Unsupervised Machine Learning

Introduction

Sebastian Sternberg & Marcel Neunhoeffer

27.03.2019

Introduction

What will we cover today?

- What is unsupervised machine learning?
- What are real world applications?
- What are the most commonly used methods, and how do they work?
- How to apply these methods on your own using *R*.

Today's schedule

Morning session (10:00 - 12:00)

- Introduction
- Principle Component Analysis
- Practical session

Afternoon session (13:00 - 16:00)

- Clustering (K-mean clustering, Hierarchical clustering)
- Practical session

Supervised versus Unsupervised Learning

- Yesterday, supervised learning methods such as classification and regression were covered.
- The supervised learning setting has a set of features (X_1, X_2, X_p) for each object and an outcome variable Y. The data is labeled. Predicting Y is done using the features.
- In the unsupervised learning setting, we only observe the features X_1, X_2, X_p. We cannot predict Y, because there is no outcome variable Y.

Supervised versus Unsupervised Learning

Supervised Learning:

- Dataset has clear set of Predictors (X) and Outcomes (Y).
- Goal is to understand the relationship between predictors and outcomes.
- End goal is either inference or prediction.
- Either Regression or Classification is used.

Unsupervised Learning:

- Dataset consists only of Predictors (X).
- We seek to understand the relationships between the variables/observations in the dataset.
- Either Clustering or Dimensionality Reduction is used.

The Goals of Unsupervised Learning

Unsupervised learning is also called exploratory data analysis or knowledge discovery:

- Is there an informative way to visualize the data?
- Are there subgroups among the variables and/or the observations?

Unsupervised learning often aims at dimensionality reduction:

- Many supervised algorithms that work fine in low dimensions become intractable when the input is high-dimensional (*curse of dimensionality*).
- Unsupervised learning can also be a *data pre-processing step*.

Unsupervised learning examples

Real world examples:

- **Netflix recommendations**
- **Friend recommendations in social networks**
- **Amazon: People who buy product X are also interested in product Y**

Economics/Business examples:

- **Online banking fraud detection** (Cabanes, Bennani and Grozavu, 2013)
- **Credit fraud detection** (Vikrant Agaskar et al., 2017)
- **Market segmentation** (Moutinho and Brownlie, 1989)
- **Clustering banking sectors** (Moutinho and Brownlie, 1989)

Challenges of unsupervised learning

Unsupervised learning can be more challenging than supervised learning:

- It is more subjective: there is no such "simple" goal as forecasting a certain outcome.
- Result evaluation is more difficult: there are no universally accepted mechanisms such as cross validation or other ways to validate results.
  - In *supervised learning*, we train a model on some training data, and then evaluate its predictive capability using test data where we know the true outcomes.
  - In *unsupervised learning*, there is no way to check our work because we do not know the true outcome.

The strengths of unsupervised learning

Unsupervised learning also has a couple of strengths:

- It is often easier to obtain unlabeled data than labeled data, which often require human intervention.
- It can be a faster alternative to supervised methods
- It helps to detect pattern prior to running any supervised algorithm.

With the rise of Big Data and automated data collection, unsupervised learning becomes more and more important.

Overview of Unsupervised Learning Methods

- Association rules
- Principal component analysis
- Multidimensional scaling
- Factor analysis
- K-means clustering
- Hierarchical clustering
- Network analysis

There are many more related but specific hybrid forms of these methods for particular problems.

Overview of Unsupervised Learning Methods

Today we will focus on the following methods:

- Association rules
- Principal component analysis
- Multidimensional scaling
- Factor analysis
- K-means clustering
- Hierarchical clustering
- Network analysis

References

Cabanes, Guenael, Younes Bennani and Nistor Grozavu. 2013. "Unsupervised Learning for Analyzing the Dynamic Behavior of Online Banking Fraud." *2013 IEEE 13th International Conference on Data Mining Workshops* (December):513–520.

Moutinho, Luiz and Douglas T. Brownlie. 1989. "Customer Satisfaction with Bank Services: A Multidimensional Space Analysis." *International Journal of Bank Marketing* 7(5):23–27.

Vikrant Agaskar, Professor, Megha Babariya, Shruthi Chandran and Namrata Giri. 2017. "Unsupervised Learning for Credit Card fraud detection." *International Research Journal of Engineering and Technology* pp. 2395–56.

Unsupervised Machine Learning Part I

Sebastian Sternberg & Marcel Neunhoeffer

27.03.2019

University of Mannheim, Graduate School of Economic and Social Sciences

Introduction to Principal Component Analysis

Principal Component Analysis

Principal Component Analysis (PCA) generates linearly uncorrelated dimensions that can be used to understand the underlying structure of the data.

Main goal of PCA is dimensionality reduction:

- Summarize many (correlated) features into a few, uncorrelated dimensions.
- Especially useful in Big Data context: saves computational time and reveals patterns.
- Helps to gain a better macro-level understanding of our data.

Examples of PCA usage in the field of economics

PCA is an old but still widely used method:

- Investigating the linkage between different attributes of central bank independence and inflation performance (Banaian, Burdekin and Willett, 1998).
- The effect of oil price on world food prices (Esmaeili and Shokoohi, 2011).
- Fraud classification (Brockett et al., 2002).

Own work example: collaboration with market research institute

Large data set (10.5 Mio. data points) on the internet behavior of individuals across a long time span, collected by a large market research institute.

- Data set contained every single internet usage (visited, apps used etc.) with a time stamp and context information.
- Goal was to develop a classifier that predicts voting behavior of individuals
- Not possible to run machine learning classifiers without some feature engineering.
- PCA was used to collapse the big data set into only 15 features.

Why do we need principal components?

PCA has two strengths:

- dimension reduction and
- data inspection

... with the same method

What are principle components?

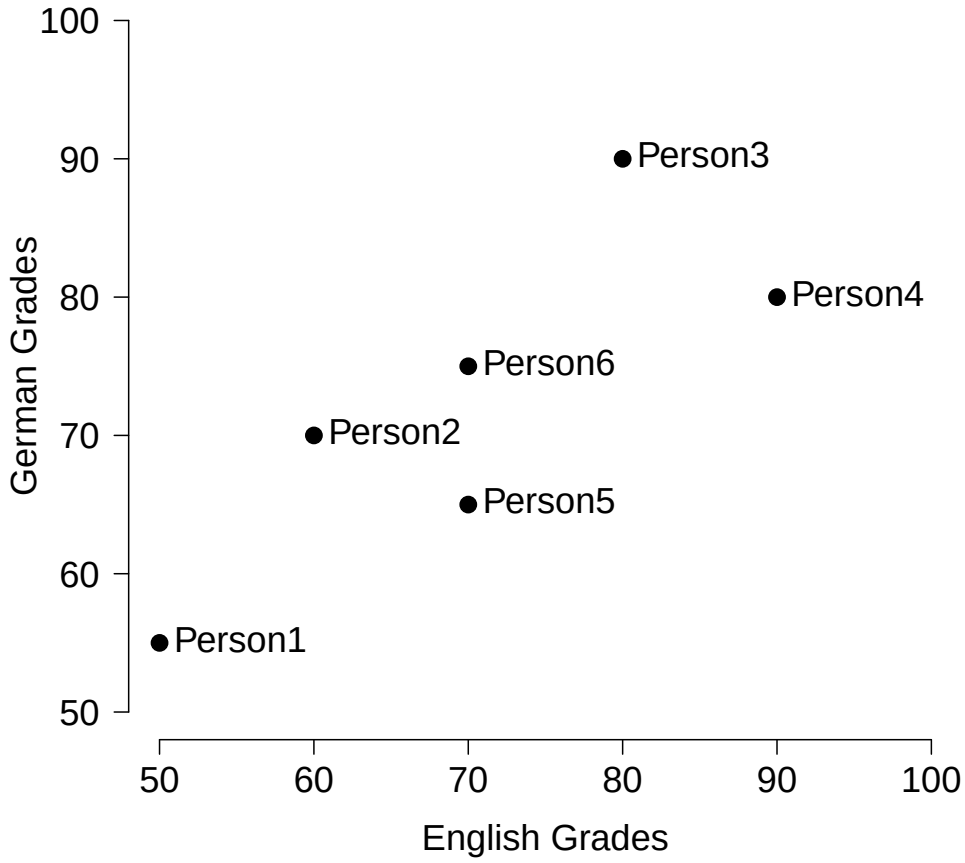
Given a set of data on *n* dimensions, PCA aims to find a linear subspace of dimension *d* lower than *n* such that the data points lie mainly on this subspace.

- Combine a correlated group of variables into a new characteristic (component).
- We therefore “throw away” information by combining variables into new components.
- The better the components are in explaining the variance of all variables, the better job PCA did in summarizing the variables.

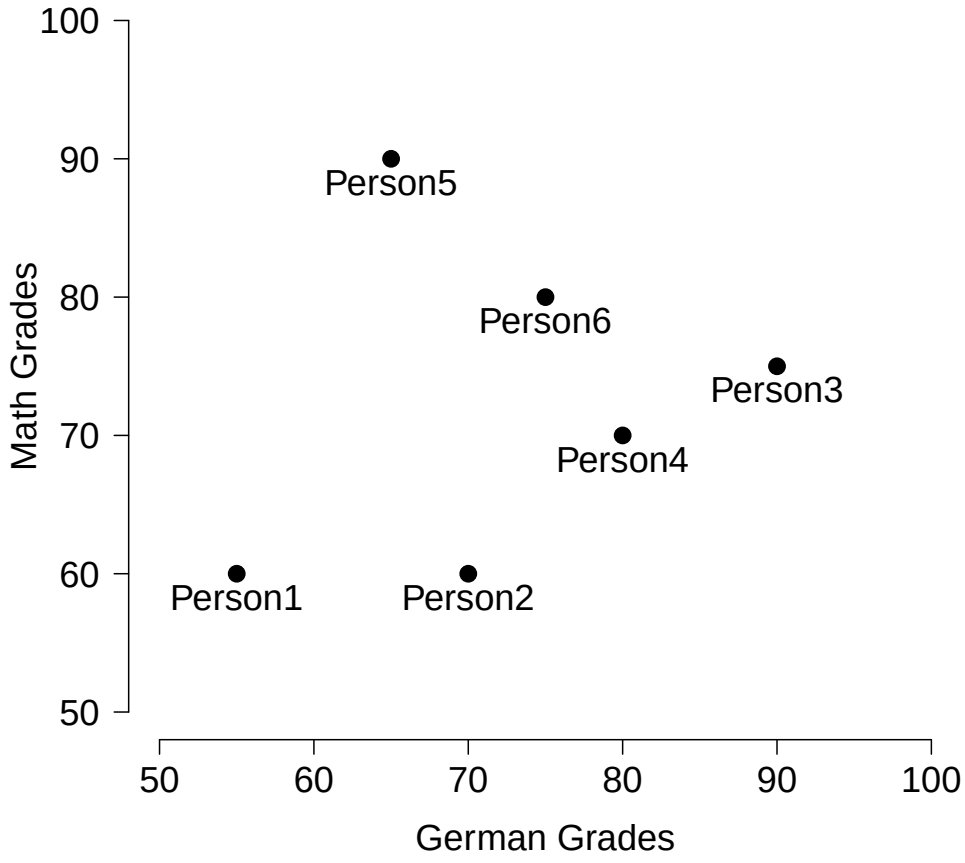
Quick demonstration using school grades

Name	Maths	Physics	English	German	Music	Art
Person1	80	90	50	55	60	55
Person2	90	95	60	70	60	65
Person3	60	65	80	90	75	80
Person4	65	65	90	80	70	65
Person5	65	70	70	65	90	80
Person6	60	60	70	75	80	90

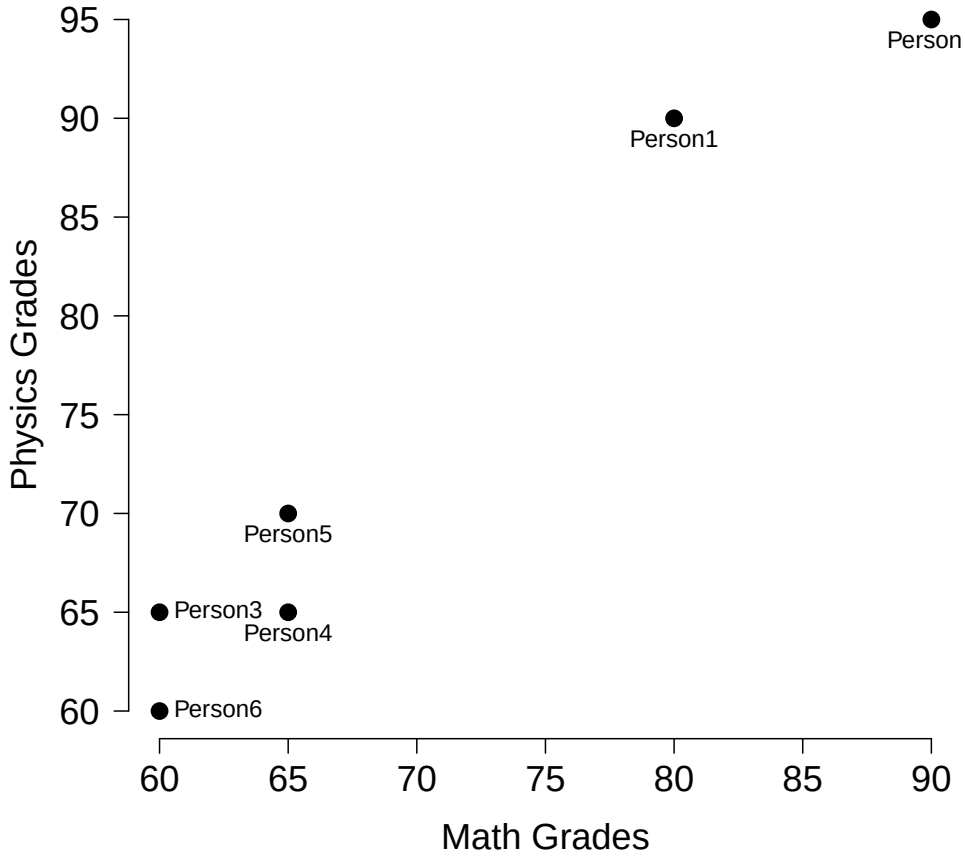
Scatterplot English-German Grades



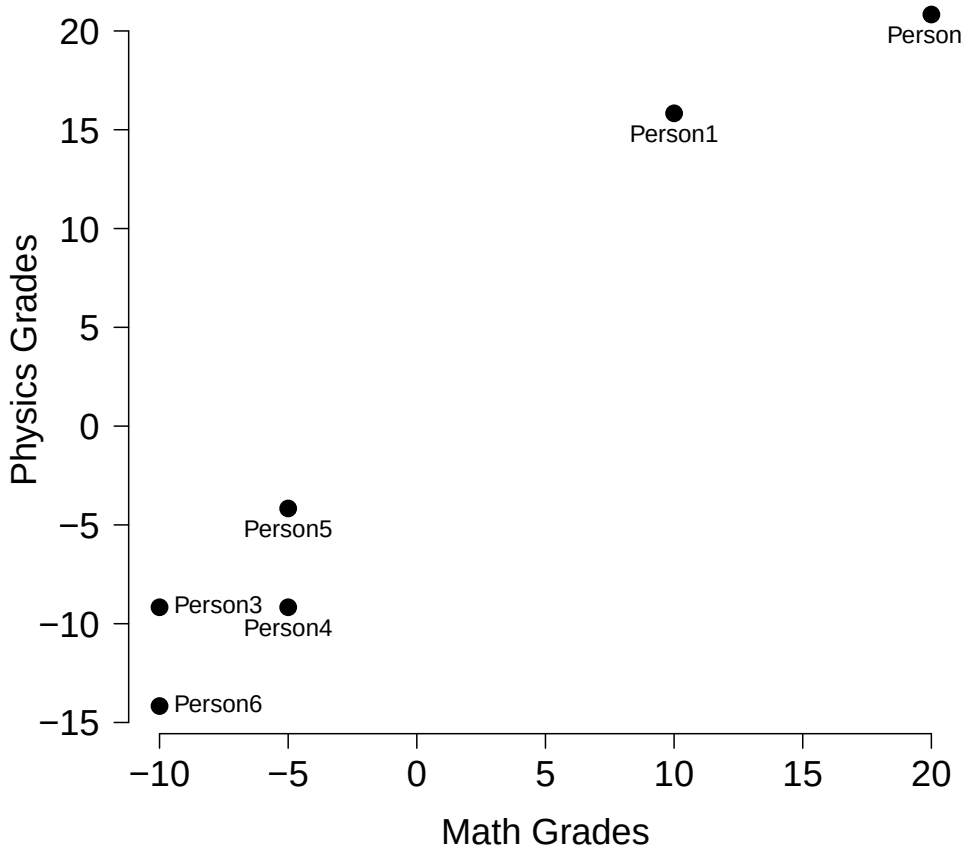
Scatterplot German-Math Grades



A Two-Dimensional Example: Math and Physics Grades

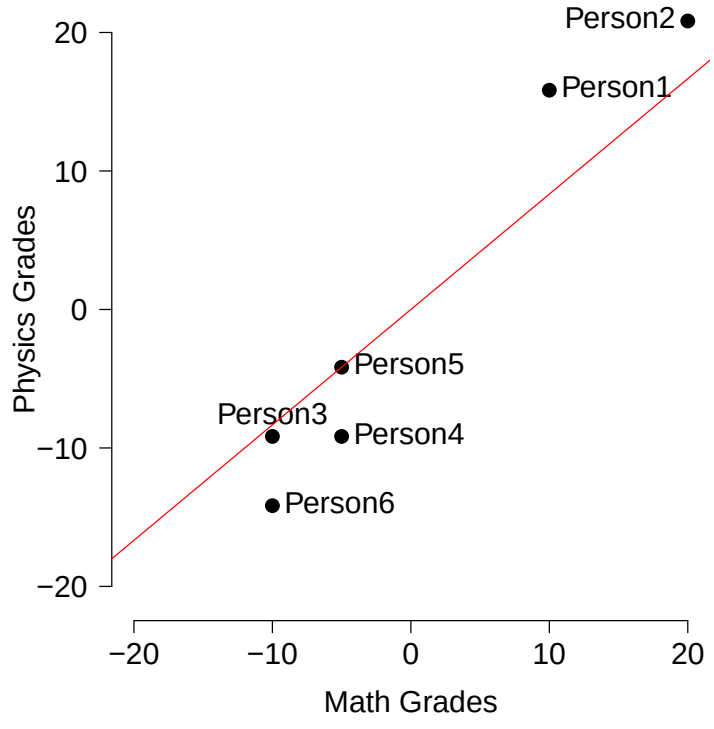


After normalization

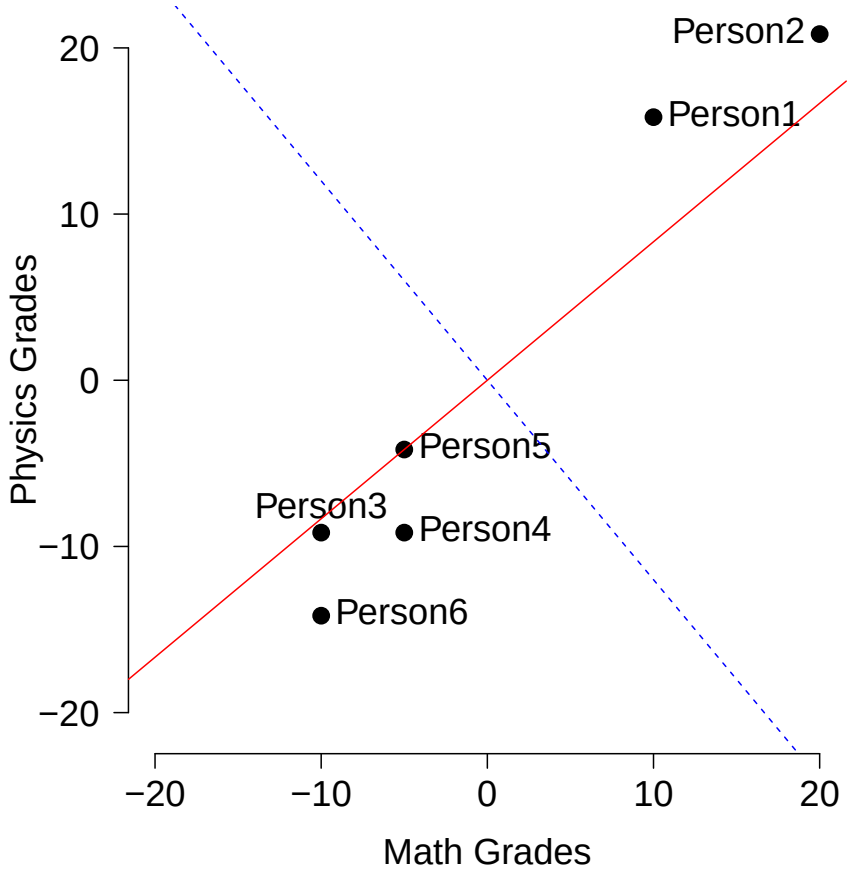


Finding the first principle component

- The first principal component is the best linear approximation for the data and the direction with the maximum variance.
- This results in a line which minimizes the sum of squared distances between a data point and the line.

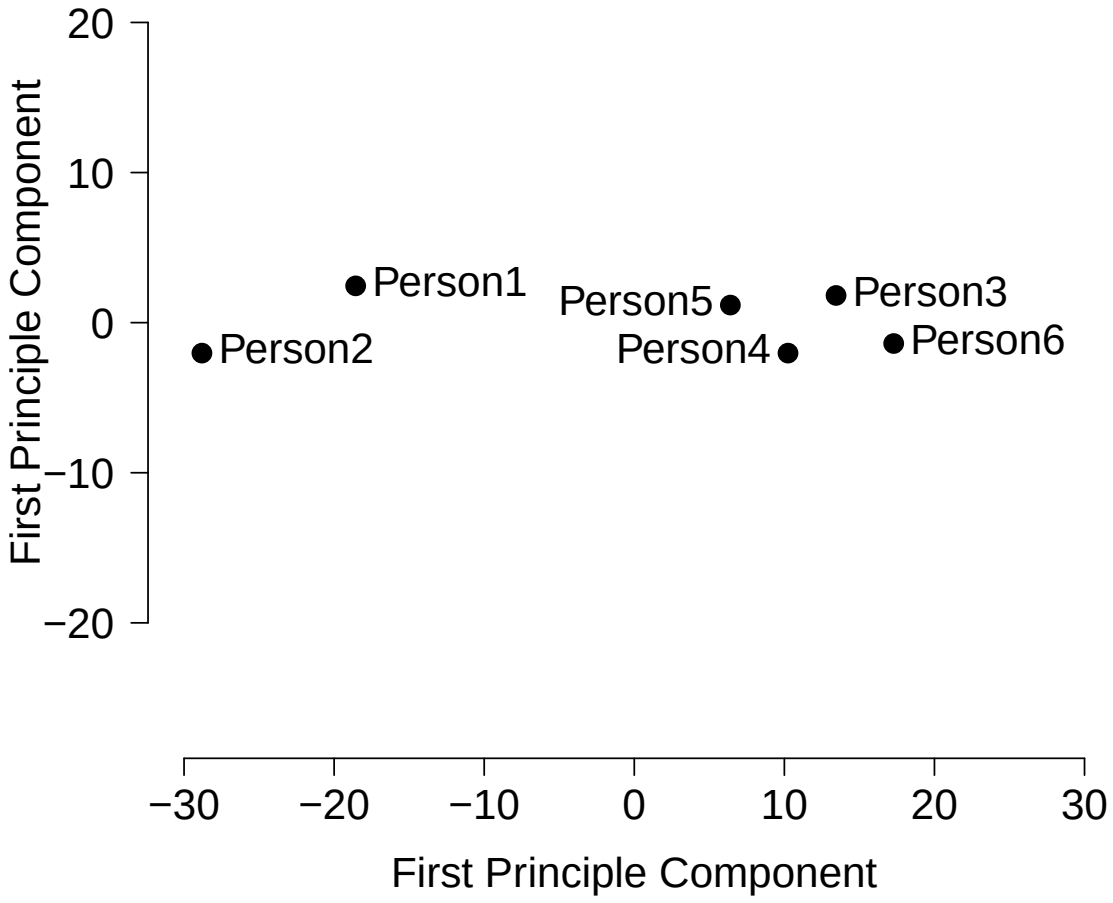


Finding the second principle component



Reoriented data

- These two principle components can be used as new dimensions.



Computation

Computation of Principle Components

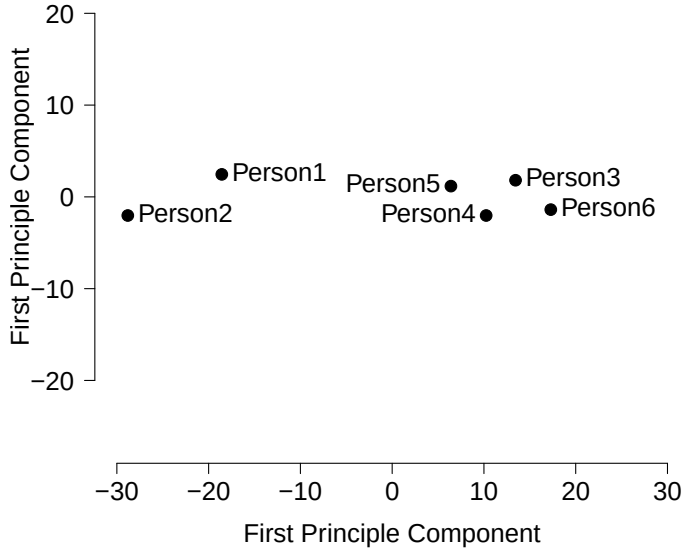
Principle components are calculated as: $Z = P \times A^{\intercal}$, where P are so called scores, A is a matrix of loadings (eigenvectors), and Z the original (standardized) data matrix.

1. Standardize the variables.
2. Calculate covariance matrix.
3. Perform eigen decomposition to find *eigenvectors* and *eigenvalues* of covariance matrix. The eigenvector with the highest eigenvalue is the first principal component.
4. Reorient the data by multiplying the original data with the eigenvectors. This gives the scores.

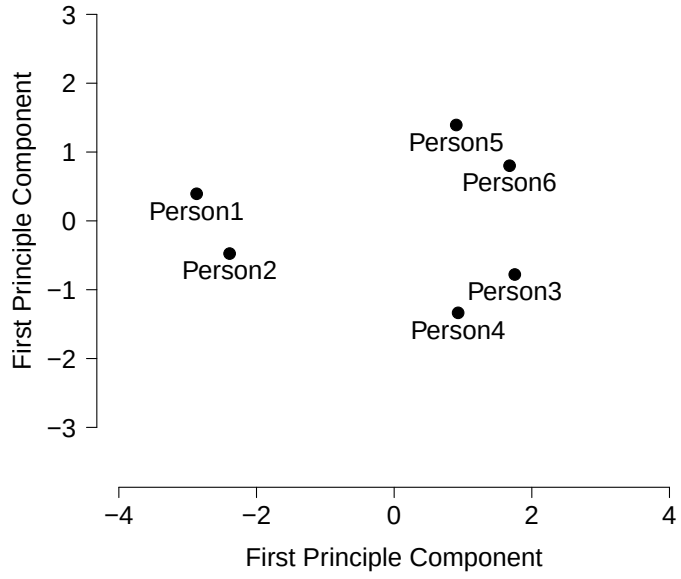
The positions of each observation in this new coordinate system of principal components are the scores and are calculated as the linear combinations of the original variables and the respective loadings.

Reduction of two and six dimensions

- Just math and physic grades were used so far, but dimension reduction can also be done for all six school subjects.

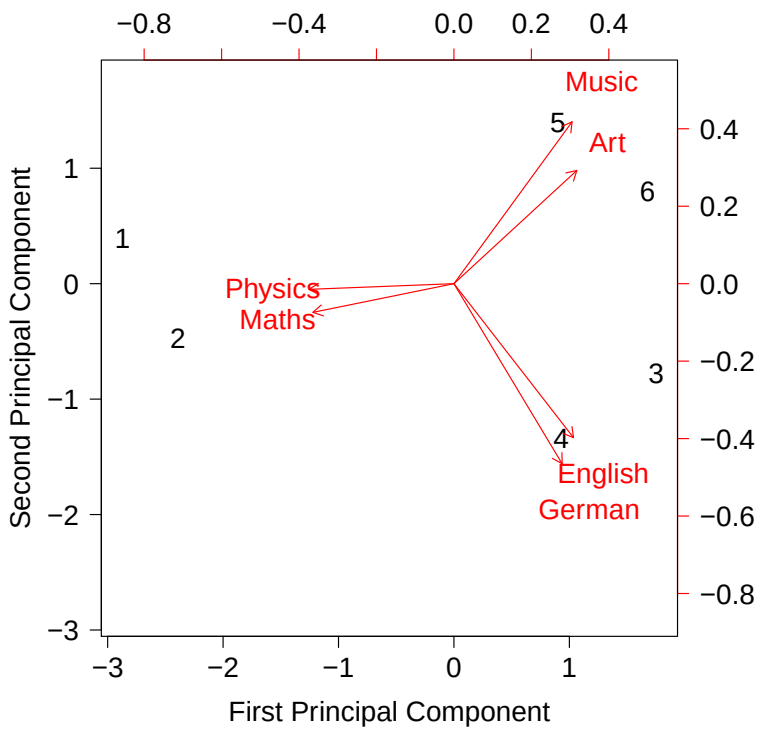


(a) 2 Dimensional Reduction



(b) 6 Dimensional Reduction

Introducing Biplots



- Biplots are often used to visualize the PCA results using the first two PCs.
- "Biplot" because both the *scores* and the *loadings* are displayed in one figure.
- The red arrows indicate the first two principle component loading vectors.
- The black numbers represent the *scores* for the first two principle components.

A Textbook Illustration

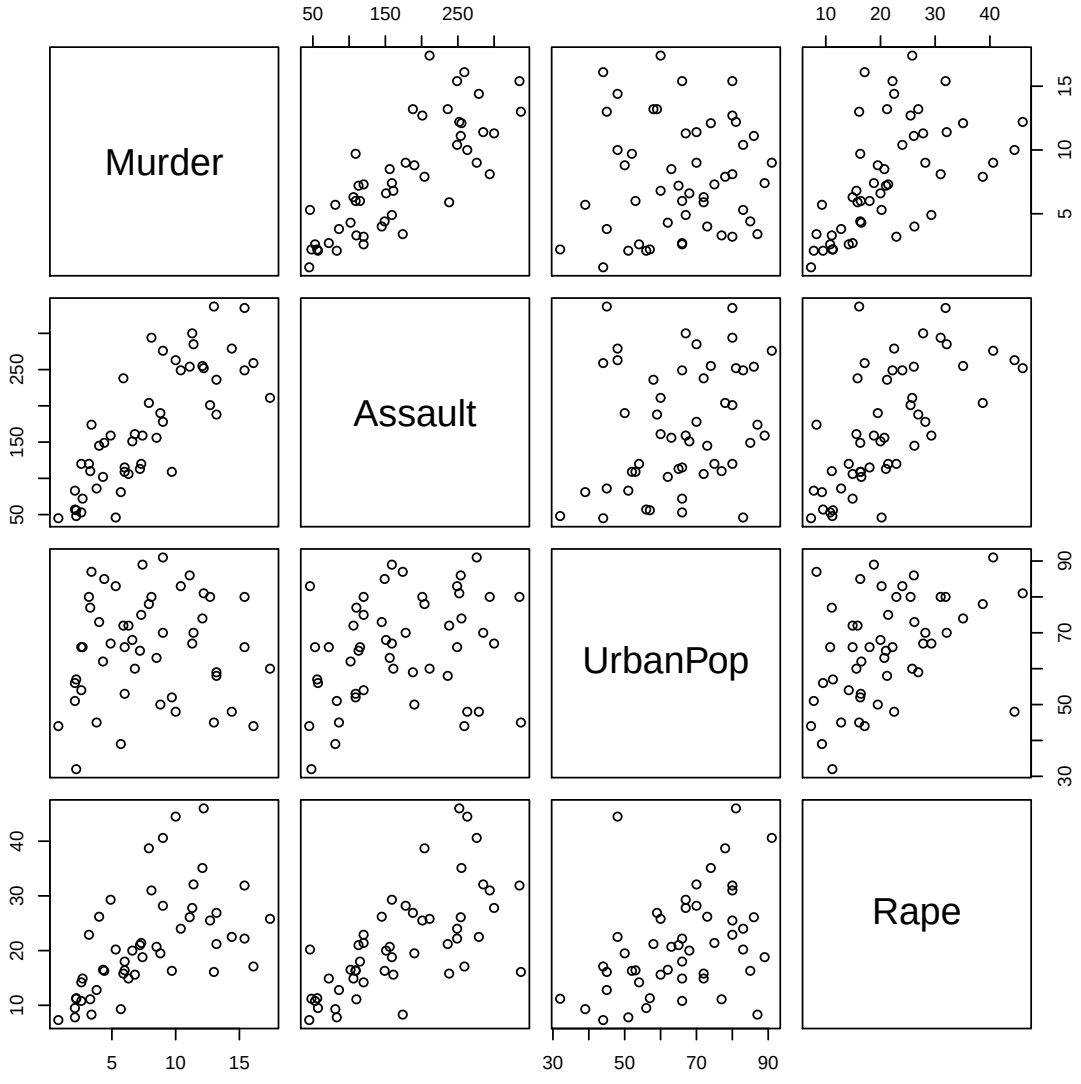
A textbook illustration: USArrests data

We use a textbook example: USArrests data (James et al., 2013)

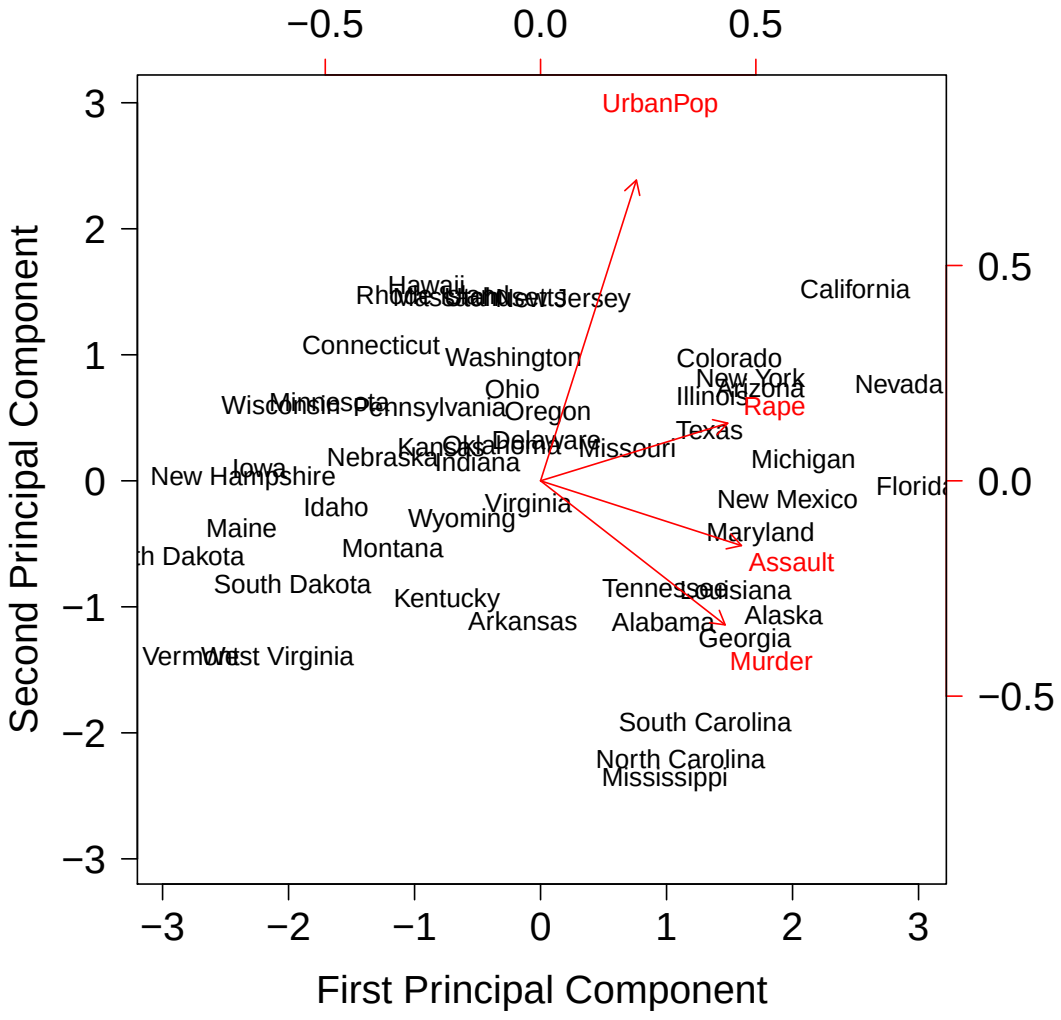
- For each of the fifty states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas).
- Our principal component score vectors thus have length $n = 50$, and the principle component loading vectors have length $p = 4$
- PCA was performed after standardizing each variable.

This is just a teaching data set with some nice properties. In the real world, PCA is used when there are way more observations and variables (see practical session).

Scatterplots of the included variables



USArrests data: PCA biplot



Interpreting USArrests results - Part I

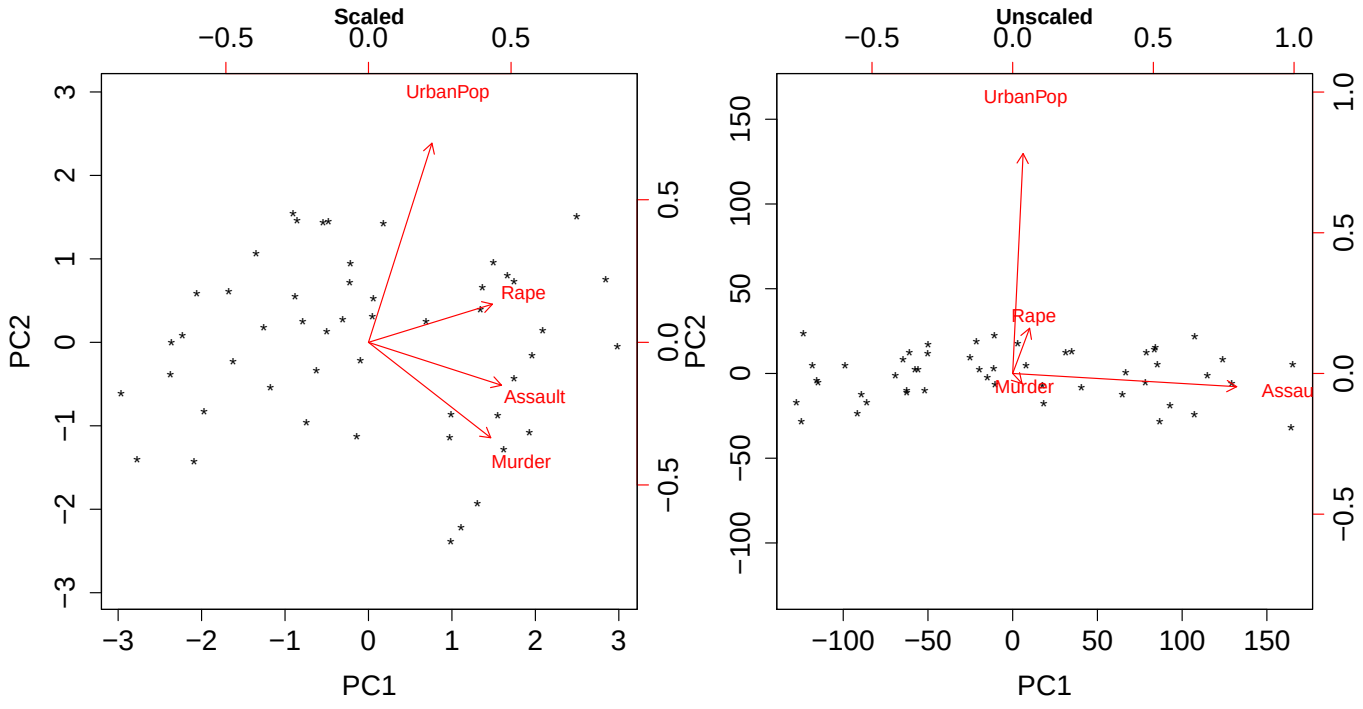
The first two principal components for the USArrests data.

- The black state names represent the scores for the first two principal components.
- The red arrows indicate the first two principal component loading vectors.
- For example, the loading for *Rape* on the first component is 0.54, and its loading on the second principal component 0.17 (the word *Rape* is centered at the point (0.54, 0.17)).

Interpreting USArrests results - Part II

- The first principle component vector places approximately equal weight on *Assault*, *Murder*, and *Rape*, and less weight on *UrbanPop*.
  - This component corresponds to the overall rates of serious crime.
- The second principle component vector places most of its weight on *UrbanPop* and much less weight on the other three features.
  - This component corresponds to the level of urbanization of the state.
- Crime-related variables are located close to each other; UrbanPop is distant.
  - Crime-related variables are correlated with each other.
  - States with high murder rates tend to have high assault and rape rates; UrbanPop variable is less correlated with the other three.

Why scaling is important



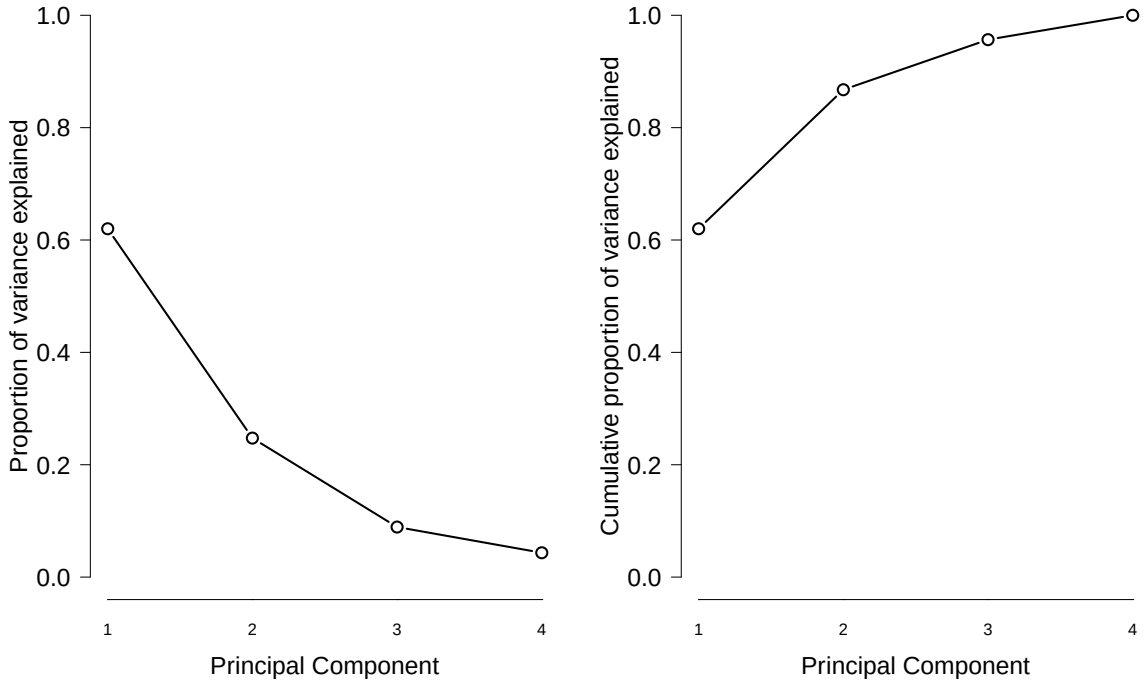
- If the variables are in different units, standardizing is recommended.

Proportion of variance explained

- We summarized 50 observations and 4 variables using the first two principal component score vectors and the first two principal component loading vectors.
- How much of the information in a given data set is lost by projecting the observations onto the first few principal components?
- The contribution of each component is given by the proportion of variance explained (PVE).
- Eigenvalues in PCA indicate how much variance can be explained by their associated eigenvectors.

Proportion of variance explained

- The PVE of a component is between 0 and 1 (by construction, all eigenvalues sum up to 1).
- The first two principal components explain almost 87% of the variance in the data, and the last two principal components explain only 13% of the variance.



How many principle components should we use?

- If we use principal components as a summary of our data, how many components are sufficient?
- There is no single (or simple!) answer to the question how many PC are needed.
- Decision is based on “eyeballing” the scree plot and looking for the “elbow”.

Wrap-up PCA

PCA is a standard tool in big data context: **dimension reduction** and **data inspection**.

Direct usage:

- Summarize many (correlated) features into a few, uncorrelated dimensions.
- Detecting interesting patterns in data for a deeper investigation.

Indirect usage:

- PCA as data pre-processing: only use first few principle components instead of all features.
- These can be used for prediction (saves computation time), or as new outcome variables.

Practical Session PCA

In the practical session you will...

- learn how to do PCA in *R* replicating the US-Arrest example.
- apply PCA to the Immoscout data set for data exploration and data pre-processing.

References

Banaian, King, Richard C. K. Burdekin and Thomas D. Willett. 1998. "Reconsidering the Principal Components of Central Bank Independence: The More the Merrier?" *Public Choice* 97(1/2):1–12.

Brockett, P., R. Derrig, L. Golden, a. Levine and M. Alpert. 2002. "Fraud Classification Using Principal Component Analysis of RIDITs." *Journal of Risk and Insurance* 69(3):341–371.

Esmaeili, Abdoulkarim and Zainab Shokoohi. 2011. "Assessing the effect of oil price on world food prices: Application of principal component analysis." *Energy Policy* 39(2):1022–1025.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Unsupervised Machine Learning Part II

Cluster Methods

Sebastian Sternberg & Marcel Neunhoeffer

27.03.2019

Outline

Introduction to Clustering

Clustering

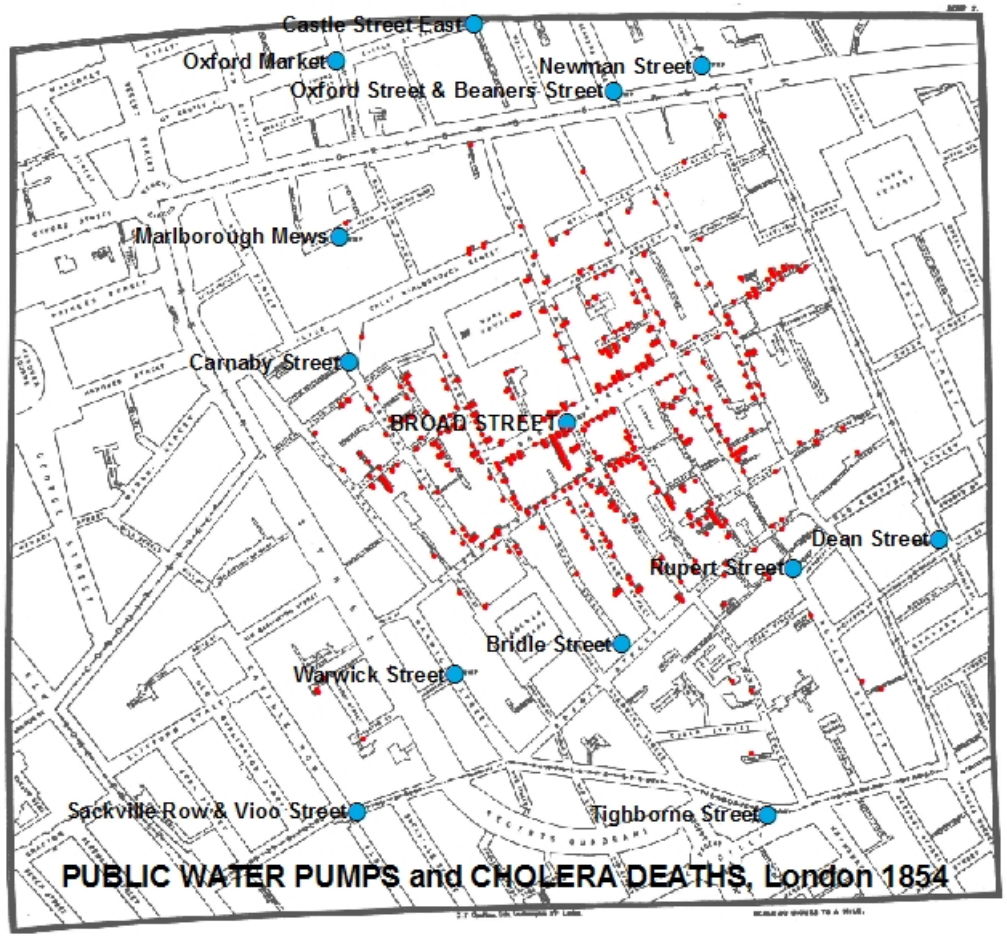
- Clustering refers to a very broad set of techniques for finding subgroups (clusters) in a data set.
- A cluster can be defined as a group of similar objects (cases, points, observations, members, customers, locations etc.).
- The goal of clustering methods is to produce clusters with high intra-cluster (within) similarity and low inter-cluster (between) similarity.

Clustering is different to PCA:

PCA aims at clustering features, whereas cluster methods aim at clustering objects.

Historic application of clustering

Map by John Snow showing the clusters of cholera cases in the London epidemic of 1854.



Economic applications of cluster analysis

- Clustering analysis of world banking sector (Dias and Ramos, 2014)
- Customer segmentation in bank marketing (Machauer and Morgner, 2001)
- Clustering consumer of Internet banking services (Katariina Mäenpää, 2006)

Dias and Ramos (2014): The aftermath of the subprime crisis: a clustering analysis of world banking sector

- Goal: study the behavior of the banking sector of 40 countries during the period 2007-2010 to identify groups of countries with similar profiles.
- Data: banking sector indexes from 40 countries, from the period from July 2007 to October 2010, with a total of 847 end of the day observations per countries.

Dias and Ramos (2014): The aftermath of the subprime crisis: a clustering analysis of world banking sector

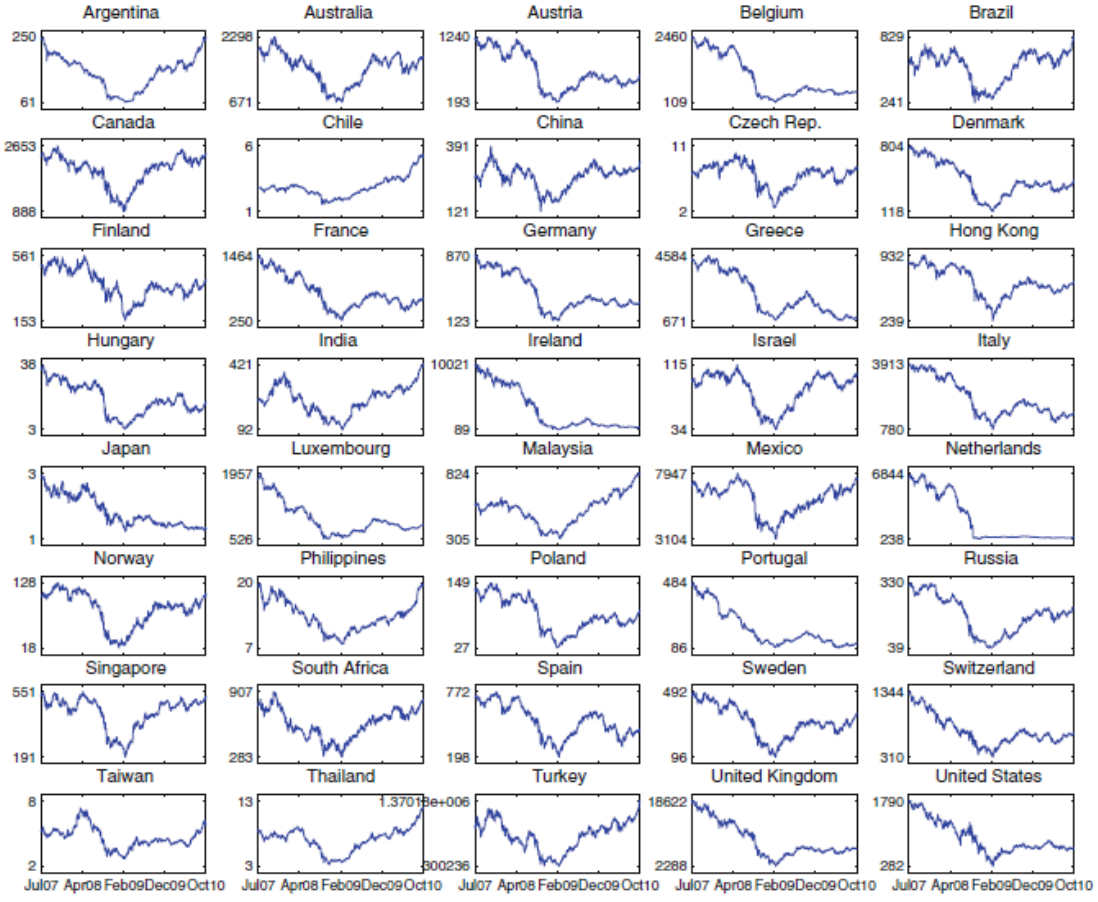
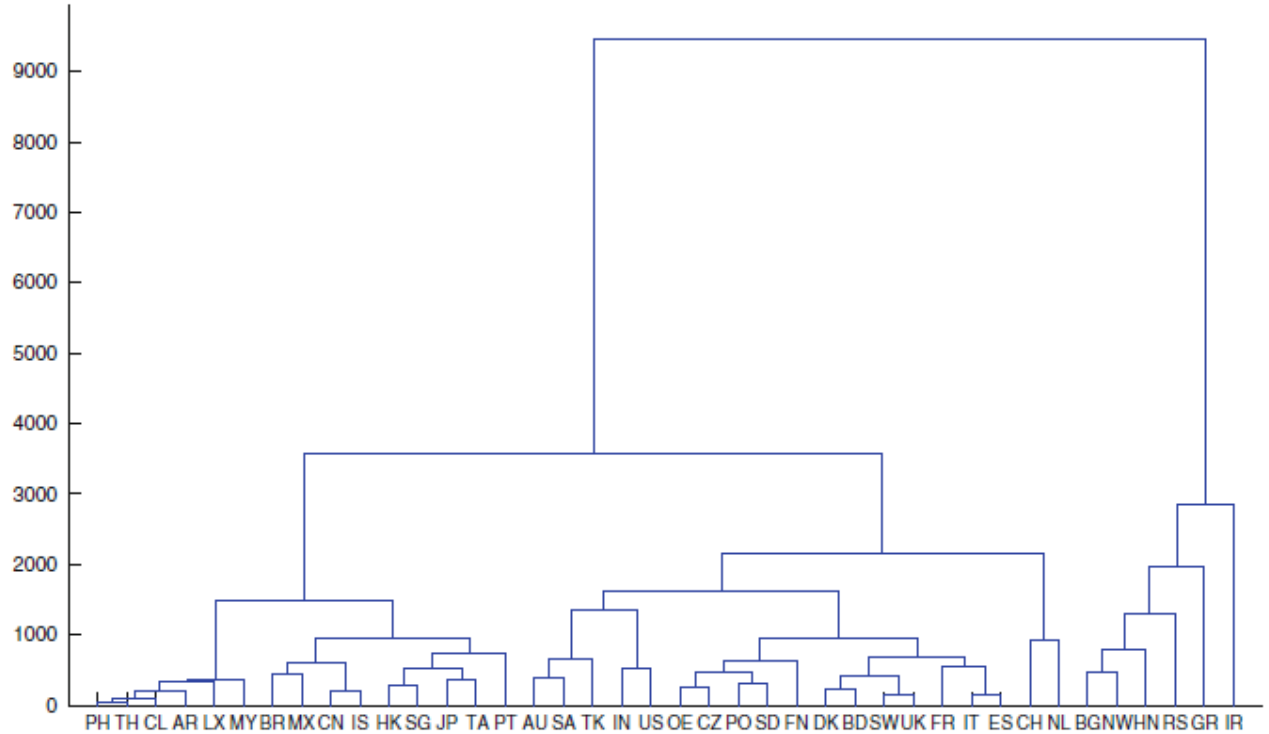


Fig. 1 Time series of banking sector indexes (in USD)

Dias and Ramos (2014): The aftermath of the subprime crisis: a clustering analysis of world banking sector

Dendogram of World banking sector using hierarchical clustering:



Overview cluster methods

A variety of cluster methods exist:

- K-Means clustering
- Hierarchical clustering
- Expectation-maximization clustering
- Mean shift clustering
- Spectral clustering
- ...

K-means and Hierarchical clustering

Here, we focus on the two most commonly used cluster methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram.

K-means and hierarchical clustering are the most commonly used cluster methods, but there exists are variety of domain specific cluster solutions for specific problems.

K-Means Clustering

K-means clustering

- K-means is the simplest and the most common cluster algorithm.
- K-means algorithm is fast and easy to use; it is thus a good solution in a Big Data context.
- Is often used as a preprocessing step for other algorithms, for example to find a starting configuration.
- K-means also has some drawbacks which should be kept in mind.

Algorithm Outline k-means clustering

Algorithm 1: K-Means Clustering

1. Randomly choose k data points (seeds) to be the initial centroids/cluster centers ;
2. Assign each data point to the closest centroid (Euclidean distance) ;
3. Re-compute the centroids using the current cluster memberships. The *kth* cluster centroid is the vector of the p feature means for the observations in the *kth* cluster ;
4. Assign each data point to the closest centroid;
5. Go back to step 3 until no reclassification is necessary;

K-means example, Step 1

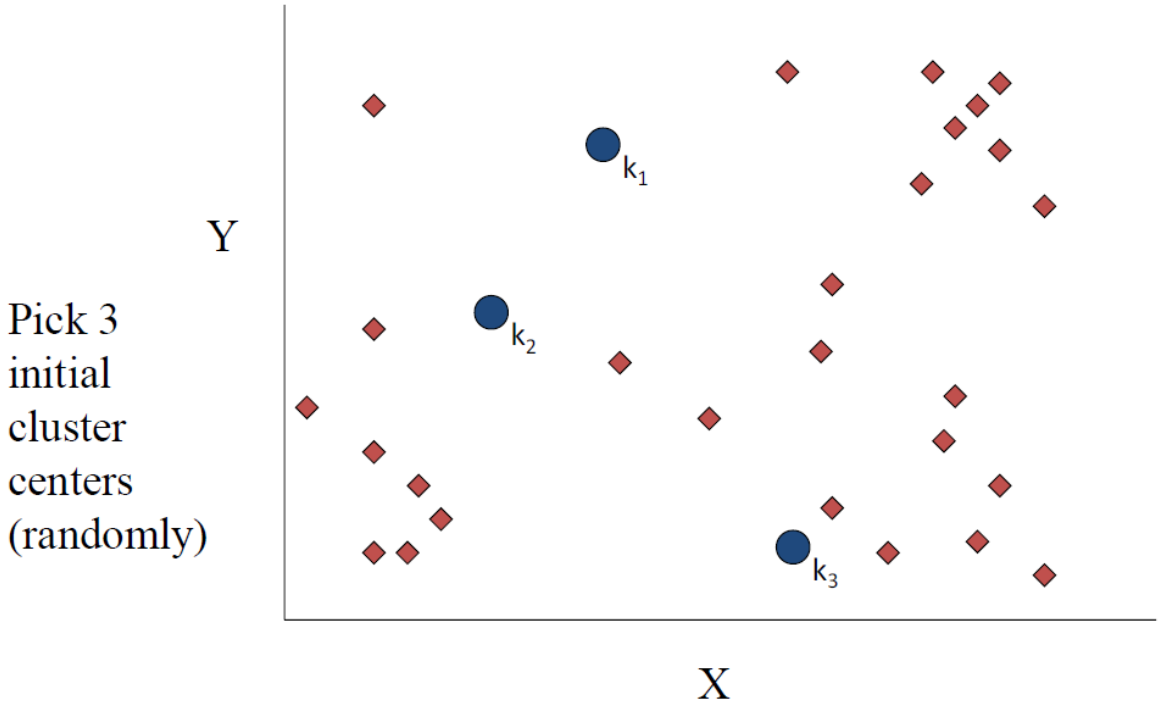


Figure 1: Ghani 2017

K-means example, Step 2

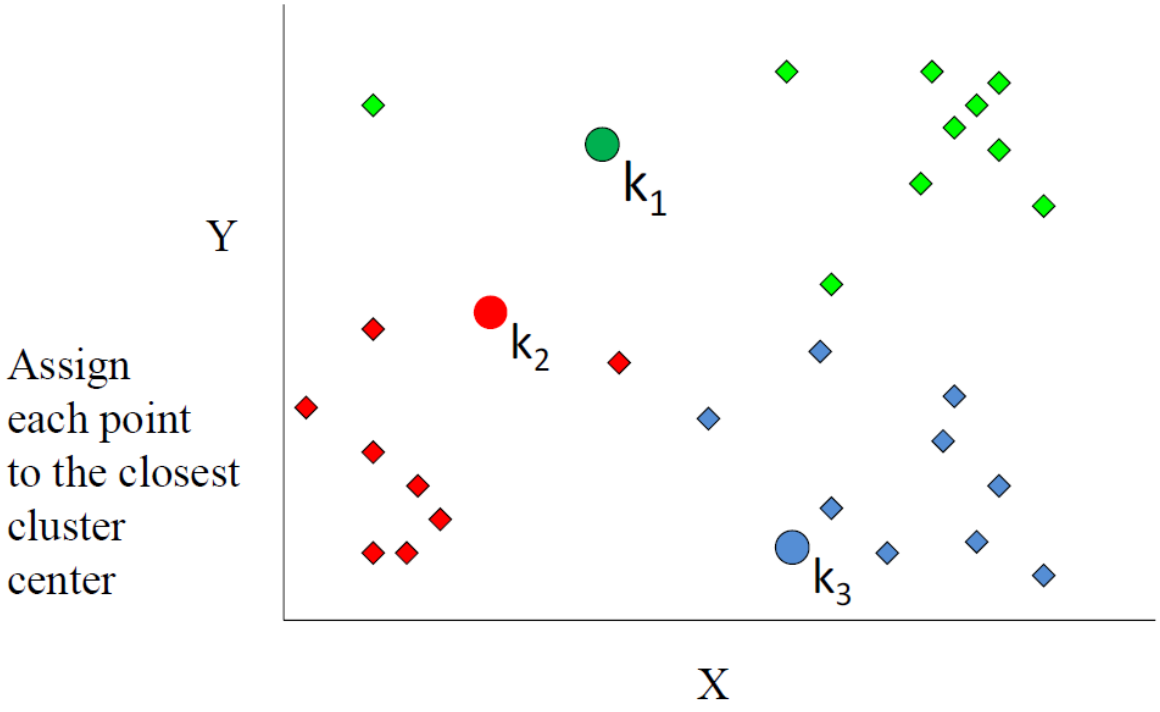


Figure 2: Ghani 2017

K-means example, Step 3

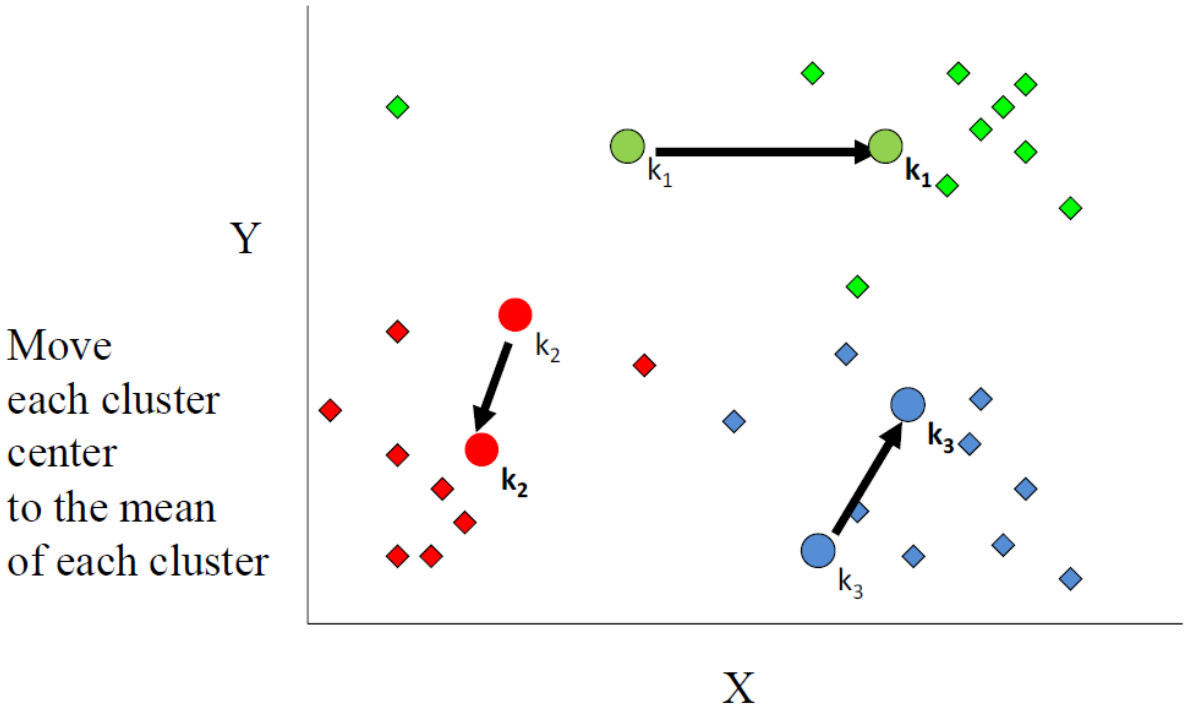


Figure 3: Ghani 2017

K-means example, Step 4a

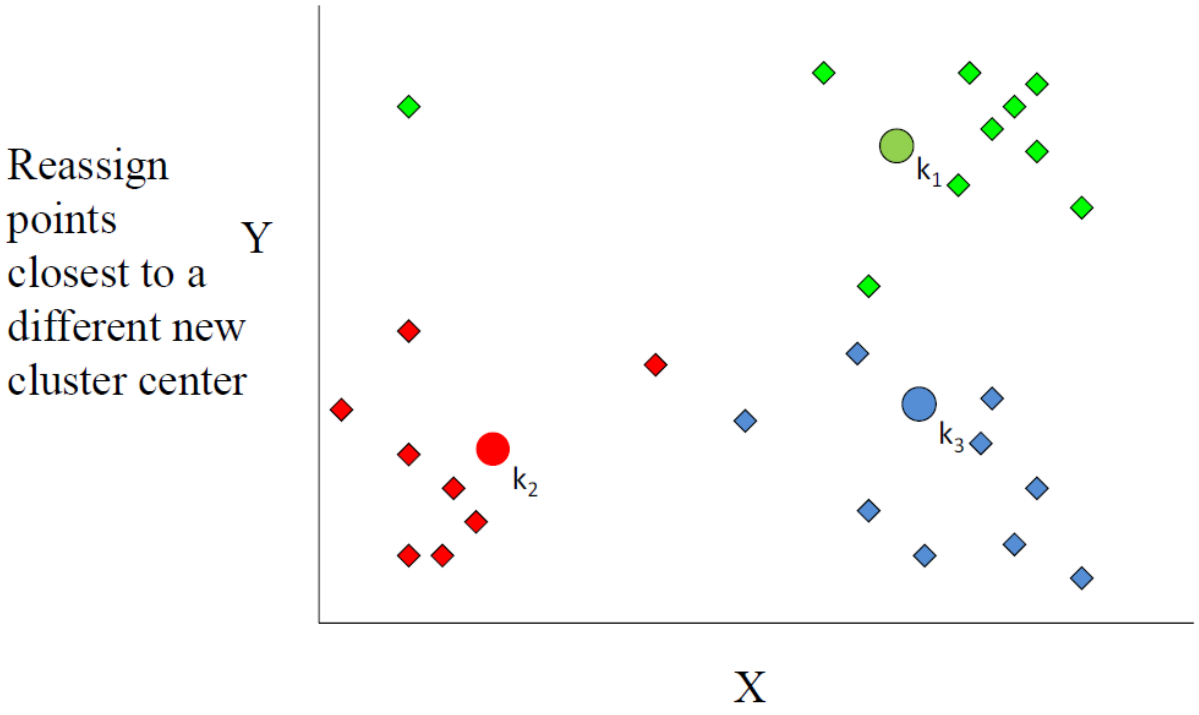


Figure 4: Ghani 2017

K-means example, Step 4a

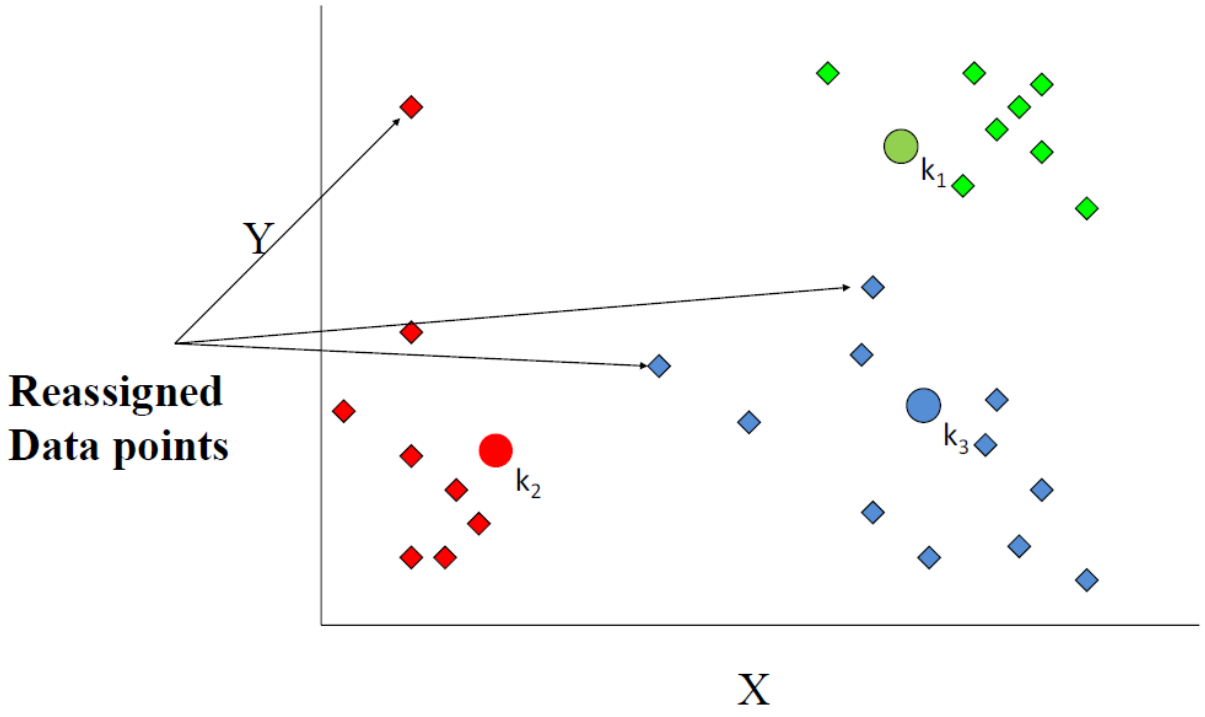


Figure 5: Ghani 2017

K-means example, Step 4b

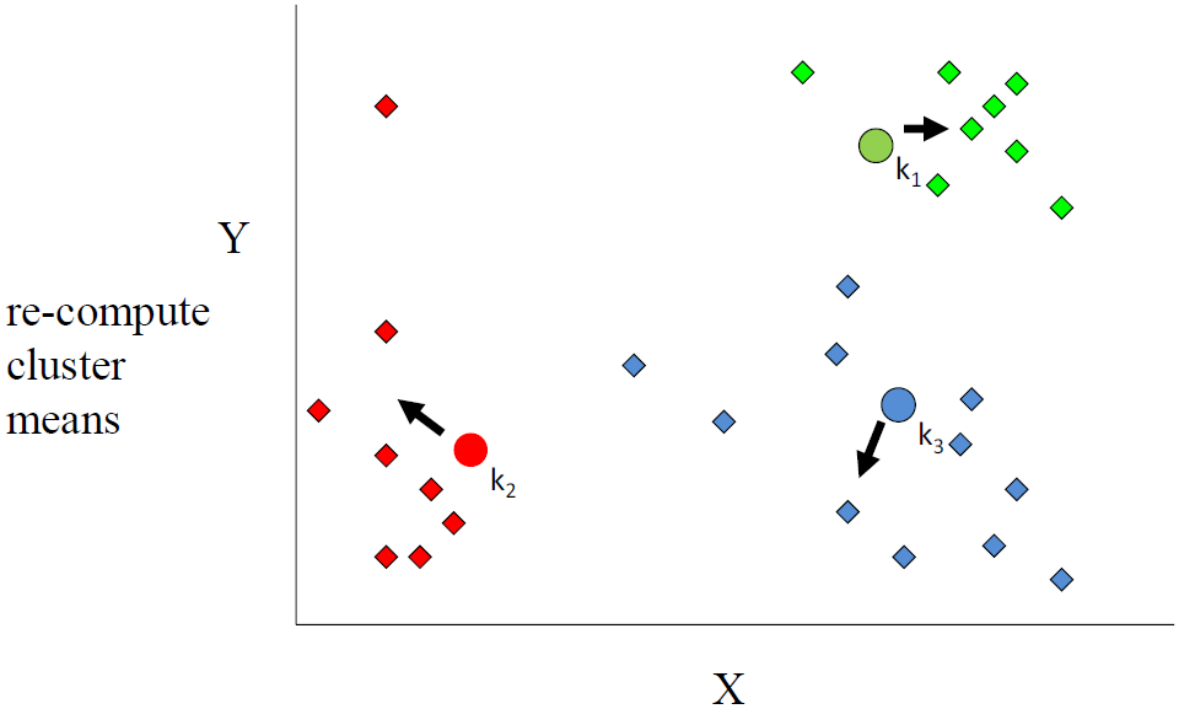


Figure 6: Ghani 2017

K-means example, Step 5

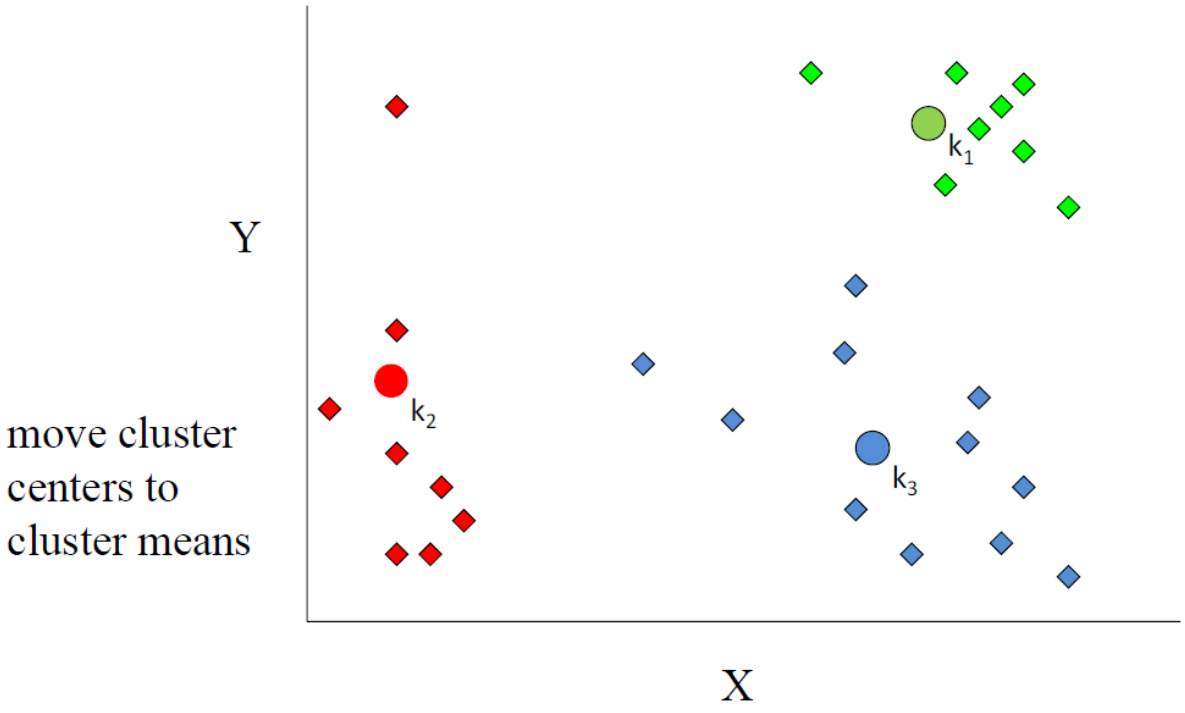


Figure 7: Ghani 2017

Drawback 1: K-means for different values of k

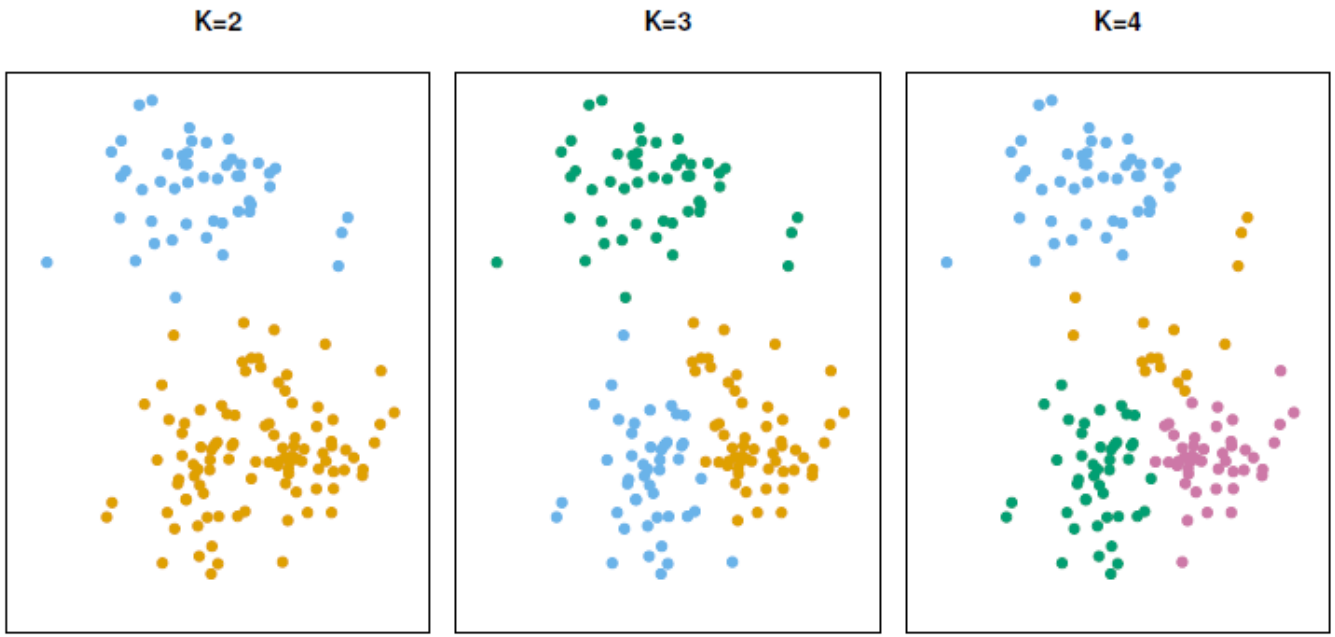


Figure 8: Consequences of different values for k (James et al. 2013)

Drawback 2: K-means for different starting values

Figure 9: Consequences of different starting values (James et al. 2013)

How to deal with these drawbacks?

Practical solution for choice of starting values:

- Run *k*-means several times with different starting values, and take the best solution.

Practical solution for initial choice of *k*:

- Run k-means with different values of *k*.
- Diagnostic checks such as the elbow-method (ratio of the between-group variance to the total variance) .
- Cross validation: does the same *k* yield to good solutions for different subsets of the original data?

Hierarchical Clustering

Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters *K*. This can be a disadvantage.
- Hierarchical clustering does not require that we commit to a particular choice of *K*.
- Here we focus on bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering
- It refers to the fact that a tree-based representation (called *dendogram*) is built starting from the leaves and combining clusters up to the trunk.
- They key idea is to repeatedly combine the two nearest clusters into a larger cluster.

The idea of hierarchical clustering

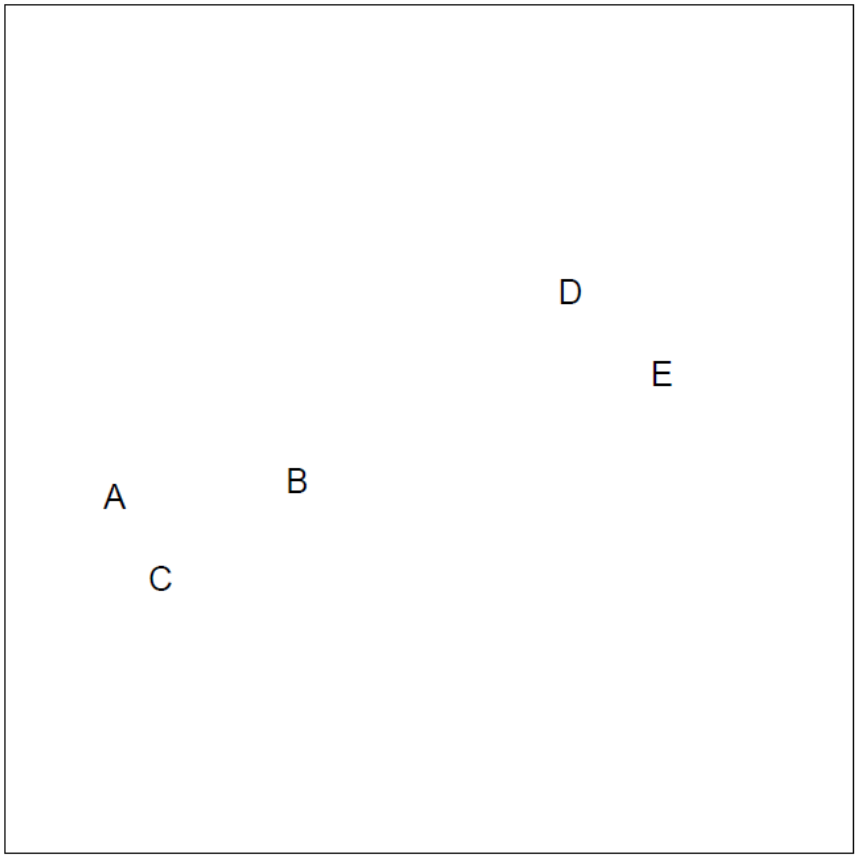


Figure 10: James et al. 2013

The idea of hierarchical clustering

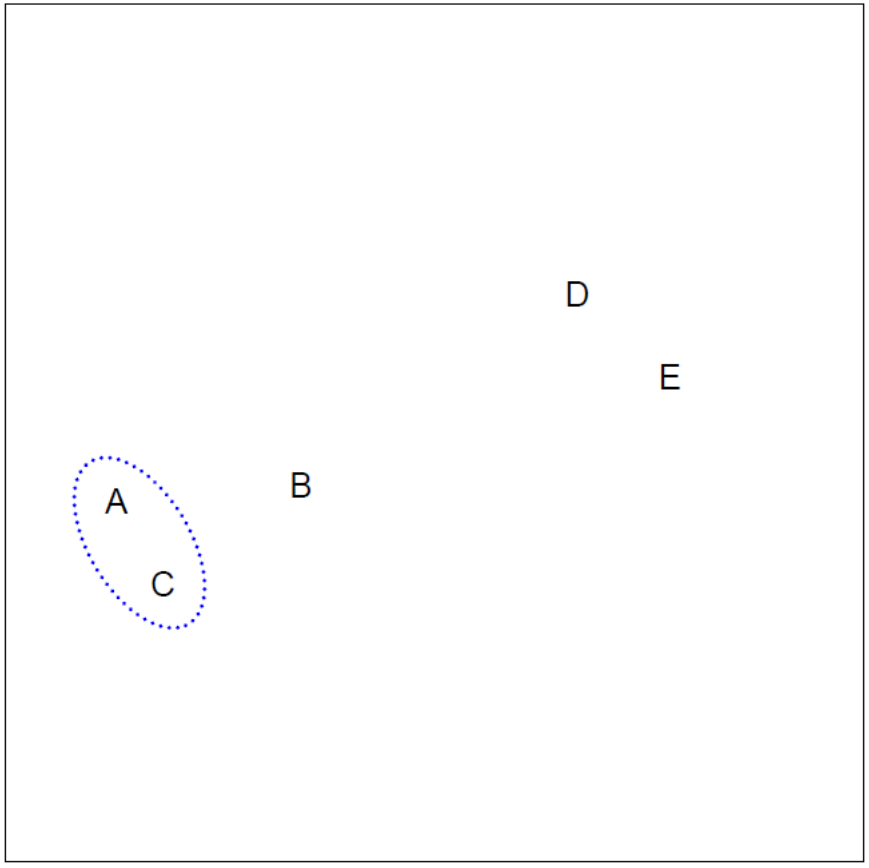


Figure 11: James et al. 2013

The idea of hierarchical clustering

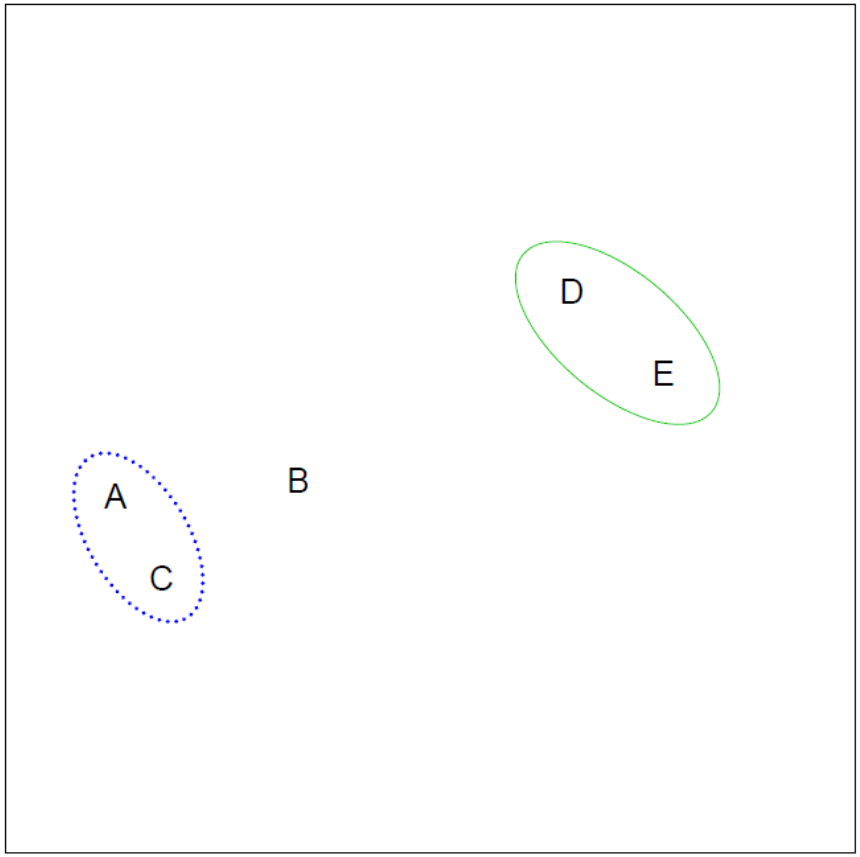


Figure 12: James et al. 2013

The idea of hierarchical clustering

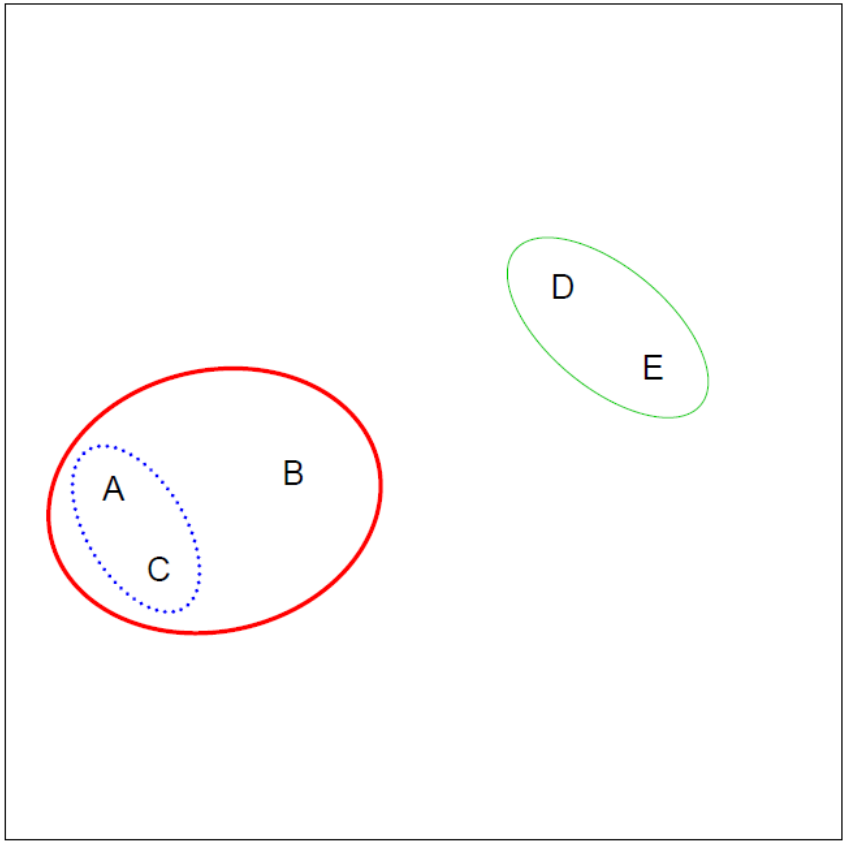


Figure 13: James et al. 2013

The idea of hierarchical clustering

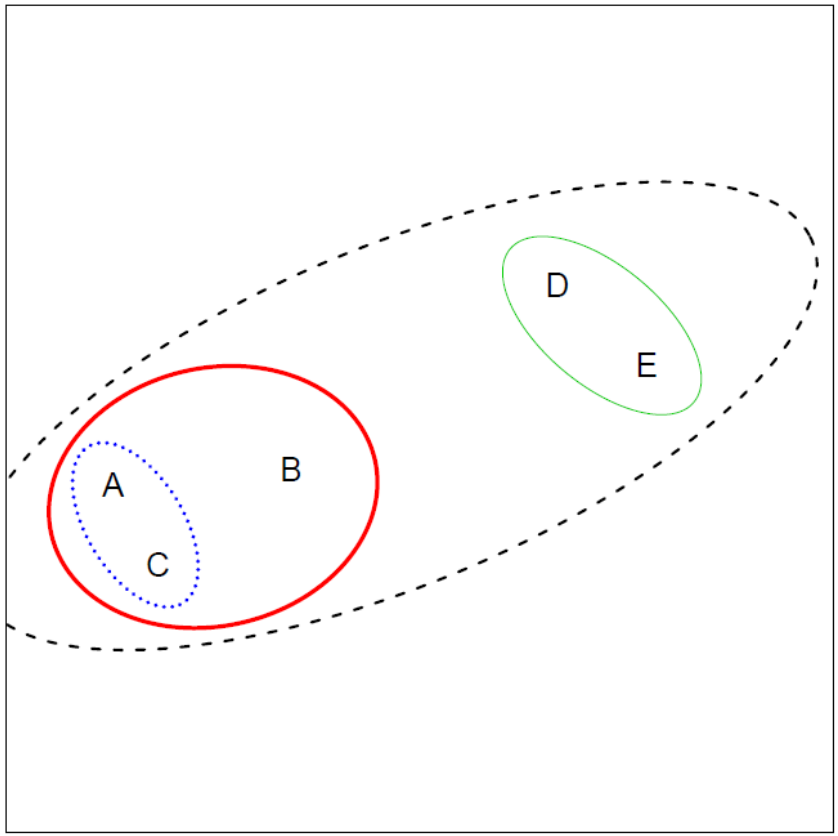
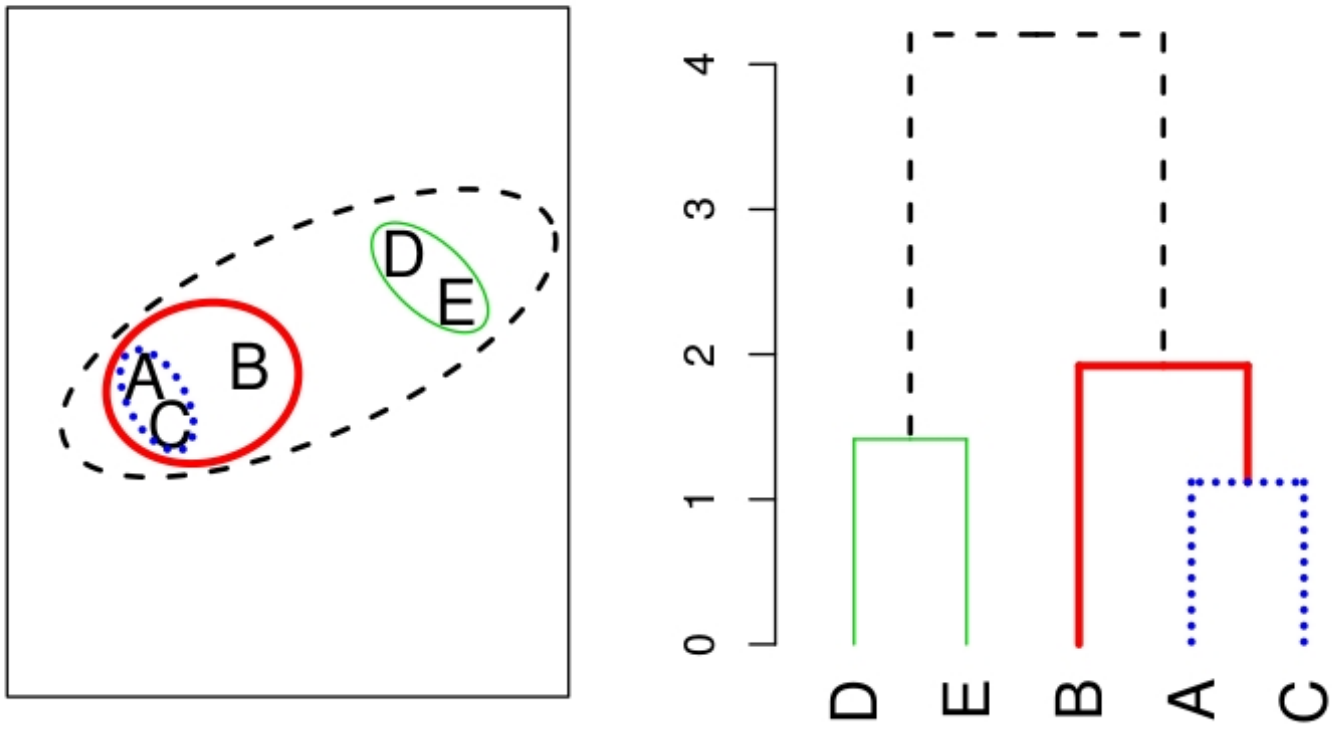


Figure 14: James et al. 2013

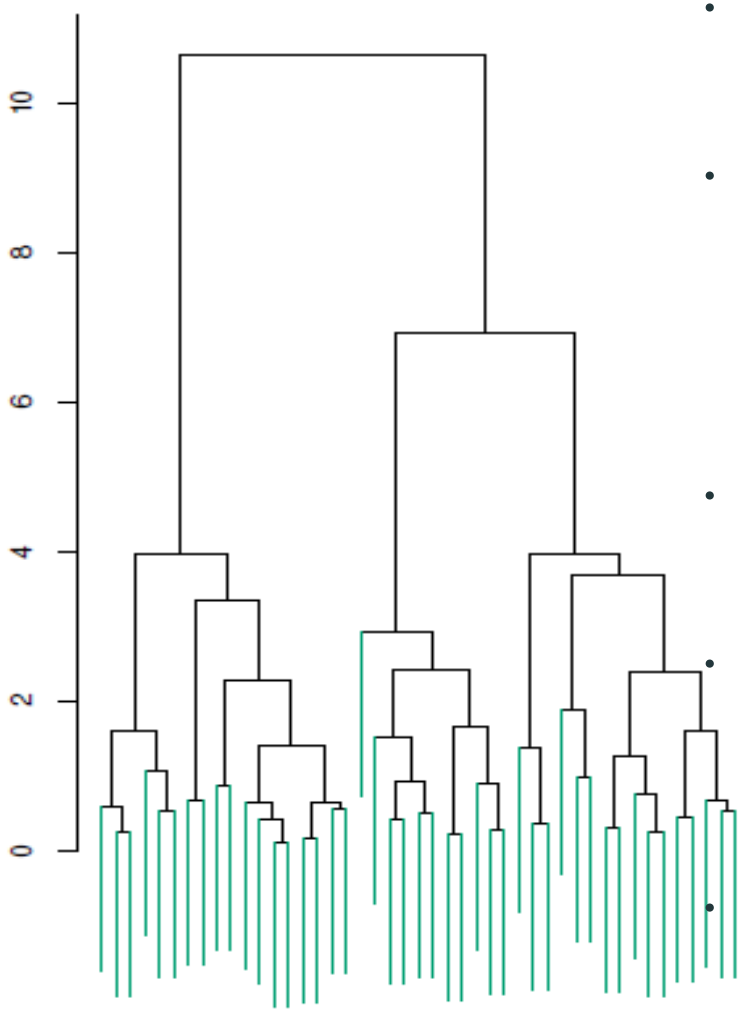
Hierarchical Clustering Algorithm

1. Start by calculating the distance between every pair of observation.
2. Assign each point to its own cluster.
3. Identify the closest two clusters (pair of observations) and merge them to a new cluster.
4. Recompute the distance between the new cluster and the remaining ones.
5. Repeat until all data points are in a single cluster.

Hierarchical Clustering Algorithm

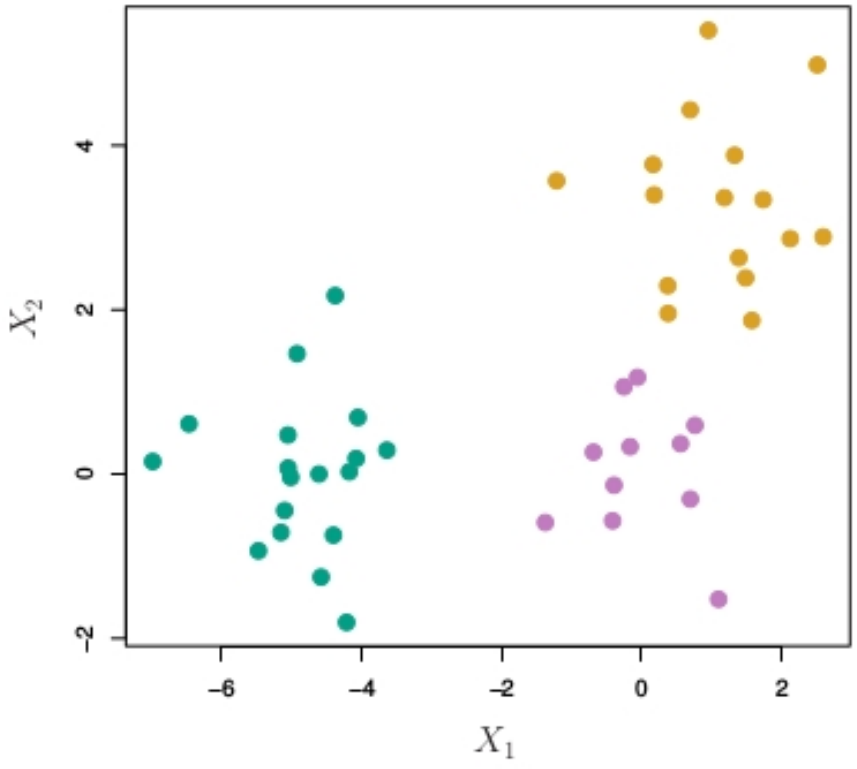


How to interpret a dendogram



- Each leaf represents one observation.
- Moving up the tree, some leaves begin to fuse into branches. These observations are similar to each other.
- As we move higher up the tree, branches themselves fuse.
- The earlier the fusions occur, the more similar the groups of observations are to each other.
- Observations that fuse later are quite different.

A simple example



Different cuts result in different clusters

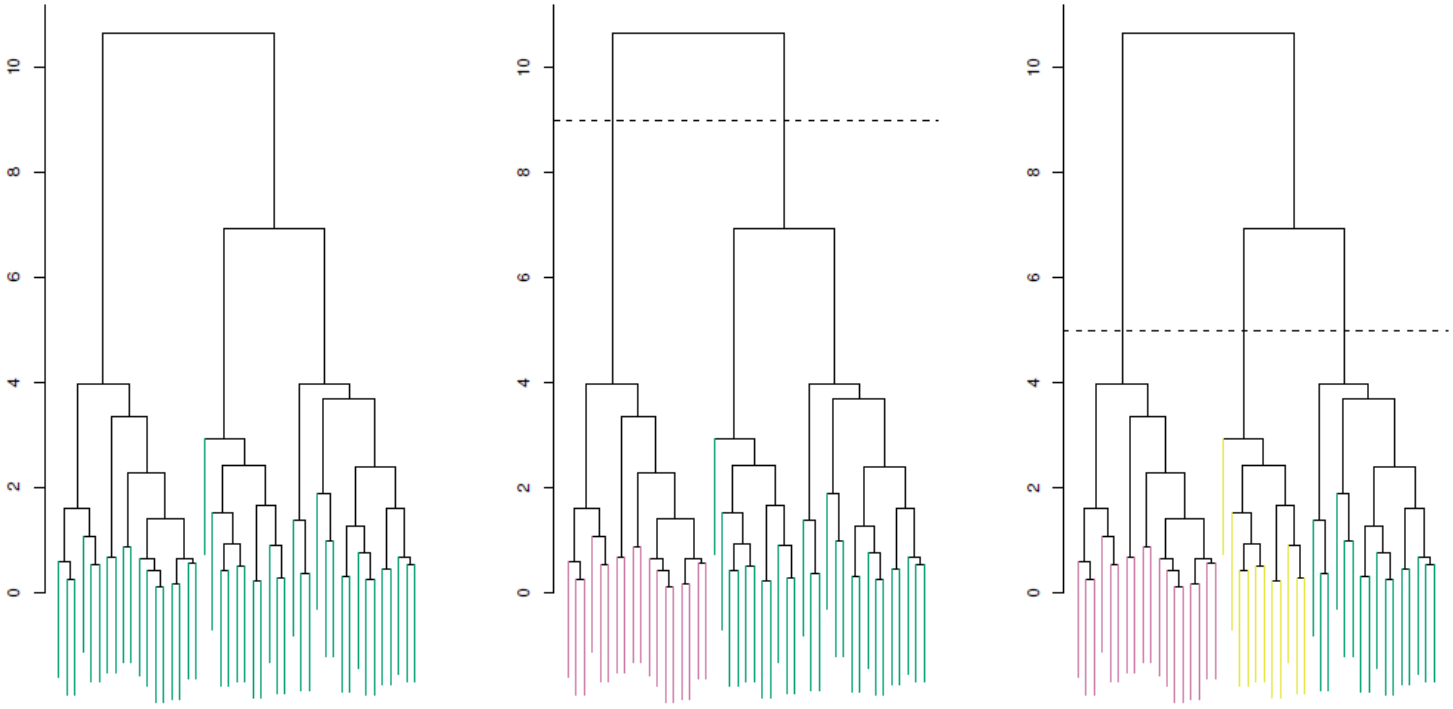
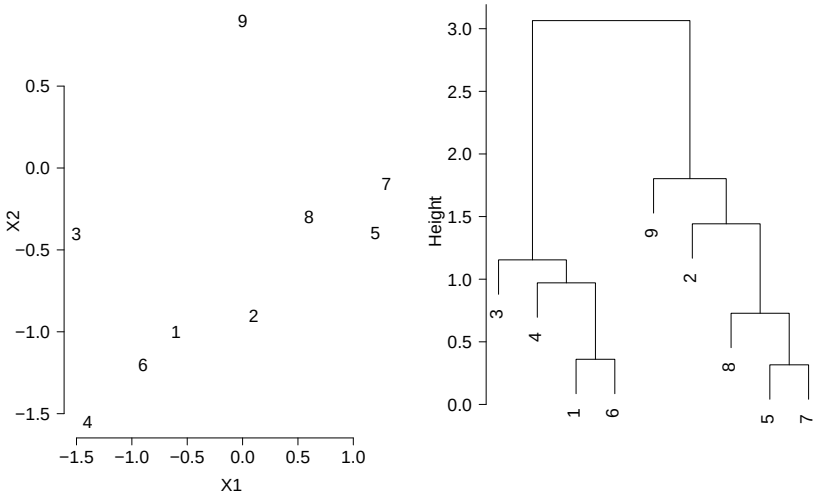


Figure 15: James et al. 2013

A more detailed example



- Observations 5, 7 are similar to each other (so are observ. 1 and 6)
- Observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7, because all fuse with observation 9 at the same height, approximately 1.8.
- Similarity of observations based on proximity of vertical axis, not horizontal axis.

How to identify clusters based on dendrogram?

- Cuts at a certain height result in clusters.
- The height of the cut to the dendrogram serves the same role as the K in K-means clustering: it controls the number of clusters obtained.
- Attractive aspect of hierarchical clustering: one single dendrogram can be used to obtain any number of clusters.
- Decision about number of clusters must be made by researcher (based on visual inspection).

Merges in previous example

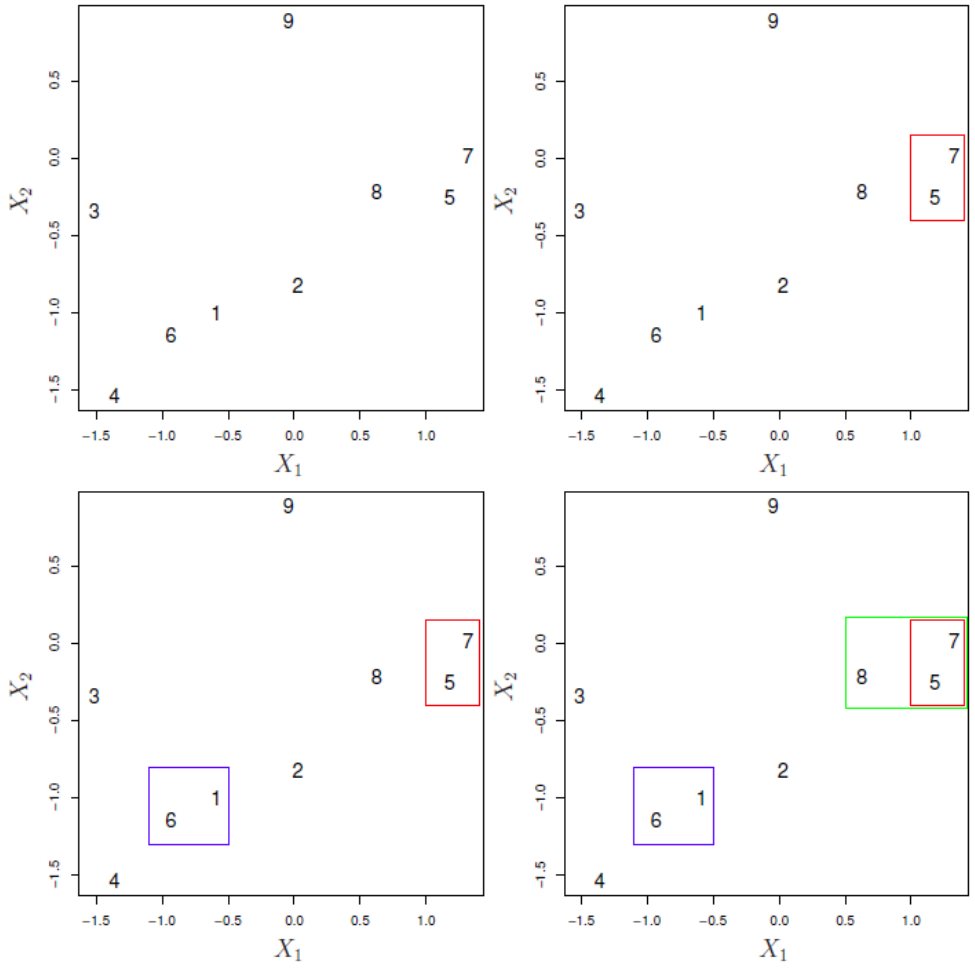


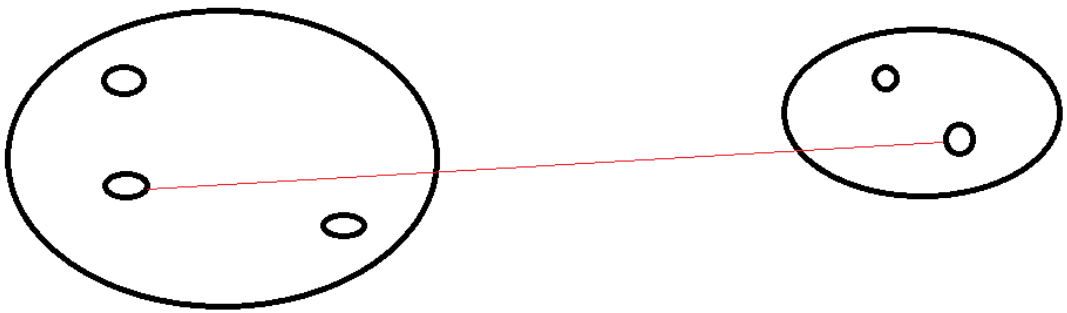
Figure 16: James et al. 2013

Types of Linkage: Cluster Dissimilarity

- How did we determine that the cluster $\{5, 7\}$ should be fused with the cluster $\{8\}$? How can we measure dissimilarity between clusters containing multiple observations?
- The concept of dissimilarity between a pair of observations needs to be extended to a pair of groups of observations. This extension is achieved by developing the notion of linkage, which defines the dissimilarity between two groups of observations.
- Average, complete, and single linkage are most popular choices.
- Average and complete linkage are generally preferred over single linkage, as they tend to result in more balanced dendrograms.

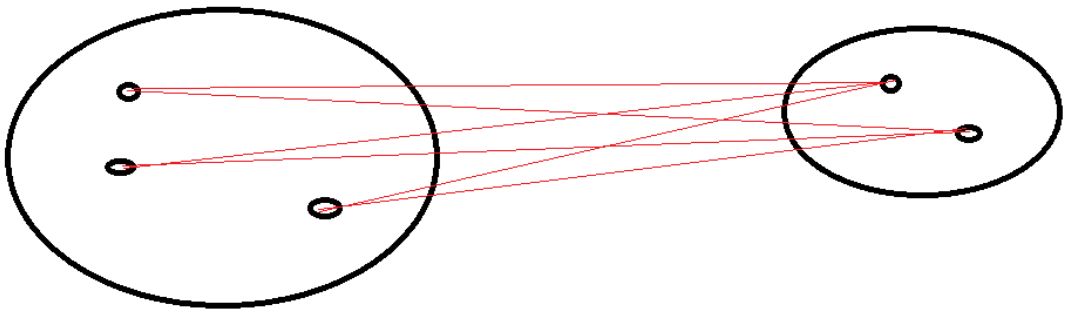
Types of Linkage: Complete linkage

Complete linkage: calculates the maximum distance between clusters before merging.



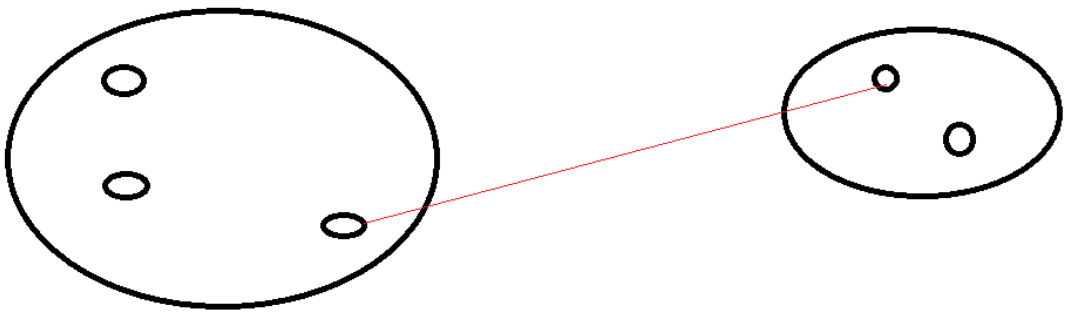
Types of Linkage: Average linkage

Average linkage: calculates the average distance between clusters before merging.



Types of Linkage: Single linkage

Single linkage: calculates the minimum distance between the clusters before merging.



The visual difference of different linkages

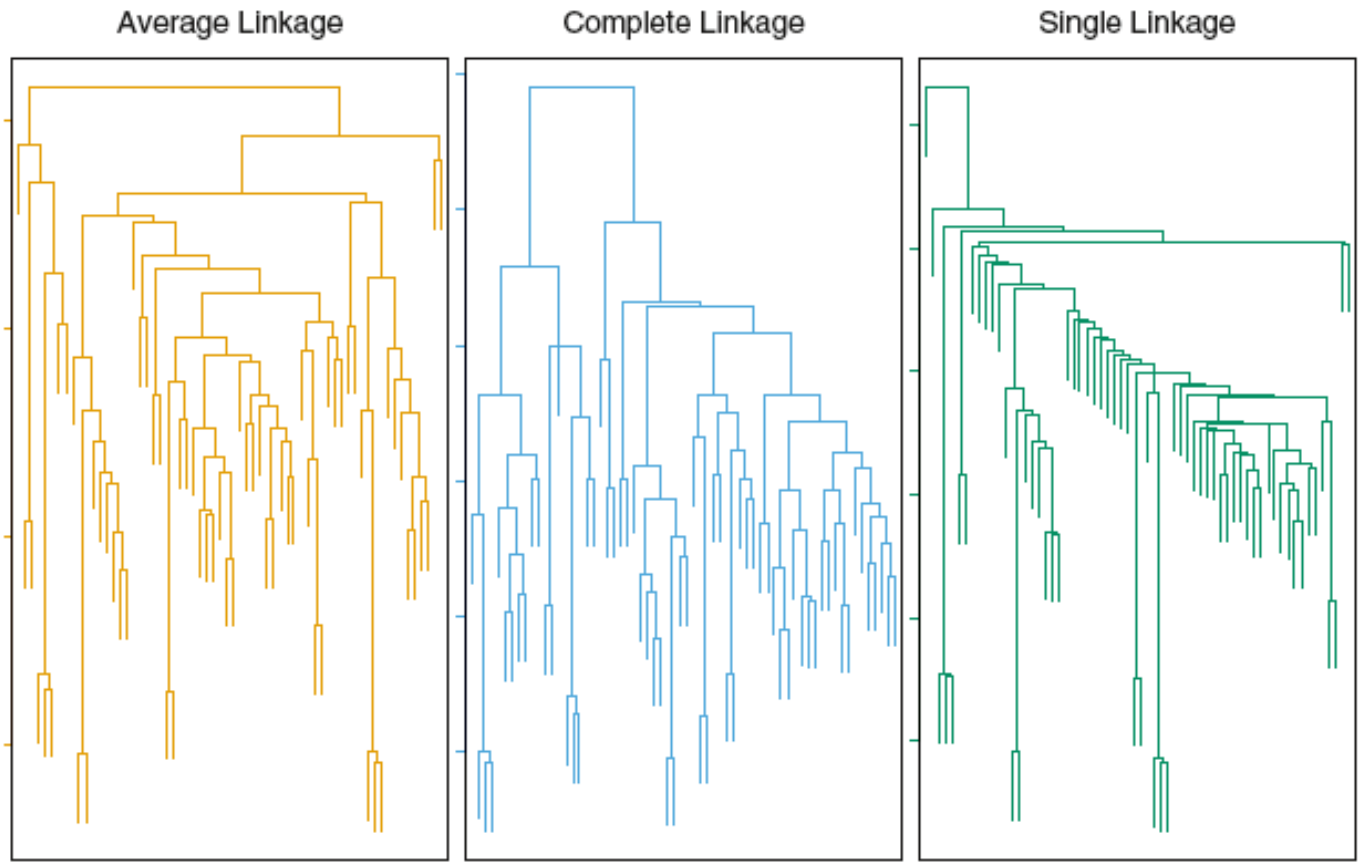


Figure 17: James et al. 2013

Choice of Dissimilarity Measure

- Euclidean distance is most frequently used.
- An alternative is the correlation-based distance which considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance.
- This is an unusual use of correlation: normally it is computed between variables; here it is computed between observation profiles for each pair of observations.
- Choice of dissimilarity measure is very important, as it has a strong effect on the resulting dendrogram.

An example where it made a difference

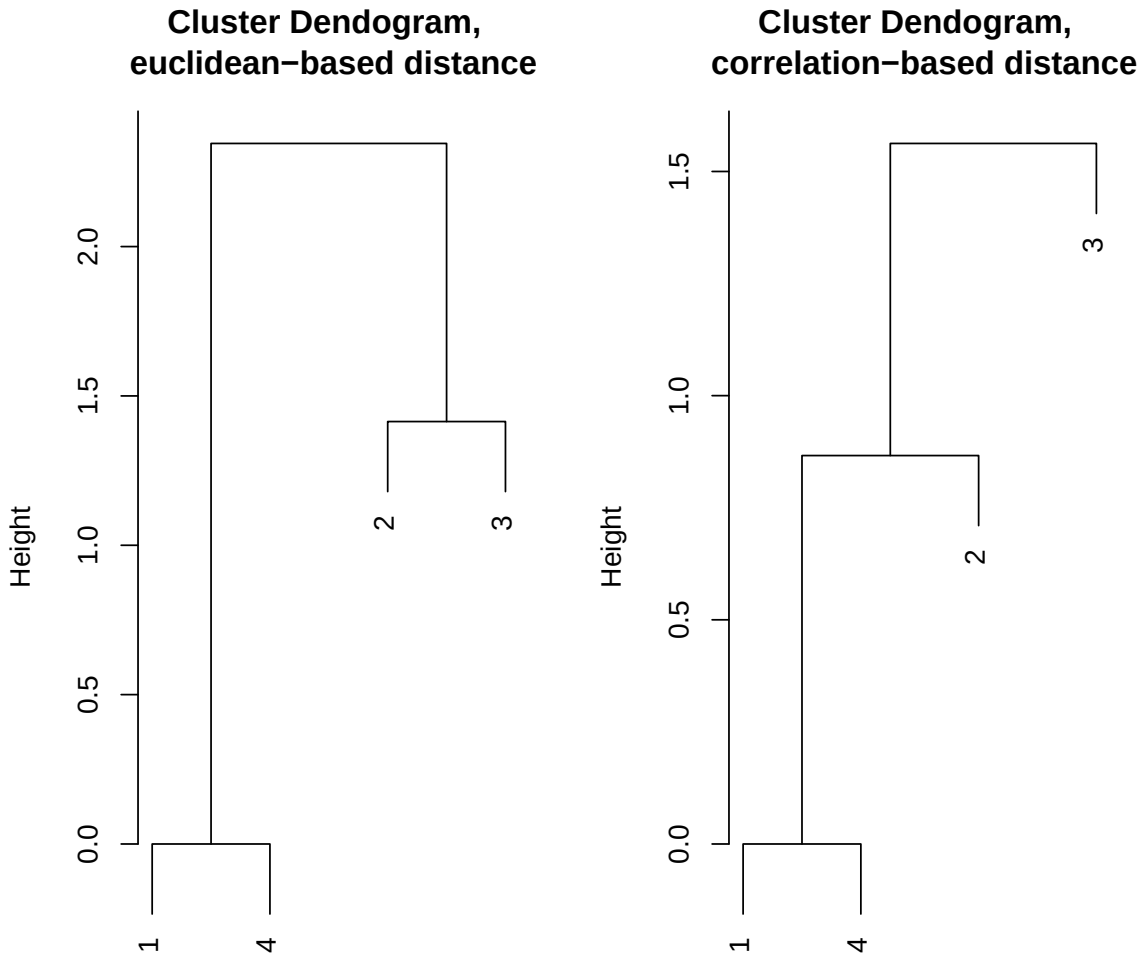
- Consider an online retailer interested in clustering shoppers based on their past shopping histories.
- Goal: identify subgroups of similar shoppers, so that shoppers within each subgroup can be shown items and advertisements that are likely to interest them.
- Rows are the shoppers, columns are the items; the elements of the data matrix indicate the number of times a given shopper has purchased a given item.

	V1	V2	V3	V4	V5
id.1	2	0	0	1	1
id.2	0	0	0	1	0
id.3	0	0	1	0	0
id.4	2	0	0	1	1

An example where it made a difference

- If Euclidean distance is used, then shoppers who have bought very few items overall (i.e. infrequent users of the online shopping site) will be clustered together. This may not be desirable.
- If correlation-based distance is used, then shoppers with similar preferences (e.g. shoppers who have bought items A and B but never items C or D) will be clustered together, even if some shoppers with these preferences are higher-volume shoppers than others.
- Therefore, for this application, correlation-based distance may be a better choice.

Choice of Dissimilarity Measure



Practical issues with cluster analysis

Practical issues with cluster analysis

- Should the observations or features first be standardized in some way?
- In the case of hierarchical clustering:
  - Which dissimilarity measure should be used?
  - What type of linkage should be used?
- In the case of K-means clustering, what is the best K for the data set?
- Did cluster algorithm really found true subgroups, or are the obtained clusters a result of clustering the noise?
- Outliers can heavily distort clusters, because the belong to no group but are forced into one.

There is no single right answer to these questions. Clustering should be performed with different choices of parameters, and for different subsets of the data.

Practical Session Clustering

In the practical session you will...

- learn how k-means and hierarchical cluster analysis is done in R.
- apply both cluster methods to the Immoscout data set.

Resources I

Resources I

- Books
  - Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.
  - James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.
  - Kassambara, Alboukadel. 2017. *Practical Guide To Cluster Analysis in R*. STHDA Publishing
- Book chapter
  - Ghani, R., Schierholz, M. (2017). Machine Learning. In: Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., Lane, J. (Eds.). (2017). *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press Taylor & Francis Group.

References

Dias, José G. and Sofia B. Ramos. 2014. "The aftermath of the subprime crisis: A clustering analysis of world banking sector." *Review of Quantitative Finance and Accounting* 42(2):293–308.

Katariina Mäenpää. 2006. "Clustering the consumers on the basis of their perceptions of the Internet banking services." *Internet Research* 16(3):304–322.

Machauer, Achim and Sebastian Morgner. 2001. "Segmentation of bank customers by expected benefits and attitudes." *International Journal of Bank Marketing* 19(1):6–18.