

Lecture 9: Statistical Inference

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

May 9, 2023

Outline

1 Overview of Statistical Inference

2 Classical Statistical Inference

3 Bayesian Statistical Inference

一个中心 Bayes

4 Conditional Expectation: Useful Tools

conditional
probability
expectation

5 Prediction & Estimation

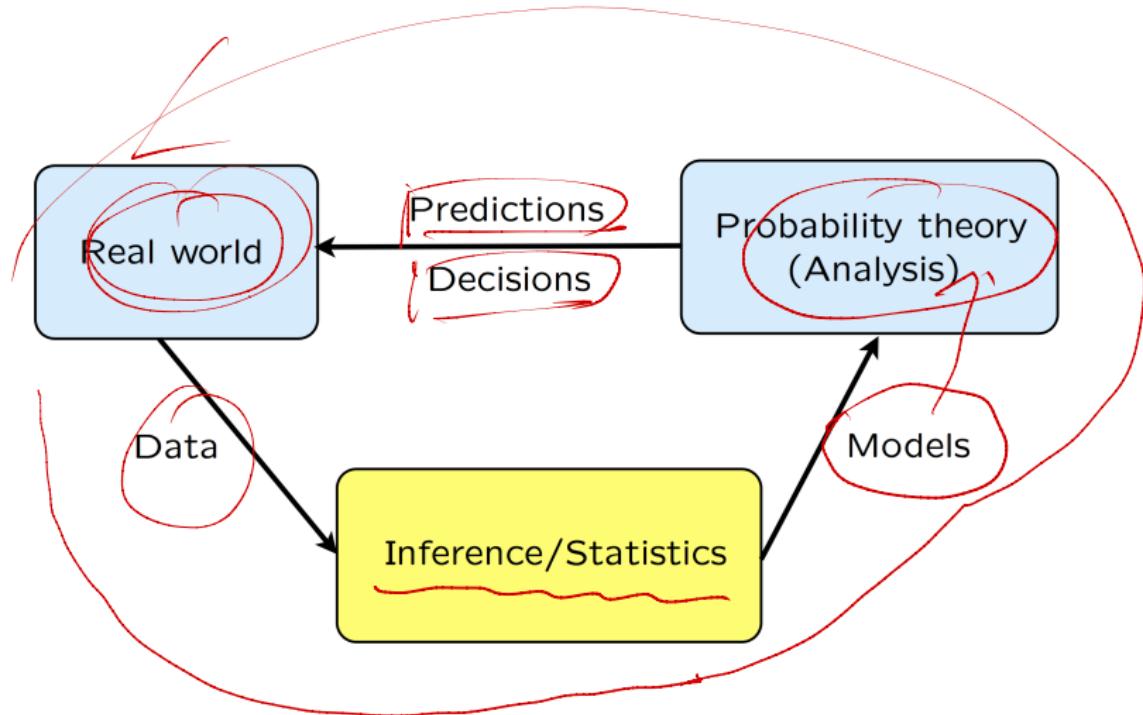
预测
估计

6 Application Case: Kalman Filter

Outline

- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

Inference



Statistical Inference

统计推断：从数据中获取信息。

- The process of extracting information about an unknown variable or an unknown model from available data
- Called “Learning” in CS
- Called “Signal Processing” in EE
- A typical question is: given a sample $X_1, \dots, X_n \sim F$, how do we infer F or some features of F ?

Core Parts of Statistical Inference

- Point Estimation
- Interval Estimation (Confidence Interval)
- Hypothesis Testing

Point Estimation

点估值

Point Estimation refers to providing a single “best guess” of some quantity of interest such as

- a parameter θ (possibly multi-dimensional) in a parametric model
- a CDF F
- a probability density function f
- a prediction for a future value Y of some random variable

Interval Estimation

区间估值

Definition

An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \dots, x_n)$ and $U(x_1, \dots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. When $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an interval estimator.

Confidence Interval

Definition

A $1 - \alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \dots, X_n)$ and $b = b(X_1, \dots, X_n)$ are functions of the data such that

样本函数

$$\underline{P_\theta(\theta \in C_n) \geq 1 - \alpha, \forall \theta \in \Theta.}$$

In words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the coverage of the confidence interval. If θ is a vector, then we use a confidence set instead of an interval.

Hypothesis Testing

假设检验

Start with a finite number of competing hypotheses and use the available data to decide which of them is true.

- Example 1: given a noisy picture, decide whether there is a person in the picture or not
- Example 2: given a noisy received signal, decide whether symbol 1 or 0 was sent by the transmitter
- Example 3: given a set of trials with three alternative medical treatments, decide which treatment is the most effective

Hypothesis Testing

- A hypothesis is a statement about the data
- The two complementary hypothesis in a hypothesis testing problem are called the Null Hypothesis (denoted by H_0) and the Alternative Hypothesis (denoted by H_1).

Example: Testing if a coin is Fair

- $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ are n independent coin flips.
 - H_0 : the coin is fair (Null Hypothesis)
 - H_1 : the coin is not fair (Alternative Hypothesis)
 - $H_0 : p = \frac{1}{2}$ versus $H_1 : p \neq \frac{1}{2}$
 - It seems reasonable to reject H_0 if $T = |\hat{p}_n - \frac{1}{2}|$ is large
- reject $H_0 \neq$ accept H_1
- \hat{p}_n 样本平均

Bayesian versus Classical Statistical Inference

Bayesian : 未知的变量 (prior) 一切在变

Classic : 未知的常数 逼近真实

- Difference relates to the nature of the unknown models or variables
- Treated as random variables with prior (known) distributions:
Bayesian approach
- Treated as an unknown constants: **classical/frequentist approach**

Bayesian vs. Classical

有先验的随机变量

- Inference using the Bayes rule:

unknown Θ and observation X are both random variables

— Find $p_{\Theta|X}$

- Classical statistics: unknown constant θ



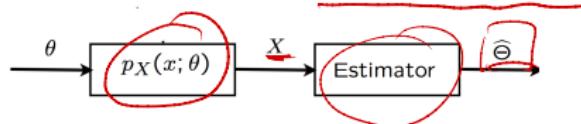
- also for vectors X and θ : $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$
- $p_X(x; \theta)$ are NOT conditional probabilities; θ is NOT random
- mathematically: many models, one for each possible value of θ

Outline

- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

Classical Statistical Inference

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
- Estimation: design an **estimator** $\hat{\theta}$, to “keep estimation error $\hat{\theta} - \theta$ small”

Inference Rule: Maximum Likelihood Estimation

最大似然估値

(MLE) PDF $f(x; \theta)$ = PDF; a function of x with θ fixed;
Likelihood is a function of θ with x fixed;

- Joint distribution of the vector of observations

$X = (X_1, \dots, X_n)$: PMF $P_X(x; \theta)$ (or PDF $f_X(x; \theta)$)

- θ : unknown (scalar or vector) parameter θ .

- We observe a particular value $x = (x_1, \dots, x_n)$ of X , then a maximum likelihood estimate (MLE) is a value of the parameter that maximizes the numerical function

$P_X(x_1, \dots, x_n; \theta)$ (or $f_X(x_1, \dots, x_n; \theta)$) over all θ :

θ fixed

给定数据,
找 θ ,
使似然 max

$$\hat{\theta}_n = \arg \max_{\theta} P_X(x_1, \dots, x_n; \theta)$$

$$\hat{\theta}_n = \arg \max_{\theta} f_X(x_1, \dots, x_n; \theta)$$

概率的观点:

x fixed

统计的角度.

同一东西, 不同视角

MLE under Independent Case

- Observations X_i are independent, and we observe a particular value $x = (x_1, \dots, x_n)$ of X .
- We define the log-likelihood function as follows:

$$\underbrace{\log[P_X(x_1, \dots, x_n; \theta)]}_{\text{log-likelihood}} = \log \prod_{i=1}^n P_{X_i}(x_i; \theta) = \sum_{i=1}^n \underbrace{\log[P_{X_i}(x_i; \theta)]}_{\text{log-likelihood}}$$

$$\underbrace{\log[f_X(x_1, \dots, x_n; \theta)]}_{\text{log-likelihood}} = \log \prod_{i=1}^n f_{X_i}(x_i; \theta) = \sum_{i=1}^n \underbrace{\log[f_{X_i}(x_i; \theta)]}_{\text{log-likelihood}}$$

MLE under Independent Case

- Thus a **maximum likelihood estimate** (MLE) under independent case is a value of the parameter that maximizes the numerical function $P_X(x_1, \dots, x_n; \theta)$ (or $f_X(x_1, \dots, x_n; \theta)$) over all θ :

$$\left\{ \begin{array}{l} \hat{\theta}_n = \arg \max_{\theta} \sum_{i=1}^n \log[P_{X_i}(x_i; \theta)] \\ \hat{\theta}_n = \arg \max_{\theta} \underbrace{\sum_{i=1}^n \log[f_{X_i}(x_i; \theta)]}_{n} \end{array} \right.$$

Example: Biased Coin Problem

Coin : P

1°. n coin tosses; n independent Bernoulli trials.
 $x_1, \dots, x_n \sim \text{Bern}(p)$.

2°. $x_1, \dots, x_n \in \mathbb{R}$ (observation value)
 $x_i = 0$ or 1 .

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = 0) = 1-p$$

$$\begin{aligned} P_X(x; p) &= \prod_{i=1}^n P_{X_i}(x_i; p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n-\sum_{i=1}^n x_i} \\ \text{r.v. } s_n &= x_1 + \dots + x_n : \# \text{ of heads in } n \text{ coin tosses.} \end{aligned}$$

$$\begin{aligned} 3°. \log P_X(x; p) &= \underbrace{s_n \log p + (n-s_n) \log (1-p)}_{f(p)} = f(p), \quad [f'(p)=0; \\ \hat{p}_{MLE} &= \arg \max_p f(p) \quad f''(p) \leq 0] \end{aligned}$$

$$\Rightarrow \hat{p}_{MLE} = \frac{1}{n} s_n = \frac{1}{n} (x_1 + \dots + x_n)$$

误差的分布：正态分布

Example: Biased Coin Problem

Outline

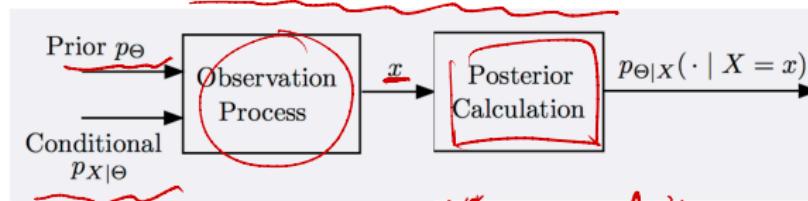
- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

The Bayesian Statistical Inference Framework



- Unknown Θ
 - treated as a random variable
 - prior distribution p_Θ or f_Θ
- Observation X
 - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot | X = x)$ or $f_{\Theta|X}(\cdot | X = x)$

- Where does the prior come from?
 - symmetry
 - known range
 - earlier studies
 - subjective or arbitrary

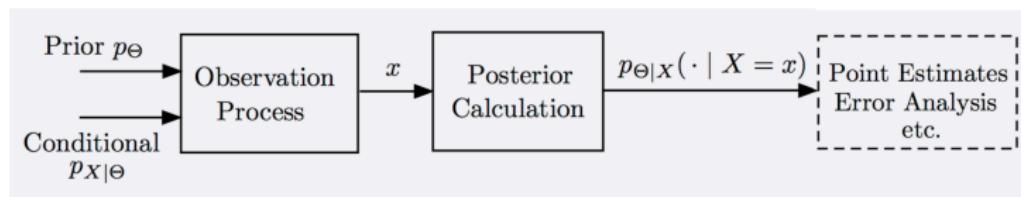
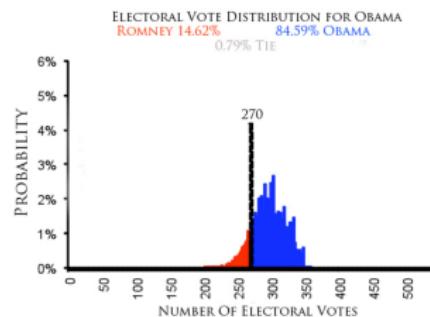
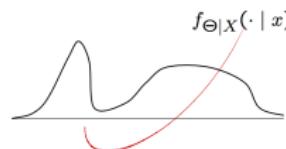
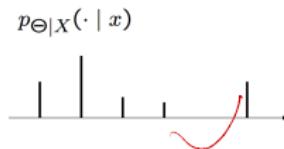


计算后验为核心
↓
算力要求

The Output of Bayesian Statistical Inference

The complete answer is a posterior distribution:

PMF $p_{\Theta|X}(\cdot | x)$ or PDF $f_{\Theta|X}(\cdot | x)$



General LOTP

	Y discrete	Y continuous
X discrete	$P(X = x) = \sum_y P(X = x Y = y)P(Y = y)$	$P(X = x) = \int_{-\infty}^{\infty} P(X = x Y = y)f_Y(y)dy$
X continuous	$f_X(x) = \sum_y f_X(x Y = y)P(Y = y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X Y}(x y)f_Y(y)dy$

General Bayes' Rule

	Y discrete	Y continuous
X discrete	$P(Y = y X = x) = \frac{P(X=x Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y X = x) = \frac{P(X=x Y=y)f_Y(y)}{P(X=x)}$
X continuous	$P(Y = y X = x) = \frac{f_X(x Y=y)P(Y=y)}{f_X(x)}$	$f_{Y X}(y x) = \frac{f_{X Y}(x y)f_Y(y)}{f_X(x)}$

General Bayes' Rule: Another Perspective

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta | x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x | \theta)}{\sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x | \theta')}.$$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta | x) = \frac{p_\Theta(\theta)f_{X|\Theta}(x | \theta)}{\sum_{\theta'} p_\Theta(\theta')f_{X|\Theta}(x | \theta')}.$$

- Θ continuous, X discrete:

$$f_{\Theta|X}(\theta | x) = \frac{f_\Theta(\theta)p_{X|\Theta}(x | \theta)}{\int f_\Theta(\theta')p_{X|\Theta}(x | \theta') d\theta'}.$$

- Θ continuous, X continuous:

$$f_{\Theta|X}(\theta | x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x | \theta)}{\int f_\Theta(\theta')f_{X|\Theta}(x | \theta') d\theta'}.$$

Example of Inference Rule: The Maximum A Posteriori Probability (MAP)

最大后验概率

- Given the observation value x , the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|x}(\theta|x)$ (if Θ is discrete) or $f_{\Theta|x}(\theta|x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ :
 - $p_{\Theta}(\theta)p_{X|\Theta}(x|\theta)$: if both Θ and X are discrete
 - $p_{\Theta}(\theta)f_{X|\Theta}(x|\theta)$: if Θ is discrete and X is continuous
 - $f_{\Theta}(\theta)p_{X|\Theta}(x|\theta)$: if Θ is continuous and X is discrete
 - $f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)$: if both Θ and X are continuous

$$\text{Bayes' Rule: } p_{\Theta|x}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{P_X(x)}$$

↓
CONST

The MAP Rule for Point Estimates

- An estimator is a random variable of the form $\hat{\theta} = g(X)$, for some function g . Different choices of g correspond to different estimators.
- An estimate is the value $\hat{\theta}$ of an estimator, as determined by the realized value x of the observation X .
- Once the value x of X is observed, the MAP estimator, sets the estimate $\hat{\theta}$ to a value that maximizes the posterior distribution over all possible values of θ .

Example: Inference of A Biased Coin

1^o. $\Theta \sim \text{Unif}(0,1) = \text{Beta}(1,1)$; # of heads $X | \Theta = \theta \sim \text{Bin}(n, \theta)$.

By Beta-Binomial Conjugacy, $\Theta | X=k \sim \text{Beta}(1+k, 1+n-k)$

Another Inference Rule $\hat{\theta} = E[\Theta | X=k] = \frac{k+1}{k+1+n-k} = \frac{k+1}{n+2}$; $\hat{\theta} = \left(\frac{x+1}{n+2}\right)$. \Rightarrow 順道: 平均值

We wish to estimate the probability of landing heads, denoted by θ , of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \text{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MAP estimator of Θ .

2^o. MAP: $f_{\Theta|X=k}(\theta) \propto \frac{\theta^k (1-\theta)^{n-k}}{\theta f(0,1)}$. Beta distribution

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}} f_{\Theta|X=k}(\theta) = \underset{\theta}{\operatorname{argmax}} \theta^k (1-\theta)^{n-k} \\ &= \underset{\theta}{\operatorname{argmax}} [\log(\theta^k (1-\theta)^{n-k})]\end{aligned}$$

Solution

$$g(\theta) = \log(\theta^k (1-\theta)^{n-k}) = k \log \theta + (n-k) \log(1-\theta)$$

$$g'(\theta) = 0 \text{ ; } g''(\theta) \leq 0.$$

$$\Rightarrow \theta^* = \frac{k}{n} :$$

$$\Rightarrow \hat{\theta}_{MAP|x=k} = \frac{k}{n}$$

估
值

estimate .. (real number.)

$$\hat{\theta}_{MAP|X} = \frac{X}{n} : \frac{\text{estimator}}{(r.v.)} \text{ 估值器}$$

MAP + Unif prior distribution = MLE,
Bayesian. Classical.

Solution

Outline

- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
条件期望
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

event
↑
Given A $E[X|A]$: real number.

Given r.v. Y $E[X|Y]$: r.v.

Conditional PMF

- Let A be an event with positive probability. If X is a discrete r.v., then the *conditional PMF of X given A* is

$$\underline{P_{X|A}(x)} = P(X = x|A) = \frac{P(\{X = x\} \cap A)}{P(A)}.$$

- Bayes' Rule:

$$\underline{P_{X|A}(x)} = P(X = x|A) = \frac{\underline{P(A|X = x)} P(X = x)}{\underline{P(A)}}.$$

- LOTP: with a partition A_1, \dots, A_n , each A_i with a positive probability $P(A_i) > 0$, $i = 1, 2, \dots, n$:

$$\underline{P(X = x)} = \sum_{i=1}^n P_{X|A_i}(x) P(A_i).$$

Conditional PDF

- Let A be an event with positive probability. If X is a continuous r.v., then the *conditional PDF of X given A* is

$$f_{X|A}(x) = \underline{(P(X \leq x|A))'}$$

$$F'_{X|A}(x)$$

- LOTP: with a partition A_1, \dots, A_n , each A_i with a positive probability $P(A_i) > 0$, $i = 1, 2, \dots, n$:

$$\underline{f_X(x)} = \sum_{i=1}^n P(A_i) f_{X|A_i}(x).$$

$$\underline{P(X \leq x)} = \sum_{i=1}^n P(A_i) \underline{(P(X \leq x|A_i))'}$$

Conditional PDF $f_{X|A}(x) = \lim_{\delta \rightarrow 0} \frac{P(x \leq X \leq x+\delta | A)}{\delta}$.

- Bayes' Rule: given an event A with $P(A) > 0$, then

$$f_{X|A}(x) = \frac{P(A|X=x)}{P(A)} \cdot f_X(x).$$

- Bayes' Rule: given event $A = "a \leq X \leq b"$ and $P(A) > 0$, then

$$\lim_{\delta \rightarrow 0} \frac{P(A|x \leq X \leq x+\delta) \cdot P(x \leq X \leq x+\delta)}{\delta \cdot P(A)} f_{X|A}(x) = \frac{\mathbf{1}_{x \in [a,b]}}{P(A)} \cdot f_X(x)$$

$$= \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$= \lim_{\delta \rightarrow 0} \frac{1}{P(A)} \left[\underbrace{P(A|x \leq X \leq x+\delta)}_{\delta} \cdot \underbrace{\frac{P(x \leq X \leq x+\delta)}{\delta}}_{f_{X|A}(x)} \right]$$

$$= \frac{1}{P(A)} \cdot P(A|X=x) \cdot f_X(x)$$

Conditional Expectation Given An Event

Definition

Let A be an event with positive probability. If Y is a discrete r.v., then the *conditional expectation of Y given A* is

$$\underline{E(Y|A)} = \sum_y y \cdot P(Y=y|A) = \sum_y y \cdot \underline{P_{Y|A}(y)},$$

where the sum is over the support of Y . If Y is a continuous r.v. with PDF f , then

$$E(Y|A) = \int_{-\infty}^{\infty} y \cdot \underline{f_{Y|A}(y)} dy.$$

LOTUS Given An Event

Definition

Let A be an event with positive probability and g is a function from \mathbf{R} to \mathbf{R} . If Y is a discrete r.v., then the *conditional expectation of $g(Y)$ given A* is

$$E(\underline{g(Y)}|A) = \sum_y \underline{g(y)} P_{Y|A}(y), \quad \sum_y \underline{g(y)}(P_{Y|A}(y))$$

where the sum is over the support of Y .

If Y is a continuous r.v. with PDF f , then

$$E(g(Y)|A) = \int_{-\infty}^{\infty} g(y) \cdot \underline{f_{Y|A}(y)} dy.$$

Example 1^o. event $A = "X > 1"$; $P(A) = P(X > 1)$

(Method 1:

$$f_{X|A}(x) = \begin{cases} \frac{f(x)}{P(A)} & = \lambda e^{-\lambda(x-1)} \\ & \text{if } x > 1. \\ 0 & \text{otherwise.} \end{cases}$$

$$E(X) = \frac{1}{\lambda}$$

$$\text{PDF. } f(x) = \lambda e^{-\lambda x}, x > 0. \quad E(X|X > a) = a + \frac{1}{\lambda}$$

Let $X \sim \text{Expo}(\lambda)$, find $E(X|X > 1)$ and $\text{Var}(X|X > 1)$.

$$X \sim \text{Expo}(\lambda).$$

$$E(X) = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

$$2^o. \quad E(X|X > 1) = \int_1^\infty x \cdot f_{X|A}(x) dx = \int_1^\infty x \cdot \lambda e^{-\lambda(x-1)} dx = 1 + \frac{1}{\lambda}$$

$$E(X^2|X > 1) = \int_1^\infty x^2 \cdot f_{X|A}(x) dx = \int_1^\infty x^2 \cdot \lambda e^{-\lambda(x-1)} dx$$

$$\Rightarrow \text{Var}(X|X > 1) = E(X^2|X > 1) - (E(X|X > 1))^2 = \frac{\lambda^2 + 2\lambda + 2}{\lambda^2}$$

$$= \frac{1}{\lambda^2}$$

Solution Method 2 : ① $P(X > s+t | X > t) = P(X > s)$ Memoryless. $s, t > 0$

? $E[X | X < 1]$

Let $t=1 \Rightarrow P(X > s+1 | X > 1) = P(X > s).$

$\Rightarrow P(X-1 > s | X > 1) = P(X > s).$

条件期望也是期望 \Rightarrow linearity

$$E[X | X > m] = m + \frac{1}{\lambda}.$$

$$E[X-1 | X > 1] = E[X]. \checkmark$$

② $E[X | X > 1] = E[(X-1)+1 | X > 1] = E[X-1 | X > 1] + 1$

③ $Var(X | X > 1)$

$$= E[X] + 1$$

$$= Var(X-1 | X > 1) = Var(X) = \frac{1}{\lambda^2}$$

~~Geom(p)~~ 类似 \Rightarrow memory less

$$= \frac{1}{\lambda^2} + 1$$

If $X \sim Geom(p)$; $E[X | X > 1]$? $Var(X | X > 1)$?

Motivation of Conditional Expectation

- Conditional expectation is a powerful tool for calculating expectations: first-step analysis
- Conditional expectation allows us to predict or estimate unknowns based on whatever evidence is currently available.
- Conditional Expectation given an event: $E(Y|A)$ *Real Number*
- Conditional Expectation given a random variable: $E(Y|X)$

Intuition for $E(Y|A)$

① Y , r.v.

$$E(Y) \approx \frac{1}{n} \sum_{j=1}^n y_j$$

y_1, \dots, y_n

②

$E[Y|A]$

$$\approx \frac{\sum_{j=1}^n I_j \cdot y_j}{\sum_{j=1}^n I_j}$$

n measurements.

where event A occurs.

I_j : in the j th measure ments.

Intuition for $E(Y|A)$

Principle

$E(Y|A)$ is approximately the average of Y in a large number of simulation runs in which A occurred.

Life Expectancy

T : Life span.

$$E(T) = 70 \text{ ;}$$

$$\underline{E[T | T \geq 20]} \stackrel{?}{=} E[T]$$

if $T \sim \text{Expo}(\lambda)$;

$$E[T | T \geq 20] = 20 + E[T] = 90$$

Law of Total Expectation

LoTE

$$E[X|X \leq 1] ?$$

$$E[X|X > 1]$$

$$X \sim \text{Exp}(\lambda)$$

$$E[X] = E[X|X > 1] \cdot P(X > 1) + E[X|X \leq 1] \cdot P(X \leq 1)$$

Theorem $\frac{1}{\lambda} = [1 + \frac{1}{\lambda}] \cdot e^{-\lambda} + E[X|X \leq 1] \cdot (1 - e^{-\lambda})$

Let A_1, \dots, A_n be a partition of a sample space, with $P(A_i) > 0$ for all i , and let Y be a random variable on this sample space. Then

$$E(Y) = \sum_{i=1}^n E(Y|A_i) P(A_i).$$

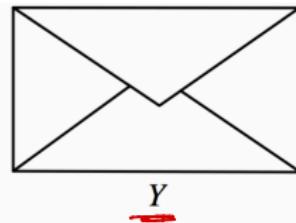
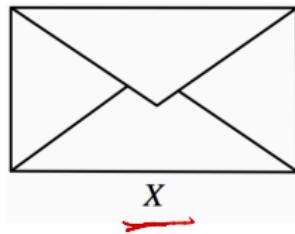
$$\Rightarrow E[X|X \leq 1] = \frac{\frac{1}{\lambda} - (1 + \frac{1}{\lambda})e^{-\lambda}}{1 - e^{-\lambda}} = \frac{1}{\lambda} - \frac{e^{-\lambda}}{1 - e^{-\lambda}}$$

Two-envelope Paradox

$$\textcircled{1} \quad Y = 2X \text{ or } \frac{X}{2}.$$

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

A stranger presents you with two identical-looking, sealed envelopes, each of which contains a check for some positive amount of money. You are informed that one of the envelopes contains exactly twice as much money as the other. You can choose either envelope. Which do you prefer: the one on the left or the one on the right? (Assume that the expected amount of money in each envelope is finite—certainly a good assumption in the real world!) $\textcircled{2} \quad X = \frac{Y}{2}$



$$\text{or } X = 2Y$$

$$E(X) = \frac{5}{4} E(Y)$$

$$> E(Y).$$

FIGURE 9.1

Two envelopes, where one contains twice as much money as the other. Either $Y = 2X$ or $Y = X/2$, with equal probabilities. Which would you prefer?

Solution

$$Y = \begin{cases} 2X & \text{w.p. } 0.5 \\ \frac{X}{2} & \text{w.p. } 0.5 \end{cases}$$

NOTE : $E(Y) = \underbrace{E[Y|Y=2X]}_{\neq E[2X]} P(Y=2X) + E[Y|Y=\frac{X}{2}] P(Y=\frac{X}{2})$

$$E[Y|Y=2X] = \underbrace{E[2X|Y=2X]}_{\text{2X与Y=2X不独立}} \neq E[2X]$$

2X与Y=2X不独立

Geometric Expectation Redux

$X \sim \text{Geom}(p)$, Find $E[X]$?

① First-Step Analysis;

Conditioning on the outcome of the first toss;

$$O_1 = H \text{ or } T$$

$$\begin{aligned} E[X] &\stackrel{\text{Lotto}}{=} \underbrace{E[X | O_1=H]}_{0 \cdot p} \underbrace{p(O_1=H)}_{+} + \underbrace{E[X | O_1=T]}_{+ [1+E(X)] \cdot (1-p)} \cdot p(O_1=T) \\ &+ [1+E(X)] \cdot (1-p) \end{aligned}$$

$$\Rightarrow E[X] = \frac{1}{p}$$

Coin Tosses.

of tosses

before the first Head

H: Head

T: Tail.

Time until HH vs. HT

H: Head, T: Tail.

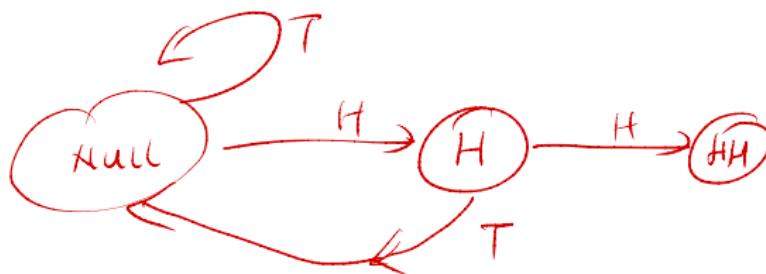
- ① HT : partial progress



{ PGF
LOTE.
Markov chain
Renewal process
Martingale

You toss a fair coin repeatedly. What is the expected number of tosses until the pattern HT appears for the first time? What about the expected number of tosses until HH appears for the first time?

- ② HH :



Solution

W_{HT} : # of tosses until "HT" for the first time.

$$W_{HT} = W_1 + W_2$$

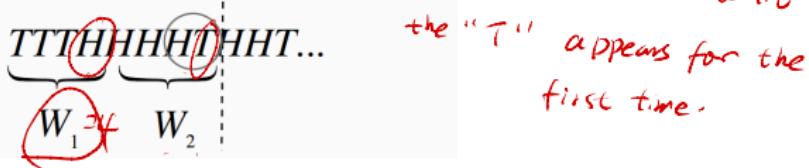
$X \sim F_{Sc(p)}$.

$$E(X) = \frac{1}{p};$$

FS : Memoryless.

W_1 : # of tosses until the "H" appears for the first time.

W_2 : after "H", # of additional tosses until the "T" appears for the first time.



$$W_1 \sim F_{Sc}(1/2); \quad W_2 \sim F_{Sc}(1/2).$$

$$\Rightarrow E(W_1) = 2; \quad E(W_2) = 2$$

$$\Rightarrow E(W_{HT}) = E(W_1 + W_2) = E(W_1) + E(W_2) \\ = 2 + 2 = 4.$$

Solution $E(W_{HH})$: First-step Analysis.

O_1 : outcome of 1st toss
 O_2 : ... 2nd toss

$$1^0. E(W_{HH}) \stackrel{LOTE}{=} E(W_{HH} | O_1=H) P(O_1=H) + E(W_{HH} | O_1=T) P(O_1=T)$$

$$\downarrow ? [2 + \frac{1}{2} E(W_{HH})] \quad \frac{1}{2} + [1 + E(W_{HT})] \cdot \frac{1}{2}.$$

LOTE with extra conditioning.

THTTTT(HHHHH...)
↑
start over

$$2^0. E[W_{HH} | O_1=H] = E[W_{HH} | O_1=H, O_2=H] \cdot P(O_2=H | O_1=H)$$

$$= 2 \cdot \frac{1}{2} + (2 + E(W_{HT})) \cdot \frac{1}{2}$$

$$= 2 + \frac{1}{2} E(W_{HT}).$$

$$+ E[W_{HH} | O_1=H, O_2=T] \cdot P(O_2=T | O_1=H)$$

$$(2 + E(W_{HT})) \cdot \frac{1}{2}.$$

Solution

$$3^\circ. \quad E[W_{HH}] = [2 + \frac{1}{2}E[W_{HH}]] \cdot \frac{1}{2} + [1 + E[W_{HH}]] \cdot \frac{1}{2}$$
$$= \frac{3}{2} + \frac{3}{4}E[W_{HH}]$$

$$\Rightarrow E[W_{HH}] = 6 > E[W_{HT}] = 4$$

$HHTHHTTTHHHHHHTHTHTHTTHTT$

$HHTHHTTTHHHHHHTHTHTHTTHTT$

Conditional Expectation Given An R.V.

$$\underline{g(x)} = E[Y | X=x] : \text{Real Number. estimate.}$$

$$g(X) = \underline{E[Y|X]} : \text{Random Variable. estimator.}$$

Definition $A = "X=x"$,

Let $\underline{g(x)} = E(Y|X=x)$. Then the *conditional expectation of Y given X*, denoted $E(Y|X)$, is defined to be the random variable $\underline{g(X)}$. In other words, if after doing the experiment X crystallizes into x , then $E(Y|X)$ crystallizes into $\underline{g(x)}$.

$X \rightarrow x$ 基于事件 \star



$$X \rightarrow \boxed{g(\)} \rightarrow \hat{Y} = \underline{g(X)} : \text{estimator.}$$

$$X = \underline{(x)}.$$

$$\underline{g(x)} \text{ estimate.}$$

Remark

- $E(Y|X)$ is a function of X , and it is a random variable.
- It makes sense to computer $E(E(Y|X))$ and $\text{Var}(E(Y|X))$.

Example: Stick Length



① $X \sim \text{unif}(0,1)$.

$$Y|_{X=x} \sim \text{unif}(0,x)$$



$$\Rightarrow E[Y|_{X=x}] = \frac{x}{2} = g(x) \quad \Rightarrow E[Y|X] = g(X) = \frac{X}{2}.$$

Suppose we have a stick of length 1 and break the stick at a point X chosen uniformly at random. Given that $X = \underline{\underline{X}}$, we then choose another breakpoint Y uniformly on the interval $[0, x]$. Find $E(Y|X)$, and its mean and variance.

② $E[\underline{\underline{E}}[Y|X]] = E\left[\frac{X}{2}\right] = \frac{1}{2}E[X] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$$\text{Var}[E[Y|X]] = \text{Var}\left[\frac{X}{2}\right] = \frac{1}{4}\text{Var}(X) = \frac{1}{4} \cdot \frac{1}{12} = \frac{1}{48}.$$

Find $E[Y|X]$ \Rightarrow ① $g(x) = E[Y|X=x]$

② $g(x) \rightarrow g(X)$ $(x \rightarrow X)$

Solution

Dropping What's Independent

$\forall x \in \mathbb{R}$ *

$$\underline{g(x)} = \underline{E[Y | X=x]} = E[Y]$$

$$\underline{g(x)} = E[Y] \quad \times$$

Theorem

If X and Y are independent, then $E(Y|X) = E(Y)$.

$$1^\circ E(XY) = E(X)E(Y)$$

$$2^\circ E(X|Y) = E(X)$$

Taking Out What's Known

$$\textcircled{1} \quad g(x) = E[h(x) \cdot Y | X=x]$$

$$= E[\underline{h(x)} \cdot Y | X=x]$$

$$\textcircled{2} \quad \underline{g(x)} = E[h(x) \cdot Y | X]$$

$$= \underline{h(x) E[Y|X]}$$

Theorem

For any function h ,

$$= h(x) \cdot \underline{E[Y|X=x]}$$

$$E(\underline{h(X) \cdot Y|X}) = \underline{h(X)} \underline{E(Y|X)}$$

给定 X 的信息 $\Rightarrow h(x) : \text{const}$

$$\underline{\underline{E(X|X) = X}}$$

Linearity

$$\textcircled{1} \quad g(x) = E[Y_1 + Y_2 | X=x]$$

$$= E[Y_1 | X=x] + E[Y_2 | X=x]$$

$$\textcircled{2} \quad g(x) = E[Y_1 | X] + E[Y_2 | X]$$

Theorem

$$\underline{E(Y_1 + Y_2 | X)} = E(Y_1 | X) + E(Y_2 | X).$$

$$E(Y | X_1 + X_2) \neq E(Y | X_1) + E(Y | X_2)$$

Example

1^o. By symmetry, $E[X_1|S_n] = E[X_2|S_n] = \dots = E[X_n|S_n]$

2^o. By Linearity - $E[X_1|S_n] + E[X_2|S_n] + \dots + E[X_n|S_n]$
 $= \overbrace{E[X_1 + \dots + X_n|S_n]}$

Let X_1, \dots, X_n be i.i.d., and $S_n = X_1 + \dots + X_n$. Find $E(X_1|S_n)$.

$$= \overbrace{E[S_n|S_n]}$$

$$= \underline{S_n}.$$

$$E[S_n|S_n=s] = s = g(s).$$

$$E[S_n|S_n] = g(S_n)$$

$$= S_n$$

3^o. $E[\star|S_n] = \frac{1}{n}S_n$.

Adam's Law

The Law of Iterated Expectation

迭代期望定律

Theorem

For any r.v.s X and Y ,

$$E(E(Y|X)) = E(Y).$$

Proof

$$\frac{E[E[Y|X]] = E[Y]}{\text{LHS} \quad \checkmark \quad \text{RHS.}}$$

W.L.O.G. X and Y are both discrete r.v.s.

$$1^{\circ}. \underline{g(X) = E[Y|X]}, \quad g(x) = E[Y|X=x] = \sum_y y \cdot p(Y=y|X=x)$$

$$2^{\circ}. \text{ LHS. } E[\underline{E[Y|X]}] = E[g(x)] = \sum_x g(x) \cdot p(X=x)$$

$$= \sum_x \left[\sum_y y \cdot p(Y=y|X=x) \cdot p(X=x) \right]$$

$$= \sum_y y \cdot \left[\sum_x p(Y=y|X=x) \cdot p(X=x) \right]$$

$$= \sum_y y \cdot \left[\sum_x p(Y=y, X=x) \right] = \sum_y y \cdot p(Y=y) \\ = EC(Y) \quad \text{RHS.}$$

Adam's Law and LOTE

discrete X .

$$1^{\circ} \text{. LOTE . } E[Y] = \sum_x E[Y|X=x] \cdot P(X=x)$$

$$2^{\circ} \text{. Adam's Law . } g(x) = \underline{E[Y|X=x]} ; \quad g(X) = \underline{E[Y|X]}$$

$$\Rightarrow E[E[Y|X]] = E[g(X)]$$

$$= \sum_x \underline{g(x)} \cdot P(X=x)$$

$$= \underline{\sum_x E(Y|X=x) \cdot P(X=x)}$$

$$\stackrel{\text{LOTE}}{=} E[Y]$$

Adam's Law with Extra Conditioning

① $\hat{P}(\cdot) = P(\cdot|Z)$ $\rightarrow \hat{E} = E(\cdot|Z)$... 条件期望也是期望

② Adam's Law. $E[\underline{E}[Y|X]] = \underline{E}[Y]$

Theorem

③ $\hat{E}[\hat{E}[Y|X]] = \underline{E}[Y]$

For any r.v.s X, Y, Z , we have

$$\begin{aligned} E(E(Y|X, Z)|Z) &= \underline{E}(Y|Z) \quad \stackrel{2^o}{=} E(Y|Z) \\ E(E(X|Z, Y)|Y) &= E(X|Y) \end{aligned}$$

2^o. $\hat{E}(Y|X) = E(Y|X, Z)$

3^o. $\hat{E}[\hat{E}(Y|X)] = E(\hat{E}(Y|X)|Z)$

$$= \underline{E}[E(Y|X, Z)|Z]$$

Conditional Variance

$$1^{\circ} \cdot \text{Var}(Y) = \underline{E[(Y - E(Y))^2]}$$

$$\hat{E}(\cdot) = E(\cdot | X)$$

$$\hat{E}(Y) = E(Y|X).$$

Definition

$$\underline{\text{Var}(Y|X)} = \hat{E}[(Y - \hat{E}(Y))^2]$$

The conditional variance of Y given X is

$$= E[(Y - \hat{E}(Y))^2 | X]$$

$$\text{Var}(Y|X) = \underline{E((Y - E(Y|X))^2 | X)} = \underline{E[(Y - E(Y|X))^2 | X]}$$

This is equivalent to

$$2^{\circ} \cdot \text{Var}(Y) = E(Y^2) - (\underline{E(Y)})^2$$

$$\text{Var}(Y|X) = \underline{E(Y^2 | X)} - (\underline{E(Y|X)})^2.$$

$$\begin{aligned} \text{Var}(Y|X) &= \hat{E}(Y^2) - (\underline{\hat{E}(Y)})^2 \\ &= E(Y^2 | X) - (\underline{E(Y|X)})^2 \end{aligned}$$

全方差定律 / 方差分解公式

$$X \Rightarrow x$$

Theorem

For any r.v.s X and Y ,

$$\underline{\text{Var}}(Y) = \underline{E}(\underline{\text{Var}}(Y|X)) + \underline{\text{Var}}(\underline{E}(Y|X)).$$

The ordering of E 's and Var 's on the right-hand side spells EVVE, whence the name Eve's law. Eve's law is also known as the law of total variance or the variance decomposition formula.

Proof $\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E(Y|X)).$

① $g(X) = E[Y|X] ; \text{ By Adam's Law, } E[g(X)] = E[E(Y|X)] = \underline{E[Y]}.$

② $E[\text{Var}(Y|X)] = E[E(Y^2|X) - (\underline{E(Y|X)})^2] = E[E(Y^2|X) - g^2(x)]$
 $= \underline{E[E(Y^2|X)]} - E[g^2(x)] \stackrel{\text{Adam's Law}}{=} \underline{E[Y^2]} - \underline{E[g^2(x)]}$

③ $\text{Var}[E(Y|X)] = \text{Var}(g(x)) = E[g^2(x)] - (\underline{E(g(x))})^2$
 $= \underline{E[g^2(x)]} - E^2(Y).$

② + ③ $\Rightarrow E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)] = E[Y^2] - \underline{E^2[Y]}$
 $= \text{Var}(Y)$

Proof

Example: Random Sum 随机个随机数相加(双随机)

$$\textcircled{1} \quad E[X|N=n] = E\left[\sum_{j=1}^N X_j | N=n\right] = E\left(\sum_{j=1}^n \underline{X_j} | \underline{N=n}\right) \\ = E\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n E(X_j) = n \cdot \mu.$$

A store receives N customers in a day, where N is an r.v. with finite mean and variance. Let X_j be the amount spent by the j th customer at the store. Assume that each X_j has mean μ and variance σ^2 , and that N and all the X_j are independent of one another. Find the mean and variance of the random sum $X = \sum_{j=1}^N X_j$, which is the store's total revenue in a day, in terms of μ , σ^2 , $E(N)$, and $\text{Var}(N)$.

$$\Rightarrow E[X|N] = N \cdot \mu.$$

Adam's Law

$$\Rightarrow E[X] = E[E[X|N]] = E[N \cdot \mu] = \mu \cdot E[N]$$

Solution ② $\text{Var}(X|N=n) = \text{Var}(\sum_{j=1}^N X_j | N=n)$

$$= \text{Var}\left(\sum_{j=1}^n X_j | N=n\right) = \text{Var}\left(\sum_{j=1}^n X_j\right)$$

$$= \sum_{j=1}^n \text{Var}(X_j) = n \cdot \sigma^2.$$

$\Rightarrow \text{Var}(X|N) = N \cdot \sigma^2$

Eve's Law.

$$\Rightarrow \text{Var}(X) = E[\underbrace{\text{Var}(X|N)}_{E(N\sigma^2)}] + \text{Var}[\underbrace{E(X|N)}_{N\mu}]$$

$$= E(N\sigma^2) + \text{Var}(N\mu)$$

$$= \sigma^2 E[N] + \mu^2 \cdot \text{Var}(N)$$

Solution

Outline

- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

Basic Problem

$$\hat{Y} = g(X)$$

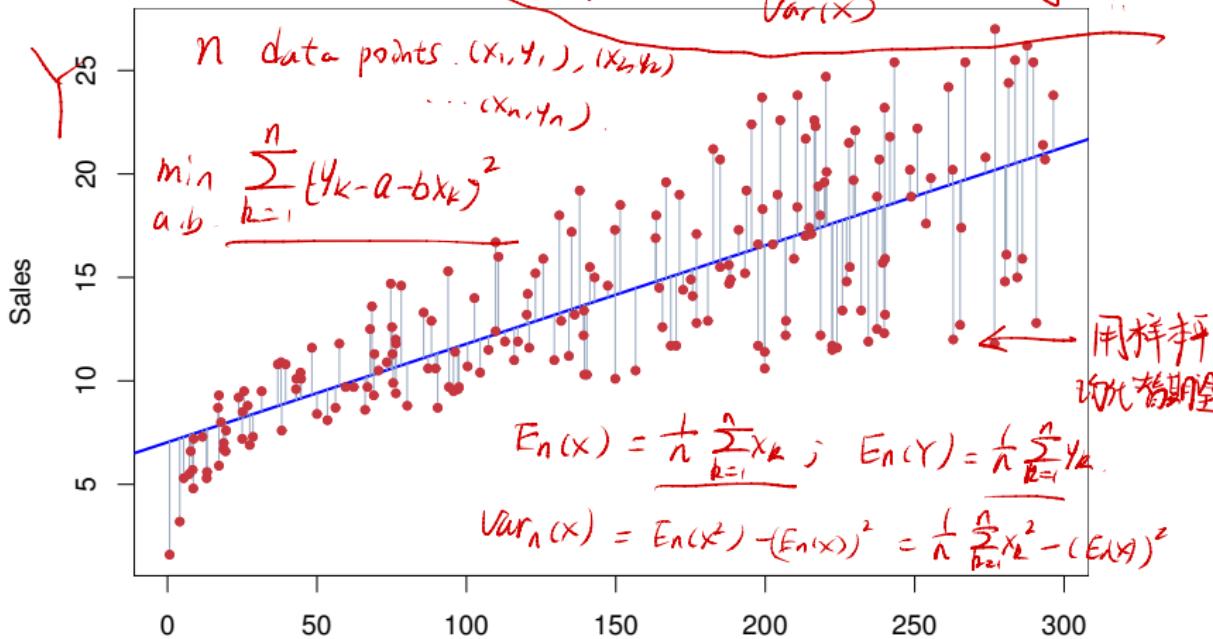
- Estimate \underline{Y} from the observed value X
- Choose the estimator (inference function) $\underline{g(\cdot)}$ to minimize the expected error $E(c(Y, \hat{Y}))$
- $c(Y, \hat{Y})$ is the cost of guessing \hat{Y} when the actually value is Y .
- When $c(Y, \hat{Y}) = \|Y - \hat{Y}\|^2$, the best guess is called “the least square estimate (LSE)” estimate of Y given X . 最小二乘估计
- Further, if the function $\underline{g(\cdot)}$ is restricted to be linear, i.e., of the form $a + bX$, it is called “the Linear Least Square Estimate (LLSE)” estimate of Y given X . 线性最小二乘估计
- Further, if the function $\underline{g(\cdot)}$ can be arbitrary, it is called “the Minimum Mean Square Estimate (MMSE)” estimate of Y given X .

$$E\hat{Y}(X)$$

Linear Regression

$$\hat{Y} = E[Y|X] = a + bX$$

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



$$\text{Cov}_n(X, Y) = E_n(XY) - E_n(X) \cdot E_n(Y)$$

$$= \frac{1}{n} \sum_{h=1}^n x_h y_h - E_n(x) \cdot E_n(y).$$

TV

$\{E_n\} \rightarrow E$,
 $n \rightarrow \infty$; SLU

Linear Regression

An extremely widely used method for data analysis in statistics is *linear regression*. In its most basic form, the linear regression model uses a single explanatory variable X to predict a response variable Y , and it assumes that the conditional expectation of Y is *linear* in X :

$$E(Y|X) = a + bX.$$

$$\hat{Y} = \underline{a+bX}$$

$$Y = \hat{Y} + \epsilon$$

- (a) Show that an equivalent way to express this is to write $\Rightarrow \text{error}$

$$Y = a + bX + \epsilon,$$

where ϵ is an r.v. (called the *error*) with $E(\epsilon|X) = 0$. $\Rightarrow \text{不相关}$

- (b) Solve for the constants a and b in terms of $E(X)$, $E(Y)$, $\text{Cov}(X, Y)$, and $\text{Var}(X)$.

Solution (a) 1°. if $Y = a + bX + \varepsilon$, $E[\varepsilon | X] = 0$ $\Rightarrow E[Y|X] = a + bX$.

$$\begin{aligned}E[Y|X] &= E[a + bX + \varepsilon | X] = a + b \cdot \underbrace{E[X|X]}_{\downarrow} + \underbrace{E[\varepsilon|X]}_0 \\&= a + b \cdot X + 0 = a + bX\end{aligned}$$

2°. if $E[Y|X] = a + bX$ $\Rightarrow Y = a + bX + \varepsilon$, $E[\varepsilon|X] = 0$

Let $\varepsilon \triangleq Y - (a + bX)$. $\Rightarrow Y = a + bX + \varepsilon$.

$$\begin{aligned}E[\varepsilon|X] &= E[Y - (a + bX)|X] = E[Y|X] - (a + bX) \\&= a + bX - (a + bX) = 0.\end{aligned}$$

3°. $E[\varepsilon] = \underbrace{E[E[\varepsilon|X]]}_{\text{Adam}} = E[0] = 0$.

Solution

$$(b), \circ Y = a + bX + \varepsilon ; \quad E[Y] = a + bE[X] + E[\varepsilon] \\ \Rightarrow a = E[Y] - b \cdot E[X] ; \quad = a + bE[X]$$

$$2^\circ. \text{Cov}(X, \varepsilon) = E[\underbrace{\varepsilon X}_{\text{0}}] - E[X] \cdot E[\varepsilon] = 0 .$$

$$E[\varepsilon X] \stackrel{\text{Adam's Law}}{=} E[E[\varepsilon X | X]] = E[X \cdot \underbrace{E[\varepsilon | X]}_{\text{0}}] = 0 .$$

$$3^\circ. \text{Cov}(X, Y) = \text{Cov}(X, a + bX + \varepsilon) = \underbrace{\text{Cov}(X, a)}_{\text{0}} + \underbrace{\text{Cov}(X, bX)}_{\text{0}} + \underbrace{\text{Cov}(X, \varepsilon)}_{\text{0}} = b \text{Cov}(X, X) = b \text{Var}(X)$$

$$\Rightarrow b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$a = E[Y] - bE[X] = E[Y] - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot E[X]$$

$$\Rightarrow \hat{Y} = a + bX = E[Y] + \frac{\text{Cov}(X, Y)}{\text{Var}(X)} (X - E[X]).$$

Linear Least Square Estimate $f(a,b) = E[(Y-a-bX)^2]$

$$\min_{a,b} E[(Y-a-bX)^2] = \min_{a,b} f(a,b)$$

$$(a^*, b^*) \in \arg\min_{(a,b)} f(a,b) \Leftrightarrow \begin{aligned} \textcircled{1} \quad & \frac{\partial f(a,b)}{\partial a} \Big|_{a=a^*, b=b^*} = 0 \\ \textcircled{2} \quad & \frac{\partial f(a,b)}{\partial b} \Big|_{a=a^*, b=b^*} = 0. \end{aligned}$$

Theorem

The Linear Least Square Estimate (LLSE) of Y given X , denoted by $L[Y|X]$, is the linear function $a + bX$ that minimizes $E[(Y - a - bX)^2]$.
In fact,

$$\hat{Y} = L[Y|X] = E(Y) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - E(X))$$

③ Hessian Matrix.

positive semidefinite.

$$\begin{pmatrix} \frac{\partial^2 f(a,b)}{\partial a^2} & \frac{\partial^2 f(a,b)}{\partial a \partial b} \\ \frac{\partial^2 f(a,b)}{\partial b \partial a} & \frac{\partial^2 f(a,b)}{\partial b^2} \end{pmatrix} \succcurlyeq 0$$

Proof

Proof

Minimum Mean Square Error Estimator

MMSE

$$\text{LLSE} = \min_{\hat{Y}} E[(Y - \hat{Y})^2] \quad . \quad \hat{Y} = a + bX \Rightarrow \hat{Y}^* = E[\hat{Y}|X]$$

$$\text{MMSE} = \min_{\hat{Y}} E[(Y - \hat{Y})^2]. \quad \hat{Y} = g(X) \Rightarrow \hat{Y}^* = E[Y|X]$$

Theorem

The MMSE of Y given X is given by

$$g(X) = E[Y|X]$$

$$\hat{Y} = g(X) \text{ 不限制线性}$$

Geometric Perspective

X, Y : r.v.

③ X and Y are "orthogonal" 正交

If $E[XY] = 0$, $\underline{X \perp Y}$

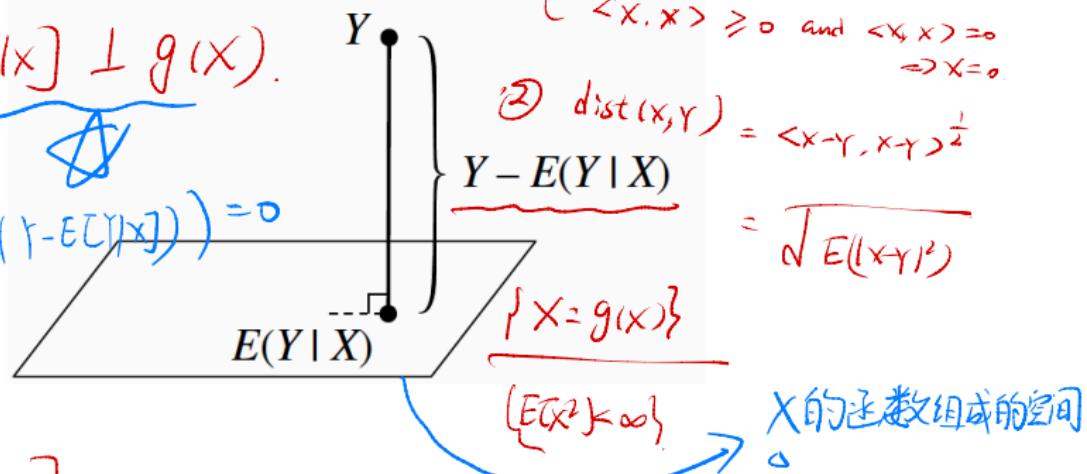
① Inner product $\langle X, Y \rangle = E(XY)$

$$\begin{cases} \langle X, Y \rangle = \langle Y, X \rangle, \\ \langle \alpha X + \beta Y, Z \rangle = \alpha \langle X, Z \rangle + \beta \langle Y, Z \rangle \\ \langle X, X \rangle \geq 0 \text{ and } \langle X, X \rangle = 0 \Rightarrow X = 0. \end{cases}$$

④ $Y - E(Y|X) \perp g(X)$.



$$E(g(X)(Y - E(Y|X))) = 0$$



⑤ $E[Y|X]$: a projection of Y onto the space of arbitrary functions of X .

条件期望就是投影

Projection Interpretation

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$$

if $E(X)$ or $E(Y) = 0$ or both

$$\Rightarrow \text{Cov}(X, Y) = 0$$

\Rightarrow Uncorrelated \Leftrightarrow orthogonal.

Theorem

For any function h , the r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$.

Equivalently,

$$E((Y - E(Y|X))h(X)) = 0.$$

$$Y - E(Y|X) \perp h(X)$$

(This is equivalent since $E(Y - E(Y|X)) = 0$, by linearity and Adam's law.)

$$\begin{aligned} E[Y - E(Y|X)] &= E(Y) - E[E(Y|X)] \\ &= E(Y) - E[Y] = 0 \end{aligned}$$

Adam

Proof

$$Y - E(Y|X) \perp h(X) \quad \text{✓}$$

$$\underline{E[(Y - E(Y|X)) \cdot h(X)]} = 0 \quad \text{✓}$$

$$= E[Yh(X)] - h(X) \cdot E(Y|X)$$

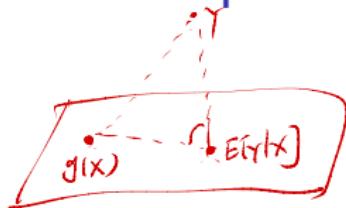
$$= E[Yh(X)] - E[h(X) \cdot \underline{E(Y|X)}]$$

$$= \underline{E[Yh(X)]} - \underline{E[E[h(X)Y|X]]} \quad \text{Adam's Law}$$

$$= E[h(X) \cdot Y] - E[h(X) \cdot Y] = 0$$

Proof

Prediction Perspective



$$E[(Y - E(Y|X))^2]$$

$$\leq E[(Y - g(X))^2] \quad \underline{g(\cdot)}$$

投影点是最近的点

- Predict or estimate the future observations or unknown parameters based on data
- $E(Y|X)$ is our **best predictor** of Y based on X . 均方误差
- Best means it is the function of X with the lowest mean squared error (expected squared difference between Y and prediction of Y).

Proof 1^o. \hat{Y} : estimator of Y . ($\hat{Y} = g(X)$).

$$E[(Y - \hat{Y})^2] = E[(Y - g(X))^2] \quad ; \quad Y - g(X) = \underbrace{(Y - E[Y|X])}_A + \underbrace{(E[Y|X] - g(X))}_B$$

$$\begin{aligned} 2^o. \quad \underbrace{E[(Y - g(X))^2]}_{\text{Var}(Y)} &= \underbrace{E[(Y - E[Y|X])^2]}_{\text{Var}(E[Y|X])} + \underbrace{E[(E[Y|X] - g(X))^2]}_{\text{Var}(g(X))} \\ &\quad + 2 \underbrace{E[(Y - E[Y|X])(E[Y|X] - g(X))]}_{\text{Cov}(Y, g(X))} \end{aligned}$$

3^o. $E[Y|X] - g(X)$ is still a function of X .

$$Y - E[Y|X] \perp (E[Y|X] - g(X))$$

$$\Rightarrow E[(Y - E[Y|X])(E[Y|X] - g(X))] = 0$$

\cancel{A} $E[Y|X]$ 也是 X 的函数 $\Leftarrow \int_Y f_{Y|X}(Y|X)Y dY$ 与 Y 无关

Proof 4°. $E[(Y - g(x))^2] = E[(Y - E[Y|x])^2] + \underbrace{E[(E[Y|x] - g(x))^2]}_{\text{dist}^2(Y, E[Y|x]) + \text{dist}^2(g(x), E[Y|x]) \geq 0}$

Hilbert 空间上勾股定理

$$g^*(x) = E[Y|x]$$

$$\Rightarrow E[(Y - g(x))^2] \geq \underbrace{E[(Y - E[Y|x])^2]}$$

$$\underline{\hat{Y}^* = g^*(x) = E[Y|x]} \quad \text{MSE.}$$

Proof

Orthogonality Property of MMSE

Theorem

(a) For any function $\phi(\cdot)$, one has

$$E[(Y - E[Y|X])\phi(X)] = 0$$

(b) Moreover, if the function $g(X)$ is such that

$$E[(Y - g(X))\phi(X)] = 0, \forall \phi(\cdot).$$

then $g(X) = \boxed{E(Y|X)}$

投影点存在且唯一

Proof

$$E[(Y - g(x))\phi(x)] = 0 \quad \forall \phi(\cdot)$$

$$\underbrace{E[(g(x) - E[Y|x])^2]}_{=} = 0 \Rightarrow g(x) = \underline{E[Y|x]}$$

$$= E[(g(x) - E[Y|x])(g(x) - Y + Y - E[Y|x])]$$

$$= E[\underbrace{(g(x) - E[Y|x])(g(x) - Y)}_{} + \underbrace{(g(x) - E[Y|x])(Y - E[Y|x])}_{}]$$

$$= E[\underbrace{(g(x) - E[Y|x])(g(x) - Y)}_{\xrightarrow{\text{X}} 0} + \underbrace{(g(x) - E[Y|x])(Y - E[Y|x])}_{0}]$$

$$\phi(x) = g(x) \Rightarrow E[Y|x]$$

$$Y - E[Y|x] \perp \phi(x)$$

$$g(x) - E[Y|x]$$

$$= 0$$

Proof

MMSE for Jointly Gaussian Random Variables

联合正态分布 LLSE : Suboptimal ; low complexity. 次优，复杂度低

MMSE : optimal ; high complexity. 最优，复杂度高

对线性解

Theorem 的投影

Let X, Y be jointly ^{Normal} Gaussian random variables. Then

$$E[Y|X] = L[Y|X] = E(Y) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - E(X)).$$

对所有解
的投影

Optimal / low complexity.

LLSE : a projection of Y onto the space of arbitrary linear
function of X .
 $(aX+b)$

Proof

proof sketch:

$$1^{\circ}. \underline{Y - L(Y|X)} \perp \underline{X}.$$

$$2^{\circ}. E[Y - L(Y|X)] = 0 \Leftrightarrow \text{Uncorrelation}$$

3^o. $Y - L(Y|X)$ is uncorrelated with X .

4^o. $Y - L(Y|X)$ and X Joint Normal.

$$t_1(Y - L(Y|X)) + t_2X \sim \text{Normal}.$$

$$L(Y|X) = \underline{ax + b} \quad \underline{t_1 Y + t_2' X} \sim \text{Normal}$$

5^o. $\underline{Y - L(Y|X)}$ is independent of X . \Rightarrow independent of $\phi(X)$. 二元正态不相关必独立

6^o. $\underline{Y - L(Y|X)}$ is uncorrelated of $\phi(X)$.

$$7^{\circ}. \underline{Y - L(Y|X)} \perp \underline{\phi(X)}.$$

$$8^{\circ}. \underline{L(Y|X)} = \underline{E[Y|X]}$$

Proof

Example: Revisit Biased Coin Problem

1^o. MMSE : $E[\Theta | X=k]$; $\Theta \sim \text{unif}(0,1) = \text{Beta}(1,1)$

$E[\Theta | X]$ # of heads $\times | \Theta = \theta \sim \text{Bin}(n, \theta)$.

By Beta-Binomial Conjugacy , $\Theta | X=k \sim \text{Beta}(1+k, 1+n-k)$

We wish to estimate the probability of landing heads, denoted by θ , of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \text{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta | X)$ and LLSE $L(\Theta | X)$. 除正态分布外，某些特别也满足

$$\Rightarrow E[\Theta | X=k] = \frac{k+1}{1+k+1+n-k} = \frac{k+1}{n+2} \quad E(\Theta | X) = L(\Theta | X)$$

估值
↓
估值器

$$\Rightarrow E[\Theta | X] = \frac{X+1}{n+2}, \quad \text{MMSE}$$

Solution 2°. LLSE. $L(\theta|x) = E[\theta] + \frac{\text{cov}(\theta, x)}{\text{var}(x)}(x - E(x))$.

$$\theta \sim \text{unif}(0,1). \Rightarrow E(\theta) = \frac{1}{2}, \text{var}(\theta) = \frac{1}{12}, E(\theta^2) = \frac{1}{3}.$$

$$X|\theta=\theta \sim \text{Bin}(n, \theta) \Rightarrow E[X|\theta=\theta] = n\theta. \Rightarrow E[X|\theta] = n\theta.$$

$$\text{Var}[X|\theta=\theta] = n\theta(1-\theta). \Rightarrow \text{Var}[X|\theta] = n\theta(1-\theta)$$

$$\Rightarrow E[X] = E[E[X|\theta]] = E[n\theta] = nE[\theta] = \frac{n}{2} \quad \text{Adam}$$

$$\text{Var}(X) = E[\text{Var}(X|\theta)] + \text{Var}[E[X|\theta]]$$

$$= E[n\theta(1-\theta)] + \text{Var}[n\theta]$$

$$= n(E(\theta) - E(\theta^2)) + n^2 \text{Var}(\theta) \quad \checkmark \text{Eve}$$

$$= n\left(\frac{1}{2} - \frac{1}{3}\right) + n^2 \cdot \frac{1}{12} = \frac{n}{12}(n+2)$$

Solution

$$\begin{aligned}\text{Cov}(x, \theta) &= \underbrace{E[\theta x]}_{\Downarrow \text{Adam}} - E[\theta] \cdot E[x] \\ &= E[E[\theta x | \theta]] - E[\theta] \cdot E[x] \\ &= E[\theta \underbrace{E[x | \theta]}_{\text{}}] - \underbrace{E[\theta]}_{\text{}} \cdot \underbrace{E[x]}_{\text{}} \\ &= E[\theta \cdot n \theta] - \frac{1}{2} \cdot \frac{n}{2} \\ &= n E[\theta^2] - \frac{n}{4} \\ &= n \cdot \frac{1}{3} - \frac{n}{4} = \frac{1}{12}n\end{aligned}$$

$$\begin{aligned}\Rightarrow \text{LLSE} \quad L(\theta | x) &= E[\theta] + \frac{\text{Cov}(\theta, x)}{\text{Var}(x)} [x - E(x)] \\ &= \frac{1}{2} + \frac{\frac{1}{12}n}{\frac{n}{12}(n+2)} \left(x - \frac{n}{2} \right) = \left(\frac{x+1}{n+2} \right) = E[\theta | x]\end{aligned}$$

Solution

Remark: Statistical Learning Perspective

$$E[\theta|x]$$

- In general, MMSE is a highly nonlinear function.
- Adoption of various approximation methods leads to various learning methods
 - ▶ Linear regression
 - ▶ Logistic regression
 - ▶ Polynomial regression
 - ▶ Regression with Spline functions
 - ▶ Neural network

Outline

- 1 Overview of Statistical Inference
- 2 Classical Statistical Inference
- 3 Bayesian Statistical Inference
- 4 Conditional Expectation: Useful Tools
- 5 Prediction & Estimation
- 6 Application Case: Kalman Filter

Milestones in Statistics & Signal Processing

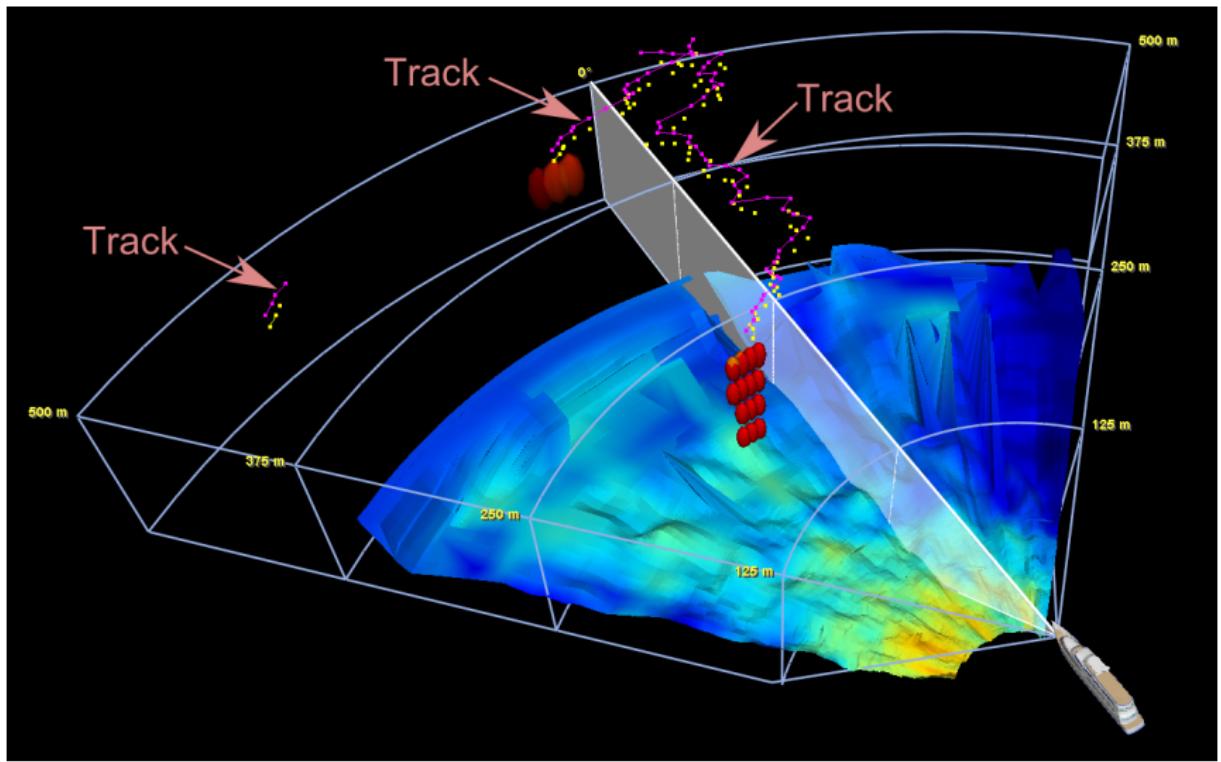
- 1960: Rudolph Emil Kalman (1930-2016) introduced what is known as Kalman filter.



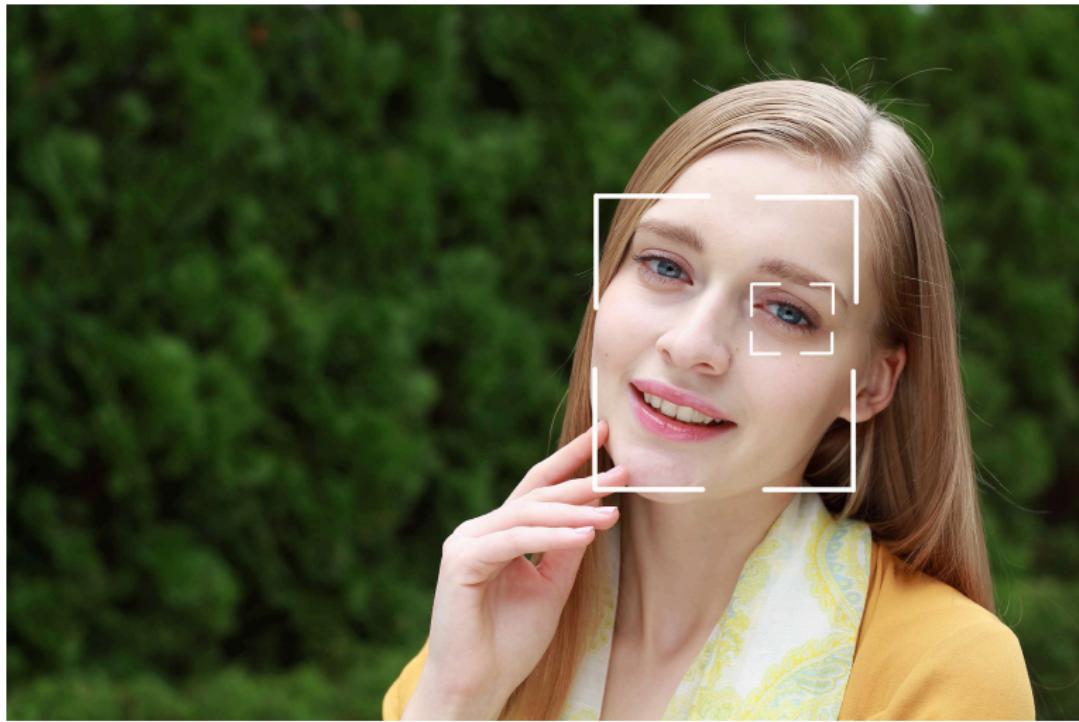
Widely Applications: Location & Navigation & Map Building



Widely Applications: Radar Tracking



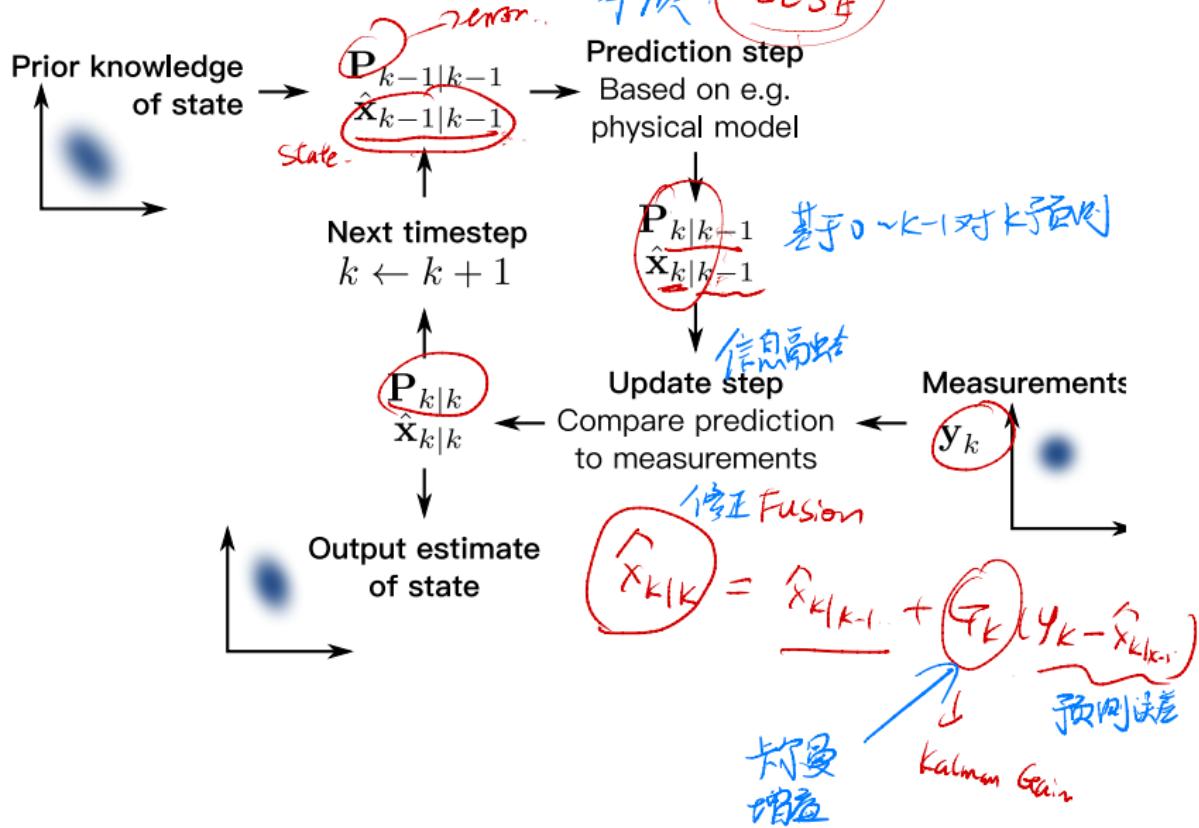
Widely Applications: Human Face & Eye Detection Autofocus



Widely Applications: Animal Eye Detection Autofocus



Essence of Kalman Filter



Essence of Kalman Filter

†

Reasons for Popularity of Kalman Filter

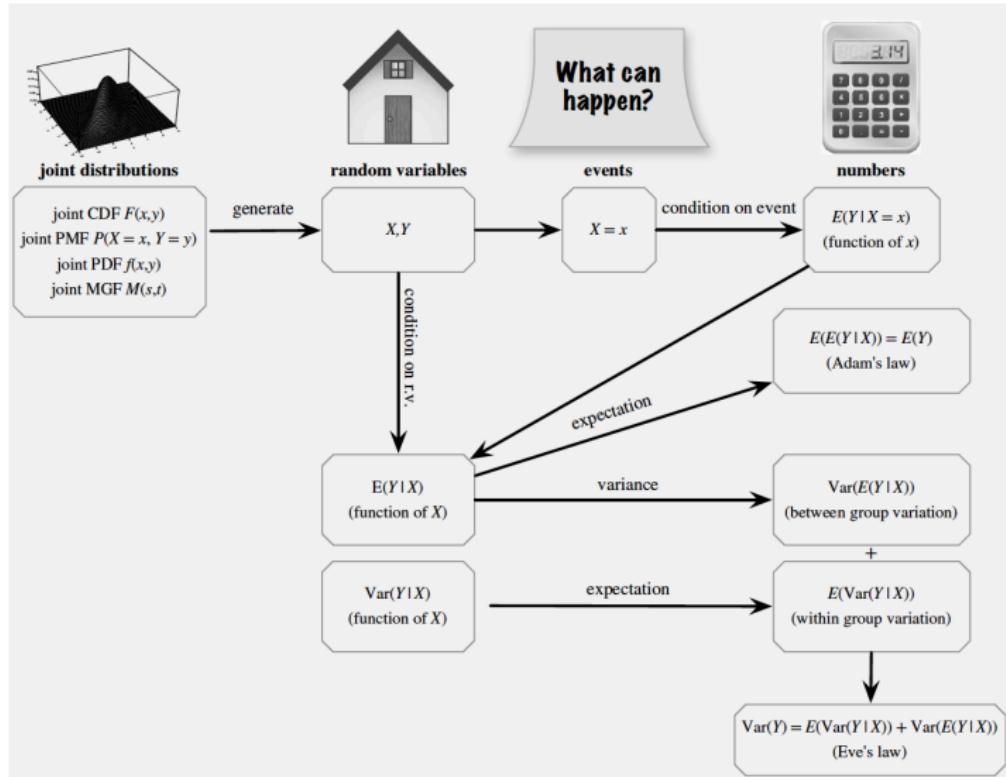
- Good results in practice due to optimality and structure: LLSE estimation in general, MMSE estimation under the setting of Gaussian noise.
Gaussian; LLSE = MMSE
- Convenient form for online real time processing: recursive equations.
- Easy to formulate and implement given a basic understanding.

Why Use The Word “Filter”

最优估值 \leftrightarrow 滤掉 noise

- The process of finding the “best estimate” from noisy data amounts to “filtering out” the noise.
- Estimation (statistical perspective) vs. Filtering (signal processing perspective)
- A Kalman filter not only cleans up the data measurements
- A Kalman filter also projects these measurements onto the state estimate

Summary 1



References

- Chapters 9 of **BH**
- Chapters 4 & 6 & 8 of **BT**

↗