Section 8: Bayesian Inference

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30–9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM–12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30–11:30 AM in Cabot D-Hall (Emily).

2 Big Picture

This week we turn to *Bayesian inference*, a fundamentally different philosophy from the frequentist approach we have been using. Most of Harvard Stats faculty (like Joe) are Bayesian; I think you will soon find out why!

In the *frequentist* framework, θ is an unknown, but fixed. We build estimators and confidence intervals whose performance we assess over hypothetical repeated experiments. However, in the *Bayesian* framework, θ is treated as a random variable, and we use probability to quantify our uncertainty about it.

The workflow is:

- *Prior* $\pi(\theta)$: our beliefs about θ before seeing data, encoded as a probability distribution. This is the most controversial part about Bayesian statistics.
- *Likelihood* $L(\theta; y) = f(y \mid \theta)$: how likely the observed data are under each value of θ. Both Bayesians and frequentists agree on this!
- *Posterior* $\pi(\theta \mid y)$: our updated beliefs about θ after seeing data. Get familiar with the term, this is the core object of Bayesian inference.

We will also discuss *conjugate priors* (which keep the posterior in a tractable family), *credible intervals* (the Bayesian analog of confidence intervals, with a more direct probability interpretation), and *Bayesian point estimation* (posterior mean, median, and MAP, each optimal under a different loss). We will then discuss *hierarchical models* that give new insight into frequentist problems like the inadmissibility of the MLE (Stein's paradox).

3 Prior to Posterior

Definition 1 (Prior, posterior, marginal likelihood)**.** Consider a parametric model $f(y \mid \theta)$ with unknown parameter θ. In the Bayesian approach we posit a **joint distribution** for (Y, θ). The **prior** $\pi(\theta)$ is the marginal distribution of θ, encoding our beliefs before seeing data. The **posterior** $\pi(\theta \mid y)$ is the conditional distribution of θ given the observed data y. The **marginal likelihood** (or *prior predictive distribution*) is $f(y) = \int f(y \mid \theta)\pi(\theta)\, d\theta$.

Theorem 1 (Bayes' rule)**.**

$$\pi(\theta \mid y) = \frac{L(\theta; y)\, \pi(\theta)}{f(y)} \propto L(\theta; y)\, \pi(\theta).$$

We almost always work up to proportionality (normalization is difficult), since $f(y)$ does not depend on θ.

• **Strategy**: Write $\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$, simplify, and pattern-match to a known distribution to identify the posterior without computing $f(y)$.

• ☣: **Cromwell's rule.** If $\pi(\theta_0) = 0$ for some θ_0, then $\pi(\theta_0 \mid y) = 0$ no matter what the data say. Never assign prior probability of exactly 0 or 1 to something unless it is logically impossible or certain!

• **As** $n \to \infty$: intuitively, the likelihood dominates the prior (the data overwhelm prior beliefs), and the posterior will concentrate near the MLE. For small n, the prior plays a large role based on weight!

Concept Checker 1. Let $Y \mid \theta \sim \text{Bern}(\theta)$ and $\theta \sim \text{Beta}(2, 2)$.

1. Write out $\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$ and identify the posterior distribution.
2. What does the $\text{Beta}(2, 2)$ prior encode about our beliefs for θ?
3. What happens to the posterior as we observe more and more data?

Solution

4 Point Estimation

Definition 2 (Posterior mean, median, and mode)**.** Let θ have a continuous posterior density $\pi(\theta \mid y)$. Then:

$$\text{Posterior mean} = \mathbb{E}[\theta \mid y] = \int \theta\, \pi(\theta \mid y)\, d\theta$$

$$\text{Posterior median} = Q_{\theta|y}(0.5)$$

$$\text{Posterior mode (MAP)} = \arg\max_{\theta}\, \pi(\theta \mid y) = \arg\max_{\theta} \{\log L(\theta; y) + \log \pi(\theta)\}$$

Theorem 2 (Optimal loss functions)**.**

- *Squared error loss $(\theta - \hat{\theta})^2$: minimized by the **posterior mean** $\mathbb{E}[\theta \mid y]$.*
- *Absolute error loss $|\theta - \hat{\theta}|$: minimized by the **posterior median** $Q_{\theta|y}(0.5)$.*
- *0–1 loss (in a limit): minimized by the **posterior mode (MAP)**.*

• **MAP vs. MLE.** MAP = MLE + $\log \pi(\theta)$ in the optimization. With a flat (Uniform) prior on a bounded interval, MAP = MLE (why?). With an informative prior, MAP regularizes the MLE toward the prior mean.

Regularization is a good mix between frequentist and Bayesian approaches–if you are curious, it's worth looking more into here:

• **MAP and LASSO.** With a Laplace prior $\pi(\theta) \propto e^{-d|\theta|}$, the MAP is the LASSO estimator—it thresholds $\bar{y}$ exactly to 0 if $|\bar{y}| < c$. With a Normal prior, the MAP/posterior mean is ridge regression—it *shrinks* $\bar{y}$ toward 0 but never exactly to 0.

• ☣: MAP is **not** invariant to reparameterization. If $\hat{\theta}_{\text{MAP}}$ maximizes $\pi(\theta \mid y)$, then $g(\hat{\theta}_{\text{MAP}})$ does *not* generally maximize $\pi(g(\theta) \mid y)$.

Concept Checker 2. For the Beta-Binomial model $Y \mid p \sim \text{Bin}(n, p)$, $p \sim \text{Beta}(a, b)$, write down the posterior mean, median, and MAP explicitly. For large n, what do all three converge to?

Solution

5 Credible Intervals

Definition 3 (Credible interval)**.** Let $0 < \alpha < 1$. A $100(1-\alpha)\%$ **credible interval** (or **posterior probability interval**) for θ is an interval $[a(y), b(y)]$ such that

$$P(a(y) \le \theta \le b(y) \mid y) = 1 - \alpha.$$

The standard choice is the equal-tailed interval $[Q_{\theta|y}(\alpha/2),\, Q_{\theta|y}(1-\alpha/2)]$.

• **Direct probability interpretation.** A 95% credible interval means: given the data, there is a 95% probability that θ lies in the interval. This is the statement we *want* to make about confidence intervals, but can't (frequentist intervals are about the procedure, not the parameter).

• **Average frequentist coverage.** A 95% credible interval also has *on average* 95% frequentist coverage, where the average is over both θ and Y. By Adam's law: letting $I = \mathbf{1}(\theta \in C(Y))$,

$$P(I = 1) = \mathbb{E}[I] = \mathbb{E}[\mathbb{E}[I \mid Y]] = \mathbb{E}[P(I = 1 \mid Y)] = \mathbb{E}[0.95] = 0.95.$$

• ☣: A 95% credible interval is *not* guaranteed to be a 95% confidence interval for a specific fixed θ. Coverage at a *fixed* θ may be higher or lower than 95%, depending on the prior since we include uncertainty in the prior. The averaging above is over the prior distribution of θ, not a fixed θ.

Concept Checker 3. Suppose $\theta \sim \mathcal{N}(0, 2^2)$ and $Y \mid \theta \sim \mathcal{N}(\theta, 1)$. We observe $Y = y$.

1. Find the posterior distribution $\theta \mid y$.
2. Write down a 95% credible interval for θ.
3. If the true $\theta = 1$, is the coverage of this credible interval exactly 95%? Explain intuitively why or why not.

Solution

6 Conjugate Priors

Idea. A conjugate prior is one where the posterior stays in the same distributional family as the prior. So we only need to update the parameters, not the family itself—which is very convenient for computation and gives us known results.

Definition 4 (Conjugate prior)**.** A family of priors is **conjugate** for a particular likelihood if choosing a prior in the family always results in a posterior in the same family.

6.1 Beta–Binomial

Theorem 3 (Beta-Binomial conjugacy)**.** *If $p \sim Beta(a, b)$ and $Y \mid p \sim Bin(n, p)$, then*

$$p \mid (Y = y) \sim Beta(a + y,\ b + n - y).$$

Posterior mean: $\dfrac{a+y}{a+b+n}$*. Interpret $a-1$ as prior successes, $b-1$ as prior failures.*

6.2 Gamma–Poisson

Theorem 4 (Gamma-Poisson conjugacy)**.** *If $\lambda \sim Gamma(r_0, b_0)$ (rate b_0) and $Y_1, \ldots, Y_n \mid \lambda \overset{i.i.d.}{\sim} Pois(\lambda)$, then with $S = \sum y_i$:*

$$\lambda \mid y \sim Gamma(r_0 + S,\ b_0 + n).$$

Posterior mean: $\dfrac{r_0+S}{b_0+n}$*.* *Predictive:* $\tilde{Y} \mid y \sim NBin\left(r_0 + S,\ \dfrac{b_0+n}{b_0+n+1}\right)$*.*

6.3 Normal–Normal

Theorem 5 (Normal-Normal conjugacy, general sample size)**.** *Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$. Then*

$$\mu \mid y \sim \mathcal{N}(\mu_n, \tau_n^2), \qquad \tau_n^{-2} = n\sigma^{-2} + \tau_0^{-2}, \qquad \mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right).$$

Writing $b_n = \tau_n^2/\tau_0^2 = \sigma^2/(\sigma^2 + n\tau_0^2)$ (shrinkage factor):

$$\mu_n = (1 - b_n)\bar{y} + b_n\mu_0.$$

The posterior mean is a ***precision-weighted average*** *of the sample mean and the prior mean. As $n \to \infty$, $b_n \to 0$ and $\mu_n \to \bar{y}$.*

• **Intuition.** The posterior mean compromises between $\bar{y}$ (the data's estimate) and μ_0 (the prior's estimate), weighted by their relative precisions. More data ⇒ less shrinkage toward the prior.

• **Normal-Normal predictive.** By Adam's and Eve's laws:

$$\tilde{Y} \mid y \sim \mathcal{N}(\mu_n,\ \sigma^2 + \tau_n^2).$$

The extra τ_n^2 reflects *parameter uncertainty*—even if we knew μ exactly, $\tilde{Y}$ would have variance σ^2; our uncertainty about μ adds τ_n^2 on top.

Theorem 6 (Normal with heteroskedasticity)**.** *If* $Y_j \mid \mu \overset{ind.}{\sim} \mathcal{N}(\mu, \sigma_j^2)$ *and* $\mu \sim \mathcal{N}(m_0, \tau_0^2)$*, then*

$$\tau_n^{-2} = \tau_0^{-2} + \textstyle\sum_j \sigma_j^{-2}, \qquad \mu_n = \tau_n^2\Big(\tau_0^{-2} m_0 + \textstyle\sum_j \sigma_j^{-2} y_j\Big).$$

Observations with smaller variances receive more weight. Reduces to the standard case when all $\sigma_j^2 = \sigma^2$*.*

Theorem 7 (Bayesian linear regression)**.** *If* $Y_j \mid (X = x, \theta) \overset{ind.}{\sim} \mathcal{N}(\theta x_j, \sigma^2)$ *and* $\theta \mid X \sim \mathcal{N}(m_0, \tau_0^2)$*, then* $\theta \mid y, x \sim \mathcal{N}(m_n, \tau_n^2)$ *where*

$$\tau_n^{-2} = \tau_0^{-2} + \sigma^{-2} \textstyle\sum_j x_j^2, \qquad m_n = \tau_n^2\Big(\tau_0^{-2} m_0 + \sigma^{-2} \textstyle\sum_j x_j y_j\Big).$$

Concept Checker 4. A manufacturer claims their product weights are $\mathcal{N}(\theta, 100)$ grams (so $\sigma^2 = 100$). Your prior for θ is $\mathcal{N}(200, 400)$. You weigh $n = 25$ items and find $\bar{y} = 190$.

1. Find the posterior distribution of θ.
2. Interpret the shrinkage factor b_n: how much does the posterior mean shrink toward the prior?
3. Find the posterior predictive distribution for a new item's weight.

Solution

7 Practice Problems

Problem 1 (Normal posterior predictive)**.** A random sample of n students is drawn from a population whose weights are $\mathcal{N}(\theta, 400)$ (so $\sigma = 20$) with unknown mean θ. The sample mean is $\bar{y} = 150$. Use prior $\theta \sim \mathcal{N}(180, 1600)$ (so $\tau_0 = 40$).

(a) Find the posterior distribution $\theta \mid y$ as a function of n.

(b) Find the posterior predictive distribution $\tilde{y} \mid y$ for a new student's weight, justifying the parameters using Adam's and Eve's laws.

(c) For $n = 10$: give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.

Solution

Problem 2 (Posterior as a compromise)**.** Let Y be the number of heads in n coin flips with unknown probability θ.

(a) With a Unif$(0, 1)$ prior, derive the prior predictive distribution $P(Y = k)$ for each $k = 0, \ldots, n$.

(b) With $\theta \sim \text{Beta}(\alpha, \beta)$ and y heads observed, show algebraically that the posterior mean always lies strictly between the prior mean $\frac{\alpha}{\alpha+\beta}$ and the observed frequency $\frac{y}{n}$.

(c) Show that if the prior is Unif$(0, 1)$, the posterior variance is always less than the prior variance.

(d) Give an example of a Beta(α, β) prior and data (n, y) where the posterior variance *exceeds* the prior variance.

Solution

Problem 3 (Bayesian persuasion)**.** Let $R \sim \text{Unif}(0, 1)$ be the probability of a successful persuasion. Leo will stop resisting once he experiences s successes; let S_{total} be the total number of attempts needed.

(a) Show that $\mathbb{E}[S_{\text{total}}]$ is infinite.

(b) After n total attempts with exactly s successes, find the posterior distribution of R.

Solution

Problem 4 (Censored and uncensored data)**.** Suppose $Y \mid \theta \sim \text{Expo}(\theta)$ with conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$.

(a) We observe $Y \geq 100$ (but not the exact value). Find $\pi(\theta \mid Y \geq 100)$, and the posterior mean and variance.

(b) Now suppose we are told $Y = 100$ exactly. Find $\pi(\theta \mid Y = 100)$, and the posterior mean and variance.

(c) Why is the posterior variance *higher* in (b), even though more information was given?

Solution