MONASH University

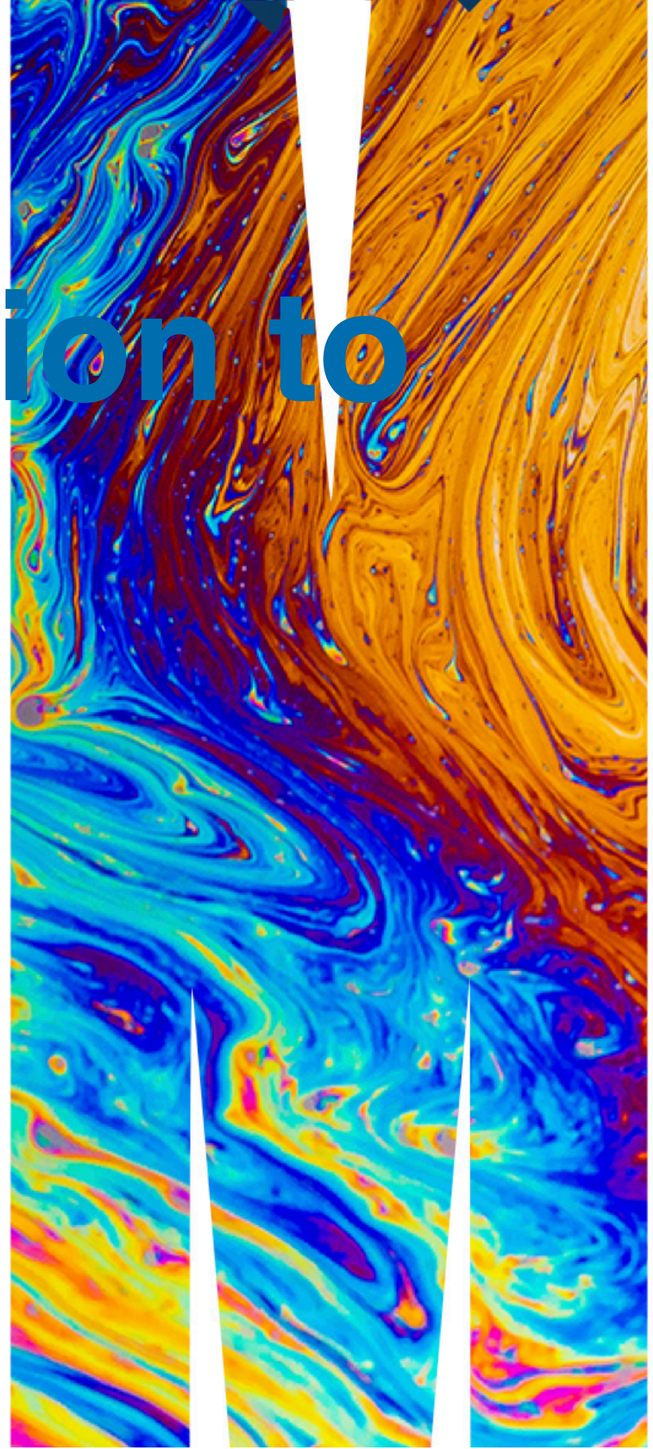
ETC3250/5250 Introduction to Machine Learning

Week 11: Evaluating your clustering model

Professor Di Cook

etc3250.clayton-x@monash.edu

Department of Econometrics and Business Statistics



Overview

We will cover:

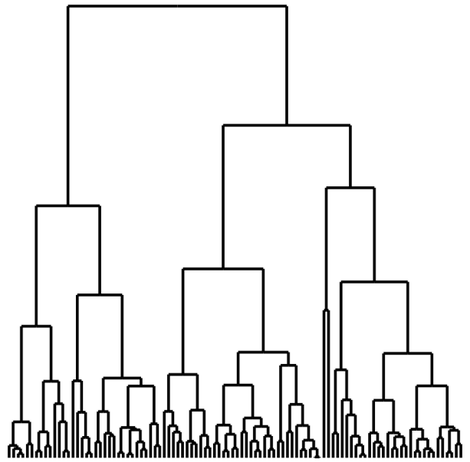
- Confusion tables
- Cluster metrics
- Numerical summaries of solution
- Visual summaries
- Low-dimensional representations

Evaluating your clustering

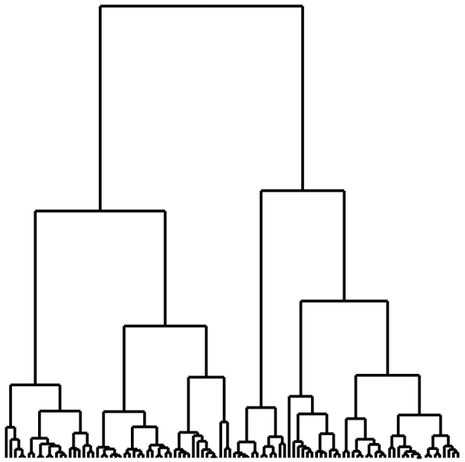
Confusion tables: comparing results (1/3)

- Two clusterings can be compared with a confusion matrix, similarly to supervised classification.
- The main difference though is cluster labels are not consistent between methods. Comparison is better if the labels are matched optimally, as a first step.

Method 1



Method 2



The dendrograms look similar. Are the clusters similar?

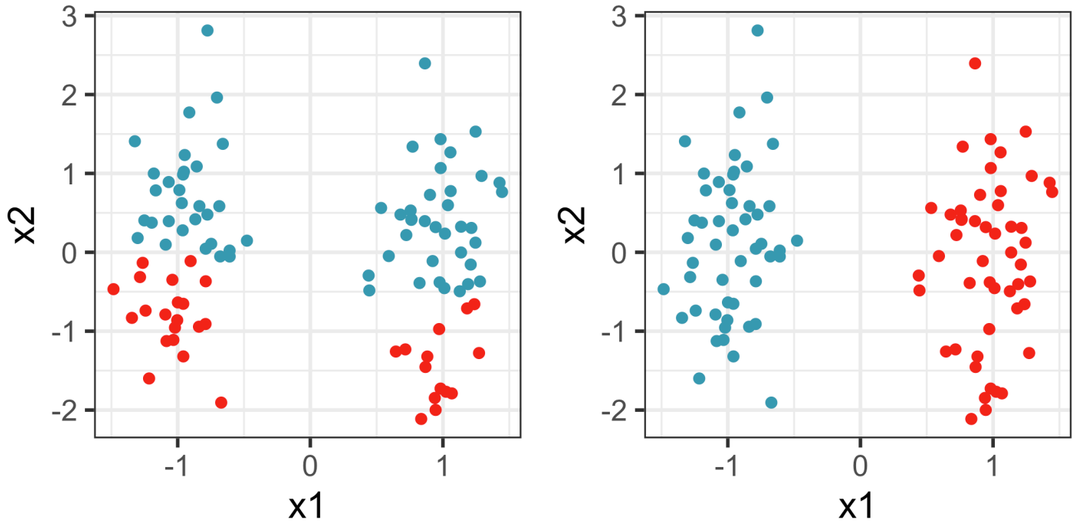
Confusion tables: comparing results (2/3)

```
# A tibble: 2 × 3
  cl1     `1`   `2`
  <fct> <int> <int>
1 1        30    36
2 2        20    14
```

Match labels: What method 1 calls cluster 1, method 2 calls cluster 2. Rearrange the confusion matrix, or re-label one methods clusters, so that large counts are on (top-left to bottom-right) diagonal.

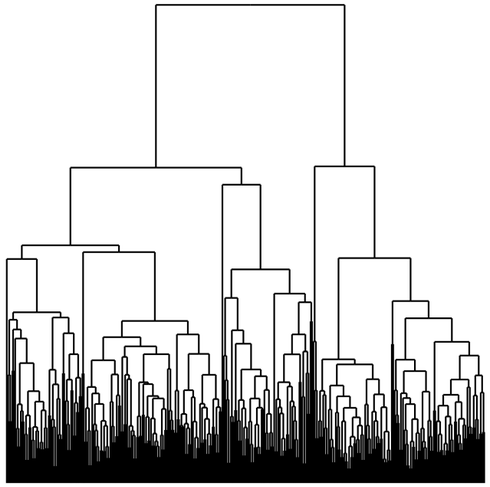
```
# A tibble: 2 × 3
# Groups:   cl1 [2]
  cl1     `1`   `2`
  <fct> <int> <int>
1 2        20    14
2 1        30    36
```

The two clusterings are indeed quite different. If you look at the model in the data space, it is clear.

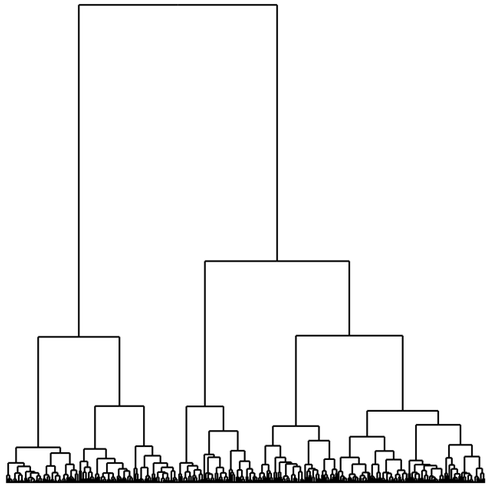


Confusion tables: comparing results (3/3)

Average linkage



Wards linkage



```
# A tibble: 5 × 6
  cl1      `1`   `2`   `5`   `3`   `4`
  <fct> <int> <int> <int> <int> <int>
1 1       102    52     0     0     0
2 2         3     5    56     0     0
3 3         0     0     0    52    67
4 4         0     0     0     0     4
5 5         0     0     1     0     0
```

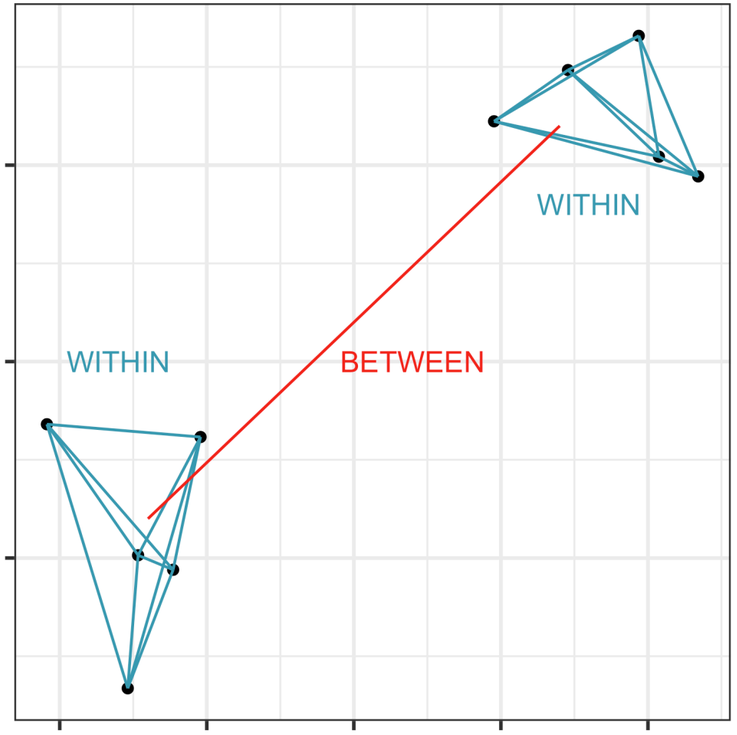
- Method 1: 2 or 5 clusters?
- Method 2: 2, 3, 4 or 5 clusters?

- You might think that average linkage would give similar results to Wards linkage, but results here are very different.
- Average divides the data into three large clusters with two clusters with few observations.
- Wards divides the data into five large clusters, with one having twice the points as the others.

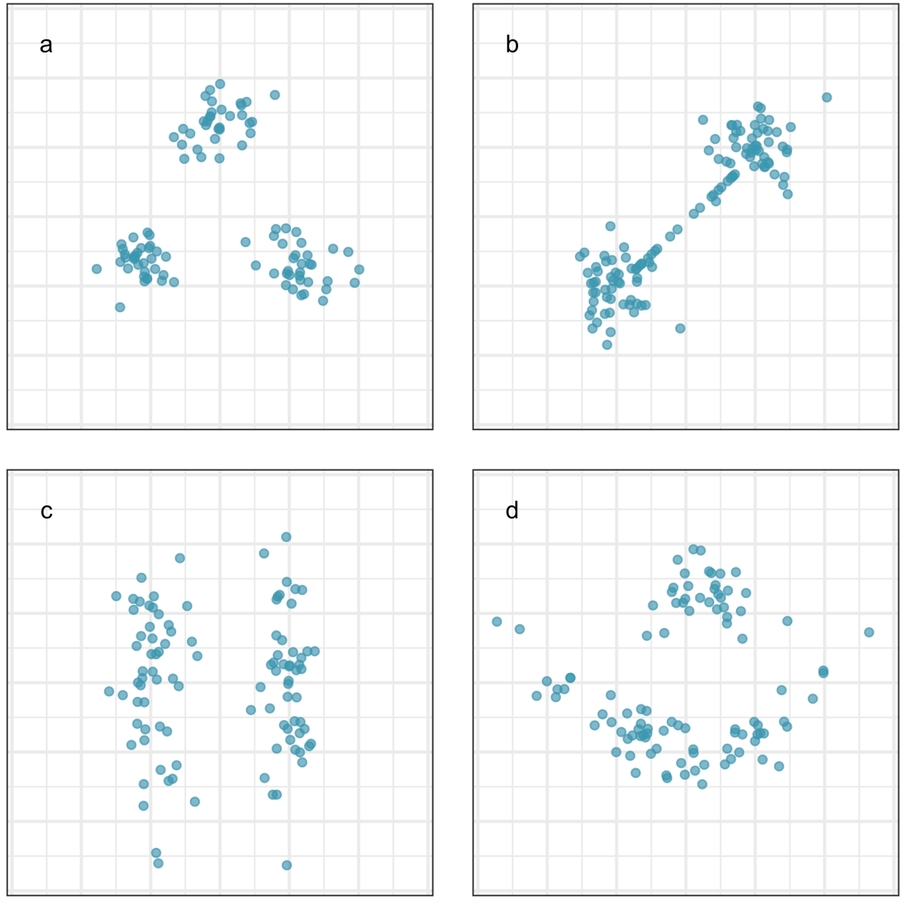
Cluster metrics (1/2)

We covered: within.cluster.ss, WBRatio, Hubert Gamma, Dunn, Calinski-Harabasz Index.

They measure:



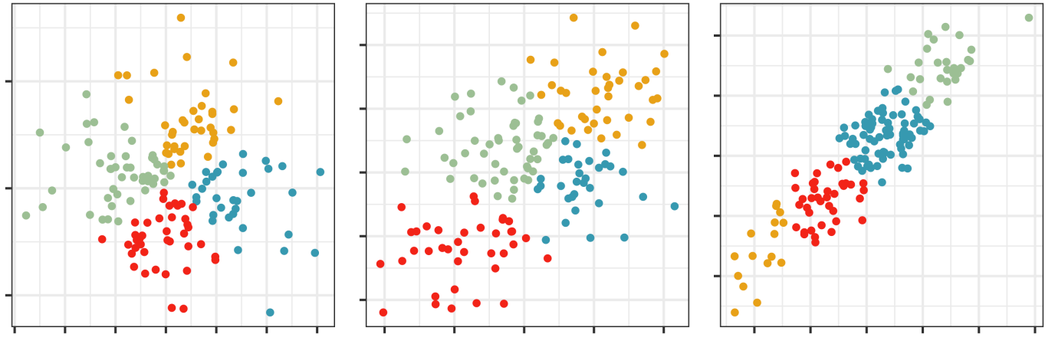
Remember the world has many shapes:



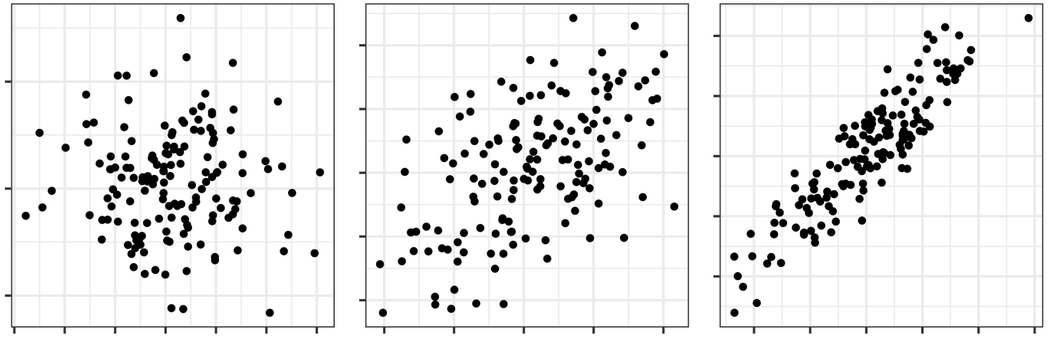
In an idealistic setting, these are all perfect metrics.

Cluster metrics (2/2)

Clustering algorithms not primarily looking for gaps. Gaps can be local or global in size, or just not exist. Clustering that partitions data can still be useful. Know how an algorithm will partition.



What would happen to these data sets?



The shape (strictly correlation here) strongly affects how an algorithm divides or partitions a blob.

Numerical summaries

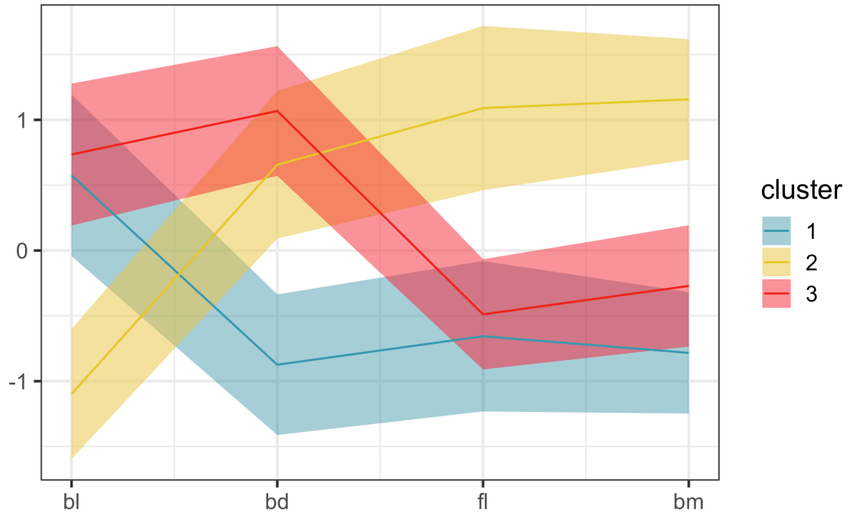
- Means
- Standard deviations
- Variances and covariances: but unwieldy to print all
- Cluster size

		Clusters		
Var	**Stat**	**1**	**2**	**3**
bl	mean	-0.87	0.66	1.07
	sd	*0.54*	*0.56*	*0.50*
bd	mean	0.58	-1.10	0.73
	sd	*0.62*	*0.50*	*0.54*
fl	mean	-0.78	1.16	-0.27
	sd	*0.46*	*0.46*	*0.46*
bm	mean	-0.66	1.09	-0.49
	sd	*0.58*	*0.63*	*0.42*
	n	162.00	123.00	57.00

Visual summaries

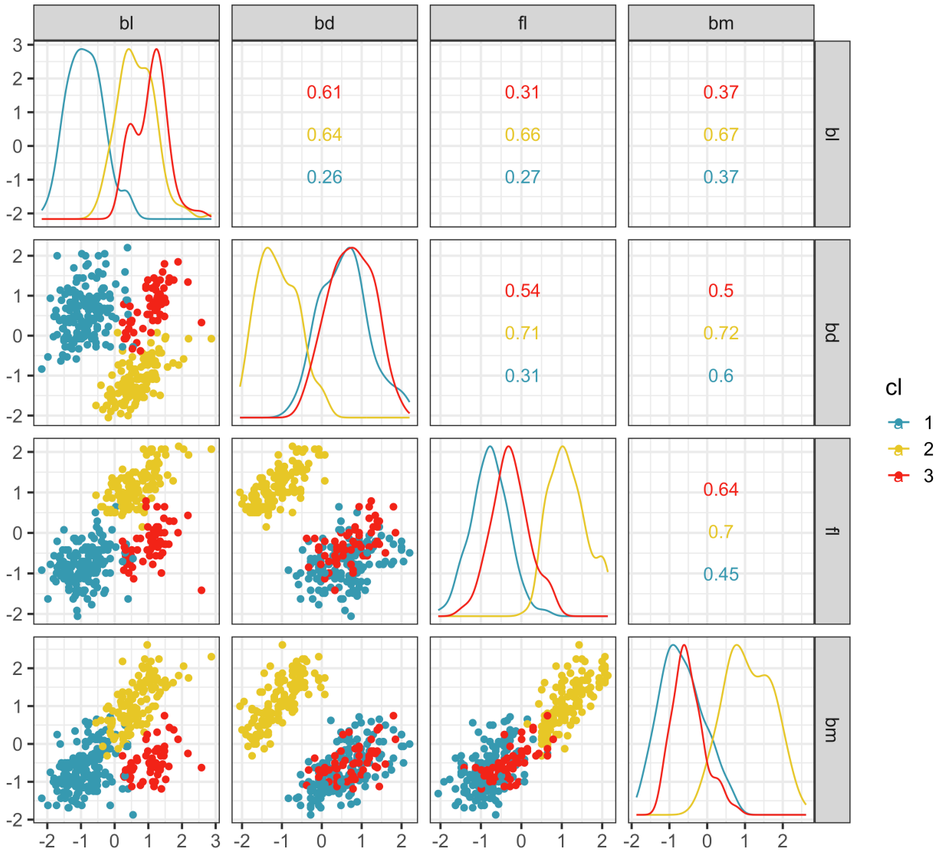
- Statistics: Show the mean and variances for each cluster, as a parallel coordinate plot.
- Model in the data space:
  - Colour observations by their cluster label, and use multivariate plots.
  - Dendrograms: Draw the connections of points being joined to form clusters as a high-d graph.
  - Ellipses: As done in summarising model-based fit.
  - Net: As done for SOM.
- Convex hulls: Bounding shapes of each cluster, can be done in high-D, too.

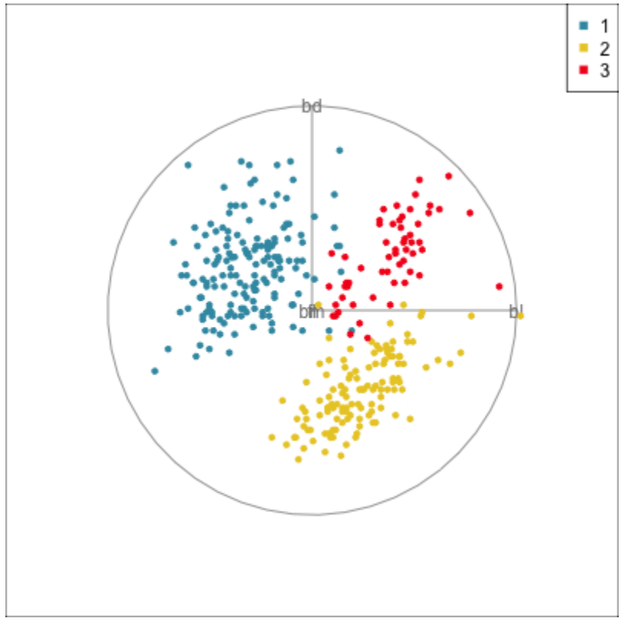
Statistics



- Cluster 1 has high values of `bl` and low values of `bd`, `fl` and `bm`.
- Cluster 2 has low values of `bl`, and high values on `bd`, `fl` and `bm`.
- Cluster 3 has high values of `bl`, `bd` and low values of `fl` and `bm`.

Model-in-the-data-space (1/3)





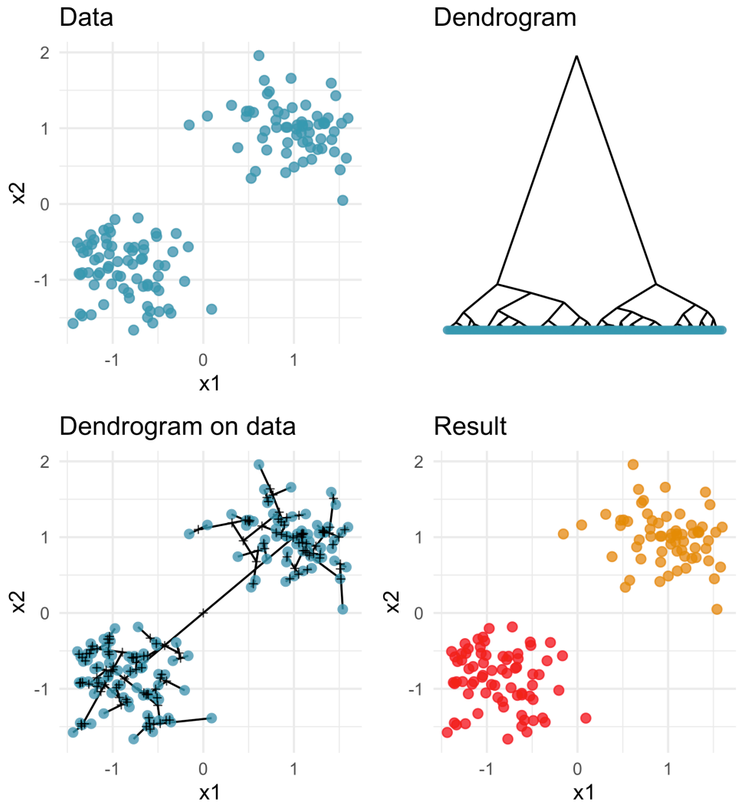
Cluster 2 is neatly separated. Clusters 1 and 3 are distinct but there is no gap.

The clusters are neatly distinguished in different pairs of variables.

Model-in-the-data-space (2/3)

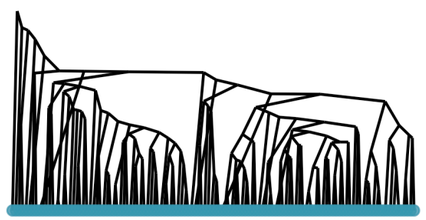
The dendrogram can be drawn in the data space.

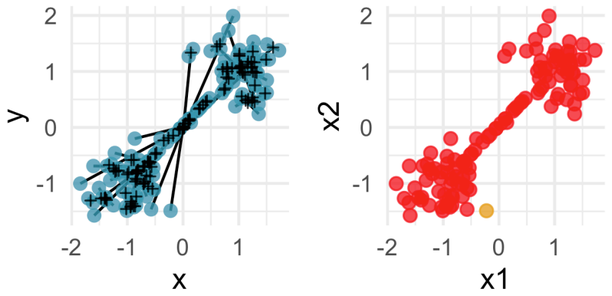
- Add extra points at average between clusters indicating joins.
- Create variable specifying point type, “data” or “node”.
- Add edges data, specifying which points are connected to make dendrogram.



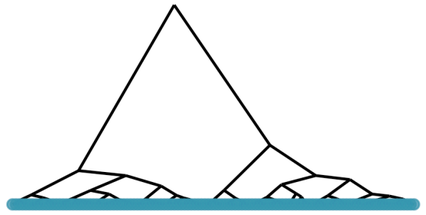
Why?

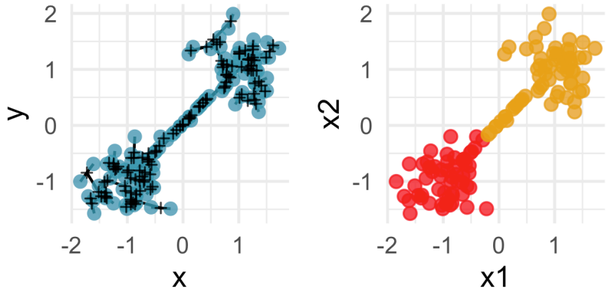
Single linkage





Ward's linkage



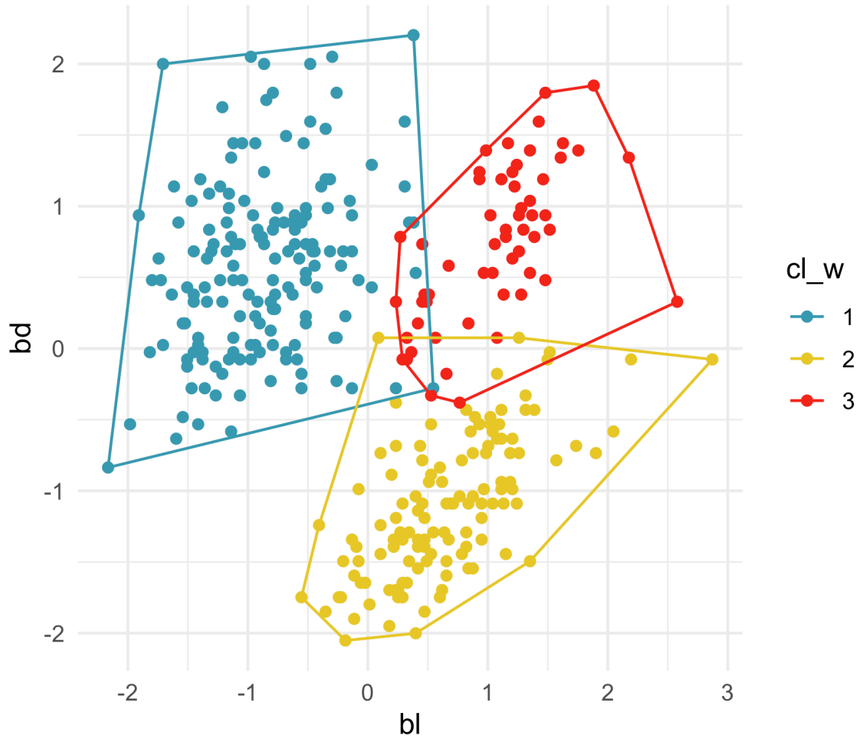


Because it shows us how the algorithm is sequentially joining the observations, relative to their values.

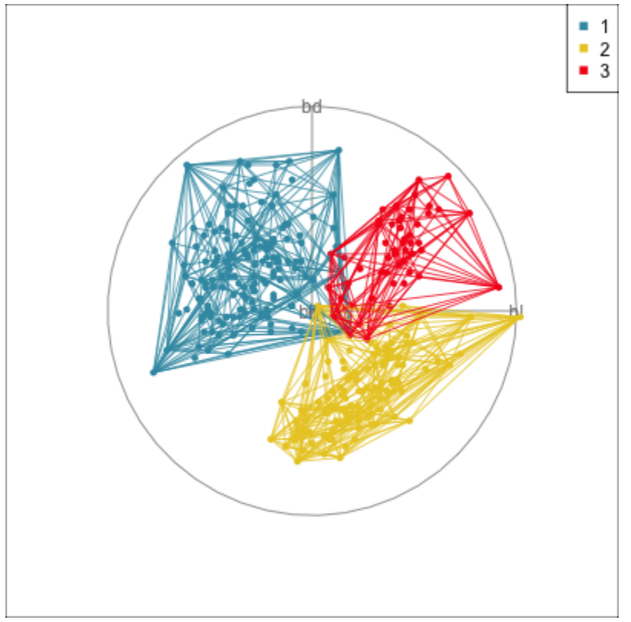
Model-in-the-data-space (3/3)

Convex hulls

Convex hulls are a common method for visually displaying the region in the data space corresponding to each cluster. It avoids making any assumption about the shapes.



Convex hulls are defined in p dimensions also.



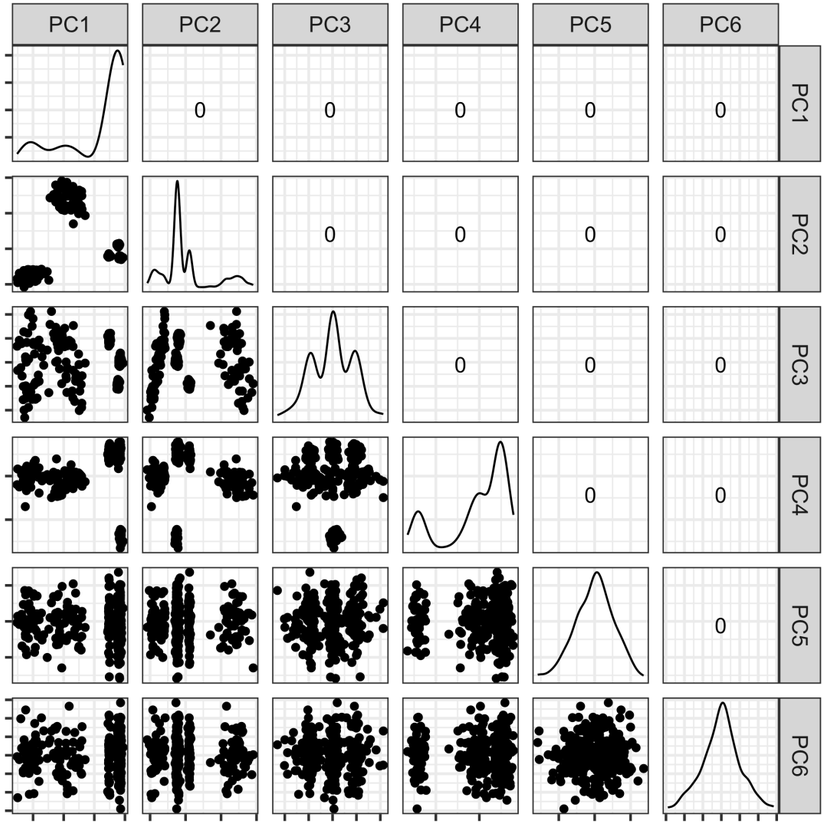
Are clusters 1 and 3 separated/distinct?

Low-dimensional representations

- PCA, but not PCA
- Perhaps, projection pursuit
- Nonlinear dimension reduction
- Self-organising map
- When you have a clustering, linear discriminant analysis

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     2.802 1.720 0.4204 0.3470 0.05296 0.04880
Proportion of Variance 0.707 0.266 0.0159 0.0108 0.00025 0.00021
Cumulative Proportion  0.707 0.973 0.9887 0.9995 0.99979 1.00000
```

Proportion of total variance is 0.97 with 2 PCs. But if you only used 2 PCs, you would miss the clusters visible only with PC4.



How do you know you got it right?

All “models” are wrong by some are useful.

- Are the subsets useful, eg if you market differently to the different subsets do you get more purchases than marketing the same way to all.
- Can you reduce the number of parameters needed in a high-dimensional regression model, by grouping genomes into a smaller set?
- Does it match the shape of the data?
- Bootstrap or CV
- Domain knowledge and interpretability

Wrapping up

What we have covered

- Computational and statistical models for a categorical response: logistic, discriminant, trees, forests, support vectors, multilayer perceptron neural network
- Discovering class labels by clustering: k-means, hierarchical, model-based, self-organising
- Dimension reduction: princi**pal** components, multidimensional scaling t-SNE, UMAP
- Visualising your high-dimensional data and models

Not covered

- *Continuous, numerical response*: ⟶ ETC2420/5242, ETC3580/5580
  - Fit statistics easier: MSE, AIC, R^2, but not all applicable for categorical response
  - Visualisation a little harder
- *Full range of neural networks*: convolutional, recurrent, autoencoder, generative adversarial, recursive, … ⟶ ETC5555

Philosophy

Some technicians brag that they can pull an appliance into parts. That is not as impressive as the technician who can put it back together, and make it work. I would hire this person in a heartbeat. *~Raymond Bruce Cook*

Preparing for the exam

The exam is like the quizzes, tutorial and assignment exercises all together, without the programming elements.

It is open. You can use whatever resources you like during the exam, including AI.

It is supervised. You can only take one device.

Open exams need solid preparation.

Ways to prepare:

- *Review your quizzes*.
- *Review the tutorial and assignment materials*: what results mean, what problems are detected in the model, conceptual thinking, and explanations of results and diagnostics.
- Resources provided have several books with exercises at the end of chapters. Work your way through some where you feel you don't know the topic well enough.
- *Form a small study group*: make up questions for each other to tackle, compare and contrast your answers.
- *Try asking GAI questions*, and evaluate the answers as useful or misguided based on your learning.

Next: What the ETC5250 students have learned from the kaggle challenge