Clustering
Hierarchical and K-means

Erik-Jan van Kesteren & Daniel L. Oberski

Last weeks

- Dimension reduction
- PCA, SVD
- ICA, NMF, Factor analysis
- “reconstruction error”, e.g., $\left(X - \hat{X}\right)^2$

Today

- K-means clustering
- Hierarchical clustering
- Distance metrics
- Evaluating cluster solutions
- Sparse hierarchical clustering

- Goal: understand, apply, and assess clustering methods

Reading materials

- ISLR section 12.4 on clustering
- Introduction to data mining section 7.5 cluster validation
- SLS sections 8.5.1 and 8.5.2.

Clustering

Clustering

Find subgroups (clusters) of similar examples in a database

Why clustering?

- We expect clusters in our data, but weren't able to measure them
  - potential new subtypes of cancer tissue

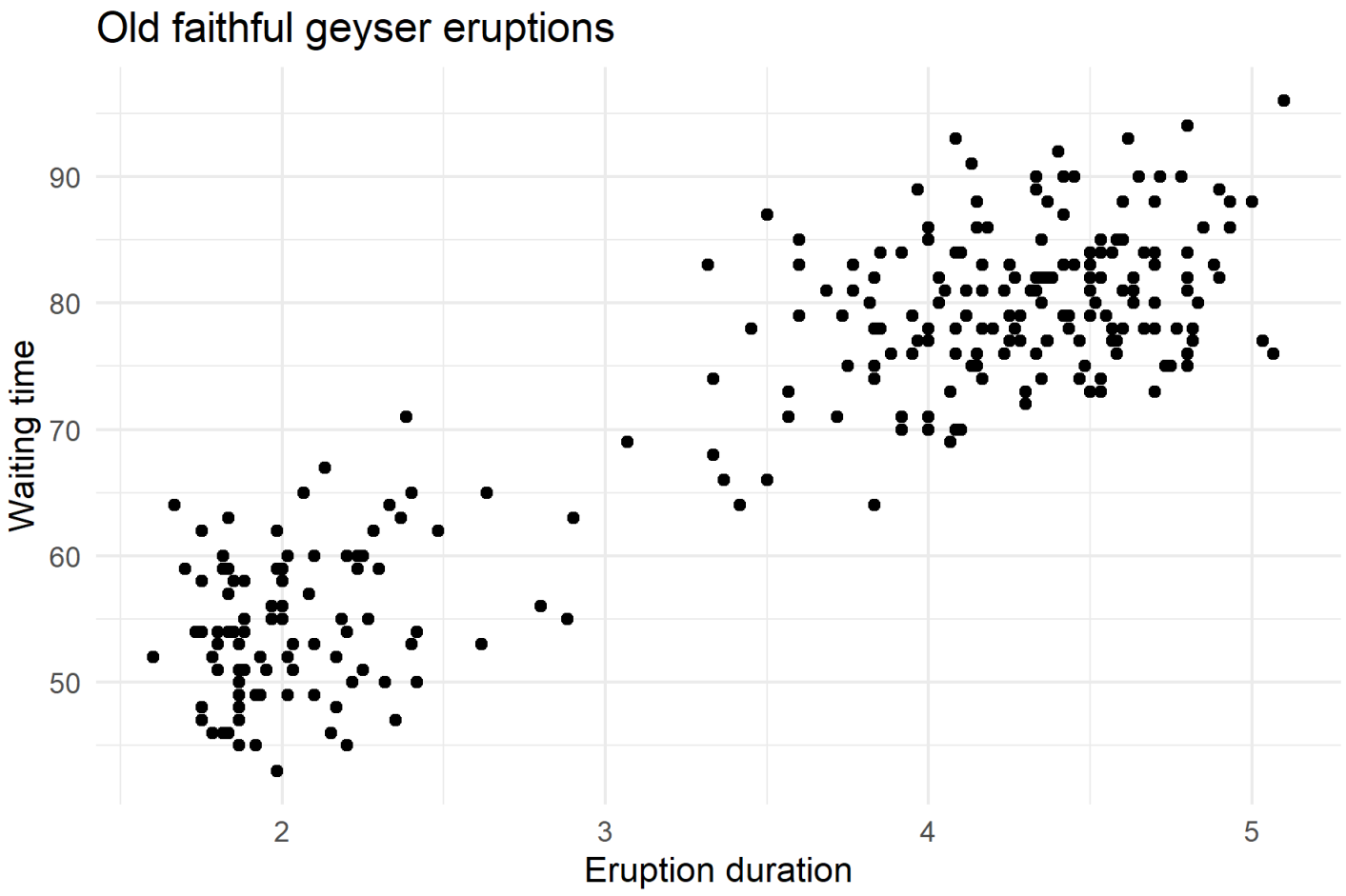
- We want to summarize features into a categorical feature to use in further decisions/analysis
  - subgrouping customers by their spending types

Clustering: parallels

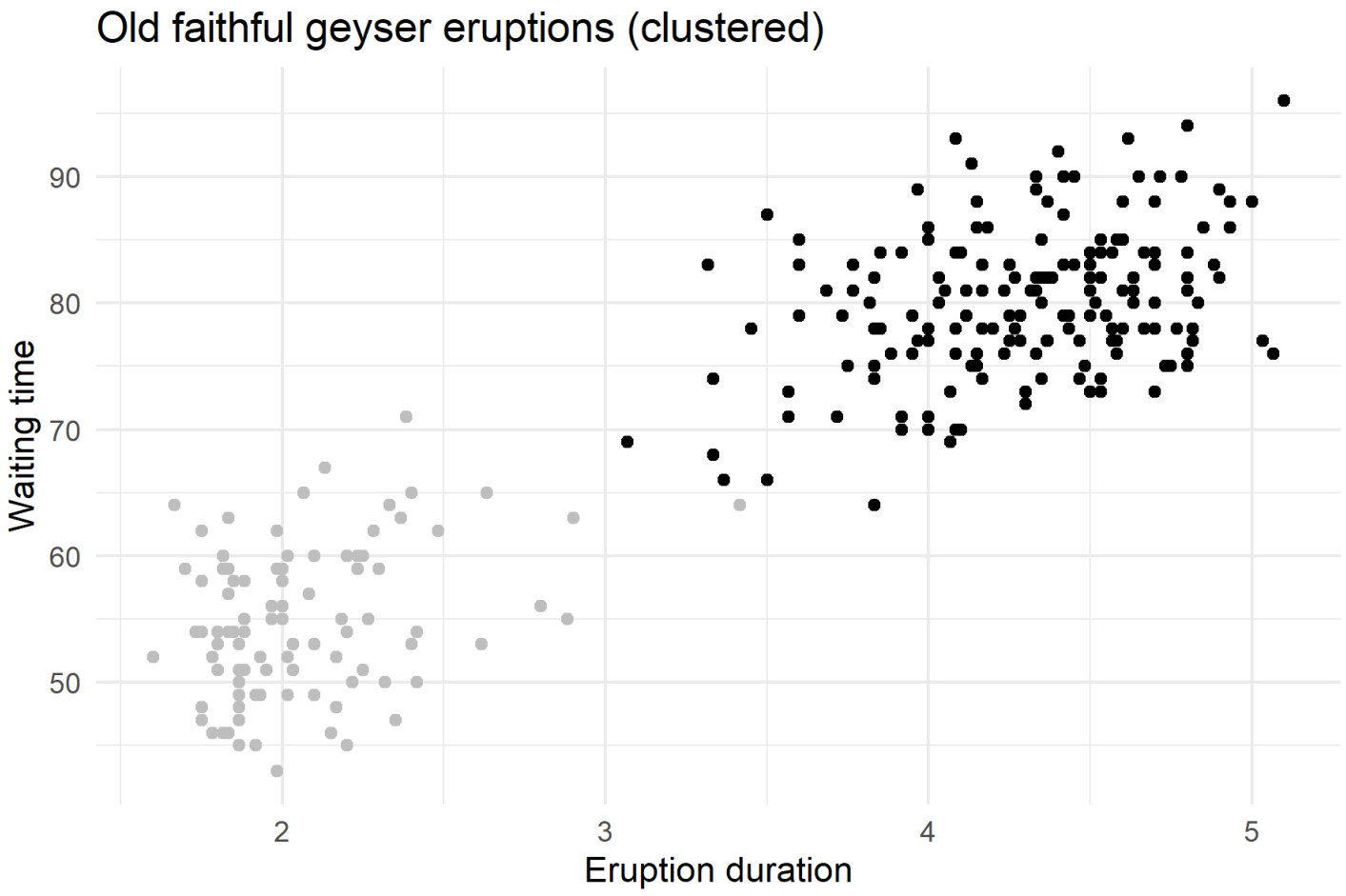
- Clustering is (unsupervised) classification
  - Predicting unobserved class!
- Clustering is dimension reduction
  - Reducing P variables to a single categorical variable!

- You may notice these parallels at several points during these two weeks

Old faithful: two types of eruption?



Old faithful: two types of eruption?

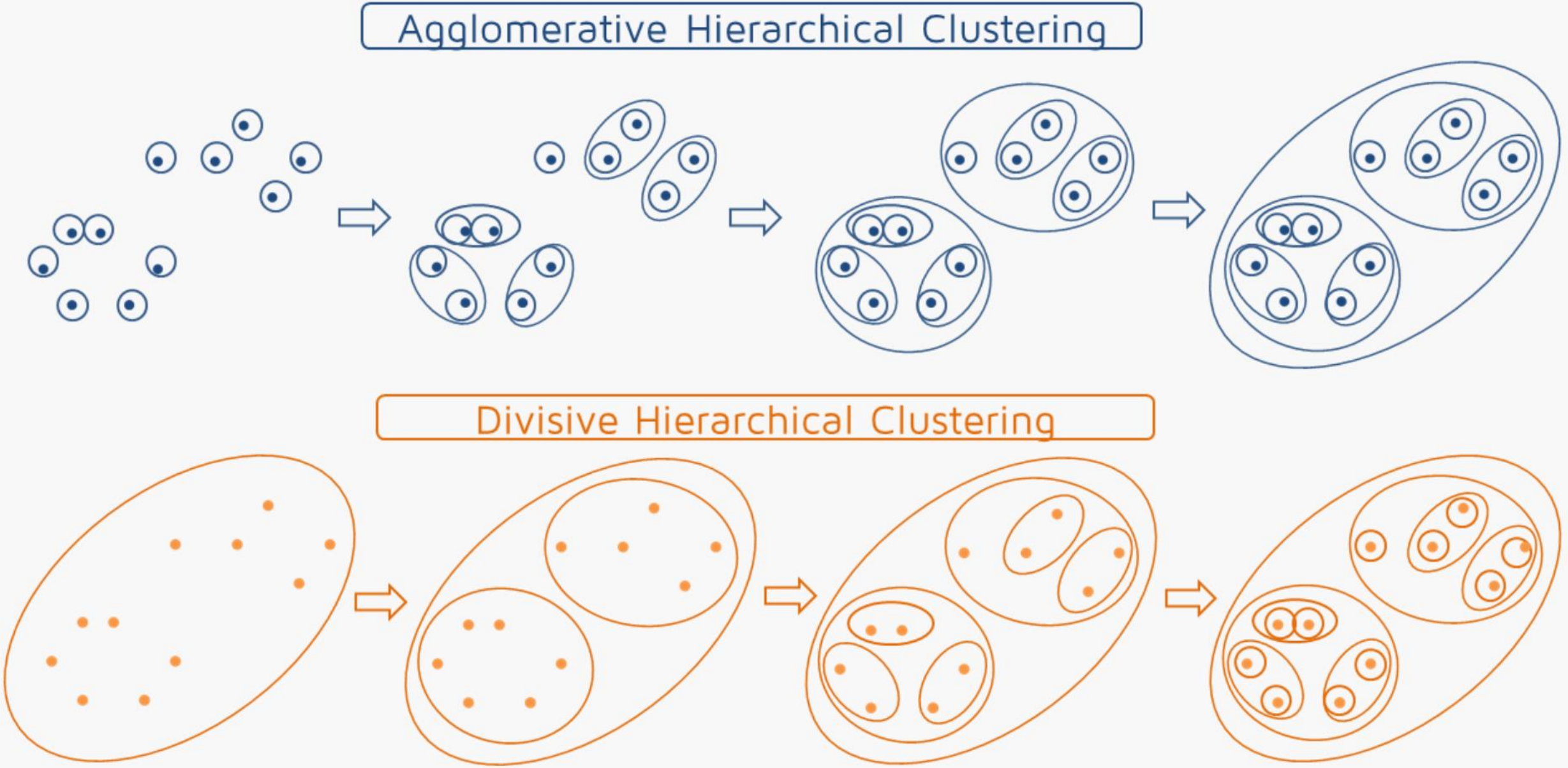


Hierarchical clustering

Hierarchical clustering

- Bottom-up agglomerative clustering
  - For each observation, compute the distance to all other observations
  - Assign all examples to their individual cluster
  - Combine most similar cluster
  - Keep combining clusters until there is only one cluster left
  - Select number of clusters for the final solution

Divisive: start with one and keep splitting most different



https://quantdare.com/hierarchical-clustering/

Hierarchical clustering

- Distance and cluster similarity are predefined hyperparameters
- **Distance**: how do we measure the multivariate distance between two examples?
  - euclidean, maximum, manhattan, canberra, binary, minkowski, correlation, cosine similarity
- **Cluster similarity** (linkage): how do we summarise the difference between clusters?
  - Complete, single, average, centroid

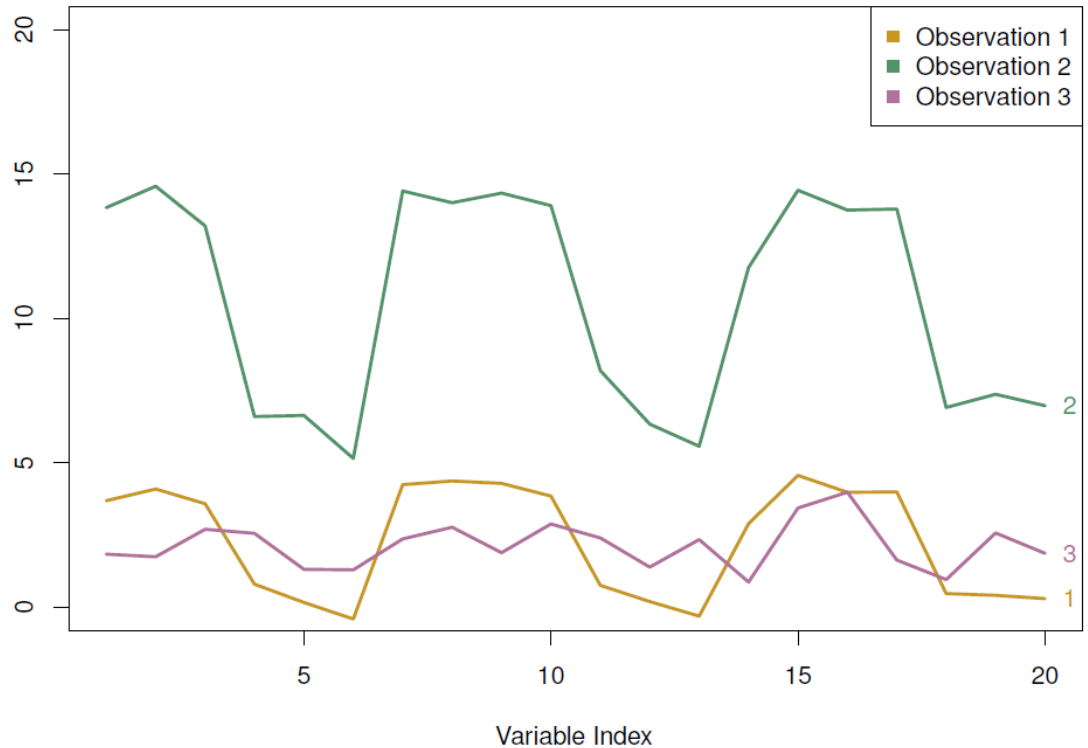


FIGURE 12.15. *Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations 1 and 2 have quite different values for each variable, and so there is a large Euclidean distance between them. But they are highly correlated, so there is a small correlation-based distance between them.*

klcluster: Center-based Clustering of Trajectories

Kevin Buchin*
Eindhoven Technical University
Eindhoven, the Netherlands

Anne Driemel*
Hausdorff Center for Mathematics
University of Bonn
Bonn, Germany

Natasja van de L'Isle*
Eindhoven Technical University
Eindhoven, the Netherlands

André Nusser*
Max Planck Institute for Informatics
Graduate School of Computer Science
Saarbrücken, Germany

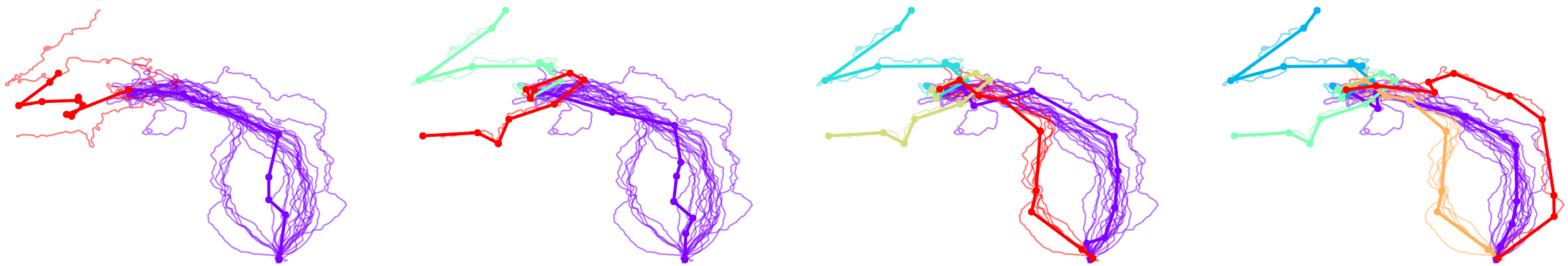
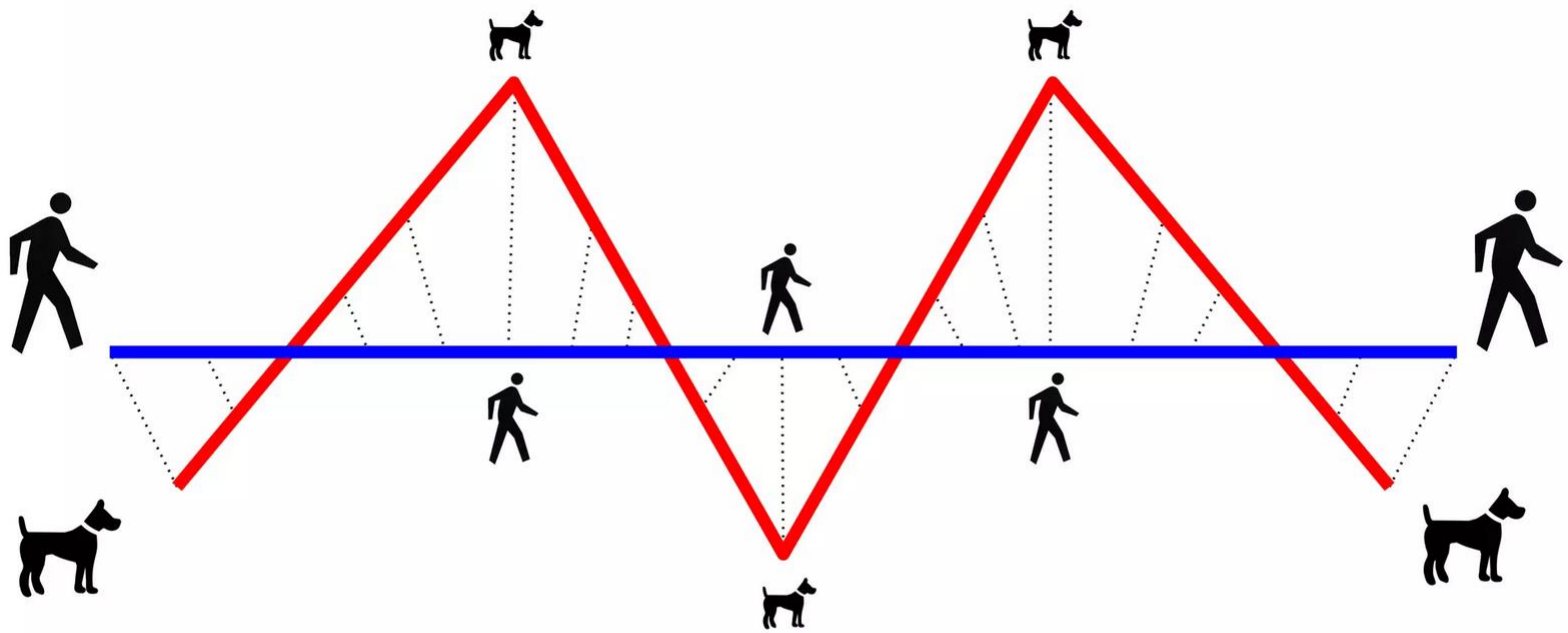


Figure 1: Example of a (k, ℓ)-clustering for the flight paths of a pigeon with the number of clusters k increasing from 2 (left) until 5 (right) and the complexity of the clusters being $\ell = 10$. Trajectories belonging to the same cluster are shown in the same color. For each cluster, a center trajectory generated by the algorithm is shown using thick lines of the same color.

Walking your dog



The **Fréchet distance** between the curves is the minimum leash length that permits such a walk

https://www.slideshare.net/shripadthite/frechettalk

Hierarchical clustering

- Distance and cluster similarity are predefined hyperparameters
- **Distance**: how do we measure the multivariate distance between two examples?
  - euclidean, maximum, manhattan, canberra, binary, minkowski, correlation, cosine similarity
- **Cluster similarity** (linkage): how do we summarise the difference between clusters?
  - Complete, single, average, centroid

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

TABLE 12.3. *A summary of the four most commonly-used types of linkage in hierarchical clustering.*

Hierarchical clustering

In R:

```
distances <- dist(faithful,     method = "euclidean")
result    <- hclust(distances, method = "average")
```

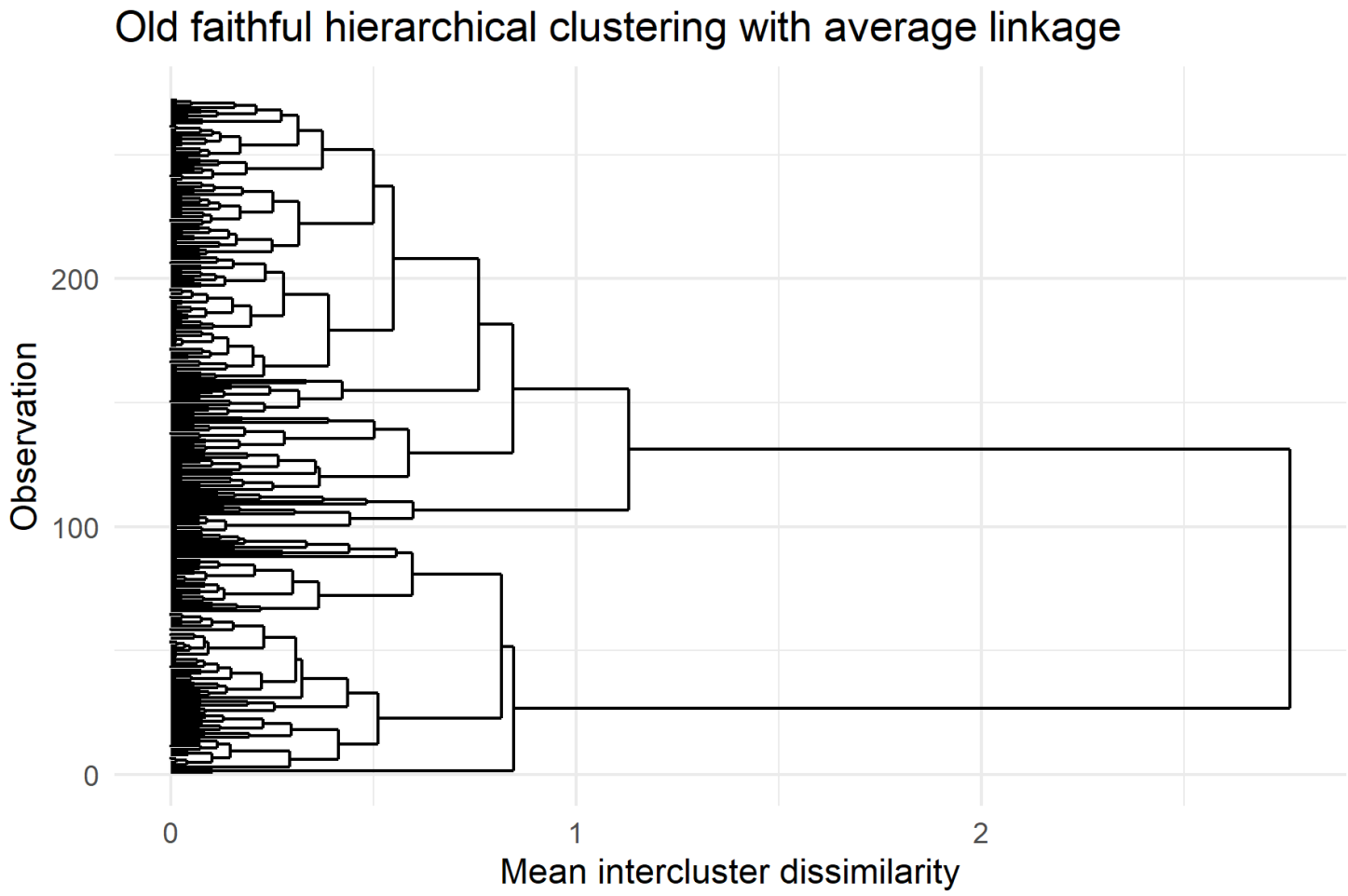
Then we can plot the dendrogram with `plot()` or `ggdendrogram`

```
library(ggdendro)
ggdendrogram(result)
```

Then, select the number of clusters using a cutoff point

```
cutree(result, h = 2)
```

Hierarchical clustering



Hierarchical clustering

In R:

```
distances <- dist(faithful,     method = "euclidean")
result    <- hclust(distances, method = "average")
```

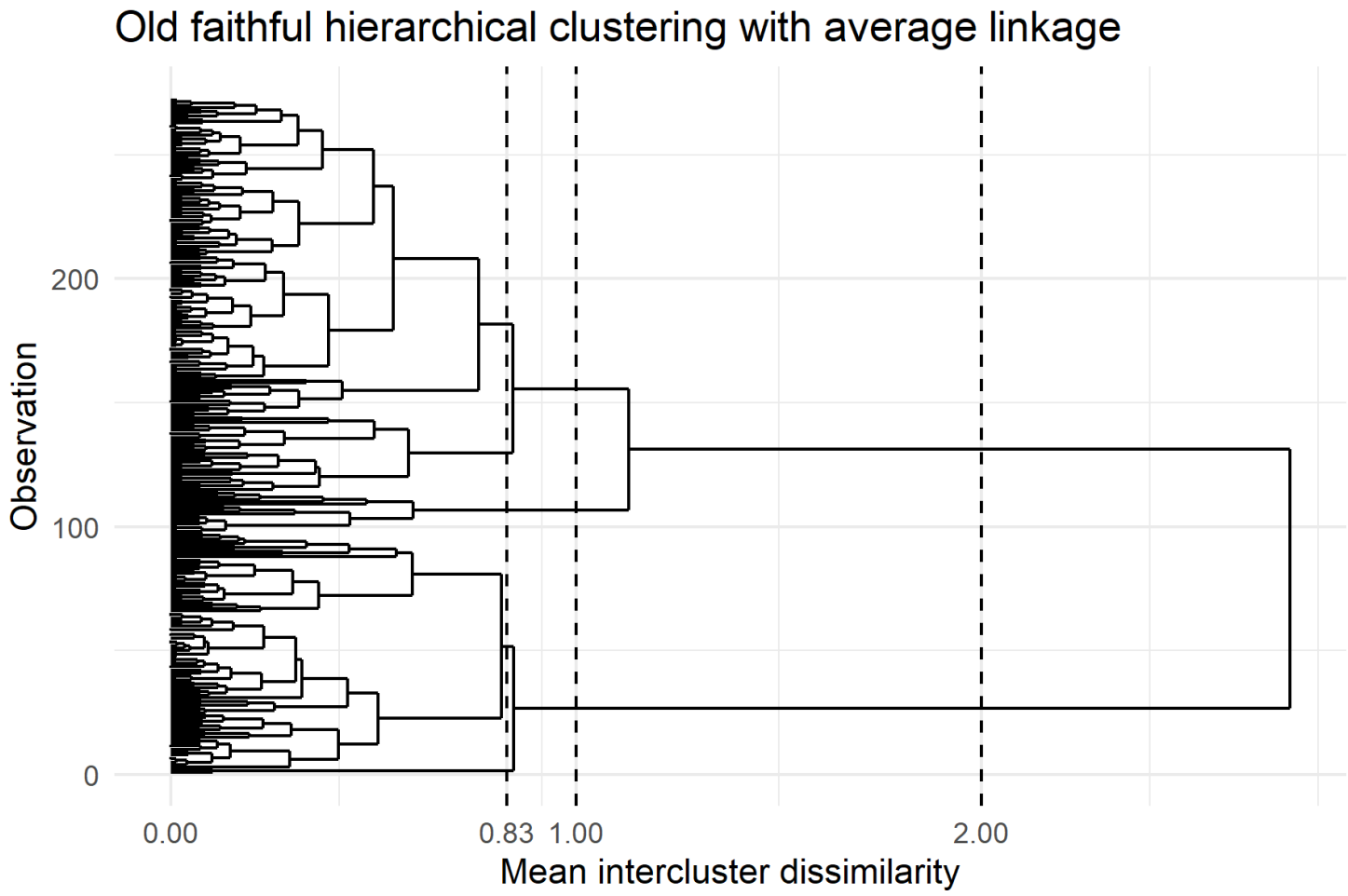
Then we can plot the dendrogram with `plot()` or `ggdendrogram`

```
library(ggdendro)
ggdendrogram(result)
```

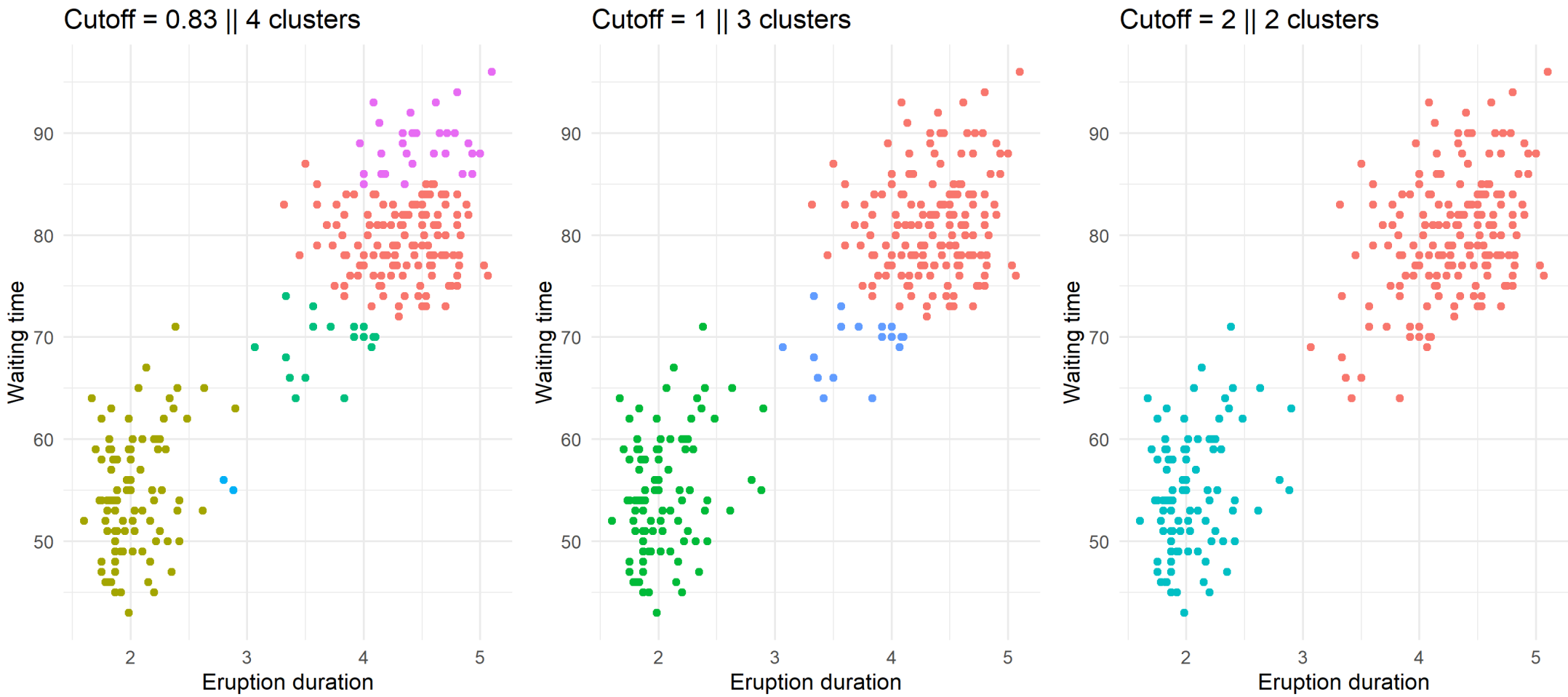
Then, select the number of clusters using a cutoff point

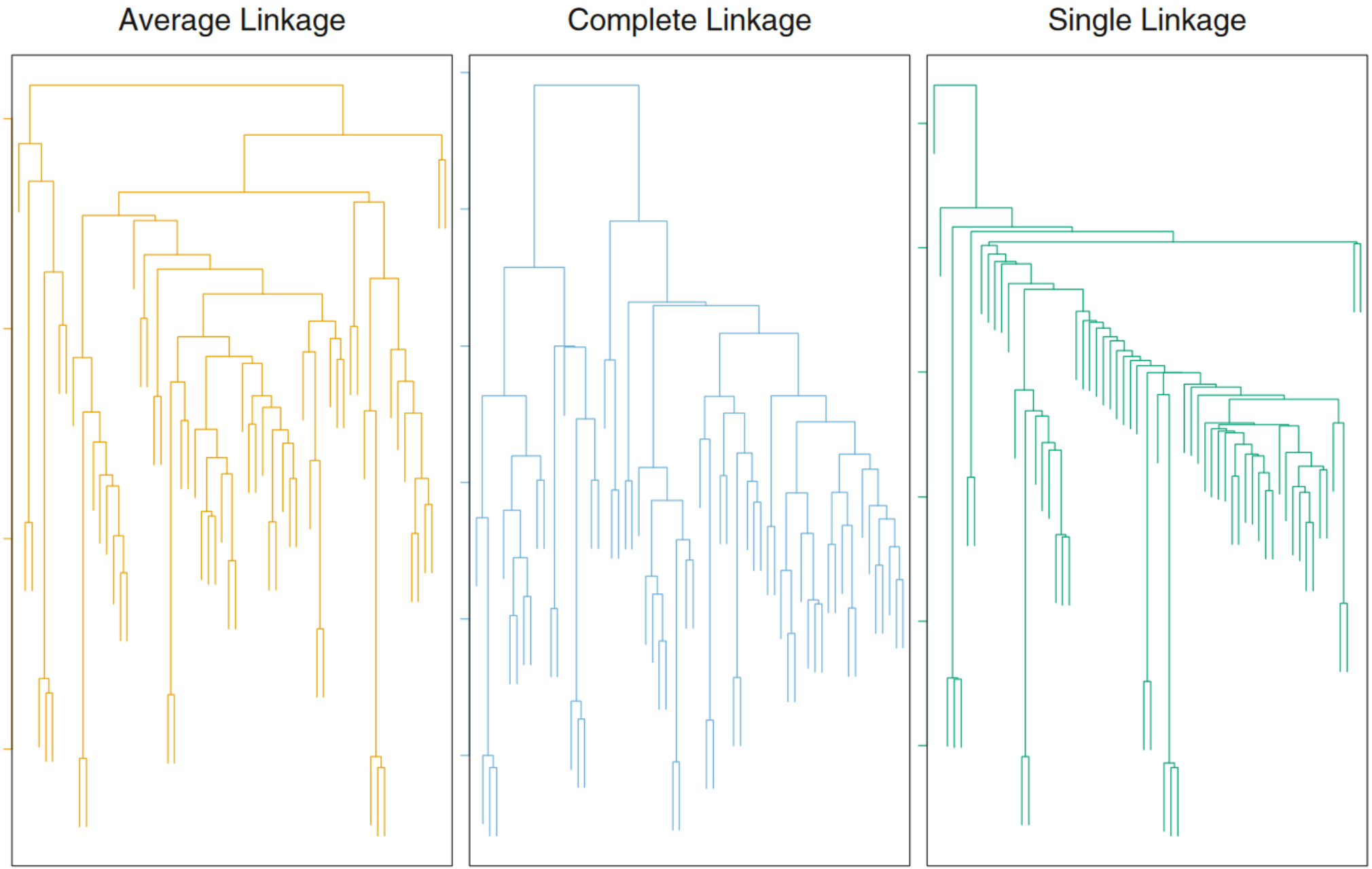
```
cutree(result, h = 2)
```

Hierarchical clustering



Hierarchical clustering




Average Linkage
Complete Linkage
Single Linkage

Note: scaling

- Measure your features in the same scale for clustering
  - Otherwise, height in cm will be more important than width in meters
- This can be done by standardization, or z-transformation
  - subtract the mean from each feature and divide by its observed standard deviation
- Changes the interpretation of the values, but not their association

Hierarchical clustering: conclusion

- Tree-based representation - dendrogram
- Determine number of clusters afterwards
- Different distance metrics possible
- Different agglomeration methods possible

Break

K-means clustering

K-means clustering

- Predefine the number of clusters (K)
- Apply an algorithm to assign observations to clusters
- Top-down method

K-means clustering algorithm

1. Randomly assign examples to K clusters

2a. Calculate the centroid (per-feature mean) for each cluster

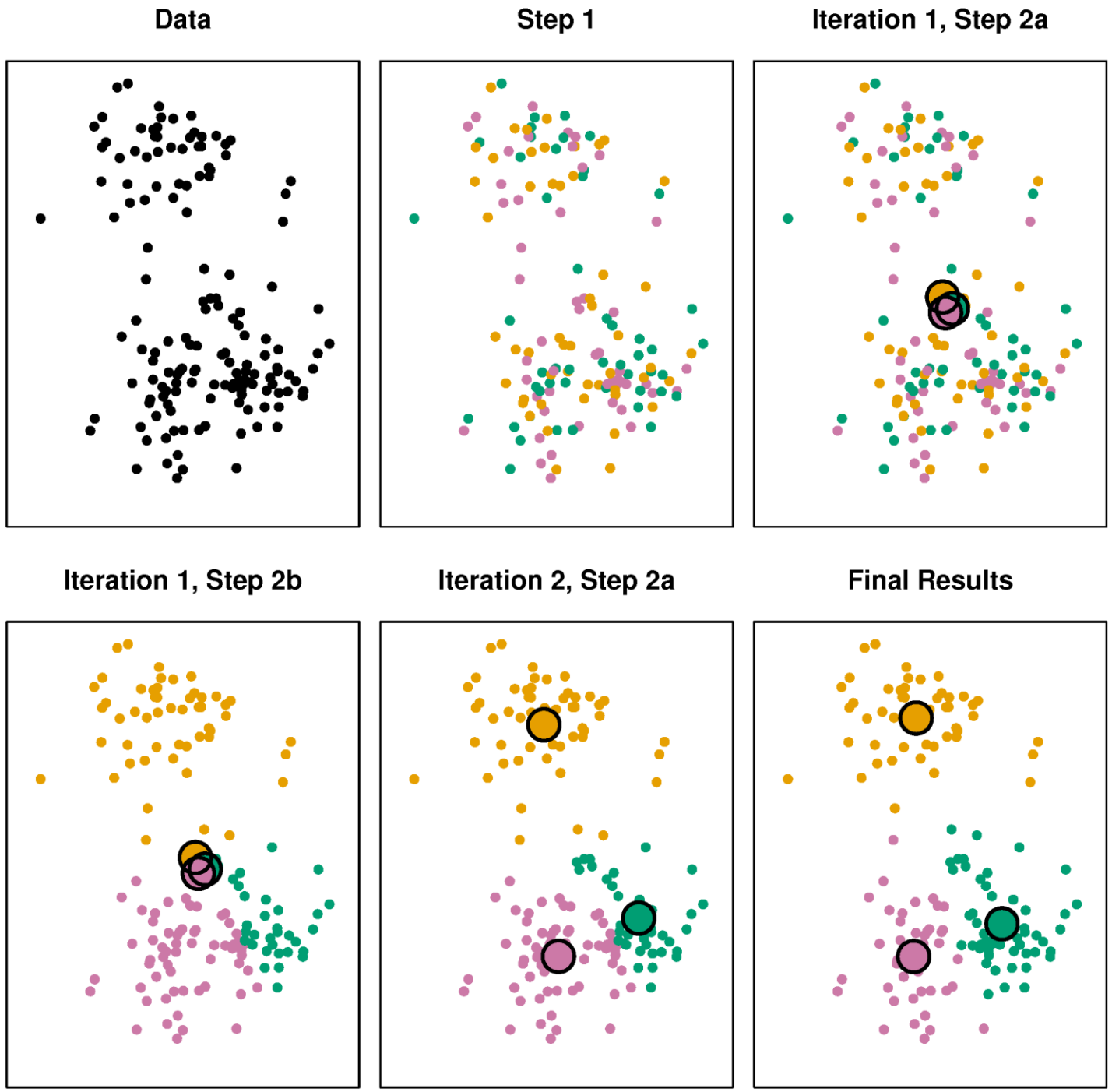
```
faithful %>% group_by(cluster) %>% summarize_all(mean)

#>   cluster eruptions waiting
#>     <int>     <dbl>   <dbl>
#> 1       1      3.69    73.4
#> 2       2      3.30    68.6
```

2b. Assign each example to the cluster belonging to its closest centroid

3. If the assignments changed, go to step 2a, else stop

K-means clustering



ISLR2 fig 12.8

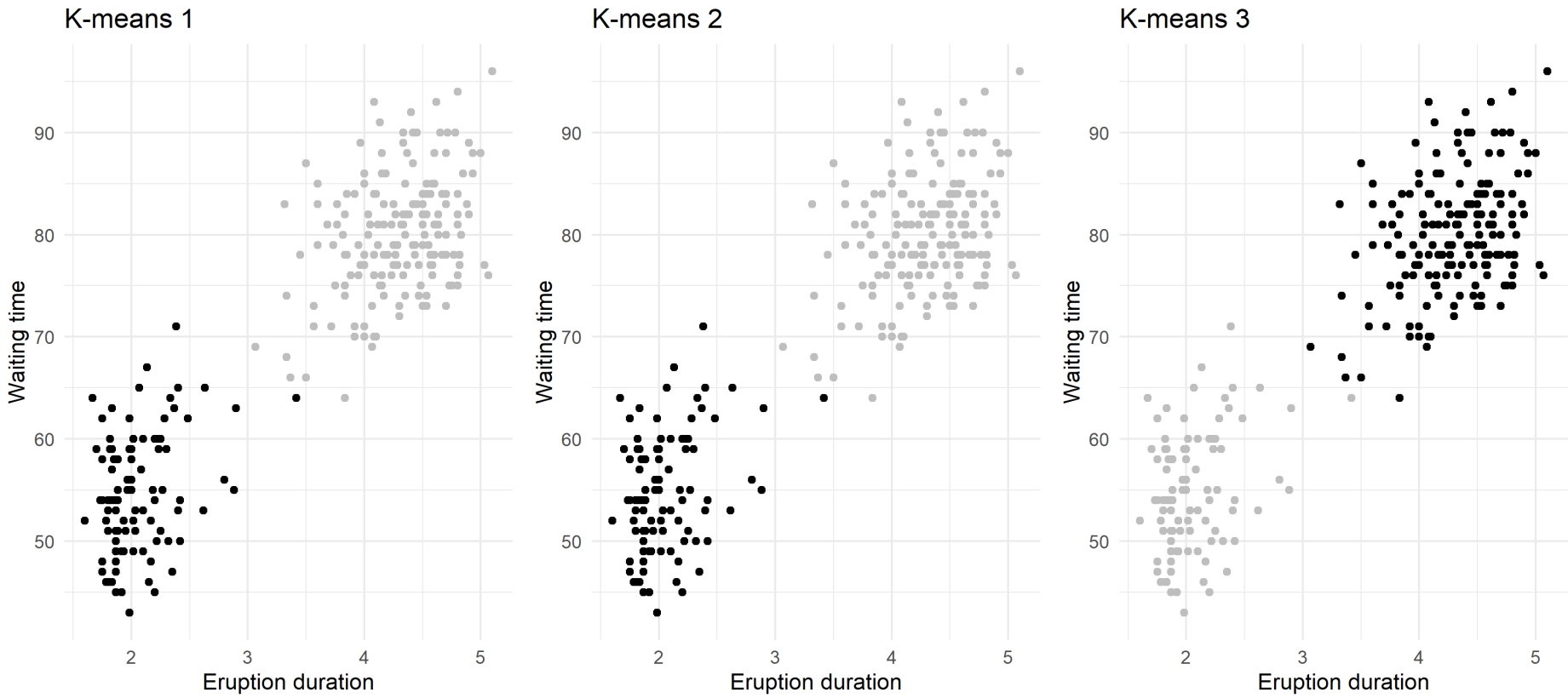
K-means clustering

- K and the distance metric are hyperparameters
- Distance metric can be anything, just like in hierarchical clustering
  - Determined by the structure / content of the data
  - Usually euclidian distance
- K is determined in advance by the data scientist based on knowledge about the data or the goal of the analysis
  - Perhaps there are generally 2 types of geyser eruption because of physics
  - We may have resources to approach customers in at most 3 different ways

K-means clustering

- Because the initialization is random, the result is random
  - Label switching: cluster 1, 2, 3 may end up in each other's locations
  - Some examples at the boundary may end up in different clusters altogether
  - Use multiple starts to obtain the best solution

K-means clustering



K-means clustering

- Vector quantization (Lloyd's algorithm) is k-means clustering applied to image processing.
- Goal: image compression -- less storage!
- Cluster blocks of 4 pixels, then replace blocks by their cluster centroid

K-means clustering

Pixel 1	Pixel 2	Pixel 3	Pixel 4
113	234	40	230
198	164	86	237
47	222	96	224
93	28	226	22
39	240	207	21
242	35	212	113
…	…	…	…

K-means clustering

Original (1MB)

K = 200 (0.2375 MB)

K = 4 (0.0625 MB)

ESL fig 14.9

Evaluating cluster results

Evaluating cluster results

Many options (hyperparameters) in clustering methods

- They need to be tuned to obtain “good” clusters.
- “small decisions with big consequences” (ISLR2, section 12.4.3)
- In other words: clusters are sensitive to hyperparameter choice

Three methods to assess whether obtained clusters are “good”

- Stability
- External validity
- Internal indices and model fit (next lecture)

Cluster stability

How stable are the clusters?
Two different ways

Bootstrap stability
- Compute cluster solutions on M bootstrap samples (Hennig, 2007)
- Compare the similarity of the M cluster solutions to the original solution
- More similar is better!

Leave-one-feature-out
- Leave a feature out and perform clustering
- Compare the similarity of the cluster solutions to the original solution
- More similar is better!

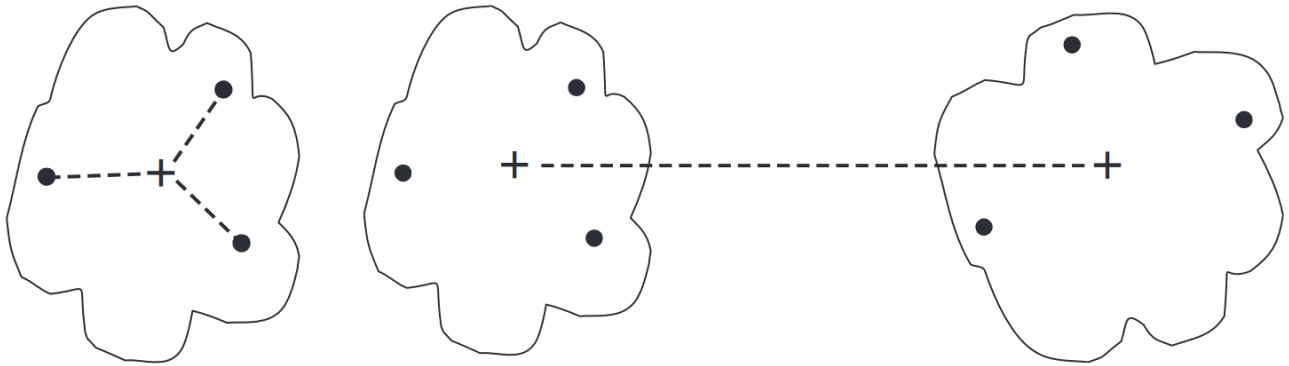
External validity

- Are the clusters associated with external feature Y?
- Making unsupervised supervised (a little cheating)

- Examples:
  - Are my customer segments based on spending associated with the demographics of the customers?
  - Are the geyser eruption types strongly correlated with water pressure or temperature?
  - Can I recognize the person in the vector quantized picture?

Internal indices

- Quantify the cohesion and separation of the clusters
  - Cohesion: how similar are observations within cluster?
  - Separaton: how dissimilar are the clusters?



(a) Cohesion. (b) Separation.

Figure 7.28. Prototype-based view of cluster cohesion and separation.

Internal indices

- Silhouette coefficient
  - a_i = avg. distance to fellow cluster members (cohesion)
  - b_i = min. distance to member from different cluster (separation)

$$s_i = \frac{(b_i - a_i)}{\max(a_i,\ b_i)}$$

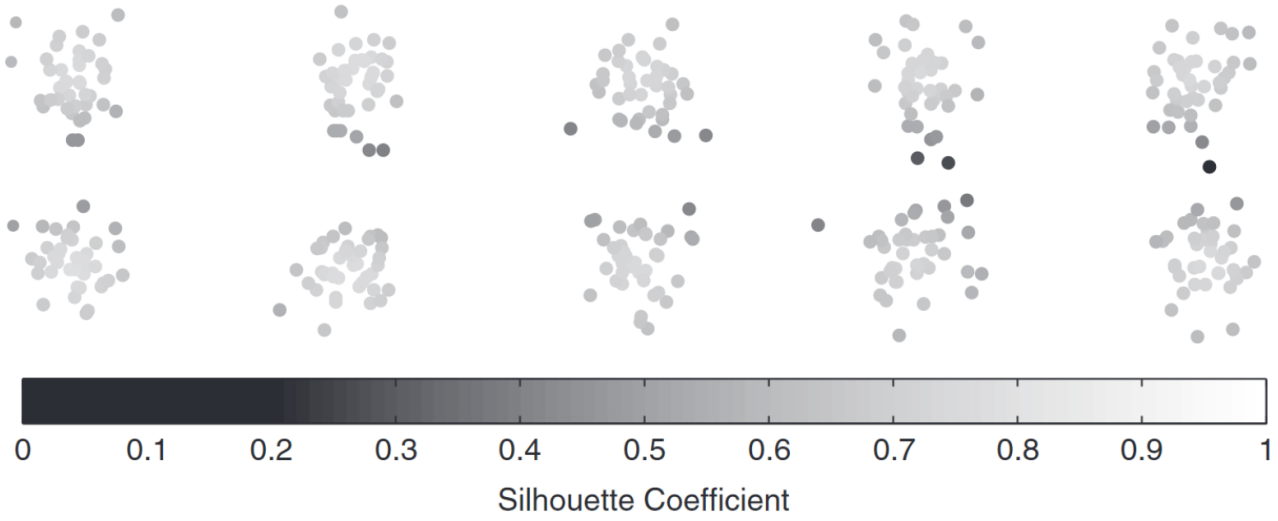


Figure 7.29. Silhouette coefficients for points in ten clusters.

Sparse hierarchical clustering

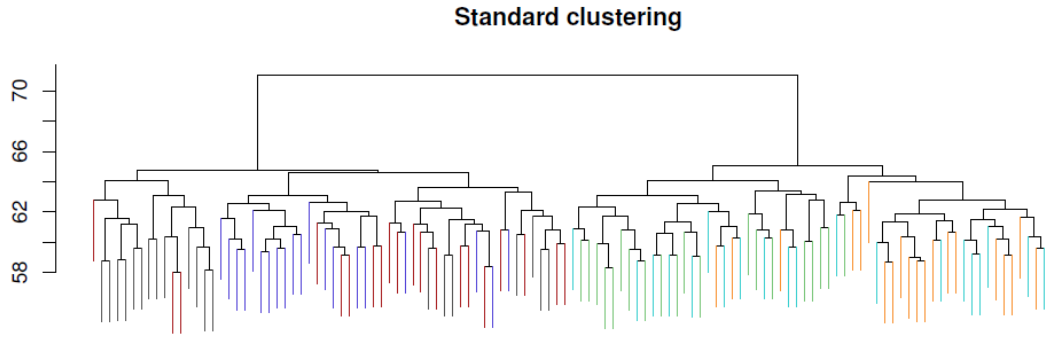
Sparse hierarchical clustering

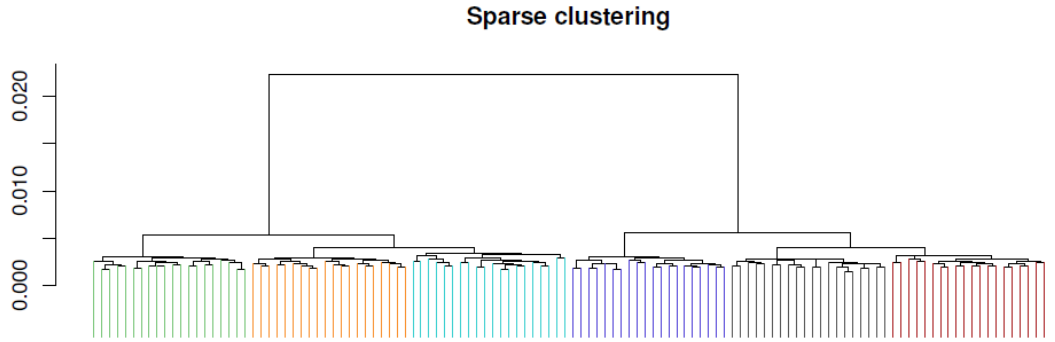
- Insight: distance matrix $\boldsymbol{U}$ can be computed as sum of per-feature distance matrices $\boldsymbol{U} = \sum_{p=1}^{P} \boldsymbol{U}_p$
- Weighted distance matrix can be computed using weights $\boldsymbol{U}' = \sum_{p=1}^{P} \boldsymbol{U}_p \cdot w_p$
- Transform each $\boldsymbol{U}_p$ into a vector, and create a $N^2 \times P$ matrix $\boldsymbol{\Delta}$
- Estimate $\boldsymbol{w}$ to get sparse $\widehat{\boldsymbol{w}}$
- Create $\widehat{\boldsymbol{U}'} = \boldsymbol{\Delta}\widehat{\boldsymbol{w}}$
- Perform hierarchical clustering as normal on $\widehat{\boldsymbol{U}'}$

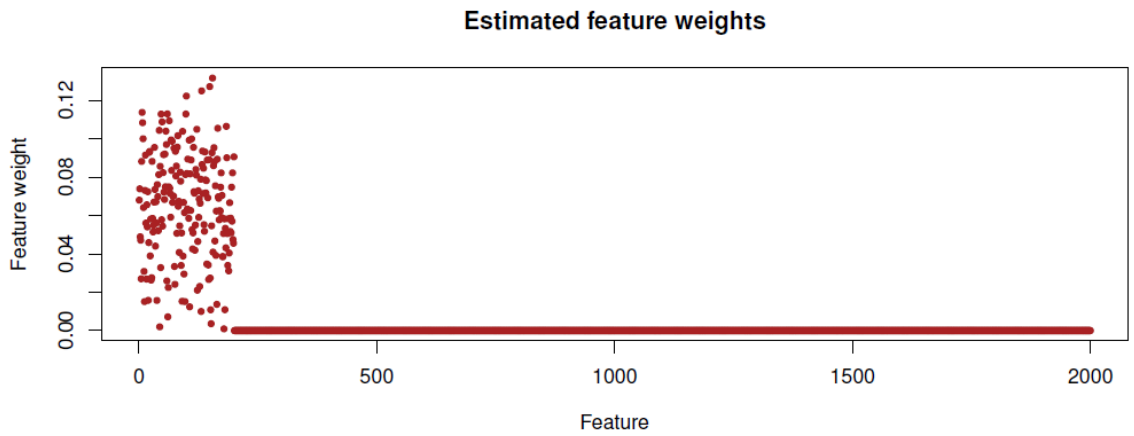
$$\underset{\mathbf{u}\in\mathbb{R}^{N^2},\, w\in\mathbb{R}^{p}}{\text{maximize}} \left\{\mathbf{u}^T \boldsymbol{\Delta} w\right\} \text{ subject to } \|\mathbf{u}\|_2 \le 1,\ \|w\|_2 \le 1, \quad \|w\|_1 \le s, \text{ and } w \succeq 0. \tag{8.43}$$

Sparse hierarchical clustering

- On simulated data with P = 2000
- Sparsity: only first 200 features varied across 6 underlying classes
- Sparse clustering groups real underlying classes together (colours) while standard clustering does not







Sparse hierarchical clustering

- Solves this problem:
  - We want to cluster observations $n = 1 \ldots N$ on their P features
  - $P > N$
  - Only a few features are relevant

- This works well but it is very difficult!
- Can you think of some alternatives?

Sparse hierarchical clustering

- Nice website (using UMAP to dimension reduce MNIST):

https://grantcuster.github.io/umap-explorer/

Conclusion

- Clustering seeks to obtain similar subgroups of observations
- There are two basic clustering methods:
  - hierarchical clustering (bottom-up agglomerative)
  - partitioning (k-means clustering)
- They each have hyperparameters which need tuning
- Assess with stability, external validity, internal indices
- Hi-dim:
  - Make a sparse representation of the distance matrix
  - Alternative: first reduce dimension and then perform clustering

Next lecture
Model-based clustering / mixture modeling

Practical: perform k-means and hierarchical clustering

Take-home exercises: 1-6

Challenge question: create your own k-medians clustering algorithm

Questions?