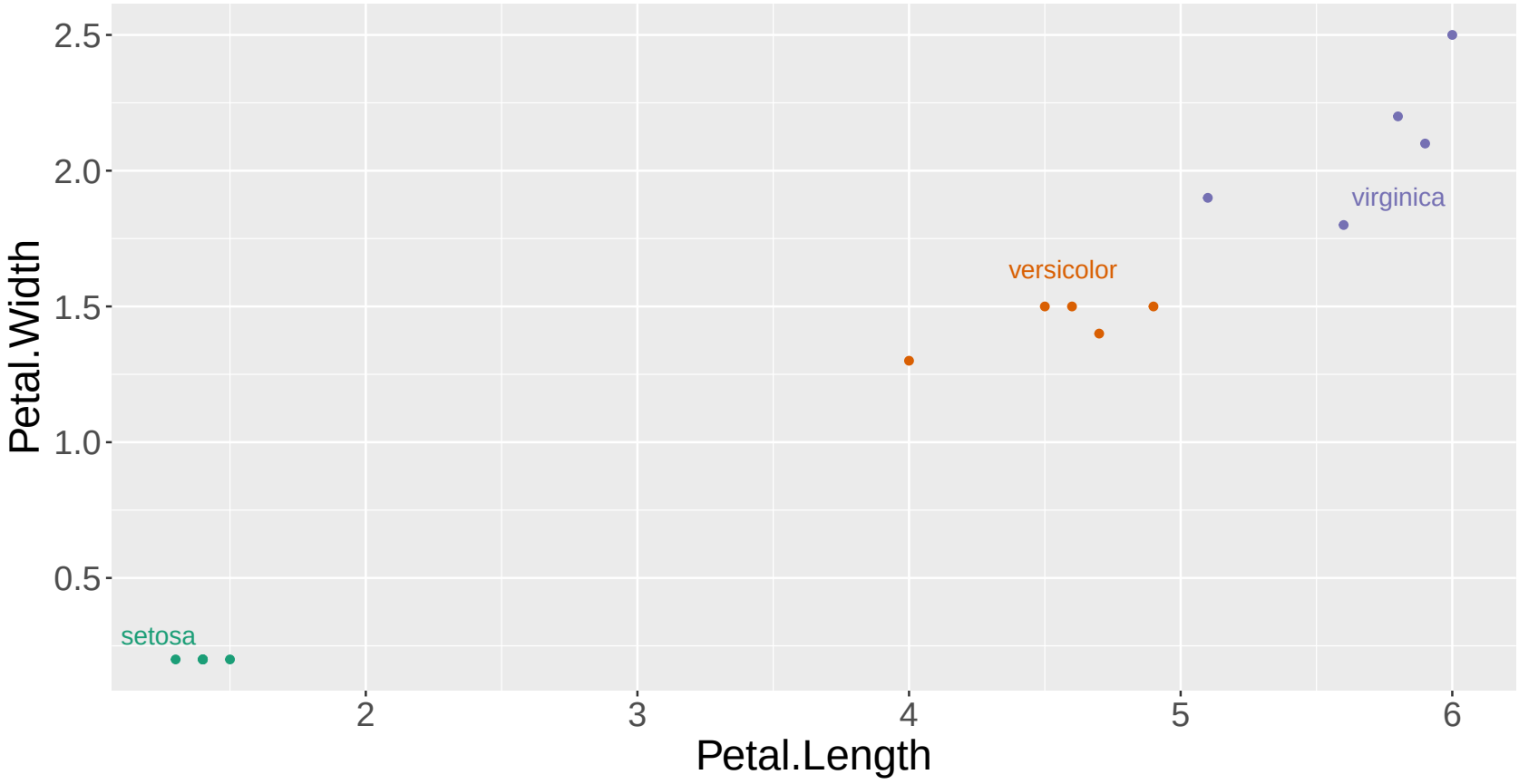
Hierarchical Clustering

Toby Dylan Hocking

Visualize iris data with labels



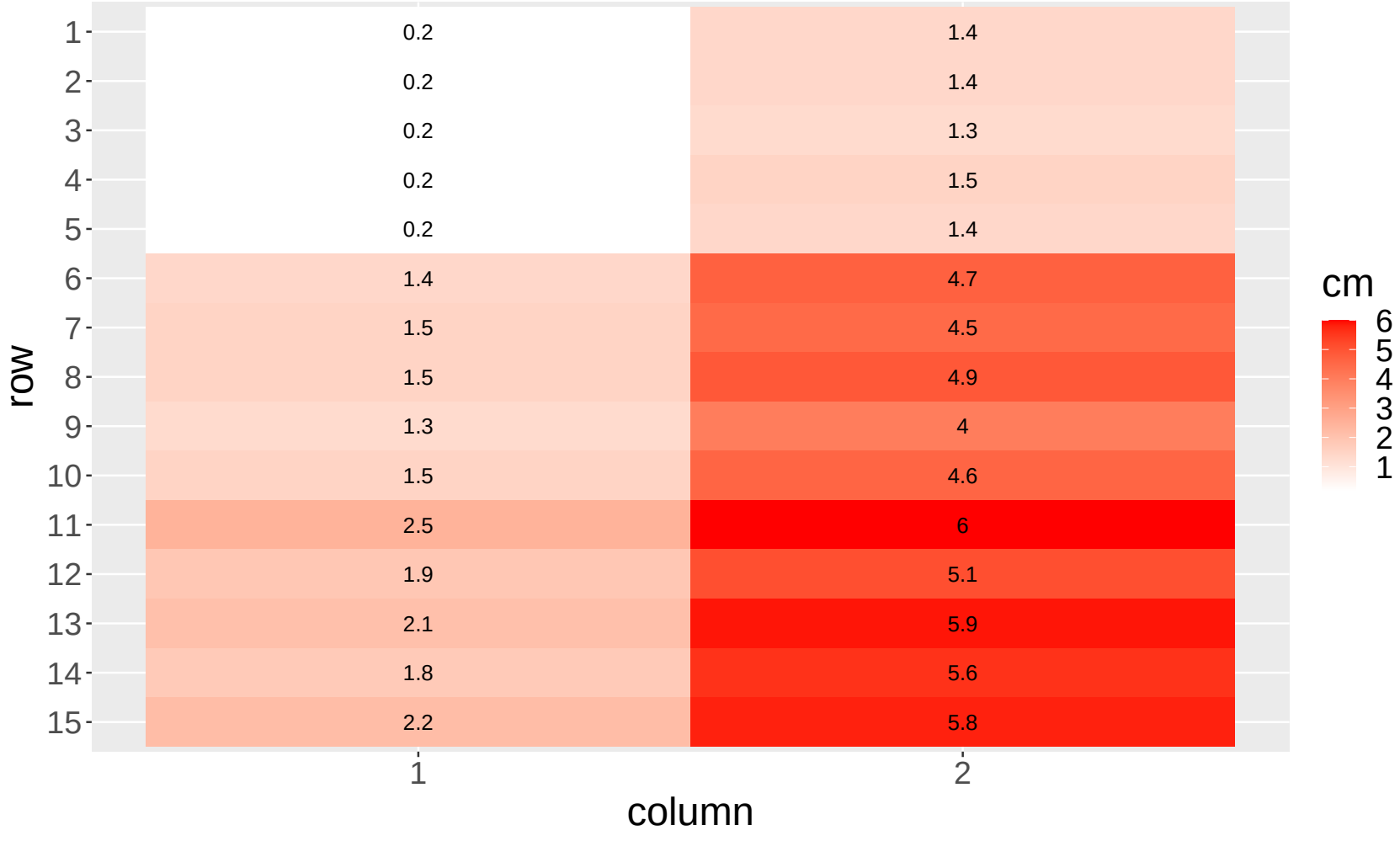
Visualize iris data without labels

- Let $X = [x_1 \cdots x_n]^\intercal \in \mathbb{R}^{n \times p}$ 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)

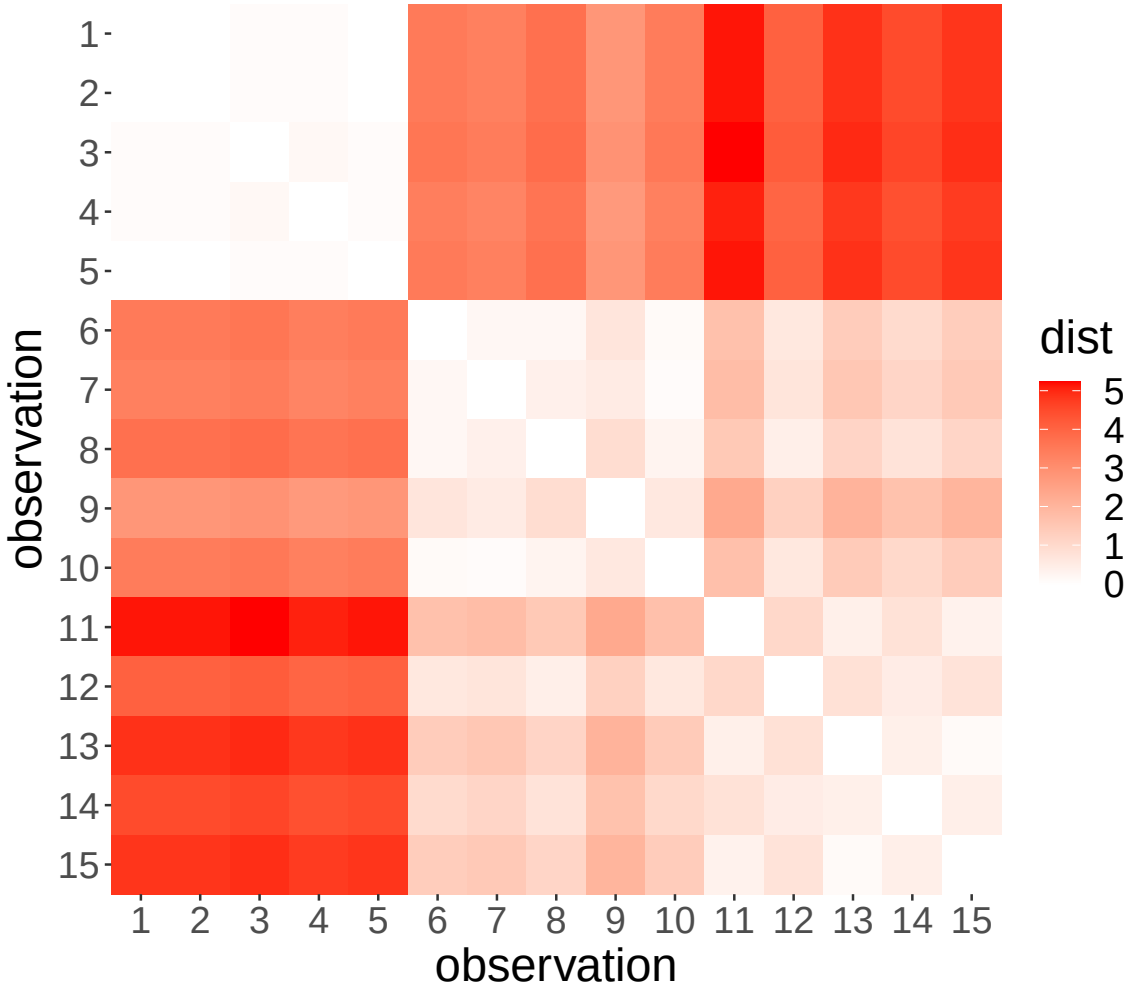
How to compute similarity/distance between rows?

- 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: $||x - x'||_2 = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2}$.

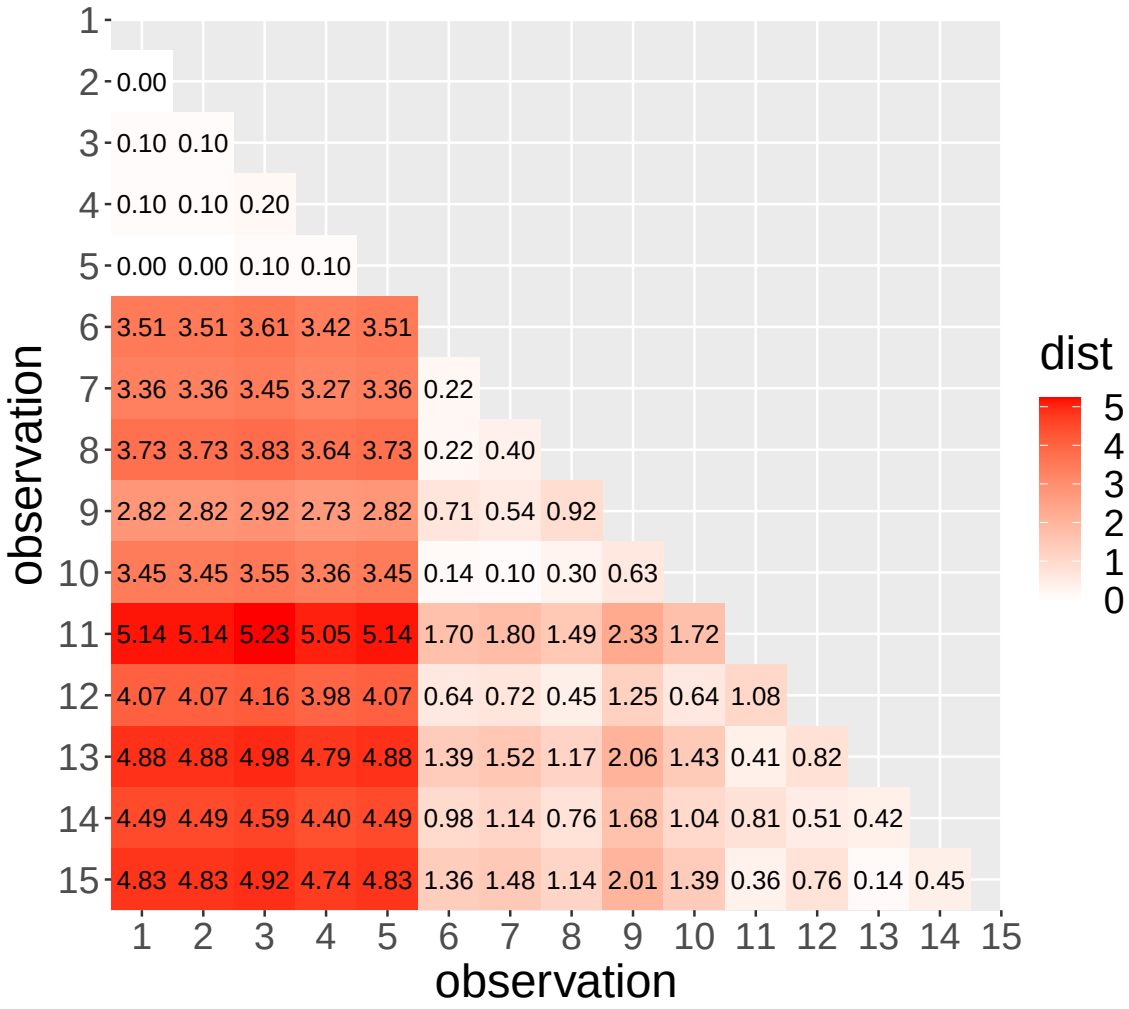
How to compute distance with a group/cluster? There are several rules, or agglomeration methods:

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

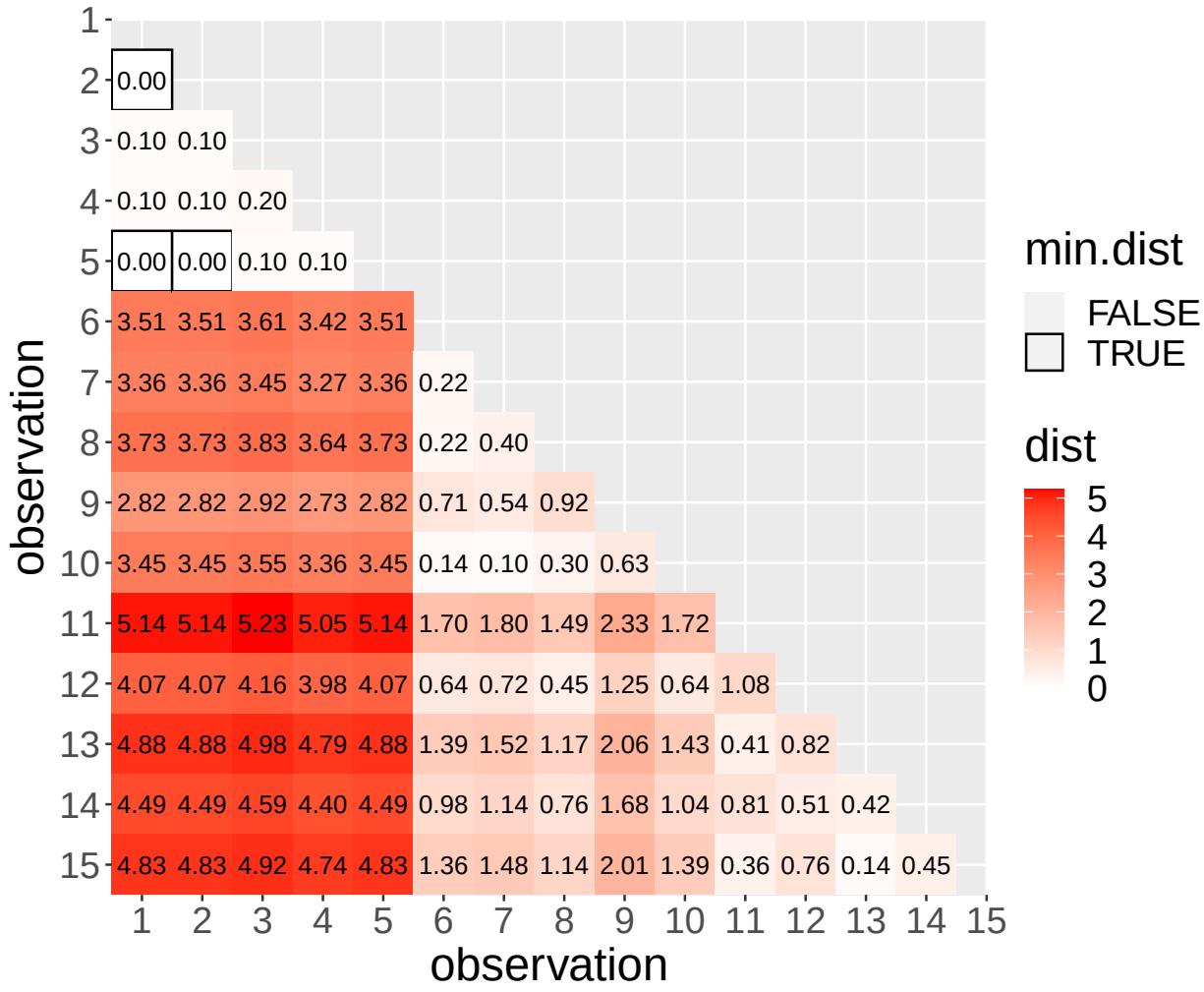
Hierarchical clustering inputs a pairwise distance matrix



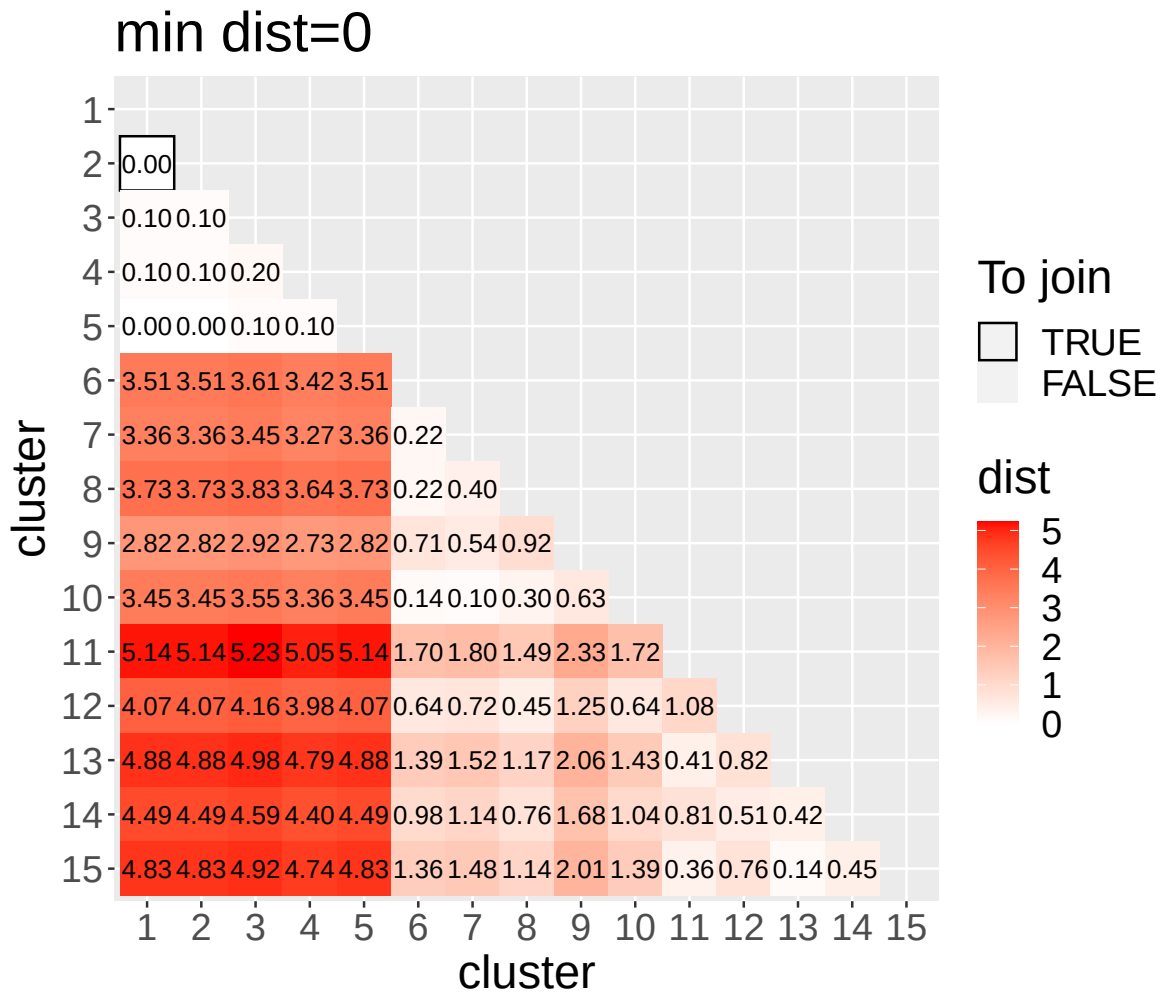
Only need lower triangle (symmetry)



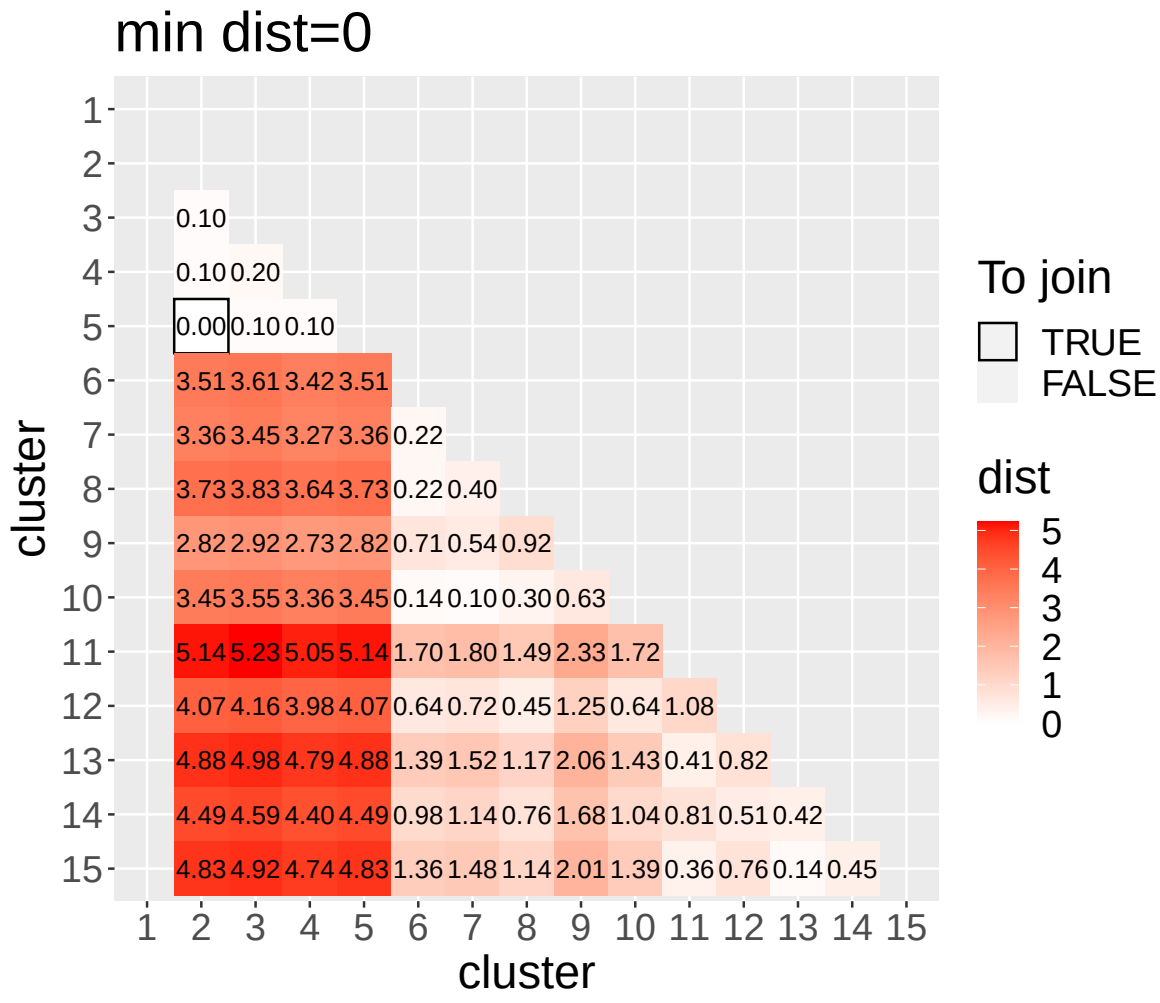
Find the closest pairs



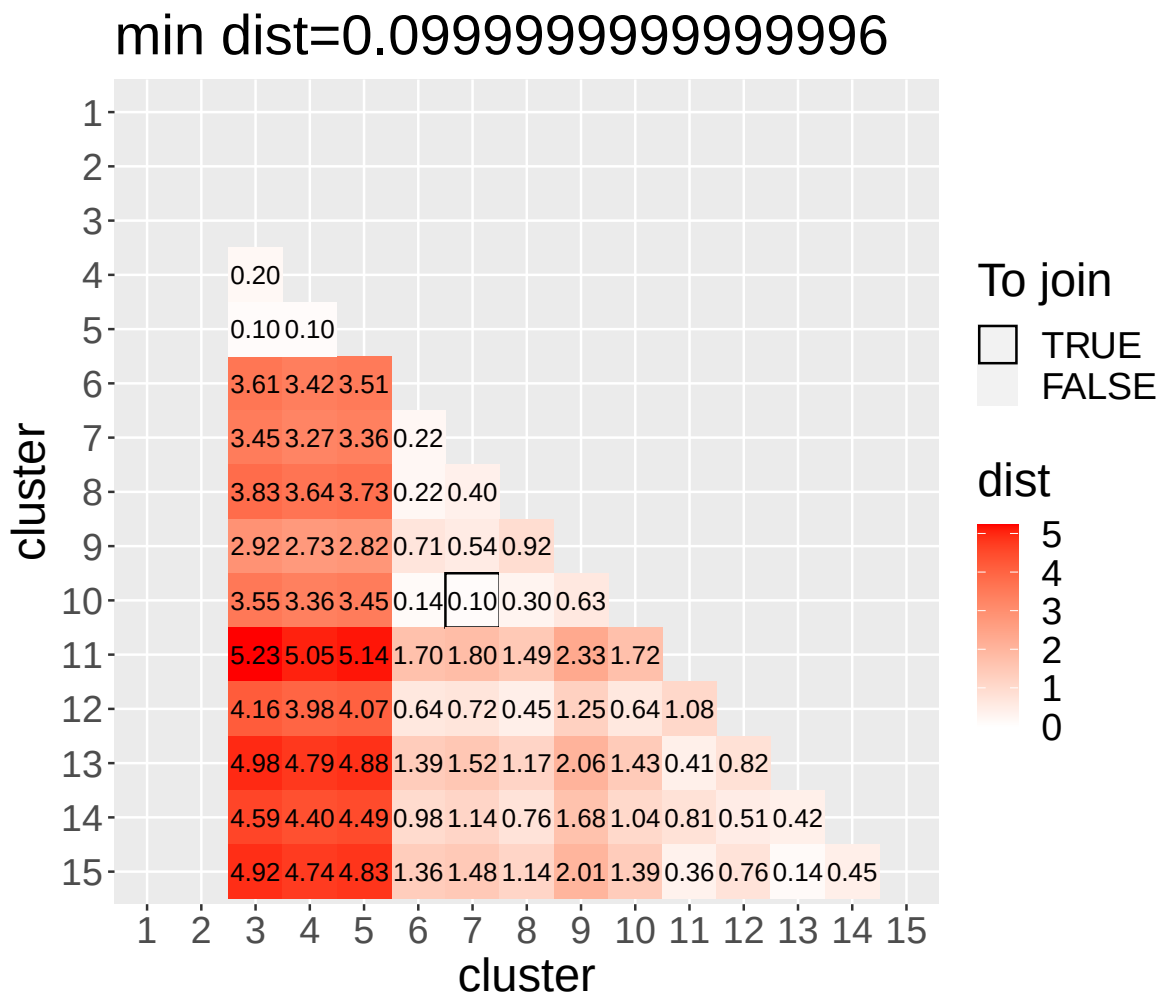
Join one of the closest pairs (iteration 1)



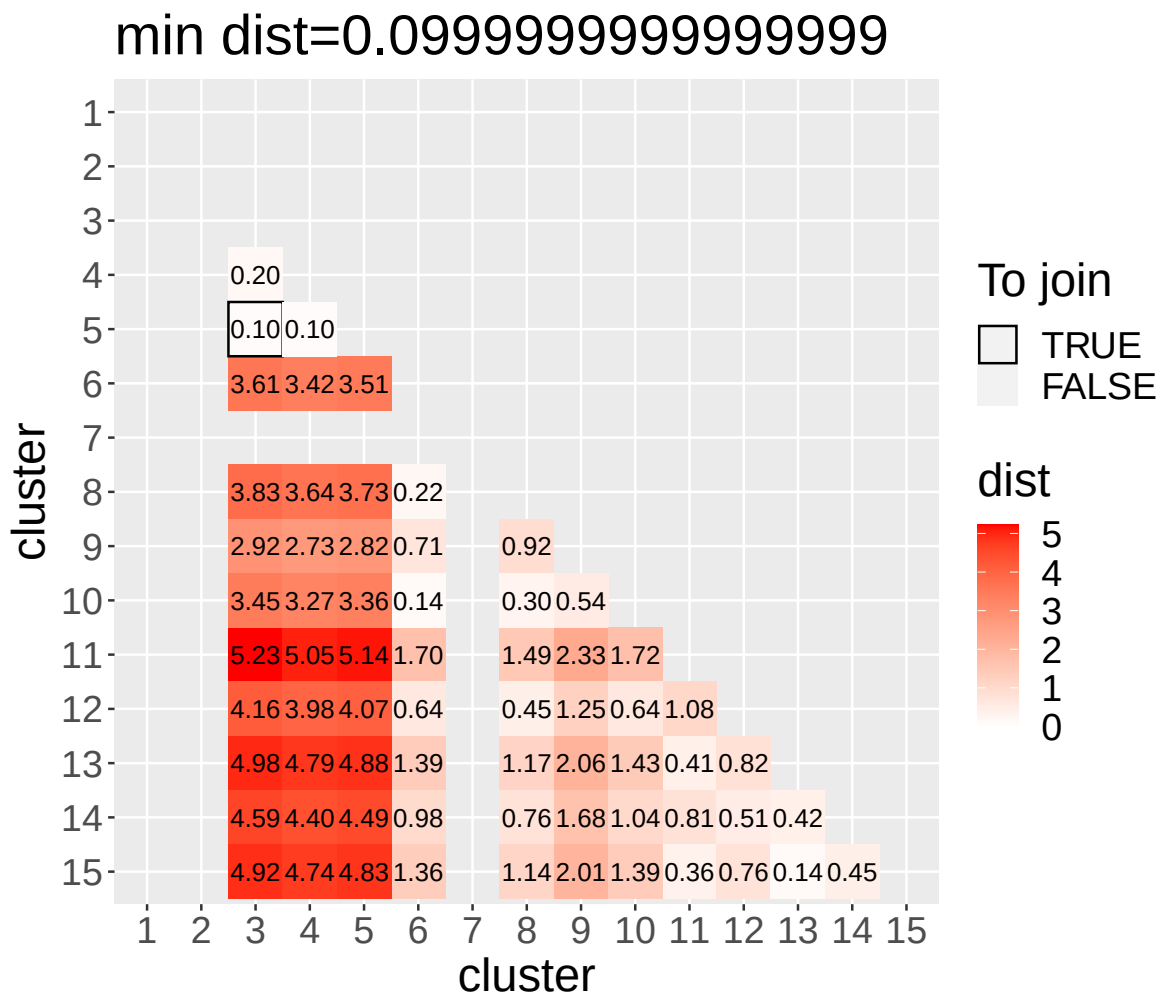
iteration 2


min dist=0
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.10
0.10 0.20
0.00 0.10 0.10
3.51 3.61 3.42 3.51
3.36 3.45 3.27 3.36 0.22
3.73 3.83 3.64 3.73 0.22 0.40
2.82 2.92 2.73 2.82 0.71 0.54 0.92
3.45 3.55 3.36 3.45 0.14 0.10 0.30 0.63
5.14 5.23 5.05 5.14 1.70 1.80 1.49 2.33 1.72
4.07 4.16 3.98 4.07 0.64 0.72 0.45 1.25 0.64 1.08
4.88 4.98 4.79 4.88 1.39 1.52 1.17 2.06 1.43 0.41 0.82
4.49 4.59 4.40 4.49 0.98 1.14 0.76 1.68 1.04 0.81 0.51 0.42
4.83 4.92 4.74 4.83 1.36 1.48 1.14 2.01 1.39 0.36 0.76 0.14 0.45
cluster
To join
TRUE
FALSE
dist
5
4
3
2
1
0

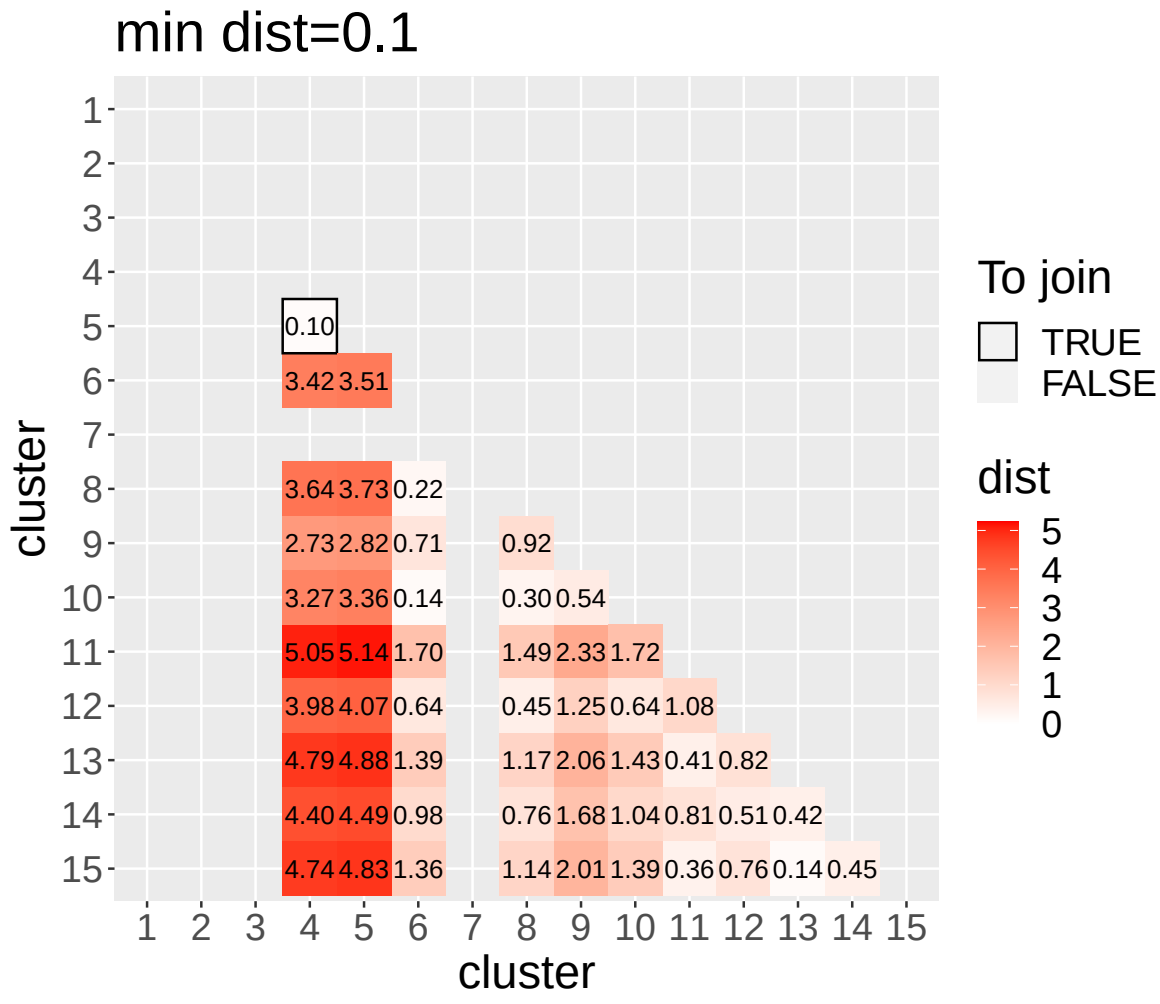
iteration 3


min dist=0.0999999999999996
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.20
0.10 0.10
3.61 3.42 3.51
3.45 3.27 3.36 0.22
3.83 3.64 3.73 0.22 0.40
2.92 2.73 2.82 0.71 0.54 0.92
3.55 3.36 3.45 0.14 0.10 0.30 0.63
5.23 5.05 5.14 1.70 1.80 1.49 2.33 1.72
4.16 3.98 4.07 0.64 0.72 0.45 1.25 0.64 1.08
4.98 4.79 4.88 1.39 1.52 1.17 2.06 1.43 0.41 0.82
4.59 4.40 4.49 0.98 1.14 0.76 1.68 1.04 0.81 0.51 0.42
4.92 4.74 4.83 1.36 1.48 1.14 2.01 1.39 0.36 0.76 0.14 0.45
cluster
To join
TRUE
FALSE
dist
5
4
3
2
1
0

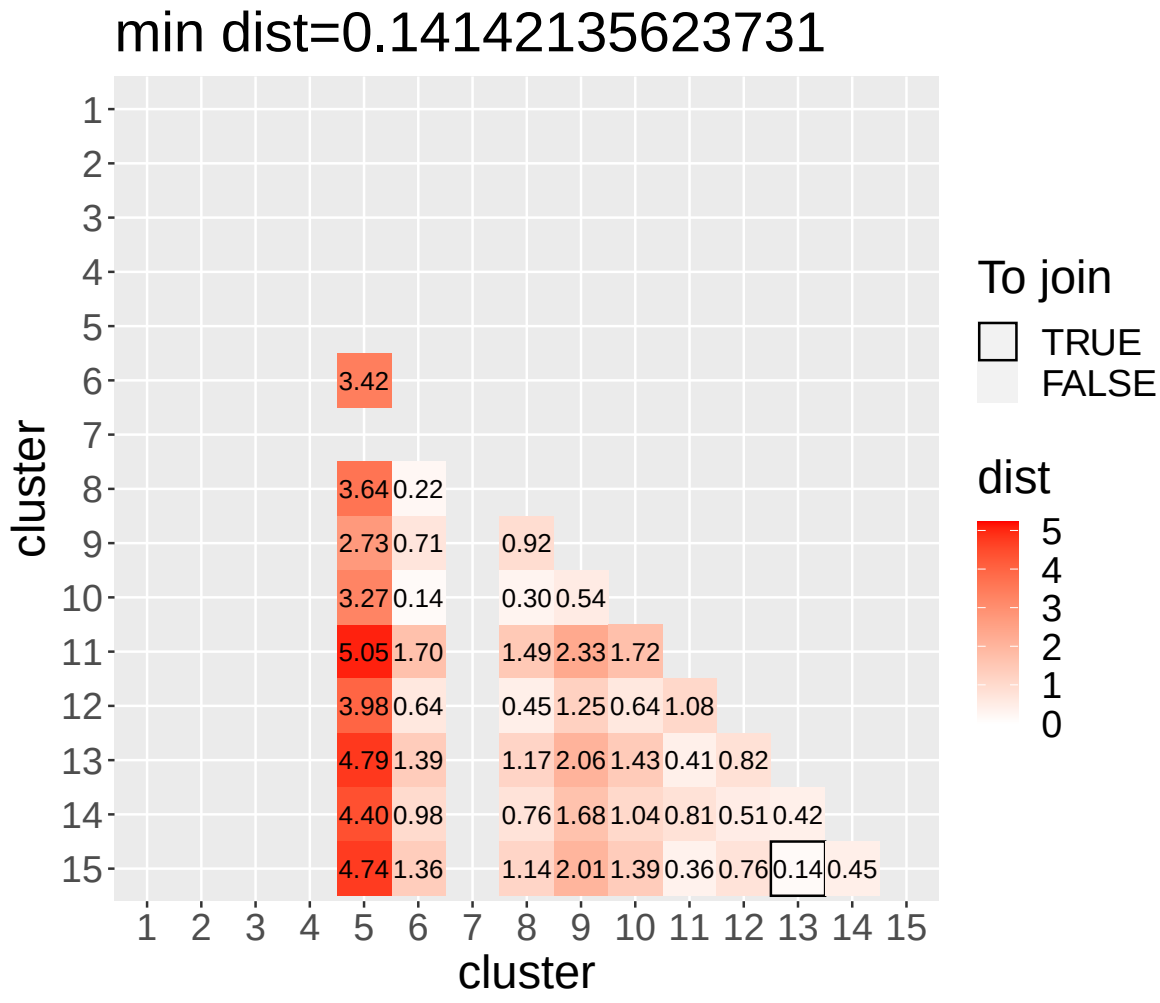
iteration 4


min dist=0.0999999999999999
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.20
0.10 0.10
3.61 3.42 3.51
3.83 3.64 3.73 0.22
2.92 2.73 2.82 0.71 0.92
3.45 3.27 3.36 0.14 0.30 0.54
5.23 5.05 5.14 1.70 1.49 2.33 1.72
4.16 3.98 4.07 0.64 0.45 1.25 0.64 1.08
4.98 4.79 4.88 1.39 1.17 2.06 1.43 0.41 0.82
4.59 4.40 4.49 0.98 0.76 1.68 1.04 0.81 0.51 0.42
4.92 4.74 4.83 1.36 1.14 2.01 1.39 0.36 0.76 0.14 0.45
cluster
To join
TRUE
FALSE
dist
5
4
3
2
1
0

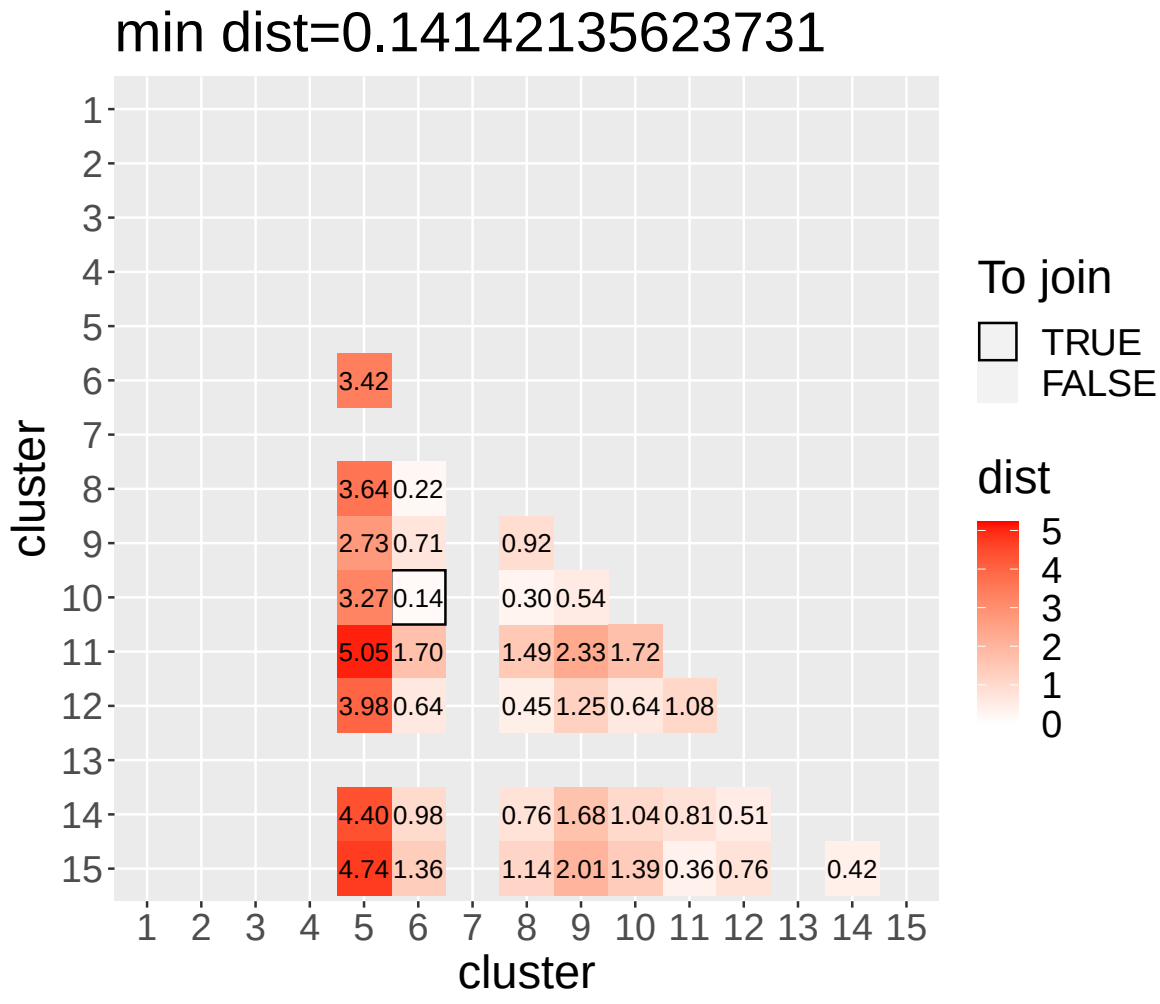
iteration 5


min dist=0.1
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.10
3.42 3.51
3.64 3.73 0.22
2.73 2.82 0.71 0.92
3.27 3.36 0.14 0.30 0.54
5.05 5.14 1.70 1.49 2.33 1.72
3.98 4.07 0.64 0.45 1.25 0.64 1.08
4.79 4.88 1.39 1.17 2.06 1.43 0.41 0.82
4.40 4.49 0.98 0.76 1.68 1.04 0.81 0.51 0.42
4.74 4.83 1.36 1.14 2.01 1.39 0.36 0.76 0.14 0.45
To join
TRUE
FALSE
dist
5
4
3
2
1
0

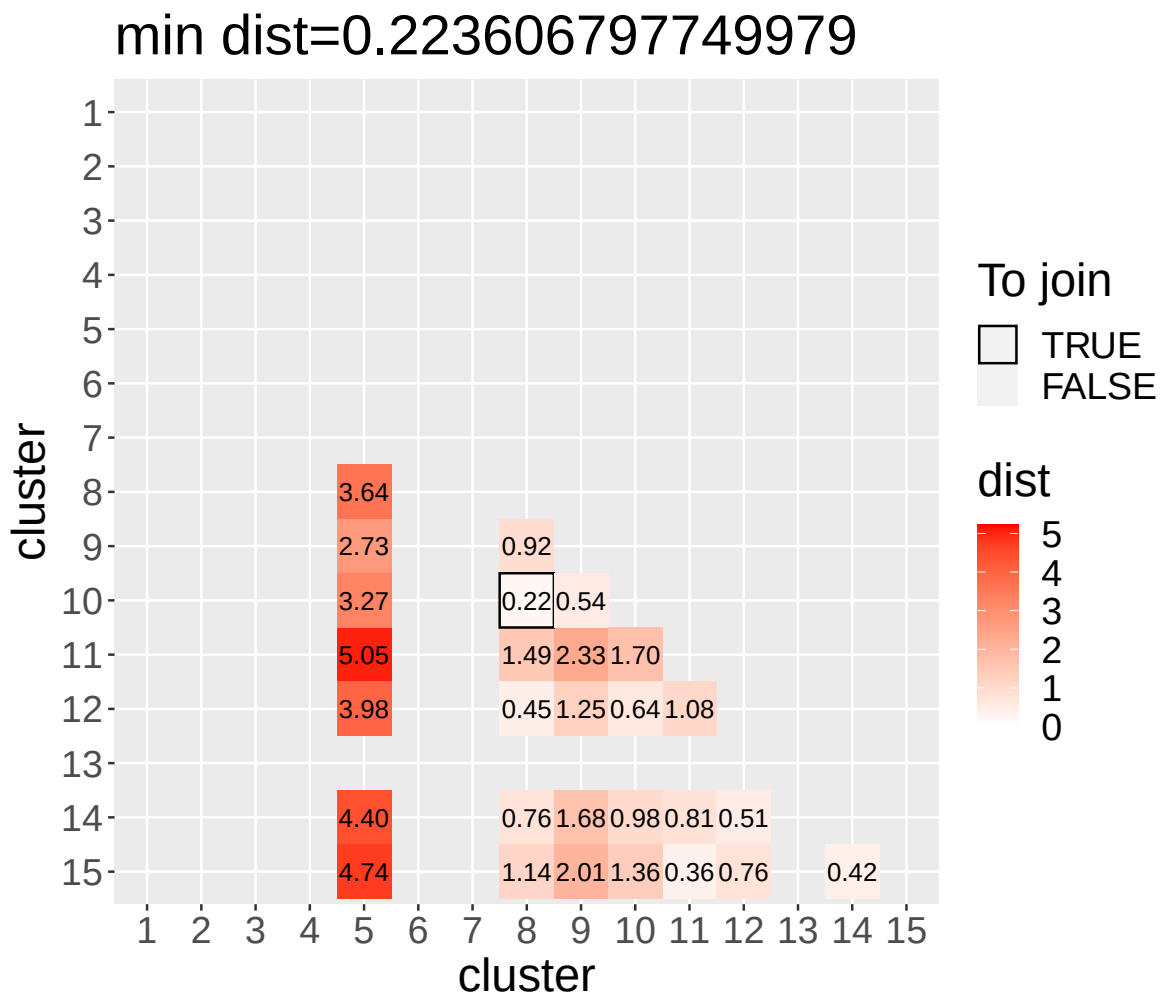
iteration 6


min dist=0.14142135623731
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
3.42
3.64 0.22
2.73 0.71 0.92
3.27 0.14 0.30 0.54
5.05 1.70 1.49 2.33 1.72
3.98 0.64 0.45 1.25 0.64 1.08
4.79 1.39 1.17 2.06 1.43 0.41 0.82
4.40 0.98 0.76 1.68 1.04 0.81 0.51 0.42
4.74 1.36 1.14 2.01 1.39 0.36 0.76 0.14 0.45
cluster
To join
TRUE
FALSE
dist
5
4
3
2
1
0

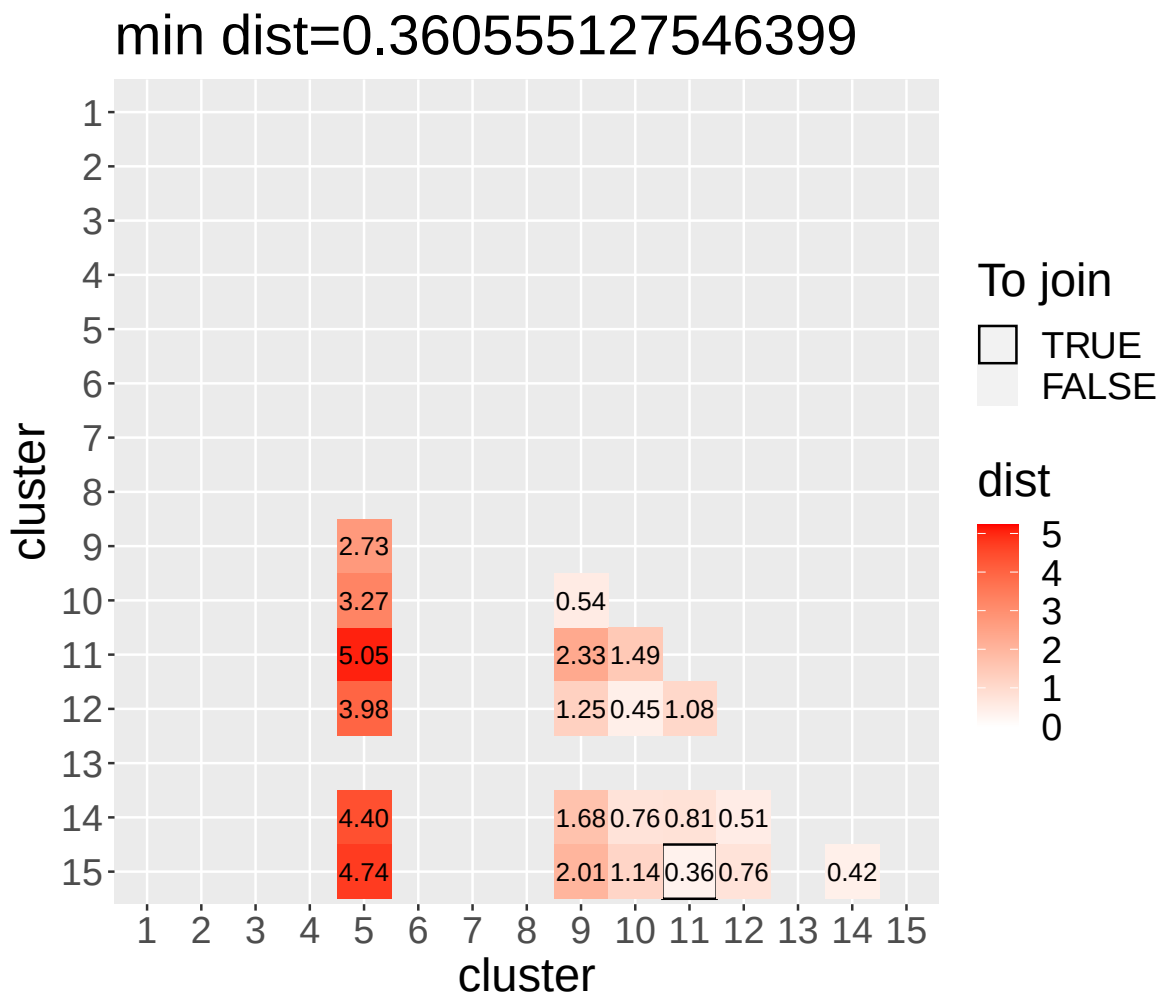
iteration 7


min dist=0.14142135623731
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
3.42
3.64 0.22
2.73 0.71 0.92
3.27 0.14 0.30 0.54
5.05 1.70 1.49 2.33 1.72
3.98 0.64 0.45 1.25 0.64 1.08
4.40 0.98 0.76 1.68 1.04 0.81 0.51
4.74 1.36 1.14 2.01 1.39 0.36 0.76 0.42
cluster
To join
TRUE
FALSE
dist
5
4
3
2
1
0

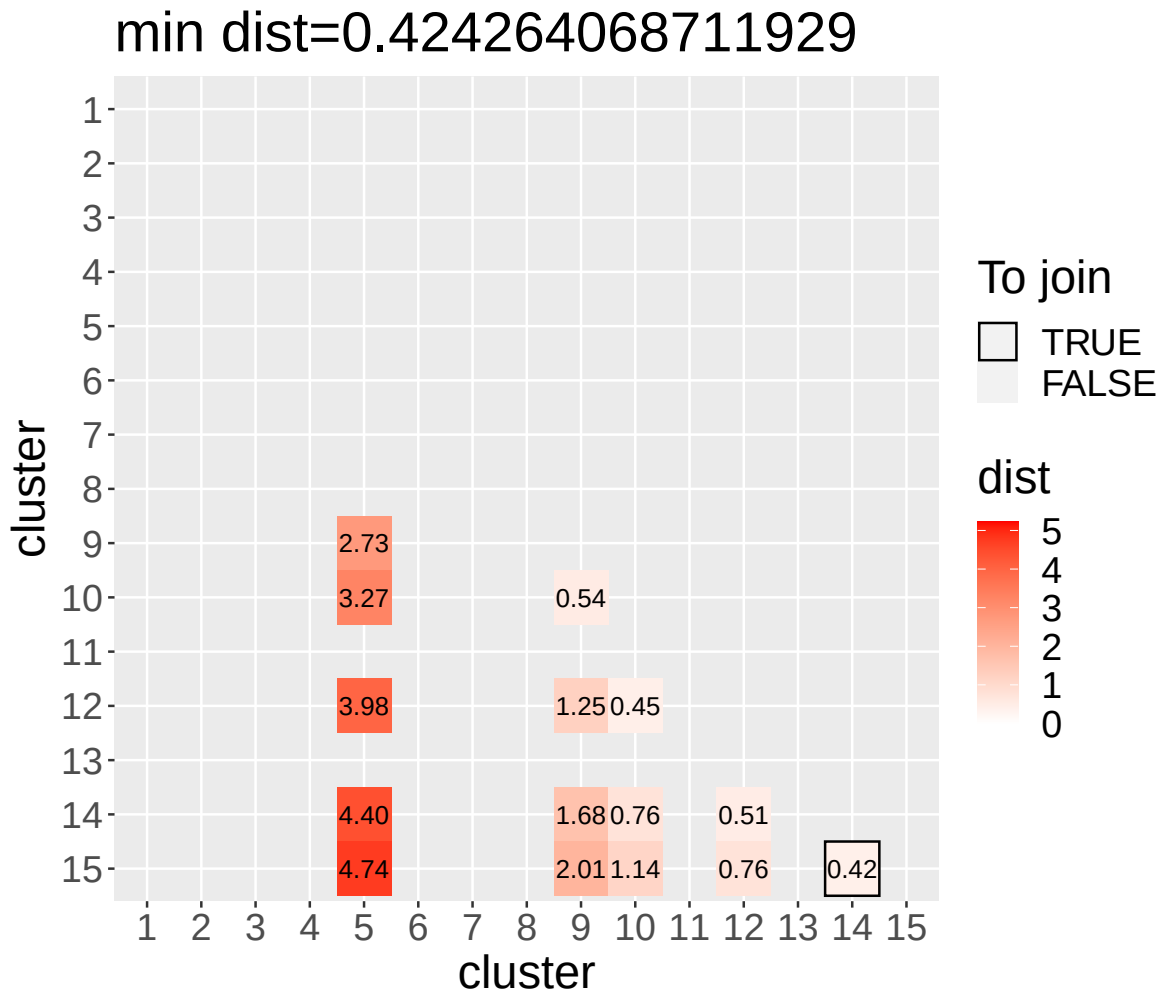
iteration 8


min dist=0.223606797749979
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
3.64
2.73
3.27
5.05
3.98
4.40
4.74
0.92
0.22 0.54
1.49 2.33 1.70
0.45 1.25 0.64 1.08
0.76 1.68 0.98 0.81 0.51
1.14 2.01 1.36 0.36 0.76 0.42
To join
TRUE
FALSE
dist
5
4
3
2
1
0

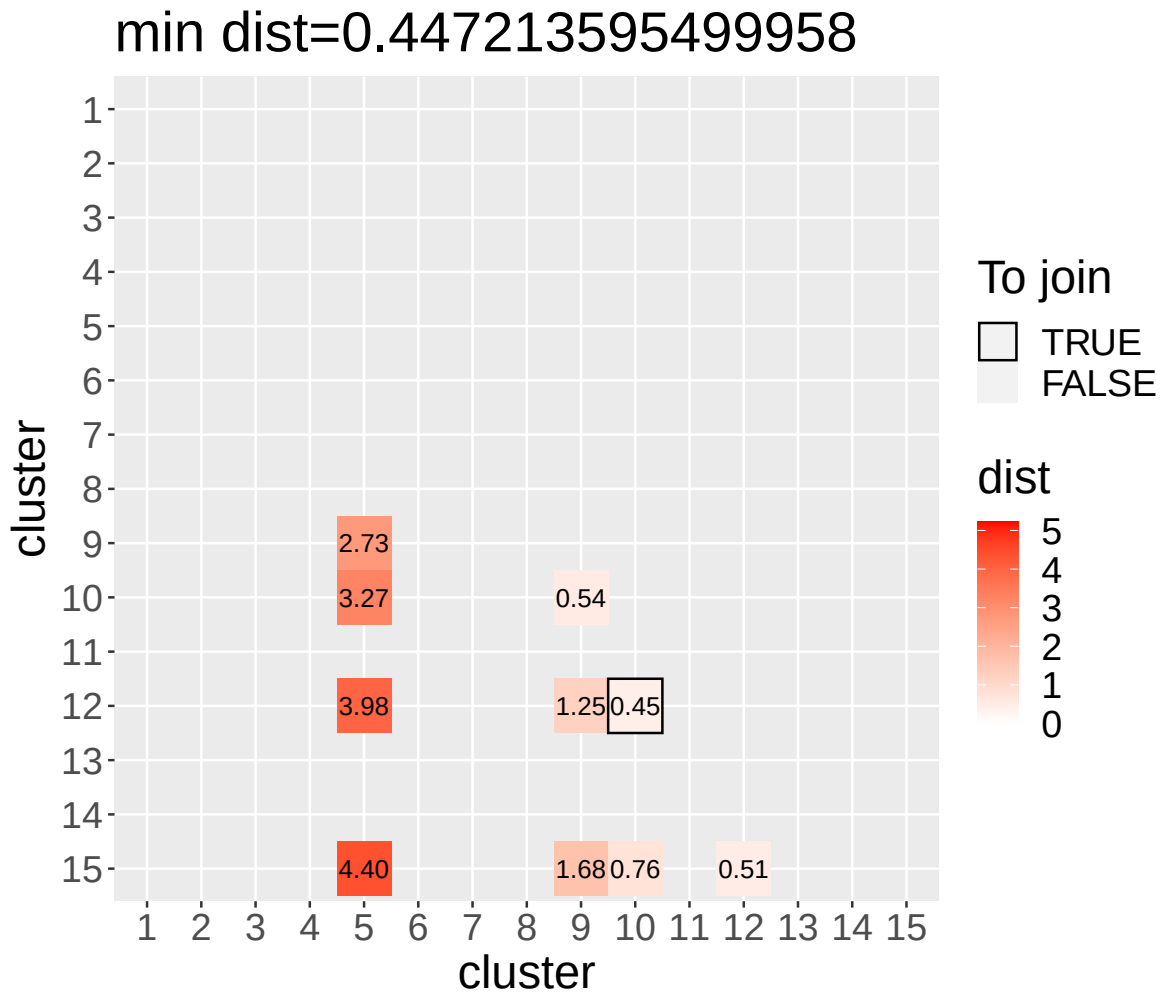
iteration 9


min dist=0.360555127546399
cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
2.73
3.27
5.05
3.98
4.40
4.74
0.54
2.33
1.49
1.25
0.45
1.08
1.68
0.76
0.81
0.51
2.01
1.14
0.36
0.76
0.42
To join
TRUE
FALSE
dist
5
4
3
2
1
0

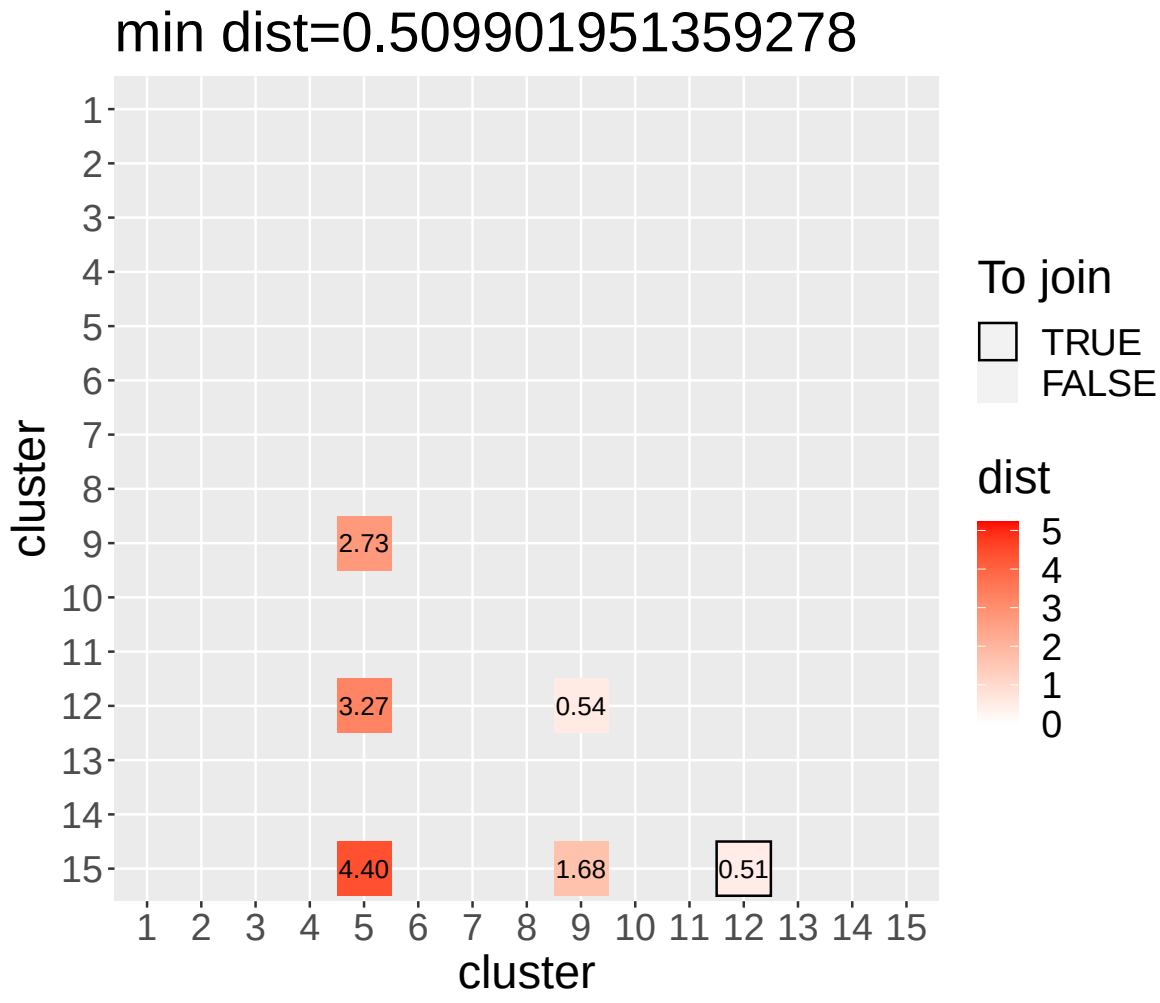
iteration 10



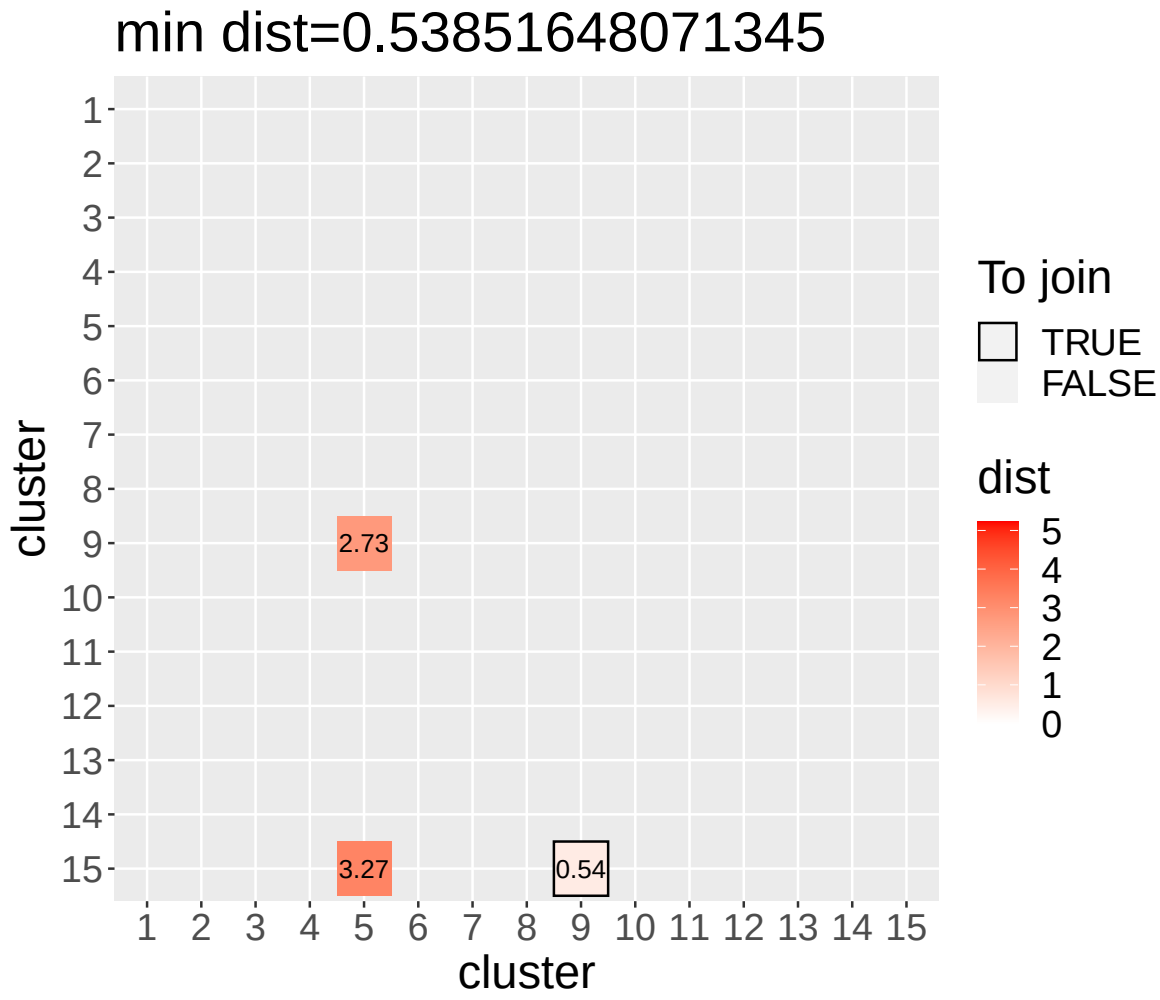
iteration 11



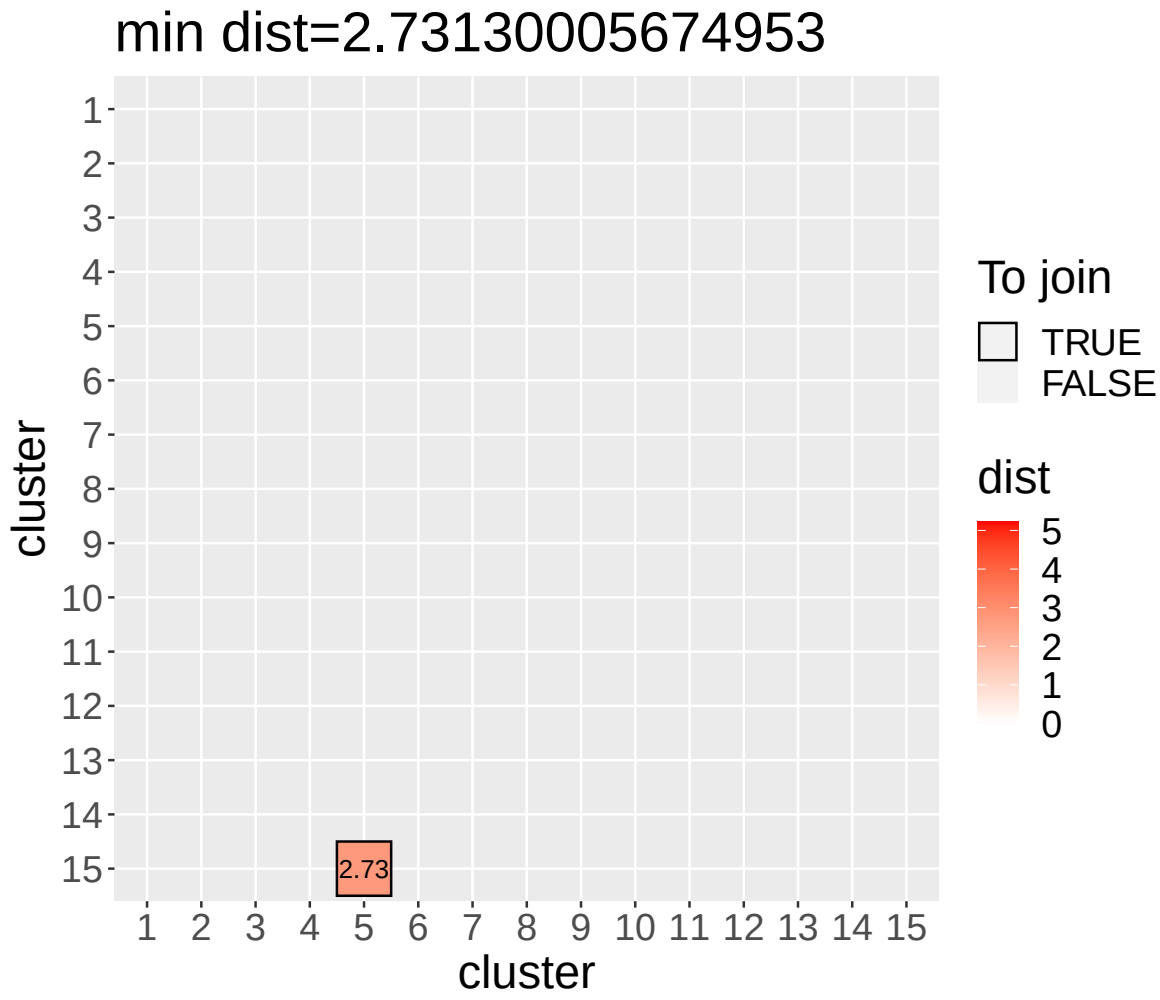
iteration 12



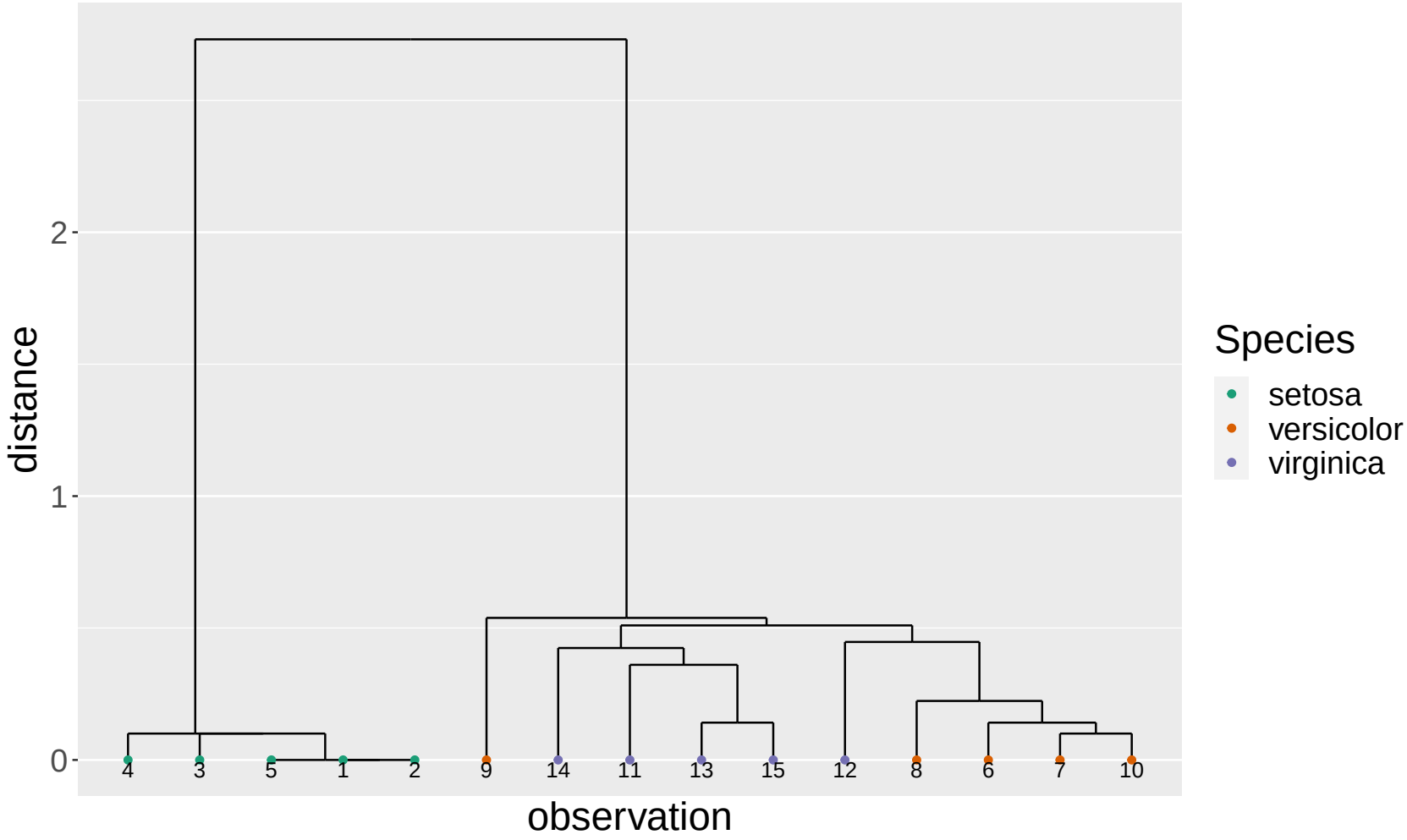
iteration 13



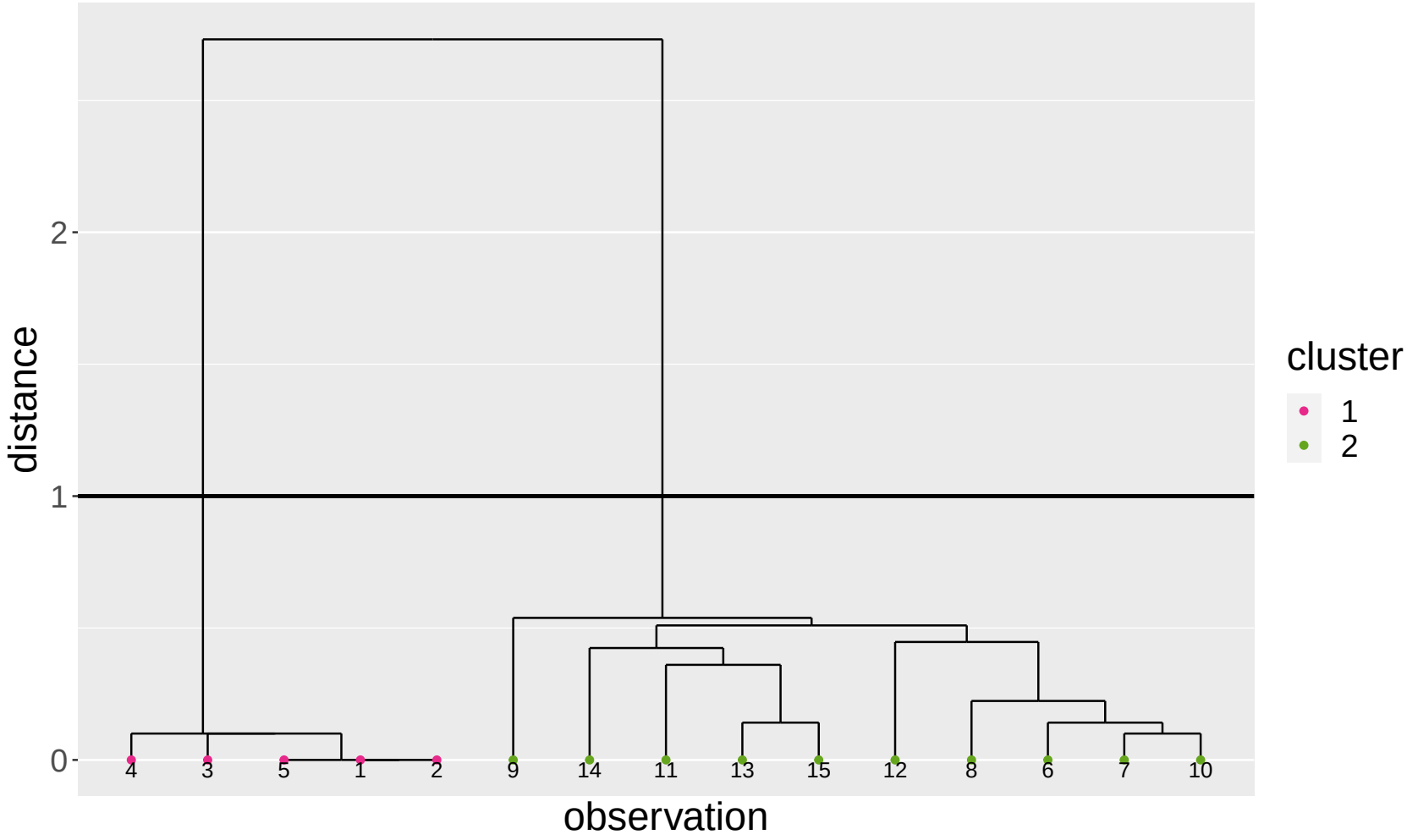
iteration 14



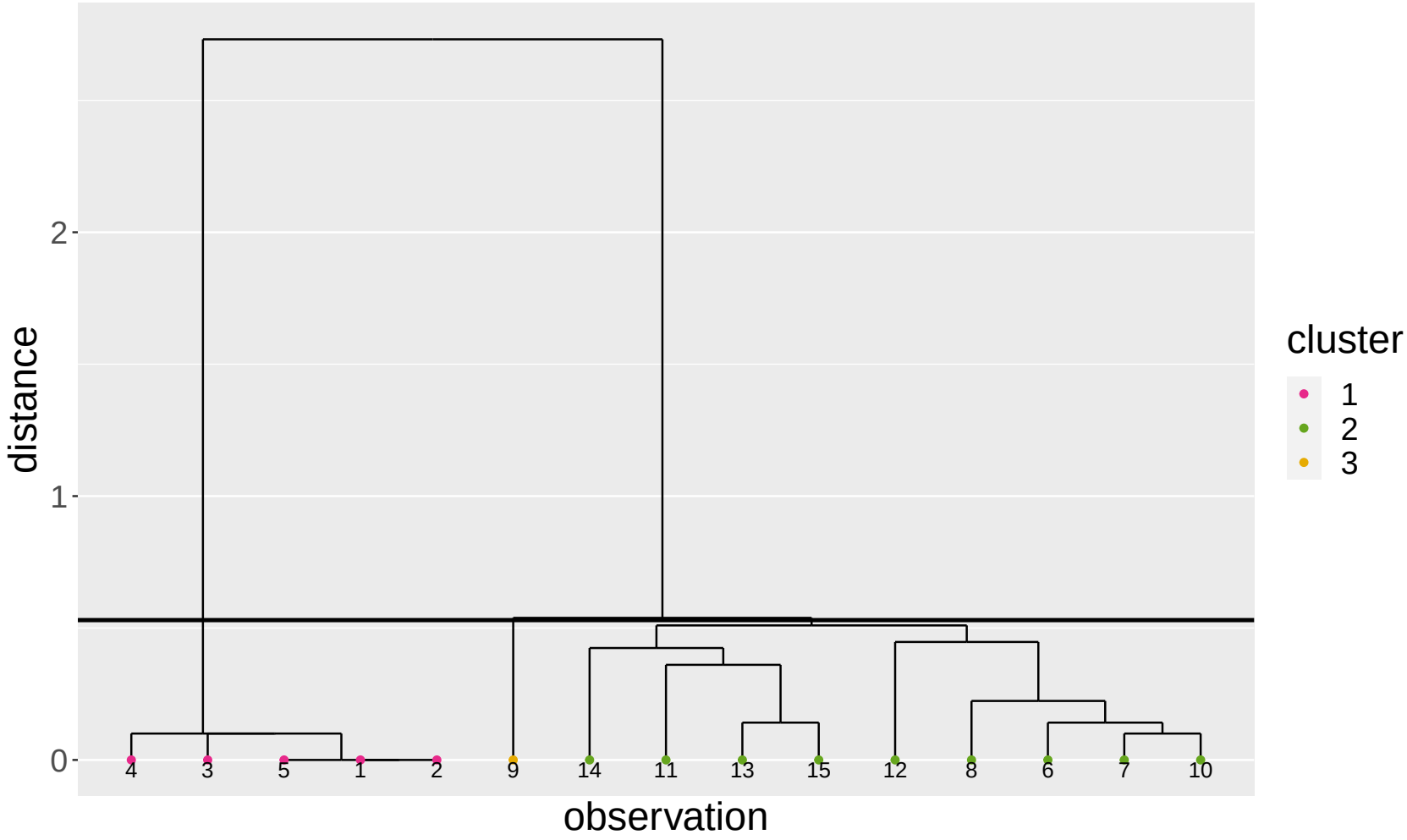
Visualization of dendrogram (tree diagram)



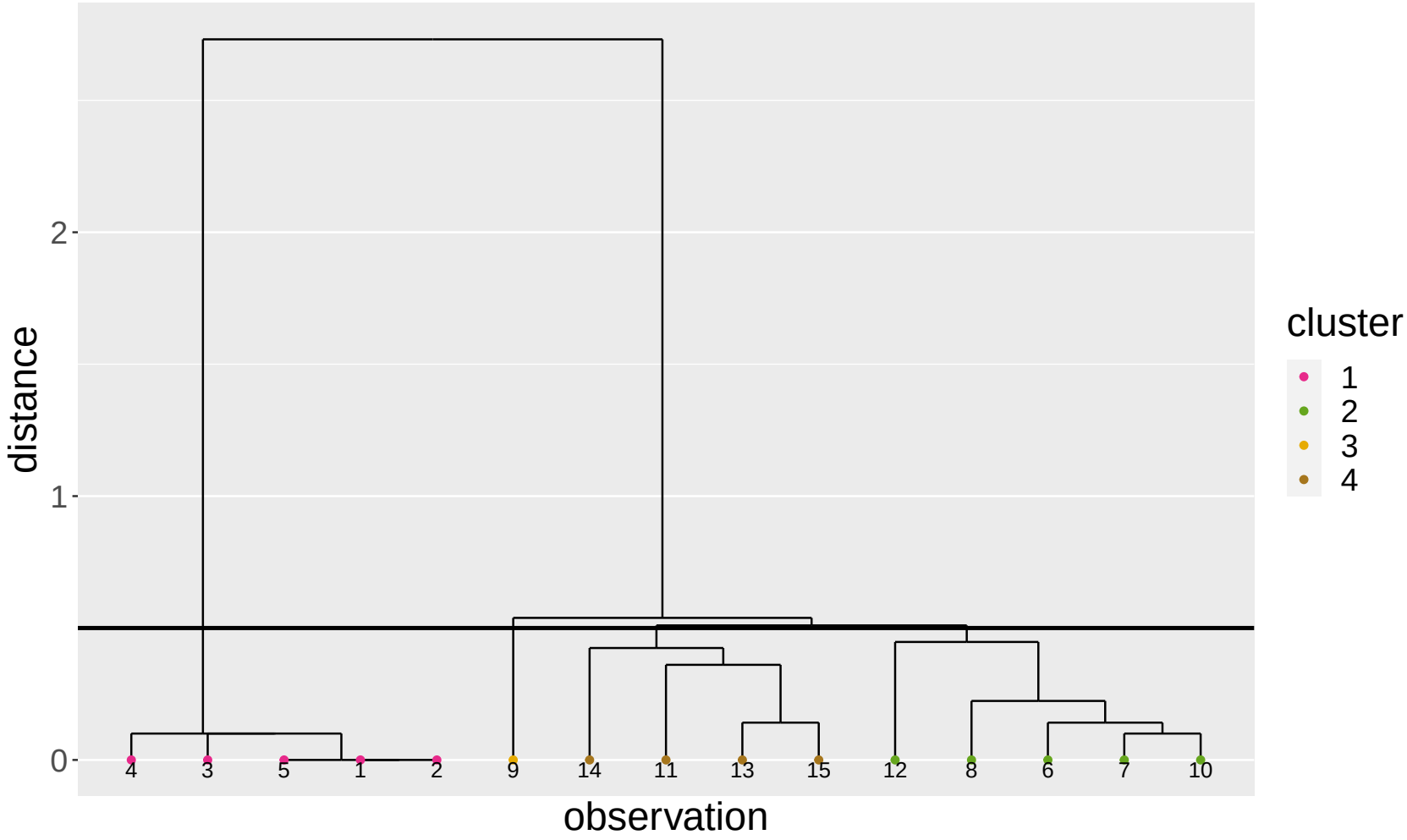
Cutting the tree to get two clusters



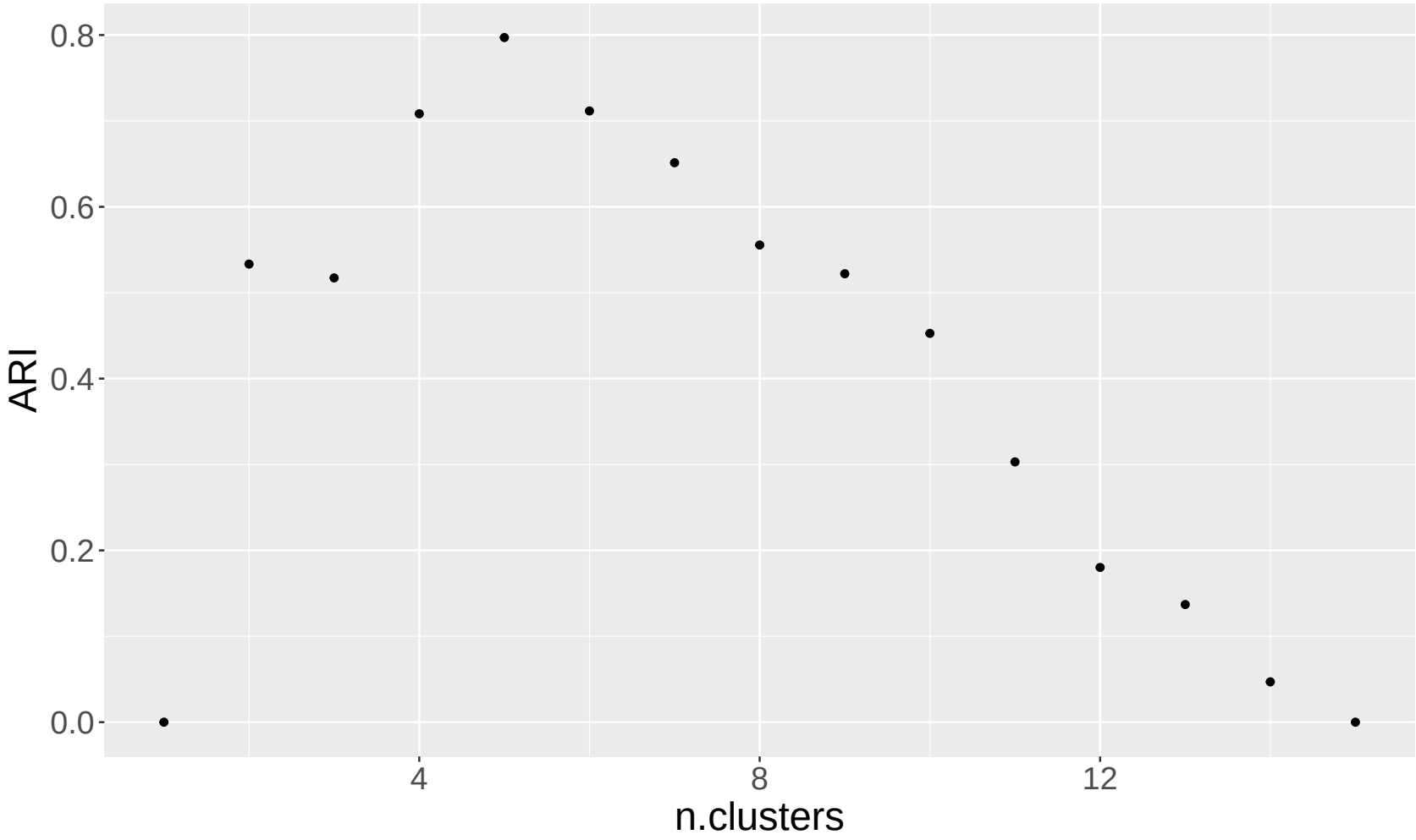
Cutting the tree to get three clusters



Cutting the tree to get four clusters



ARI computation



TODO full data set

TODO other choices, compare trees

Possible Exam Questions

TODO