Applied Statistics and Data Analysis

7. Unsupervised methods

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine

paolo.vidoni@uniud.it

Partly based on Chapter 10 of *An Introduction to Statistical Learning: with Applications in R* by G. James et al.

Table of contents

Summary

- Introduction to unsupervised learning
- Principal components analysis
- Cluster analysis

Introduction to unsupervised learning

- Most statistical learning problems fall into one of two categories: supervised or unsupervised.
- The problems discussed so far belong to the **supervised learning** framework: for each observation $\mathbf{x}_i$, $i = 1, \ldots, n$, of the p predictor variables there is an observation y_i for the associated response variable.

  A model that relates the response to the predictors is fitted and it can be considered for both interpretation and prediction purposes. Linear regression and logistic regression models operate in this context.
- The **unsupervised learning** framework is, in some sense, more challenging, since for every observation $i = 1, \ldots, n$, only a vector of measurements $\mathbf{x}_i$ is given, with no associated response y_i.

  This situation is unsupervised because a response variable, that can supervise the analysis, is not available. There is no way to check the results by seeing how well the model predicts a response.

- The focus here is on unsupervised learning, which is concerned with the *joint* study of a set of p variables, $X_1, \ldots, X_p$, observed for a sample of size n.

  The **data matrix** is then given by $\mathbf{X}$ and it has size $n \times p$; the element x_{ij} corresponds to the i-th observation on the j-th variable.

  There is no response variable, and all the p variables are treated on an equal footing.
- The goal is to discover interesting things about the measurements on the p variables $X_1, \ldots, X_p$ and, in particular, if there is an informative way to visualize and to summarize the data and/or if there are subgroups among the variables or among the observations.
- The statistical techniques considered in this framework are typically **explorative** in nature. They were classically referred to as *multivariate analysis* techniques, but this name is perhaps too vague.

- Unsupervised learning is often performed as part of an exploratory data analysis.

  This exercise tends to be somewhat subjective, since there is no simple goal for the analysis, such as prediction of a response.
- Like any other statistical techniques, the first step for analyzing multivariate data is given by data description and visualization, that can be far from easy in high dimensions.
- For a first look at the data, scatterplot matrices and low-dimensional view are always useful, yet to consider a number of plots can miss important structures in the data.

  More advanced techniques, such as **dynamic graphics**, can be quite effective, as they allow for rotation of the point cloud and projection in lower dimensions.
- There are many unsupervised techniques, and the focus here on two classes of methods: **Principal Components Analysis** and **Cluster analysis**.

Table of contents

Introduction to Principal Components Analysis

- **Principal Components Analysis** (**PCA**) replaces the input variables by a set of new derived variables, called **principal components**.
- These components are ordered according to the amount of variation of the original variables they are able to explain.
- The method is useful for understanding multivariate data and it also serves as a tool for data visualization.

  Plots of the first principal components are often insightful, leading to effective **dimension reduction**, useful in particular when p is large.

  PCA gives a low-dimensional representation of the data, based on small number of interesting dimensions, that captures as much of the information as possible.
- This idea is at times useful in regression, where a large number of candidate explanatory variables may be replaced by the first few principal components, provided they synthesize adequately the information in the candidate variables.

The principal components

- Given a set of p variables $X_1, \ldots, X_p$, the **first principal component** Z_1 is the *normalized* linear combination

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

  with the *largest variance*.
- The weights $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})^T$ are called **loadings** and they are normalized, namely $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

  This constraint is required since setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.
- The principal components depend on the scaling of the variables, then it is important that the variables are in comparable units. Thus, it is typically recommendable to **standardize** the variables before applying the method.

- Obtaining the loadings for the first principal component is a simple linear algebra task.
- Given the $n \times p$ data matrix $\mathbf{X}$, the aim is to derive the loadings $\phi_{11}, \ldots, \phi_{p1}$ such that the linear combination of the n sample values

$$z_{i1} = \phi_{11}\, x_{i1} + \phi_{21}\, x_{i2} + \ldots + \phi_{p1}\, x_{ip}$$

has largest variance, subject to the constraint $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

Since the variables are centered (that is the column means of $\mathbf{X}$ are zero), the sample variance is (the sample mean is zero)

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$$

- After the loading vector ϕ_1 is found, the observed values for the first principal component Z_1 are obtained.

These values, one for each observation, are z_{i1}, $i = 1, \ldots, n$, and they are referred to as the **scores** of the first principal component.

- The **second principal component** Z_2 is the *normalized* linear combination of $X_1, \ldots, X_p$ that has maximal variance out of all linear combinations that are *uncorrelated* with Z_1.
- The second principal component scores $z_{12}, \ldots, z_{n2}$ take the form

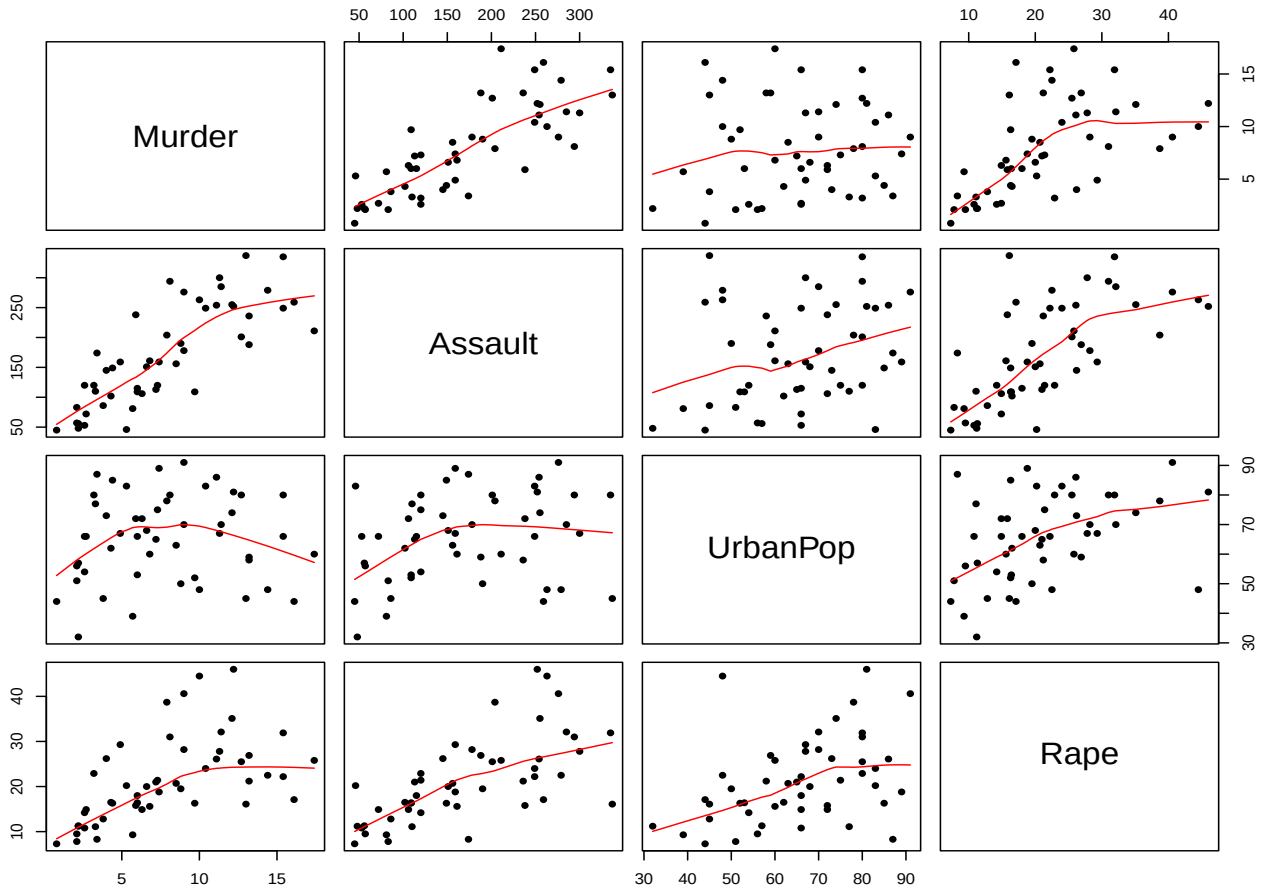
$$z_{i2} = \phi_{12}\, x_{i1} + \phi_{22}\, x_{i2} + \ldots + \phi_{p2}\, x_{ip}$$

  where the second principal loading vector $\phi_2 = (\phi_{12}, \ldots, \phi_{p2})^T$ is such that $\sum_{j=1}^{p} \phi_{j2}^2 = 1$ and $\sum_{j=1}^{p} \phi_{j1}\phi_{j2} = 0$ (uncorrelation).
- The procedure can be iterated, up to a $m = \min(n-1, p)$ principal components.
- **Geometric interpretation**: the first loading vector ϕ_1 defines a direction in the variable space along which the data vary the most; projection of the n data points onto this direction gives the first principal component scores.

  The second loading vector ϕ_2 defines a direction, orthogonal (perpendicular) to the direction ϕ_1, along which the variability is maximized; data projections onto this further direction gives the second principal component scores.

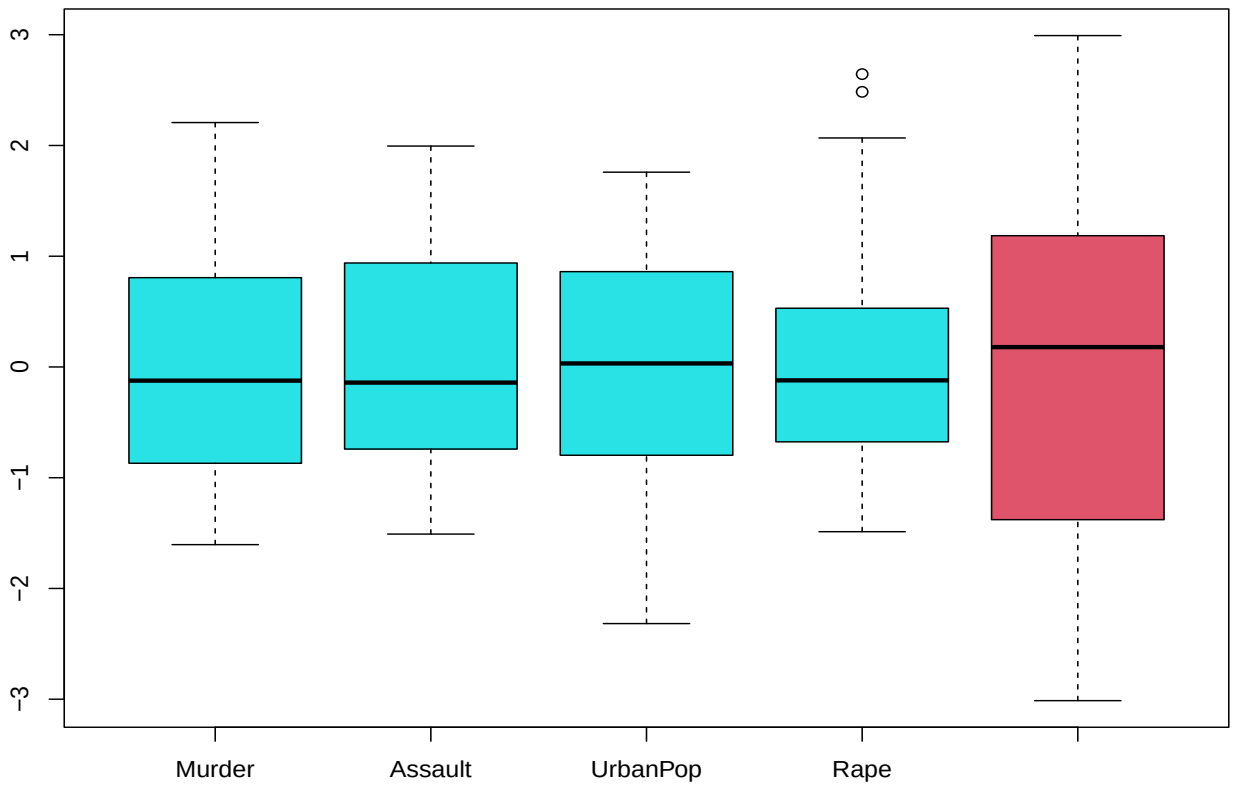
Example: US arrest data

Data set containing statistics, in arrests per 100,000 residents for `Assault`, `Murder`, and `Rape` in each of the 50 US states in 1973, and the percent of the population (`UrbanPop`) living in urban areas (4 variables and $n = 50$ observations).



A boxplot of the four (standardized) variables compared with the scores from the first principal component readily shows that the latter is more variable.

Using the first principal component, some global information has been extracted.



The first two principal component loading vectors, ϕ_1 and ϕ_2, are given below

	Murder	Assault	UrbanPop	Rape
PC1	0.5358995	0.5831836	0.2781909	0.5434321
PC2	-0.4181809	-0.1879856	0.8728062	0.1673186

The first loading vector places approximately equal weight on `Murder`, `Assault`, and `Rape`, with much less weight on `UrbanPop`.

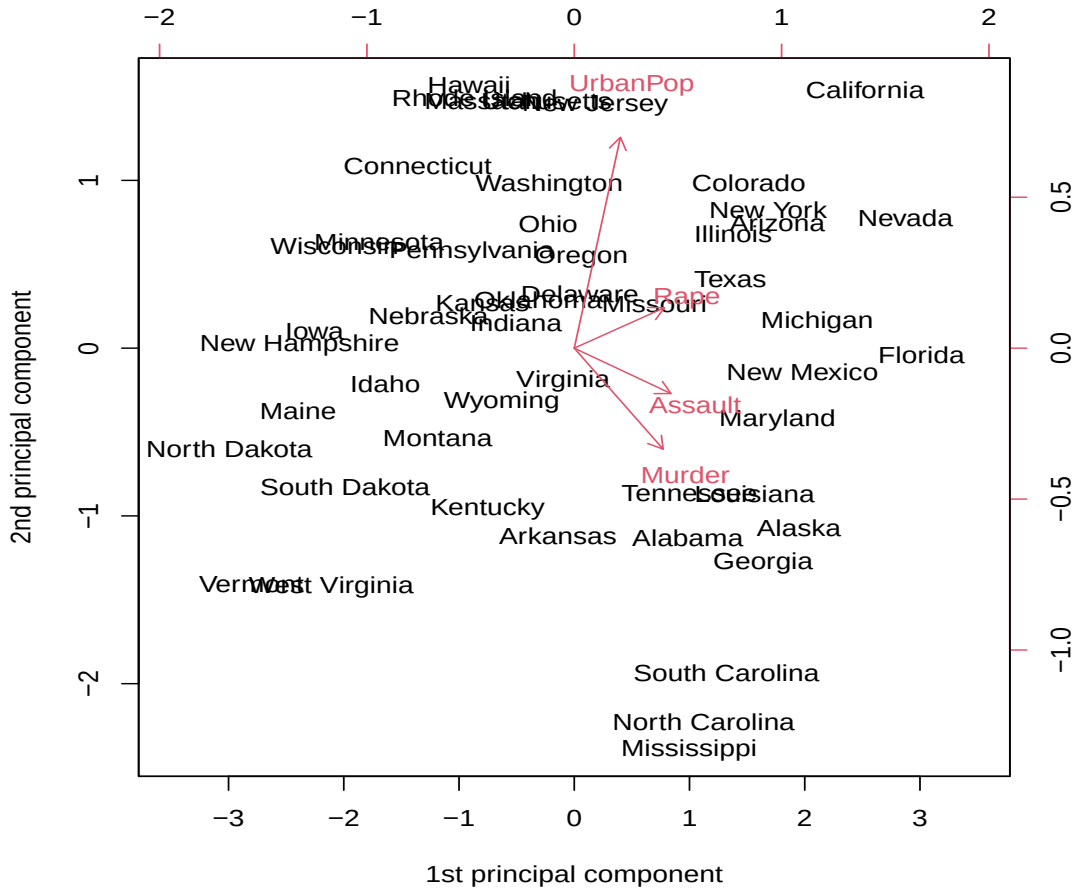
Then the first principal component Z_1 roughly corresponds to a measure of overall rates of serious crimes.

The second loading vector places most of its weight on `UrbanPop` and much less weight on the other three variables.

Hence, the second principal component Z_2 roughly corresponds to the level of urbanization of the state.

The **biplot** is a graphical summary which displays both the scores and the loadings associated to the first two principal components.

The **state names** represent the scores for the first two principal components (left and down axes), while red arrows indicate the first two loadings associated to the four variables (top and right).



The crime-related variables (`Murder`, `Assault` and `Rape`) are located close to each other, and the `UrbanPop` variable is far from the other three.

This indicates that the crime-related variables are correlated with each other (states with high murder rates tend to have high assault and rape rates) and that the `UrbanPop` variable is less correlated with the other three.

States with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states like North Dakota, with negative scores on the first component, have low crime rates.

California also has a high score on the second component, indicating a high level of urbanization.

States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

Further comments

- **Uniqueness of the principal components**: each component is unique up to a sign flip.
  - Flipping the sign to all the estimated loadings gives totally equivalent results, since the direction specified in the p-dimensional space is the same.
  - Similarly, the score vectors are unique up to a sign flip, since the variance of Z is the same as the variance of $-Z$.
- **Alternative geometrical interpretation**: the first r principal components score and loading vectors provide the best r-dimensional approximation (using the Euclidean distance in $\mathbf{R}^p$) to the i-th observation, namely $x_{ij} \approx \sum_{s=1}^{r} z_{is}\phi_{js}$; the approximation is exact with $r = \min(n-1, p)$.
- **How many principal components**: the first component is the most informative one-dimensional linear summary, and the first two provide the most informative two-dimensional graphical summary.

  Under this respect, a key point is how much of the variance in the data is extracted by a given number of principal components.

The proportion of variance explained

- Assuming that the observed variables have mean zero, the **proportion of variance explained** by the s-th component is

$$\text{PVE}_s = \frac{(1/n)\sum_{i=1}^{n} z_{is}^2}{\sum_{j=1}^{p}(1/n)\sum_{i=1}^{n} x_{ij}^2},$$

  The numerator is the (estimated) variance of the s-th component and the denominator estimates the total variance $\sum_{j=1}^{p} V(X_j)$.

  The cumulative proportion of variance explained for the first r components is then $\sum_{s=1}^{r} \text{PVE}_s$, and if $r = \min(n-1, p)$ such value is just 1.

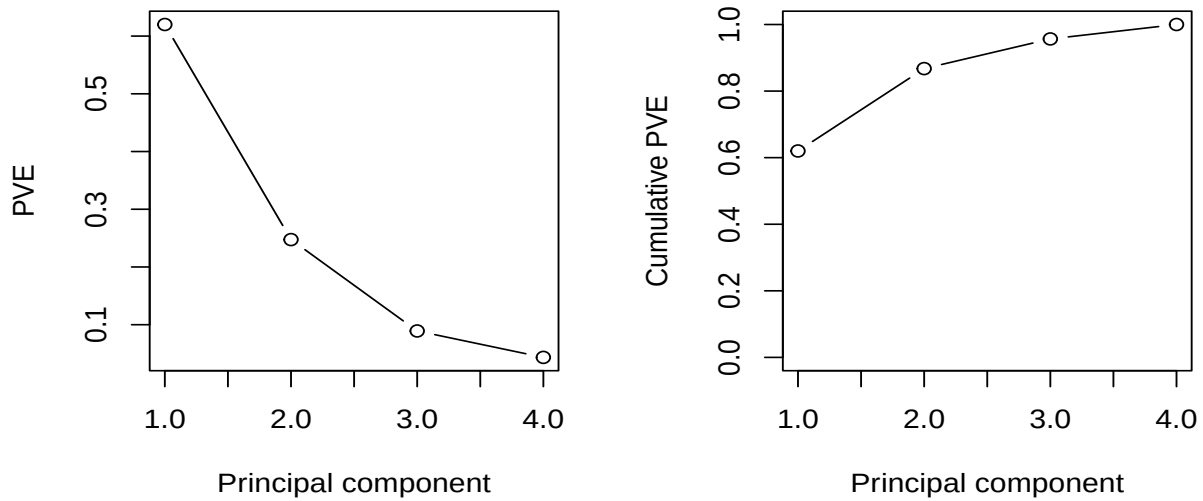
- A useful graphical representation is given by the **scree plot**, which describes the values of PVE_s, for $s = 1, \ldots, \min(n-1, p)$.

  There is no well-accepted objective way to decide how many principal components to consider, since it depends on the specific application.

  It is customary to consider the scree plot and to look for a point where the proportion drops off, the so-called **elbow**.

Example: US arrest data

The screeplot (left panel) and the cumulative proportion of variance explained (right panel) are given below



The first component explains more than 60% of the total variance, and the second one about 25%: they explain almost 87% of the variance in data, and provide a very useful summary using only two dimensions.

The scree plot displays an elbow after the second principal component.

Table of contents

Clustering methods

- Clustering methods refer to a very broad class of techniques for finding subgroups, or clusters, in a data set.
- When the aim is to cluster the observations of a data set, the task is to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
- For example, in a *market segmentation* study the available data concerns p characteristics (variables) for n potential customers and the aim is to classify the customers into different groups, such as big spenders versus low spenders, without the knowledge of the customer's spending patterns.
- Indeed, the n observations may correspond to tissue samples for patients with breast cancer, and the p variables are measurements collected for each *tissue sample*. Clustering could be used to find subgroups related to different unknown subtypes of breast cancer.

- Clustering aims at discovering groups in data, and in particular in the n rows of the $n \times p$ data matrix $\mathbf{X}$.
- In general, the objective is to cluster the observations on the basis of the features (variables) in order to identify subgroups among the observations, but it is as well possible to cluster features on the basis of the observations in order to discover subgroups among the features.

  In what follows, the focus is on clustering observations, though the converse can be performed by simply transposing the data matrix.
- Both clustering and PCA seek to simplify the data via a small number of summaries, but their mechanisms are different.
- Cluster analysis is widely used, but it is useful to remember that, in many cases, visualization based on specialized software may be a compelling alternative.

Clustering algorithms

- Clustering methods are usually based on a measure of *dissimilarity* between units, but other than that they greatly differ among them.

  A rough classification of clustering methods is as follows.

  1. **Hierarchical clustering**, that seeks to build a hierarchy of clusters. There are **agglomerative** (bottom-up) or **divisive** (top-down) approaches. They typically produce a graphical output called *dendrogram*, and need to be tailored to the data at hand.
  2. **Partitioning clustering**, which tries to iteratively optimize the allocation of units to clusters. The most commonly used method of this class is K-**means clustering**, which requires to specify the number of clusters in advance and it is suitable for continuous data.
  3. **Model-based clustering** fits an actual model to the data. Methods of this kind are usually computationally harder, but when the model is sensible for the data at hand they produce a quite reliable output.

- Different clustering methods may give different answers, and there is some risk of over-interpretation. Clustering methods are better used **in conjunction with visualization techniques**, and principal component analysis is very useful in that respect.

Measure of dissimilarity

- Many methods for cluster analysis starts from a measure of **dissimilarity** (or of **similarity**) between observations or cases (rows of the data matrix $\mathbf{X}$.
- A *dissimilarity coefficient* d has the following three properties. For each $\mathbf{a} = (a_1, \ldots, a_p)$, $\mathbf{b} = (b_1, \ldots, b_p)$ and $\mathbf{c} = (c_1, \ldots, c_p)$,
  - $d(\mathbf{a}, \mathbf{b}) \geq 0$ (*non-negativity*)
  - $d(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$ (*identity of indiscernibles*)
  - $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$ (*symmetry*)
- Furthermore, for a *metric dissimilarity* (**distance**)

$$d(\mathbf{a}, \mathbf{c}) \leq d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c})$$

  and for an *ultrametric dissimilarity*

$$d(\mathbf{a}, \mathbf{c}) \leq \max\{d(\mathbf{a}, \mathbf{b}), d(\mathbf{b}, \mathbf{c})\}$$

- A dissimilarity need not be a distance.

Given two numeric vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$, the following distance measures can be defined:

- **Euclidean distance**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

- **Manhattan (taxicab) distance**

$$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p} |a_i - b_i|$$

- **maximum distance**

$$d(\mathbf{a}, \mathbf{b}) = \max_i |a_i - b_i|$$

For non-numeric vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$, the following distance (dissimilarity) measures can be defined:

- **binary distance** (for binary vectors, with only 0s and 1s): it is the percentage of nonzero coordinates (namely, other than (0,0)) that differ.

  It is a special case of the **Jaccard distance** (for categorical variables with a preferred level): the proportion of such variables with one of the cases at the preferred level (level 1 in case of binary distance) in which the cases differ. In general, given the sets A and B, it is

  $$d(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

  where $|\cdot|$ is the size of a given set.
- **Hamming distance** (for strings or categorical data): it is the number of coordinates where the two strings differ.
- **Gower dissimilarity** (for mixed, numeric-categorical, data): it is based on a more involved formula. It is a non-metric dissimilarity, very useful for real data sets.

Hierarchical clustering

- Hierarchical clustering algorithms connect "objects", to form "clusters", based on their distance. It results in an attractive tree-based representation of the observations, called a dendrogram.
- There are **agglomerative hierarchical methods** (*bottom-up*) and **divisive hierarchical methods** (*top-down*).
- Agglomerative methods produce a set of clusterings, starting with one cluster for each observation, and then merging pairs of clusters as moving up the hierarchy.

  It is the most common type of hierarchical clustering.
- Divisive methods also produce a set of clusterings, starting from a single cluster and making successive splitting.

  They are computationally harder, and may be attractive when grouping into a few large clusters is of interest.
- In some situations, the assumption of a hierarchical structure might be unrealistic (for example, a group of people classified by gender or by nationality).

Linkage criteria

- Hierarchical clustering is a whole family of methods that differ by the way dissimilarities are computed.
- Furthermore, in order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a suitable **dissimilarity measure between sets of observations** is employed.
- **Linkage criteria** define the dissimilarity between two groups of observations, starting from a notion of dissimilarity between a pair of observations. Some common choices are:
  - **complete linkage**: the dissimilarity between clusters is the maximum of the dissimilarities between their members;
  - **single linkage**: the dissimilarity between clusters is the minimum of the dissimilarities between their members;
  - **average linkage**: the dissimilarity between clusters is the average of the dissimilarities between their members;
  - **centroid linkage**: the dissimilarity between clusters is defined as the dissimilarity between their centroids (mean vectors).

Dendrogram

- Hierarchical methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters.

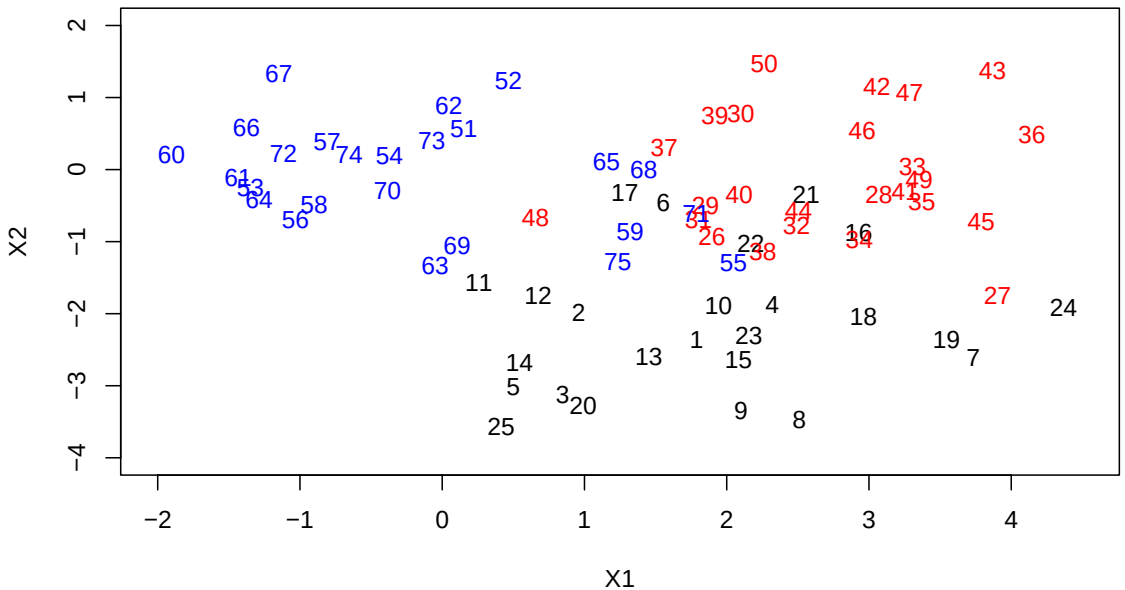
  They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (namely, the "chaining phenomenon", in particular with single linkage).
- The result of hierarchical clustering is typically displayed by means of tree diagrams referred to as **dendrograms**.

  In a dendrogram, the y-axis indicates the distance at which the clusters merge, while the objects are placed along the x-axis .
- The hierarchy of clusters in a dendrogram is obtained by cutting it at different heights.

  There are many proposals, but no general and effective guidelines for performing such a task: cluster analysis is exploratory in nature, and the optimal number of clusters is context-specific.

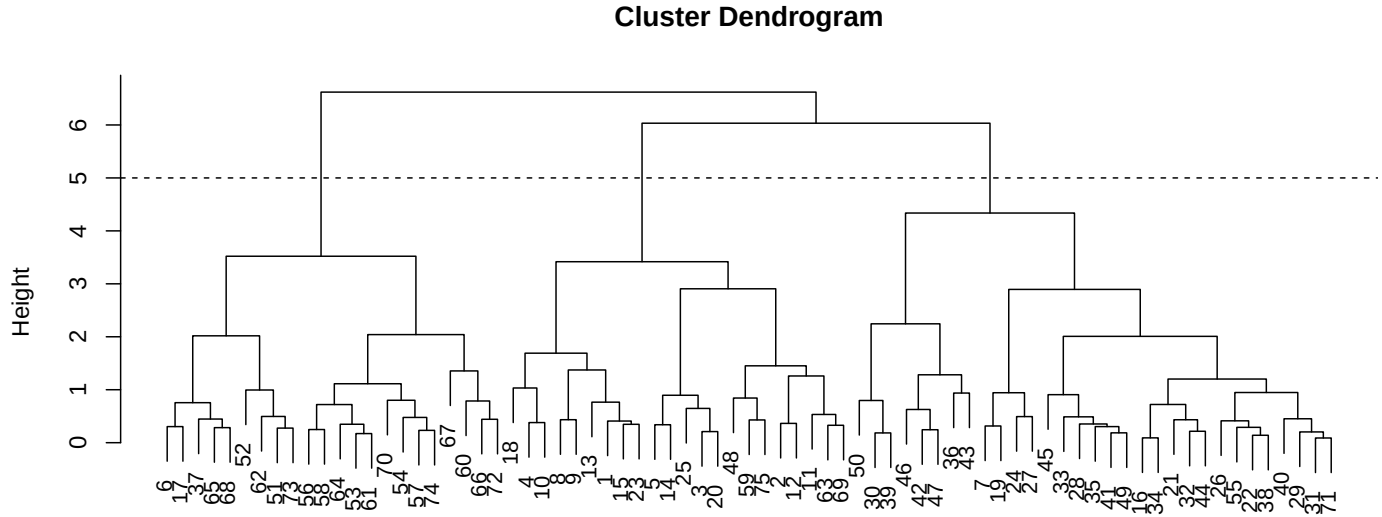
Example: three clusters simulated data

Simulated data set with $n = 75$ observations of the variables X_1 and X_2; 25 bivariate observations in each cluster: **black**, red and blue numbers.



The class labels are treated as unknown and the aim is to cluster the observations in order to discover classes from the data.

Hierarchical clustering, with complete linkage and Euclidean distance, is performed. Different results can be obtained with alternative dissimilarity and linkage criterion.



Each *leaf* of the dendrogram represents one of the 75 observations. When moving up the tree, some leaves begin to fuse into *branches*. These correspond to observations that are similar.

The height of this fusion is measured on the y-axis. Thus, observations that fuse at the very bottom of the tree are quite similar.

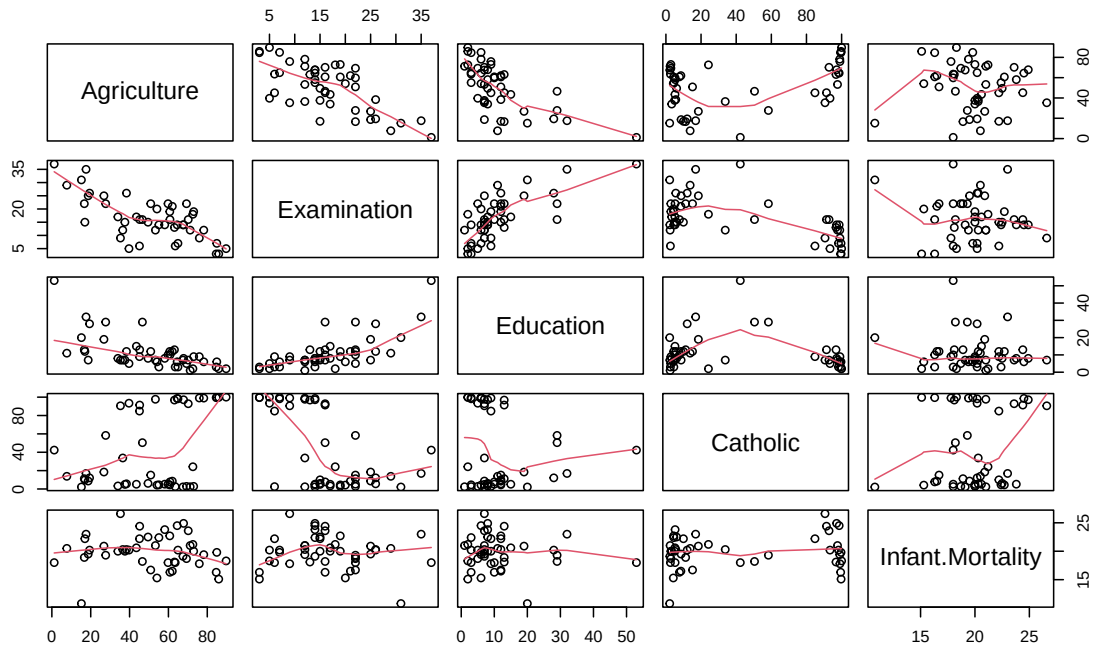
Conclusions about the similarity of two observations should not be based on their proximity along the x-axis. Rather, on the location on the y-axis where branches containing those two observations first are fused.

If the dendrogram is cut at the height of 5, three distinct clusters are specified. Alternative cut points may give different clustering results.

Example: Swiss socioeconomic indicators

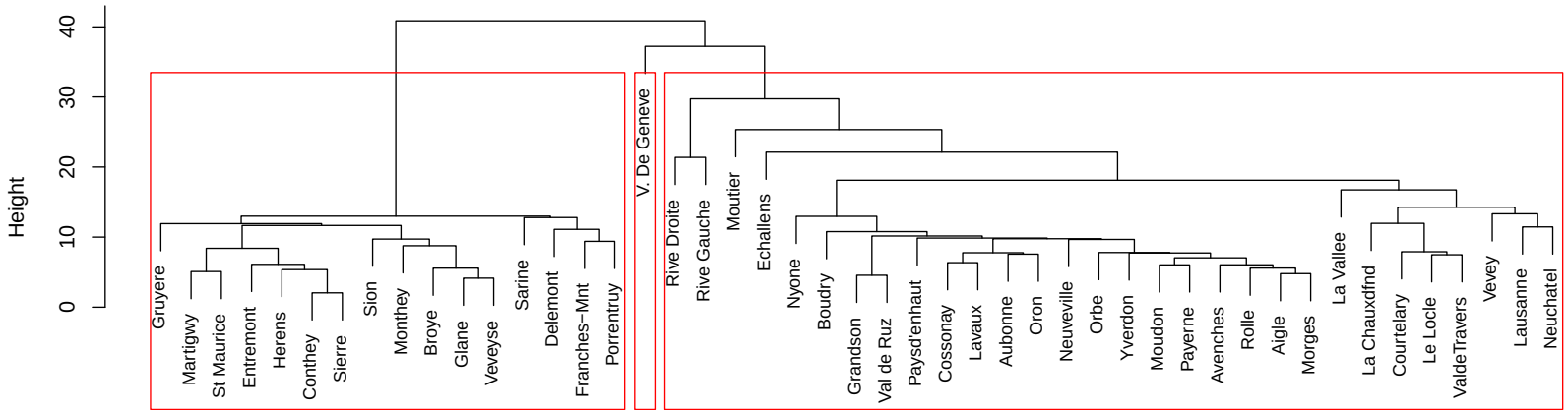
Data from *Modern Applied Statistics with S* by W.N. Venables and B.D. Ripley

Data set on socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888; $n = 47$ observations on $p = 5$ variables, each of which is in percent.

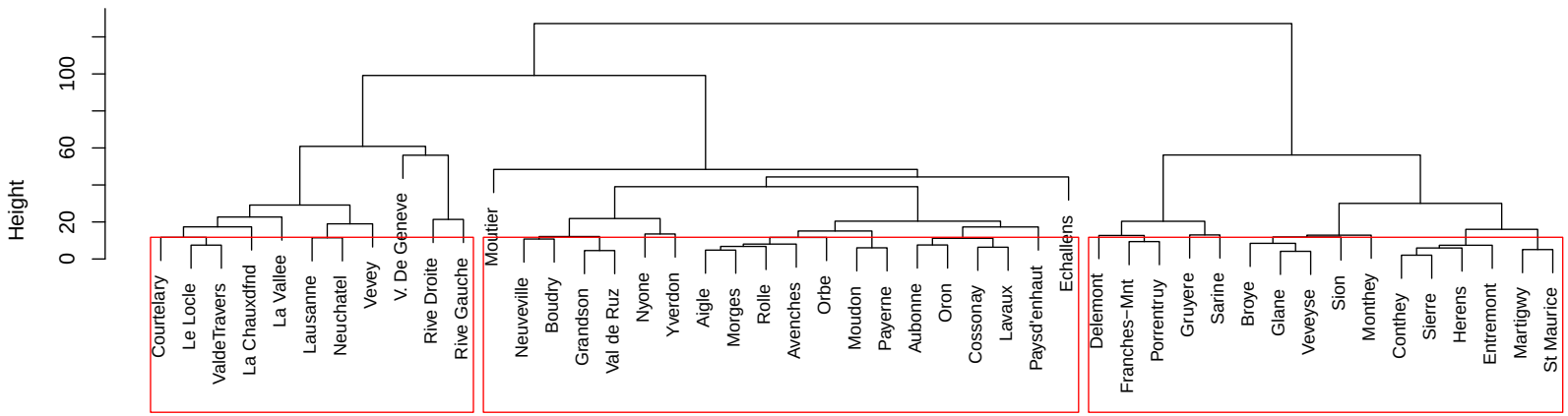


Since data are percentages, then the Euclidean distance is a reasonable choice as dissimilarity between pairs of observations.

Single linkage, agglomerative clustering is performed. The tree is cut into three clusters, indicated by red rectangles. There are two main groups, with a single point, well separated from them.



Divisive clustering is also performed, giving the following three clusters.



Partitioning clustering

- Partitioning methods aim at classifying the observations into $K > 0$ distinct, non-overlapping clusters.
- These methods require to fix in advance the number K of clusters.

  Although there are criteria for choosing K in an iterative fashion, this selection problem is far from simple.

  In partitioning clustering, it is required the decision on how many clusters are expected in the data, as in hierarchical clustering it is necessary to cut the dendrogran, in order to obtain clusters.
- An initial cluster assignment is required for the observations.
- There are different partitioning algorithms and several alternative optimality criteria, some of them are based on probabilistic models.

K-means clustering

- K-**means** is by far the most commonly used partitioning clustering method. It aims at minimizing the *within-cluster variation*.

  It is most appropriate for continuous variables, suitably scaled; customary implementations use the Euclidean distance.

- Given the sets (clusters) $C_1, \ldots, C_K$, containing the indices of observations and satisfying these two properties
  - $C_1 \cup \cdots \cup C_K = \{1, \ldots, n\}$ (each observation belongs to at least one cluster)
  - $C_r \cap C_s =$, for all $r \neq s$ (no observation belongs to more than one cluster)

  the aim is to solve the following optimization problem

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{r=1}^{K} W(C_r) \right\}$$

  where $W(C_r)$ measures the within-cluster variation of C_r (the amount by which its observations differ from each other).

- There are many possible ways to define this concept, but by far the most common choice is

$$W(C_r) = \frac{1}{|C_r|} \sum_{r,s \in C_r} \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$$

namely, the sum of all of the pairwise squared Euclidean distances between the observations in C_r, divided by the total number of observations in the cluster.

- The K-means clustering algorithm involves the following steps:

1. a number from 1 to K is randomly assigned to each of the observations (initial cluster assignments for the observations);
2. until the cluster assignments stop changing:

   a. the centroid (the vector of the p feature means) for each of the K clusters is computed;

   b. each observation is assigned to the cluster whose centroid is closest (according to the Euclidean distance).

- Alternatively, the starting point of the algorithm may be given by the centroids identified by group-average agglomerative hierarchical clustering.
- Differently from hierarchical clustering, K-means requires the entire data matrix, not just the matrix of dissimilarities .
- The optimization problem is usually very difficult to solve precisely: there are almost K^n ways to partition n observations into K clusters.
- The K-means algorithm finds a local rather than a global optimum: the result depends on the initial (random) cluster assignment.

  For this reason, it is important to run the algorithm multiple times from different random initial configurations.

  Then one selects the best solution, namely, that for which the objective function is smallest.

K-medoids clustering

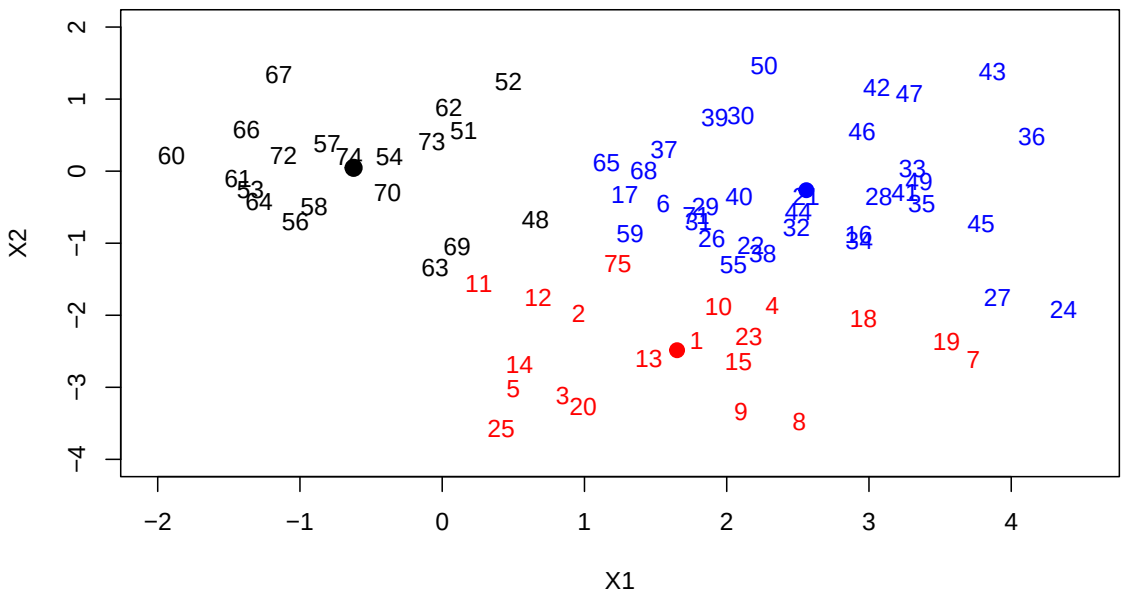
- K-medoids methods are a variation of K-means clustering.
- As K-means, K-medoids algorithms are partitioning methods, attempting to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster.

  In contrast, K-medoids methods choose datapoints as centers, and work with an arbitrary dissimilarity between points, rather than the Euclidean distance.
- Then, only the matrix of dissimilarities is required and the results are more robust to noise and outliers, as compared to those obtained with K-means.
- There are several K-medoids variants. The most common is the **Partitioning Around Medoids** (**PAM**) algorithm, with notable applications to genomic data.

  It uses a greedy search procedure which may not find the optimum solution, but it is faster than exhaustive search algorithms.

Example: three clusters simulated data

K-means clustering, with $K = 3$, is applied to the simulated data set with $n = 75$ bivariate observations. The resulting partition is given below, with indication of the centroids of the three clusters.



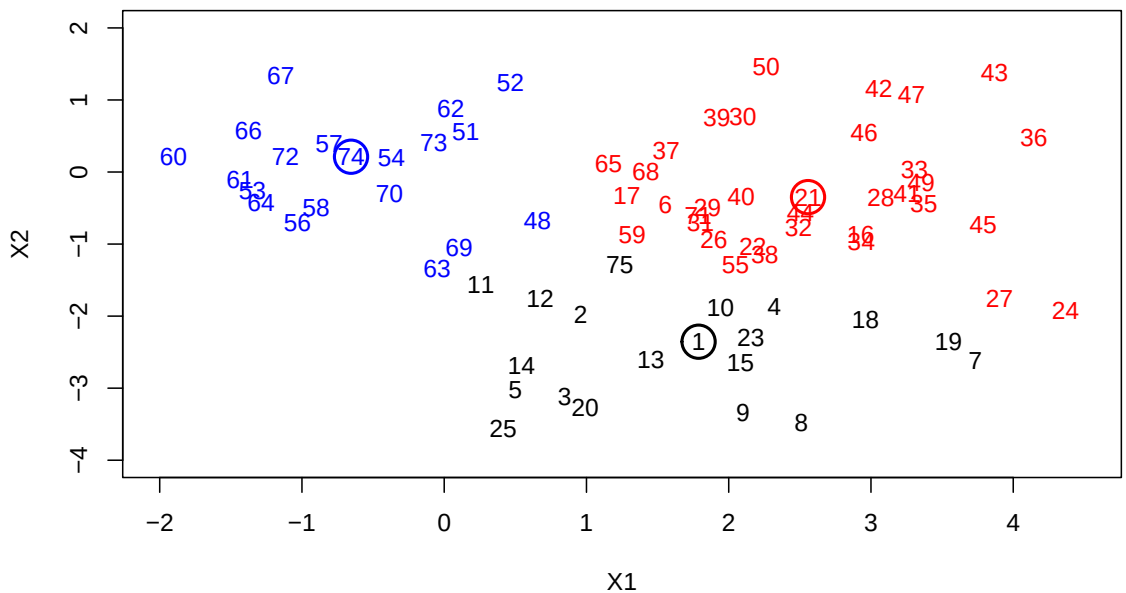
Multiple (in this case 20) initial random cluster assignments are considered and the best solution is selected.

This procedure is strongly recommend, since otherwise an undesirable local optimum may be obtained.

Alternatively, K-medoids methods can be considered. In this case, the three clusters corresponds exactly to those ones obtained using the K-means procedure.

The PAM algorithm is considered. It is based on the search for $K = 3$ representative objects or medoids among the observations of the data set. These observations should represent the structure of the data.

Then the clusters are constructed by assigning each observation to the nearest medoid. The clusters are given below, with indication of the three medoids (observations used as cluster centers)



Model-based clustering

- Model-based clustering provides a more thorough methodology, which includes criteria for choosing the number of clusters and for assessing the goodness of the solution found.
- For continuous data, the employed models are *mixture of multivariate normal distributions*, which are capable of approximating well a broad array of multivariate distributions.

  The model is estimated by maximum likelihood estimation or by the Bayesian approach.
- Unfortunately, there are many algorithmic parameters to tune in order to use the method successfully.

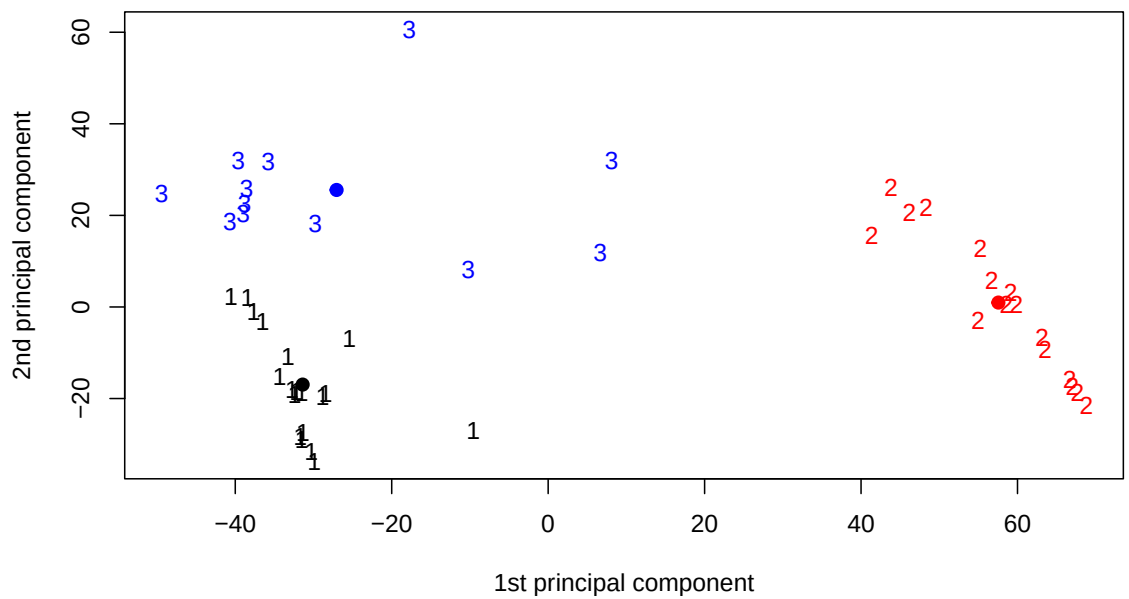
  Furthermore, the method needs a deeper understanding of the underlying theory with respect to other clustering methods, and the theory for model-based clustering is far more complex.

Example: Swiss socioeconomic indicators

With regard to the data set on Swiss provinces, the K-means clustering, with $K = 3$, is considered.

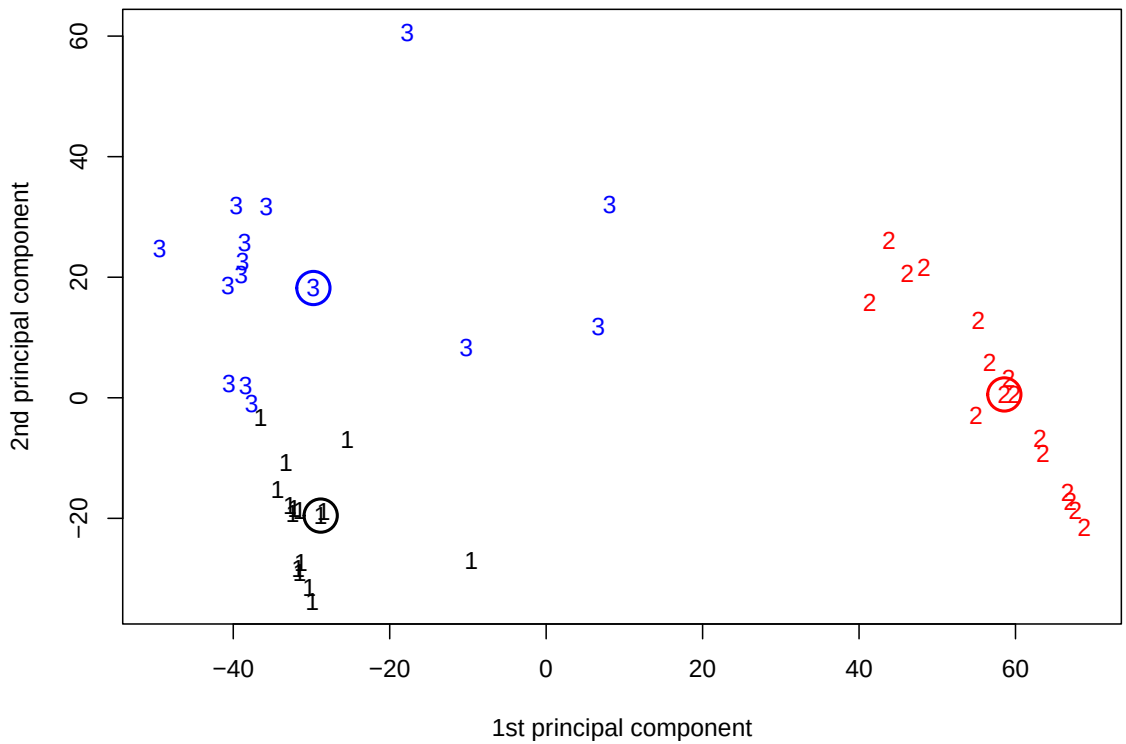
As a starting point, the means of the three clusters identified by hierarchical clustering, with average linkage, are taken into account. The result is given below, with indication of the centroids of the three clusters.

Since there are $p > 2$ variables, the observations are plotted by considering their first two principal components.



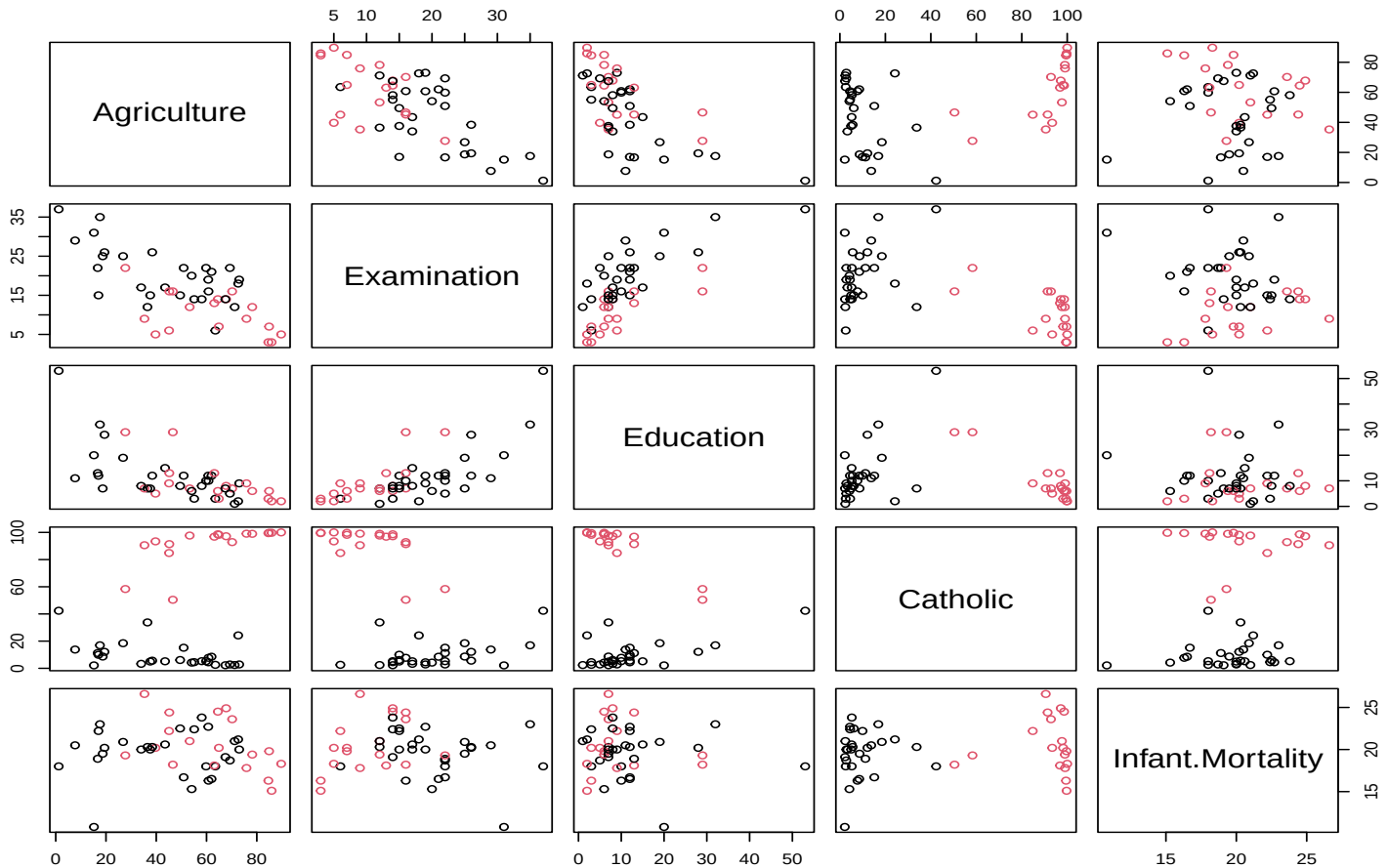
A similar result is obtained by considering multiple initial random cluster assignments.

Furthermore, K-medoids methods can be applied. Using the PAM algorithm, with $K = 3$, the following clusters are specified, with indication of the three medoids (observations used as cluster centers)



All the obtained clusters tend to isolate the counties with a prevalence of Catholics, splitting the remaining counties in two groups.

The scatterplot matrix describes the relationships among the 5 variables. In red the counties with Catholic majority, which displays a peculiar behavior, as pointed out using clustering techniques.



Practical issues

- In order to perform clustering, some crucial decisions must be made, with particular regard to:
  - the standardization of the variables;
  - the choice of the dissimilarity, the linkage criterion and the position where to cut the dendrogram (hierarchical clustering);
  - the number K of clusters to be considered (partitioning clustering).
- Each of these decisions can have a strong impact on the results obtained and, usually, there is no single right answer.

  In practice, several different choices should be tried in order to look for the one giving the most useful or interpretable result.
- There are advanced methods for selecting the number of clusters (and the most effective ones are based on simulations), or for comparing the similarity of two alternative cluster solutions.
- However, it is very hard to try to understand whether the clusters that have been found represent true subgroups in the data, or whether they are simply a result of clustering the noise.

- In addition, clustering methods generally are not very robust to perturbations to the data.
- Since clustering methods force every observation into a cluster, the clusters found may be heavily distorted due to the presence of outliers, not belonging to any cluster.

  Model-based clustering and mixture models are an attractive approach for accommodating the presence of such outliers.
- Cluster analysis is a very useful methodology, which can often add some information to other statistical analyses, even if it is not prudent to view the results as the absolute truth about a data set.
- As there are so many clustering methods available, the recommended strategy is to try more than one method, and assess whether the resulting outcomes are similar.