Week05 - Summary

Unsupervised Learning

Dimension Reduction

- High-dimensional data refers to data set with more features p than observations n
    - Examples: in genetic data, we can easily measure 500k individual DNA mutations (human genome have ~3 billion base pairs of DNA), but experiments generally have <1000 people, e.g., p ~ 500k, n ~ 1000
- Many algorithms and methods have been designed for low-dimensional data and would not work well for high-dimensional data
- To build a linear regression model data with 500k features will result in 500k parameters. This problem is underdetermined if we only have 1000 observations

Dimension Reduction Strategies

- Eliminate or remove features
    - Need to decide which features to be eliminated? Keep ones with high variance?
- Select features
    - Stepwise selection or Lasso
- Build or construct new features from existing ones
    - Replace many existing features with a single one
    - PCA and t-SNE

Principal Component Analysis (PCA)

- Suppose we have a data matrix X with n observations and p features
    - Can we plot the data in a two-dimensional plot?

- Principal component analysis (PCA) finds a way to represent the data in a different space
  - It is still p dimensional, albeit a different coordinate space
  - Aims to explain most variation in the first few dimensions

- The goal of principal component analysis (PCA) is two project information from a high-dimensional space into a smaller number of dimensions
- Find orthogonal linear combinations of variables that explain large proportions of the variation in the data

Principal Components

- Start with a data matrix X, assume it has mean zero

$$\boxed{\mathbf{X} = (X_1 \quad X_2 \quad \dots \quad X_p)}$$

- The first principal component is the normalised linear combination of the features that **maximises the variance** in the new component

$$\boxed{Z = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p = \sum_{i=1}^{p} \phi_{i1}\, X_i = \boldsymbol{\phi}_1^T \mathbf{X}}$$

- The elements $\Phi_i 1$ are known as the loadings of the first principal component
  - By normalised, we mean the squared loadings have to sum to 1, i.e.,

$$\boxed{\textstyle\sum_{i=1}^{p} \phi_{i1}^2 = 1 \Leftrightarrow \boldsymbol{\phi}_1^T \boldsymbol{\phi}_1 = 1}$$

  - Also, it is desired to maximise

$$\boxed{\mathbb{V}\mathrm{ar}(Z_1) = \mathbb{V}\mathrm{ar}(\boldsymbol{\phi}_1^T \mathbf{X}) = \sum_{i=1}^{p} \phi_{i1}^2\, \mathbb{V}\mathrm{ar}(X_i) + \sum_{i \neq j} \phi_{i1}\, \phi_{j1} \mathbb{C}\mathrm{ov}(X_i, X_j)}$$

Principal Component Scores

- Given the principal component loadings, we can project our data matrix X onto the principal component space
  - The projection is a linear combination of the sample feature values:

$$\boxed{z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}}$$

- This is known as the principal component **score**
- The first principal component score vector is:

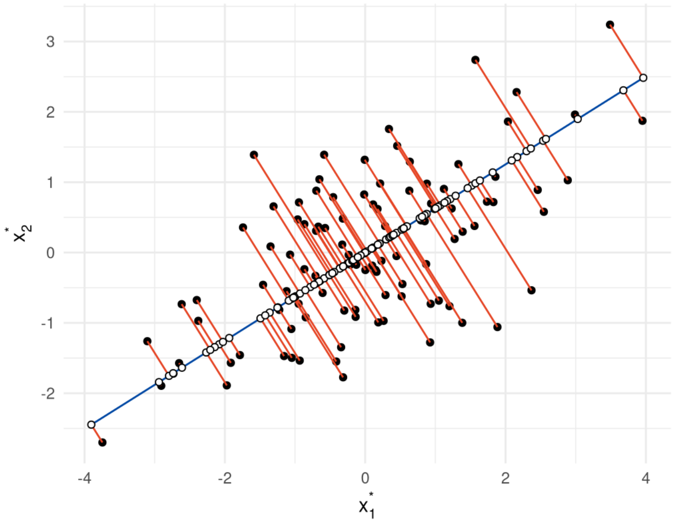
$$\boxed{\mathbf{Z}_1 = (z_{11}, z_{21}, \dots, z_{n1})}$$

- The principal component score vectors are **uncorrelated**

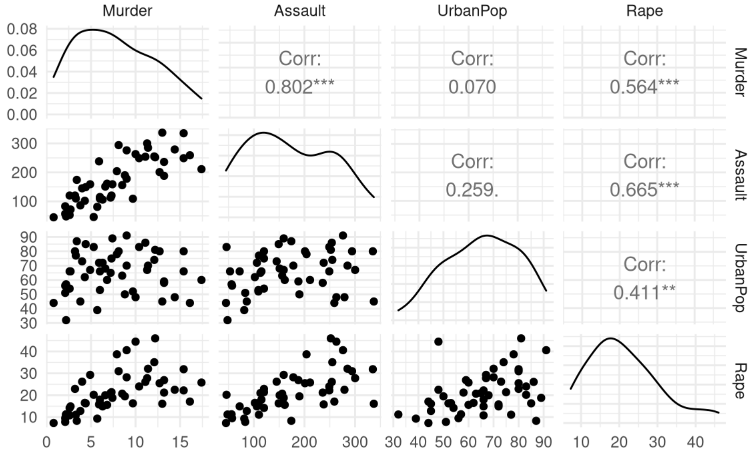
The Details

- The **components** are the new variables. These are sorted by their ability to describe large proportions of the variation in the data
- The **loadings** are a description of how each variable contributes to the new PCs
- The eigenvalues measure the proportion of the variance explained by PCs

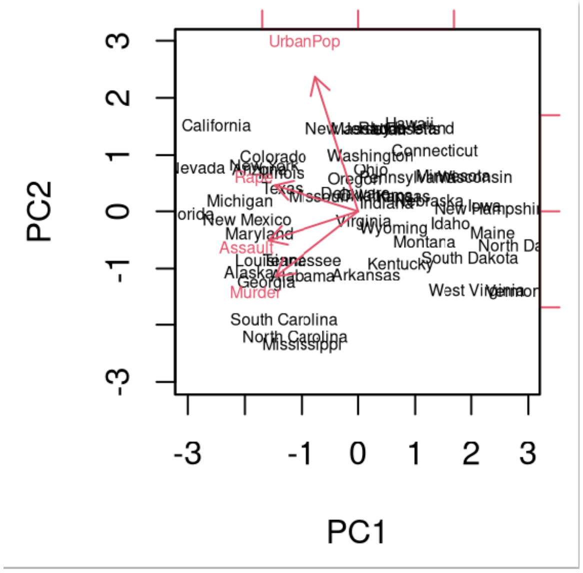
Geometric Interpretation



US Arrests Example



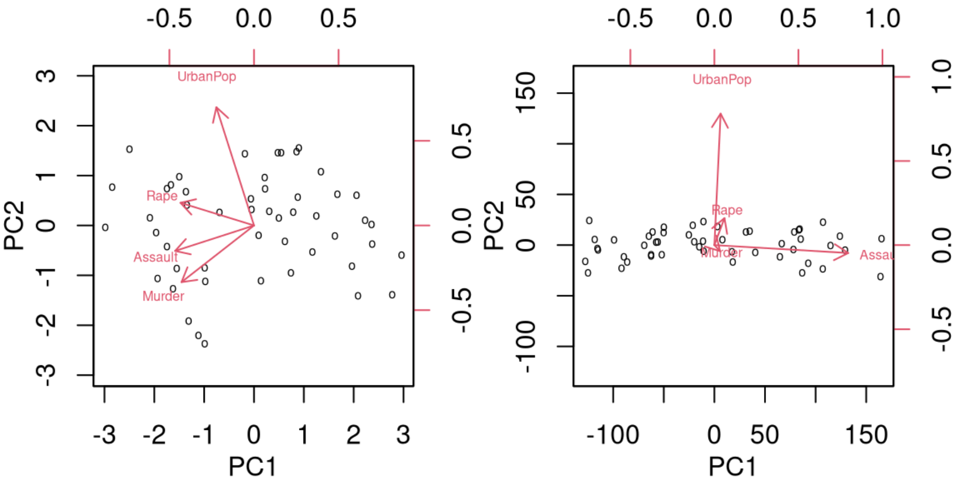
Biplot of the US Arrests



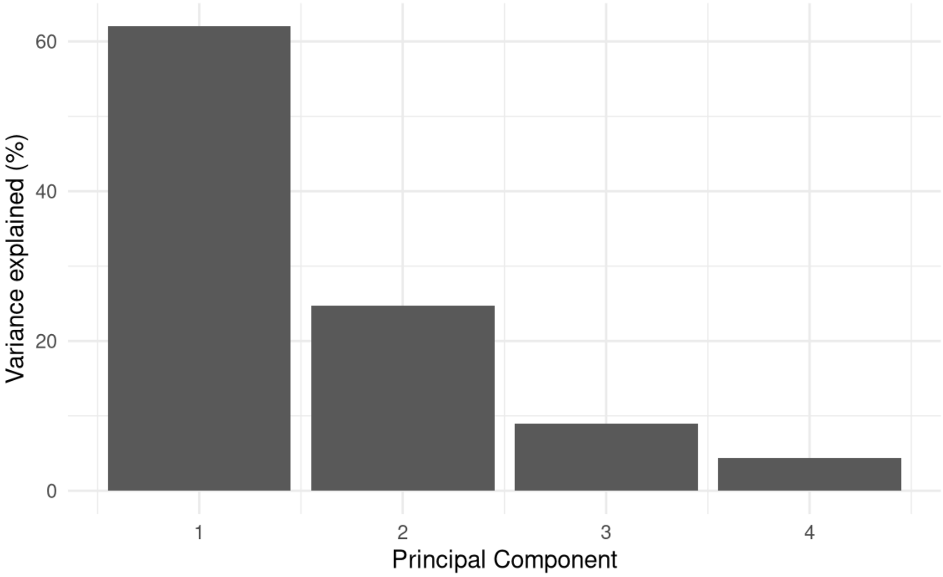
Scaling the Variables

- In a PCA analysis, it is common to centre the variable by removing the mean
- You can also standardise the data to make all the variables have a standard deviation of 1
- If the variables have different units (e.g., in the USArrests data set, murder is measured as number per 100,000 people, but UrbanPop is the percentage of population that lives in urban area), the variance would be very different
- The loadings will put more weight on variables with higher varaince
- However, if all the variables share the same unit, then standardisation may not be necessary

Effect of Scaling (Left) vs Unscaled (Right)



Scree Plot



PCA with Clustering

- Very common approach to deal with high-dimensional data

- Use the first M principal component scores as inputs into the k-means algorithm (M < p)
- Can help improve the clustering model if the signal in the data can be captured in a few principal components

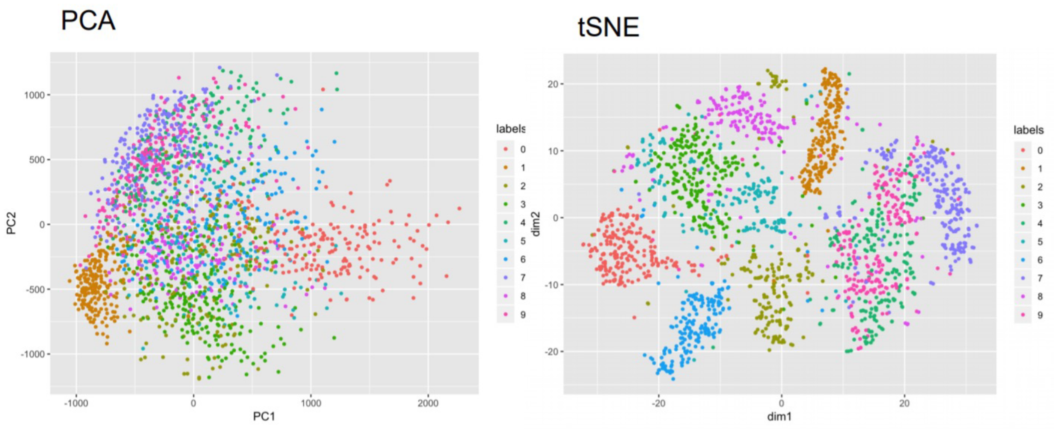
PCA with Regression

- Use the first M principal component scores as the predictors in a linear regression model
- We are assuming that a small number of principal components can explain most of the variability in the data as well as the response
- PCR is useful when variables in the data are highly correlated (i.e., collinear)

t-Distributed Stochastic Neighbour Embedding (t-SNE)

- Nonlinear technique developed for visualising high-dimensional data sets
- Uses local structure in the data to find a low-dimensional representation

MNIST Example



Three steps in t-SNE

1. Constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked
2. Defines a similar probability distribution over the points in the low-dimensional map
3. Minimises the Kullback-Leibler divergence between the two distributiosn with respect to the locations of the points in the map

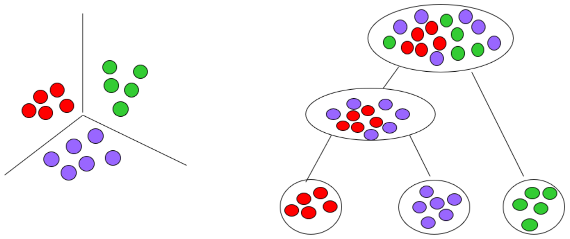
t-SNE vs PCA

- t-SNE is a probabilistic method: it will give you a different representation every time you run it
- PCA is defined by a mathematical formula
- t-SNE is mostly a visualisation method. The PCs from PCA can be interprted whereas t-SNE representation cannot be used for inference
- PCA is a linear method so can only capture linear relationships whereas t-SNE can find more complicated nonlinear relationships

Clustering

Typical Methods

- Partitioning
  - Pre-specified number K of mutually exclusive and exhaustive groups
  - Iterate until criteria is met
- Hierarchical methods
- Two paradigms
  1. Agglomerative: bottom-up
  2. Divisive: top-down



k-Means

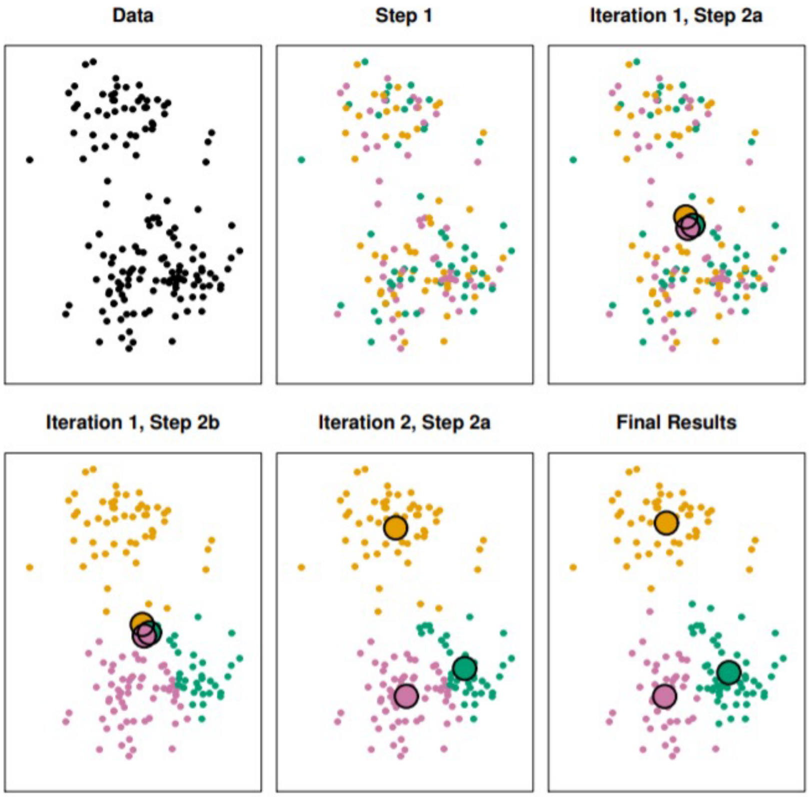
- Iterate each observation at random to a cluster
- Iterate the following until convergence

1. Find cluster means with cluster memberships fixed

$$\bar{x}_j = \text{argmin}_m \sum_{cluster(i)=j} ||\,x_i - m||^2$$

2. Find cluster memberships with cluster means fixed

$$cluster(i) = \text{argmin}_k ||x_i - \bar{x}_k||^2$$



Choosing K

- For cluster C_k can define within-group sum of squares are:

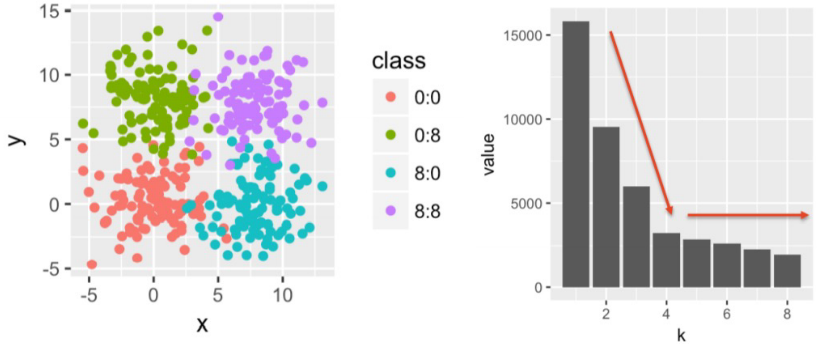
$$\boxed{WSS_k = \frac{1}{|C_k|} \sum_{i,j \in C_k} || x_i - x_j||^2}$$

- This is the sum of all the pairwise squared Euclidean distances between observations in the kth cluster, divided by total number of observations in the kth cluster
- The total within sum of squares criterion aggregates this metric across

$$\boxed{WSS_{Total} = \sum_{k=1}^{K} W\,SS_k}$$

- The total within sum of square criterion will decrease as k increases

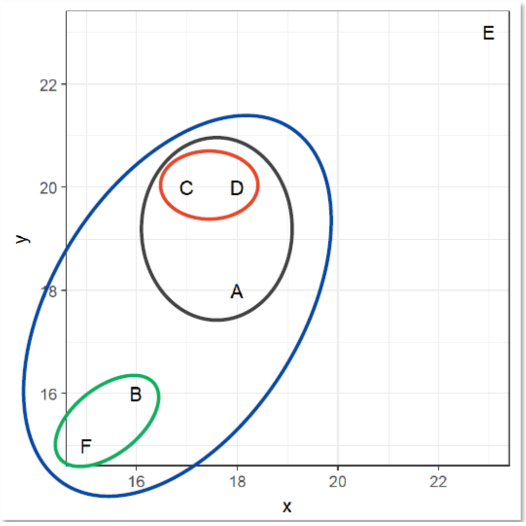
Elbow Plot

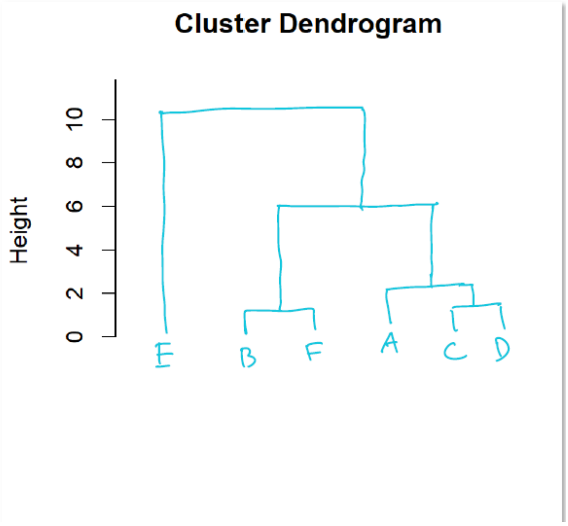


Hierarchical Clustering

- Hierarchical clustering methods produce a tree or dendrogram
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level
- The tree can built in two distinct ways:
  1. Bottom-up: agglomerative clustering
  2. Top-down: divisive clustering

Agglomerative Clustering

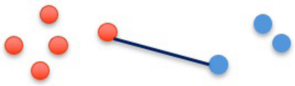




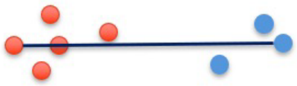
Between cluster similarity measures

Linkages: measure of dissimilarity between two sets of objects that determine how two sets of objects are merged

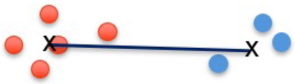
- Single linkage
- Complete linkage
- Average linkage



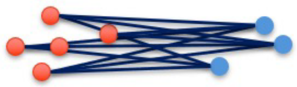
Single (minimum)



Complete (maximum)



Distance between centroids



Average (mean) linkage