Probability and Distributions

CS115 - Math for Computer Science

TS. Lương Ngọc Hoàng
TS. Dương Việt Hằng

September 9, 2023

Roadmap

- Construction of a Probability Space
- Discrete and Continuous Probabilities
- Sum Rule, Product Rule, and Bayes' Theorem
- Summary Statistics and Independence
- Gaussian Distribution

Construction of a Probability Space

Modeling: Approximate reality with a simple (mathematical) model

- Experiment
  - Flip two coins
- Observation: a random outcome
  - for example, (H, H)
- All outcomes
  - $\{(H, H), (H, T), (T, H), (T, T)\}$

- Our goal: Build up a $\boxed{\text{probabilistic model}}$ for an experiment with random outcomes
- Probabilistic model?
  - Assign a number to each outcome or a set of outcomes
  - Mathematical description of an uncertain situation
- Which model is good or bad?

Why probabilities modeling?

Inferences from data are intrinsically uncertain.

Probability theory: Model uncertainty instead of ignoring it

Inferences or prediction can be done by using probabilities

Goal: Build up a probabilistic model

The first thing: What are **the elements** of a probabilistic model?

Elements of Probabilistic Model

1. All outcomes of my interest: $\boxed{\text{Sample Space } \Omega}$
2. Assigned numbers to each outcome of Ω: $\boxed{\text{Probability Law } \mathbb{P}(\cdot)}$

Sample Space Ω

The set of all outcomes of $\boxed{\text{my interest}}$

1. Mutually exclusive
2. Collectively exhaustive
3. At the $\boxed{\text{right granularity}}$ (not too concrete, not too abstract)

1. Toss a coin. What about this? $\Omega = \{H, T, HT\}$
2. Toss a coin. What about this? $\Omega = \{H\}$
3. (a) Just figuring out prob. of H or T. $\Longrightarrow \Omega = \{H, T\}$

   (b) The impact of the weather (rain or no rain) on the coin's behavior.

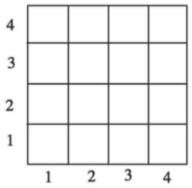
   $$\Longrightarrow \Omega = \{(H, R), (T, R), (H, NR), (T, NR)t\},$$

   where R(Rain), NR(No Rain).

Example: Sample Space Ω

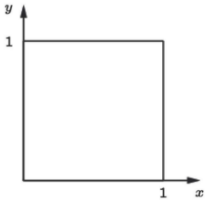
- *Discrete case:* Two rolls of a tetrahedral die

  - $\Omega = \{(1,1),(1,2),\ldots,(4,4)\}$



- *Continuous case:* Dropping a needle in a plain

  - $\Omega = \{(x,y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq 1\}$



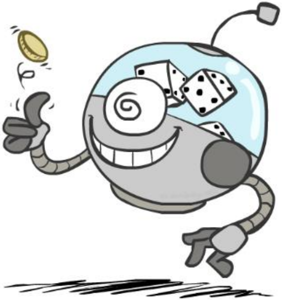
Probability Law

- Assign numbers to each subset of Ω: A subset of Ω: an event
- $\mathbb{P}(A)$: Probability of an event A.
  - This is where probability meets set theory.
  - Roll a dice. What is the probability of odd numbers?

    $\mathbb{P}(\{1,3,5\})$, where $\{1,3,5\} \subset \Omega$ is an event.
- Event space $\mathcal{A}$: The collection of subsets of Ω. For example, in the discrete case, the power set of Ω.
- Probability Space $(\Omega, \mathcal{A}, \mathbb{P}(\cdot))$

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?
- We denote random variables with capital letters
- Random variables have domains
  - R in {true, false}
  - T in {hot, cold}
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}



Conditional Probability

- Definition.

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \textit{for} \quad \mathbb{P}(B) > 0.$$

  - Note that this is a definition, not a theorem.

- All other properties of the law $\mathbb{P}(\cdot)$ is applied to the conditional law $\mathbb{P}(\cdot|B)$.

- For example, for two disjoint events A and C,

$$\mathbb{P}(A \cup C \mid B) = \mathbb{P}(A \mid B) + \mathbb{P}(C \mid B)$$

Discrete Random Variables

- The values that a random variable X takes is discrete (i.e., finite or countably infinite).
- Then, $p_X(x) := \mathbb{P}(X = x) := \mathbb{P}\Big(\{\omega \in \Omega \mid X(w) = x\}\Big)$, which we call probability mass function (PMF).
- Examples: Bernoulli, Uniform, Binomial, Poisson, Geometric

Bernoulli X with parameter p ∈ [0, 1]

- Only binary values

$$X = \begin{cases} 0, & \text{w.p.}^1 \quad 1-p, \\ 1, & \text{w.p.} \quad p \end{cases}$$

In other words, $p_X(0) = 1 - p$ and $p_X(1) = p$ from our PMF notation.

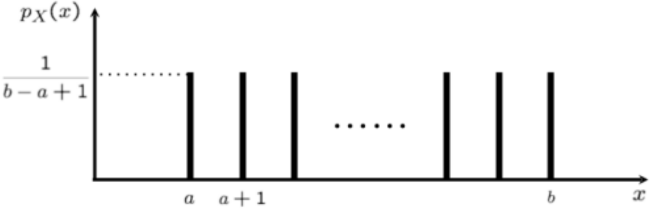
- Models a trial that results in binary results, e.g., success/failure, head/tail
- Very useful for an $\boxed{\text{indicator rv}}$ of an event A. Define a rv 1_A as:

$$1_A = \begin{cases} 1, & \text{if } A \text{ occurs}, \\ 0, & \text{otherwise} \end{cases}$$

[1] with probability

Uniform X with parameter a, b

- integers a, b, where $a \leq b$
- Choose a number of $\Omega = \{a, a+1, \ldots, b\}$ uniformly at random.
- $p_X(i) = \frac{1}{b-a+1}$, $i \in \Omega$.



- Models complete ignorance (I don't know anything about X)

Binomial X with parameter n, p

- Models the number of successes in a given number of independent trials
- n independent trials, where one trial has the success probability p.

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Poisson X with parameter λ

- $Binomial(n, p)$: Models the number of successes in a given number of independent trials with success probability p.
- Very large n and very small p, such that $np = \lambda$

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

- Is this a legitimate PMF?

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!}\ldots\right) = e^{-\lambda}e^{\lambda} = 1$$

- Prove this:

$$\lim_{n\to\infty} p_X(k) = \binom{n}{k}(1/n)^k(1 - 1/n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

Joint PMF

- $\boxed{\text{Joint PMF.}}$ For two random variables X, Y, consider two events $\{X = x\}$ and $\{Y = y\}$, and

$$p_{X,Y}(x,y) := \mathbb{P}\Big(\{X = x\} \cap \{Y = y\}\Big)$$

- $\sum_x \sum_y p_{X,Y}(x,y) = 1$

- $\boxed{\text{Marginal PMF.}}$

$$p_X(x) = \sum_y p_{X,Y}(x,y),$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Conditional PMF

- Conditional PMF

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

for y such that $p_Y(y) > 0$.

- $\sum_x p_{X|Y}(x|y) = 1$
- Multiplication rule.

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x|y) \\ &= p_X(x)p_{Y|X}(y|x) \end{aligned}$$

- $p_{X,Y,Z}(x,y,z) =$
$p_X(x)p_{Y|X}(y|x)p_{Z|X,Y}(z|x,y)$

Continuous RV and Probability Density Function (PDF)

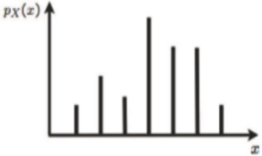
- Many cases when random variable have "continuous values", e.g., velocity of a car

Continuous Random Variable

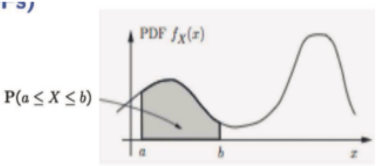
A rv X is continuous if $\exists$ a function f_X, called probability density function (PDF), s.t.

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx$$

- All of the concepts and methods (expectation, PMFs, and conditioning) for discrete rvs have continuous counterparts

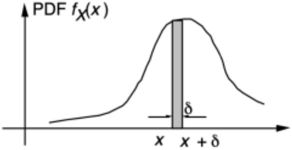


- $\mathbb{P}(a \leq X \leq b) = \sum_{x: a \leq x \leq b} p_X(x)$
- $p_X(x) \geq 0$, $\sum_x p_X(x) = 1$



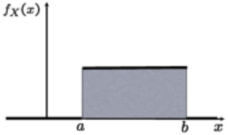
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$
- $f_X(x) \geq 0$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$

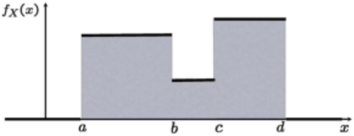
PDF and Examples



- $\mathbb{P}(a \le X \le a+\delta) \approx \boxed{f_X(a) \cdot \delta}$
- $\mathbb{P}(X = a) = 0$

Examples





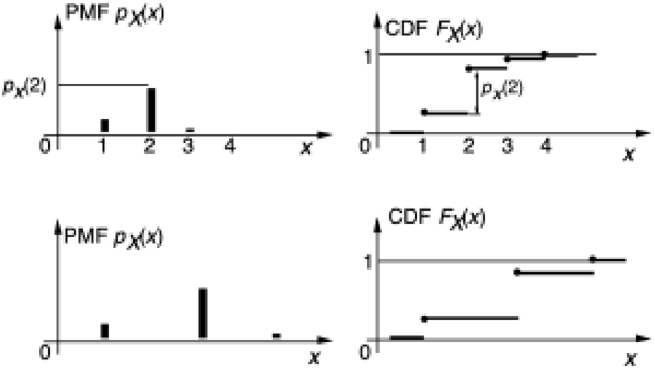
Cumulative Distribution Function (CDF)

- Discrete: PMF, Continuous: PDF
- Can we describe all rvs with a single mathematical concept?

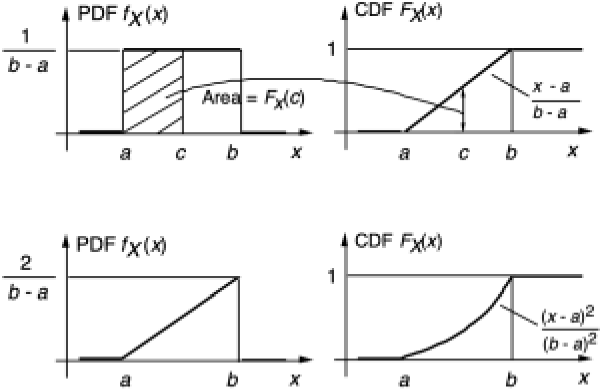
$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{discrete} \\ \int_{-\infty}^{x} f_X(t)dt, & \text{continuous} \end{cases}$$

- always well defined, because we can always compute the probability for the event $\{X \leq x\}$
- CCDF (Complementary CDF): $\mathbb{P}(X > x)$

Cumulative Distribution Function (CDF)



Cumulative Distribution Function (CDF)



CDF Properties

- Non-decreasing
- $F_X(x)$ tends to 1, as $x \to \infty$
- $F_X(x)$ tends to 0, as $x \to -\infty$

Continuous: Joint PDF and CDF

Jointly Continuous

Two continuous rvs are jointly continuous if a non-negative function $f_{X,Y}(x,y)$ (called joint PDF) satisfies: for every subset B of the two dimensional plane,

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)dxdy$$

1. The joint PDF is used to calculate probabilities

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)dxdy$$

Our particular interest: $B = \{(x,y) \mid a \le x \le b, c \le y \le d\}$

Continuous: Joint PDF and CDF

2. The marginal PDFs of X and Y are from the joint PDF as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

3. The joint CDF is defined by $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$, and determines the joint PDF as:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{x,y}}{\partial x \partial y}(x,y)$$

4. A function $g(X,Y)$ of X and Y defines a new random variable, and

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

Continuous: Conditional PDF given a RV

- $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$
- Similarly, for $f_Y(y) > 0$,

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

- Remember: For a fixed event A, $\mathbb{P}(\cdot|A)$ is a legitimate probability law.
- Similarly, For a fixed y, $f_{X|Y}(x|y)$ is a legitimate PDF, since

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx}{f_Y(y)} = 1$$

Sum Rule and Product Rule

- Sum Rule

$$p_X(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) & \text{if discrete} \\ \int_{y \in \mathcal{Y}} f_{X,Y}(x,y)dy & \text{if continuous} \end{cases}$$

  - Generally, for $X = (X_1, X_2, \ldots, X_D)$,

$$p_{X_i}(x_i) = \int p_X(x_1, \ldots, x_i, \ldots, x_D) d\mathbf{x}_{-i}$$

  - Computationally challenging, because of high-dimensional sums or integrals

Sum Rule and Product Rule

- Product Rule

$$p_{X,Y}(x,y) = p_X(x) \cdot p_{Y|X}(y|x)$$

joint dist. = marginal of the first × conditional dist. of the second given the first

  - Same as $p_Y(y) \cdot p_{X|Y}(x|y)$

Bayes Rule

- X: state/cause/original value $\rightarrow$ Y: result/resulting action/noisy measurement
- Model: $\mathbb{P}(X)$ (prior) and $\mathbb{P}(Y|X)$ (cause $\rightarrow$ result)
- Inference: $\mathbb{P}(X|Y)$?

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)p_{Y|X}(y|x) \\ &= p_Y(y)p_{X|Y}(x|y) \end{aligned}$$

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

$$p_Y(y) = \sum_{x'} p_X(x')p_{Y|X}(y|x')$$

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)f_{Y|X}(y|x) \\ &= f_Y(y)f_{X|Y}(x|y) \end{aligned}$$

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

$$f_Y(y) = \int f_X(x')f_{Y|X}(y|x')dx'$$

Bayes Rule

- X: state/cause/original value $\rightarrow$ Y: result/resulting action/noisy measurement
- Model: $\mathbb{P}(X)$ (prior) and $\mathbb{P}(Y|X)$ (cause $\rightarrow$ result)
- Inference: $\mathbb{P}(X|Y)$?

$$\underbrace{p_{X|Y}(x|y)}_{\text{posterior}} = \frac{\overbrace{p_{Y|X}(y|x)}^{\text{likelihood}}\,\overbrace{p_X(x)}^{\text{prior}}}{\underbrace{p_Y(y)}_{\text{evidence}}}$$

Bayes Rule for Mixed Case

K: discrete, Y: continuous

- Inference of K given Y

$$p_{K|Y}(k|y) = \frac{p_K(k) f_{Y|K}(y|k)}{f_Y(y)}$$

$$f_Y(y) = \sum_{k'} p_K(k') f_{Y|K}(y|k')$$

- Inference of Y given K

$$f_{Y|K}(y|k) = \frac{f_Y(y) p_{K|Y}(k|y)}{p_K(k)}$$

$$p_K(k) = \int f_Y(y') p_{K|Y}(k|y') dy'$$

Independence

- Occurrence of A provides no new information about B. Thus, knowledge about A does no change my belief about B.

$$\mathbb{P}(B|A) = \mathbb{P}(B)$$

- Using $\mathbb{P}(B|A) = \mathbb{P}(B \cap A)/\mathbb{P}(A)$,

Independence of A and B, $\boxed{A \perp\!\!\!\perp B}$

$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$

Conditional Independence

Conditional Independence of A and B given C, $\boxed{A \perp\!\!\!\perp B | C}$

$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \times \mathbb{P}(B|C)$

Independence for Random Variable

- Two rvs

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y), \quad \text{for all } x, y$$

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

$$\mathbb{P}(\{X = x\} \cap \{Y = y\} | C) = \mathbb{P}(X = x | C) \cdot \mathbb{P}(Y = y | C), \quad \text{for all } x, y$$

$$p_{X,Y|C}(x, y) = p_{X|C}(x) \cdot p_{Y|C}(y)$$

- Notation: $X \perp\!\!\!\perp Y$ (independence), $X \perp\!\!\!\perp Y | Z(conditionalindependence)$

Expectation/Variance

- Expectation

$$\mathbb{E}[X] = \sum_x x p_X(x), \quad \mathbb{E}[X] = \int_x x f_X(x) dx$$

- Variance, Standard deviation

- Measures how much the spread of PMF/PDF is

$$\text{var}[X] = \mathbb{E}[(X-\mu)^2]$$
$$\sigma_X = \sqrt{\text{var}[X]}$$

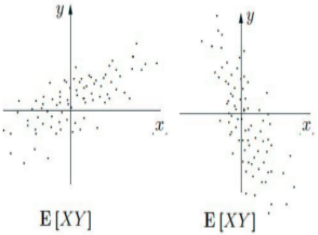
Properties

- $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $\text{var}[aX + b] = a^2 \text{var}[X]$
- $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$ if $X \perp\!\!\!\perp Y$ (generally not equal)

Covariance

- Goal: Given two rvs X and Y, quantify the degree of their dependence
  - Dependent: Positive (If $X \uparrow$, $Y \uparrow$) or Negative (If $X \uparrow$, $Y \downarrow$)
  - Simple case: $\mathbb{E}[X] = \mu_x = 0$ and $\mathbb{E}[Y] = \mu_Y = 0$

- What about $\mathbb{E}[XY]$? Seems good.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 0$ when $X \perp\!\!\!\perp Y$
- More data points (thus increases) when $xy > 0$ (both positive or negative)



Independence

Covariance

$$\mathrm{cov}(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\Big]$$

- After some algebra, $\mathrm{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $X \perp\!\!\!\perp Y \implies \mathrm{cov}(X, Y) = 0$
- $\mathrm{cov}(X, Y) = 0 \implies X \perp\!\!\!\perp Y$? NO.
- When $\mathrm{cov}(X, Y) = 0$, we say that X and Y are uncorrelated.

Properties

$$\text{cov}(X, X) = 0$$

$$\text{cov}(aX + b, Y) = \mathbb{E}[(aX + b)Y] - \mathbb{E}[aX + b]\mathbb{E}[Y] = a \cdot \text{cov}(X, Y)$$

$$\text{cov}(X, Y + Z) = \mathbb{E}[X(Y + Z)] - \mathbb{E}[X]\mathbb{E}[Y + Z] = \text{cov}(X, Y) + \text{cov}(X, Z)$$

$$\text{var}[X + Y] = \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 = \text{var}[X] + \text{var}[Y] - 2\text{cov}(X, Y)$$

Extension to Random Vectors X

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

Expectation, Covariance, Variance

- $\mathbb{E}(\boldsymbol{X}) := \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$
- Covariance of $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$

$$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}\boldsymbol{Y}^\mathsf{T}) - \mathbb{E}(\boldsymbol{X})\mathbb{E}(\boldsymbol{Y})^\mathsf{T} \in \mathbb{R}^{n \times m}$$

- Variance of $\boldsymbol{X}$: $\operatorname{var}(\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$, often denoted by $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ (or simply $\boldsymbol{\Sigma}$):

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} := \operatorname{var}[\boldsymbol{X}] = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots \operatorname{cov}(X_1, X_n) \\ \vdots & \vdots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots \operatorname{cov}(X_n, X_n) \end{pmatrix}$$

  - We call $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ covariance matrix of $\boldsymbol{X}$.

Data Matrix and Data Covariance Matrix

- N: number of samples, D: number of measurements (or original features)
- iid dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ whose mean is 0 (well-centered), where each $\mathbf{x}_i \in \mathbb{R}^D$, and its corresponding data matrix

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,N} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,N} \\ & & \vdots & \\ x_{D,1} & x_{D,2} & \ldots & x_{D,N} \end{pmatrix} \in \mathbb{R}^{D \times N}$$

Data Covariance Matrix

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top \in \mathbb{R}^{D \times D}$$

Relation between covariance matrix and data covariance matrix?

- Covaiance matrix for a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)^{\mathsf{T}}$,

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \begin{pmatrix} \mathrm{cov}(Y_1, Y_1) & \mathrm{cov}(Y_1, Y_2) & \cdots \mathrm{cov}(Y_1, Y_D) \\ \vdots & \vdots & \vdots \\ \mathrm{cov}(Y_D, Y_1) & \mathrm{cov}(Y_n, Y_2) & \cdots \mathrm{cov}(Y_D, Y_D) \end{pmatrix}$$

- Data convariance matrix $\boldsymbol{S} \in \mathbb{R}^{D \times D}$
  - Each Y_i has N samples $\begin{pmatrix} x_{i,1} & \cdots & x_{i,N} \end{pmatrix}$

$$\begin{aligned} \boldsymbol{S}_{ij} = \mathrm{cov}(Y_i, Y_j) &= \frac{1}{N} \sum_{k=1}^{N} x_{i,k} \cdot x_{j,k} \\ &= \text{average covariance (over samples) btwn feastures } i \text{ and } j \end{aligned}$$

Properties

For two random vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$,

- $\mathbb{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y}) \in \mathbb{R}^n$
- $\mathrm{var}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{var}(\boldsymbol{X}) + \mathrm{var}(\boldsymbol{Y}) \in \mathbb{R}^{n \times n}$
- Assume $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$.
  - $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{A}\mathbb{E}(\boldsymbol{X}) + \boldsymbol{b}$
  - $\mathrm{var}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\,\mathrm{var}(\boldsymbol{X})\boldsymbol{A}^\top$
  - $\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \Sigma_X \boldsymbol{A}^\top$ (Please prove)

Normal (also called Gaussian) Random Variable

- Standard Normal $\mathcal{N}(0,1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- $\mathbb{E}[X] = 0$
- $\text{var}[X] = 1$

- General Normal $\mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $\mathbb{E}[X] = \mu$
- $\text{var}[X] = \sigma^2$

Gaussian Random Vector

- $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ with the mean vector $\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$ and the covariance matrix $\boldsymbol{\Sigma}$.
- A Gaussian random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ has a joint pdf of the form:
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right),$$
where $\boldsymbol{\Sigma}$ is symmetric and positive definite.
- We write $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or $p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$.