Multivariate Data Analysis

Grégoire Clarté

Winter-Spring 2024

About the course

Welcome

- Grégoire Clarté [ˈgʁegwaʁ̥ ˈkl̥aʁ̥te]
- Lecturer in Statistics
- Office 5608. Office hours friday 2:30-3:00.
- Contact: `gclarte@ed.ac.uk`
- GitHub: `https://github.com/GClarte/MultivariateDataAnalysis`
- All the material will be on Learn

Course structure

- Lectures: Monday 14:00-16:00;
- Workshops: Friday every other week two sessions per day: 10:00:10:50 and 15:10-16:00

Course Structure

- Assessment: 85%, written exam 2h;
- Coursework: 15% two assignments to hand over on *gradescope*, the first one (23/2) is theoretical the second (12/4) practical.

Syllabus

Five parts:

- Intro to multivariate analysis;
- Principal component analysis (PCA);
- Discriminant Analysis (DA)
- Distance and clustering analysis;
- Correspondance analysis (CA).

Textbooks and softwares

- Applied Multivariate Data Analysis, Everitt & Dunn;
- Applied Multivariate Statistical Analysis, Johnson & Wichern;
- An introduction to Applied Multivariate Analysis with R, Everitt & Hothorn;
- Multivariate Statistical Methods, Manly & Navarro Alberto.
- R (RStudio).

Introduction

Mathematical context

Observations in $\mathbb{R}^d$:

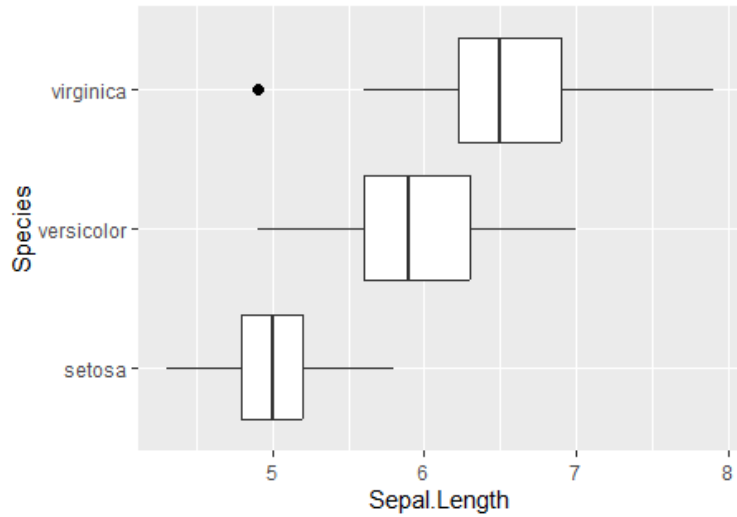
$$X = (X_1, \ldots, X_n) = \overbrace{\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{d1} & \cdots & x_{dn} \end{pmatrix}}^{n \text{ individuals}} \Bigg\} d \text{ variables}$$

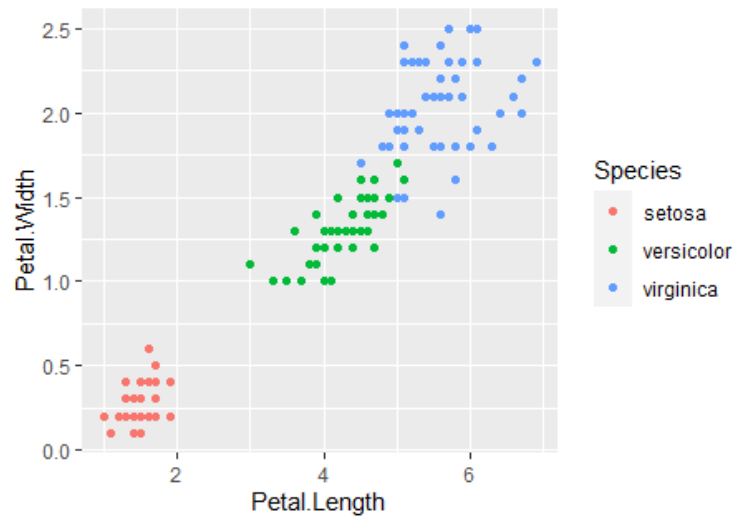
Attention: the notation can change in some textbooks

Example in R

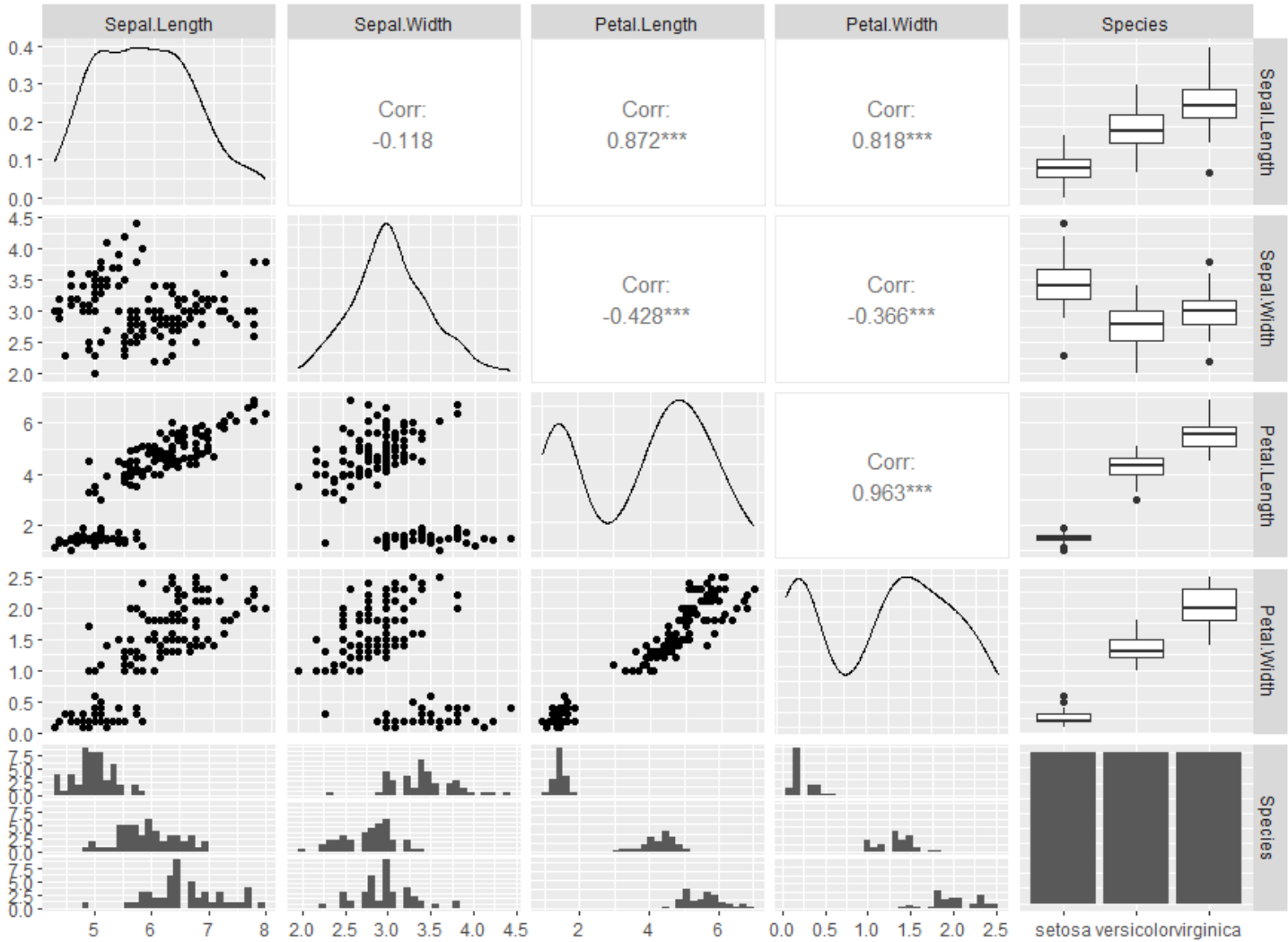
```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
26           5.0         3.0          1.6         0.2     setosa
27           5.0         3.4          1.6         0.4     setosa
28           5.2         3.5          1.5         0.2     setosa
29           5.2         3.4          1.4         0.2     setosa
30           4.7         3.2          1.6         0.2     setosa
31           4.8         3.1          1.6         0.2     setosa
32           5.4         3.4          1.5         0.4     setosa
33           5.2         4.1          1.5         0.1     setosa
34           5.5         4.2          1.4         0.2     setosa
35           4.9         3.1          1.5         0.2     setosa
36           5.0         3.2          1.2         0.2     setosa
37           5.5         3.5          1.3         0.2     setosa
38           4.9         3.6          1.4         0.1     setosa
39           4.4         3.0          1.3         0.2     setosa
40           5.1         3.4          1.5         0.2     setosa
41           5.0         3.5          1.3         0.3     setosa
42           4.5         2.3          1.3         0.3     setosa
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

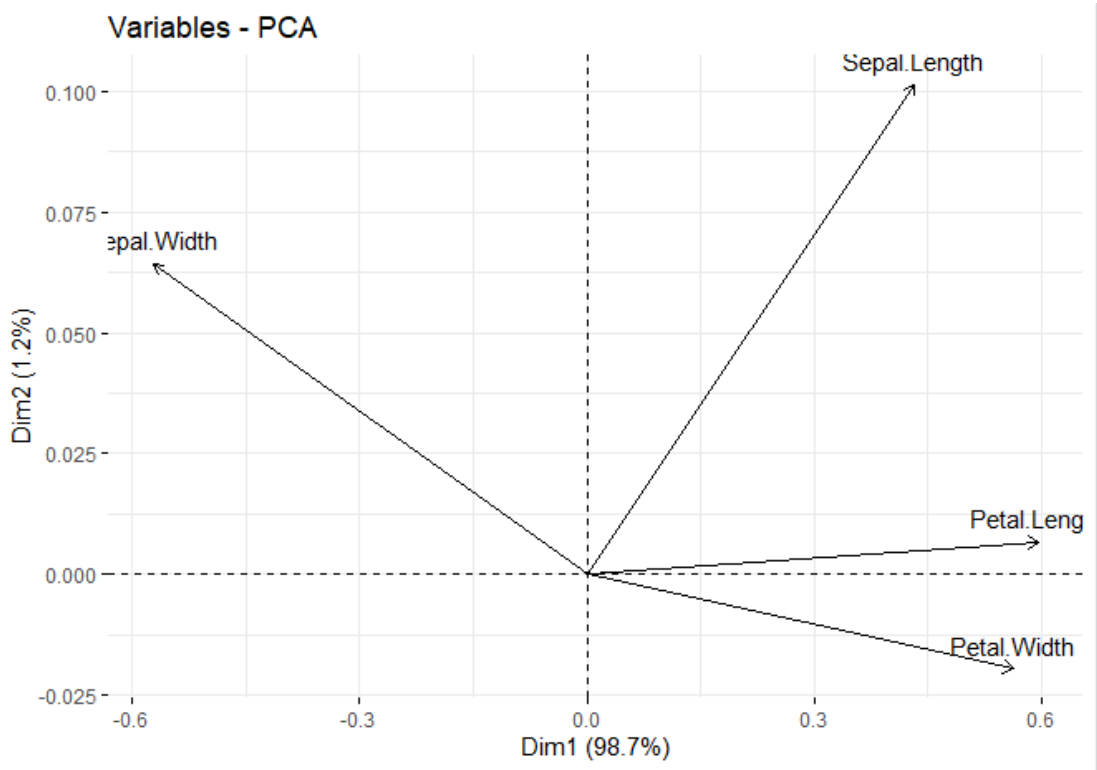




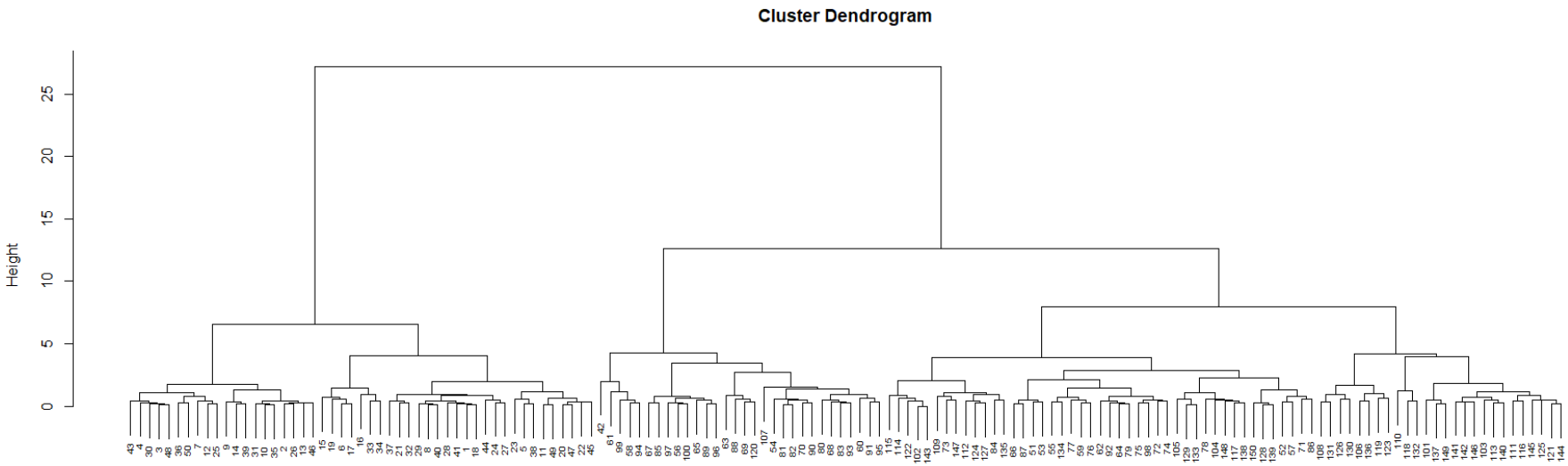
Example in R


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Corr:
-0.118
Corr:
0.872***
Corr:
0.818***
Corr:
-0.428***
Corr:
-0.366***
Corr:
0.963***
setosa versicolorvirginica

Example in R


Variables - PCA
Sepal.Length
epal.Width
Petal.Leng
Petal.Width
0.100
0.075
0.050
0.025
0.000
-0.025
Dim2 (1.2%)
-0.6
-0.3
0.0
0.3
0.6
Dim1 (98.7%)

Examples in R



Which method to choose?

Depends on the goals:

- What is the nature of the variables?
- what are the dependencies?
- what are the variables of interest?

Several techniques are of general use:

- Dimension reduction, variable selection (finance, ecology, etc.);
- correlations (i.e. relationships) between variables;
- classification/prediction.

What is MDA?

Descriptive or inferential ?

- Descriptive: sample proportion, means, standard deviation, correlations, covariances, etc.
- Inferential: equivalent of the univariate models but in larger dimension (MANOVA, (generalised) linear regression, etc.)

Representation of multivariate data is a challenge.
Main issue: correlation of variables $\Rightarrow$ you (the statistician) have to be aware of that.

MDA is the base of Data Mining.

Goals of MDA

- Simplify the data: selection of important variables, makes it more readable;
- sorting and grouping: classification of objects;
- studying the dependencies: independent variables, conditionally independent variables, correlated variables, etc.
- inference/prediction: determine relationships, carry inference on a model, extrapolate from it;
- hypothesis testing: tests in higher dimension.

Revision: Matrices

Notations

- $d \times 1$ vector $X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$;
- transposed vector $X' = X^\top = (x_1, \ldots, x_d)$;
- $p \times q$ matrix: $A = (a_{ij})_{ij} = \begin{pmatrix} a_{11} & \cdots & a_{1q} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pq} \end{pmatrix}$;
- transposed matrix $A^\top = (a_{ji})_{ij}$.

Definitions

- If $p = q$ the matrix is *square*;
- if $\forall i, j,\ a_{ij} = 0$, $A = 0$, the zero (null) matrix;
- If $p = q$, and if $a_{ij} = 0, i \neq j;\ a_{ii} = 1, \forall i$ then $A = I_d$ the identity matrix;
- the set of $p \times q$ matrices with coefficient in $\mathbb{R}$ is written $\mathcal{M}_{p,q}(\mathbb{R})$.

For A a square matrix:

- A is symmetric iif $A^\top = A$;
- A is antisymmetric (skew-symmetric) iif $A^\top = -A$;
- A is diagonal iif $a_{ij} = 0, \forall i \neq j$.

Definitions

- Matrix multiplication: if $A \in \mathcal{M}_{p,q}(\mathbb{R})$, $B \in \mathcal{M}_{q,r}(\mathbb{R})$: $C = AB := (c_{i,j}) \in \mathcal{M}_{p,r}(\mathbb{R})$, where $c_{ij} = \sum_{k=1}^{q} a_{ik} b_{kj}$.

Property: $(AB)^\top = B^\top A^\top$, $(A^\top)^\top = A$.
For a square matrix A:

- A is orthogonal iif $AA^\top = A^\top A = I$;
- A is idempotent of order ℓ iif $A^\ell = \underbrace{A \times \cdots \times A}_{\ell \text{ times}} = A$
- if B is such that $AB = I$, then B is the inverse matrix of A, written A^{-1}.

The trace of the matrix A is $\mathrm{Tr}(A) = \sum_{i=1}^{p} a_{ii}$.
The trace is linear (i.e. $\mathrm{Tr}(A + \lambda B) = \mathrm{Tr}(A) + \lambda \mathrm{Tr}(B)$);
$\mathrm{Tr}(A^\top) = \mathrm{Tr}(A)$, and $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Eigenvalues

Definition 1

λ is an *eigenvalue* of A a square matrix if there exists $x \neq 0$ such that:

$$Ax = \lambda x \Leftrightarrow (A - \lambda I)x = 0.$$

We call x the *eigenvector*, the set of eigenvectors for an eigenvalue is called the *eigenspace*, and the set of eigenvalues is the *spectrum* of the matrix.

All the eigenvalues are roots of the characteristic polynomial in λ, $p(\lambda) = \mathrm{Det}(A - \lambda I)$.

Eigenvalues

If we can write:

$$p(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i)^{n_i},$$

where n_i is the *algebraic multiplicity*, N_i the number of *unique* eigenvalues and $\sum_i n_i = d$.
When solving $(A - \lambda_i)e_i = 0$, there will be $m_i \leq n_i$ linearly independent solutions, m_i is the *geometric* multiplicity.
In general algebraic multiplicity is larger than geometric multiplicity. In general not all matrices have eigenvalues (c.f. rotation matrices).

Example

$A = \begin{pmatrix} -1 & 2 \\ -3 & 4 \end{pmatrix}$, what are the eigen(value/vector/space)?
Characteristic polynomial:

$$p_A(\lambda) = \lambda^2 - 3\lambda + 2 = (\lambda - 1)(\lambda - 2).$$

After solving $\begin{pmatrix} -2 & 2 \\ -3 & 3 \end{pmatrix} x = 0$, and $\begin{pmatrix} -3 & 2 \\ -3 & 2 \end{pmatrix} x = 0$, we get eigenvectors:
$e_1 = (1, 1)$ and $e_2 = (2, 3)$.

$A = \begin{pmatrix} 1 & -1 \\ 2 & 2 \end{pmatrix}$
Which has characteristic polynomial: $p_A(\lambda) = \lambda^2 - 3\lambda + 4$.
This polynomial is irreducible (in $\mathbb{R}$).

Properties

- A and $A^\top$ have the same eigenvalues;
- A is singular (i.e. non-invertible) iif $0 \in Sp(A)$;
- $\mathrm{Tr}(A) = \sum_{\lambda \in Sp(A)} \lambda$;
- $\mathrm{Det}(A) = \prod_{\lambda \in Sp(A)} \lambda$;
- if A is invertible, with eigenvalues λ_i, the eigenvalues of A^{-1} are λ_i^{-1} with the same eigenvectors and multiplicities.

Mutlidimensional RV

Multidimensional RV

Notations:

- Random Vector: $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = (X_1, \ldots, X_p)^\top$.
- $x \in \mathbb{R}^p$ is a realisation of X.

Reminder:
A random vector is a function from $(\Omega, \mathcal{F}, P)$ a probability space to $\mathbb{R}^p$.

Revisions

Definition

A probability mass function (probability density function) with respect to the Lebesgue measure in $\mathbb{R}^d$ is a function:

$$f : \begin{cases} \mathbb{R}^p \to \mathbb{R} \\ x \mapsto f(x) \end{cases}$$

such that, for any $O \subset \mathbb{R}^p$ measurable subset (e.g. Open),

$$\int_O f(x)\mathrm{d}x = P(X \in O).$$

In the discrete case (i.e. with respect to the counting measure on $\mathbb{R}^p$), $f(x) = P(X = x)$.

Note: Proper definition of these objects require measure theory, the use of these objects do not require measure theory in the simple cases.

Properties

Let f be a pdf.

- $\forall x \in \mathbb{R}^p,\ f(x) \leq 0$;
- $\int_{\mathbb{R}^p} f(x)\mathrm{d}x = 1$, and in the discrete case $\sum_x f(x) = 1$.

If we write $X = (Y, Z)$, $Y \in \mathbb{R}^s$, $Z \in \mathbb{R}^t$, $t + s = p$, the marginal pdf of Y and Z are $f_Y(y) = \int_{\mathbb{R}^t} f(y, z)\mathrm{d}z$, and resp. for f_Y.
Once again, note the discrete case:
$f_Y(y) = \sum_z f(y, z)$.

The cumulative density function (cdf) is defined for a *real* random variable X of density f by:
$F_X(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x$.

Conditional distributions

We write $X = (Y, Z)$, $Y \in \mathbb{R}^s$, $Z \in \mathbb{R}^t$, $t + s = p$. The conditional density is *defined* as:

$$f(y \mid z) = \frac{f(y, z)}{f_Z(z)}.$$

In the discrete case:

$$P(Y = y \mid Z = z) = \frac{P(Y = y, Z = z)}{P(Z = z)}$$

Theorem 2

$$f(y, z) = f_Y(y) f(z \mid y).$$

This is the total probabilities theorem.

(In)dependencies

Two r.v. Y and Z are independent iif their density writes:

$$f(y, z) = f_Y(y) f_Z(z).$$

If Y and Z are independent, then:

$$f(y \mid z) = f_Y(y).$$

Expectation

Let g be a measurable function, we define for f the density of X:

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^p} g(x)f(x)\mathrm{d}x$$

in the continuous (Lebesgue) case, and

$$\mathbb{E}[g(X)] = \sum_x g(x)P(X = x)$$

in the discrete (counting) case.

Properties:

- $\mathbb{E}[k] = k,\ k \in \mathbb{R}^p$,
- $E[\alpha g(X) + \beta h(X)] = \alpha\mathbb{E}[g(X)] + \beta\mathbb{E}[h(X)]$, for g and h measurable.

The mean is an element of the space of destination of g. In particular, $E[X] \in \mathbb{R}^p$ is called the *mean vector*.

(Co)variance

If Y is a real random variable, its variance is defined as:

$$\sigma_Y^2 = \mathbb{E}[(Y - \mathbb{E}[Y])^2].$$

Let $X = (Y, Z)$, with $Y, Z \in \mathbb{R}$ (i.e. bidimensional rv).

$$\sigma_{Y,Z} = Cov(Y, Z) = \mathbb{E}[(Y - \mathbb{E}[Z])(Z - \mathbb{E}[Z])].$$

Note that $\sigma_{Y,Z} = \sigma_{Z,Y}$, and that $\sigma_{Y,Z} = \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]$.
The covariance matrix of X is defined as:

$$\Sigma_X = \begin{pmatrix} \sigma_Y^2 & \sigma_{Y,Z} \\ \sigma_{Z,Y} & \sigma_Z^2 \end{pmatrix}.$$

- Σ_X is a square, symmetric, positive semidefinite matrix.
- Y and Z are independent $\overset{\Rightarrow}{\not\Leftarrow}$ $\sigma_{Y,Z} = 0$.

Properties of Covariance

- $Cov(Y, \alpha) = 0$;
- the covariance is bilinear:

$$Cov(\alpha Y + \beta Z, W) = \alpha Cov(Y, W) + \beta Cov(Z, W);$$

- $Var(Y + Z) = Cov(Y + Z, Y + Z) =$
  $V(Z) + V(Y) + 2Cov(Y, Z)$.

Correlation coefficient

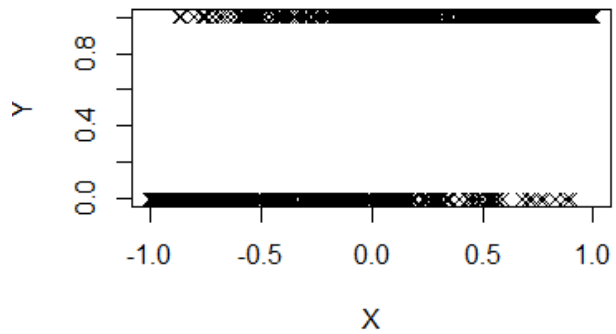
The correlation coefficient is defined as:

$$\rho_{Y,Z} = \frac{\sigma_{Y,Z}}{\sigma_Y \sigma_Z}.$$

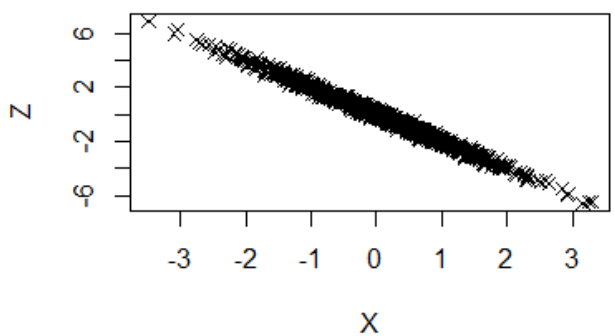
- It accounts for the degree of linear relationship between the two r.v.
- $-1 \leq \rho_{Y,Z} \leq 1$;
- $\rho = 0$ no linear relation, no correlation
- $\rho = 1$ exact direct linear relationship
- $\rho = -1$ exact inverse linear relationship

Example

- If $X_i \sim \mathcal{U}(-1,1)$, and $Y_i \mid X_i \sim \mathcal{B}\left(\frac{arctan(X_i)-arctan(1)}{arctan(1)-arctan(-1)}\right)$, we have $\rho_{X,Y} \simeq 0.65$.

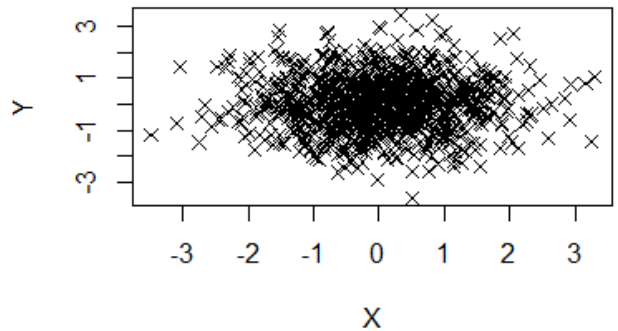


- if $X_i \sim \mathcal{N}(0,1)$, and $Z_i \mid X_i \sim \mathcal{N}(-2X_i, 0.2)$, $\rho_{X,Z} \simeq -0.99$.

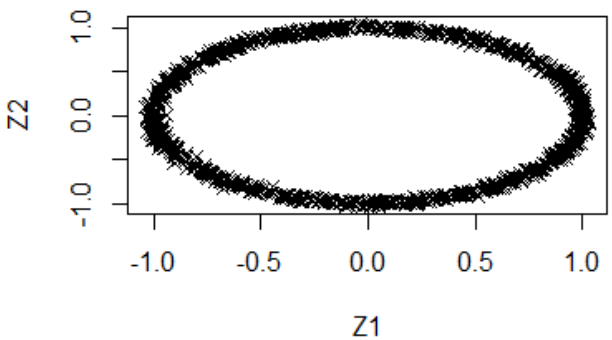


Example

- if $X_i, Y_i \sim \mathcal{N}(0,1)$ independent: $\rho \simeq 0$



- $X_i \sim \mathcal{N}(1, 0.02)$, $Y_i \sim \mathcal{U}(0, 2\pi)$ independent of X_i. We define $Z_i^1 = X_i \cos(Y_i)$ and $Z_i^2 = X_i \sin(Y_i)$.



With a correlation coefficient: $\rho \simeq -0.01$

Higher dimension

All these properties and definition can be extended to any dimension problem.
In $\mathbb{R}^p$ the covariance matrix Σ is $p \times p$, the mean is a vector in $\mathbb{R}^p$, etc.
Let A be a matrix of size $m \times p$, and X a random vector in $\mathbb{R}^p$, linearity writes:
$\mathbb{E}[AX] = A\mathbb{E}[X]$, and $\Sigma_{AX} = A\Sigma_X A^\top$.

Gaussian RV

$$p = 1 \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$

$$p > 1 \quad f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)\right).$$

Note:
$(x-m)^\top \Sigma^{-1}(x-m)$ is called Mahalanobis distance between x and m associated with Σ.

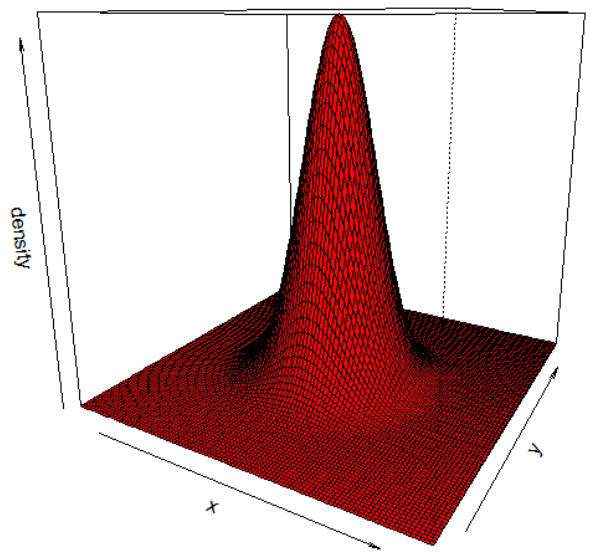
Figure: Prasanta Chandra Mahalanobis (1893-1972)

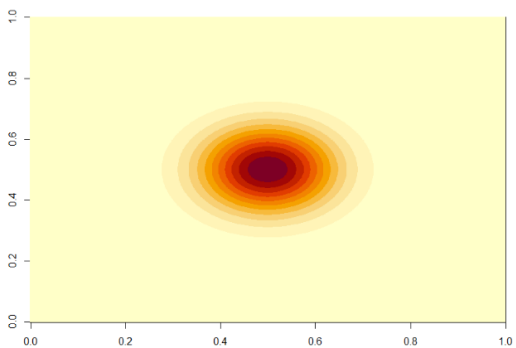
Gaussian RV

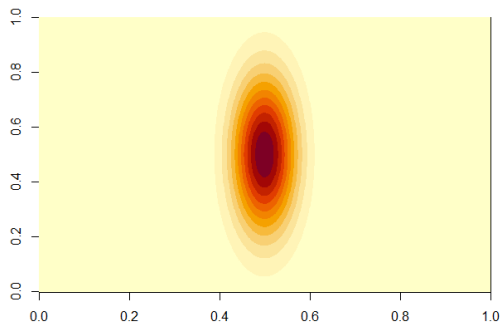
If $X \sim \mathcal{N}(m, \Sigma)$

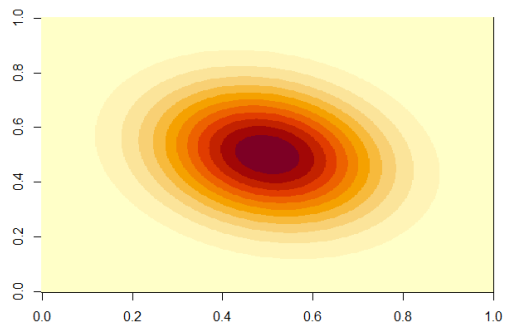
- $\mathbb{E}[X] = m$, $Cov(X) = \Sigma$ (thus, Σ is symmetric, definite positive).
- If the variables are higly correlated: the eigenvalues of Σ are close to 0, $|\Sigma|$ is small.
- Any linear combination of X is Gaussian: $AX + \beta$ is also normal. Compute its mean and covariance (**exercice**).
- Special cases if A is $1 \times p$: projection of the vector X, $AX \in \mathbb{R}$.
- We can reduce any Gaussian RV to a **centered reduced** gaussian distribution (i.e. with mean 0 and with variance I_p) (**exercice**).

Some plots









Related distribution: Chi-square

- If $X \sim \mathcal{N}(0, I_p)$, then $\|X\|^2 = \sum_{i=1}^{p} X_i^2 = X^\top X \sim \chi^2_p$.

Conditional distributions

All marginal and conditional distributions are also **normal**.

- If we take $A = (1, 0, \ldots, 0)$, AX corresponds to the 1st marginal.
- same for any subset of dimensions $A = (0, \ldots, 1, \ldots, 1, \ldots, 0)$.

Consequence, if X and Y are Gaussian, we can write their joint distribution:

$$\mathbb{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \; Cov(X, Y) = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}.$$

Conditional distributions

Note: For Gaussian distributions, Independence $\Leftrightarrow$ Uncorrelation. That is, if $\Sigma_{XY} = 0$ then X and Y are independent.

How to compute the conditional distribution of X and Y?

- we have the joint distribution and the marginal that are all Gaussian.
- After computation it is multivariate normal:

$$\mathbb{E}[X \mid Y] = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y)$$
$$Cov(X, Y) = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$$

Geometric interpretation of the Covariance matrix

- Σ is symmetrical definite positive.
- Level lines of the Gaussian density are of the form $x^\top \Sigma x = \alpha$.
- What does $\{x \in \mathbb{R}^p \mid x^\top \Sigma x = 1\}$ look like?
- Σ is symmetric definite positive...