# STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 5A: Choosing Priors

# Noinformative Priors

- Idea: let the data speak for themselves. Suggestions: $\mathrm{Beta}(0,0)$, $\mathrm{Beta}(1,1)$, $\mathrm{G}(1,1)$.
- A way to "guarantee" prior distributions play a minimal role in the posterior distribution.
- One option: flat improper priors.
  - $y \sim \mathrm{N}(\mu, 1)$ with $\pi(\mu) \propto 1$
  - $y \sim \mathrm{N}(0, \sigma^2)$ with $\pi(\sigma^2) \propto 1/\sigma^2$
- Other approaches:
  - Maximize the information entropy of the parameter's probability distribution:

$$\arg\max_{\theta \in \Omega} h(\theta) = \arg\max_{\theta \in \Omega} \mathrm{E}[-\log \pi(\theta)] = \arg\max_{\theta \in \Omega} \int_\Omega -\pi(\theta) \log(\pi(\theta)) d\theta$$

  - Jeffreys' invariant prior (will look at next)
  - Weakly information priors. Eg: $\mathrm{G}(2,1)$.

# Jeffreys' Invariance Principle

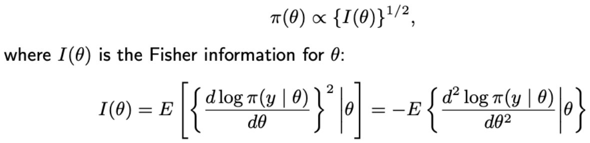
- Another information-based approach is based on Jeffrey's invariance principle
- Consider monotone one-to-one transformations of the parameter $\phi = h(\theta)$ where $h$ is a strictly increasing continuous differentiable function with inverse $\theta = g(\theta)$.
- By the change of variable formula, the prior density $\pi(\theta)$ is equivalent, in terms of expressing the same beliefs, to the following prior density on $\phi$:

$$\pi(\phi) = \pi(g(\theta))|g'(\theta)| \tag{1}$$

- Jeffreys' Principle: Any rule for determining the prior density $\pi(\theta)$ should yield an equivalent result if applied to the transformed parameter $\phi = h(\theta)$ for any one-to-one transformation $h$. Hence,
  - Determine $\pi(\theta)$ using $\pi(y|\theta)$
  - Then determine $\pi(\phi)$ using $\pi(y|\phi)$
  - While satisfying (1)

# Jeffreys' Prior

- Jeffreys' prior is given by

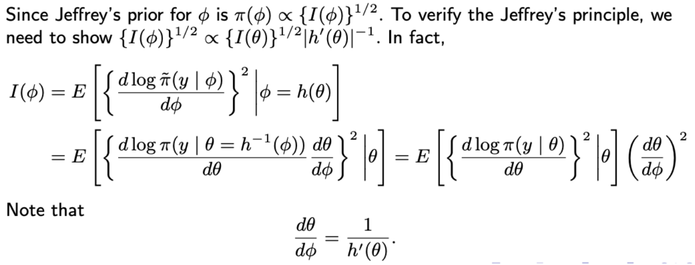


$$\pi(\theta) \propto \{I(\theta)\}^{1/2},$$

where $I(\theta)$ is the Fisher information for $\theta$:

$$I(\theta) = E\left[\left\{\frac{d\log\pi(y \mid \theta)}{d\theta}\right\}^2 \middle|\theta\right] = -E\left\{\frac{d^2\log\pi(y \mid \theta)}{d\theta^2}\middle|\theta\right\}$$

- Can we verify that Jeffreys' principle is satisfied for Jeffreys' prior selection method?



Since Jeffrey's prior for $\phi$ is $\pi(\phi) \propto \{I(\phi)\}^{1/2}$. To verify the Jeffrey's principle, we need to show $\{I(\phi)\}^{1/2} \propto \{I(\theta)\}^{1/2}|h'(\theta)|^{-1}$. In fact,

$$\begin{aligned} I(\phi) &= E\left[\left\{\frac{d\log\tilde{\pi}(y \mid \phi)}{d\phi}\right\}^2 \middle|\phi = h(\theta)\right] \\ &= E\left[\left\{\frac{d\log\pi(y \mid \theta = h^{-1}(\phi))}{d\theta}\frac{d\theta}{d\phi}\right\}^2 \middle|\theta\right] = E\left[\left\{\frac{d\log\pi(y \mid \theta)}{d\theta}\right\}^2 \middle|\theta\right]\left(\frac{d\theta}{d\phi}\right)^2 \end{aligned}$$

Note that

$$\frac{d\theta}{d\phi} = \frac{1}{h'(\theta)}.$$

# Jeffreys' Prior

- The choice of a prior depending on Fisher information is justified by the fact that $I(\theta)$ is widely accepted as an indicator of the amount of information brought by the model (or the observation) about $\theta$ (Fisher (1956)).
- Therefore, it seems intuitively justified that the values of $\theta$ for which $I(\theta)$ is larger should be more likely for the prior distribution.
- In other words, $I(\theta)$ can evaluate the ability of the model to discriminate between $\theta$ and $\theta + d\theta$ through the expected slope of $\log f(y|\theta)$.
- To favor the values of $\theta$ for which $I(\theta)$ is large is equivalent to minimizing the influence of the prior distribution and is therefore as non informative as possible.
- In fact, the Jeffreys prior is usually improper

# Example: Binomial Model

- Let $y \sim \text{Binomial}(n, \theta)$:

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Then,

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

- And,

$$\begin{aligned} I(\theta) &= -\mathrm{E}\left[ -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2} \right] \\ &= \frac{\mathrm{E}[y]}{\theta^2} + \frac{n-\mathrm{E}[y]}{(1-\theta)^2} \\ &= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\ &= \frac{n}{\theta} + \frac{n}{(1-\theta)} \\ &= \frac{n}{\theta(1-\theta)} \end{aligned}$$

- Hence the Jeffreys' prior is: $\pi(\theta) \propto [\theta(1-\theta)]^{-1/2}$
- This is a $\text{Beta}(1/2, 1/2)$ prior.

# Example: Normal Model

- Now, let $x \sim \mathrm{N}(\mu, \sigma^2)$.
- When $\sigma^2$ is known, what is the Jeffreys' prior for $\mu$?

$$\pi(\mu) \propto \sqrt{I(\mu)} = \sqrt{\mathrm{E}\left\{\left(\frac{x-\mu}{\sigma^2}\right)^2\right\}} \propto 1$$

- This is an improper prior.
- How about the Jeffreys' Prior for $\pi(\mu^3)$?
- This would be $\pi(\mu^3) \propto \frac{1}{\mu^2}$
- When $\mu$ is known, what is the Jeffreys' prior for $\sigma^2$?

$$\pi(\sigma^2) \propto \sqrt{I(\sigma^2)} \propto \sqrt{\mathrm{E}\left\{\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^4}\right)^2\right\}} \propto \frac{1}{\sigma^2}$$

- What is the Jeffreys' prior for $\log(\sigma^2)$?
- This would be $\pi(\log(\sigma^2)) \propto 1$

# Difficulties with Noninformative Priors

- Searching for a prior distribution that is always vague seems misguided: If the likelihood is truly dominant in a given problem, then the choice among a range of relatively flat prior densities cannot matter.

- For many problems, there is no clear choice for a vague prior distribution, since a density that is flat or uniform in one parameterization will not be in another. For example, for normal model $y \sim \mathrm{N}(\mu, 1)$, we can assume $\pi(\mu) \propto 1$, i.e., a uniform flat prior. How about $\pi\{\exp(\mu)\}$? still uniform?

- Noninformative priors are often useful when it does not seem to be worth the effort to quantify one's real prior knowledge as a probability distribution, as long as one is willing to perform the mathematical work to check that the posterior density is proper

# Weakly Informative Priors

Prior distribution is weakly informative if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge.

Examples: $\mu \sim \mathrm{N}(0, 10^6)$, $\sigma^2 \sim \mathrm{G}^{-1}(0.001, 0.001)$.

Principles for setting up the weakly informative priors.

Start with some version of a noninformative prior distribution and then add enough information so that inferences are constrained to be reasonable

Start with a strong, highly informative prior and broaden it to account for uncertainty in ones' prior beliefs and in the applicability of any historically based prior distribution to new data.

# Multiparameter Models

- Almost every practical problem involves more than one unknown parameters
- In many problems, we are only interested in one or two parameters. Although to have a realistic probability model we may have more parameters than we are ultimately interested in.
  - These extra parameters are often called nuisance parameters which can cause difficulty in classical statistics.
  - They are easily handled within the Bayesian framework. How?

# Multiparameter Models

Suppose we have a vector of parameters $\theta$

Divide $\theta$ into two subvectors: $\theta = (\theta_1, \theta_2)$

- $\theta_2$ is a vector of nuisance parameters and we are only interested in $\theta_1$,
- All of $\theta$ necessary for probabilistic modeling.

In Bayesian inference, $\pi(y \mid \theta)$ where $y$ is a vector of observations. The prior is $\pi(\theta)$. The posterior is

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

To obtain $\pi(\theta_1 \mid y)$, the *marginal posterior* of $\theta_1$, we integrate $\theta_2$ out of the posterior distribution of $\theta$

$$\begin{aligned}\pi(\theta_1 \mid y) &= \int \pi(\theta_1, \theta_2 \mid y) d\theta_2 \\ &= \int \frac{\pi(y \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2)}{\pi(y)} d\theta_2\end{aligned}$$

# Dirichlet-Multinomial Model

The multinomial distribution is a generalization of the binomial distribution

Let $y$ denote the $k$ vector of counts of observations in $k$ categories with $y_i$ the number of counts in category $i$ and let $n = \sum_i y_i$ and $\theta = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ with $\theta_i > 0$ and $\sum_i \theta_i = 1$. The density of $y$ given $\theta$ is

$$\pi(y \mid \theta) = \frac{n!}{\prod_i y_i!} \theta_1^{y_1} \cdots \theta_n^{y_n}.$$

Recall for binomial data, the conjugate prior was the beta distribution. The multivariate generalization of the beta is the Dirichlet distribution. Let $\theta_i > 0$ for all $i$ and $\sum_i \theta_i = 1$. Then

$$\theta \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$

If

$$\pi(\theta) = \frac{\Gamma(\sum \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

What is the posterior distribution?

$$\theta \mid y \sim \mathrm{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k).$$

# Example

- Suppose that you are interested in the probabilities that a six sided die will land on each of its six possible faces.
- Write down the uniform prior in terms of its Dirichlet form.
- Suppose that you toss the die five times observing: two 1's, one 3, and two 4's.
- Write down posterior distribution in its Dirichlet form.