STATS 769

Clustering

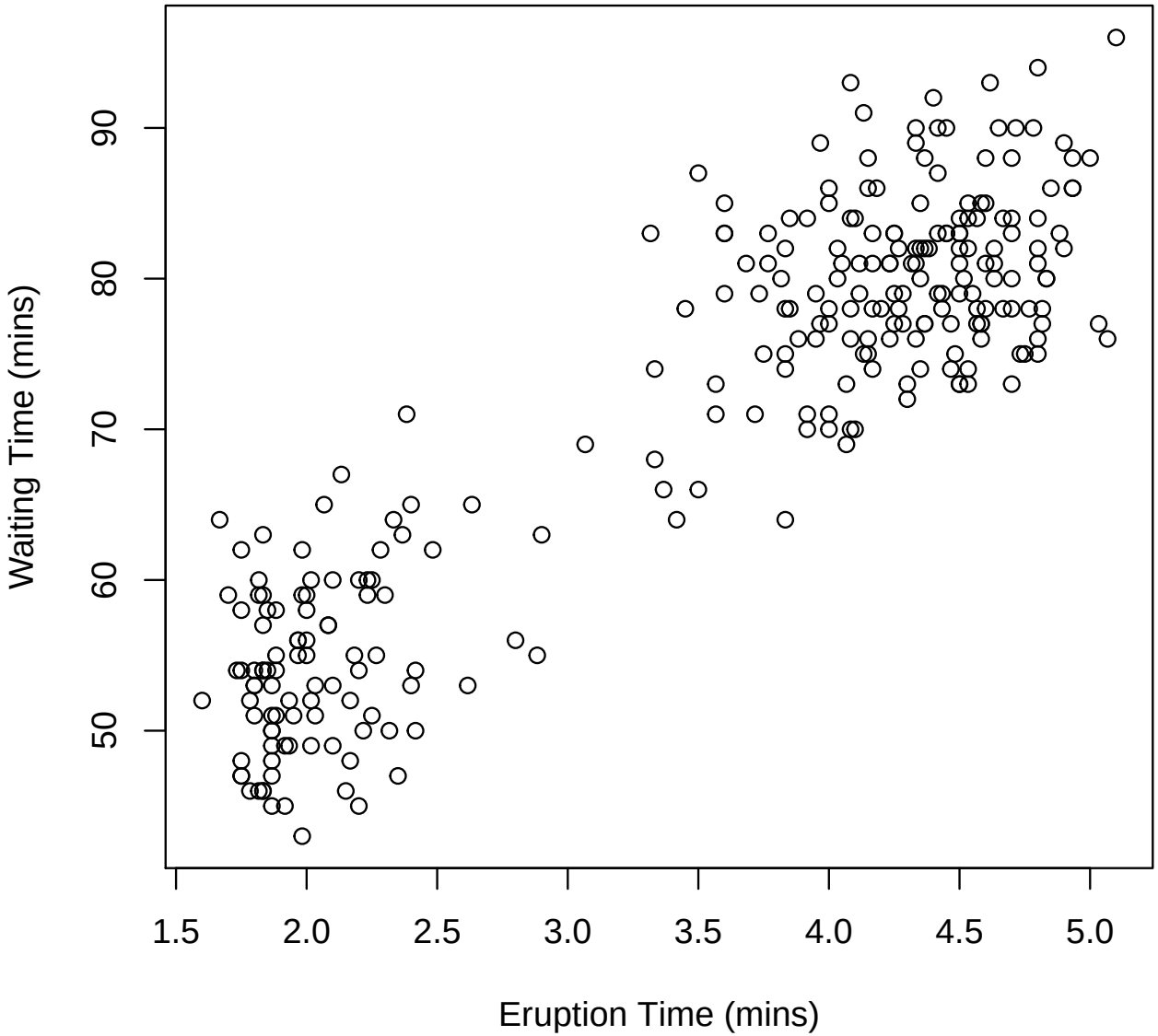
Yong Wang
Department of Statistics
The University of Auckland

2021

Clustering

- Clustering is to partition observations into groups/clusters.
- It is unsupervised learning, as there is no response variable.
- Many applications in practice
  - Taxonomy: evolution, language, library
  - Business and economics: corporations, stocks, customers
- What is a cluster?

Old Faithful Geyser



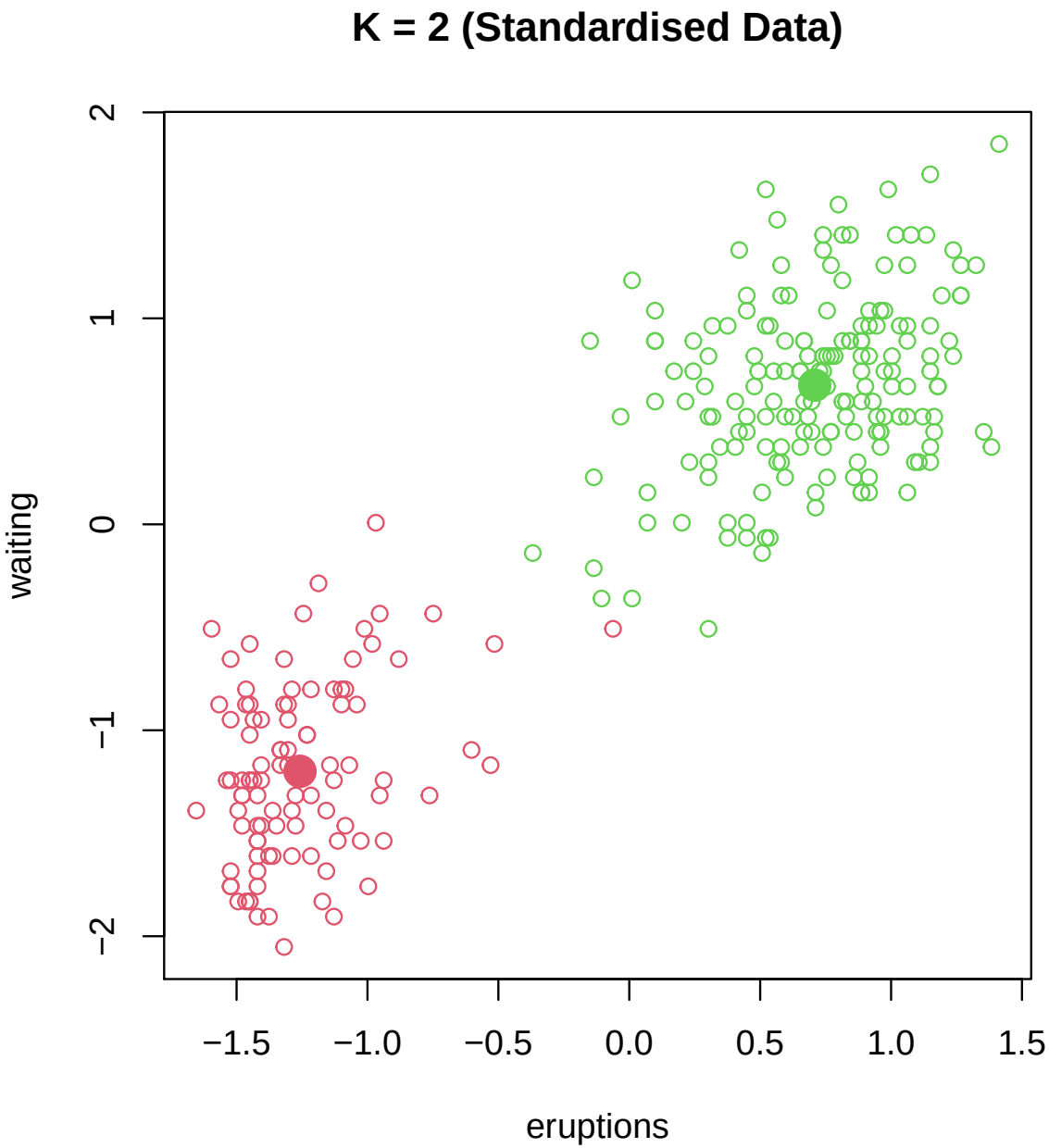
The K-means Method

- For K given, the method partitions observations into k clusters.
- It iterates between the following two steps until convergence:
  1. Given cluster centers, allocate each observation to its nearest cluster (in Euclidean distance to the cluster center).
  2. Update each cluster center with the mean of observations in the cluster.
- Aim to minimise (over K clusters)

$$\sum_{j=1}^{K} |C_j| \sum_{x \in C_j} [d(x, \bar{x}_j)]^2,$$

  where C_j denotes the jth cluster of size $|C_j|$ and with mean (vector) $\bar{x}_j$, and $d(\cdot, \cdot)$ the Euclidean distance.
- Method may converge to a local minimum.
- May need to standardise the data first.

K-means Clustering



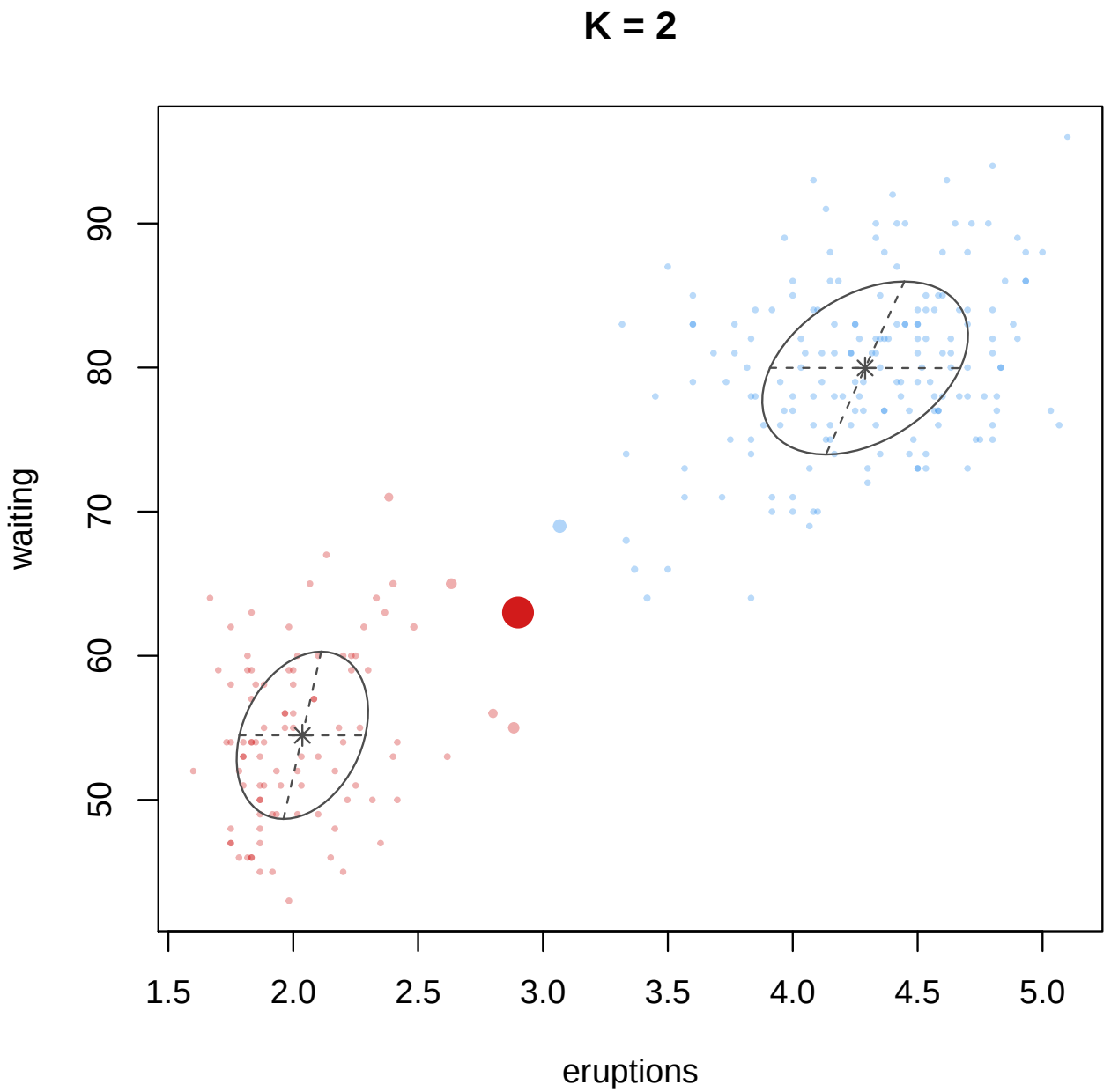
Mixture-based Clustering

- Also known as model-based clustering.
- Fit a (normal) mixture

$$f(x) = \pi_1 f_1(x) + \cdots + \pi_K f_K(x)$$

  to the data, where each component represents a cluster.
- An observation x is of cluster j, if $\pi_j f_j(x)$ is the largest, i.e., the same rule we used for classification.
- Each cluster needs to have an approximately normal distribution.
- One can choose to use mixture subfamilies, such as equal or varying variances.
- Invariant of data scaling.

Mixture-based Clustering



Hierarchical Clustering

- Hierarchical clustering gives a tree structure, with each observation down at the bottom of the tree (i.e., a leaf), forming a cluster by itself.
- The clustering tree is typically constructed in a bottom-up or agglomerative fashion.
- As moving up the tree (towards the root), in turn two clusters are fused into one, according to some dissimilarity measure.
- All observations or their clusters will eventually be fused into one final cluster at the root.
- No need to pre-specify the number of clusters, but may need to decide the number of clusters afterwards.

Dissimilarity

- First, we need to have a dissimilarity measure $d(x, x')$ between two observations x and x'.
  - This can just be Euclidean distance.
  - It also be a correlation-based distance, which would be more appropriate for variables.
- Then we need to decide on a dissimilarity measure $d(A, B)$ between two clusters A and B
  - This depends how we prefer the clusters to be linked together, hence the notion of linkage.
- If clusters are well separated, then different linkage methods will produce similar results.
- Otherwise, the results can be very much different.

Linkage Methods

- Complete linkage:

$$d(A, B) = \max\{d(x, x') : x \in A, x' \in B\}.$$

  - Use the largest pairwise observation dissimilarity.
  - Tend to produce compact clusters.
  - Some observations in the cluster may be closer to the observations in another cluster.

- Single linkage:

$$d(A, B) = \min\{d(x, x') : x \in A, x' \in B\}.$$

  - Use the smallest pairwise observation dissimilarity.
  - Close, immediate observations are linked first.
  - May produce extended, chain-shaped clusters, but almost no gaps between the observations.

Linkage Methods II

- Average linkage:

$$d(A, B) = \text{mean}\{d(x, x') : x \in A, x' \in B\}.$$

  - It is a compromise between complete linkage and single linkage.
  - Tend to produce relatively compact and relatively far apart clusters.

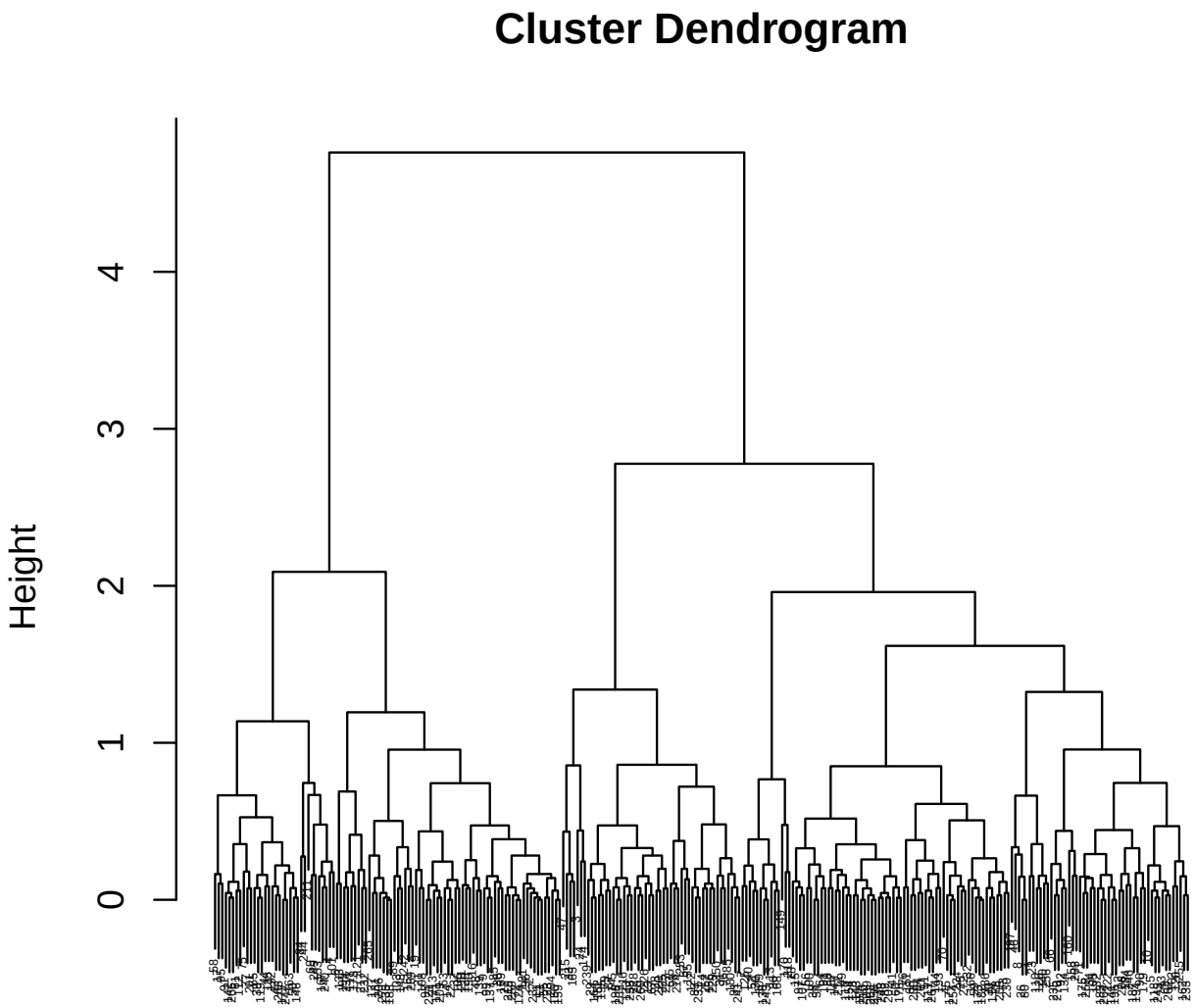
- Centroid linkage:

$$d(A, B) = d(\bar{x}_A, \bar{x}_B),$$

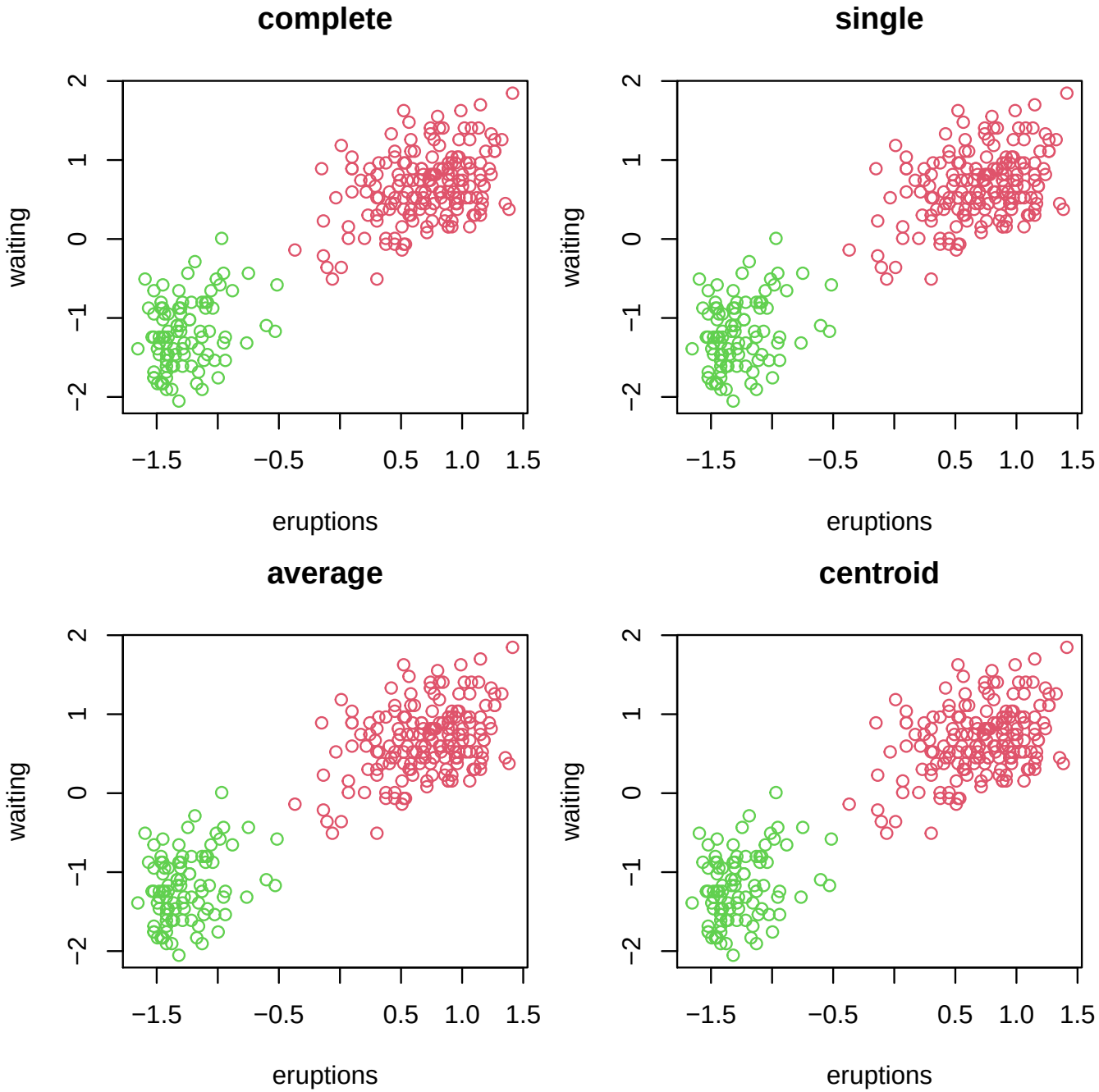
  where $\bar{x}_C$ denotes the sample mean (vector) of a cluster C.

  - Use dissimilarity between centers.

Hierarchical Clustering



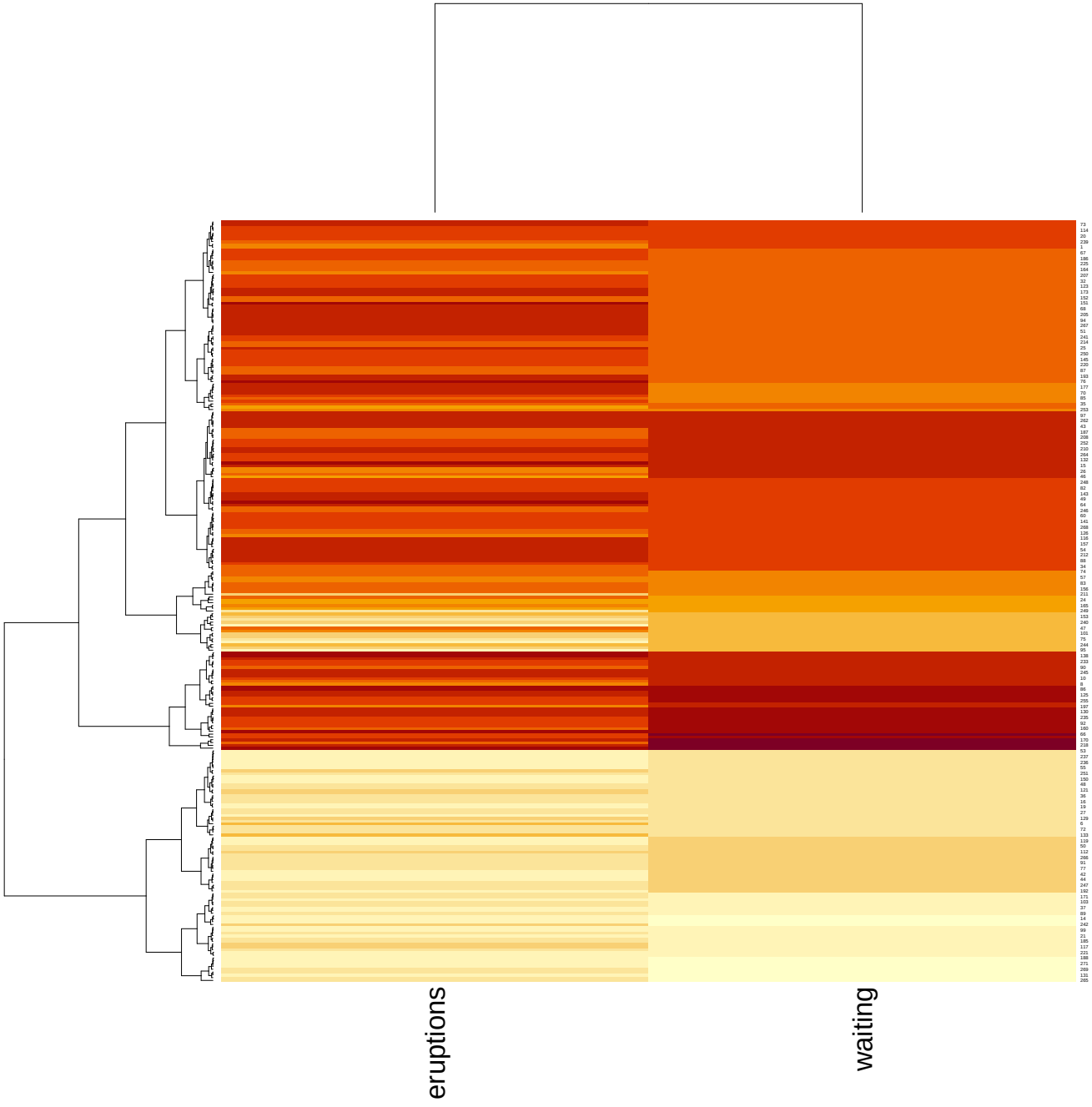
Linkage Methods ($K = 2$)



Heatmaps

- We can perform a hierarchical clustering of observations/rows and rearrange the rows according to the clustering results.
- Then perform a hierarchical clustering of variables/columns and rearrange the columns according to the clustering results.
- Then create a heatmap plot.
- In the heatmap plot shown below, high values are in red and low values are in yellow.

A Heatmap for Faithful Data Set



Adjusted Rand Index

- A clustering is a partition of observations.
- One can be interested in evaluating how similar two partitions are.
- The adjusted Rand index is such a measure, computed by evaluating whether or not each pair of observations are sent to one cluster in one partition, as well as to one cluster in the other partition.
- Its value is between -1 and 1, where 0 means a random allocation and 1 means a perfect agreement between the two partitions.

Recommended Readings

ISLv2 (basics):

- Section 12.4
- Labs: Section 12.5.3

ESL (advanced):

- Sections 14.3.6, 14.3.7, 14.3.12