Probability & Statistics for DS & AI

Joint distributions

Michele Guindani

Jio Institute

Summer 2022

What are joint distributions?

Joint distributions are high-dimensional PDF (or PMF or CDF).

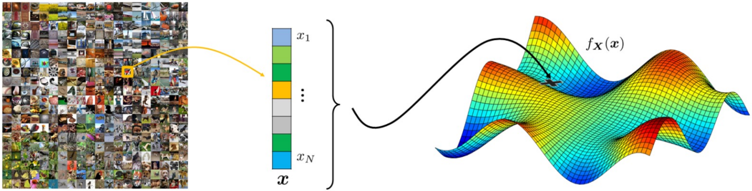
$$\underbrace{f_X(x)}_{\text{one variable}} \implies \underbrace{f_{X_1,X_2}(x_1,x_2)}_{\text{two variables}} \implies \underbrace{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}_{\text{three variables}}$$
$$\implies \ldots \implies \underbrace{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)}_{N \text{ variables}}$$

Notation:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$$

Why study joint distributions?

- Joint distributions are ubiquitous in modern data analysis.
- For example, an image from a dataset can be represented by a high-dimensional vector $\boldsymbol{x}$.
- Each vector has certain probability to be present.
- Such probability is described by the high-dimensional joint PDF $f_X(\boldsymbol{x})$.



Joint PMF

Definition

Let X and Y be two discrete random variables. The joint PMF of X and Y is defined as

$$p_{X,Y}(x,y) = \mathbb{P}[X = x \quad \text{and} \quad Y = y]$$

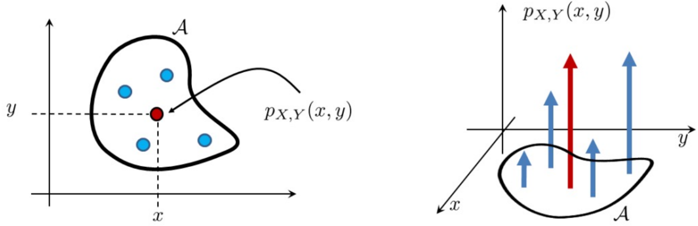


Figure: A joint PMF for a pair of discrete random variables consists of an array of impulses. To measure the size of the event $\mathcal{A}$, we sum all the impulses inside $\mathcal{A}$.

Example (Coin and die)

Let X be a coin flip, Y be a dice. Find the joint PMF.

Solution: The sample space of X is $\{0, 1\}$. The sample space of Y is $\{1, 2, 3, 4, 5, 6\}$. The joint PMF is

	Y					
X = 0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
X = 1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Or written in equation:

$$p_{X,Y}(x, y) = \frac{1}{12}, \quad x = 0, 1, \quad y = 1, 2, 3, 4, 5, 6$$

Example (Coin and die - contd.)

In the previous example, define $\mathcal{A} = \{X + Y = 3\}$ and $\mathcal{B} = \{\min(X, Y) = 1\}$. Find $\mathbb{P}[\mathcal{A}]$ and $\mathbb{P}[\mathcal{B}]$.

Solution:

$$\begin{aligned}\mathbb{P}[\mathcal{A}] &= \sum_{(x,y)\in\mathcal{A}} p_{X,Y}(x, y) = p_{X,Y}(0, 3) + p_{X,Y}(1, 2) \\ &= \frac{2}{12}\end{aligned}$$

$$\begin{aligned}\mathbb{P}[\mathcal{B}] &= \sum_{(x,y)\in\mathcal{B}} p_{X,Y}(x, y) \\ &= p_{X,Y}(1, 1) + p_{X,Y}(1, 2) + \ldots + p_{X,Y}(1, 5) + p_{X,Y}(1, 6) \\ &= \frac{6}{12}\end{aligned}$$

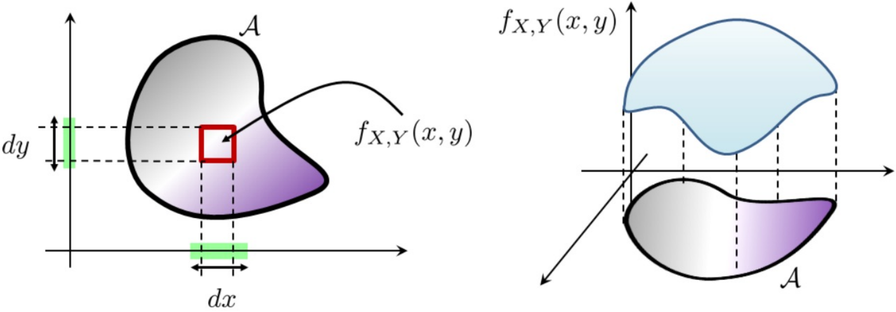
Joint PDF

Definition

Let X and Y be two continuous random variables. The joint PDF of X and Y is a function $f_{X,Y}(x, y)$ that can be integrated to yield a probability:

$$\mathbb{P}[\mathcal{A}] = \int_{\mathcal{A}} f_{X,Y}(x, y)\,dxdy$$

for any event $\mathcal{A} \subseteq \Omega_X \times \Omega_Y$



Example

Consider a uniform joint PDF $f_{X,Y}(x,y)$ defined on $[0,2]^2$ with $f_{X,Y}(x,y)=\frac{1}{4}$. Let $\mathcal{A}=[a,b]\times[c,d]$. Find $\mathbb{P}[\mathcal{A}]$.

Solution:

$$\begin{aligned}\mathbb{P}[\mathcal{A}] &= \mathbb{P}[a \le X \le b, \quad c \le X \le d] \\ &= \int_c^d \int_a^b f_{X,Y}(x,y)\,dxdy \\ &= \int_c^d \int_a^b \frac{1}{4}\,dxdy = \frac{(d-c)(b-a)}{4}\end{aligned}$$

Suppose $[a,b]\equiv[0,1]$, $[c,d]\equiv[0.5,1.5]$, then

$$\mathbb{P}[\mathcal{A}] = \mathbb{P}[0 \le X \le 1, \quad 0.5 \le Y \le 1.5] = \frac{1}{4}$$

Marginal PMF and PDF

The marginal PMF is defined as

$$p_X(x) = \sum_{y \in \Omega_Y} p_{X,Y}(x,y) \text{ and } p_Y(y) = \sum_{x \in \Omega_X} p_{X,Y}(x,y)$$

and the marginal PDF is defined as

$$f_X(x) = \int_{\Omega_Y} f_{X,Y}(x,y)dy \quad \text{ and } f_Y(y) = \int_{\Omega_X} f_{X,Y}(x,y)dx$$

Independence of random variables

Definition

If two random variables X and Y are independent, then

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{and} \quad f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Definition

If a sequence of random variables $X_1, \ldots, X_N$ are independent, then their joint PDF (or joint PMF) can be factorized:

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \prod_{n=1}^{N} f_{X_n}(x_n) \tag{1}$$

Example

- Consider a uniform joint PDF $f_{X,Y}(x,y)$ defined on $[0,2]^2$ with $f_{X,Y}(x,y) = \frac{1}{4}$. Let $\mathcal{A} = [a,b] \times [c,d]$. Find $\mathbb{P}[\mathcal{A}]$

  In this case,

$$f_{X,Y}(x,y) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = f_X(x)\, f_Y(y) \quad 0 \leq x, y \leq 2$$

🤔 Why is this important?

👉 It is easy to compute in Python the probability of events of vectors of random variables (multivariate distributions) if we know the independence structure between the random variables.💞

Example (Python example)

- Let $[a, b] \equiv [0, 1], [c, d] \equiv [0.5, 1.5]$. Compute $\mathbb{P}[\mathcal{A}] = \mathbb{P}[0 \leq X \leq 1, \quad 0.5 \leq Y \leq 1.5]$.
- We can use Monte Carlo approximation.

```
import numpy as np
import random
np.random.seed(12345)
nreps=1000000
x= np.random.uniform(low=0, high=2, size=nreps)
y= np.random.uniform(low=0, high=2, size=nreps)

# Probability of A
condition=np.zeros(nreps)
for rep in range(nreps):
condition[rep]=(((x[rep]>0) and (x[rep]<1)) and ((y[rep]>0.5) and (y[rep]<1

count=sum(condition)
freq=count/nreps
print(freq)  #0.250293
```

Example

In the previous example, $\mathcal{B} = \{X + Y \leq 2\}$. Find $\mathbb{P}[\mathcal{B}]$

```
import numpy as np
import random
np.random.seed(12345)
nreps=1000000
x= np.random.uniform(low=0, high=2, size=nreps)
y= np.random.uniform(low=0, high=2, size=nreps)

# Probability of B
sumxy=np.zeros(nreps)
for rep in range(nreps):
sumxy[rep]=x[rep]+y[rep]

freq=sum(sumxy<=2)/nreps
print(freq) #0.4998
```

Not all r.v. are independent!!

- Consider two random variables X and Y with a joint PDF given by

$$\begin{aligned} f_{X,Y}(x,y) &\propto \exp\left\{-(x-y)^2\right\} \\ &= \exp\left\{-x^2+2xy-y^2\right\} \\ &= \underbrace{\exp\left\{-x^2\right\}}_{f_X(x)} \underbrace{\exp\{2xy\}}_{\text{extra term}} \underbrace{\exp\left\{-y^2\right\}}_{f_Y(y)} \end{aligned}$$

- This PDF cannot be factorized into a product of two marginal PDFs. Therefore, the random variables are dependent.

An interesting case

Example

- A joint Gaussian random variable (X, Y) has a joint PDF given by:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{\left((x-\mu_X)^2 + (y-\mu_Y)^2\right)}{2\sigma^2} \right\}$$

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$

- **Solution:**

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{\left((x-\mu_X)^2 + (y-\mu_Y)^2\right)}{2\sigma^2} \right\} dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu_X)^2}{2\sigma^2} \right\} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu_Y)^2}{2\sigma^2} \right\} dy \end{aligned}$$

Example

- Recognizing that the last integral is equal to unity because it integrates a Gaussian PDF over the real line, it follows that

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma^2}\right\}$$

- Similarly, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma^2}\right\}$$

- It's immediate to see that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, hence X, Y are independent.

Independent and Identically Distributed (i.i.d.)

A collection of random variables $X_1, \ldots, X_N$ are called independent and identically distributed (i.i.d.) if

- All $X_1, \ldots, X_N$ are independent;
- All $X_1, \ldots, X_N$ have the same distribution, i.e., $f_{X_1}(x) = \ldots = f_{X_N}(x)$.

⚠ **Why is i.i.d. so important?**

- If a set of random variables are i.i.d., then the joint PDF can be written as a product of PDFs.
- Integrating a joint PDF is not fun. Integrating a product of PDFs is a lot easier.

Example

Let $X_1, X_2, \ldots, X_N$ be a sequence of i.i.d. Gaussian random variables where each X_i has a PDF

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

- The joint PDF of $X_1, X_2, \ldots, X_N$ is

$$\begin{aligned} f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) &= \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left\{-\sum_{i=1}^{N} \frac{x_i^2}{2}\right\} \end{aligned}$$

Joint CDF

- We can extend the definition of CDF also to vectors of r.v.'s.

Let X and Y be two random variables. The joint CDF of X and Y is the function $F_{X,Y}(x, y)$ such that

$$F_{X,Y}(x, y) = \mathbb{P}[X \leq x \cap Y \leq y]$$

- We won't say anything more about joint CDFs (see textbook)

Joint Expectation

- Similarly, for the expectation, we can define the joint expectation:

$$\mathbb{E}[XY] = \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X} xy \cdot pX,\, Y(x, y)$$

if X and Y are discrete, or

$$\mathbb{E}[XY] = \int_{y \in \Omega_Y} \int_{x \in \Omega_X} xy \cdot f_{X,Y}(x, y) dxdy$$

if X and Y are continuous.

Covariance

Let X and Y be two random variables. Then the covariance of X and Y is

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

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Remark:

$$\mathrm{Cov}(X, X) = \mathbb{E}\left[(X - \mu_X)(X - \mu_X)\right] = \mathrm{Var}[X]$$

Theorem

Let X and Y be two random variables. Then,

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Interesting properties

- For any X and Y

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}[Y]$$

Correlation coefficient

Definition

Let X and Y be two random variables. The correlation coefficient is

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

- ρ is always between 0 and 1 , i.e., $1 \leq \rho \leq 1$. This is due to the cosine angle definition.
- When $X = Y$ (fully correlated), $\rho = +1$.
- When $X = -Y$ (negatively correlated), $\rho = -1$.
- When X and Y uncorrelated then $\rho = 0$.

Independence

Theorem

If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

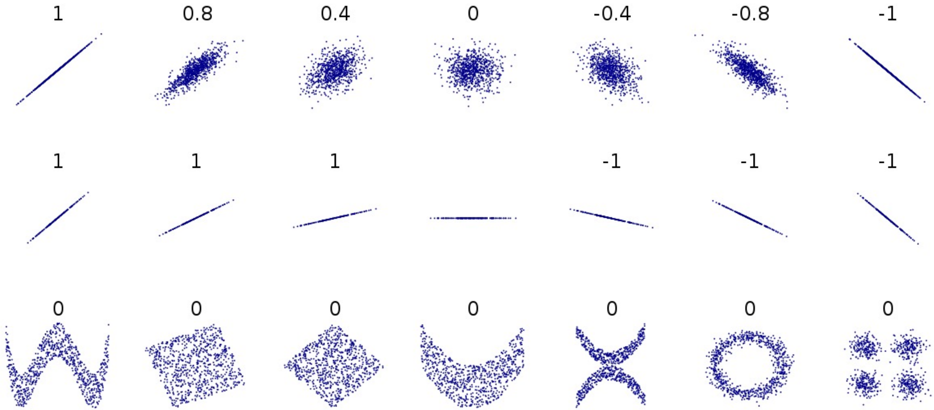
- Consider the following two statements:

a. X and Y are independent;

b. $\mathrm{Cov}(X, Y) = 0$.

⇨ It holds that (a) implies (b), but (b) does not imply (a). Thus, independence is a stronger condition than correlation.

Independence



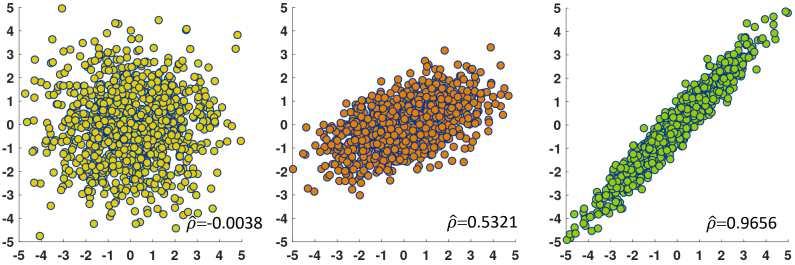
Ideal vs Empirical

Theory:

$$\rho = \frac{\mathbb{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}.$$

Practice:

$$\widehat{\rho} = \frac{\frac{1}{N}\sum_{n=1}^{N} x_n y_n - \overline{x}\,\overline{y}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \overline{y})^2}},$$



Probability & Statistics for DS & AI

Conditional Distribution

Michele Guindani

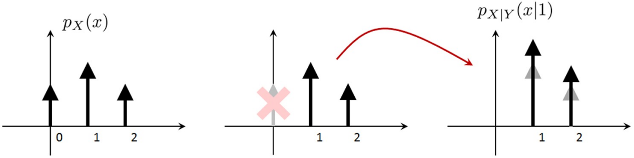
Jio Institute

Summer 2022

Conditional PMF

Let X and Y be two discrete random variables. The conditional PMF of X given Y is

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$



Suppose X is the sum of two coins with PMF $0.25, 0.5, 0.25$. Let Y be the first coin. When X is unconditioned, the PMF is just $[0.25, 0.5, 0.25]$. When X is conditioned on $Y = 1$, then "$X = 0$" cannot happen. Therefore, the resulting PMF $p_{X|Y}(x|1)$ only has two states. After normalization we obtain the conditional PMF $[0, 0.66, 0.33]$.

See examples 5.17; 5.7; 1.18 in your textbook

Conditional PDF

Let X and Y be two continuous random variables. The conditional PDF of X given Y is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Conditional Expectation

The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_x x \, p_{X|Y}(x \mid y)$$

for the discrete random variables, and

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \, f_{X|Y}(x \mid y) dx$$

- **What is conditional expectation?**
- $\mathbb{E}[X \mid Y = y]$ is the expectation using $f_{X|Y}(x \mid y)$.
- The integration is taken w.r.t. x, because $Y = y$ is given and fixed.

Law of total expectation

Example

- Suppose there are two classes of cars. Let X be the speed and C be the class.
- When $C = 1$, we know that $X \sim$ Gaussian (μ_1, σ_1). We know that $\mathbb{P}[C = 1] = p$.
- When $C = 2, X \sim$ Gaussian (μ_2, σ_2).
- Also, $\mathbb{P}[C = 2] = 1 - p$.
- Suppose you see a car on the freeway, what is its average speed?

Law of Total expectation

The problem has given us everything we need. In particular, we know the conditional PDFs of X, and the marginal pmf of C:

$f_{X|C}(x \mid 1) =$

$f_{X|C}(x \mid 2) =$

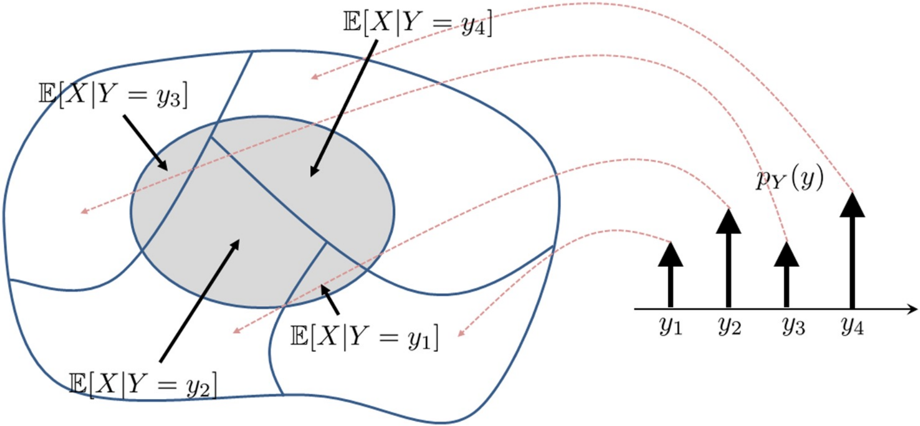
Conditioned on C, we have two expectations:

$\mathbb{E}[X \mid C = 1] =$

$\mathbb{E}[X \mid C = 2] =$

The overall expectation is:

Law of total expectation



Multivariate Gaussian

A d-dimensional joint Gaussian has a PDF

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

where d denotes the dimensionality of the vector $\boldsymbol{x}$.

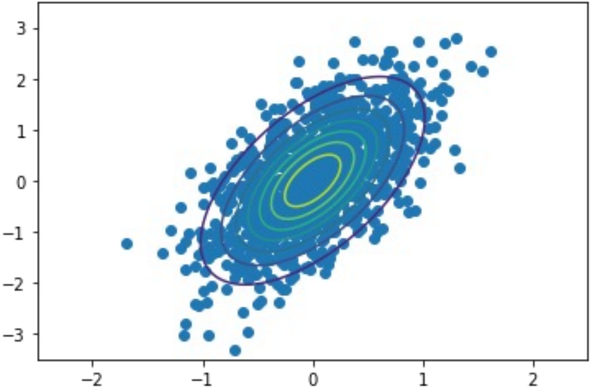
Multivariate Gaussian

- Random vector: $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix}$, and $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$
- Mean Vector:

$$\boldsymbol{\mu} \stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix}$$

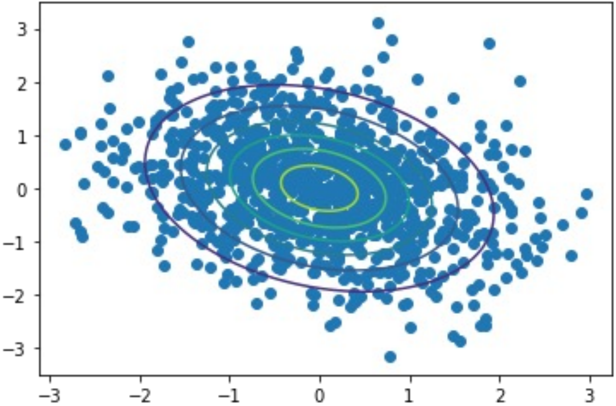
- Covariance:

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \text{Cov}(\boldsymbol{X}) = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_d) \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \dots & \text{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \text{Cov}(X_d, X_2) & \dots & \text{Var}[X_d] \end{bmatrix}$$

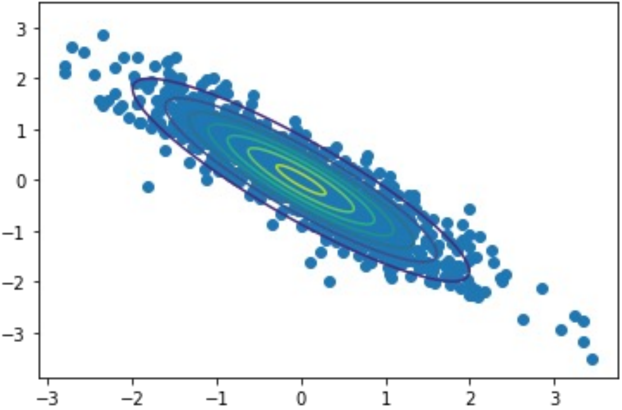


$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$

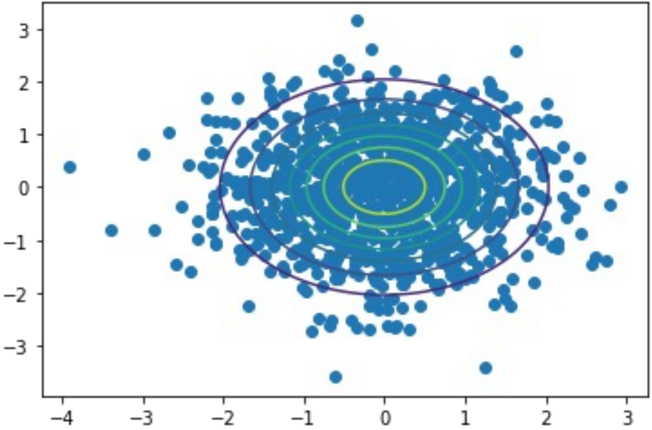
```python
# Python code: Overlay random numbers with the Gaussian contour.
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

X = stats.multivariate_normal.rvs([0,0],[[0.25,0.3],[0.3,1.0]],1000)
x1 = np.arange(-2.5, 2.5, 0.01)
x2 = np.arange(-3.5, 3.5, 0.01)
X1, X2 = np.meshgrid(x1,x2)
Xpos = np.empty(X1.shape + (2,))
Xpos[:,:,0] = X1
Xpos[:,:,1] = X2

F = stats.multivariate_normal.pdf(Xpos,[0,0],[[0.25,0.3],[0.3,1.0]])

plt.scatter(X[:,0],X[:,1])
plt.contour(x1,x2,F)
```



$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.25 \\ -0.25 & 1 \end{bmatrix}$$



$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$



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

Probability & Statistics for DS & AI

Estimation

Michele Guindani

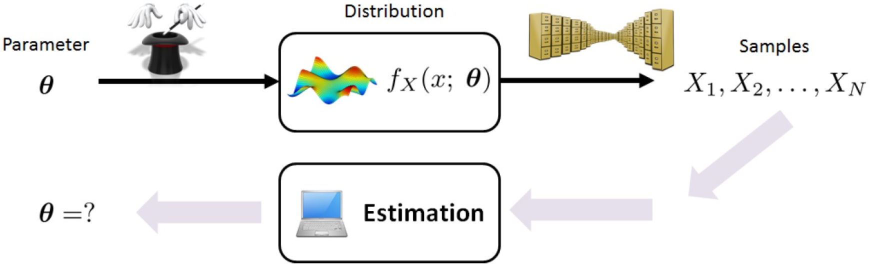
JioInstitute

Summer 2022

Estimation

- Estimation is an inverse problem with the goal of recovering the underlying parameter $\boldsymbol{\theta}$ of a distribution $f_X(x;\boldsymbol{\theta})$ based on the observed samples $X_1, \ldots, X_N$

8



Estimation is an inverse problem of recovering the unknown parameters that were used by the distribution. In this figure, the PDF of X using a parameter $\boldsymbol{\theta}$ is denoted as $f_X(x;\boldsymbol{\theta})$. The forward data-generation process takes the parameter $\boldsymbol{\theta}$ and creates the random samples $X_1, \ldots, X_N$. Estimation takes these observed random samples and recovers the underlying model parameter $\boldsymbol{\theta}$.

What are parameters?

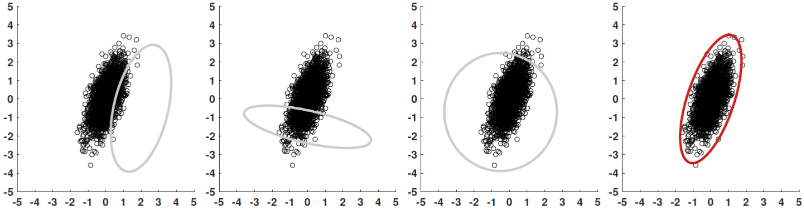
- All probability density functions (PDFs) have parameters.
- A Bernoulli random variable is characterized by a parameter p that defines the probability of obtaining a "head"
- A Gaussian random variable is characterized by two parameters: the mean μ and variance σ^2:

$$f_{X_n}(x_n; \underbrace{\boldsymbol{\theta}}_{=(\mu,\sigma)}) =| \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

If we know that $\sigma = 1$, then the PDF is

$$f_{X_n}(x_n; \underbrace{\theta}_{=\mu}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_n-\mu)^2}{2}\right\}$$

where θ is the mean



Bad estimate	Bad estimate	Bad estimate	Good estimate
$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -0.5 \end{bmatrix}$	$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ -1.5 \end{bmatrix}$	$\boldsymbol{\mu} = \begin{bmatrix} -0.5 \\ -0.7 \end{bmatrix}$	$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
$\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.2 \\ 0.2 & 1 \end{bmatrix}$	$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 0.1 \end{bmatrix}$	$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$

An estimation problem. Given a set of 1000 data points drawn from a Gaussian distribution with unknown mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we propose several candidate Gaussians and see which one would be the best fit to the data. Visually, we observe that the right-most Gaussian has the best fit. The goal of this chapter is to develop a systematic way of solving estimation problems of this type.

Estimation methods

- We will be looking at two estimation methods:

1. Maximum Likelihood methods
2. Maximum a posteriori method (Bayesian but used a lot in ML)

Probability & Statistics for DS & AI

Maximum Likelihood

Michele Guindani

Jio Institute

Summer 2022

Likelihood function

Consider a set of N data points $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$.

Since we have N data points, based on the problem at hand, we can postulate a data generating model:

$$X_1, \ldots, X_N \sim f(\boldsymbol{x}; \boldsymbol{\theta})$$

which means $\boldsymbol{x} = (x_1, \ldots, x_N)$, $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})$ = PDF of the random vector $\boldsymbol{X}$ with parameter $\boldsymbol{\theta}$.

When you express the joint PDF as a function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, you have two variables to play with:

- observation $\boldsymbol{x}$, given by the measured data (known)
- parameter $\boldsymbol{\theta}$ ⇨ our interest in an estimation problem.

- GOAL: find value of $\boldsymbol{\theta}$ that offers the "best explanation" to data $\boldsymbol{x}$

⇨ maximize the likelihood

Likelihood function

Let $\boldsymbol{X} = [X_1, \ldots, X_N]^T$ be a random vector drawn from a joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})$, and let $\boldsymbol{x} = [x_1, \ldots, x_N]^T$ be the realizations. The likelihood function is a function of the parameter $\boldsymbol{\theta}$ given the realizations $\boldsymbol{x}$:

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \stackrel{\text{def}}{=} f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})$$

- ⚠ $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ is not a conditional PDF because $\boldsymbol{\theta}$ is not a random variable.

  The correct way to interpret $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ is to view it as a function of $\boldsymbol{\theta}$.

  This function changes its shape according the observed data $\boldsymbol{x}$. We will return to this point shortly.

Independent observations

- If we measure the interarrival times of a bus for several days, it is quite likely the measurements are not correlated
- Assumption: the data points are independent and drawn from an identical distribution $f_X(x)$:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\boldsymbol{\theta}) = \prod_{n=1}^{N} f_{X_n}(x_n;\boldsymbol{\theta})$$

- so the likelihood is

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) \stackrel{\text{def}}{=} \prod_{n=1}^{N} f_{X_n}(x_n;\boldsymbol{\theta})$$

- and the log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \log f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{n=1}^{N} \log f_{X_n}(x_n;\boldsymbol{\theta})$$

Example (Bernoulli)

Find the log-likelihood of a sequence of i.i.d. Bernoulli random variables $X_1, \dots, X_N$ with parameter θ.

Solution: If $X_1, \dots, X_N$ are i.i.d. Bernoulli random variables, we have

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{n=1}^{N} \left\{ \theta^{x_n} (1-\theta)^{1-x_n} \right\}$$

Taking the log on both sides of the equation yields the log-likelihood function:

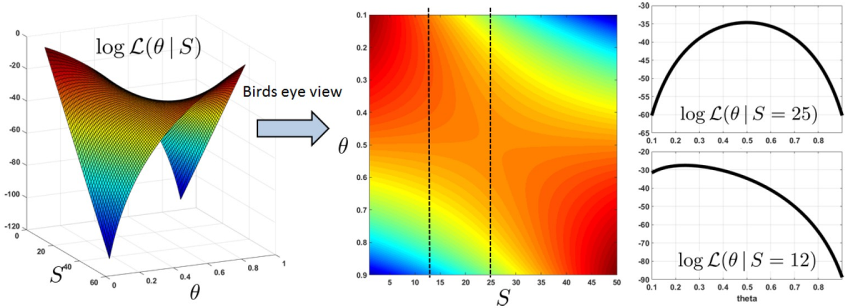
$$\begin{aligned}
\log \mathcal{L}(\theta \mid \boldsymbol{x}) &= \log \left\{ \prod_{n=1}^{N} \left\{ \theta^{x_n} (1-\theta)^{1-x_n} \right\} \right\} \\
&= \sum_{n=1}^{N} \log \left\{ \theta^{x_n} (1-\theta)^{1-x_n} \right\} \\
&= \sum_{n=1}^{N} x_n \log \theta + (1-x_n) \log(1-\theta) \\
&= \left(\sum_{n=1}^{N} x_n \right) \cdot \log \theta + \left(N - \sum_{n=1}^{N} x_n \right) \cdot \log(1-\theta)
\end{aligned}$$

- We can write

$$\log \mathcal{L}(\theta \mid \boldsymbol{x}) = \underbrace{\left(\sum_{n=1}^{N} x_n\right)}_{S} \cdot \log \theta + \underbrace{\left(N - \sum_{n=1}^{N} x_n\right)}_{N-S} \cdot \log(1-\theta)$$

That is: $\log \mathcal{L}(\theta \mid S) = S \log \theta + (N - S) \log(1-\theta)$.

- We plot the surface of $L(\theta \mid S)$ as a function of S and θ, assuming that $N = 50$



We plot the log-likelihood function as a function of $S = \sum_{n=1}^{N} x_n$ and θ. [Left] We show the surtace plot of $\mathcal{L}(\theta|S) = S \log \theta + (N - S) \log(1 - \theta)$. Note that the surface has a saddle shape. [Middle] By taking a bird's-eye view of the surface plot, we obtain a 2-dimensional contour plot of the surface, where the color code matches the height of the log-likelihood function. [Right] We take two cross sections along $S = 25$ and $S = 12$. Observe how the shape changes.

Example (Gaussian)

Find the log-likelihood of a sequence of i.i.d. Gaussian random variables $X_1, ., X_N$ with mean μ and variance σ^2

Solution Since the random variables $X_1, \ldots, X_N$ are i.i.d. Gaussian, the PDF is

$$f_{\boldsymbol{X}}\left(\boldsymbol{x}; \mu, \sigma^2\right) = \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \right\}$$

Taking the log on both sides yields the log-likelihood function:

$$\begin{aligned}
\log \mathcal{L}\left(\mu, \sigma^2 \mid \boldsymbol{x}\right) &= \log f_{\boldsymbol{X}}\left(\boldsymbol{x}; \mu, \sigma^2\right) \\
&= \log \left\{ \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \right\} \right\} \\
&= \sum_{n=1}^{N} \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \right\} \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{(x_n-\mu)^2}{2\sigma^2} \right\} \\
&= -\frac{N}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2
\end{aligned}$$

Maximum Likelihood estimate

Let $\mathcal{L}(\boldsymbol{\theta})$ be the likelihood function of the parameter $\boldsymbol{\theta}$ given the measurements $\boldsymbol{x} = [x_1, \ldots, x_N]^T$. The maximum-likelihood estimate of the parameter $\boldsymbol{\theta}$ is a parameter that maximizes the likelihood:

$$\widehat{\boldsymbol{\theta}}_{ML} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$$

Example (Bernoulli)

Find the ML estimate for a set of i.i.d. Bernoulli random variables $\{X_1, \ldots, X_N\}$ with $X_n \sim \text{Bernoulli}(\theta)$ for $n = 1, \ldots, N$

Solution. The log-likelihood function of a set of i.i.d. Bernoulli random variables is

$$\log \mathcal{L}(\theta \mid \boldsymbol{x}) = \left(\sum_{n=1}^{N} x_n\right) \cdot \log\theta + \left(N - \sum_{n=1}^{N} x_n\right) \cdot \log(1-\theta)$$

Thus, to find the ML estimate, we need to solve the optimization problem

$$\widehat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} \left\{ \left(\sum_{n=1}^{N} x_n\right) \cdot \log\theta + \left(N - \sum_{n=1}^{N} x_n\right) \cdot \log(1-\theta) \right\}$$

Taking the derivative with respect to θ and setting it to zero,

$$\frac{\left(\sum_{n=1}^{N} x_n\right)}{\theta} - \frac{N - \sum_{n=1}^{N} x_n}{1-\theta} = 0$$

Rearranging the terms yields

$$\widehat{\theta}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

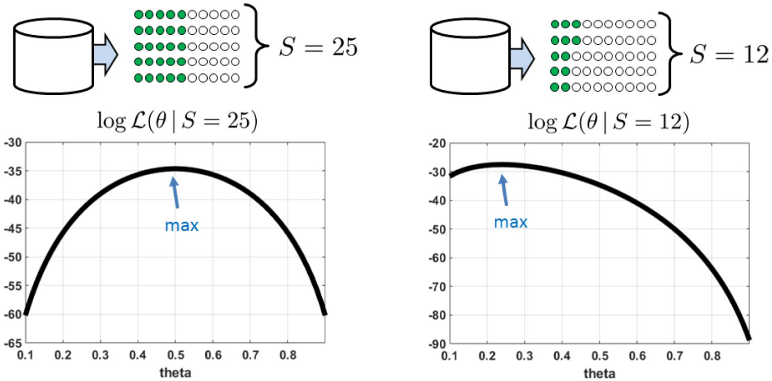
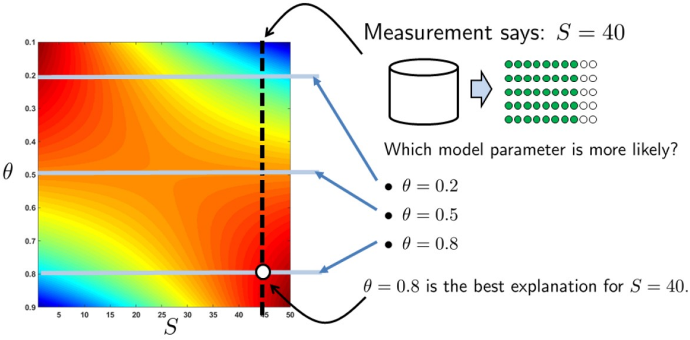


Illustration of how the maximum-likelihood estimate of a set of i.i.d. Bernoulli random variables is determined. The subfigures above show two particular scenarios at $S = 25$ and $S = 12$, assuming that $N = 50$. When $S = 25$, the likelihood function has a quadratic shape centered at $\theta = 0.5$. This point is also the peak of the likelihood function when $S = 25$. Therefore, the ML estimate is $\widehat{\theta}_{\mathrm{ML}} = 0.5$. The second case is when $S = 12$. The quadratic likelihood is shifted toward the left. The ML estimate is $\widehat{\theta}_{\mathrm{ML}} = 0.24$.



Suppose that we have a set of measurements such that $S = 40$. To determine the ML estimate, we look at the vertical cross section at $S = 40$. Among the different candidate parameters, e.g., $\theta = 0.2$, $\theta = 0.5$ and $\theta = 0.8$, we pick the one that has the maximum response to the likelihood function. For $S = 40$, it is more likely that the underlying parameter is $\theta = 0.8$ than $\theta = 0.2$ or $\theta = 0.5$.

Example (Social Network Analysis)

- **Recall:** The Erdos-Renyi graph is one of the simplest models for social networks. The Erdos-Renyi graph is a single-membership network that assumes that all users belong to the same cluster. Thus the connectivity between users is specified by a single parameter:

$$X_{ij} \sim \text{Bernoulli}(p)$$

- In other words, the edge X_{ij} linking user i and user j in the network is either $X_{ij} = 1$ with probability p, or $X_{ij} = 0$ with probability $1 - p$.
- The resulting matrix $\boldsymbol{X} \in \mathbb{R}^{N \times N}$ as the adjacency matrix, with the (i, j) th element being X_{ij}.