

Outline

9/26/2023

- 1) Interpretations of Probability
- 2) Where does prior come from?
- 3) Examples

Interpretations of Probability

Why do we model anything as random?

What does "probability" mean in the real world?

1) Long-run frequency over repeated trials

Ex. repeatedly flipping a coin

shooting electrons at a double slit

"you can never step into the same river twice"

2) Systematic random sampling from a population

Ex. survey of 500 random voters

random assignment to treatment/control

Applies in controlled experiments

3) Subjective uncertainty about an outcome

chance that...

- President Biden is re-elected

- Higgs boson has a given mass

- $P = NP$

- 100th digit of π is 5

Could be broad intersubjective agreement

These are often intertwined:

Ex. What if survey sampling is pseudo-random?

Probably relying on shared ignorance

Where does Λ come from?

(Bayesian rejoinder: Where does \mathcal{P} come from?)

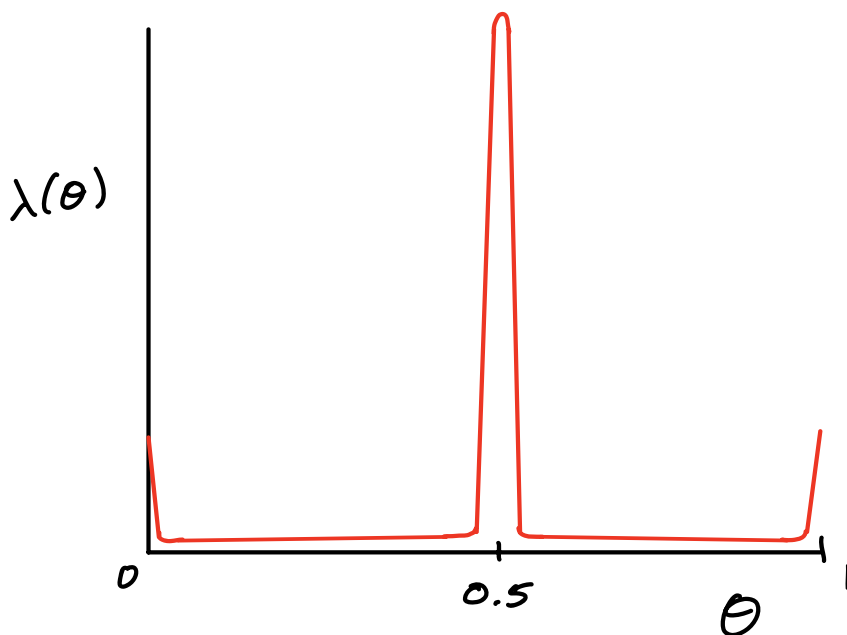
Four main sources for prior on θ

Source #1: Subjective beliefs

Pro: Brings all relevant info. to bear
Straightforward interp. of posterior

Con: Posterior is therefore subjective
Embarrassing to write "I think" in abstract
Hard if θ high-dim or \mathcal{P} nonparametric

Ex: Flip coin 20 times, get 7 heads
0.5 probably a better estimate than 0.35
My subjective prior on coins:



Source #2: "Objective" or "vague" prior

Using default prior removes subjectivity

(But then what does the posterior mean?)

Flat prior $\lambda(\theta) \propto_{\theta} 1$ on Θ

"Indifference" (in θ parameterization)

Often improper ($\int \lambda(\theta) = \infty$) but usually ok

Ex: $\theta \sim$ flat prior on \mathbb{R}

$$X|\theta \sim N(\theta, \sigma^2)$$

$$\begin{aligned}\lambda(\theta|x) &\propto_{\theta} p_{\theta}(x) \\ &= \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2\sigma^2} \\ &\propto_{\theta} N(x, \sigma^2)\end{aligned}$$

Jeffreys prior $\lambda(\theta) \propto_{\theta} |J(\theta)|^{1/2}$

Higher density where P_{θ} "changing faster"

Invariant to parameterization (HW 5)

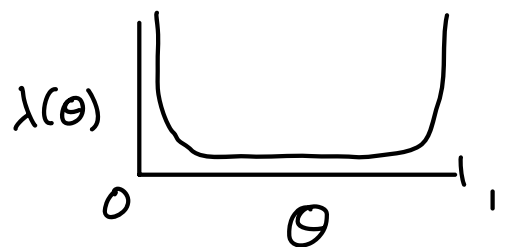
Ex. $X|\theta \sim \text{Binom}(n, \theta)$

$$\lambda(\theta) \propto_{\theta} J(\theta)^{1/2} = \left(\frac{n}{\theta(1-\theta)}\right)^{1/2} \propto_{\theta} \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$\lambda(\theta) \rightarrow \infty$ as $\theta \rightarrow 0$ or 1 :

$$D_{KL}(0.001 || 0.01) \stackrel{(33x)}{\gg} D_{KL}(0.49 || 0.5)$$

$7n \cdot 10^{-3}$ $2n \cdot 10^{-4}$



Intersubjective Agreement

Data may effectively rule out most θ values
 \leadsto Makes posterior uncontroversial

Ex. $X \sim \text{Binom}(10^4, \theta)$, observe $X = 3000$

$$SD_{\theta}(X/n) = \sqrt{\frac{\theta(1-\theta)}{n}} \leq 0.005$$

$$\Rightarrow \text{Lik}(\theta; X) \approx 0 \text{ outside } C = [0.29, 0.31]$$

All "reasonable" priors may be \approx flat on C

$$\Rightarrow \lambda(\theta | X) \propto_{\theta} \text{Lik}(\theta; X)$$

$$\propto_{\theta} \exp\left\{-\frac{J(0.3)}{2} (\theta - 0.3)^2\right\}$$

$$\propto_{\theta} N(0.3, J(0.3)^{-1}) \quad (\text{HW 5})$$

Data "swamps" everyone's prior

Gaussian sequence model

$$X|\theta \sim N_d(\mu, I_d) \quad \mu \in \mathbb{R}^d$$

Jeffereys prior is flat: $\lambda(\mu) \propto 1$

$$\lambda(\mu|x) = N_d(x, I_d) \Rightarrow \mathbb{E}[\mu|x] = x$$

Same as UMVU

What about $\rho^2 = \|\mu\|^2$? Recall

$$\mu \sim N_d(x, I_d) \Rightarrow \mathbb{E}[\|\mu\|^2 | x] = \|x\|^2 + d$$

$$\text{Recall } \delta_{\text{umvu}}(x) = \|x\|^2 - d \Rightarrow \delta_\lambda(x) = \delta_{\text{umvu}}(x) + 2d$$

$$\begin{aligned} \text{MSE}(\theta; \delta_\lambda) &= \text{Var}_\theta(\delta_\lambda) + \text{Bias}_\theta(\delta_\lambda)^2 \\ &= \text{Var}_\theta(\delta_{\text{umvu}}) + 4d^2 \end{aligned}$$

What went wrong? Examine Jeffereys prior:

$$\mathbb{P}(\rho^2 \leq t) = \text{Vol}(\text{Ball of radius } \sqrt{t})$$

$$= \text{const}(d) \cdot t^{d/2}$$

$$\Rightarrow \lambda(\rho^2) \propto_{\rho^2} (\rho^2)^{d/2-1} = \rho^{d-2}$$

Grows rapidly! Prior "expects" ρ^2 to be huge

Source #3: Prior or concurrent experience

May have many "copies" of same problem

Assume corresp. θ values drawn from a population

→ Hierarchical Bayes / empirical Bayes

Can be hard to choose right reference class

Ex. Estimate batting average

Player i has n_i at-bats, "true" batting avg. θ_i

Hierarchical model: players $i = 1, \dots, m$

"hyperparameters" $\alpha, \beta \sim \lambda_0 \leftarrow$ "hyperprior"

$$\theta_i | \alpha, \beta \stackrel{iid}{\sim} \text{Beta}(\alpha, \beta)$$

$$X_i | \alpha, \beta, \theta \stackrel{ind.}{\sim} \text{Binom}(n_i, \theta_i)$$

$$\mathbb{E}[\theta_i | X, \alpha, \beta] = \frac{X_i + \alpha}{n_i + \alpha + \beta}$$

$$\mathbb{E}[\theta_i | X] = \iint \frac{X_i + \alpha}{n_i + \alpha + \beta} \lambda_0(\alpha, \beta | X) d\alpha d\beta$$

If m large α, β may be "almost known"

→ choice of λ_0 doesn't matter much

Reference class problem: Which players to include?

Flexibility of Bayes

Any $\Omega, \mathcal{P}, L, g(\theta)$: δ_Ω defined straightforwardly

$$\delta_\Omega(x) = \operatorname{argmin}_d \int L(\theta, d) \lambda(\theta|x) d\theta$$

Problem reduced to (possibly hard) computation

Posterior is "one stop shop" for all answers

No need for:

- special family structure (exp. fam. / complete s.s.)
- special estimator (U-estimable)
- convex or nice L

⇒ Highly expressive modeling & estimation

Caveat: Limited by ability to do computations
(Topic of next lecture)

Source # 4: Convenience Priors

Choosing conjugate or other "nice" priors

↪ much faster computations esp. in high-dim.

(But what does the posterior mean?)

\mathbb{E}_X . $X_1, \dots, X_n \stackrel{iid}{\sim} \rho$, ρ unknown density on \mathbb{R}

Estimand: $m = \text{median}(\rho)$

Estimator $\delta(X) = \text{median}(X)$

good estimator: robust, nonparametric

large n : $\delta(X) \approx N(m, (4np(m))^{-1})$

not Bayes for any realistic prior

Bayes approach

Step 1. Define prior over ρ (infinite-dim!)

Step 2. Calculate posterior

Horrific unless we pick special prior

Step 3. Return e.g. $\mathbb{E}[m | X]$

If it differs substantially from $\text{median}(X)$,
do we trust it?

Gaussian Hierarchical Model:

$$\tau^2 \sim \lambda_0$$

$$\theta_i | \tau^2 \stackrel{iid}{\sim} N(0, \tau^2) \quad i \leq d$$

$$X_i | \tau^2, \theta \stackrel{ind.}{\sim} N(\theta_i, 1)$$

Posterior mean:

$$\begin{aligned} \delta(x_i) &= \mathbb{E}[\theta_i | X] \\ &= \mathbb{E}\left\{ \mathbb{E}[\theta_i | X, \tau^2] \mid X \right\} \\ &= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} X_i \mid X \right] \\ &= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X \right] \cdot X_i \end{aligned}$$

Linear shrinkage estimator,

Bayes-optimal shrinkage estimated from data

Likelihood for τ^2 : marginalize over θ_i

$$X_i | \tau^2 \sim N(0, 1+\tau^2)$$

$$\Rightarrow \frac{1}{d} \|X\|^2 \sim \frac{1+\tau^2}{d} \chi_d^2$$

$$\sim \left(1+\tau^2, \frac{2+2\tau^2}{d} \right) \quad \text{notation (mean, variance)}$$

Define $\zeta(\tau^2) = \frac{1}{1+\tau^2}$

$\Rightarrow \delta(x) = (1 - \mathbb{E}[\zeta | x]) X_i$

Conjugate prior:

$$v = \|X\|^2 | \zeta \sim \frac{1}{\zeta} \chi_d^2 = \frac{\zeta^{d/2}}{\Gamma(d/2)} v^{d/2-1} e^{-\zeta v}$$

$$\zeta \sim \frac{1}{s} \chi_k^2 = \frac{s^{k/2}}{\Gamma(k/2)} \zeta^{k/2-1} e^{-s\zeta}$$

$$\Rightarrow \zeta | \|X\|^2 \propto \zeta^{\frac{k+d}{2}-1} e^{-(s+\|X\|^2)\zeta}$$

$$\sim \frac{1}{s+\|X\|^2} \chi_{k+d}^2$$

$$\mathbb{E}[\zeta | \|X\|^2] = \frac{k+d}{s+\|X\|^2} \approx d(1+\tau^2) + O(d^{1/2})$$

[might want to truncate prior to $[0, 1]$ if d small]