§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 distributions in $\mathcal{P}$ have the same functional form of θ as the likelihood.

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:

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

Why is conjugacy useful? Because it simplifies analysis.

- Ensures posterior follows a known parametric form.
- 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) distribution mimics a binomial likelihood with $y_0 = \alpha - 1$ successes in $n_0 = \alpha + \beta - 2$ trials;
  - therefore, we can think of a 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

$$p(y \mid \theta) = f(y)g(\theta)\,\mathsf{exp}\left[h(\theta)t(y)\right] \;\; ,$$

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

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

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

$$p(\theta) \propto g(\theta)^\nu\,\mathsf{exp}\left[h(\theta)\delta\right]$$

and the posterior distribution is then

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

Example 3.2: Binomial family

Suppose we have a single observation $Y = y$, $Y \sim \text{Bin}(m, \theta)$ (so, $n = 1$).

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

So, this belongs to the exponential family:

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

Thus, the conjugate prior is of the form

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

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 $\sum x_i = \delta$.

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.

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 $\theta \sim$ 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 $\theta \sim$ Uniform$(0, 1)$:

  $$p(\theta) = \begin{cases} 1 & \text{for } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise;} \end{cases}$$

  $p(\theta)$ is proper: $\int p(\theta) d\theta = 1$.

- If support of θ is $\mathbb{R}$, then uniform prior is $\theta \sim$ Uniform$(-\infty, \infty)$:

  $$p(\theta) \propto 1 \quad \text{for } -\infty < \theta < \infty \ ;$$

  $p(\theta)$ is **improper**: $\int p(\theta) d\theta = \infty$.

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.

Example 3.3: Bayes' postulate

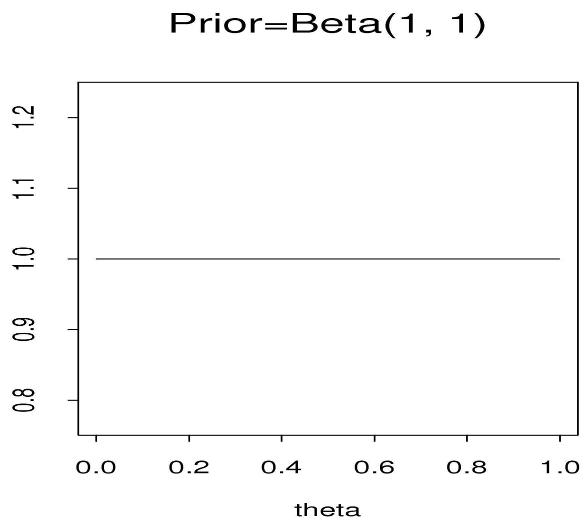
Let $Y \mid \theta \sim \text{Bin}(n, \theta)$.

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 3.4: Haldane's prior

Let $Y \mid \theta \sim \text{Bin}(n, \theta)$.

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.

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.

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 3.5: 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 Binomial likelihood is Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$. Note that this is a proper prior. (However, Jeffreys' prior is **often improper**!)

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:
$Y \sim$ Bin$(10, \theta)$,
$\theta \sim$ Beta(α, β),
$\alpha \sim$ Gamma$(4, 4)$, $\beta \sim$ 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.

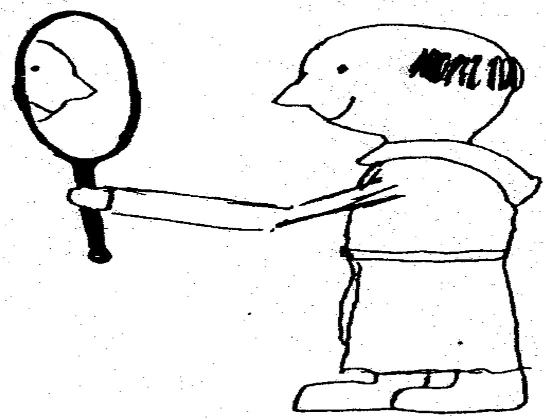
Hierarchical priors particularly useful when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\theta_1, \ldots, \theta_K$ are exchangeable, and we have data on each θ_k.

More on hierarchical models later.

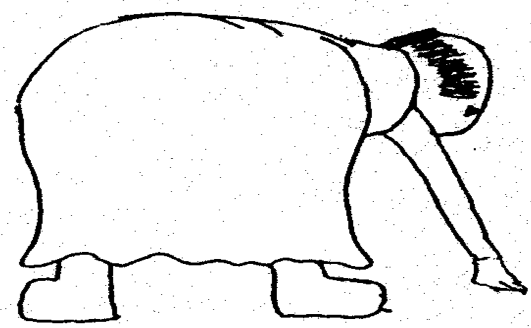
5. Summary of prior distributions

- Conjugate priors are computationally convenient, but may be restrictive. Parameters of the prior may be elicited using relevant information from past studies.
- 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, because they are often improper, or not invariant to transformation (see *Jeffreys’ priors*).
- Hierarchical models using conditionally-specified priors offer an alternative
- Sensitivity analysis to a range of priors is essential in most practical applications

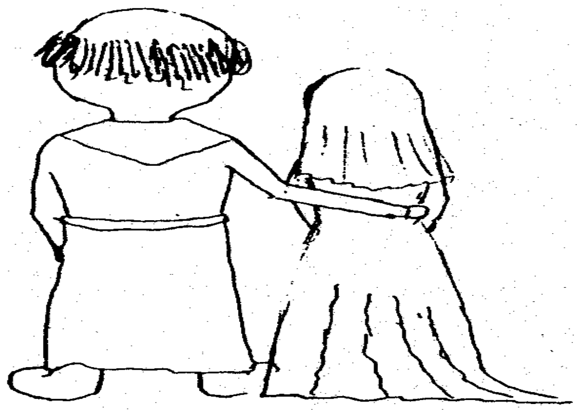
By Professor D.M. Titterington



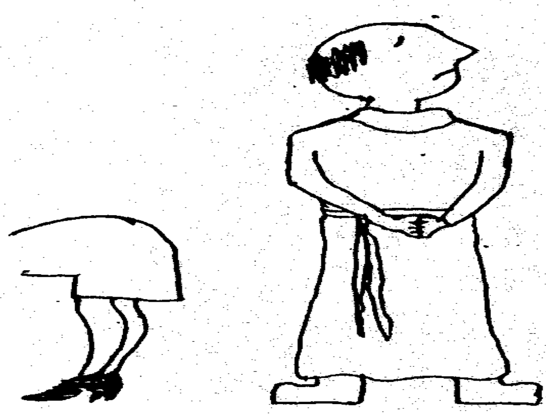
Subjective Prior



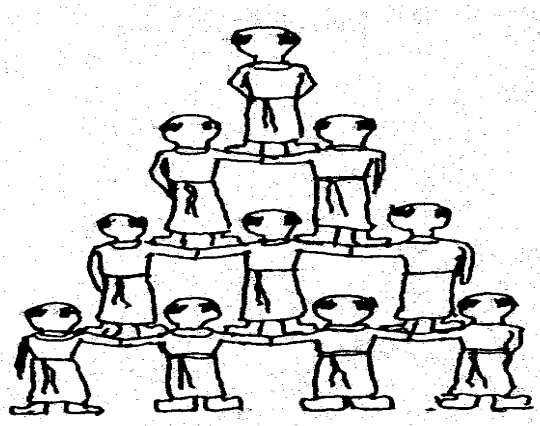
Posterior



Conjugal Prior



Proper Prior
(Discreet Prior)



Hierarchical Priors

Outline revisited

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

Next week: Hierarchical Models & Graphical Models