ORF 418 : OPTIMAL LEARNING

LECTURE 9 : October 1, 2025

ESTINATION

0) Conditional Expectation : Gaussian

1) Bayesian Estimator

2) Example

3) Maximum Likelihood Estimator

1, Gaussian Conditional Expectations.

We assume $Z = (X, Y) \in \mathbb{R}^k \times \mathbb{R}^d$ is jointly Gaussian and want to compute $f_{X|Y}$ is terms of means of X, Y and their covariances. General formulae are given in the lecture notes. Here we consider a simple example.

Example: We take $k = d = 1$, $\mu_X = 1, \mu_Y = 2$ and

$$\mathrm{cov}(Z) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} =: \Sigma$$

This implies that $\sigma_X = 1$, $\sigma_Y = 2$ and the correlation between X and Y is $\rho_{X,Y} = 1/2$. Also,

$$\Sigma^{-1} = \frac{1}{3} \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}.$$

Then,

$$f_Z(x,y) = (2\pi \det(\Sigma))^{-1} \exp\left(-\tfrac{1}{2} Q(x,y)\right)$$

where $\det(\Sigma) = 3$ and

$$Q(x,y) = (x-1, y-2)\, \Sigma^{-1} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix}$$

$$= \frac{1}{3}\left[4(x-1)^2 + (y-2)^2 - 2(x-1)(y-2)\right]$$

Also, we know that

$$f_X(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x-1)^2\right),$$

$$f_Y(y) = (8\pi)^{-1/2} \exp\left(-\frac{1}{8}(y-2)^2\right).$$

By definition

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

$$= \text{constant} \exp\left(-\frac{1}{2}Q(x,y) + \frac{1}{8}(y-2)^2\right)$$

Note that

$$Q(x,y) = \frac{4}{3}\left[(x-1)^2 - \frac{1}{2}(x-1)^2(y-2)\right] + \frac{1}{3}(y-2)^2$$

$$= \frac{4}{3}\left[(x-1) - \frac{1}{4}(y-2)\right]^2 - \frac{1}{12}(y-2)^2 + \frac{1}{3}(y-2)^2$$

$$= \frac{4}{3}\left(x - \left[\frac{1}{2} + \frac{1}{4}y\right]\right)^2 + c(y).$$

Hence,

$$f_{X|Y}(x|y) = \text{constant}(y) \exp\left(-\frac{2}{3}\left(x - \left(\frac{1}{2} + \frac{1}{4}y\right)\right)^2\right).$$

Therefore, $f_{X|Y}$ is also Gaussian and

$$\mu_{X|Y} = \frac{1}{2} + \frac{1}{4}y$$

$$\text{var}(X|Y) = \frac{3}{4}.$$

I. BAYESIAN ESTIMATOR

Suppose we want to estimate $h(X)$ after an observation $\{Y=y\}$, where X and Y are random variables and $h(\cdot)$ is a given nonlinear function. The procedure is simple:

1. Compute the posterior distribution of X given $\{Y=y\}$.
2. Use this posterior to compute the conditional expectation of $h(X)$ given $\{Y=y\}$ and set

$$\boxed{\hat{h}(X) := \mathbb{E}[h(X) \mid Y=y]}$$

To compute the above conditional expectation, we use the Bayes formulae which requires us to have a prior distribution for X.

We detail several seperate cases:

(i) X, Y are continuous random variables.

Suppose we know $f_X(x)$, $f_{Y|X}(y|x)$. Then,

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

and $$\hat{h}(X) = \int_{-\infty}^{\infty} h(x) f_{X|Y}(x|y) \, dx .$$

(i) X continuous and Y discrete.

Suppose we know $f_X(x)$, $\mathbb{P}(Y=y|x)$. Then,

$$f_{X|Y}(x|y) = \frac{\mathbb{P}(Y=y|x)}{\int_{-\infty}^{\infty} \mathbb{P}(Y=y|x') f_X(x') \, dx'}$$

and $$\hat{h}(X) = \int_{-\infty}^{\infty} h(x) f_{X|Y}(x|y) \, dx .$$

(ii) X is discrete and Y is continuous

Suppose we know $\mathbb{P}(X=x)$, $f_{Y|X}(y|x)$. Then,

$$\mathbb{P}(X=x|Y=y) = \frac{f_{Y|X}(y|x) \, \mathbb{P}(X=x)}{\sum_{x'} f_{Y|X}(y|x') \, \mathbb{P}(X=x')}$$

and $$\hat{h}(X) = \sum_{x} h(x) \, \mathbb{P}(X=x|Y=y) .$$

(iii) X and Y are discrete.

Suppose we know $\mathbb{P}(X=x)$, $\mathbb{P}(Y=y \mid X=x)$. Then,

$$\mathbb{P}(X=x \mid Y=y) = \frac{\mathbb{P}(Y=y \mid X=x)\,\mathbb{P}(X=x)}{\sum_{x'} \mathbb{P}(Y=y \mid X=x')\,\mathbb{P}(X=x')}$$

and

$$\hat{h}(X) = \sum_{x} h(x)\,\mathbb{P}(X=x \mid Y=y).$$

II. EXAMPLES

(a) Gaussian Suppose X and Y are jointly Gaussian with

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \quad \mu_X = 1, \quad \mu_Y = 2.$$

Hence, $\mathrm{var}(X) = \sigma_X^2 = 1$, $\mathrm{var}(Y) = \sigma_Y^2 = 4$ and

$$\rho = \mathrm{correlation}(X,Y) = \frac{\sigma_{X,Y}^2}{\sigma_X \sigma_Y} = \frac{1}{1 \cdot 2} = \frac{1}{2}$$

Then, $\sigma_{X|Y}^2 = \mathrm{var}(X \mid Y=2) = \sigma_X^2 (1-\rho^2) = \frac{3}{4}$, and

$$\mathbb{E}[X \mid Y=2] = \mu_X + \frac{\rho\,\sigma_X}{\sigma_Y}(2-\mu_Y) = 1.$$

Hence

$$f_{X|Y}(x|2) = \frac{1}{\sqrt{2\pi}\,\sigma_{X|Y}} \exp\left(-\frac{(x-1)^2}{2\,\sigma_{X|Y}^2}\right)$$

$$= \sqrt{\frac{2}{3\pi}} \exp\left(-\frac{2(x-1)^2}{3}\right).$$

(a) $h(x) = x$

$$\hat{X} = \int_{-\infty}^{\infty} x f_{X|Y}(x|2)\,dx = 1 = \mathbb{E}[X|Y=2].$$

(b) $h(x) = x^2$

$$\widehat{X^2} = \int_{-\infty}^{\infty} x^2 f_{X|Y}(x|2)\,dx$$

$$= \mathbb{E}[X^2|Y=2] = \mathbb{E}[(X-1)^2|Y=2] + 1$$

$$= \mathrm{Var}(X|Y=2) + 1 = \frac{3}{4} + 1 = \frac{7}{4} \neq (\hat{X})^2 = 1.$$

---

b) Coin Toss. From last lecture (Example 3.4 in the Lecture Notes). 100 coin-flips with 52 heads and P is the probability of a Head taken as a random variable. We assume that prior on P is uniform. Then it is show that posterior is a Beta distribution with $a=53$ and $b=49$. It is known that the mean of this distribution is $a/a+b = 53/102$.

So

$$\hat{X} = \frac{53}{102}.$$

General Bayesian Estimation. See the lecture notes.

Definition is

$$\widehat{h(X)} = \underset{a}{\operatorname{argmin}}\ \mathbb{E}\left[\, \ell(a, h(X)) \mid Y=y \right]$$

where ℓ is a loss function. We treated the case

$$\ell(a, h(X)) = |a - h(X)|^2 .$$

III. MAXIMUM LIKELIHOOD ESTIMATOR (MLE)

Here the task is the same: find an estimate of $h(X)$ after observing the event $\{Y=y\}$. Instead of a prior distribution we postulate a parametric likelihood function $L(x,y)$. In all our examples, likelihood function is the conditional density or probability of Y given X. Then, the MLE is the maximizer of the map $x \mapsto L(x,y)$.

example. (Coin Flip) Here P the head probability is like X in above discussions and Y is the number of heads. We observe the event $\{Y=52\}$.

The conditional expectation is

$$L(p,y) = \mathbb{P}(Y=y\,;p) = \binom{100}{y} p^{y} (1-p)^{100-y}.$$

Then, by definition the estimate given $\{Y=52\}$ is

$$\hat{p} = \underset{p}{\operatorname{argmax}} \binom{100}{52} p^{52} (1-p)^{48}$$

Since $\ln(\cdot)$ is an increasing function,

$$\underset{p}{\operatorname{argmax}}\, L(p,52) = \underset{p}{\operatorname{argmax}}\, \ln(L(p,52)).$$

Hence,

$$\hat{p} = \operatorname{argmax}\; \ln\binom{100}{52} + \underbrace{52\ln(p) + 48\ln(1-p)}_{=:\, \mathcal{L}(p),\ \text{log-likelihood.}}$$

To find the maximizer we differentiate:

$$0 = \mathcal{L}'(p) = \frac{52}{p} - \frac{48}{1-p} \Rightarrow 52(1-p) = 48p \Rightarrow \hat{p} = \frac{52}{100}$$

No prior $f_P(p)$ is used.