MONASH University

# ETC3250/5250: Introduction to Machine Learning

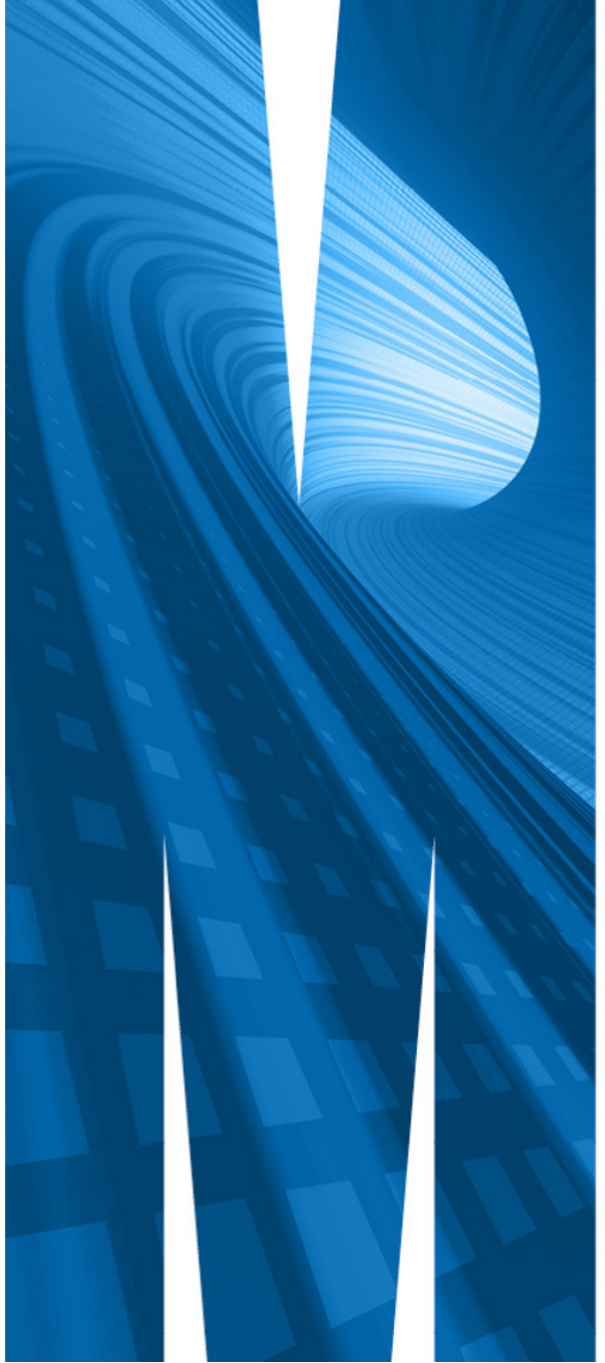
## Assessing clustering results

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

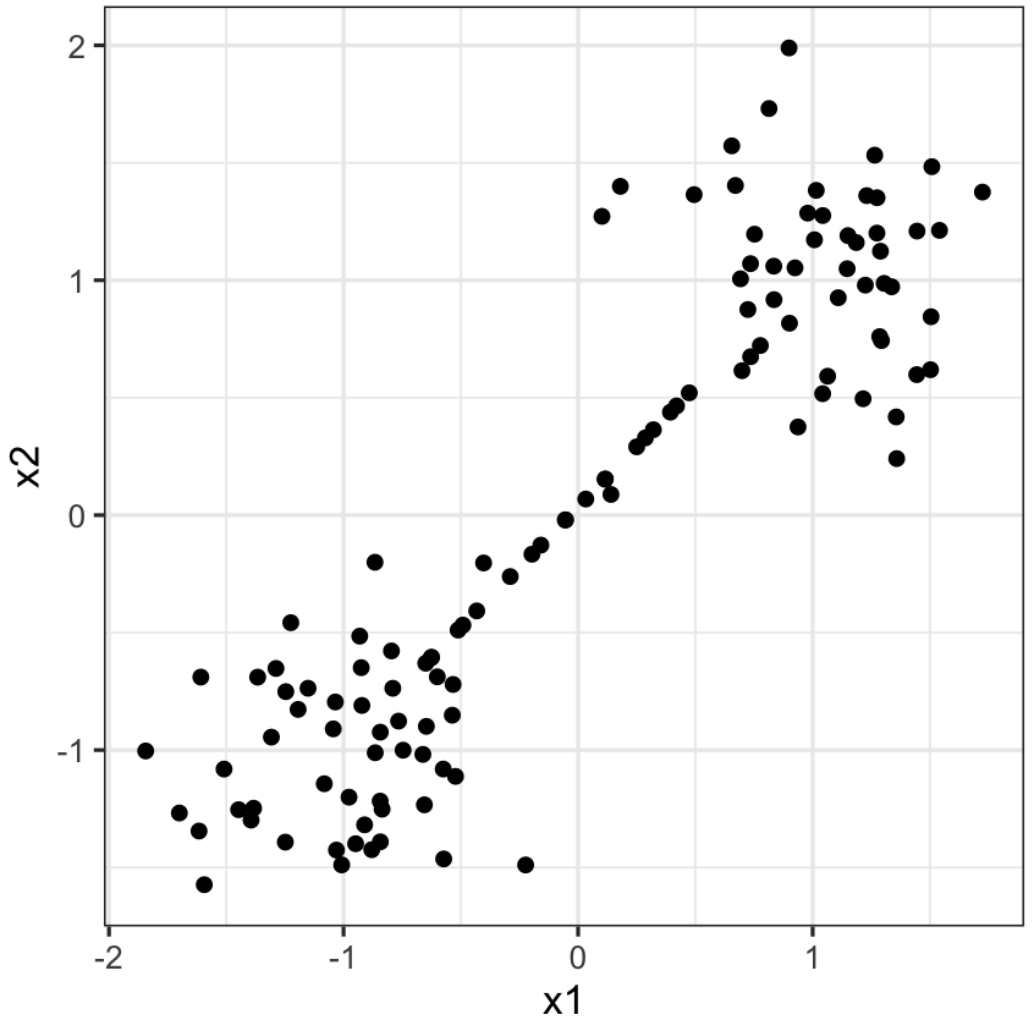
ETC3250.Clayton-x@monash.edu

Week 10b



# Where cluster algorithms can be tripped up

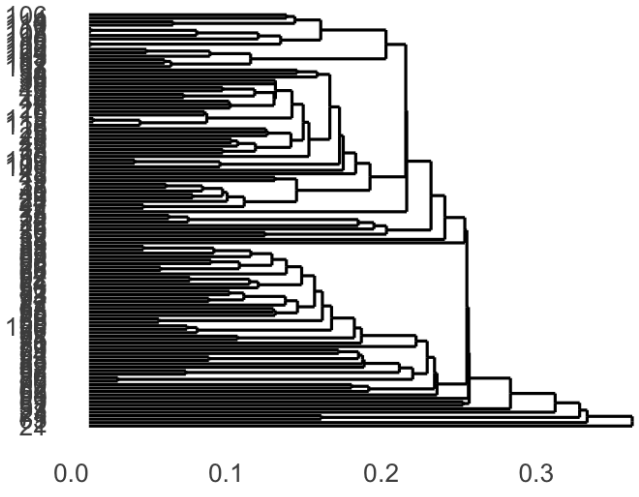
# Inlier-outlier observations



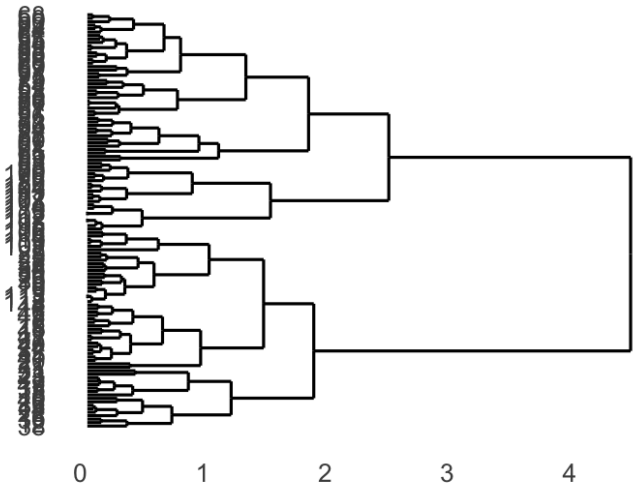

Nuisance cases "Hansel and Gretel data"

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.

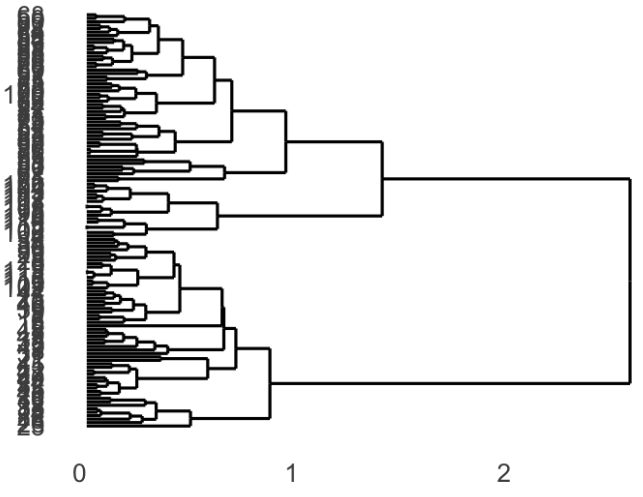
single



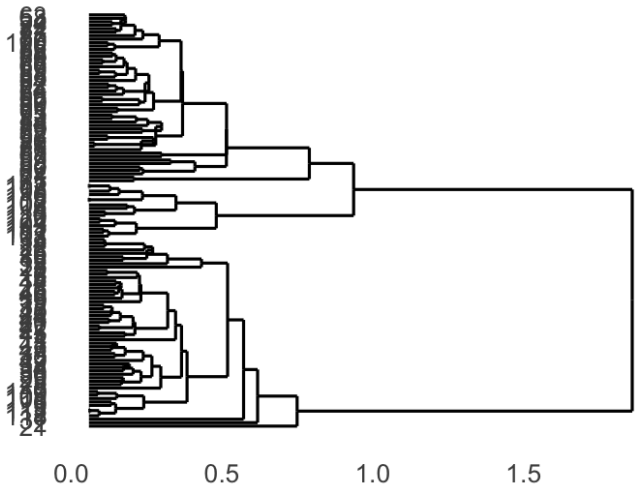

complete



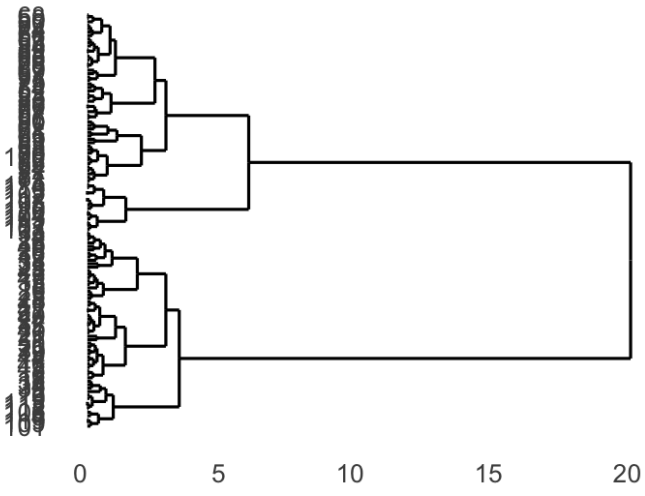

average



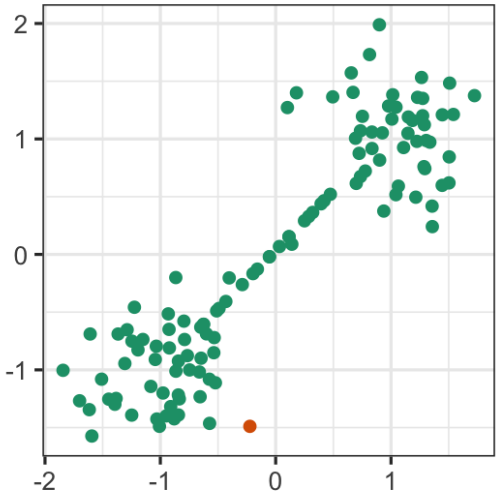

centroid



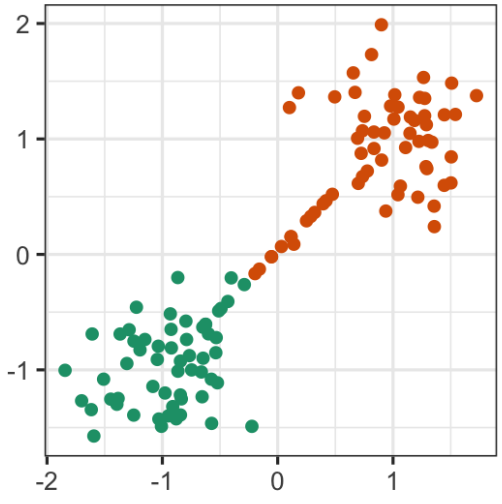

wards



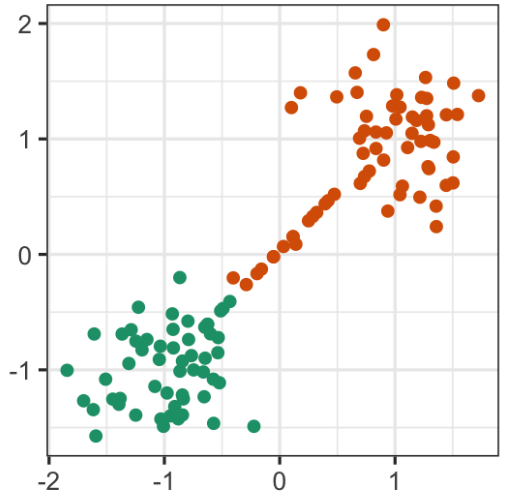
single



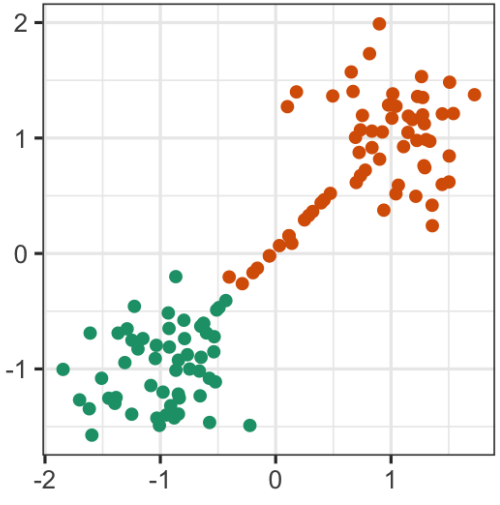

complete



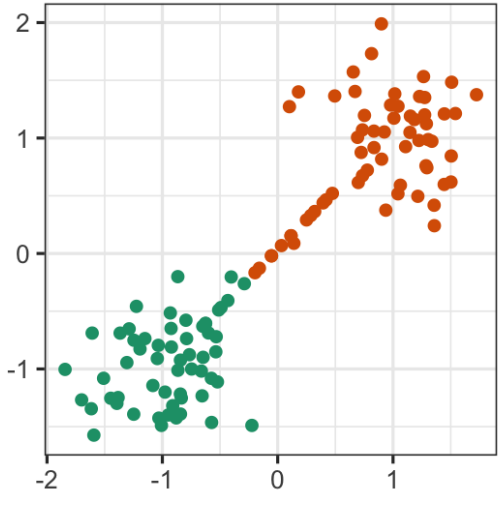

average



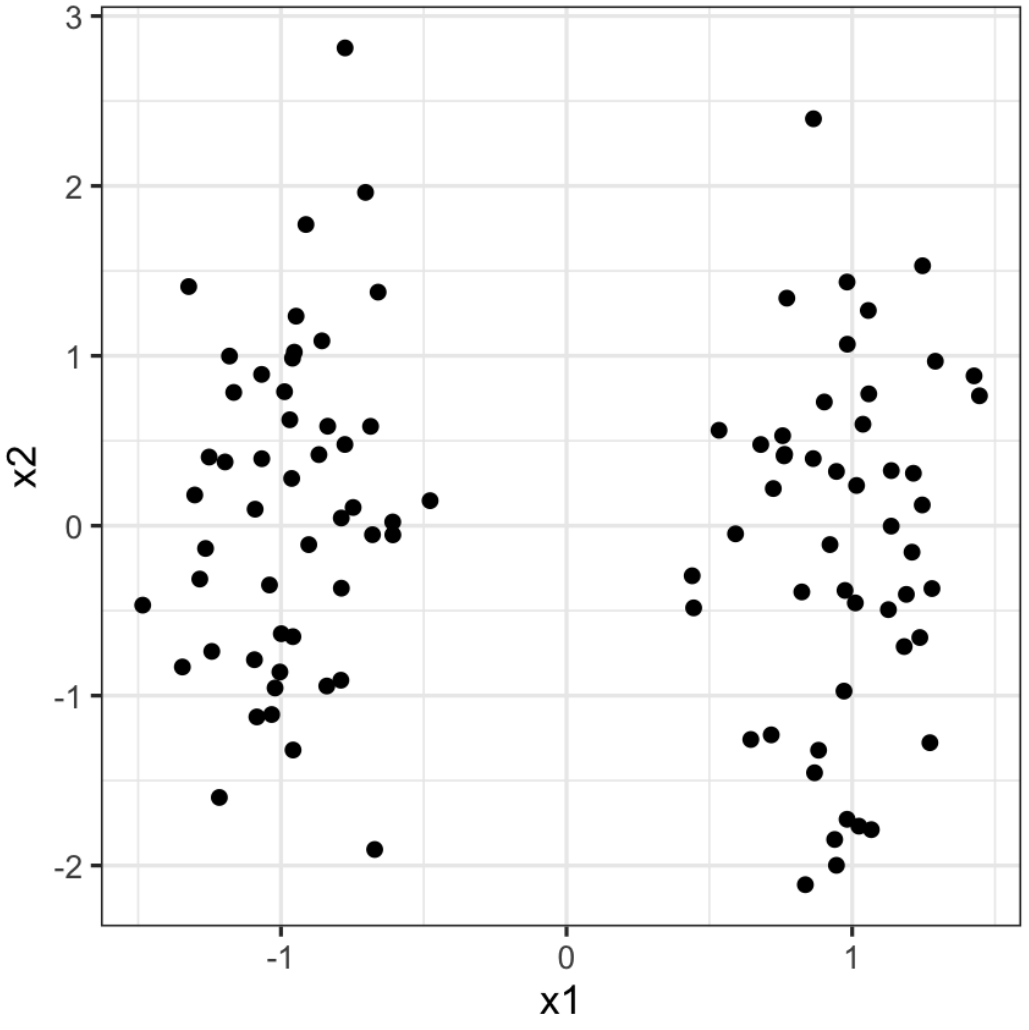

centroid




wards



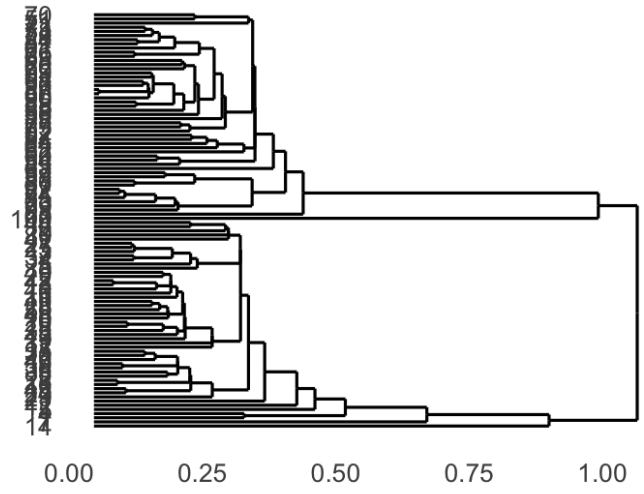
# Nuisance variables



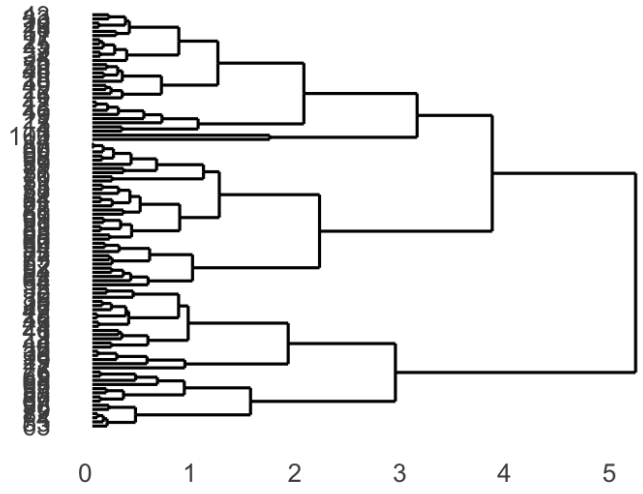

Variables that don't contribute to the clustering but are included in the distance calculations.

x2 is a nuisance variable.

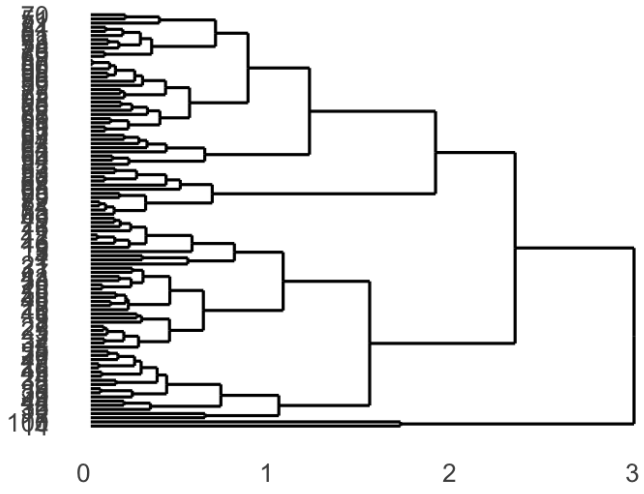
single



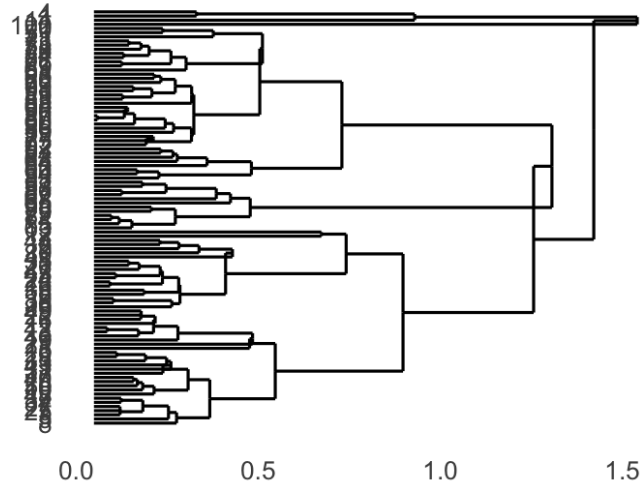

complete



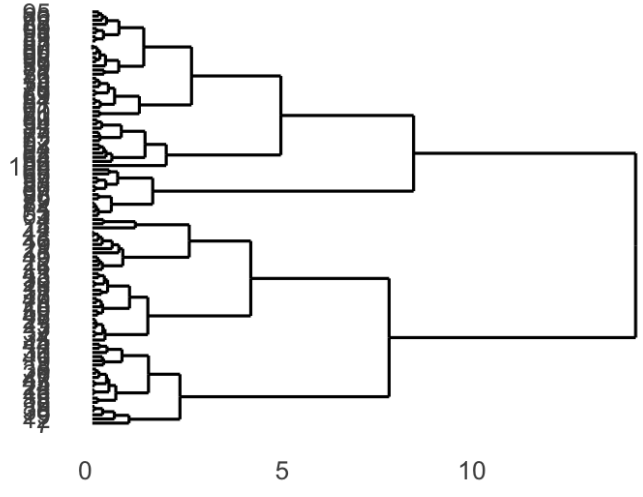

average

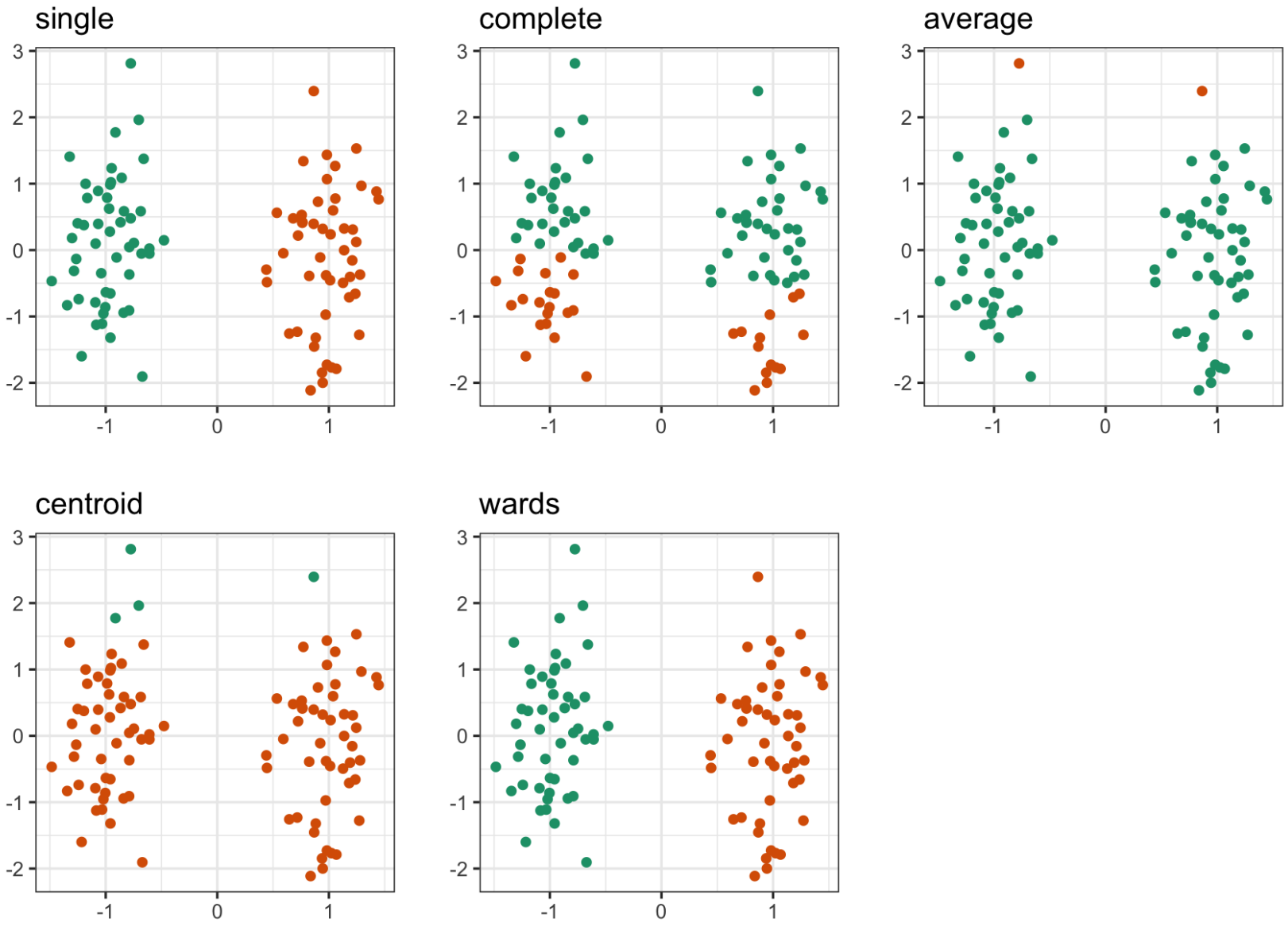



centroid




wards

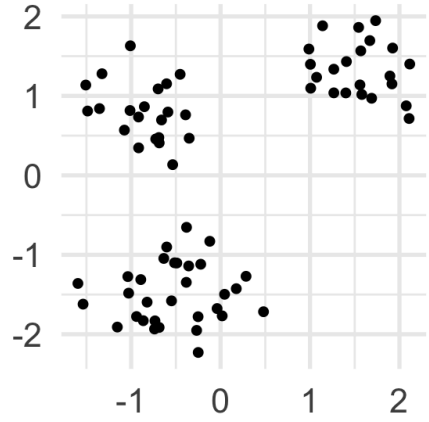



single
complete
average
centroid
wards
3
2
1
0
-1
-2
-1
0
1

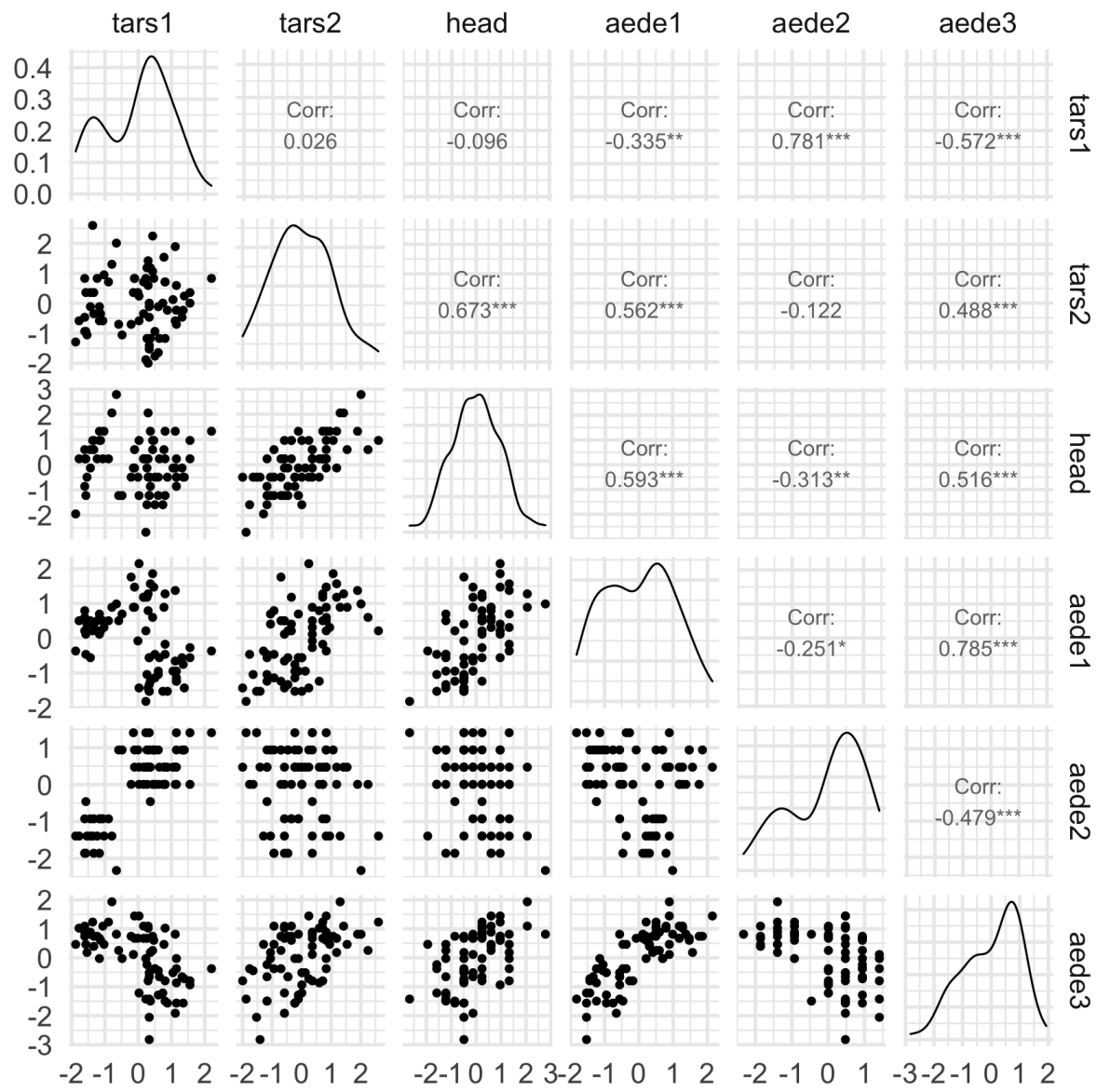
# Example - flea data

6 variables, 74 cases. Three very clear clusters.

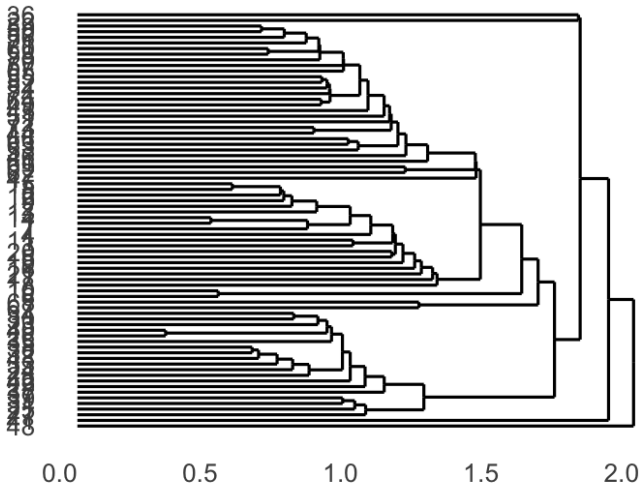
Data has a mix of nuisance variables and nuisance observations.



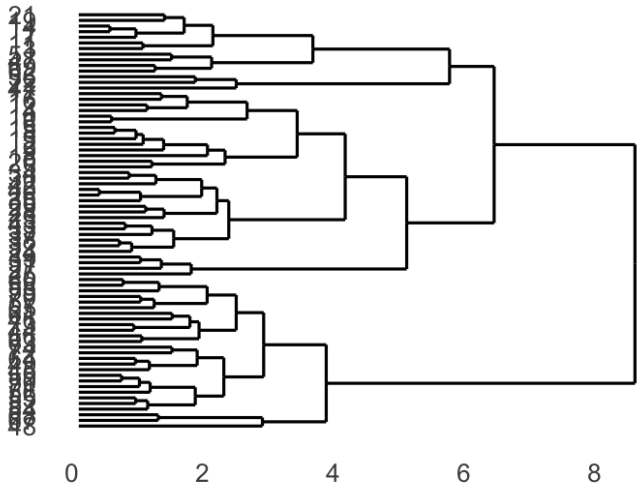
Above is the 2D projection pursuit dimension reduction, using LDA index with true class.



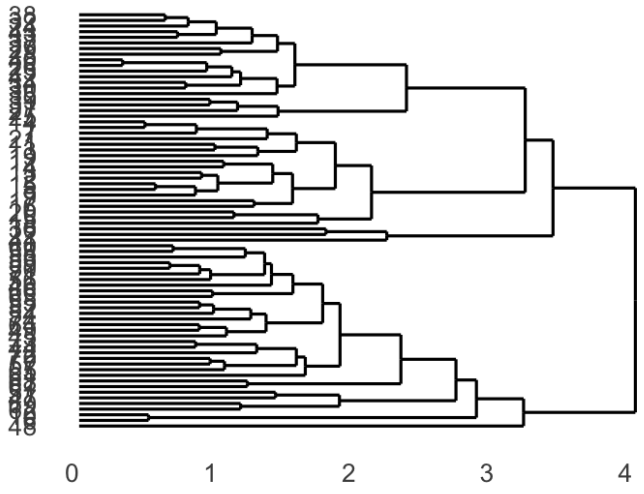
single



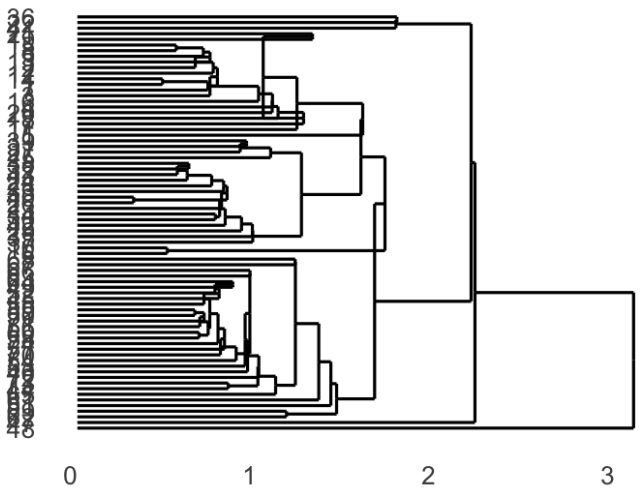

complete



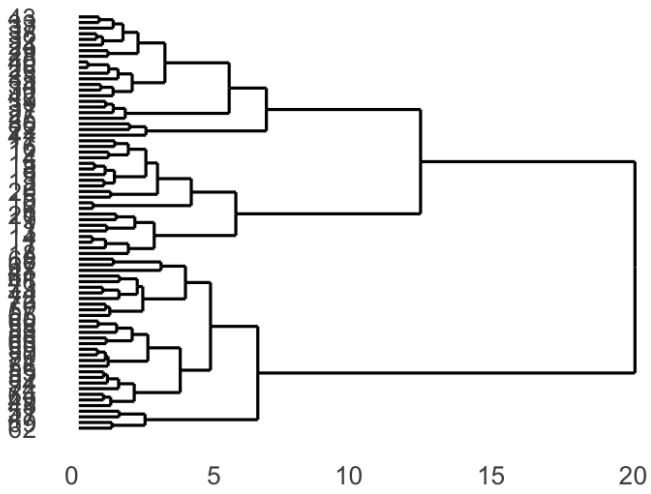

average



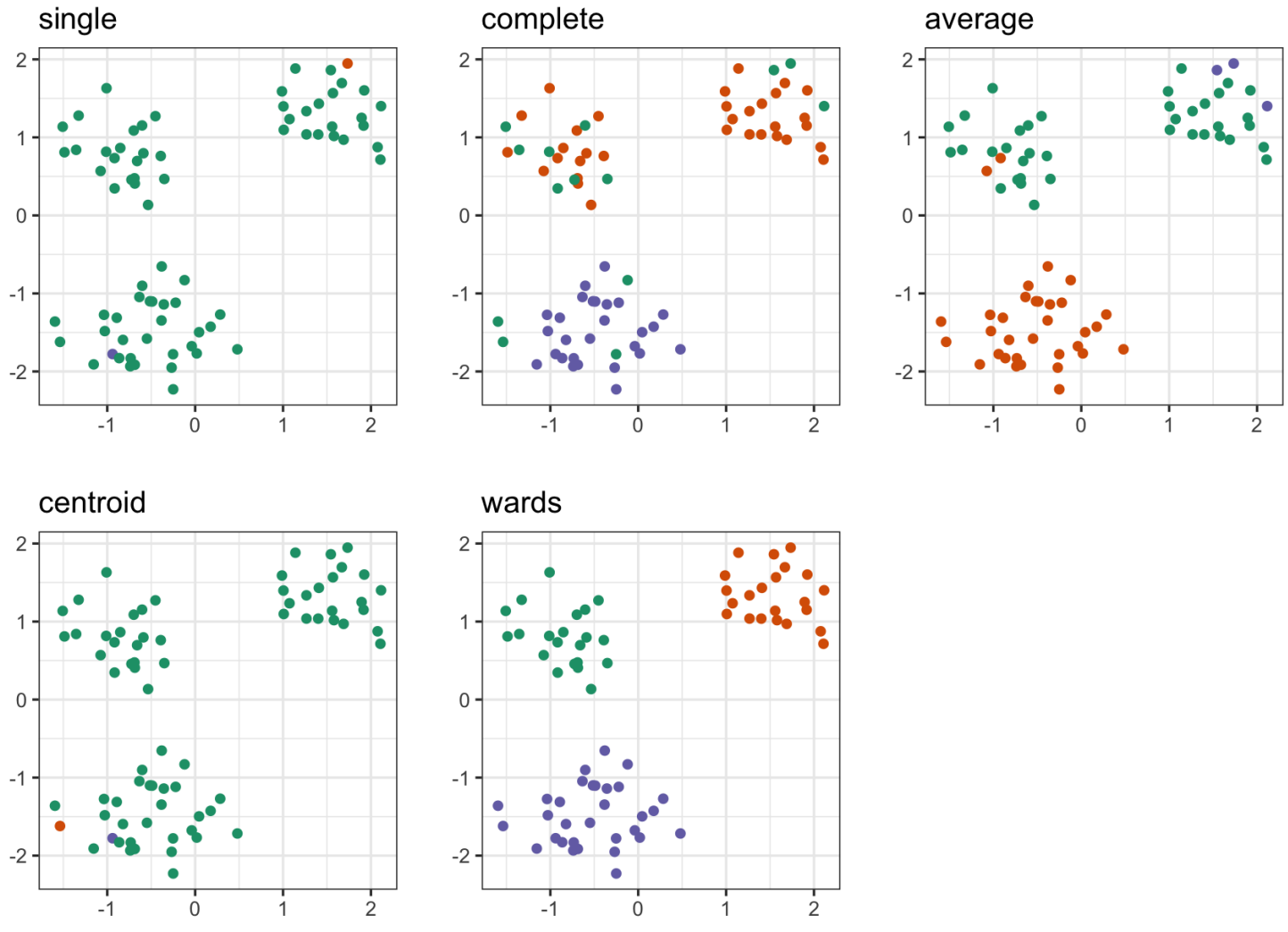

centroid




wards



# Cluster solutions plotted in 2D projection pursuit dimension reduction, using LDA index with true class



Note that, if clustering is conducted on the 2D projection, where clusters are well-separated, all linkage methods produce the three true clusters.

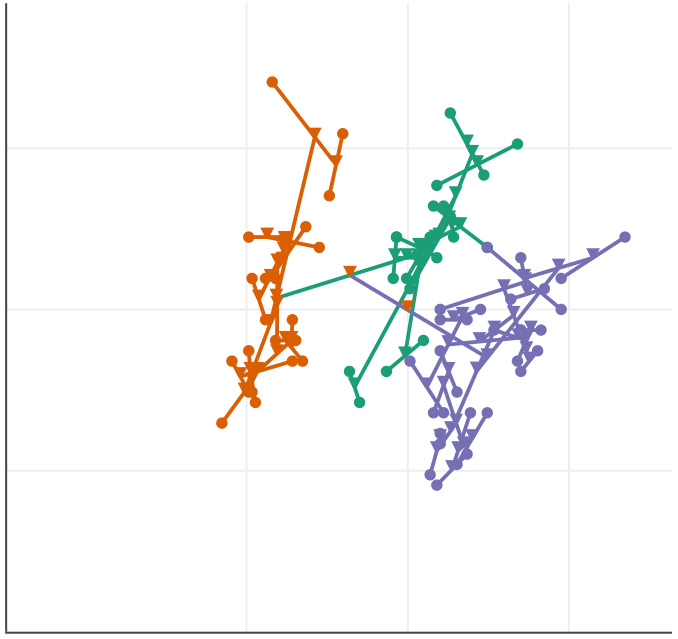
If we could do dimension reduction to remove nuisance variables prior to clustering, all would be so much easier. But this is hard.

# Dendrogram in p-space

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

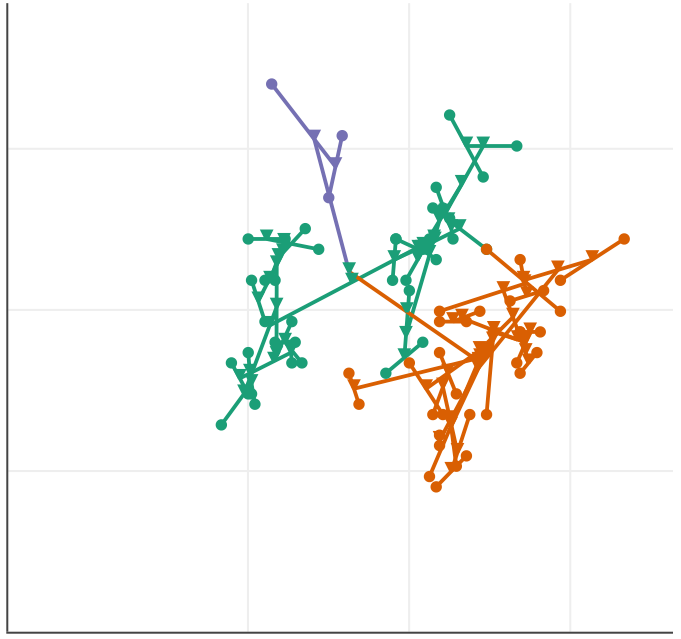
1. Add points to the data to provide the places where the leaves join. These are the nodes in the dendrogram.
2. Create a data set of edges, indicating which points should be connected.

# Dendrogram in $p$ dimensions (Wards and average linkage)



Play

11 26 41 56 71 86 101 116 131 14



Play

11 26 41 56 71 86 101 116 131 14

# Comparing cluster solutions

# Confusion table

Ward's linkage solution in columns, and average linkage in rows.

| cl_av | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 19 | 19 | 0 |
| 2 | 2 | 0 | 31 |
| 3 | 0 | 3 | 0 |

Solutions agree on 31 observations. Named differently: Wards labels group "3", and average labels it "2".

Re-number the labels. Change average "2" to "3"

| cl_av | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 19 | 19 | 0 |
| 2 | 0 | 3 | 0 |
| 3 | 2 | 0 | 31 |

Now agreement can be viewed as numbers on main diagonal, as used in a labelled class confusion matrix.

Methods agree on 19+3+31 out of 74 observations, 71.6%.

# Summarising a clustering

# Clustering summaries

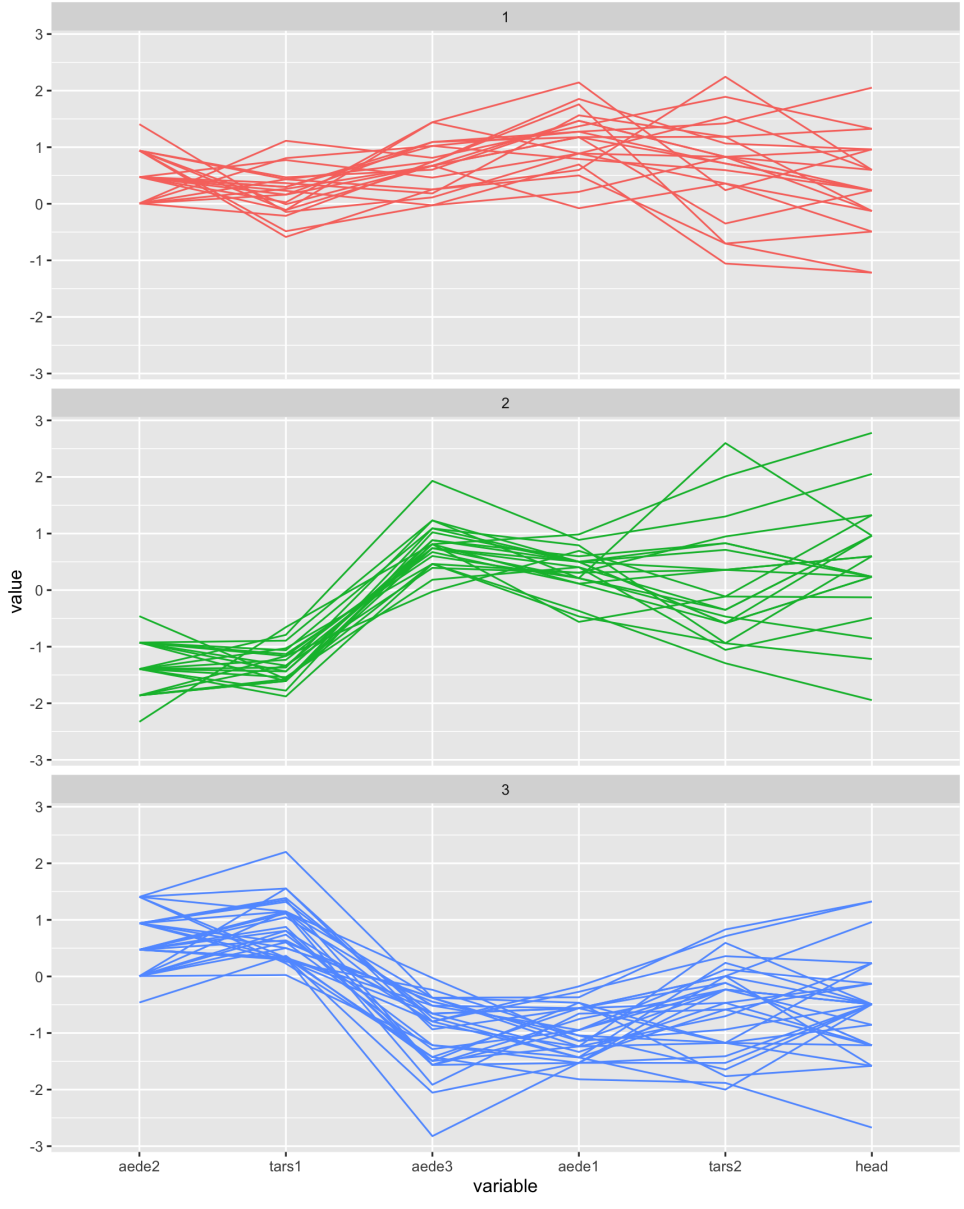
Once you have cluster labels, the data can be treated like data encountered in supervised classification

- Report means and standard deviations, and sample size of clusters
- Compute important variables, eg using random forests
- Dimension reduction using LDA (or PCA), to examine clusters
- Plot using colour for cluster label, using tour, parallel coordinates, scatterplot matrix

# Example

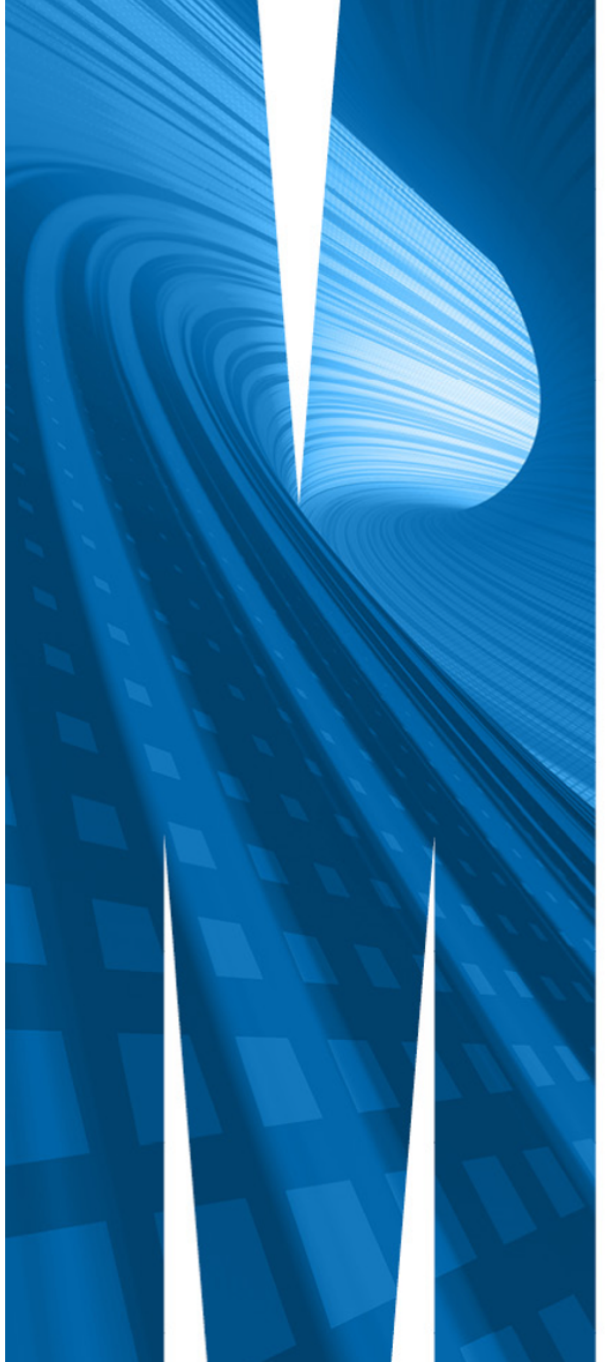
Summary statistics by cluster, in original units.

| cl5 | stat | aede2 | tars1 | aede3 | aede1 | tars2 | head |
|---|---|---|---|---|---|---|---|
| 1 | m | 14.10 | 183.10 | 104.86 | 146.19 | 129.62 | 51.24 |
| 1 | s | 0.89 | 12.14 | 6.18 | 5.63 | 7.16 | 2.23 |
| 1 | n | 21.00 | 21.00 | 21.00 | 21.00 | 21.00 | 21.00 |
| 2 | m | 10.09 | 138.23 | 106.59 | 138.27 | 125.09 | 51.59 |
| 2 | s | 0.97 | 9.34 | 5.85 | 4.14 | 8.55 | 2.84 |
| 2 | n | 22.00 | 22.00 | 22.00 | 22.00 | 22.00 | 22.00 |
| 3 | m | 14.29 | 201.00 | 81.00 | 124.65 | 119.32 | 48.87 |
| 3 | s | 1.10 | 14.90 | 8.93 | 4.62 | 6.65 | 2.35 |
| 3 | n | 31.00 | 31.00 | 31.00 | 31.00 | 31.00 | 31.00 |



MONASH University





Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 10b