Statistical Learning

Multidimensional Data

Pedro Galeano
Department of Statistics
UC3M-BS Institute on Financial Big Data
Universidad Carlos III de Madrid
pedro.galeano@uc3m.es

Academic year 2016/2017

Master in Big Data Analytics

uc3m | Universidad Carlos III de Madrid

Introduction

- Statistics: Wide-ranging discipline with certain doses of Mathematics, Empirical Science, Computer Science and Philosophy, among many others.
- Definition: The art and science of learning from data.
- Birth: Discipline dates to 1763, with the publication of the Bayes' rule as an argument for the existence of God.
- Classical Statistics (1763–1960 aprox.): Develop methods, mostly based in Probability Theory, to make inference and prediction.
- Modern Statistics (1960–2000 aprox.): Develop algorithms with the same goals taking advantage of the fact that computation became faster and easier.
- Computer Age Statistics (2000–): Almost all topics in 21th-century Statistics are now computer dependent.

Introduction

- Usual questions treated by Statistics:
  - Does smoking really cause cancer?
  - Does unemployment cause return migration?
  - What will be the maximum temperature of tomorrow?
- Statistical approach: Collect, process and analyze data.
- How to analyze data?: Develop appropriate statistical models and/or methods, perform inference, and predict (or forecast) future outcomes.
- Limitations: Computation has been the traditional bottleneck of statistical applications.
- However: Times are changing...

Introduction

TECHNOLOGY

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR AUG. 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

Introduction

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by 2012, according to a projection by IDC, a research firm.

Yet data is merely the raw material of knowledge. "We're rapidly entering a world where everything can be monitored and measured," said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology's Center for Digital Business. "But the big problem is going to be the ability of humans to use, analyze and make sense of the data."

Introduction

The new breed of statisticians tackle that problem. They use powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data. The applications are as diverse as improving Internet search and online advertising, culling gene sequencing information for cancer research and analyzing sensor and location data to optimize the handling of food shipments.

Even the recently ended Netflix contest, which offered $1 million to anyone who could significantly improve the company's movie recommendation system, was a battle waged with the weapons of modern statistics.

Though at the fore, statisticians are only a small part of an army of experts using modern statistical techniques for data analysis. Computing and numerical skills, experts say, matter far more than degrees. So the new data sleuths come from backgrounds like economics, computer science and mathematics.

Introduction

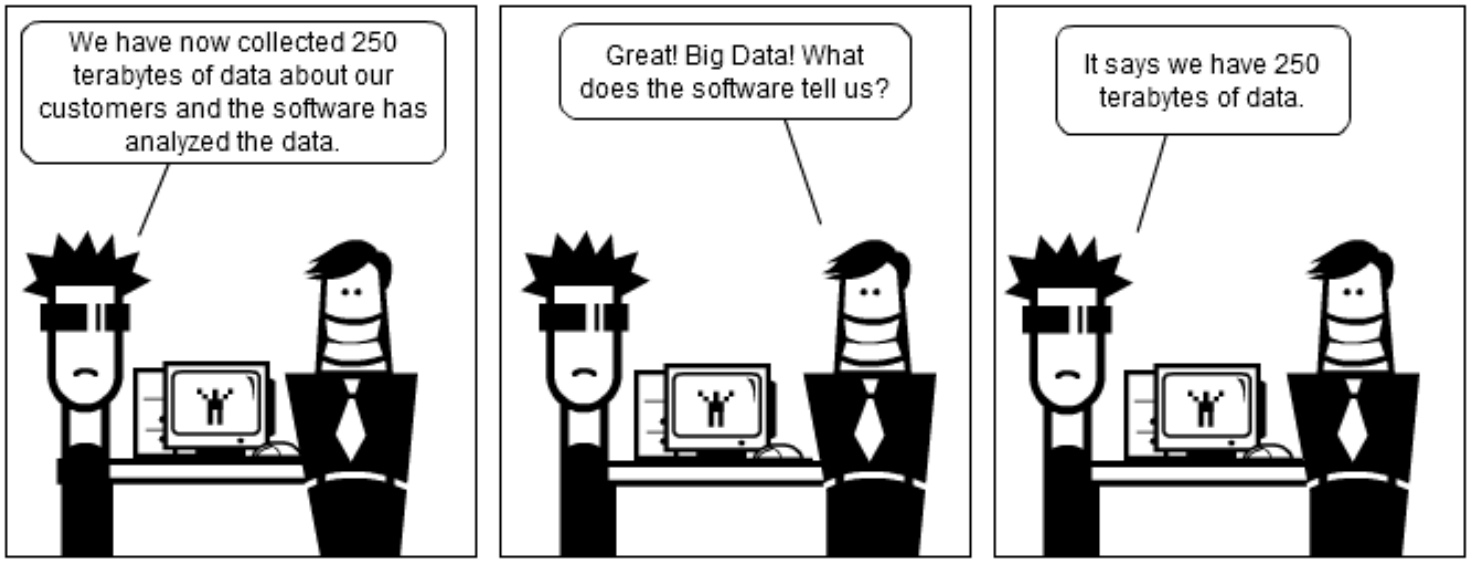
- Nowadays, the world is awash with data coming from:
  - Digital media (books, newspapers,...).
  - E-commerce.
  - Genome.
  - Images and videos.
  - Internet.
  - Medical devices.
  - Social networks.
  - Speech recognition.
  - ...

Introduction

- Think in your day-to-day activities:
  - Read your personal email: Find recommendations to buy products from certain stores based on your previous purchases.
  - Go to the bank: Everyone is subject of a scoring model that predicts whether he/she will default on his/her loan in the near future.
  - Use your smartphone: Your provider analyzes your calling behavior to predict whether you are going to churn in the near future.
  - Read an internet newspaper: You get advertisements based on pages you use to visit.
  - Buy products in the supermarket: I use my credit card, then my provider needs to know if this is a legitimate transaction or not.
  - ...

Introduction

- Nevertheless: "We are drowning in information and starving for knowledge" (Rutherford D. Rogers):



- Crucial need: Take advantage of the large amount of information available today to achieve a thorough understanding of it, by means of data analysis.
- The signal and the noise: The key problem is how to separate the signal from the noise, because only a small percentage of data available is useful.

Introduction

- How to do it?: Many problems found in practice can be analyzed with classical, modern and computer-age statistical techniques.
- Crucial task: It is necessary to do things properly.
- Otherwise: We will kill the goose that lays the golden eggs:
  - http://science.sciencemag.org/content/343/6176/1203
  - http://elpais.com/elpais/2016/11/07/talento_digital/1478535225_341110.html
- Parallel approaches: In machine learning. See the courses "Big Data Intelligence: Methods and Technologies" and "Machine Learning" of the Master's program.
- Examples: The next slides show practical problems that are handled with very well known statistical techniques.

Introduction

- Business analytics: credit risk modeling
  - Reform measures such as Basel III: Banks need to develop systems in an attempt to model the credit risk arising from important aspects of their business lines.
  - How?: Take different snapshots of information: application and credit bureau information at loan origination, default status information,...
  - Problem: Classify a consumer as good or bad.
  - Technique: Use supervised classification methods taking into account that a labeled data set with good and bad clients (thousands, up to millions) is available.

Introduction

- Business analytics: fraud detection
  - Typical examples: Credit card fraud, insurance claim fraud, money laundering, tax evasion, product warranty fraud, and click fraud, among others.
  - Problem: Label certain operations as fraudulent or not.
  - Techniques:
    - ⋆ Use supervised classification methods taking into account that a labeled data set with fraud objects is available.
    - ⋆ Use unsupervised classification and outlier detection procedures to detect clusters of abnormal operations.
    - ⋆ Use social network analysis to analyze the relationships between fraudsters. Exploiting relational information provides some interesting insights in criminal patterns and activities.

Introduction

- Business analytics: net lift modeling
  - Net lift modeling: Deepen customer relationships by means of targeted or win-back campaigns: mail catalog, email, coupon,...
  - Purpose: Identify customers most likely to respond based on demographic and relationship variables, social network information, and RFM (recency, frequency and monetary) variables.
  - Problem: Classify consumer attitudes (not buying, buying, buying if motivated correctly,...)
  - Technique: Use unsupervised classification to build consumers attitudes and supervised classification to perform future classifications.

Introduction

- Genomics:
  - Microarrays: There are more than 500000 microarrays that are publicly available with each array containing tends of thousands of expression values of molecules.
  - Goal: The large amount of genome sequencing data now make possible to uncover genetic markers of rare disorders and find associations between diseases and rare sequence variants.
  - Characteristics: In such microarrays, the number of variables scales in thousands and is usually larger than the number of individuals involved in the experiments.
  - Problem: Large-scale hypothesis testing is frequently used to pick important genes or proteins.

Introduction

- Image (and video) analysis:
  - Images and videos: Ones of the most frequent sources of information that are nowadays available for analysis.
  - Examples: Medical images, astrophysical images, surveillance images,...
  - Pixels: Each image can be represented by millions of pixels.
  - Techniques:
    - ⋆ Dimension reduction techniques are frequently used to reduce the size of image and video files.
    - ⋆ Supervised and unsupervised classification methods are used to find similar images (or videos).

Introduction

- Social networks:
  - Social network data: Massive amount of data are being produced by Twitter, Facebook, Youtube,...
  - Uses: These data have been exploited to predict influenza epidemic, stock market trends, and box-office revenues for movies, among many others.
  - Techniques: Use supervised classification, regression and time series models for prediction and forecasting.

Introduction

- Structured data sets: In this course, we will be concerned with techniques for handling well structured data sets.
- Well structured data set: After a data processing exercise, the idea is to analyze a certain collection of characteristics in a certain set of objects.
- Data base processing: A large amount of work should be usually done in order to get a well structured data set.

Introduction

- Non-structured data sets: Some non-structured data sets can be converted to well structured data sets.
- For instance: Digital media, like books or newspapers, can be converted into certain structured data sets through text mining techniques (see the courses "Big Data Intelligence: Methods and Technologies" and "Machine Learning" of the Master's program.).
- Examples of well structured data sets: See, next, a few examples of well structured data sets.

Introduction

- The Spam data set (see Chapter 1.R script):
  - The Spam data set from Hewlett-Packard Labs: Contains the values of 57 variables indicating the frequency of certain words in 4601 e-mails.
  - The e-mails are labeled: Spam (1813 out of 4601) or non-spam (2788 out of 4601).
  - Example of: Multidimensional data.
  - Problem: Is it possible to predict if a new email is spam or not using the information provided by this data set?
  - Chapters: 1, 2 and 3.

Introduction

- The NCI60 data set (see Chapter 1.R script):
  - The data set: Contains expression levels on 6830 gene expression measurements from 64 cancer cell lines.
  - Example of: Multidimensional data.
  - Problem: Are there groups among the cell lines based on their gene expression measurements?
  - Chapters: 1, 2 and 3.

Introduction

- The Births2006 data set (see Chapter 1.R script):
  - The data set: Contains information on 427323 babies born in the United States during 2006 (it is a random ten percent sample).
  - Example of: Multidimensional data.
  - Problem: Is it possible to fill missing values in the data set?
  - Chapters: 1, 2 and 3.

Introduction

- The CanadianWeather data set (see Chapter 1.R script):
  - The data set: Contains daily temperature and precipitation at 35 different locations in Canada in the period from 1960 to 1994.
  - Example of: Functional data.
  - Problems: Several problems can arise including predicting precipitation in terms of temperature or cluster cities with similar weather behavior.
  - Difference with previous data sets: Here, we observe two processes over time.
  - Chapter: 4.

Introduction

- Chapter 2: Dimension reduction techniques
  - Noise: The essential structure of large dimensional data sets is obscured by noise.
  - Dimension reduction: It becomes vital to reduce the original data set in such a way that the interesting structure in the data is preserved while irrelevant features are removed.
  - Principal Component Analysis: Simple, elegant, and surprisingly powerful dimension reduction tool.
  - Complex methods: As time moves on, more complex methods are being developed, although PCA have not lost its appeal.

Introduction

- Chapter 3: Supervised and unsupervised classification methods
  - Supervised classification: Consists on building a statistical model for predicting a qualitative response based on one or more predictors.
    - ⋆ It is assumed that we have a set of well classified observations, i.e., a set of observations with known associated qualitative response.
    - ⋆ The problem consists in make use of this information to construct a classifier of new observations with unknown qualitative response.
    - ⋆ It is possible to check our work by seeing how well our model predicts the qualitative response on observations not used in fitting the statistical model.
  - Unsupervised classification: Also known as clustering, consists on discovering unknown subgroups in data.
    - ⋆ Unsupervised classification is often much more challenging than supervised classification because there is no way to know the true answer.

Introduction

- Chapter 4: Functional data analysis
  - Functional data: Consists on random functions or curves observed discretely at a finite interval.
  - In a conceptual sense: Functional data are intrinsically infinite dimensional and thus, methods designed for multidimensional data sets are no longer applicable.
  - Functional techniques: Solve similar problems to those of multidimensional data but taking into account the functional nature of the data.

Introduction

- The rest of this chapter is devoted to:
  - Introduce the general structure of multidimensional data sets.
  - Present some data quality problems.
  - Show some useful plots.
  - Introduce several interesting descriptive measures.
  - Consider simple linear transformations of the data.
  - Summarize some useful concepts of multidimensional distributions and inference.
  - Illustrate some problems related with correlations and some ideas about how to deal with them.

Multidimensional data sets

- Multidimensional data sets: Multiple measurements or observations obtained on a collection of selected variables.
- Data: Usually, numerical and categorical values after appropriate transformations.
- Data matrix: Large rectangular arrays where rows represent observations and columns represent variables.
- Data matrix of massive sizes: Even if they are stored and manipulated in special database systems, we can still think in terms of data matrices.
- Response vector: In supervised classification problems, additionally to the data matrix, there is a response vector where rows represent values of an indicator variable.

Multidimensional data sets

- Data matrix: The most important object in multidimensional analysis.
- Usually: The data matrix, denoted by X, contains n multidimensional observations taken on p variables:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Size: $n \times p$.
- Sample size: n.
- Dimension: p.

Multidimensional data sets

- Generic element of X: x_{ij}, represents the value of the $j-$th univariate variable over the i-th individual.
- Values of the $j-$th univariate variable: $x_{1j}, \ldots, x_{nj}$, for $j = 1, \ldots, p$, summarized in the column vector given by $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})$.
- Values of the $i-$th observation: $x_{i1}, \ldots, x_{ip}$, for $i = 1, \ldots, n$, summarized in the row vector given by $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})$.

Multidimensional data sets

- Response vector: Useful in supervised classification problems.
- Column vector: The response vector, denoted by Y, contains n multidimensional observations taken on a single indicator variable:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Size: $n \times 1$.
- Sample size: n.
- Generic element of Y: y_i, represents the value of the response variable over the i-th individual.

Multidimensional data sets

- Chapter 1.R script:
  - Identify: Sample sizes and dimensions for the Spam, NCI60 and births2006 data sets.

Multidimensional data sets

- Two main data types:
  - Quantitative variables: Their value is a measurable quantity, such as the variables in the NCI60 data set.
  - Qualitative variables: Their value can be attributed to a category, such as the variable spam in the spam data set or the variable sex in the births2006 data set.

Multidimensional data sets

- Quantitative variables:
  - Continuous: Their values can be read as an interval, such as the expression levels in the NCI60 data sets.
  - Discrete: Their values are distinct and separate, such as the variables capitalLong and capitalTotal in the spam data set.
- Qualitative variables:
  - Binaries: There are only two possible values, such as the variable spam in the spam data set.
  - General: There are more than two possible values, such as the variable DMEDUC in the births2006 data set.

Multidimensional data sets

- Chapter 1.R script:
  - Identify: The previous variables in the Spam, NCI60 and births2006 data sets.

Multidimensional data sets

- Numerical codification: Qualitative variables can be coded numerically.
- For example: The variable spam in the spam data set can be coded with 0, if spam, and 1, if non-spam, or vice versa.
- More information will be given later.

Data quality problems

- Problems in data sets: Problems of all kinds exists.
- Data cleaning: Problems easy to detect will most likely be found at the data cleaning stage.
- Data analysis: Quite resistant problems might only be discovered during data analysis.
- Examples:
  - Inconsistencies: Matching data coming from different sources can create inconsistencies.
  - Outliers: Observations that do not appear to fit the pattern of the other data values.
  - Missing data: Incomplete or totally missed observations.

Data quality problems

- Inconsistencies:
  - Example: Consider different registers from the same person because his name is recorded in different ways: John, Johnny or Jack.
  - Example: Assign a wrong code to a product that already exists.
- Unfortunately: These mistakes are very difficult to find.
- Solution: Check the behavior of the variables carefully and try to correct everything wrong.
- Also: Sometimes, these problems are found when data analysis is performed.

Data quality problems

- Outliers:
  - Gross errors: Outliers can occur for many different reasons but should not be confused with gross errors that are cases where something went wrong, such as human or mechanical errors.
  - Outliers in single variables: Are usually easy to detect as they are values that are very large or very small compared with others in the sample.
  - Multidimensional outliers: Much more difficult to detect.
  - Indeed: Multidimensional outliers are not necessary outliers in the single variables.
  - Methods: There are several multidimensional outlier detection procedures available, although most of them are based on Gaussian assumptions on the variables.
  - Idea: Obtain low-dimensional visual displays of the data and try to detect the most obvious outliers.
  - More information will be given later.

Data quality problems

- Missing data:
  - Missing values: Very frequent in databases.
  - Marks of missing data: In **R**, missing values are flag as NA, in **SQL**, as null,...
  - Complete case analysis: Delete those observations with missing values in variables of interest.
  - However: Only acceptable if the number of observations with missing values is small relative to the size of the data set and if the missing data mechanism is independent of the variables.
  - Example of dependency: A survey where participants older than a certain age refuse to answer a particular survey question and age is measured in the study.

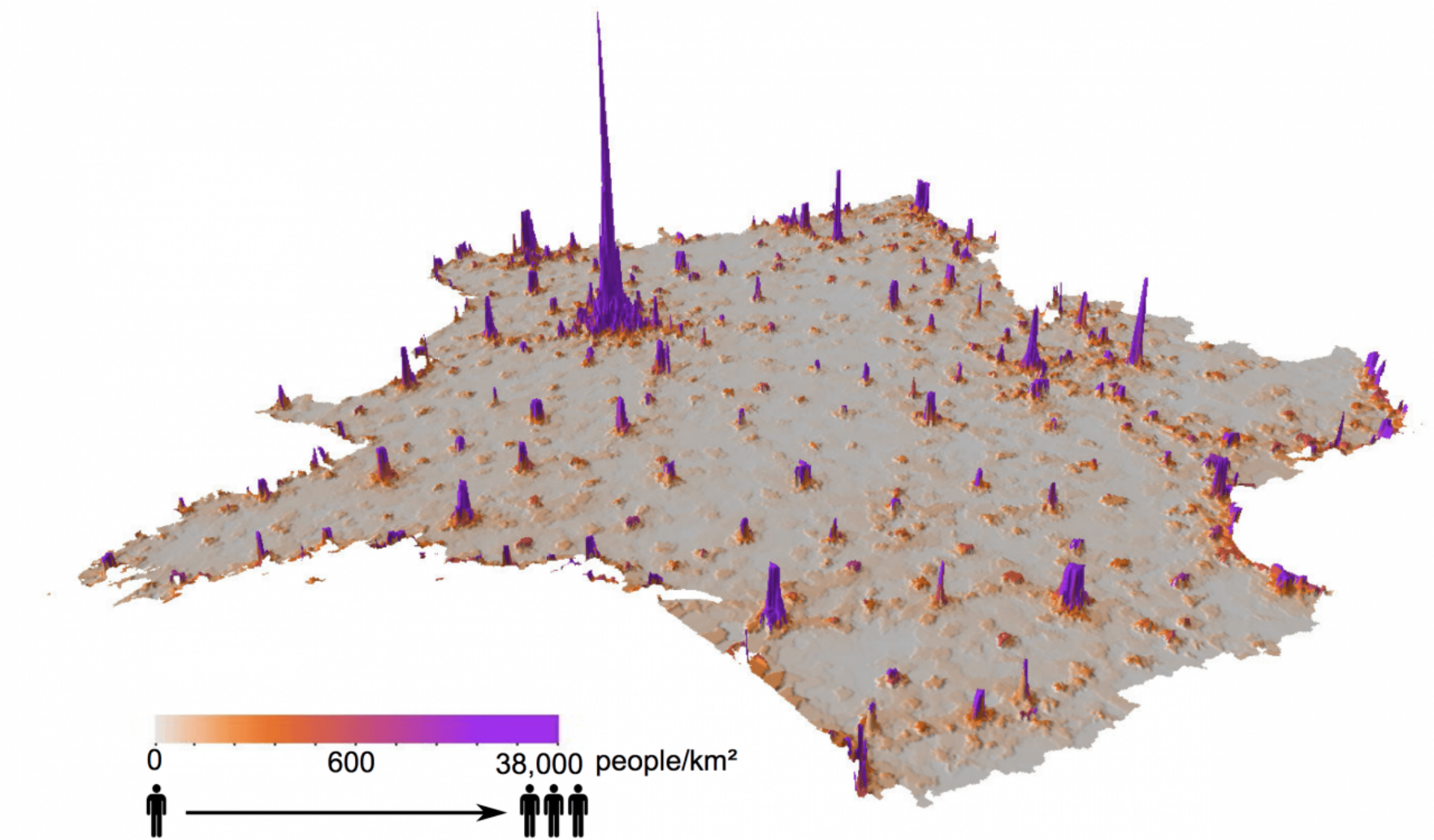
Data quality problems

- Missing data:
  - Single imputation: Fill an estimated value for each missing observation.
    - ⋆ Hot-deck imputation: A missing value is imputed by substituting a value from a similar but complete record in the data set.
    - ⋆ Mean imputation: A single missing value is imputed by substituting the sample mean of all the completely recorded values for that variable.
  - Multiple imputation: Fill estimated values for each missing observation leading to several data sets, then analyze each data set and, finally, combine the results.
    - ⋆ How to do it?: Usually using complex regression and Bayesian methods.
    - ⋆ Large data sets: Multiple imputation is not very well suited for very large data sets.
- More information will be given later.

Visualizing multidimensional data sets

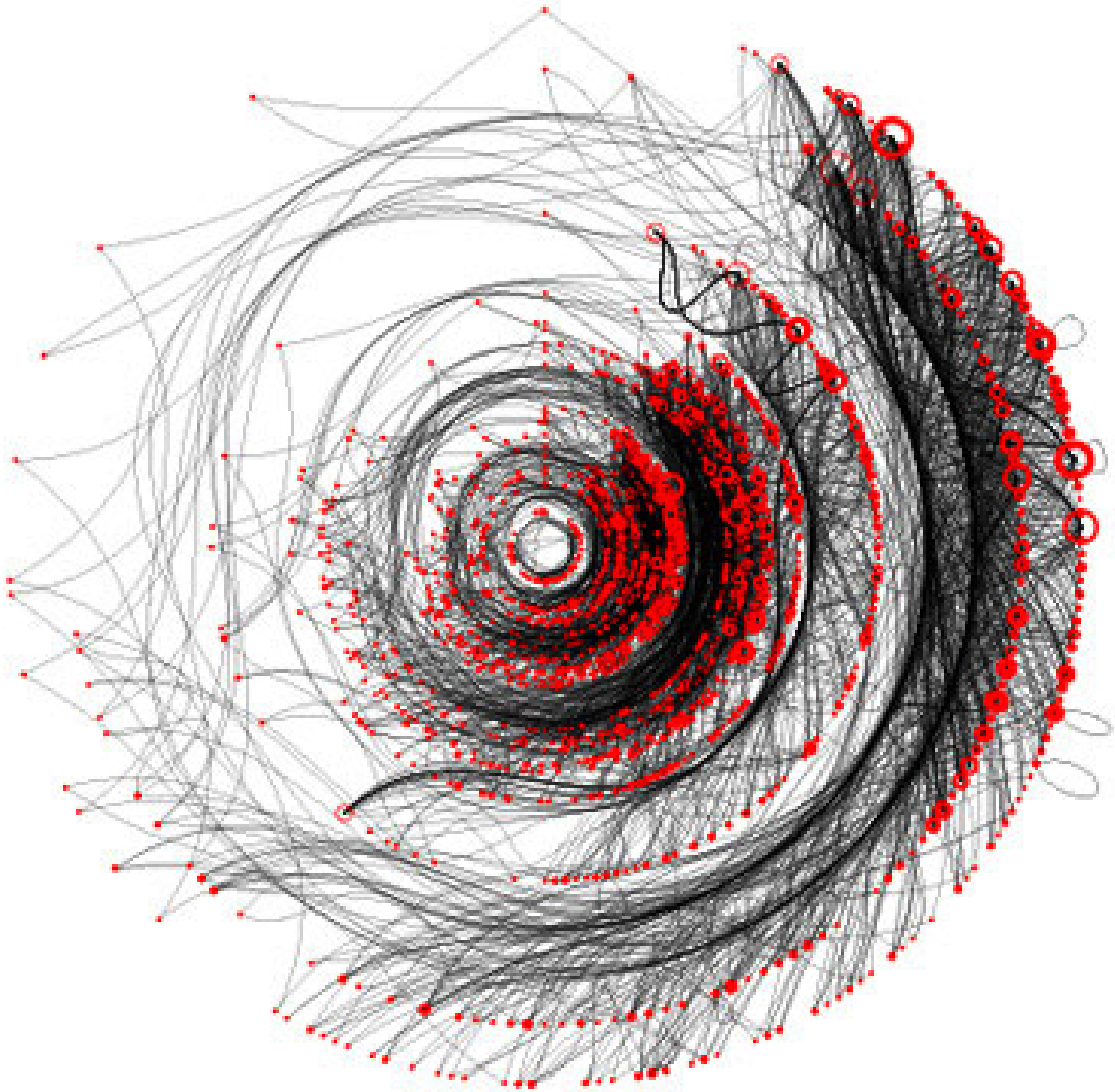
- Graphical displays: Exploratory data analysis tools which can help to understand the data.
- Data visualization: One of the most important statistical research areas nowadays.
- Why?: Because we need to summarize a large bunch of information.
- Important: Graphics strongly depends on the data structure.
- Some examples of modern plots are given in the next slides.

Visualizing multidimensional data sets



Visualizing multidimensional data sets

Visualizing multidimensional data sets



Visualizing multidimensional data sets

- See course: Network analysis and data visualization.
- Here: We are going to focus on informative plots for multidimensional data sets.
- Plots for single variables:
  - Qualitative variables: Often we get an idea of frequencies of appearance of the categories defined by the variables.
  - Quantitative variables: Often we get useful features from plots including:
    - skewness.
    - multimodality.
    - outliers.
    - distint groupins.
    - ...

Visualizing multidimensional data sets

- Plots for multiple variables:
  - One difficulty: The human perceptional system.
  - Point clouds in two dimensions: Are easy to understand and to interpret.
  - See real time 3D rotations: Allow us to perceive three-dimensional data.
  - Qualitative jump: Dimensions greater than or equal to 4.
  - Dimension reduction techniques: Provide low dimensional data sets more easily interpretable.
  - Particular techniques: There are many plots developed for particular techniques.
- Next: Some basic descriptive and graphical techniques allowing simple exploratory data analysis.

Visualizing multidimensional data sets

- Plots for one or two qualitative variables:
  - Barplots and piecharts: Usual plots for single qualitative variables.
  - Absolute frequencies: These plots show the absolute frequencies of the observed values of the variables.
  - Consequently: Shows the proportions of data in each defined category.
  - Problem: When the number of classes is very large, it is recommendable to join classes.
  - Joint barplots: These are barplots that show the proportions of values of two qualitative variables.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Barplots and piecharts: Variable spam in the spam data set and variable DMEDUC in the births2006 data set.
  - Joint barplot: For the variables DMEDUC and SEX in the births2006 data set.

Visualizing multidimensional data sets

- Plots for one quantitative variables:
  - Barplots: Also used for discrete variables, although if the number of different values is very large, it is sometimes advisable to use some of the plots shown below.
  - Boxplots: Simple univariate device that detects outliers variable by variable and that can compare distributions of the data among different groups.
  - Histograms and kernel densities: Basic techniques to estimate density functions of continuous variables, thus providing a quick insight into the shape of the distribution of the data.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Barplot: Variable capitalLong in the spam data set.

Visualizing multidimensional data sets

- Boxplots: Graphical representation of five statistics of the variable:
  - The sample minimum, $x_{(1)}$: The minimum observed value of the variable.
  - The sample lower quartile, Q_L: The value that separates the smallest 25% observed values of the variable from the largest 75%.
  - The sample median, M: The value that separates the smallest 50% observed values of the variable from the largest 50%.
  - The sample upper quartile, Q_U: The value that separates the smallest 75% observed values of the variable from the largest 25%.
  - The sample maximum, $x_{(n)}$: The minimum observed value of the variable.
- Usefulness: See the location, spread, skewness, tail length and outliers.

Visualizing multidimensional data sets

- Summary of boxplot construction:
  1. Draw a box with borders at Q_L and Q_U (i.e., 50% of the data are in this box).
  2. Draw the sample median as a solid line.
  3. Draw *whiskers* from each end of the box to the most remote point that is not an outlier.
  4. Show outliers with special characters.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Boxplots: Several variables in the NCI60 data set.
  - Boxplot: Variable capitalAve in the spam data set.
  - Boxplot: Logarithm of the variable capitalAve in the spam data set.
  - Boxplot: Logarithm of the variable capitalAve in the spam data set in terms of the variable spam.

Visualizing multidimensional data sets

- Histograms: Are density estimates, i.e., histograms provides with an estimate of the data distribution.
- In contrast to boxplots: Histograms show possible multimodality of the data.
- Idea: Represent locally the data density by counting the number of observations in a sequence of consecutive bins.
- Then: the total area of histogram bars is normalised to unity.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Histograms: Several variables in the NCI60 data set.
  - Histogram: Variable capitalAve in the spam data set.
  - Histogram: Logarithm of the variable capitalAve in the spam data set.
  - Histogram: Logarithm of the variable capitalAve in the spam data set in terms of the variable spam.

Visualizing multidimensional data sets

- Kernal densities: Smooth the histogram replacing a box with a smooth function.
- Smooth function: Centred directly over each observation.
- General form of a kernel density:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where:

  - $K(\cdot)$ is a kernel function.
  - h is called the bandwidth.

Visualizing multidimensional data sets

- Commonly used kernels:
  - Uniform kernel: $K(u) = \frac{1}{2} \mathbf{I}\{|u| \leq 1\}$.
  - Triangle kernel: $K(u) = (1 - |u|) \mathbf{I}\{|u| \leq 1\}$.
  - Epanechnikov kernel: $K(u) = \frac{3}{4}\left(1 - u^2\right) \mathbf{I}\{|u| \leq 1\}$.
  - Quartic (Biweight) kernel: $K(u) = \frac{15}{16}\left(1 - u^2\right)^2 \mathbf{I}\{|u| \leq 1\}$.
  - Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$
- Different kernels: Generate different shapes of the estimated density, although the difference between them is usually small.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Kernel densities: Several variables in the NCI60 data set.
  - Kernel density: Variable capitalAve in the spam data set.
  - Kernel density: Logarithm of the variable capitalAve in the spam data set.
  - Kernel density: Logarithm of the variable capitalAve in the spam data set in terms of the variable spam.

Visualizing multidimensional data sets

- Plots for several quantitative variables:
  - Scatterplots: Bivariate plots of one variable against another that help us to understand the relationship among the two variables and allow for the detection of groups or clusters of points.
  - Three dimensional scatterplots: Three-variate plots against each other.
  - Scatterplot matrix: Draw all possible two-dimensional scatteplots for the variables allowing for building knowledge about dependencies and structures.
  - Parallel coordinate plots: Useful to detect outliers and/or groups.
- Dimensionality problem: Any of the previous plots have problems when we have many variables to plot.
- Suggestion: Dimension reduction techniques.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Scatterplot: Two first variables in the NCI60 data set.
  - Three dimensional scatterplot: Three first variables in the NCI60 data set.
  - Scatterplot matrix: First ten variables in the NCI60 data set.
  - Scatterplot matrix: First ten variables in the NCI60 data set.

Visualizing multidimensional data sets

- Parallel Coordinates Plots (PCP): Method for representing high-dimensional data.
- Parallel axes: Instead of plotting observations in an orthogonal coordinate system, PCP draws coordinates in parallel axes and connects them with straight lines.
- Variables: Drawn into the horizontal axis,
- Values of the variables: Mapped onto the vertical axis.
- Sensitive to the order of the variables: Certain trends in the data can be shown more clearly in one ordering than in another.

Visualizing multidimensional data sets

- Chapter 1.R script:
  - Parallel Coordinates Plots: Variables in the NCI60 data set.

Standard descriptive measures for multivariate data sets

- Simple graphical devices: Help to understand the structure and dependency of multidimensional data sets.
- However: Many graphical tools are extremely useful in a modelling step but do not give the full picture of the data set.
- Why?: Graphical tools capture only certain dimensions of the data and do not concentrate on those dimensions or parts of the data under analysis that carry the maximum structural information.
- Chapter 2: Will present powerful tools for reducing the dimension of a data set.
- As a starting point: Use simple and basic tools to describe dependency.
- In the following of this chapter: Assume that the data matrix only contains quantitative variables.

Standard descriptive measures for multivariate data sets

- Sample mean: Given the data matrix X, the sample mean of the j-th variable in X is given by:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

- Location: The sample mean provides an estimate of the center of the data.
- Sample mean vector: Summarises the p univariate sample means as follows:

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix} = \frac{1}{n} X' 1_n$$

where:

  - $\overline{x}_j$ denotes the sample mean of the j-th variable in X, for $j = 1, \ldots, p$.
  - $1_n = (1, 1, \ldots, 1)'$ is the $n \times 1$ vector of $1'$s.

Standard descriptive measures for multivariate data sets

- Chapter 1.R script:
  - Sample mean vector: Variables in the NCI60 data set.
  - Sample mean vector: Variables in the spam data set.

Standard descriptive measures for multivariate data sets

- Sample variance: Given the data matrix X, the sample variance of the j-th variable in X is given by:

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

- Spread: The sample variance provides an estimate of the dispersion of the data with respect to its sample mean.

- Sample standard deviation: The square root of the variance, denoted by s_j, thus has the same unit of measurement than the variable x_j.

Standard descriptive measures for multivariate data sets

- Sample covariance of x_j and x_k in X:

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

- Linear dependency: The sample covariance measures the linear dependency between the observations of x_j and x_k.
- Importantly: s_{jk} depends on the measurement units of x_j and x_k.
- Consequently: The sample covariance is a quantity that it is usually very difficult to interpret.

Standard descriptive measures for multivariate data sets

- Solution: Standardize the variables first (i.e., subtract the mean and divide by the standard deviation), then compute the sample covariance between the standardized variables.
- Sample correlation coefficient:

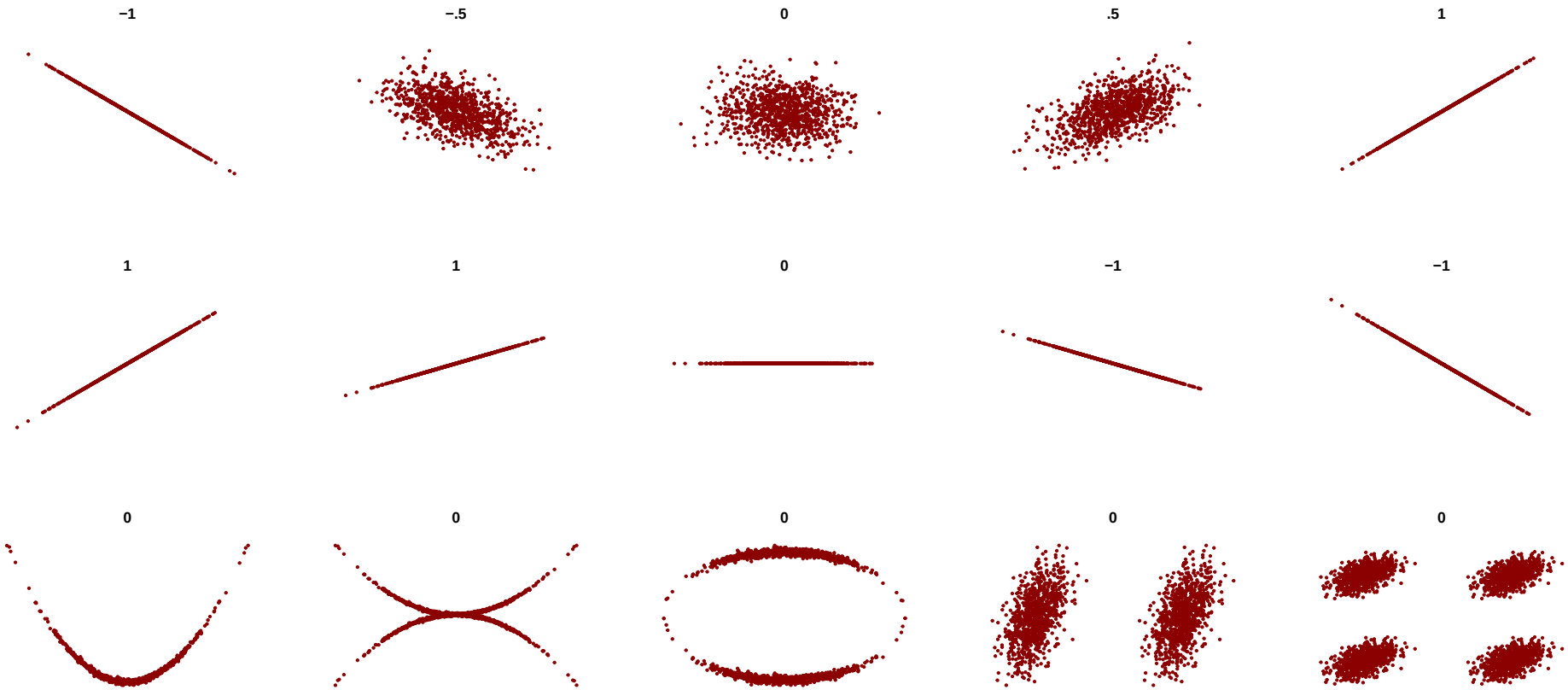
$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

- Linear dependency: The sample correlation coefficient also measures the linear dependence between the observations of the variables x_j and x_k.
- However: r_{jk} does not depend on the units of measurement of x_j and x_k.

Standard descriptive measures for multivariate data sets

- Interpretation: Note that $|r_{jk}| \leq 1$ such that:
  - The closer r_{jk} to 1, the more positive linearly dependent the observations of x_j and x_k.
  - The closer r_{jk} to -1, the more negative linearly dependent the observations of x_j and x_k.
  - The closer r_{jk} to 0, the less linearly dependency between the observations of x_j and x_k.
- In particular: If $r_{jk} = 0$, we say that the observations of x_j and x_k are uncorrelated.
- Important: Understand properly the correlation coefficient.

Standard descriptive measures for multivariate data sets



Standard descriptive measures for multivariate data sets

- Sample covariance matrix of X:

$$S_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i\cdot} - \overline{x})(x_{i\cdot} - \overline{x})' = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \ddots & s_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

  where $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$.

- Summary: S_x contains the variances as well as the covariances of the variables in X.

- Therefore: S_x contains all the information about the spread of the variables and the linear dependency of every pair of variables in X.

Standard descriptive measures for multivariate data sets

- Three important properties of S_x:
  - Symmetry: S_x is a symmetric matrix because $s_{jk} = s_{kj}$.
  - Alternative definition: S_x can be written as follow:

$$S_x = \frac{1}{n-1}\widetilde{X}'\widetilde{X}$$

    where:

    - ⋆ $\widetilde{X} = X - 1_n\overline{x}'$ is the centered data matrix.
  - Non-negative eigenvalues: The eigenvalues of S_x are non-negative.

Standard descriptive measures for multivariate data sets

- Sample correlation matrix of X:

$$R_x = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \ddots & r_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

- Summary: R_x contains the correlation coefficients of the variables in X.
- Therefore: R_x contains all the information about the linear dependency of every pair of variables in X.

Standard descriptive measures for multivariate data sets

- Three important properties of R_x:
  - Symmetry: R_x is a symmetric matrix because $r_{jk} = r_{kj}$.
  - Alternative definition: R_x can be written as follow:

$$R_x = D_x^{-1/2} S_x D_x^{-1/2}$$

    where:

    - ⋆ D_x is the $p \times p$ diagonal matrix containing the elements of the main diagonal of S_x, i.e., the variances $s_1^2, \ldots, s_p^2$.
  - Non-negative eigenvalues: The eigenvalues of R_x are non-negative.

Standard descriptive measures for multivariate data sets

- Chapter 1.R script:
  - Sample covariance matrix: Variables in the spam data set and NCI60 data set.
  - Sample correlation matrix: Variables in the spam data set and NCI60 data set.

Standard descriptive measures for multivariate data sets

- Useful tools: The sample covariance and correlation matrices are extremely useful tools in multidimensional data analysis for a number of purposes.
- Nevertheless: If p is large, $p \simeq n$ or $p > n$, then both the sample covariance and correlation matrices might have certain non-desirable characteristics.
- More information will be given later: We will get back to this problem and will review alternative matrices more adequate to these cases.

Linear transformations

- Transformations: Are very frequent in real data analysis.
- How sample quantities change after a transformation?: Depends on the type of transformation carried out.
- Linear transformations: The most simple transformations that are very frequently used with real data (e.g., principal components, standardization).
- Next: See, how sample quantities change after a linear transformation.

Linear transformations

- Data matrix: X, with size $n \times p$.
- Data transformation: C, with size $p \times r$, where r can be or not be p.
- Linear transformation: New data matrix, Y, with size $n \times r$:

$$Y = XC$$

- Sample mean vector of Y: $\overline{y} = C'\overline{x}$.
- Sample covariance matrix of Y: $S_y = C'S_xC$.
- Sample correlation matrix of Y: $R_y = D_y^{-1/2}S_yD_y^{-1/2}$.

Linear transformations

- Standardization of X:

$$Y = \widetilde{X} D_x^{-1/2}$$

  where:

  - $\widetilde{X}$ is the centered data matrix.
  - D_x is the $p \times p$ diagonal matrix formed by the elements of the principal diagonal of S_x, i.e., the variances $s_1^2, \ldots, s_p^2$.

- Sample mean vector: $\overline{y} = 0_p$.
- Sample covariance matrix: $S_y = D_x^{-1/2} S_x D_x^{-1/2} = R_x$.
- Sample correlation matrix: $R_y = D_y^{-1/2} S_y D_y^{-1/2} = R_x$.
- Therefore: $S_y = R_y = R_x$.

Linear transformations

- Chapter 1.R script:
  - Standardization of: The variables in the spam data set.

Multidimensional distributions and inference

- For future developments: We will need some probabilistic concepts.
- Particularly: We need to introduce the concept of multidimensional distributions.
- Multivariate Gaussian distribution: Canonical example of multidimensional distribution.
- Maximum likelihood estimation: Usual method to estimate parameters of multidimensional distributions.
- Curse of dimensionality: When the dimension of the data set is large, estimation of model parameters becomes problematic.
- Sparse estimation methods: Restrict the number of parameters to estimate, thus avoiding estimation error.

Multidimensional distributions and inference

- Observe: n observations of p single random variables, say $x_1, \ldots, x_p$.
- Multidimensional random variable: The random vector $x = (x_1, \ldots, x_p)'$.
- Types of multidimensional random variables:
  - Continuous: If the variables $x_1, \ldots, x_p$ are continuous.
  - Discrete: If the variables $x_1, \ldots, x_p$ are discrete.
  - Mixed: If there are continuous as well as discrete variables.
- Simplicity: Focus on the continuous case.

Multidimensional distributions and inference

- Cumulative distribution function (CDF) of x at point x^0:

$$F_x\left(x^0\right) = \Pr\left(x \leq x^0\right) = \Pr\left(x_1 \leq x_1^0, \ldots, x_p \leq x_p^0\right)$$

  where:

  - $x = (x_1, \ldots, x_p)'$.
  - $x^0 = \left(x_1^0, \ldots, x_p^0\right)'$.

- Probability density function (PDF) of x at point x^0:

$$F_x\left(x^0\right) = \int_{-\infty}^{x_p^0} \cdots \int_{-\infty}^{x_1^0} f_x\left(x_1, \ldots, x_p\right) dx_1 \cdots dx_p$$

- Property: f_x is a continuous function such that:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_x\left(x_1, \ldots, x_p\right) dx_1 \cdots dx_p = 1$$

Multidimensional distributions and inference

- Expectation or mean vector of x:

$$E[x] = \begin{pmatrix} E[x_1] \\ \vdots \\ E[x_p] \end{pmatrix}$$

where:

  - $E[x_j]$ is the expectation or mean of the j-th random variable, x_j, for $j = 1, \ldots, p$.

Multidimensional distributions and inference

- Covariance matrix of x:

$$Cov\left[x\right] = E\left[\left(x - E\left[x\right]\right)\left(x - E\left[x\right]\right)'\right] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ddots & \sigma_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

- Diagonal elements of $Cov\left[x\right]$: Variances of the components of x, denoted by σ_j^2.
- Off-diagonal elements of $Cov\left[x\right]$: Covariances between pairs of components of x, denoted by σ_{jk}, for $j, k = 1, \ldots, p$ and $j \neq k$.

Multidimensional distributions and inference

- Correlation matrix of x:

$$Cor\left[x\right] = \Delta_x^{-1/2} Cov\left[x\right] \Delta_x^{-1/2} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \ddots & \rho_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

  where:

  - Δ_x is a diagonal matrix with elements the variances of the components of x.

- Off-diagonal elements of $Cor\left[x\right]$: Correlations coefficients between pairs of components of x, denoted by ρ_{jk}, for $j, k = 1, \ldots, p$ and $j \neq k$ and given by:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

Multidimensional distributions and inference

- Multidimensional Gaussian distribution: Generalization to two or more dimensions of the univariate Gaussian (or Normal) distribution.
- Bell curve: The MGD is often characterized by its resemblance to the shape of a bell.
- Importance: The MGD is used extensively in both theoretical and applied statistics.
- Data are rarely Gaussian: Although it is well known that real data rarely obey the dictates of the MGD, this deception does provide us with a useful approximation to reality.

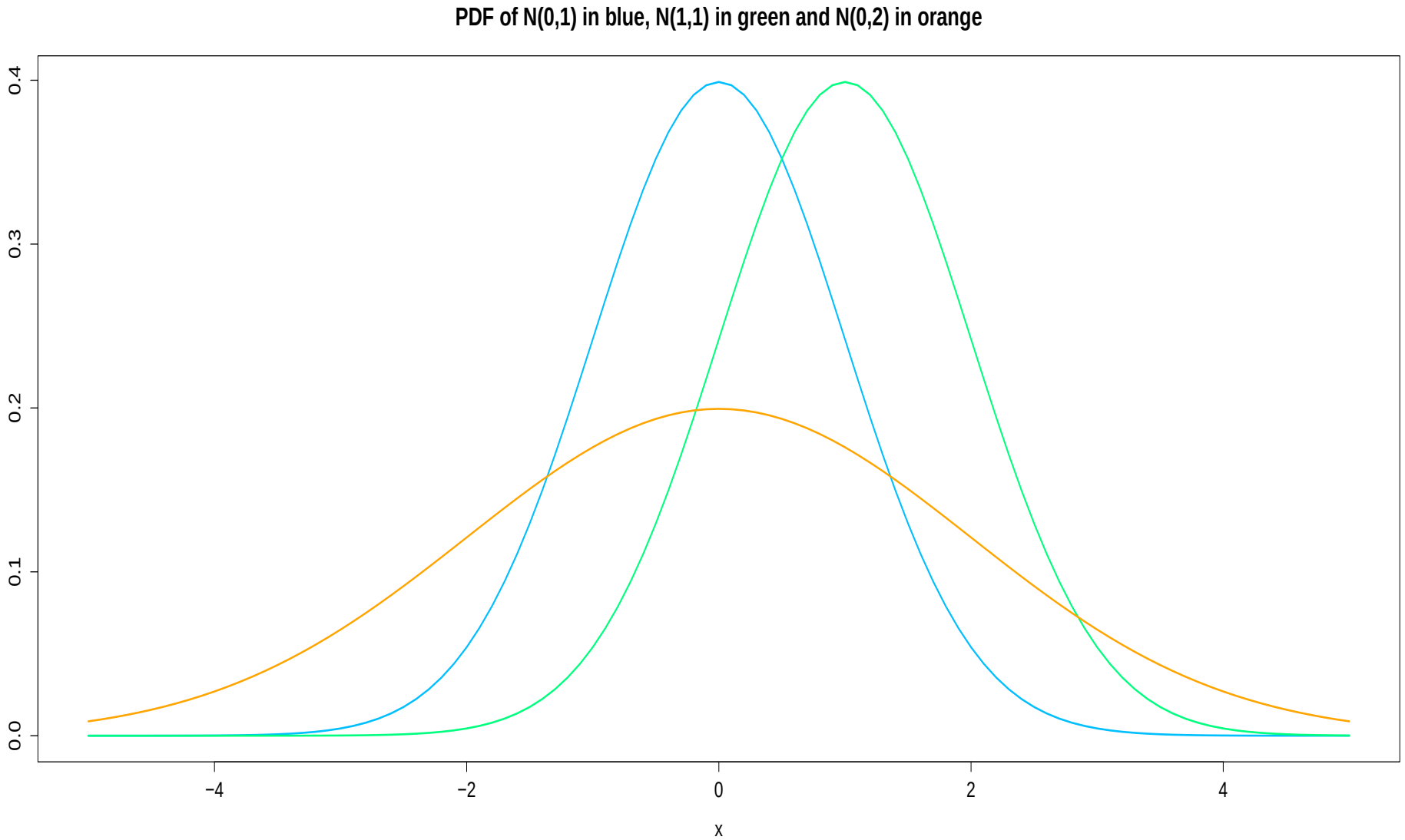
Multidimensional distributions and inference

- Univariate Gaussian distribution: $x \sim N\left(\mu_x, \sigma_x^2\right)$, where $\mu_x = E[x]$ and $\sigma_x^2 = Var[x]$, has PDF:

$$f_x(x) = \left(2\pi\sigma_x^2\right)^{-1/2} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right) \qquad -\infty < x < \infty$$

- Important: Note that μ_x and σ_x^2 completely characterize the density.

Multidimensional distributions and inference



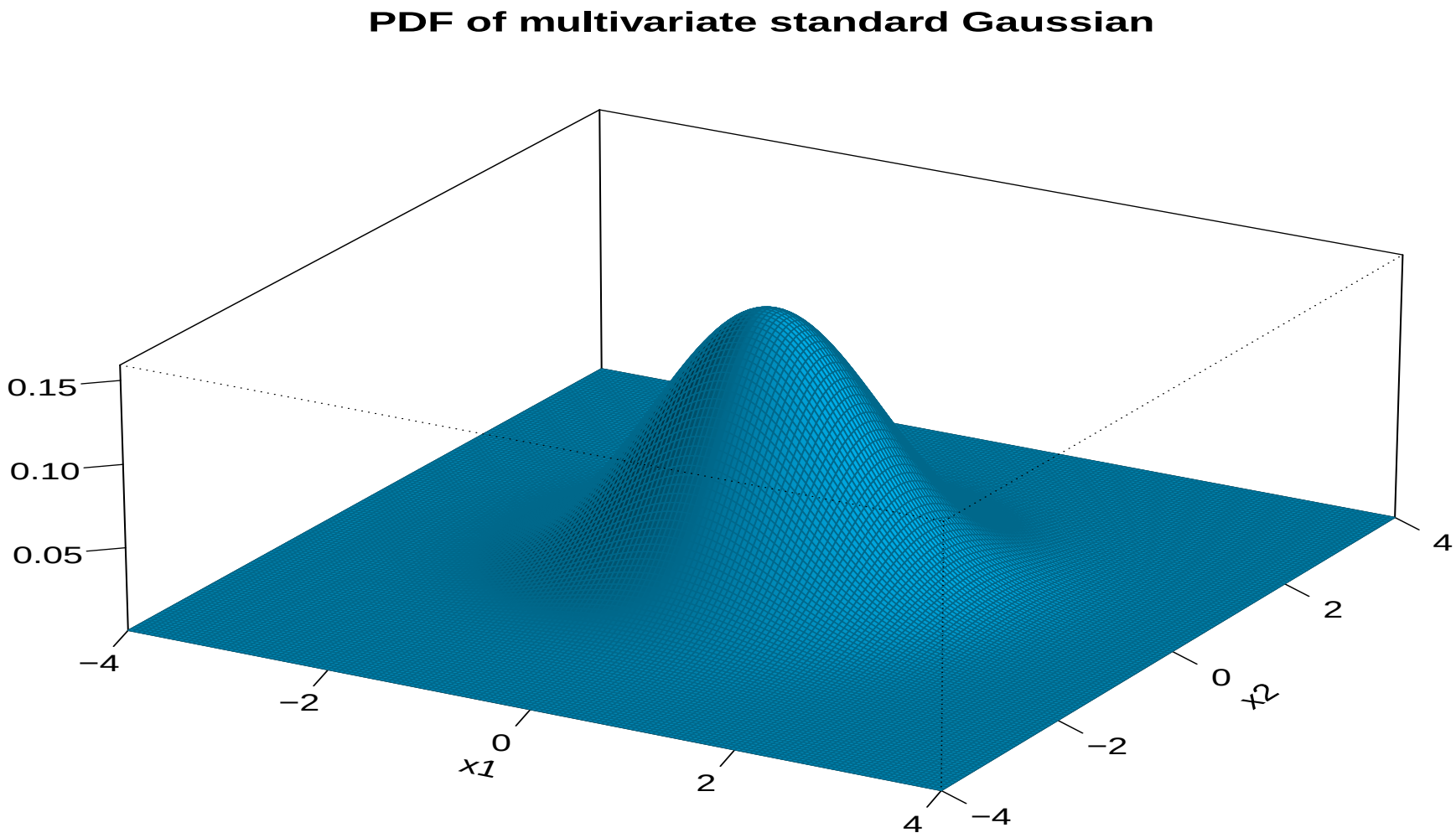
Multidimensional distributions and inference

- Multidimensional Gaussian distribution: $x \sim N_p(\mu_x, \Sigma_x)$, where $\mu_x = E[x]$ and $\Sigma_x = Cov[x]$, has PDF:

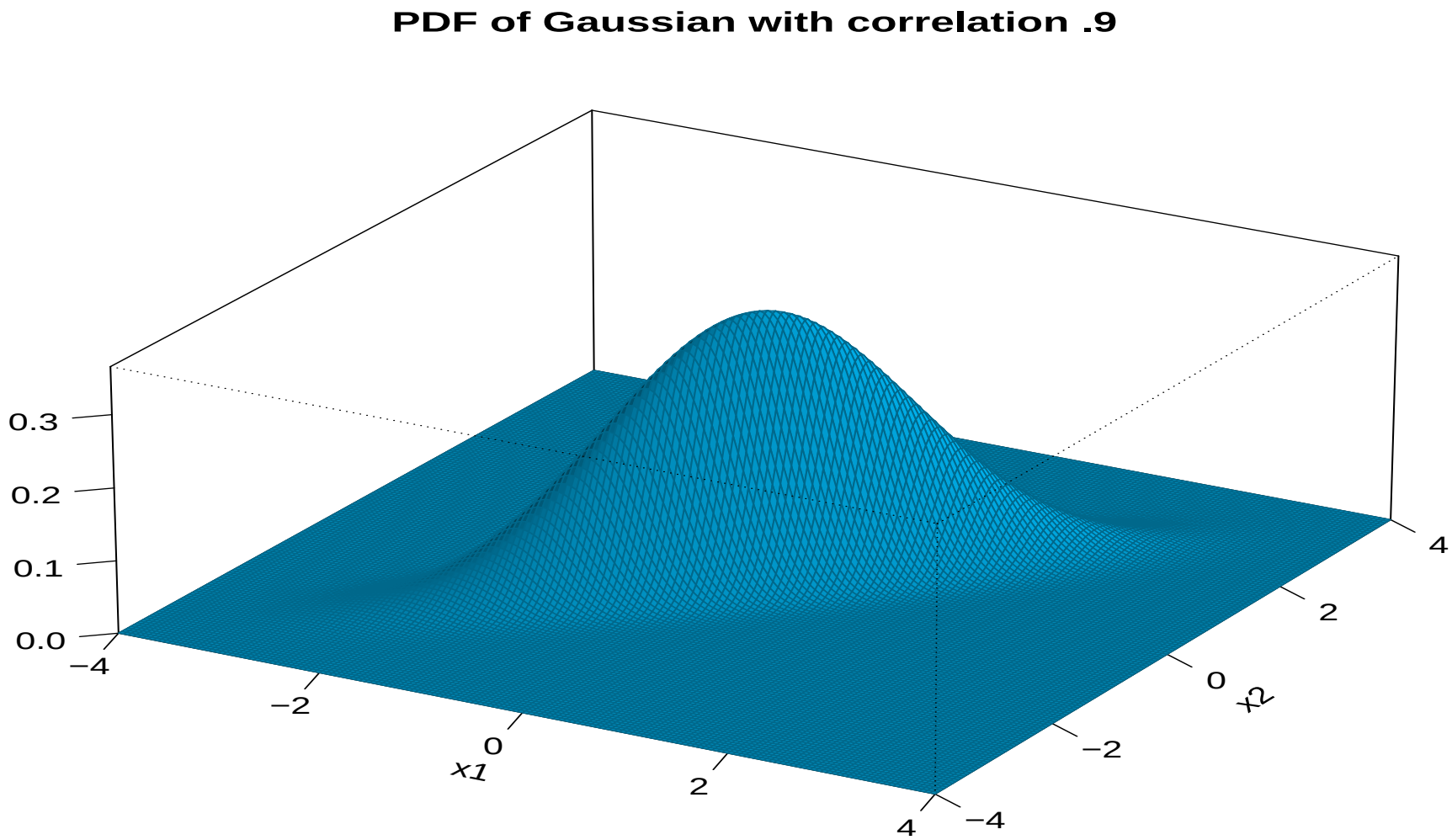
$$f_x(x) = (2\pi)^{-p/2} |\Sigma_x|^{-1/2} \exp\left(-\frac{(x-\mu_x)' \Sigma_x^{-1} (x-\mu_x)}{2}\right) \quad -\infty < x_j < \infty$$

- Examples: The next slides show some examples of PDFs of bivariate Gaussian distributions.

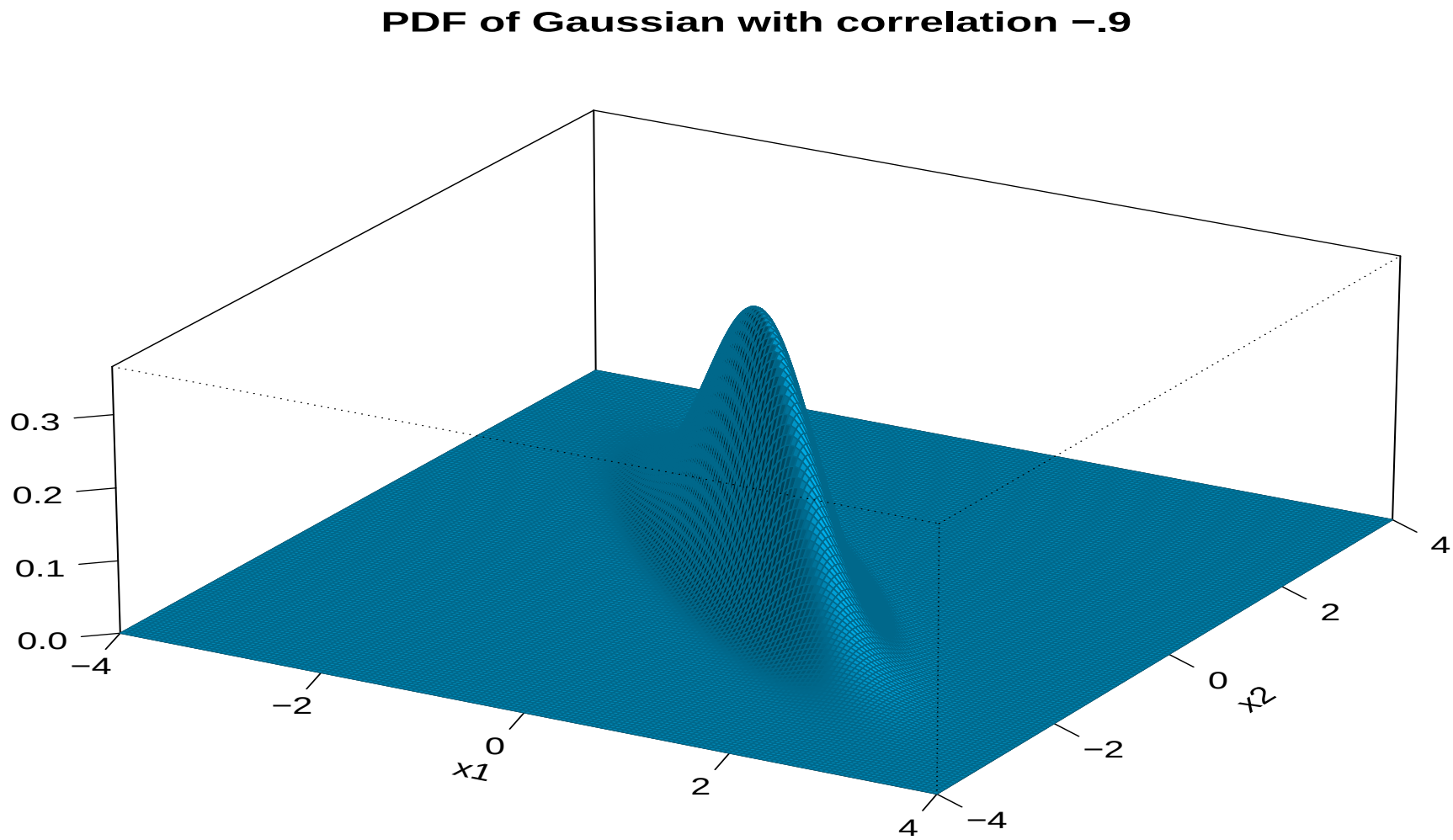
Multidimensional distributions and inference



Multidimensional distributions and inference



Multidimensional distributions and inference



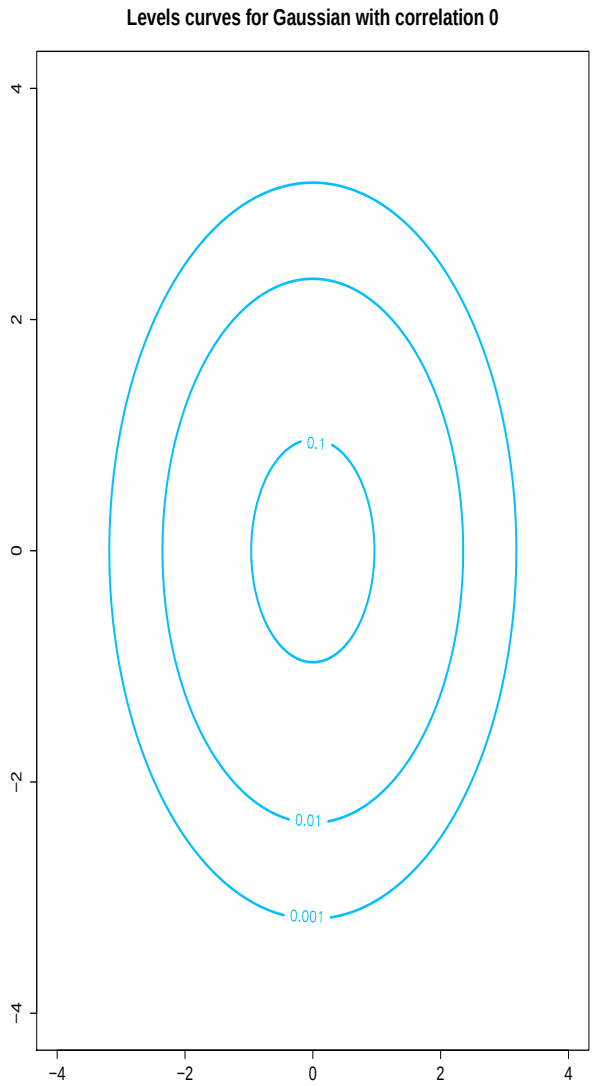
Multidimensional distributions and inference

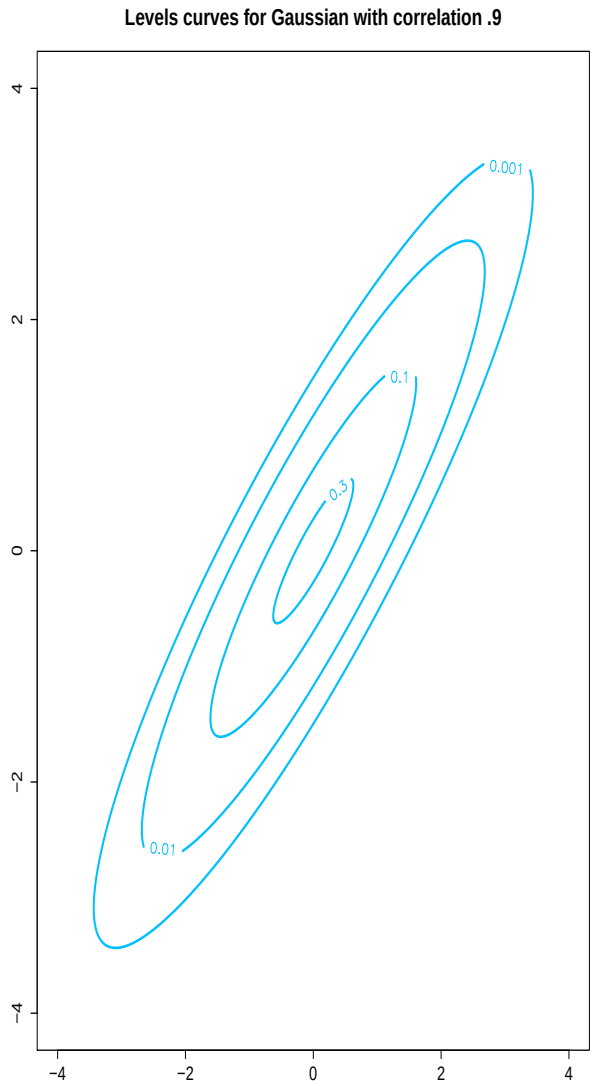
- Contours: Points with the same density value, i.e., $\{x_0 : f_x(x_0) = c\}$, for a certain constant c.
- Level curves: In two dimensions, contours are called level curves and are obtained by cutting the PDF by parallel hyperplanes.
- Multidimensional Gaussian distribution: Contours are given by:

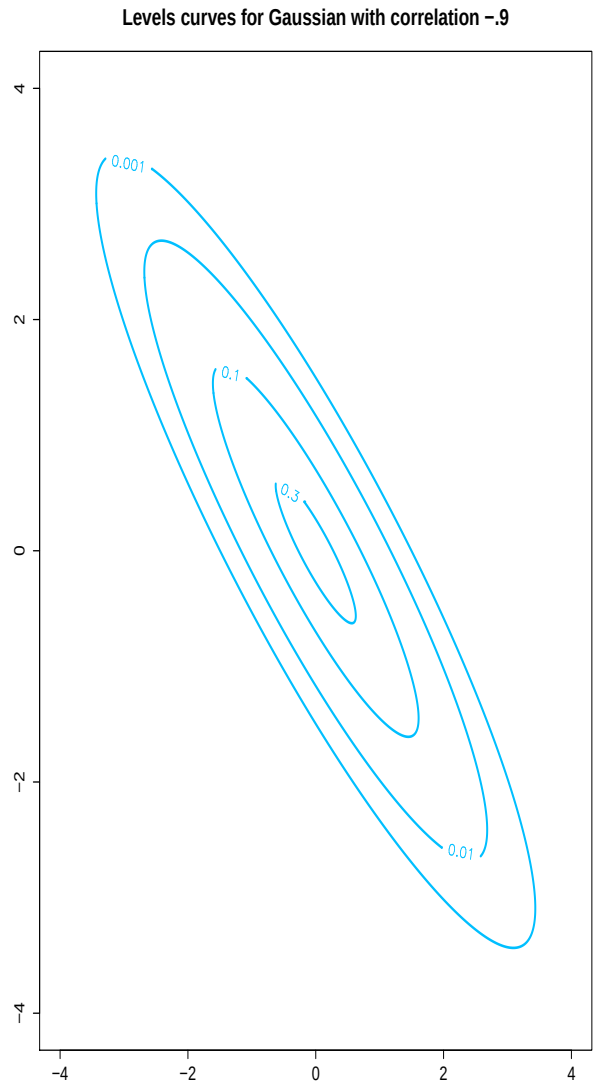
$$(x - \mu_x)' \Sigma_x^{-1} (x - \mu_x) = c^*$$

  for a certain constant c^*.
- Consequence: Contours of multivariate Gaussian distributions are ellipsoids.
- Examples: The next two slides show level curves for GDs with and without a sample of 100 points generated from these distributions.

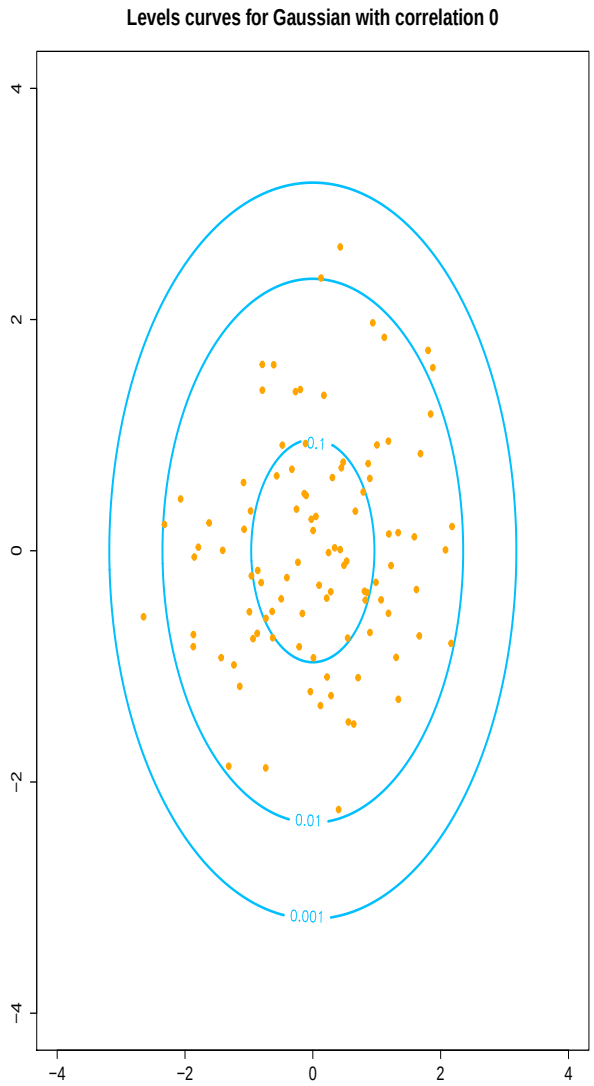
Multidimensional distributions and inference

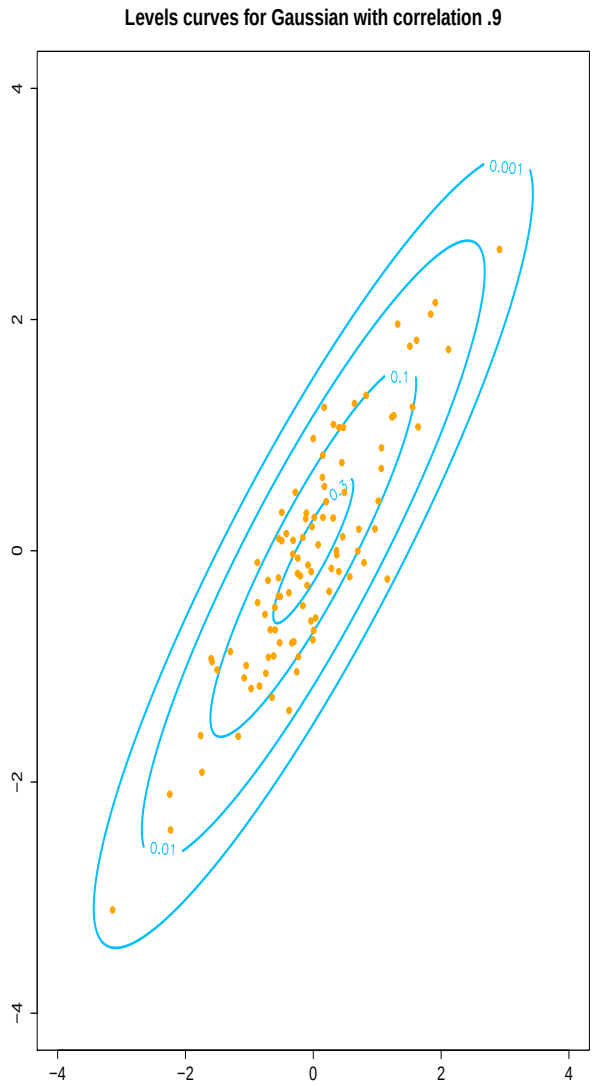


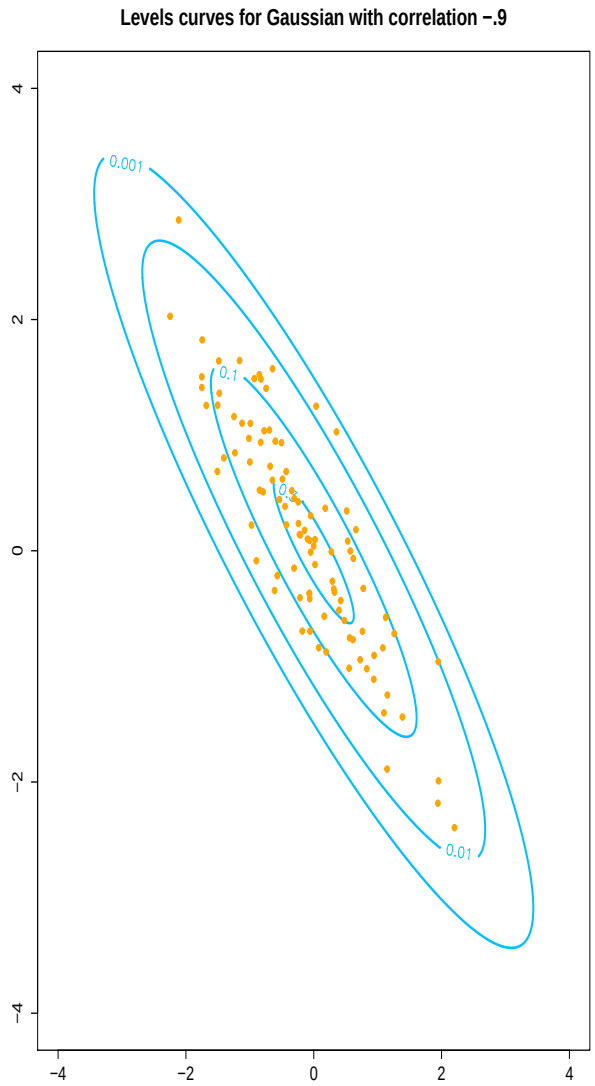




Multidimensional distributions and inference







Multidimensional distributions and inference

- Contours: Points with the same density.
- Idea: Assume that all points belonging to the same contour are at the same distance from the center of the distribution.
- Mahalanobis distance between x and μ_x: Implied by contours of the MGD:

$$D_M(x, \mu_x)^2 = (x - \mu_x)' \Sigma_x^{-1} (x - \mu_x)$$

- Important role: The Mahalanobis distance plays an important role in many problems such as outlier detection, classification, clustering and so on.

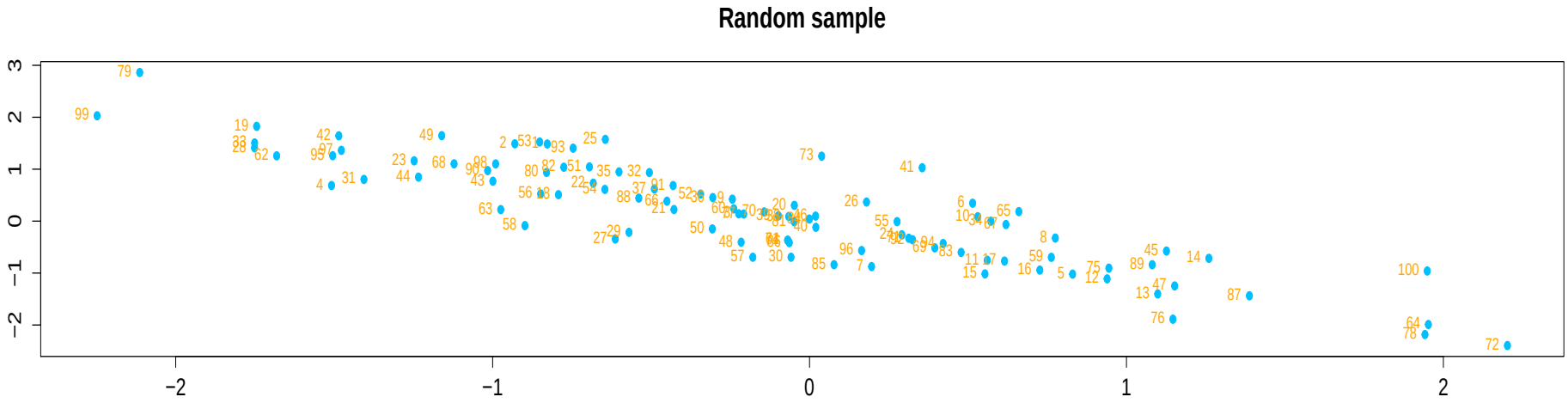
Multidimensional distributions and inference

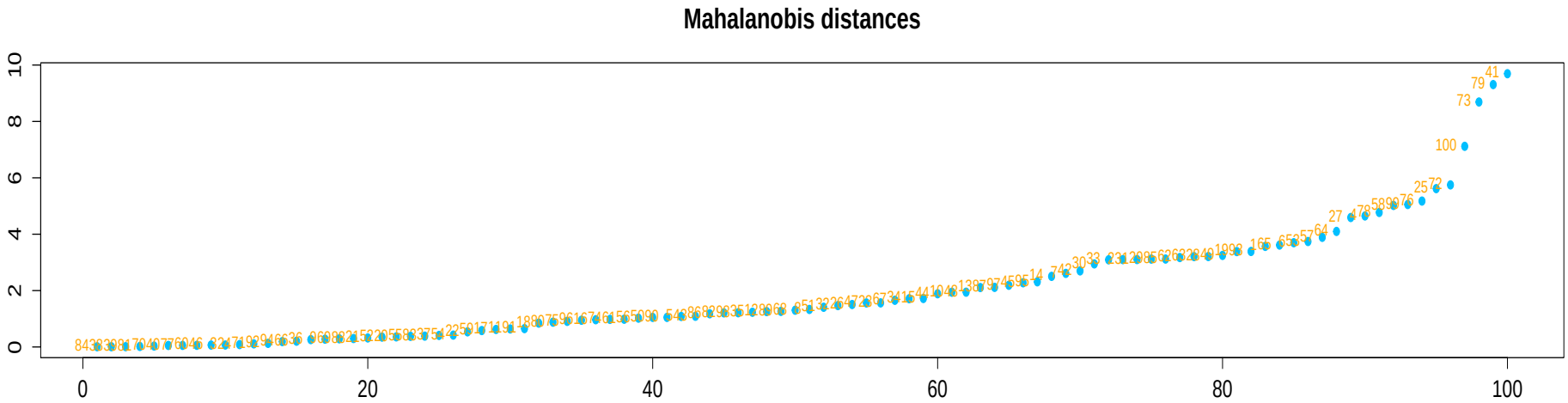
- In practice: The Mahalanobis distance is computed for data which is not necessarily multivariate Gaussian distributed.
- Given a data matrix X of dimension $n \times p$: We can compute the Mahalanobis distance between each observation $x_{i\cdot}$ and the sample mean vector of X, $\overline{x}$, by replacing Σ_x with S_x:

$$D_M\left(x_{i\cdot}, \overline{x}\right)^2 = \left(x_{i\cdot} - \overline{x}\right)' S_x^{-1} \left(x_{i\cdot} - \overline{x}\right)$$

- Example: Mahalanobis distances between 100 points generated from a bivariate Gaussian distribution and the corresponding sample mean vector.

Multidimensional distributions and inference





Multidimensional distributions and inference

- Multidimensional outliers: The Mahalanobis distance has been routinely used to detect outliers.
- Nevertheless: The Mahalanobis distance has two main drawbacks for detecting outliers:
  1. It is mainly appropriate for approximately symmetric data sets.
  2. The sample mean vector and the sample covariance matrices are largely influenced by the outliers.
- Robust estimates: Instead of the sample mean vector and the sample covariance matrices, it is advisable to use robust estimates not influenced by outliers.
- Large dimensions: It is advisable to use first a dimension reduction of the data set, as we will see in Chapter 2.

Multidimensional distributions and inference

- Chapter 1.R script:
  - Mahalanobis distance: The variables in the College data set (more information in the R script).

Multidimensional distributions and inference

- The marginal distribution of x_j: Each univariate random variable in x is a continuous random variable with its own CDF and PDF, denoted by F_{x_j} and f_{x_j}, respectively.
- Relationship between f_x and the f_{x_j}:

$$f_{x_j}(x_j) = \int_{-\infty}^{\infty} \underset{\neq j}{\cdots} \int_{-\infty}^{\infty} f_x(x_1, \ldots, x_p)\, dx_1 \underset{\neq j}{\cdots} dx_p$$

- Extension: This argument can be extended to any subset of the elements of x, say $\left(x_{i_1}, \ldots, x_{i_j}\right)'$.

Multidimensional distributions and inference

- Two multivariate random variables:
  - $x = (x_1, \ldots, x_p)'$ with PDF f_x.
  - $y = (y_1, \ldots, y_q)'$ with PDF f_y.
- Conditional density function of y given $x = x^0$:

$$f_{y|x=x^0}\left(y|x = x^0\right) = \frac{f_{x,y}\left(x^0, y\right)}{f_x\left(x^0\right)}$$

- Interpretation: The distribution of the random variable y use to change if we have information provided by another related random variable x.

Multidimensional distributions and inference

- Independency: $x = (x_1, \ldots, x_p)'$ and $y = (y_1, \ldots, y_q)'$ are independent if:

$$f_{y|x=x^0}\left(y|x = x^0\right) = f_y\left(y\right)$$

and,

$$f_{x|y=y^0}\left(x|y = y^0\right) = f_x\left(x\right)$$

- Interpretation: Knowing $x = x^0$ does not change the probability assessments on y and knowing $y = y^0$ does not change the probability assessments on x.

- Consequence: x and y are independent if, and only if:

$$f_{x,y}\left(x, y\right) = f_x\left(x\right) f_y\left(y\right)$$

Multidimensional distributions and inference

- Missing values: Sometimes we have data sets with missing values for certain variables.
- Two alternatives: Single or multiple imputation methods.
- Simple and faster approach: Single imputation.
- Two different ways: Unconditional and conditional approaches.

Multidimensional distributions and inference

- Unconditional approach:
  - Consider the marginal distributions of the variables with missing values, thus ignore the information provided by the other variables.
  - Replace missing values with the sample mean of the observed values.
- Conditional approach:
  - Consider the conditional distributions of the variables with missing values given the information given by the other variables.
  - Replace missing values with predicted values obtained from regression models.
- Qualitative variables: It is possible to replace missing values in qualitative values with the sample mode, i.e., the most repeated value, but this is not the best idea.
- Alternatively: Use conditional approaches, such as logistic regression.

Multidimensional distributions and inference

- Chapter 1.R script:
  - Missing data: The variables in the birth2006 data set (more information in the R script).

Multidimensional distributions and inference

- Two multivariate random variables:
  - $x = (x_1, \ldots, x_p)'$ with PDF f_x, mean vector $E[x]$ and covariance matrix $Cov[x]$.
  - $y = (y_1, \ldots, y_q)'$ with PDF f_y, mean vector $E[y]$ and covariance matrix $Cov[y]$.
- Covariance matrix between x and y: The $p \times q$ matrix given by:

$$Cov[x, y] = E\left[(x - E[x])(y - E[y])'\right]$$

- Correlation matrix between x and y: The $p \times q$ matrix given by:

$$Cor[x, y] = \Delta_x^{-1/2} Cov[x, y] \Delta_y^{-1/2}$$

where:

- Δ_x is the diagonal matrix with elements the diagonal elements of $Cov[x]$.
- Δ_y is the diagonal matrix with elements the diagonal elements of $Cov[y]$.

Multidimensional distributions and inference

- Linear transformation: Let $x = (x_1, \ldots, x_p)'$ with mean vector $E[x]$ and covariance matrix $Cov[x]$ and let $y = (y_1, \ldots, y_q)'$ such that:

$$y = Ax + b$$

where A is a $q \times p$ matrix and b is a $q \times 1$ column vector.

- Therefore:
  - $E[y] = AE[x] + b$.
  - $Cov[y] = ACov[x]A'$.

Multidimensional distributions and inference

- Many other multidimensional distributions:
  1. Elliptical distributions: Their level curves are ellipsoids.
  2. Heavy-tailed distributions: Have higher probability density in its tail area compared with a Gaussian distribution with the same mean vector and covariance matrix.
  3. Copula distributions: Based on determining the marginals and then couple them through a certain multivariate function called the copula function.
  4. Mixture distributions: Weighted linear combinations of several distributions.

Multidimensional distributions and inference

- Data matrix: The data matrix, X, contains a sample $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$ of a multidimensional random variable $x = (x_1, \ldots, x_p)'$.
- Statistical inference: Given X, we want to analyse the properties of the population random variable x.
- Assume we know the distribution of x: The main goal is to estimate the parameters of this distribution.
- PDF of x: $f_x(\cdot|\theta)$, where $\theta = (\theta_1, \ldots, \theta_r)'$ is the vector of parameters.
- How to estimate θ from X: The most popular method to carry out this task is the maximum likelihood estimation (MLE) method.

Multidimensional distributions and inference

- Joint PDF of the sample $x_{1\cdot}, \ldots, x_{n\cdot}$:

$$f_{(x_{1\cdot},\ldots,x_{n\cdot})}\left(x_{1\cdot}, \ldots, x_{n\cdot}|\theta\right) = \prod_{i=1}^{n} f_x\left(x_{i\cdot}|\theta\right)$$

- Important: The sample is known (X, the data matrix) but θ is unknown.
- Then: MLE considers that that θ is a variable and X is fixed, leading to the likelihood function:

$$L\left(\theta|X\right) = \prod_{i=1}^{n} f_x\left(x_{i\cdot}|\theta\right)$$

- Likelihood function: Can be seen as the PDF of $\theta|X$.

Multidimensional distributions and inference

- The maximum likelihood estimate (MLE) of θ: Denoted by $\widehat{\theta}$, is the value of θ that maximizes $L(\theta|X)$, i.e.:

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|X)$$

- In other words: The MLE, $\widehat{\theta}$, is the value of θ that maximizes the probability of obtaining the sample under study.
- It is easier to maximize the log-likelihood or support function:

$$\ell(\theta|X) = \log L(\theta|X)$$

- Hence:

$$\widehat{\theta} = \arg\max_{\theta} \ell(\theta|X) = \arg\max_{\theta} L(\theta|X)$$

Multidimensional distributions and inference

- In almost all the cases: Maximizing $L(\theta|X)$ or $\ell(\theta|X)$ involves the use of nonlinear optimization techniques (see the course Optimization for large-scale data).
- The multivariate Gaussian distribution: $x \sim N(\mu_x, \Sigma_x)$.
- The MLE can be derived analitically.
- The support function (up to a constant):

$$\ell(\mu_x, \Sigma_x|X) = -\frac{n}{2}\log|\Sigma_x| - \frac{n}{2}\left(Tr\left(\Sigma_x^{-1}\widetilde{S}_x\right) + (\mu_x - \overline{x})'\Sigma_x^{-1}(\mu_x - \overline{x})\right)$$

where:

$$\widetilde{S}_x = \frac{1}{n}\sum_{i=1}^{n}(x_{i\cdot} - \overline{x})(x_{i\cdot} - \overline{x})' = \frac{n-1}{n}S_x$$

Multidimensional distributions and inference

- The support function (up to a constant):

$$\ell\left(\mu_x, \Sigma_x | X\right) = -\frac{n}{2}\log|\Sigma_x| - \frac{n}{2}\left(Tr\left(\Sigma_x^{-1}\widetilde{S}_x\right) + \left(\mu_x - \overline{x}\right)'\Sigma_x^{-1}\left(\mu_x - \overline{x}\right)\right)$$

- Thus: $\ell\left(\mu_x, \Sigma_x | X\right)$ only depends on μ_x in the last term and that this is maximized if $\left(\overline{x} - \mu_x\right)'\Sigma_x^{-1}\left(\overline{x} - \mu_x\right) = 0$.

- Consequence: The MLE of μ_x is $\widehat{\mu}_x = \overline{x}$.

Multidimensional distributions and inference

- MLE of Σ_x:

$$\ell\left(\Sigma_x | X, \widehat{\mu}_x = \overline{x}\right) = -\frac{n}{2}\log|\Sigma_x| - \frac{n}{2} Tr\left(\Sigma_x^{-1}\widetilde{S}_x\right)$$

- Much more complicated: After some algebra it is possible to show that the MLE of Σ_x is:

$$\widehat{\Sigma}_x = \widetilde{S}_x = \frac{n-1}{n} S_x$$

- Consequence: The MLE of Σ_x is not S_x, but a re-scaled version of it.
- Unbiased estimators: $E\left[\overline{x}\right] = \mu_x$ and $E\left[S_x\right] = \Sigma_x$.
- Thus: $E\left[\widetilde{S}_x\right] = \frac{n-1}{n}\Sigma_x$.

Problems with correlations

- Correlations: Two main problems arise when using correlations in big data sets.
  - Correlation does not imply causality: Causation and correlation are very different things but the differences are usually completely ignored.
  - The spurious correlation problem: When the dimension p scales with the sample size n, neither S_x nor R_x are reliable estimators of $Cov[x]$ and $Cor[x]$, respectively.

Problems with correlations

- Correlation does not imply causality:
  - Causation: A causes B.
  - Correlation: A and B are usually observed simultaneously.
- http://www.information-age.com/causation-and-correlation-big-data-headache-123460611/
- Solution: This problems affects mainly to predictive problems (see course "Predictive modeling").
- Also: Try to use the common sense.

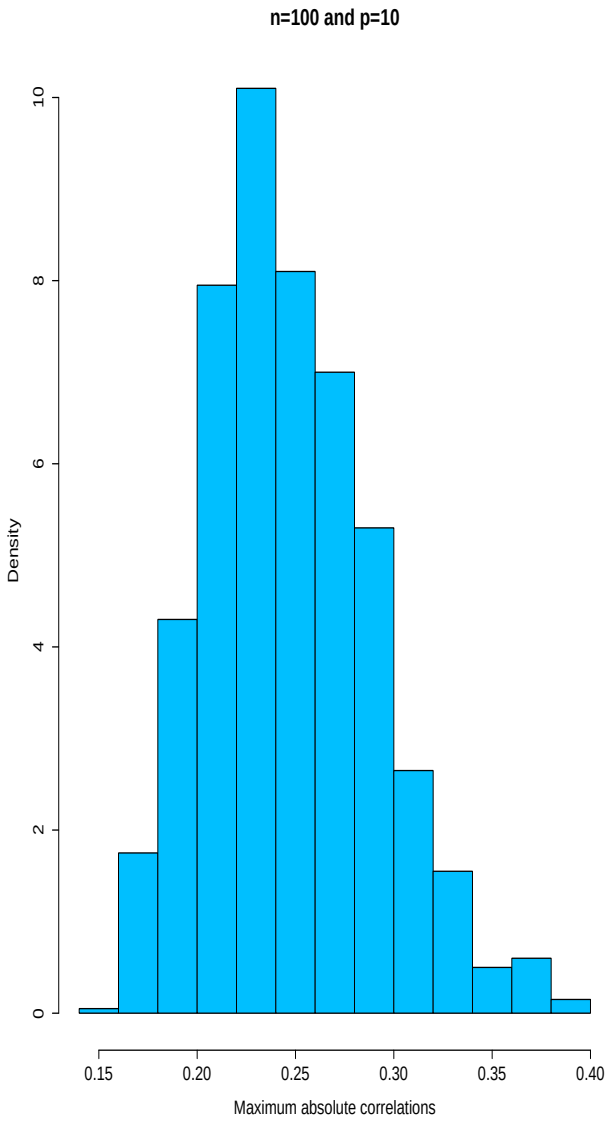
Problems with correlations

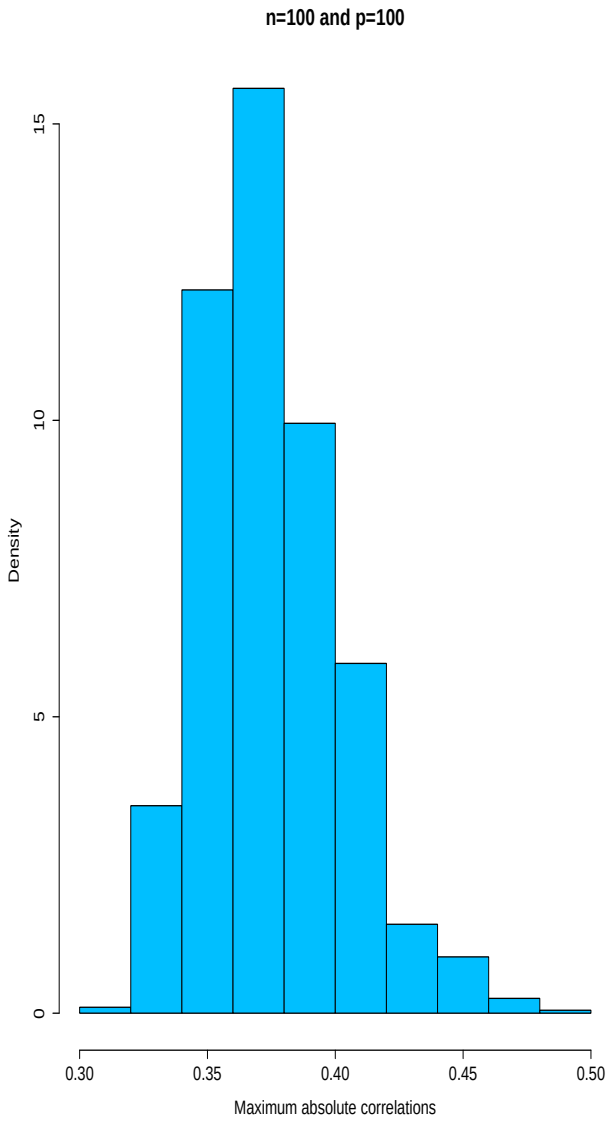
- Spurious correlation: One of the most important causes of false scientific discoveries and wrong statistical inferences.
- Examples: http://www.tylervigen.com/spurious-correlations
- Big data sets: Bring spurious correlation because many uncorrelated random variables may have high sample correlations in high dimensions.
- Simulation: Consider a simple example to illustrate this phenomenon.

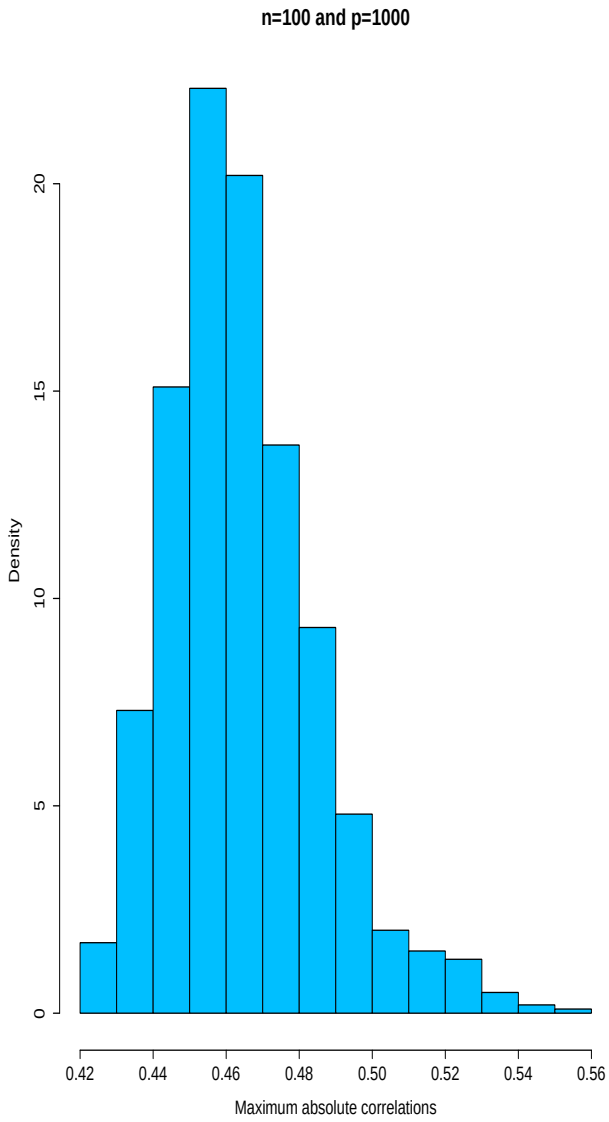
Problems with correlations

- Generation of: 1000 data sets with n observations from a $N(0_p, I_p)$, for the three pairs $(n, p) = (100, 10)$, $(n, p) = (100, 100)$ and $(n, p) = (100, 1000)$.
- For each data set: Obtain the sample correlation matrix and get the maximum absolute correlation.
- Figure in the next slide: Shows the histograms of the 1000 maximum absolute correlations obtained in the three situations.
- Consequence: The larger the dimension, the larger the maximum absolute correlations.
- Thus: Uncorrelated random variables may have high sample correlations in high dimensions.

Problems with correlations







Problems with correlations

- Sparse methods: Are becoming very popular for handling multidimensional data sets.
- Sparse statistical model: One having only a small number of parameters.
- Basic idea under sparse modeling is that of simplicity: A sparse model can be much easier to estimate and interpret than a dense model.
- Many work on sparse modeling: Mainly in regression models but recent proposals include sparse covariance matrix estimation, sparse methods for principal component analysis and sparse supervised and unsupervised classification, among others.
- Here: Focus on sparse covariance matrix estimation.

Problems with correlations

- Covariance matrix of x: $Cov[x]$ contains $\frac{p(p+1)}{2}$ parameters (variances and covariances).
- Thus: The number of parameters to estimate grows with the square of the dimension p.
- Leading to: Inefficient estimation.
- Idea: Impose sparsity in $Cov[x]$ by assuming that several covariances are just 0.
- Consequently: The number of parameters to estimate can be reduced substantially which decreases the estimation error.
- Problem: How to identify which are the covariances that can be assumed to be 0?

Problems with correlations

- Next: Present one of the most popular approaches to perform sparse covariance matrix estimation.
- Remember: Under the Gaussian likelihood, the support function (up to a constant) once μ_x has been replaced with its MLE, $\widehat{\mu}_x = \overline{x}$:

$$\ell\left(\Sigma_x | X, \widehat{\mu}_x = \overline{x}\right) = -\frac{n}{2}\log|\Sigma_x| - \frac{n}{2} Tr\left(\Sigma_x^{-1}\widetilde{S}_x\right)$$

- The MLE of Σ_x: $\widehat{\Sigma}_x = \widetilde{S}_x$, obtained by maximizing $\ell\left(\Sigma_x | X, \widehat{\mu}_x = \overline{x}\right)$ with respect to Σ_x.

Problems with correlations

- Sparse estimator of Σ_x: Obtained after maximizing:

$$-\frac{n}{2}\log|\Sigma_x| - \frac{n}{2}Tr\left(\Sigma_x^{-1}\widetilde{S}_x\right) - \lambda\|P * \Sigma_x\|_1$$

where:

1. λ is a certain penalization parameter.
2. P is a $p \times p$ matrix given by:

$$P = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \cdots & 0 \end{pmatrix}$$

3. $\|P * \Sigma_x\|_1$ is the L_1 norm of a matrix such that $\|A\|_1 = \sum_{jk}|A_{jk}|$.
4. $*$ denotes elementwise multiplication.

Problems with correlations

- Therefore: The idea after maximizing the previous expression is to penalize the value of the covariances.
- Resolution: To solve this problem is necessary to rely in an optimization algorithm (generalized gradient descent).
- Key point: Select an appropriate value of the parameter λ.
- Best choice: Consider several values of λ and select the most stable solution.

Multidimensional distributions and inference

- Chapter 1.R script:
  - Sparse covariance matrix estimation: Spam data set.

1. Introduction
2. Multidimensional data sets
3. Data quality problems
4. Visualizing multidimensional data sets
5. Standard descriptive measures for multivariate data sets
6. Linear transformations
7. Multidimensional distributions and inference
8. Problems with correlations