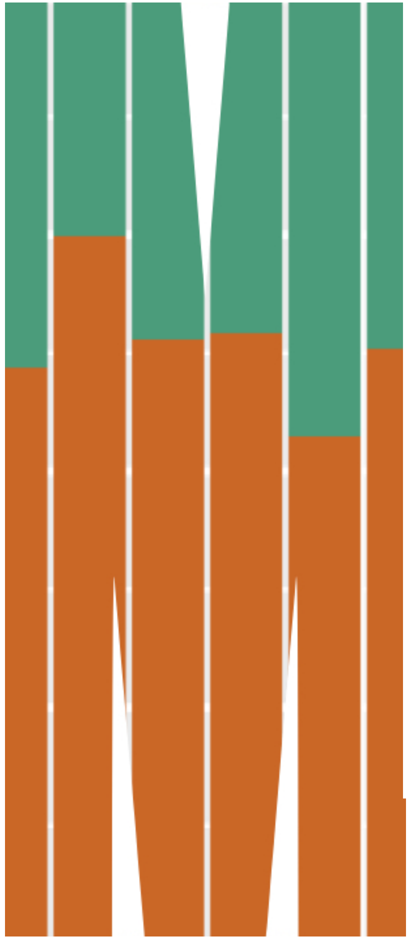


ETC3250/5250: Hierarchical clustering

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

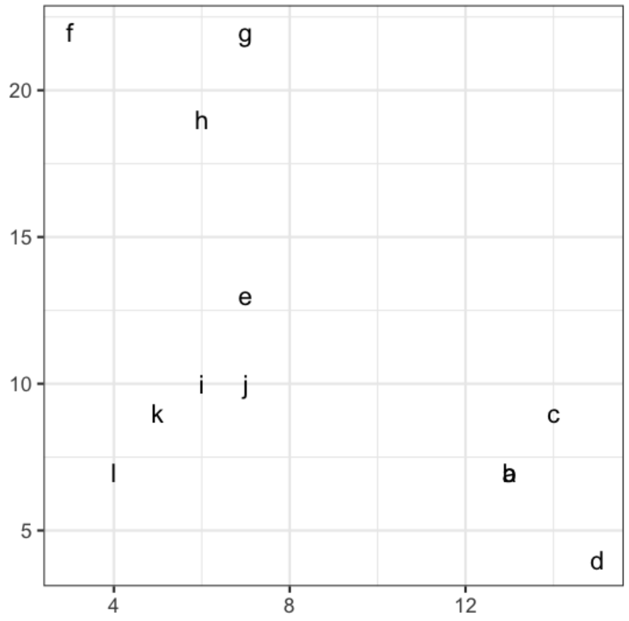
Week 11 (b)

Hierarchical clustering

- Agglomeration: Begin with all observations in singleton clusters. Sequentially join points into clusters, until all are in one cluster.
- Divisive: Begin with all observtions in one cluster, adn sequentially divide until all observations are in singleton clusters.
- Produces a tree diagram illustrating the process, called a dendrogram.

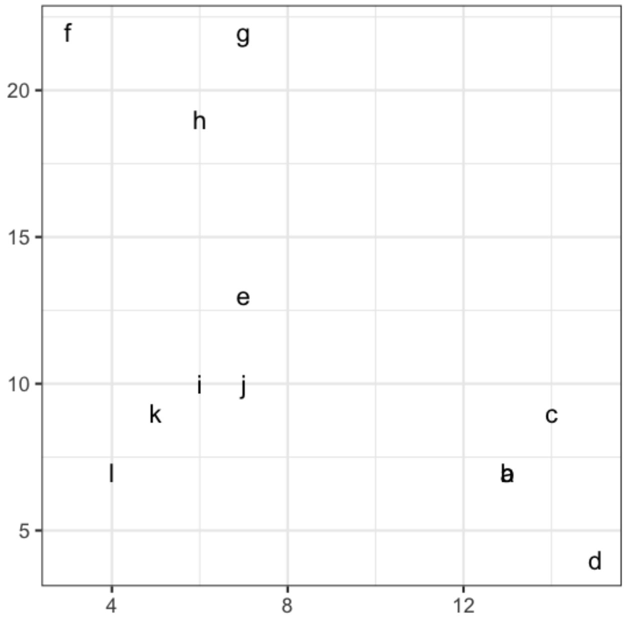
Some new data 🤸

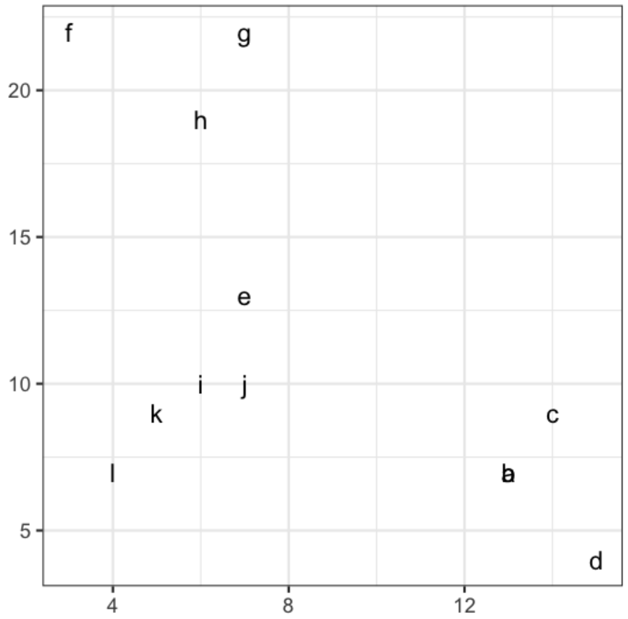
lbl	x1	x2
a	13	7
b	13	7
c	14	9
d	15	4
e	7	13
f	3	22
g	7	22
h	6	19
i	6	10
j	7	10
k	5	9
l	4	7

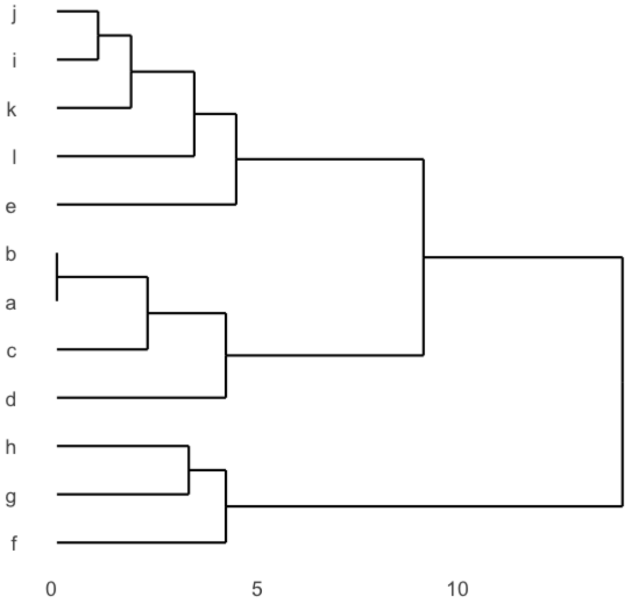


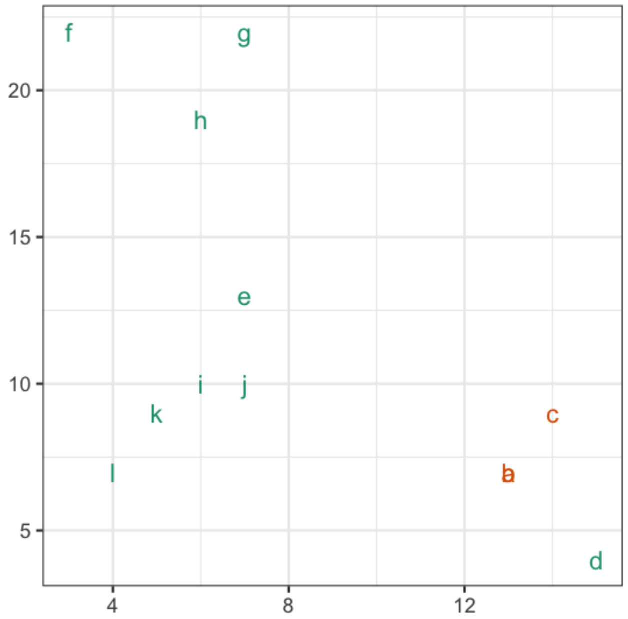
$n \times n$ distance matrix 📏

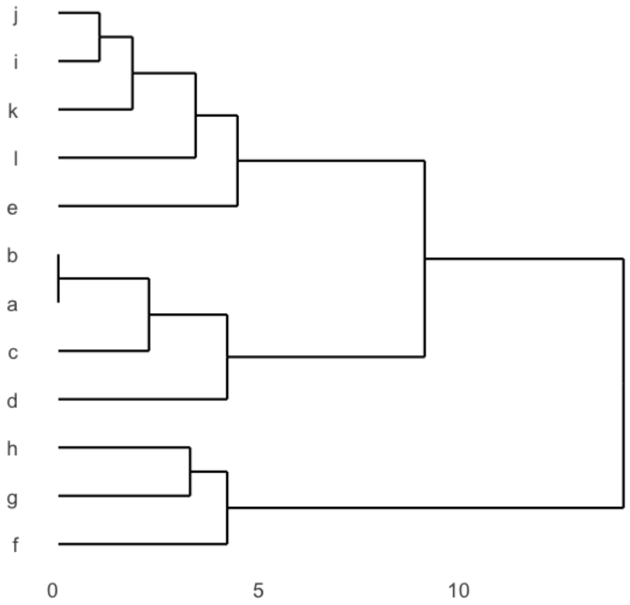
	a	b	c	d	e	f	g	h	i	j	k	l
a	0.0	0.0	2.2	3.6	8.5	18.0	16.2	13.9	7.6	6.7	8.2	9.0
b	0.0	0.0	2.2	3.6	8.5	18.0	16.2	13.9	7.6	6.7	8.2	9.0
c	2.2	2.2	0.0	5.1	8.1	17.0	14.8	12.8	8.1	7.1	9.0	10.2
d	3.6	3.6	5.1	0.0	12.0	21.6	19.7	17.5	10.8	10.0	11.2	11.4
e	8.5	8.5	8.1	12.0	0.0	9.8	9.0	6.1	3.2	3.0	4.5	6.7
f	18.0	18.0	17.0	21.6	9.8	0.0	4.0	4.2	12.4	12.6	13.2	15.0
g	16.2	16.2	14.8	19.7	9.0	4.0	0.0	3.2	12.0	12.0	13.2	15.3
h	13.9	13.9	12.8	17.5	6.1	4.2	3.2	0.0	9.0	9.1	10.0	12.2
i	7.6	7.6	8.1	10.8	3.2	12.4	12.0	9.0	0.0	1.0	1.4	3.6
j	6.7	6.7	7.1	10.0	3.0	12.6	12.0	9.1	1.0	0.0	2.2	4.2
k	8.2	8.2	9.0	11.2	4.5	13.2	13.2	10.0	1.4	2.2	0.0	2.2
l	9.0	9.0	10.2	11.4	6.7	15.0	15.3	12.2	3.6	4.2	2.2	0.0




f
g
20
h
15
e
10
i
j
k
c
l
b
a
5
d
4
8
12


j
i
k
l
e
b
a
c
d
h
g
f
0
5
10


f
g
20
h
15
e
10
i
j
k
c
l
b
a
5
d
4
8
12


j
i
k
l
e
b
a
c
d
h
g
f
0
5
10

🤔

What is the distance between the new cluster (a, b,c) and all of the other observations?

Linkage

Between points in the cluster to points not in the cluster.

- Single: minimum distance between points in the different clusters
- Complete: maximum distance between points in the different clusters
- Average: average of distances between points in the different clusters
- Centroid: distances between the average of the different clusters
- Wards: minimizes the total within-cluster variance

	a	b	c	d	e	f	g	h	i	j	k	l
a	0.0	0.0	2.2	3.6	8.5	18.0	16.2	13.9	7.6	6.7	8.2	9.0
b	0.0	0.0	2.2	3.6	8.5	18.0	16.2	13.9	7.6	6.7	8.2	9.0
c	2.2	2.2	0.0	5.1	8.1	17.0	14.8	12.8	8.1	7.1	9.0	10.2
d	3.6	3.6	5.1	0.0	12.0	21.6	19.7	17.5	10.8	10.0	11.2	11.4
e	8.5	8.5	8.1	12.0	0.0	9.8	9.0	6.1	3.2	3.0	4.5	6.7
f	18.0	18.0	17.0	21.6	9.8	0.0	4.0	4.2	12.4	12.6	13.2	15.0
g	16.2	16.2	14.8	19.7	9.0	4.0	0.0	3.2	12.0	12.0	13.2	15.3
h	13.9	13.9	12.8	17.5	6.1	4.2	3.2	0.0	9.0	9.1	10.0	12.2
i	7.6	7.6	8.1	10.8	3.2	12.4	12.0	9.0	0.0	1.0	1.4	3.6
j	6.7	6.7	7.1	10.0	3.0	12.6	12.0	9.1	1.0	0.0	2.2	4.2
k	8.2	8.2	9.0	11.2	4.5	13.2	13.2	10.0	1.4	2.2	0.0	2.2
l	9.0	9.0	10.2	11.4	6.7	15.0	15.3	12.2	3.6	4.2	2.2	0.0

Distance (linkage) between (a,b,c) and d

Single: 3.6 or 5.1 $\longrightarrow$ 3.6
Complete: 3.6 or 5.1 $\longrightarrow$ 5.7
Average: (3.6 + 3.6 + 5.1)/3 = 4.1
Centroid: 4.1
mean of (a,b,c) is (13.3, 7.7)
mean of d (15, 4)

$$\sqrt{(13.3-15)^2+(7.7-4)^2}$$

Wards: Your turn to calculate it 🤷‍♀️

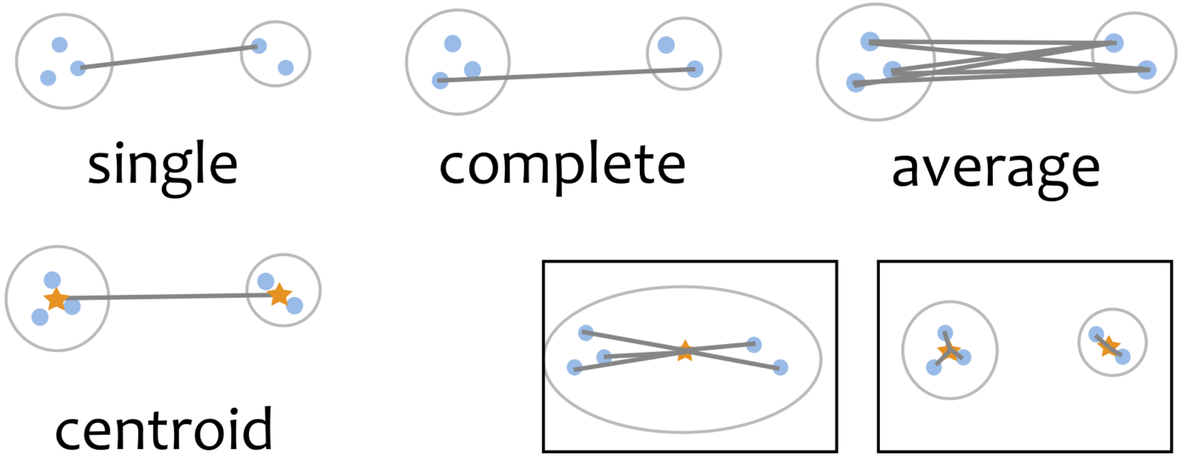
lbl	x1	x2	cl11
a	13	7	2
b	13	7	2
c	14	9	2
d	15	4	1
e	7	13	1
f	3	22	1
g	7	22	1
h	6	19	1
i	6	10	1
j	7	10	1
k	5	9	1
l	4	7	1

Distance (linkage) between (a,b,c) and d

Single: 3.6 or 5.1 → 3.6
Complete: 3.6 or 5.1 → 5.7
Average: (3.6 + 3.6 + 5.1)/3 = 4.1
Centroid: 4.1
mean of (a,b,c) is (13.3, 7.7)
mean of d (15, 4)

$$\sqrt{(13.3-15)^2+(7.7-4)^2}$$

Wards: Your turn to calculate it 🤷‍♀️



single

complete

average

centroid

wards

Single linkage reduced distance matrix

	(a,b,c)	d	e	f	g	h	i	j	k	l
(a,b,c)	0.0	3.6	8.1	17.0	14.8	12.8	7.6	6.7	9.0	9.0
d	3.6	0.0	12.0	21.6	19.7	17.5	10.8	10.0	11.2	11.4
e	8.1	12.0	0.0	9.8	9.0	6.1	3.2	3.0	4.5	6.7
f	17.0	21.6	9.8	0.0	4.0	4.2	12.4	12.6	13.2	15.0
g	14.8	19.7	9.0	4.0	0.0	3.2	12.0	12.0	13.2	15.3
h	12.8	17.5	6.1	4.2	3.2	0.0	9.0	9.1	10.0	12.2
i	7.6	10.8	3.2	12.4	12.0	9.0	0.0	1.0	1.4	3.6
j	6.7	10.0	3.0	12.6	12.0	9.1	1.0	0.0	2.2	4.2
k	9.0	11.2	4.5	13.2	13.2	10.0	1.4	2.2	0.0	2.2
l	9.0	11.4	6.7	15.0	15.3	12.2	3.6	4.2	2.2	0.0

Now which distance is the smallest?

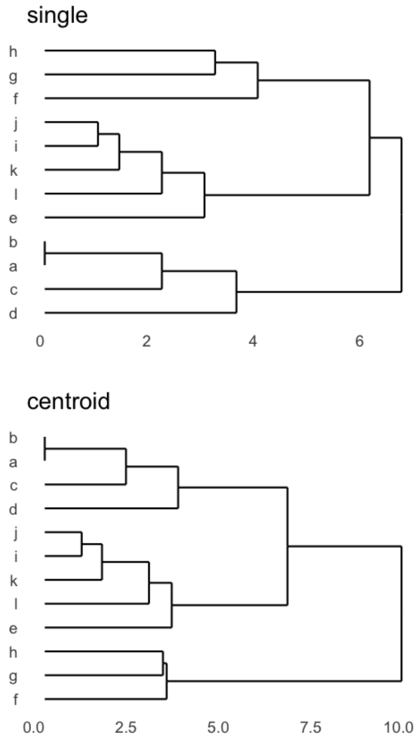
i, j would be joined at the next step

Average linkage reduced distance matrix

	(a,b,c)	d	e	f	g	h	i	j	k	l
(a,b,c)	0.0	4.1	8.4	17.7	15.7	13.5	7.7	6.8	9.4	9.0
d	4.1	0.0	12.0	21.6	19.7	17.5	10.8	10.0	11.2	11.4
e	8.4	12.0	0.0	9.8	9.0	6.1	3.2	3.0	4.5	6.7
f	17.7	21.6	9.8	0.0	4.0	4.2	12.4	12.6	13.2	15.0
g	15.7	19.7	9.0	4.0	0.0	3.2	12.0	12.0	13.2	15.3
h	13.5	17.5	6.1	4.2	3.2	0.0	9.0	9.1	10.0	12.2
i	7.7	10.8	3.2	12.4	12.0	9.0	0.0	1.0	1.4	3.6
j	6.8	10.0	3.0	12.6	12.0	9.1	1.0	0.0	2.2	4.2
k	9.4	11.2	4.5	13.2	13.2	10.0	1.4	2.2	0.0	2.2
l	9.0	11.4	6.7	15.0	15.3	12.2	3.6	4.2	2.2	0.0

Now which distance is the smallest?

Its still i, j that would be joined at the next step



single

complete

average

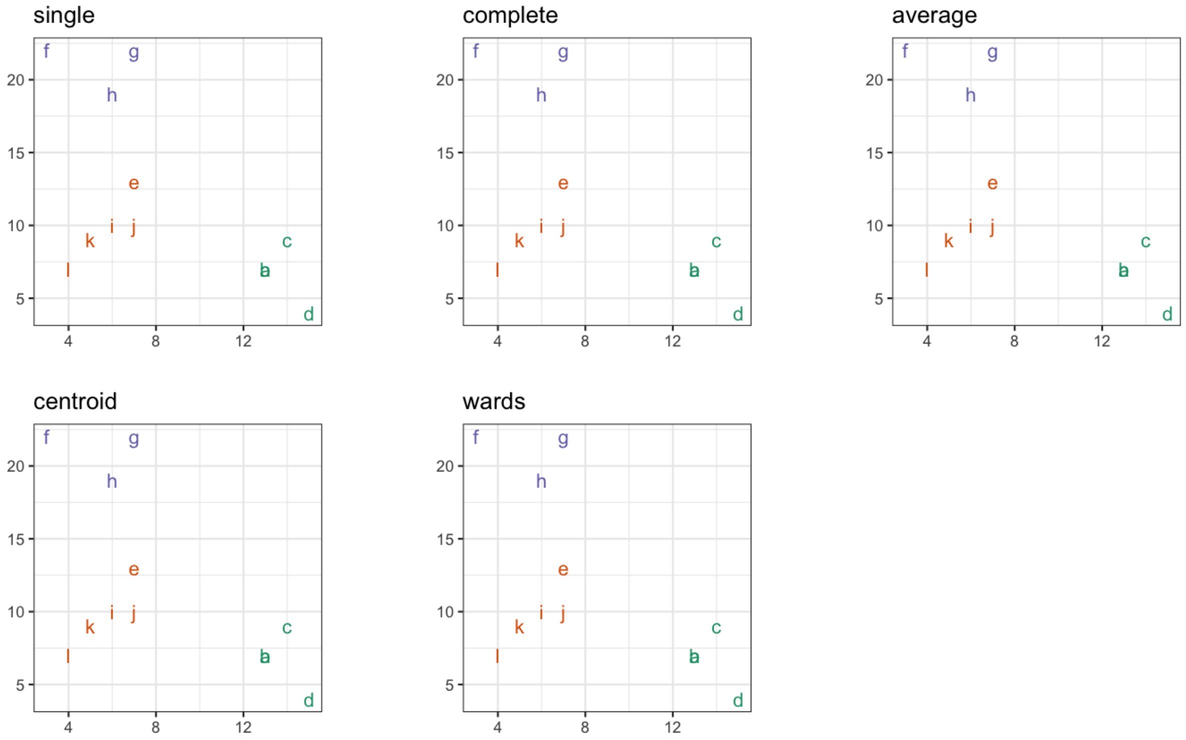
centroid

wards

Dendrogram

- Each leaf of the dendrogram represents one observation
- Leaves fuse into branches and branches fuse, either with leaves or other branches.
- Fusions lower in the tree mean the groups of observations are more similar to each other.

Cut the tree to partition the data into k clusters.



single

complete

average

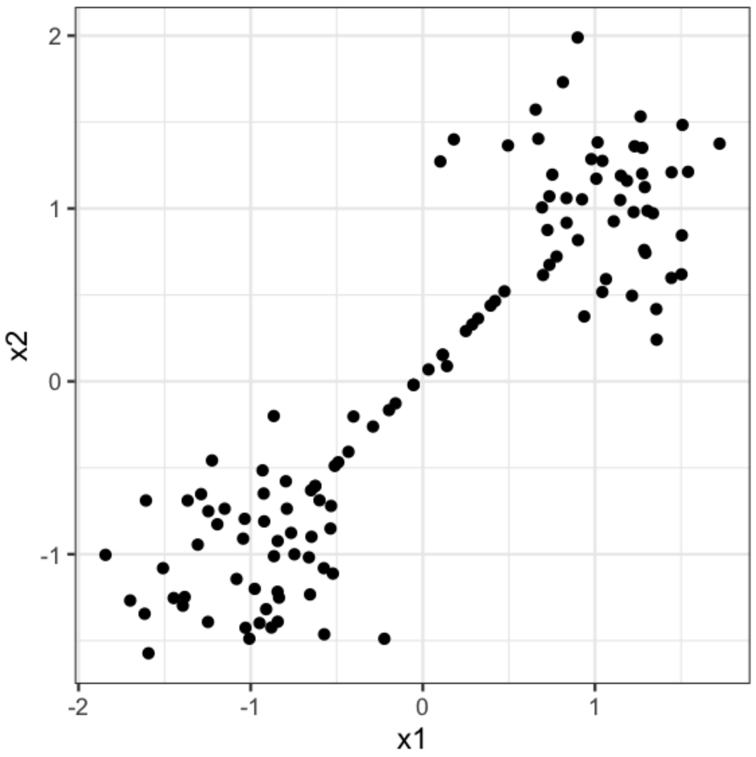
centroid

wards

Pros and cons

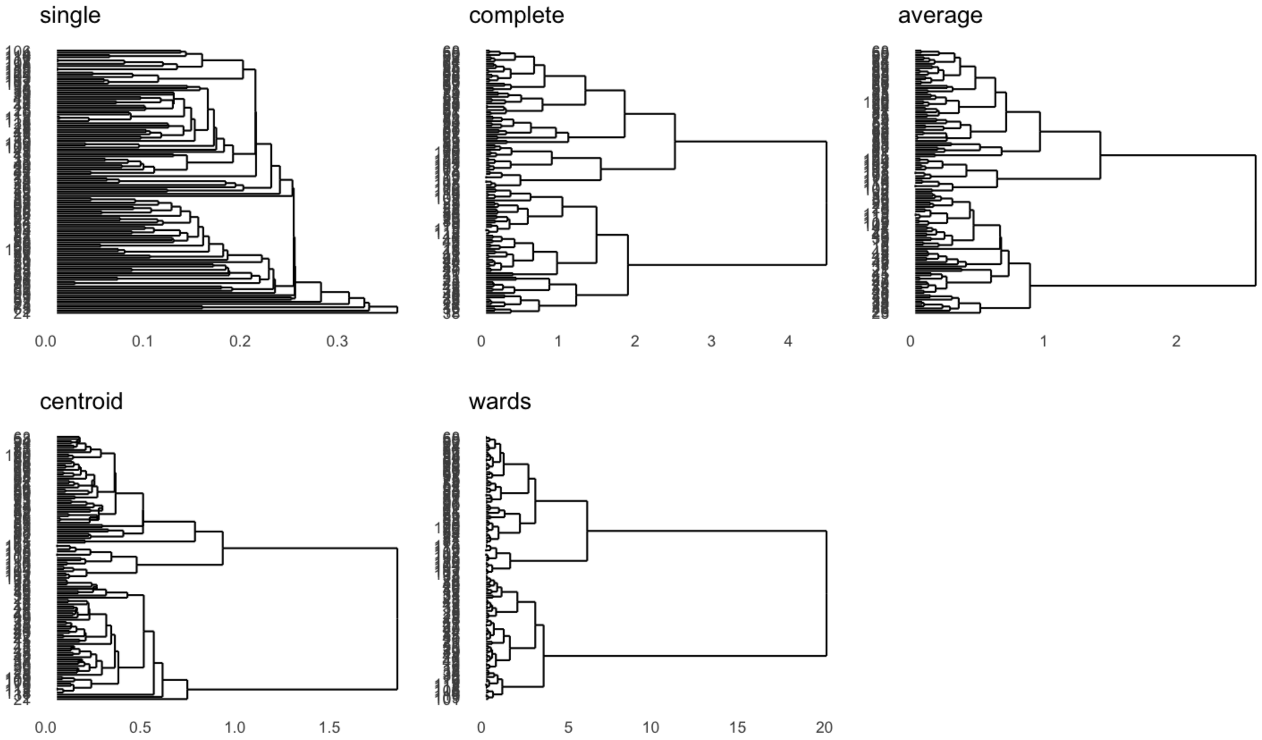
- Single linkage tends to "chain" the data into long stringy clusters, can avoid confusion from nuisance variables but gets confused by "inliers" (outliers between clusters)
- Complete linkage tends to be confused by nuisance variables, but not by inliers
- Wards tends to create spherical homogeneously shaped clusters, a little similar to k-means

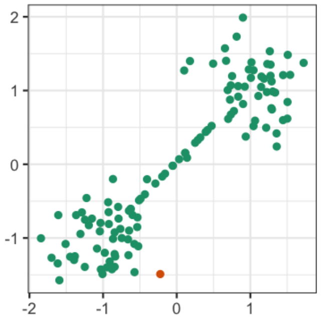
No one perfect method for all problems, but Wards tends to be a good starting point.



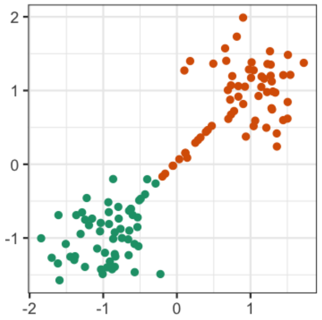
Nuisance cases "Hansel and Gretel data"

Points that are between major clusters of data. This affects some linkage methods, eg single, which will tend to "chain" through the data grouping everything together.

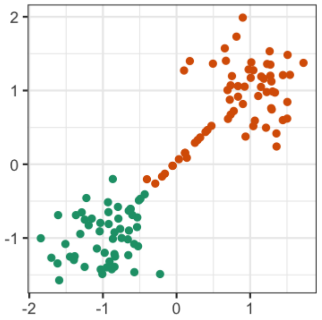

single
0.0
0.1
0.2
0.3
complete
0
1
2
3
4
average
0
1
2
centroid
0.0
0.5
1.0
1.5
wards
0
5
10
15
20



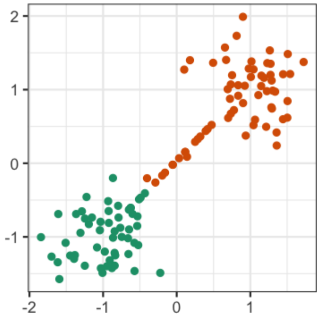
single



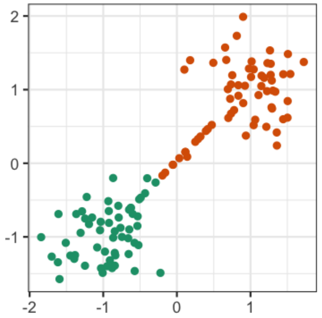
complete



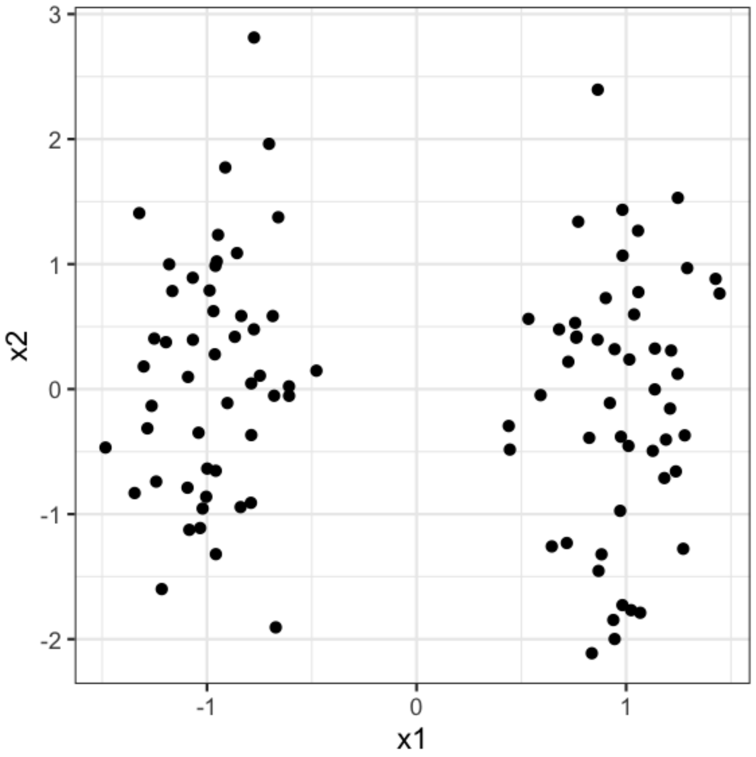
average



centroid

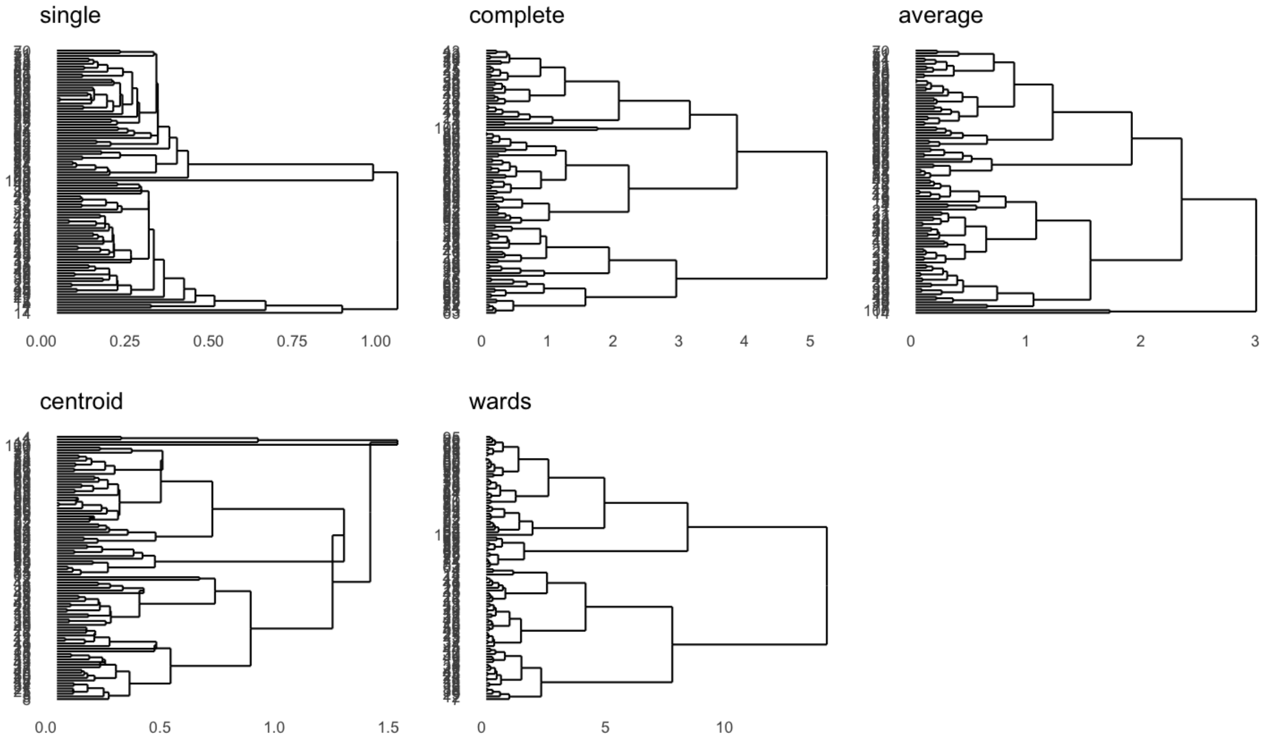


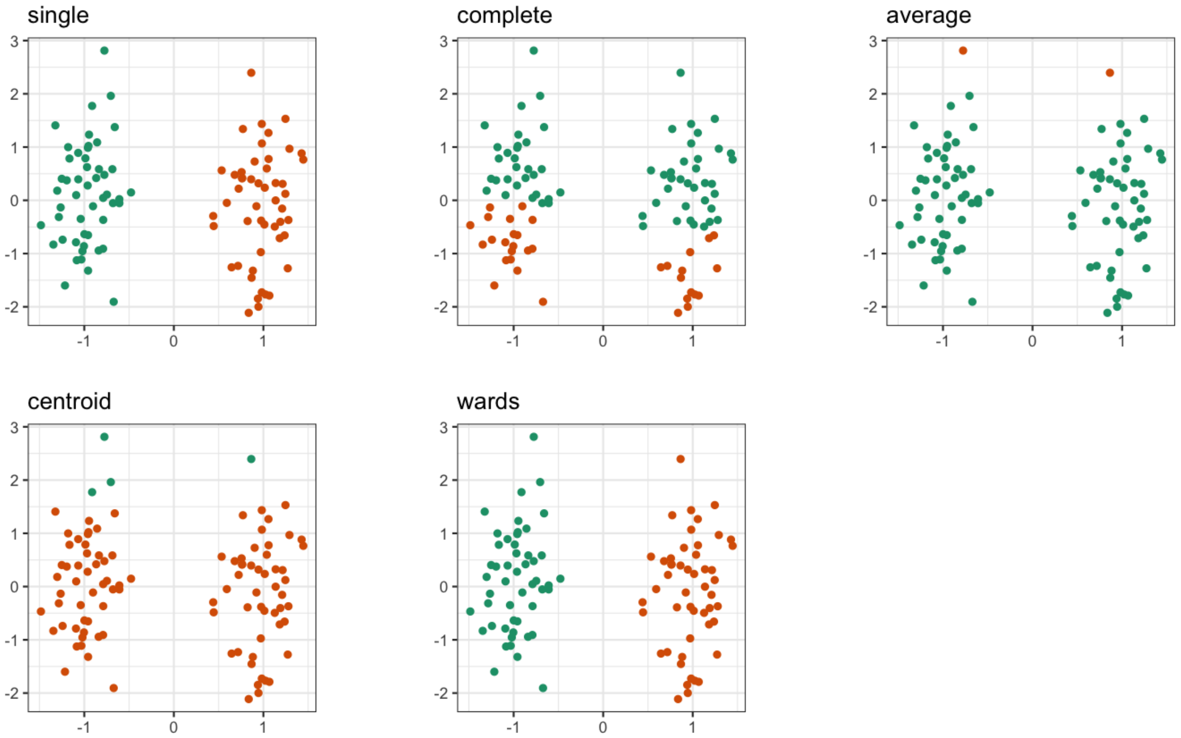
wards



Nuisance variables

Variables that don't contribute to the clustering but are included in the distance calculations.

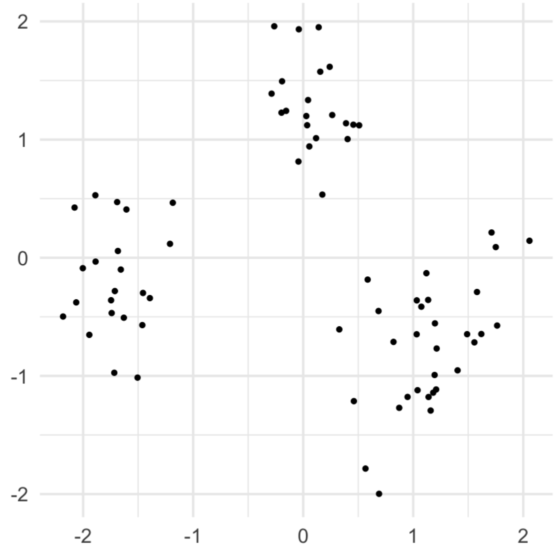

single
0.00
0.25
0.50
0.75
1.00
complete
0
1
2
3
4
5
average
0
1
2
3
centroid
0.0
0.5
1.0
1.5
wards
0
5
10

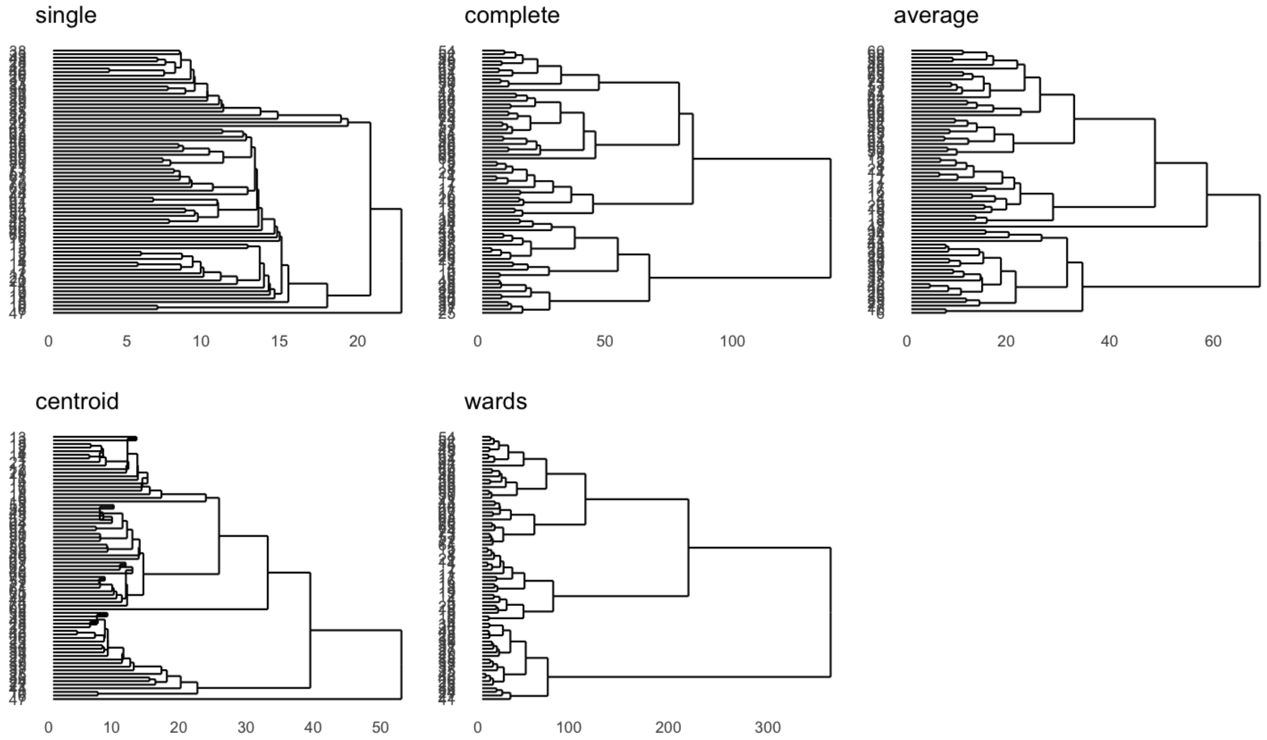

single
complete
average
centroid
wards
3
2
1
0
-1
-2
-1
0
1

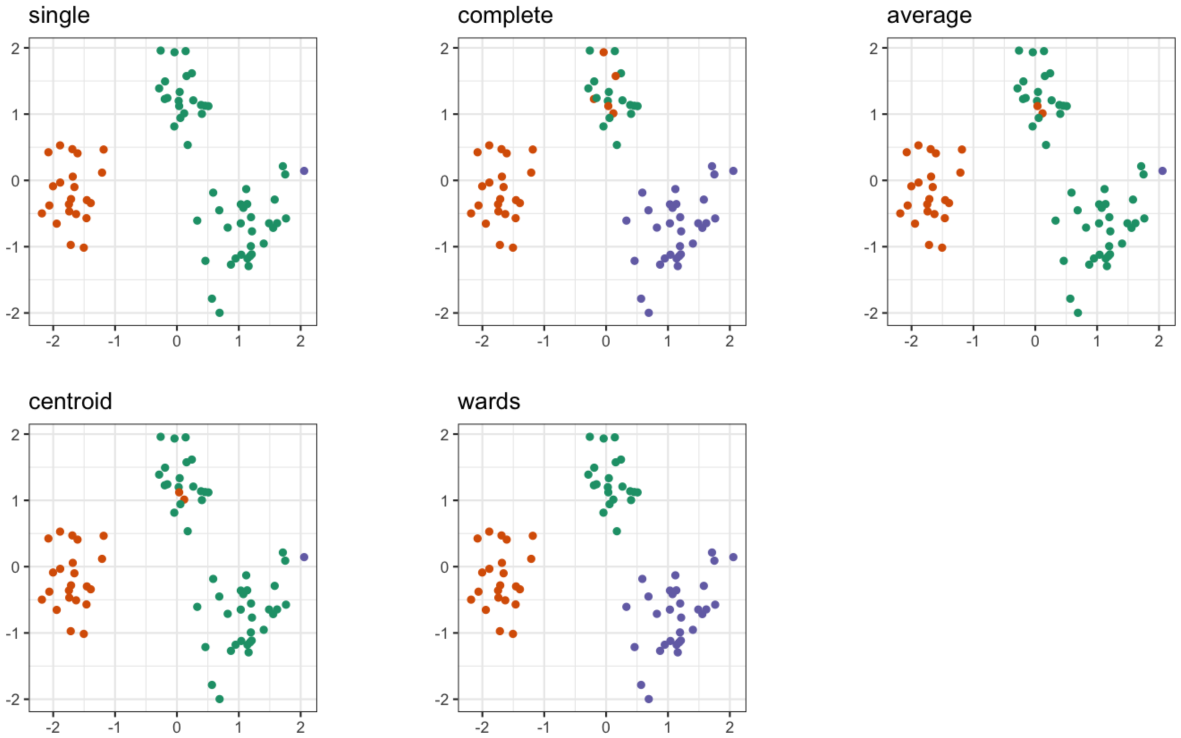
Example - flea data

Flea data: 6 variables, 74 cases. Shown here is the best 2D projection showing the three true clusters.

Data has a mix of nuisance variables and observations.




single
0
5
10
15
20
complete
0
50
100
average
0
20
40
60
centroid
0
10
20
30
40
50
wards
0
100
200
300


single
complete
average
centroid
wards
2
1
0
-1
-2
-2
-1
0
1
2

Dendrogram in p-space

Examining the dendrogram in the high-dimensional data space can be done using the tour. You need to

1. Add points to the data to provide the places where the leaves join. These are the nodes in the dendrogram.
2. Create a data set of edges, indicating which points should be connected.

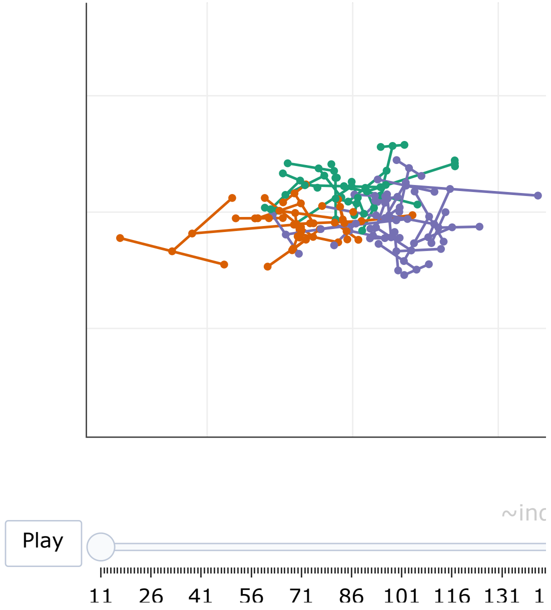
👩‍💻 Made by a human with a computer

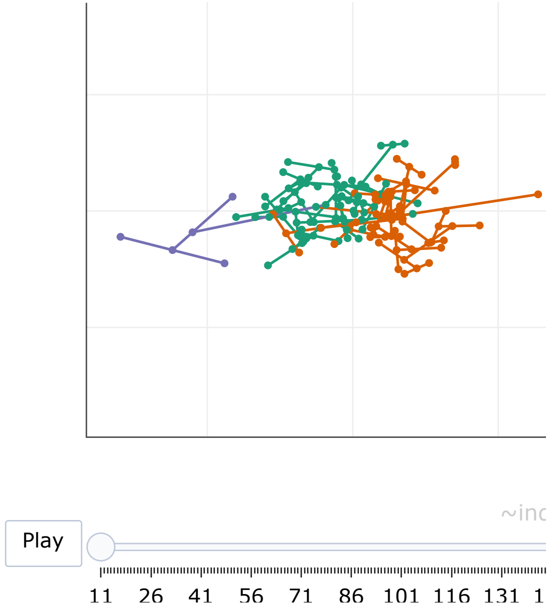
Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.

Dendrogram in p-dimensions (Wards and average linkage)


~inc
Play


~inc
Play