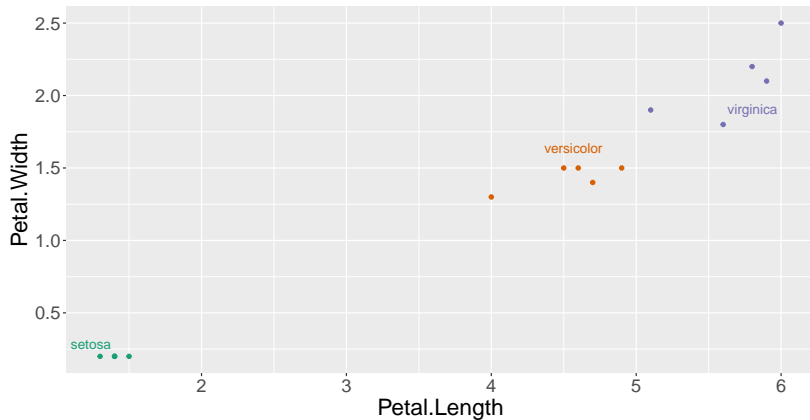


Hierarchical Clustering

Toby Dylan Hocking

Visualize iris data with labels



Visualize iris data without labels

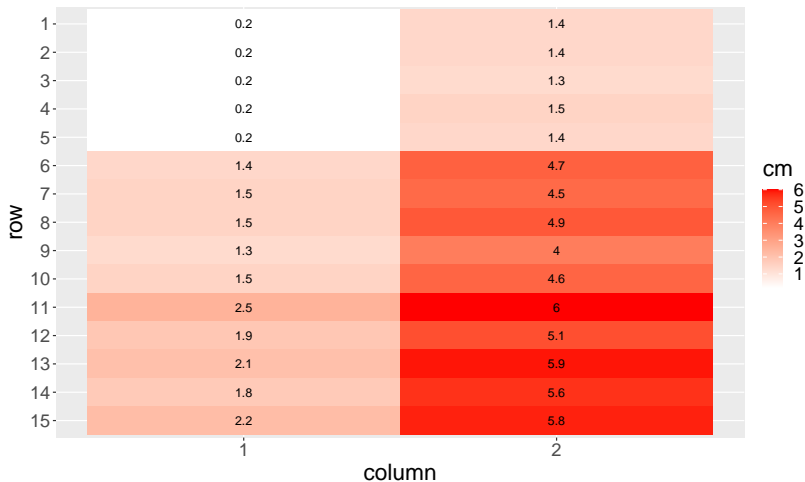
- ▶ Let $X = [x_1 \cdots x_n]^T \in \mathbb{R}^{n \times p}$ be the data matrix (input for clustering), where $x_i \in \mathbb{R}^p$ is the input vector for observation i .
- ▶ Example iris $n = 150$ observations, $p = 2$ dimensions.

##	Petal.Width	Petal.Length
## [1,]	0.2	1.4
## [2,]	0.2	1.4
## [3,]	0.2	1.3
## [4,]	0.2	1.5



Which pair of rows is most similar?

This is a visualization of 15 rows and two columns from the iris data.



Hyper-parameter choices (must be fixed prior to learning)

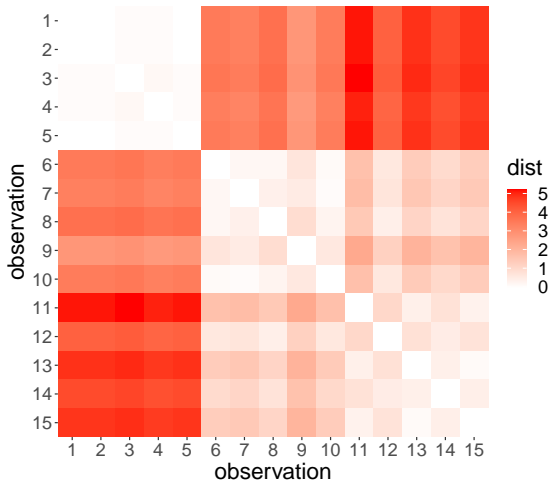
How to compute similarity/distance between rows?

- ▶ Let $x, x' \in \mathbb{R}^p$ be two feature vectors (rows of data matrix).
- ▶ L1/manhattan distance: $\|x - x'\|_1 = \sum_{j=1}^p |x_j - x'_j|$.
- ▶ L2/euclidean distance: $\|x - x'\|_2 = \sqrt{\sum_{j=1}^p (x_j - x'_j)^2}$.

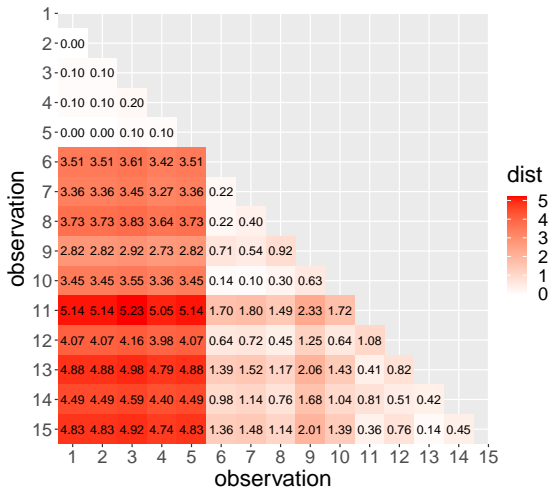
How to compute distance with a group/cluster? There are several rules, or agglomeration methods:

- ▶ single: min distance from any point,
- ▶ complete: max distance from any point,
- ▶ average: mean distance over all points,
- ▶ there are others.

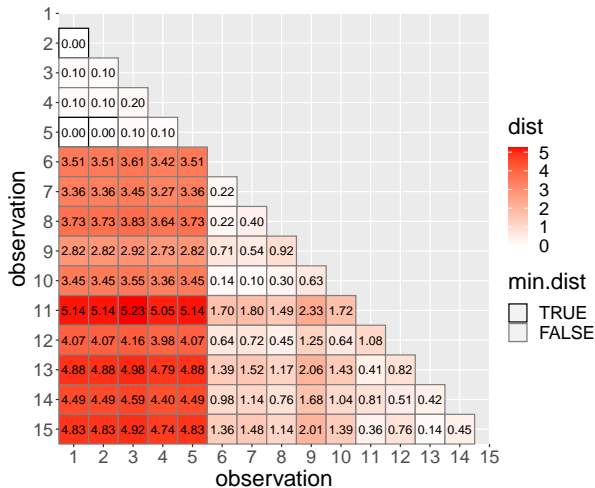
Hierarchical clustering inputs a pairwise distance matrix



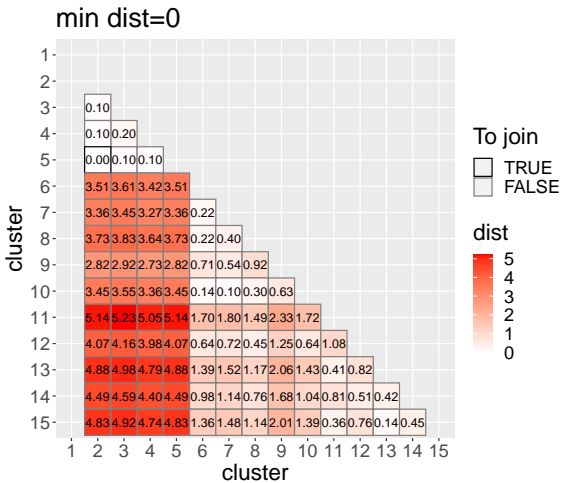
Only need lower triangle (symmetry)



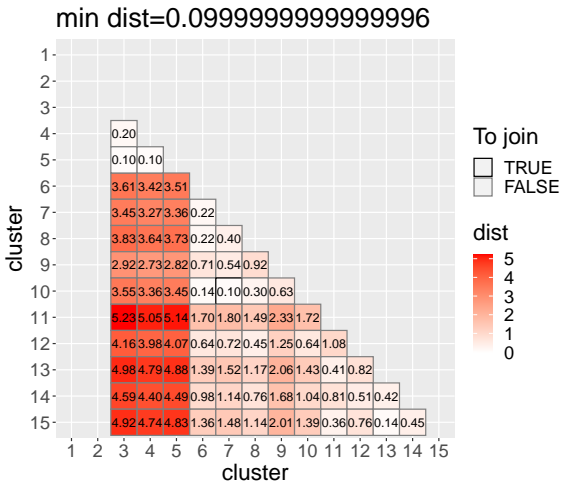
Find the closest pairs



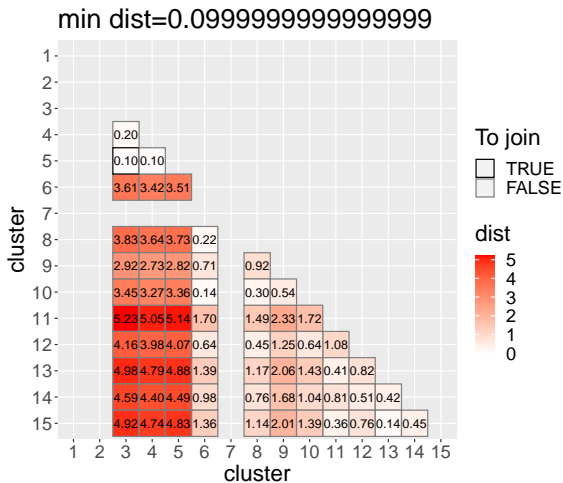
iteration 2



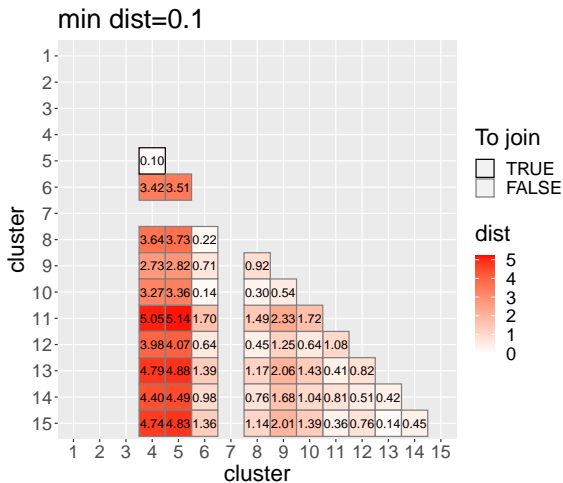
iteration 3



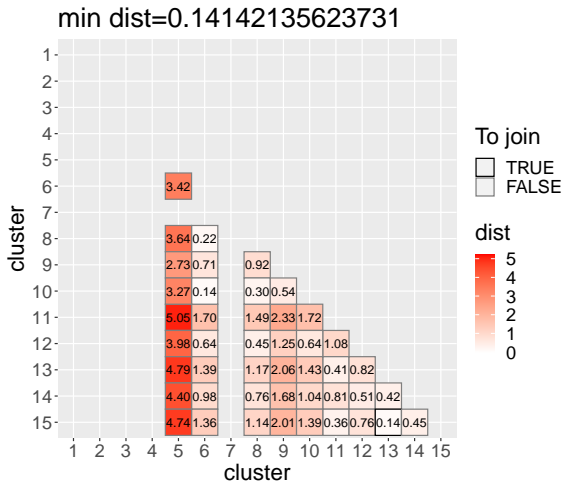
iteration 4



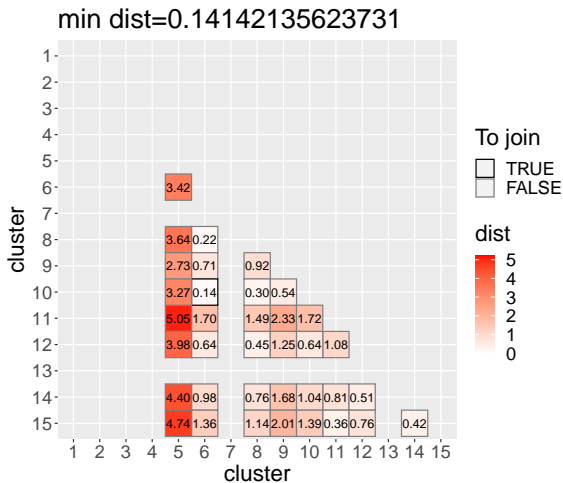
iteration 5



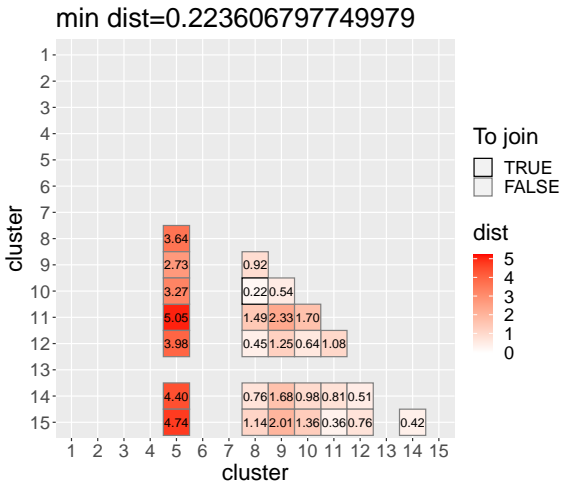
iteration 6



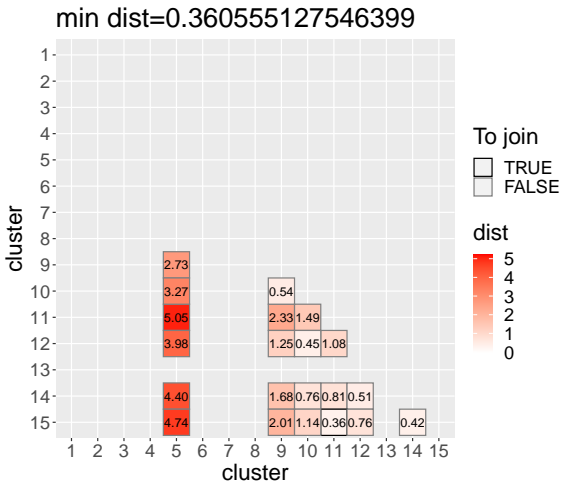
iteration 7



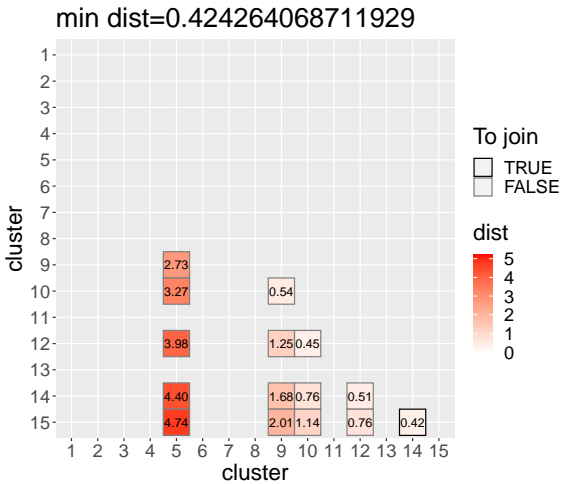
iteration 8



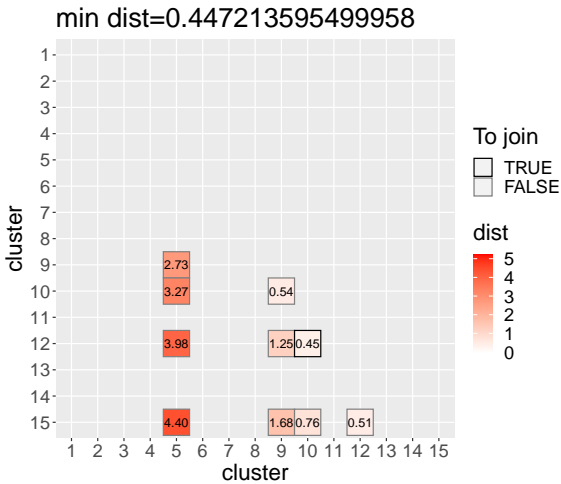
iteration 9



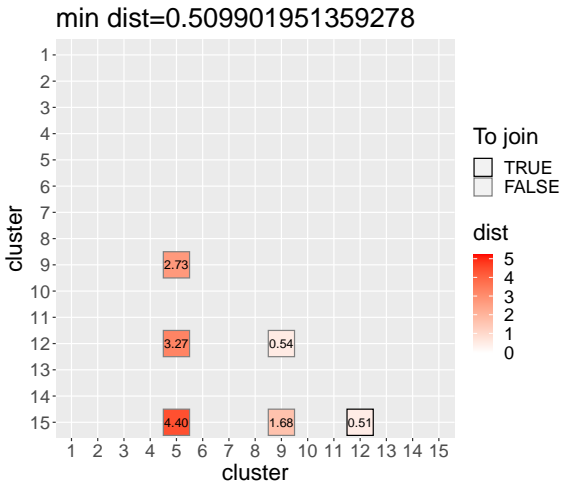
iteration 10



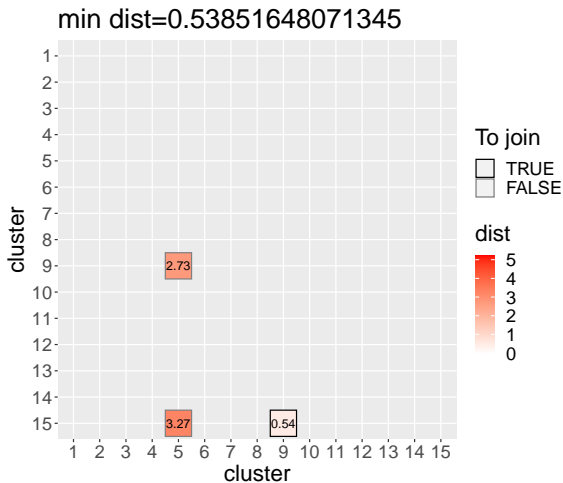
iteration 11



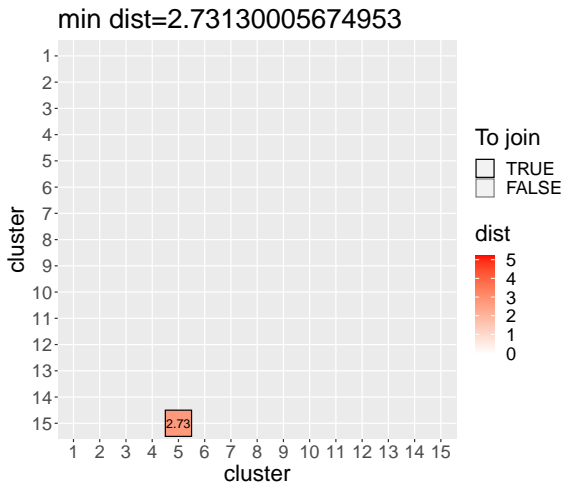
iteration 12



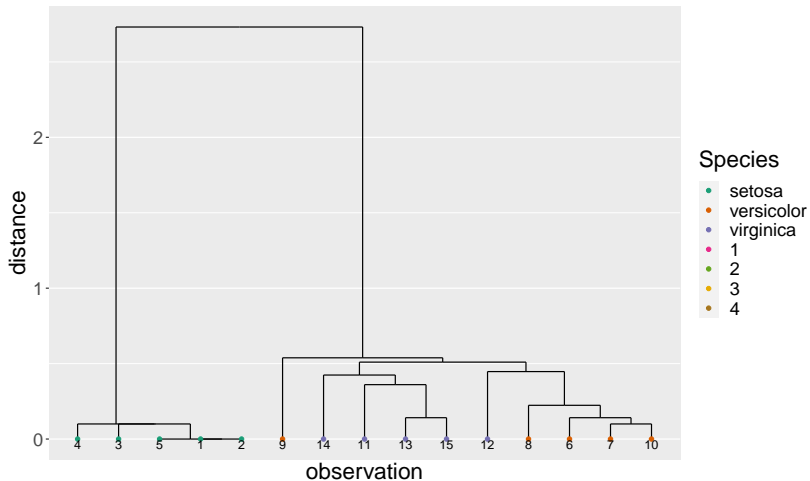
iteration 13



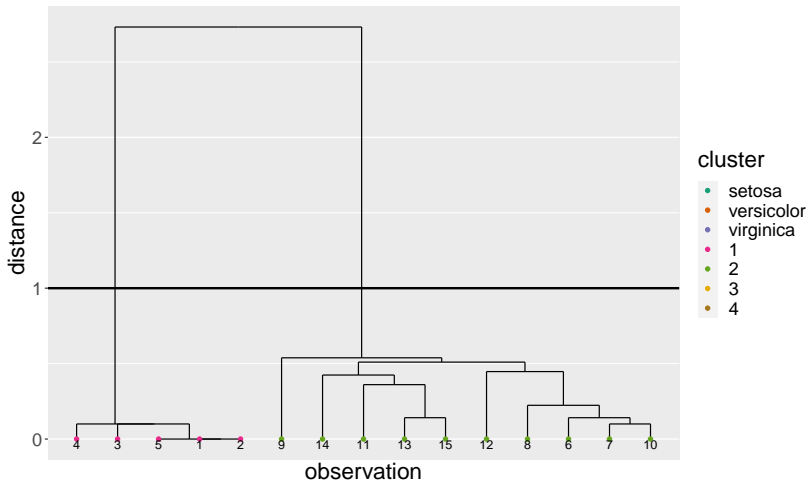
iteration 14



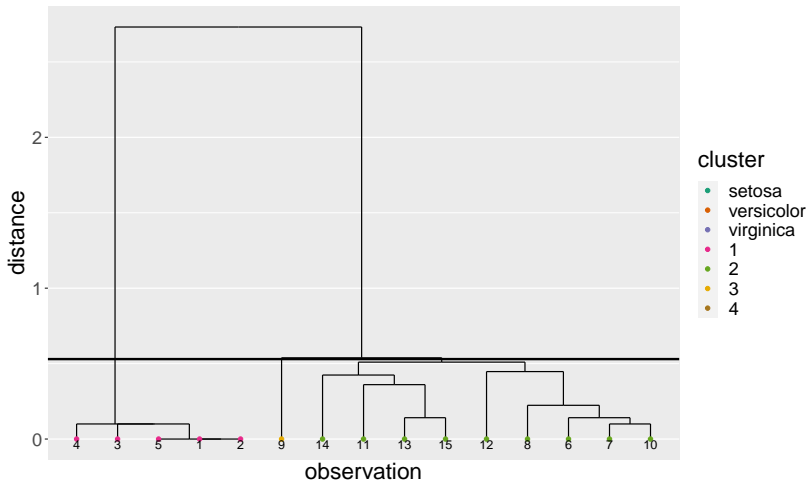
Visualization of dendrogram (tree diagram)



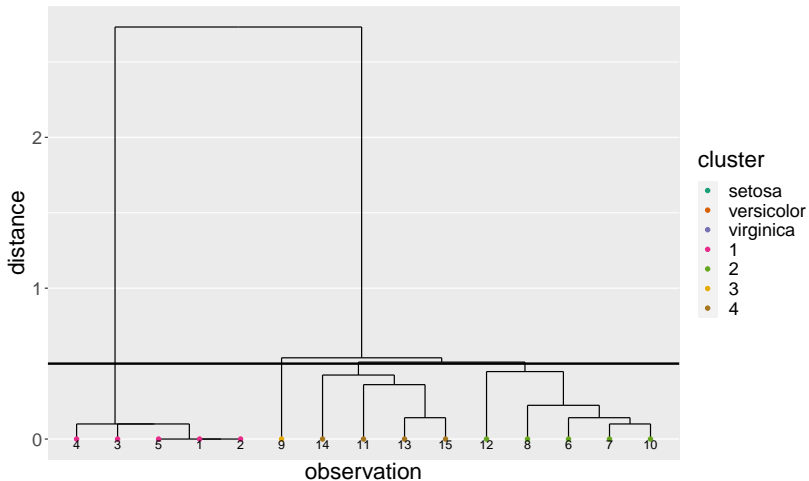
Cutting the tree to get two clusters



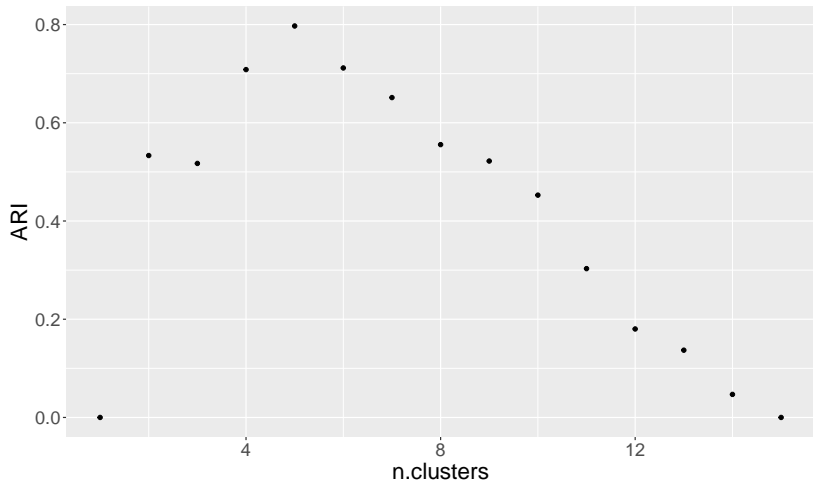
Cutting the tree to get three clusters



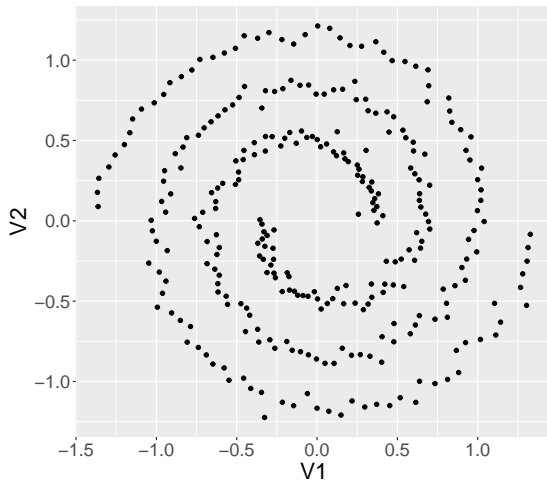
Cutting the tree to get four clusters



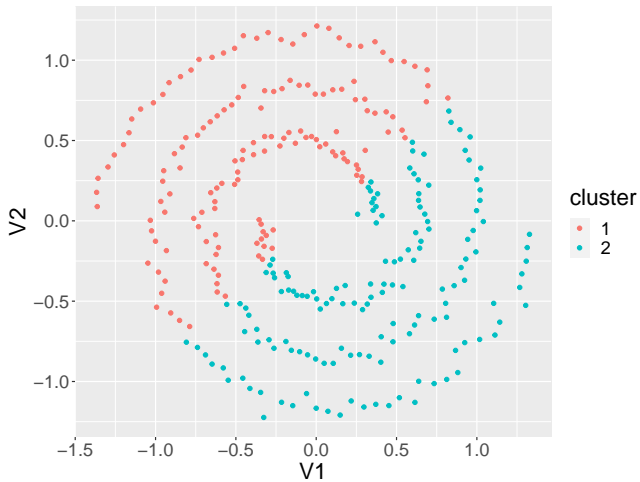
ARI computation



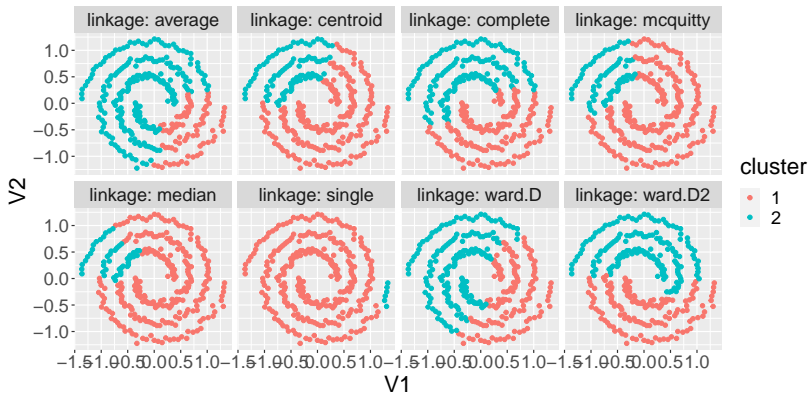
Another 2d data set



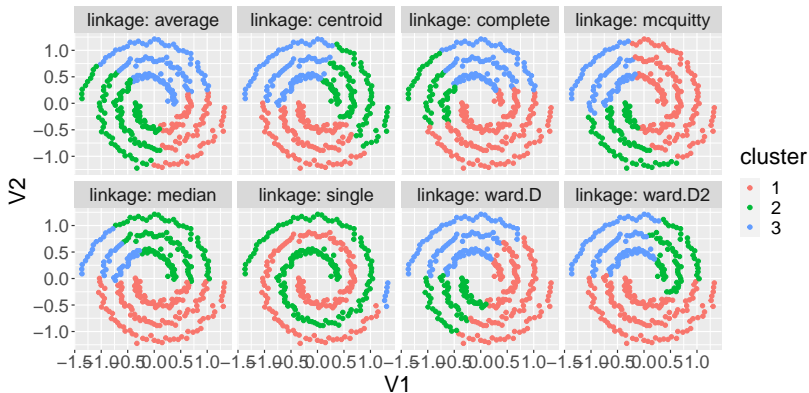
K-means fails



Hierarchical with two clusters fails

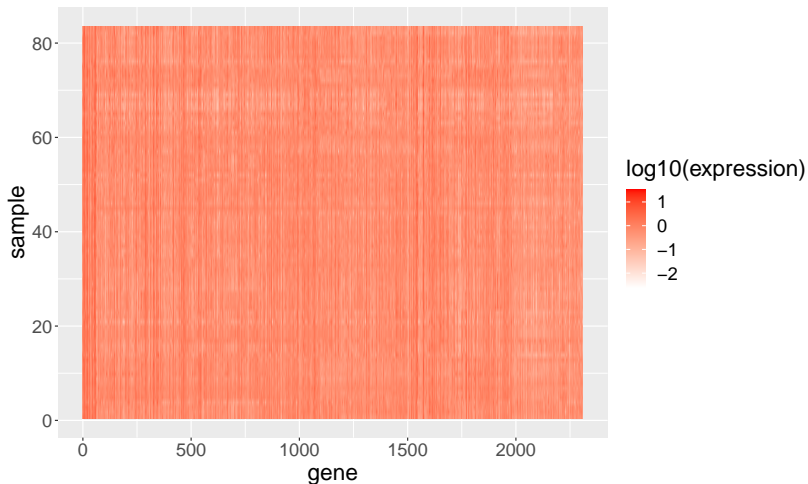


Hierarchical with single linkage and 3 clusters better

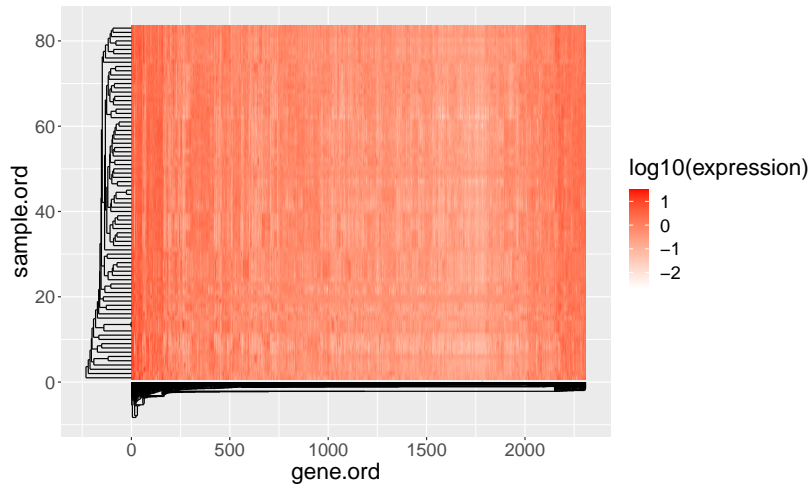


Gene expression clustering (cancer classes)

Small Round Blue Cell Tumors (SRBCT) of childhood cancer study of Khan et al. (2001).



Clustering samples



Possible Exam Questions

What is the big O notation asymptotic time complexity of the following algorithms in terms of N (number of data observations/rows), P (number of data features/columns), and K (number of clusters).

- ▶ K-means.
- ▶ Gaussian mixture model with diagonal covariance matrix.
- ▶ Gaussian mixture model with unconstrained covariance matrix.
- ▶ Hierarchical clustering with single linkage.

Possible Exam Questions 2

- ▶ What are the two hyper-parameters that must be chosen before running the hierarchical clustering algorithm?
- ▶ For a data set with $N=200$ observations/rows, how large is the pairwise distance matrix? How many iterations of the cluster joining occur?