Statistical Learning

Week 2 - Multidimensional Data (II)

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

Academic Year 2017/2018

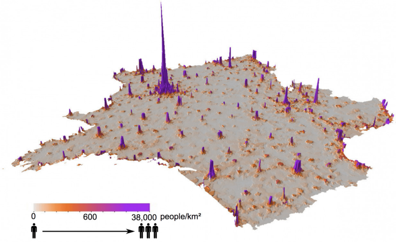
Master in Big Data Analytics

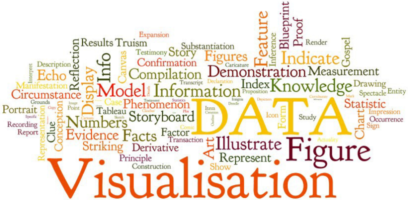
uc3m | Universidad Carlos III de Madrid

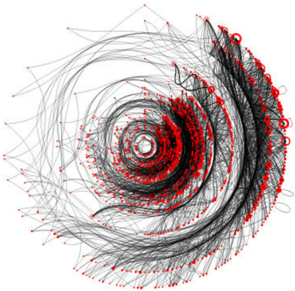
Visualizing multidimensional data sets

- Graphical displays: Tools that help to understand better the contents of the data sets and to find important features.
- Important: Graphics strongly depends on the data structure.
- Examples: See the next slide.
- See course: Network analysis and data visualization.
- Here: We focus on informative plots for multidimensional data sets in a broad sense.

Visualizing multidimensional data sets







Visualizing multidimensional data sets

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

Visualizing multidimensional data sets

- Week 2.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 a single quantitative variable:
  - 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 described below.
  - Boxplots: Simple univariate devices 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

- Week 2.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 (data below $Q_L - 1.5 \times (Q_U - Q_L)$ and data over $Q_U + 1.5 \times (Q_U - Q_L)$).
  4. Show outliers with special characters.

Visualizing multidimensional data sets

- Week 2.R script:
  - Boxplots: Second gene 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: Estimates of the density function of the random variable.
  - Idea: Represent locally the density of the variable by counting the number of observations in a sequence of consecutive bins.
  - Then: The total area of histogram bars is normalised to unity (again, they are density estimates).

Visualizing multidimensional data sets

- Week 2.R script:
  - Histograms: Second gene 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.
  - Idea: Replace the box in the histogram with a smooth function.
  - 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 and h is called the bandwidth.
  - Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ leads to:

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right)$$

Visualizing multidimensional data sets

- Week 2.R script:
  - Kernel density: The second gene 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 and outliers.
  - 3-D scatterplots: Three-variate plots against each other.
  - Scatterplot matrix: Draw all possible two-dimensional scatterplots 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

- Week 2.R script:
  - Scatterplot: Income and Rating in the Credit data set.
  - Three dimensional scatterplot: Income, Limit and Rating in the Credit data set.
  - Scatterplot matrix: Quantitative variables in the Credit data set.

Visualizing multidimensional data sets

- Parallel Coordinates Plots (PCP):
  - Idea: 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

- Week 2.R script:
  - Parallel Coordinates Plots: Quantitative variables in the Credit 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.
- Topic 2: Will present powerful tools for reducing the dimension of a data set for doing this.
- As a starting point: Use simple and basic tools to describe dependency.
- In the following of this topic: Assume that the data matrix only contains quantitative variables or binary variables (maybe).

Standard descriptive measures for multivariate data sets

- Goal: Measure the center of the observations of the variable x_j.
- Sample mean of x_j in X:

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

- Goal: Measure the center of the observations of the data matrix X.
- Sample mean vector of X:

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

where $1_n = (1, 1, \ldots, 1)'$ is the $n \times 1$ vector of $1'$s.

Standard descriptive measures for multivariate data sets

- Week 2.R script:
  - Sample mean vector: Balance and income in the Default data set.

Standard descriptive measures for multivariate data sets

- Goal: Measure the dispersion of the observations of x_j with respect to $\overline{x}_j$.
- Sample variance of x_j in X:

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

- Sample standard deviation of x_j in X: Square root of s_j^2, denoted by s_j.
- Thus: s_j has the same unit of measurement than the variable x_j.

Standard descriptive measures for multivariate data sets

- Goal: Measure the linear dependency between the observations of x_j and x_k in X.
- Sample covariance of x_j and x_k in X:

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

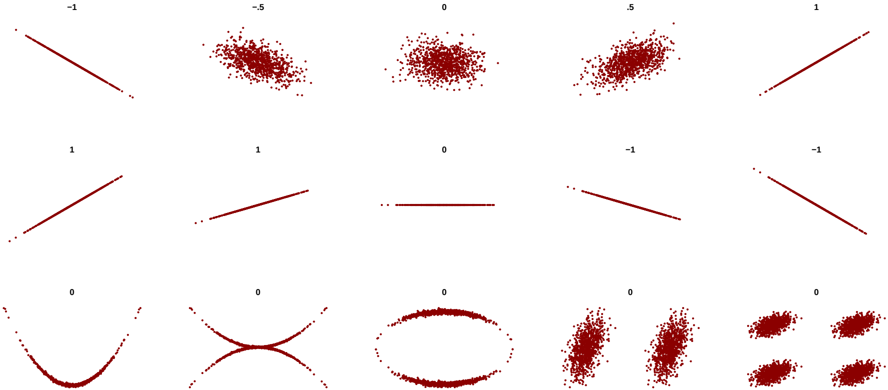
- Importantly: s_{jk} depends on the measurement units of x_j and x_k, so it is usually very difficult to interpret.
- Solution: Standardize the variables first (i.e., subtract the sample mean and divide by the sample standard deviation).
- Sample correlation coefficient of x_j and x_k in X:

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

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.
- Question: What is the sample correlation coefficient between a quantitative variable and a binary variable?

Standard descriptive measures for multivariate data sets



Standard descriptive measures for multivariate data sets

- Week 2.R script:
  - Correlation coefficient: Income and student in the Default data set.

Standard descriptive measures for multivariate data sets

- Centered data matrix: $\widetilde{X} = X - 1_n \overline{x}'$.
- Sample covariance matrix of X:

$$S_x = \frac{1}{n-1}\widetilde{X}'\widetilde{X} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \ddots & s_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

- Therefore: S_x contains all the information about the dispersion of the variables and the linear dependency of every pair of variables in X.
- Symmetry: S_x is a symmetric matrix because $s_{jk} = s_{kj}$.
- Eigenvalues of S_x: All are non-negative.

Standard descriptive measures for multivariate data sets

- Matrix of sample variances: D_x is a diagonal matrix with elements $s_1^2, \ldots, s_p^2$.
- Individual standardization of X: $Y = \widetilde{X} D_x^{-1/2}$.
- Sample mean vector of Y: $\overline{y} = 0_p$.
- Sample covariance matrix of Y:

$$S_y = D_x^{-1/2} S_x D_x^{-1/2} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \ddots & r_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} = R_x$$

- Sample correlation matrix of X: R_x.
- Symmetry: R_x is a symmetric matrix because $r_{jk} = r_{kj}$.
- Eigenvalues of R_x: All are non-negative.

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 $p \simeq n$ or $p > n$, then both the sample covariance and correlation matrices might have certain non-desirable characteristics.
- Two solutions:
  - Dimension reduction: If there are many variables, try to reduce its number.
  - Alternative matrices: We will review alternative matrices more adequate to these cases later.

Standard descriptive measures for multivariate data sets

- Week 2.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.

Multidimensional distributions and inference

- For future developments: We will need some probabilistic concepts.
- Particularly: We need 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

- Main assumption: We 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 $x = \left(x_1, \ldots, x_p\right)'$ and $x^0 = \left(x_1^0, \ldots, x_p^0\right)'$.

- 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 and non-negative function such that:

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

- 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.

Multidimensional distributions and inference

- Expectation or mean vector of x:

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

where $E[x_1], \ldots, E[x_p]$ are the expectations or means of the univariate random variables $x_1, \ldots, x_p$.

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 is Gaussian distributed, the MGD provide us with a useful approximation to reality.

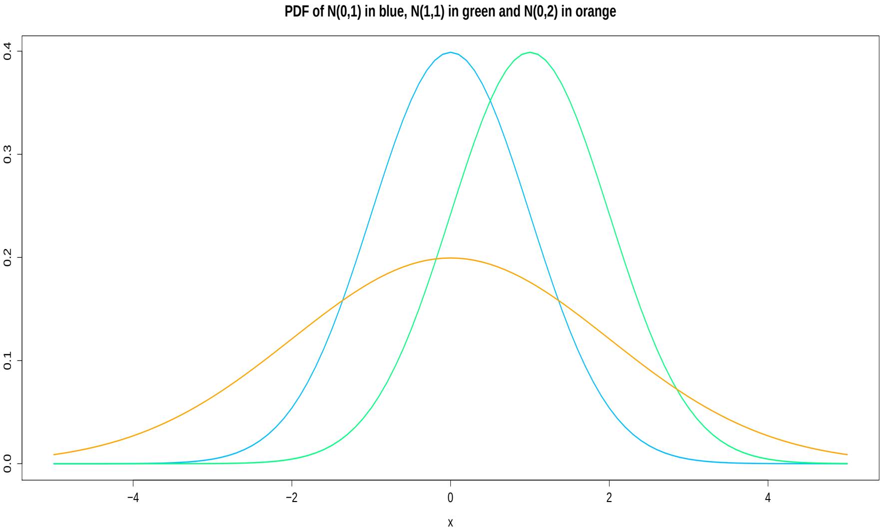
Multidimensional distributions and inference

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

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

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

Multidimensional distributions and inference



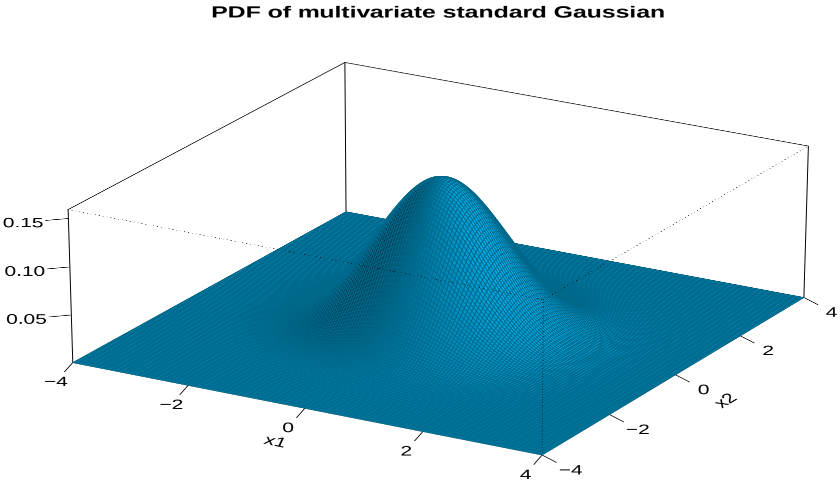
Multidimensional distributions and inference

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

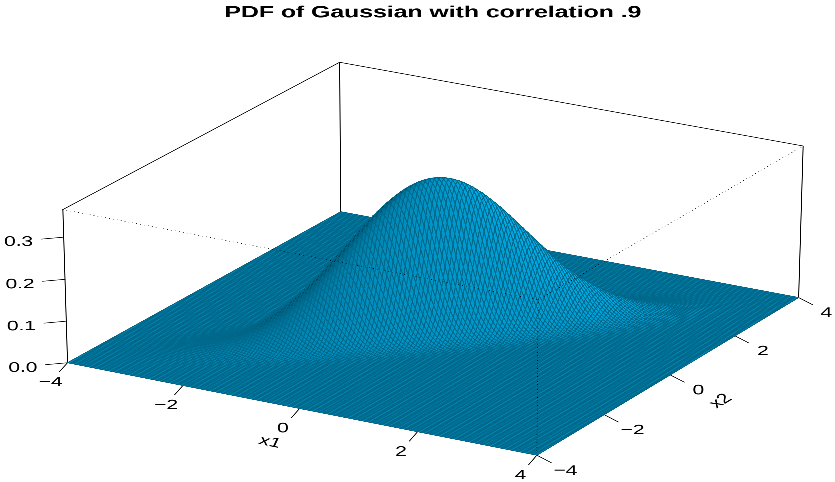
$$f_x(x) = (2\pi)^{-p/2} |\Sigma_x|^{-1/2} \exp\left(-\frac{(x-\mu_x)' \Sigma_x^{-1} (x-\mu_x)}{2}\right) \quad -\infty < x_j < \infty$$

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

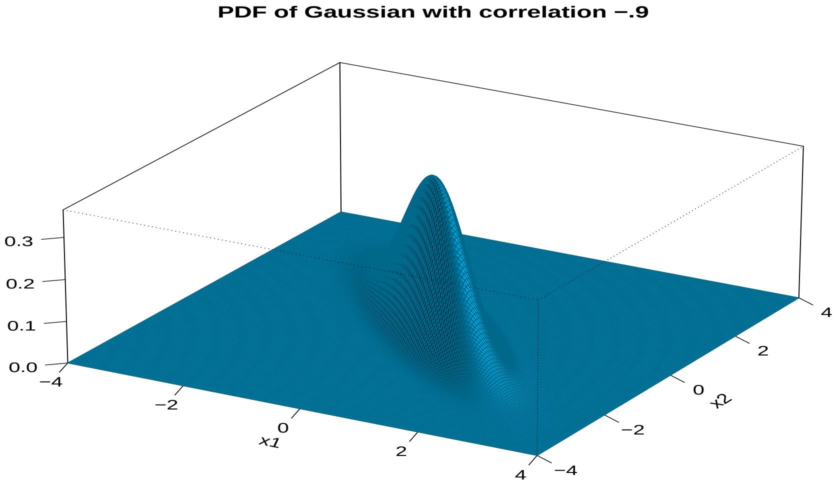
Multidimensional distributions and inference



Multidimensional distributions and inference



Multidimensional distributions and inference



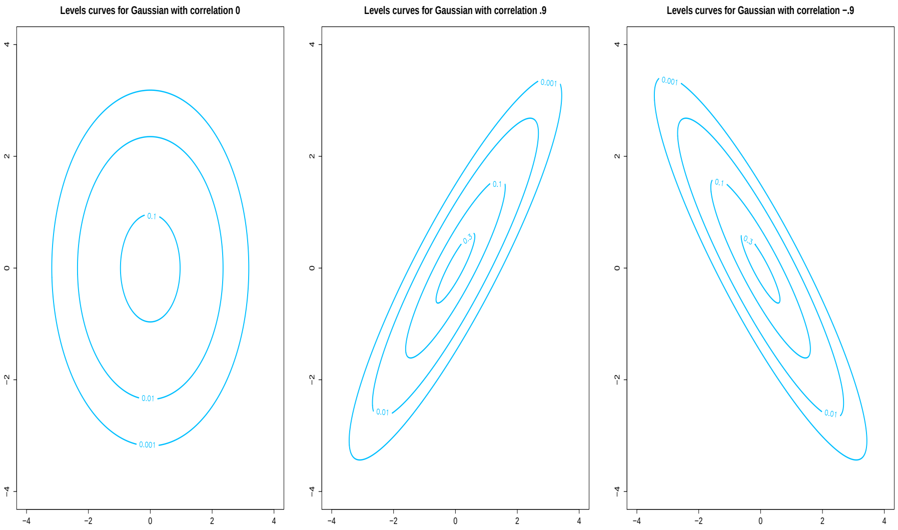
Multidimensional distributions and inference

- 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:

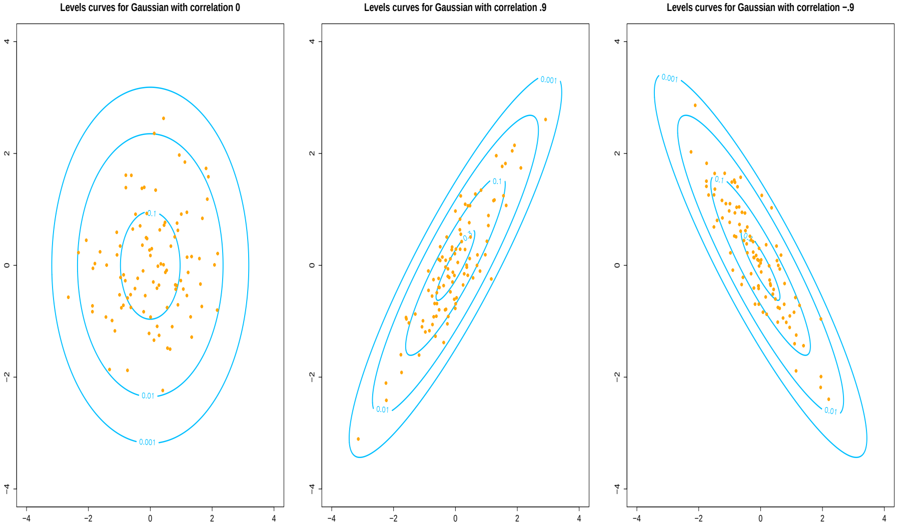
$$(x - \mu_x)' \Sigma_x^{-1} (x - \mu_x) = c^*$$

  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.
- Squared 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.

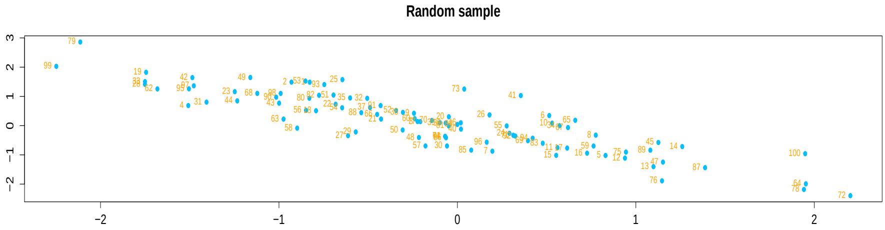
Multidimensional distributions and inference

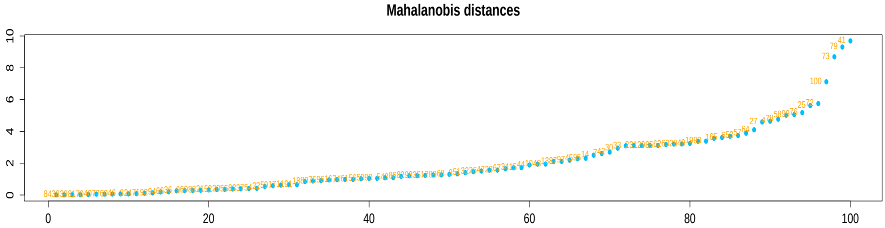
- In practice: The Mahalanobis distance is computed for data which is not necessarily multivariate Gaussian distributed.
- Given a data matrix X of dimension $n \times p$: We can compute the Mahalanobis distance between each observation $x_{i\cdot}$ and the sample mean vector of X, $\overline{x}$, by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

- Solutions:

  1. Try to transform the variables highly non-Gaussian (for instance, a very skewed variable) to something more symmetric (use logarithms).
  2. Replace the sample mean vector and the sample covariance matrices with robust estimates not influenced by outliers.

Multidimensional distributions and inference

- What to do with outliers?:
  - If an outlier is a gross errors due to data handling or something similar, the observation should be deleted from the data matrix for posterior analyses.
  - If an outlier is a good observation but different than others in the data set, the observation should be kept in the analysis but then it is necessary to consider methods robust to its presence.

Multidimensional distributions and inference

- Week 2.R script:
  - Outlier detection: The variables in the College 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.
  - $y = (y_1, \ldots, y_q)'$ with PDF f_y.
- Joint probability density function: $f_{x,y}(x, y)$.
- 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

- One application: Imputation of missing values.
- Missing values: Some data is missing for some reason.
- Two different ways to impute missing values:
  - Unconditional approach.
  - Conditional approach.

Multidimensional distributions and inference

- Unconditional approach for quantiative variables:
  - 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 or the sample median of the observed values of the variables.
- Qualitative variables: It is possible to replace missing values in qualitative values with the sample mode.
- However: This is not the best idea.
- Alternatively: Use conditional approaches, such as logistic regression (see Topic 4).

Multidimensional distributions and inference

- 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:
    - Use the complete observations to estimate a linear regression of x_j on the rest of variables ($x_j = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$), producing an estimate $\widehat{\beta}$ with covariance matrix $\Sigma_{\widehat{\beta}}$.
    - Draw a random sample of $N\left(\widehat{\beta}, \Sigma_{\widehat{\beta}}\right)$, denoted by $\widehat{\beta}^*$.
    - Use $\widehat{\beta}^*$ to predict all (including observed) the values of the variable x_j.
    - For each missing value of x_j, select the observations whose predicted values are close to the predicted value for the missing value.
    - Impute the missing value of x_j with one of those observations randomly chosen.
    - Repeat steps 2 through 5 with all the missing values.

Multidimensional distributions and inference

- Week 2.R script:
  - Missing data: The variables in the birth2006 data set (more information in the R script).

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 (useful for supervised and unsupervised classification problems, see Topics 3 and 4).

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 $x = (x_1, \ldots, x_p)'$.
- PDF of x: $f_x(\cdot|\theta)$, where $\theta = (\theta_1, \ldots, \theta_r)'$ is the vector of parameters.
- Goal: Estimate θ based on X.
- How to do this?: The most popular method to carry out this task is the maximum likelihood estimation (MLE) method.

Multidimensional distributions and inference

- Key point: The sample is known (X, the data matrix) but θ is unknown.
- Idea behind MLE: Estimates θ with the value of the parameters that maximizes the PDF of $\theta|X$.
- Thus: θ is treated as a variable.
- In other words: The MLE, denoted by $\widehat{\theta}$, is the value of θ that maximizes the probability of obtaining X.

Multidimensional distributions and inference

- The PDF of $\theta|X$ is called the Likelihood function:

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

- The MLE of θ, $\widehat{\theta}$:

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

- The log-likelihood or support function:

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

  is easier to maximize.

- Note that:

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

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: The MLE can be derived analytically.
- Assume: $x \sim N(\mu_x, \Sigma_x)$.
- Data matrix: X.

Multidimensional distributions and inference

- The support function (up to a constant):

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

where:

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

- Note: $\ell\left(\mu_x, \Sigma_x | X\right)$ only depends on μ_x in the last term.
- Moreover: In terms of μ_x, $\ell\left(\mu_x, \Sigma_x | X\right)$ 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 the sample mean vector $\widehat{\mu}_x = \overline{x}$.

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: $E\left[\widetilde{S}_x\right] = \frac{n-1}{n} \Sigma_x$.

Multidimensional distributions and inference

- Uses of MLE in this course:
  - Unsupervised classification: Model-based clustering in Topic 3.
  - Supervised classification: Bayes classifiers in Topic 4.

Important facts about correlations in big data sets

- A popular mantra in big data analytics: Look for highly correlated variables if you want to study common effects or trends.
- For instance: https://www.google.com/trends/correlate/
- Correlations: Two important facts about sample correlations in big data sets.
  - Spurious correlations: Two variables can be highly correlated due to just simply a coincidence or the presence of another variable not taken into account.
  - Correlation does not imply causality: Causation and correlation are different things (this is usually completely ignored):
    - ⋆ Causation: A causes B.
    - ⋆ Correlation: A and B are usually observed simultaneously.

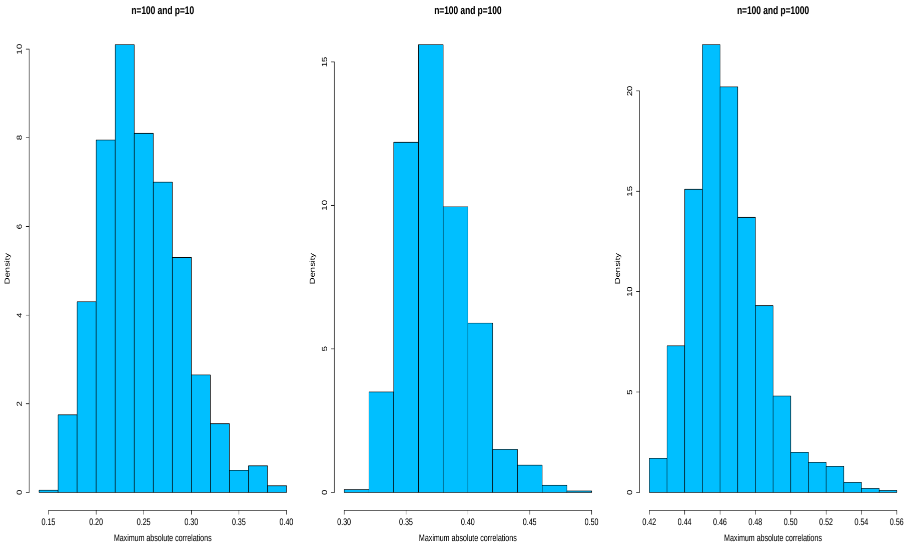
Important facts about correlations in big data sets

- 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: Try to understand the next simulation exercise that illustrates this phenomenon.

Important facts about correlations in big data sets

- 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: True uncorrelated random variables may have high sample correlations in high dimensions.

Important facts about correlations in big data sets



Important facts about correlations in big data sets

- Possible solutions:
  - Dimension reduction: Once again, one option is to reduce the dimension of the data set (see Topic 2).
  - Sparse methods: Reduce the number of correlations to estimate by shrinking to 0 those corresponding to true uncorrelated variables.

Important facts about correlations in big data sets

- Sparse methods: Becoming very popular for handling multidimensional data sets.
- Sparse model: Tries to explains many data with few 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.
- Examples: 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.

Important facts about correlations in big data sets

- $Cov[x]$: 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 the covariances of true uncorrelated variables 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?

Important facts about correlations in big data sets

- 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.

- Note that: $\widehat{\Sigma}_x = \widetilde{S}_x$ is obtained after estimating all the variances and covariances.

Important facts about correlations in big data sets

- Sparse estimator of Σ_x: Obtained after maximizing:

$$\widetilde{\ell}\left(\Sigma_x | X, \widehat{\mu}_x = \overline{x}\right) = -\frac{n}{2} \log |\Sigma_x| - \frac{n}{2} Tr\left(\Sigma_x^{-1} \widetilde{S}_x\right) - \lambda \| P * \Sigma_x \|_1$$

where:

1. λ is a penalization parameter (a positive number).
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. $*$ denotes elementwise multiplication.
4. $\|P * \Sigma_x\|_1 = \sum_{jk} |\sigma_{jk}|$ is the L_1 norm of the matrix $P * \Sigma_x$.

Important facts about correlations in big data sets

- 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.

Important facts about correlations in big data sets

- Week 2.R script:
  - Sparse covariance matrix estimation: Spam data set.