Rotman

INTRO TO R PROGRAMMING

R Tutorial (RSM456) – Session 3

January 20, 2025 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-intro-2025-winter/

Rotman School of Management
UNIVERSITY OF TORONTO

Plan

- K-Means Clustering
- Principal Components Analysis

K-Means Cluster Analysis

- An **unsupervised** learning method to partition n observations into k clusters (i.e., subgroups)
  - e.g., market segmentation: an online shopping site try to identify groups of shoppers with similar browsing and purchase histories
- A **cluster** refers to a collection/subgroup of data points aggregated together because of certain similarities
  - Similarity based on a distance measure
- Need to set k
  - There are methods to help you decide the value of k

K-Means Cluster Analysis: A Simple Example

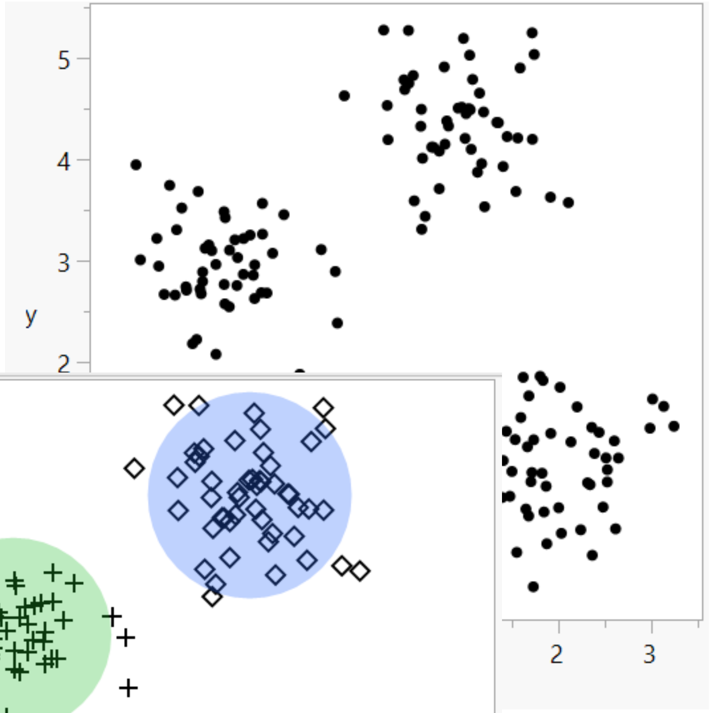
- Observations: 150 2-d points
- Set $k = 3$
  - partition each observation into one of the 3 clusters $S = \{S_1, S_2, S_3\}$
- Run K-means clustering algorithm
  - find 3 clusters such that total WSS is minimized

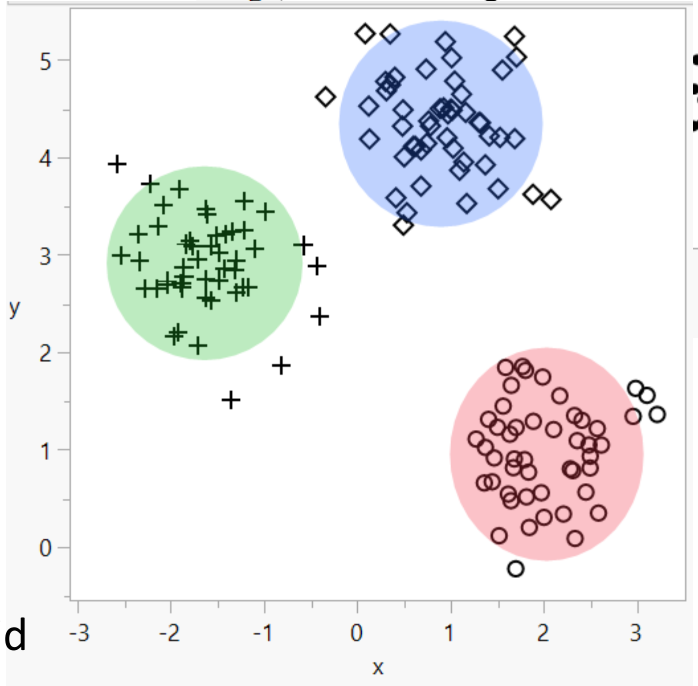
Within-cluster **S**um of **S**quares (WSS)

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{3} \sum_{x \in S_i} \|x - \mu_i\|^2$$

center/centroid of cluster i

Total WSS





K-means Clustering Algorithm

- An iterative algorithm
- Clustering result can depend on initial random cluster assignment
  - So, important to run the algorithm multiple times from different initial configurations, and then select the best one

Algorithm 12.2 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Source: ISLR2 chapter 12

K-means in R – the Country Risk Exercise

- Import the `country_risk.xlsx` data

- Prepare the data for k-means clustering
  - Perform correlation analysis and choose features
  - Standardize the features
    - an important step for k-means clustering unless all features are already on the same scale or the differences in scales are irrelevant for your specific application

- Perform a K-means cluster analysis (`kmeans()` in base R stats package)
  - Determine *k* using the "elbow" method
  - Run k-mean clustering algorithm for a chosen *k*
  - Interpret/name the clusters

Principal Components Analysis (PCA)

- A dimensionality reduction technique
  - transform data from a high-dimensional space into a low-dimensional space while retaining the most important properties of the original data

- PCA reduces dimension (number of variables/features) while preserving as much **variance** as possible

- Why PCA?
  - Data visualization
  - Machine learning data preprocessing
    - Why: 1) prevent overfitting; 2) improve computational efficiency;
  - Data imputation (i.e., filling in missing values)

PCA – Some Details

- Setup: a set of p variables/features $X_1, X_2, \dots, X_p$

- The first principal component (PC) is the *normalized* linear combination of $X_1, X_2, \dots, X_p$ that has the largest variance

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p,$$

  where $\sum_j \phi_{j1}^2 = 1$. $\phi_{11}, \phi_{21}, \dots \phi_{p1}$ are called **loadings** of the first PC.

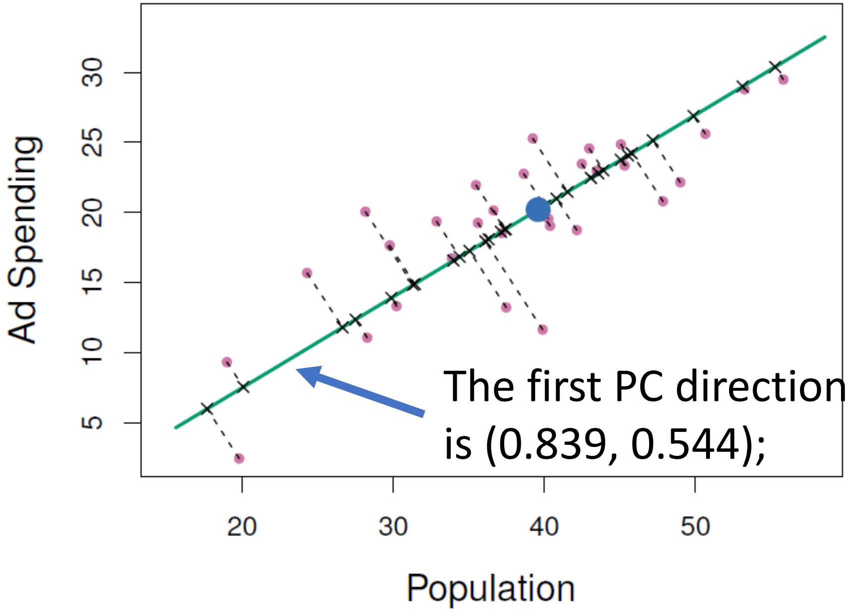
- $\phi_1 = (\phi_{11}, \phi_{21}, \dots \phi_{p1})$ defines a direction along which the data vary the most.

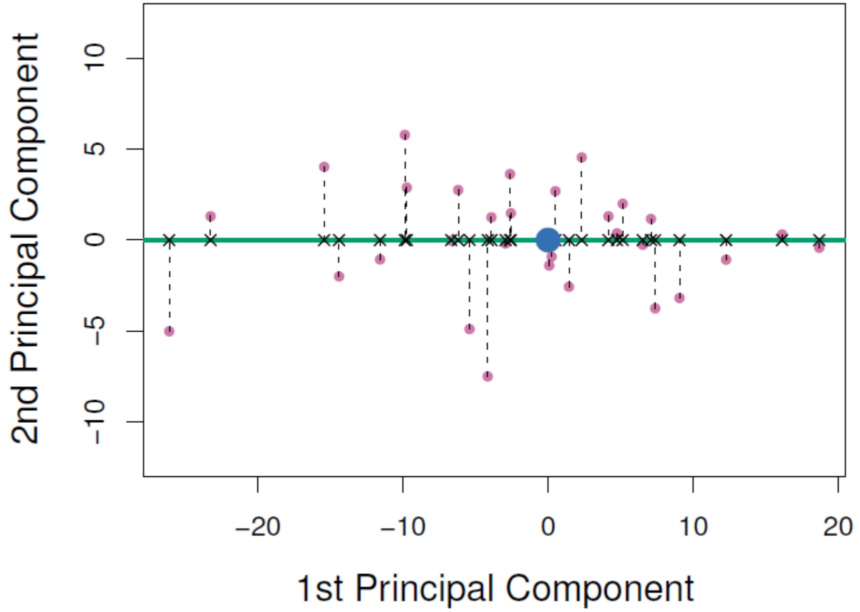
- The second PC is the normalized linear combination of $X_1, X_2, \dots, X_p$ that has the largest variance out of all combinations that are *uncorrelated* with Z_1.

Reference: ISLR2 chapter 12.

PCA – An Example

- Two variables/features: Population & Ad Spending
- **Loadings** of the first principal component $(\phi_{11}, \phi_{21}) = (0.839, 0.544)$
  - (ϕ_{11}, ϕ_{21}) is the direction in feature space along which the data vary the most
- **Scores** of the first PC: projection of data onto the direction of (ϕ_{11}, ϕ_{21})
  - $z_{i1} = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$ for observation i





Source: ISLR2 chapter 6 Figure 6.15.

PCA in R - the Country Risk Exercise

- Import the `country_risk.xlsx` data
- Quickly explore the data
  - Discuss the importance of scaling variables before performing PCA
- Perform PCA (`prcomp()` in base R stats package)
  - Run the PCA algorithm (scale prior to PCA)
  - Examine and understand the results
  - Visualize the first two principal components & variance explained

Reference: ISLR2 chapter 12 lab.