Mathematical Statistics II

The Bayesian Approach to Parameter Estimation

Jesse Wheeler

Contents

1 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.

2 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 $f(x|\theta)$.

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:

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

- There are a few things to note in the equation above. First, the likelihood function $L(\theta) = f(x^*|\theta)$ makes its presence on the right hand side of the equation.
- Next, the denominator is not a function of θ. As a result, it is just a normalizing constant to ensure that the posterior is a proper probability distribution.
- With this in mind, we often say that the posterior distribution $\pi_{\Theta|X}(\theta|x^*)$ is a product of the likelihood $L(\theta)$ and the prior $\pi_\Theta(\theta)$.
- There is a large number of notations that are often used. For instance, the symbol f is often used instead of π as a function. The most common is perhaps:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, d\theta}.$$

- The above notation does a bit of "function overload", but it is often clear from context and the symbols used as input what is meant.
- As before, f is taken to be either a pmf or pdf, depending on the problem.

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.

- The probability model describing the data is Binomial($N = 10, p = \theta$), which has mass function:

$$f_{X|\Theta}(n|\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}.$$

- After picking the model for the data $f(x|\theta)$, the next step is to define our prior belief about the coin, characterized by $\pi_\Theta(\theta)$.
- A natural prior might be: "I know nothing about the coin, all probabilities are possible". Then, our prior would be uniform:

$$\pi_\Theta(\theta) = 1(0 \le \theta \le 1)$$

- Thus, we previously found the posterior to be:

$$\begin{aligned}\pi(\theta|n) &= \frac{\binom{N}{n}\theta^n(1-\theta)^{N-n}}{\int_0^1 \binom{N}{n}\theta^n(1-\theta)^{N-n}\, d\theta}1[0 \le \theta \le 1] \\ &= \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)}\theta^n(1-\theta)^{N-n}1[0 \le \theta \le 1]\end{aligned}$$

- We then demonstrated that this corresponds to a beta distribution. Thus:

$$\Theta|X = n \sim \text{Binom}(\text{Beta}(n + 1, N - n + 1))$$

- Now our belief about Θ has been updated using the data $X = n$. This belief is represented not be a single point, but an entire distribution.
- There are multiple ways to get a single point estimation.
- One idea is the mean of the posterior, $E[\Theta|X = n]$. In this case, the mean of the Beta distribution is known, and we find:

  $$E[\Theta|X = n] = \frac{n + 1}{N + 2}.$$

  As mentioned, this is like a regularized version of the MLE (n/N), as the estimate is "pulled" toward the center $\theta = 0.5$.
- Another common approach is similar to what we did with the MLE: if the posterior represents our updated belief as a distribution, why don't we let our point estimate be the *maximum* of that belief? In this setting, the maximum of a probability density is the *mode*. The mode of the Beta(α, β) distribution is given by:

  $$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

- For our specific posterior, this implies that the mode is:

  $$\hat{\theta} = \frac{(n + 1) - 1}{(n + 1) + (N - n + 1) - 2} = \frac{n}{M}.$$

- Note that the mode in this case matches the MLE!
- The mode of the posterior distribution is called the Maximum A Posteriori (MAP) estimate. This is a common choice for a point estimate, in particular by Frequentists who use Bayesian methods to solve a given problem. There are some advantages and disadvantages of this approach, one being that it is not a properly weighted version of our belief; it also lacks some of the proprieties and guarantees that Bayesian statisticians like. The MAP can also be difficult (or impossible) to compute in many situations, whereas the posterior mean and median can readily be approximated using samples from a distribution.
- Another possible point estimate is the median of the posterior distribution. There's not a closed-form expression for the median of a Beta-distribution, but it can be calculated via software.

- 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.

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

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

- A proof sketch is given in class, but is left as an exercise in these notes.

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

3 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^*.

- First we find the density of $X|\Lambda = \lambda$, which begins with finding the density of a single observation $X_i|\Lambda = \lambda$:
$$f_{X_i|\Lambda}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, 2, \ldots\}.$$
- Under the IID assumption, the joint density of X is:
$$\begin{aligned} f_{X|\Lambda}(x|\lambda) &= \prod_{i=1}^{n} f_{X_1|\Lambda}(x_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$
- Now we want to find the posterior of $\Lambda|X$, which is given by:
$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_{i=1}^{n} x_i^*} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}.$$
  Above, $f_\Lambda(\lambda)$ is the prior distribution of Λ, and the product $\prod_{i=1}^{N} x_i!$ canceled out as it was in both the numerator and denominator.
- Now there are two remaining steps, the parts that are often challenging in Bayesian estimations: (1) choosing the prior, (2) computing the integral in the denominator.
- We will consider two different approaches for picking the prior. The first is the traditional / orthodox Bayesian, who takes very seriously the philosophy that the prior distribution captures their prior opinion.
- In this orthodox approach, the prior density $f_\Lambda(\lambda)$ should be specified *before* ever seeing the data (the whole point is this is our prior belief before observing data).
- This itself is not an easy task; even in this scenario, we may pick a prior based both on belief, and convenience.

- That is, suppose that we believe the prior mean $E[\Lambda] = \mu = 15$, with variance $\text{Var}(\Lambda) = \sigma^2 = 25$. There's a lot of distributions out there that have these features, but we will pick the Gamma distribution because it will be mathematically convenient.

- Since the Gamma(α, β) has mean $15 = \alpha/\beta$ and variance $25 = \alpha/\beta^2$, we can solve and get our prior density for Λ as:
$$\Lambda \sim \text{Gamma}(\alpha = 9, \beta = 3/5).$$

- Note that the choice of the mean and variance can (and should) be aided by plotting the function, and typically the Gamma distribution has parameter values α and λ, but we're already using λ for something else here.

- Thus, the prior density is:
$$f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0.$$

- After canceling constants (anything not involving λ) and combining like-terms, we get:
$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda}}{\int_0^\infty \lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda}\, d\lambda}$$

- Now we encounter a common trick (and the reason we picked a Gamma prior). Note that the denominator is *only* a function of x, and not λ. Thus,
$$\begin{aligned} f_{\Lambda|X}(\lambda|x^*) &\propto \lambda^{\sum_{i=1}^n x_i^* + \alpha - 1} e^{-(n+\beta)\lambda} \\ &= \lambda^{\text{something}} e^{-\text{something}\lambda}. \end{aligned}$$

  Here, λ is the variable of interest (not a constant). Thus, we want to compare this statement to other *kernels* that look like:
$$f(x) \propto x^a e^{-b\lambda}.$$

- This kernel matches that of the standard Gamma(γ, ζ) distribution:
$$f(x) \propto x^{\gamma-1} e^{-\zeta x}$$

  (again, swapping variables γ, ζ for α and λ for obvious reasons).

- We can immediately conclude that the posterior MUST be a Gamma distribution (since it will integrate to one). What is left is to pick the corresponding parameter values. To do this, we must have:
$$\gamma - 1 = \sum_{i=1}^n x_i^* + \alpha - 1 \implies \gamma = \sum_{i=1}^n x_i^* + \alpha,$$

  and
$$-\zeta = -(n + \beta) \implies \zeta = n + \beta.$$

  Thus, the posterior distribution is:
$$\Lambda|X = x^* \sim \text{Gamma}\big(\sum_{i=1}^n x_i^* + \alpha, n + \beta\big).$$

- Then all we need to do is plug is our specific data values x_i^*, and the specific values of our prior $\alpha = 9$, $\beta = 3/5$.

- From this, we can get various point estimates: posterior mean, MAP, or posterior median. We can also talk about posterior variance if we want.

- **This trick of avoiding calculating the integral in the denominator is extremely common, and it will appear again. Make sure this makes sense to you**.

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

- The setup for the problem is the exact same, so note that the posterior distribution is:
$$f_{\Lambda|X}(\lambda|x^*) = \frac{\lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}.$$
- Now suppose we really don't have a good guess for the parameter λ, or we want to more utilitarian / noncommittal approach. Now what?
- The default is a uniform probability, but the possible values of λ is the interval $(0, \infty)$; we can't have a uniform prior on this interval (it doesn't exist)
- Instead, we will feign confidence that λ must be smaller than some fixed number based on the problem at hand. For instance, maybe $0 < \lambda \leq 100$ is reasonable.
- Then, the prior would be:
$$f_\Lambda(\lambda) = \frac{1}{100} 1(0 < \lambda \leq 100),$$
and the posterior would be:
$$\begin{aligned} f_{\Lambda|X}(\lambda|x^*) &= \frac{\frac{1}{100} \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}}{\frac{1}{100} \int_0^{100} \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}\, d\lambda} 1(0 < \lambda \leq 100) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}}{\int_0^{100} \lambda^{\sum_{i=1}^n x_i^*} e^{-n\lambda}\, d\lambda} 1(0 < \lambda \leq 100). \end{aligned}$$
- In this case, we can't just do the denominator-integration trick! It looks very similar, because the denominator is still a constant, and the kernel in the numerator looks very similar to a Gamma kernel, **but** we have a new bound $0 < \lambda \leq 100$ that makes it distinct from the Gamma distribution, since the Gamma distribution has support on $(0, \infty)$.
- Unfortunately, there is not an easy way to compute the integral either (partly because of the bound). Thus, the integral needs to be computed numerically.
- For now, we will just use the `integrate` function in R, which is really good at integrating univariate functions.
- The posterior mean can similarly be found using a numeric integration technique, and posterior mode can be found using numeric optimization strategies covered in the last set of lecture notes.

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.

```
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
```

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

4 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.

5 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*.

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.
- 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

6 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 $\pi(\theta; \psi)$.
- How do we choose ψ? Why don't we put a prior on these 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).
- Philosophically, this situation naturally arises if we want to pick a conjugate prior for Θ, but are not committal about the *hyperparameters* Ψ that define the distribution of Θ. Then, Ψ should also be modeled as a random variable.
- This leads to a hierarchical model:
$$X|\Theta = \theta, \Psi = \psi \sim f(x|\theta, \psi),$$
and
$$\Theta|\Psi = \psi = g(\theta|\psi), \quad \Psi \sim \pi(\psi).$$
- Why would we want to do this? Do we need a prior now hyper-parameters for ψ?
  1. We'll see that this is sometimes a very useful way of thinking about a problem.
  2. We could, but this is not usually useful.

Trivial case: hierarchical Normal-Normal
Suppose that the data X_i are iid $N(\theta, 1)$. Set a prior for θ as $\Theta|M = \mu \sim N(\mu, 1)$, and $M \sim N(0, 1)$.

- This model seems rather odd, unless there is a good reason to get the posterior $(\Theta, M)|X = x^*$.
- Otherwise, we can show that the *hyperparameter* M can be eliminated from the model.
- In particular, consider the marginal distribution of Θ:

$$\pi_\Theta(\theta) = \int \pi_{\Theta|M}(\theta|\mu)\pi_M(\mu)\, d\mu,$$

- Some algebra (i.e., completing the square in the exponential terms), shows that this implies: $\Theta \sim N(0, 2)$.
- Thus, if we're not interested in the hyper-parameters themselves, then what we are interested in is:

$$\Theta|X = x^*,$$

  which can just be calculated by setting the prior $\Theta \sim N(0, 2)$, and following the standard Bayesian approach.

7 Empirical Bayes

Empirical Bayes
TODO

8 Uncertainty quantification

Uncertainty in Bayes estimates

- TODO

Acknowledgments

- Compiled on January 29, 2026 using R version 4.5.2.

- We acknowledge students and instructors for previous versions of this course / slides.

References

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1