Hierarchical clustering

Problem Statement

Identifying groups of similar data points is useful when analysing high-dimensional data, to understand how observations relate to each other

In hierarchical clustering, an algorithm groups similar data points into clusters. Unlike K-means clustering, hierarchical clustering does not require the number of clusters to be specified prior to analysis

The dendrogram is a key feature of hierarchical clustering. This tree allows the arrangement of clusters produced by analysis to be illustrated

Create a dendrogram

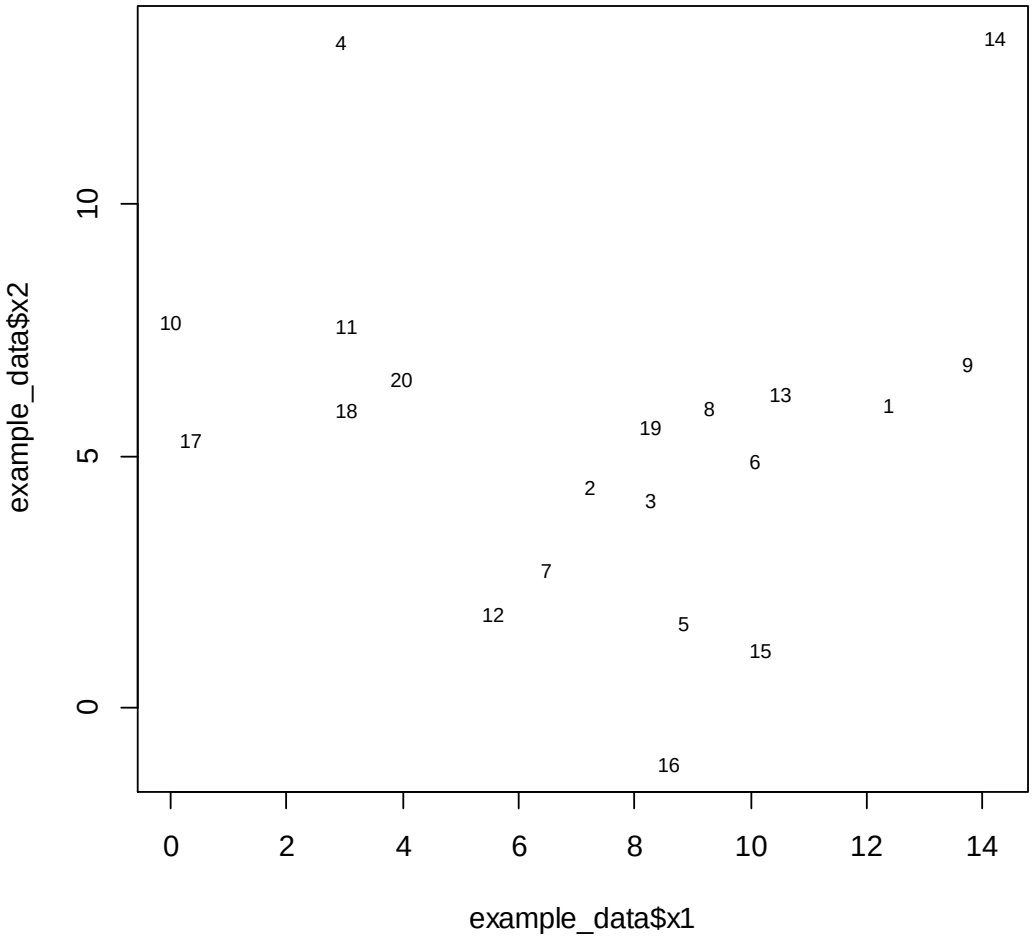
We can plot dendrograms in R using the ‘hclust’ function which takes a distance matrix as input and creates the associated tree using hierarchical clustering

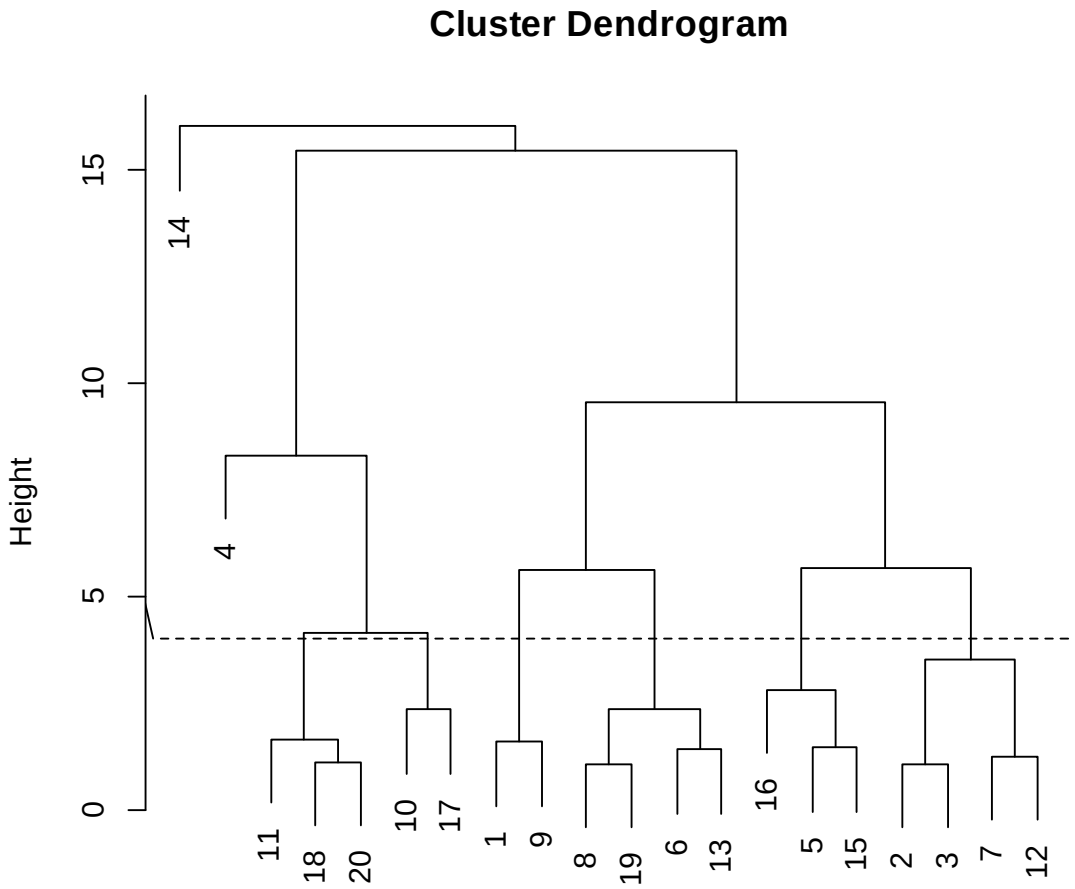
20 data points are generated for two variables (x1 and x2). Hierarchical clustering carried out on the data can be used to produce a dendrogram. But how do we interpret this dendrogram?

Challenge 1

Use the ‘clust’ function to implement hierarchical clustering using the distance matrix ‘dist_m’ and the ‘complete’ method and plot the results as a dendrogram using ‘plot’

Challenge 1

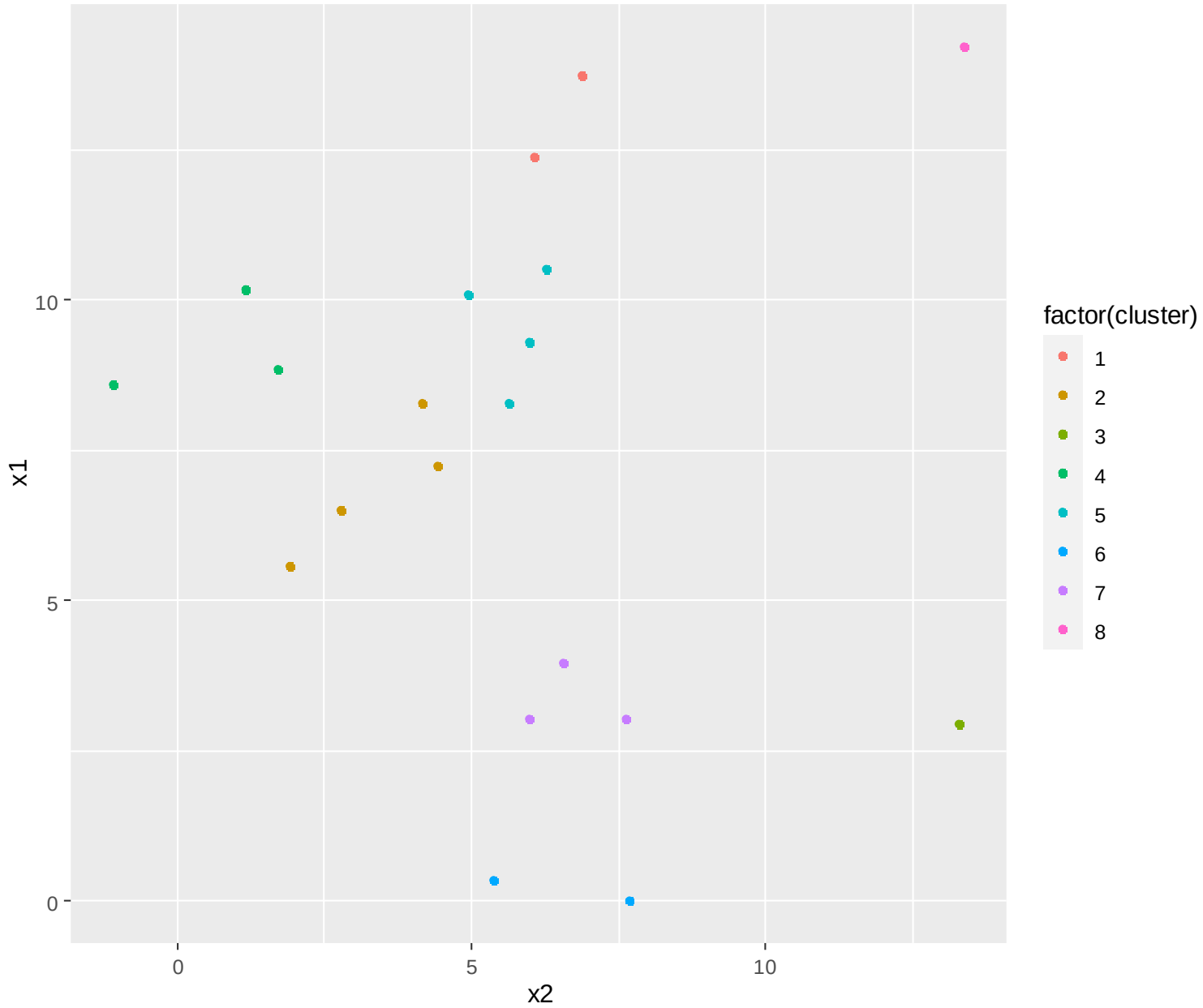

example_data$x2
example_data$x1
0
5
10
0
2
4
6
8
10
12
14
4
14
10
11
20
18
13
8
1
19
9
17
6
2
3
7
12
5
15
16


Cluster Dendrogram
Height
0
5
10
15
14
4
11
18
20
10
17
1
9
8
19
6
13
16
5
15
2
3
7
12

Number of clusters

We can cut the dendrogram to determine number of clusters at different heights using the function ‘cutree’. This function can be used to identify how many clusters occur at different parts of the tree

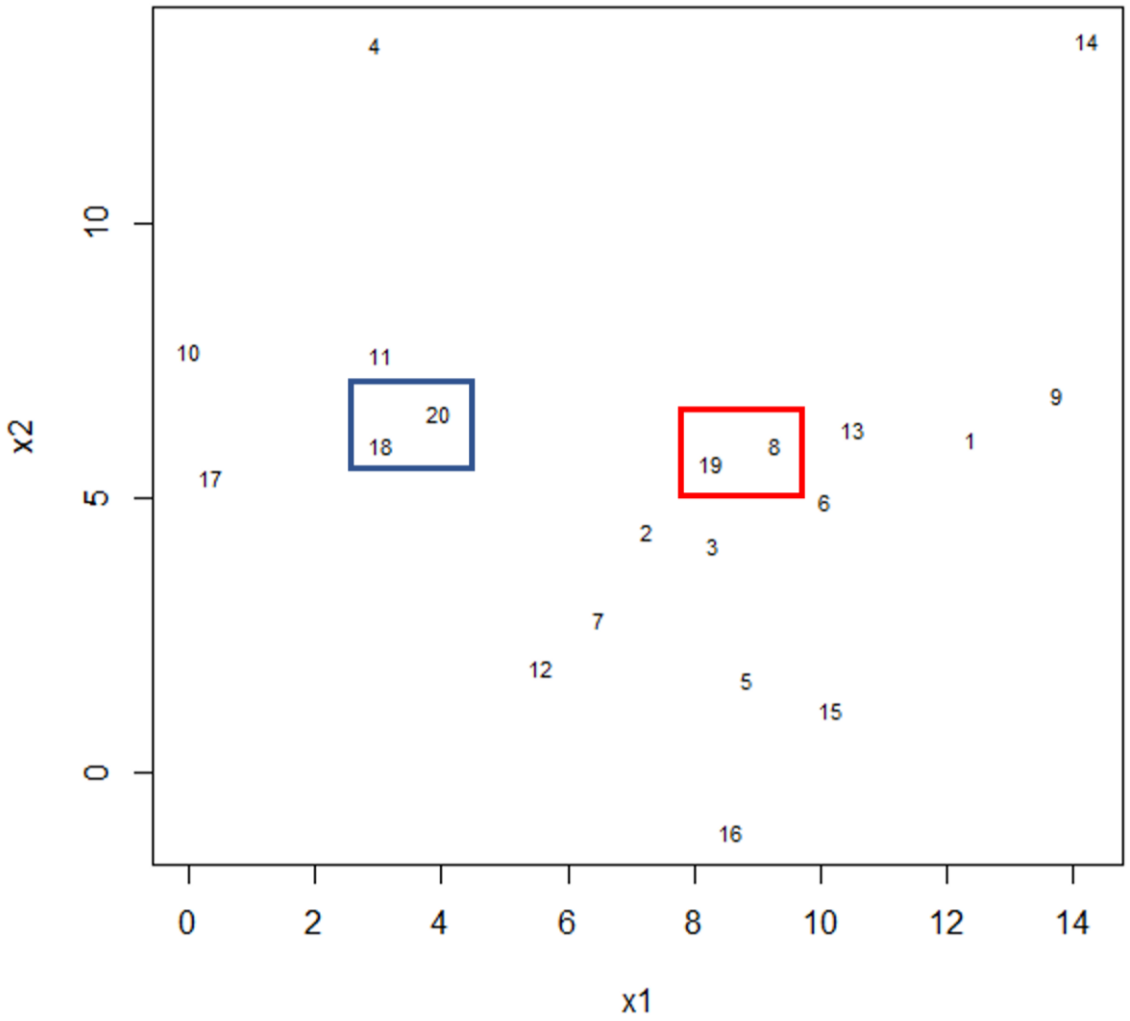
Cutting the tree at h = 4 produces 8 clusters

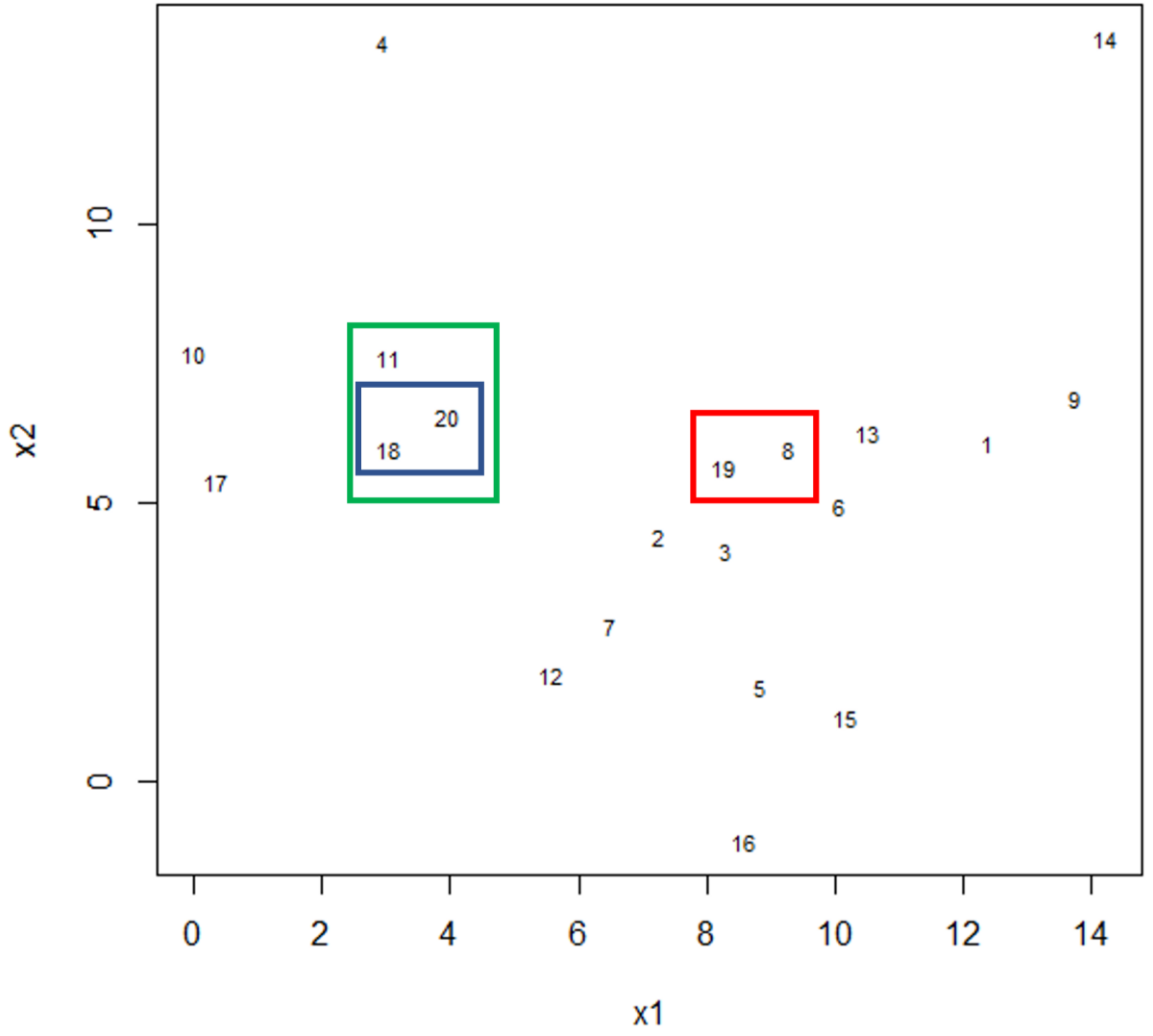

x1
10
5
0
0
5
10
x2
factor(cluster)
1
2
3
4
5
6
7
8

Challenge 2

Identify the value of k in the ‘cutree’ function that gives the same output as h = 5

Hierarchical clustering algorithm





Distance matrix

Hierarchical clustering is performed in two steps: calculating the distance (or dissimilarity) matrix and applying clustering using this matrix

Distance matrix can be calculated by:

- specifying distance matrix as a pre-defined option using the ‘method’ argument in the ‘dist()’ function
- create a self-defined function which calculates distance from a matrix or from two vectors

The type of distance matrix used in hierarchical clustering can have a big effect on the resulting tree. The decision of which distance matrix to use before carrying out hierarchical clustering depends on type of data and question to be addressed

Linkage method

The second step in performing hierarchical clustering is determining how to fuse different clusters

Linkage is used to define dissimilarity between groups of observations (or clusters) and is used to create the hierarchical structure in the dendrogram. Different linkage methods of creating a dendrogram are available

Complete linkage works by computing all pairwise dissimilarities between data points in different clusters, using the largest pairwise dissimilarity to decide which cluster will be fused. Clusters with smallest value of are fused

Challenge 3

Carry out hierarchical clustering on the small version of the methylation dataset using different linkage methods and compare resulting dendrograms.

Do any of the methods produce similar dendrograms? Do some methods appear to produce more realistic dendrograms than others? Discuss in groups

Validating clusters

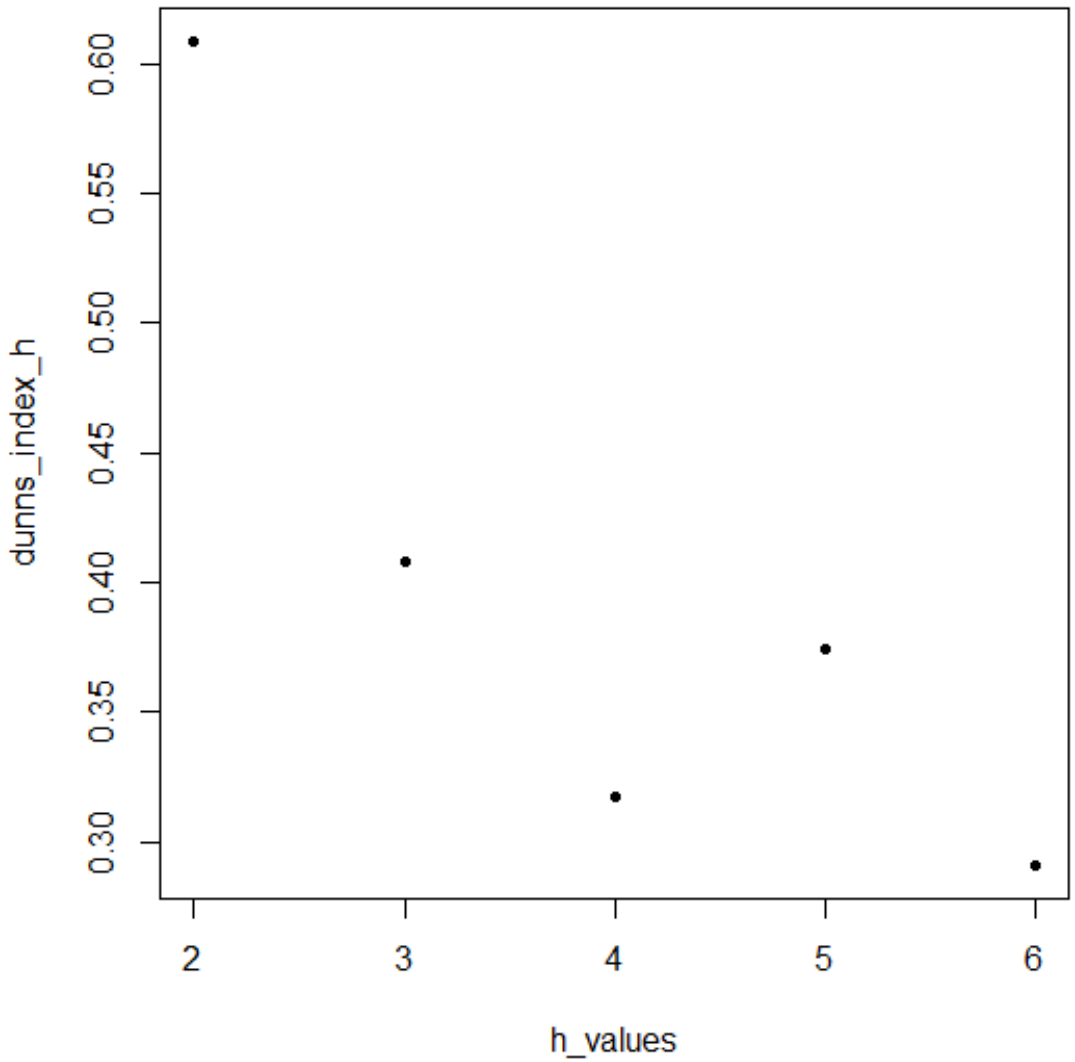
Now that we know how to carry out hierarchical clustering, how do we know how many clusters are optimal for the dataset?

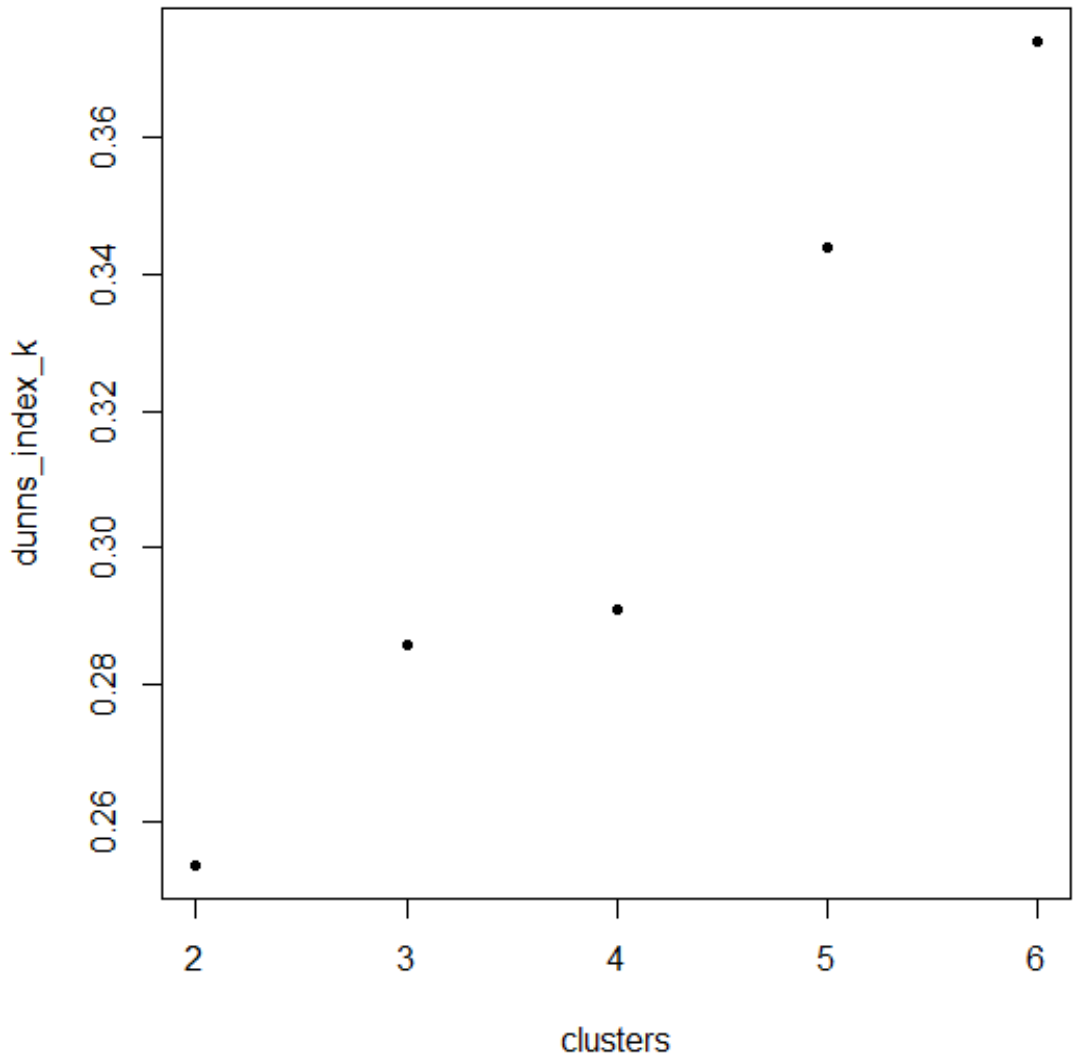
We need to be able to determine whether identified clusters represent true groups in the data, or whether clusters have been identified just due to chance. There are some statistical tests that can determine the optimal number of clusters in the data by assessing whether there is more evidence for a cluster than we would expect due to chance

The Dunn index is a ratio of the smallest distance between observations not located within the same cluster to the largest intra-cluster distance found within any cluster. The higher the Dunn index, the better defined the clusters

Challenge 4

Examine how changing the h or k arguments in the ‘hclust’ function affects the value of the Dunn index


dunns_index_h
0.30
0.35
0.40
0.45
0.50
0.55
0.60
2
3
4
5
6
h_values


dunns_index_k
0.26
0.28
0.30
0.32
0.34
0.36
2
3
4
5
6
clusters

Further reading

Dunn, J. C. (1974) Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4(1):95–104

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001) On clustering validation techniques. Journal of Intelligent Information Systems 17(2/3):107-145

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Section 10.3.2 (Hierarchical Clustering)

Understanding the concept of Hierarchical clustering Technique. towards data science blog https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec