$Y = f(X) + \varepsilon$

Unsupervised has no target variable

Unsupervised Learning

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk and Actuarial Applications

- Can we summarise or find relationships in the raw data X?

UNSW SYDNEY

Overview

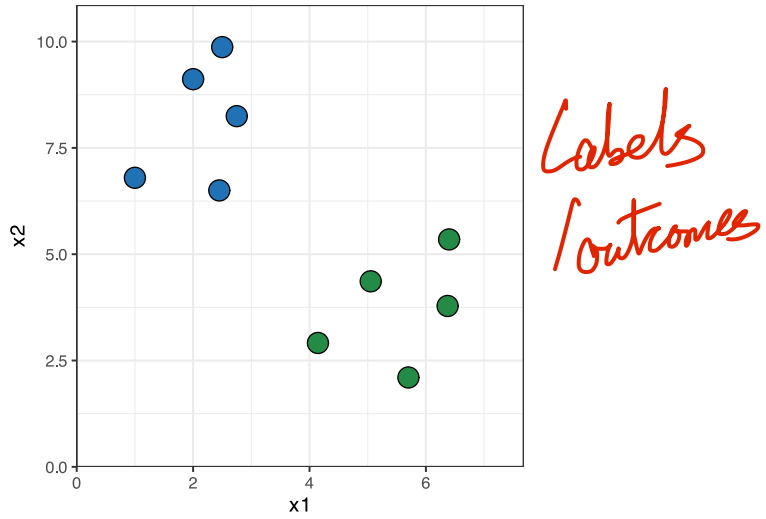
- Challenge of Unsupervised Learning
- K-means clustering
- Hierarchical clustering
- Dimension Reduction (PCA)

Grouping

– Dimension reduction (summarise)

Supervised vs Unsupervised Learning

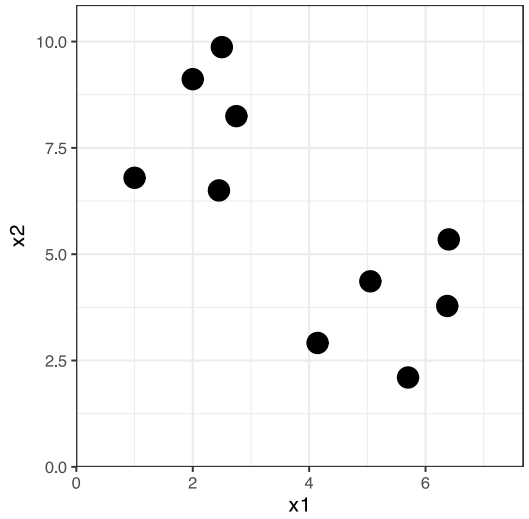
Supervised



Labels
Outcomes

- Data: $X_1, X_2, \ldots, X_p, Y$
- Goal: Predict Y using $X_1, X_2, \ldots, X_p$

Unsupervised



- Data: $X_1, X_2, \ldots, X_p$
- Goal: Discover interesting things using $X_1, X_2, \ldots, X_p$

UNSW SYDNEY

Challenge of Unsupervised Learning

- Typical questions
  - Is there an informative way to visualize the data?
  - Can we discover subgroups among the variables?
- More subjective than supervised learning
  - no simple goal for the analysis
- Hard to assess the results obtained from unsupervised learning methods
  - no universally accepted mechanism for performing cross-validation or validating results on an independent data set

– Even methods today are not conclusive

Today is an exercise in seeing "what sticks"

Clustering vs. PCA

- Both seek to simplify the data via a small number of summaries
- Different mechanisms
  - Clustering: find homogeneous subgroups among the observations
  - PCA: find a low-dimensional representation of the observations that explain a good fraction of the variance
- Both useful for visualisation

• Both are not conclusive -

Clustering Methods

- A very broad set of techniques for finding subgroups, or clusters, in a data set
- The observations within each group are quite similar to each other
- Need to specify what it means for two or more observations to be similar or different
  - often a domain-specific consideration
- Two Clustering Methods
  - K-means clustering
    - partition the observations into a pre-specified number of clusters
  - Hierarchical clustering
    - do not know in advance how many clusters we want
    - dendrogram, a tree-like visual representation of the observations

Applications of Clustering

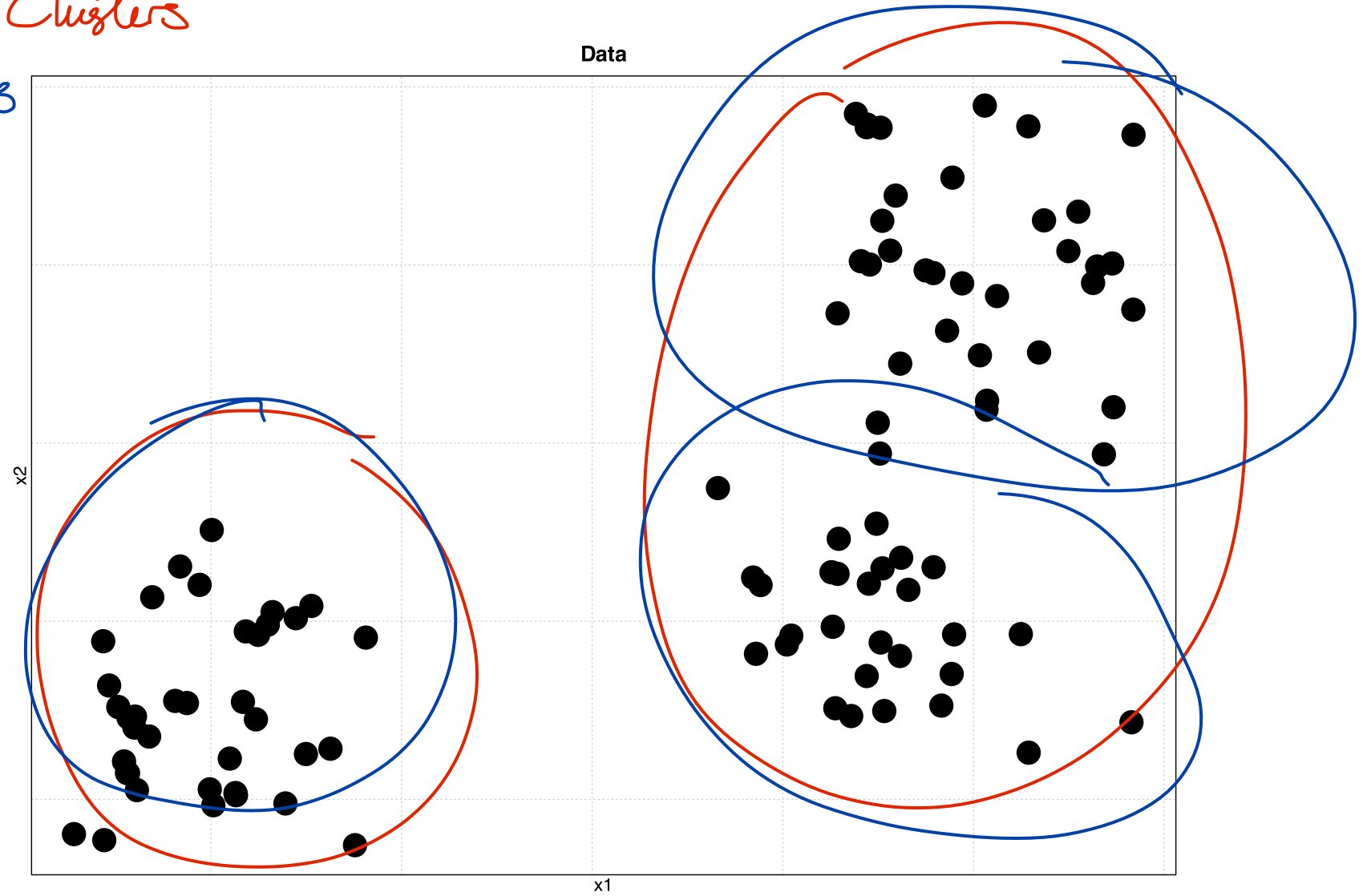
- Market segmentation
- Fraud detection
- Group patients by medical condition (e.g Types of Diabetes)
- Clustering of documents by type
- Compression of information – representative policies

K-Means Clustering

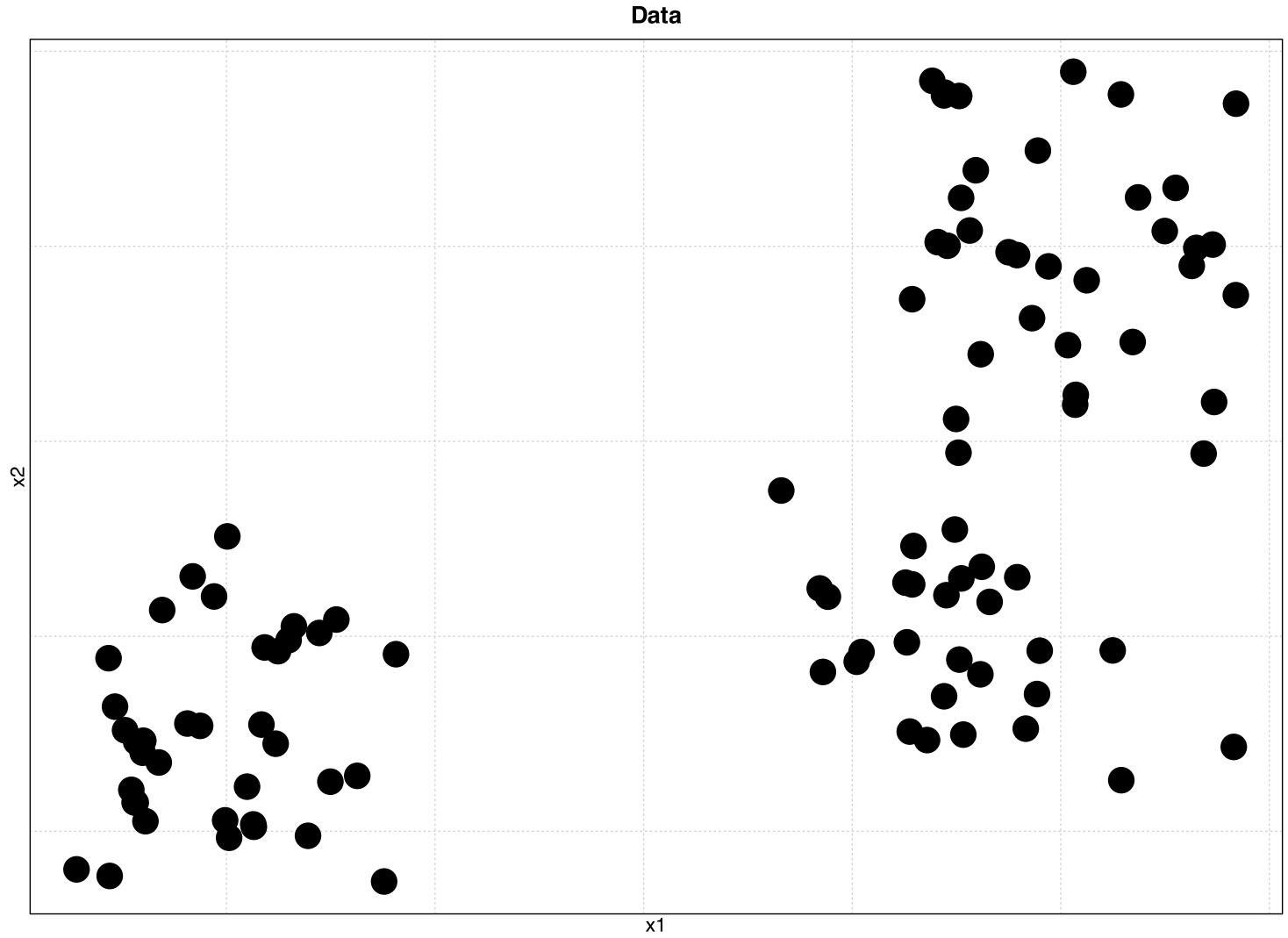
K-means clustering: Demonstration I

2 - Clusters

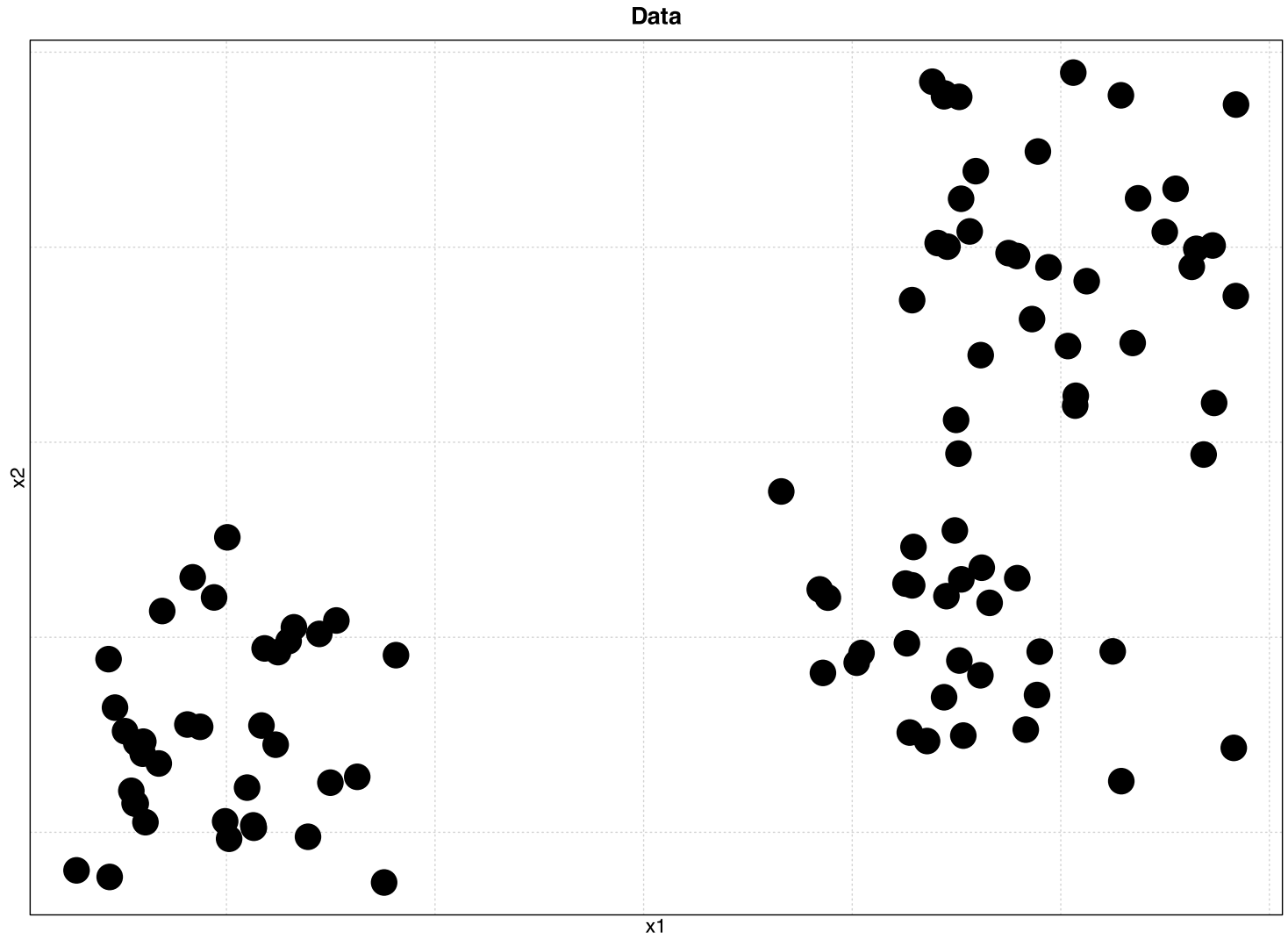
3 - Clusters



K-means clustering: Demonstration II



K-means clustering: Demonstration III



K-Means Clustering

$C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster

- Each observation belongs to at least one of the K clusters

$$C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$$

Outliers will be forced into a cluster

- The clusters are non-overlapping, or no observation belongs to more than one cluster

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

UNSW SYDNEY

Mathematical formulation: Clustering

A good clustering is one for which the within-cluster variation is as small as possible

Mathematically:

Distance of each obs to each other in cluster K.

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

The most common choice of $W(\cdot)$

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

– Average of squared distance

where $|C_k|$ is the number of observations in the kth cluster

UNSW SYDNEY

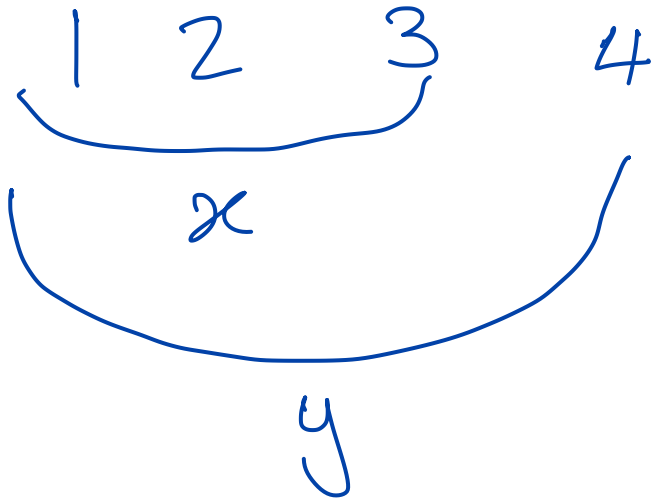
K-Means Clustering

The optimisation problem that defines K-means clustering

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

- A difficult problem to solve precisely
- Exist a very simple algorithm that provides a local optimum

Means we rerun multiple times and take the min.



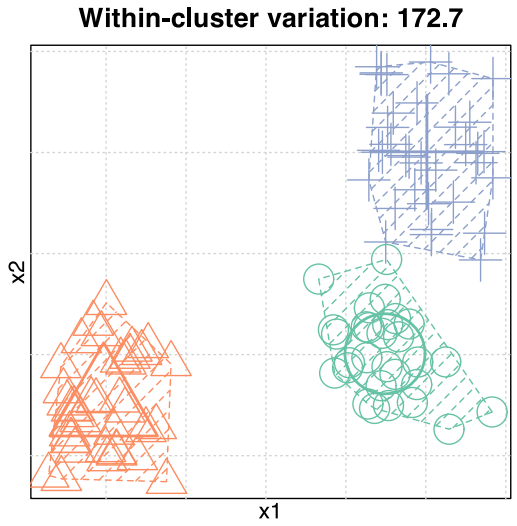
$y \leq x$

K-means clustering Algorithm – To find local min.

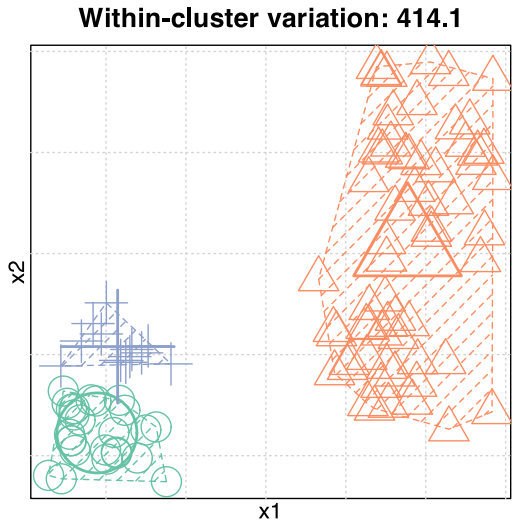
1. Randomly initialise K cluster centres/centroids
2. Assign each observation to the cluster whose centroid is closest
   - closest defined using Euclidean distance
3. For each of the K clusters, compute the cluster centroid
   - centroid is the vector of the means for the observations in the kth cluster
4. Repeat 2 & 3 until convergence

UNSW SYDNEY

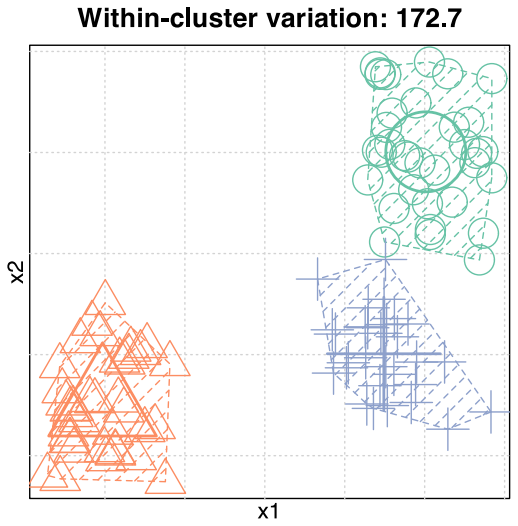
K-means clustering: Local Optima



Clusters after a seed of 1



Clusters after a seed of 2



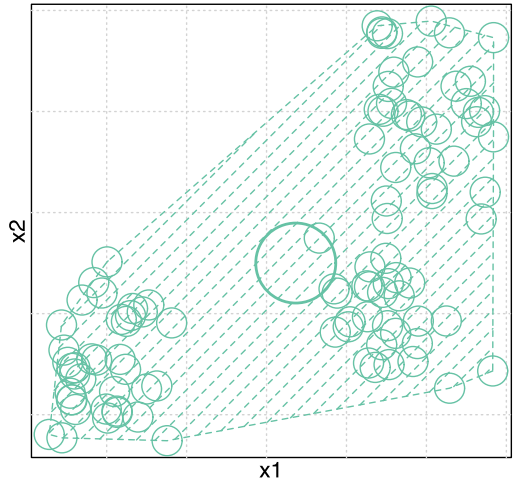
Clusters after a seed of 9

- The algorithm finds a local rather than a global optimum
- Results depend on initial centroids used
- Important to run the algorithm multiple times and select the best solution (minimum within-cluster variation)

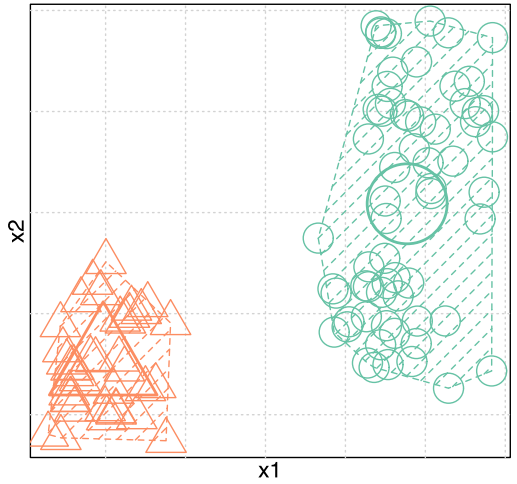
UNSW SYDNEY

n clusters ⇒ Each point is its own cluster

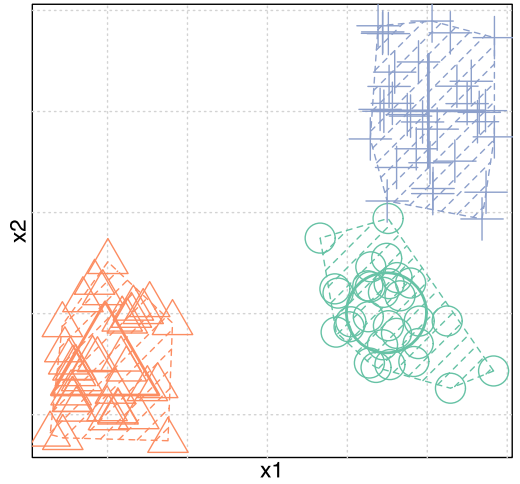
Within-cluster variation: 2022



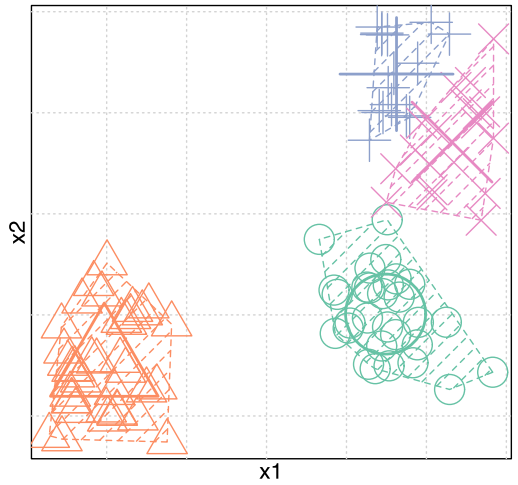
Within-cluster variation: 440



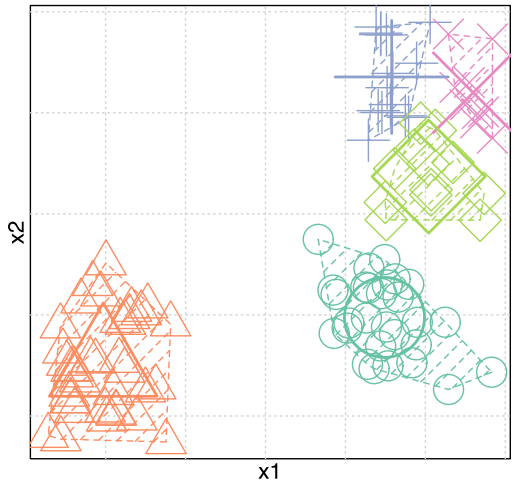
Within-cluster variation: 172.7



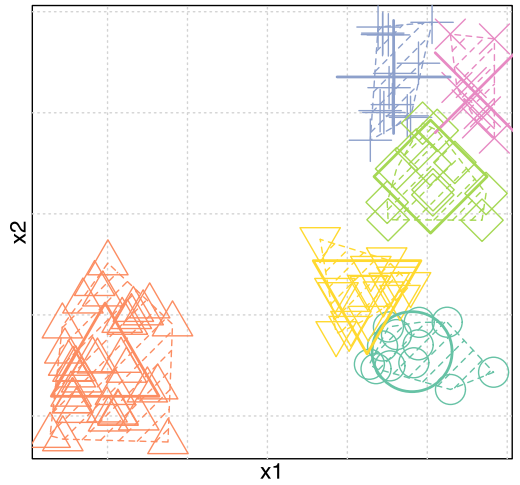
Within-cluster variation: 141.4



Within-cluster variation: 124.1



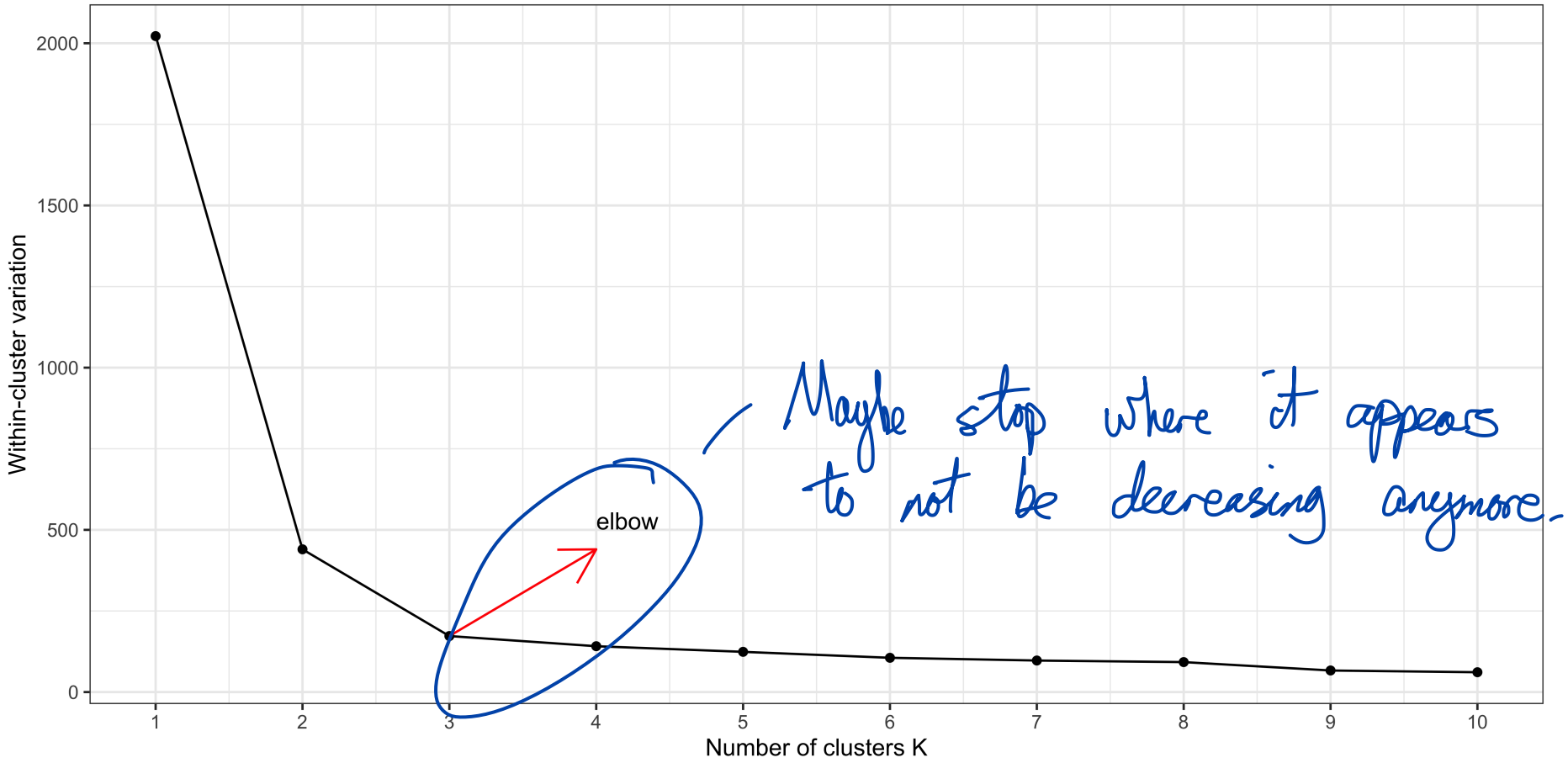
Within-cluster variation: 105.7



UNSW SYDNEY

What is the right value of K?

– Adhoc



Hierarchical Clustering

Hierarchical Clustering

— for a given linkage and distance measure the tree is unique.

- No need to specify the number of clusters K
- Result is a tree-based representation, called a dendrogram
- Allows user to choose any distance metric
  - K-means restricted us to Euclidean distance
- Focus on bottom-up or agglomerative clustering
  - start from the leaves
  - combine the clusters up to the trunk

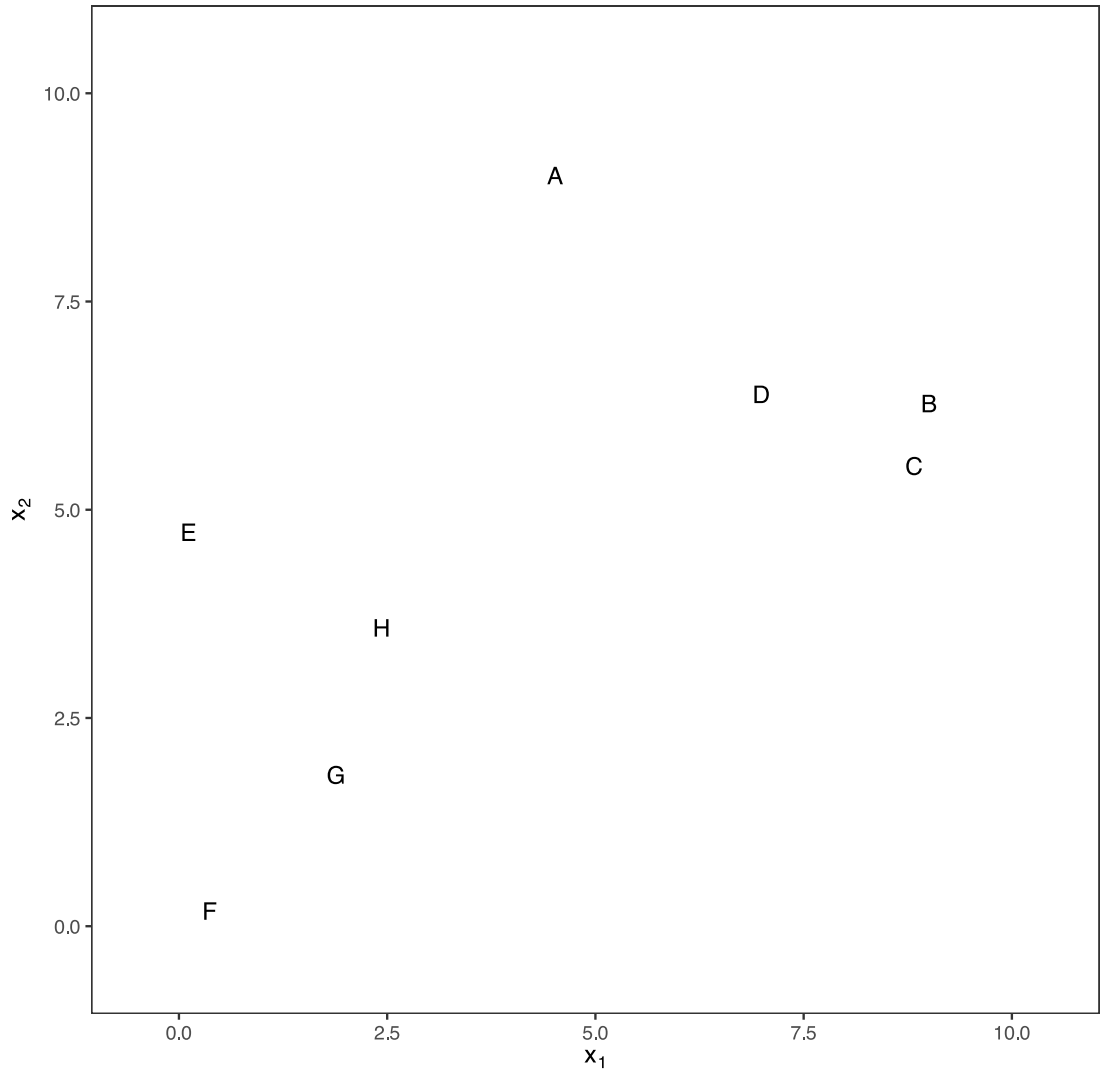
Algorithm:

1. Treat each of the n observations as its own cluster
2. For $i = n, n-1, \ldots, 2$:
   1. Compute the pairwise inter-cluster dissimilarities among the i clusters
   2. Identify the pair of clusters that are least dissimilar and merge them

Linkage / distance measure.

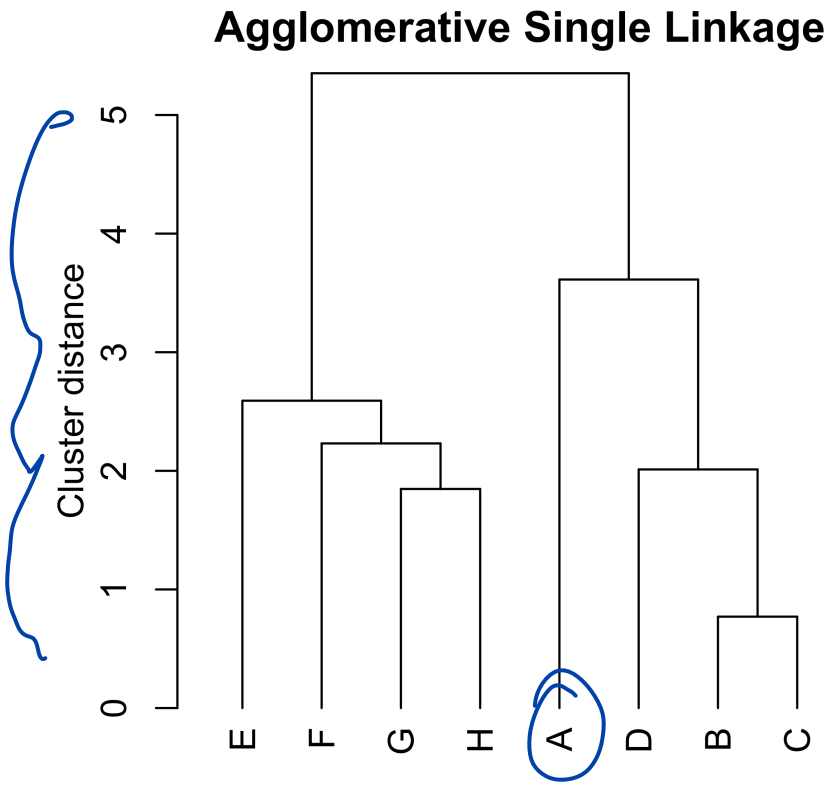
Merge one cluster at a time.

Hierarchical Agglomerative Clustering



Hierarchical Clustering: The dendogram

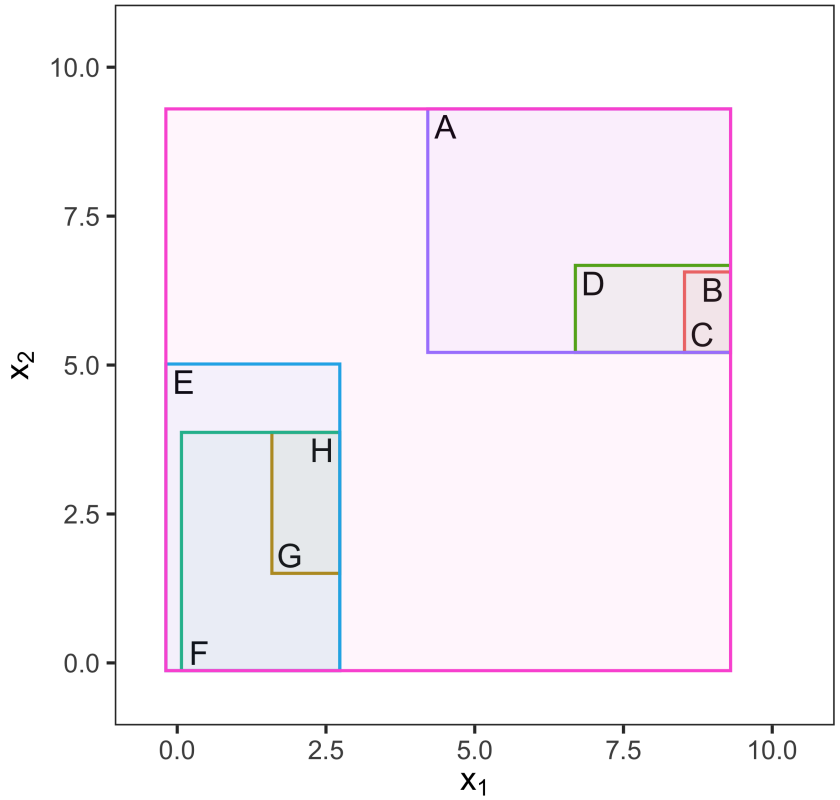
A is as different to B+C as it is to D.



How dissimilar each cluster is.

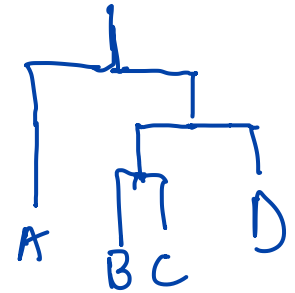
The x-axis in dendograms must be read with the y-axis.

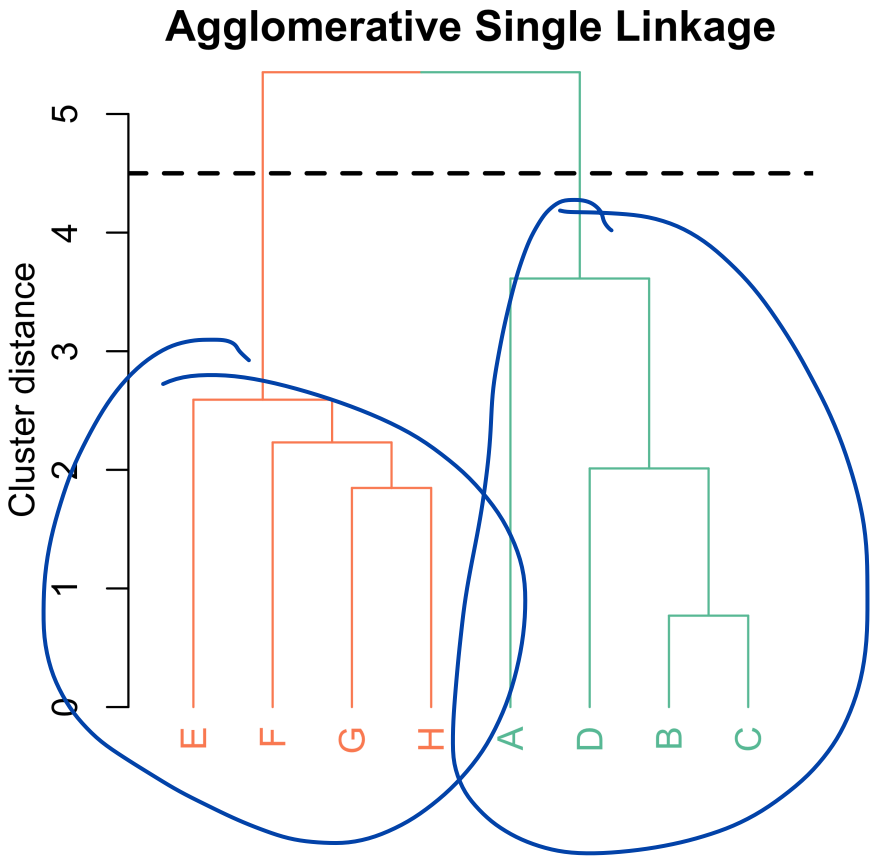
Dendogram



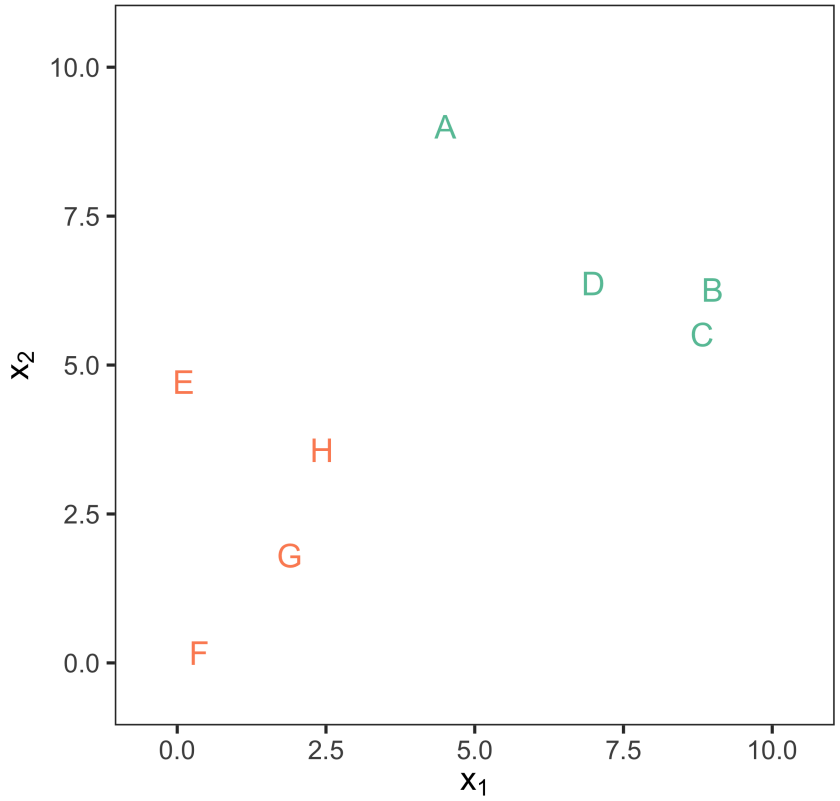
Clusters

Choose a max distance I



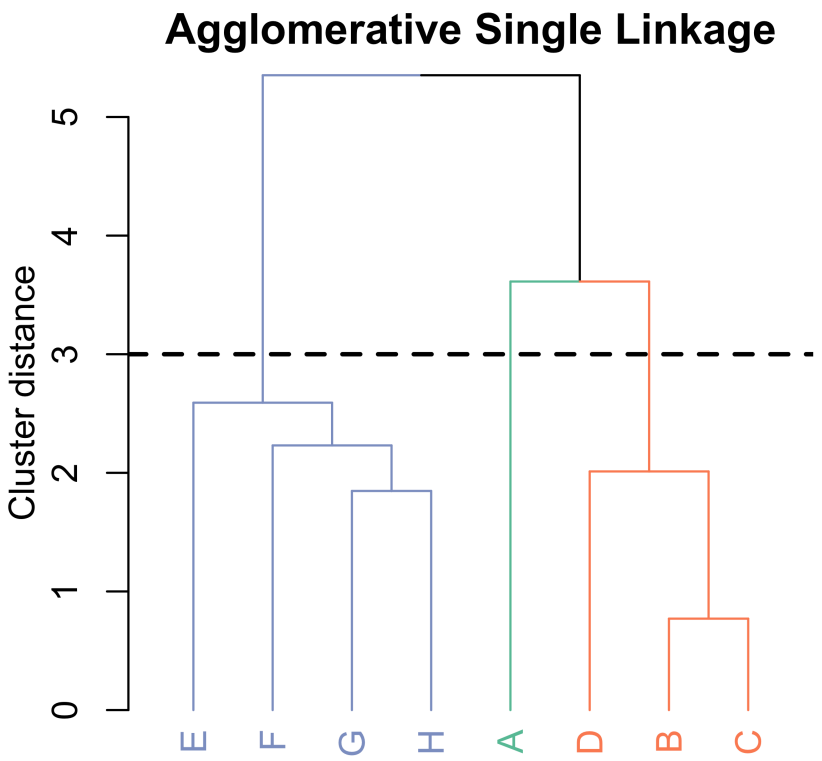


Cut at K = 2

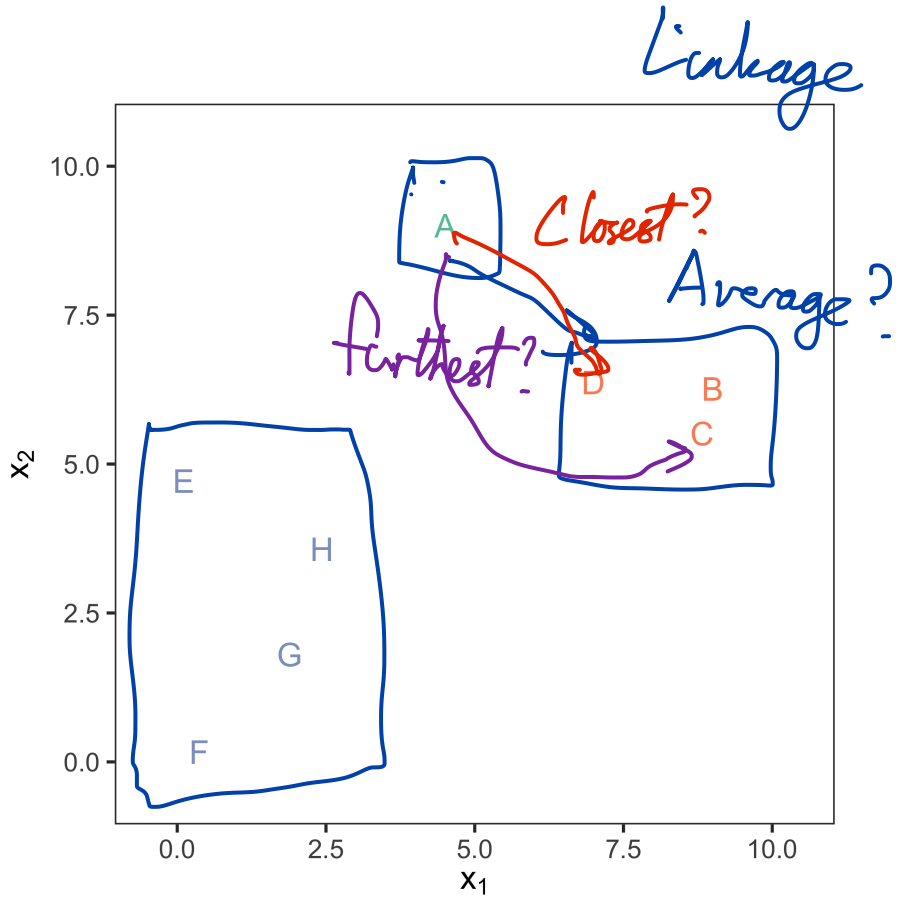


K = 2 clusters

Choose a max distance II

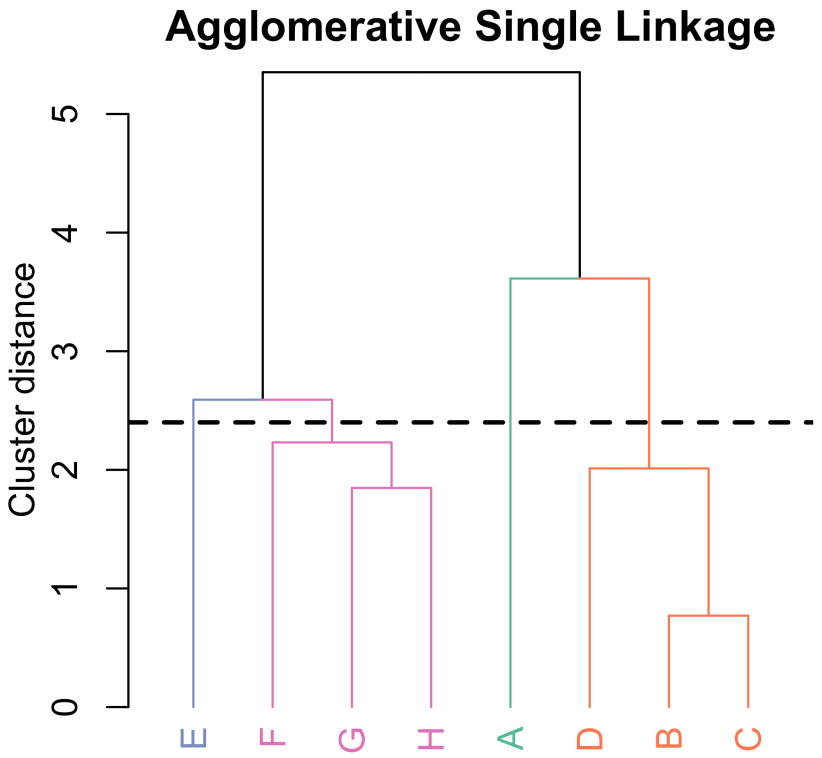


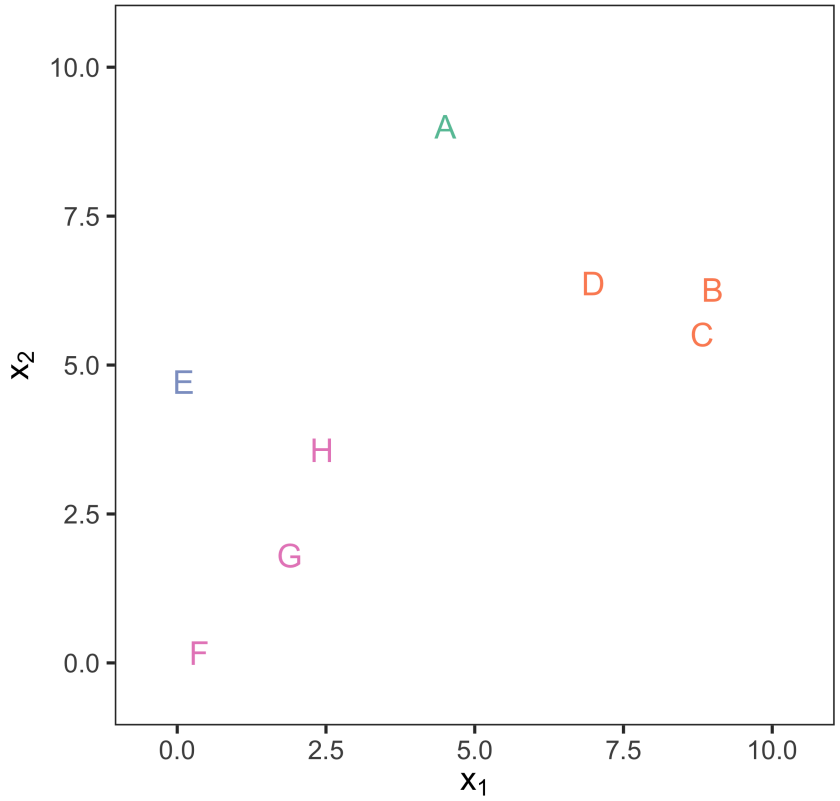
Cut at K = 3



K = 3 clusters

Choose a max distance III





Cut at K = 4

K = 4 clusters

Choice of Dissimilarity Measure — Within each cluster.

- Euclidean distance

$$\sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2}$$

- Simple matching

$$\frac{1}{p}\sum_{j=1}^{p} I(x_{ij} \neq x_{i'j})$$

- Manhattan distance

$$\sum_{j=1}^{p} |x_{ij} - x_{i'j}|$$

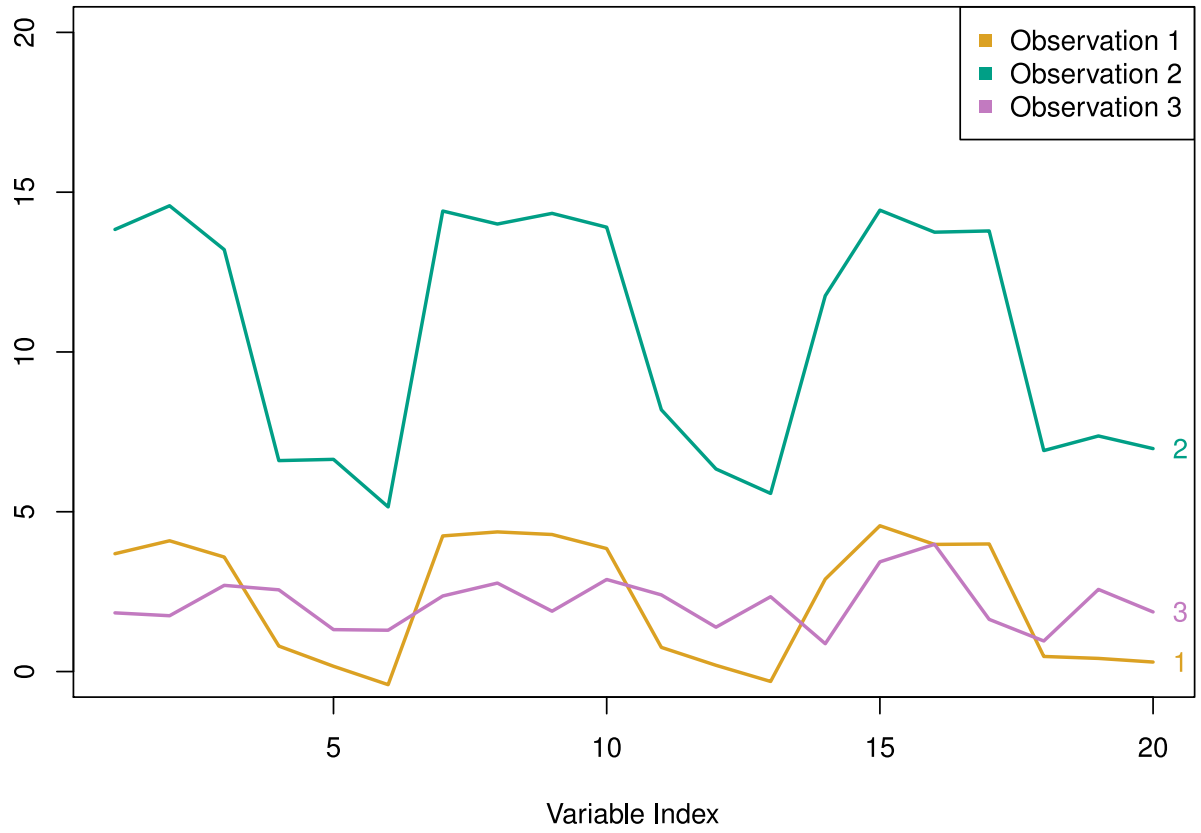
- Combination of numerical and categorical?

— Dummy coding

Note that we need to consider how to compare groups as well.

UNSW SYDNEY

Distance between time series



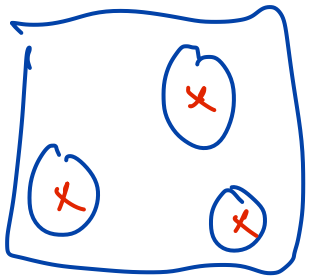
Correlation distance instead of Euclid distance might be better.

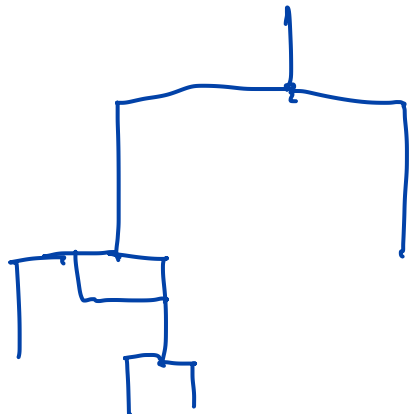
Observations 1 and 3 have a small Euclidean distance between them, but are very weakly correlated. Observations 1 and 2 have a large Euclidean distance between them, but a small correlation-based distance between them.

Source: James et al. (2021), An Introduction to Statistical Learning, Figure 12.15.

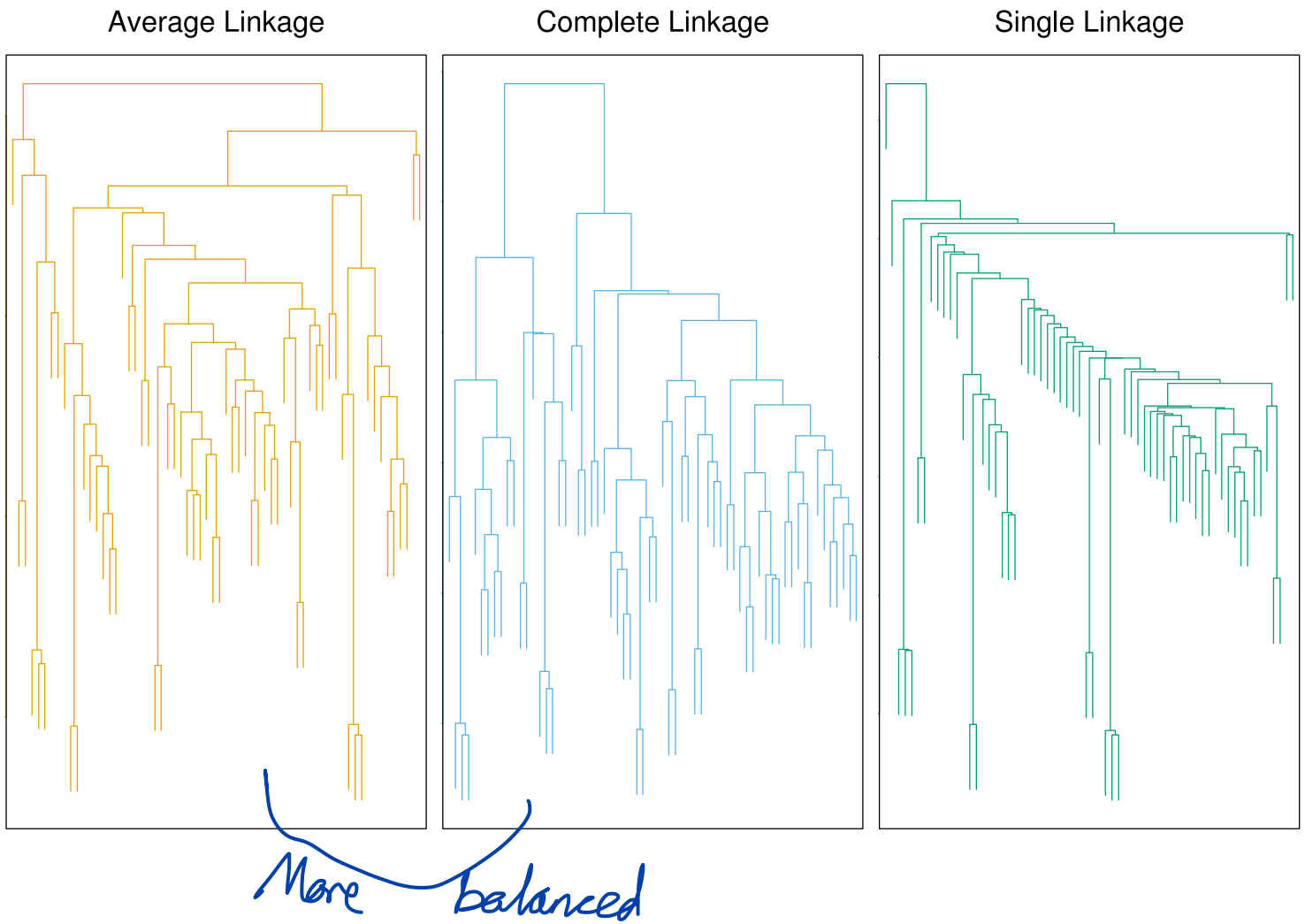
Distance between clusters (linkage)

- Complete
  - maximal inter-cluster dissimilarity
  - compute all pairwise dissimilarities between clusters A and B and take **largest**.
- Single
  - minimal inter-cluster dissimilarity
  - compute all pairwise dissimilarities between clusters A and B and take **smallest**.
- Average
  - mean inter-cluster dissimilarity
  - compute all pairwise dissimilarities between clusters A and B and take **average**.
- Centroid
  - dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B
  - an inversion can occur





Same Data, Different Linkage



Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.

Source: James et al. (2021), An Introduction to Statistical Learning, Figure 12.14.

Practical Issues

- Should the observations / features be standardised in some way?
- Hierarchical clustering
  - dissimilarity measure?
  - type of linkage? — Balanced or less balanced?
  - where to cut the dendrogram?
- K-means clustering
  - how many clusters?
- Validate the clusters obtained
  - does the clusters represent true subgroups in the data?
- Robustness
  - Don't rely on one single answer
  - Try different assumptions/data and check consistency of message

UNSW SYDNEY

Dimension Reduction

Can you memorise these in 30 secs?

112358132134

248163264128

203048154248

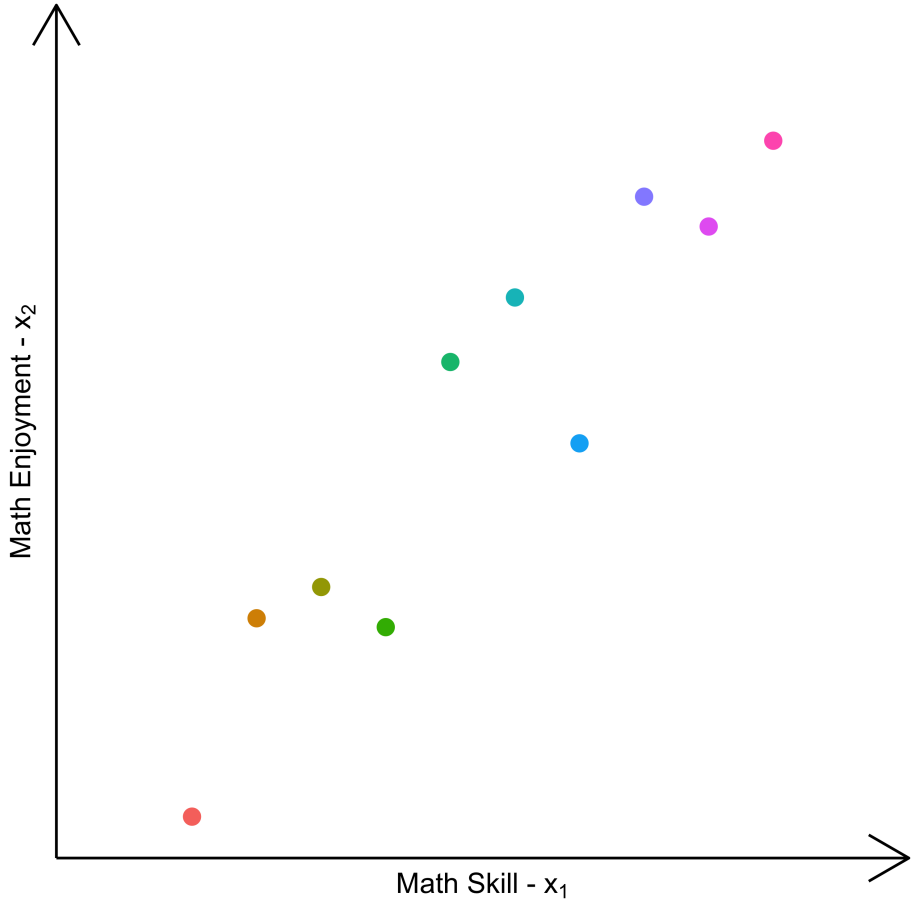
Principal Components Analysis

- Produce derived variables for supervised learning
  - of smaller size than the original data set (i.e. dimension reduction)
  - explain most of the variability in the original set
  - mutually uncorrelated
- A tool for data visualisation

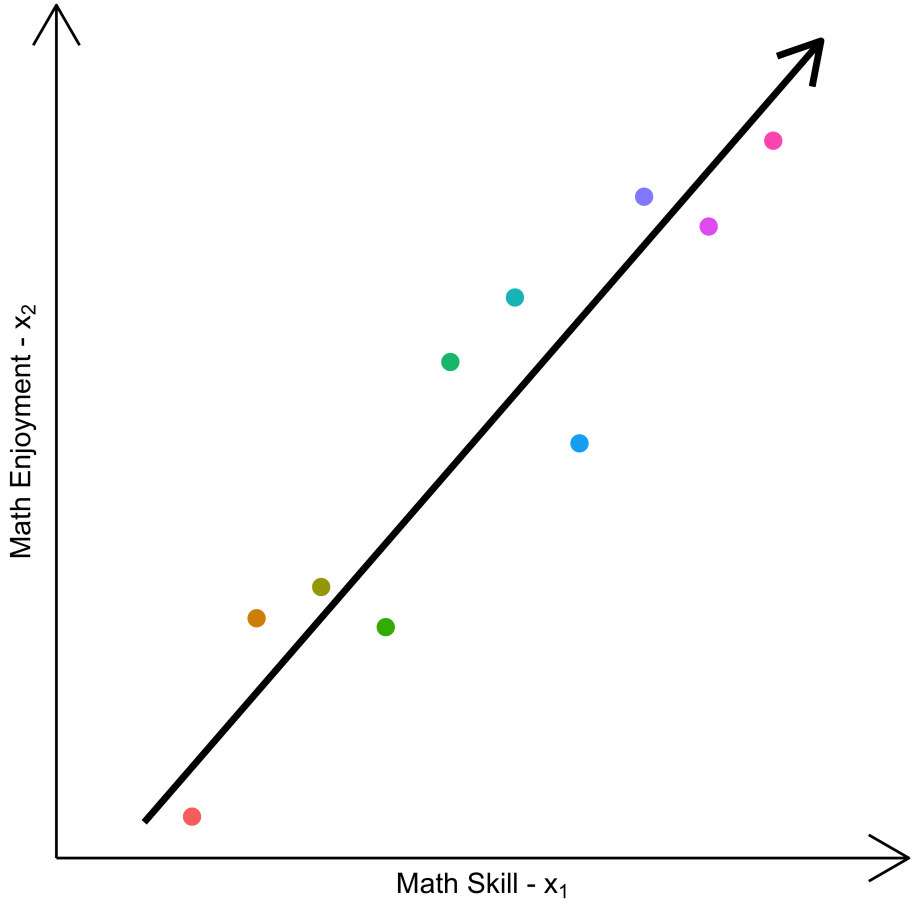
> "… our brains are sort of bad at looking at columns of numbers, but absolutely ace at locating patterns and information in a two-dimensional field of vision" Jordan Ellenberg

Source: Jordan Ellenberg (2014), How Not to Be Wrong: The Power of Mathematical Thinking

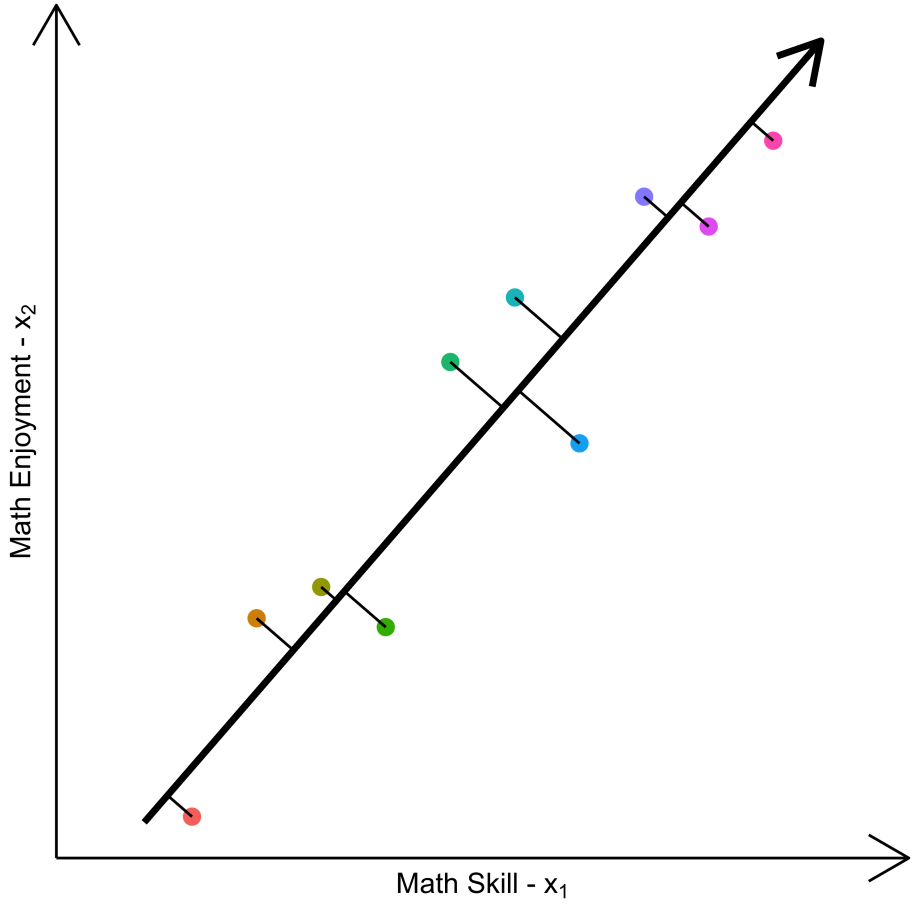
PCA Motivation: Data compression I



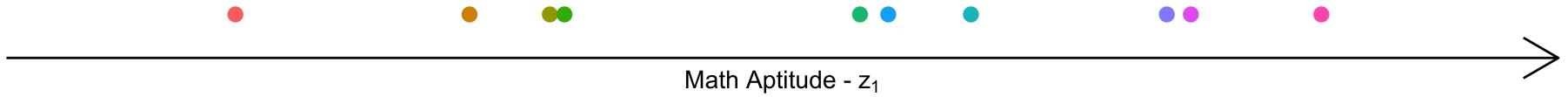
PCA Motivation: Data compression II



PCA Motivation: Data compression III



PCA Motivation: Data compression IV



PCA Motivation: Data compression V

Reduce data from 2D to 1D

$$x^{(1)} \in \mathcal{R}^2 \rightarrow z^{(1)} \in \mathcal{R}$$

$$x^{(2)} \in \mathcal{R}^2 \rightarrow z^{(2)} \in \mathcal{R}$$

$$\vdots$$

$$x^{(n)} \in \mathcal{R}^2 \rightarrow z^{(n)} \in \mathcal{R}$$

OR

Take n-dimensions down to 2 for example.

Principal Components

The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalised linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p, \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

first principal comp. — linear combinations — Otherwise we have infinite sols.

that has the largest variance

$\phi_{11}, \ldots, \phi_{p1}$	loadings of the first principal component
$\phi_1 = (\phi_{11}, \ldots, \phi_{p1})^T$	principal component loading vector
$\sum_{j=1}^{p} \phi_{j1}^2 = 1$	constraint to prevent an arbitrarily large variance

$$Z_1 = \begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{pmatrix} = \phi_{11} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \phi_{21} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \ldots$$

Try out this interactive demo or this demo.

Further Principal Components

The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance and are uncorrelated with Z_1

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip}, \quad i = 1, 2, \ldots, n$$

$\phi_2 = (\phi_{12}, \ldots, \phi_{p2})^T$ is the second principal component loading vector

Geometry of PCA

- The loading vector ϕ_1 defines a direction in feature space along which the data vary the most
- The projection of the n data points $x_1, \ldots, x_n$ onto this direction are the principal component scores $z_{11}, \ldots, z_{n1}$

Is PCA the same as linear regression? Why or why not?

No, PCA is unsupervised.

Another Interpretation of PC

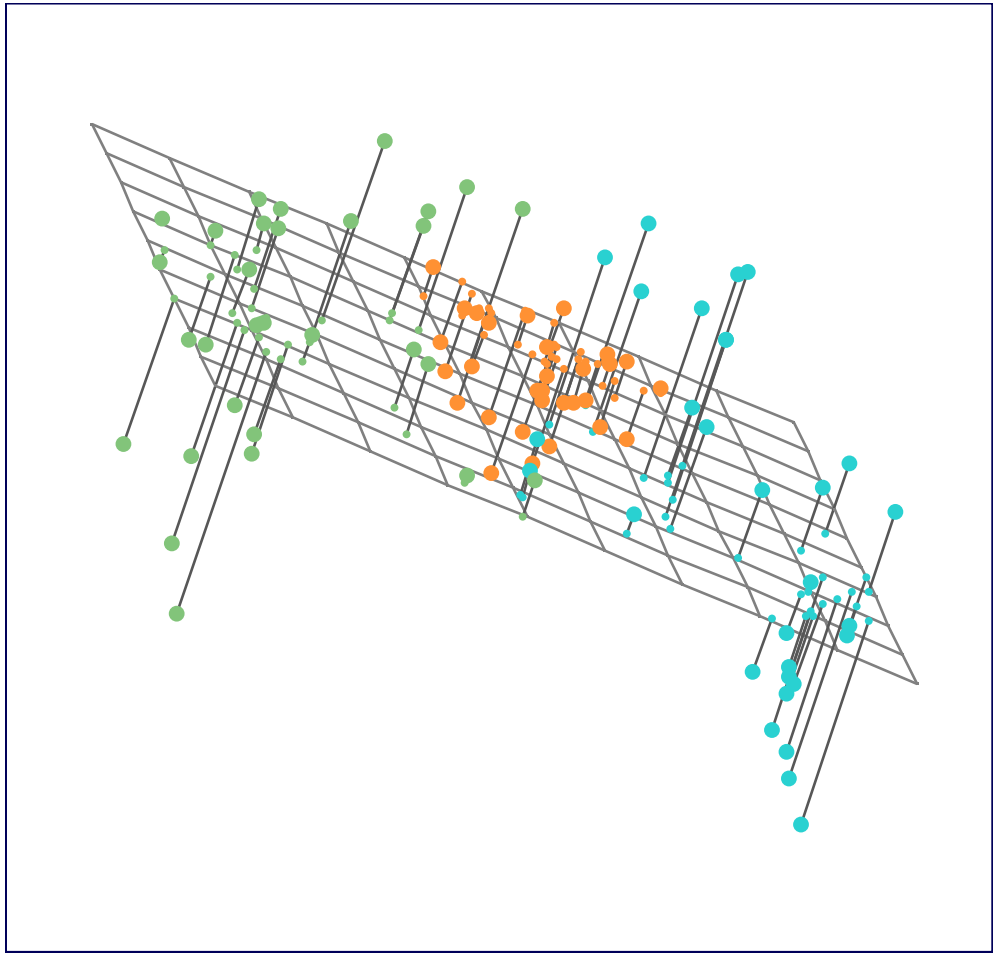
– PC are all orthogonal directions

- The first principal component loading vector
  - the line in p-dimensional space that is closest to the n observations
- Extends beyond the first principal component
  - the first two principal components of a data set span the plane that is closest to the n observations
  - the first three principal components of a data set span the hyperplane that is closest to the n observations
  - and so forth

In 3 dimensions, first two PCs:

- Plane spans the first two principal component directions.
- Minimises the sum of square distances from each point to the plane.

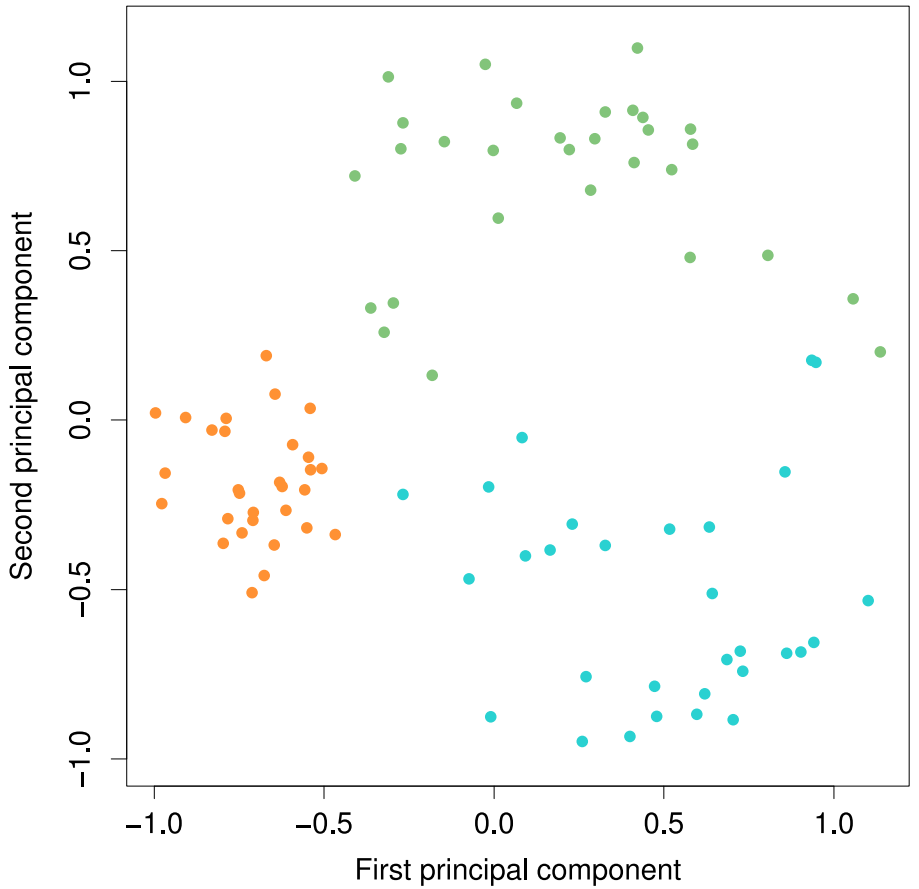
2 principal component directions I



Ninety observations simulated in three dimensions. The observations are displayed in color for ease of visualization. The first two principal component directions span the plane that best fits the data. The plane is positioned to minimize the sum of squared distances to each point.

Source: James et al. (2021), An Introduction to Statistical Learning, Figure 12.2a.

2 principal component directions II



The first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane.

Source: James et al. (2021), An Introduction to Statistical Learning, Figure 12.2b.

UNSW SYDNEY

More on PCA

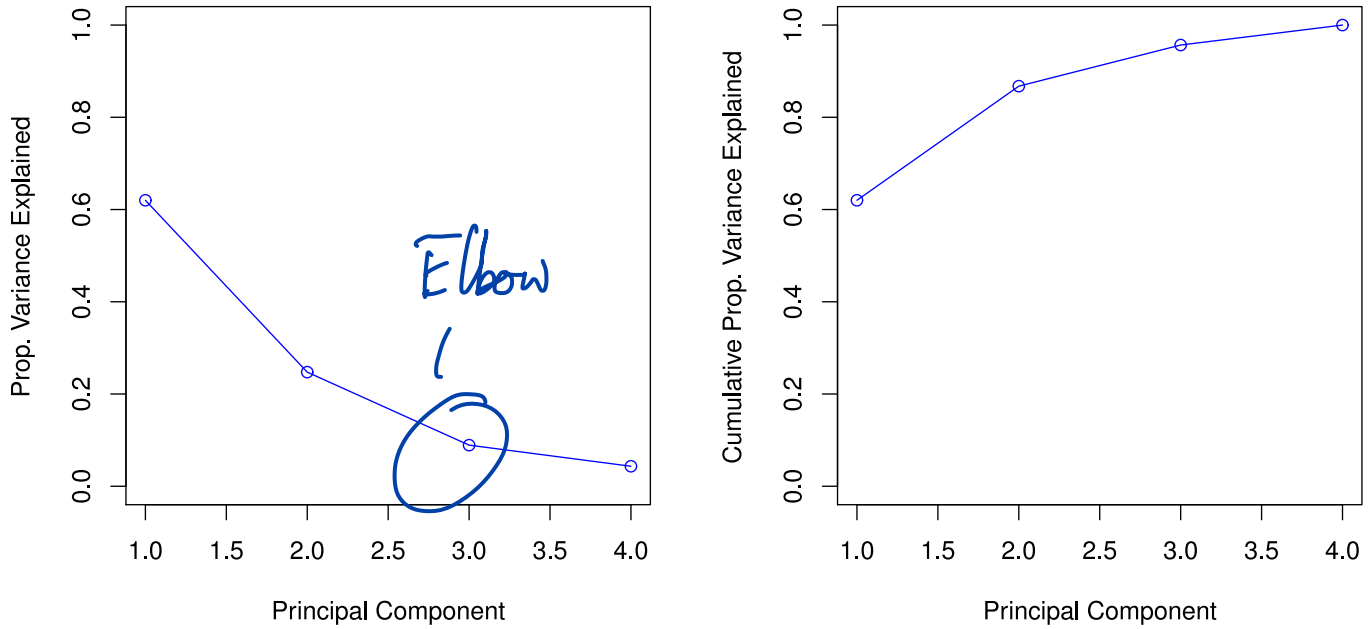
- Scaling the variables
  - typically scale each variable to have standard deviation one before performing PCA
  - may not be necessary if variables are measured in the same units
- Uniqueness of the principal components
  - each principal component loading vector is unique up to a sign flip
- The proportion of variance explained (PVE)
  - the PVE of the mth principal component is given by

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

Typically center around 0

UNSW SYDNEY

How many principal components to use?



Left: a scree plot depicting the proportion of variance explained by each of the four principal components in the USArrests data. Right: the cumulative proportion of variance explained by the four principal components in the USArrests data.

Source: James et al. (2021), An Introduction to Statistical Learning, Figure 12.3.