3 Clustering

Clustering refers to a broad set of techniques for finding *subgroups* in a data set.

We seek to partition observations into distinct groups so that
- observations within a group are similar
- observations in different groups are dissimilar

need to define
Depend on domain!

For instance, suppose we have a set of n observations, each with p features. The n observations could correspond to tissue samples for patients with breast cancer and the p features could correspond to measurements collected for each tissue sample:

- clinical measurements, e.g. tumor stage or grade
- gene expression measurements.

diverse in character

We may have reason to believe there is heterogeneity among the n observations.

e.g. different unknown subtype of cancer.

This is *unsupervised* because

We are trying to discover structure (distinct clusters) in the absence of a response.

vs.

Supervised problems we have the goal of prediction of a response.

Both clustering and PCA seek to simplify the data via a small number of summaries.

- PCA – finds a low dimensional representation of the observations that explain a good fraction of the variance.
- Clustering – finds homogenous subgroups among observations

Since clustering is popular in many fields, there are many ways to cluster.

We will focus on 2 best-known clustering approaches.

- K-means clustering

  Seeks to partition the observations into a pre-specified # of clusters.

- Hierarchical clustering

  We do not know in advance how many clusters we want.

  We obtain clusterings for $1, \ldots, n$ # of clusters

  ↳ can view in a tree-like visualization called a "dendrogram"

In general, we can ① cluster observations on the basis of features or ② we can cluster features on the basis of observations.

① ↓ identify subgroups among observations

② ↓ identify subgroups among the features.

We will focus on ①

But we can perform ② by transposing the data matrix.

$X \longrightarrow X^T \longrightarrow$ clustering.

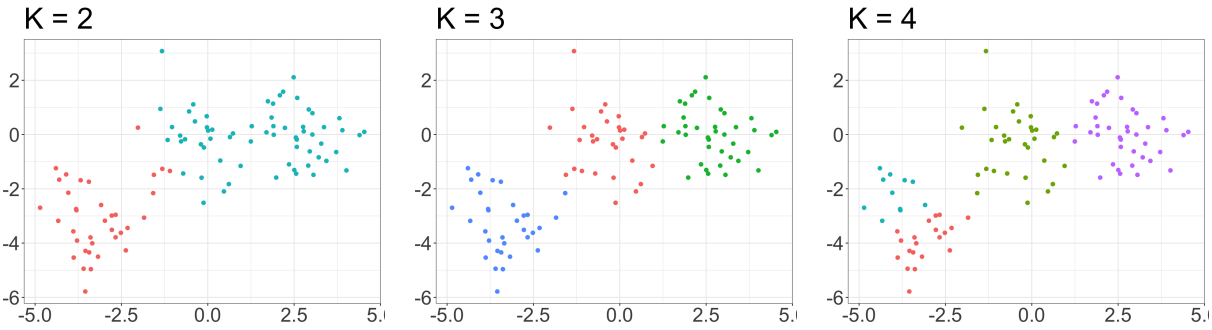
$n \times p$ $\quad p \times n$

3.1 K-Means Clustering

Simple and elegant approach to parition a data set into K distinct, non-overlapping clusters.

We must first specify how many clusters K.
Then K-means assigns each observation to one of the K clusters.

eg. clustering $n=100$ observations into K clusters using $p=2$ features.



The K-means clustering procedure results from a simple and intuitive mathematical problem. Let $C_1, \ldots, C_K$ denote sets containing the indices of observations in each cluster. These satisfy two properties:

e.g. if observation i is in cluster k, $i \in C_k$

1. $C_1 \cup \cdots \cup C_K = \{1, 2, \ldots, n\}$

   each observation belongs to one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset \quad \forall k \neq k'$

   The clusters are non overlapping.

Idea: "good clustering" is one for which the within-cluster variation is as small as possible.

The *within-cluster variation* for cluster C_k is a measure of the amount by which the observations within a cluster differ from each other.

Call this $W(C_k)$.

Then we want to solve the problem:

$$\underset{C_1,\dots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

← we want to partition observations into K clusters s.t. total within-cluster variation is minimized.

To solve this, we need to define within-cluster variation.

Many way we can do that.

Most common way: squared euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

$|C_k|$ ~ # obs in k^{th} cluster.

This results in the following optimization problem that defines K-means clustering:

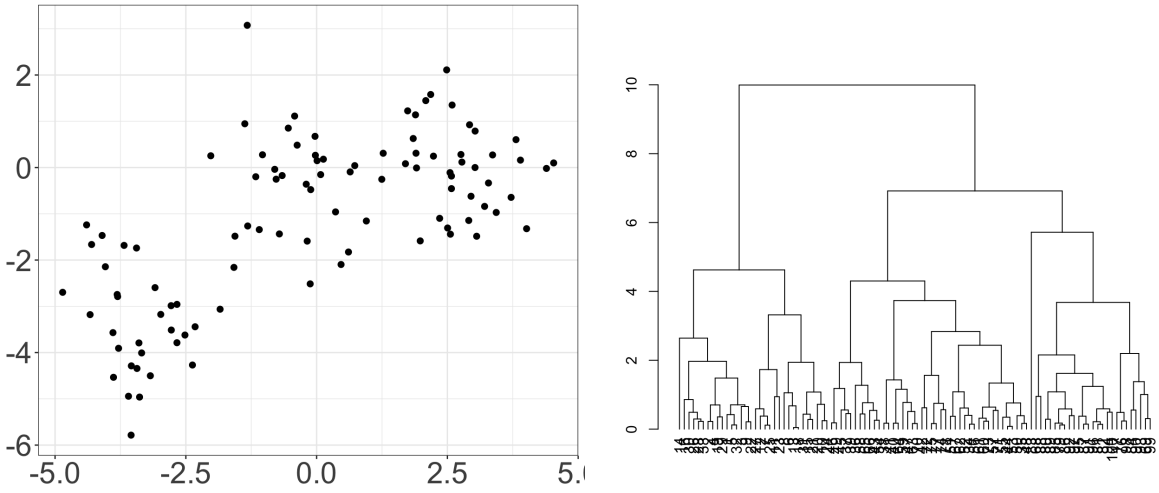
A very simple algorithm has been shown to find a local optimum to this problem:

3.2 Hierarchical Clustering

One potential disadvantage of K-means clustering is that it requires us to specify the number of clusters K. *Hierarchical clustering* is an alternative that does not require we commit to a particular K.

We will discuss *bottom-up* or *agglomerative* clustering.

3.2.1 Dendrograms

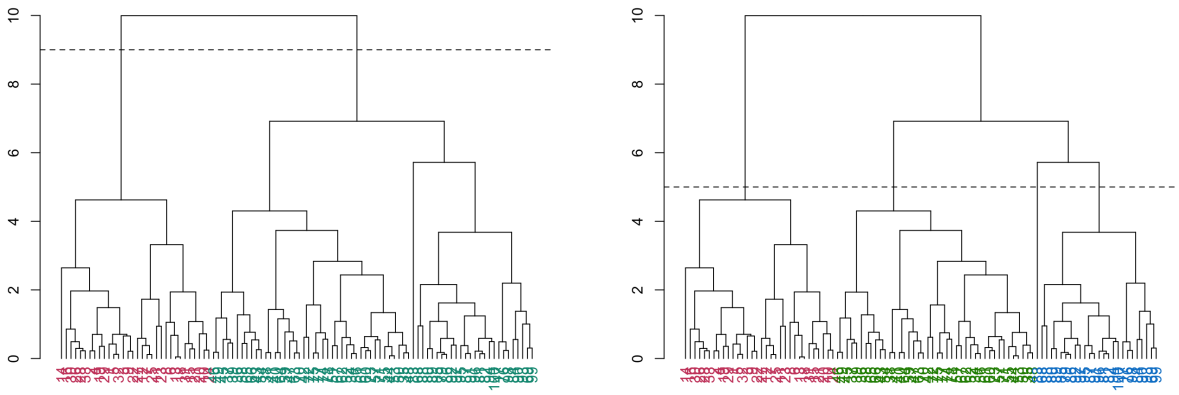


Each *leaf* of the dendrogram represents one of the 100 simulated data points.

As we move up the tree, leaves begin to fuse into branches, which correspond to observations that are similar to each other.

For any two observations, we can look for the point in the tree where branches containing those two observations are first fused.

How do we get clusters from the dendrogram?

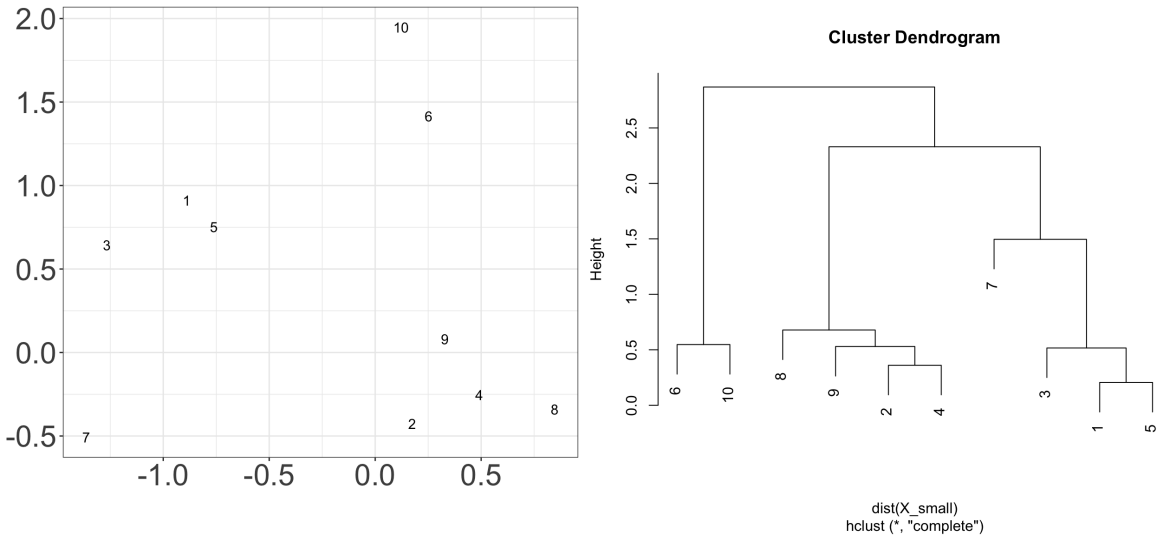


The term *hierarchical* refers to the fact that clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at a greater height.

3.2.2 Algorithm

First, we need to define some sort of *dissimilarity* metric between pairs of observations.

Then the algorithm proceeds iteratively.



More formally,

One issue has not yet been addressed.

How do we determine the dissimilarity between two clusters if one or bother of them contains multiple observations?

1.

2.

3.

4.

3.2.3 Choice of Dissimilarity Metric

3.3 Practical Considerations in Clustering

In order to perform clustering, some decisions should be made.

-
-
-

Each of these decisions can have a strong impact on the results obtained. What to do?