ML01 – Introduction to Machine Learning

Lecture 2: Basics of Matrix Algebra, Probability and Statistics

Thierry Denœux

tdenoeux@utc.fr

https://www.hds.utc.fr/~tdenoeux

Université de Technologie de Compiègne

Spring 2019

Overview

1. Matrix algebra
   - Basic definitions
   - Matrix operations
2. Probability
   - Basic notions
   - Random vectors
   - Multivariate normal distribution
3. Statistical inference
   - Random sample
   - Estimation

Matrix

Definition

A matrix $\mathbf{A}$ *is a rectangular array of numbers. If* $\mathbf{A}$ *has n rows and p columns we say it is of order* $n \times p$*. For example, n observations on p variables are arranged in this way.*

Notation: we write matrix $\mathbf{A}$ of order $n \times p$ as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix},$$

where a_{ij} is the element in row i and column j of the matrix $\mathbf{A}$. Sometimes. we write $(\mathbf{A})_{ij}$ for a_{ij}.

Transpose

Definition

The transpose of a matrix $\mathbf{A}(n \times p)$ *is formed by interchanging the rows and columns:*

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1p} & a_{2p} & \dots & a_{np} \end{bmatrix}$$

Its order is $p \times n$.

Property: $(\mathbf{A}^T)^T = \mathbf{A}$

Vectors

Definition

A matrix with column-order one is called a column vector. Thus,

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix},$$

Row vectors are matrices of row-order one. They can be written as column vectors transposed, i.e.

$$\mathbf{a}^T = (a_1, \ldots, a_n)$$

Particular matrices

Name	Definition	Notation	Examples
Column vector	$p = 1$	$\mathbf{a}, \mathbf{b}, \ldots$	$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$
Unit vector	$(1, \ldots, 1)^T$	$\mathbf{1}$ or $\mathbf{1}_p$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
Square	$p = n$	$\mathbf{A}(p \times p)$	$\begin{pmatrix} 1 & 3 \\ 4 & 5 \end{pmatrix}$
Diagonal	$p = n,\ a_{ij} = 0,\ i \neq j$	$\mathrm{diag}(a_{ii})$	$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$
Identity	$\mathrm{diag}(\mathbf{1})$	$\mathbf{I}$ or $\mathbf{I}_p$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Scalar	$c\mathbf{I}$		$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$
Symmetric	$a_{ij} = a_{ji}$		$\begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix}$

Overview

1. Matrix algebra
   - Basic definitions
   - Matrix operations
2. Probability
   - Basic notions
   - Random vectors
   - Multivariate normal distribution
3. Statistical inference
   - Random sample
   - Estimation

Arithmetic operations

- The sum and difference of matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order are

$$\mathbf{A}+\mathbf{B}=(a_{ij}+b_{ij}) \quad \text{and} \quad \mathbf{A}-\mathbf{B}=(a_{ij}+b_{ij})$$

- If $\mathbf{A}$ has order $n \times p$ and $\mathbf{B}$ has order $p \times q$, the product $\mathbf{AB}$ is the matrix of order $n \times q$ defined by

$$(\mathbf{AB})_{ij}=\sum_{k=1}^{p} a_{ik} b_{kj}$$

  Remark: in general, $\mathbf{AB} \neq \mathbf{BA}$.

- Properties:

$$(\mathbf{A}+\mathbf{B})^T=\mathbf{A}^T+\mathbf{B}^T, \quad (\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$$

Determinant

Definition

The determinant of a square matrix $\mathbf{A}$ *is defined as*

$$|\mathbf{A}| = \sum_{\tau} (-1)^{|\tau|} a_{1\tau(1)} \ldots a_{p\tau(p)}$$

where the summation is taken over all permutations τ *of* $(1, 2, \ldots, p)$ *and* $|\tau|$ *equals* $+1$ *or* -1 *depending on whether* τ *can be written as the product of an even or odd number of transpositions.*

For $p = 2$, $|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$.

Some properties of the determinant

Proposition

- *If* $\mathbf{A}$ *is diagonal,*

$$|\mathbf{A}| = \prod_i a_{ii}$$

- $|c\mathbf{A}| = c^p|\mathbf{A}|$
- $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$

Non-singular matrix

Definition (Linear independence)

Veclors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *are called linearly dependent if there exist numbers* $\lambda_1, \ldots, \lambda_k$ *not all zero such that*

$$\lambda_1 \mathbf{x}_1 + \ldots + \lambda_k \mathbf{x}_k = 0$$

Otherwise the k *vectors are linearly independent.*

Definition (Nonsingular matrix)

A square matrix is nonsingular if its column vectors (or, equivalently, its row vectors) are linearly independent; otherwise it is singular.

Proposition

The square matrix $\mathbf{A}$ *is nonsingular iff* $|\mathbf{A}| \neq 0$.

Inverse

Definition

The *inverse* *of square matrix* $\mathbf{A}$ *is the unique matrix* $\mathbf{A}^{-1}$ *satisfying*

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

The inverse exists if and only if $\mathbf{A}$ *is non-singular, that is, if and only if* $|\mathbf{A}| \neq 0$.

Proposition

- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Matrix differentiation

Definition

The ***derivative*** *of* $f(\mathbf{X})$ *with respect to* $\mathbf{X}(n \times p)$ *is the matrix*

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\right)$$

Proposition

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}, \quad \frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}, \quad \frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}\,\mathbf{y}$$

Matrix definition and operations in R

```
> D<-diag(c(1,1)) # diagonal matrix
> D
 1   0
 0   1
> A<-matrix(c(1,2,3,4),2,2)
> A
 1   3
 2   4
> det(A) # determinant
-2
```

Matrix definition and operations in R (continued)

```
> B<-solve(A) # inverse
B
 -2   1.5
  1  -0.5
> A * B # Hadamard (pointwise) product
 -2   4.5
  2  -2.0
> A %*% B # matrix multiplication
 1  0
 0  1
```

Overview

1. Matrix algebra
   - Basic definitions
   - Matrix operations
2. Probability
   - Basic notions
   - Random vectors
   - Multivariate normal distribution
3. Statistical inference
   - Random sample
   - Estimation

Probability space

Definition (Probability space)

A probability space is a model of a random experiment. It is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, *where*

- Ω *is the set of outcomes*
- $\mathcal{A}$ *is a set of events*
- $\mathbb{P}$ *is a mapping (called a probability measure) from* $\mathcal{A}$ *to* $[0,1]$ *that assigns to each event* $A \in \mathcal{A}$ *its probability* $\mathbb{P}(A)$.

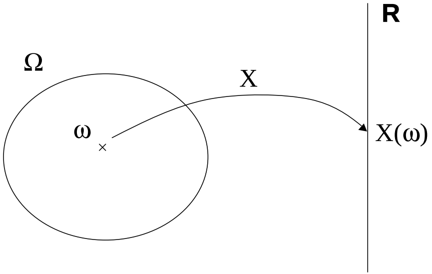
Mapping $\mathbb{P}$ must verify $\mathbb{P}(\Omega) = 1$ and, for any countable family of events $(A_i)_{i \in I}$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$,

$$\mathbb{P}\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \mathbb{P}(A_i)$$

Random variable

Definition (Random variable)

A random variable *(r.v.) is a quantity that depends on the outcome of a random experiment. Formally, it is a mapping from a probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ *to* $\mathbb{R}$*.*

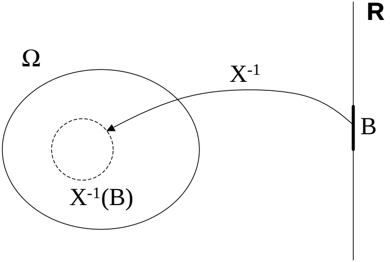


Probability distribution

- Given a subset B of the real line, the probability that $X \in B$ is

$$\mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\})$$

- The mapping $B \to \mathbb{P}(X \in B)$ is called the probability distribution of X.



Discrete random variable

Definition (Discrete r.v.)

X is discrete if it takes countably many values $\{x_1, x_2, \ldots\}$. We define the probability function of X by

$$p_X(x) = \mathbb{P}(X = x)$$

We have

$$\sum_i p_X(x_i) = 1$$

and, for any subset B of the real line,

$$\mathbb{P}(X \in B) = \sum_{x_i \in B} p_X(x_i).$$

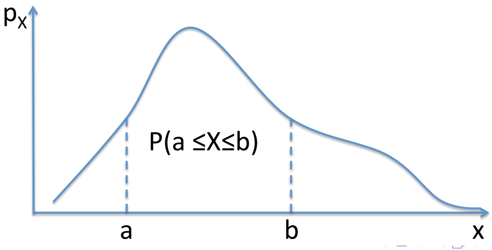
Continuous random variable

Definition (Continuous r.v.)

A r.v. X is continuous if there exists a function $p_X : \mathbb{R} \mapsto \mathbb{R}_+$ such that $\int_{-\infty}^{+\infty} p_X(x)dx = 1$, and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b p_X(x)dx.$$

Function p_X is called the probability density function (pdf) of X.



Expectation

Definition (Expectation)

The expectation μ *of r.v.* X *is a one-number summary of* X*. It is defined by*

$$\mu = \mathbb{E}(X) = \begin{cases} \displaystyle\sum_{x \in V_X} x\, p_X(x) & \text{if } X \text{ is discrete} \\ \displaystyle\int_{\mathbb{R}} x\, p_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

(if these quantities exist).

Property: For any constants a and b,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Variance

Definition (Variance)

The *variance* *of r.v. X is a measure of variability of X. It is defined by*

$$\sigma^2 = Var(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

(if this quantity exists). The *standard deviation* *of X is $\sigma = \sqrt{Var(X)}$.*

Properties:

- $\mathrm{Var}(X) \geq 0$
- For any constants a and b,

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

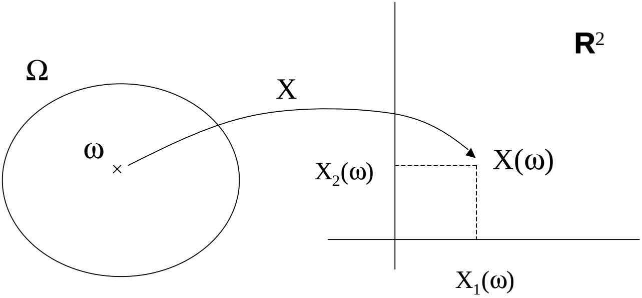
Overview

1. Matrix algebra
   - Basic definitions
   - Matrix operations
2. Probability
   - Basic notions
   - Random vectors
   - Multivariate normal distribution
3. Statistical inference
   - Random sample
   - Estimation

Random vector

Definition

A *random vector* *is a vector* $\mathbf{X} = (X_1, \ldots, X_p)^T$ *whose components* X_j *are random variables. Formally, it is a mapping from a probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ *to* $\mathbb{R}^p$.



Discrete random vector

Definition

The random vector $\mathbf{X}$ *is discrete if it takes countably many values* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$. *We define its probability function by*

$$p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}).$$

We write $p(\mathbf{x})$ *when there is no ambiguity.*

We have

$$\sum_i p_{\mathbf{X}}(\mathbf{x}_i) = 1$$

and, for every subset B of $\mathbb{R}^p$,

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{\mathbf{x}_i \in B} p_{\mathbf{X}}(\mathbf{x}_i).$$

Continuous random vector

Definition

The random vector $\mathbf{X}$ *is* *continuous* *if there exists a function* $p_{\mathbf{X}} : \mathbb{R}^p \mapsto \mathbb{R}_+$ *such that, for every subset* B *of* $\mathbb{R}^p$*,*

$$\mathbb{P}(\mathbf{X} \in B) = \int_B p_{\mathbf{X}}(\mathrm{x}) d\mathrm{x}$$

Function $p_{\mathbf{X}}$ *is called the* *(joint) probability density function (pdf)* *of* $\mathbf{X}$*. We write* $p(\mathrm{x})$ *when there is no ambiguity.*

Marginal distribution

Definition

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ *a random vector, and* $J \subset \{1, \ldots, p\}$. *The probability distribution of the random sub-vector* $\mathbf{X}_J = (X_j)_{j \in J}$ *is called the marginal distribution of* $\mathbf{X}_J$.

- The probability or density function $p(\mathbf{x}_J)$ is obtained by summing $p(\mathbf{x})$ over the components $j \notin J$.
- For instance, for a 2-D random vector $\mathbf{X} = (X_1, X_2)$,

$$p(x_1) = \begin{cases} \int_{-\infty}^{+\infty} p(x_1, x_2) dx_2 & \text{if } \mathbf{X} \text{ is continuous} \\ \sum_{x_2} p(x_1, x_2) & \text{if } \mathbf{X} \text{ is discrete} \end{cases}$$

Conditional distribution

Definition

Assume for simplicity that $p = 2$. The conditional distribution of X_1 given $X_2 = x_2$ is defined by the following probability or density function:

$$p(x_1 \mid X_2 = x_2) = \frac{p(x_1, x_2)}{p(x_2)},$$

which is defined iff $p(x_2) \neq 0$.

Bayes' theorem

- We have defined

$$p(x_1 \mid X_2 = x_2) = \frac{p(x_1, x_2)}{p(x_2)},$$

- Symmetrically,

$$p(x_2 \mid X_1 = x_1) = \frac{p(x_1, x_2)}{p(x_1)}.$$

- Hence

$$p(x_1 \mid X_2 = x_2) = \frac{p(x_2|X_1 = x_1)p(x_1)}{p(x_2)}.$$

This formula is called Bayes' theorem.

Independence

Definition

The r.v.'s $X_1, \ldots, X_n$ are said to be independent if, for any events $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1; \ldots; X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).$$

(The joint distribution is the product of the marginal distributions).

Equivalent condition:

$$p(x_1, \ldots, x_n) = p(x_1) \ldots p(p_n),$$

where $p(\cdot)$ denotes probability or density functions.

Expectation

Definition

The expectation of random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ *is the vector*

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))^T.$$

Proposition

For any constant matrix $\mathbf{A}(q \times p)$ *and any constant vector* $\mathbf{b} \in \mathbb{R}^q$, *we have*

$$\mathbb{E}(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\,\mathbb{E}(\mathbf{X}) + \mathbf{b}.$$

In particular, if $\mathbf{u} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\mathbb{E}(\mathbf{X})$.

Covariance and correlation

Definition (Covariance and correlation)

Let (X, Y) be a random vector. The covariance and the correlation between X and Y are defined, respectively, by

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right].$$

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Properties:

- $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}(X)\mathbb{E}(Y)$
- $-1 \leq \rho(X, Y) \leq 1$. If $Y = aX + b$ for some constants a and b, then $\rho(X, Y) = 1$ if $a > 0$ and $\rho(X, Y) = -1$ if $a < 0$
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$. The converse is not true in general.

Variance and correlation matrices

Definition

The variance matrix of random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ *is the matrix* $\mathbf{\Sigma}(p \times p)$ *with diagonal terms* $(\mathbf{\Sigma})_{ii} = Var(X_i)$ *and off-diagonal terms*

$$(\mathbf{\Sigma})_{ij} = Cov(X_i, X_j), \quad i \neq j.$$

The correlation matrix of random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ *is the matrix* $\mathbf{R}(p \times p)$ *with diagonal terms* $(\mathbf{R})_{ii} = 1$ *and off-diagonal terms*

$$(\mathbf{R})_{ij} = \rho(X_i, X_j), \quad i \neq j.$$

Remark: If the r.v.'s $X_1, \ldots, X_n$ are independent then $\mathbf{\Sigma}$ is diagonal and $\mathbf{R} = \mathbf{I}$.

Properties of the variance

Proposition

- *The variance matrix can be written in matrix form as*

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

- *Matrix $\boldsymbol{\Sigma}$ is symmetric.*
- *For any constant matrix $\mathbf{A}(q \times p)$ and vector $\mathbf{b} \in \mathbb{R}^q$, we have*

$$Var(\mathbf{A}\mathbf{X}+\mathbf{b}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T.$$

- *In particular, for any vector $\mathbf{u} \in \mathbb{R}^p$*

$$Var(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}.$$

Properties of the variance (continued)

- The equality $\text{Var}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\mathbf{\Sigma}\mathbf{u}$ shows that, for any $\mathbf{u} \neq \mathbf{0}$, we have

$$\mathbf{u}^T\mathbf{\Sigma}\mathbf{u} > 0 \tag{1}$$

  unless there exists a deterministic relation $\mathbf{u}^T\mathbf{X} = c$ for some constant vector $\mathbf{u}$ and scalar c.
- A symmetric matrix $\mathbf{\Sigma}$ verifying (1) for any $\mathbf{u} \neq \mathbf{0}$ is said to be positive definite, and we write $\mathbf{A} > 0$.
- It can be shown that a positive definite matrix is nonsingular.
- Consequently, $\mathbf{\Sigma}^{-1}$ exists, except if there is a linear relation $\mathbf{u}^T\mathbf{X} = c$ among the variables in $\mathbf{X}$.

Overview

1 Matrix algebra
- Basic definitions
- Matrix operations

2 Probability
- Basic notions
- Random vectors
- Multivariate normal distribution

3 Statistical inference
- Random sample
- Estimation

Definition of the multivariate normal distribution

Definition

Way say that random vector $\mathbf{X}$ *has a multivariate normal distribution if it has the following density function:*

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Notation: $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Property:

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}.$$

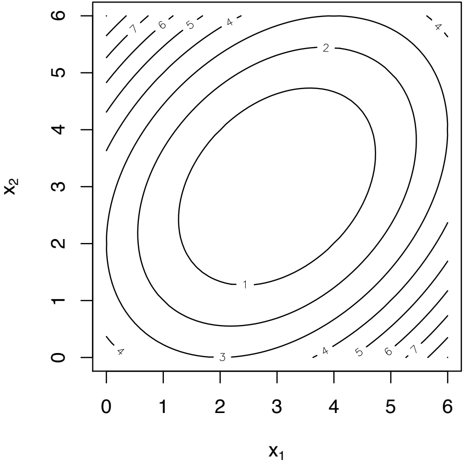
Properties of the multivariate normal distribution

- When $p = 1$, we have the univariate normal distribution with $\sigma^2 = \mathbf{\Sigma}$.
- Matrix Σ is diagonal iff r.v.'s $X_1, \ldots, X_p$ are independent.
- Any sub-vector of $\mathbf{X}$ has a normal distribution. In particular, the components X_i have normal distributions $\mathcal{N}(\mu_i, \sigma_i^2)$ with $\sigma_i^2 = (\mathbf{\Sigma})_{ii}$.
- The multivariate normal distribution has constant density on ellipses or ellipsoids of the form

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

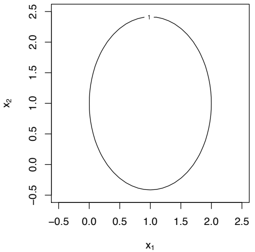
  c being a constant. These ellipsoids are called the contours the distribution. For $\boldsymbol{\mu} = 0$ these contours are centered at the origin and when $\mathbf{\Sigma} = a\mathbf{I}$ the contours are circles or, in higher dimensions, spheres or hyperspheres.

Example with $\boldsymbol{\mu} = (3, 3)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$

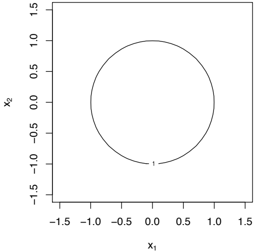


More examples

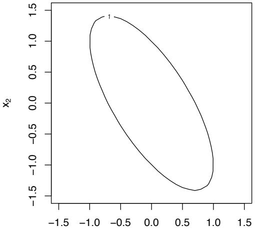
$$\boldsymbol{\mu} = (1,1)^T,\ \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$



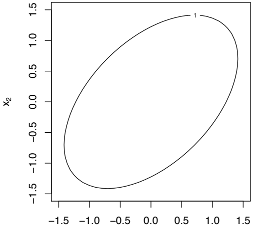
$$\boldsymbol{\mu} = (0,0)^T,\ \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{\mu} = (0,0)^T,\ \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

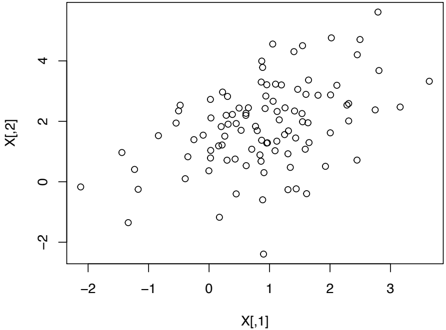


$$\boldsymbol{\mu} = (0,0)^T,\ \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$



Multivariate normal random vector generation in R

```
library(mvtnorm)
mu<-c(1,2)
Sigma<-matrix(c(1,0.5,0.5,2),2,2)
X<-rmvnorm(100,mu,Sigma)
plot(X)
```



Overview

1 Matrix algebra
- Basic definitions
- Matrix operations

2 Probability
- Basic notions
- Random vectors
- Multivariate normal distribution

3 Statistical inference
- Random sample
- Estimation

Modeling the data-generating process

- We have seen that, in machine learning, we wish to make predictions based on past observed data (a training set).
- For that purpose, we need a model of data-generating process (the way data are generated).
- In this section, we introduce two important notions:
  - Random sample
  - Statistical model

Introductory example

- Let X be the weight of a student picked at random in the population of Chinese male students. It is a random variable.
- Assume that we pick n students. We can denote by X_1 the weight of the 1st student, X_2 the weight of the 2nd student, etc.
- The n variables $X_1, \ldots, X_n$ are independent, and they all have the same distribution.
- We say that the random vector $(X_1, \ldots, X_n)$ is an independent and identically distributed (iid) random sample.

Random sample

Definition (iid sample)

If $X_1, \ldots, X_n$ are independent and each has the same marginal distribution with probability or density function p_X, we say that $X_1, \ldots, X_n$ are independent and identically distributed (iid) and we write

$$X_1, \ldots, X_n \sim p_X$$

We also call $X_1, \ldots, X_n$ a random sample of size n from p_X.

Remarks on random samples

- A random sample represents the data-generating process.
- An actual dataset $x_1, \ldots, x_n$ is called a realization of the random sample $X_1, \ldots, X_n$.
- The individual observations can be vectors. In that case we have a random sample of n random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Statistical inference

- Statistical inference, or learning as it is called in computer science, is the process of using data to infer (approximate) the distribution that generated the data.
- A typical statistical inference question is:

  Given a dataset $x_1, \ldots, x_n$ assumed to be a realization of a random sample $X_1, \ldots, X_n \sim p_X$, how do we infer (approximate) p_X, or some property of p_X such as its mean?

Statistical inference

Example

- Assume that we have observed the weights (in kg) of 10 male students

$$64.0, 52.7, 73.4, 78.2, 50.0, 75.7, 83.3, 62.0, 63.3, 44.3$$

- What can we say about
  - The mean weight $\mathbb{E}(X)$ in the whole population of male Chinese students?
  - The probability that the weight of a random student will be greater that 80 kg?
- To answer such questions, we start from a statistical model – a set of probability distributions that are postulated to contain p_X (or a good approximation of it).

Statistical model

Definition (Statistical model)

A statistical model is a set of probability distributions. A parametric model is a set that can be parameterized by a finite number of parameters.

Examples:

- We can assume that the weights of male Chinese students have a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2. The weight of a randomly picked student is then $X \sim \mathcal{N}(\mu, \sigma^2)$.
- We can assume that the heights and weights of male Chinese students have a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The height and weight of a randomly picked student is then $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Overview

1. Matrix algebra
   - Basic definitions
   - Matrix operations
2. Probability
   - Basic notions
   - Random vectors
   - Multivariate normal distribution
3. Statistical inference
   - Random sample
   - Estimation

Point estimation

- Point estimation refers to providing a single "best guess" of some quantity of interest. The quantity of interest could be
  - A parameter in a parametric model
  - A probability density function p_X
  - A regression function $f(x) = \mathbb{P}(Y \mid X = x)$
  - A prediction for a future value Y of some random variable.
- By convention, we denote a (point) estimator of a parameter θ by $\widehat{\theta}$. Remember that θ is a fixed, unknown quantity. The estimator $\widehat{\theta}$ depends on the data so it is a random variable.

Point estimator

Definition

Let $X_1, \ldots, X_n$ be n iid data points from some distribution $p_X(x;\theta)$ depending on some parameter θ. A point estimator $\widehat{\theta}$ of θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta} = g(X_1, \ldots, X_n).$$

We say that

- *$\widehat{\theta}$ is unbiased if $\mathbb{E}(\widehat{\theta}) = \theta$*
- *$\widehat{\theta}$ is consistent if it converges to the true parameter value θ as we collect more and more data ($N \to \infty$).*

Example

- Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be an iid sample. Assume we want to estimate $\theta = \mathbb{E}(X)$.
- Let $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\mathbf{1}^T\mathbf{X}$.
- We have

$$\mathbb{E}(\widehat{\theta}) = \mathbb{E}\left(\frac{1}{n}\mathbf{1}^T\mathbf{X}\right) = \frac{1}{n}\mathbf{1}^T\mathbb{E}(\mathbf{X}) = \frac{1}{n}\mathbf{1}^T(\theta\mathbf{1}) = \frac{\theta}{n}\underbrace{\mathbf{1}^T\mathbf{1}}_{n} = \theta.$$

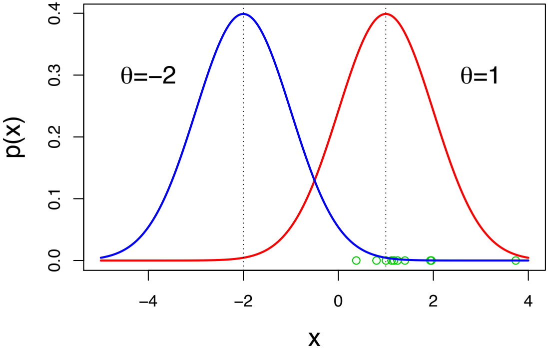
$$\text{Var}(\widehat{\theta}) = \text{Var}\left(\frac{1}{n}\mathbf{1}^T\mathbf{X}\right) = \frac{1}{n^2}\mathbf{1}^T\text{Var}(\mathbf{X})\mathbf{1} = \frac{1}{n^2}\mathbf{1}^T(\sigma^2\mathbf{I})\mathbf{1} = \frac{\sigma^2}{n}.$$

- So, $\widehat{\theta}$ is an unbiased estimator of θ. As $\text{Var}(\widehat{\theta}) \to 0$ when $n \to \infty$, we can show that $\widehat{\theta}$ tends to θ when $n \to \infty$ (it is consistent).

Maximum Likelihood estimation

Example

- Consider the statistical model: $X \sim \mathcal{N}(\theta, 1)$ with $\theta \in \{-2, 1\}$.
- Given the 10 green data points below, which value of θ is more likely?



Maximum Likelihood estimation

Definition

Definition

Given the model model $X \sim p(x;\theta)$ with $\theta \in \Theta$, and an iid sample $X_1, \ldots, X_n$, the likelihood function is the mapping

$$\begin{aligned} L: \quad & \Theta \mapsto \mathbb{R}_+ \\ & \theta \to L(\theta; x_1, \ldots, x_n) = p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta). \end{aligned}$$

The log-likelihood is the function $\ell(\theta; x_1, \ldots, x_n) = \ln L(\theta; x_1, \ldots, x_n)$.

Definition

The maximum likelihood estimator (MLE) of θ is the estimator $\widehat{\theta}$ that maximizes the likelihood (or log-likelihood) function:

$$\ell(\widehat{\theta}; x_1, \ldots, x_n) = \max_{\theta \in \Theta} \ell(\theta; x_1, \ldots, x_n)$$

Maximum Likelihood estimation

Example

- Assume $X \sim \mathcal{N}(\theta, 1)$.
- We have

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right)$$
$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$
$$\ell(\theta; x_1, \ldots, x_n) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

- To find the MLE of θ, we solve the equation $\ell'(\theta; x_1, \ldots, x_n) = 0$.
- The solution is the estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.