

Exercise 18

Consider again the data set `agriculture` in the package `cluster`.

- (a) Perform an agglomerative hierarchical clustering (of the countries) with the function `hclust()`. Use the Euclidean distance, and perform the clustering based on Single-, Complete- as well as Average-Linkage. Compare the results of the different linkage methods (e.g. based on Dendrograms).
- (b) Choose a clustering from part (a). Plot the (abbreviated) country names into a coordinate system (see Exercise Sheet 2). Denote the (chosen) clusters by different colors.
- (c) Perform an agglomerative hierarchical clustering (of the countries) with the function `agnes()` (in the package `cluster`). Are there any differences between the results when using `agnes()` and `hclust()`?
- (d) Perform a divisive hierarchical clustering (of the countries) with the function `diana()` (in the package `cluster`). Compare with the result in (c) (and/or (a)).

Exercise 19

Consider again the data set `jet.txt` introduced in Exercise 15.

- (a) Use K-means based on the variables `SPR`, `RGF`, `PLF` and `SLF` to cluster the jets into 2, 3, or 4 groups, respectively. Look at the sum of within-group variances across all clusters. Visualize the clusters by appropriate scatter plots of the variables, where the individual clusters should be denoted by different colors. Which clustering seems most appropriate here?
- (b) Now visualize the clusters by projecting the (original) data into the space spanned by the first two principal components, i.e. draw a scatter plot of the scores on the first two PCs (see Exercise 15), and indicate cluster membership in an appropriate way in the plot.
- (c) Compare your results from (a) with those obtained by an agglomerative hierarchical clustering or by a mixture distribution approach.

Exercise 20

Consider the data set `AutoBi` in the package `insuranceData`.

- (a) Load the data set into R and delete the variables `CASENUM` and `MARITAL` from the data frame.
- (b) There is an extreme outlier in the data set. Try to identify it and delete the outlier observation from the data set.
- (c) Perform a cluster analysis with a method of your choice, and make an attempt to interpret the resulting cluster.

Exercise 21

In the following the heights (in m) of 10 persons are given.

1.56 1.71 1.54 1.66 1.58 1.90 1.85 2.05 1.72 1.67

Perform a hierarchical clustering approach (of your choice) as well as a K-means approach (with $K = 2$) with Paper and Pencil, and compare your results to those obtained with R.