STAT7630: Bayesian Statistics

Lecture Slides # 6

The Bayesian Prior (Other Prior Families and Types)

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,

Updated: September, 2024

Outline

Chapter 4 - The Bayesian Prior

- In Bayesian analysis, a prior distribution must be specified.
- The choice of prior can significantly influence posterior conclusions, particularly with small sample sizes.
- Next, we will explore several key methods for determining prior distributions.

Conjugate Priors

- Conjugacy refers to a prior and likelihood combination where the posterior distribution maintains the same form as the prior, with updated parameters.
- Examples of conjugate priors:
  - Bernoulli likelihood $\rightarrow$ Beta prior for p
  - Poisson likelihood $\rightarrow$ Gamma prior for λ
  - Normal (with unknown μ) likelihood $\rightarrow$ Normal prior for μ
  - Normal (with unknown σ^2) likelihood $\rightarrow$ Inverse Gamma prior for σ^2

Outline

Conjugate Priors: Other Examples

- Additional examples of conjugate priors:
  - Multinomial($p_1, p_2, \ldots, p_k$) likelihood $\rightarrow$ Dirichlet prior for $p_1, p_2, \ldots, p_k$
  - Negative Binomial(r, p) (with unknown p) likelihood $\rightarrow$ Beta prior for p
  - Uniform($0, \theta$) likelihood $\rightarrow$ Pareto prior for the upper limit θ
  - Exponential(λ) likelihood $\rightarrow$ Gamma prior for λ
  - Gamma(α, β) (with unknown α) likelihood $\rightarrow$ Gamma prior for α
  - Pareto(β, λ) (with unknown β) likelihood $\rightarrow$ Gamma prior for β
  - Pareto(β, λ) (with unknown λ) likelihood $\rightarrow$ Pareto prior for λ

Conjugate Priors: Exponential Family

- Consider the one-parameter exponential family of distributions.
- This family includes any distribution whose pdf (or pmf) can be written as:

$$f(x|\theta) = e^{t(x)u(\theta)} r(x) s(\theta)$$

  where $t(x)$ and $r(x)$ do not depend on the parameter θ, and $u(\theta)$ and $s(\theta)$ do not depend on x.
- This density can also be expressed as:

$$f(x|\theta) = e^{t(x)u(\theta) + \log r(x) + \log s(\theta)}$$

Conjugate Priors: Exponential Family

- For an i.i.d. sample $X_1, \ldots, X_n$, the joint density of the data is:

$$f(\mathbf{x}|\theta) = e^{u(\theta)\sum_{i=1}^{n} t(x_i) + \sum_{i=1}^{n} \log r(x_i) + n \log s(\theta)}$$

- Consider a prior for θ (with prior parameters k and η) of the form:

$$p(\theta) = c(k, \eta) e^{ku(\theta) + k \log s(\theta)}$$

Conjugate Priors: Exponential Family

- The posterior distribution is proportional to:

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto p(\theta)L(\theta|\mathbf{x}) \\ &\propto \exp\left(u(\theta)\sum t(x_i) + n\log s(\theta) + ku(\theta)\eta + k\log s(\theta)\right) \\ &= \exp\left(u(\theta)\left(\sum t(x_i) + k\eta\right) + (n+k)\log s(\theta)\right) \\ &= \exp\left((n+k)u(\theta)\left(\frac{\sum t(x_i)+k}{n+k}\right) + (n+k)\log s(\theta)\right) \end{aligned}$$

- This simplifies to:

$$\exp\left((n+k)u(\theta)\left(\frac{\sum t(x_i)+k\eta}{n+k}\right) + (n+k)\log s(\theta)\right)$$

- The posterior is of the same form as the prior, but with updated parameters: $k \to n+k$ and $\eta \to \frac{\sum t(x_i)+k\eta}{n+k}$
- Thus, if data are i.i.d. from a one-parameter exponential family, a conjugate prior will exist.

Conjugate Priors

- Conjugate priors are mathematically convenient.
- They can be flexible, depending on the choice of hyperparameters.
- However, they reflect very specific prior knowledge, so caution is advised when using them unless that prior knowledge is actually available.

Outline

Uninformative Priors

- **Uninformative:** Not providing particularly useful or interesting information.
- **Noninformative:** Lack or absence of information.
- These priors are designed to provide minimal specific information about the parameter(s).
- A classic uninformative prior is the *uniform* prior.
- A *proper* uniform prior integrates to a finite quantity.
- **Example:** For Bernoulli(θ) data, a uniform prior on θ is:

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1$$

A uniform prior is proper when the parameter θ has bounded support.

Uninformative Priors

- **Example 2:** For $N(0, \sigma^2)$ data, it is "reasonable" to assume, that, say $\sigma^2 < 100$, we could use the uniform prior:

$$p(\sigma^2) = \frac{1}{100}, \quad 0 \leq \sigma^2 \leq 100$$

  (even though σ^2 is not intrinsically bounded).
- An improper uniform prior integrates to infinity.
- **Example 3:** For $N(\mu, 1)$ data with:

$$p(\mu) = 1, \quad -\infty < \mu < \infty$$

  An improper prior is okay as long as the resulting posterior is proper.
- **Caveat:** Sometimes an improper prior can yield an improper posterior. See `S7630_S6_PriorFam_Sup1.pdf`.

Other Uninformative Priors

- Other methods for constructing uninformative priors include:
  - **Bernardo's reference prior:** A prior that maximizes the discrepancy between the prior and the posterior, and minimizes the discrepancy between the likelihood and the posterior (a "dominant likelihood prior").
  - **Improper prior:** A prior where $\int_\Theta p(\theta)d\theta = \infty$.
  - **Highly diffuse proper prior:** For example, for normal data with μ unknown, using $N(0, 1000000)$ as a prior for μ—this is very close to the improper prior $p(\mu) \propto 1$.

Improper Priors: Why They Are Okay

- **Improper Priors:** These are priors that do not integrate to a finite value over the parameter space.
- **Why Use Improper Priors?**
  - They are often used to represent a "uninformative" or "vague" prior when no strong prior information is available.
  - Improper priors can still lead to valid, proper posteriors if the data is sufficient to "regularize" the posterior.
  - Examples include:
    - $p(\mu) \propto 1$ for a normal mean, representing no prior knowledge about μ.
    - $p(\theta) \propto \frac{1}{\theta}$ with $\theta > 0$ for scale parameters, reflecting an uninformative prior over scales.
- **Key Point:** As long as the posterior is proper (i.e., it integrates to 1), using an improper prior is acceptable and can simplify analyses by reflecting minimal prior knowledge.

Outline

Invariance Property

- A limitation of the uniform prior is that its "lack of information" is not invariant under transformation.
- **Example 1:** Consider the odds of success, $\tau = \frac{\theta}{1-\theta}$.
- If the prior for θ is $p(\theta) = 1$, the Jacobian is:

$$|J| = \left| \frac{d}{d\tau} \left(\frac{\tau}{1+\tau} \right) \right| = \frac{1}{(1+\tau)^2}$$

- This gives $p_\tau(\tau) = \frac{1}{(1+\tau)^2}$ for $0 < \tau < \infty$.

Invariance Property: Prior on the Odds of Success

- A prior on the odds of success τ transforms the uniform prior into an "informative" prior for τ.
- Visual representation:

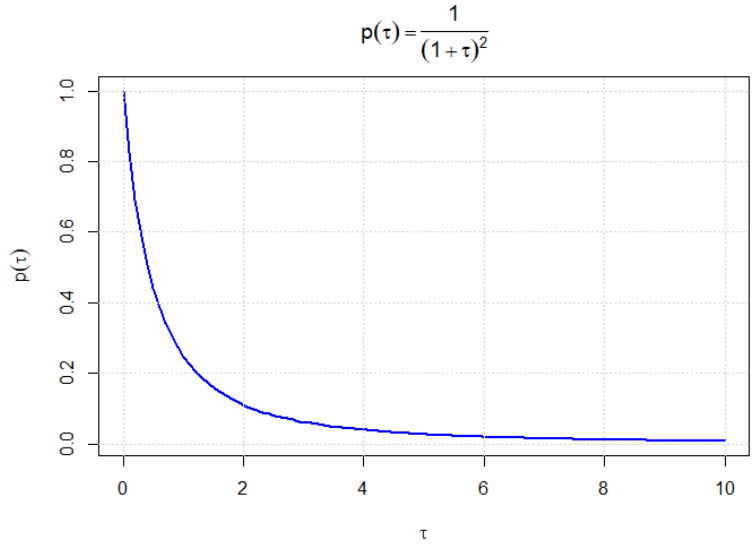

Figure 1: A Prior on the Odds of Success: $p_\tau(\tau)$ for $0 \leq \tau \leq 10$

- Despite $p_\tau(\tau)$ being an "informative" prior, note that:

$$P(0 < \tau < 1) = P(\tau > 1) = 0.5$$

Jeffreys Prior

- Jeffreys (1961) introduced a class of priors that are invariant under transformation $\tau = g(\theta)$.
- For a single parameter θ and data with joint density $f(\mathbf{x}|\theta)$, the Jeffreys prior is:

$$p_\theta^J(\theta) \propto \left[-\mathbf{E}\left(\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{X}|\theta) \right) \right]^{1/2} = [I(\theta)]^{1/2}$$

  where $I(\theta)$ is the Fisher information.
- For a parameter vector $\boldsymbol{\theta}$, the Jeffreys prior is:

$$p_\theta^J(\theta) \propto \left[-\mathbf{E}\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)' \left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right) \right]^{1/2}$$

Jeffreys Prior

Jeffreys Prior

Jeffreys Prior: Example 1

- Consider $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, with:

$$f(\mathbf{x}|\theta) = \theta^y (1-\theta)^{n-y}, \quad 0 \leq \theta \leq 1$$

  where $y = \sum_{i=1}^{n} x_i$.

- The log-likelihood is:

$$\log f(\mathbf{x}|\theta) = y \log(\theta) + (n-y) \log(1-\theta)$$

- First derivative:

$$\frac{\partial}{\partial \theta} \log f(\mathbf{x}|\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

- Second derivative:

$$\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta) = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

Jeffreys Prior: Example 1 (Continued)

- The expectation of the second derivative is:

$$-\mathbf{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

- The Jeffreys prior for θ is:

$$p_\theta^J(\theta) \propto \left[\frac{n}{\theta(1-\theta)}\right]^{1/2}$$

- Simplifying, we get:

$$p_\theta^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} = \theta^{1/2-1}(1-\theta)^{1/2-1}$$

Jeffreys Prior: Beta Distribution

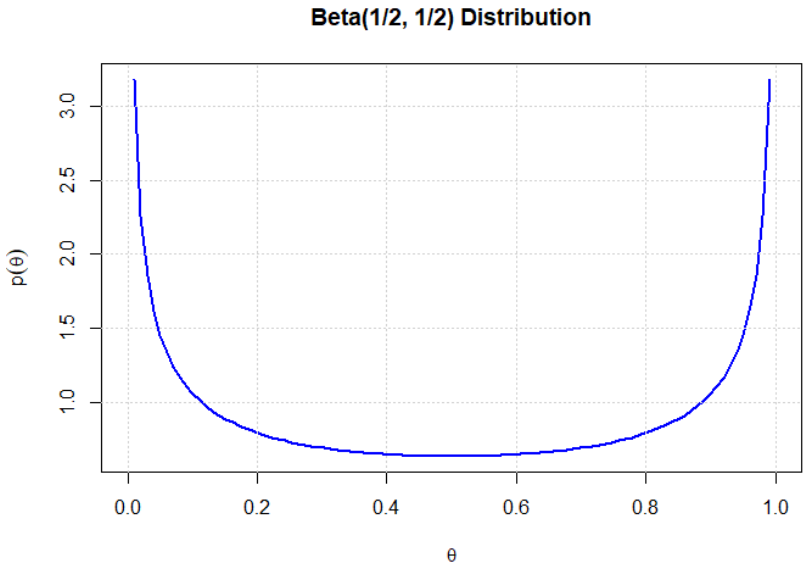

Figure 2: The Jeffreys prior for θ is a Beta$(1/2, 1/2)$ distribution.

This prior reflects more uncertainty around the extreme values of θ, making it suitable for modeling a success probability.

Jeffreys Prior: Invariance

- **Invariance Property:** If $\theta \sim p_\theta^J(\theta)$, which is the Jeffreys prior distribution for θ (i.e. $p_\theta^J(\theta) \propto I(\theta)^{1/2}$), then the transformed parameter $\tau = g(\theta)$ has the distribution $\pi(\tau)$:

$$\pi(\tau) = p_\theta^J(g^{-1}(\tau)) \left| \frac{d\theta}{d\tau} \right|$$

- It turns out that this is also equal to $I(\tau)^{1/2}$ (which is the Jeffreys prior distribution for τ).
- That is, if the (prior) distribution of θ is Jeffreys prior, then the (prior) distribution of the transformed parameter $\tau = g(\theta)$ is the Jeffreys prior for τ:

$$\theta \sim p_\theta^J(\theta) \propto \sqrt{I(\theta)} \quad \Rightarrow \quad \tau \sim p_\tau^J(\tau) \propto \sqrt{I(\tau)}$$

- **Question:** So, in what sense is Jeffreys prior invariant?
  It is invariant under transformations of the parameter(s) with respect to Fisher information.

Jeffreys Prior: Example 1 Revisited

- **Example 1 Revisited:** For $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, recall that Jeffreys prior for θ is:

$$\theta \sim p_\theta^J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$$

- The Jeffreys prior for the odds ratio $\tau = g(\theta) = \frac{\theta}{1-\theta}$:

$$\tau \sim p_\tau^J(g^{-1}(\tau)) \left| \frac{d\theta}{d\tau} \right| \propto \frac{1}{\sqrt{\tau}(1-\tau)}$$

- See `S7630_S6_PriorFam_Sup2.pdf` for the verification of the invariance property of Jeffreys prior in this setting:

$$p_\theta^J(\theta) \propto \sqrt{I(\theta)} \quad \Rightarrow \quad p_\tau^J(\tau) \propto \sqrt{I(\tau)}$$

Outline

Example 1: Quantile-Based Interval (Cabinet Duration)

- Suppose $X_1, \ldots, X_n$ are the durations of cabinets for a sample from Western European countries.
- We assume the X_i's follow an exponential distribution:

$$p(x|\theta) = \theta e^{-\theta x}, \quad x > 0$$

- The likelihood function is:

$$L(\theta|\mathbf{x}) = \theta^n e^{-\theta \sum_{i=1}^n x_i}$$

- Suppose the prior distribution for θ is:

$$p(\theta) \propto \frac{1}{\theta}, \quad \theta > 0$$

- This reflects that larger values of θ are less likely a priori.

Example 1: Quantile-Based Interval (Continued)

- The posterior distribution is:

$$p(\theta|\mathbf{x}) \propto p(\theta)L(\theta|\mathbf{x}) \propto \frac{1}{\theta}\theta^n e^{-\theta\sum x_i} = \theta^{n-1}e^{-\theta\sum x_i}$$

- This is the **kernel** of a **Gamma** distribution with:
  - "shape" parameter: n
  - "rate" parameter: $\sum_{i=1}^{n} x_i$
- Including the normalizing constant, the posterior distribution is:

$$p(\theta|\mathbf{x}) = \frac{(\sum x_i)^n}{\Gamma(n)}\theta^{n-1}e^{-\theta\sum x_i}, \quad \theta > 0$$

Example 1: Quantile-Based Interval (Continued)

- Given the observed data $x_1, \ldots, x_n$, we can calculate any quantiles of the Gamma distribution.
- The 0.05 and 0.95 quantiles provide a 90% credible interval for θ.
- **Note:** Refer to `Slide6_RCode` on Canvas for detailed calculations.
- **Question:** What would the posterior be if the prior were $p(\theta) \propto \theta^k$ for $\theta > 0$ and $k > n + 1$?

Example 1: Quantile-Based Interval (Noninformative Prior)

- Suppose we feel that $p(\theta) = \frac{1}{\theta}$ is too subjective, favoring small values of θ.
- Instead, consider the uninformative prior:

$$p(\theta) = 1, \quad \theta > 0$$

  which favors all values of θ equally.
- The posterior distribution is:

$$p(\theta|\mathbf{x}) \propto p(\theta)L(\theta|\mathbf{x}) = (1)\theta^n e^{-\theta \sum x_i} = \theta^{(n+1)-1} e^{-\theta \sum x_i}$$

- This is a Gamma distribution with parameters $(n+1)$ and $\sum x_i$.
- We can find the equal-tail credible interval similarly.

Informative Prior Forms: Power Priors

- **Informative priors** are typically based on expert opinion or previous research regarding the parameter(s) of interest.
- **Power Priors:**
  - Suppose we have access to previous data $\mathbf{x}_0$ that is analogous to the data we will gather.
  - The "power prior" is given by:

$$p(\theta|\mathbf{x}_0, a_0) \propto p(\theta)[L(\theta|\mathbf{x}_0)]^{a_0}$$

where $p(\theta)$ is a standard prior, and $a_0 \in [0, 1]$ is a parameter that measures the influence of the previous data.

Power Priors

- As $a_0 \to 0$, the influence of the previous data decreases.
- As $a_0 \to 1$, the influence of the previous data increases.
- The posterior, given new data $\mathbf{x}$, is:

$$p(\theta|\mathbf{x}, \mathbf{x}_0, a_0) \propto p(\theta|\mathbf{x}_0, a_0) L(\theta|\mathbf{x})$$

- To avoid specifying a single value for a_0, we could place a distribution on a_0, such as a Beta distribution, and average over values of a_0:

$$p(\theta|\mathbf{x}_0) \propto \int_0^1 p(\theta)[L(\theta|\mathbf{x}_0)]^{a_0} p(a_0)\, da_0$$