Lecture 8: Conditional Expectation

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

December 10, 2024

Outline

Outline

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

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

- Bayes' Rule:

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

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

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

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

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

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

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

Conditional PDF

- 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 $A =$ "$a \le X \le b$" and $P(A) > 0$, then

$$\begin{aligned} f_{X|A}(x) &= \frac{\mathbf{1}_{x \in [a,b]}}{P(A)} \cdot f_X(x) \\ &= \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

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

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

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

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

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

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

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

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

Example

Method 1: 1°. event $A = $ "$X>1$". $P(A) = P(X>1) = \int_1^\infty \lambda e^{-\lambda x} dx = e^{-\lambda} > 0$

$$f_{X|A}(x) = \frac{f(x)}{P(A)} = \lambda e^{-\lambda(x-1)}, \quad x>1.$$

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

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

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

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

$E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$;

Method 2:

$\forall s,t>0$, $P(X>s+t \mid X>t) = P(X>s)$; $t=1$

$\Rightarrow P(X>s+1 \mid X>1) = P(X>s) \Rightarrow P(X-1>s \mid X>1) = P(X>s)$

$X-1 \mid X>1 \sim X$

1° $E[X|X>1] = E[X-1+1|X>1] = 1+E[X-1|X>1]$

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

$E(X-1|X>1) = E(X)$

2°. $Var(X|X>1) = Var(X-1|X>1)$

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

$Var(X-1|X>1) = Var(X)$

Solution

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: $E(Y|A)$
- Conditional Expectation given a random variable: $E(Y|X)$

Life Expectancy

T : Life span.

$E(T) = 70$;

$E[T \mid T \geq 20] \neq E[T]$

$T \sim \mathrm{Expo}(\cdot)$ $\Rightarrow$ $20 + E[T]$

Law of Total Expectation

LOTE

W.L.O.G. Discrete.

$E(Y|A_i) = \sum y\, P(Y=y|A_i)$

$\Rightarrow \sum_{i=1}^{n} E(Y|A_i) \cdot p(A_i)$

$= \sum_{i=1}^{n} y \cdot \sum_y P(Y=y|A_i) \cdot P(A_i)$

$= \sum_y y \cdot \sum_{i=1}^{n} P(Y=y, A_i)$

$= \sum_y y \cdot P(Y=y) = E(Y)$

LOTE $\to$ LOTP : $Y = I_B$

$\Rightarrow P(B) = E(I_B) = E(Y) \overset{LOTE}{=} \sum_{i=1}^{n} E(Y|A_i) \cdot p(A_i)$

$= \sum_{i=1}^{n} E(I_B|A_i) \cdot P(A_i) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$

Theorem

Let $A_1, ..., A_n$ 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

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

$X \sim Expo(\lambda)$; Find $E[X | X \le 1]$?

Memoryless

$X - 1 | X > 1 \sim X$

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

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

Geometric Expectation Redux

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

Coin Tosses
# of tosses before the first Head's appearance
H: Head ; T: Tail.

① First step Analysis. Conditioning on the outcome of the first toss.

$O_1 = H$ or T

② $$E(X) \overset{\text{LOTE}}{=} \underbrace{E(X \mid O_1 = H)}_{0} \cdot P(O_1 = H) + E(X \mid O_1 = T) \cdot P(O_1 = T)$$

$$= 0 \cdot p + [1 + E[X]) \cdot (1-p)$$

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

Outline

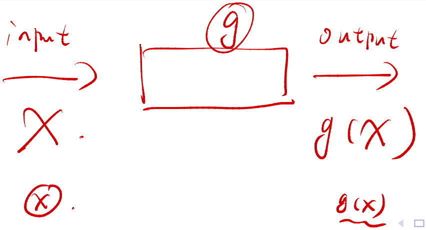
Conditional Expectation Given An R.V.

$g(x) = E[Y|X=x]$ real Number. estimate.

$g(X) = E[Y|X]$ R.V. estimator.

Definition

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



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

Example: Stick Length

① $X \sim \text{unif}(0,1)$: $Y|X=x \sim \text{unif}(0,x) \Rightarrow E[Y|X=x] = \frac{x}{2}$ $= g(x)$.

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

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

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

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

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

② $g(x) \to g(X) = E[Y|X]$

$x \to X$.

Solution

Dropping What's Independent

Theorem

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

$g(x) = E[Y|X=x] = E[Y]$, $\forall x$.

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

$\Rightarrow E[Y|X] = E[Y]$

Taking Out What's Known

1°. $g(x) = E[h(X)Y \mid X=x] = E[h(x)Y \mid X=x]$

$= h(x) \cdot E[Y \mid X=x]$

2°. $g(X) = h(X) \cdot E[Y \mid X]$

Theorem

For any function h,

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

Linearity

$g(x) = E[Y_1 + Y_2 | X = x] = E[Y_1 | X = x] + E[Y_2 | X = x]$

$\Rightarrow g(X) = E[Y_1 | X] + E[Y_2 | X]$

Theorem

$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X).$

Example

1°. By i.i.d. $E[X_1|S_n] = E[X_2|S_n] = \cdots = E[X_n|S_n]$

2°. By Linearity. $E[X_1|S_n] + E[X_2|S_n] + \cdots + E[X_n|S_n]$

$= E[X_1 + X_2 + \cdots + X_n | S_n]$

$= E[S_n|S_n] = S_n$.

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

$g(S_n) = S_n$.

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

3°. $n E[X_1|S_n] = S_n$

$\Rightarrow E[X_1|S_n] = \frac{1}{n} S_n$

Adam's Law

The Law of Iterated Expectation

The Tower Rule.

The Smoothing Theorem

Theorem

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

For any r.v.s X and Y, 1°. $g(X) = E[Y|X]$; $g(x) = E[Y|X=x]$

$$\underbrace{E(E(Y|X))}_{LHS} = \underbrace{E(Y)}_{RHS}.$$

$= \sum_y y \cdot P(Y=y|X=x)$

2°. LHS. $E[E[Y|X]] = E[g(X)] = \sum_x g(x) \cdot P(X=x)$

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

$\sum_x P(Y=y, X=x)$

$= \sum_y y \cdot P(Y=y) = E[Y]$ RHS.

Proof

Adam's Law with Extra Conditioning

$\tilde{P}(\cdot) = P(\cdot|Z)$; $\tilde{E} = E(\cdot|Z)$

Adam's Law : $E[E[Y|X]] = E[Y]$

$\tilde{E}[\tilde{E}[Y|X]] = \tilde{E}[Y]$

$\Downarrow$

$E(Y|Z)$

Theorem

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

$$E(E(Y|X,Z)|Z) = E(Y|Z)$$
$$E(E(X|Z,Y)|Y) = E(X|Y)$$

$\tilde{E}(Y|X)$

$= E(Y|X,Z)$

Conditional Variance

1°. $Var(Y) = E[(Y - E(Y))^2]$

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

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

Definition

The *conditional variance of Y given X* is

$$\mathrm{Var}(Y|X) = E\left((Y - E(Y|X))^2 \,|X\right).$$

This is equivalent to

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

2°. $Var(Y) = E(Y^2) - E^2(Y)$

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

Eve's law

Theorem

For any r.v.s X and Y,

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

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

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$$

① $g(X) = E[Y|X] \xRightarrow{\text{Adam's Law}} E[g(X)] = E[E[Y|X]] = E[Y]$

② $E[\mathrm{Var}(Y|X)] = E[\,E[Y^2|X] - (E(Y|X))^2] = E[E[Y^2|X] - g^2(X)]$

$= E[E[Y^2|X]] - E[g^2(X)] \overset{\text{Adam's Law}}{=} E[Y^2] - \cancel{E[g^2(X)]}$

③ $\mathrm{Var}[E[Y|X]] = \mathrm{Var}(g(X)) = E[g^2(X)] - E^2[g(X)]$

$= \cancel{E[g^2(X)]} - E^2(Y)$

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

$= \mathrm{Var}(Y)$

Proof

Example: Random Sum

N, X_j independent.

① $E[X]$: $E[X|N=n] = E[\sum_{j=1}^{N} X_j | N=n] = E[\sum_{j=1}^{n} X_j | N=n]$

$\Rightarrow E[X|N] = g(N) = N \cdot \mu$. $= E[\sum_{j=1}^{n} X_j] = \sum_{j=1}^{n} E[X_j] = n \cdot \mu = g(n)$

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

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 jth 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 $\text{Var}(N)$.

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

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

$\Rightarrow Var(X|N) = N \cdot \sigma^2$.

Solution

(3) By Eve's Law

$$\mathrm{Var}(X) = E[\underbrace{\mathrm{Var}(X|N)}] + \mathrm{Var}[\underbrace{E[X|N]}]$$

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

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

Solution

Outline

Basic Problem

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

- Estimate Y from the observed value X
- Choose the estimator (inference function) $g(\cdot)$ to minimize the expected error $E(c(Y, g(X)))$
- $c(Y, \hat{Y})$ 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.
- Further, if the function $g(\cdot)$ 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.
- Further, if the function $g(\cdot)$ can be arbitrary, it is called "the Minimum Mean Square Estimate (MMSE)" estimate of Y given X.

Linear Least Square Estimate

① $f(a,b) = E[(Y-a-bX)^2] = a^2 + E[Y^2] + b^2 E[X^2] - 2aE[Y] + 2abE[X] - 2bE[XY]$

$1^\circ.\ \frac{\partial f(a,b)}{\partial a} = 2a - 2E[Y] + 2bE[X] = 0 \Rightarrow a + bE[X] = E[Y]$

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,

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

$2^\circ.\ \frac{\partial f(a,b)}{\partial b} = 2bE[X^2] + 2aE[X] - 2E[XY] \Rightarrow aE[X] + bE[X^2] = E[XY]$

Proof

$$\Rightarrow b = \frac{Cov(X,Y)}{Var(X)}, \quad a = E[Y] - \frac{Cov(X,Y)}{Var(X)} \cdot E[X]$$

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

Hessian Matrix:

$$H = \begin{bmatrix} \frac{\partial^2 f(a,b)}{\partial a^2} & \frac{\partial^2 f(a,b)}{\partial a \partial b} \\ \frac{\partial^2 f(a,b)}{\partial a \partial b} & \frac{\partial^2 f(a,b)}{\partial b^2} \end{bmatrix} \succeq 0$$

$$H \succeq 0 \Leftrightarrow Z^T H Z \geq 0$$

$$H = \begin{bmatrix} 2 & 2E[X] \\ 2E[X] & 2E[X^2] \end{bmatrix} = 2\begin{bmatrix} 1 & E[X] \\ E[X] & E[X^2] \end{bmatrix}; \quad Z = (z_1, z_2)^T$$

$$Z^T H Z = (z_1, z_2) H \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 2[z_1^2 + 2z_1 z_2 E[X] + z_2^2 E[X^2]]$$

$$= 2[(z_1 + z_2 E[X])^2 + z_2^2 (E(X^2) - E^2(X))] = 2[(z_1 + z_2 E(X))^2 + z_2^2 Var(X)] \geq 0$$

Proof

Data-Driven LLSE: Linear Regression

$L_k[Y|X]$

- $L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$
- Now we only have k i.i.d samples: $(X_1, Y_1), \ldots, (X_k, Y_k)$
- Use sample mean to replace expectation

$$E(X) \leftarrow E_k(X) = \frac{1}{k}\sum_{j=1}^{k} X_k$$

$$E(Y) \leftarrow E_k(Y) = \frac{1}{k}\sum_{j=1}^{k} Y_k$$

$E \leftarrow$ Sample mean.

$k \to \infty$

SSLN ;

LR $\to$ LLSE

$$Cov(X,Y) = E(XY) - E(X)E(Y) \leftarrow \frac{1}{k}\sum_{j=1}^{k} X_k Y_k - E_k(X)E_k(Y)$$

$$Var(X) = E(X^2) - (E(X))^2 \leftarrow \frac{1}{k}\sum_{j=1}^{k} X_k^2 - (E_k(X))^2$$

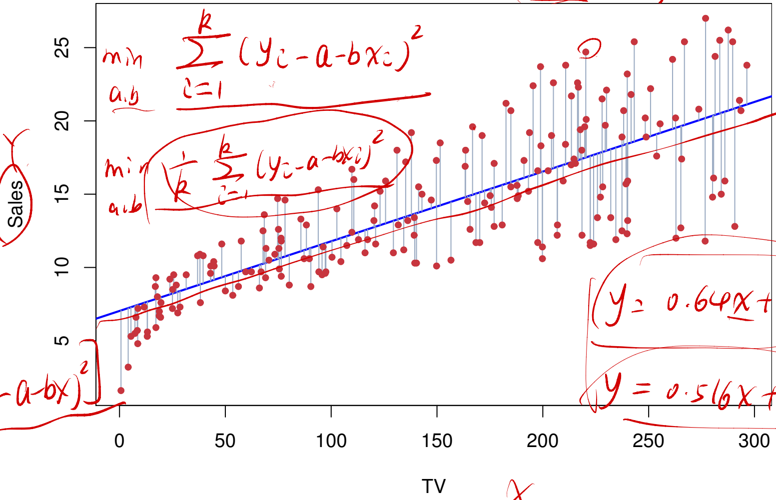
Data-Driven LLSE: Linear Regression

$y = 0.512X + 0.62$

Son's height

Father's height

(x_i, y_i) data



$$\min_{a,b} \sum_{i=1}^{k} (y_i - a - bx_i)^2$$

$$\min_{a,b} \frac{1}{k} \sum_{i=1}^{k} (y_i - a - bx_i)^2$$

$E(Y - a - bX)^2$

Y

X

$y = 0.64X + 22.37$

$y = 0.516X + 33.4$

Minimum Mean Square Error Estimator

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

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

Theorem

The MMSE of Y given X is given by

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

Geometric Perspective of Conditional Expectation

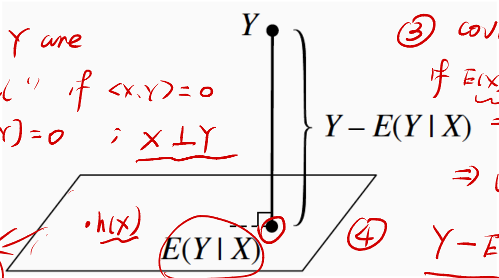
$E(X^2) < \infty$

① inner product $\langle X, Y\rangle = E[X\cdot Y]$, $\|X\| = \sqrt{\langle X, X\rangle} = \sqrt{E(X^2)}$.

$\cos\theta = \dfrac{\langle X, Y\rangle}{\|X\|\cdot\|Y\|}$; $\operatorname{dist}(X,Y) = \sqrt{\langle X-Y, X-Y\rangle} = \sqrt{E(|X-Y|^2)}$

② X and Y are "orthogonal" if $\langle X, Y\rangle = 0$

$\Leftrightarrow E[X\cdot Y] = 0$; $X \perp Y$



③ $\operatorname{cov}(X,Y) = E(X\cdot Y) - E(X)\cdot E(Y)$

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

$\Rightarrow \operatorname{cov}(X,Y) = E(X\cdot Y)$

$\Rightarrow$ uncorrelated $\Leftrightarrow$ orthogonal

④ $Y - E[Y|X] \perp h(X)$

$\forall h(\cdot)$

Space of $h(X)$ $\forall h(\cdot)$;

⑤ $E[Y|X]$: a projection of Y onto the space of $h(X)$ $\forall h(\cdot)$

⑥ $L[Y|X]$: a projection of Y onto the $\mathcal{L}(X) = \{a+bX,\ a,b\in\mathbb{R}\}$

LLSE

Projection Interpretation

1°. $E[Y - E[Y|X]] = E[Y] - E[E[Y|X]]$

$= E[Y] - E[Y]$

$= 0$

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

Theorem

uncorrelated $\Leftrightarrow$ orthogonal

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

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

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

2°. Show $Y - E[Y|X] \perp h(X)$

Proof

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

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

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

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

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

Adam's Law

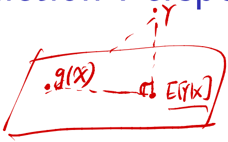
$$= E[Y \cdot h(X)] - E[h(X) \cdot Y]$$

$$= 0$$

Proof

Prediction Perspective

$$E[(Y-E[Y|X])^2] \le E[(Y-\overset{\hat{Y}}{g(X)})^2] \quad \forall g(\cdot)$$



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

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

Proof

1°. $\hat{Y}$: estimator of Y based on X $(\hat{Y} = g(X))$

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

2°. $$E[(Y-g(X))^2] = E[(A+B)^2] = E[A^2] + 2E[AB] + E[B^2]$$

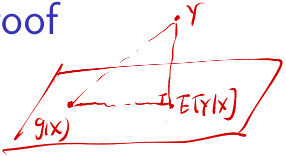
$$= E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2] + \cdots$$

$$2E[(Y-E[Y|X])\cdot(E[Y|X]-g(X))]$$

3°. $h(X) \triangleq E[Y|X] - g(X)$; $Y - E[Y|X] \perp h(X)$ $\Rightarrow E[(Y-E[Y|X])\cdot h(X)] = 0$

$$\Rightarrow E[(Y-g(X))^2] = E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2]$$

Proof



$$\mathrm{dist}^2(Y, E[Y|X]) + \mathrm{dist}^2(g(X), E[Y|X])$$
$$= \mathrm{dist}^2(Y, g(X))$$

5°. $E[(Y-\hat{Y})^2] = E[(Y-g(X))^2]$

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

≥ 0

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

$$\hat{Y}^* = g^*(X) = \underline{E[Y|X]}$$

MMSE

Proof

MMSE for Jointly Normal Random Variables

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

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

Theorem

Let X, Y be jointly Normal random variables. Then

σ_Y^2 σ_X^2

MMSE LLSE

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

$$\Rightarrow E[Y|X] - E[Y] = \frac{\rho\sigma_X\sigma_Y}{\sigma_X^2}(X - E[X])$$

$$\Rightarrow \boxed{\frac{E[Y|X] - E[Y]}{\sigma_Y} = \rho \cdot \frac{X - E[X]}{\sigma_X}} \Rightarrow$$

X : parent's height.

Y : child's height.

$X - E[X] = t\sigma_X$

Remark: Statistical Learning Perspective

X : Parent's height

Y : Child's height

$\rho > 0$

$0 < \rho < 1$

$X - E[X] = t\sigma_X$

$\Rightarrow \frac{E[Y|X] - E[Y]}{\sigma_Y} = \rho t < t \Rightarrow E[Y|X] - E[Y] < t\sigma_Y$

- 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

$g(x) = a + bx$

$g(x) = \frac{1}{1+e^{-(a+bx)}}$

$\log \frac{g(x)}{1-g(x)} = a + bx$

logit

$g(x) = a + bx + cx^2 + dx^3 + \cdots + fx^6$

$g(x)$

piecewise polynomial

$g(x)$

regression towards Mean

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

regression function

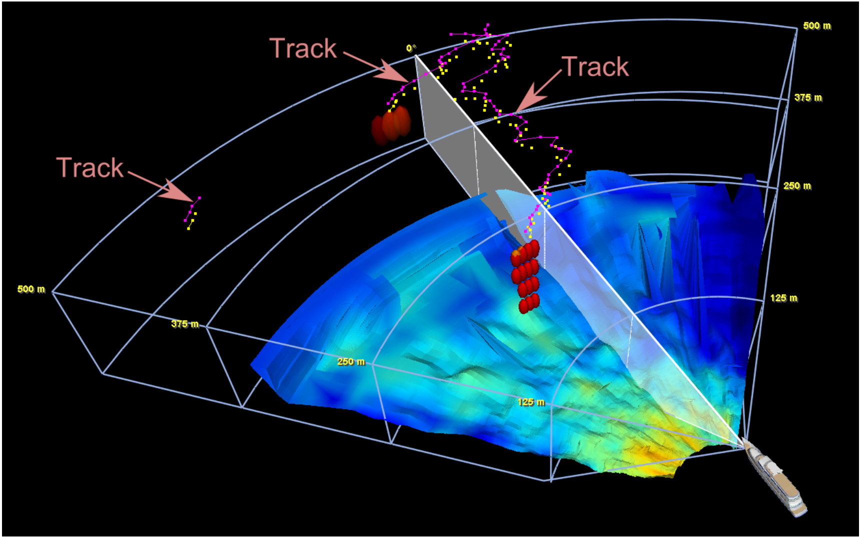
Outline

Milestones in Statistics & Signal Processing

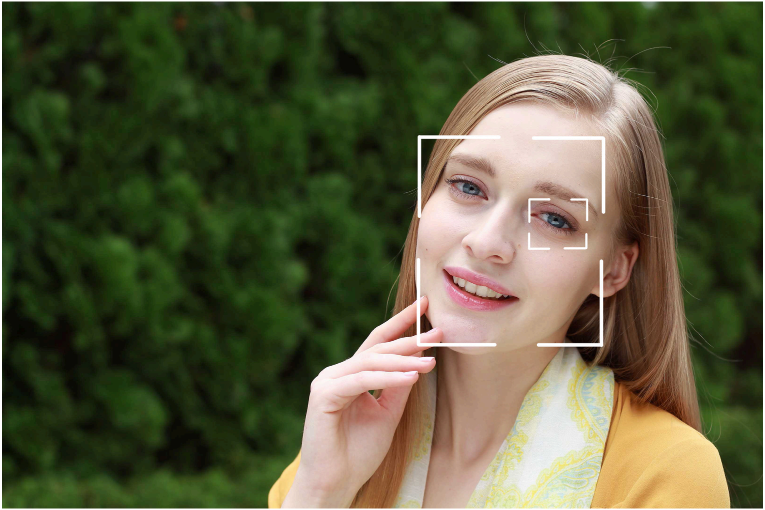
- 1960: Rudoph Emil Kalman (1930-2016) introduced what is known as Kalman filter.

Widely Applications: Location & Navigation & Map Building

Widely Applications: Radar Tracking



Widely Applications: Human Face & Eye Detection Autofocus

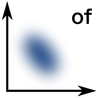


Widely Applications: Animal Eye Detection Autofocus

Essence of Kalman Filter

Online LLSE

(MMSE)



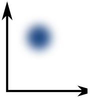
Prior knowledge of state → $\mathbf{P}_{k-1|k-1}$, $\hat{\mathbf{x}}_{k-1|k-1}$ → **Prediction step** Based on e.g. physical model

↓

$\mathbf{P}_{k|k-1}$, $\hat{\mathbf{x}}_{k|k-1}$

↓

Update step Compare prediction to measurements ← **Measurements** $\mathbf{y}_k$

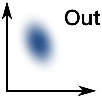


← $\mathbf{P}_{k|k}$, $\hat{\mathbf{x}}_{k|k}$

↑ **Next timestep** $k \leftarrow k+1$ ↑ $\mathbf{P}_{k-1|k-1}$, $\hat{\mathbf{x}}_{k-1|k-1}$

$\mathbf{P}_{k|k}$, $\hat{\mathbf{x}}_{k|k}$ ↓

Output estimate of state



n : index of time.

$$M_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

offline

$$nM_n = (X_1 + \cdots + X_{n-1}) + X_n$$

$$= (n-1)M_{n-1} + X_n$$

$$\Rightarrow M_n = \left(1 - \frac{1}{n}\right)M_{n-1} + \frac{1}{n}X_n$$

Online

$$\hat{S}_{n|n} = (1-k_n)\hat{S}_{n|n-1} + k_n X_n$$

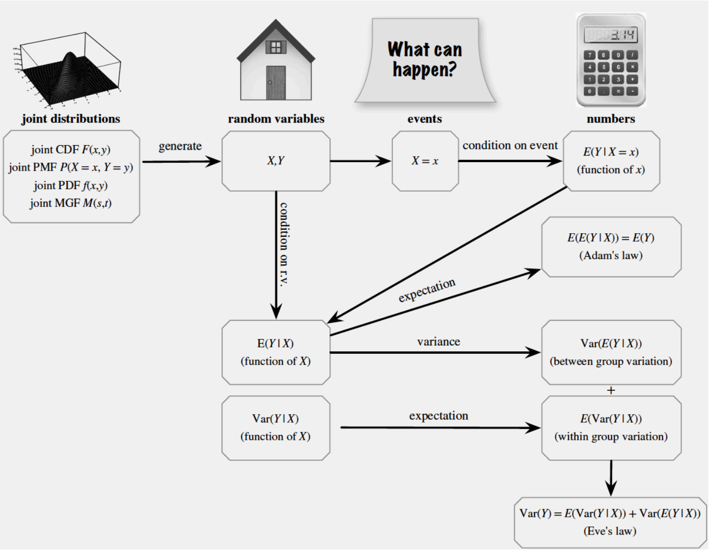
Reasons for Popularity of Kalman Filter

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

Why Use The Word "Filter"

- 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



What can happen?
joint distributions
random variables
events
numbers
joint CDF F(x,y)
joint PMF P(X = x, Y = y)
joint PDF f(x,y)
joint MGF M(s,t)
generate
X,Y
X = x
condition on event
E(Y | X = x)
(function of x)
condition on r.v.
E(E(Y | X)) = E(Y)
(Adam's law)
expectation
E(Y | X)
(function of X)
variance
Var(E(Y | X))
(between group variation)
+
Var(Y | X)
(function of X)
expectation
E(Var(Y | X))
(within group variation)
Var(Y) = E(Var(Y | X)) + Var(E(Y | X))
(Eve's law)

References

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