Clustering

Workshop on RNA-Seq

Nima Rafati

NBIS, SciLifeLab

Clustering

- What is clustering?
  - Clustering is an approach to classify/group data points.
- Why do we use clustering?
  - For exploring the data
  - To discover patterns in our data set
  - Identify outliers

Clustering Methods

- Centroid-based
- Density-based
- Distribution-based
- Hierarchical-based

Steps:

In short all clustering approach follows these steps:

- Calculate distance between data points
- Group | cluster the data based on similarities

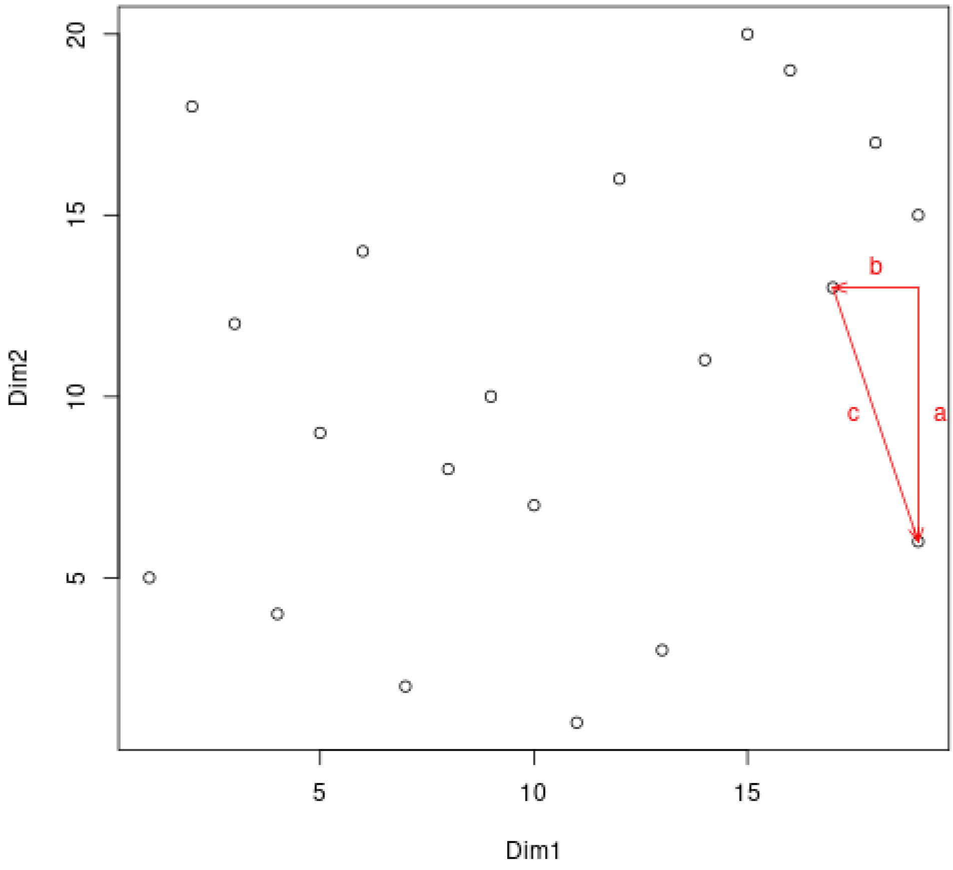
Distance can be measured in:

- In multidimensional space (raw data)
- In reduced space (i.e. top PCs)

Euclidean distance

- Euclidean distance is a straight line between two points

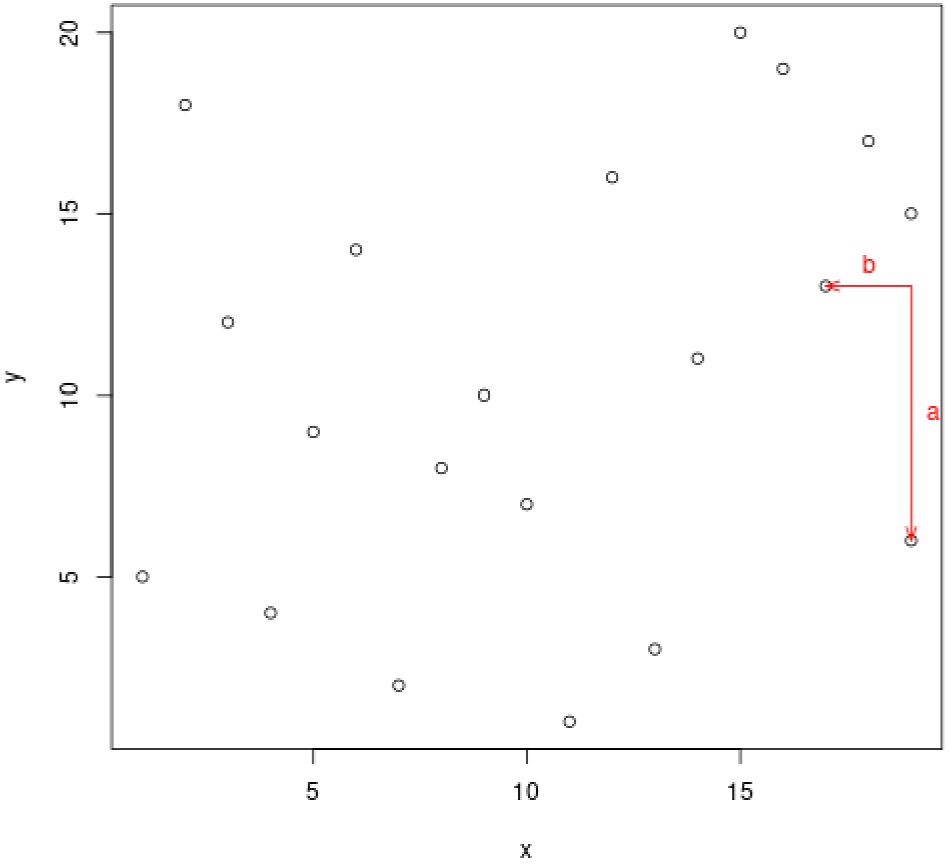
$$c^2 = a^2 + b^2$$



Manhattan distance

- Manhattan distance

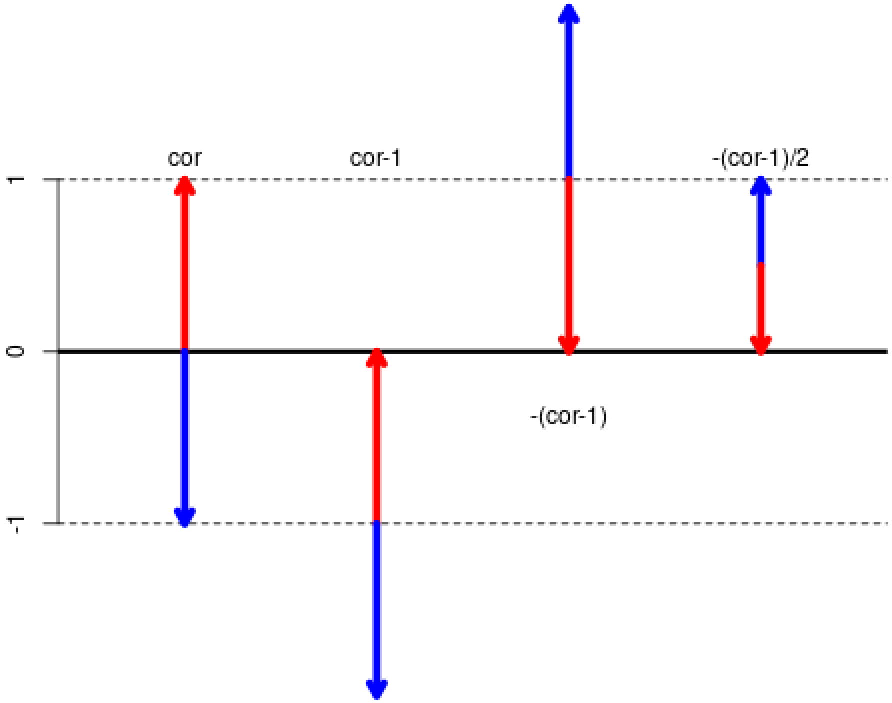
$$a + b$$



Inverted pairwise correlations

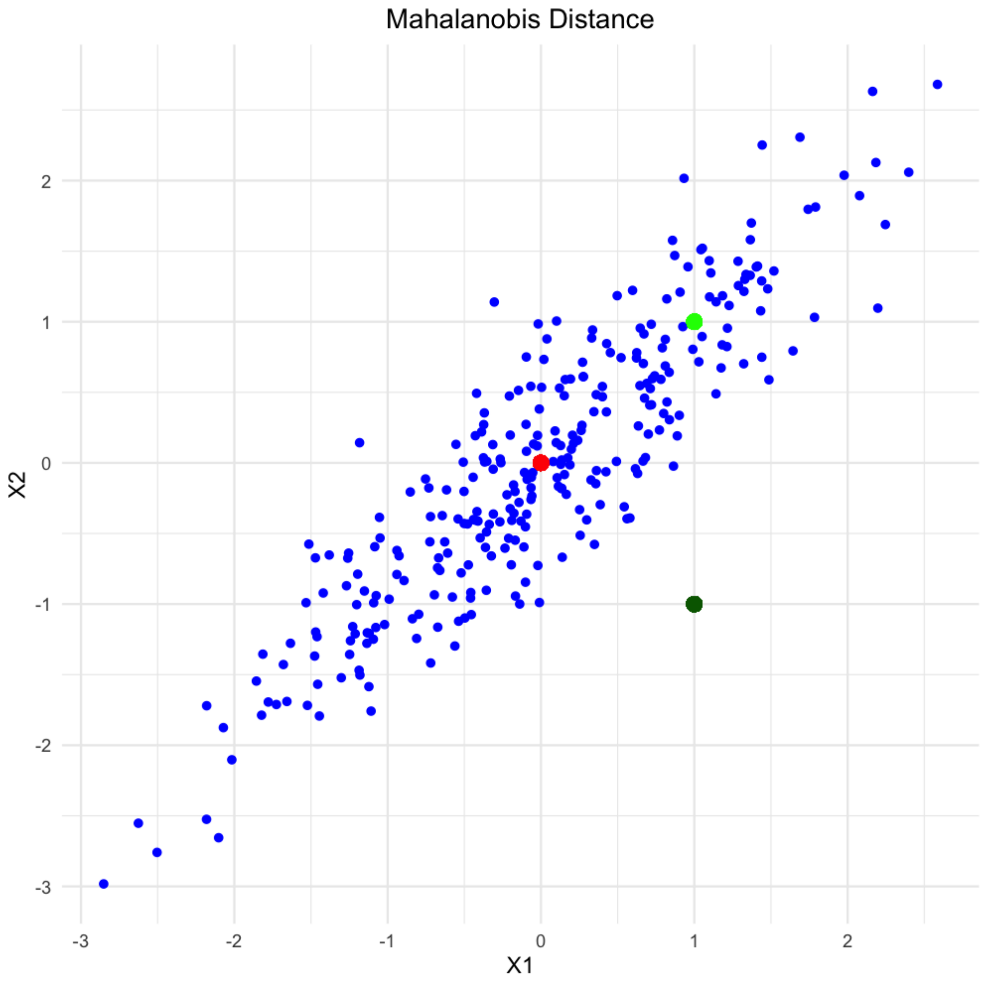
- Inverted pairwise correlations

$$dist = -(cor - 1)/2$$



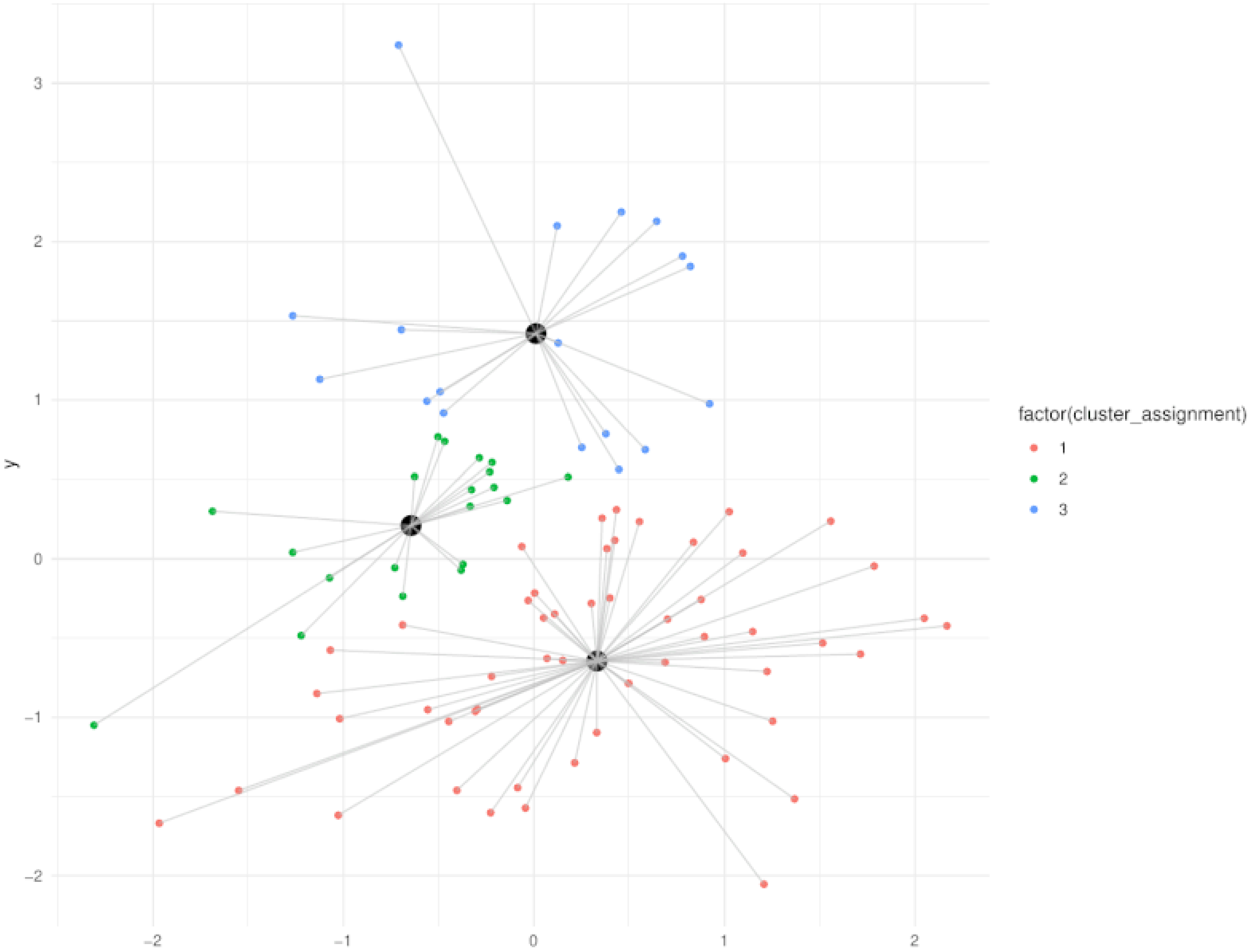
Mahalanobis Distance

- Despite of previous approach which was based on distance between data points, this method measures the distance between a data point and a distribution.



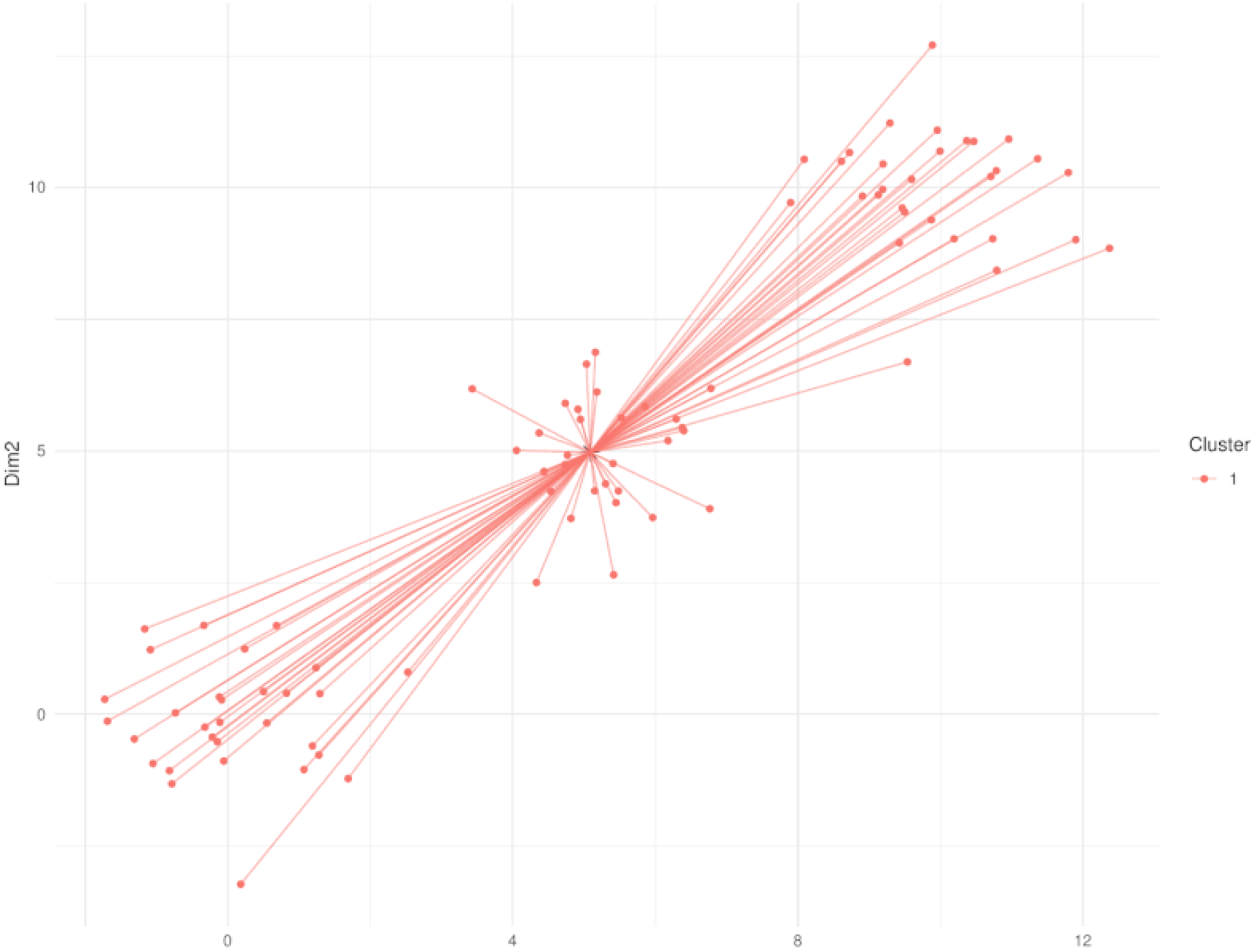
Centroid-based: K-means clustering

- One of the most commonly used clustering methods
- In this method the distance between data points and centroids is calculated
- Each data point is assigned to a cluster based on Euclidean distance from centroid.
- Dependent on number of K (clusters) new centroids are created


3
2
1
0
−1
−2
y
−2
−1
0
1
2
factor(cluster_assignment)
● 1
● 2
● 3

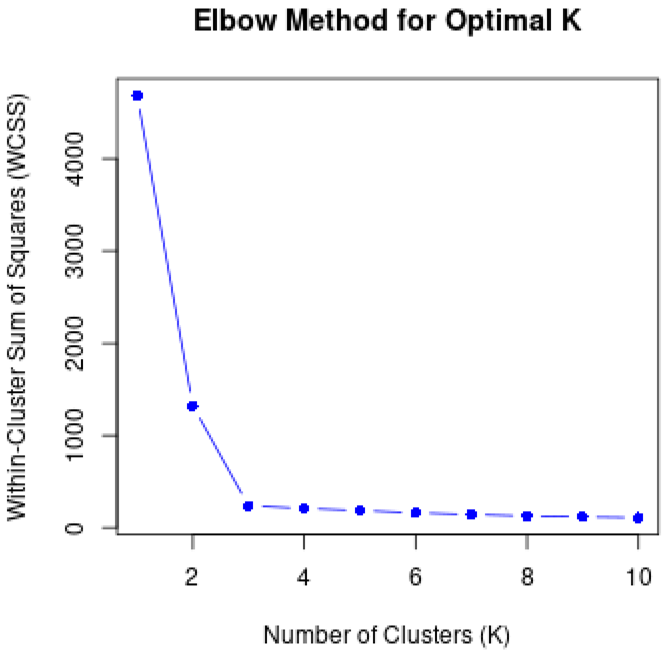
Centroid-based: K-means clustering

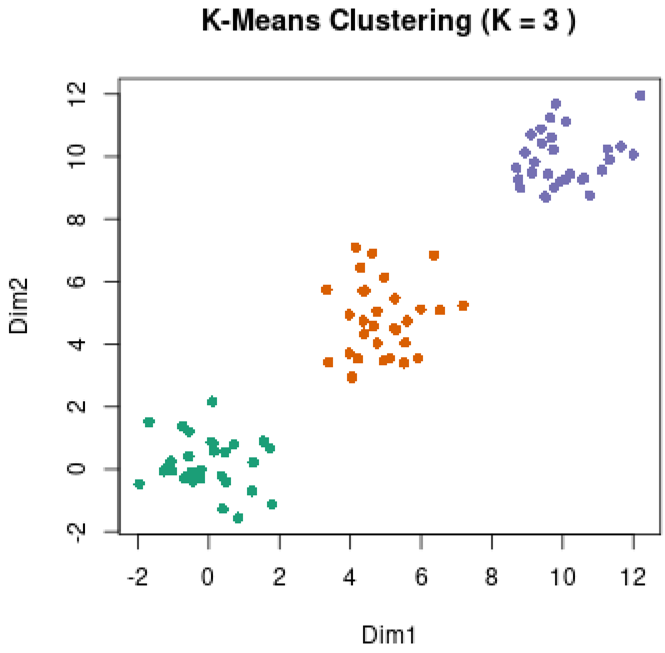
- One of the most commonly used clustering methods
- In this method the distance between data points and centroids is calculated
- Each data point is assigned to a cluster based on Euclidean distance from centroid.
- Dependent on number of K (clusters) new centroids are created


10
5
0
Dim2
0
4
8
12
Cluster
● 1

What is optimal K?

- The user needs to define the number of clusters:
  - **Elbow method.**
  - Gap statistics.
  - Average Silhouette method





Density-based clustering: DBSCAN

- This method identifies regions within your data distribution that exhibits high density of data points.

Distribution-based clustering: Guassian Mixture Model (GMM)

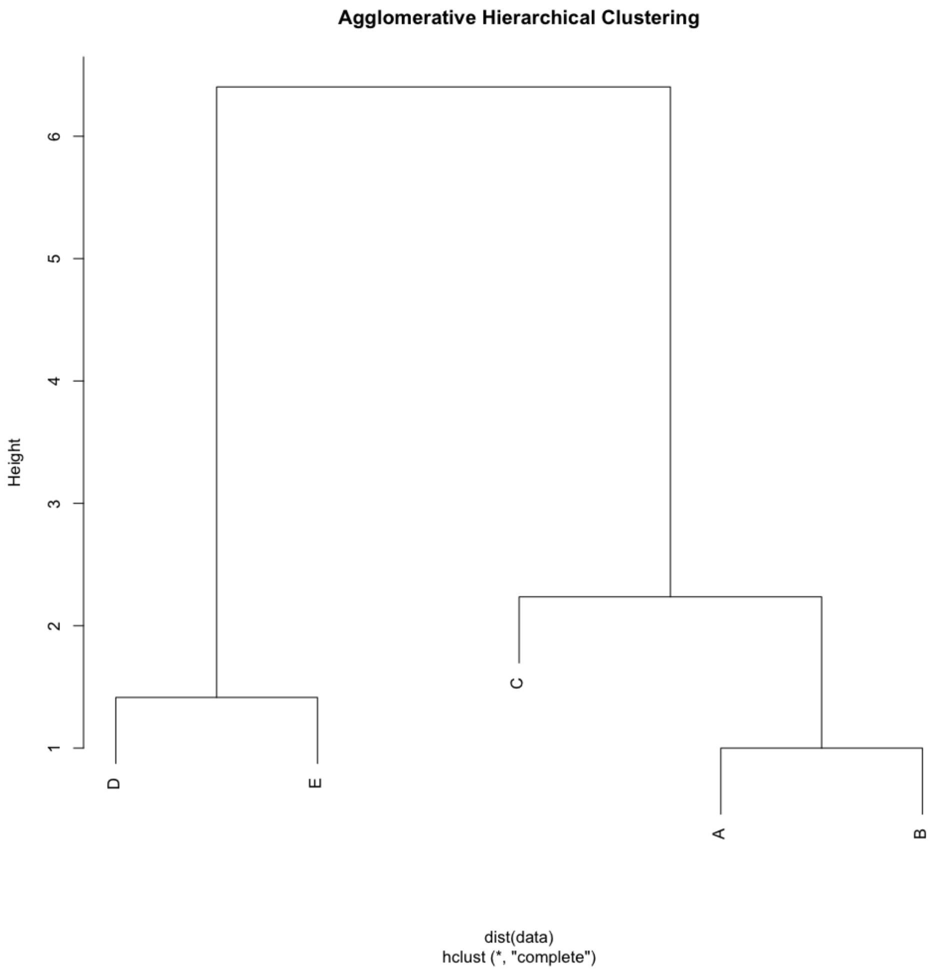
- It involves modeling the data points with probability distribution.
- In this method prior knowledge on distribution of your data is required. If you do not know the distribution of your data try another approach.
- You need to specify number of clusters.

Hierarchical-based clustering

- This approach creates a tree of clusters.
- Well suited for hierarchical data (e.g. taxonomies).
- Final output is a dendrogram representing the order decisions at each merge/division of clusters.
- Two approaches:
    - Agglomerative (Bottom-up): All data points are treated as clusters and the joins similar ones.
    - Divisive (Top-down): All data points are in one large clusters and recursively splits the most heterogeneous clusters.
- Number of clusters are decided after generating the tree.

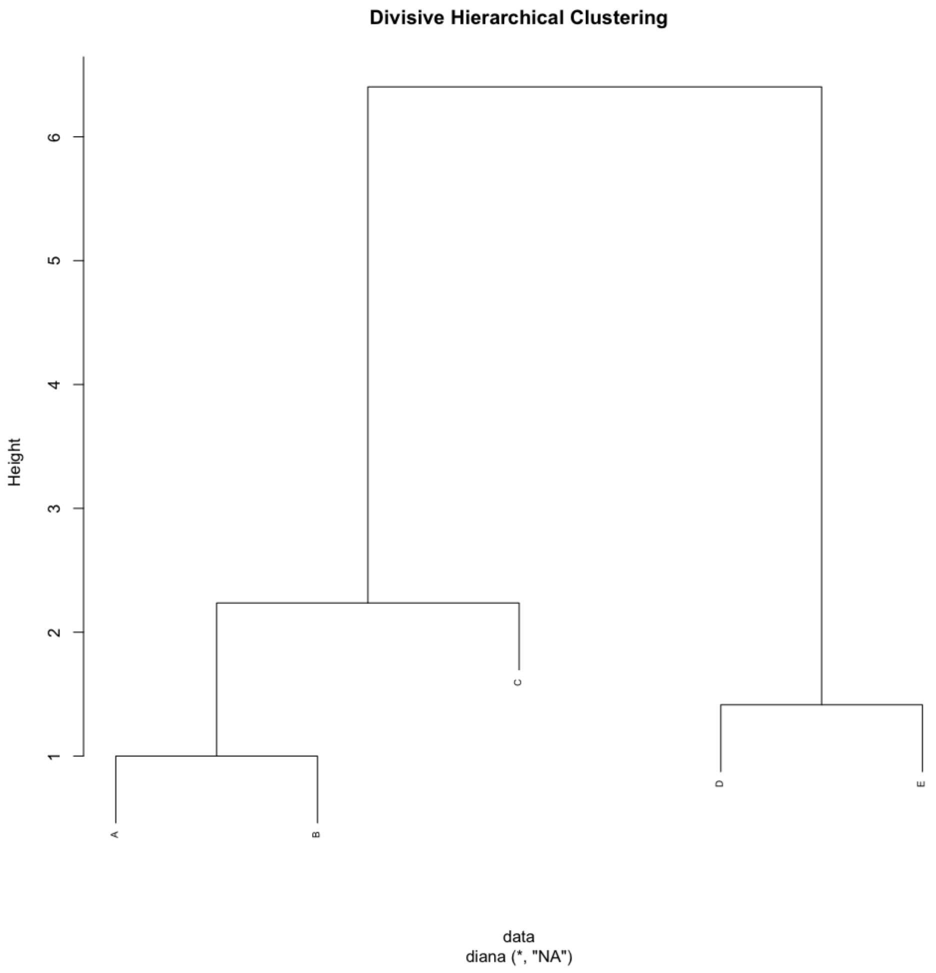
Hierarchical-based clustering

- Agglomerative clustering



Hierarchical-based clustering

- Divisive clustering

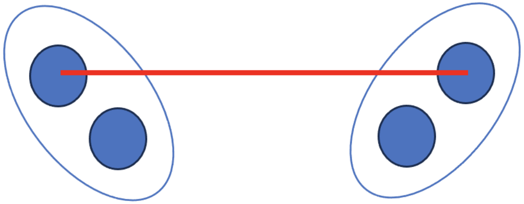


Linkage methods.

To combine clusters, it's essential to establish their positions relative to one another. The technique used to determine these positions is known as **Linkage**.

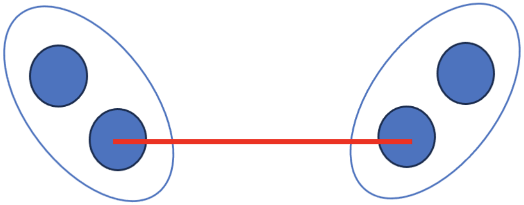
Complete Linkage

Maximum distance



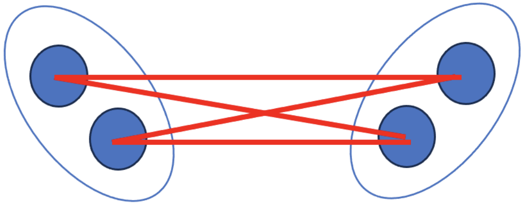
Single Linkage

Minimum distance



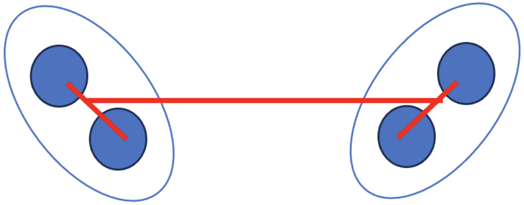
Average Linkage

Average of all points



Ward's Linkage

Minimize within cluster variance



Summary

- For bulk RNASeq you can perform clusteirng on raw or Z-Score scaled data.
- For the sample size is large (>10,000) you can perform clustering on PC. For instance in scRNASeq data.
- You always need to tune some parameters.
- K-means performs poorly on unbalanced data.
- On hierarchical clustering, some distance metrics need to be used with a certain linkage method.
- Checking clustering Robustness (a.k.a Ensemble perturbations):
  - Most clustering techniques will cluster random noise.
  - One way of testing this is by clustering on parts of the data (clustering bootstrapping)
  - Read more in Ronan et al (2016) Science Signaling).

Thank you. Questions?