Statistics 210A Lecture 9 Notes

Daniel Raban

September 23, 2021

1 Priors in Bayesian Estimation

1.1 Recap: Bayesian estimation

Last time, we introduced Bayes estimation, where we want to minimize the **Bayes risk**

$$\begin{aligned} R_{\text{Bayes}}(\Lambda; s) &= \int_\Omega R(\theta; s)\, d\Lambda(\theta) \\ &= \mathbb{E}[L(\Theta; \delta(X))], \end{aligned}$$

where $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$.

The **Bayes estimator** $\delta_\Lambda(x)$ minimizes

$$\mathbb{E}[L(\Theta; d) \mid X = x]$$

in d. If we have a **prior** density $\lambda(\theta)$ and a likelihood $p_\theta(x)$, then we get the **posterior** density

$$\lambda(\theta \mid x) = \frac{\lambda(\theta) p_\theta(x)}{\int \lambda(\theta) p_\theta(x)\, dx}.$$

Example 1.1 (Beta-Binomial)**.** In this example, $X \mid \theta \sim \text{Binom}(n, \theta) = \theta^x (1-\theta)^{1-x} \binom{n}{x}$ with the prior $\theta \sim \text{Beta}(\alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The posterior distribution is

$$\begin{aligned} \lambda(\theta \mid x) &\propto_\theta \theta^{x+\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \text{Beta}(\alpha + x - 1, \beta + n - x - 1) \end{aligned}$$

It follows that

$$\mathbb{E}[\Theta \mid X] = \frac{X + \alpha}{n + \alpha + \beta}$$

is the Bayes estimator for the squared error loss.

We also had a normal location family with a normal prior which gave us a normal posterior, as well.

1.2 Conjugate priors

Definition 1.1. If the posterior is from the same family as the prior, we say the prior (family) is **conjugate** to the likelihood.

Suppose $X_i \mid \eta \overset{\text{iid}}{\sim} p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$ for $i = 1, \dots n$, with $\eta \in \Xi_1 \subseteq \mathbb{R}^s$. For some carrier density $\lambda_0(\eta)$, define the $(s+1)$-parameter exponential family.

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

The sufficient statistic is $\begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix}$ with natural parameter $\begin{bmatrix} k\mu \\ k \end{bmatrix}$. If we take $\lambda_{k\mu,k}$ as our prior, then

$$\begin{aligned} \lambda(\eta \mid X_1, \dots, X_n) &\propto_\eta e^{k\mu^\top \eta - kA(\eta)} \lambda_0(\eta) \cdot \prod_{i=1}^n e^{\eta^\top T(x_i) - A(\eta)} \\ &= \exp\Big((k\mu + n\overline{T}(x))^\top \eta - (k+n) A(\eta) \Big) \lambda_0(\eta) \\ &\propto_\eta \lambda_{k\mu + n\overline{T}, k+n}(\eta). \end{aligned}$$

Here is the interpretation:

1. Suppose we take the prior $\lambda_{k\mu,k}$ and observe X_1. Then the posterior is $\lambda_{k\mu+X_1,k+1}$.
2. Now observe X_2 and update the posterior to get $\lambda_{k\mu+X_1+X_2,k+2}$.
3. ...

If we have a (possibly improper) prior λ_0 and make $k+n$ observations with $\sum_i T(X_i) = k\mu + s$, this is the same as if we had the prior $\lambda_{k\mu,k}$ and observe n observations with $\sum_i T(X_i) = s$.

Example 1.2. Here is a list of some conjugate priors:

Likelihood	Prior
$\text{Binom}(n, \theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$
$N(\theta, \sigma^2)$	$\theta \sim N(\mu, \tau^2)$
$\text{Pois}(\theta)$	$\theta \sim \text{Gamma}(\nu, s)$

People will say that the Beta, for example, is *the* conjugate prior to the Binomial. There can be more than one conjugate prior, which we can get just by changing the carrier distribution.

1.3 Types of priors

Bayesian estimation requires us to have a prior distribution we believe in. In what ways do we do this?

1. **Direct prior or parallel experience:** We can estimate the prior from data. If there is a broad agreement on the prior, corresponding to many observations, the prior may be more meaningful. This gives rise to the following types of Bayesian estimation:
   - Hierarchical Bayes
   - Empirical Bayes
2. **Subjective beliefs:**[1] Here, the prior represents epistemic uncertainty, and the posterior is uncertainty ex post, after observing data and rationally updating.
3. **Convenience prior:** Generally, we have to calculate posteriors. If $\dim(\Omega)$ is large, the posterior is ≈ 0 for most of Ω. This can make it computationally difficult to perform Bayesian estimation, so we might pick a prior which makes the calculation easier, such as a conjugate prior.
4. **"Objective" prior:** We may try to pick a prior which seems to not represent our individual opinion.

   Example 1.3. Suppose $X_i \mid \theta \sim N(\theta, 1)$ for $i = 1, \dots, n$. We could try to use a **flat prior**: $\lambda(\theta) \propto_\theta 1$. This prior is is not a probability distribution, but we can still use it because it gives a valid posterior:

   $$\begin{aligned}\lambda(\theta)p_\theta(x) &\propto_\theta e^{\theta\sum_i x_i - n\theta^2/2} \\ &\propto_\theta N(\overline{x}, 1/n).\end{aligned}$$

   The Bayes estimator is $\overline{X}$. The posterior arises naturally as taking taking a limit of priors: $\lim_{\tau^2\to\infty} N(0, \tau)$.

   The issue with a flat prior is that this is not invariant to reparameterization of the model.

   Example 1.4. Let $X \sim \text{Binom}(n, \Theta)$ with $\Theta \sim U[0, 1]$. Then

   $$\mathbb{P}(\Theta \in [0.5, 0.51]) = \mathbb{P}(\Theta \in [0.0001, 0.0101]) = 0.01.$$

   If we let $\eta = \log \frac{\Theta}{1-\Theta}$, then

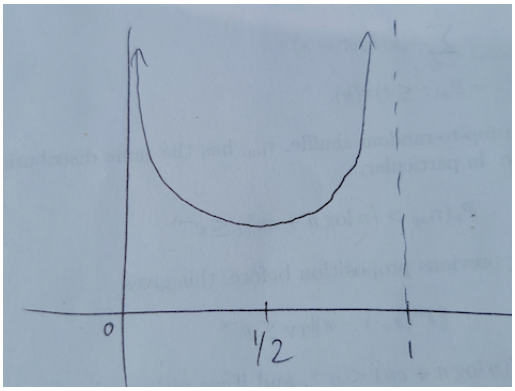
   $$\mathbb{P}(\Theta \in [0.5, 0.51]) \approx \mathbb{P}(\eta \in [0, 0.01]),$$

   while

   $$\mathbb{P}(\Theta \in [0.0001, 0.0101]) = \mathbb{P}(\eta \in [\log 0.001, \log 0.1]).$$

[1] One may call this the "hardcore" Bayesian perspective.

Jeffreys proposed using $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$. This is called the **Jeffreys prior**, which is invariant under reparameterization. However, the Jeffreys prior can have less of a claim to being agnostic. In the normal case, the Jeffreys prior is the flat prior, but in the binomial case, the Jeffreys prior looks like this:



Remark 1.1. There has been some controversy about Bayesian vs frequentist statistics. Historically, frequentist statisticians tend to give objections of the form "The object of interest (such as the number of elephants in Africa[2]) is not actually random!" However, if you flip a coin and don't yet look at the result, even though the outcome is certain, there is still epistemic uncertainty about the result.

The Bayesian perspective has the advantage (and disadvantage) of being able to express vague intuitions. Ultimately, making a decision in government may require different statistics from writing a scientific paper. But subjective beliefs and intuitions can often be incorrect.

A practical issue is that it is very difficult to express an opinion of a joint distribution of many random variables.

[2]The elephants in Africa are just standing around, waiting to be counted.