Prior

Anita Wang

2017 - 2018

Binomial example with different prior distributions

- In the binomial example, we have considered the uniform prior distribution for θ that the prior predictive distribution for y (given n) is uniform on the discrete set {0, 1, . . . , n}.
- We can also parameterize a prior density as θ ~ Beta(α, β):

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

As the likelihood is

$$p(y|\theta) \propto \theta^{a}(1-\theta)^{b}.$$

Comparing p(θ) and p(y|θ) suggests that this prior density is equivalent to α −1 prior successes and β −1 prior failures.

conjugacy

- The parameters of the prior distribution are often referred to as **hyperparameters,** which means we can specify a particular prior distribution by fixing two features of the distribution, for example its mean and variance.
- The posterior density for θ is

$$
\begin{aligned}
p(\theta|y) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\
&= \mathrm{Beta}(\theta|\alpha+y, \beta+n-y).
\end{aligned}
$$

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**; the beta prior distribution is **a conjugate family** for the binomial likelihood.

Conjugate prior

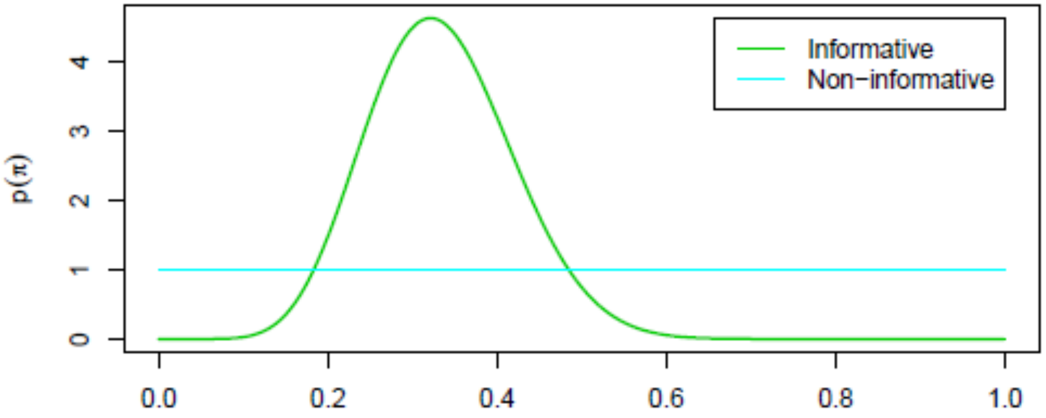
- The conjugate family is mathematically convenient in that the posterior distribution follows a known parametric form.
- If information is available that contradicts the conjugate parametric family it may be necessary to use a more realistic prior distribution.
- Definition: If F is a class of sampling distributions $p(y|\theta)$, and P is a class of prior distributions for θ, then the class P is conjugate for F if

  $p(\theta|y) \in P$ for all $p(\cdot|\theta) \in F$ and $p(\cdot) \in P$.
- Conjugate prior distributions have the practical advantage, in addition to computational convenience, of being interpretable as **additional data**.

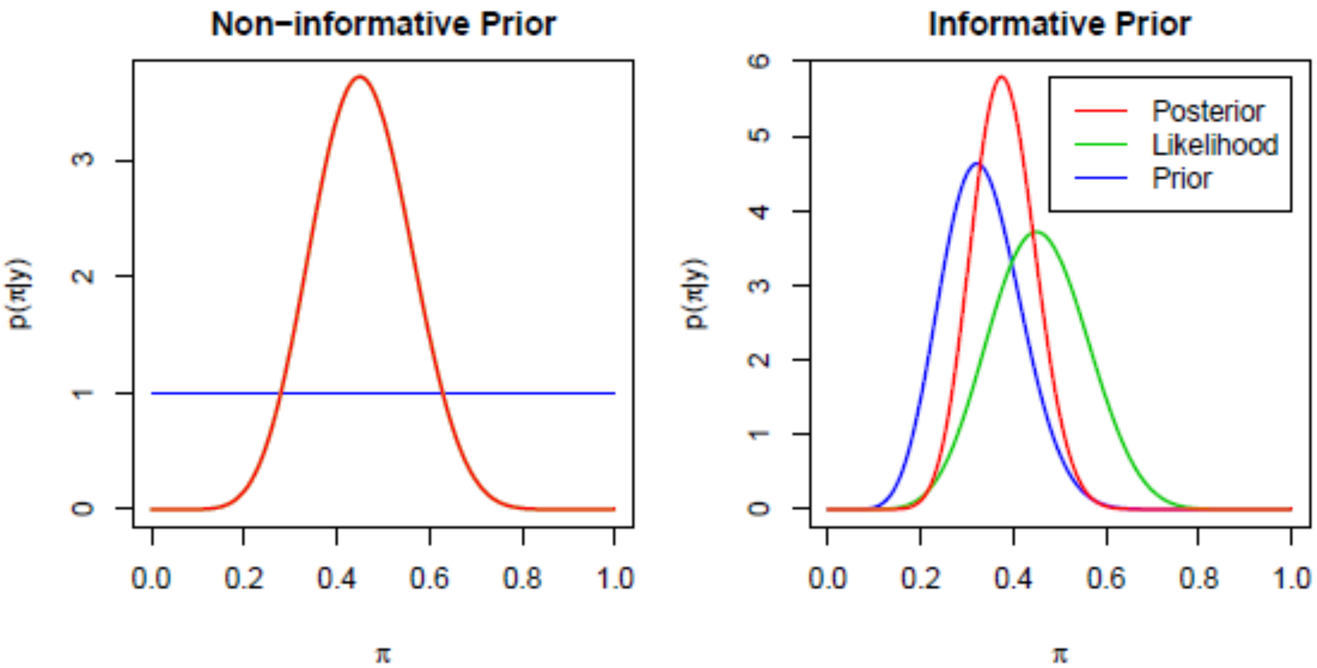
Noninformative prior distributions

- When prior distributions have no population basis, they can be difficult to construct
- There has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution
- Such distributions are sometimes called 'reference prior distributions,' and the prior density is described as vague, flat, diffuse or **noninformative**.
- The rationale for using noninformative prior distributions is often said to be 'to let the data speak for themselves,' so that inferences are unaffected by information external to the current data.

Informative vs Non-informative

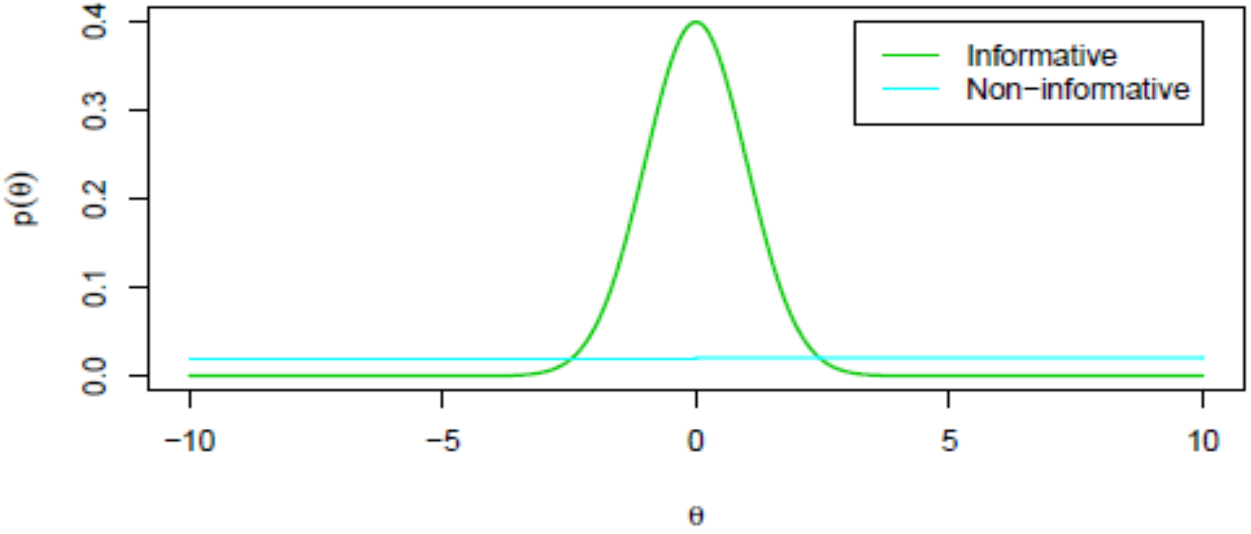


For this example, with the non-informative prior, Posterior=Likelihood



Improper prior distributions

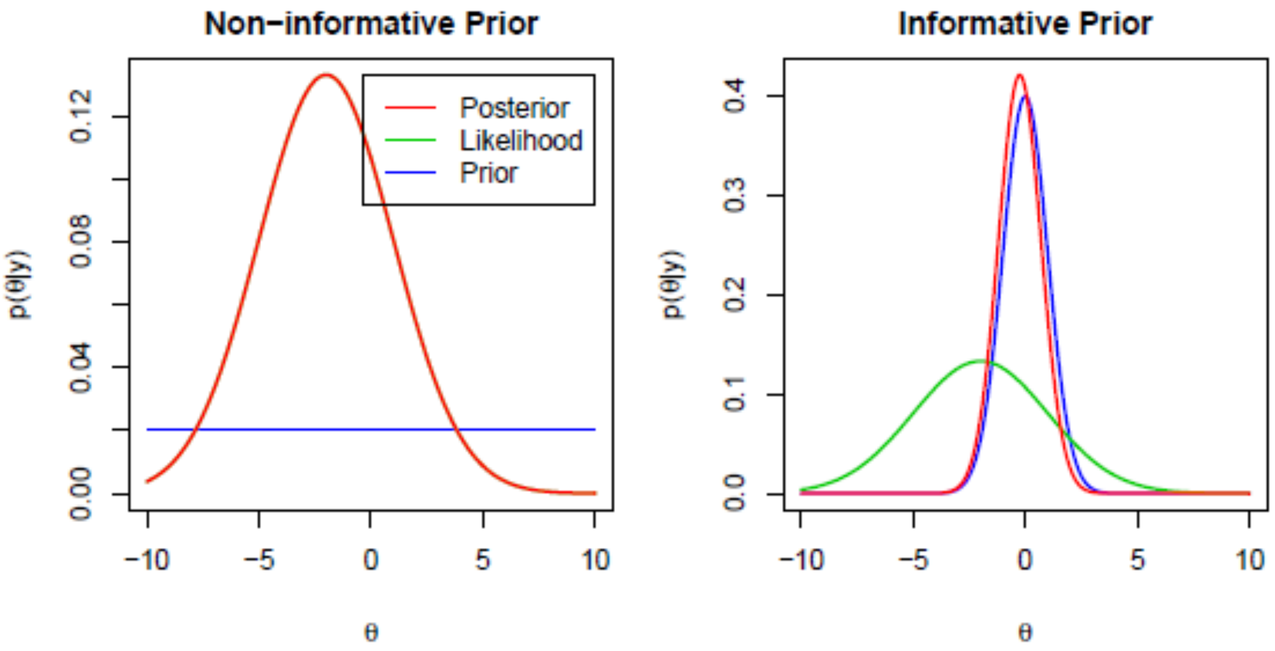
- However for a parameter that occurs on an infinite interval (e.g. a normal mean μ), using a uniform prior on μ is problematic.
- For the normal mean example, lets use the conjugate prior $N(\mu_0, \tau_0^2)$, but with a very big variance τ_0^2



Non-informative Prior

- The posterior mean and precision are

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$



Non-informative Prior

- So if we let $\tau_0^2 \to \infty$, then

$$\mu_n \to \bar{y} \text{ and } \frac{1}{\tau_n^2} \to \frac{n}{\sigma^2}$$

- This equivalent to the posterior being proportional to the likelihood, which is what we get if $p(\theta) \propto 1$(e.g. uniform).

- This does not describe a valid probability density as

$$\int_{-\infty}^{\infty} d\theta = \infty$$

Proper

- A prior is called proper if it is a valid probability distribution

$$p(\theta) \geq 0, \forall\theta \in \Theta \text{ and } \int p(\theta)\, d\theta = 1$$

(If p(θ) integrates to any positive finite value, it is called an **unnormalized density** and can be renormalized—multiplied by a constant—to integrate to 1.

If a prior is proper, so must the posterior.

Improper

- A prior is called improper if

$$p(\theta) \geq 0, \forall\theta \in \Theta \text{ and } \int p(\theta)\,d\theta = \infty$$

- The prior distribution is improper in this example, but the posterior distribution is proper i.e., $\int p(\theta|y)d\theta$ is finite for all y, given at least one data point.
- Posterior distributions obtained from improper prior distributions must be interpreted with great care—one must always check that the posterior distribution has a finite integral and a sensible form.

Another example of noninformative prior

- The normal model with known mean but unknown variance

- The prior degrees of freedom, $\nu_0 \rightarrow 0$

$$p(\sigma^2|y) \approx \text{Inv-}\chi^2(\sigma^2|n, \nu).$$

- This can also be derived by defining the prior density for σ^2 as $p(\sigma^2) \propto 1/\sigma^2$, which is improper as

$$\int_0^\infty 1/\sigma^2 d\sigma = -\infty$$

Jeffreys' Priors

- One approach that is sometimes used to define noninformative prior distributions was introduced by Jeffreys, based on considering one-to-one transformations of the parameter: φ = h(θ).

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1} \text{ where } \theta = h^{-1}(\phi)$$

- Jeffreys' general principle is that any rule for determining the prior density p(θ) should yield an equivalent result if applied to the transformed parameter
- that is, p(φ) computed by determining p(θ) should match the distribution that is obtained by determining p(φ) directly using the transformed model, p(y, φ) = p(φ)p(y|φ)

Jeffreys' Priors

- Jeffreys’ principle leads to defining the noninformative prior density

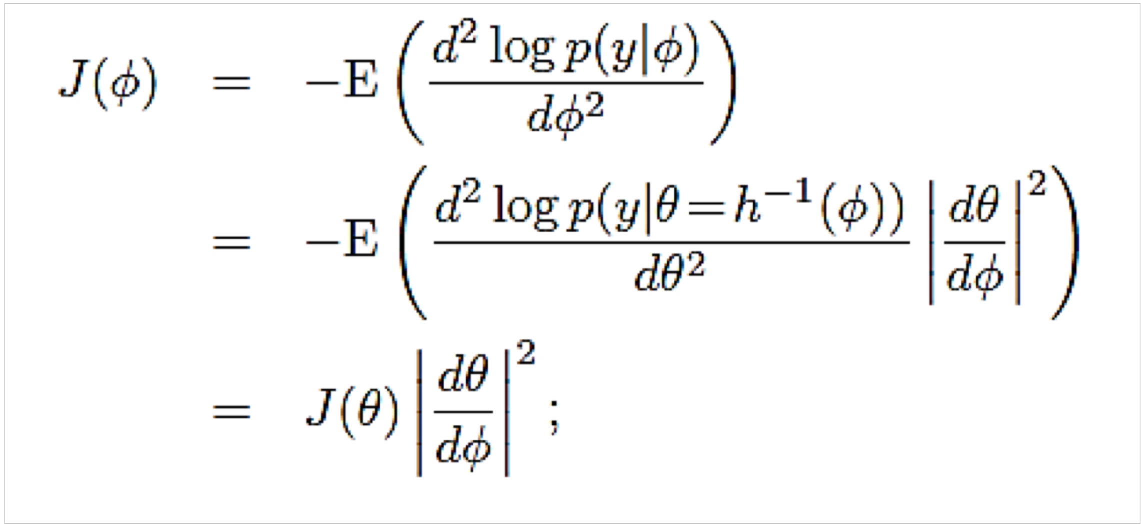
$$p(\theta) = [J(\theta)]^{1/2}$$

where $J(\theta)$ is the *Fisher information* for θ

$$J(\theta) = E\left[\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\frac{d^2\log p(y|\theta)}{d\theta^2} |\theta\right]$$

Jeffreys' Priors

To see that Jeffreys' prior model is invariant to parameterization



$$
\begin{aligned}
J(\phi) &= -\mathrm{E}\left(\frac{d^2 \log p(y|\phi)}{d\phi^2}\right) \\
&= -\mathrm{E}\left(\frac{d^2 \log p(y|\theta = h^{-1}(\phi))}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) \\
&= J(\theta)\left|\frac{d\theta}{d\phi}\right|^2;
\end{aligned}
$$

thus,

$$J(\phi)^{1/2} = J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$$

Example: binomial distribution

- log-likelihood of y ~ Bin(n, θ)

$$\log p(y|\theta) = \text{constant} + y \log \theta + (n - y) \log(1 - \theta).$$

$$J(\theta) = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} |\theta\right] = \frac{n}{\theta(1-\theta)}$$

- Jeffreys' prior density is then $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ is a Beta(1/2, 1/2) density.
- the uniform prior density, which can be expressed as θ ~ Beta(1, 1)
- p(logit(θ)) ∝ constant corresponds to the improper θ ~ Beta(0, 0).

Exercise: Normal distribution

- For the normal example with unknown variance, prove the Jeffreys prior for the standard deviation σ is

$$p(\sigma) \propto \frac{1}{\sigma}$$

- Alternative descriptions under different parameterizations for the variability are

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$p(\log \sigma^2) \propto p(\log \sigma) \propto 1$$

Pivotal Quantities: location parameter

- If the density of y is such that $p(y-\theta|\theta)$ is a function that is free of θ and y, say $f(u)$, where $u = y - \theta$, then $y - \theta$ is a **pivotal quantity**, and θ is called a pure **location parameter**.

- A noninformative prior distribution should lead to posterior distribution $p(y-\theta|y)$ still be free of θ and y, $f(y-\theta)$. $y - \theta$ should still be a pivotal quantity.

- $p(y-\theta|y) \propto p(\theta)p(y-\theta|\theta)$ implying $p(\theta) \propto$ constant, a uniform distribution on θ

Pivotal Quantities: scale parameter

- If the density of y is such that p(y/θ |θ) is a function that is free of θ and y—say, g(u), where u = y/θ—then u = y/θ is a **pivotal quantity** and θ is called a pure **scale parameter**
- A noninformative prior distribution should lead to posterior distribution p(y/θ|y) still be free of θ and y, g(y/θ).
- By transformation of variables

$$p(y|\theta) = \frac{1}{\theta} p(u|\theta)$$

and similarly,

$$p(\theta|y) = \frac{y}{\theta^2} p(u|y)$$

Pivotal Quantities

- Letting both $p(u|\theta)$ and $p(u|y)$ equal $g(u)$,

$$p(\theta|y) = \frac{y}{\theta} p(y|\theta)$$

which implies $p(\theta) \propto \frac{1}{\theta}$

- Equivalently, $p(\log\theta) \propto 1$ or $p(\theta^2) \propto \frac{1}{\theta^2}$.

- The standard deviation from a normal distribution and the mean of an exponential distribution are scale parameters.

Thanks