§3 Prior Distributions

Outline

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

1. Basic considerations

The only requirement for the prior distribution is that it should represent the knowledge about θ *before* observing the current data.

Therefore, the prior distribution can

- be specified entirely subjectively
- depend on past data
- be weak or non-informative

Choosing a prior involves

1. Choosing the functional form of the distribution
2. Specifying values for the parameters of that distribution

The functional form chosen for $p(\theta)$ must take into account the support of θ.

- If the support of θ is $(-\infty, \infty)$, e.g. θ is the mean of a normally distributed rv, or a regression coefficient, then suitable priors $p(\theta)$ might include normal or Student-t prior distributions
- If support of θ is $(0, \infty)$, e.g. θ is a precision parameter or mean of a Poisson rv, then suitable priors $p(\theta)$ might include gamma or log-normal distributions
- If support of θ is $(0, 1)$, e.g. θ is a proportion or the success probability of a binomial rv, then suitable priors $p(\theta)$ might include beta distributions

More complex functional forms can be specified by taking *mixtures* of standard distributions, but we shall not consider mixture priors here.

2. Conjugate priors

A convenient way to choose the functional form of the prior is by use of conjugate distributions.

Definition

Let $l(\theta) = p(\mathbf{x} \mid \theta)$ be a likelihood function. A class $\mathcal{P}$ of prior distributions $p(\theta)$ is said to form a conjugate family (*for this likelihood function*) if the posterior distribution $p(\theta \mid \mathbf{x})$ is also in the class $\mathcal{P}$ for all data $\mathbf{x}$.

That is: **the prior** $p(\theta)$ **and the posterior** $p(\theta \mid \mathbf{x})$ **belong to the same class** $\mathcal{P}$.

Some difficulties with this definition:

- If $\mathcal{P}$ = all distributions, then $\mathcal{P}$ is always conjugate whatever the likelihood function is
- If $\mathcal{P}$ consists only of *point mass* priors

$$p(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

  then $\mathcal{P}$ is always conjugate whatever the likelihood function is

In practice, we are also interested in *natural conjugate priors*: A natural conjugate prior is (i) a conjugate prior, ie the prior and the posterior belong to the same class $\mathcal{P}$, and (ii) the likelihood has the same functional form of θ as the distributions in $\mathcal{P}$.

Example 3.1: Binomial likelihood

The likelihood is

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

The beta prior Beta(α, β) for θ is

$$\begin{aligned} p(\theta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

So the posterior is

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

- Is this beta prior a conjugate prior of θ for the binomial likelihood?
- Is it also a natural conjugate prior of θ?
  - The natural conjugate prior must have the same functional form of θ as the likelihood
  - Here, the likelihood is of the form of θ:

$$\theta^a (1-\theta)^b$$

Example 3.2: Normal, known precision

The likelihood for $Y_i \mid \theta \sim \text{Normal}(\theta, \tau^{-1})$ is

$$p(\mathbf{y} \mid \theta) \propto \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \theta)^2\right]$$

The normal prior Normal(μ_0, ϕ_0^{-1}) for θ is

$$p(\theta) \propto \exp\left[-\frac{\phi_0}{2}(\theta - \mu_0)^2\right]$$

So the posterior is Normal(μ_1, ϕ_1^{-1}):

$$p(\theta \mid \mathbf{y}) \propto \exp\left[-\frac{\phi_1}{2}(\theta - \mu_1)^2\right]$$

- Is this normal prior a conjugate prior of θ for the normal likelihood?
- Is it also a natural conjugate prior of θ?

Why is conjugacy useful? Because it simplifies analysis.

- Ensures posterior follows a known parametric form of θ.
- Every new observation leads only to a change in the values of the parameters of the distribution for θ, as indicated by the sequential learning in §1; no new algebra needed.
- An objective meaning can be attached to the parameters of the prior distribution, e.g.
  - the Beta(α, β) prior mimics a binomial likelihood with $y_0 = \alpha - 1$ successes in $n_0 = \alpha + \beta - 2$ trials;
  - therefore, we can think of Beta(α, β) as representing information equivalent to having observed $\alpha - 1$ successes in $\alpha + \beta - 2$ trials of a hypothetical prior experiment.

Exponential family likelihoods

Many of the common likelihoods we come across belong to the exponential family.

A density is from the one-parameter exponential family if it has the form

$$p(y \mid \theta) = f(y)g(\theta)\exp[h(\theta)t(y)] \ ,$$

for some functions $f(y)$ and $t(y)$ of data y only and some functions $g(\theta)$ and $h(\theta)$ of parameter θ only.

Then the likelihood of n independent observations $\mathbf{y} = (y_1, \ldots y_n)$ is

$$p(\mathbf{y} \mid \theta) = \prod p(y_i \mid \theta) \propto g(\theta)^n \exp\left[h(\theta)\sum t(y_i)\right] \ ,$$

and we say that the likelihood function comes from the one-parameter exponential family.

The conjugate family $\mathcal{P}$ for a likelihood belonging to the exponential family is the class of distributions of the form

$$p(\theta) \propto g(\theta)^\nu \exp[h(\theta)\delta]$$

and the posterior distribution is then

$$p(\theta \mid \mathbf{y}) \propto g(\theta)^{n+\nu} \exp\left[h(\theta)(\sum t(y_i) + \delta)\right]$$

Note: θ in the likelihood and posterior associates with the data $\mathbf{y}$ only through the statistic $\sum t(y_i)$. We say that $\sum t(y_i)$ is a *sufficient statistic*.

How to interpret prior parameters δ and ν?

Notice that

$$g(\theta)^\nu \exp[h(\theta)\delta]$$

can be viewed as the likelihood of ν independent observations $\mathbf{x} = (x_1, \ldots, x_\nu)$ with $\sum t(x_i) = \delta$.

So, we can think of

$$p(\theta) \propto g(\theta)^\nu \exp[h(\theta)\delta]$$

as corresponding to the following prior information:
We have observed a hypothetical 'prior' sample of ν observations, $\mathbf{x} = (x_1, \ldots, x_\nu)$, with sufficient statistic δ.

Example 3.3: Binomial family

Suppose we have a single ($n = 1$) binomial observation $Y = y$: $Y \sim \text{Bin}(m, \theta)$ (ie containing y successes out of m Bernoulli trials).

$$\begin{aligned} p(y \mid \theta) &= \binom{m}{y}\theta^y(1-\theta)^{m-y} \\ &= \binom{m}{y}(1-\theta)^m \exp\left[y\log\left(\frac{\theta}{1-\theta}\right)\right] \\ &\propto g(\theta)\exp[h(\theta)t(y)] \end{aligned}$$

So, this belongs to the exponential family:

$$g(\theta) = (1-\theta)^m; \ \ h(\theta) = \log\left(\tfrac{\theta}{1-\theta}\right); \ \ t(y) = y.$$

Thus, the conjugate prior is of the form

$$\begin{aligned} p(\theta) &\propto g(\theta)^\nu \exp[h(\theta)\delta] \\ &= (1-\theta)^{m\nu}\exp\left[\left\{\log\left(\frac{\theta}{1-\theta}\right)\right\}\delta\right] \\ &= (1-\theta)^{m\nu}\theta^\delta(1-\theta)^{-\delta} \\ &= \theta^\delta(1-\theta)^{m\nu-\delta} \\ \theta &\sim \text{Beta}(\delta+1, m\nu-\delta+1) \end{aligned}$$

This prior represents a hypothetical 'prior' sample of ν independent observations, $x_1, \ldots, x_\nu$, from the $\text{Bin}(m, \theta)$ distribution, with total number of successes $\sum x_i = \delta$.

Example 3.4: Normal, known precision

$$\begin{aligned} p(\mathbf{y} \mid \theta) &= \left(\frac{\tau}{2\pi}\right)^{n/2}\exp\left[-\frac{\tau}{2}\sum_{i=1}^n(y_i-\theta)^2\right] \\ &\propto \exp\left[-\frac{\tau n\theta^2}{2}\right]\exp\left[\tau\theta\sum_i y_i\right] \end{aligned}$$

So, this belongs to the exponential family:

$$g(\theta) = \exp\left[-\tfrac{\tau\theta^2}{2}\right]; \ \ h(\theta) = \tau\theta; \ \ t(y_i) = y_i.$$

Thus, the conjugate prior is of the form

$$\begin{aligned} p(\theta) &\propto g(\theta)^\nu \exp[h(\theta)\delta] \\ &= \left\{\exp\left[-\frac{\tau\theta^2}{2}\right]\right\}^\nu \exp[\tau\theta\delta] \\ &= \exp\left[-\frac{\tau\nu}{2}\left(\theta^2 - \frac{2\theta\delta}{\nu}\right)\right] \\ &\propto \exp\left[-\frac{\tau\nu}{2}\left(\theta - \frac{\delta}{\nu}\right)^2\right] \\ \Rightarrow \theta &\sim \text{Normal}\left(\mu_0 = \frac{\delta}{\nu}, \ \phi_0^{-1} = (\nu\tau)^{-1}\right) \end{aligned}$$

We see that ν represents prior sample size; δ represents sum of y in prior sample. (So, δ/ν represents the prior sample mean μ_0; also see '§2 Bayesian Inference' p5 where $\kappa_0 = \nu$ is the prior sample size.)

So, in general, the parameters of conjugate priors for exponential family likelihoods have a natural interpretation as *observing a 'prior' sample of size ν with the sufficient statistic of this 'prior' sample being equal to δ.*

This can be used as an aid to eliciting prior parameters

- by imagining a hypothetical experiment that corresponds to your prior beliefs, or

- by 'converting' previous data into a suitable prior distribution.

Alternatives to eliciting prior parameters and to conjugate priors

- Specify particular features of your prior beliefs and find parametric distribution that approximately matches these (and has the right support), e.g.
  - mean of θ
  - variance of θ
  - mode of θ (most likely value)
  - median of θ (central value)
  - central 95% interval of θ

    E.g. If we think a normal prior for θ is reasonable and a plausible range for θ is $[3.5, 4.4]$, then we might set $\theta \sim \text{Normal}(\mu, \sigma^2)$ and choose μ and σ such that $\mu - 1.96\sigma = 3.5$ and $\mu + 1.96\sigma = 4.4$. This way, $P(\theta \in [3.5, 4.4]) = 0.95$.
- Allow prior distribution itself to depend on unknown parameters (*hyperparameters*) and assign these *hyperprior* distributions. This leads to a *hierarchical model*.

- Choose a non-informative prior, but why?

3. Non-informative priors

Two statisticians may use different priors reflecting their different subjective beliefs, then produce different posteriors.

Idea of non-informative priors is that:

- If the inference is based on a minimum of subjective prior belief, more likely that statisticians (and everyone else) can agree, or
- at the least, posterior from a non-informative prior provides a reference, against which posteriors using subjective, informative priors can be compared (part of sensitivity analysis).

Non-informative priors are also known as *vague, flat, diffuse* or *reference priors*.

Uniform priors

If $\theta \sim \text{Uniform}$, then $p(\theta) \propto 1$: 1) no value of θ is more probable than any other value; 2) $p(\theta \mid y) \propto p(y \mid \theta)$.

Thus, the likelihood *dominates* the prior, ie posterior depends on the data (the likelihood) as much as possible.

- If support of θ is $(0, 1)$, then uniform prior is $\theta \sim \text{Uniform}(0, 1)$:
$$p(\theta) = \begin{cases} 1 & \text{for } 0 \le \theta \le 1 \\ 0 & \text{otherwise;} \end{cases}$$
$p(\theta)$ is proper: $\int_0^1 p(\theta) d\theta = 1$.

- If support of θ is $\mathbb{R}$, then uniform prior is $\theta \sim \text{Uniform}(-\infty, \infty)$:
$$p(\theta) \propto 1 \quad \text{for } -\infty < \theta < \infty \;;$$
$p(\theta)$ is **improper**: $\int_{-\infty}^{\infty} p(\theta) d\theta = \infty$.

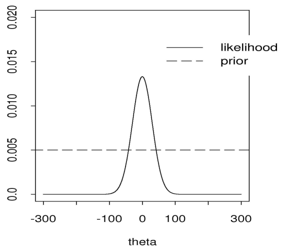
Improper priors *may* give improper posteriors; however, sometimes an improper prior *may* still lead to a *proper* posterior (examples soon). Therefore, check posteriors derived from improper priors.

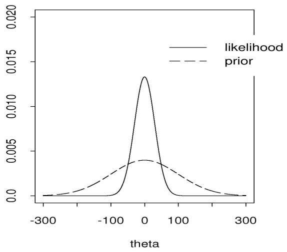
An alternative to using improper uniform priors is to use *locally uniform* proper priors:

1. when the likelihood $p(y|\theta)$ is non-negligible at some values of θ, let $p(\theta)$ not change much over these values of θ;
2. when the likelihood $p(y|\theta)$ is negligible at some values of θ, let $p(\theta)$ not assume large values at these values of θ.

As such,

- the prior $p(\theta)$ will be dominated by the likelihood $p(y \mid \theta)$, and thus
- no risk of improper posterior $p(\theta \mid y)$.





Example 3.5: Normal, known precision

As seen earlier, if $Y \mid \theta \sim$ Normal(θ, τ^{-1}) $(i = 1, \ldots, n)$ and $\theta \sim$ Normal(μ_0, ϕ_0^{-1}), then

$$\theta \mid y \sim \text{Normal}(\mu_1, \phi_1^{-1})$$

where

$$\begin{aligned} \mu_1 &= \frac{\mu_0\phi_0 + n\bar{y}\tau}{\phi_0 + n\tau} \\ \phi_1 &= \phi_0 + n\tau \end{aligned}$$

(see '§2 Bayesian Inference' p4)

- For non-informative prior, we could take $\phi_0 = 0$. But this prior is improper although the posterior is proper.
- If we choose ϕ_0 small but > 0, then the prior is locally uniform and proper; in this case, the proper posterior for $\phi_0 = 0$ is the limit of posteriors as $\phi_0 \to 0$.
- We can often think of improper prior as mathematical device: its posterior is the limit of posteriors from a sequence of proper priors.

Example 3.6: Bayes' postulate

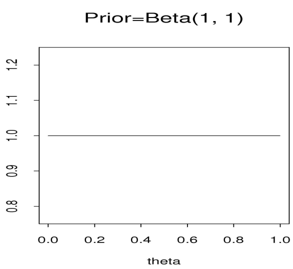
Let $Y \mid \theta \sim$ Bin(n, θ).

Uniform prior $p(\theta)$ for θ is Beta$(1,1) \propto 1$, ie Beta$(\alpha = 1, \beta = 1) \equiv$ Uniform$(0,1)$. The prior is proper.

Then, as seen earlier, posterior $p(\theta \mid y)$ is Beta$(y+\alpha, n-y+\beta)$ =Beta$(y+1, n-y+1)$.

A 'natural' estimate for θ is $\frac{y}{n}$. And we know that the mode of Beta(α,β) is $\frac{\alpha-1}{\alpha+\beta-2}$, for $\alpha, \beta > 1$. So, here, the mode of $p(\theta \mid y)$ is $\frac{y}{n}$.

However, the mean of $p(\theta \mid y)$ here is $\frac{y+1}{n+2}$ as the mean of Beta(α,β) is $\frac{\alpha}{\alpha+\beta}$.



Example 3.7: Haldane's prior

Let $Y \mid \theta \sim$ Bin(n, θ).

Haldane's prior for θ is Beta$(0,0)$, given by

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

Then, the posterior $p(\theta \mid y)$ is Beta$(y+\alpha, n-y+\beta)$ =Beta$(y, n-y)$

Therefore, when $\alpha = \beta = 0$ for Beta(α,β) prior, we have $E[\theta \mid y] = \frac{y}{n}$, the 'natural' estimate for θ.

Furthermore, Beta(α,β) prior becomes more and more informative as α and β increase. Therefore, it could be argued that taking $\alpha = \beta = 0$ corresponds to minimum possible prior information.

However, Beta$(0,0)$ is an improper prior.