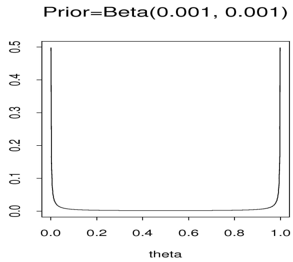
Comments

- For α, $\beta > 0$, we know
$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$
But for $\alpha = 0$ or $\beta = 0$ we have
$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \infty$$
So, there is no normalising constant such that $\int p(\theta)d\theta = 1$. Hence, Beta($\alpha, \beta$) is improper when $\alpha = 0$ or $\beta = 0$.
- If $y > 0$ and $n - y > 0$, the posterior $p(\theta \mid y)$ =Beta($y+\alpha, n-y+\beta$) =Beta($y, n-y$) is proper. That is, the improper prior has given a proper posterior.
- However, if $y = 0$ or $y = n$ (so $n - y = 0$), Beta($y, n-y$) is improper. The improper prior has given an improper posterior.



Choosing a Beta(ϵ, ϵ) prior, with ϵ very small will give a proper prior. It is nearly uniform (except for extreme values of θ). If the likelihood at extreme values of θ is negligible, the likelihood will dominate the prior.

Beware: if likelihood is NOT negligible at $\theta = 0$ or $\theta = 1$, such a prior will be highly informative because it places very high probability mass at $\theta = 0$ and $\theta = 1$.

E.g. $Y \sim$ Bin(n, θ), $\theta \sim$ Beta(0.001, 0.001) and $y = 0$, $n = 10$.

$$\begin{aligned} \theta \mid y &\sim \text{Beta}(0.001,\ 10.001) \\ E[\theta \mid y] &= 0.0001 \\ P[\theta < 0.0001 \mid y] &= 0.994 \end{aligned}$$

A posteriori, we are almost certain that θ is very small, despite having observed only 10 trials.

Jeffreys' prior

In addition to being often improper, uniform priors may not remain uniform under transformation.

Suppose we claim to know nothing about θ, and so say all values are equally likely: $p(\theta) \propto 1$. If we know nothing about θ, we should know nothing about $\phi = g(\theta)$, where $\phi = g(\theta)$ is a one-to-one transformation.

However, the prior for ϕ is
$$p_\Phi(\phi) = p_\Theta(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \left|\frac{d\theta}{d\phi}\right|$$
which is constant only if $\left|\frac{d\theta}{d\phi}\right|$ is constant, ie only if $g()$ is a linear transformation.

Thus, when $g()$ is NOT a linear transformation, our non-informative prior for θ is equivalent to that some values of ϕ are more likely than others; ie, we know something about ϕ!

E.g. Let $\phi = 1/\theta$. $\left|\frac{d\theta}{d\phi}\right| = 1/\phi^2$. So, $p(\phi) \propto 1/\phi^2 \Rightarrow$ small values of ϕ more likely than large values.

Therefore, one statistician might use uniform prior for θ, claiming this is non-informative, while another statistician might use uniform prior for $\phi = g(\theta)$, claiming this is non-informative.

Jeffreys (1960s) proposed a different rule for selecting non-informative prior: $p(\theta) \propto I(\theta)^{1/2}$, where $I(\theta)$ is the *Fisher Information*.

Fisher Information

The expected information about θ provided by an observable rv Y with distribution $p(Y \mid \theta)$ was defined by Fisher (1925) as
$$I(\theta) = -E_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(Y \mid \theta)\right] = E_{Y|\theta}\left[\left(\frac{\partial}{\partial\theta}\log p(Y \mid \theta)\right)^2\right]$$
(See Lee p.83 for proof of second form)

Comments

- The expectation is w.r.t. distribution $p(Y|\theta)$, so $I(\theta)$ depends on this distribution rather than any particular value of Y.
- If Y_k $(k = 1, \ldots, n)$ are iid random variables with distribution $p(Y|\theta)$ then the total information is $\sum_{k=1}^n I(\theta) = nI(\theta)$.

Jeffreys' Rule

Choose a non-informative prior for θ as $p(\theta) \propto I(\theta)^{1/2}$. This is called Jeffreys' prior for θ.

Theorem

Jeffreys' prior is invariant to reparametrisation, ie $p(\theta) \propto I(\theta)^{1/2} \Longleftrightarrow p(\phi) \propto I(\phi)^{1/2}$.

Proof

If $\phi = g(\theta)$ is a one-to-one transformation,

$$\frac{d}{d\phi}\log p(y \mid \phi) = \frac{d}{d\theta}\log p(y \mid \theta) \times \frac{d\theta}{d\phi}$$

Squaring and taking expectations gives:

$$I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2$$

So, if $p(\theta) \propto I(\theta)^{1/2}$, we have

$$\begin{aligned}
p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \\
&\propto I(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right| \\
&= I(\phi)^{1/2}
\end{aligned}$$

Example 3.8: Normal, known precision

Suppose $Y \sim \text{Normal}(\theta, \tau^{-1})$ with τ known. Then

$$\begin{aligned}
p(y \mid \theta) &\propto \exp\left[-\frac{\tau}{2}(y-\theta)^2\right] \\
\log p(y \mid \theta) &= -\frac{\tau}{2}(y-\theta)^2 + \text{const} \\
\frac{d}{d\theta}\log p(y \mid \theta) &= \tau(y-\theta) \\
\frac{d^2}{d\theta^2}\log p(y \mid \theta) &= -\tau \\
I(\theta)^{1/2} &= \sqrt{-E_{Y|\theta}(-\tau)} = \sqrt{\tau} \propto 1 \\
p(\theta) &\propto I(\theta)^{1/2} \propto 1
\end{aligned}$$

So Jeffreys' prior for θ is the uniform distribution. Note that this is an improper prior. (The prior is equivalent to Normal$(0, \infty)$, see Example 3.5 when $\phi_0 = 0$).

Example 3.9: Normal, known mean

Suppose $Y \sim \text{Normal}(\theta, \tau^{-1})$ with θ known.

$$\begin{aligned}
p(y \mid \tau) &\propto \sqrt{\tau}\exp\left(-\frac{\tau}{2}(y-\theta)^2\right) \\
\log p(y \mid \tau) &= \frac{1}{2}\log\tau - \frac{\tau}{2}(y-\theta)^2 + \text{const} \\
\frac{d}{d\tau}\log p(y \mid \tau) &= \frac{1}{2\tau} - \frac{1}{2}(y-\theta)^2 \\
\frac{d^2}{d\tau^2}\log p(y \mid \tau) &= -\frac{1}{2\tau^2} \\
I(\tau)^{1/2} &= \left[\frac{1}{2\tau^2}\right]^{\frac{1}{2}} \propto 1/\tau \\
p(\tau) &\propto I(\tau)^{1/2} \propto 1/\tau
\end{aligned}$$

So, Jeffreys' prior for τ is $p(\tau) \propto \tau^{-1}$.
Note this is a Gamma$(0, 0)$ distribution, $p(\tau) \propto e^{-0\tau}\tau^{0-1}$. This is an improper distribution.

Given Jeffreys' prior $p(\tau) \propto \tau^{-1}$ for $\tau > 0$, what transformation $\phi = g(\tau)$ will give a uniform Jeffreys' prior $p(\phi)$ for ϕ (ie $p(\phi) \propto 1$)?

$$\begin{aligned}
p(\tau) &= p(\phi)\left|\frac{d\phi}{d\tau}\right| \\
\Leftrightarrow \tau^{-1} &= \left|\frac{d\phi}{d\tau}\right| \\
\Leftrightarrow \frac{d\phi}{d\tau} &= \pm\tau^{-1} \\
\Leftrightarrow \phi &= \pm\int \tau^{-1} d\tau \\
\Leftrightarrow \phi &= \pm\log|\tau| \\
\Leftrightarrow \phi &= \pm\log\tau
\end{aligned}$$

So, Jeffreys' prior for τ is equivalent to a uniform prior on $\log\tau$ (or $-\log\tau$).

Example 3.10: Binomial

Suppose we observe y successes in n independent Bernoulli trials. So, $Y \sim \text{Bin}(n, \theta)$.

$$
\begin{aligned}
p(y \mid \theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\
\log p(y \mid \theta) &= \log\binom{n}{y} + y\log\theta + \\
&\quad (n-y)\log(1-\theta) \\
\frac{d}{d\theta}\log p(y \mid \theta) &= \frac{y}{\theta} - \frac{n-y}{1-\theta} \\
\frac{d^2}{d\theta^2}\log p(y \mid \theta) &= -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2} \\
I(\theta) &= -E\left[-\frac{Y}{\theta^2} - \frac{n-Y}{(1-\theta)^2}\right] \\
&= \frac{E(Y)}{\theta^2} + \frac{n-E(Y)}{(1-\theta)^2} \\
&= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\
&= \frac{n}{\theta(1-\theta)} \\
I(\theta)^{\frac{1}{2}} &\propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}
\end{aligned}
$$

So, Jeffreys' prior for success probability θ of the binomial likelihood is Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$. Note that this is a proper prior.

Limitations of Jeffreys' prior

1. Jeffreys' prior is often improper
2. Violates likelihood principle — Jeffreys' prior is different for equal likelihoods coming from different study designs
3. In the multiparameter case $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, Jeffreys' prior is given by
$$p(\boldsymbol{\theta}) \propto \sqrt{\det\, I(\boldsymbol{\theta})}$$
However, can lead to inconsistencies.

E.g. Suppose $Y \sim \text{Normal}(\theta, \tau^{-1})$, with both θ and τ unknown.

A. Using the above rule gives a joint prior $p(\theta, \tau) \propto \tau^{-1/2}$.

B. Deriving prior $p(\theta, \tau)$ as product of independent Jeffreys' priors for θ and τ gives
$$p(\theta, \tau) = p(\theta)p(\tau) \propto 1 \times \tau^{-1} = \tau^{-1} .$$

Comments: In most cases, judgement about mean would not be affected by anything you were told about variance, or vice versa, so seems reasonable to take a prior which is the product of the priors for the mean and variance separately.

Note

We have three different 'non-informative' priors for θ when $Y \sim \text{Bin}(n, \theta)$:

$$
\begin{aligned}
\theta &\sim \text{Beta}(0,\ 0) \\
\theta &\sim \text{Beta}(0.5,\ 0.5) \\
\theta &\sim \text{Beta}(1,\ 1)
\end{aligned}
$$

When there is much data, it makes very little difference: likelihood dominates the prior. E.g. $y = 50$, $n = 200$

$$
\begin{aligned}
\theta \mid y &\sim \text{Beta}(50,\ 150) \\
\theta \mid y &\sim \text{Beta}(50.5,\ 150.5) \\
\theta \mid y &\sim \text{Beta}(51,\ 151)
\end{aligned}
$$

The problem is when there is little data. E.g. $y = 0$, $n = 10$.

There is no real solution to this.

- Consider using your knowledge to formulate informative prior
- In some cases, hierarchical priors can be useful.

4. Hierarchical priors

A strategy sometimes useful for specifying the prior is to divide the model into stages and construct the prior hierarchically.

Example:
$Y \sim \text{Bin}(10, \theta)$,
$\theta \sim \text{Beta}(\alpha, \beta)$,
$\alpha \sim \text{Gamma}(4, 4)$, $\beta \sim \text{Gamma}(5, 10)$.

Suppose we have a model for the data $p(\mathbf{y} \mid \boldsymbol{\theta})$ and wish to specify a prior $p(\boldsymbol{\theta})$.

If we are unsure what values to specify for the parameters $\boldsymbol{\alpha}$ of this prior $p(\boldsymbol{\theta})$, then we could represent this uncertainty by assigning $\boldsymbol{\alpha}$ a probability distribution, $p(\boldsymbol{\alpha})$. Then,

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha} \\
p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto \int p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}
\end{aligned}
$$

The parameters $\boldsymbol{\alpha}$ are often called *hyperparameters*. The prior distribution for $\boldsymbol{\alpha}$ is often called a *hyperprior*.

In principle, we could introduce yet more levels into the prior (e.g. specifying $p(\boldsymbol{\alpha})$ conditional on further parameters, and so on). However, it is often hard to interpret higher-level parameters.

Hierarchical priors particularly useful when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\theta_1, \ldots, \theta_K$ are exchangeable, and we have data on each θ_k.

More on hierarchical models later.

5. Summary of prior distributions

- Conjugate priors are computationally convenient, but may be restrictive
- Parameters of the prior may be elicited using
  - relevant information from past studies; or
  - matching prior beliefs to appropriate moments and quantiles of a parametric distribution.
- Non-informative priors aim to provide an analysis with minimal subjective input.
  - Useful to provide a 'reference' for comparing with results obtained from using informative priors.
  - But, should be used with care!

- Problems with non-informative priors include:
  - often improper: use *locally uniform proper priors*, usually preferable to flat improper priors
  - not invariant to transformation (e.g. be informative on a different scale): use *Jeffreys' priors*; but they can be problematic (depend on study design; inconsistent in multiparameter setting)
- Hierarchical models using conditionally-specified priors offer an alternative
- Sensitivity analysis to a range of priors is essential in most practical applications

Outline revisited

1. Basic considerations
2. Conjugate priors
3. Non-informative priors
4. Hierarchical priors
5. Summary of prior distributions

Next week: Graphical Models