Mathematical Statistics II

The Bayesian Approach to Parameter Estimation

Jesse Wheeler

Introduction

Bayesian Estimation

- Much of this work is based on Rice (2007, Section 8.6).
- We have already discussed the philosophy of Bayesian statistics.
- We start with a prior belief about parameter values, and update these beliefs using observed data.
- The resulting distribution is called the *posterior*, and it represents our updated belief after observing data.
- This is very natural idea that is closely related to the idea of likelihood: likelihood quantifies some degree of belief about a parameter value.

Review

Some Review

- Before we begin, we will first do a bit of review.
- In the context of Bayesian inference, we treat unknown parameter vectors as random variables, which I will denote Θ.
- Thus, our probability model can be expressed as $f(x|\Theta = \theta)$, which we often shorten to $f(x|\theta)$.

Some Review II

Bayes' Theorem

Let X be the random vector representing observed data, and Θ the random parameter vector, and x^* the observed data. Bayes Theorem states:

$$\begin{aligned}\pi_{\Theta|X}(\theta|x^*) &= \frac{f_{X|\Theta}(x^*|\theta)\pi_\Theta(\theta)}{f_X(x^*)} \\ &= \frac{f_{X|\theta}(x^*|\theta)\pi_\Theta(\theta)}{\int f_{X|\Theta}(x^*|\tau)\pi_\Theta(\tau)\,d\tau}\end{aligned}$$

- As before, f is taken to be either a pmf or pdf, depending on the problem.

Some Review III

Flipping 10 coins

Our friend hands us a coin from another country, and we want to estimate $\theta = p$, the probability that the coin lands heads.

Suppose we flip a coin 10 times, and see n heads. Find a Bayesian estimate for θ.

Some Review IV

- Even in the simple problem above, we see two of the primary challenges with Bayesian parameter estimation:
  - How do we choose the prior distribution $\pi(\theta)$? A generally safe and accepted approach is a uniform prior. However, this formally only exists if θ is bounded, which is not always the case. Also, it represents a prior belief: given a new coin, do we really think all values of p are equally likely, or maybe values close to $p = 0.5$ are more likely than extreme values $p = 0, 1$? Since the prior represents our beliefs about θ, is a uniform prior actually appropriate? If it isn't appropriate, how exactly should we specify the prior?
  - Even in this very simple model and prior, the denominator $f(x)$ was difficult to compute. What about more complex models and priors? A large amount of Bayesian computation and theory is dedicated to solving this problem.

Some Review V

Proposition: the MAP and MLE

Let θ be a parameter of interest, and x^* the observed data. If our prior distribution is proportional to 1, i.e., $\pi(\theta) \propto 1$ (which is effectively a uniform prior on a bounded interval), then

$$\hat{\theta}_{\mathsf{MAP}} = \hat{\theta}_{\mathsf{MLE}}.$$

- This is true for the Coin-tossing example; look back at the likelihood function and posterior, and use R to plot them both.

Examples

Bayesian point-estimate examples

Poisson model

Suppose we have observations n observations, which we wish to model as IID Poisson(λ). Find a Bayesian estimate of $\Lambda = \lambda$ given the observed data x^*.

Bayesian point-estimate examples II

Poisson posterior, uniform prior

Revisit the Poisson(λ) model, while taking the alternative approach of using a uniform prior.

Real-data example: Poisson Distribution

- Now let's look at a real-data example. These data are the 23 observations from the asbestos-filter problem.

```r
x <- c(
  31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18,
  26, 27, 27, 18, 24, 22, 28, 24, 21, 17, 24
)
x
```

```
 [1] 31 29 19 18 31 28 34 27 34 30 16 18 26 27 27 18 24 22
[19] 28 24 21 17 24
```

Real-data example: Poisson Distribution II

Comparing Estimates

Using the data above, compare estimates using the MoM, MLE, and the two Bayesian approaches, as well as the corresponding errors related to these estimates.

Posteriors and Likelihood

- In the problem above, we saw that we get very similar estimates using MLE or Bayesian approaches, regardless of which prior we picked.
- We can argue why this will often be the case, especially for IID data.
- Previously, we saw:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- When n gets large, the likelihood dominates in this equation. In the IID case:

$$\text{likelihood} = \prod_{i=1}^{n} f(x_i^*|\theta).$$

Posteriors and Likelihood II

- In particular, each new data point scales the likelihood larger and larger, to the point where the prior has little impact on the posterior distribution.
- See the accompanying Lecture 4 R code for a visual demonstration of this using the Poisson distribution.

Introduction to Numeric Integration

Numeric Integration

- As we saw in the previous example, one of the primary challenges of Bayesian estimation is the integration in the denominator of the posterior.
- Bayesian statistics has really exploded since the late 20th century, largely thanks to improved computational tools that help with the numeric integration.
- For this set of lectures, we only briefly introduce this topic. Depending on time and interest, we can explore this topic more later in the semester.

Choice of Priors

Conjugate priors

- The Poisson(λ) example showed two main approaches:
  - The traditional (subjective) Bayesian, who takes seriously the choice of prior, and chose a Gamma density to aid computations.
  - The utilitarian (objective) Bayesian, who picked an uninformative prior.
- The former approach was aided by what is known as a conjugate prior.

Conjugate priors II

Definition: Conjugate priors

Suppose the prior distribution belongs to a family of distributions, G, and the data come from a family of distributions H.

G is said to be conjugate to H if the posterior is in the family G.

- Example: If the data-model is Poisson(λ), then the family H is the family of Poisson distributions. The Gamma family (G) of distributions is conjugate to the Poisson family, because if Gamma is selected as the prior distribution, then the posterior distribution (under data model H) is still Gamma (G), with updated parameters.

Conjugate priors III

- Much of the Bayesian statistics of the 20th century relied on conjugate priors to help with integration, or were confined to models with very few parameters.
- Recent developments in computing, both hardware, software, and theory of Bayesian computing, has enabled fitting much more complex models using arbitrary priors.
- Still, it's worth discussing conjugate priors, and we will provide a few examples.

Jeffrey's priors

- TODO

Hierarchical Bayes

Hierarchical Bayes

- The idea behind Hierarchical Bayes is simple: our model f depends on parameters θ.
- We can get a prior for θ, $\pi(\theta)$.
- The prior itself depends on parameters, say $\pi(\theta; \theta_1)$.
- How do we choose θ_1? Sometimes we might know θ_1, but sometimes not.
- In a pure Bayesian paradigm, if we don't know the value of θ_1, then it is also a random variable Θ_1, and we should put a prior on this as well!
- In some way, this allows us to be less-committal about the parameters in the prior model, and instead allow the data to inform our choice of priors (to some degree).

Hierarchical Bayes II

- Philosophically, this situation naturally arises if we want to pick a conjugate prior for Θ, but are not committal about the hyperparameters Θ_1 that define the distribution of Θ.
- We could continue doing this many times if we wanted!
- The prior for Θ_1 might depend on parameters θ_2, which we model as a random variable Θ_2, $\ldots$.
- This leads to a model for $(X, \Theta, \Theta_1, \ldots, \Theta_N)$.
- However, there is a conditional structure to this model:

$$\Theta_N \longrightarrow \Theta_{N-1} \longrightarrow \ldots \longrightarrow \Theta \longrightarrow X.$$

- Thus, X depends only on Θ, and Θ_n only on Θ_{n+1}:

$$X|\Theta = \theta \sim f(x|\theta),\ \Theta|\Theta_1 = \theta_1 \sim \pi_1(\theta|\theta_1)\ \ldots\ \Theta_N \sim \pi_N(\theta_n).$$

Hierarchical Bayes III

- Using rules of marginal probability and conditional probability, then

$$
\begin{aligned}
\pi(\theta) &= \int \pi(\theta, \theta_1, \ldots, \theta_N)\, d\theta_{1:N} \\
&= \int \pi(\theta|\theta_{1:N})\pi(\theta_{1:N})\, d\theta_{1:N} \\
&= \int \pi(\theta|\theta_1)\pi(\theta_1|\theta_{2:N})\pi(\theta_{2:N})\, d\theta_{1:N} \\
&= \vdots \\
&= \int \pi(\theta|\theta_1)\pi(\theta_1|\theta_2)\ldots, \pi(\theta_{N-1}|\theta_N)\pi(\theta_N)\, d\theta_{1:N}
\end{aligned}
$$

Hierarchical Bayes IV

- Thus, the hierarchical model is functionally equivalent to the standard Bayesian model:

$$X|\Theta = \theta \sim f(x|\theta) \quad \Theta \sim \pi(\theta),$$

  where $\pi(\theta)$ is given by the integral above.
- Why would we want to do this?
  1. Sometimes the data / problem give rise to a natural hierarchical structure, and this idea will be useful. Here, we might actually be interested in the hyperparameters $\theta_1, \ldots, \theta_N$.
  2. We can now be less committal about our priors, while still using desirable structures.
  3. It can sometimes aid computations.

Hierarchical Bayes V

Trivial case: hierarchical Normal-Normal

Suppose that the data X_i are iid $N(\theta, 1)$. Set a prior for θ as $\Theta | \Theta_1 = \theta_1 \sim N(\theta_1, 1)$, and $\Theta_1 \sim N(0, 1)$.

Hierarchical Bayes VI

More realistic example: Coin-toss experiment

Suppose your friend gives you a coin from another country, and you want to estimate $\theta = p$, the probability of heads. Thus, in N tosses, the natural model for X, the number of heads, $X \sim \text{Bin}(N, \theta)$. You're believe that the proportion is close to $1/2$, but not quite sure. A nice prior would be the Beta(α, β)-distribution, since it is conjugate for the binomial family.

If $\Theta \sim \text{Beta}(\alpha, \beta)$, then $E[\Theta] = \frac{\alpha}{\alpha+\beta}$. Thus, if I want a prior centered at $1/2$, I can pick: $\theta_1 = \alpha = \beta$, and $E[\Theta] = \theta_1/2\theta_1 = 1/2$. We can now give a prior for Θ_1.

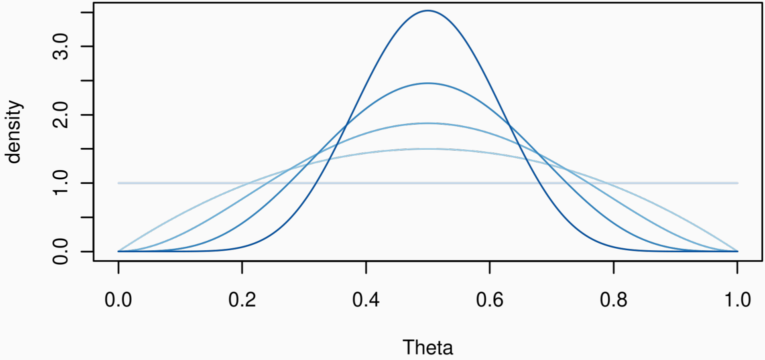
Hierarchical Bayes (continued)

- For now, we will restrict our choices of Θ_1 to be integers.

```
Theta <- seq(1e-8, 1-1e-8, length.out = 1000)
B1 <- dbeta(Theta, 1, 1)
B2 <- dbeta(Theta, 2, 2)
B3 <- dbeta(Theta, 3, 3)
B5 <- dbeta(Theta, 5, 5)
B10 <- dbeta(Theta, 10, 10)

plot(x = Theta, y = B1, type = 'l', ylim = c(0, 3.5), col = "#c6
lines(x = Theta, y = B2, type = 'l', col = '#9ecae1')
lines(x = Theta, y = B3, type = 'l', col = '#6baed6')
lines(x = Theta, y = B5, type = 'l', col = '#3182bd')
lines(x = Theta, y = B10, type = 'l', col = '#08519c')
```

Hierarchical Bayes (continued) II



- As θ_1 grows, the variance of Θ shrinks at a rate $O(\theta_1)$.
- Thus, to be noncommittal about our prior on Θ, we will set a hyperprior on Θ_1 that has more weight on smaller values of k:

$$\pi_{\Theta_1}(k) = \frac{1}{2\log(2)k(2k-1)}, \quad k \in \{1, 2, \ldots\}$$

- This hyper-prior was selected somewhat out of convenience (The Catalan Numbers), which will allow us to get the marginal prior of Θ:

$$\pi(\theta) = \sum_{k=1}^{\infty} \pi_{\Theta|\Theta_1}(\theta|k)\pi_{\Theta_1}(k) = \frac{1-|1-2\theta|}{4\log(2)\theta(1-\theta)}, \quad 0 < \theta < 1$$

Hierarchical Bayes (continued) IV

- In this case, we can get a closed-form expression for $\pi(\theta)$, but as you can tell, it can often get very difficult to do this mathematically.
- Thus, while the hierarchical structure is equivalent to just setting $\pi(\theta)$ as our prior (and not worrying about hierarchical model), this additional structure can aid in computations.
- If we are looking to estimate, for instance, the posterior mean:

$$E_{\Theta|X}[\Theta],$$

Then the law of total expectation gives:

$$E_{\Theta|X}[\Theta|X] = E_{\Theta_1|X}\left[E_{\Theta|\Theta_1,X}[\Theta|\Theta_1, X]\right].$$

Hierarchical Bayes (continued) V

- Thus, the calculation of the posterior mean of $\Theta|X$ can be done without needing explicit form of the posterior $\Theta|X$, which can simplify the problem.
- Our particular choice of likelihood and prior makes it easy to calculate the marginal-likelihood of $\Theta_1 = k$:

$$\begin{aligned} \pi_{X|\Theta_1}(x|k) &= \int_0^1 f(x|\theta, k)\pi_{\Theta|\Theta_1}(\theta; k)\, d\theta \\ &= \binom{N}{x} \frac{B(x+k, N-x+k)}{B(k,k)}. \end{aligned}$$

- Also, The Beta distribution was picked because it is conjugate, so the posterior mean $\Theta|\Theta_1 = k, X$ is readily available:

$$E_{\Theta|\Theta_1=k,X} = \mu_k = \frac{x+k}{N+2k}.$$

Hierarchical Bayes (continued) VI

- Now we need to take the expectation of this, with respect to the marginal posterior (un-normalized weights) $\pi_{\Theta_1|x}(k|x)$:

$$\begin{aligned}\pi_{\Theta_1|x}(k|x) &\propto w_k \\ &= \pi_{X|\Theta_1}(x|k)\pi_{\Theta_1}(k) \\ &= \frac{B(x+k, N-x+k)}{2\log(2)B(k,k)k(2k-1)}.\end{aligned}$$

- Then, the normalized weights are:

$$\bar{w}_k = \frac{w_k}{\sum_j w_j} = p(k|x),$$

and the posterior mean is:

$$E[\Theta|x] = \sum_{k=1}^{\infty} \bar{w}_k \mu_k.$$

Hierarchical Bayes (continued) VII

- For this particular example, the sum can be calculated exactly. However, we can also approximate this using software by taking the first K partial sums. Check out the provided HB-code R code.

Empirical Bayes

Empirical Bayes

TODO

Uncertainty quantification

Uncertainty in Bayes estimates

- TODO

References and Acknowledgements

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA.

- Compiled on January 30, 2026 using R version 4.5.2.
-
- We acknowledge students and instructors for previous versions of this course / slides.