Bayesian Inference

Ricardo Ehlers
ehlers@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Prior Distributions

Conjugate Priors

In some standard models, the posterior and predictive distributions can be found in closed form.

Definition

If $F = \{p(x|\theta), \theta \in \Theta\}$ is a family of sampling distributions then a class P of distributions is a conjugate family with respect to F if

$$\forall\ p(x|\theta) \in F \quad \text{and} \quad p(\theta) \in P \Rightarrow p(\theta|x) \in P.$$

So, prior and posterior distributions belong to the same class.

In practice, the following steps determine the class of conjugate priors.

1. Identify the class P of distributions for θ such that $p(x|\theta)$ is proportional to a member of this class.
2. Verify whether P is closed under multiplication, i.e. if $\forall$ $p_1, p_2 \in P$ $\exists$ k such that $kp_1p_2 \in P$.

If also there exists a constant k such that $k^{-1} = \int p(x|\theta)d\theta < \infty$ and all $p \in P$ is defined as $p(\theta) = k\ p(x|\theta)$ then P is the natural conjugate family with respect to this sampling model.

Example. Let $X_1, \ldots, X_n \sim$ Bernoulli(θ). The joint sampling density is,

$$p(\mathbf{x}|\theta) = \theta^t (1-\theta)^{n-t}, \quad 0 < \theta < 1 \quad \text{where} \quad t = \sum_{i=1}^{n} x_i$$

and by Bayes theorem it follows that,

$$p(\theta|\mathbf{x}) \propto \theta^t (1-\theta)^{n-t} p(\theta).$$

Note that $p(\mathbf{x}|\theta)$ is proportional to the density of a Beta($t+1, n-t+1$) distribution.

Also, if p_1 and p_2 are the densities of Beta(a_1, b_1) and Beta(a_2, b_2) then

$$p_1 p_2 \propto \theta^{a_1+a_2-2}(1-\theta)^{b_1+b_2-2},$$

which is proportional to the density of a Beta($a_1 + a_2 - 1, b_1 + b_2 - 1$) distribution.

- We conclude that the family of Beta distributions with integer parameters is the natural conjugate to the Bernoulli family.
- In practice, this class can be extended to include all Beta distributions, i.e. for all positive parameters.

Binomial Model

Let $X|\theta \sim \text{Binomial}(n, \theta)$. Then,

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

The natural conjugate family is the Beta(α, β) distribution,

$$p(\theta) \quad = \quad \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \alpha > 0, \beta > 0$$

- The Beta function is defined as,

$$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

  $y \in (0,1), a > 0, b > 0.$

- The Gamma function is defined as,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx.$$

- Properties,
  - Integrating by parts,

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \alpha > 0.$$

  - $\Gamma(1) = 1.$
  - $\Gamma(1/2) = \sqrt{\pi}.$
  - For n positive integer,

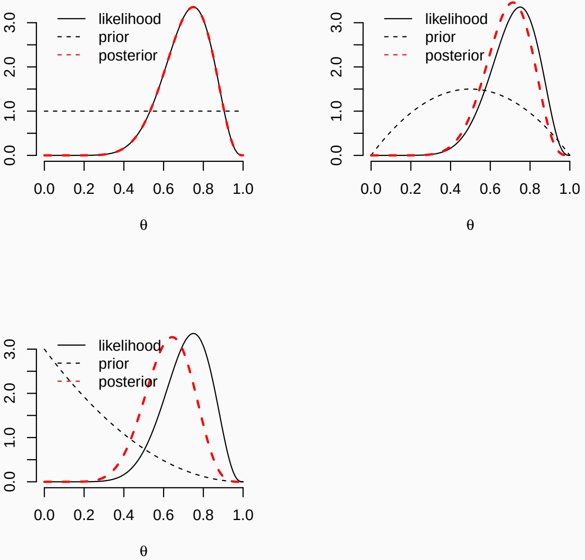
$$\Gamma(n+1) = n!$$
$$\Gamma\left(n+\frac{1}{2}\right) = \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\dots\frac{3}{2}\,\frac{1}{2}\,\sqrt{\pi}$$

The posterior distribution is also Beta with parameters $\alpha + x$ and $\beta + n - x$,

$$p(\theta|x) \quad \propto \quad \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

Beta(1,1), Beta(2,2) and Beta(1,3) priors, posterior and normalized likelihood for $n = 12$ and $X = 9$.



The predictive distribution is given by,

$$p(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} d\theta,$$

$x = 0, 1, \ldots, n$.

Then, solving the integral we have,

$$\begin{aligned} p(x) &= \binom{n}{x} B^{-1}(\alpha,\beta) B(\alpha+x, \beta+n-x) \\ &= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)}, \quad x = 0, 1, \ldots, n. \end{aligned}$$

This is called a Beta-Binomial distribution.

Predictive probabilities $P(X = k)$ for $n = 12$ associated with Beta(1,1), Beta(2,2) and Beta(1,3) conjugate priors.

k	Beta(1,1)	Beta(2,2)	Beta(1,3)
0	0.0769	0.0286	0.2000
1	0.0769	0.0527	0.1714
2	0.0769	0.0725	0.1451
3	0.0769	0.0879	0.1209
4	0.0769	0.0989	0.0989
5	0.0769	0.1055	0.0791
6	0.0769	0.1077	0.0615
7	0.0769	0.1055	0.0462
8	0.0769	0.0989	0.0330
9	0.0769	0.0879	0.0220
10	0.0769	0.0725	0.0132
11	0.0769	0.0527	0.0066
12	0.0769	0.0286	0.0022

Normal Model with Known Variance

For a random sample $X_1, \ldots, X_n$ from a $N(\theta, \sigma^2)$ with σ^2 known, the likelihood function is,

$$
\begin{aligned}
p(\mathbf{x}|\theta) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right\}
\end{aligned}
$$

This has the same form as the likelihood based on a single observation replacing x by $\overline{x}$ and σ^2 by σ^2/n.

Therefore, the previous results hold with appropriate substitutions. The posterior distribution of θ given $\mathbf{x}$ is $N(\mu_1, \tau_1^2)$ where,

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{x}}{\tau_0^{-2} + n\sigma^{-2}} \quad \text{and} \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}.$$

The posterior mean can be rewritten as,

$$\mu_1 = w\mu_0 + (1 - w)\overline{x}$$

where,

$$w = \frac{\tau_0^{-2}}{\tau_0^{-2} + n\sigma^{-2}}.$$

Poisson Model

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter θ. The joint probability function is given by,

$$p(\mathbf{x}|\theta) = \frac{e^{-n\theta}\theta^t}{\prod x_i!} \propto e^{-n\theta}\theta^t, \quad \theta > 0, \quad t = \sum_{i=1}^{n} x_i.$$

The likelihood kernel is of the form $\theta^a e^{-b\theta}$ which characterizes the Gamma family of distributions.

This family is closed under multiplication (check this!).

The natural conjugate prior for θ is Gamma with positive parameters α and β, i.e.

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}, \quad \alpha > 0,\ \beta > 0,\ \theta > 0.$$

The posterior density is then given by,

$$p(\theta|x) \propto \theta^{\alpha+t-1} \exp\left\{-(\beta+n)\theta\right\},$$

which corresponds (up to a constant) to the density of a Gamma$(\alpha+t, \beta+n)$ distribution, i.e.

$$\theta|x \sim \text{Gamma}(\alpha+t, \beta+n).$$

The posterior mean can be rewritten as,

$$\begin{aligned} E(\theta|\mathbf{x}) = \frac{\alpha+t}{\beta+n} &= \left(\frac{\alpha}{\beta}\right)\frac{\beta}{\beta+n} + \left(\frac{t}{n}\right)\frac{n}{\beta+n}, \\ &= E(\theta)\frac{\beta}{\beta+n} + \overline{x}\frac{n}{\beta+n}. \end{aligned}$$

which is a compromise between prior and sample means.

Note that,

- When $n \to \infty$,

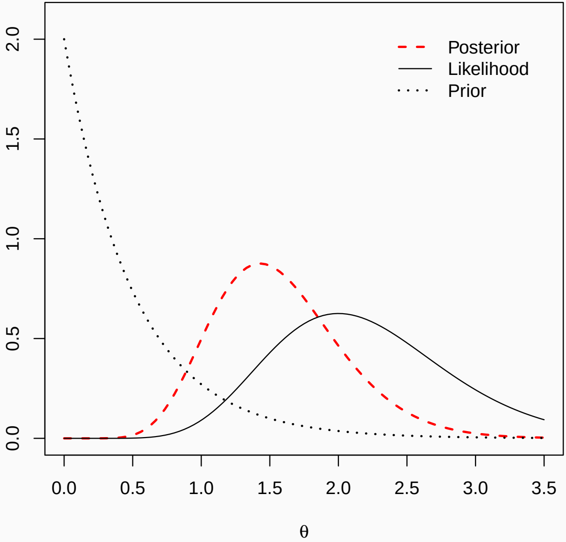
$$E(\theta|\mathbf{x}) \to \overline{x}$$

- When $\alpha \to 0$ and $\beta \to 0$ also,

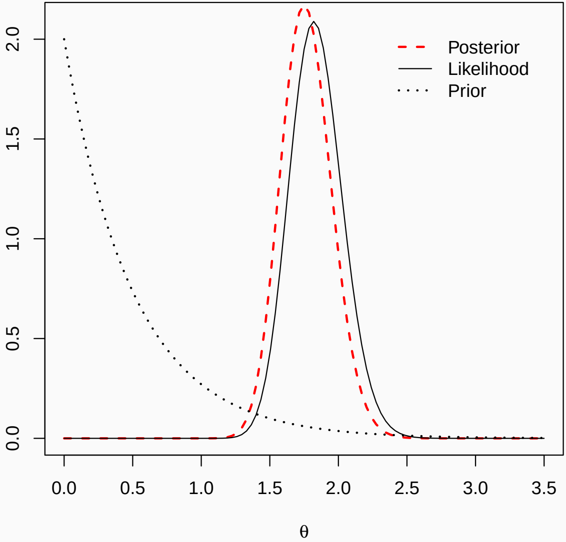
$$E(\theta|\mathbf{x}) \to \overline{x}$$

  but this would imply a limiting prior $p(\theta) \propto \theta^{-1}$ which is improper.

Gamma(1,2) prior, posterior and normalized likelihood for $n = 5$ and $t = 10$.



Gamma(1,2) prior, posterior and normalized likelihood for $n = 50$ and $t = 91$.



The predictive distribution is also easily obtained as,

$$
\begin{aligned}
p(\mathbf{x}) &= \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{\alpha+t-1} e^{-(\beta+n)\theta} d\theta \\
&= \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+t)}{(\beta+n)^{\alpha+t}}.
\end{aligned}
$$

For a single observation x and α integer valued it follows that,

$$\begin{aligned}
p(x) &= \frac{1}{x!}\,\frac{\beta^{\alpha}\,\Gamma(\alpha+x)}{\Gamma(\alpha)\,(\beta+1)^{\alpha+x}} \\
&= \frac{1}{x!}\,\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x}\frac{(\alpha+x-1)!}{(\alpha-1)!} \\
&= \binom{\alpha+x-1}{x}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x}.
\end{aligned}$$

This is the probability function of a Negative-Binomial distribution with parameters α and β.

Mean and variance are easily obtained as,

$$\begin{aligned}
E(X) &= E[E(X|\theta)] = E(\theta) = \alpha/\beta \\
Var(X) &= E[Var(X|\theta)] + Var[E(X|\theta)] \\
&= E(\theta) + Var(\theta) = \frac{\alpha(\beta+1)}{\beta^2}.
\end{aligned}$$

Therefore, a future observation X (after observing $x_1, \ldots, x_n$) has a Negative-Binomial distribution with parameters $\alpha + t$ and $\beta + n$.

$$p(x|x_1, \ldots, x_n) \quad = \quad \binom{\alpha + t + x - 1}{x} \left(\frac{\beta + n}{\beta + n + 1} \right)^{\alpha + t} \left(\frac{1}{\beta + n + 1} \right)^{x}.$$

Multinomial Distribution

In this model we denote the number of ocurrences in each of p categories in n independent trials by $\mathbf{X} = (X_1, \dots, X_p)$ and the associated unknown probabilities by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$.

There are $p-1$ parameters since $\sum_{i=1}^{p} \theta_i = 1$.

The restriction $\sum_{i=1}^{p} X_i = n$ also applies.

Definition

We say that $\mathbf{X}$ has a multinomial distribution with parameters n and $\boldsymbol{\theta}$ and the joint probability function of $\mathbf{X}$ is given by,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{p} x_i!} \prod_{i=1}^{p} \theta_i^{x_i}, \quad x_i = 0, \dots, n, \quad \sum_{i=1}^{p} x_i = n$$

for $0 < \theta_i < 1$ and $\sum_{i=1}^{p} \theta_i = 1$.

- This is clearly a generalization of the Binomial model which has only 2 categories.
- The marginal distribution of each X_i is Binomial with parameters n and θ_i, with

$$E(X_i) = n\theta_i, \quad V(X_i) = n\theta_i(1-\theta_i), \quad \text{and} \quad Cov(X_i, X_j) = -n\theta_i\theta_j.$$

Definition

The random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ follows a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_p$, if its joint density function is given by,

$$p(\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_p) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \ldots, \Gamma(\alpha_p)} \theta_1^{\alpha_1 - 1} \ldots \theta_p^{\alpha_p - 1}, \quad \sum_{i=1}^{p} \theta_i = 1,$$

for $\alpha_1, \ldots, \alpha_p k > 0$ and $\alpha_0 = \sum_{i=1}^{p} \alpha_i$.

Marginal moments,

$$\begin{aligned} E(\theta_i) &= \frac{\alpha_i}{\alpha_0}, \quad Var(\theta_i) = \frac{(\alpha_0 - \alpha_i)\alpha_i}{\alpha_0^2(\alpha_0 + 1)}, \\ Cov(\theta_i, \theta_j) &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

The Dirichlet family with parameters $\alpha_1, \ldots, \alpha_p$ is the natural conjugate prior for the multinomial model.

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \ldots, \Gamma(\alpha_p)} \theta_1^{\alpha_1 - 1} \ldots \theta_p^{\alpha_p - 1}, \quad \sum_{i=1}^{p} \theta_i = 1,$$

with $\alpha_1, \ldots, \alpha_p > 0$ and $\alpha_0 = \sum_{i=1}^{p} \alpha_i$.

The posterior density is given by,

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \prod_{i=1}^{p} \theta_i^{x_i} \prod_{i=1}^{p} \theta_i^{\alpha_i - 1} = \prod_{i=1}^{p} \theta_i^{x_i + \alpha_i - 1}.$$

which is the density of a Dirichlet distribution with parameters $x_i + \alpha_i$, $i = 1, \ldots, p$

- The Dirichlet distribution is a generalization of the Beta distribution.
- The Beta distribution is obtained as a particular case for $p = 2$.
- So, we are extending the conjugate analysis for binomial samples with Beta prior.

The marginal posterior means are,

$$\begin{aligned} E(\theta_i|x_i) &= \frac{\alpha_i + x_i}{\alpha_0 + n} \\ &= \frac{\alpha_0}{\alpha_0 + n} E(\theta_i) + \frac{n}{\alpha_0 + n} x_i. \end{aligned}$$

Let $X_1, \ldots, X_n$ a random sample from a $N(\theta, \sigma^2)$ distribution with θ known and $\phi = \sigma^{-2}$ unknown.

In this case,

$$p(\mathbf{x}|\theta, \phi) \propto \phi^{n/2} \exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}.$$

The kernel has the same form as that of a Gamma distribution.

Since the Gamma family is closed under multiplication this is our natural conjugate prior for ϕ,

$$\phi \sim Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Define $ns_0^2 = \sum_{i=1}^n (x_i - \theta)^2$ and apply Bayes theorem to obtain,

$$\begin{aligned} p(\phi|\mathbf{x}) &\propto \phi^{n/2} \exp\left\{-\frac{\phi}{2} ns_0^2\right\} \phi^{n_0/2-1} \exp\left\{-\frac{\phi}{2} n_0\sigma_0^2\right\} \\ &= \phi^{(n_0+n)/2-1} \exp\left\{-\frac{\phi}{2}(n_0\sigma_0^2 + ns_0^2)\right\}. \end{aligned}$$

Then,

$$\phi|\mathbf{x} \sim Gamma\left(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2 + ns_0^2}{2}\right).$$

Equivalently,

$$
\begin{aligned}
n_0\sigma_0^2\phi &\sim \chi^2_{n_0} \\
(n_0\sigma_0^2 + ns_0^2)\phi \mid \mathbf{x} &\sim \chi^2_{n_0+n}
\end{aligned}
$$

Also, the posterior mean,

$$\frac{n_0 + n}{n_0\sigma_0^2 + ns_0^2} \to \frac{1}{s_0^2} = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2\right]^{-1}$$

when $n \to \infty$.

Definition

A continuous random variable X follows an Inverse Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, if its density function is given by,

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{-(\alpha+1)} \, e^{-\beta/x}, \quad x > 0.$$

Mean and variance are given by,

$$\begin{aligned} E(X) &= \frac{\beta}{\alpha - 1}, \; \alpha > 1 \\ V(X) &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \; \alpha > 2. \end{aligned}$$

- This is the distribution of $1/X$ when $X \sim Ga(\alpha, \beta)$.
- Check that this is the natural conjugate prior distribution for σ^2 in the previous problem.

Mixtures of conjugate priors

Let ϕ a discrete random variable assuming values $\phi_1, \ldots, \phi_k$ and suppose that we can assign a conjugate distribution for θ given each value of ϕ, i.e. we can specify $p(\theta|\phi_i)$, $i = 1, \ldots, k$.

Then, the prior distribution of θ is a mixture of distributions,

$$p(\theta) = \sum_{i=1}^{k} p(\theta|\phi_i)p(\phi_i).$$

It can be verified that the posterior distribution is still a mixture of distributions.

Applying the Bayes theorem we obtain,

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\theta)p(x|\theta)d\theta} = \frac{\sum_{i=1}^{k} p(x|\theta)p(\theta|\phi_i)p(\phi_i)}{\sum_{i=1}^{k} p(\phi_i) \int p(x|\theta)p(\theta|\phi_i)d\theta}.$$

Also, by Bayes theorem,

$$p(\theta|x, \phi_i) = \frac{p(x|\theta)p(\theta|\phi_i)}{\int p(x|\theta)p(\theta|\phi_i)d\theta} = \frac{p(x|\theta)p(\theta|\phi_i)}{m(x|\phi_i)}$$

or equivalently, $p(x|\theta)p(\theta|\phi_i) = p(\theta|x, \phi_i)m(x|\phi_i)$

Again by Bayes theorem, the posterior distribution of ϕ_i is obtained as,

$$p(\phi_i|x) = \frac{m(x|\phi_i)p(\phi_i)}{p(x)}.$$

Finally, we can write the posterior distribution of θ as,

$$p(\theta \mid x) = \frac{\sum_{i=1}^{k} p(\theta|x, \phi_i) m(x|\phi_i) p(\phi_i)}{\sum_{i=1}^{k} m(x|\phi_i) p(\phi_i)} = \sum_{i=1}^{k} p(\theta|x, \phi_i) p(\phi_i|x)$$

As a consequence, the predictive distribution is also a mixture of conditional predictive distributions,

$$p(x) = \sum_{i=1}^{k} m(x|\phi_i) p(\phi_i).$$

Example. If $\theta \in (0, 1)$, the family of $Beta(a, b)$ prior distributions is convenient but these are unimodal and left or right skewed (if $a \neq b$). Other interesting forms which might be more suitable to our prior information can be obtained by mixing 2 or 3 elements from this family.

Suppose that

$$\theta \sim 0.25\text{Beta}(3, 8) + 0.75\text{Beta}(8, 3)$$

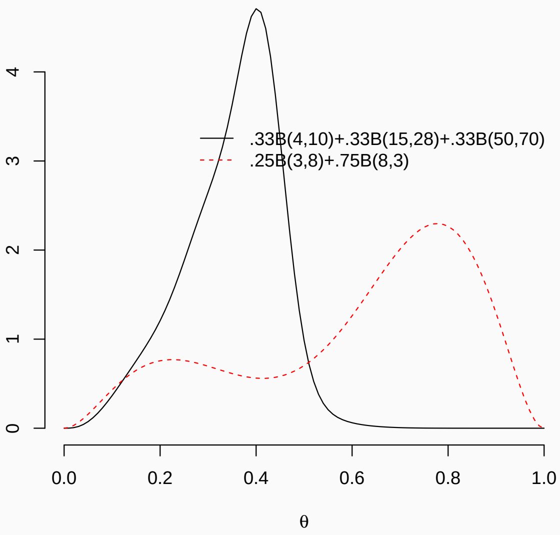
. Then,

- $\theta \in (0.5, 0.95)$ with high probability (0.714).
- $\theta \in (0.1, 0.4)$ with moderate probability (0.2).
- The modes are 0.23 and 0.78.

On the other hand,

$$\theta \sim 0, 33\text{Beta}(4, 10) + 0, 33\text{Beta}(15, 28) + 0, 33\text{Beta}(50, 70)$$

tells us that $\theta > 0.6$ with negligible probabbility and $E(\theta) = 0.35$.


.33B(4,10)+.33B(15,28)+.33B(50,70)
.25B(3,8)+.75B(8,3)
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
θ

Normal Model with Unknown Mean and Variance

Let $X_1, \ldots, X_n$ a random sample from a $N(\theta, \sigma^2)$ distribution with θ and $\phi = \sigma^{-2}$ unknown.

Suppose we assume the following prior distribution for (θ, ϕ),

$$\begin{aligned} \theta|\phi &\sim N(\mu_0, \tau_0^2\phi^{-1}) \\ \phi &\sim \text{Gamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right). \end{aligned}$$

What is the marginal prior distribution of θ?

$$
\begin{aligned}
p(\theta) &= \int p(\theta|\phi)p(\phi)d\phi \\
&\propto \int_0^\infty \phi^{(n_0+1)/2-1} \exp\left\{-\frac{\phi}{2}[n_0\sigma_0^2 + \tau_0^{-2}(\theta-\mu_0)^2]\right\} d\phi \\
&\propto \left[\frac{n_0\sigma_0^2 + \tau_0^{-2}(\theta-\mu_0)^2}{2}\right]^{-\frac{n_0+1}{2}} \\
&\propto \left[1 + \frac{\tau_0^{-2}(\theta-\mu_0)^2}{n_0\sigma_0^2}\right]^{-\frac{n_0+1}{2}},
\end{aligned}
$$

Then,

$$
\theta \sim t_{n_0}(\mu_0, \sigma_0^2\tau_0^2).
$$

Also, combining likelihood function with priors we obtain,

$$\begin{aligned} \theta|\phi, \mathbf{x} &\sim N(\mu_1, \tau_1^2\phi^{-1}) \\ \phi|\mathbf{x} &\sim \text{Gamma}\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right). \end{aligned}$$

where,

$$\begin{aligned} \mu_1 &= \frac{\tau_0^{-2}\phi\mu_0 + n\phi\bar{x}}{\tau_0^{-2}\phi + n\phi} = \frac{\tau_0^{-2}\mu_0 + n\bar{x}}{\tau_0^{-2} + n} \\ \tau_1^{-2} &= \tau_0^{-2} + n \\ n_1 &= n_0 + n \\ n_1\sigma_1^2 &= n_0\sigma_0^2 + \sum(x_i - \bar{x})^2 + \tau_0^{-2}n(\mu_0 - \bar{x})^2/(\tau_0^{-2} + n). \end{aligned}$$

Then,

$$\theta \sim t_{n_1}(\mu_1, \sigma_1^2\tau_1^2).$$

Hierachical Priors

Suppose now that ϕ is a continuous random vector which contains the parameters in the prior distribution of θ (the hyperparameters).

Then we specify $p(\theta|\phi)$ and $p(\phi)$ to obtain the joint prior distribution $p(\theta, \phi)$.

The marginal prior is obtained as,

$$p(\theta) = \int p(\theta|\phi) p(\phi) d\phi.$$

Applying Bayes theorem, we obtain the joint posterior distribution as,

$$\begin{aligned} p(\theta, \phi|\mathbf{x}) &\propto p(\mathbf{x}|\theta, \phi)\ p(\theta|\phi)\ p(\phi) \\ &\propto p(\mathbf{x}|\theta)\ p(\theta|\phi)\ p(\phi) \end{aligned}$$

The marginal posterior distribution of θ is obtained by integration,

$$p(\theta|\mathbf{x}) = \int p(\theta, \phi|\mathbf{x}) d\phi.$$

- The prior specification was split in stages.
- Instead of fixing the value of ϕ we assign a prior distribution completing the second stage in the hierarchy.
- There is no theoretical limit for the number of stages, but in practice 2 or 3 stages are employed in general.

Example. Let $X_1, \ldots, X_n$ such that $X_i \sim N(\theta_i, \sigma^2)$ with σ^2 known and we need to specify a prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

In the first stage we can set $\theta_i \sim N(\mu, \tau^2)$, $i = 1, \ldots, n$. Fixing the value $\tau^2 = \tau_0^2$ and assuming that μ is normaly distributed then $\boldsymbol{\theta}$ follows a multivariate normal distribution.

Now, fixing the value $\mu = \mu_0$ and assuming that τ^{-2} follows a Gamma distribution will imply a multivariate Student-t distribution for $\boldsymbol{\theta}$.

Jeffreys Prior

Intuitively, thinking of all possible values of θ as equally likely seems to be a natural choice to represent complete ignorance.

Bayes and Laplace used a uniform distribution for estimating $\theta \in (0, 1)$, i.e. $\theta \sim \text{Beta}(1, 1)$.

In general, if $p(\theta) \propto k$ for $\theta \in \Theta \subset \mathbb{R}$ then no particular set of values of θ is preferable.

This choice brings some technical difficulties,

- If the parameter space Θ is unbounded the distribution is improper,

$$\int p(\theta) d\theta = \infty.$$

- If $\phi = g(\theta)$ is a nonlinear monotone reparameterization of θ then $p(\phi)$ is non-uniform since,

$$p(\phi) = p_\theta(g^{-1}(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right|.$$

But clearly, if you are completely ignorant about θ you should be completely ignorant about any function of θ.

Jeffreys Prior

Harold Jeffreys' idea to specify a prior was motivated by the desire that inference should not depend on how a model is parameterized.

Jeffreys (1961) proposed a class of priors that is invariant to 1-1 transformations, although generally improper.

Definition

For one observation X with probability (density) function $p(x|\theta)$, the expected Fisher information measure of θ through X is defined as,

$$I(\theta) = E\left[-\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right].$$

If $\boldsymbol{\theta}$ is a vector the expected Fisher information matrix is defined as,

$$\mathbf{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right].$$

- The concept of informatios is associated to a kind of mean curvature of the likelihood function (the more curvature the more precise is the information).
- The mean curvature measured by the second derivative is in most cases negative, therefore the minus sign.
- The expectation is with respect to the distribution of $X|\theta$.

Definition

If X has probability (density) function $p(x|\theta)$, the Jeffreys prior for θ is given by,

$$p(\theta) \propto [I(\theta)]^{1/2}.$$

If $\boldsymbol{\theta}$ is a vector of parameters then,

$$p(\boldsymbol{\theta}) \propto |\det \mathbf{I}(\boldsymbol{\theta})|^{1/2}.$$

Definition

X follows a location model if there exist a function f and a quantity θ such that $p(x|\theta) = f(x-\theta)$ and θ is called location parameter.

The definition is also valid if θ is a vector of parameters.

Examples are the normal distribution with known variance and the multivariate normal with variance-covariance matrix known

The Jeffreys prior for a location model is,

$$p(\theta) \propto \text{constante.}$$

Definition

X follows a scale model if there exist a function f and a quantity σ such that $p(x|\sigma) = (1/\sigma)f(x/\sigma)$ and σ is called scale parameter.

Examples are the exponential distribution with parameter θ and scale parameter $\sigma = 1/\theta$ and the $N(\theta, \sigma^2)$ with known mean and scale σ.

The Jeffreys prior for a scale parameter is,

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

Definition

X follow a location-scale model if there exist a function f and quantities θ and σ such that,

$$p(x|\theta, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right).$$

In this case θ is called location parameter and σ is the scale parameter.

Examples are the uni and multivariate normal and the Cauchy distributions.

The Jeffreys prior for a location-scale parameter usually assumes independence between θ and σ, so that

$$p(\theta, \sigma) = p(\theta)p(\sigma) \propto \sigma^{-1}.$$

Invariance of Jeffreys prior

For a 1 to 1 transformation $\phi = g(\theta)$, a direct application of the change of variables theorem shows that,

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

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with μ and σ^2 unknown. In this case,

$$p(x|\mu, \sigma^2) \propto \frac{1}{\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

so (μ, σ) is the location-scale parameter and $p(\mu, \sigma) \propto \sigma^{-1}$ is the prior.

By the invariance property, the prior for (μ, σ^2) in the normal model is $p(\mu, \sigma^2) \propto \sigma^{-2}$.

Example. Let $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$. Then,

$$\log p(\mathbf{x}|\theta) = -n\theta + \sum_{i=1}^{n} x_i \log \theta - \log \prod_{i=1}^{n} x_i!$$

and taking the second derivative it follows that,

$$\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left[-n + \frac{\sum_{i=1}^n x_i}{\theta}\right] = -\frac{\sum_{i=1}^n x_i}{\theta^2}$$

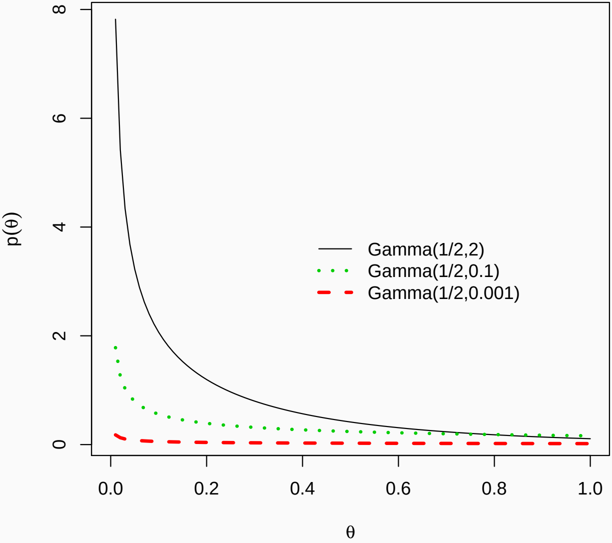
and then,

$$I(\theta) = \frac{1}{\theta^2} E\left[\sum_{i=1}^{n} x_i\right] = \frac{n}{\theta} \propto \theta^{-1}.$$

So, the Jeffreys prior for θ in the Poisson model is,

$$p(\theta) \propto \theta^{-1/2}.$$

- This is also obtained taking the natural conjugate Gamma(α, β) and setting $\alpha = 1/2$ and $\beta \to 0$.
- Is this proper?


Gamma(1/2,2)
Gamma(1/2,0.1)
Gamma(1/2,0.001)
p(θ)
θ
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

Example. Let $X_1, \ldots, X_n \sim \text{Exponential}(\lambda)$. The Fisher information is given by,

$$I(\lambda) = \frac{n}{\lambda^2}$$

so that the Jeffreys prior is,

$$p(\lambda) \propto \lambda^{-1}.$$

This prior is improper (why?)

Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. Then,

$$\log p(\mathbf{x}|\theta) = \sum_{i=1}^{n} x_i \log(\theta) + (n - \sum_{i=1}^{n} x_i) \log(1-\theta)$$

and taking the second derivative,

$$\begin{aligned} \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left[\frac{\sum_{i=1}^n x_i}{\theta} - \frac{(n - \sum_{i=1}^n x_i)}{1-\theta}\right] \\ &= -\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{(n - \sum_{i=1}^n x_i)}{(1-\theta)^2}. \end{aligned}$$

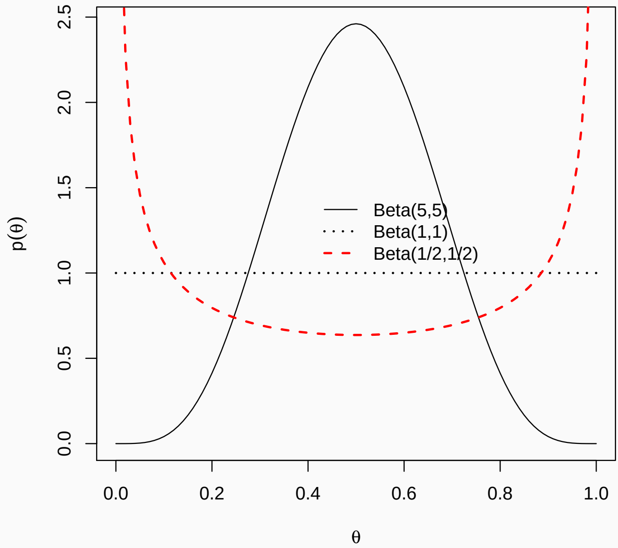
Then,

$$\begin{aligned} I(\theta) &= \frac{1}{\theta^2}\, E\left[\sum_{i=1}^{n} X_i\right] + \frac{1}{(1-\theta)^2}\left(n - E\left[\sum_{i=1}^{n} X_i\right]\right) \\ &= \frac{n}{\theta(1-\theta)} \propto \theta^{-1}(1-\theta)^{-1}. \end{aligned}$$

The Jeffreys prior for θ in the Bernoulli model is,

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

This is also obtained taking the natural conjugate Beta(α, β) and setting $\alpha = 1/2$ and $\beta = 1/2$.


Beta(5,5)
Beta(1,1)
Beta(1/2,1/2)
p(θ)
θ
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0

Example. Multinomial model. The number of ocurrences in each of p categories in n independent trials is denoted by $\mathbf{X} = (X_1, \ldots, X_p)$ and the associated unknown probabilities by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

The joint probability function of $\mathbf{X}$ is,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{p} x_i!} \prod_{i=1}^{p} \theta_i^{x_i}, \quad x_i = 0, \ldots, n, \quad \sum_{i=1}^{p} x_i = n$$

for $0 < \theta_i < 1$ and $\sum_{i=1}^{p} \theta_i = 1$.

The parameter space is given by,

$$\Theta = \{\boldsymbol{\theta} : 0 < \theta_i < 1,\ i = 1, \ldots, p, \sum_{i=1}^{p} \theta_i = 1\}$$

A natural noninformative prior for $\boldsymbol{\theta}$ is to take $\alpha_1 = \ldots, \alpha_p = 1$ in the Dirichlet conjugate prior,

$$p(\boldsymbol{\theta}) \propto 1, \ \boldsymbol{\theta} \in \Theta.$$

What would be the Jeffreys prior in this case?

Recall that,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{p} x_i!} \prod_{i=1}^{p} \theta_i^{x_i}, \quad x_i = 0, \ldots, n, \quad \sum_{i=1}^{p} x_i = n$$

so that,

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{p} x_i \log \theta_i + C$$

$$\frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i} = \frac{x_i}{\theta_i}$$

$$\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -x_i/\theta_i^2,$$

for $i = j$ and zero otherwise.

The Fisher information matrix is thus diagonal with diagonal elements,

$$\frac{1}{\theta_i^2} E(X_i) = \frac{n\theta_i}{\theta_i^2} = \frac{n}{\theta_i}.$$

Then,

$$|I(\boldsymbol{\theta})| = n\theta_1^{-1} \dots \theta_p^{-1},$$

and the Jeffreys prior is,

$$p(\boldsymbol{\theta}) \propto \theta_1^{-1/2} \dots \theta_p^{-1/2}.$$

This is proportional to a Dirichlet density with $\alpha_1 = \cdots = \alpha_p = 1/2$, which is a proper prior.

To sum up

- Jeffreys prior violates the likelihood principle since the Fisher information depends on the sampling distribution.
- Jeffreys prior is widely accepted for single parameter models, but somewhat more controversial and often subject to modification, in multiparameter models.
- Jeffreys priors are usually improper.
- In a few models, the use of improper priors can result in improper posteriors.
- Use of improper priors makes model selection and hypothesis testing difficult.
- General purpose packages WinBUGS and JAGS do not allow the use of improper priors.