

Data Analysis and Knowledge Discovery

Data Understanding

prof. Jukka Heikkonen

University of Turku
Department of Information Technology

Jukka.Heikkonen@utu.fi

Goals of data understanding

- ▶ Gain general insight on the data (independent of the project goal).
- ▶ Checking the assumptions made during the project understanding phase.
(representativeness, informativeness, data quality, presence/absence of external factors, dependencies, ...)
- ▶ Checking the specified domain knowledge.
- ▶ Suitability of the data for the project goals.

Rule of thumb: never trust any data before some plausibility tests

Attribute understanding

We (often) assume that the data set is provided in the form of one or more simple tables.

	attribute ₁	...	attribute _m
record ₁			
⋮			
record _n			

- ▶ The rows of the table are called **instances**, **records**, **samples** or **data objects**.
- ▶ The columns of the table are called **attributes**, **features** or **variables**.

Data matrix

Here we are interested on numerical data matrix $R^{n \times m}$ of type

$$\begin{pmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}$$

where the data set consist of n samples (records) and m attributes (variables).

Commonly, to simply the mathematical notations, the data matrix is expressed in the following form (transpose of the above):

$$\begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \vdots & \vdots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix}$$

Types of attributes

categorical (nominal): finite domain

The values of a categorical attribute are often called **classes** or **categories**.

Examples: {female,male}, {pine,spruce,birch}

ordinal: finite domain with a linear ordering on the domain.

Examples: {B.Sc.,M.Sc.,Ph.D.}

numerical: values are numbers.

discrete: categorical attribute or numerical attribute whose domain is a subset of the integer number.

continuous: numerical attribute with values in the real numbers or in an interval

Scales for numerical attributes

interval scale: The definition of the value 0 is arbitrary. Ratios are meaningless.

Examples: date, temperature measured in Celsius or Fahrenheit degrees.

ratio scale: 0 has a canonical meaning. Ratios make sense.

Examples: distance, duration

absolute scale: Domain with a unique measurement unit.

Examples: any kind of counting process (number of children, number of visits to the doctor)

Specific problems of categorical attributes

- ▶ Different levels of **granularity** might be definable.

Examples: Product categories/types:

- ▶ General category: drinks, food, clothes,...
- ▶ More refined categories for drinks: water, beer wine,...
- ▶ Further refinement for water based on the producer.
- ▶ Further refinement of the water of each producer based on the bottle size (0.33 l, 0.5 l, 1 l, 1.5 l).

The most refined level provides the most detailed information, but might not help to discover general associations like *Wine and cheese are often bought together*.

It is crucial to choose an appropriate level of granularity to support the analysis goals.

Specific problems of categorical attributes

- ▶ **dynamic domains:** The possible values of the domain might change over time.
Example: Certain product categories or products might not be sold anymore.
New product categories or products are introduced.

The analysis of such data can cause problems, for instance:

- ▶ Products that have just entered the market will not show significant (accumulated) sales numbers compare to products that have been sold for years.

Coding of categorical attributes

Numerical values are sometimes used for coding categorical attributes:

- ▶ The attribute *general product category* of three possible values *food, drinks, nonfood* may be coded by numbers 1, 2, and 3.

Coding does not make categorical attributes numerical.

- ▶ Does not make sense to count the data mean and say that the average general product category we sell is 2.6.

Sometimes binary code is used to provide equal distances for categorical attributes, e.g. 001, 010, 100 where 3 categories produce 3 input variables.

Data quality

- ▶ Low data quality makes it impossible to trust analysis results:
“Garbage in, garbage out”
- ▶ Mistakes made in the data are most often very difficult to recover by computational methods.

Accuracy: Closeness between the value in the data and the true value.

- ▶ Reason of low accuracy of **numerical attributes**: noisy measurements, limited precision, wrong measurements, transposition of digits (when entered manually).
- ▶ Reason of low accuracy of **categorical attributes**: erroneous entries, typos.

Syntactic and semantic accuracy

Data quality: syntactic accuracy

Syntactic accuracy is violated if an entry does not belong to the domain of the attribute.

Examples:

- ▶ The entry *female* for the categorical attribute *gender* violates syntactic accuracy.
- ▶ Text entries for numerical attributes violate syntactic accuracy.
- ▶ Values out of the range for numerical attributes violate syntactic accuracy (negative numbers for weight, distance, counting processes,...).

Syntactic accuracy can be checked quite easily.

Data quality: semantic accuracy

Semantic accuracy is violated if an entry is not correct although it belongs to the domain of the attribute.

Example:

- ▶ The entry *female* for the categorical attribute *gender* in the record with name entry *John Smith* is within the domain of the attribute *gender*, but obviously incorrect given that the name is correct.
(It could also be that the gender is correct, but the name is misspelled as John instead of Joan.)

Semantic accuracy is more difficult to check than syntactic accuracy, and sometimes even impossible to check.

Semantic accuracy can only be checked based on “business rules” (e.g. only women can be pregnant) and plausibility checks.

Data quality: completeness

Completeness is violated if an entry is missing.

- ▶ w.r.t. **attribute values**: Fraction of null entries for an attribute. Note that missing values are not always marked explicitly as missing, for instance in the case of default entries.
- ▶ w.r.t. **records**: Complete records might be missing because
 - ▶ three years ago SAP was introduced and not all customer data were transferred to the new system.
 - ▶ the data set is biased and non-representative. (A bank might have rejected customers with no income.)

Data quality: unbiased and representative

The data should always be unbiased and representative, i.e. it should contain all information about the inherent patterns and rules in the data.

In many applications we do not have that sort of data

- ▶ **Machine condition monitoring:** A lot of examples when machine is running normally. Sometimes not possible to obtain interesting data, such as in the case of nuclear power plant.
- ▶ **Natural disasters:** E.g. no earthquake data from a certain area where future earthquake probability should be estimated.
- ▶ **Mortgage/insurance etc. analysis:** Certain types of customers are totally missing, e.g. we may only have information about customers who have been granted a loan.

Data quality: unbiased and representative

The biasness and non-representativeness of the data is one of the most difficult problems in data analysis.

Sometimes it is not even known what is needed, e.g.

- ▶ **Electricity load prediction:** What information should be used to estimate the electricity load for tomorrow.
- ▶ **Pattern recognition tasks:** Often impossible to describe where the recognition should be based.

Data quality: unbalanceness and timeliness

Unbalanced data: The data set might be biased extremely to one type of records.

Example: Production line for goods including quality control. Defective goods will be a very small fraction of all records.

Timeliness: Are the available data up to date to be considered to be representative? This is related to the non-stationarity of the domain where only the recently collected data provide relevant information.

Example: Many industrial processes change dynamically and are non-stationary in nature. Hence too old data may not be much of use for analysing current and future states of the process.

Data visualisation is one of the most important steps for

- ▶ Data understanding and preliminary data quality evaluation
- ▶ Learning from the domain

Visualisation as a test

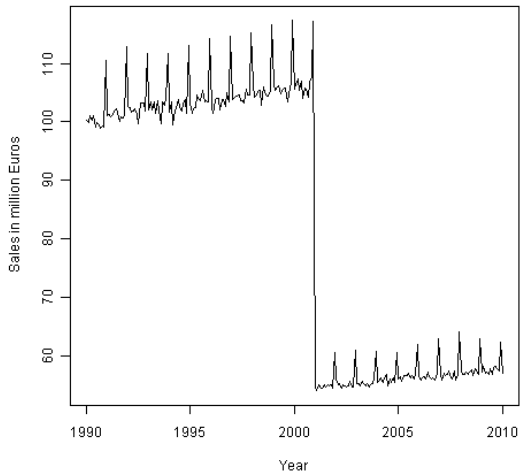
- ▶ When visualisations reveal patterns or exceptions, then there is “something” in the data set.
- ▶ When visualisations do not indicate anything specific, there might still be patterns or structures in the data that cannot be revealed by the corresponding (simple) visualisation techniques.

There are infinitely many possibilities for visualisation

Here we cover some very useful data visualisation techniques

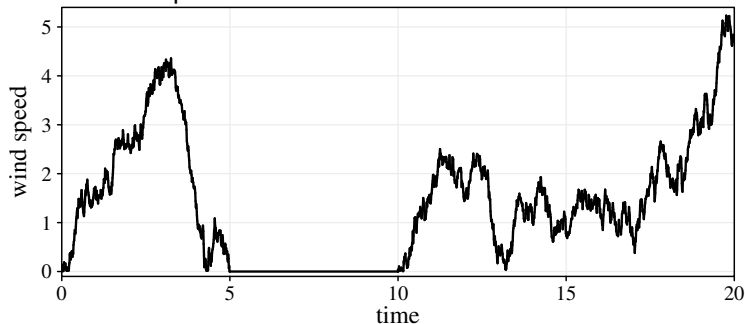
Data visualisation

There is no excuse for failing to plot and look.



Data visualisation: hidden missing values

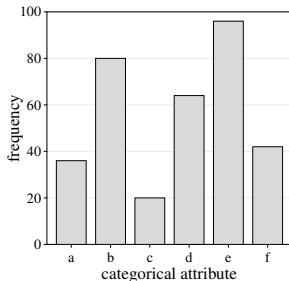
Measured wind speeds.



The zero values might come from a broken or blocked sensor and might be considered as missing values.

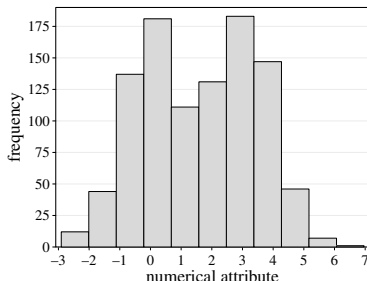
Bar charts

A **bar chart** is a simple way to depict the frequencies of the values of a categorical attribute.

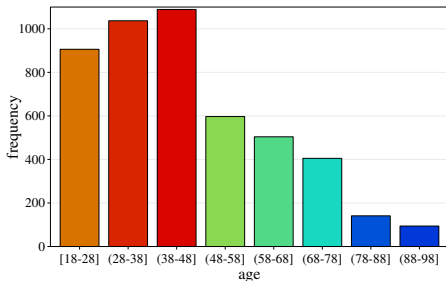


Histograms

A **histogram** shows the frequency distribution for a numerical attribute. The range of the numerical attribute is discretized into a fixed number of intervals (called **bins**), usually of equal length. For each interval the (absolute) frequency of values falling into it is indicated by the height of a bar.

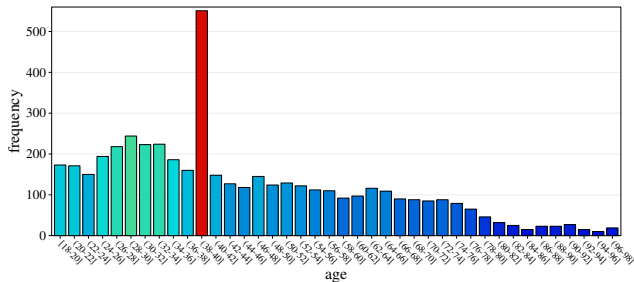


Histograms: number of bins



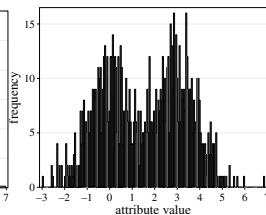
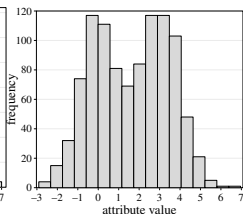
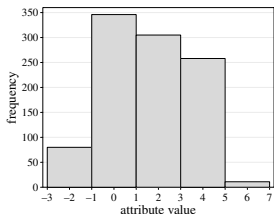
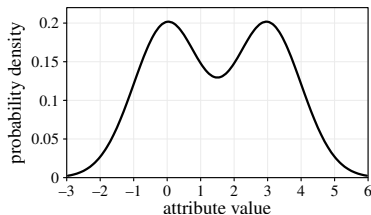
Distribution of the age of the customers in 2010

Histograms: number of bins



Distribution of the age of the customers in 2010

Histograms: number of bins



Three histograms with 5, 17 and 200 bins for a sample from the same bimodal distribution.

Histograms: number of bins

The choice of the number of bins clearly influences the results.

- ▶ There is no *best* number of bins
- ▶ Different bin sizes can reveal different features of the data.
- ▶ Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution.
- ▶ Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate

Experimentation is usually needed to determine an appropriate width.

Histograms

In a more general mathematical sense, a histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (known as bins)

Thus, if we let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^k m_i \quad (1)$$

Histograms: number of bins

There are various useful guidelines for selecting the number of bins.

The number of bins k can be calculated from a suggested bin width h as:

$$k = \left\lceil \frac{\max_i \{x_i\} - \min_i \{x_i\}}{h} \right\rceil ; , \quad (2)$$

where x_1, \dots, x_n is the sample to be displayed and the braces indicate the ceiling (flooring) function.

Sturges' rule:

$$k = \lceil \log_2(n) + 1 \rceil , \quad (3)$$

where n is the sample size.

Sturges' rule is suitable for data from normal distributions and from data sets of moderate size ($n > 30$)

Histograms: number of bins

Scott's rule:

$$h = \frac{3.5s}{n^{1/3}}, \quad (4)$$

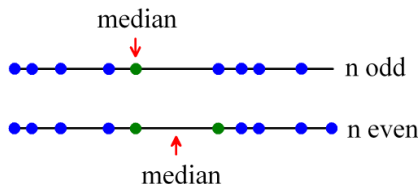
where s is the sample standard deviation.

Square-root choice:

$$k = \sqrt{n} \quad (5)$$

Used by Excel histograms and many others (Wikipedia).

Reminder: Median, quantiles, quartiles, interquartile range



Median: The value in the middle (for the values given in increasing order).

$q\%$ -quantile ($0 < q < 100$): The value for which $q\%$ of the values are smaller and $100-q\%$ are larger.
The median is the 50%-quantile.

Quartiles: 25%-quantile (1st or lower quartile), median (2nd quartile), 75%-quantile (3rd or upper quartile).

Interquartile range (IQR): 3rd quartile - 1st quartile.

Histograms: number of bins

Freedman-Diaconis' choice

$$k = 2 \frac{\text{IQR}(x)}{n^{1/3}}, \quad (6)$$

where IQR is interquartile range of the sample.

Histograms: number of bins

All the above methods are highly sensitive for outliers:

- ▶ They divide the range between the smallest and largest value of the sample into bins of equal size.

To avoid the effect of outliers one can leave out extreme values from the sample, e.g. based on certain percents.

- ▶ Also variable bin width histograms can be employed.

Example data set: Iris data

Collected by E. Anderson in 1935

Contains measurements of four real-valued variables:

Sepal length, sepal widths, petal lengths and petal width of 150 iris flowers of types Iris Setosa, Iris Versicolor, Iris Virginica (50 each)

The fifth attribute is the name of the flower type.

Example data set: Iris data



iris setosa



iris versicolor



iris virginica

Example data set: Iris data

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

5.1 3.5 1.4 0.2 Iris-setosa

...

...

5.0 3.3 1.4 0.2 Iris-setosa

7.0 3.2 4.7 1.4 Iris-versicolor

...

...

5.1 2.5 3.0 1.1 Iris-versicolor

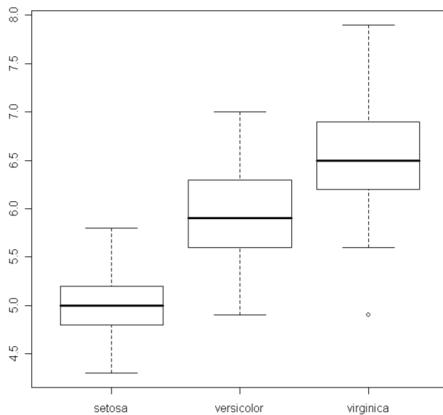
5.7 2.8 4.1 1.3 Iris-versicolor

...

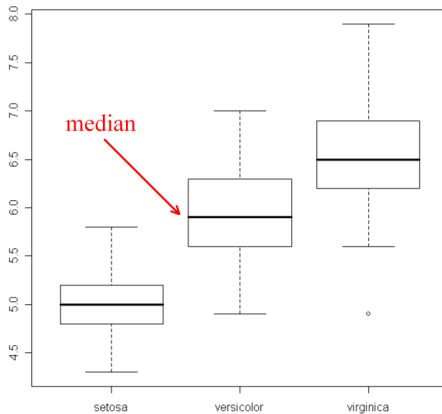
...

5.9 3.0 5.1 1.8 Iris-virginica

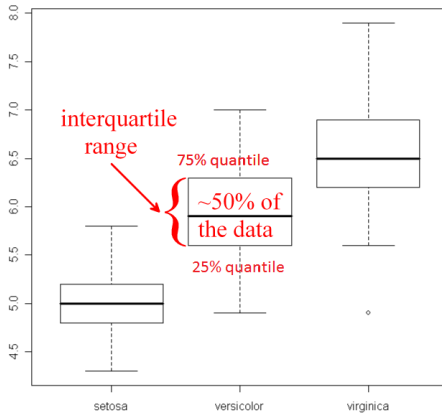
Boxplots



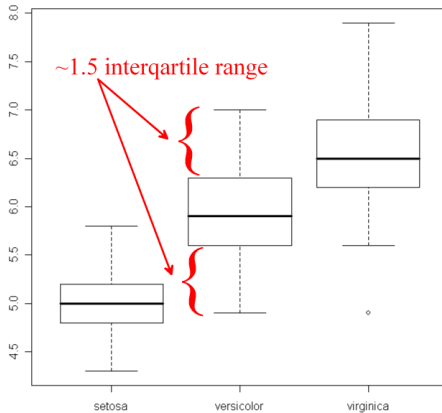
Boxplots



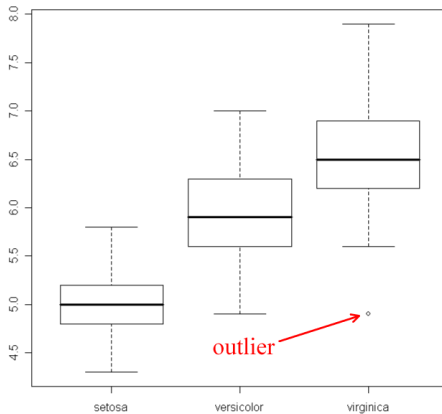
Boxplots



Boxplots



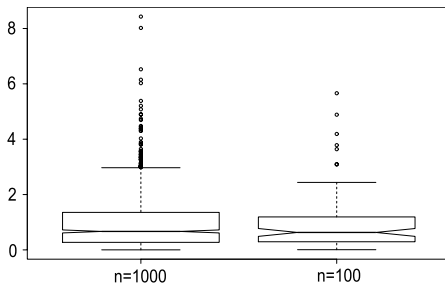
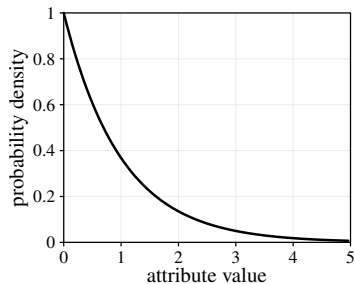
Boxplots



Boxplot construction

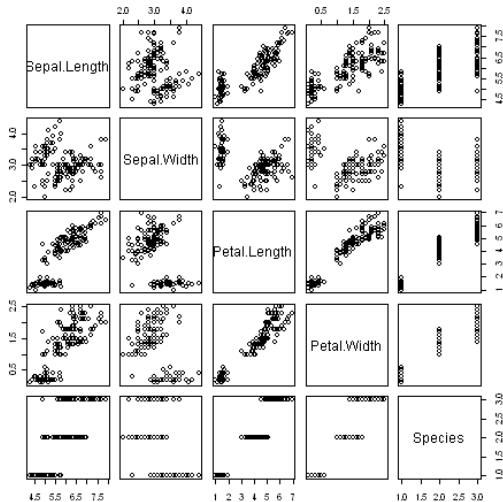
1. Determine the median. Draw a thick line at the position of the median.
2. Determine the 25%- and the 75%-quartiles q_1 and q_3 for the sample. Draw a box limited by these two quartiles. The other dimension of the box can be chosen arbitrarily.
3. $iqr = q_3 - q_1$ is the interquartile range. The inner fence is defined by the two values $f_1 = q_1 - 1.5 \cdot iqr$ and $f_3 = q_3 + 1.5 \cdot iqr$.
4. Find the smallest data point greater than f_1 and the largest data point smaller than f_3 . Add "whiskers" to the box extending to these two data points.
5. Data points lying outside the box and the whiskers are called outliers. Enter these data points in the diagram, for instance by circles.
6. Sometimes, extreme outliers (out of the outer fence defined by $F_1 = q_1 - 2 \cdot 1.5 \cdot iqr$ and $F_3 = q_3 + 2 \cdot 1.5 \cdot iqr$) are drawn in a different way than mild outliers outside the whiskers, but within the inner fence.

Boxplots for asymmetric distribution



Scatter plots

Scatter plots visualise two variables in a two-dimensional plot. Each axes corresponds to one variable.

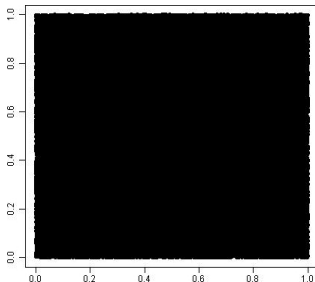


Scatter plots

For large data sets, points are plotted over each other.

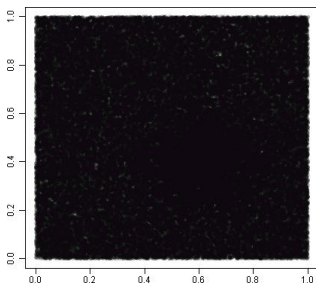
Density information is lost.

In the worst case, all information is lost.



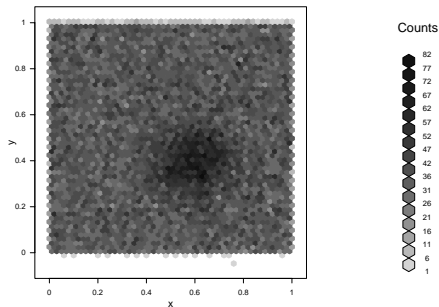
A scatter plot for a data set with 100000 objects.

Scatter plots



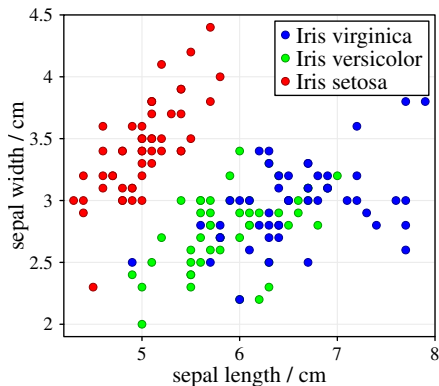
A density plot. Instead of solid points, semitransparent points are plotted.

Scatter plots



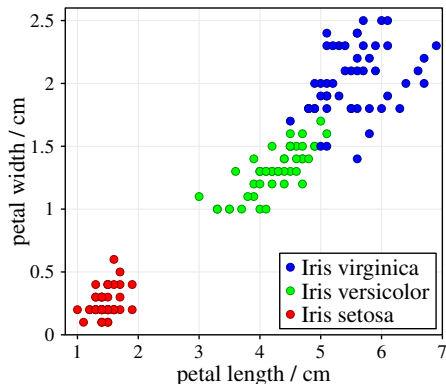
A scatter plot based on hexagonal binning. The grey intensity in each bin indicates the number of points falling into the bin.

Scatter plots



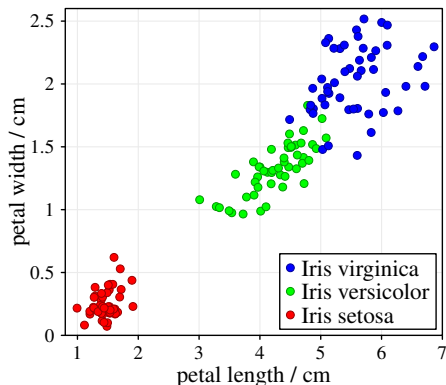
Scatter plots can be enriched with additional information: Colour or different symbols to incorporate a third attribute in the scatter plot.

Scatter plots



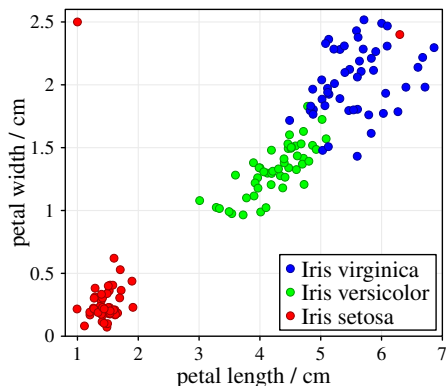
The two attributes petal length and width provide a better separation of the classes Iris versicolor and Iris virginica than the sepal length and width.

Scatter plots: jitter



Data objects with the same values cannot be distinguished in a scatter plot. To avoid this effect, jitter is used, i.e. before plotting the points, small random values are added to the coordinates. Jitter is essential for categorical attributes.

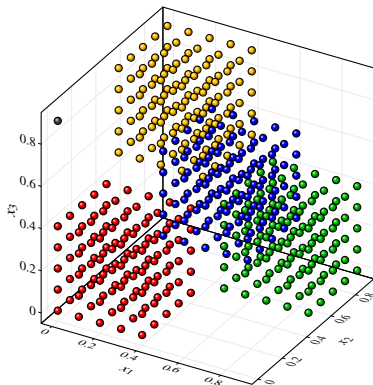
Scatter plots



The Iris data set with two (additional artificial) outliers. One is an outlier for the whole data set, one for the class Iris setosa.

3D scatter plots

For data sets of moderate size, scatter plots can be extended to three dimensions.



A 3D scatter plot of an artificial data set filling a cube in a chessboard-like manner with one outlier.

Methods for higher-dimensional data

- ▶ A display or plot is by definition two-dimensional, so that only two axes (attributes) can be incorporated.
- ▶ 3D techniques can be used to incorporate three axes (attributes).
- ▶ The number of possible scatter plots grows in a quadratic fashion with the number of attributes. For m attributes there are $\binom{m}{2} = m(m-1)$ possible scatter plots. For 50 attributes there are 2450 scatter plots.

Methods for higher-dimensional data

Principle approach for incorporating all attributes in a plot:

- ▶ Try to preserve as much of the “structure” of the high-dimensional data set when representing (plotting) the data in two (or three) dimensions.
- ▶ Define a measure that evaluates lower-dimensional representations (plots) of the data in terms of how well a representation preserves the original “structure” of the high-dimensional data set.
- ▶ Find the representation (plot) that gives the best value for the defined measure.

There is no unique measure for “structure” preservation.

Higher-dimensional data

Assume that we have m -dimensional input $x \in R^m$ represented as random variable X

How we characterize $p(x)$ assuming it is normally distributed?

- ▶ For 1-dimension it is **mean** (μ) and **variance** (σ^2)
 - ▶ Mean = $E[X]$
 - ▶ Variance = $E[(X - \mu)^2]$
- ▶ For d -dimensions we need
 - ▶ m -dimensional **mean** vector
 - ▶ $m \times m$ dimensional **covariance** matrix

Matrix transpose

Given the matrix A , the transpose of A is denoted by A^T .

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad B = \begin{bmatrix} 3 & 5 \\ 2 & 7 \\ 6 & 9 \\ 1 & 0 \\ 5 & 2 \end{bmatrix}$$
$$A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \quad B^T = \begin{bmatrix} 3 & 2 & 6 & 1 & 5 \\ 5 & 7 & 9 & 0 & 2 \end{bmatrix}$$

Some properties of transpose

- ▶ $(A^T)^T = A$
- ▶ $(A + B)^T = A^T + B^T$
- ▶ $(rA)^T = rA^T$, where r is any scalar
- ▶ $(AB)^T = B^T A^T$
- ▶ $A^T B = B^T A$, where A and B are vectors

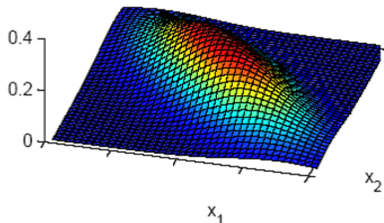
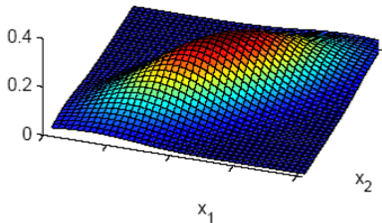
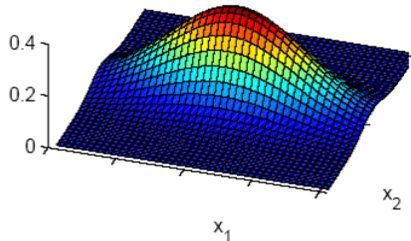
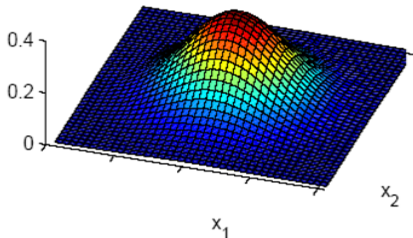
Matrix derivation

- ▶ If $y = \mathbf{x}^T \mathbf{A} \mathbf{x}$ where \mathbf{A} is a square matrix and \mathbf{x} vector
 - ▶ $\partial y / \partial \mathbf{x} = \mathbf{A} \mathbf{x} + \mathbf{A}^T \mathbf{x}$
- ▶ If $y = \mathbf{x}^T \mathbf{A} \mathbf{x}$ where \mathbf{A} is symmetric ($\mathbf{A} = \mathbf{A}^T$)
 - ▶ $\partial y / \partial \mathbf{x} = 2\mathbf{A} \mathbf{x}$
- ▶ If $y = \mathbf{x}^T \mathbf{x}$
 - ▶ $\partial y / \partial \mathbf{x} = 2\mathbf{x}$

Eigenvectors and eigenvalues

- ▶ Given a matrix \mathbf{A} , a non-zero vector \mathbf{x} is defined to be an eigenvector of the transformation if it satisfies the eigenvalue equation
 - ▶ $\mathbf{Ax} = \lambda\mathbf{x}$, for some scalar λ .
- ▶ In this situation, the scalar λ is called an **eigenvalue** of \mathbf{A} corresponding to the **eigenvector** \mathbf{x} .
- ▶ Eigenvectors and eigenvalues are got from the equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0 \Rightarrow \det(\mathbf{A} - \lambda\mathbf{I})$ must be 0.
 - ▶ Gives the characteristics polynomial whose roots are the eigenvalues of \mathbf{A} .

How to characterize differences between these distributions



Multivariate Parameters: Mean, Covariance

Each record (column) of data matrix $R^{m \times n}$

$$\begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \vdots & \vdots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix}$$

can be considered to be a realization of m -dimensional random variable $\mathbf{X} = [X_1, \dots, X_m]^T$.

Mean of attributes: $E[\mathbf{X}] = \boldsymbol{\mu} = [\mu_1, \dots, \mu_m]^T$

Covariance:

$$\boldsymbol{\Sigma} \equiv \text{Cov}(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

$$\sigma_{ij} = \text{Cov}(\mathbf{X}_i, \mathbf{X}_j) = E[(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)^T]$$

Multivariate Parameters: Mean, Covariance

Variance: How much a random variable varies around the expected value

Covariance is the measure the strength of the linear relationship between two random variables

- ▶ Covariance becomes more positive for each pair of values which differ from their mean in the same direction.
- ▶ Covariance becomes more negative with each pair of values which differ from their mean in opposite directions.
- ▶ If two variables are independent, then their covariance/correlation is zero (converse is not true).

Covariance matrix

Shape and orientation of the hyper-ellipsoid centered at μ is defined by Σ .

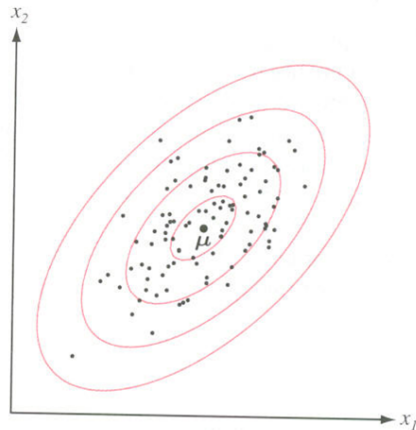
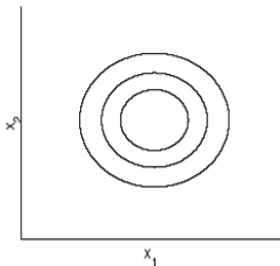


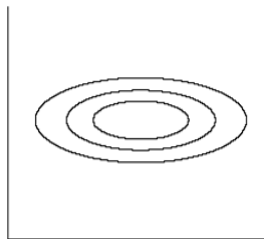
FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ . The ellipses show lines of equal probability density of the Gaussian.

Covariance matrices

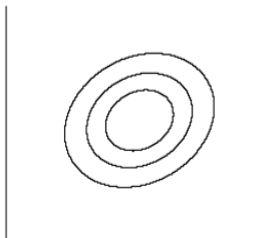
$$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) = \text{Var}(x_2)$$



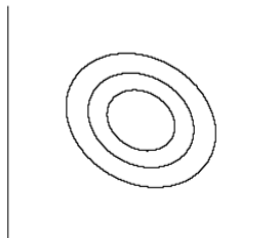
$$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) > \text{Var}(x_2)$$



$$\text{Cov}(x_1, x_2) > 0$$

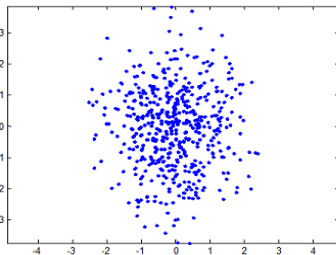


$$\text{Cov}(x_1, x_2) < 0$$

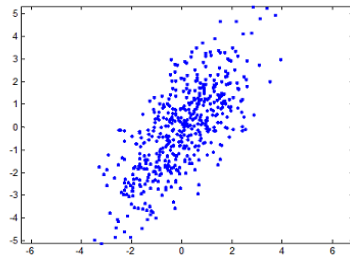


Covariance matrices

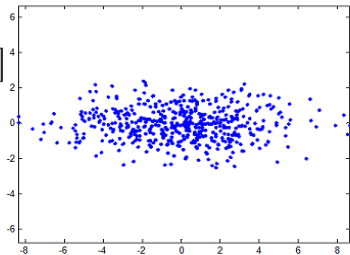
$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



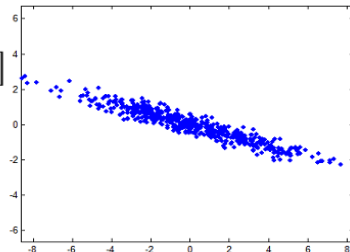
$$\begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 10 & -3 \\ -3 & 1 \end{bmatrix}$$



Properties of Σ

- ▶ A small value of $|\Sigma|$ indicates that samples are close to μ .
- ▶ Small $|\Sigma|$ may also indicate that there is a high correlation between variables
- ▶ If some of the variables are linearly dependent, or if the variance of one variable is 0, then Σ is singular and $|\Sigma|$ is 0.
 - ▶ Dimensionality should be reduced to get a positive definite matrix

Properties of Σ

- ▶ The covariance matrix Σ is symmetrical and it can always be diagonalized as

- ▶ $\Sigma = \Phi \Lambda \Phi^T$,

where

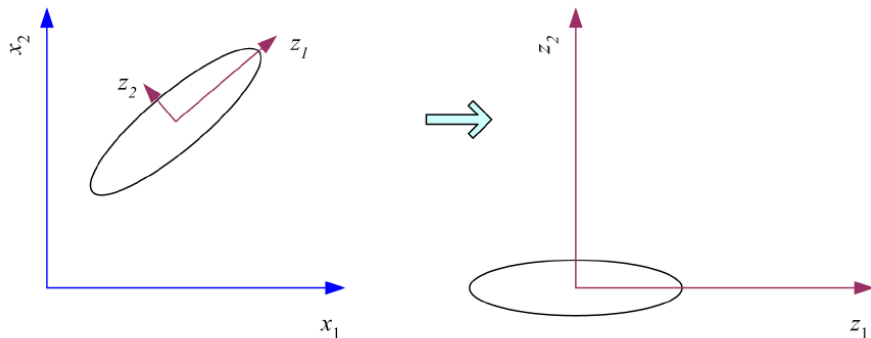
- ▶ $\Phi = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ is the column matrix consisting of the eigenvectors of Σ .
 - ▶ $\Phi^T = \Phi^{-1}$
 - ▶ Λ is the diagonal matrix whose elements are the eigenvalues of Σ .

Principal component analysis

Principal component analysis (PCA) uses the variance in the data as the structure preservation criterion.

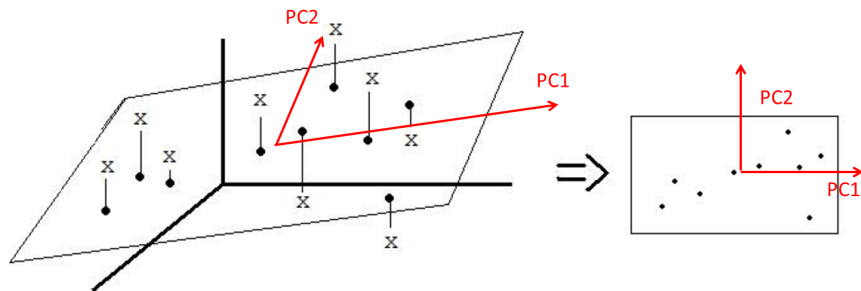


PCA tries to preserve as much of the original variance of the data when projected to a lower-dimensional space.



PCA example: 3D to 2D

The 2D plane where the 3D points are projected is defined by 2 first principal components.



Principal component analysis

PCA constructs a projection from the high-dimensional space to a lower-dimensional space (plane or hyperplane).

The data points are first centered around the origin by subtracting the mean values.

Aim: Find a projection in the form of a linear mapping $\mathbb{R}^m \rightarrow \mathbb{R}^q$ (for visualisation purposes choose $q = 2$ or $q = 3$) given by

$$\mathbf{y} = M^T \cdot (\mathbf{x} - \bar{\mathbf{x}})$$

where M is a $m \times q$ (projection) matrix such that the variance of the projected data $\mathbf{y}_i = M^T \cdot (\mathbf{x}_i - \bar{\mathbf{x}})$ is as large as possible.

Principal component analysis

Problem: Without restriction for the matrix M , the entries in M can be chosen arbitrary large, so that the data are not only projected but also stretched, leading to an arbitrary large variance of the projected data.

Therefore: Introduce constraints such that the matrix M is only a projection.

Constraints: The column \mathbf{v}_i of the matrix

$$M = (\mathbf{v}_1, \dots, \mathbf{v}_q)$$

must be normalised i.e. $\|\mathbf{v}_i\| = 1$.


Principal component analysis

Solution of the constraint optimisation problem:

The projection matrix M for PCA is given by

$$M = (\mathbf{v}_1, \dots, \mathbf{v}_q)$$

where the **principal components** $\mathbf{v}_1, \dots, \mathbf{v}_q$ are the normalized eigenvectors of the **covariance matrix** of the data

$$C = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$


for the q largest eigenvalues $\lambda_1 \geq \dots \geq \lambda_q$.

λ is called an eigenvalue of a matrix C , if there is a non-zero vector \mathbf{v} such that $C\mathbf{v} = \lambda\mathbf{v}$ holds. The vector \mathbf{v} is called eigenvector to the eigenvalue.

Principal component analysis

Another look for principal components

The projection of \mathbf{x} on the direction of \mathbf{v} is: $y = \mathbf{v}^T \mathbf{x}$.

$$\text{Var}(y) = \text{Var}(\mathbf{v}^T \mathbf{x}) = E[(\mathbf{v}^T \mathbf{x} - \mathbf{v}^T \boldsymbol{\mu})^2] \quad (7)$$

$$= E[(\mathbf{v}^T \mathbf{x} - \mathbf{v}^T \boldsymbol{\mu})(\mathbf{v}^T \mathbf{x} - \mathbf{v}^T \boldsymbol{\mu})] \quad (8)$$

$$= E[\mathbf{v}^T (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{v}] \quad (9)$$

$$= \mathbf{v}^T E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \mathbf{v} = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} \quad (10)$$

Find \mathbf{v} such that $\text{Var}(y)$ is maximized subject to $\|\mathbf{v}\| = 1$

Principal component analysis

For the 1. principal component \mathbf{v}_1 maximize $\text{Var}(y_1)$ subject to $\|\mathbf{v}_1\| = 1$:

$$\max_{\mathbf{v}_1} \{ \mathbf{v}_1^T \sum \mathbf{v}_1 - \alpha (\mathbf{v}_1^T \mathbf{v}_1 - 1) \} \quad (11)$$

- ▶ Taking the derivative w.r.t to \mathbf{v}_1 and setting it equal to 0, we get $\sum \mathbf{v}_1 + \sum^T \mathbf{v}_1 = 2\alpha \mathbf{v}_1$ which leads to $\sum \mathbf{v}_1 = \alpha \mathbf{v}_1$.
- ▶ That is, \mathbf{v}_1 is an eigenvector of \sum .
- ▶ Choose the eigenvector with the largest eigenvalue for 1. principal component to maximize $\text{Var}(y_1)$.

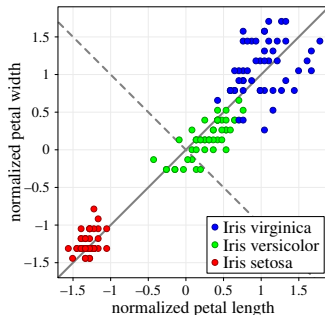
Principal component analysis

Second principal component \mathbf{v}_2 maximize $\text{Var}(y_2)$ subject to $\|\mathbf{v}_2\| = 1$ and orthogonal to \mathbf{v}_1

$$\max_{\mathbf{v}_2} \{ \mathbf{v}_2^T \Sigma \mathbf{v}_2 - \alpha(\mathbf{v}_2^T \mathbf{v}_2 - 1) - \beta(\mathbf{v}_2^T \mathbf{v}_1 - 0) \} \quad (12)$$

- ▶ Similar analysis shows that $\Sigma \mathbf{v}_2 = \alpha \mathbf{v}_2$
- ▶ That is, \mathbf{v}_2 is another eigenvector of Σ .

Principal component analysis



PCA applied to the Iris data set restricted to the (zscore normalised) petal length and width.

The principal components (1. solid line, 2. dashed line) are always orthogonal.

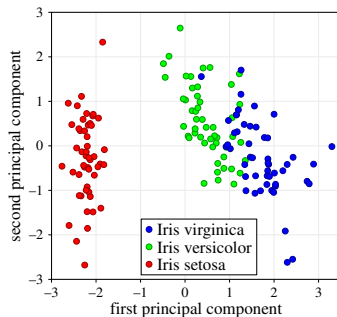
Principal component analysis: Normalisation

Usually, the data should be **z-score standardised** $x \mapsto \frac{x - \hat{\mu}_X}{\hat{\sigma}_X}$ to ensure that all attributes contribute equally to the overall variance, where $\hat{\mu}_X$ and $\hat{\sigma}_X$ are the mean value and the sample standard deviation (the square root of the sample variance) of attribute X .

When we change the measurement of the petal length from centimetres to metres, but leave the measurement of the petal width in centimetres, the first principal component becomes the vector $(0.0223, 0.9998)^\top$ without z-score standardisation.

The variance of the petal length becomes negligible compared to the variance of the petal width.

Principal component analysis



Projection to the first two principal components of PCA for the Iris data set taking all four numerical attributes into account.

Principal component analysis: Dimension reduction

Let $\lambda_1 \geq \dots \geq \lambda_m$ be the eigenvalues of the covariance matrix.

When we project the data to the first q principal components v_1, \dots, v_q corresponding to the eigenvalues $\lambda_1, \dots, \lambda_q$, this projection will preserve a fraction of

$$\frac{\lambda_1 + \dots + \lambda_q}{\lambda_1 + \dots + \lambda_m}$$

of the variance of the original data.

Principal component analysis: Dimension reduction

	Principal component			
	PC1	PC2	PC3	PC4
Proportion of variance	0.73	0.229	0.0367	0.00518
Cumulative proportion	0.73	0.958	0.9948	1.00000

Preservation of the variance of the Iris data set depending on the number of principal components.

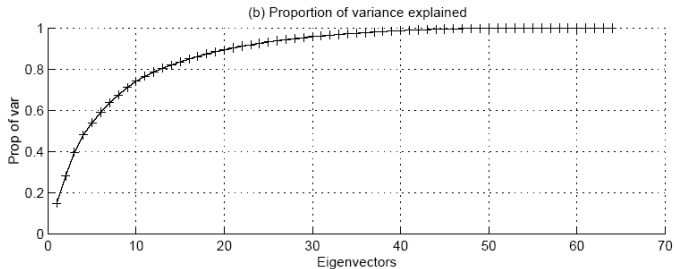
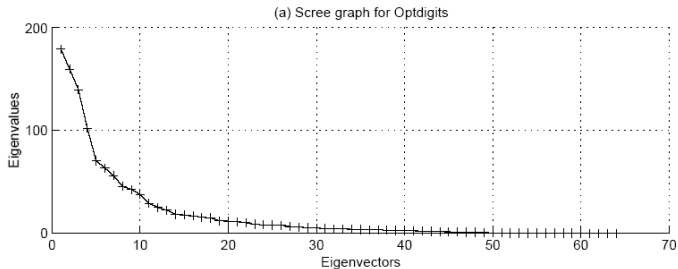
Principal component analysis: Dimension reduction

As noted dimension of the original data can be reduced by projecting the original m -dimensional data for q -first principal components $v_j, j = 1, \dots, q, (q < m)$, and use these projections for further processing.

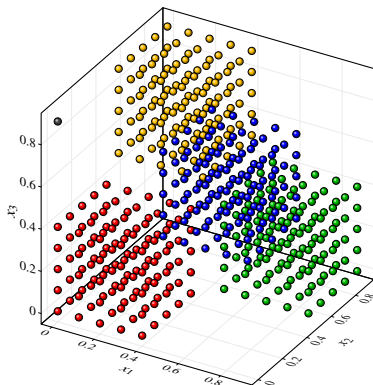
The selection of the value q is based on the preservation of the variance of the original data (for instance, 95%).

Note that PCA is a linear method for dimension reduction.

Principal component analysis: Dimension reduction

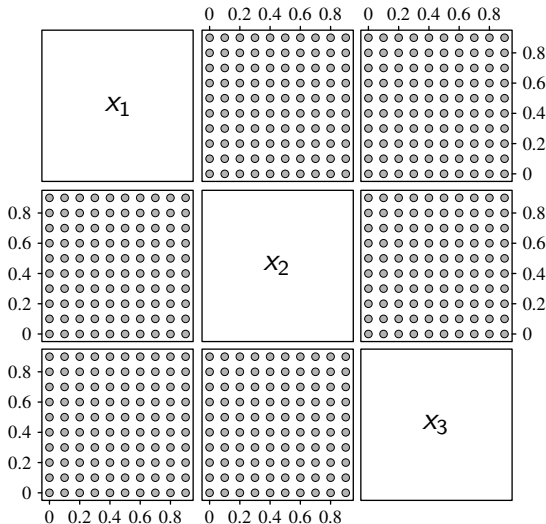


Principal component analysis



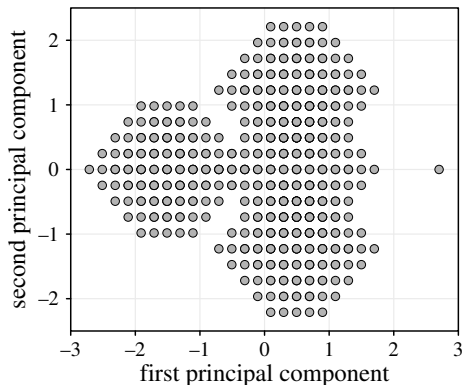
An artificial data set filling a cube in a chessboard-like manner with one outlier.

Principal component analysis



Scatter plots of the “cube data”.

Principal component analysis



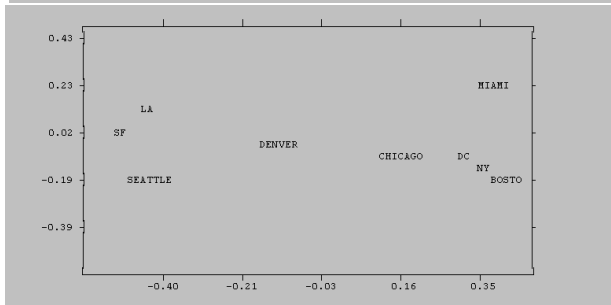
Projection of the “cube data” to the first two principal components

Multidimensional scaling (MDS)

- ▶ Multidimensional scaling (MDS) is not restricted to mappings in the form of simple projections.
- ▶ In contrast to PCA, MDS does not even construct an explicit mapping from the high-dimensional space to the low-dimensional space.
 - ▶ It only positions the data points in the low-dimensional space.
- ▶ The representation of the data in the low-dimensional space constructed by MDS aims at preserving the distances between the data points
 - ▶ Not like PCA the variance in the data set.

Multidimensional scaling (MDS)

		1	2	3	4	5	6	7	8	9
		BOST	NY	DC	MIAM	CHIC	SEAT	SF	LA	DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0



Projection of the city distances to 2D by MDS

Multidimensional scaling

MDS requires a distance matrix $\left[d_{ij}^{(X)} \right]_{1 \leq i, j \leq n}$ where $d_{ij}^{(X)}$ is the distance between data object i and data object j .

- ▶ The distances should be non-negative: $d_{ij}^{(X)} \geq 0$.
- ▶ The distance matrix should be symmetric: $d_{ij}^{(X)} = d_{ji}^{(X)}$.
- ▶ The entries in the diagonal should be zero: $d_{ii}^{(X)} = 0$.
(Each data object has distance zero to itself.)

Usually, the distances are the Euclidean distances of the data objects (after normalization) in the high-dimensional space.

Multidimensional scaling

MDS must define a point $\mathbf{y}_i \in \mathbb{R}^q$ (usually $q = 2$, sometimes also $q = 3$) for each data object \mathbf{x}_i such that the distances $d_{ij}^{(Y)}$ between the points \mathbf{y}_i and \mathbf{y}_j ($i \in \{1, \dots, n\}$) are (roughly) the same as the distances $d_{ij}^{(X)}$ between the original data objects \mathbf{x}_i and \mathbf{x}_j .

Usually

$$d_{ij}^{(Y)} = \| \mathbf{y}_i - \mathbf{y}_j \| .$$

Multidimensional scaling: Objective functions

$$E_0 = \sum_{i=1}^n \sum_{j=i+1}^n \left(d_{ij}^{(Y)} - d_{ij}^{(X)} \right)^2 \quad (\text{sum of squared error})$$

$$E_1 = \frac{1}{\sum_{i=1}^n \sum_{j=i+1}^n \left(d_{ij}^{(X)} \right)^2} \sum_{i=1}^n \sum_{j=i+1}^n \left(d_{ij}^{(Y)} - d_{ij}^{(X)} \right)^2$$

(normalised sum of squared error)

- ▶ The normalisation factor $\frac{1}{\sum_{i=1}^n \sum_{j=i+1}^n \left(d_{ij}^{(X)} \right)^2}$ does not have an influence on the location of the minimum of the objective function.
- ▶ In contrast to E_0 , the value of E_1 does neither depend on the number of data objects nor on the magnitude of the original distances.

Multidimensional scaling: Objective functions

$$E_2 = \sum_{i=1}^n \sum_{j=i+1}^n \left(\frac{d_{ij}^{(Y)} - d_{ij}^{(X)}}{d_{ij}^{(X)}} \right)^2 \quad (\text{relative sum of squared error})$$

$$E_3 = \frac{1}{\sum_{i=1}^n \sum_{j=i+1}^n d_{ij}^{(X)}} \sum_{i=1}^n \sum_{j=i+1}^n \frac{\left(d_{ij}^{(Y)} - d_{ij}^{(X)} \right)^2}{d_{ij}^{(X)}}$$

(mixture between relative and absolute sum of squared error)

MDS based on E_3 is called **Sammon mapping**. The value of E_3 is called **stress**.

Multidimensional scaling

- ▶ MDS represents a non-linear optimisation problem with $q \cdot n$ ($2n$ for $q = 2$) parameters to be optimised.

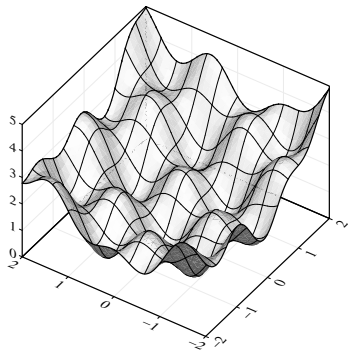
Even for a small data set like the Iris data set, a two-dimensional MDS representation requires the optimisation of 300 parameters.

- ▶ Since the problem is non-linear, a [gradient descent method](#) is used to minimise the objective function for MDS.

Reminder: Gradient descend method

Given a differentiable objective function $f : \mathbb{R}^k \rightarrow \mathbb{R}$ with k parameters for which we want to find the minimum.

f defines a $(k + 1)$ -dimensional “landscape” in which we want to find the lowest point.



Note that the landscape cannot be plotted for $k > 2$.

Reminder: Gradient descend method: Algorithm

1. Start with an arbitrary initial solution \mathbf{x}_0 .

2. Compute the gradient of f at \mathbf{x}_0 .

The gradient is the vector of partial derivatives

$$\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_k} \right)^\top.$$

The gradient points in the direction of steepest ascend.

3. From \mathbf{x}_0 go a fixed (or variable) step width in the opposite direction of the gradient (the direction of steepest descend) to find the next solution \mathbf{x}_1 .

4. Continue in the same way with \mathbf{x}_1 (steps 2 and 3), \mathbf{x}_2, \dots until a stop criterion is satisfied.

((almost) no change of the \mathbf{x}_i , maximum number of iteration steps, no improvement of the value for the objective function, ...)

Multidimensional scaling: Gradient for E_1

$$\frac{\partial E_1}{\partial \mathbf{y}_k} = \frac{2}{\sum_{i=1}^n \sum_{j=i+1}^n \left(d_{ij}^{(X)}\right)^2} \sum_{j \neq k} \left(d_{kj}^{(Y)} - d_{kj}^{(X)}\right) \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}^{(Y)}}$$

where

$$\frac{\partial d_{ij}^{(Y)}}{\partial \mathbf{y}_k} = \frac{\partial}{\partial \mathbf{y}_k} \|\mathbf{y}_i - \mathbf{y}_j\| = \begin{cases} \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}^{(Y)}} & \text{if } i = k, \\ 0 & \text{otherwise} \end{cases}$$

was used.

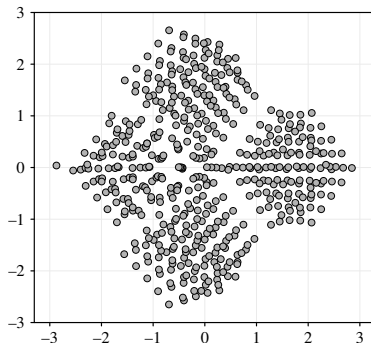
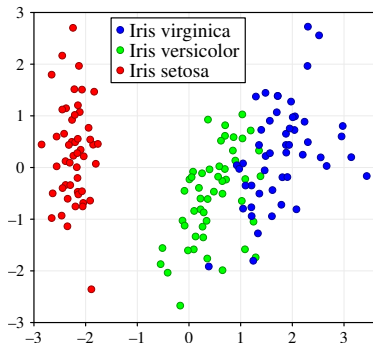
Multidimensional scaling: Algorithm

Algorithm MDS(\mathcal{D})

input: data set $\mathcal{D} \subset \mathbb{R}^m$ with $|\mathcal{D}| = n$ or distance matrix $[d_{i,j}^{(X)}]_{1 \leq i,j \leq n}$
parameter: dimension q for the representation, stepwidth $\alpha > 0$, stop criterion SC
output: set Y of n points in \mathbb{R}^q

```
1 Initialize  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \subset \mathbb{R}^q$  randomly or better with a PCA projection
2 If the input is a data set, compute the distances  $d_{ij}^{(X)}$ 
  between the data objects
3 do
4     Compute  $d_{i,j}^{(Y)} = \|\mathbf{y}_i - \mathbf{y}_j\|$  (for all  $i, j = 1, \dots, n$ )
5     Compute  $\partial E_1 / \partial \mathbf{y}_k$  (for all  $k = 1, \dots, n$ )
6     update  $\mathbf{y}_k^{\text{new}} = \mathbf{y}_k^{\text{old}} - \alpha \cdot \partial E_1 / \partial \mathbf{y}_k$  (for all  $k = 1, \dots, n$ )
7 while SC is not satisfied
```

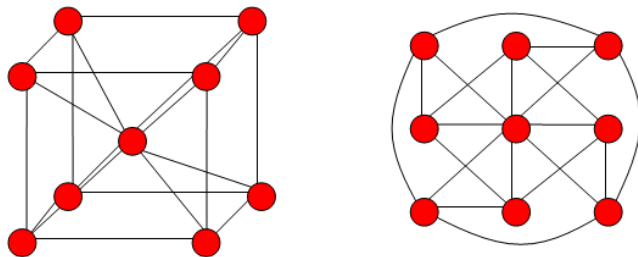
Multidimensional scaling



MDS (Sammon mapping) for the Iris and the “cube” data sets

Multidimensional scaling

Notice that in general MDS cannot preserve all distances of the original space in a lower dimensional space.



MDS from 3D to 2D: Corner points and center point of a cube.

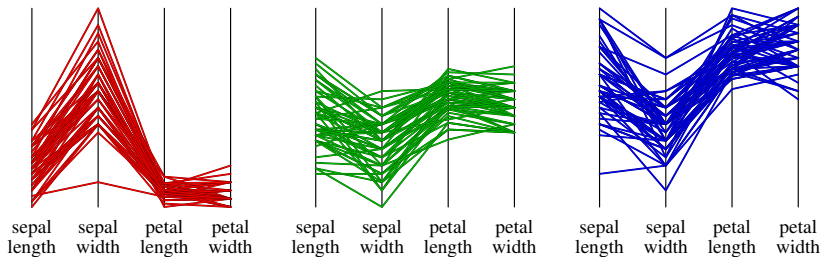
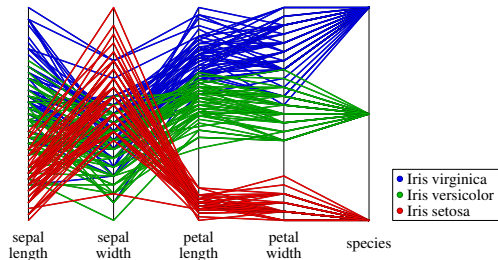
PCA versus MDS

- ▶ PCA is based on variance preservation, whereas MDS tries to preserve distances
- ▶ PCA provides an explicit mapping from the original space to a lower-dimensional space
- ▶ MDS provides explicit representation of the data in the lower dimensional space
 - ▶ New data need recalculation
- ▶ PCA has lower computational complexity
- ▶ PCA provides a linear mapping, whereas MDS can find nonlinear properties

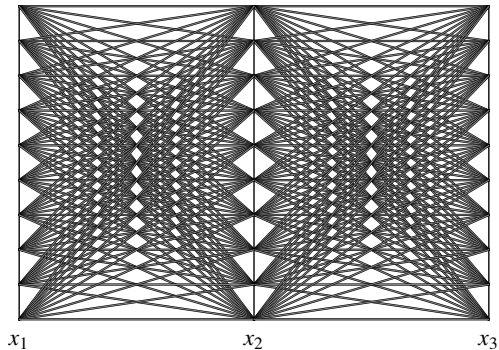
Parallel coordinates draw the coordinate axes parallel to each other, so that there is no limitation for the the number of axes to be displayed.

For a data object, a polyline is drawn connecting the values of the data object for the attributes on the corresponding axes.

Parallel coordinates: Iris data

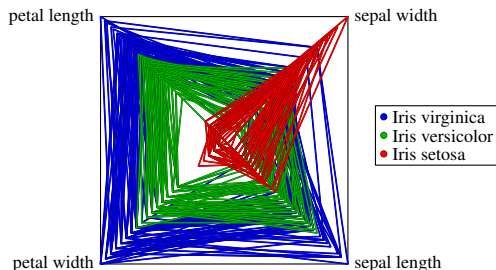


Parallel coordinates: “Cube data”



Radar plots

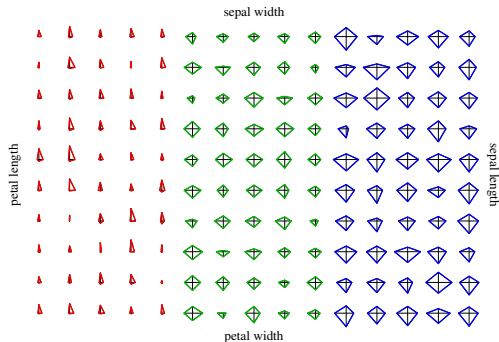
Radar plots are based on a similar idea as parallel coordinates with the difference that the coordinate axes are drawn as parallel lines, but in a star-like fashion intersecting in one point.



Radar plot for the Iris data set

Star plots

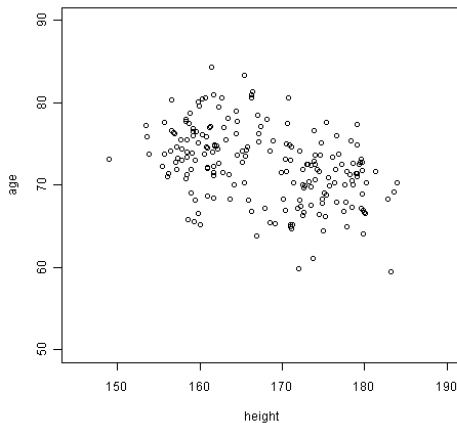
Star plots are the same as radar plots where each data object is drawn separately.



Star plot for the Iris data set

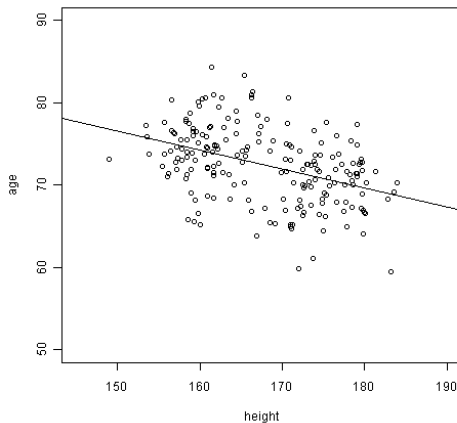
Hidden variables

Do smaller persons live longer?



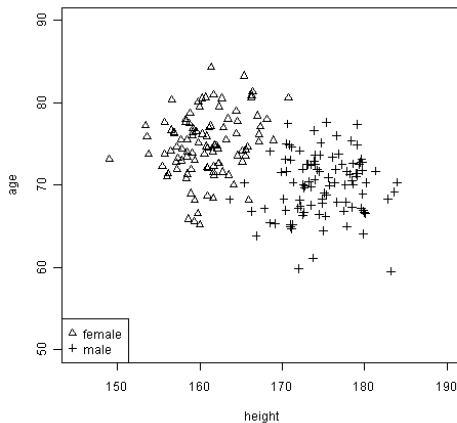
Hidden variables

Do smaller persons live longer?



Hidden variables

Do smaller persons live longer?



Scatter plots can “visually” reveal correlations or dependencies between two attributes.

Statistical measures for correlation are a more formal approach to data analysis and can be carried out automatically.

Remember the limitations of linear correlations!

Pearson's correlation coefficient

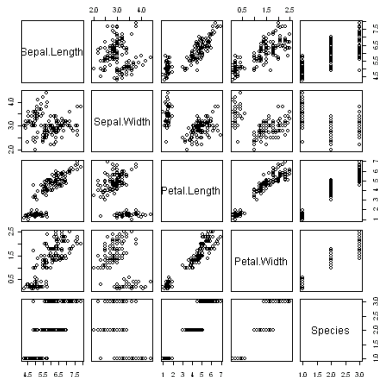
The (sample) Pearson's correlation coefficient is a measure for a linear relationship between two numerical attributes X and Y and is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where \bar{x} and \bar{y} are the mean values of the attributes X and Y , respectively. s_x and s_y are the corresponding (sample) standard deviations.

- ▶ $-1 \leq r_{xy} \leq 1$
- ▶ The larger the absolute value of the Pearson correlation coefficient, the stronger the linear relationship between the two attributes.
For $|r_{xy}| = 1$ the values of X and Y lie exactly on a line.
- ▶ Positive (negative) correlation indicates a line with positive (negative) slope.

Pearson's correlation coefficient: Iris data set



	sepal length	sepal width	petal length	petal width
sepal length	1.000	-0.118	0.872	0.818
sepal width	-0.118	1.000	-0.428	-0.366
petal length	0.872	-0.428	1.000	0.963
petal width	0.818	-0.366	0.963	1.000

Rank correlation coefficients

- ▶ Pearson's correlation coefficient measures linear correlation.
- ▶ Even for monotone functional, but non-linear relationship Pearson's correlation coefficient will not be -1 or 1 .
- ▶ It can even be close to zero despite a monotone functional relationship.

Rank correlation coefficients avoid this problem by ignoring the exact numerical values of the attributes and considering only the ordering of the values.

Rank correlation coefficients intend to measure monotonous correlations between attributes where the monotonous function does not have to be linear.

Spearman's rank correlation coefficient (Spearman's rho)

Spearman's rank correlation coefficient (Spearman's rho) is defined as

$$\rho = 1 - 6 \frac{\sum_{i=1}^n (r(x_i) - r(y_i))^2}{n(n^2 - 1)},$$

where $r(x_i)$ is the rank of value x_i when we sort the list (x_1, \dots, x_n) in increasing order. $r(y_i)$ is defined analogously.

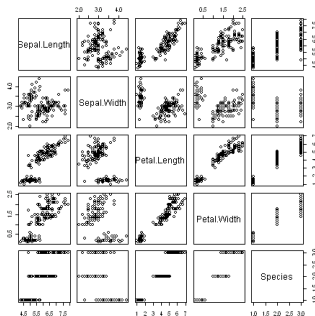
- ▶ When the rankings of the x - and y -values are exactly in the same order, Spearman's rho will yield the value 1 (e.g. 1,2,3,4 vs. 1,2,3,4).
- ▶ If they are in reverse order, we will obtain the value -1 (e.g. 1,2,3,4 vs. 4,3,2,1).

Spearman's rho: example

- ▶ Variable x = success in school (1=bad, 5=good)
- ▶ Variable y = enjoying school (1=no, 5=yes)
- ▶ $\rho = 1 - \frac{6 \cdot 35.5}{10^3 - 10} \approx 0.78$

x	y	$R(x)$	$R(y)$	$d=R(x)-R(y)$
1	2	1.5	5	-3.5
3	2	6.5	5	1.5
4	4	8.5	8	0.5
2	1	4	2	2
1	1	1.5	2	-0.5
2	3	4	7	-3
4	5	8.5	9.5	-1
5	5	10	9.5	0.5
2	1	4	2	2
3	2	6.5	5	1.5

Spearman's rho: Iris data set



	sepal length	sepal width	petal length	petal width
sepal length	1.000	−0.167	0.882	0.834
sepal width	−0.167	1.000	−0.289	−0.289
petal length	0.882	−0.289	1.000	0.938
petal width	0.834	−0.289	0.938	1.000

Kendall's tau rank correlation coefficient (Kendall's tau)

Kendall's tau rank correlation coefficient (Kendall's tau) is defined as

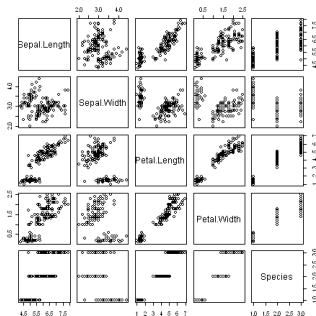
$$\tau_a = \frac{C - D}{\frac{1}{2}n(n-1)}$$

where C and D denote the numbers of concordant and discordant pairs, respectively.

$$C = |\{(i, j) \mid x_i < x_j \text{ and } y_i < y_j\}|$$

$$D = |\{(i, j) \mid x_i < x_j \text{ and } y_i > y_j\}|$$

Kendall's tau: Iris data set



	sepal length	sepal width	petal length	petal width
sepal length	1.000	-0.077	0.719	0.655
sepal width	-0.077	1.000	-0.186	-0.157
petal length	0.719	-0.186	1.000	0.807
petal width	0.655	-0.157	0.807	1.000

In statistics, an **outlier** is a value or data object that is far away or very different from all or most of the other data.

Grubbs, F. E.: Procedures for detecting outlying observations in samples. Technometrics 11, 1-21, 1969:

- ▶ An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

Outliers may be indicative of data points that belong to a different population than the rest of the sample set (example from Wikipedia):

- ▶ One is calculating the average temperature of 10 objects in a room, and most are between 20 and 25 degrees Celsius, but an oven is at 175 °C.
- ▶ The median of the data may be 23 °C but the mean temperature will be about 40 °C.
- ▶ In this case, the median better reflects the temperature of a randomly sampled object than the mean.
- ▶ However, naively interpreting the median as "a typical sample", is incorrect.

Causes for outliers:

- ▶ Data quality problems (erroneous data coming from wrong measurements or typing mistakes)
 - ▶ Exceptional or unusual situations/data objects.
-
- ▶ Outliers coming from erroneous data should be excluded from the analysis.
 - ▶ Even if the outliers are correct (exceptional data), it is sometime useful to exclude them from the analysis.
 - ▶ For example, a single extremely large outlier can lead to completely misleading values for the mean value

Outlier detection: Single attributes

Categorical attributes: An outlier is a value that occurs with a frequency extremely lower than the frequency of all other values.

In some cases, the outliers can even be the target objects of the analysis.

Example: Automatic quality control system

Goal: Train a classifier, classifying the parts as correct or with failures based on measurements of the produced parts.

The frequency of the correct parts will be so high that the parts with failure might be considered as outliers.

Outlier detection: Single attributes

Numerical attributes:

- ▶ Outliers in boxplots.
 Problems: Asymmetric distribution, large data sets
- ▶ Statistical tests, for example **Grubb's test**:

Grubbs' test is based on the assumption of normality (check the normality first)

Grubbs' test is defined for the hypothesis:

- ▶ **H0:** There are no outliers in the data set
- ▶ **H1:** There is at least one outlier in the data set

Define the statistic $G = \frac{\max\{|x_i - \bar{x}| \mid 1 \leq i \leq n\}}{s}$ where x_1, \dots, x_n is the sample, \bar{x} its mean value and s its empirical standard deviation.

Outlier detection: Single attributes

This is the two-sided version of the test.

To test if the minimum value is an outlier, the test statistic is

$$G = \frac{\bar{x} - \min \{x\}}{s} .$$

To test if the maximum value is an outlier, the test statistic is

$$G = \frac{\bar{x} - \max \{x\}}{s} .$$

For the two-sided test, the hypothesis of no outliers is rejected at significance level α if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n), n-2}^2}{n-2 + t_{\alpha/(2n), n-2}^2}}$$

with $t_{\alpha/(2n), n-2}$ denoting the upper critical value of the t -distribution with $(n-2)$ degrees of freedom.

Outlier detection: Single attributes

Grubb's test applied to the Iris data set:

attribute	p-value
sepal length	0.92
sepal width	0.13
petal length	1.0
petal width	1.0

- ▶ The p-values do not indicate any outliers:
 - ▶ Traditionally, one rejects the null hypothesis if the p-value is less than or equal to the significance level α , which normally is 0.05 or 0.01.
- ▶ Note that the assumption of normal distribution is not correct. The attributes from one species might follow a normal distribution, but not the values from all species together.

Outlier detection for multidimensional data

- ▶ Scatter plots for (visually detecting) outliers w.r.t. two attributes.
- ▶ PCA or MDS plots for (visually detecting) outliers.
- ▶ Cluster analysis techniques

Missing values

For some instances values of single attributes might be missing.

Causes for missing values:

- ▶ broken sensors
- ▶ refusal to answer a question
- ▶ irrelevant attribute for the corresponding object
(pregnant (yes/no) for men)

Missing value might not necessarily be indicated as missing (instead: zero or default values).

Types of missing values

Consider the attribute X_{obs} . A missing value is denoted by $?$.

X is the true value of the considered attribute, i.e. we have

$$X_{\text{obs}} = X, \quad \text{if } X_{\text{obs}} \neq ?$$

Let Y be the (multivariate) (random) variable denoting the other attributes apart from X .

Types of missing values

Missing completely at random (MCAR): The probability that a value for X is missing does neither depend on the true value of X nor on other variables.

$$P(X_{\text{obs}} = ?) = P(X_{\text{obs}} = ? \mid X, Y)$$

Example: The maintenance staff sometimes forgets to change the batteries of a sensor, so that the sensor sometimes does not provide any measurements.

MCAR is also called **Observed At Random (OAR)**.

Types of missing values

Missing at random (MAR): The probability that a value for X is missing does not depend on the true value of X .

$$P(X_{\text{obs}} = ? \mid Y) = P(X_{\text{obs}} = ? \mid X, Y)$$

Example: The maintenance staff does not change the batteries of a sensor when it is raining, so that the sensor does not always provide measurements when it is raining.

Types of missing values

Nonignorable: The probability that a value for X is missing depends on the true value of X .

Example: A sensor for the temperature will not work when there is frost.

In the cases of MCAR and MAR, the missing values can be estimated – at least in principle, when the data set is large enough – based on the values of the other attributes.

In the extreme case of the sensor for the temperature, it is impossible to make provide any statement concerning temperatures below 0°C .

Types of missing values

- ▶ In the case of MCAR, it can be assumed that the missing values follow the same distribution as the observed values of X .
- ▶ In the case of MAR, the missing values might not follow the distribution of X . But by taking the other attributes into account, it is possible to derive reasonable imputations for the missing values.
- ▶ In the case of nonignorable missing values it is impossible to provide sensible estimations for the missing values.

Types of missing values

If it is not known based on domain knowledge which kind of missing values can be expected, the following strategy can be applied.

1. Turn the considered attribute X into a binary attribute:
 - ▶ Replace all measured values by the values *yes* and all missing values by the value *no*.
2. Build a classifier with binary attribute X as the target attribute and use all other attributes for the prediction of the class values *yes* and *no*.
3. Determine the misclassification rate. The misclassification rate is the proportion of data objects that are not assigned to the correct class by the classifier.

Types of missing values

- ▶ In the case of **OAR**, the other attributes should not provide any information, whether X has a missing value or not.
 - ▶ Therefore, the misclassification rate of the classifier should not differ significantly from pure guessing, i.e. if there 10% missing values for the attribute X , the misclassification rate of the classifier should not be much smaller than 10%.
- ▶ If, however, the misclassification rate of the classifier is significantly better than pure guessing, this is an indicator that there is a correlation between missing values for X and the values of the other attributes. The missing values are not **OAR**.
- ▶ **MAR** and **nonignorable** cannot be distinguished in this way.

A checklist for data understanding

- ▶ There are general and specific goals for data understanding
- ▶ One important part of data understanding is to get an idea of the data quality.
 - ▶ There are standard data quality problems like syntactic accuracy which are easy to check.
- ▶ There are various methods to support the identification of outliers:
 - ▶ Methods exclusively designed for outlier detection
 - ▶ Visualization techniques like boxplots, histograms, scatter plots, projections based on PCA and MDS that can help to find outliers, but are also useful for other purposes.

A checklist for data understanding

- ▶ Missing values are another concern of data quality.
- ▶ When there are explicit missing values, i.e. entries that are directly marked as missing, then try to find out of which type—OAR, MAR or non-ignorable—they are.
- ▶ Use domain knowledge, but also classification methods can be applied.
- ▶ Be aware of the possibility of hidden missing values that are not explicitly marked as missing. The simplest case might be hidden missing values that have a default value.
- ▶ Histograms might help to identify candidates for such hidden missing values when there are unusual peaks.
- ▶ However, there is no standard test or technique to identify possible hidden missing values.
 - ▶ Therefore, whenever we see something unexpected in the data, hidden missing values of a specific type might be one explanation.

A checklist for data understanding

- ▶ Data understanding should also help to discover new or confirm expected dependencies or correlations between attributes.
 - ▶ Correlation analysis to solve this task.
 - ▶ Scatter plots can show correlations between pairs of attributes.
- ▶ Specific application dependent assumptions should be checked
 - ▶ For instance, the assumption that specific attribute follows a normal distribution
- ▶ Representativeness of the data cannot always be checked just based on the data, but we have to compare the statistics with our expectations.
 - ▶ If we suspect that there is a change in a numerical attribute over time, we can compare histograms or boxplots for different time periods.
 - ▶ We can do the same with bar charts for categorical attributes.

A checklist for data understanding

- ▶ Check the distributions for each attribute whether there are unusual or unexpected properties like outliers.
 - ▶ Are the domains or ranges correct?
 - ▶ Do the medians of numerical attributes look correct?
 - ▶ Histograms and boxplots for continuous attributes
 - ▶ Bar charts for categorical attributes.
- ▶ Check correlations or dependencies between pairs of attributes with scatter plots which should be density-based for larger data sets.
 - ▶ For small numbers of attributes, inspect scatter plots for all pairs of attributes.
 - ▶ For higher numbers of attributes, do not generate scatter plots for all pairs, but only for those ones where independence or a specific dependency is expected.
 - ▶ Generate in addition scatter plots for some randomly chosen pairs.