Hierarchical Clustering

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Clustering

Clustering refers to a very broad set of techniques for finding groups, or clusters, in a data set.

Clustering (from Tuffery 2011)

Clustering is the statistical operation of grouping objects (individuals or variables) into a limited number of groups known as clusters.

On the one hand, groups are not defined in advanced by the analyst, but are discovered during the analysis, unlike the classes used in classification.

On the other hand, the clusters are combinations of objects having similar characteristics, which are separated from objects having different characteristics (in other clusters).

Clustering Idea

Group a set of n objects in K groups such that:

- each group is as much homogeneous as possible
  i.e. *within-groups homogeneity*.
- groups are as distinct as possible among them
  i.e. *between-groups heterogeneity*

Clustering Idea

- We seek a partition of the data into distinct groups.
- We want the observations within each group to be quite similar to each other.
- We must define what it means for two or more observations to be similar or different.
- This is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Assumptions

We will assume that the rows of the data matrix correspond to the individuals to be clustered (although you could also cluster variables).

We will assume that the individuals are embeded in an euclidean space (e.g. quantitative variables, or output of dimension reduction method)

Hierarchical Clustering

Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) produce sequences of nested partitions of increasing heterogeneity.

- AHC can be used if there is a concept of distance.
- We must define the distance of two objects.
- We also need to define the distance of two clusters.

Hierarchical Clustering

The general form of the algorithm is as follows:

1. The initial clusters are the observations.
2. The distances between clusters are calculated.
3. The two clusters which are closet together are merged and replaced with a single cluster.
4. We start again at step 2 until there is only one cluster, which contains all the observations.

Visual display with Dendrograms

The sequence of partitions is presented in what is known as a tree diagram also known as **dendrogram**.

This tree can be cut at a greater or lesser height to obtain a smaller or larger number of clusters.

The number of clusters can be chosen by optimizting certain statistical quality criteria. The main criterion is the loss of between-cluster sum of squares.

Common Distances

The distance between two individual observations tends to be obvious:

- Euclidean
- Manhattan

As soon as a cluster has more than one element, the distance between two clusters is less obvious. It can be defined in many ways but the most usual ones are:

Common Distances between Clusters

Minimum Distance

- Also known as single linkage or nearest neighbor.
- Sensitive to the "chain effect" (or chaining): if two widely separated clusters are linked by a chain of individualsm they me grouped together.

Maximum distance

- Also known as complete linkage or farthest-neighbor technique.
- Tends to generate clusters of equal diameter.
- Sensitive to outliers.

Common Distances between Clusters

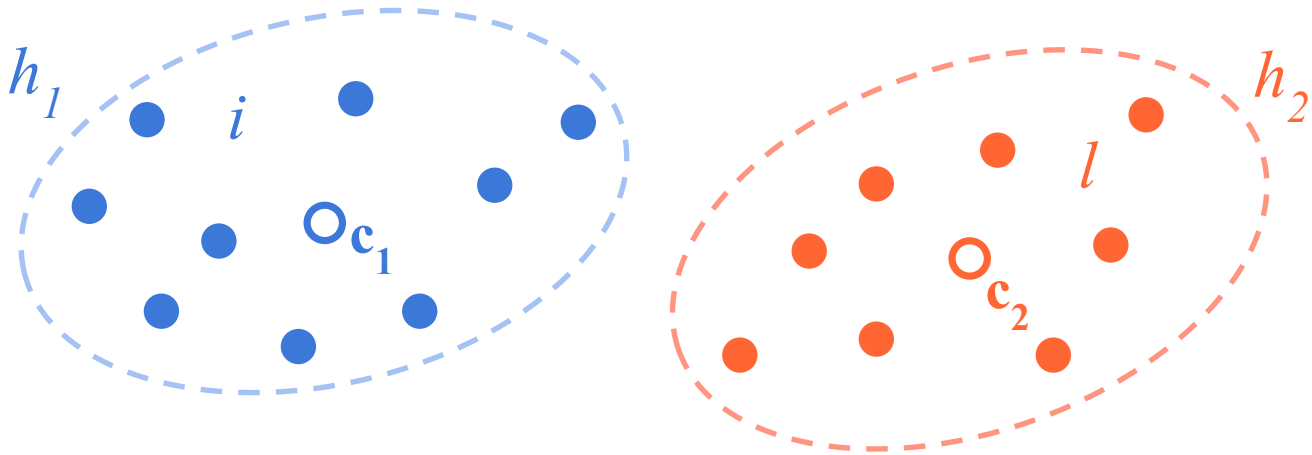
Mean Distance

- Also known as average linkage.
- Intermediate between the minimum distance and the maximum distance methods.
- Tends to produce clusters having similar variance.

Ward method

- Matches the purpose of clustering most closely.
- Ward distance defined as the reduction in between-cluster sum of squares.

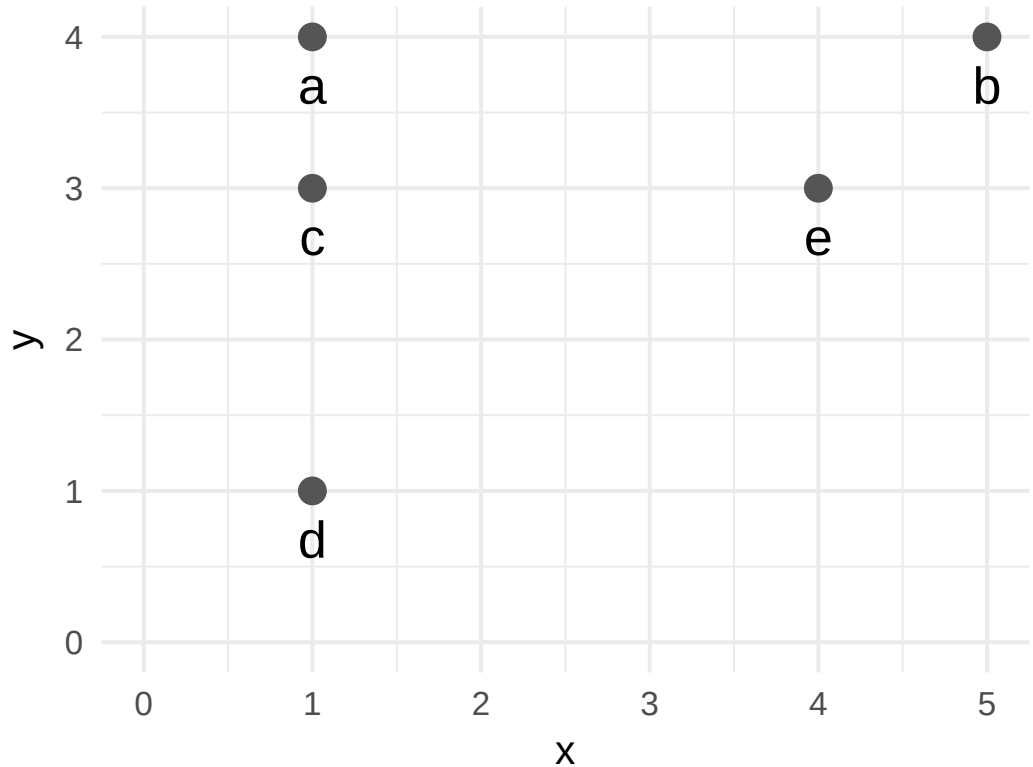
Aggregation Criteria



Single linkage	$d(h_1, h_2) = min\ \{\ d(i, l)\ \}$
Average linkage	$d(h_1, h_2) = avg\ \{\ d(i, l)\ \}$
Complete linkage	$d(h_1, h_2) = max\ \{\ d(i, l)\ \}$
Centroid	$d(h_1, h_2) = d(\boldsymbol{c}_1, \boldsymbol{c}_2)$
Ward	$d(h_1, h_2) = Inertia(\boldsymbol{c}_1, \boldsymbol{c}_2)$

Toy Data Set

	x	y
a	1	4
b	5	4
c	1	3
d	1	1
e	4	3



Clustering

Distance matrix (squared euclidean distances)

```
   a  b  c  d  e
a  0 16  1  9 10
b 16  0 17 25  2
c  1 17  0  4  9
d  9 25  4  0 13
e 10  2  9 13  0
```

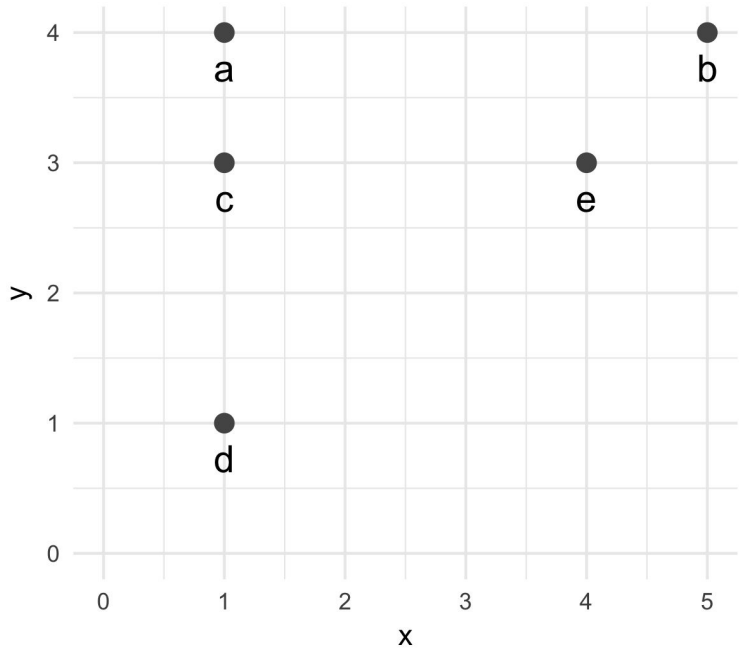
Clustering

Distance values in increasing order

Distance	Pairs Formed
0.400	a-b
0.500	d-e
0.700	c-d
0.786	b-e
0.800	c-e
0.929	b-c
0.937	b-d
1.000	a-c
1.000	a-d
1.000	a-e

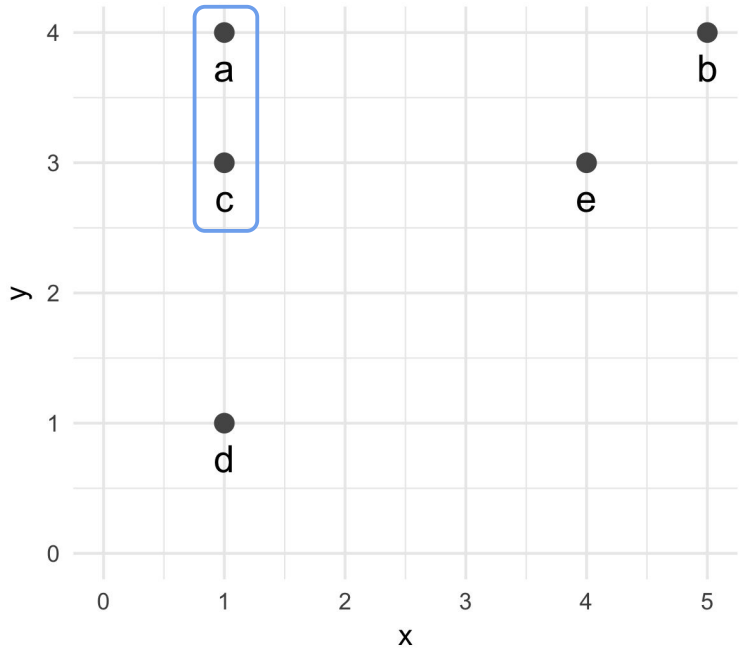
Single linkage

	a	b	c	d	e
a	0				
b	16	0			
c	1	17	0		
d	9	25	4	0	
e	10	2	9	13	0



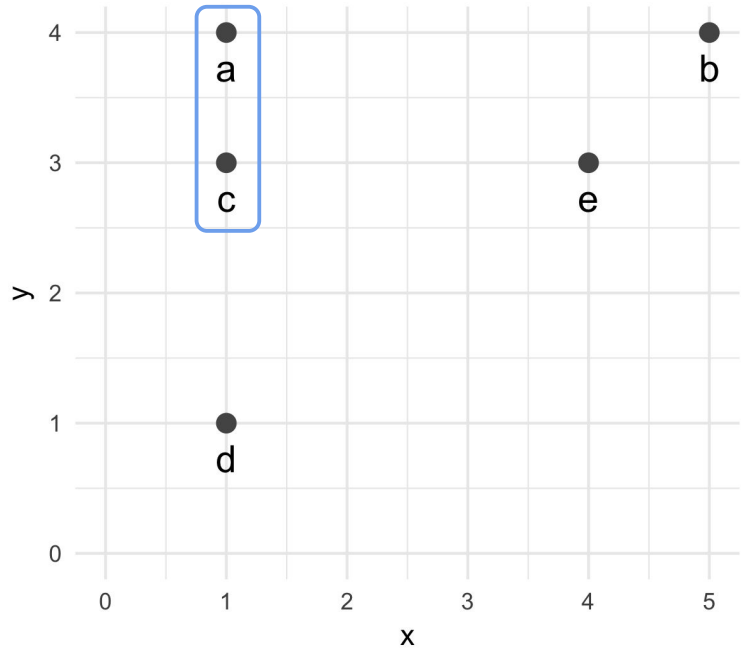
Single linkage

	a	b	c	d	e
a	0				
b	16	0			
c	**1**	17	0		
d	9	25	4	0	
e	10	2	9	13	0



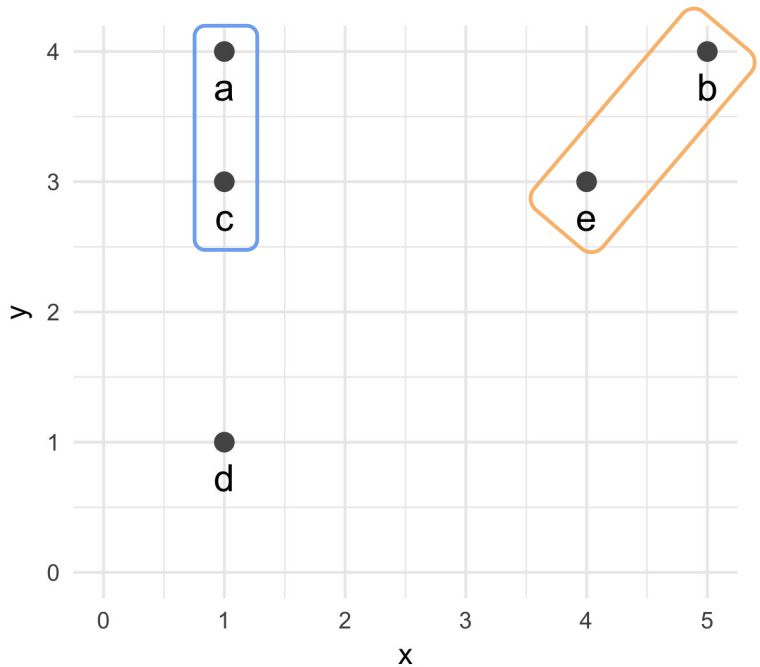
Single linkage

	a,c	b	d	e
a,c	0			
b	16	0		
d	4	25	0	
e	9	2	13	0



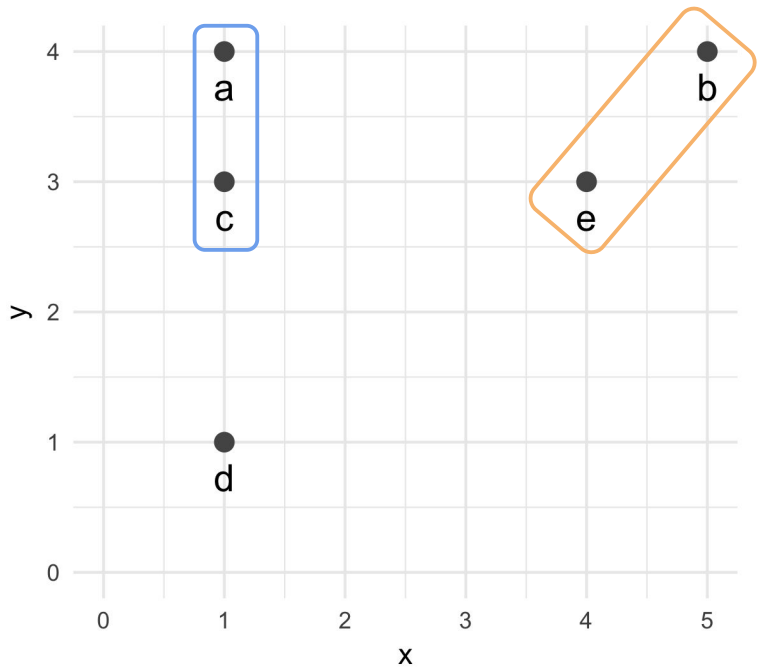
Single linkage

	a,c	b	d	e
a,c	0			
b	16	0		
d	4	25	0	
e	9	**2**	13	0



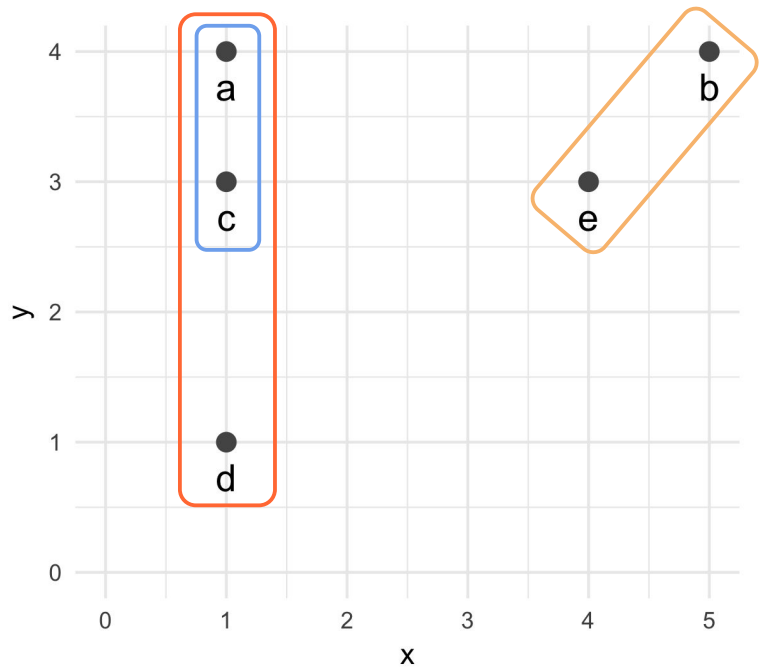
Single linkage

	a,c	b,e	d
a,c	0		
b,e	9	0	
d	4	13	0



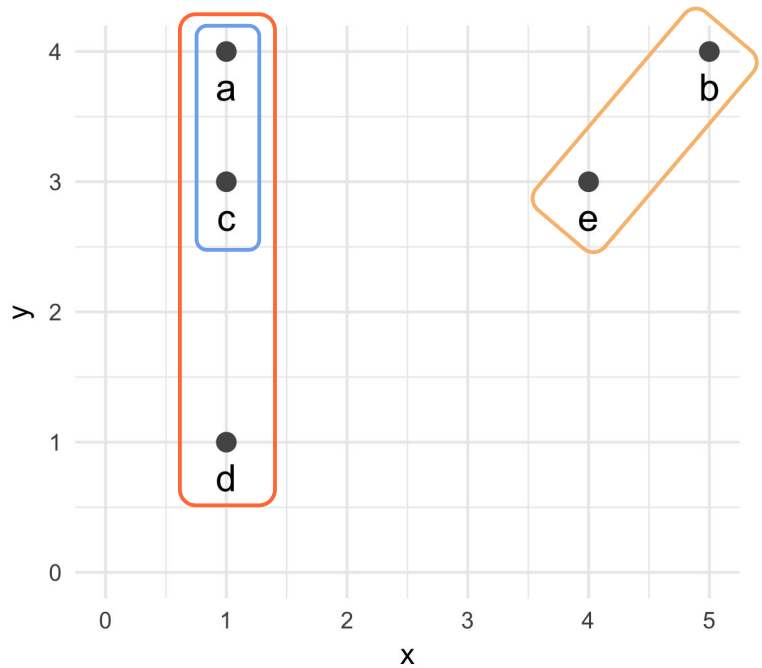
Single linkage

	a,c	b,e	d
a,c	0		
b,e	9	0	
d	**4**	13	0



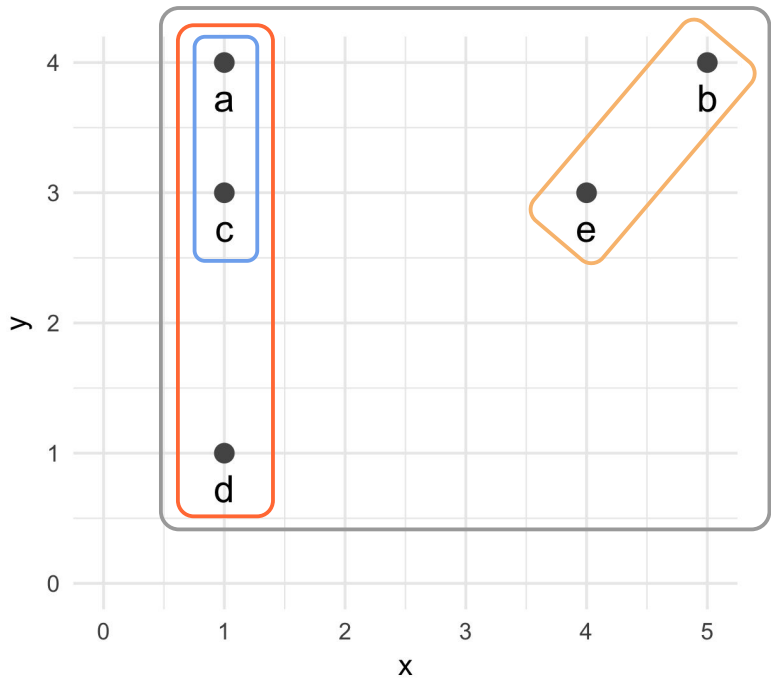
Single linkage

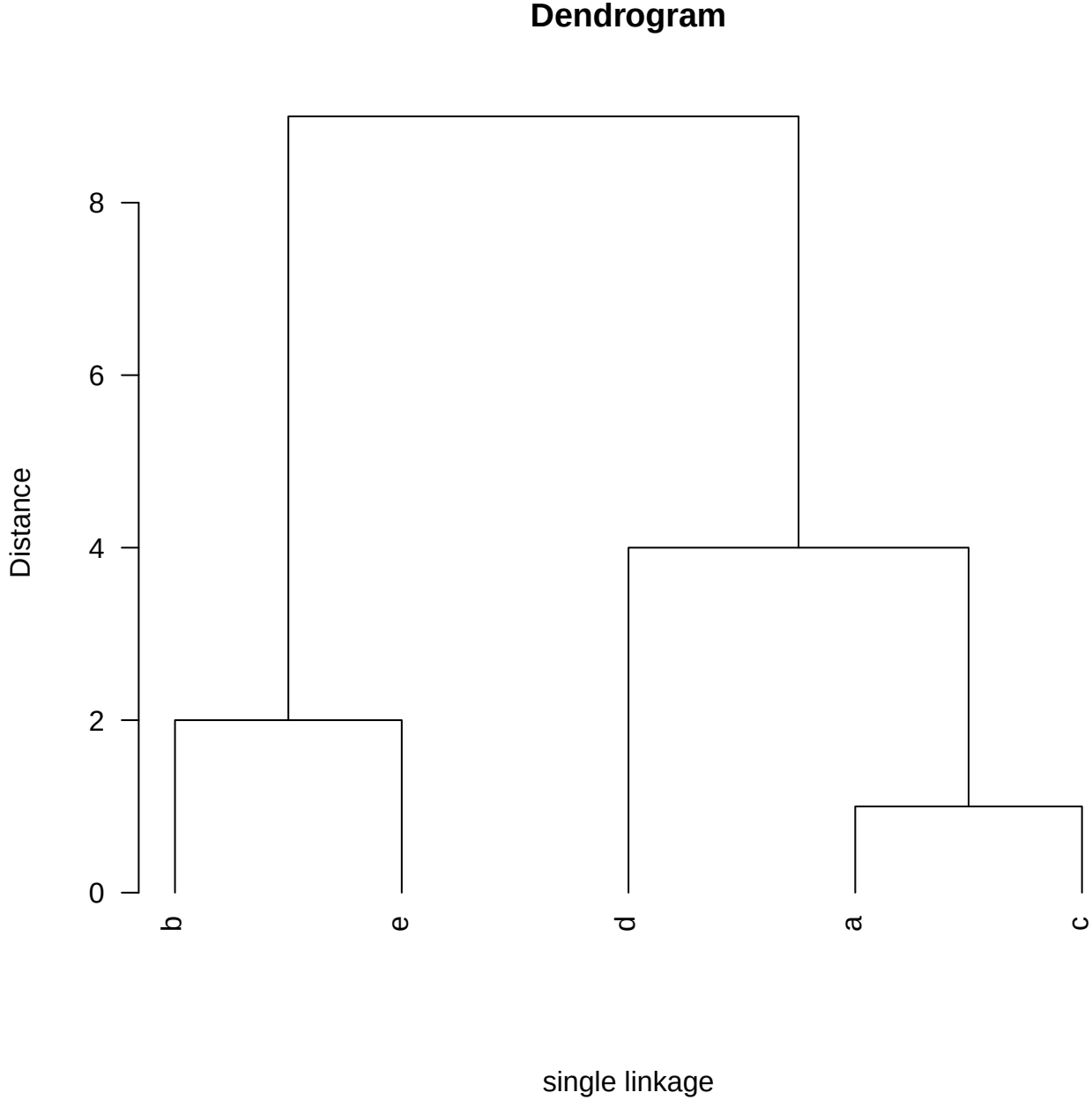
	a,c,d	b,e
a,c,d	0	
b,e	9	0

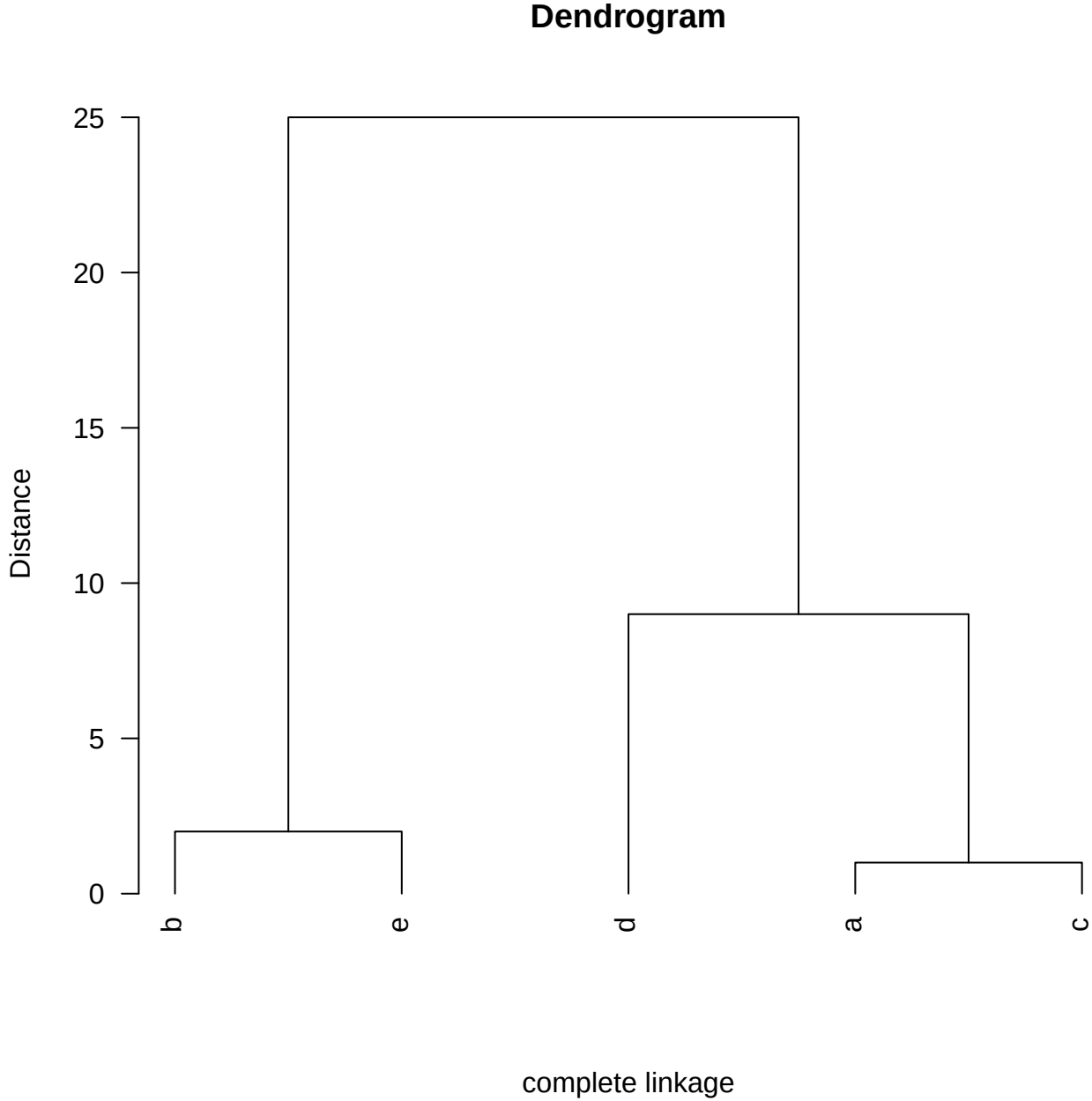


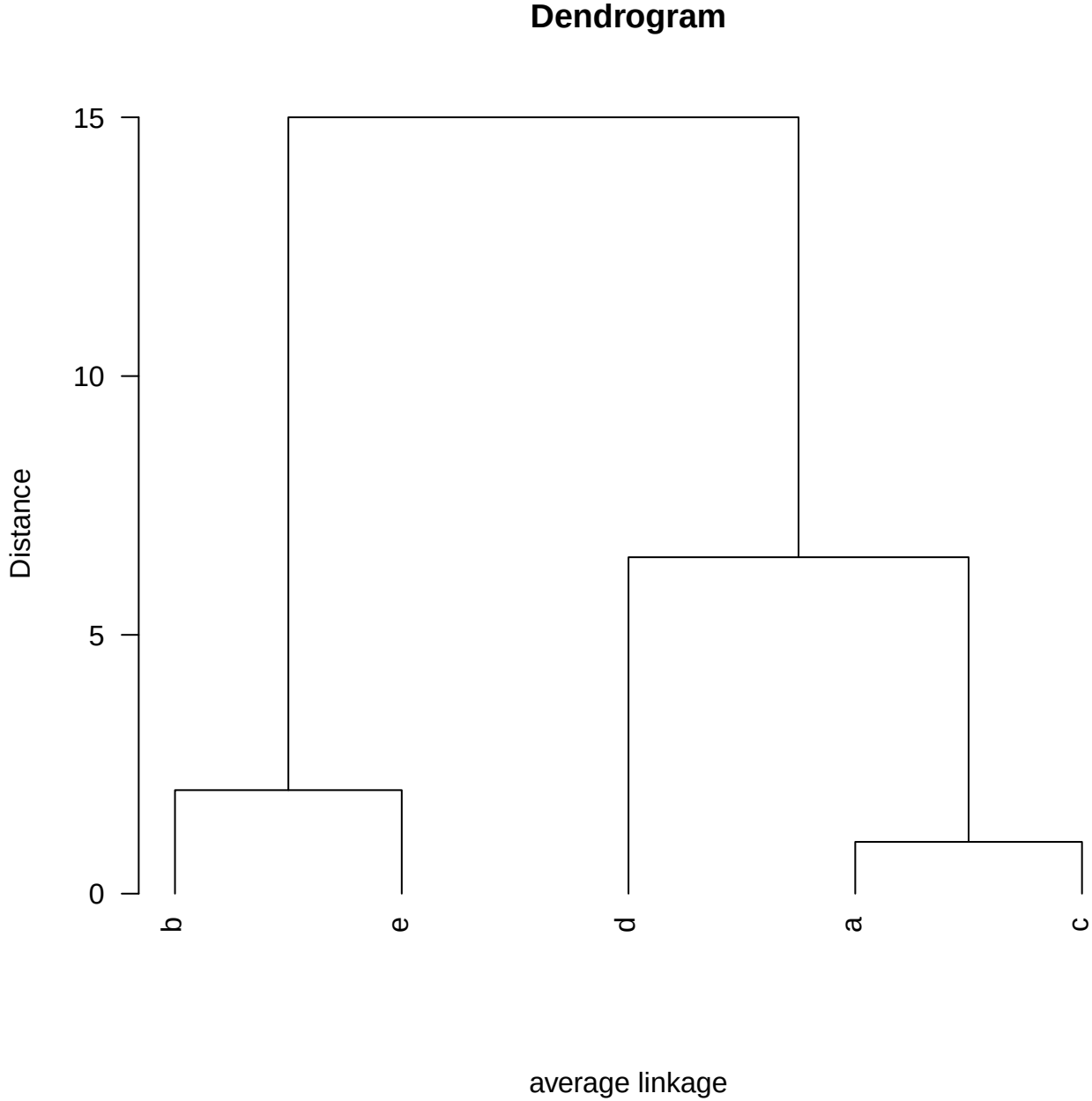
Single linkage

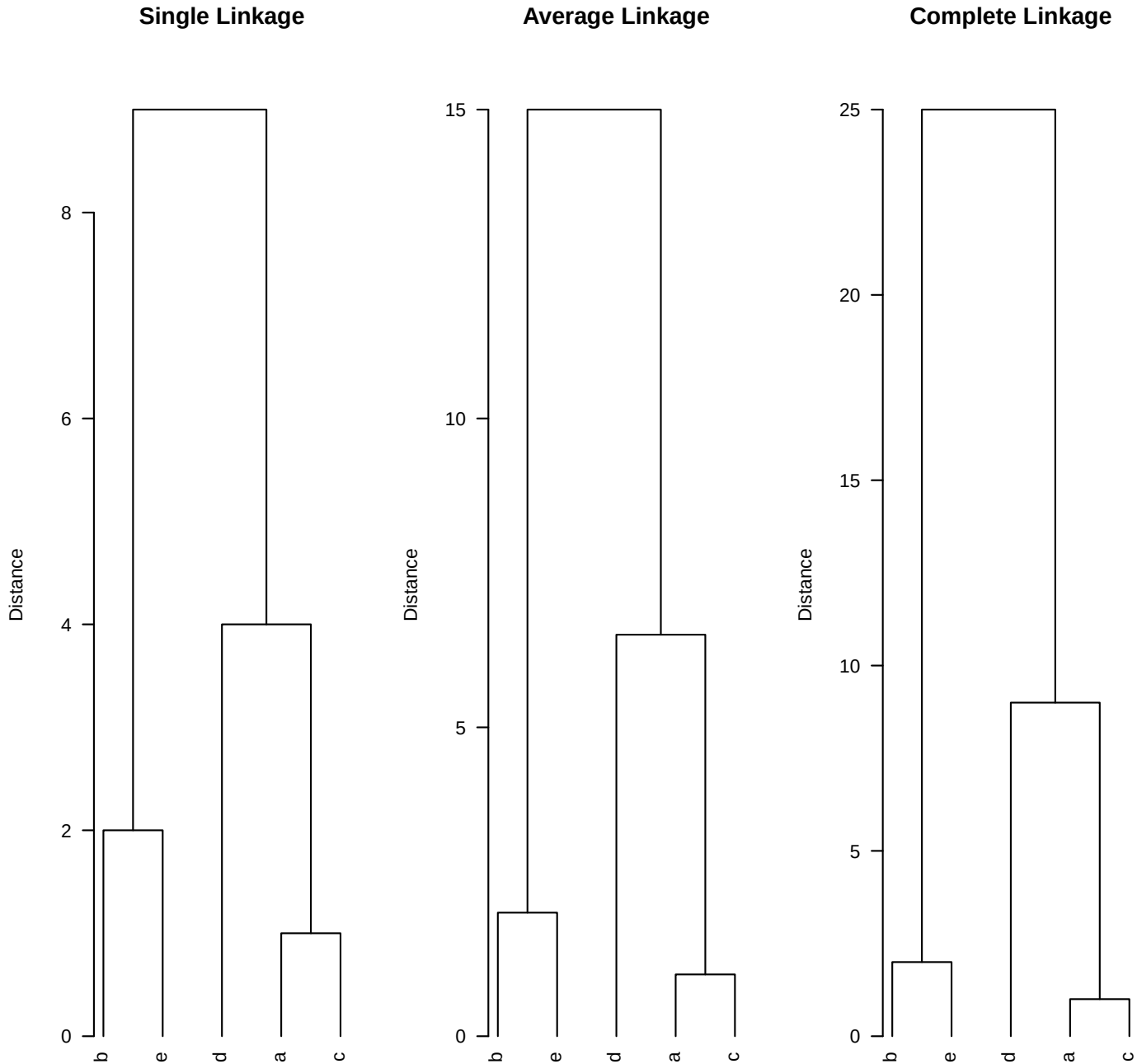
	a,c,d	b,e
a,c,d	0	
b,e	**9**	0




Dendrogram
Distance
8
6
4
2
0
b
e
d
a
c
single linkage


Dendrogram
Distance
25
20
15
10
5
0
b
e
d
a
c
complete linkage


Dendrogram
15
10
5
0
Distance
b
e
d
a
c
average linkage


Single Linkage
Average Linkage
Complete Linkage
Distance
0
2
4
6
8
0
5
10
15
0
5
10
15
20
25
b
e
d
a
c

```
     single average complete
[1,]      1     1.0        1
[2,]      2     2.0        2
[3,]      4     6.5        9
[4,]      9    15.0       25
```

Pros and Cons

Hierarchical Clustering

- Quadratic Cost
- The treee informs about the whole process of aggregation
- Gives clues about the number of groups
- Suboptimal partition (overlapping groups)

Partition Methods

- Linear Cost
- Number of groups must be predetermined
- Local optimal partition

Sequential (hybrid) Clustering

Taking advantage of both approaches, you can get a consolidation:

- Perform a hierarchical clustering
- Determine the number of groups
- Calculate the corresponding centroids
- Perform a k-means partition using as seeds the centroids previously calculated

Clustering Very Large Data Sets

- Perform $m = (2, 3)$ times a k-means algorithm (with $k = 10$)
- Form the crosstable of the m obtained partitions
- Calculate centroids of the non-empty cells of the crosstable.
- Perform a hierarchical clustering of the centroids, weighed with the number of individuals per cell
- Determine the number of groups
- Consolidate your clustering

Bibliography

- **Modern Multivariate Statistical Techniques** by Julian Izenman (2008). *Chapter 12: Cluster Analysis*. Springer.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 9: Cluster Analysis*. Wiley.

French literature

- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2006). *Chapter 11: Methodes de classification*. Editions Technip, Paris.
- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 2, section 2.1: Agregation autour des centres mobiles*. Dunod, Paris.
- **Approche pragmatique de la classification** by Nakache and Confais (2005). *Chapter 4: Classification par partition*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 9: Analyses des proximites, des preferences et typologie*. Editions Technip, Paris.