

# Bayes Estimation

9/21/2023

## Outline

- 1) Bayes risk, Bayes estimator
- 2) Examples
- 3) Conjugate priors

# Frequentist Motivation

Model  $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$  for data  $X$

Loss  $L(\theta, d)$ , Risk  $R(\theta; \delta) = \mathbb{E}_\theta[L(\theta, \delta(x))]$

The Bayes risk is the average-case risk, integrated wrt some measure  $\Lambda$ , called prior

For now, assume  $\Lambda(\Theta) = 1$  (prob. meas.)

Later we will allow to be improper ( $\Lambda(\Theta) = \infty$ )

- Note:
- $\Lambda$  and  $c\Lambda$  for  $c > 0$  functionally equiv.
  - avg risk makes sense even if we don't "believe"  $\theta \sim \Lambda$

$$\begin{aligned} r_\Lambda(\delta) &= \int_{\Theta} R(\theta, \delta) d\Lambda(\theta) \\ &= \mathbb{E}_{\theta \sim \Lambda} \left[ \underbrace{R(\theta, \delta)}_{\mathbb{E}[L(\theta, \delta(x)) | \theta]} \right] \text{ where } \theta \sim \Lambda \\ &= \mathbb{E} \left[ L(\theta, \delta(x)) \right] \text{ where } \theta \sim \Lambda \\ &\quad \underbrace{x | \theta \sim P_\theta} \end{aligned}$$

$\mathbb{E}$  now means wrt joint distr. of  $(\theta, x)$

An estimator  $\delta$  minimizing  $R_{\text{Bayes}}(\Lambda; \cdot)$  is called Bayes (a Bayes estimator). Dep. on  $\mathcal{F}, \Lambda, L$

$$r_\Lambda(\delta) = \mathbb{E} \left[ \mathbb{E} \left[ \underbrace{L(\theta, \delta(x))}_{\text{we get to choose this!}} \mid x \right] \right]$$

we get to choose this!

# Prior, Posterior

Usual interp. of  $\Delta$  is "prior belief about  $\theta$  before seeing the data"

Conditional dist.  $\Delta(\theta | X)$  called posterior dist.  
"belief after seeing the data"

Epistemic uncertainty:

"I think there is a 50% chance that..."

More on this next time

Densities: prior  $\lambda(\theta)$ , likelihood  $p_{\theta}(x)$   
or  $p(x|\theta)$

Joint density  $\lambda(\theta) p_{\theta}(x)$

Marginal density  $q(x) = \int_{\Theta} \lambda(\theta) p_{\theta}(x) d\theta$

Posterior density  $\lambda(\theta|x) = \frac{\lambda(\theta) p_{\theta}(x)}{q(x)}$

Bayes estimator depends on posterior:

$$\delta_{\Delta}(x) = \operatorname{argmin}_d \mathbb{E}[L(\theta, d) | X]$$

$$= \operatorname{argmin}_d \int_{\Theta} L(\theta, d) \lambda(\theta|x) d\theta$$

Solve for Bayes estimator "one  $x$  at a time"

# Bayes Estimator

Suppose  $X/\theta \sim P_\theta$ ,  $L(\theta, d) \geq 0$

$r_\Delta(\delta_0) < \infty$  for some  $\delta_0(x)$

Then  $\delta_\Delta(x)$  is Bayes with  $r_\Delta(\delta_\Delta) < \infty$

iff  $\delta_\Delta(x) \in \operatorname{argmin}_d \mathbb{E}[L(\theta, d) | X=x]$  a.e.  $x$

$\mathbb{P}(\delta_\Delta(x) \in \operatorname{argmin}) = 0$

Proof ( $\Rightarrow$ ) Let  $\delta$  be any other estimator

$$r_\Delta(\delta) = \mathbb{E} \left[ \mathbb{E} [L(\theta, \delta(x)) | X=x] \right]$$

$$\geq \mathbb{E} \left[ \mathbb{E} [L(\theta, \delta_\Delta(x)) | X=x] \right]$$

$$= r_\Delta(\delta_\Delta)$$

$$< \infty \quad (\text{take } \delta = \delta_0)$$

( $\Leftarrow$ ) Define  $E_x(d) = \mathbb{E}[L(\theta, d) | X=x]$

$$\text{Let } \delta^*(x) = \begin{cases} \delta_\Delta(x) & \text{if } \delta_\Delta(x) \in \operatorname{argmin} E_x \\ \delta_0(x) & \text{if } E_x(\delta_0(x)) < E_x(\delta_\Delta(x)) \\ d^*(x) & \text{otherwise, where } E_x(d^*) < E_x(\delta_\Delta(x)) \end{cases}$$

Then  $E_x(\delta^*(x)) \leq \min(E_x(\delta_0(x)), E_x(\delta_\Delta(x))) \quad \forall x$

with ineq. strict on a set of measure  $> 0$ .  $\square$

## Posterior Mean

If  $L(\theta, d) = (g(\theta) - d)^2$  then the Bayes estimator is the posterior mean:

$$\begin{aligned}\mathbb{E}\left[(g(\theta) - d)^2 \mid X\right] \\ &= \mathbb{E}\left[\left(g(\theta) - \mathbb{E}[g(\theta) \mid X] + \mathbb{E}[g(\theta) \mid X] - d\right)^2 \mid X\right] \\ &= \text{Var}(g(\theta) \mid X) + \left(\mathbb{E}[g(\theta) \mid X] - d\right)^2\end{aligned}$$

(why is the cross-term 0?)

$$\Rightarrow \delta_{\Delta}(x) = \mathbb{E}[g(\theta) \mid X=x]$$

Weighted sq. error:

$$L(\theta, d) = w(\theta) (g(\theta) - d)^2 \quad \text{e.g. } \left(\frac{\theta-d}{\theta}\right)^2$$

sq. rel. error

$$\begin{aligned}\mathbb{E}\left[(d - g(\theta))^2 w(\theta) \mid X\right] \\ &= d^2 \mathbb{E}[w(\theta) \mid X] - 2d \mathbb{E}[w(\theta)g(\theta) \mid X] \\ &\quad + \mathbb{E}[w(\theta)g(\theta)^2 \mid X] \quad \text{no dep. on } d\end{aligned}$$

$$\text{min at } d = \frac{\mathbb{E}[w(\theta)g(\theta) \mid X]}{\mathbb{E}[w(\theta) \mid X]} \quad (= \delta_{\Delta}(x))$$

# Example: Beta-Binomial

$$X | \theta \sim \text{Binom}(n, \theta) = \theta^x (1-\theta)^{n-x} \binom{n}{x}$$

$$\theta \sim \text{Beta}(\alpha, \beta) = \theta^{\alpha-1} (1-\theta)^{\beta-1} \underbrace{\frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha+\beta)}}_{\text{normalizing const.}}$$

$\theta$  is r.v. here

Marginal dist. of  $X$  called Beta-Binomial

Posterior:

$$\lambda(\theta | x) = \lambda(\theta) p_{\theta}(x) / q(x)$$

we can drop factors that don't depend on  $\theta$   $\rightarrow$

$$\propto_{\theta} \theta^{\alpha-1} (1-\theta)^{\beta-1} \theta^x (1-\theta)^{n-x}$$
$$= \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}$$

$$\Rightarrow \theta | X=x \sim \text{Beta}(x+\alpha, n-x+\beta)$$

$$\mathbb{E}[\theta | x] = \frac{x+\alpha}{n+\alpha+\beta}$$

convex combo of  $\frac{x}{n}$ ,  $\frac{\alpha}{\alpha+\beta}$   $\rightarrow$

$$\overset{\text{UMVU}}{=} \frac{x}{n} \cdot \frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \cdot \frac{\alpha+\beta}{n+\alpha+\beta}$$

Prior Expectation  $\left(1 - \frac{n}{n+\alpha+\beta}\right)$

Interp.:  $k = \alpha + \beta$  "pseudo-trials,"  $\alpha$  successes

(Recall  $\frac{x+3}{n+6}$  from Lec. 2)

## Example: Normal mean

$$X|\theta \sim N(\theta, \sigma^2) \propto_{\theta} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

$$\sim N(\mu, \tau^2) \propto_{\theta} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

$$\lambda(\theta|x) \propto_{\theta} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\tau^2}\right\}$$

$$\propto_{\theta} \exp\left\{\frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{\theta^2}{2\tau^2} + \frac{\theta\mu}{\tau^2}\right\}$$

$$= \exp\left\{\theta\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \theta^2\left(\frac{\sigma^{-2} + \tau^{-2}}{2}\right)\right\}$$

Complete square:

$$a^2\theta^2 - b\theta = \left(\theta a - \frac{b}{2a}\right)^2 - c(a, b)$$

$$= \left(\theta - \frac{b}{2a^2}\right)^2 \cdot a^2 - c$$

$$\rightarrow \propto_{\theta} \exp\left\{-\left(\theta - \frac{x\sigma^{-2} + \mu\tau^{-2}}{\sigma^{-2} + \tau^{-2}}\right)^2 / 2(\sigma^{-2} + \tau^{-2})^{-1}\right\}$$

$$\propto_{\theta} N\left(\frac{x\sigma^{-2} + \mu\tau^{-2}}{\sigma^{-2} + \tau^{-2}}, \frac{1}{\sigma^{-2} + \tau^{-2}}\right)$$

precision-weighted  
average of  $x, \mu$

harmonic mean  
of  $\sigma^2, \tau^2$

$$\mathbb{E}[\theta|x] = x \cdot \frac{\sigma^{-2}}{\sigma^{-2} + \tau^{-2}} + \mu \cdot \frac{\tau^{-2}}{\sigma^{-2} + \tau^{-2}}$$

## Gaussian iid sample

$$\theta \sim N(\mu, \tau^2), \quad X_i | \theta \stackrel{iid}{\sim} N(\theta, \sigma^2), \quad i=1, \dots, n$$

$$\bar{X} | \theta \sim N(\theta, \frac{\sigma^2}{n})$$

$$\Rightarrow \mathbb{E}[\theta | X] = X \cdot \frac{n\sigma^{-2}}{n\sigma^{-2} + \tau^{-2}} + \mu \cdot \frac{\tau^{-2}}{n\sigma^{-2} + \tau^{-2}}$$

$$= X \cdot \frac{n}{n + \sigma^2/\tau^2} + \mu \cdot \frac{\sigma^2/\tau^2}{n + \sigma^2/\tau^2}$$

Interp:  $k = \sigma^2/\tau^2$  pseudo-observations, mean  $\mu$

If  $n \gg k$ , "data swamps prior"

If  $n \ll k$ , "prior swamps data"

[ Note in both examples :

- Prior & Likelihood have similar form
- Posterior comes from same exp. fam. as prior ]

If the posterior is from the same family as the prior, we say the prior is conjugate to the likelihood.

Most common in exp. families



# Conjugate Priors

Suppose  $X_i | \eta \stackrel{\text{iid}}{\sim} p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$   $\eta \in \Xi \subseteq \mathbb{R}^s$   
 $i=1, \dots, n$

For carrier  $\lambda_0(\eta)$ , define s+1-dim family:

$$\lambda_{k, \mu, k}(\eta) = e^{k\mu' \eta - kA(\eta) - B(k, \mu, k)} \lambda_0(\eta)$$

Suff. stat  $\begin{pmatrix} \eta \\ -A(\eta) \end{pmatrix} \in \mathbb{R}^{s+1}$  Nat. param.  $\begin{pmatrix} k\mu \\ k \end{pmatrix}$

$$\begin{aligned} \Rightarrow \lambda(\eta | x_1, \dots, x_n) &\propto_\eta \left( \prod_{i=1}^n e^{\eta' T(x_i) - A(\eta)} h(x_i) \right) \\ &\cdot e^{k\mu' \eta - kA(\eta) - B(k, \mu, k)} \lambda_0(\eta) \\ &= \alpha_\eta e^{(k\mu + \sum T(x_i))' \eta - (k+n)A(\eta)} \lambda_0(\eta) \\ &= \lambda_{k\mu + n\bar{T}, k+n}(\eta) \end{aligned}$$

where  $\bar{T}(x) = \frac{1}{n} \sum_{i=1}^n T(x_i)$

Often  $\frac{k\mu + n\bar{T}}{k+n}$  is Bayes est. for  $\mathbb{E}_\eta T$ ,

then

$$\hat{\mu}_{\text{post}} = \bar{T} \cdot \frac{n}{k+n} + \mu \cdot \frac{k}{k+n}$$

$\uparrow$   $\uparrow$   
 UMVUE from data "UMVUE" from "pseudo data"

# Conjugate Prior Examples

Likelihood	Prior
$X_i   \theta \sim \text{Binom}(n, \theta)$ $= \theta^x (1-\theta)^{n-x} \binom{n}{x}$	$\theta \sim \text{Beta}(\alpha, \beta)$ $= \theta^{\alpha-1} (1-\theta)^{\beta-1} \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha+\beta)}$
$X_i   \theta \sim N(\theta, \sigma^2) \quad (\sigma^2 \text{ known})$ $= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-x)^2}{2\sigma^2}}$	$\theta \sim N(\mu, \tau^2)$ $= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$
$X_i   \theta \sim \text{Pois}(\theta) \quad x=0,1,\dots$ $= \frac{\theta^x e^{-\theta}}{x!}$	$\theta \sim \text{Gamma}(v, s) \quad \theta > 0$ $= \frac{1}{\Gamma(v) s^v} \theta^{v-1} e^{-\theta/s}$

## Gamma / Poisson :

$$\lambda(\theta | x) \propto_{\theta} \theta^{v-1 + \sum x_i} e^{-(s' + n)\theta}$$

$$= \text{Gamma}(v + \sum x_i, (s' + n)^{-1})$$

$$\Rightarrow k = s^{-1}, \quad \mu = vs$$

$$\lambda_0(\theta) = \theta^{-1} \quad (\text{not normalizable})$$