Basic Frequentist Motivation for Bayesian Statistics

Thomas Wiemann
University of Chicago

TA Discussion # 1
Econ 31740

January 18, 2022

Outline

1. Logistics

2. General Advice for the Problem Sets

3. Basics Frequentist Motivation for Bayesian Statistics

   - Brief Review of Bayesian Statistics

   - The Bernstein-van Mises Theorem

Logistics

TA office hours on Fridays, 9-10am, in the grad lounge (SHFE 201).

TA sessions will typically typically:

- Explore extensions and applications of material studied in class, or
- Discuss tools you may need for subsequent problem sets, or
- Review key results and related literature.

I'll post my material on canvas a few hours before class. (May still contain typos and could be updated after class.)

General Advice for the Problem Sets

Problem sets for this class can be labor intensive.

- Start working on them as early as possible.

Cooperate with peers, but submit your own work only.

- Don't copy answers or code from your peers (it is not worth it).

The four Cs of ~~diamonds~~ problem sets:

- **C**omprehensive – answer all (sub) parts of the questions. All necessary derivations should be included.
- **C**oncise – a few lines of math speak a thousand words.
- **C**lear – define all variables and parameters clearly. Highlight when you use non-trivial results.
- **C**orrect – avoid errors. (Typically the most difficult.)

General Advice for the Problem Sets (Contd.)

For the coding exercises:

- Carefully (but not excessively) comment your code.

```
# Good
beta <- solve(crossprod(X)) %*% crossprod(X, y) # OLS coefficient

# Bad
beta[1] < 0 # checks whether beta[1] is less than zero
```

- Do not exceed 80 characters per line of code! Most IDEs have an option to display a vertical line at 80 characters. (Use it.)

- Don't reinvent the wheel: make use of style guides for Julia (e.g., Blue), R (e.g., Hadley Wickham), or Python (e.g., PEP 8).

- Code smart: think about design patterns.

General Advice for the Problem Sets (Contd.)

Only submit LaTeX tables. (Never screenshots from software output. Maybe this helps: excel2latex) Your tables should be well structured and contain all necessary annotations. For example:

Table 1: Parameters and Prior Distributions

	β_1	β_2	β_3	β_4	σ_b^{-2}	σ_ε^{-2}
True Value	-2	-1	0	1	$\frac{1}{4}$	$\frac{1}{2}$
$\mathrm{E}_\pi[\theta]$	0	0	0	0	1	1
$\mathrm{Var}_\pi(\theta)$	100	100	100	100	100	100

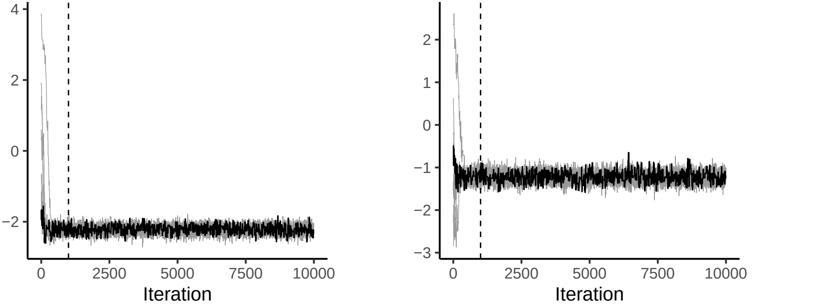
Notes. Independent Gaussian and Gamma priors are chosen for slope parameters (β_1 to β_4) and precision parameters (σ_b^{-2} and σ_ε^{-2}), respectively.

(I'm happy to share my LaTeX templates! Message me or check here.)

General Advice for the Problem Sets (Contd.)

Submit clean, self-contained plots. When possible, include vectorized figures (e.g., pdf rather than png). For example:

Figure 1: MCMC Trace



(a) β_1

(b) β_2

Notes. First 10,000 iterations of 10 MH-MCMC samplers with random initializations. Dashed lines indicate the burn-in cutoff at 1,000 iterations.

Questions

Any questions?

Next up: Basic Frequentist Motivation for Bayesian Statistics

Brief Review of Bayesian Statistics

Consider a family of probability distributions – a statistical model – defined by

$$\mathcal{F} = \{P_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}, \tag{1}$$

where $\mathcal{X}$ is the sample space of the observed data x, and Θ is the parameter space of the unobserved parameter θ.

Both frequentists and Bayesians observe a sample x from the random variable X with distribution P_θ and infer a value of the unknown parameter θ. Their key difference lies in how θ is modeled.

While frequentist statistics views θ is a fixed parameter, Bayesian analyzes model θ itself as a random draw from a so-called prior distribution $\pi(\theta)$. This approach allows for formulation of a joint distribution of X and θ characterized by the conditional distribution of the data and the prior:

$$(X, \theta) \sim P(X, \theta) = P_\theta(X) \times \pi(\theta). \tag{2}$$

Brief Review of Bayesian Statistics (Contd.)

The key object of interest in Bayesian inference is the *posterior distribution* of θ given the observed data x (often referred to as simply the "posterior"). It follows directly from Bayes rule:

$$\begin{aligned} \pi(\theta|x) &= \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta)\pi(\theta)d\theta} \\ \Rightarrow \quad \pi(\theta|x) &\propto p(x|\theta)\pi(\theta), \end{aligned} \tag{3}$$

where $p(x|\theta)$ denotes the conditional density of the data.

Note that the *marginal likelihood* of the data – $\int_\Theta p(x|\theta)\pi(\theta)d\theta$ – is invariant in θ and can therefore be omitted for notational convenience.

Brief Review of Bayesian Statistics: A Simple Example

To get some intuition, consider a simple Bayesian inference example. Suppose we observe iid data $\{x_i\}_{i=1}^n$ from a normal random variable X with unit variance and unknown mean θ – i.e., $X \sim \mathcal{N}(\theta, 1)$, Let $\theta \sim \mathcal{N}(\mu, \tau^2)$ with μ and τ being known prior parameters. Then

$$\begin{aligned}
\pi(\theta|x) &\propto \left[\prod_{i=1}^n p(x_i|\theta)\right]\pi(\theta) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\right)\exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\sum_{i=1}^n (x_i-\theta)^2 + \frac{1}{\tau^2}(\theta-\mu)^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2(n+\tau^{-2})^{-1}}\left[\theta^2 - 2\theta\left(\frac{\frac{\mu}{\tau^2}+\sum_{i=1}^n x_i}{n+\tau^{-2}}\right)\right]\right) \\
&\propto \exp\left(-\frac{1}{2\tilde{\tau}^2}(\theta-\tilde{\mu})^2\right).
\end{aligned} \tag{4}$$

Brief Review of Bayesian Statistics: A Simple Example (Contd.)

It thus holds that $\pi(\theta|x) \overset{d}{=} \mathcal{N}(\tilde{\mu}, \tilde{\tau}^2)$ where

$$\tilde{\mu} = \frac{\frac{1}{\tau^2}\mu + \sum_{i=1}^{n} x_i}{n + \frac{1}{\tau^2}}, \qquad \tilde{\tau}^2 \quad = \frac{1}{n + \frac{1}{\tau^2}}. \tag{5}$$

Note that the posterior mean and variance of θ depend on both the observed data $\{x_i\}_{i=1}^{n}$ as well as the hyperparameters μ and τ^2. In particular, the smaller τ^2 relative to the sample size n, the larger the relative importance of the prior.

Strong influence of the prior on the posterior parameters can be viewed as undesirable. When this is the case, one approach that may (or may not!) reduce prior influence is to consider increasing prior variance (e.g., by increasing τ^2 in the above example). In the extreme, this results in ”flat” priors that place equal mass on the parameter space.

Brief Review of Bayesian Statistics: A Simple Example (Contd.)

Consider a flat prior on θ – i.e., $\theta \sim \pi(\theta) \propto 1$ with support on $\mathbb{R}$. Then

$$\begin{aligned}\pi(\theta|x) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\\ &\propto \exp\left(-\frac{1}{2n^{-1}}\left[\theta^2 - 2\theta\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\right]\right)\\ &\propto \exp\left(-\frac{1}{2\tilde{\tilde{\tau}}^2}\left(\theta-\tilde{\tilde{\mu}}\right)^2\right),\end{aligned}\tag{6}$$

where $\tilde{\tilde{\mu}}$ and $\tilde{\tilde{\tau}}^2$ are the posterior parameters (under a flat prior on θ).

Notice that in this example, $\pi(\theta|x)$ is a proper probability density, even when $\pi(\theta)$ was improper. This need not always be the case! Example

Brief Review of Bayesian Statistics: A Simple Example (Contd.)

How does prior influence change with sample size n?

From the previous analysis with a normal prior $\pi(\theta) \overset{d}{=} \mathcal{N}(\mu, \tau^2)$, we have $\pi(\theta|x) \overset{d}{=} \mathcal{N}(\tilde{\mu}, \tilde{\tau}^2)$ where

$$\begin{aligned} \tilde{\mu} &= \frac{\frac{1}{\tau^2}\mu + \sum_{i=1}^{n} x_i}{n + \frac{1}{\tau^2}} \to 0 + E_{P_{\theta_0}}[x_i], \\ \tilde{\tau}^2 &= \frac{1}{n + \frac{1}{\tau^2}} \to 0, \end{aligned} \tag{7}$$

as $n \to \infty$.

So prior choice is asymptotically irrelevant here.

The next result shows that posterior consistency and (approximate) normality of the asymptotic posterior distribution can be generalized.

The Bernstein-von Mises Theorem

Bernstein-von Mises Theorem (as taken from van der Vaart, 2000)

Let the experiment $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at the true parameter value θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that $\forall \varepsilon > 0$, there exists a sequence of tests ϕ_n of $H_0 : \theta = \theta_0$ against $H_1 : \|\theta - \theta_0\| \geq \varepsilon$ such that

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|\geq\varepsilon} P^n_\theta(1-\phi_n) \to 0. \tag{8}$$

Furthermore, let the prior distribution be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$\|P_{\sqrt{n}(\bar{\Theta}_n-\theta_0)|\{x_i\}_{i=1}^n} - \mathcal{N}(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})\|_{TV} \overset{P^n_{\theta_0}}{\to} 0, \tag{9}$$

where $\bar{\Theta}_n$ denotes the random parameter conditional on n data samples and $\Delta_{n,\theta_0} := \frac{1}{\sqrt{n}}\sum_{i=1}^n I_{\theta_0}^{-1}\frac{\partial \log p(x_i|\theta_0)}{\partial\theta_0}$.

The Bernstein-von Mises Theorem (Contd.)

The Bernstein-von Mises Theorem rationalizes Bayesian inference from a frequentist point of view: Under assumptions, the asymptotic posterior distribution eventually aligns with the sampling distribution of the maximum likelihood estimator.

This provides the rational in this course for conducting Bayesian inference. It motivates the use of computational methods such as MCMC samplers targeting the posterior distribution even when interested in frequentist interpretations of estimates.

(Note that there are many other excellent motivations for conducting – both "subjective" and "objective" – Bayesian inference, ranging from deeply philosophical to entirely pragmatic. We won't cover them here. If you're interested, Robert (2007) and Gelman et al. (2014) are excellent resources to learn more.)

The Bernstein-von Mises Theorem (Contd.)

Despite it's practical attractiveness, it's important to be aware of settings in which the Bernstein-von Mises Theorem (and hence this basic frequentist motivation of Bayesian inference) does not apply.

To name a few, the asymptotic posterior distribution may not align with the asymptotic distribution of the maximum likelihood estimator when the target parameter is not identified, the posterior distribution is improper, the dimension of the target parameter depends on the sample size, or the support of the prior distribution does not include the MLE. (See Gelman et al. (2014) Chapter 4.3 for illustrations and additional examples. This blog post is also entertaining.)

A further key difficulty lies in misspecification of the likelihood $p(x|\theta)$. When misspecified, the posterior mode may still align with the MLE, but the asymptotic posterior variance my not be the Fisher information matrix (Kleijn and van der Vaart, 2012). Interesting papers that address Bayesian inference under misspecification from different angles are Müller (2013) and Lyddon et al. (2019).

References

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 3. Chapman and Hall/CRC.

Kleijn, B. J. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Lyddon, S., Holmes, C., and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Example with Improper Posterior

Back to Main Deck

Suppose we observe iid data $\{x_i\}_{i=1}^n$ from a normal random variable X with unknown mean *and unknown variance* θ – i.e., $X \sim \mathcal{N}(\theta, \sigma^2)$. Consider a joint flat prior $\pi(\theta, \sigma^2) \propto 1$ with support $\mathbb{R} \times \mathbb{R}_+$.

The posterior distribution is given by

$$\pi(\theta, \sigma^2 | x) \propto \frac{1}{\sigma} exp\left(\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \theta)^2\right)\right). \tag{10}$$

Is this a proper probability distribution? We have

$$\begin{aligned} &\int_{\mathbb{R}_+}\int_{\mathbb{R}} \frac{1}{\sigma} exp\left(\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \theta)^2\right)\right) d\theta d\sigma^2 \\ = &\int_{\mathbb{R}_+} 2\pi d\sigma^2 = \infty, \end{aligned} \tag{11}$$

so that $\pi(\theta, \sigma^2 | x)$ is *not* a probability distribution.