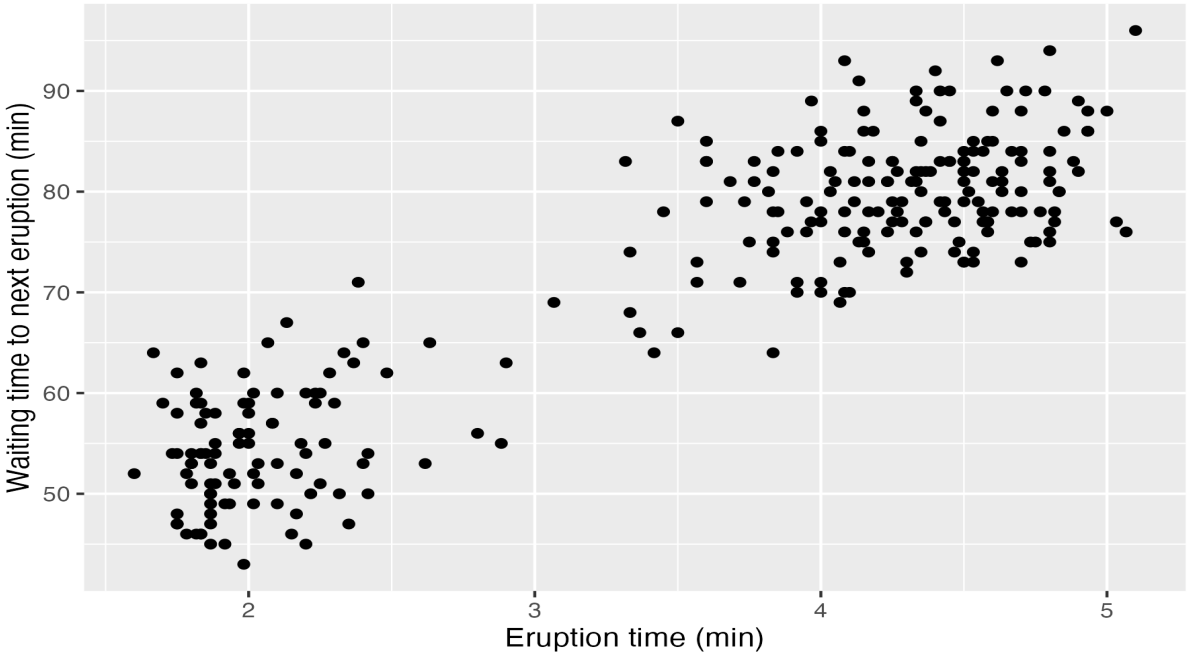
Cluster Analysis

Objective: Discover reasonable groupings of subjects, or members of some population

- Group medical patients with respect to how frequently different types of medical service are used
- Group potential customers with respect product preferences
- Group environmental habitats (areas of land) with respect to percentages of area covered by forest, grassland, water, marsh land, cultivated cropland, etc...
- Group foods relative to the content of various nutrients

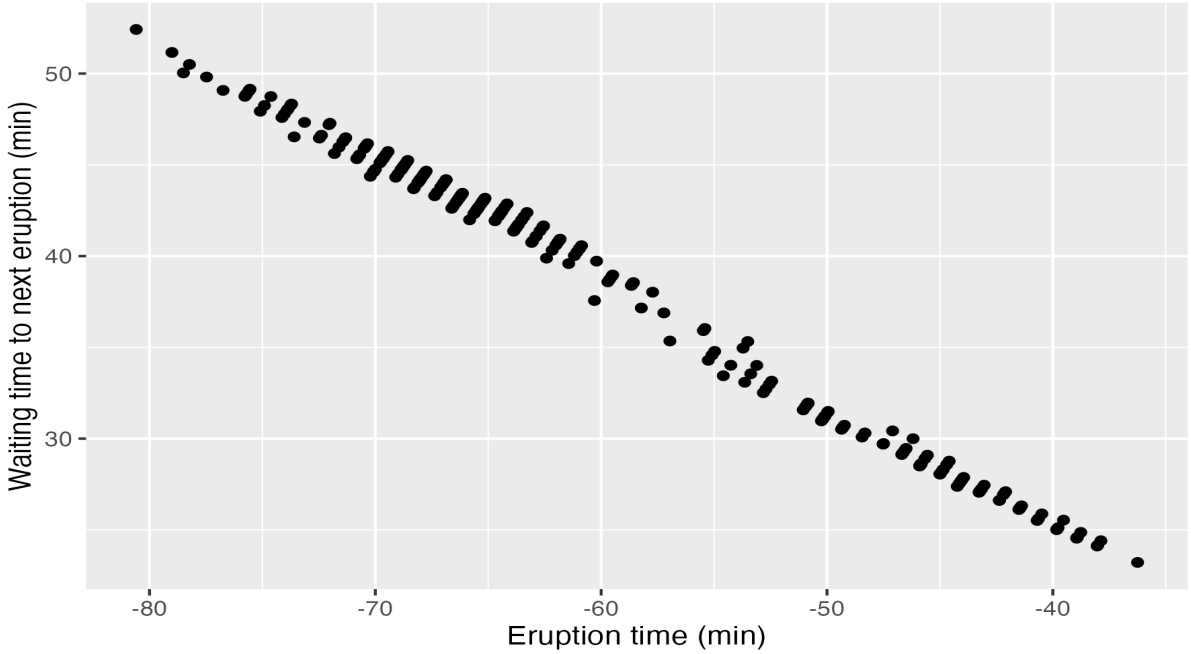
Example: Old Faithful Geyser

How many clusters/groups are there in the dataset?



Example: Old Faithful Geyser

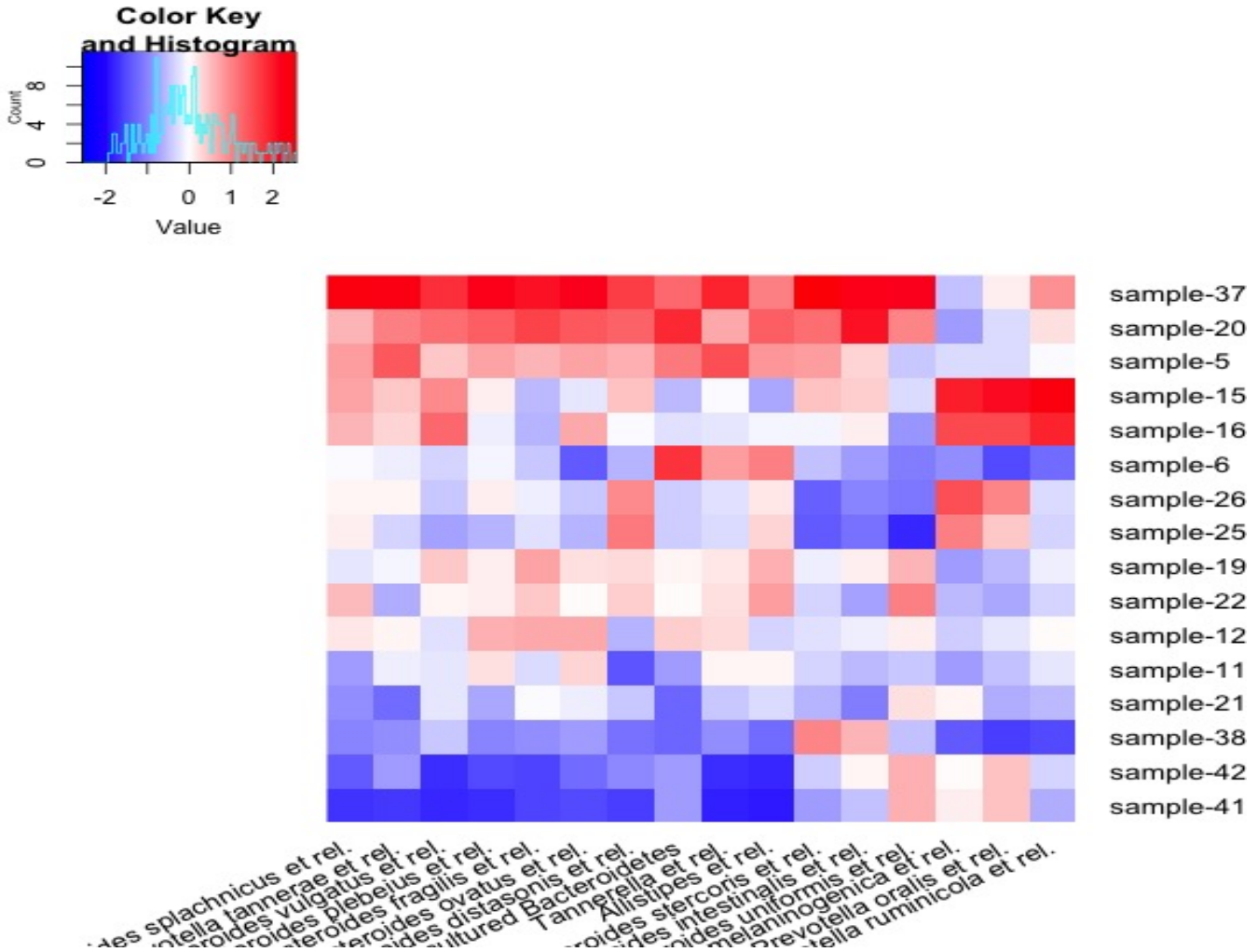
- Same data (after rotation), but with different axes.
- Is human perception reliable?



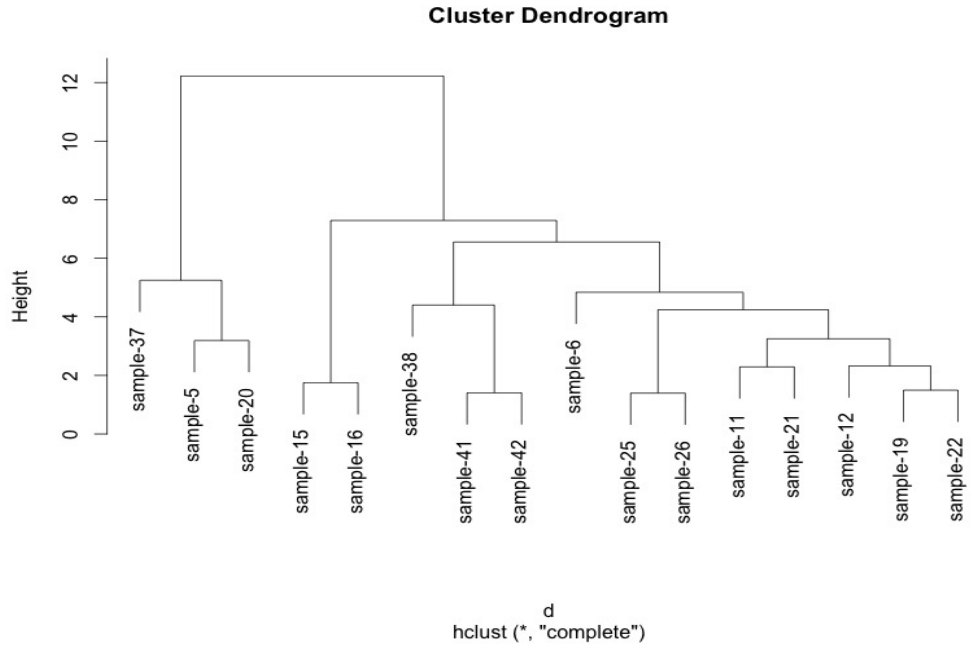
Three Main Clustering Methods

- Hierarchical clustering
  - Agglomerative (bottom-up)
  - Divisive (top-down)
- K-means clustering
- Model-based clustering

Heatmap


Color Key
and Histogram
Count
0
4
8
-2
0
1
2
Value
sample-37
sample-20
sample-5
sample-15
sample-16
sample-6
sample-26
sample-25
sample-19
sample-22
sample-12
sample-11
sample-21
sample-38
sample-42
sample-41

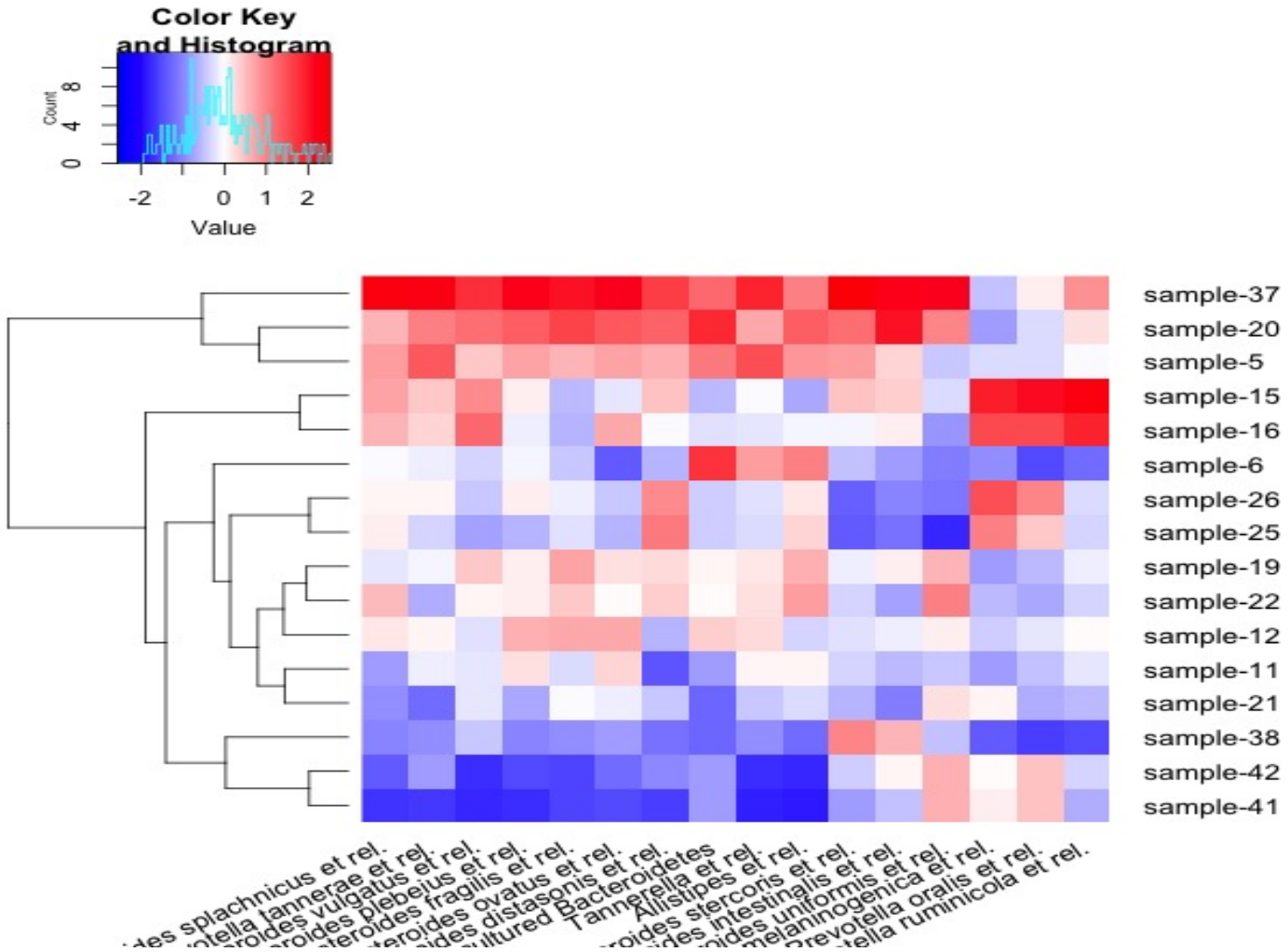
Dendrogram



Dendrogram: Graphical representation of hierarchical sequence of clustering assignments.

- Vertical axis: *distance* between clusters
- Horizontal axis: observations
- Dendrogram is a binary tree:
  - each node represents a cluster
  - each leaf node is the observation
  - root node is a cluster with all observations

Hierarchical Clustering



Cluster Analysis

Suppose there are n individuals (or cases) to be clustered, and a set of p traits is measured on each case.

$$\mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{p1} \end{bmatrix} \; \mathbf{X}_2 = \begin{bmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{p2} \end{bmatrix} \; \cdots \; \mathbf{X}_n = \begin{bmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{pn} \end{bmatrix}$$

- Not necessarily a random sample
- Possibly the entire population
- **Better data $\Rightarrow$ better clusters**: The decision on which variables to collect information is important. There may be 10 variables that can jointly produce very useful clusters. If your data set does not include most of them, then your cluster analysis may give disappointing results.

Cluster Analysis

Procedures for constructing clusters share some basic features:

(1) A measure of **distance** (dissimilarity) between each pair of cases.

(2) A **linkage** procedure for determining the distance (or similarity) between two groups of individuals.

Distance Between Observations/Cases

There are many distance measures between pairs of cases that could be used. The choice of the distance (or similarity) measure will affect the results of the cluster analysis.

Euclidean distance: $$d(\mathbf{X}_i,\ \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j)}$$

Statistical distance: $$d(\mathbf{X}_i,\ \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)'\mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mathbf{X}_j)}$$

Standardized Distance: $$d(\mathbf{X}_i,\ \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)'\left[diag(S)\right]^{-1}(\mathbf{X}_i - \mathbf{X}_j)}$$

Absolute Values: $$d(\mathbf{X}_i,\ \mathbf{X}_j) = \frac{1}{p}\sum_{k=1}^{p}|X_{ki} - X_{kj}|$$

Distance Between Observations/Cases

Correlations:	$d(item_i,\ item_j) = 1 - \lvert r_{ij} \rvert$
Correlations:	$d(item_i,\ item_j) = 1 - r_{ij}^2$
Rankings:	How many judges have both items among their top 5
Kendall's Tau	$d(item_i,\ item_j) = 1 - \lvert \tau_{ij} \rvert$

In this case you want to group items with respect to their level of association, or correlation.

Hierarchical Clustering

- Start with a dissimilarity matrix
- Join the 2 most closely related objects
- Remove the joined objects from the matrix
- Add a new object that represents the join group (e.g., complete, average, single)
- Repeat until no objects remain in the matrix

Key R function: hclust

Linkage methods

Linkage is a function $d(C_1, C_2)$ that takes two groups/clusters C_1, C_2 and return a dissimilarity score between them.

- Single linkage (nearest-neighbor linkage)

$$d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d_{ij}$$

  where $d_{ij} := d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between the i-th element in C_1 and the j-th element in C_2.

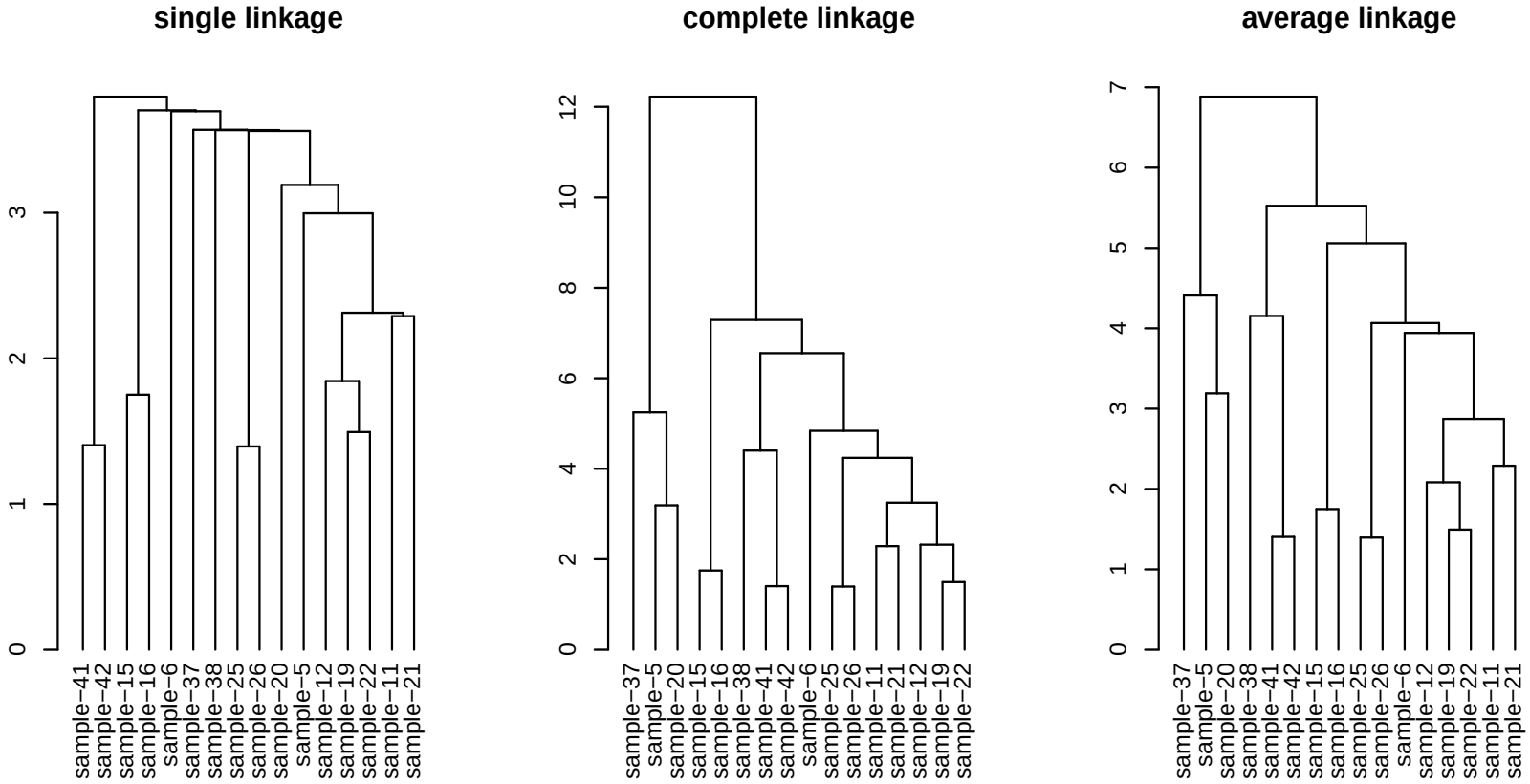
- Complete linkage (furthest-neighbor linkage)

$$d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d_{ij}$$

- Average linkage

$$d(C_1, C_2) = \text{Average}_{i \in C_1, j \in C_2} d_{ij} = \frac{1}{|C_1| \cdot |C_2|} \sum_{i \in C_1, j \in C_2} d_{ij}$$

Linkage Types: Example



Example: Clustering Foods

Cluster $n = 8$ foods based on $p = 5$ traits. The foods are:

- Braised Beef (BB)
- Hamburger (HR)
- Beef Roast (BR)
- Beef Steak (BS)
- Canned Beef (BC)
- Broiled Chicken (CB)
- Canned Chicken (CC)
- Beef Heart (BH)

Example: Clustering Foods

The traits for a three ounce portion are:

1. $X_1 =$ Energy (calories)
2. $X_2 =$ Protein (grams)
3. $X_3 =$ Fat (grams)
4. $X_4 =$ Calcium (mg)
5. $X_5 =$ Iron (mg)

Clustering Foods: Data

	BB	HR	BR	BS	BC	CB	CC	BH
X_1	340	245	420	375	180	115	170	160
X_2	20	21	15	19	21	20	25	26
X_3	28	17	39	32	10	3	7	5
X_4	9	9	7	9	17	8	12	14
X_5	2.6	2.7	2.0	2.6	3.7	1.4	1.5	5.9

Clustering Foods: Data

Using Euclidean distance to measure dissimilarity between foods
$d(BB, CB) =$

$$\sqrt{(304-115)^2+(20-20)^2+(28-3)^2+(9-8)^2+(2.6-1.4)^2}$$
$$= 226.4$$

The distance matrix is

$$\begin{bmatrix}
 & BB & HR & BR & BS & BC & CB & CC & BH \\
BB & 0 & 95.6 & 80.9 & 35.2 & 161.2 & 226.4 & 171.4 & 181.7 \\
HR & & 0 & 176.5 & 130.9 & 65.9 & 130.8 & 75.8 & 86.2 \\
BR & & & 0 & 45.8 & 242.1 & 307.2 & 252.3 & 262.6 \\
BS & & & & 0 & 196.4 & 261.6 & 201.6 & 216.9 \\
BC & & & & & 0 & 66.1 & 12.2 & 21.3 \\
CB & & & & & & 0 & 55.5 & 46.1 \\
CC & & & & & & & 0 & 11.3 \\
BH & & & & & & & & 0
\end{bmatrix}$$

Single Linkage

Step 1: Merge the two nearest neighbors (CC and BH) and construct a new matrix of distances

	BB	HR	BR	BS	BC	CB	(CC, BH)
BB	0	95.6	80.9	35.2	161.2	226.4	171.4
HR		0	176.5	130.9	65.9	130.8	75.8
BR			0	45.8	242.1	307.2	252.3
BS				0	196.4	261.6	201.6
BC					0	66.1	12.2
CB						0	46.1
(CC, BH)							0

Single Linkage

Step 2: Merge the two nearest clusters (CC and BH) and (BC) construct a new matrix of distances

	BB	HR	BR	BS	CB	(BC, CC, BH)
BB	0	95.6	80.9	35.2	226.4	161.2
HR		0	176.5	130.9	130.8	65.9
BR			0	45.8	307.2	242.1
BS				0	261.6	196.4
CB					0	46.1
(BC, CC, BH)						0

Single Linkage

Step 3: Merge (BB) with (BS) and compute new distances

$$\begin{bmatrix} & (BB,BS) & HR & BR & CB & (BC,CC,BH) \\ (BB,BS) & 0 & 95.6 & 45.8 & 226.4 & 161.2 \\ HR & & 0 & 176.5 & 130.8 & 65.9 \\ BR & & & 0 & 307.2 & 242.1 \\ CB & & & & 0 & 46.1 \\ (BC,CC,BH) & & & & & 0 \end{bmatrix}$$

Step 4: Merge (BR) with (BB, BS), compute new distances

$$\begin{bmatrix} & (BB,BS,BR) & HR & CB & (BC,CC,BH) \\ (BB,BS,BR) & 0 & 95.6 & 226.4 & 161.2 \\ HR & & 0 & 130.8 & 65.9 \\ CB & & & 0 & 46.1 \\ (BC,CC,BH) & & & & 0 \end{bmatrix}$$

Single Linkage

Step 5: Merge (CB) with (BC, CC, BH), compute new distances

$$\begin{bmatrix} & (BB,BS,BR) & HR & (BC,CB,CC,BH) \\ (BB,BS,BR) & 0 & 95.6 & 161.2 \\ HR & & 0 & 65.9 \\ (BC,CB,CC,BH) & & & 0 \end{bmatrix}$$

Step 6: Merge (HR) with (BC, CB, CC, BH)

$$\begin{bmatrix} & (BB,BS,BR) & (HR,BC,CB,CC,BH) \\ (BB,BS,BR) & 0 & 95.6 \\ (HR,BC,CB,CC,BH) & & 0 \end{bmatrix}$$

Step 7: Merge (BB, BS, BR) with (HR, BC, CB, CC, BH)

Clustering Foods: Data

The measurements for the eight nutrients are now recorded as percentage of minimum daily requirements for an average adult provided by a three ounce portion

	BB	HR	BR	BS	BC	CB	CC	BH
X_1	11	8	13	12	6	4	5	5
X_2	29	30	21	27	31	29	36	37
X_3	28	17	39	32	10	3	7	5
X_4	1	1	1	1	2	1	2	2
X_5	26	27	20	26	37	14	15	59

Clustering Foods: Data

Using Euclidean distance to measure dissimilarities, the distance matrix is

	BB	HR	BR	BS	BC	CB	CC	BH
BB	0	11.5	15.0	4.6	21.8	28.6	25.5	41.5
HR		0	25.3	15.8	12.5	19.5	17.0	35.0
BR			0	11.1	35.8	38.4	36.6	54.8
BS				0	25.7	32.5	29.6	44.4
BC					0	24.2	22.8	23.4
CB						0	8.3	45.8
CC							0	44.1
BH								0

Single Linkage

Step 1: Merge the two nearest neighbors (BB) and (BS) and construct a new matrix of distances

	(BB,BS)	HR	BR	BC	CB	CC	BH
(BB,BS)	0	11.5	15.0	21.8	28.6	25.5	41.5
HR		0	25.3	12.5	19.5	17.0	35.0
BR			0	35.8	38.4	36.6	54.8
BC				0	24.2	22.8	23.4
CB					0	8.3	45.8
CC						0	44.1
BH							0

Single Linkage

Step 2: Merge (CB) with (CC) and construct a new matrix of distances

$$\begin{bmatrix}
 & (BB,BS) & HR & BR & BC & (CB,CC) & BH \\
(BB,BS) & 0 & 11.5 & 15.0 & 21.8 & 25.5 & 41.5 \\
HR & & 0 & 25.3 & 12.5 & 17.0 & 35.0 \\
BR & & & 0 & 35.8 & 36.6 & 54.8 \\
BC & & & & 0 & 22.8 & 23.4 \\
(CB,CC) & & & & & 0 & 44.1 \\
BH & & & & & & 0
\end{bmatrix}$$

Single Linkage

Step 3: Merge (HR) with (BB, BS) and compute new distances

$$\begin{bmatrix} & (BB,BS,HR) & BR & BC & (CB,CC) & BH \\ (BB,BS,HR) & 0 & 15.0 & 12.5 & 17.0 & 35.0 \\ BR & & 0 & 35.8 & 36.6 & 54.8 \\ BC & & & 0 & 22.8 & 23.4 \\ (CB,CC) & & & & 0 & 44.1 \\ BH & & & & & 0 \end{bmatrix}$$

Step 4: Merge (BC) with (BB, BS, HR), compute new distances

$$\begin{bmatrix} & (BB,BS,HR,BC) & BR & (CB,CC) & BH \\ (BB,BS,HR,BC) & 0 & 15.0 & 17.0 & 23.4 \\ BR & & 0 & 36.6 & 54.8 \\ (CB,CC) & & & 0 & 44.1 \\ BH & & & & 0 \end{bmatrix}$$

Single Linkage

Step 5: Merge (BR) with (BB, BS, HR, BC), compute new distances

$$\begin{bmatrix} & (BB,BS,HR,BC,BR) & (CB,CC) & BH \\ (BB,BS,HR,BC,BR) & 0 & 17.0 & 23.4 \\ (CB,CC) & & 0 & 44.1 \\ BH & & & 0 \end{bmatrix}$$

Step 6: Merge (BB, BS, HR, BC, BR) with (CB, CC)

$$\begin{bmatrix} & (BB,BS,HR,BC,BR,CB,CC) & BH \\ (BB,BS,HR,BC,BR,CB,CC) & 0 & 23.4 \\ BH & & 0 \end{bmatrix}$$

Step 7: Merge (BB, BS, HR, BC, BR, CB, CC) with (BH)

Clustering Foods: Data

Using Euclidean distance to measure dissimilarity between foods (using the original measurements)

The distance matrix is

	BB	HR	BR	BS	BC	CB	CC	BH
BB	0	95.6	80.9	35.2	161.2	226.4	171.4	181.7
HR		0	176.5	130.9	65.9	130.8	75.8	86.2
BR			0	45.8	242.1	307.2	252.3	262.6
BS				0	196.4	261.6	201.6	216.9
BC					0	66.1	12.2	21.3
CB						0	55.5	46.1
CC							0	11.3
BH								0

Complete Linkage

Step 1: Merge the two nearest neighbors (CC) and (BH) and construct a new matrix of distances

	BB	HR	BR	BS	BC	CB	(CC, BH)
BB	0	95.6	80.9	35.2	161.2	226.4	181.7
HR		0	176.5	130.9	65.9	130.8	86.2
BR			0	45.8	242.1	307.2	262.6
BS				0	196.4	261.6	216.9
BC					0	66.1	21.3
CB						0	55.5
(CC, BH)							0

Complete Linkage

Step 2: Merge (CC, BH) with (BC) construct a new matrix of distances

	BB	HR	BR	BS	CB	(CC, BH, BC)
BB	0	95.6	80.9	35.2	226.4	181.7
HR		0	176.5	130.9	130.8	86.2
BR			0	45.8	307.2	262.6
BS				0	261.6	216.9
CB					0	66.1
(CC, BH, BC)						0

Complete Linkage

Step 3: Merge (BB) with (BS) and compute new distances

	(BB, BS)	HR	BR	CB	(CC, BH, BC)
(BB, BS)	0	130.9	80.9	261.6	216.9
HR		0	176.5	130.8	86.2
BR			0	307.2	262.6
CB				0	66.1
(CC, BH, BC)					0

Step 4: Merge (CB) with (CC, BH, BC), compute new distances

	(BB, BS)	HR	BR	(CB, CC, BH, BC)
(BB, BS)	0	130.9	80.9	262.6
HR		0	176.5	130.8
BR			0	307.2
(CB, CC, BH, BC)				0

Complete Linkage

Step 5: Merge (BR) with (BB, BS), compute new distances

$$\begin{bmatrix} & (BB,BS,BR) & HR & (CB,CC,BH,BC) \\ (BB,BS,BR) & 0 & 176.5 & 307.2 \\ HR & & 0 & 130.8 \\ (CB,CC,BH,BC) & & & 0 \end{bmatrix}$$

Step 6: Merge (HR) with (CB, CC, BH, BC)

$$\begin{bmatrix} & (BB,BS,BR) & (HR,CB,CC,BH,BC) \\ (BB,BS,BR) & 0 & 307.2 \\ (HR,CB,CC,BH,BC) & & 0 \end{bmatrix}$$

Step 7: Merge (BB, BS, BR) with (HR, CB, CC, BH, BC)

Average Linkage

The distance between cluster A and cluster B is a weighted average of the distances between pairs of individuals, with one member of the pair taken from each cluster

$$d(A, B) = \sum_{i \in A} \sum_{j \in B} w_{ij} d(\mathbf{X}_{Ai},\ \mathbf{X}_{Bj})$$

where w_{ij} is some set of weights.

Equal weights are most commonly used:

$$w_{ij} = \frac{1}{n_A n_B} = \frac{1}{\text{number of pairs}}$$

Ward's Method

- Initially each individual is in a cluster of size one.
- At each step, compute the within cluster sum of squared Euclidean distances from the mean vector (or centroid) of the cluster. For the i-th cluster this is

$$SS_i = \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)'(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)$$

- Then, compute the total within cluster variation as

$$SST = \sum_{i=1}^{I} SS_i$$

- Merge the two clusters that result in the smallest increase in SST

Ward's Method

- The square of the distance between Cluster A and Cluster B is
  (SST after combining clusters A and B)
  – (SST before combining clusters A and B)

$$= \frac{n_A n_B}{n_A + n_B}(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)'(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)$$

- Merge the two clusters that result in the smallest increase in SST
- Note that the factor $\frac{n_A n_B}{n_A + n_B}$ causes Ward's method to pull in small clusters and tend to make clusters with more equal numbers of individuals in the clusters.
- This is not a good method for isolating outliers.

Centroid Method

- The square of the distance between Cluster A and Cluster B is simply the distance between the centroids for the two clusters

$$d(\bar{\mathbf{X}}_A, \bar{\mathbf{X}}_B) = (\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)'(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)$$

- At each step, merge the two clusters with the closest centroids

- Relative to Ward's method, the centroid method tends to make clusters with more unequal numbers of individuals in the clusters.

- This is a good method for isolating outliers.

Non-hierarchical Clustering

- At each step of the clustering algorithm, new clusters may be formed by breaking apart clusters from the previous step and merging some of the pieces

- No tree structure exists

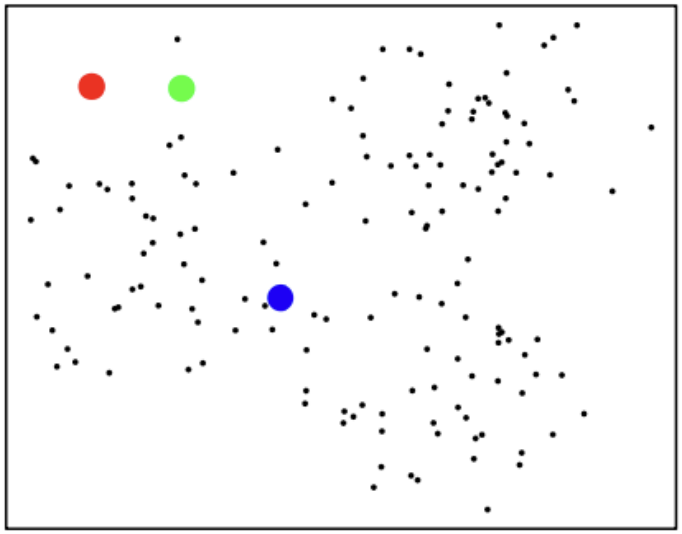
- No monotone measure of the strength of clustering

- Generally requires an iterative algorithm with potential to use substantial amounts of computer time.

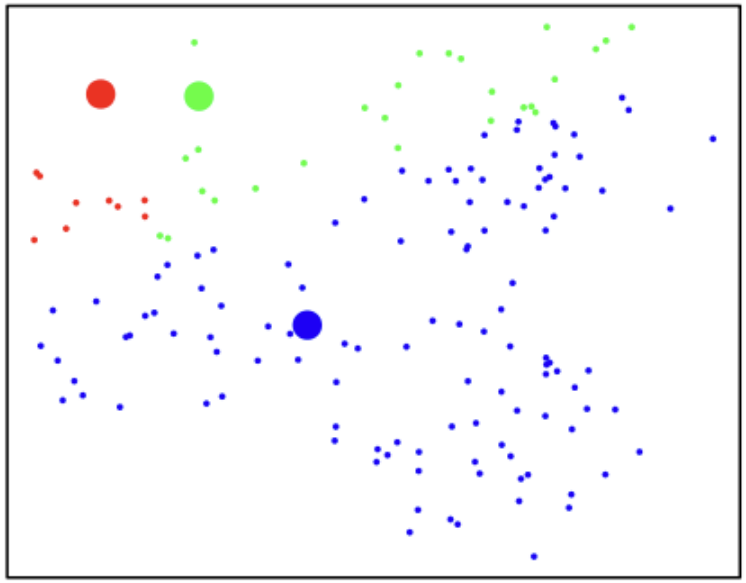
- Hundreds of non-hierarchical clustering methods have been developed. The "k-means" strategy is most frequently used.

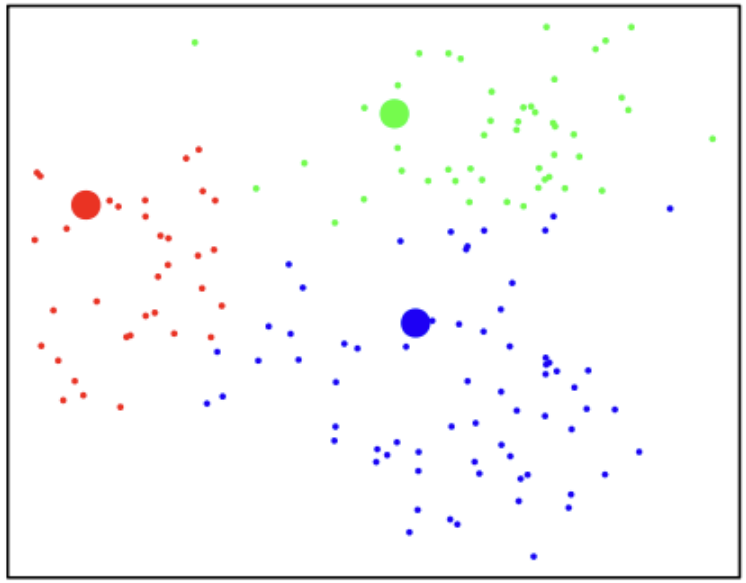
K-means Clustering

- **Initialize**: Pick K random points as cluster centers
- **Iterate**:
  - Assign points to closest cluster center
  - Update cluster center location to the mean of the assigned points
- **Stop** when no points change cluster assignment (convergence)

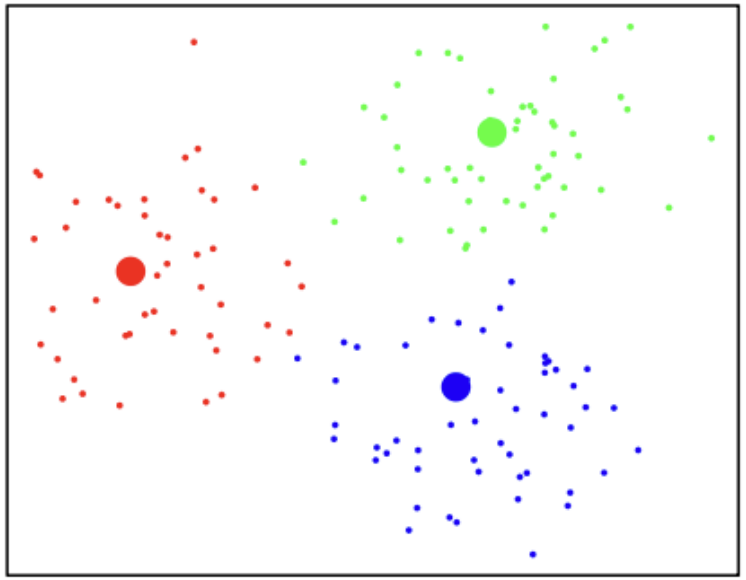


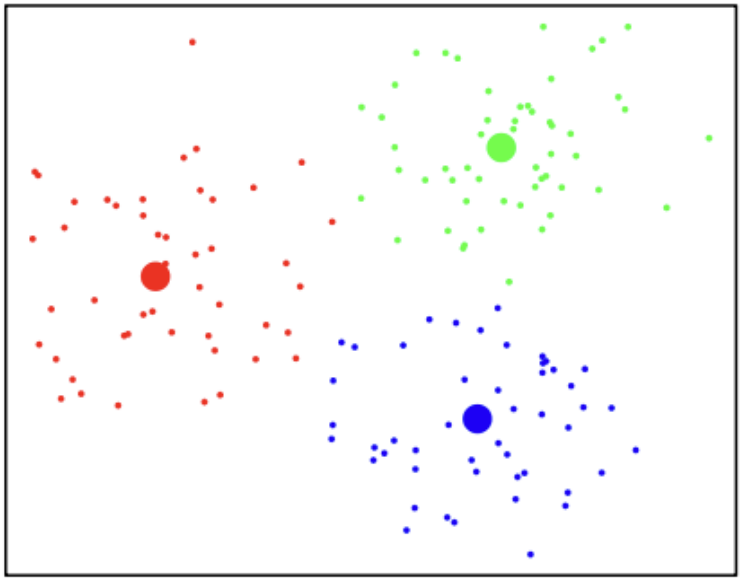
K-means Clustering



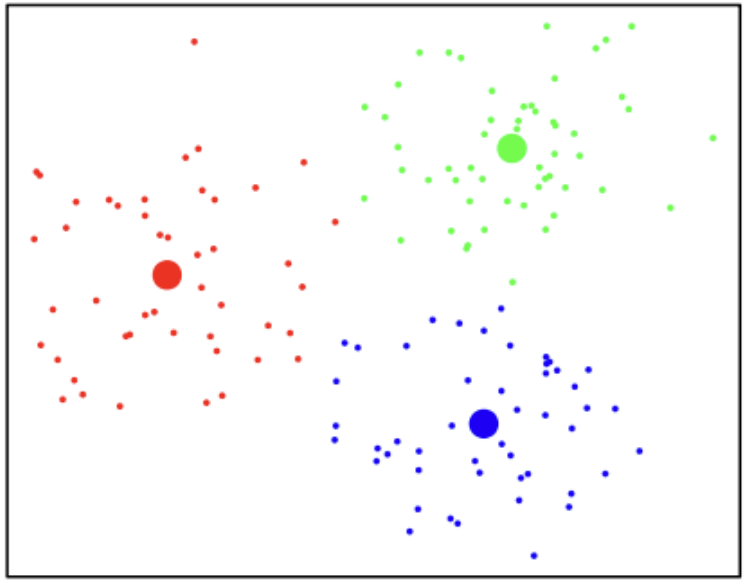


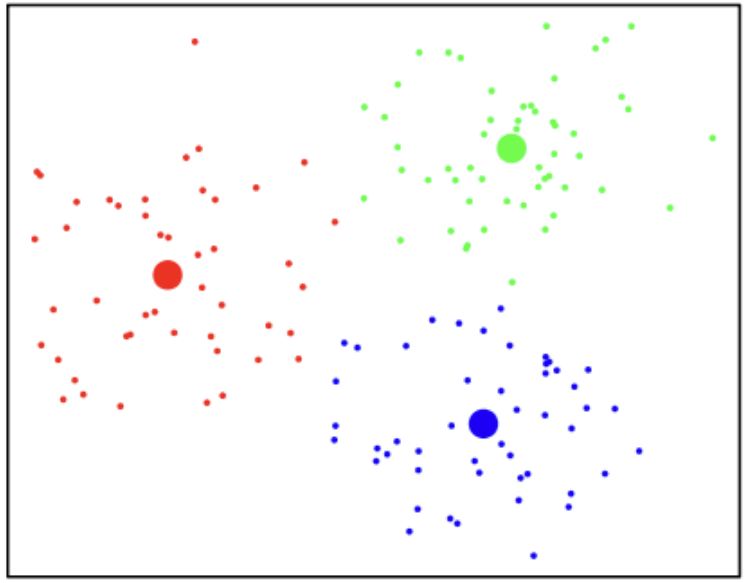
K-means Clustering





K-means Clustering





K-means Clustering Example

This small data set has only four cases ($\mathbf{A}, \mathbf{B}, \mathbf{C},$ and $\mathbf{D}$) and two variables are measure on each case. The data vectors are

$$\mathbf{A} = \begin{bmatrix} 5 \\ 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} -3 \\ -2 \end{bmatrix}$$

- Start with initial clusters ($\mathbf{A}$, $\mathbf{B}$) and ($\mathbf{C}$, $\mathbf{D}$)
- The centroids for the initial clusters are

$$\mathbf{C}_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \text{ and } \mathbf{C}_2 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

K-means Clustering Example

- Compute the distance of each case from each centroid. Use Euclidean distance:

$$d(\mathbf{A}, \mathbf{C}_1) = \sqrt{(5-2)^2 + (3-2)^2} = \sqrt{10} \qquad d(\mathbf{A}, \mathbf{C}_2) = \sqrt{(5-(-1))^2 + (3-(-2))^2} = \sqrt{61}$$

$$d(\mathbf{B}, \mathbf{C}_1) = \sqrt{(-1-2)^2 + (1-2)^2} = \sqrt{10} \qquad d(\mathbf{B}, \mathbf{C}_2) = \sqrt{(-1-(-1))^2 + (1-(-2))^2} = \sqrt{9}$$

$$d(\mathbf{C}, \mathbf{C}_1) = \sqrt{(1-2)^2 + (-2-2)^2} = \sqrt{17} \qquad d(\mathbf{C}, \mathbf{C}_2) = \sqrt{(1-(-1))^2 + (-2-(-2))^2} = \sqrt{4}$$

$$d(\mathbf{D}, \mathbf{C}_1) = \sqrt{(-3-2)^2 + (-2-2)^2} = \sqrt{41} \qquad d(\mathbf{D}, \mathbf{C}_2) = \sqrt{(-3-(-1))^2 + (-2-(-2))^2} = \sqrt{4}$$

- Make new clusters: $(\mathbf{A})$ and $(\mathbf{B},\ \mathbf{C},\ \mathbf{D})$ and compute their centroids $\mathbf{C}_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ and $\mathbf{C}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

K-means Clustering Example

- Compute the distance of each case from each of the new centroids.

$$d(\mathbf{A}, \mathbf{C}_1) = \sqrt{(5-5)^2 + (3-3)^2} = \sqrt{0} \qquad d(\mathbf{A}, \mathbf{C}_2) = \sqrt{(5-(-1))^2 + (3-(-1))^2} = \sqrt{52}$$

$$d(\mathbf{B}, \mathbf{C}_1) = \sqrt{(-1-5)^2 + (1-3)^2} = \sqrt{40} \qquad d(\mathbf{B}, \mathbf{C}_2) = \sqrt{(-1-(-1))^2 + (1-(-1))^2} = \sqrt{4}$$

$$d(\mathbf{C}, \mathbf{C}_1) = \sqrt{(1-5)^2 + (-2-3)^2} = \sqrt{41} \qquad d(\mathbf{C}, \mathbf{C}_2) = \sqrt{(1-(-1))^2 + (-2-(-1))^2} = \sqrt{5}$$

$$d(\mathbf{D}, \mathbf{C}_1) = \sqrt{(-3-5)^2 + (-2-3)^2} = \sqrt{89} \qquad d(\mathbf{D}, \mathbf{C}_2) = \sqrt{(-3-(-1))^2 + (-2-(-1))^2} = \sqrt{5}$$

- Make new clusters: $(\mathbf{A})$ and $(\mathbf{B},\ \mathbf{C},\ \mathbf{D})$. The clusters did not change; centroids are still $\mathbf{C}_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ and $\mathbf{C}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

- Stop: the K-means algorithm has converged to produce clusters $(\mathbf{A})$ and $(\mathbf{B},\ \mathbf{C},\ \mathbf{D})$

K-means Clustering Example

Some K-means algorithms continuously update the clusters and their centroids, instead of making a complete pass through all cases in the data to update the clusters. This is illustrated below.

- Start with initial clusters ($\mathbf{A}$, $\mathbf{B}$) and ($\mathbf{C}$, $\mathbf{D}$)
- The centroids for the inital clusters are

$$\mathbf{C}_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \text{ and } \mathbf{C}_2 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

- Compute the distance of case $\mathbf{A}$ from each centroid. Use Euclidean distance:

$$d(\mathbf{A}, \mathbf{C}_1) = \sqrt{(5-2)^2 + (3-2)^2} = \sqrt{10}$$

$$d(\mathbf{A}, \mathbf{C}_2) = \sqrt{(5-(-1))^2 + (3-(-1))^2} = \sqrt{52}$$

Case $\mathbf{A}$ remains in cluster 1, and the clusters and their centroids are unchanged.

K-means Clustering Example

- Next compute the distance of case $\mathbf{B}$ from each centroid.

$$d(\mathbf{B}, \mathbf{C}_1) = \sqrt{(-1-2)^2 + (1-2)^2} = \sqrt{10}$$

$$d(\mathbf{B}, \mathbf{C}_2) = \sqrt{(-1-(-1))^2 + (1-(-2))^2} = \sqrt{9}$$

Case $\mathbf{B}$ immediately moves to cluster 2. We have new clusters $(\mathbf{A})$ and $(\mathbf{B},\ \mathbf{C},\ \mathbf{D})$ and the new centroids are

$\mathbf{C}_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ and $\mathbf{C}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

K-means Clustering Example

- Next compute the distance of case $\mathbf{C}$ from each of the updated centroids.

$$d(\mathbf{C}, \mathbf{C}_1) = \sqrt{(1-5)^2 + (-2-3)^2} = \sqrt{41}$$

$$d(\mathbf{C}, \mathbf{C}_2) = \sqrt{(1-(-1))^2 + (-2-(-1))^2} = \sqrt{5}$$

Case $\mathbf{C}$ stays in cluster 2. the clusters are still $(\mathbf{A})$ and $(\mathbf{B},\ \mathbf{C},\ \mathbf{D})$ with centroids $\mathbf{C}_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ and $\mathbf{C}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

- Next compute the distance of case $\mathbf{D}$ from each centroid.

$$d(\mathbf{D}, \mathbf{C}_1) = \sqrt{(-3-5)^2 + (-2-3)^2} = \sqrt{89}$$

$$d(\mathbf{D}, \mathbf{C}_2) = \sqrt{(-3-(-1))^2 + (-2-(-1))^2} = \sqrt{5}$$

Case $\mathbf{D}$ stays in cluster 2. The clusters are unchanged.

How to select the number of clusters?

- Use measures of how good the clusters describe the structure of the data for varying number of clusters.
- F-statistic: Calinski-Harabasz index
- Gap statistic: a metric based on within group distances defined using permutations

Some Notations

- n is the number of observations in the data.
- p is the number of variables measured on each observation.
- C_k represents the k-th cluster at a particular step of the clustering algorithm and n_k is the number of obs. in C_k.
- SS_T is the total sum of squared distances from the overall vector of means for all n obs. in the data set:

$$SS_T = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}})$$

Some Notations

- Define by the sum of squared distances within the cluster C_k:

$$W_k = \sum_{i \text{ in } C_k} (\mathbf{X}_i - \bar{\mathbf{X}}_k)'(\mathbf{X}_i - \bar{\mathbf{X}}_k).$$

- Let G be the number of clusters at current step.

- Define

$$W_G = \sum_{k=1}^{G} W_k.$$

- Define by the "between" clusters sum of squared distances

$$B_G = SS_T - W_G = \sum_{j=1}^{G} n_j(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})'(\bar{\mathbf{X}}_j - \bar{\mathbf{X}}).$$

Calinski-Harabasz Index

- The Calinski-Harabasz Index is a pseudo F-ratio:

$$ch_G = \frac{\frac{B_G}{G-1}}{\frac{W_G}{n-G}}$$

- Compute ch_G for steps with G= 2, 3, 4, 5, 6, ... clusters. If values go up and then go down as G increases, choose the number of clusters at a peak. If there is no peak, then there is no natural set of clusters.

- In general, larger index value indicates better clustering.

- ch_G is not distributed as a random variable with an F-distribution. This is a guideline, not a test

- Milligan and Cooper (1985), Psychometrika, 50, 159-179.

Gap Statistic

- Define $D_k = \sum_{i,i' \in C_k} d_{ii'}$ to be the sum of pairwise distances for all points in cluster C_k.
- Set $W_G = \sum_{k=1}^{G} \frac{1}{2n_k} D_k$.
- Compute $\text{Gap}_n(r) = E_n(\log W_r) - \log(W_r)$ using "random" clustering.

Larger values of gap statistic correspond to better clustering.

Key R Commands

- `heat/heatmap.2`

  generates heatmaps.

- `hcluster`

  performs hierarchical clustering.

- `kmeans`

  performs k-means clustering.

- `fviz_nbclust`

  gives different k-mean clustering results using different measures

- R code demonstration