Fall 2021 ME424 Modern Control and Estimation

Lecture Note 7: Kalman Filter - Probability Review

- Kalman filter: application
  - control system
  - signal processing
  - communications
  - networking
  - Computer vision (tracking)
  - robotics
  - civil

Prof. Wei Zhang
Department of Mechanical and Energy Engineering
SUSTech Institute of Robotics
Southern University of Science and Technology

zhangw3@sustech.edu.cn
https://www.wzhanglab.site/

Kalman Filer Preview:

- Given stochastic linear system described by

Gaussian

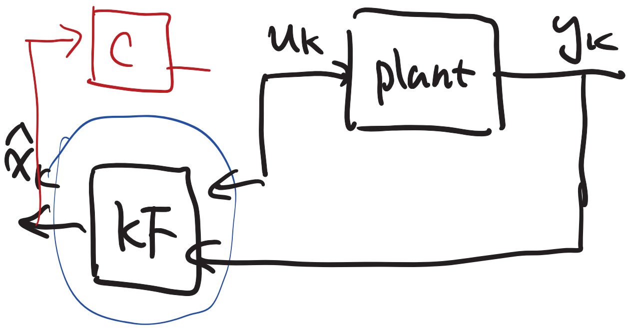
noise

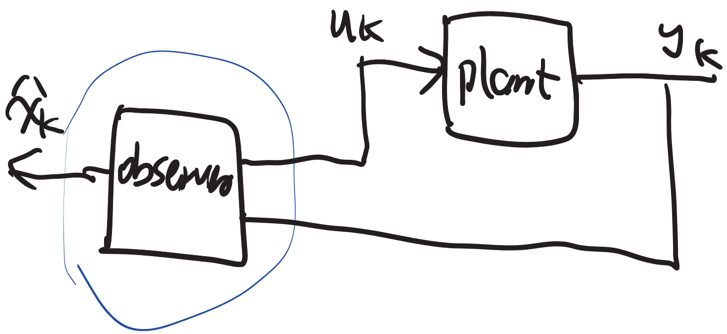
linear system corrupted by noise

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k + w_k \\ y_k = C_k x_k + D_k u_k + v_k \end{cases}$$

model no longer "deterministic": x_{next} is not known for sure, has some probabilistic distribution

- **Kalman filter**: compute the "best" estimate of x_k given input-output data history $\{u_j, y_j\}_{j=0}^{k}$





- **From Luenberger to Kalman:**

① **Deterministic to probabilistic model**

② **Stable observer to optimal observer/filter**

for observer gain L

$|\text{eig}(A-LC)|<1$

Kalman Filer Preview: Luenberger observer vs. Kalman filter

$\in \mathbb{R}^1$

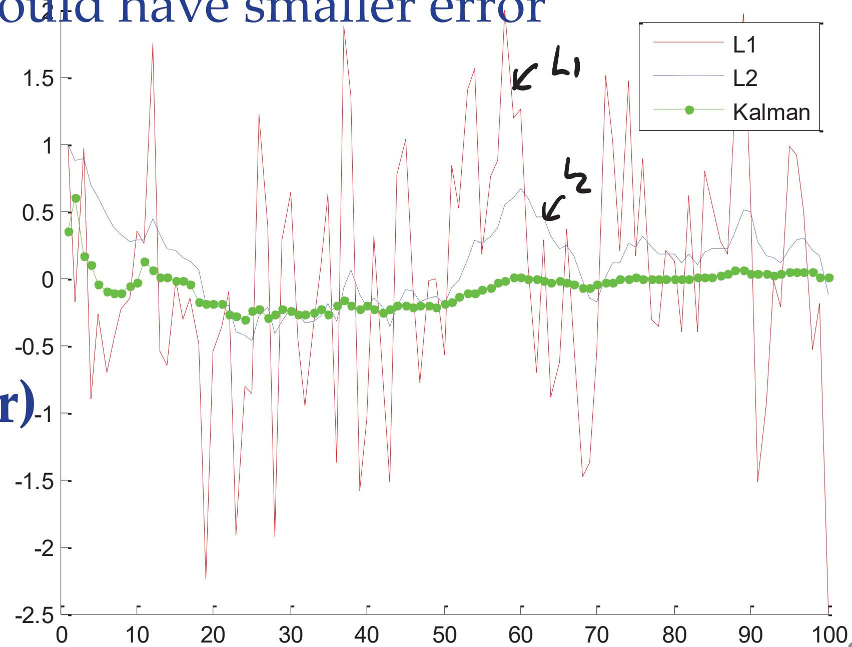
- Example: $x_{k+1} = x_k,\ y_k = x_k + v_k$, where v_k is white noise with
$$cov(v_k, v_m) = \begin{cases} 1, & \text{if } k = m \\ 0, & \text{otherise} \end{cases}$$

$A=1,\ B=0,\ C=1,\ D=0$

  - Ignoring noise, we have deterministic model $x_{k+1} = x_k,\ y_k = x_k,$
  - Luenberger type observer: $\hat{x}_{k+1} \overset{A}{=} \hat{x}_k + L(y_k - \hat{x}_k)$
  - Estimator error dynamics: $e(k+1) = (A - LC)e(k) = (1 - L)e(k)$
  - E.g.: $L_1 = 0.9$ and $L_2 = 0.1$, both provide stable error dynamics
  - According to deterministic model, L_1 should have smaller error

  - However, with noise, both L_1 and L_2 perform poorly, L_1 is worse than L_2

  - **The optimal observer (Kalman filter) is much better**



Kalman Filer Preview:

- Given stochastic linear system described by

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k + w_k \\ y_k = C_k x_k + D_k u_k + v_k \end{cases}$$

- **Kalman filter**: compute the "best" estimate of x_k given input-output data history $\{u_j, y_j\}_{j=0}^{k}$

code: — <200 lines

- **Kalman Filter Solution**: $\boxed{\hat{x}_k = \boldsymbol{E}(\boldsymbol{x}_k | \boldsymbol{y}_0, \boldsymbol{y}_1, \dots, \boldsymbol{y}_k)}$

KF is a recurssive way of computing conditional expectation

- **Our goal: in-depth understanding of the assumptions, derivations of Kalman filter**

— probability ** conditional prob / Expectation

— Minimum Mean Squared Estimation (MMSE)

Outline

- **Probability and Conditional Probability**
- Random Variables and Random Vectors
- Jointly Distributed Random Vectors and Conditional Expectation
- Covariance Matrix

What is probability?

- A formal way to quantify the uncertainty of our knowledge about the physical world

there is no right/wrong probability.

- Formalism: Probability Space $(\Omega, \mathcal{F}, P)$
  - Ω : **sampling space**: a set of all possible outcomes (maybe infinite)
  - $\mathscr{F}$: **event space**: collection of events of interest (event is a subset of Ω)
  - P: $\mathcal{F} \rightarrow [0,1]$ probability measure: assign event in $\mathscr{F}$ to a real number between 0 and 1

e.g. toss a coin, $\Omega = \{0, 1\}$, $\mathcal{F} = \{\phi, \{0\}, \{1\}, \{0,1\}\}$

toss a die. $\Omega = \{1, 2, \dots, 6\}$, $\mathcal{F} = \{\phi, \{1\}, \{2\}, \dots, \{6\}, \{1,2\}, \{1,3\} \dots \{1,2,3, \dots 6\}\}$

Event $A \in \mathcal{F}$, e.g: $A = \{2,4,6\}$ even

$\text{prob}(A) = \frac{1}{2}$ ✓

$\frac{1}{3}$ (ϕ could be right) up to the modeler

Axioms of probability:

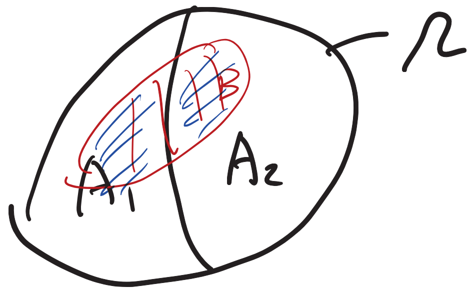
- $P(A) \geq 0$
- $P(\Omega) = 1$
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ disjoint

- **Important consequences:**
  - $P(\emptyset) = 0$ $\Leftarrow 1 = P(\Omega) = P(\Omega \cup \phi) = P(\Omega) + \underbrace{P(\phi)}_{\geq 0} = 1 \Rightarrow P(\phi) = 0$
  - $\Leftrightarrow$ Law of total probability: $P(B) = \sum_i^n P(B \cap A_i)$, for any partitions $\{A_i\}$ of Ω
    - Recall a collection of sets $A_1, \ldots, A_n$ is called a partition of Ω if
      - $A_i \cap A_j = \emptyset$, for all $i \neq j$ (mutually exclusive)
      - $A_1 \cup A_2 \cdots \cup A_n = \Omega$

divide & conquer

e.g.

$P(B) = P(B \cap \underbrace{(A_1 \cup A_2)}_{\Omega})$ $\leftarrow$ $P(B)$ $= P(B \cap A_1) + P(B \cap A_2)$



Conditional probability

- Probability of event A happens given that event B has already occurred

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- We assume $P(B) > 0$ in the above definition

Given origin

- **What does it mean?**
  - Conditional probability is a probability: $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$
  - **"Conditional" means, $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ the is derived from an original probability space $(\Omega, \mathcal{F}, P)$ given some event has occurred**
  - After B occurred we are uncertain only about the outcomes inside B

start with $(\Omega, \mathcal{F}, P)$

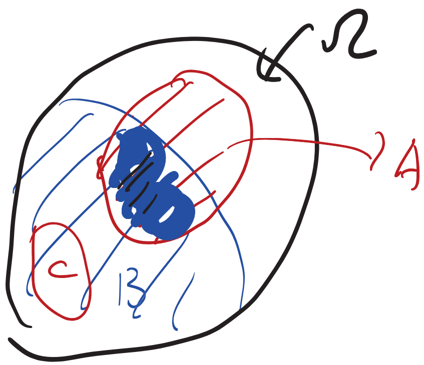
e.g. $\Omega = \{1, 2, \dots 6\}$

$B = \{2, 4, 6\}$, $A = \{1, 2, 3\}$

$\Longrightarrow$ event B occurred $\Longrightarrow$ new space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$

· now all possible outcomes is B i.e. $\tilde{\Omega} = B$

$\tilde{\mathcal{F}}$ changes (e.g. $A \in \mathcal{F}$, now $A \notin \tilde{\mathcal{F}}$, $A \cap B \in \tilde{\mathcal{F}}$)



$\tilde{\mathcal{F}}$: contains all subsets of $\tilde{\Omega}$

- $\tilde{P}$? - we want to be consistent with the original P

e.g. suppose $C \subseteq B^{\tilde{\Omega}}$, $C \in \tilde{\mathcal{F}}$

$$\tilde{P}(C) = P(C) \quad \times \quad \Rightarrow \text{ then } \tilde{P}(\tilde{\Omega}) = P(B) < 1$$

a possible way

$$= \frac{P(C)}{P(B)} \quad \Rightarrow \quad \tilde{P}(\tilde{\Omega}) = \frac{P(B)}{P(B)} = 1$$

if $A \nsubseteq B$, then $\tilde{P}(A) = \dfrac{P(A \cap B)}{P(B)}$

$$\triangleq P(A \mid B)$$

- Bayes rule: relate $P(A \mid B)$ to $P(B|A)$

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(B|A) \cdot P(A)$$

$$= \frac{P(A|B) \cdot P(B)}{P(A)}$$

- Events A and B are called (statistically) independent if
  - $P(A|B) = P(A)$
  - Or equivalently: $P(A \cap B) = P(A)P(B)$

$A \perp B$

$$P(A \cap B) = P(A|B) \cdot P(B)$$

$$= P(A) \cdot P(B)$$

- **Example of conditional probability**: A bowl contains 10 chips of equal size: 5 red, 3 white, and 2 blue. We draw a chip at random and define the event:

$$A = \text{the draw of a red or a blue chip}$$

Suppose you are told the chip drawn is not blue, what is the new probability of A

$(\Omega, \mathcal{F}, P)$: sampling space $\Omega = \{r_1, r_2 \ldots r_5, w_1, w_2, w_3, b_1, b_2\}$

$A = \{r_1, r_2 \ldots r_5, b_1, b_2\}$, $B = \{r_1\ r_2 \ldots r_5, w_1, w_2, w_3\}$

$P(A) = \frac{7}{10}$, $P(B) = \frac{4}{5}$

$$\tilde{P}(A) = P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{5}{10}}{\frac{4}{5}} = \frac{5}{8}$$

Outline

- Probability and Conditional Probability
- **Random Variables and Random Vectors**
- Jointly Distributed Random Vectors and Conditional Expectation
- Covariance Matrix

- **What is random variable and random vector?**
  - Deterministic variable/ constant
    - e.g. z is deterministic variable
      - $\underbrace{z}_{\text{can only}}$ takes one $\underbrace{\text{value}}_{\text{possible}}$ (single-value variable) which may or may not be known
  - Random variable:
    - ~~r.v.~~ random variable/vector is multi-valued variable
      - it takes multiple (or even infinite) possible values each occurs with certain probability.
    - e.g. - $X = \{1, 2\}$ with probabilities $\frac{1}{2}$, $\frac{1}{2}$; $X = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}$ with probabilities 0.3, 0.7
    - continuouse r.v. $X \in [0, 1]$

How to specify probability measure

- Discrete random variable: probability mass function (pmf)

  e.g. toss a coin or die

  $\Omega = \{1, 2, 3, 4, 5, 6\}$ pmf: $p(i) = \text{Prob}(i\text{th outcome})$

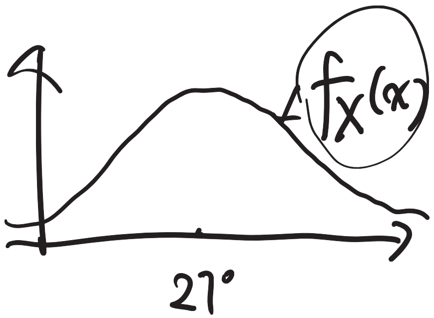
  $p(1)$ $p(2)$

  $\mathcal{F} = \{\phi, \{1\}, \dots \{6\}, \{1,2\}, \{1,3\}, \dots \{1,2\dots 6\}\}$

  For any $A \in \mathcal{F}$, its measure: $P(A) = \sum_{i \in A} p(i)$

- Continuous random variable: probability density function (pdf)

  e.g. temperature density



$X \in \mathbb{R}$, $\Omega = \mathbb{R}$, $A = [26, 26.5]$

we need pdf $f_X(x) \approx$ "prob" of $X = x$

$P(A) = \int_{x \in A} f_X(x)\,dx$

How to specify probability measure

- Random vector: scalar random variables listed according to certain order

- n-dimensional random vector: $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$

- Notation: We typically use capital to denote random variables (vectors) and lower case letter to denote specific values the random variable takes

- density function: $f(x)$, $x \in \mathrm{R}^n$. $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$f(x)$ is short hand notation for $f(x_1, x_2 \ldots x_n)$

- probability evaluation: $P(X \in A) = \int_A f(x)dx$

$P(A) = \int_A f(x) dx$

$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

$A = [1\ 2] \times [2\ 3]$

$\int_1^2 \int_2^3 f(x_1, x_2) dx_2 dx_1$

$X \in \mathbb{R}^{n \times m}$

Expectation of a random vector $X \in R^n$:

mean

$\mathbb{R}^n$ scalar

Continuous random vector: $E(X) \triangleq \int_{R^n} x f(x)\,dx$;

Discrete random vector: $E(X) \triangleq \sum_x x \cdot Prob(X = x)$.

- Expectation: $E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}$

$X \in \mathbb{R}^n$

e.g. $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $E(X) = \int_{\mathbb{R}^2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} f_X(x_1, x_2)\, dx_1\, dx_2$

V.F.Y. $= \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix}$

- Examples: Let $X \in R^2$ be discrete random variable with $Prob\left(X = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \frac{1}{2}$, $Prob\left(X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \frac{1}{3}$, $Prob\left(X = \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = \frac{1}{6}$. Compute $E(X)$

① $E(X) = \frac{1}{2}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ \frac{4}{3} \end{bmatrix}$

If we look at X_2 :

$P(X_2) \backslash X_2$	1	2
	$\frac{2}{3}$	$\frac{1}{3}$

$\Rightarrow E(X_2) = \frac{2}{3} 1 + \frac{1}{3} 2 = \frac{4}{3}$

Linearity of Expectation:

- Expectation of AX with deterministic constant $A \in R^{m\times n}$ matrix:

$$\boldsymbol{E(AX) = AE(X)}$$

if A is determinist .

$$E(AX) = \int (A\cdot x)\cdot f_X(x)\, dx = A \cdot \int x f_X(x) dx = A\cdot E(X)$$

- More generally, $E(AX + BY) = AE(X) + BE(Y)$

if A, B are deterministic matrices

- Example: Suppose $X \in R^2, Y \in R^3$, with $E(X) = \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix}$, , $E(Y) = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$,

$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, Compute $E(AX + BY)$

$$E(AX+BY) = AE(X) + BE(Y) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.85 \\ 0.55 \end{bmatrix}$$

Outline

- Probability and Conditional Probability
- Random Variables and Random Vectors
- **Jointly Distributed Random Vectors and Conditional Expectation**
- Covariance Matrix

Jointly distributed random vectors: $X \in R^n, Y \in R^m$

- Completely determined by joint density (mass) function: $(X, Y) \sim f_{XY}(x, y)$ $\approx$ "prob" of $X=x$, $Y=y$ $\Longleftrightarrow$ $Z=\begin{bmatrix} X \\ Y \end{bmatrix}$, $f_Z(z)$ $\downarrow$ (x, y)

  Compute probability:

  $$P((X,Y)\in A) = \int_A f_{XY}(x,y)\,dx\,dy$$

- marginal density: $X \sim f_X(x), Y \sim f_Y(y)$, where

  $$f_X(x) = \int_{R^m} f_{XY}(x, y) dy, \qquad f_Y(y) = \int_{R^n} f_{XY}(x, y) dx,$$

  $$\text{“}=\text{”} \sum_{\text{all possible } y} \text{prob}(X=x, Y=y)$$

- Example: $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $Prob\left(X = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \frac{1}{2}$, $Prob\left(X = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \frac{1}{3}$, $Prob\left(X = \begin{bmatrix} -1 \\ 1 \end{bmatrix}\right) = \frac{1}{6}$
  - This is joint distribution for X_1, X_2

marginal for X_2:

	1	2
	$\frac{2}{3}$	$\frac{1}{3}$

- The conditional "density": $(X, Y) \sim f_{XY}(x, y)$
  - Quantify how the observation of a value of Y, $Y = y$, affects your belief about the density of X
  - The conditional probability definition implies (nontrivially)

$$P(A \mid B)=P(A \cap B)/P(B) \Rightarrow p_{X|Y}(X=i|Y=j) = \frac{p_{XY}(X=i, Y=j)}{\sum_i p_{xy}(X=i, Y=j)}$$

$\sum_i p_{xy}(X=i, Y=j) \Rightarrow \text{Prob}(Y=j)$

marginal $f_X(x)$

$$① \longleftarrow f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

density of x

X: shenzhen temperature

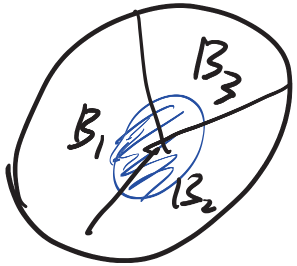
Y: guangzhou ... = 26°C

- Law of total probability: $P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

$$f_X(x) = \int_{R^m} f_{X|Y}(x|y) f_Y(y) dy$$

$$f_Y(y) = \int_{R^n} f_{Y|X}(y|x) f_X(x) dx$$

$$f_X(x) \overset{?}{=} \int f_{XY}(x,y)\, dy \overset{①}{=} \int f_{X|Y}(x|y) \cdot f_Y(y)\, dy$$



- ***X* is independent of *Y***, denoted by $X \perp Y$,

if and only if $f_{XY}(x, y) = f_X(x) f_Y(y)$

$$f_{X|Y}(x|y) = f_X(x) \quad ①$$

- **Conditional expectation:**
  - The conditional mean of $X|Y=y$ is

$$E(X|Y=y) \overset{\Delta}{=} \int_{R^n} x f_{X|Y}(x|y)dx$$

$$E(X|Y=y) = \sum_i i \cdot Prob(X=i|Y=y)$$

$(X,Y) \sim$

		X				
		2	3	4	5	6
	→1	1/4	1/8	1/8		
Y	2		1/6	1/12	1/12	
	3			1/12	1/24	1/24

- Example 1:
  - E(X | Y = 1)

Define $Z = (X|Y=1)$

$X|Y=1 \Leftrightarrow Z$	2	3	4
prob	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

$prob(Z=2) = Prob(X=2|Y=1)$

$= \frac{Prob(X=2, Y=1)}{Prob(Y=1)}$

$= \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{8}+\frac{1}{8}} = \frac{1}{2}$

$E(X|Y=1) = E(Z) = 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4}$

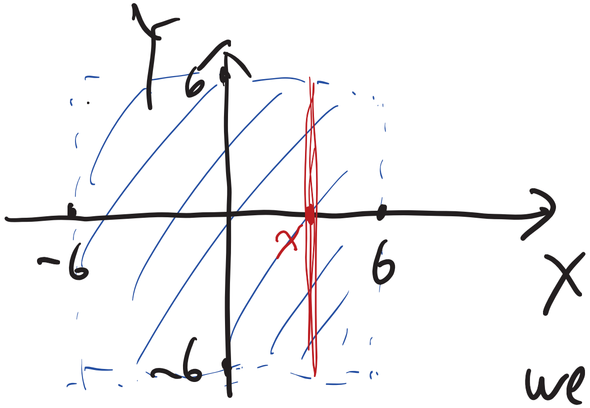
$= \frac{11}{4}$

  - E(X | Y = 2)

similar

- E(X|Y=3) → similar

- **Example 2**: Suppose that (X, Y) is uniformly distributed on the square $S = \{(x, y): -6 \leqslant x \leqslant 6, -6 \leqslant y \leqslant 6\}$. Find $E(Y|X = x)$.



we know $f_{XY}(x,y) = \begin{cases} \frac{1}{12\times 12} = \frac{1}{144}, & \text{if } (x,y)\in S \\ 0, & \text{otherwise} \end{cases}$

we need to compute $\boxed{E(Y|X=x) = \int_{-\infty}^{+\infty} y\, f_{Y|X}(y|x)\, dy}$

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

$$f_X(x) = \int \underbrace{f_{XY}(x,y)}\, dy = \begin{cases} 0, & \text{if } x \notin [-6,6] \\ \int_{-6}^{6} \frac{1}{144}\, dy & x \in [-6,6] \end{cases} \Rightarrow f_X(x) = \begin{cases} \frac{1}{12}, & x\in[-6,6] \\ 0, & \text{otherwise} \end{cases}$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{\frac{1}{144}}{\frac{1}{12}} = \frac{1}{12}, & \text{if } (x,y)\in S \\ 0 & \text{otherwise} \end{cases} \Rightarrow E(Y|X=x) = \int_{-\infty}^{+\infty} y\, f_{Y|X}(y|x)\, dy = 0$$

- Law of total probability implies:
  - $E(X) = \sum_y E(X|Y = y) \cdot p_Y(Y = y)$

divide & conquer

$$E(X) = \int x \cdot f_X(x)\,dx = \int x \cdot \left(\int f(x,y)\,dy\right) \cdot dx = \iint x \cdot \underbrace{f_{X|Y}(x|y) \cdot f_Y(y)}\,dy\,dx$$

  - $E(g(X,Y)) = \sum_y E(g(X,Y)|Y = y) \cdot p_Y(Y = y)$

$$= \int \Big(\underbrace{\int x f_{X|Y}(x|y)\,dx}_{E(X|Y=y)} \Big) \cdot f_Y(y)\,dy$$

any function of X, Y e.g. $g(X,Y) = X^2 + e^Y$

- Continue Example 1:

		X				
		2	3	4	5	6
Y	1	1/4	1/8	1/8		
	2	0	1/6	1/12	1/12	
	3	0		1/12	1/24	1/24

- compute $E(X)$

Method 1°: compute marginal of X

X	2	3	4	5	6
prob(·)	$\frac{1}{4}$	$\frac{7}{24}$	$\frac{7}{24}$	$\frac{3}{24}$	$\frac{1}{24}$

$\Rightarrow E(X) = 2\cdot\frac{1}{4} + 3\cdot\frac{7}{24} + \cdots = $ same # 1

Method 2°: $E(X) = \underbrace{E(X|Y=1)\cdot prob(Y=1)} + \underbrace{E(X|Y=2)\cdot prob(Y=2)} + \underbrace{E(X|Y=3)\cdot}_{prob(Y=3)}$

$= \frac{11}{4}\cdot\frac{1}{2}$ + please finish by yourself $+ \cdots$

$=$ same # 2

same

(X, Y)

- Example 3.: outcomes with equal chance: (1,1), (2, 0), (2,1), (1,0), (1,-1), (0,0), with $g(X, Y) = X^2Y^2$. Compute $E(X^2Y^2)$

Method 1: $E(g(X,Y)) = E(X^2Y^2) = 1^2 \cdot (-1)^2 \cdot \frac{1}{6} + 1^2 \cdot 1^2 \cdot \frac{1}{6} + 2^2 \cdot 1^2 \cdot \frac{1}{6} = 1$

Method 2: conditioning on values of $Y = -1, 0, 1$

$$E(X^2Y^2) = 1^2 1^2 \cdot \frac{1}{6} + 2^2 \cdot 0^2 \cdot \frac{1}{6} + 2^2 \cdot 1^2 \cdot \frac{1}{6} + \cdots = 1$$

Let's condition on Y,

$Y=0 \Rightarrow X^2Y^2 = 0$

$Y=-1 \Rightarrow X^2Y^2 = 1$

$Y=1 \Rightarrow X^2Y^2 = \begin{cases} 1 \\ 4 \end{cases}$

$\text{Prob}(X=1 \mid Y=1) = \frac{1}{2}$

$\text{Prob}(X=2 \mid Y=1) = \frac{1}{2}$

$E(X^2Y^2 \mid Y=0) = 0 \cdot 1 = 0$

$E(X^2Y \mid Y=-1) = 1 \cdot 1 = 1$

$E(X^2Y^2 \mid Y=1) = 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{2} = \frac{5}{2}$

		X	
Y	0	1	2
-1	0	1/6	0
0	1/6	1/6	1/6
1	0	1/6	1/6

$E(X^2Y^2) = E(X^2Y^2 \mid Y=0) \cdot \text{Prob}(Y=0) + E(X^2Y^2 \mid Y=-1) \cdot P(Y=-1) + E(X^2Y^2 \mid Y=1)\, P(Y=1)$

$= 0 + 1 \cdot \frac{1}{6} + \frac{5}{2} \cdot \frac{1}{3} = 1$

Outline

- Probability and Conditional Probability
- Random Variables and Random Vectors
- Jointly Distributed Random Vectors and Conditional Expectation
- **Covariance Matrix**

- Covariance (Random variable case):
  - $Cov(X,Y) \triangleq E\left(\left(X-\underbrace{E(X)}_{\#}\right)\left(Y-\underbrace{E(Y)}_{\#}\right)\right)$

Example:

		X	
		1	2
Y	-1	0.25	0.1
	1	0	0.65

$cov(X,Y) = E((X-E(X))(Y-E(Y)))$

$E(X) = 1\times0.25 + 2\times0.75 = 1.75$

$E(Y) = -1\times0.35 + 1\times0.65 = 0.3$

$cov(X,Y) = (1-1.75)(-1-0.3)\times0.25 + (2-1.75)(-1-0.3)\times0.1$
$+ \cdots + \cdots = \text{some } \#$

Sample data example.

Temp: °c	Shenzhen (X_i)	$X_i - \bar{X}$	Guangzhou (Y_i)	$Y_i - \bar{Y}$	Sydney (Z_i)	$Z_i - \bar{Z}$
	20	−5	19	−5.833	27	3.833
	21	−4	20	−4.833	26	2.833
	21	−4	22	−2.833	25	1.833
	27	2	26	1.1667	23	−1.1667
	30	5	30	5.1667	20	−3.1667
	31	6	32	7.1667	18	−5.1667
mean:	$\bar{X}=25$		$\bar{Y}=24.833$		$\bar{Z}=23.166$	

- **Covariance (Random variable case):**
  - If $Cov(X,Y) > 0$, X and Y are positively correlated
    - If you see a realization of X larger than $E(X)$, it is more likely for Y to be also larger than $E(Y)$
  - If $Cov(X,Y) < 0$, X and Y are negatively correlated
    - If you see a realization of X larger than $E(X)$, it is more likely for Y to be smaller than $E(Y)$
  - If $Cov(X,Y) = 0$, X and Y are uncorrelated

- **Covariance Matrix: $X \in R^n$, $Y \in R^m$**

$$Cov(X,Y) \triangleq E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)^T\right)$$

$$E\left(\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} - \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}\right)\left(\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} - \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_m) \end{bmatrix}\right)^T\right)$$

$(n\times 1)\cdot(1\times m) \to$ $n\times m$ matrix

- It is a $n \times m$ matrix: with $(Cov(X,Y))_{ij} = \mathrm{Cov}(X_i, Y_j) = \mathrm{E}\left((X_i - E(X_i))\left(Y_j - E(Y_j)\right)\right)$

$$cov(X,Y) = \begin{bmatrix} cov(X_1,Y_1) & cov(X_1,Y_2) & \cdots & cov(X_1,Y_m) \\ cov(X_2,Y_1) & cov(X_2,Y_2) & \cdots & cov(X_2,Y_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n,Y_1) & cov(X_n,Y_2) & \cdots & cov(X_n,Y_m) \end{bmatrix}$$

$X+a \quad -E(X)-a$

$$E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)^T\right)$$

$Cov(X,X)$

$$= E\left(\left(X - E(X)\right)\left(X - E(X)\right)^T\right)$$

- **Properties of Covariance** (V.F.Y.)

var

1. $Cov(X + a, Y + b) = Cov(X, Y)$

   deterministic constant

2. $Cov(X, Y) = Cov(Y, X)^T$

3. $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$

   $\int a f_X(x)\,dx = a$

4. $Cov(AX, BY) = ACov(X, Y)B^T$

   $\Downarrow \; E((AX - E(AX))(BY - E(BY))^T)$

   $= E(A(X - E(X))(Y - E(Y))^T B^T)$

5. If $X \perp Y$, $Cov(X, Y) = 0$

6. $\boxed{Cov(X) \triangleq Cov(X, X)}$ is positive semidefinite (p.s.d.)

   $Var(X)$, $\in \mathbb{R}^{n \times n}$

Assume $E(X) = a \in \mathbb{R}$, $E(Y) = b \in \mathbb{R}$

$$E((X-a)(Y-b)) = \iint (x-a)\cdot(y-b) f_{XY}(x,y)\,dx\,dy = \int (x-a) f_X(x)\,dx \cdot \int (y-b) f_Y(y)\,dy$$

$f_X(x) f_Y(y)$, 0

$= (\int x f_X(x)\,dx) - a$

$X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$

- **Example**: Suppose you know $cov(X, Y) = \Sigma_{XY}$ $\in \mathbb{R}^{n\times m}$, $\underbrace{cov(X)}$ $= \underbrace{\Sigma_X}_{n\times n}$, $cov(Y) = \underbrace{\Sigma_Y}_{m\times m}$, what is $Cov(AX + BY)$?

deterministic constant matrices

$(a+b)(a+b)$

$$Cov(AX+BY) \triangleq cov(AX+BY, AX+BY)$$

$$\overset{③}{=} cov(AX, AX) + cov(AX, BY) + cov(BY, AX) + cov(BY, BY)$$

$$\overset{④}{=} A\,cov(X,X)A^T + A\,cov(X,Y)B^T + B\,cov(Y,X)A^T + B\,cov(Y,Y)B^T$$

$$= A\Sigma_X A^T + A\Sigma_{XY}B^T + B\Sigma_{XY}^T A^T + B\Sigma_Y B^T$$

$Z \in \mathbb{R}^3$. random vector

- **Example:** Given that $E(Z) = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, and $Cov(Z,Z) = \underbrace{\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 8 \end{bmatrix}}_{\Sigma_Z}$. Let $P = \begin{bmatrix} Z_2 \\ Z_1 \end{bmatrix}$, $Q = Z_3$

  $cov(Z_1, Z_2)$

  Compute: $Cov(P,Q),\ Cov(Q,2P)$

$$cov(P,Q) = cov\left(\begin{bmatrix} Z_2 \\ Z_1 \end{bmatrix}, Z_3\right) = \begin{bmatrix} cov(Z_2, Z_3) \\ cov(Z_1, Z_3) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$cov(Q, 2P) = 2cov(Q, P) = 2 \cdot cov(P, Q)^T = [2 \quad 0]$$

- More discussions