STAT 460/560 Class 10: Parametric Bayesian inference

Ben Bloem-Reddy

Reading: Chapter 11, [Was04]; Chapter 3, [EH21].

In contrast to the frequentist interpretation of probability, the Bayesian statistical framework asserts that probability is personal. It is sometimes described as "degree of belief," but it does not necessarily mean that. It can just describe your personal uncertainty, lack of knowledge, etc., about an event or any quantity of interest—even model parameters, which a frequentist treats as fixed.

In the Bayesian framework, one starts with a prior distribution and a likelihood, then combines them with data to produce the posterior distribution, from which all inferences are drawn.

One thing to note is that because Bayesian statistics do not rely on a long-run frequency interpretation of probability, there are no appeals to hypothetical repeated experiments; just the data at hand. As a side-effect, quantifying the performance of Bayesian methods is inherently difficult, because there is no "ground truth" in the framework. Most performance metrics can be thought of as frequentist assessments of Bayesian procedures. We'll talk more about that next time.

1. The Bayesian method

The mechanics of Bayesian inference are simple:

1. For a parameter of interest, choose a **prior distribution** (or density) $f(\theta)$ that expresses our beliefs/knowledge about θ before we see any data.
2. Choose a statistical model in the form of a likelihood, $f(x|\theta)$. (Note this is now a conditional PDF, rather than just a function of θ, $f(x;\theta)$.)
3. Observe data $x^n := (x_1, \dots, x_n)$ and use Bayes' Theorem to update our beliefs/knowledge via the **posterior distribution**,

$$f(\theta|x^n) = \frac{f(x^n, \theta)}{f(x^n)} = \frac{f(\theta)f(x^n|\theta)}{\int f(\theta)f(x^n|\theta)\ d\theta} \propto \mathcal{L}_n(\theta)f(\theta)\ . \tag{10.1}$$

Once we have the posterior, we can base all of our inferences on it.

Exercise 10.1. Suppose we are interested in λ, the time to failure of our shoelaces, measured in weeks. Suppose we have n observations, which we model as $\mathsf{Poisson}(\lambda)$ random variables, and use a $\mathrm{Gamma}(a, b)$ prior (where b is the rate parameter, not scale).

- What is the posterior distribution of $\lambda|x^n$? Do you need to compute the normalizing constant, $c_n = \int \mathcal{L}_n(\theta)f(\theta)\ d\theta$ to know the posterior distribution?
- What is the posterior mean?
- What is the posterior mode? (This is called the **maximum a posteriori (MAP)** estimator.)
- What is a posterior $1 - \alpha$ credible interval?
- Simulate a dataset of size $n = 10$, choose parameters for the prior, and simulate from the posterior distribution of $\ln(\lambda)|x^n$.

- If a replacement shoelace costs \$2, what is your expected (under the posterior) expenditure on shoelaces?

Besides easy-to-obtain quantities like posterior mean, quantities such as the mode, quantiles, or the entire posterior can be used to make inferences over functions of the parameter (as in the last part of the activity), and to make predictions. The **posterior predictive distribution** is

$$f(x_{n+1}|x^n) = \int f(x_{n+1}|\theta) f(\theta|x^n)\, d\theta \,. \tag{10.2}$$

As with the posterior distribution, the posterior predictive distribution can be used to make predictive inferences.

Exercise 10.2. Continuing the previous activity, find the posterior predictive distribution. What is the expected lifetime of your next shoelace?

2. Large-sample properties of Bayesian procedures

Before discussing the choice of prior distribution, the following theorem indicates that for very large samples, Bayesian procedures are very close to those based on the MLE.

Theorem 10.1 (Bernstein–von Mises Theorem)**.** *Assume that* $X_1, \ldots, X_n \sim_{\text{IID}} P_{\theta_*}$. *Let* $\hat{\theta}_n$ *be the MLE and* $\widehat{\mathsf{se}} = 1/\sqrt{n\mathcal{I}(\hat{\theta}_n)}$. *Under appropriate regularity conditions, the posterior distribution is approximately* $\mathcal{N}(\hat{\theta}_n, \widehat{\mathsf{se}})$. *Also, if* C_n *is the approximate* $1-\alpha$ *confidence interval, then*

$$\mathbb{P}(\theta \in C_n | X^n) \to 1 - \alpha \,. \tag{10.3}$$

The basic idea is that as long as the prior isn't concentrated away from θ_*, as n gets large the likelihood will overwhelm the prior, so that the posterior is essentially equal to the likelihood. The regularity conditions tend to be very technical, largely in terms of quantities that appear in the proof. One typical condition is that the prior places "enough" probability mass around θ_*, the underlying value of θ. (Pause and think for a few minutes about whether/when/how this makes sense, if any.) A sufficient condition is something like "the prior has full support." We won't cover it in this class, but it can be found in Chapter 10 of [van98] (which you'll have the background to understand in mid-November).

3. Prior specification

The biggest criticism of Bayesian methods is that they rely on the specification of a prior distribution, and that different priors will lead to different posteriors—hence to different inferences. There's no getting around it. To a devoted Bayesian, this is a feature rather than a bug. To others, this is a serious flaw. The philosophical disagreements run deep on this point (and others). Some things to consider: What effects do different priors have in different situations? When the sample size is large? Small? What is the effect of a prior relative to other aspects of the inferential pipeline? For example, what data to measure/collect? The likelihood model?

Let's go ahead and think about specifying a prior distribution. Section 11.6 in [Was04] doesn't really do the problem justice. It's the typical "frequentist using Bayesian methods" account of the problem, and neglects much of the nuance of the problem. For example, *prior elicitation* refers to the process of turning expert knowledge into well-defined prior distributions. See the recent review article [Mik+23] if you're interested.

I also recommend reading Robert [Rob07, Ch. 3] to get an idea of different approaches to prior specification, including various "objective" and "non-informative" priors. We'll talk about this more in the context of decision theory, but it turns out that Bayesian methods can have good (optimal) properties under frequentist standards of performance, so there is interest even among frequentists in their use, as long as one can exorcise as much subjectivism as possible.

If one is unable or unwilling to specify a prior distribution that accurately captures prior knowledge/belief, the default is to try to specify a **noninformative prior**. What does it mean for a prior to be noninformative?

We'll consider just a couple of possibilities here; for more on the problem, Kass and Wasserman [KW96] is no longer cutting-edge but stands up well.

Once we've specified our model (and thus the parameter space), one obvious possibility is to use a **flat prior**. If Θ is compact then this is just the uniform distribution on Θ.

Activity 10.1. Suppose $X_1, \ldots, X_n$ are modeled as conditionally IID Bern(p) random variables. What is the flat prior for p? What is the posterior distribution in this case?

Solution: The flat prior here is uniform, i.e., $\mathsf{Beta}(1,1)$. The posterior can be found using Bayes' Rule (for a generic beta prior with parameters a, b:

$$f(p|x_1,\ldots,x_n) \propto p^{a-1}(1-p)^{b-1}p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \quad \propto p^{a+\sum_{i=1}^n x_i-1}(1-p)^{b+n-\sum_{i=1}^n x_i-1} ,$$

which tells us that the posterior is $\mathsf{Beta}(a+\sum_{i=1}^n x_i, b+n-\sum_{i=1}^n x_i)$.

What if Θ is not compact (e.g., $\Theta = \mathbb{R}$)? In this case, a flat prior is an example of an **improper prior**—a density that does not integrate to 1. In many cases, we can still get a proper posterior distribution.

Exercise 10.3. Suppose $X_1, \ldots, X_n$ are modeled as conditionally IID $\mathcal{N}(\mu, 1)$ random variables. What is the flat prior for μ? What is the posterior distribution in this case?

Solution:

$$\begin{aligned} f(\mu|x_1,\ldots,x_n) &\propto e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2} \\ &\propto e^{-\frac{n}{2}(\mu-\bar{x}_n)^2} . \end{aligned}$$

Hence, the posterior is $\mathcal{N}(\bar{x}_n, 1/n)$.

The main criticism of flat priors is that they depend on how the model is parameterized. Suppose we have $X_1, \ldots, X_n$ conditionally IID Bern(p) random variables, except we parameterize the model with

$$\psi = \ln \frac{p}{1-p} . \tag{10.4}$$

This is a one-to-one transformation and the flat prior on p is transformed into

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2} , \tag{10.5}$$

which is not flat (see Fig. 1).

Abstractly, "flat" depends on the geometry of the parameter space, and will change when mapping between parameter spaces. The **Jeffreys prior** accounts for the geometry to yield a prior that is invariant with respect to reparameterizations. It is

$$f(\theta) \propto \det(\mathcal{I}(\theta))^{1/2} = \left(\prod_{j=1}^k \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta_j} \ln f(X|\theta) \right)^2 \right] \right)^{1/2} =^* \left(\prod_{j=1}^k \left| \mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_j^2} \ln f(X|\theta) \right] \right| \right)^{1/2} . \tag{10.6}$$

To see that this is invariant under reparameterization, observe that if $\psi = h(\theta)$ for some one-to-one transformation h, then

$$\det(\mathcal{I}_\Psi(\psi)) = \det(\mathcal{I}(h^{-1}(\psi))) \det(J_{h^{-1}}(\psi))^2 , \tag{10.7}$$

where J_h is the Jacobian of h (convince yourself this is true). Thus,

$$f_\Psi(\psi) \propto \det(\mathcal{I}_\Psi(\psi))^{1/2} = \det(\mathcal{I}(h^{-1}(\psi)))^{1/2} \det(J_{h^{-1}}(\psi)) \tag{10.8}$$

$$= f_\Theta(h^{-1}(\psi)) \det(J_{h^{-1}}(\psi)) , \tag{10.9}$$

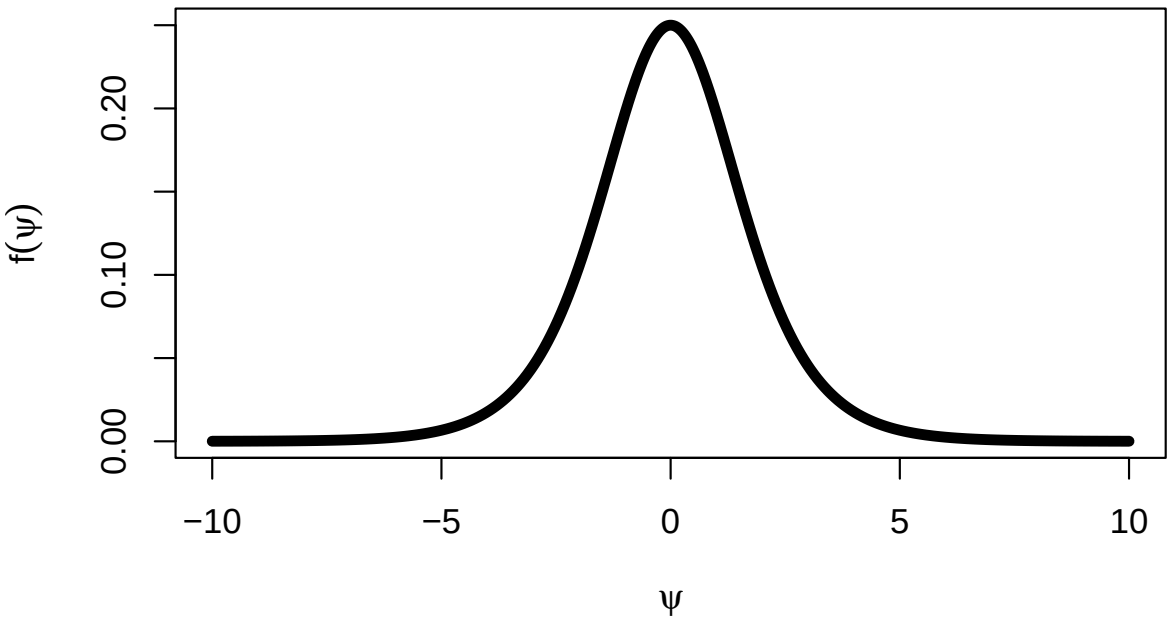

Figure 1: Transformation of flat prior on p into the prior (10.5) on ψ.

which is just the usual transformation rule of the Jeffreys prior on Θ (i.e., what one would get by specifying $f_\Theta(\theta) \propto \det(\mathcal{I}(\theta))^{1/2}$ and then transforming to $\psi = h(\theta)$). The interpretation of this is that if one starts by specifying the prior on Θ and then transforming it via $h(\theta)$, they should get the same prior as when specifying the prior on Ψ directly.

Activity 10.2. Find the Jeffreys prior for a Bern(p) likelihood model.

Solution: The Fisher information is $1/(p(1-p))$, so the Jeffreys prior is proportional to $1/\sqrt{p(1-p)} = p^{-1/2}(1-p)^{-1/2}$. This corresponds to a $\mathsf{Beta}(1/2, 1/2)$ prior.

Exercise 10.4. Find the Jeffreys prior for a $\mathcal{N}(\mu, \sigma^2)$ likelihood model.

Solution: Working through the math yields a flat prior on μ and an improper prior $\propto 1/\sigma^2$ on σ^2.

References

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.

[KW96] R. E. Kass and L. Wasserman. "The Selection of Prior Distributions by Formal Rules". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1343–1370. URL: http://www.jstor.org/stable/2291752.

[Mik+23] P. Mikkola et al. "Prior Knowledge Elicitation: The Past, Present, and Future". In: *Bayesian Analysis* (2023). URL: https://projecteuclid.org/journals/bayesian-analysis/advance-

publication/Prior-Knowledge-Elicitation-The-Past-Present-and-Future/10.1214/23-BA1381.full.

[Rob07] C. P. Robert. *The Bayesian Choice*. 2nd ed. Springer New York, 2007.

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.