Unsupervised Learning: Clustering

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MED3007
Statistical Principles in Genomics: an Introduction with Rstudio
16.01.2025

1. Introduction
   - Motivation
   - Notions of similarity
2. Hierarchical Clustering Algorithms
   - Main ideas
   - Illustrative example
   - Visualization
   - Summary
3. Partitioning clustering: k-means
   - Partitioning Clustering Algorithms
   - Main idea of the k-means
   - Illustrative example 1
   - Illustrative example 2
4. Summary

Recap from last time...

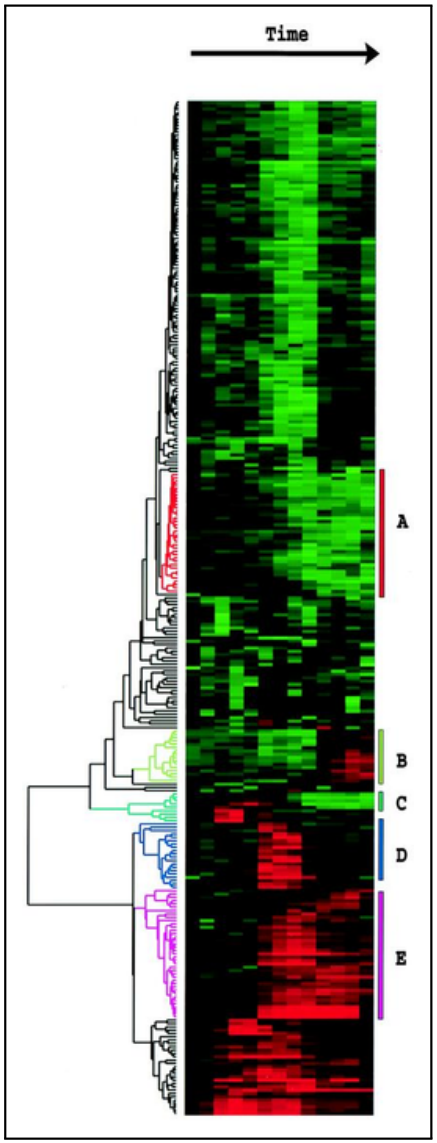
Unsupervised methods

Methods that do not make use of external information: groups / class assignments, clinical outcomes, covariates.
Aim: finding hidden structure in the data

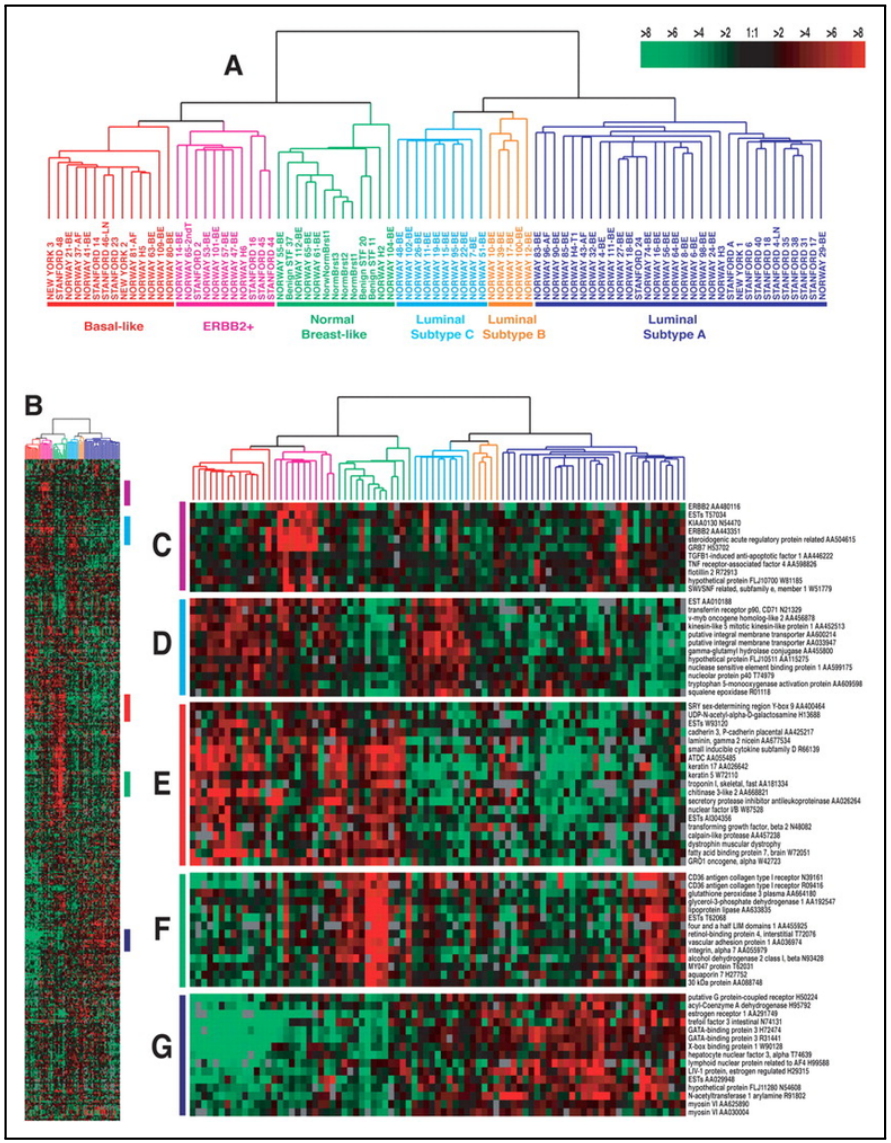
How can we find and/or visualise structure?

- Reducing dimensionality by removing noise and "uninteresting" stuff
  → TODAY – Lecture 2
- Ordering and grouping together via clustering
  → TOMORROW – Lecture 3

Heatmaps are everywhere...



Eisen et al. (1998), Figure 1



Sorlie et al. (2001), Figure 1

What is clustering?

Clustering methods

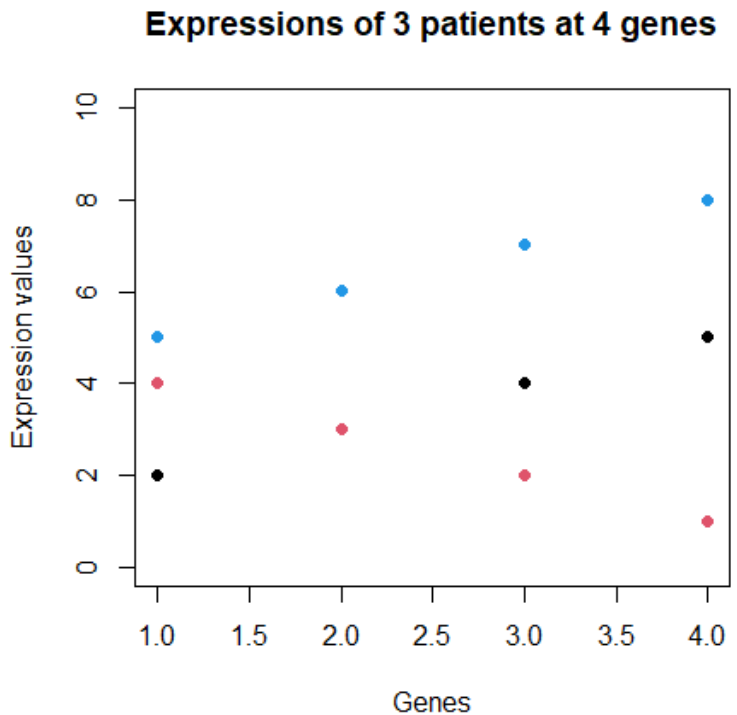
Methods that **aim at grouping** a collection of objects into groups, or clusters, such that objects within each cluster are more closely related to one another than objects assigned to a different cluster

- Distance measure
  A notion of distance or similarity of two objects: When are two objects close to each other?
- Clustering algorithm
  A procedure to minimise distances of objects within groups and/or maximise distances between groups.

What is meant when two patients are said to be “similar”?

Possibilities:

- Red and black patients are similar:
  They lie close to each other.
- Blue and black patients are similar:
  They are positively correlated.
- Red and blue patients are associated:
  They are negatively correlated.



Examples of distance measures $d(\cdot, \cdot)$

- Euclidean distance

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{m} (x_k - y_k)^2 \right)^{1/2}$$

- 1 - Pearson's correlation

$$d_{cor}(\mathbf{x}, \mathbf{y}) = 1 - r(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}}$$

- 1 - Spearman's rank correlation ($R(x_i)$= Rank of x_i)

$$\begin{aligned} d_{spear}(\mathbf{x}, \mathbf{y}) &= 1 - r_s(\mathbf{x}, \mathbf{y}) \\ &= 1 - \frac{\sum_{i=1}^{m} (R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{m} (R(x_i) - \overline{R(x)})^2 \sum_{i=1}^{m} (R(y_i) - \overline{R(y)})^2}} \end{aligned}$$

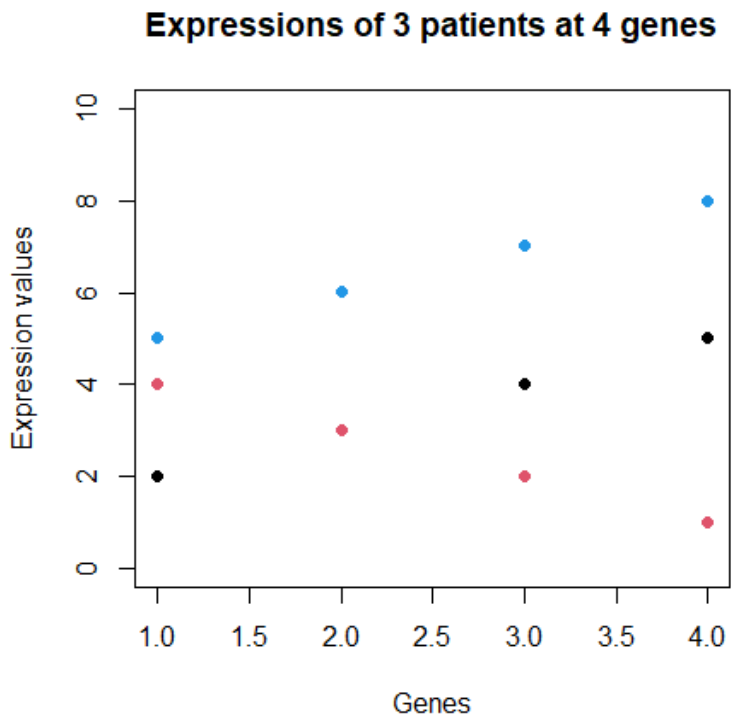
Examples of distance measures $d(\cdot, \cdot)$ – visual!

Euclidean distance:

- d(black, red) = 4.90
- d(black, blue) = 6.00
- d(blue, red) = 9.16

Spearman-correlation-distance:

- d(black, red) = 2.00
- d(black, blue) = 0.00
- d(blue, red) = 2.00



Interpretation:

- the smaller the distance, the more similar the patients' response
- different distances measure "similarity" differently

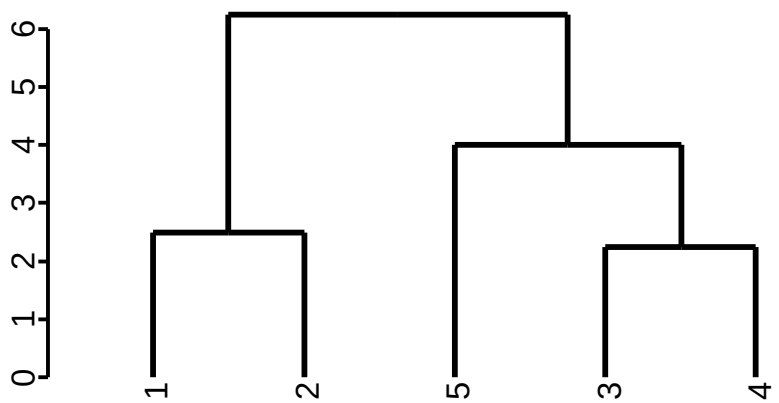
1 Introduction
- Motivation
- Notions of similarity

2 Hierarchical Clustering Algorithms
- Main ideas
- Illustrative example
- Visualization
- Summary

3 Partitioning clustering: k-means
- Partitioning Clustering Algorithms
- Main idea of the k-means
- Illustrative example 1
- Illustrative example 2

4 Summary

Hierarchical clustering - Agglomerative algorithm



- Bottom-up algorithm (top-down methods are less common)
- Start with each object assigned to its own cluster.
- In each iteration, merge the two clusters with the minimal distance from each other - until you are left with a single cluster containing all objects
- But how define the distance between two clusters?

Hierarchical clustering - Linkage

Calculation of distance d(G,H) between clusters G and H is based on pairwise distances between objects from the two clusters:

- Single linkage uses the smallest distance between the objects:

$$d_S(G, H) = min_{(i \in G; j \in H)} d_{ij}$$

  Single linkage is not commonly used.

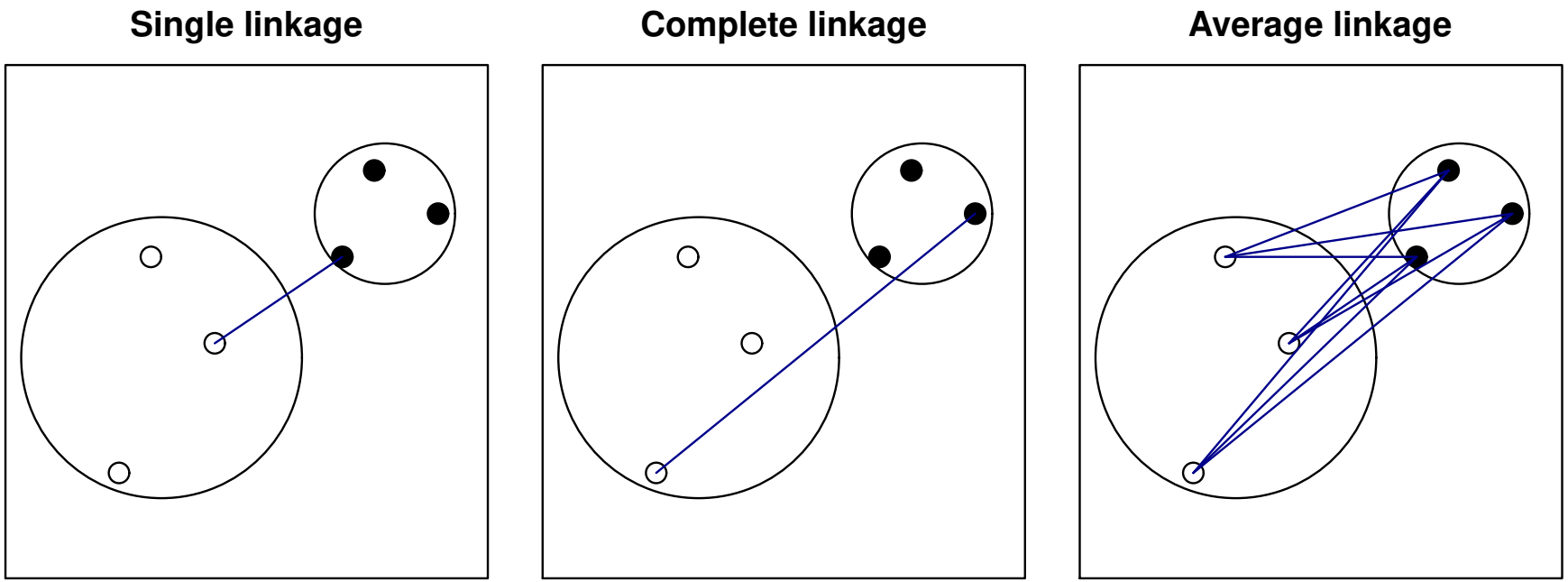
- Complete linkage uses the largest distance between the objects:

$$d_C(G, H) = max_{(i \in G; j \in H)} d_{ij}$$

- Average linkage uses the average distance between the objects:

$$d_A(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{ij}$$

Hierarchical clustering - Linkage illustration



Hierarchical clustering - Single linkage example

- Data: 5 patients 1-5 (= rows), 2 expression arrays A,B (= columns)
- Method: **Agglomerative hierarchical clustering** using **Single-Linkage**

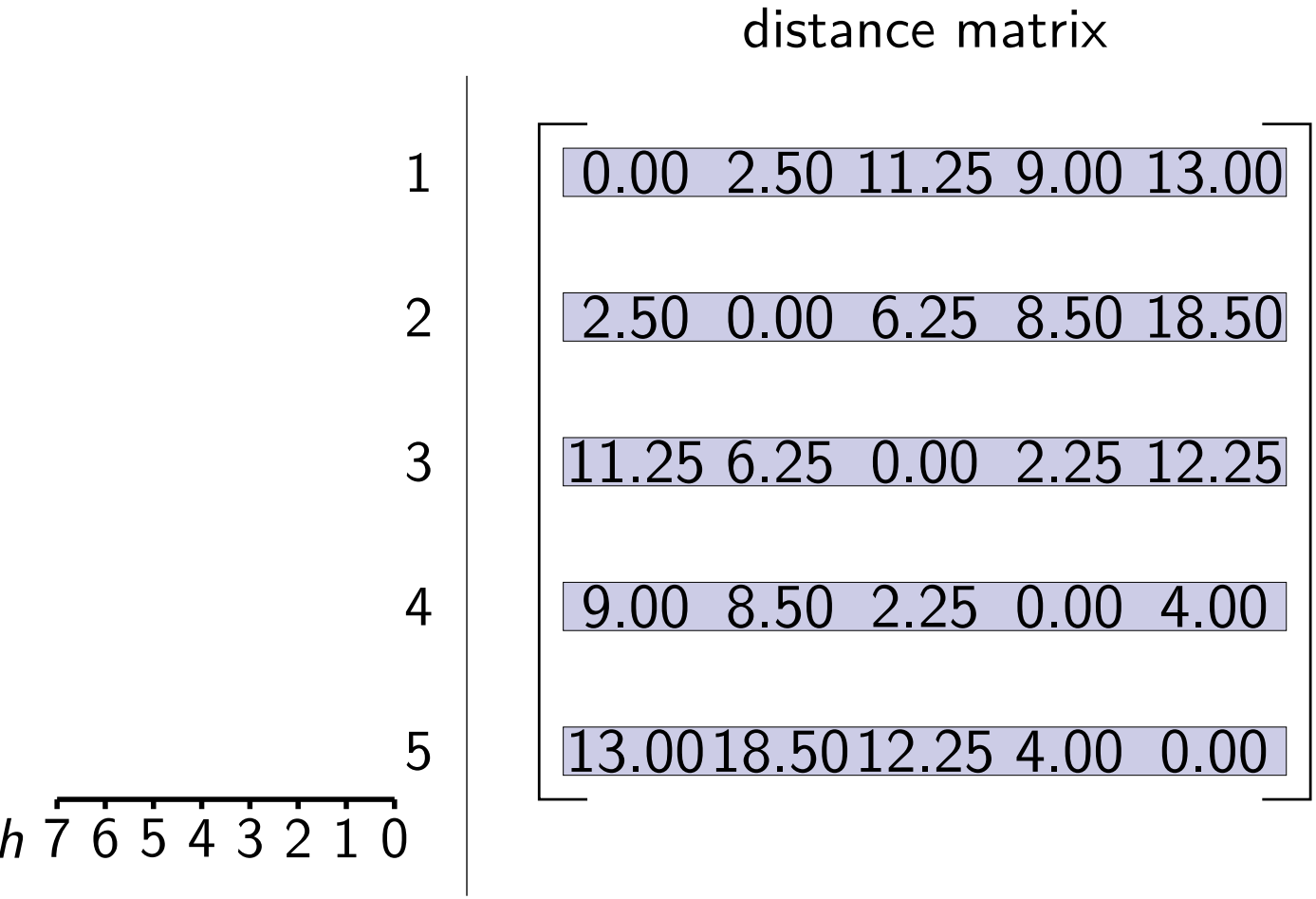
Expression matrix

	A	B
1	5.0	7.0
2	5.5	8.5
3	8.0	8.5
4	8.0	7.0
5	8.0	5.0

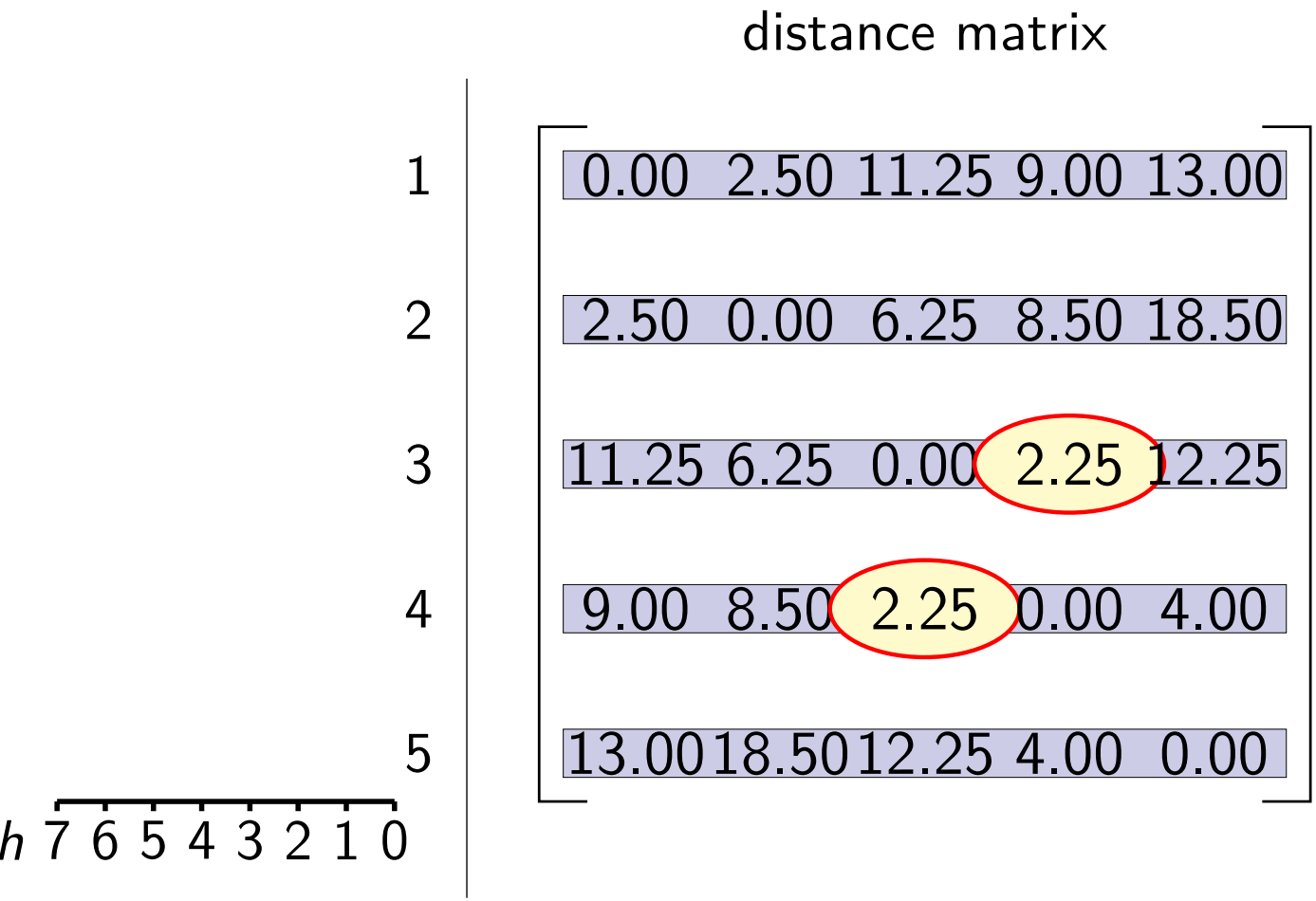
Distance matrix (**Euclidean distance**)

	1	2	3	4	5
1	0.00	2.50	11.25	9.00	13.00
2	2.50	0.00	6.25	8.50	18.50
3	11.25	6.25	0.00	2.25	12.25
4	9.00	8.50	2.25	0.00	4.00
5	13.00	18.50	12.25	4.00	0.00

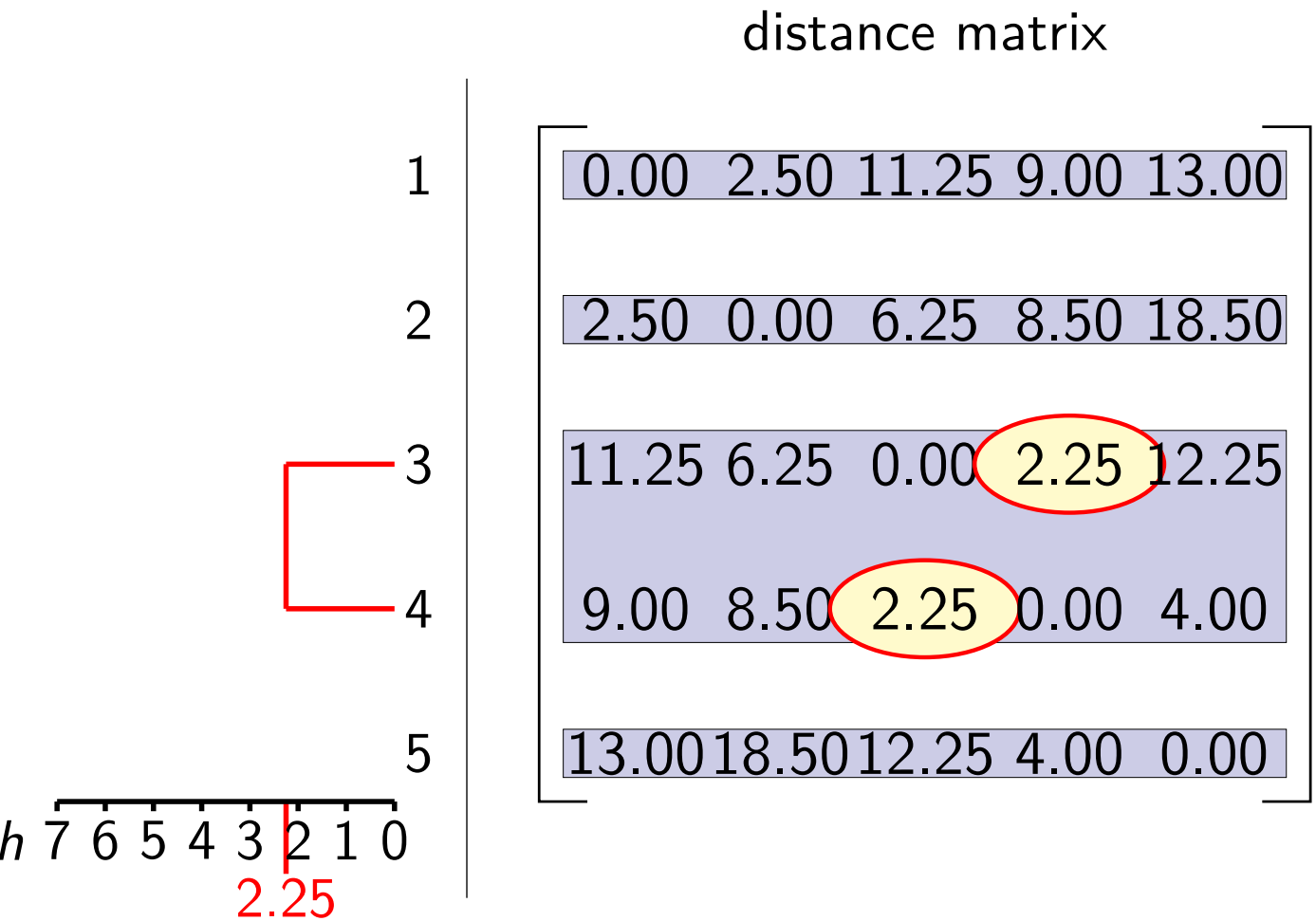
Agglomerative clustering - Single linkage example



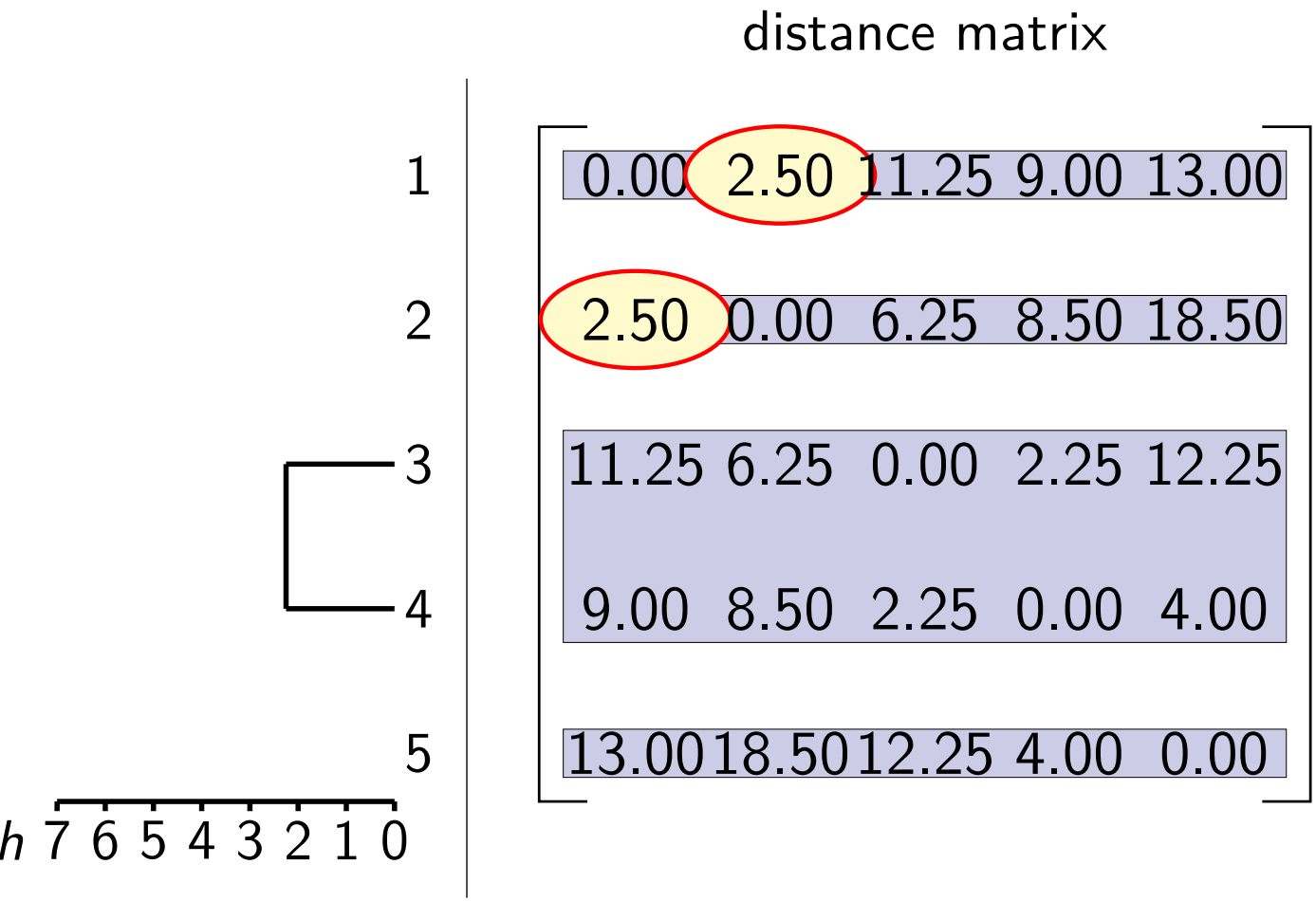
Agglomerative clustering - Single linkage example



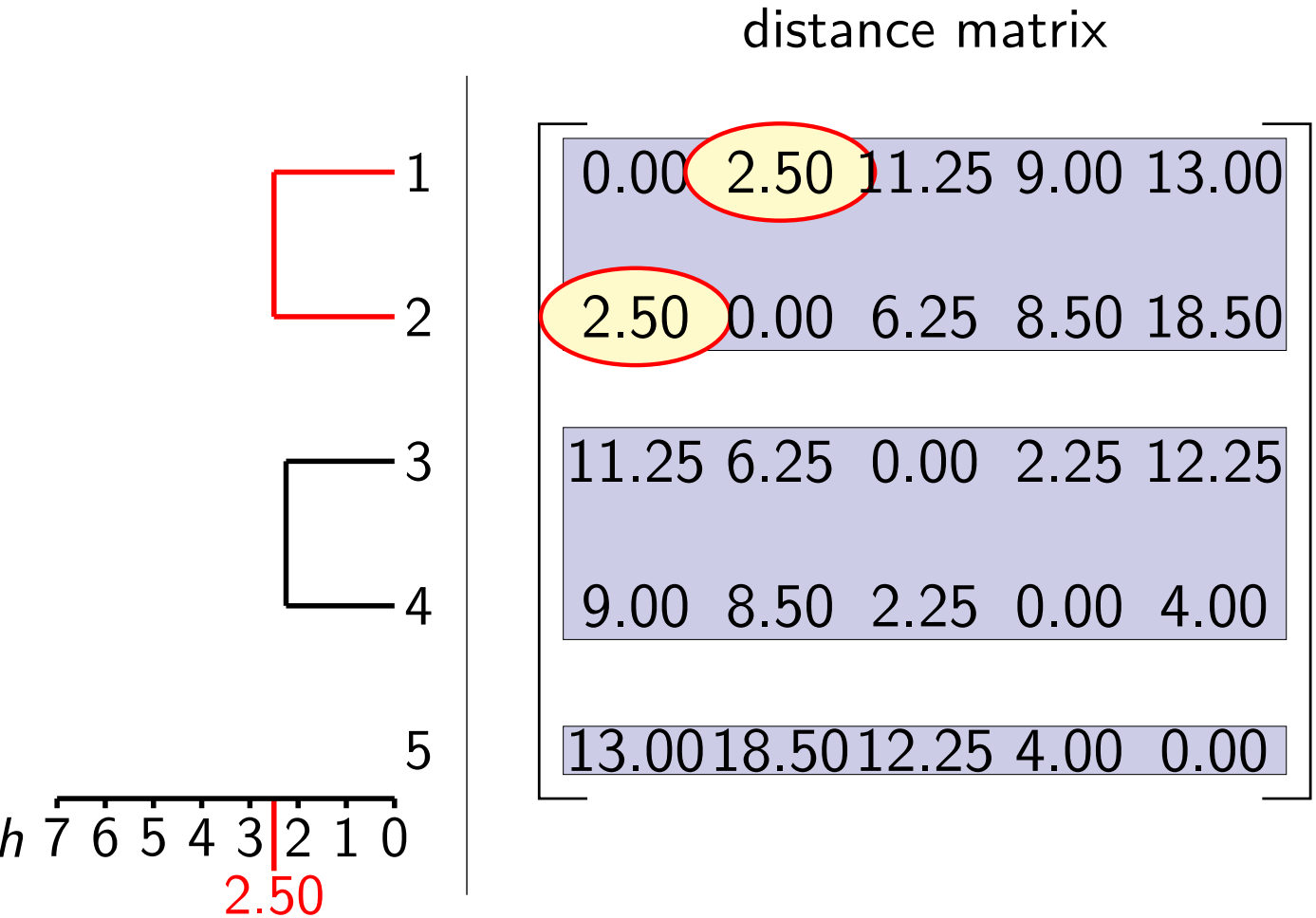
Agglomerative clustering - Single linkage example



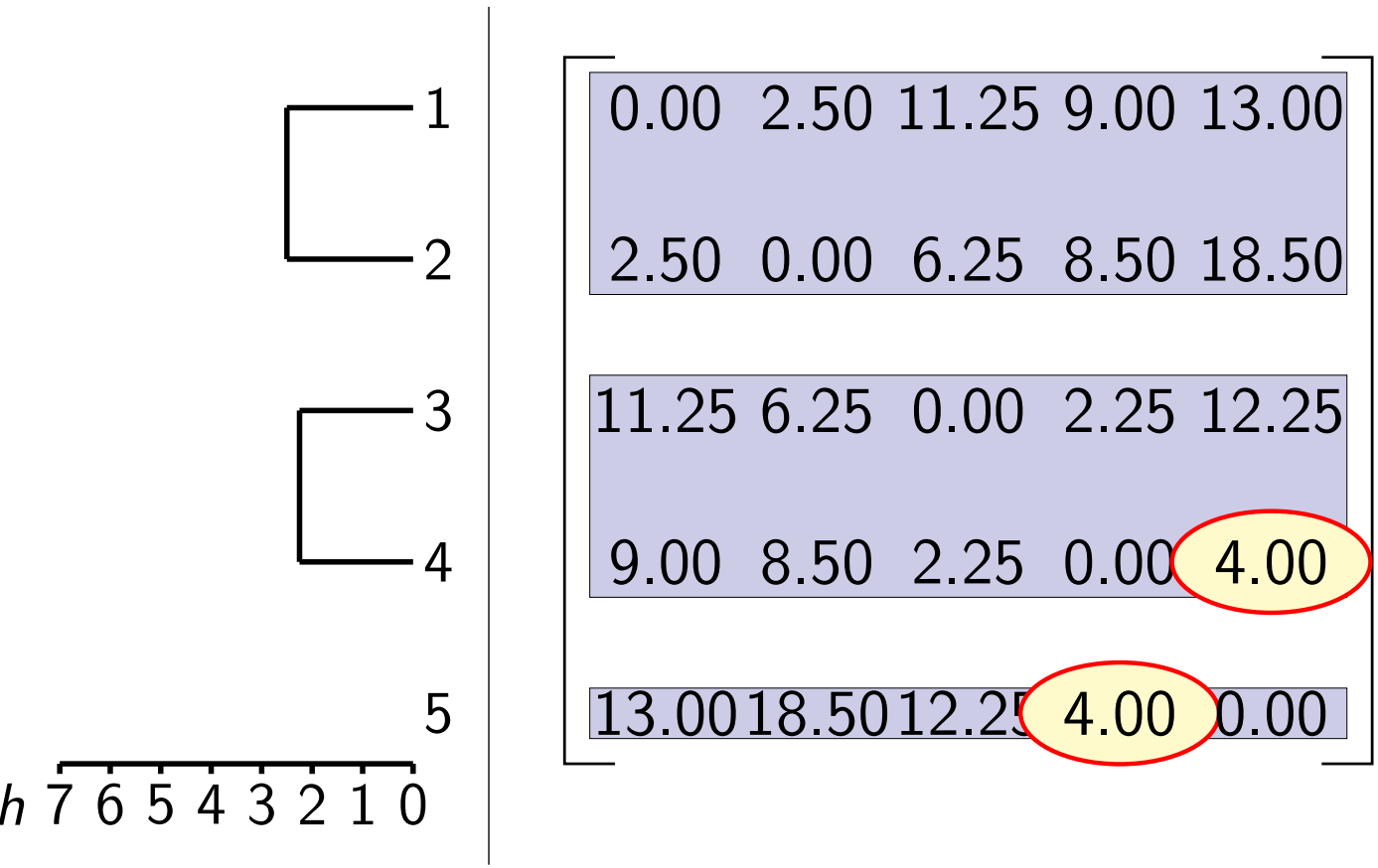
Agglomerative clustering - Single linkage example



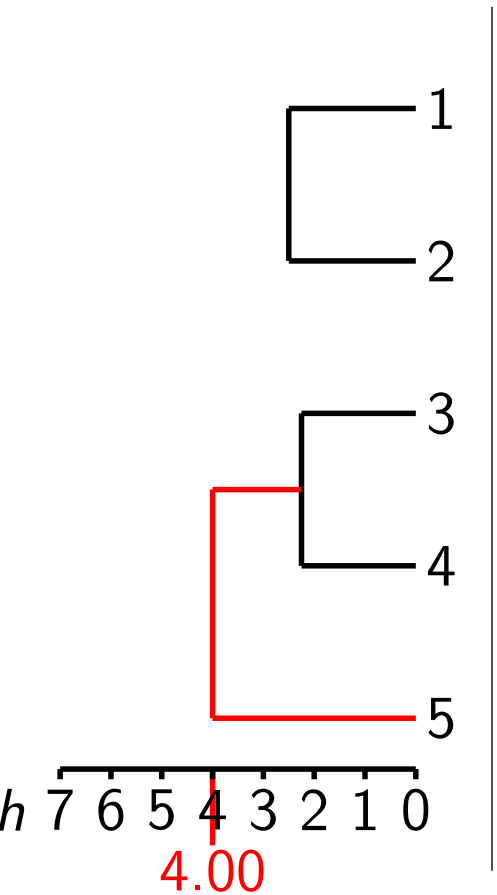
Agglomerative clustering - Single linkage example



Agglomerative clustering - Single linkage example



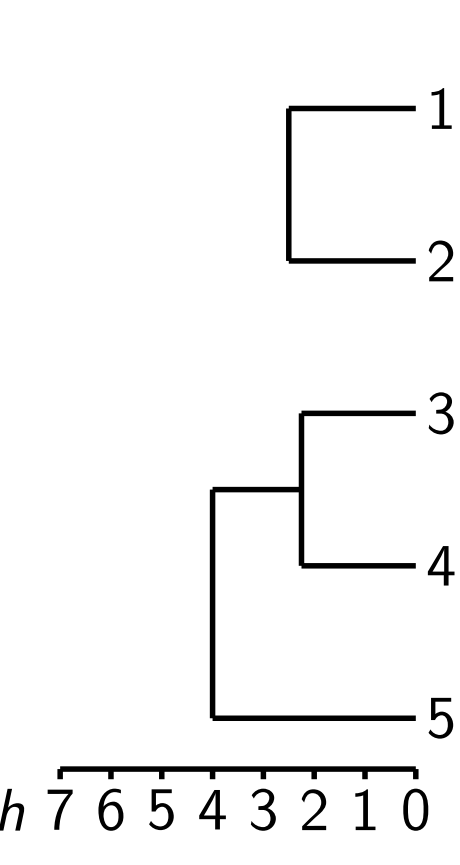
Agglomerative clustering - Single linkage example



distance matrix

0.00	2.50	11.25	9.00	13.00
2.50	0.00	6.25	8.50	18.50
11.25	6.25	0.00	2.25	12.25
9.00	8.50	2.25	0.00	4.00
13.00	18.50	12.25	4.00	0.00

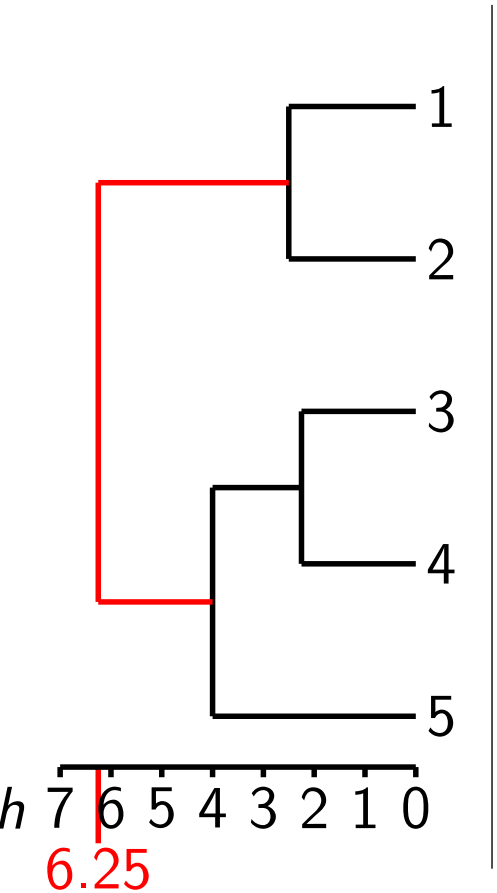
Agglomerative clustering - Single linkage example



distance matrix

0.00	2.50	11.25	9.00	13.00
2.50	0.00	6.25	8.50	18.50
11.25	6.25	0.00	2.25	12.25
9.00	8.50	2.25	0.00	4.00
13.00	18.50	12.25	4.00	0.00

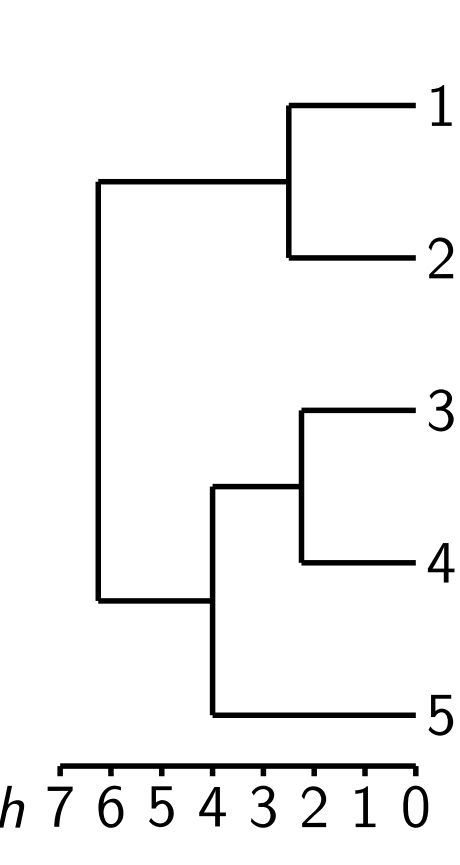
Agglomerative clustering - Single linkage example



distance matrix

0.00	2.50	11.25	9.00	13.00
2.50	0.00	6.25	8.50	18.50
11.25	6.25	0.00	2.25	12.25
9.00	8.50	2.25	0.00	4.00
13.00	18.50	12.25	4.00	0.00

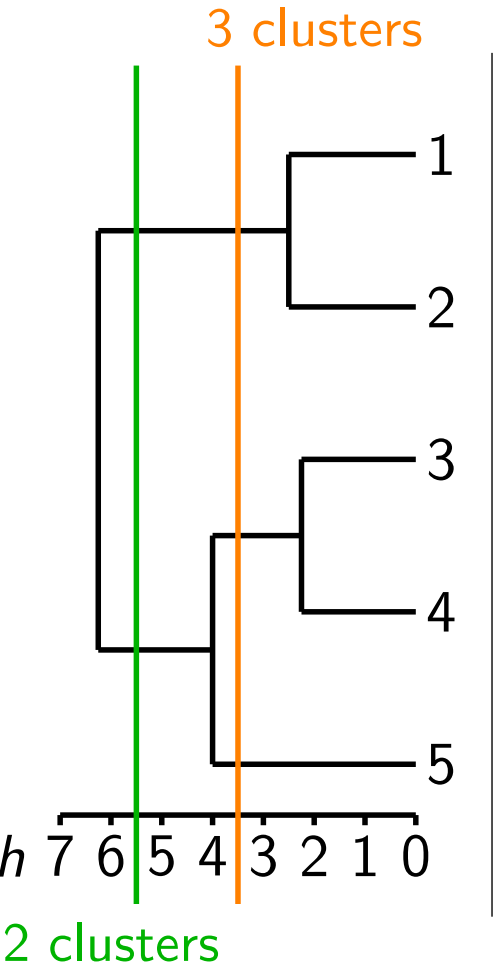
Agglomerative clustering - Single linkage example



distance matrix

$$\begin{bmatrix} 0.00 & 2.50 & 11.25 & 9.00 & 13.00 \\ 2.50 & 0.00 & 6.25 & 8.50 & 18.50 \\ 11.25 & 6.25 & 0.00 & 2.25 & 12.25 \\ 9.00 & 8.50 & 2.25 & 0.00 & 4.00 \\ 13.00 & 18.50 & 12.25 & 4.00 & 0.00 \end{bmatrix}$$

Agglomerative clustering - Single linkage example

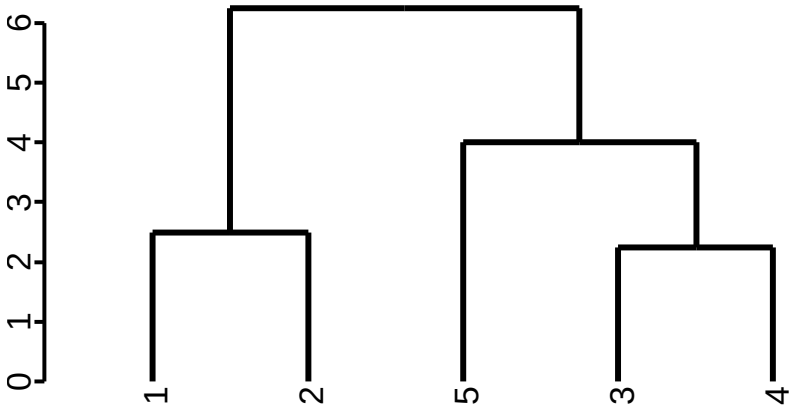


distance matrix

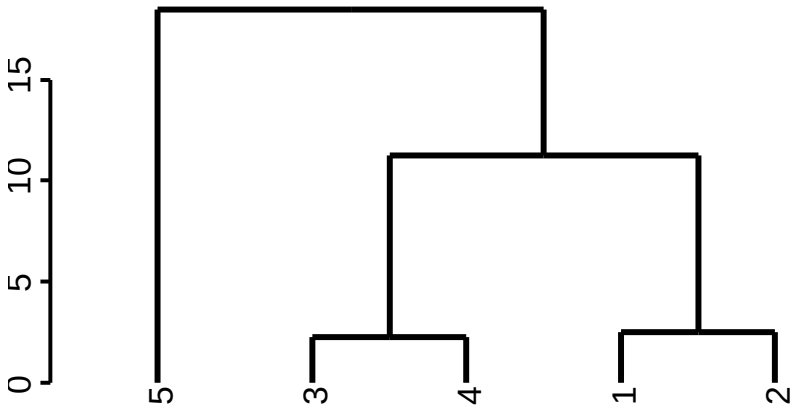
$$\begin{bmatrix} 0.00 & 2.50 & 11.25 & 9.00 & 13.00 \\ 2.50 & 0.00 & 6.25 & 8.50 & 18.50 \\ 11.25 & 6.25 & 0.00 & 2.25 & 12.25 \\ 9.00 & 8.50 & 2.25 & 0.00 & 4.00 \\ 13.00 & 18.50 & 12.25 & 4.00 & 0.00 \end{bmatrix}$$

Hierarchical clustering - Linkage methods

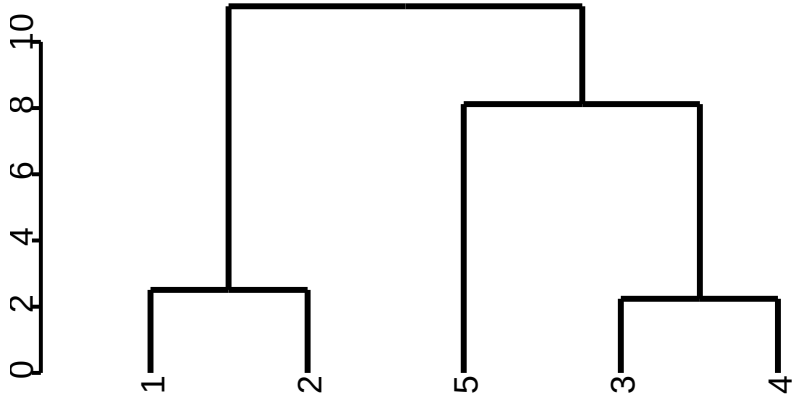
Same distance matrix!



Single Linkage

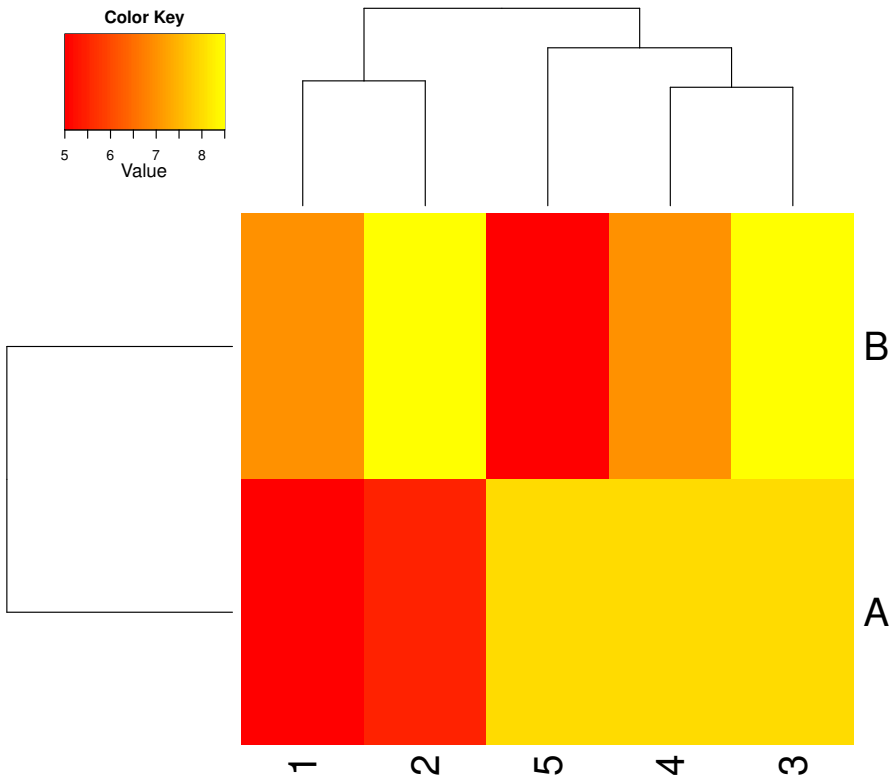


Complete Linkage



Average Linkage

Hierarchical clustering - Heatmap



- Single linkage
- Euclidean distance
- The expression values are represented as colours.

Summary & Take-home messages

- The procedure provides a hierarchy of the clustering, with the number of clusters ranging from 1 to the number of objects
- Finding a meaningful cut is similar to the problem of finding the number of clusters k for partitioning methods (NEXT SLIDES)
- An incorrect merge early in the tree cannot be changed later on
- The choice of the distance measure depends on the data and the intention of the clustering
- Even data generated at random will result in a clustering: be careful with interpretation!

1. Introduction
   - Motivation
   - Notions of similarity
2. Hierarchical Clustering Algorithms
   - Main ideas
   - Illustrative example
   - Visualization
   - Summary
3. Partitioning clustering: k-means
   - Partitioning Clustering Algorithms
   - Main idea of the k-means
   - Illustrative example 1
   - Illustrative example 2
4. Summary

Partitioning

Partitioning algorithms (= non-hierarchical methods)

- split the data into a pre-specified number k of groups
- iteratively re-allocate objects until some optimality criterion is met

k thus needs to be fixed in advance

Examples:

- k-means clustering
- Partitioning around medoids (PAM)

  generalization of k-means (allows additional optimisation criteria)
- Self-organising maps (SOM)

  similar to k-means but with additional constraints (grid-like structure)

Partitioning - k-means

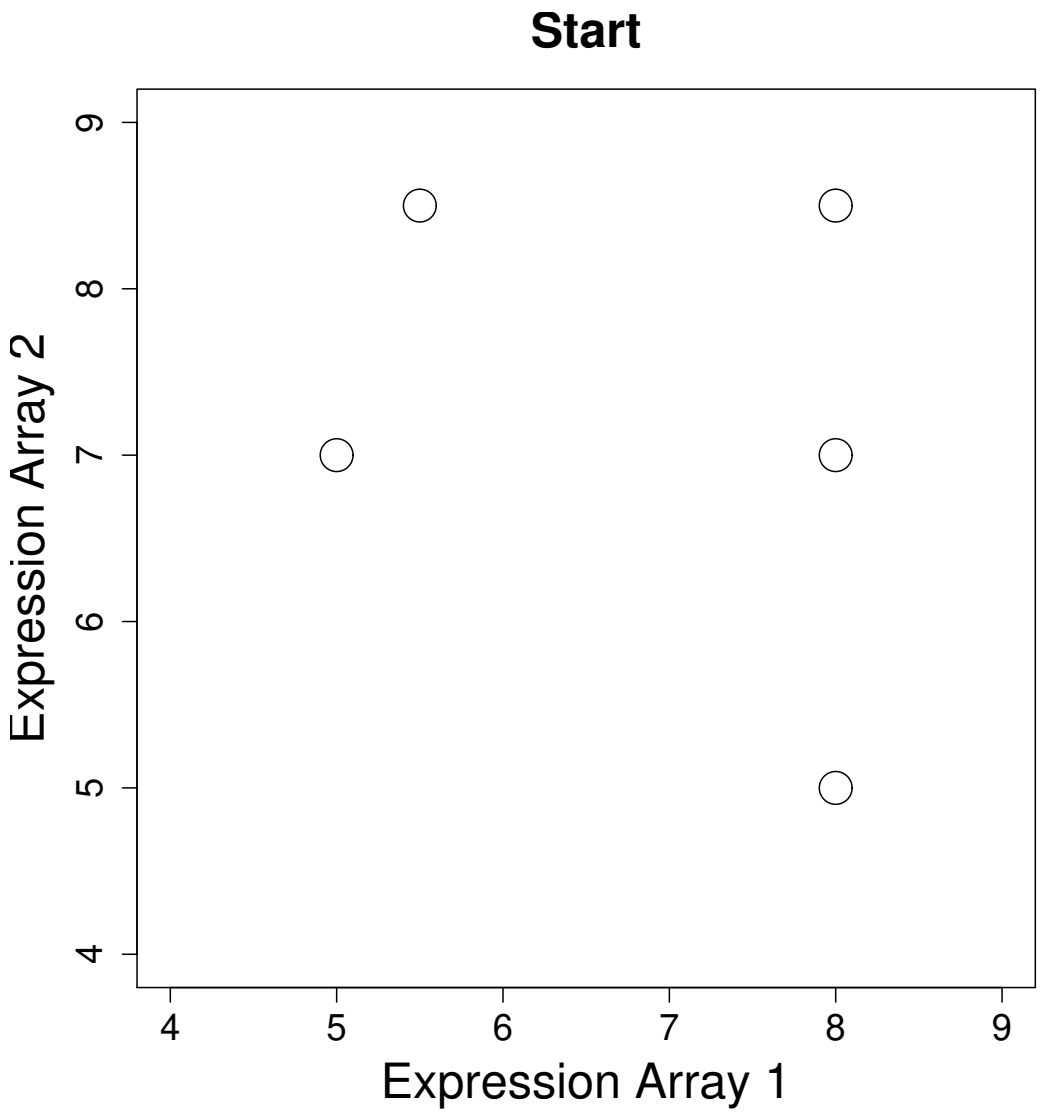
k-means

The k-means algorithm minimises the sum of within-cluster variances

It chooses a random sample of k different objects as initial cluster centroids, then alternating until convergence:

1. Assign each object to the cluster whose centroid is closest (among the k centroids) with respect to Euclidean distance
2. Calculate k new centroids as the averages of all points assigned to the same cluster

Partitioning - k-means example

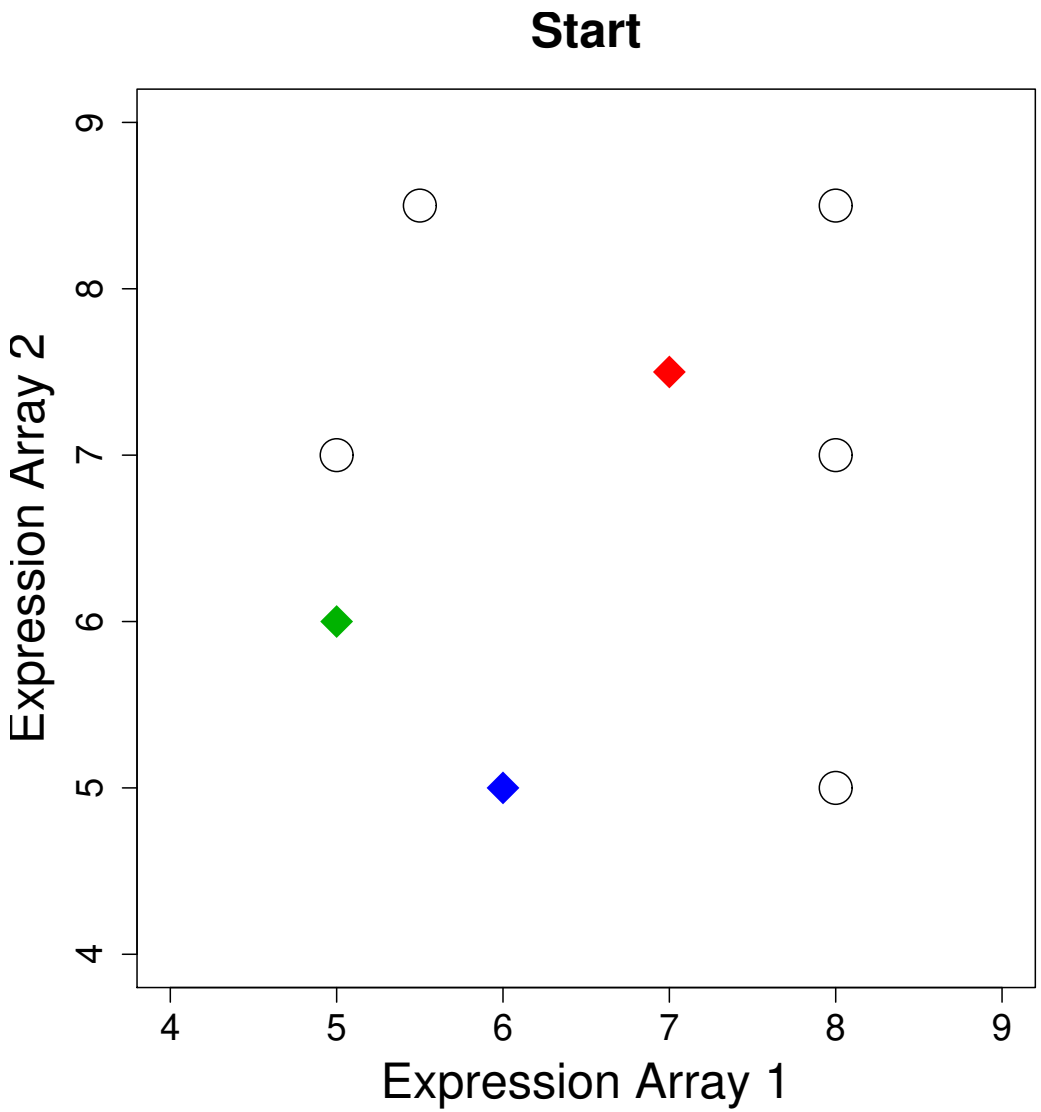


$k = 3$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



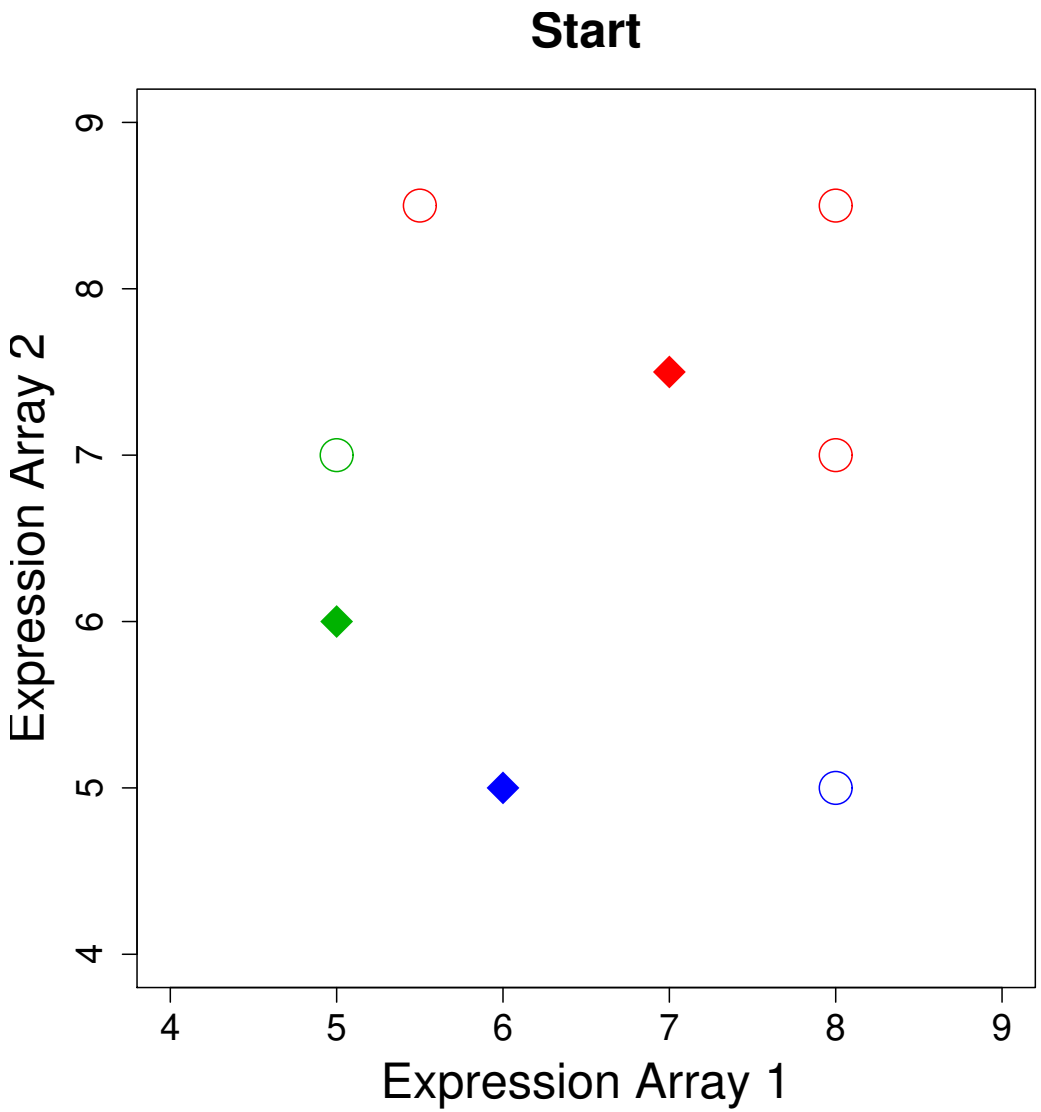
Centroid matrix

$$\begin{pmatrix} 5.0 & 6.0 \\ 7.0 & 7.5 \\ 6.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



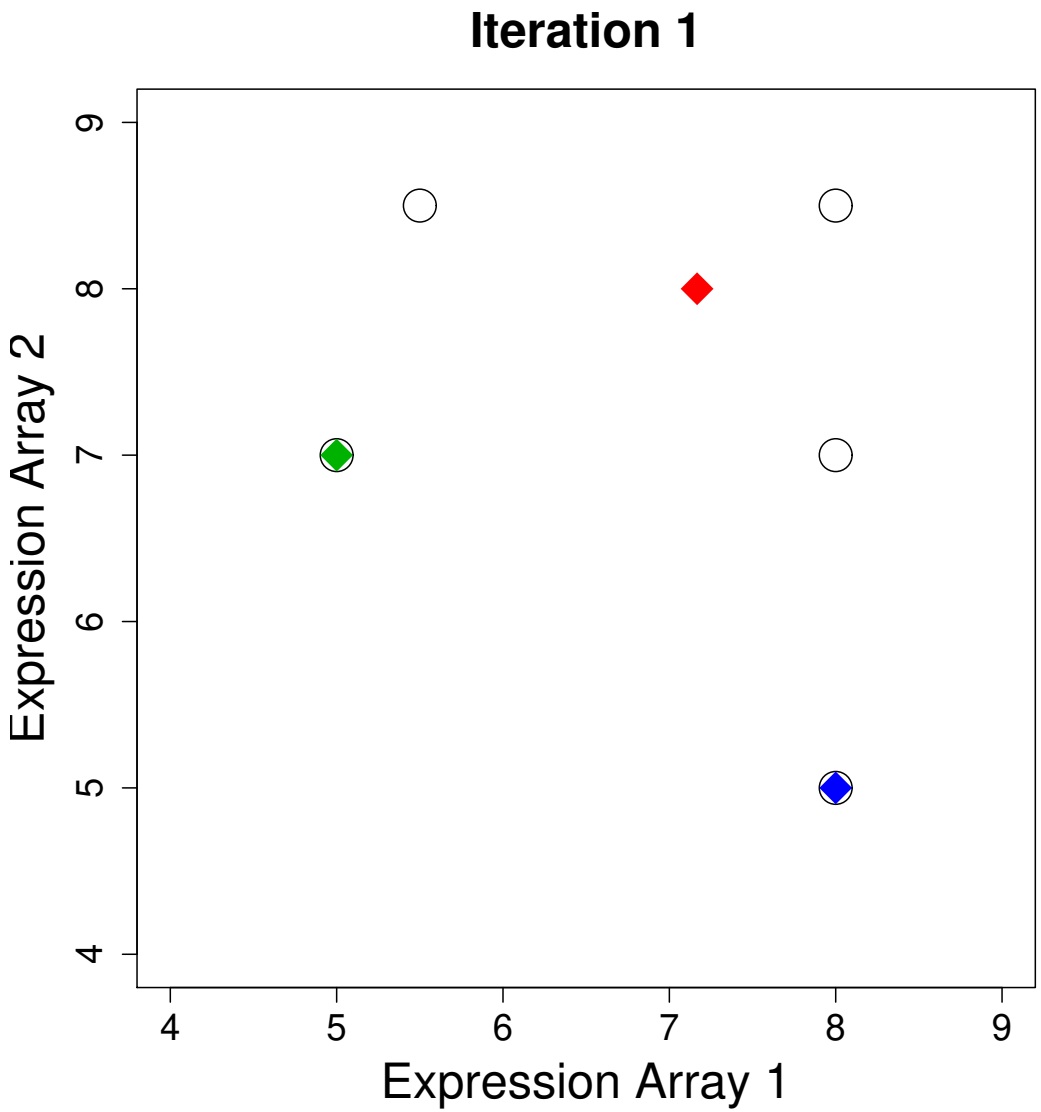
Centroid matrix

$$\begin{pmatrix} 5.0 & 6.0 \\ 7.0 & 7.5 \\ 6.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



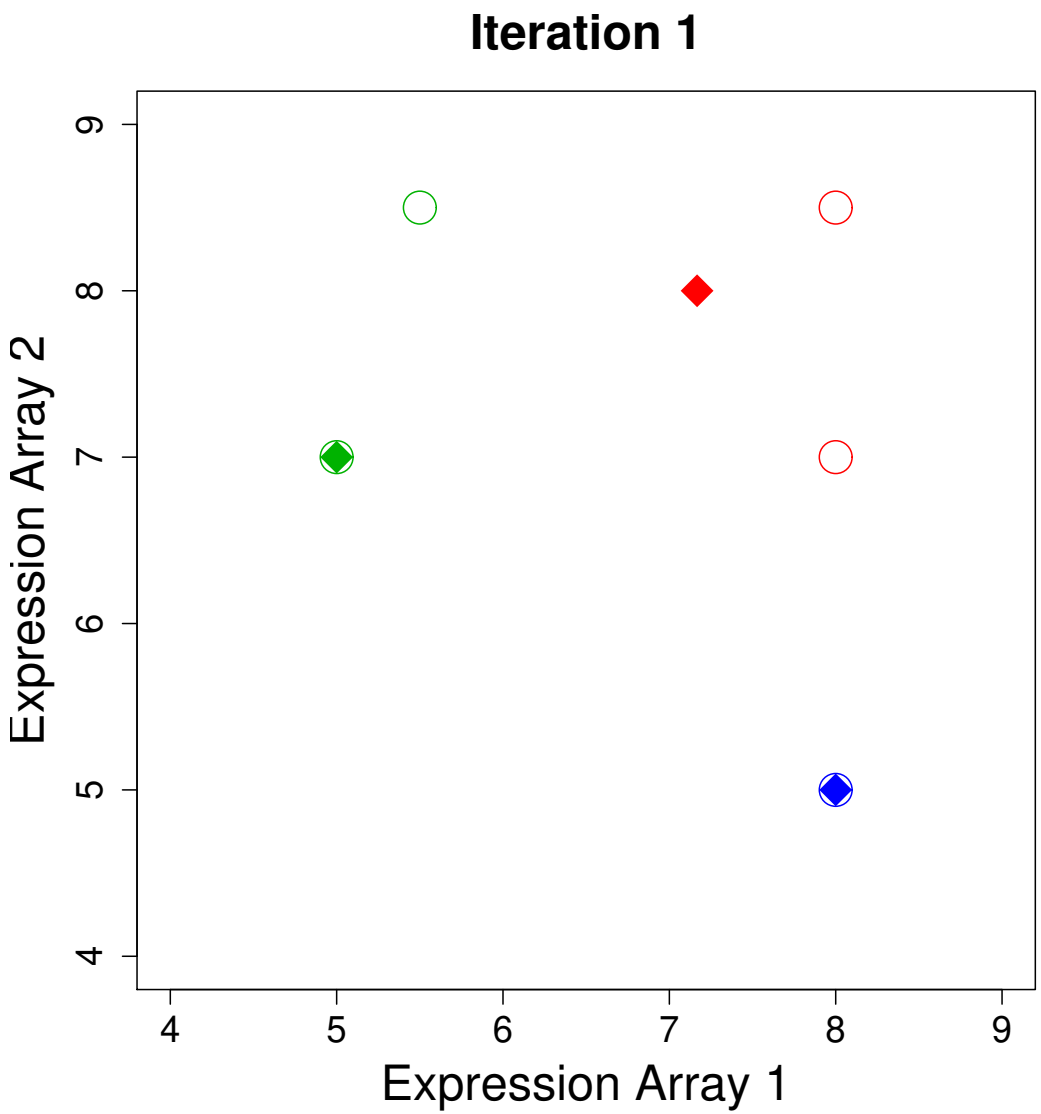
Centroid matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 7.17 & 8.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



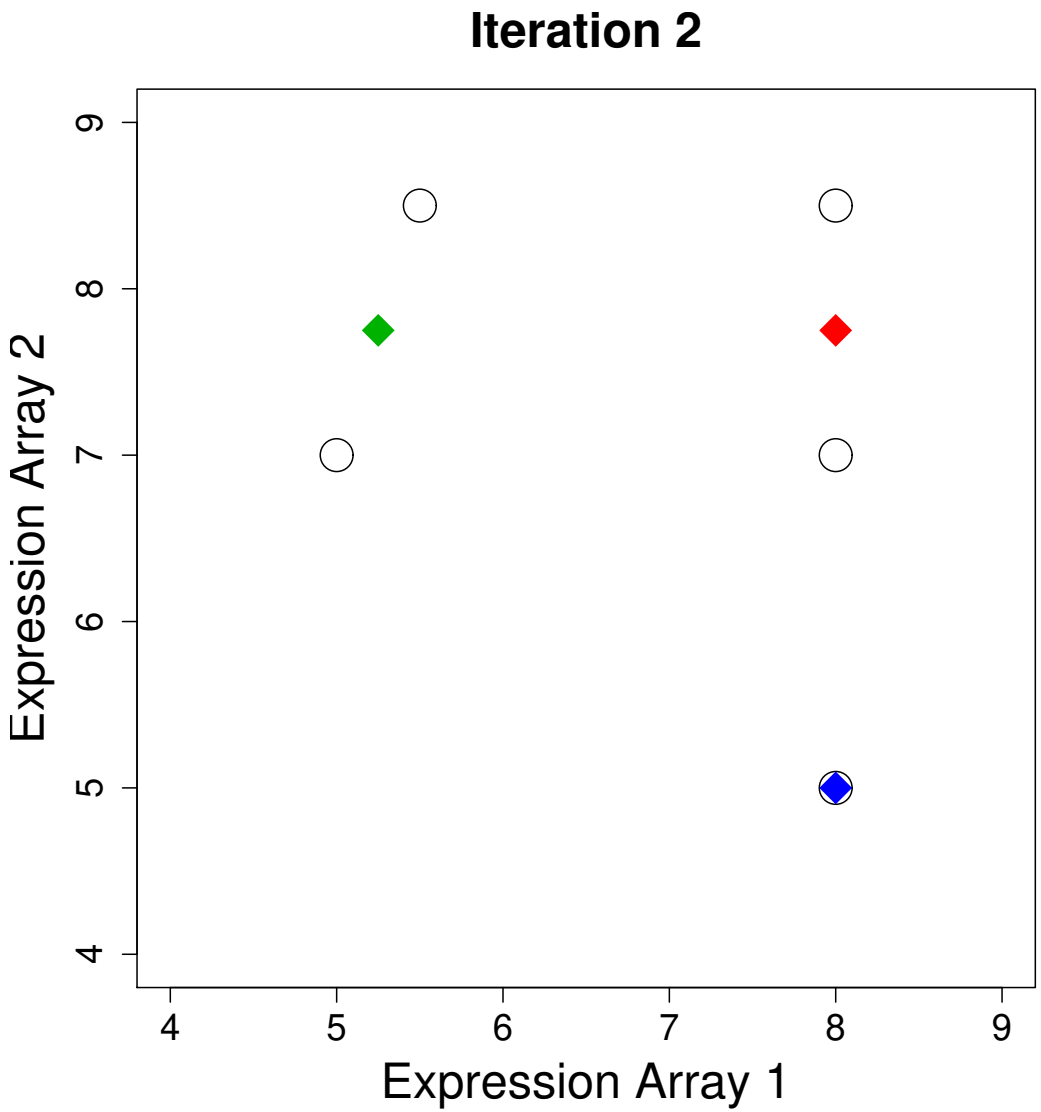
Centroid matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 7.17 & 8.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



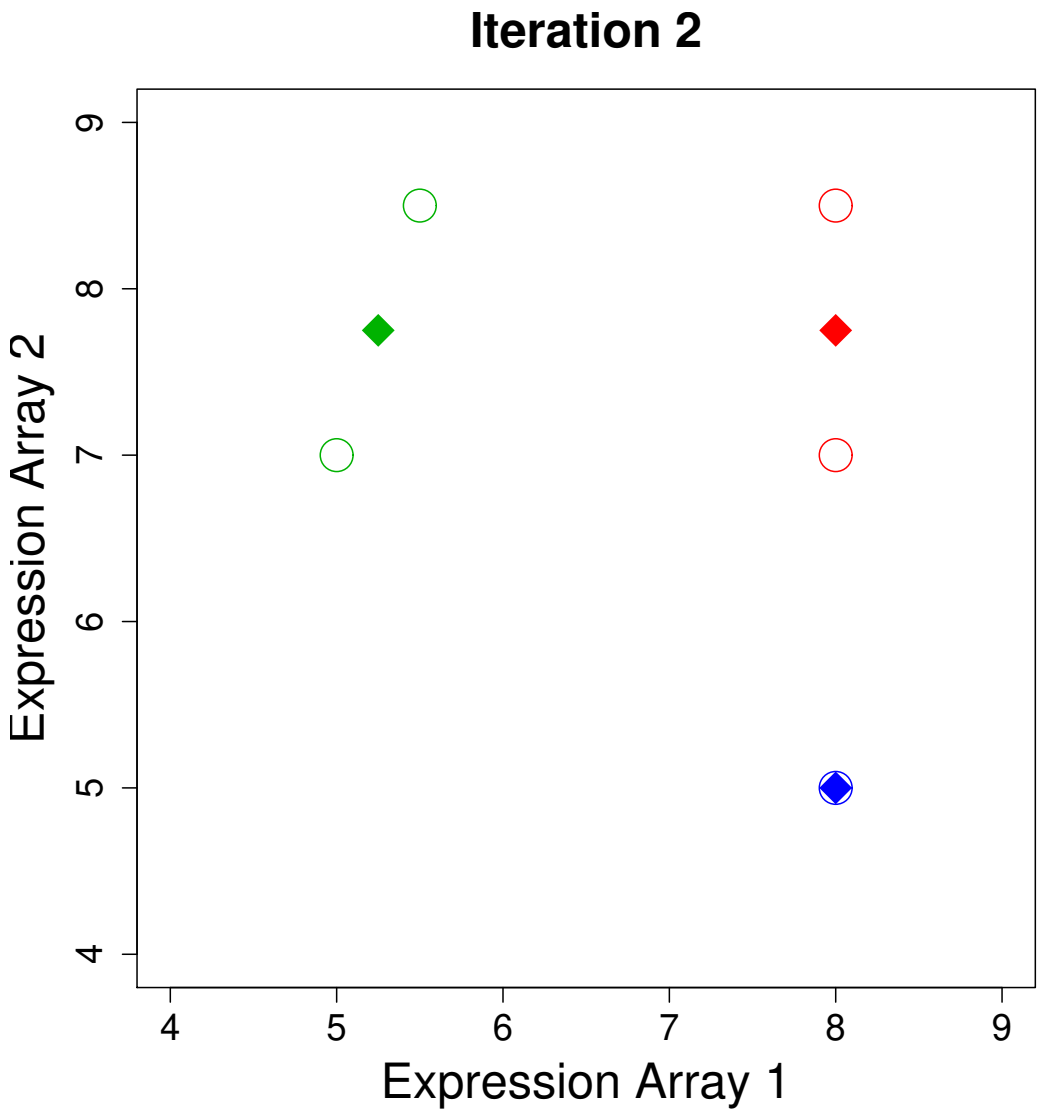
Centroid matrix

$$\begin{pmatrix} 5.25 & 7.75 \\ 8.00 & 7.75 \\ 8.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



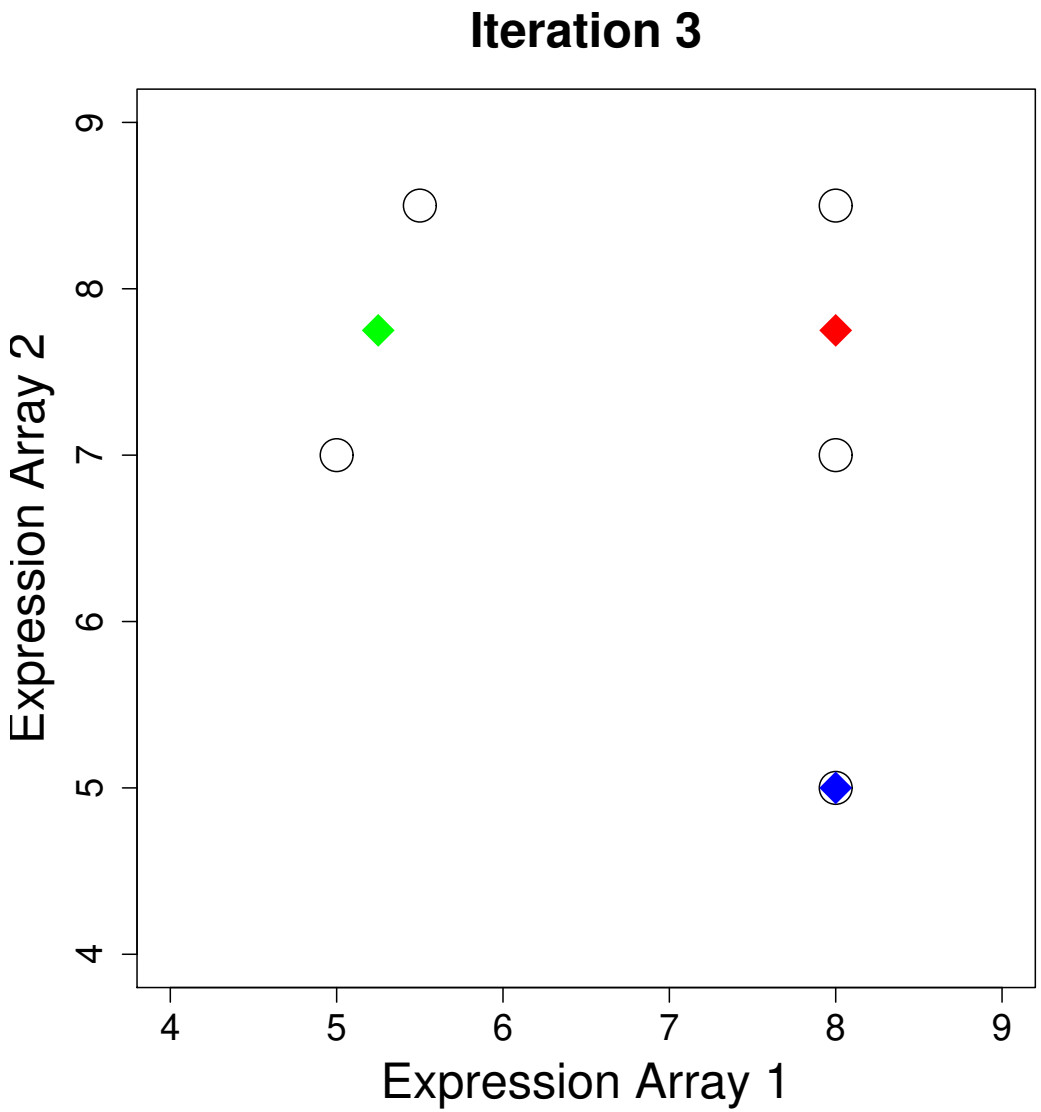
Centroid matrix

$$\begin{pmatrix} 5.25 & 7.75 \\ 8.00 & 7.75 \\ 8.0 & 5.0 \end{pmatrix}$$

Expression matrix

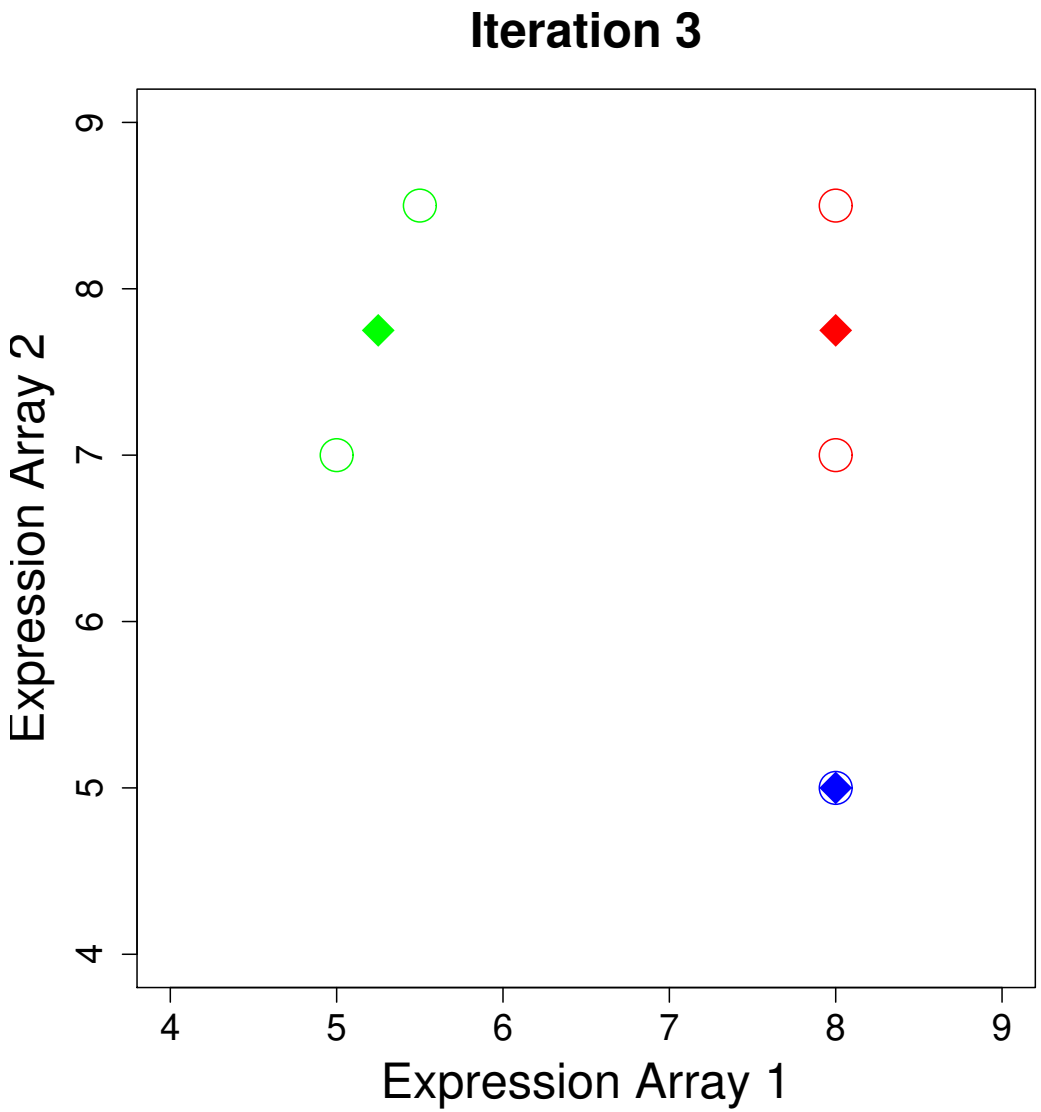
$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example



Iteration 3: No changes in centroid matrix

Partitioning - *k*-means example



Iteration 3: No changes
$\rightarrow$ We are done.

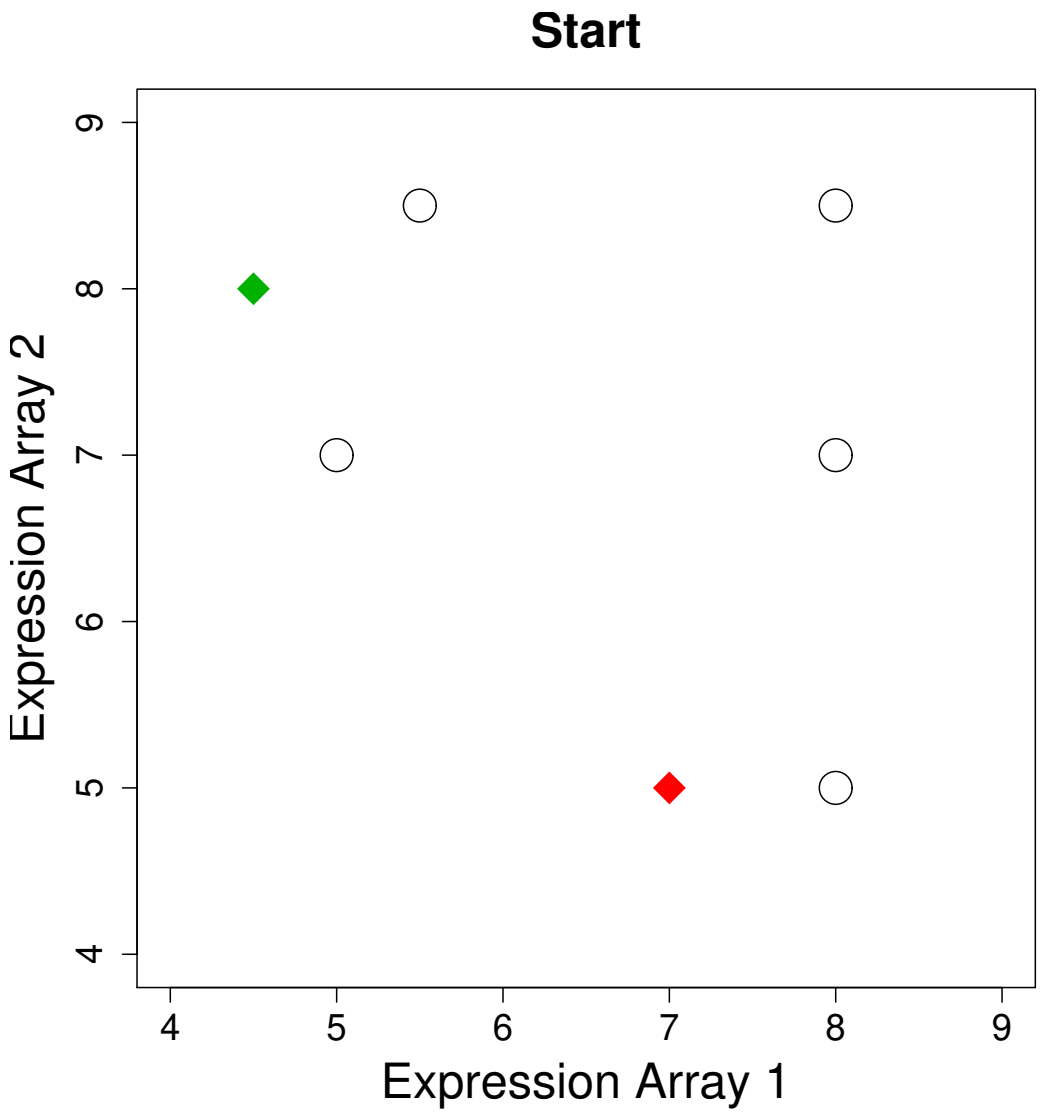
Partitioning - k-means

Choice of cluster number and initial starting values for cluster centroids:

- Changing the number of clusters can completely change the cluster structure. This is contrary to hierarchical clustering.
- k-means is a randomised algorithm: two runs usually produce different results. Thus, it has to be applied several times and the result with minimal sum of within-cluster-variances should be chosen.
  Even when doing so, we are not guaranteed to find the best solution.

Example: same data set (5 patients, 2 expression values)
Method: **Partitioning clustering** with k-**means** with $k = 2$

Partitioning - *k*-means example 2



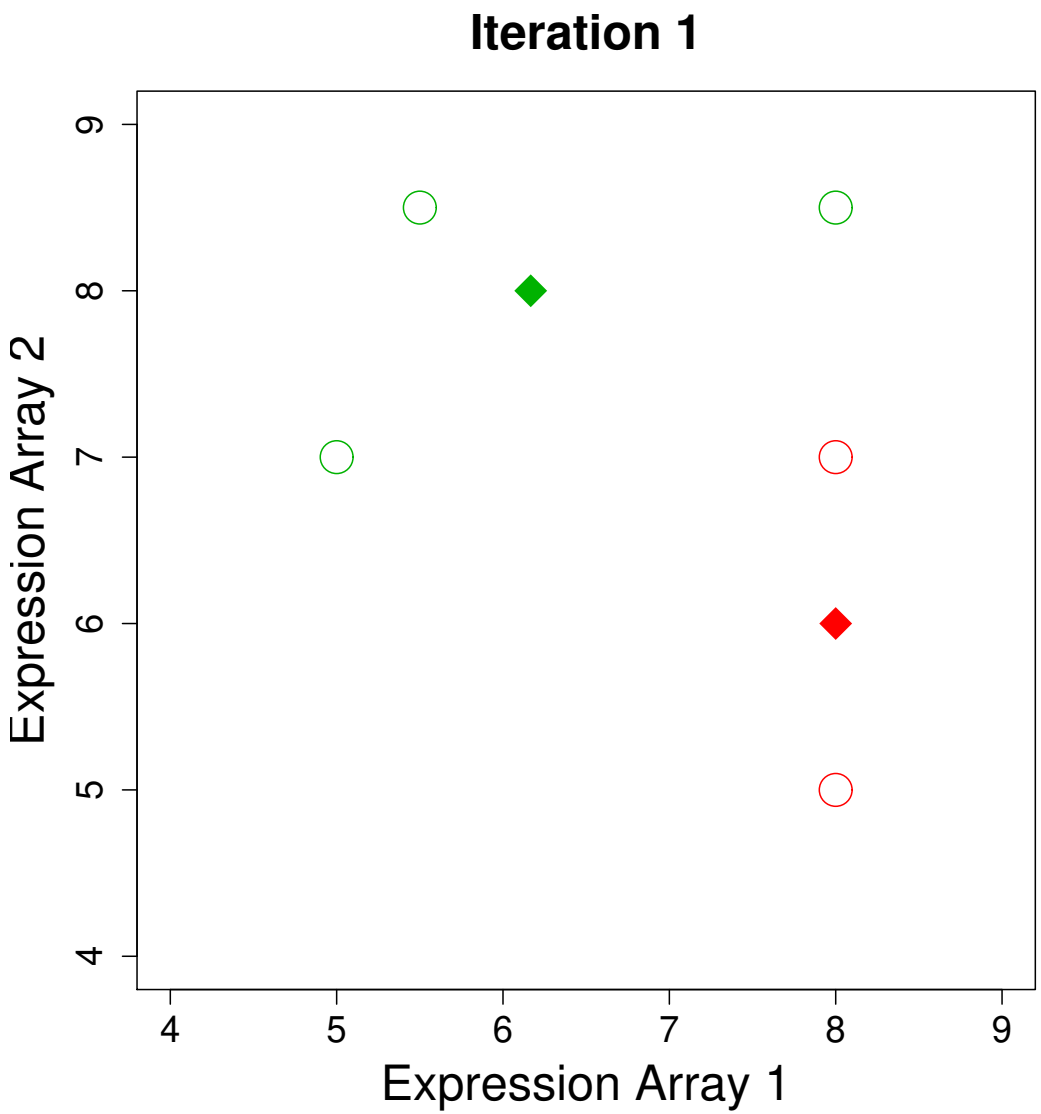
Centroid matrix

$$\begin{pmatrix} 4.5 & 8.0 \\ 7.0 & 5.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example 2



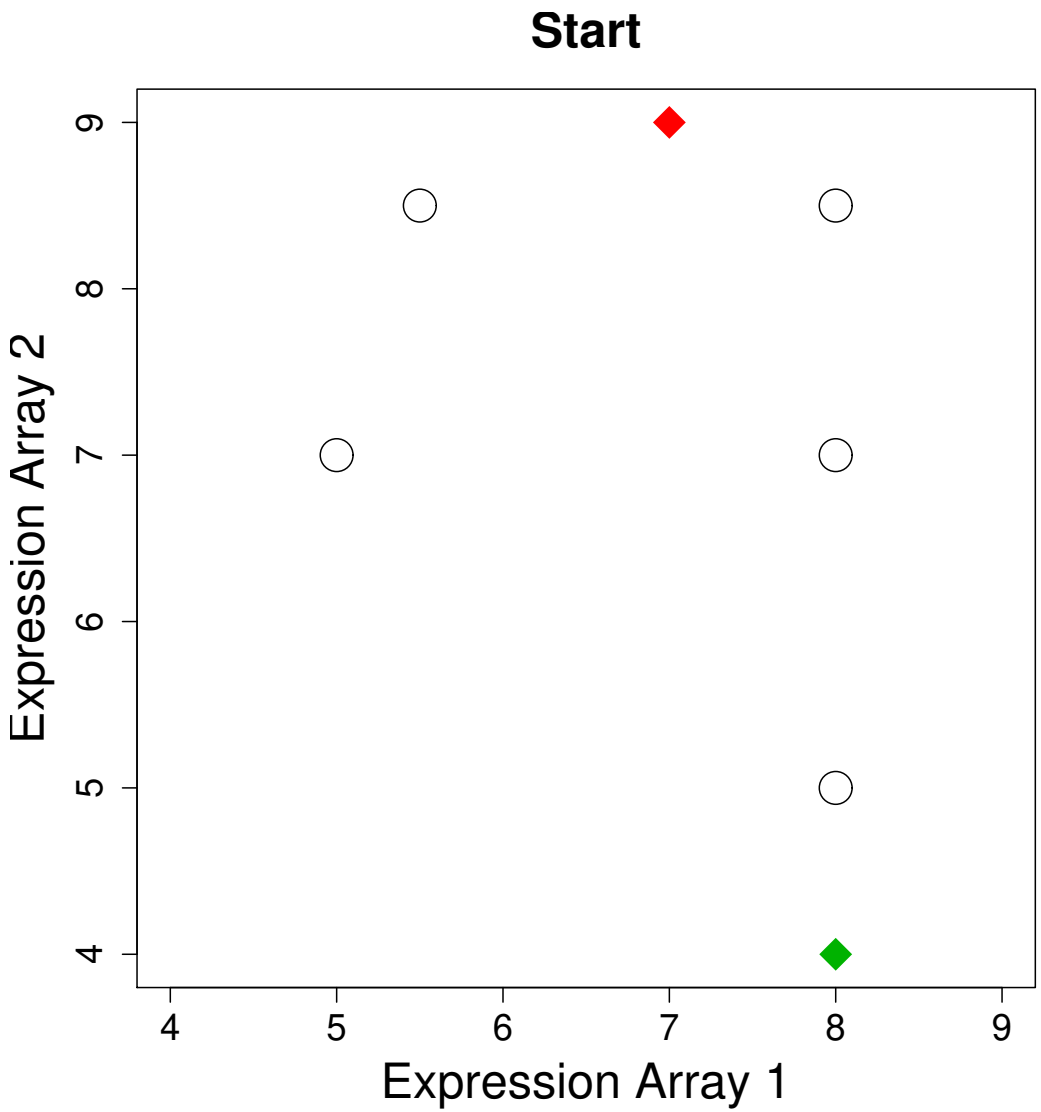
Centroid matrix

$$\begin{pmatrix} 6.17 & 8.0 \\ 8.0 & 6.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example 3



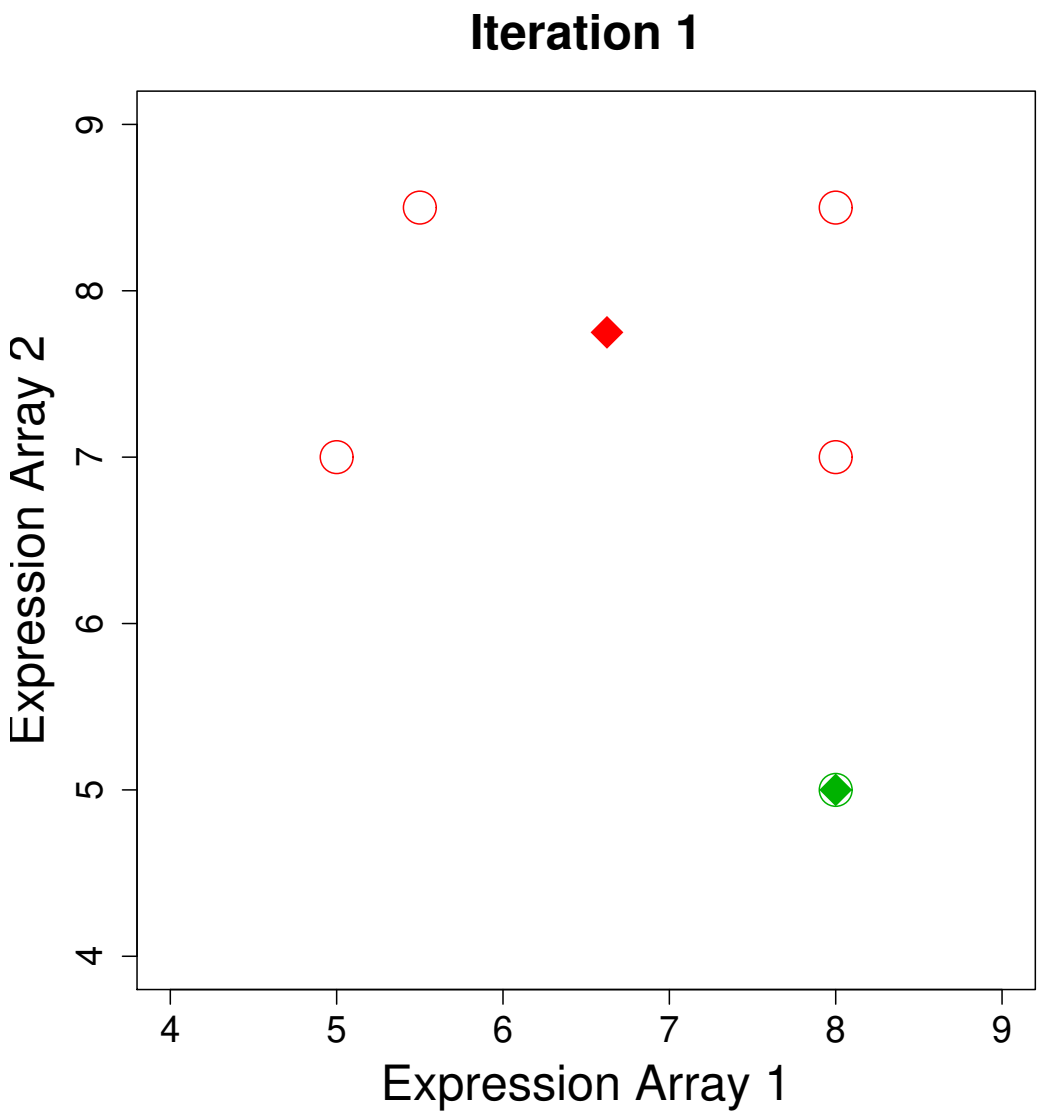
Centroid matrix

$$\begin{pmatrix} 8.0 & 4.0 \\ 7.0 & 9.0 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

Partitioning - *k*-means example 3



Centroid matrix

$$\begin{pmatrix} 8.0 & 5.0 \\ 6.63 & 7.75 \end{pmatrix}$$

Expression matrix

$$\begin{pmatrix} 5.0 & 7.0 \\ 5.5 & 8.5 \\ 8.0 & 8.5 \\ 8.0 & 7.0 \\ 8.0 & 5.0 \end{pmatrix}$$

1. Introduction
   - Motivation
   - Notions of similarity
2. Hierarchical Clustering Algorithms
   - Main ideas
   - Illustrative example
   - Visualization
   - Summary
3. Partitioning clustering: k-means
   - Partitioning Clustering Algorithms
   - Main idea of the k-means
   - Illustrative example 1
   - Illustrative example 2
4. Summary

Clustering - Summary

Hierarchical clustering:

- Procedure provides a hierarchy of the clustering, with the number of clusters ranging from 1 to the number of objects
- Incorrect merges early in the tree cannot be changed later on
- Careful with interpretation of dendrograms and resulting heatmaps! The order of objects within a cluster is arbitrary

Partitioning algorithms:

- Careful with initalization...
- How to choose the correct number of groups?

Clustering - Summary

All clustering methods:

- The choice of the distance measure depends on the data and the intention of the clustering
- Use objective measures to support the decision for number of clusters
- Even data generated at random will result in clusters
- Be careful with pre-selection of features before clustering!

Practical Issues in Clustering

- Some decisions to be made:
  - Should the variables be standardized first? Other transformations needed to achieve normally distributed data?
  - Do we need a pre-selection of variables? This needs to be unsupervised, i.e. using variances not t-tests
- In case of hierarchical clustering:
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?
- In case of partitioning clustering:
  - How many clusters should we look for?
- Try several choices, and look for clustering results that are most useful for interpretation. There is no single right answer!

R/ Bioconductor

Heatmaps

- heatmap() (stats) and many enhancements, e.g. pheatmap() (pheatmap) and Heatmap() (ComplexHeatmap)

Hierarchical clustering

- hclust() (stats), hcluster (amap)

Partitioning clustering

- kmeans() (stats), pam() (cluster)

CRAN Task View Cluster

`https://cran.r-project.org/web/views/Cluster.html`