume that Z is related to `TUCE` by the following equation :

$$P(Z = 1 \mid \texttt{TUCE}) = \frac{\exp(\alpha_1 + \alpha_2 \texttt{TUCE})}{1 + \exp(\alpha_1 + \alpha_2 \texttt{TUCE})}.$$

Fit this model to the data using function `flexmix` in package `flexmix`.