Bayesian Data Analysis Summary

Shravan Vasishth (vasishth@uni-potsdam.de)

June 9, 2014

Contents

Bayesian Data Analysis Summary Sheet
Compiled by: Shravan Vasishth (vasishth@uni-potsdam.de)
Version dated: June 9, 2014

Basic math

The gamma function is: $\Gamma(\lambda) = \int_0^\infty y^{y-1}e^{-y}\,dy$
Stirling's approximation for $\Gamma(\lambda+1)$:

$$\Gamma(\lambda+1) \approx \sqrt{2\pi}\lambda^{\lambda+1/2}\exp(-\lambda) \quad \lambda \to \infty \tag{1}$$

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2}$$

$$f(\theta \mid x) = \frac{f(x \mid \theta)f(\theta)}{f(y)} \quad f(x) = \int f(x,\theta)\,d\theta = \int f(x \mid \theta)f(\theta)\,d\theta \tag{3}$$

Independence and conditional independence

When we sample $x_i \sim N(\mu, \sigma^2)$, frequentists would assume the x_i are independent because the parameters are constants. But in the Bayesian setting, the x_i are conditionally independent: $x_i \mid \theta, \sigma^2$; this is because once we have observed x_1, we have learnt something about the parameters, our beliefs about them change, and therefore our beliefs about an unobserved x_j also change. If all parameters are known then the x_i are independent even in the Bayesian setting.

Eliciting priors

Location and dispersion elicitation

Elicit m, elicit 95% credible region, i.e., [m-a,m+a], then $a/2$ is SD, and $(a/2)^2$ (assuming normal distrn).

Bisection method

Elicit 25th, 50th, 75th percentile. More generally, to elicit an N(m,v) distribution to someone's beliefs about a parameter, in theory we need two percentiles, say $P(\theta < x_1) = p_1$ and $P(\theta > x_1) = p_2$. We can then find m and v as solutions of

$$\int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2v}(m-\theta)^2\right\} d\theta = p_1 \tag{4}$$

$$\int_{-\infty}^{x_2} \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2v}(m-\theta)^2\right\} d\theta = p_1 \tag{5}$$

Because people are inaccurate in delivering percentiles, eliciting several percentiles and minimizing the following quantity may be better:

$$\sum[\int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2v}(m-\theta)^2\right\} d\theta - p_i] \tag{6}$$

Example: correlated variables

$$X, Y \sim MVN\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho \\ \rho & \sigma_Y^2 \end{bmatrix}\right) \tag{7}$$

Elicit μ by asking for median and setting it to m. Elicit the σ^2 by obtaining the 95% percentile p and setting $s^2 = ((p-m)/1.65)^2$.
For ρ, ask for percentiles of the difference between X and Y. The difference $X - Y$ would also have a normal distribution under the bivariate normal model, and so by eliciting a 50th and 95th percentile we could determine $\sigma_{X-Y}^2 = Var(X-Y)$, then use:

$$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y) \tag{8}$$

so

$$\rho = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2} \tag{9}$$

Use feedback to check if expert agrees with the implications of their priors.

Integrating out an unknown parameter

Suppose we have a sampling distribution $p(y \mid \theta)$. This could be, for example, $y \sim N(0,1)$ (here, θ is a vector representing the mean and variance). Let our certainty about the parameters be expressed by the probability distribution $p(\theta)$. In this situation, we can produce a predictive distribution by "integrating out the unknown parameter":

$$p(y) = \int p(y,\theta)\,d\theta = \int p(y \mid \theta)p(\theta)\,d\theta \tag{10}$$

This kind of "integrating out" can be done in BUGS quite easily. Suppose we have a random variable Y that has a binomial distribution with success probability θ and sample size n, and our uncertainty about θ is expressed as the distribution Beta(a,b). For specific values of a,b, and n (see code below), we can write the model as:

```
> cat("
 model
    {
     theta ~ dbeta(3,27)
```

```
  Y ~ dbin(theta,20)
 }",
   file="JAGSmodels/integratingout.jag")
```

Then we generate samples from $p(Y)$ in the usual way.

Prior predictive distributions

The prior predictive distribution is the distribution of observations we expect before we observe any data. It is useful for assessing whether the choice of prior distribution captures our prior beliefs.

Example 1: Gamma-Exponential (Machine failure times)

Let machine failure times $t_1, \dots, t_n$ be $t_i \mid \lambda \sim Exp(\lambda)$, for $i = 1, \dots, n$, and $\lambda \sim Gamma(a, b)$.
Prove that the **prior predictive density** of $t_1, \dots, t_n$ is

$$f(t_1, \dots, t_n) = \frac{b^a \Gamma(a+n)}{\Gamma(a)(b + \sum t_i)^{a+b}} \tag{11}$$

$$\begin{aligned} f(t_1, \dots, t_n) &= \int p(t_1)p(t_2) \dots p(t_n)p(\lambda)\, d\lambda \\ &= \int \lambda \exp(-\lambda t_1) \lambda \exp(-\lambda t_2) \dots \lambda \exp(-\lambda t_n) \frac{b^a \lambda^{a-1} \exp(-b\lambda)}{\Gamma(a)}\, d\lambda \\ &= \int \lambda^{n+a-1} \exp(-\lambda \sum t_i - b\lambda) \frac{b^a}{\Gamma(a)}\, d\lambda \\ &= \int \lambda^{n+a-1} \exp(-\lambda(\sum t_i + b)) \frac{b^a}{\Gamma(a)}\, d\lambda \end{aligned} \tag{12}$$

Now, a $Gamma(n + a, \sum t_i + b)$ is:

$$f(\lambda) = \frac{(\sum t_i + b)^{a+n} \lambda^{a+n-1} \exp(-b\lambda)}{\Gamma(a+n)} \tag{13}$$

We can rewrite the above integral as follows.

$$\begin{aligned} &\int \lambda^{n+a-1} \exp(-\lambda(\sum t_i + b)) \frac{b^a}{\Gamma(a)}\, d\lambda \\ =&\frac{\Gamma(a+n) b^a}{\Gamma(a)(b + \sum t)^{a+n}} \int \left[\frac{(\sum t_i + b)^{a+n} \lambda^{a+n-1} \exp(-b\lambda)}{\Gamma(a+n)} \right] d\lambda \end{aligned} \tag{14}$$

The integral integrates to 1, leaving us with:

$$\frac{\Gamma(a+n) b^a}{\Gamma(a)(b + \sum t)^{a+n}} \tag{15}$$

as required.

Posterior predictive distributions

This section is taken from Lunn et al., Section 3.2; slightly reworded.

Once we have the posterior distribution $f(\theta \mid y)$, we can derive the predictions based on this posterior distribution using the same trick as above.

$$p(y_{pred} \mid y) = \int p(y_{pred}, \theta \mid y)\, d\theta = \int p(y_{pred} \mid \theta, y) p(\theta \mid y)\, d\theta \tag{16}$$

Assuming that past and future observations are conditionally independent given θ, i.e., $p(y_{pred} \mid \theta, y) = p(y_{pred} \mid \theta)$, we can write:

$$p(y_{pred} \mid y) = \int p(y_{pred} \mid \theta) p(\theta \mid y)\, d\theta \tag{17}$$

Note that we are conditioning y_{pred} only on y, we do not condition on what we don't know (θ); we integrate out the unknown parameters. Cf. the frequentist approach, which gives only a predictive distribution of y_{pred} given our estimate of θ (a point value).

We use this in the next example.

Example 1: Beta-Binomial

X: no. successes in n trials, $X \sim Bin(n, \theta)$, θ unknown. Prior: $Beta(a, b)$. Say we observe $X = x$. What is the distribution of Y, the number of successes y in a further m trials?

If $m = 1$, just take the posterior mean. If $m > 1$:

$$p(y_{pred} \mid X = x) = \int_0^1 p(y_{pred} \mid \theta) p(\theta \mid X = x)\, d\theta \tag{18}$$

Since $p(y_{pred} \mid \theta) = \binom{m}{y} \theta^y (1-\theta)^{y-m}$, and $p(\theta \mid X = x) = Beta(a + x, b + n - x)$, just plug in the values, and simplify. We get the beta-binomial:

$$f(Y = y \mid X = x) = \binom{n}{k} \frac{B(a + x + y, b + m + n - x - y)}{B(a + x, b + n - x)} \tag{19}$$

Example 2: Beta-Geometric

For n RVs from a geometric distribution, the likelihood is $p(x) = \theta(1-\theta)^{\sum x_i}$, where $x_i = 0, 1, 2, \ldots$. Let the prior be: Beta(a,b).
The posterior is:

$$\theta(1-\theta)^x \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \frac{1}{B(a,b)}\theta^{a+1-1}(1-\theta)^{b+\sum x-1} \tag{20}$$

Posterior: $Beta(a+1, b+\sum x)$.
Posterior predictive distribution: $P(Y = y \mid X = x)$.

$$\begin{aligned} P(Y = y \mid X = x) = &\int_0^1 p(y \mid \theta)p(\theta \mid x)\, d\theta \\ =&\frac{1}{B(a,b)}\int_0^1 \theta(1-\theta)^y\theta^{a+1-1}(1-\theta)^{b+\sum x-1}\, d\theta \\ =&\frac{1}{B(a,b)}\int_0^1 \theta^{a+2-1}(1-\theta)^{b+\sum x+y-1}\, d\theta \end{aligned} \tag{21}$$

Using the standard trick to get rid of the integral:

$$\begin{aligned} &\frac{1}{B(a,b)}\int_0^1 \theta^{a+2-1}(1-\theta)^{b+\sum x+y-1}\, d\theta \\ =&\frac{B(a+2, b+\sum x+y)}{B(a,b)}\int_0^1 \frac{\theta^{a+2-1}(1-\theta)^{b+\sum x+y-1}}{B(a+2, b+\sum x+y)}\, d\theta \\ =&\frac{B(a+2, b+\sum x+y)}{B(a,b)} \end{aligned} \tag{22}$$

The last line arises because the integral on the right, $\int_0^1 \frac{\theta^{a+2-1}(1-\theta)^{b+\sum x+y-1}}{B(a+2,b+\sum x+y)}\, d\theta$, sums to 1 as it is now a pdf.

Example 3: Inverse Gamma (from Ex. 2)

Given that $Y, x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, where μ is known. A prior on σ^2 is defined as $\sigma^2 \sim IG(d, a)$. IG above refers to Inverse Gamma, which is defined as:

$$f(\theta) = \int_{-\infty}^{\infty} \frac{a^d\theta^{-(d+1)}\exp\{-\frac{a}{\theta}\}}{\Gamma(d)}\, d\theta \tag{23}$$

Show that the posterior distribution is $\sigma^2 \sim IG(d_*, a_*)$, where

$$d_* = d + \frac{n}{2} \quad a_* = a + \frac{\sum_i^n (x_i-\mu)^2}{2} \tag{24}$$

Likelihood: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ (assume μ known). Prior: IG(d,a). Posterior is:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2) \times \frac{a^d\theta^{-(d+1)}\exp\{-\frac{a}{\theta}\}}{\Gamma(d)} \tag{25}$$

Now we just need to collect the terms together and rearrange them a bit to get the posterior as $IG(d_*, a_*)$.
Next, derive the posterior predictive distribution given the above information. Let $\theta = \sigma^2$, and let $X = x_1, \ldots, x_n$. The posterior predictive distribution is:

$$\begin{aligned} f(Y \mid X, \mu, \theta) &= \int f(Y, \theta, \mu \mid X)\, d\theta \\ &= \int f(Y \mid \theta, \mu, X) f(\theta \mid \mu, X)\, d\theta \end{aligned} \tag{26}$$

Due to the conditional independence of Y and X, we can write $f(Y \mid \theta, \mu, X)$ as $f(Y \mid \theta, \mu)$. Thus, we can expand the terms out as follows:

$$\begin{aligned} f(Y \mid X, \mu, \theta) &= \int f(Y \mid \theta, \mu) f(\theta \mid \mu, X)\, d\theta \\ &= \int \left[\frac{1}{\sqrt{2\pi}}\theta^{-1/2}\exp\{-\frac{(Y-\mu)^2}{2\theta}\}\right]\left[\frac{a_*^{d_*}\theta^{-(d_*+1)}\exp\{-\frac{a_*}{\theta}\}}{\Gamma(d_*)}\right] d\theta \end{aligned} \tag{27}$$

Rearranging terms and simplifying:

$$\int \frac{1}{\Gamma(d_*)}\frac{a_*^{d_*}}{\sqrt{2\pi}}\theta^{-(d_*+1)-\frac{1}{2}}\exp\left\{\frac{-(Y-\mu)^2}{2\theta} - \frac{a_*}{\theta}\right\} d\theta \tag{28}$$

Let $A = \frac{(Y-\mu)^2}{2} + a_*$ and let $D = d_* + \frac{1}{2}$. These abbreviations allow us to write the above as an inverse gamma PDF (adding the appropriate proportionality constant $\frac{\Gamma(D)}{A^D}$), which sums to 1:

$$\frac{1}{\Gamma(d_*)}\frac{a_*^{d_*}}{\sqrt{2\pi}}\left[\frac{\Gamma(D)}{A^D}\int \frac{A^D}{\Gamma(D)}\theta^{-(D+1)}\exp\left\{-\frac{A}{\theta}\right\} d\theta\right] \tag{29}$$

Since

$$\int \frac{A^D}{\Gamma(D)}\theta^{-(D+1)}\exp\left\{-\frac{A}{\theta}\right\} d\theta = 1 \tag{30}$$

we are left with:

$$f(Y \mid X, \mu, \theta) = \frac{1}{\Gamma(d_*)}\frac{a_*^{d_*}}{\sqrt{2\pi}}\frac{\Gamma(D)}{A^D} \tag{31}$$

As an aside, we note that with a little rearrangement, this will look like the t-distribution.

Posterior predictive density: Lik×Posterior

The posterior predictive density can be computed using the above formula:

$$f(Y \mid X, \mu, \theta) = \frac{\Gamma(D)}{\Gamma(d_*)} \frac{1}{\sqrt{2\pi}} \frac{a_*^{d_*}}{A^D} \tag{32}$$

Prior predictive density: Lik×Prior

The prior predictive density can be computed from the prior and likelihood (here, θ is σ^2):

$$p(Y) = \int p(Y \mid \mu, \theta) p(\theta \mid \mu) \, d\theta \tag{33}$$

Using the same reasoning as above, this reduces to

$$f(Y \mid \mu, \theta) = \frac{\Gamma(D)}{\Gamma(d)} \frac{1}{\sqrt{2\pi}} \frac{a^d}{A^D} \tag{34}$$

where $A = \frac{(Y-\mu)^2}{2} + a$ and $D = d + \frac{1}{2}$.

Example 4: Exponential-Gamma

See later section on Conjugate forms (page 7).

Computing posterior predictive distributions using JAGS

Suppose our prior is Beta(2,2) and the data are Beta(46,54), and the posterior is Beta(47,55).
We first define the data and model. The data includes the parameters for the prior as well (a,b), and the sample size of the predicted values.

```
> ## the data:
> data<-list(a=3,b=27,y=0,n=10,n.pred=20)
> cat("
 model
    {
     ## prior
     theta ~ dbeta(a,b)
     ## likelihood
     y ~ dbin(theta,n)
     ## predicted posterior distribution
     y.pred ~ dbin(theta,n.pred)
    }",
      file="JAGSmodels/predictionexample1.jag" )
```

Conjugate forms

Definition of conjugacy from notes:

> Given the likelihood $f(x \mid \theta)$, if the prior $f(\theta)$ results in a posterior $f(\theta \mid x)$ that has the same form as $f(\theta)$, then we call $f(\theta)$ a conjugate prior.

Conjugate priors can be found if the likelihood function is a member of the exponential family.

Beta-Binomial

P. 38 in notes. Posterior is $Beta(a + x, b + n - x)$.

The posterior predictive distribution of future data $y \mid x$ is (see lecture notes)

$$f(Y = y \mid X = x) = \binom{n}{k} \frac{B(a + x + y, b + m + n - x - y)}{B(a + x, b + n - x)} \tag{35}$$

Posterior parameters as a weighted mean:

$$E[\theta \mid x] = \frac{a}{a+b} \times \frac{w1}{w1 + w2} + \frac{x}{n} \times \frac{w2}{w1 + w2} \quad w_1 = a + b, w_2 = n \tag{36}$$

Multinomial-Dirichlet

p. 31 to-do

Normal-Normal

$$v^* = \frac{1}{\frac{1}{v} + \frac{n}{\sigma^2}} \quad m^* = v^* \left(\frac{m}{v} + \frac{n\bar{x}}{\sigma^2} \right) \tag{37}$$

$$E[\theta \mid x] = m \times \frac{w1}{w1 + w2} + \bar{x} \times \frac{w2}{w1 + w2} \quad w_1 = v^{-1}, w_2 = (\sigma^2/n)^{-1} \tag{38}$$

Two parameter normal-Normal inverse gamma

p. 32 to-do

Poisson-Gamma

$$\begin{aligned} L(\theta) =& \prod_{i=1}^{n} \frac{\exp(-\theta)\theta^{x_i}}{x_i!} \\ =& \frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!} \end{aligned} \tag{39}$$

we can rewrite the right hand side as

$$\text{Posterior} = [\frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!}][\frac{b^a \theta^{a-1}\exp(-b\theta)}{\Gamma(a)}] \tag{40}$$

Disregarding the terms $x!, \Gamma(a), b^a$, which do not involve θ, we have

$$\begin{aligned} \text{Posterior} \propto& \exp(-n\theta)\theta^{\sum_i^n x_i}\theta^{a-1}\exp(-b\theta) \\ =& \theta^{a-1+\sum_i^n x_i}\exp(-\theta(b+n)) \end{aligned} \tag{41}$$

If we equate $a^* - 1 = a - 1 + \sum_i^n x_i$ and $b^* = b + n$, we can rewrite the above as:

$$\theta^{a^*-1}\exp(-\theta b^*) \tag{42}$$

This means that $a^* = a + \sum_i^n x_i$ and $b^* = b + n$. We can find a constant k such that the above is a proper probability density function, i.e.:

$$\int_{-\infty}^{\infty} k\theta^{a^*-1}\exp(-\theta b^*) = 1 \tag{43}$$

Thus, the posterior has the form of a Gamma distribution with parameters $a^* = a + \sum_i^n x_i, b^* = b + n$. Hence the gamma distribution is a conjugate prior for the Poisson.

Exponential-Gamma

If prior is $Ga(a,b)$, given $Exp(\lambda)$ as likelihood, posterior is $Ga(a+n, \sum x + b)$.

$$\begin{aligned} Lik \times Prior =& [\lambda^n \exp\{-\lambda \sum x\}] \left[\frac{b^a \lambda^{a-1}\exp\{-b\lambda\}}{\Gamma(a)}\right] \\ =& \frac{b^a}{\Gamma(a)}\lambda^{a+n-1}\exp\{-\lambda(b+\sum x)\} \end{aligned} \tag{44}$$

The posterior is $Gamma(a+n, b+\sum x)$.

The posterior predictive distribution of Y given x

$$\begin{aligned} f(Y \mid x, \lambda) =& \int f(Y, \lambda \mid x)\, d\lambda \\ =& \int f(Y \mid \lambda, x) f(\lambda \mid x)\, d\lambda \\ =& \int f(Y \mid \lambda) f(\lambda \mid x)\, d\lambda \\ =& \int \lambda \exp\{-\lambda y\}\lambda^{a+1-1}\exp\{-\lambda(b+\sum x)\}\frac{b^a}{\Gamma(a)}\, d\lambda \\ =& \int \exp\{-\lambda y\}\lambda^{a+2-1}\exp\{-\lambda(b+\sum x + y)\}\frac{b^a}{\Gamma(a)}\, d\lambda \end{aligned} \tag{45}$$

[Note that we can write $f(Y \mid \lambda, x)$ as $f(Y \mid \lambda)$ due to conditional independence of Y and x].
Now we use the trick of finding the integral to solve this:

$$\frac{\Gamma(a+2)}{(b+\sum x+y)^{a+2}}\frac{b^a}{\Gamma(a)}\int \frac{(b+\sum x+y)^{a+2}}{\Gamma(a+2)}\lambda^{a+2-1}\exp\{-\lambda(b+\sum x+y)\}\, d\lambda \tag{46}$$

The integral on the right is the $Gamma(a+2, b+\sum x + y)$ and integrates to 1. Hence,

$$f(Y \mid x, \lambda) = \frac{\Gamma(a+2)}{(b+\sum x+y)^{a+2}}\frac{b^a}{\Gamma(a)} \tag{47}$$

If we have a Jeffrey's prior, Ga(0,0), then $a, b \to 0$.
Posterior for $\lambda \mid x$ becomes: $Ga(n, \sum x)$.
Predictive distribution for $Y \mid x$ becomes:

$$\frac{\Gamma(2)}{(\sum x + y)^2} \tag{48}$$

Plug-in prediction

In the Jeffrey's prior case, estimate λ from posterior mean: $\hat{\lambda} = \frac{n}{\sum x}$, and then plug it into $P(Y \mid x)$.

$$\begin{aligned} f(Y \mid \lambda, x) =& \lambda \exp(-\lambda y)\lambda^{a+1-1} exp(-\lambda(b+\sum x)) \\ =& \frac{1}{\bar{x}}\exp(\frac{1}{\bar{x}}(\sum x - y)) \end{aligned} \tag{49}$$

Given that x=10, y=0,10,30,50, using equation for the posterior predictive distribution 48, we have $(\frac{\Gamma(2)}{(\sum x+y)^2})$.

And using equation 49 for the plug-in posterior predictive distribution, we have $\frac{1}{\bar{x}} \exp(\frac{1}{\bar{x}(\sum x - y)})$.

```
                0      10       30          50        500
postpred 0.01000 0.0025 0.000625 0.00027778 3.8447e-06
plugin   0.27183 0.1000 0.013534 0.00183156 5.2429e-23
```

The plug-in prediction approximates the true PPD for large values of y.

Deriving full conditional distributions

Full conditional distributions are needed, e.g., for Gibbs sampling.
Deriving a full conditional using a DAG as a guide (Lunn et al): Prior on the target parameter times distribution of each child of target parameter conditional on that child's parents.
This section is based on Gilks et al.
How to get the full conditional distribution of $\theta_i \mid \theta_{-i}, \mathbf{x}$? P. 77 of lecture notes.

$$\begin{aligned} f(\theta_i \mid \theta_{-i}, \mathbf{x}) =& \frac{f(\theta_i, \theta_{-i} \mid \mathbf{x})}{f(\theta_{-i}\mathbf{x})} \\ =& \frac{f(\theta \mid \mathbf{x})}{\int f(\theta \mid \mathbf{x})\, d\theta_i} \\ \propto f(\theta \mid \mathbf{x}) \end{aligned} \tag{50}$$

So, to get $f(\theta_i \mid \theta_{-i}, \mathbf{x})$ we simply take $f(\theta \mid \mathbf{x})$ and treat all the parameters in θ_{-1} as constants.

Example 1: A simple Bayesian model

- $y_i \sim N(\mu, \tau^{-1}), i = 1, \dots, n$
- $\mu \sim N(0, 1)$
- $\tau \sim Ga(2, 1)$

y_i are conditionally independent given μ, τ, and μ, τ are themselves independent.
Let $y = (y_1, \dots, y_n)$.
The joint likelihood:

$$p(y, \mu, \tau) = \prod_{i=1}^{n} P(y_i \mid \mu, \tau) P(\mu) P(\tau) \tag{51}$$

Expanding this out:

$$p(y, \mu, \tau) = (2\pi)^{(n+1)/2} \tau^{n/2} \exp\left\{-\frac{\tau}{2} \sum (y_i - \mu)^2\right\} \exp\left\{-\frac{1}{2}\mu^2\right\} \tau e^{-\tau} \tag{52}$$

When y is observed, the joint posterior of μ, τ is:

$$p(\mu, \tau) = P(\mu, \tau \mid y) = \frac{P(y, \mu, \tau)}{\int P(y, \mu, \tau)\, d\mu\, d\tau} \tag{53}$$

The full conditional for μ is

$$p(\mu \mid \tau) = \frac{P(\mu, \tau \mid y)}{P(\tau \mid y)} = \frac{P(y, \mu, \tau)}{P(y, \tau)} \tag{54}$$

This is proportional to $P(y, \mu, \tau)$. Therefore,

$$\begin{aligned} p(\mu \mid \tau) &\propto \exp\left\{-\frac{\tau}{2} \sum (y_i - \mu)^2\right\} \exp\left\{-\frac{1}{2}\mu^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}(1 + n\tau)(\mu - \frac{t \sum y}{1 + n\tau})^2\right\} \end{aligned} \tag{55}$$

Therefore $p(\mu \mid \tau)$ is $N(\frac{t \sum y}{1+n\tau}, (1 + n\tau)^{-1})$.
Similarly,

$$\begin{aligned} p(\tau \mid \mu) \propto& \tau^{n/2} \exp\left\{-\frac{\tau}{2} \sum (y - \mu)^2\right\} \tau e^{-\tau} \\ =& \tau^{1+n/2} \exp\left\{-\tau[1 + \frac{1}{2} \sum (y_i - \mu)^2]\right\} \end{aligned} \tag{56}$$

This is the kernel for $Ga(2 + n/2, 1 + +\frac{1}{2} \sum (y_i - \mu)^2)$.

Example 2: A more complex DAG (hierarchical) model

The full conditional distribution for any parameter can be contructed from those few terms of the joint distribution which depend on it.
Consider the **Normal random effects model**:

- $y_{ij} \sim N(\alpha_i, \tau^{-1})$, where $j = 1, \dots, m_i$, and $i = 1, \dots, n$.
- $\alpha_i \sim N(\mu, \omega^{-1})$.
- $\mu \sim N(0, 1)$
- $\tau \sim Ga(2, 1)$
- $\omega \sim Ga(1, 1)$

Hyperparameters are μ and τ^{-1}, and the priors on these are the hyperpriors.
Assumptions:

- y_{ij} independent given all parameters.

- α_i independent given μ, τ, ω.
- μ, τ, ω mutually independent.

The joint probability:

$$p(y, \alpha, \mu, \tau, \omega) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m_i} P(y_{ij} \mid \alpha_i, \tau) P(\alpha_i \mid \mu, \omega) \right\} P(\mu)P(\tau)P(\omega) \tag{57}$$

Then, the full conditional for α_i is

$$p(\alpha_i \mid y, \alpha_{-1}, \mu, \tau, \omega) \propto \prod_{j=1}^{m_i} P(y_{ij} \mid \alpha_i, \tau) P(\alpha_i \mid \mu, \omega) \tag{58}$$

Note that $P(\mu)P(\tau)P(\omega)$ drop off as they are now constants. The above expands to:

$$p(\alpha_i \mid y, \alpha_{-1}, \mu, \tau, \omega) \propto \exp\left\{ -\frac{1}{2}(\omega + m_i\tau)(\alpha_i - \frac{\omega\mu + \tau\sum_{i=1}^{m_i} y_{ij}}{\omega + m_i\tau})^2 \right\} \tag{59}$$

In other words, $N(\frac{\omega\mu+\tau\sum_{i=1}^{m_i} y_{ij}}{\omega+m_i\tau}, (\omega + m_i\tau)^{-1})$.

Example 3: Meta-analysis of stroke data (old exam question)

Given data from four trials on treatment and control groups. The number of strokes in each group is counted.

Let x_i be the number of strokes in the control group, y_i the number of strokes in the treatment group, i is trial number. Total count in control is nx_i and in treatment group ny_i.
The model is

- Likelihoods:
  - $x_i \mid \gamma_i, nx_i \sim Binom(\gamma_i, nx_i)$
  - $y_i \mid \gamma_i, ny_i \sim Binom(\delta_i, ny_i)$
- Priors:
  - $\mu_i \sim N(mean = 0, variance = 1/0.001)$
    $logit(\gamma_i) \leftarrow \mu_i$
  - $\theta_i \sim N(\alpha, \beta)$
    $logit(\delta_i) \leftarrow \mu_i + \theta_i$
  - $\alpha \sim N(0, 1/0.002)$
  - $\beta \leftarrow 1/\sigma^2$
  - $\sigma^2 \sim Unif(0, 100)$

Find the full conditional distribution of $f(\alpha \mid \beta, \mu, \theta, x, y)$.
The joint posterior distribution of the parameters is:

$$p(\alpha, \beta, \mu, \theta \mid x, y) = \prod_{i=1}^{4} p(x_i \mid \gamma_i, nx_i) \prod_{i=1}^{4} p(y_i \mid \delta_i, ny_i) p(\alpha)p(\beta)p(\mu)p(\theta \mid \alpha, \beta) \tag{60}$$

To get the conditional distribution of α, we just treat all terms not involving α as constant. This gives us (up to proportionality):

$$\begin{aligned} (\alpha \mid \beta, \mu, \theta, x, y) &\propto p(\alpha)p(\theta \mid \alpha, \beta) \\ &\propto N(0, 0.002) \times N(\alpha, \beta) \end{aligned} \tag{61}$$

```
> dat<-list(y=c(5,6,8,11),
           x=c(10,6,12,11),
           ny=c(100,155,170,190),
           nx=c(100,130,150,180))
> ## prediction:
> cat("
 model
    {
 for(i in 1:4){
 x[i] ~ dbin(gamma[i],nx[i])
 y[i] ~ dbin(delta[i],ny[i])
 logit(gamma[i]) <- mu[i]
 logit(delta[i]) <- mu[i]+theta[i]
 ## priors:
 mu[i] ~ dnorm(0,0.001)
 theta[i] ~ dnorm(alpha,beta)
 }
 alpha ~ dnorm(0,0.001)
 beta<-1/sigma
 sigma ~ dunif(0,100)
 mu.pred ~ dnorm(0,0.001)
 theta.pred ~ dnorm(alpha,beta)
 probc<-exp(mu.pred)/(1+exp(mu.pred))
 probt<-exp(mu.pred+theta.pred)/(1+exp(mu.pred+theta.pred))
 logit(gamma.pred) <- mu.pred
 logit(delta.pred) <- mu.pred+theta.pred
 ## shaped like a Beta(1/2,1/2)
 st ~ dbin(gamma.pred,100)
 sc ~ dbin(delta.pred,100)
 }
 ",
     file="JAGSmodels/metaanalysisstrokes.jag" )
```

Example 4: Rare diseases example (old exam problem)

Data on rare diseases in different towns.

```
> dat<-list(alpha=c(251,229,174,143,103,101),
           x=c(20,13,11,7,6,5))
> ## load rjags library:
> library(rjags,quietly=T)
> ## prediction:
> cat("
 model
    {
 for(j in 1:6){
 x[j] ~ dpois(lambda[j])
 lambda[j] <- alpha[j]*theta[j]
 theta[j] ~ dgamma(psi,rho)
 }
 psi ~ dgamma(0.001,0.001)
 rho ~ dgamma(0.001,0.001)
 mu <- psi/rho
 ## coef of var: extent of variability
 ## (sigma/mu) in relation to mean
 ## incidence of disease
 cv <- 1/sqrt(psi)
 theta.pred ~ dgamma(psi,rho)
 lambda.pred<-259*theta.pred
 x.pred ~ dpois(lambda.pred)
 }
 ",file="JAGSmodels/rarediseasepred.jag" )
```

Assumptions: (a) λ_i, λ_j are correlated (not independent), for $i \neq j$; (b) λ_i, λ_j are exchangeable; (c) psi, rho are independent.

Example 5: Bird breeding (old exam problem)

```
> ## n=no. of birds
> ## s=no. of pairs successfully bred
> ## t=temperature
> dat<-list(n=c(26,48,16,54,45),
           s=c(18,36,10,41,34),
           t=c(7.5,8.6,6.9,9.1,8.5))

> cat("
 model
    {
 for(i in 1:5){
     s[i] ~ dbin(p[i],n[i])
     logit(p[i]) <- a + b * t[i] + r[i]
     r[i] ~ dnorm(0,tau)
     } ## likelihood
    ## priors:
    a ~ dnorm(0,0.001)
    b ~ dnorm(0.1,200)
    tau ~ dgamma(0.001,0.001)
    sigma <- 1/sqrt(tau)
 ## prediction:
    temp ~ dnorm(7,1/0.5^2)
    rnew ~ dnorm(0,tau)
    logit(p.pred) <- a + b*temp + rnew
    snew ~ dbin(p.pred,50)
 }",
     file="JAGSModels/birds.jag" )
```

Predicted prob (p.pred):

```
          Mean             SD
      0.701306       0.062171
 2.5%   25%   50%   75% 97.5%
0.571 0.669 0.705 0.739 0.810
```

Full conditional using p on logit scale:
Prior on τ is Gamma(0.001,0.001), times distribution of each child of tau conditional on that child's parents ($P(p \mid \alpha, \beta, \tau) = N(\alpha + \beta t, 1/\tau)$.
Therefore: $Gamma(0.001, 0.001) \times N(\alpha + \beta t, 1/\tau)$. Not an NIG because the two distributions are independent.

Decision theory

Components of a decision problem

1. Alternatives: acts, actions
2. Events: states, outcomes of random processes

3. Probabilities of events
4. Consequences of each event (Loss/Utility)
5. Decision rule: select best alternative

The best act is the one with highest expected utility.

Example: Lottery tickets

The expected gain of buying: $150 \times 0.5 + -100 \times 0.5 = 25$. The expected gain of not buying: 0.
People may attach different utilities to winning and losing. Determining utility:

- Give 1 to preferred outcome, 0 to dispreferred.
- Utility for Don't Buy:
  ∃ p such that indifferent between buying and not buying.
  - If p=1 win is certain, prefer to buy
  - p=0, loss is certain, prefer to not buy

  Say $p = 0.80$. This is the utility of not buying.

Expected utility for buying: $p \times 1 + (1 - p) \times 0 = p$.
Expected utility for not buying: $p \times 0.8 + (1 - p) \times 0.8 = 0.8$.
If $p = 0.5$, then Decision: Don't buy.

Formalization

- S: status
- U(S): utility of status S. $U(S_0) = u_0$, $U(S_1) = u_1$.
- $d \in D$: possible decisions.
- $S(d)$: status after decision.
- Optimal decision $d* = \text{argmax}_{d \in D}(U(D(d)))$
- Let status depend on random variable X. Then:
  - $S(X)$: uncertain status, depends on how X is realized ($S(X = Heads)$, status: won; $S(X = Tails)$, status: lost).
  - $U(S(X))$ is a random variable.
  - S_X is status when X is some specific outcome but unknown which one.
  - $U(S_X)$: a single number.
  - Utilities:

$$U(S_X(DB)) = E[U(S(DB, X))] = 0.8 \times 0.5 + 0.8 \times 0.5 = 0.8 \quad (62)$$

$$U(S_X(B)) = E[U(S(B, X))] = 1 \times 0.5 + 0 \times 0.5 = 0.5 \quad (63)$$

Optimal decision: don't buy.

In coin toss example:

- X=H or X=T.
- S(H): won; S(T): lost.
- U(S(H))=1; U(S(T))=0.
- $S_X = won$, or $S_X = lost$ (depending on what X is).
- $U(S_X) = qu_1 + (1 - q)u_0 = E[U(S(X))]$.

Risk-averseness: Utility of not buying can go down (e.g., 0.2) if the amounts are small, leading to a reversal in optimal decision (Buy).

Inference as a decision problem

Hypothesis testing

$H_0 : \theta \in T$. Compute posterior probability of H_0 being true:
$P(H \mid x) = \int f(\theta \mid x)\, d\theta$
The **loss structure/function**:

a $E[U(S(Accept, X))] = p \times 0 + (1 - p)l_1 = (1 - p)l_1$

b $E[U(S(Reject, X))] = p \times l_2 + (1 - p) \times 0 = (1 - p)l_1$

If $(a) < (b)$, accept H. I.e., accept if:

$$(1 - p)l_1 < pl_2 \Rightarrow p > \frac{l_1}{l_1 + l_2} \quad (64)$$

Point estimation

Decision d: estimate of θ.
Event: true unknown value of θ.
Let $S(d, \theta)$ be your status for an estimate d and true value θ.

Quadratic loss (optimal: mean)

$$L[S(d,\theta)] = (d-\theta)^2 \tag{65}$$

We choose d* to minimize:

$$\begin{aligned} E[S(d,\theta)] =& E[(d-\theta)^2] \\ =& E[d^2+\theta^2-2d\theta] \\ =& d^2+E[\theta^2]-2dE[\theta] \end{aligned} \tag{66}$$

Differentiate with respect to d and maximize:

$$2d-2E[\theta] \Rightarrow E[\theta] = d*. \tag{67}$$

That is, the optimal estimate of θ under quadratic loss is $E[\theta]$.
If you make the optimal decision (choose mean as your estimate of θ), the expected loss is the variance:

$$E[L(d*,\theta)] = E[(\theta-E[\theta])^2] = Var(\theta) \tag{68}$$

Absolute loss (median)

$L(S(d,\theta)) = \mid d-\theta \mid$
Minimize expected loss:
$E(S(d,\theta)) = E[\mid d-\theta \mid]$

Zero-one loss (mode)

$$L(S(d,\theta)) = \begin{cases} 0 & if \mid d-\theta \mid \leq a, \\ 1, & if \mid d-\theta \mid > a \end{cases} \tag{69}$$

Expected loss:
$P(\mid d-\theta \mid > a \mid x) = 1 - P(d-a\theta > d+a) = 1 - P(\theta \leq d+a) + P(\theta < d-a)$

Example decision analysis with data

Hormone measurement example from practice exam.

- Prior: 75% interval ("patient healthy") is [-0.3,0.3].
- Data from patient: $x = 0.2$, known $\sigma = 0.15$
- Compute posterior N(m*,v*)
- Decision problem for incoming patient:

	H	Not H
d1: healthy	0	c
d2: further screening	1	0

  Loss d1: (1-p)c; Loss d2: p

- For cost c=3, optimal decision is the one with lower loss among d1 and d2:

  $L(d1,\theta): 3-3p \quad L(d2,\theta) = p$

- Given prior, σ, and loss function, what would the value of c have to be such that the doctor always sends patients for screening, regardless of the data?

  We have to find out when this is true: $(1-p)c > p$. Compute $p = P(\theta \in H \mid x)$ and solve for c.