L3: Bayesian Theory—Priors, Part 2

1 Specifying hyperparameters

Hyperparameters are sometimes calculated on the basis of subjective specifications of summary measures for the prior distribution. One common procedure is to equate *a priori* guesses of expected values, variances, or coefficients of variances to the algebraic formulas for those values in the prior distribution[1]. Another is to specify the probability that θ lies in some range $[\theta_L, \theta_U]$; e.g., $\Pr(3 \leq \theta \leq 9) = 0.8$.

Beta prior for Binomial. For example, a Beta(α, β) prior will be used for a binomial data distribution, Binomial(n, θ), and the desired expected value of θ is 0.8 with a coefficient of variation of 0.25 ($\sqrt{V[\theta]}/E[\theta]$). This implies a desired variance of $(0.25 * 0.8)^2 = 0.04$. One can calculate α and β by solving the following system of two equations with two unknowns[2].

$$E[\theta] = \frac{\alpha}{\alpha + \beta}$$
$$V[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Gamma prior for Poisson. A similar exercise can be carried out when specifying a Gamma$(\alpha,\ \beta)$ prior for a Poisson data distribution, Poisson(θ). Desired characteristics of the prior distribution are that the expected value of θ is 10 with a coefficient of variation of 0.30 ($\sqrt{V[\theta]}/E[\theta]$). One can calculate α and β using the following[3]:

$$E[\theta] = \frac{\alpha}{\beta}$$
$$V[\theta] = \frac{\alpha}{\beta^2}$$

Normal prior for Normal μ. Instead of specifying moment-related measures of parameters, one might specify desired quantiles. For example, the data distribution is Normal$(\theta,\ \sigma^2)$, where σ^2 is known, and a normal distribution will be used as the prior for θ, i.e., $\theta \sim$ Normal$(\mu_o,\ \sigma_o^2)$. The 5th and 95th percentiles for the prior should be 15 and 30. One can calculate μ_0 and σ_0^2 using the following equations[4]:

$$15 = \mu_o - 1.960 * \sigma_o$$
$$30 = \mu_o + 1.960 * \sigma_o$$

In general, letting z_p denote the standard normal, Normal(0,1), quantile for probability p, and y_p denote the normal quantile for probability p for Normal$(\mu_o,\ \sigma_o^2)$:

$$y_{p1} = \mu_o + z_{p1} * \sigma_o$$
$$y_{p2} = \mu_o + z_{p2} * \sigma_o$$

where $p1 < p2$ (and $y_{p1} < y_{p2}$).

[1] This is referred to moment matching in Bayesian Models: A Statistical Primer for Ecologists, by Hobbs and Hooten (2015). I really like this book and note that it is written for ecologists rather than statisticians. It is available online through the University Library webpage.

[2] Try this out: given $E[\theta]$=0.8 and CV=0.25, show that α=2.4 and β=0.6.

[3] Given $E[\theta]$=12 and CV=0.3, show that α=11.11111 and β=0.9259259.

[4] With 5th and 95th percentiles equal to 15 and 30, μ_o=22.5; σ_o= 3.826531.

2 Normal Distribution Priors

Given the ubiquity and utility of the normal distribution, it is useful to thoroughly explore the posterior for the parameters of that distribution for various priors. We begin with conjugate priors for just one parameter of a univariate normal distribution, here denoted Normal(μ, σ^2), either μ or σ^2. Later we'll examine the case of specifying joint prior distributions for μ and σ^2.

To derive the conjugate posteriors, there is a fair amount of algebraic manipulation involved. It is worth understanding and being able to reproduce the following for yourself as such techniques are useful in other circumstances.

In all cases considered, the data distribution is Normal(μ, σ^2) from which there are n independent and identically distributed (iid) observations, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. The joint probability density function is

$$f(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) \tag{1}$$

Remarks.

- *Re-expressing the Normal pdf.* A useful re-expression of the normal pdf is the following.

  $$f(\mathbf{y}|\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right) \tag{2}$$

  Check the validity of this re-expression. (Hint: rewrite $(y_i - \mu)^2$ as $([y_i - \bar{y}] + [\bar{y} - \mu])^2$.)

- *Sufficient statistics.* Suppose that σ^2 is known. Note the only term in eq'n 2 containing μ includes $\bar{y}$, thus $\bar{y}$ is the relevant data for inference about μ. We say that $\bar{y}$ is a *sufficient statistic* for μ, as this statistic contains all the information in the data that is useful for estimating the unknown parameter (here it is μ).

  Now suppose that σ^2 is also unknown, letting $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$, then

  $$f(\mathbf{y}|\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)}{2\sigma^2}s^2\right) \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right) \tag{3}$$

  and $(\bar{y}, s^2)$ is the *joint sufficient statistic* for (μ, σ^2).

- *Precision.* Instead of writing the normal pdf in terms of μ and σ^2, it is sometimes written in terms of μ and τ, where τ is called the Precision, and is the inverse of the variance, $\tau = 1/\sigma^2$. Symbolically, $y \sim$ Normal$\left(\mu, \frac{1}{\tau}\right)$, and mathematically

  $$f(y|\mu, \tau) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(y - \mu)^2}{2}\right) \tag{4}$$

  Some software packages for Bayesian inference, e.g. JAGS and WinBUGS, specify the normal distribution in terms of the precision τ.

  The term precision has an intuitive interpretation, if a random variable is more precise, it is less variable, i.e., as τ increases, σ^2 decreases.

2.1 Normal distribution: unknown μ and known σ^2

While assuming that σ^2 is known is usually not a realistic situation, the methods used for inference for μ are helpful building blocks for more realistic models.

To begin we write the likelihood for μ, given that σ^2 is known, and examine the portion of the pdf in Eq'n (2) that involves μ alone:

$$f(\mathbf{y}|\mu,\sigma^2) \equiv L(\mu|\mathbf{y},\sigma^2) \propto \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{(\mu-\bar{y})^2}{2\sigma^2/n}\right) \tag{5}$$

Note that the last expression in (5) can be seen as the kernel of a Normal distribution. Thus the *conjugate distribution* for μ (when σ^2, or τ, is known) is the Normal distribution:

$$\text{Prior for } \mu\text{:} \quad \mu \sim \text{Normal}\left(\mu_0, \sigma_0^2\right)$$

where μ_0 and σ_0^2 are the hyperparameters of the prior. An alternative formulation for the prior (see Reich and Ghosh, p 47), which leads to a tidier posterior distribution, is to write $\sigma_0^2 = \sigma^2/m$, where m is some positive number. Of course if one specifies σ_0^2 first, then m is σ^2/σ_0^2.

$$\text{Prior for } \mu\text{:} \quad \mu \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) \tag{6}$$

The posterior distribution for μ given a normal prior is then:

$$\begin{aligned}
\text{Posterior:} \quad p(\mu|\mathbf{y},\sigma^2) &\propto \pi(\mu)p(\mathbf{y}|\mu) \\
&= (2\pi\sigma^2/m)^{-\frac{1}{2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma^2/m}\right) * \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{m(\mu-\mu_0)^2}{2\sigma^2}\right) \ \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left[-\frac{m\mu^2 - 2m\mu_0\mu + n\mu^2 - 2n\bar{y}\mu}{2\sigma^2}\right] \\
&= \exp\left[-\frac{(m+n)\mu^2 - 2(m\mu_0 + n\bar{y})\mu}{2\sigma^2}\right] \\
&= \exp\left[-\frac{(\mu - \frac{m\mu_0+n\bar{y}}{m+n})^2}{2\sigma^2/(m+n)}\right] \qquad (7)
\end{aligned}$$

where eq'n 7 is the kernel for a normal distribution, in other words:

$$\text{Posterior for } \mu|\mathbf{y},\sigma^2\text{:} \quad \text{Normal}\left(\frac{m\mu_0+n\bar{y}}{m+n}, \frac{\sigma^2}{m+n}\right) \tag{8}$$

The posterior mean for μ is thus a weighted combination of the prior mean, μ_0, and the sample mean, $\bar{y}$:

$$E[\mu|\mathbf{y},\sigma^2] = \frac{m\mu_0+n\bar{y}}{m+n} = \frac{m}{m+n}\mu_0 + \frac{n}{m+n}\bar{y} = (1-w)\mu_0 + w\bar{y}$$

where w=$\frac{n}{m+n}$.

Comments.

- As n increases, the posterior mean is dominated by the sample mean:

$$\lim_{n\to\infty} E[\mu|\mathbf{y},\sigma^2] = \lim_{n\to\infty}\left[\frac{m}{m+n}\mu_0 + \frac{n}{m+n}\overline{y}\right] = \overline{y}$$

  And the posterior becomes concentrated at $\overline{y}$

$$\lim_{n\to\infty} V[\mu|\mathbf{y},\sigma^2] = \lim_{n\to\infty}\frac{\sigma^2}{m+n} = 0$$

- Conversely, as m increases, σ^2/m decreases, and the posterior is dominated by the prior.

$$\lim_{m\to\infty} E[\mu|\mathbf{y},\sigma^2] = \lim_{m\to\infty}\frac{m\mu_0 + n\bar{y}}{m+n} = \mu_0$$
$$\lim_{m\to\infty} V[\mu|\mathbf{y},\sigma^2] = \lim_{m\to\infty}\frac{\sigma^2}{m+n} = 0$$

  Thus a point mass at μ_0 as σ_0^2 goes to 0.

- Conversely, as m decreases, σ^2/m increases (a "vaguer" prior), the influence of the prior decreases:

$$\lim_{m\to 0} E[\mu|\mathbf{y},\sigma^2] = \lim_{m\to 0}\frac{m\mu_0 + n\bar{y}}{m+n} = \overline{y}$$
$$\lim_{m\to 0} V[\mu|\mathbf{y},\sigma^2] = \lim_{m\to 0}\frac{\sigma^2}{m+n} = \frac{\sigma^2}{n}$$

2.1.1 Unknown μ and known τ

If express the sampling distribution for $\mathbf{y}$ in terms of precision, $\tau = 1/\sigma^2$, then the prior for μ (given known τ) is:

$$\mu \sim \text{Normal}\left(\mu_0, \frac{1}{\tau m}\right) \tag{9}$$

And the posterior for μ:

$$\text{Posterior for } \mu|\mathbf{y},\sigma^2: \quad \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{1}{\tau(m+n)}\right) \tag{10}$$

2.1.2 Posterior predictive distribution

As discussed in Lecture 1, the posterior predictive distribution is the marginal probability distribution for a new scalar value, y^{new}, given the past data, $\mathbf{y}^{old}$:

$$p(y^{new}|\mathbf{y}^{old}) = \int p(y^{new}|\theta,\mathbf{y}^{old})p(\theta|\mathbf{y}^{old})d\theta = \int p(y^{new}|\theta)p(\theta|\mathbf{y}^{old})d\theta \tag{11}$$

The second equality results from y^{new} being conditionally independent on $\mathbf{y}^{old}$ given θ.

In this normal distribution case with known σ^2, letting μ_1 and σ_1^2 denote the posterior mean and variance for μ:

$$\begin{aligned} p(y^{new}|\mathbf{y}) &= \int p(y^{new}|\mu)p(\mu|\mathbf{y}^{old})d\mu = \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y^{new}-\mu)^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right)d\mu \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma^2)}}\exp\left(-\frac{(y^{new}-\mu_1)^2}{2(\sigma_1^2+\sigma^2)}\right) \end{aligned} \tag{12}$$

Thus, $y^{new}|\mathbf{y}^{old} \sim \text{Normal}\big(\mu_1, \sigma_1^2 + \sigma^2\big)^5$. Note that the variance of y^{new} is the sum of the variance for the distribution of y if μ were known, namely, σ^2, and the variance due to the uncertainty in μ.

2.1.3 Example: inference for μ

Assume that the amount of money spent during September on food by individual students, y, is Normally distributed with an unknown mean μ and known standard deviation, σ, of £50, or a precision, τ, of $1/50^2$ = 0.0004. You would like to estimate μ and will take a simple random sample of n=20 students.

Before doing so, you specify a Normal(μ_0, $\frac{50^2}{m}$) prior for μ. You think that the average is around μ_0 = £200 and set $\mu_0 = 200$. Further, you guess that the 25th and 75th percentiles are 150 to 250. Given that the corresponding standard normal percentiles are -0.6744898 and 0.6744898, you can solve for m using the following equation:

$$-0.6744898 = \frac{150 - 200}{50/\sqrt{m}}$$

Thus $m = 0.6744898^2 = 0.4549$. Then the prior for μ:

$$\mu \sim \text{Normal}\left(200, \frac{50^2}{0.4549}\right)$$

The simple random sample of n=20 was then taken and the average was £165.

The mean and variance for the posterior distribution for μ are based on (8):

$$E[\mu|\bar{y}] = \frac{0.4549 * 200 + 20 * 165}{0.4549 + 20} = 165.778$$
$$V[\mu|\bar{y}] = \frac{50^2}{0.4549 + 20} = 122.2199 = 11.05^2$$

and the posterior for μ is

$$\mu|\bar{y} \sim \text{Normal}\left(165.778, 11.05^2\right)$$

A weight of 0.4549/20.4549, about 2%, was given to the prior mean (and 98% to the sample mean) and the posterior standard deviation for μ went from 74 in the prior ($50/\sqrt{0.4549}$)) to 11.05 in the posterior.

Figure 1 shows the prior and posterior distributions for μ.

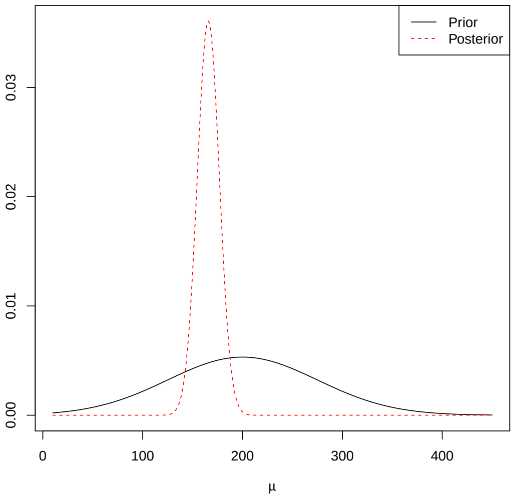
Posterior predictive distributions. If a student was randomly sampled, the prior and posterior predictive distributions for that student's food expenditures are:

$$\text{Prior } y^{new} \sim \text{Normal}\left(200, 50^2\right)$$
$$\text{Posterior } y^{new}|\bar{y}^{old} = 165 \sim \text{Normal}\left(165.778, 11.05^2 + 50^2 = 51.2^2\right)$$

Comparing the posterior to the prior, the mean has decreased by around £34 while the prediction variance of 51.2^2 is slightly larger than 50^2, with the additional variance due to the uncertainty in the value of μ.

[5] For details on the derivation see Lecture 3A: Supplement Posterior Predictive For Normal Dist'n on *Learn*.

Figure 1: Prior and posterior distribution for μ, average amount spent on food during September, assuming Normal(μ, 50^2) distribution.



2.2 Normal distribution: unknown τ or σ^2 and known μ

Again this is usually not a realistic situation, but the results are useful for more complex and realistic models.

2.2.1 Conjugate prior for τ

Before presenting results for σ^2, we begin with the precision, $\tau = 1/\sigma^2$.

We examine the likelihood for τ, keeping only terms that involve τ (see Eq'n 1).

$$f(\mathbf{y}|\mu,\tau) \equiv L(\tau|\mathbf{y},\mu) \propto (\tau)^{\frac{n}{2}} \exp\left(-\frac{\tau\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right) = (\tau)^{\frac{n}{2}} \exp\left(-\frac{\tau z^2}{2}\right) \tag{13}$$

where to reduce notation

$$z^2 = \sum_{i=1}^{n}(y_i-\mu)^2$$

The likelihood (13) is the kernel of a Gamma distribution. Thus the conjugate prior for τ when μ is known is

$$\tau \sim \text{Gamma}\left(\alpha,\beta\right) \tag{14}$$

and the posterior for τ:

$$\begin{aligned} \text{Posterior:} \quad \tau|\mathbf{y},\mu &\propto (\tau)^{\alpha-1}\exp\left(-\beta\tau\right)(\tau)^{\frac{n}{2}}\exp\left(-\frac{z^2\tau}{2}\right) \\ &= (\tau)^{\left(\alpha+\frac{n}{2}-1\right)}\exp\left(-(\beta+z^2/2)\tau\right) && (15) \\ &\Rightarrow && (16) \\ \tau|\mathbf{y},\mu &\sim \text{Gamma}\left(\alpha+\frac{n}{2},\beta+\frac{z^2}{2}\right) && (17) \end{aligned}$$

Comments.

- Recall that if $\theta \sim$ Gamma(α, β), then $E[\theta]$=α/β and $V[\theta] = \alpha/\beta^2$.
- Examining the posterior mean:

$$E[\tau|\mathbf{y}] = \frac{\alpha + n/2}{\beta + z^2/2} = \frac{\alpha}{\beta + z^2/2} + \frac{n/2}{\beta + z^2/2} = \frac{2\alpha}{2\beta + z^2} + \frac{n}{2\beta + z^2} \tag{18}$$

  Referring to $\frac{2\alpha}{2\beta+z^2}$, as n increases z^2 increases, thus the term goes to zero. Referring to $\frac{n}{2\beta+z^2}$ goes to $n/\sum_{i=1}^{n}(y_i - \mu)^2 = \frac{1}{\hat{\sigma}^2}$, where $\hat{\sigma}_2$ is the maximum likelihood estimate (mle) for σ^2. Thus as n increases $E[\tau|\mathbf{y}]$ approaches $1/\hat{\sigma}^2$ or $\hat{\tau}$, the mle for τ.

- One approach for selecting the hyperparameters for the Gamma dist'n prior is to specify approximate values for $E[\tau]$ and $V[\tau]$ and then solve for α and β. (Admittedly, specifying a value for $V[\tau]$ might be a little involved.)
- *An Aside: Re-expression as a χ^2 distribution.* The χ^2 distribution with ν degrees of freedom has the following pdf (see Appendix A of King and Ross's notes).

$$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{\frac{\nu}{2}-1} \exp\left(-\frac{\theta}{2}\right)$$

  Note that this is the same pdf as for Gamma$\left(\frac{\nu}{2}, \frac{1}{2}\right)$. Focusing on the kernel of the posterior for τ in (15):

$$\begin{aligned} p(\tau|\mathbf{y}, \mu) &\propto (\tau)^{\left(\alpha + \frac{n}{2} - 1\right)} \exp\left(-(\beta + z^2/2)\tau\right) \\ &\propto (2\beta + z^2)^{\frac{2\alpha+n}{2}-1} \tau^{\frac{2\alpha+n}{2}-1} \exp\left(-\frac{\tau(2\beta + z^2)}{2}\right) && (19) \\ &= (\tau(2\beta + z^2))^{\frac{2\alpha+n}{2}-1} \exp\left(-\frac{\tau(2\beta + z^2)}{2}\right) && (20) \end{aligned}$$

  where the multiplier $(2\beta + z^2)^{\frac{2\alpha+n}{2}-1}$ in (19) is a constant that does not affect the kernel. Then it can be seen that (20) is the kernel of a χ^2 distribution for $\tau(2\beta + z^2)$ with $2\alpha + n$ degrees of freedom:

$$\tau(2\beta + z^2) \sim \chi^2_{2\alpha+n} \tag{21}$$

  This can also be written as what is called a *scaled* χ^2 distribution[6]:

$$\tau \sim \frac{1}{2\beta + z^2} \chi^2_{2\alpha+n} \tag{22}$$

2.2.2 Conjugate prior for σ^2

Again examine the likelihood for σ^2 keeping only terms that involve σ^2.

$$f(\mathbf{y}|\mu, \sigma^2) \equiv L(\sigma^2|\mathbf{y}, \mu) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) = \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \tag{23}$$

The term on the right-hand side of Eq'n (23) is the kernel of an Inverse Gamma distribution. The pdf for an Inverse Gamma with parameters α and β (see Appendix A of King and Ross's notes):

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\theta)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\theta}\right) \tag{24}$$

[6] If a random variable θ multiplied by a constant c follows a distribution D, $c\theta \sim D$, then θ follows a scaled distribution $(1/c)D$.

Thus Inverse Gamma(α, β) is the conjugate prior for σ^2 when μ is known.

Then the posterior for σ^2:

$$\text{Posterior:}\quad \sigma^2|\mathbf{y}, \mu \propto (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right)$$

$$= (\sigma^2)^{-(\alpha+\frac{n}{2}+1)} \exp\left(-\frac{\beta + z^2/2}{\sigma^2}\right) \tag{25}$$

where eq'n 25 is the kernel for the Inverse Gamma:

$$\text{Posterior for } \sigma^2|\mu\text{:}\quad \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{z^2}{2}\right) \tag{26}$$

Comments.

- If $\theta \sim$ Inverse Gamma(α, β), $E[\theta] = \beta/(\alpha - 1)$ if $\alpha > 1$. and $V[\theta]$ is $\beta^2/((\alpha-1)^2(\alpha-2))$ if $\alpha > 2$. Thus, values for the hyperparameters, α and β, can be deduced given prior notions about the mean and the variance of σ^2.
- Relatively small values for α and β, e.g., 0.01 or 0.001, are often used in practice. Such choices are not universally considered a good idea. However, examining the sensitivity of the posterior to choices of α and β is good practice.
- As for τ, the posterior for σ^2 can be written as a scaled negative χ^2 distribution (See Appendix A of King and Ross).
- If $x \sim$ Gamma(α, β), then $y = 1/x \sim$ Inverse Gamma(α, β)[7].

2.2.3 Example: inference for σ^2

The construction timber called 2x4 has cross-piece dimensions of 1.5 inches by 3.5 inches on average. Let Y be the longer dimension and assume that $Y \sim$ Normal(μ, σ^2). The higher the quality control the closer the value of Y should be to 3.5 inches, in other words, σ^2 should be relatively small. A building company is considering purchasing timber from a new supplier but before doing so would like to see just how precisely the 2 by 4's are cut. They will take a random sample of n=10 2x4s and measure the length of the longer side. They are comfortable assuming that the average length μ is 3.5, thus the data distribution is $Y \sim$ Normal(3.5, σ^2).

Before taking the sample, they decide to use a Inverse Gamma(α, β) prior distribution for σ^2, and need to specify the hyperparameters. They would like to be cautious and will assume *a priori* that the average value of σ^2 is 0.05 with a variance of 0.02. This results in Inverse Gamma$(2.125, 0.05625)$ (check this). A simple random sample of n=10 2x4s yielded the following measurements:

```
3.527 3.387 3.466 3.382 3.612 3.680 3.471 3.475 3.603 3.680
```

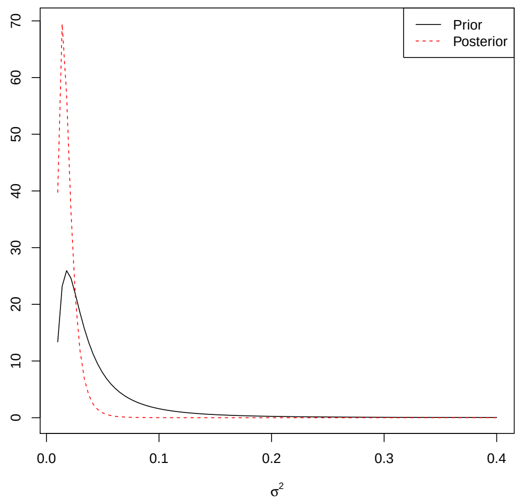
where $\sum_{i=1}^{10}(y_i - 3.5)^2 = 0.117997$. The posterior distribution for $\sigma^2|\mathbf{y}$:

$$\sigma^2|\mathbf{y} \sim \text{Inverse Gamma}\left(2.125 + \frac{10}{2}, 0.05624 + \frac{0.117997}{2}\right) = \text{Inverse Gamma}\left(7.125, 0.1152385\right)$$

Thus the posterior mean for σ^2 is $\frac{0.1152385}{7.125-1} = 0.0188$ and posterior variance is $\frac{0.1152385^2}{(7.125-1)^2*(7.125-2)} =$ 6.908e-05. Figure 2 plots the prior and posterior distributions for σ^2.

[7] This can be shown using the so-called change of variable theorem.

Figure 2: Prior and posterior distribution for σ^2, variance of the longer side of 2x4s, assuming sides are Normal(3.5, σ^2).



`R` does not have "built-in" Inverse Gamma distribution functions which could be used to calculate the quantiles of the posterior distribution for σ^2. However, we can use the quantile function for the Gamma distribution in `R`, namely qgamma which will yield quantiles for $\tau = 1/\sigma^2$, and invert the results. The posterior distribution for τ is $\Gamma(7.125, 0.1152385)$ and the 2.5 and 97.5 percentiles can be found with `qgamma(c(0.025,0.975),shape=7.125, rate=0.1152385)` = (25.10391, 114.81646). Thus

$$
\begin{aligned}
0.95 = \Pr(25.104 \leq \tau \leq 114.8) &= \Pr(25.104 \leq \frac{1}{\sigma^2} \leq 114.8) \\
&= \Pr\left(\frac{1}{114.8} \leq \sigma^2 \leq \frac{1}{25.104}\right) = \Pr\left(.00871 \leq \sigma^2 \leq 0.03983\right)
\end{aligned}
$$

Thus a 95% credible interval for σ^2 is (0.00871, 0.03983).

3 Bayes Theorem for multiple parameters

Often there will be $q > 1$ parameters:

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_q\}$$

Given data $\mathbf{y}$, Bayes theorem has the same form

$$\Pr(\Theta|\mathbf{y}) = \frac{\Pr(\mathbf{y}|\Theta)\Pr(\Theta)}{\Pr(\mathbf{y})} \propto \Pr(\mathbf{y}|\Theta)\Pr(\Theta)$$

3.1 Comments

- The posterior distribution for multiple parameters can be high dimensional, e.g., q parameters = q dimensions, and summarising a high dimensional space can be complicated.

- Often one dimensional summaries, namely marginal posterior distributions, are examined instead:

$$p(\theta_i|\mathbf{y}) = \int p(\Theta|\mathbf{y}) d\theta_1 d\theta_2 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_q$$

  This is the posterior distribution for θ_i found by "averaging" over all the other parameters, and thus "projecting" (collapsing) the q-dimensional posterior distribution on a single dimension. One can then examine a single posterior density plot, for example, calculate posterior mean, variance, and credible interval for θ_i.

- However, one dimensional summaries can fail to capture important features of the joint posterior.

- Two-dimensional graphical summaries are useful: contour plots or perspective plots, or in the case of samples from the posterior distribution, pairwise scatterplots.

- Two-dimensional numerical summaries include correlations or covariances.

- With more than two-dimensions, however, detecting patterns, if they exist, can be more difficult and complex.

3.2 Normal Dist'n: Unknown μ and σ^2

Without deriving any results, we discuss two approaches to the situation where both μ and σ^2 are unknown.

3.2.1 Joint prior constructed with independent marginal priors

One approach to arriving at a joint priors for both μ and σ^2 is to specify independent informative marginal priors for μ and σ^2 and multiply the two to yield a joint prior.

The joint posterior distribution:

$$\pi(\mu, \sigma^2|\mathbf{y}) = \frac{\pi(\mu)\pi(\sigma^2)\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right)}{\int\int \pi(\mu)\pi(\sigma^2)\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) \mathrm{d}\mu\ \mathrm{d}\sigma^2} \tag{27}$$

In general (27) will not be something that can be calculated analytically depending on the choices of $\pi(\mu)$ and $\pi(\sigma^2)$, because of the integral in the denominator. Numerical or simulation-based integration methods, which we will discuss later, can be used to yield approximate results.

Consider the 2x4 timber example, but now assume that both μ and σ^2 are unknown. Suppose one specified that the prior for μ is Lognormal(μ_0, σ_0^2) and the prior for σ^2 is Gamma(α, β). The denominator of (27) in this case:

$$\int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma_0^2}} \frac{1}{\mu} \exp\left(-\frac{(\ln(\mu)-\mu_0)^2}{2\sigma_0^2}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{\alpha-1} \exp(-\beta\sigma^2) \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) \mathrm{d}\mu\ \mathrm{d}\sigma^2$$

which is "probably" not analytically tractable (no closed form solution, I'm guessing as I've not tried to solve it).

Comment. While the joint prior distribution for μ and σ^2 was constructed by multiplying two independent marginal pdfs, the joint posterior distribution is *not* the product of two independent marginal pdfs. This is common: joint priors constructed as products of independent distributions do not usually yield a joint posterior that can be written as products of independent distributions.

3.2.2 Conjugate prior for μ and σ^2

There is a joint conjugate prior density for μ and σ^2. It is defined in terms of a marginal prior density for σ^2, which is an Inverse Gamma, and then a conditional prior density for μ *given* σ^2, which is a Normal where the variance hyperparameter is a function of σ^2. Namely,

$$\mu, \sigma^2 \sim \text{Inverse Gamma}(\alpha, \beta) \times N\left(\mu_0, \frac{\sigma^2}{\kappa}\right) \tag{28}$$

There are 4 hyperparameters, α, β, μ_0, and κ. As can be seen, conditional on the value of σ^2, prior uncertainty about the variance of μ around the expected value μ_0 increases as σ^2 increases and as κ decreases. Thus decreasing the values for α and β and the values for κ leads to a more dispersed prior for μ.

The resulting joint posterior density is then the product of an inverse gamma density and a normal density (conditional on σ^2):

$$\mu, \sigma^2 | \mathbf{y} \sim \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa n(\bar{y} - \mu_0)^2}{2(\kappa + n)}\right) \times \text{Normal}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{\sigma^2}{\kappa + n}\right) \tag{29}$$

The marginal density for σ^2 is Inverse Gamma, while the marginal density for μ conditional on σ^2 is a students' t distribution (see "Applied Bayesian Statistics", 2013, Cowles, M.K.).

3.3 Example: Multiple Parameter Inference

As a demonstration of multiple parameter inference, we fit a Bayesian linear regression of the lengths of dugongs[8] (sea cows, a type of marine mammal; see Figure 3) against their age in years. There were n=27 dugongs measured and the sampling model was the following.

$$Length_i \overset{iid}{\sim} \text{Normal}\left(\beta_0 + \beta_1 \ln(age), \sigma^2\right)$$

The relationship is shown in Figure 4.

Figure 3: Dugong (image from Wikipedia)

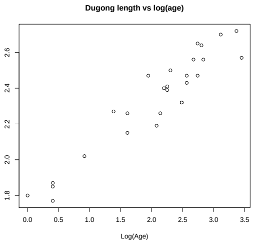


Figure 4: Dugong lengths plotted against log(age).

There are three parameters, $\Theta = \left\{\beta_0, \beta_1, \sigma^2\right\}$. The following independent marginal prior distributions were chosen to construct a joint prior distribution.

$$\beta_0, \beta_1 \sim \text{Uniform}\left(-50, 50\right), \quad \sigma^2 \sim \text{Uniform}\left(0.01, 20\right)$$

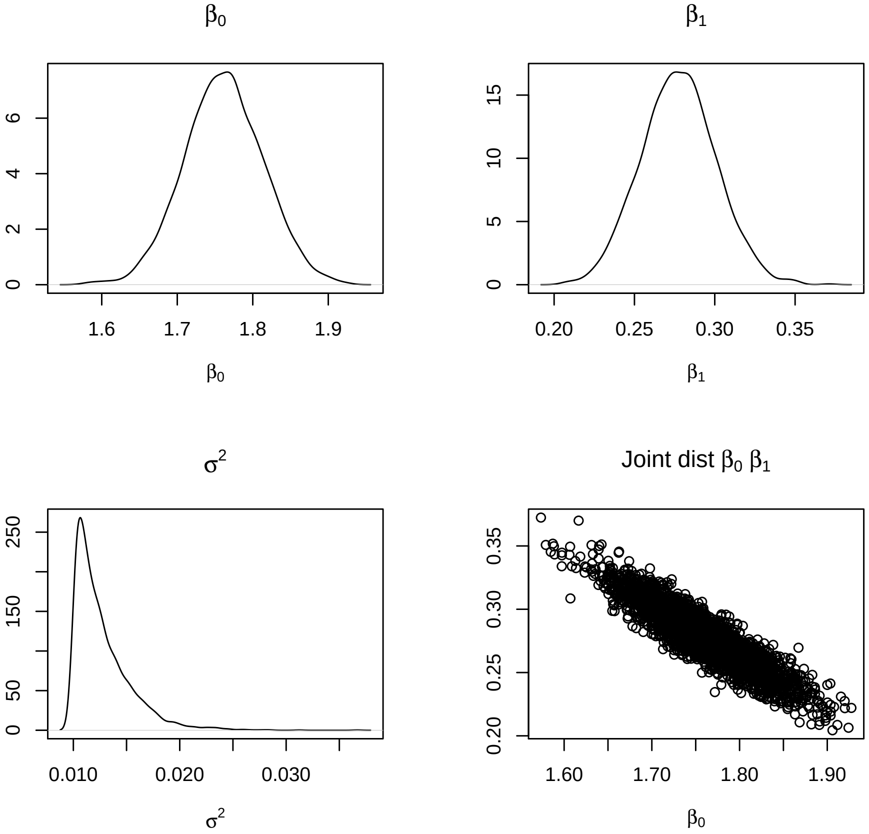
The resulting (estimated) posterior distribution was found using JAGS (code in the Appendix). The posterior means and standard deviation are shown below, along with the maximum likelihood estimates (except $\hat{\sigma}^2$ is bias-correction of the mle).

[8] Example motivated from Bayesian Methods for Data Analysis, 2009. Carlin and Louis.

	Bayesian		Frequentist	
	Mean	SD	mle	std error
β_0	1.76098	0.053089	1.762	0.0424
β_1	0.27757	0.023531	0.277	0.0188
σ^2	0.01277	0.002719	0.0081	

Figure 5 shows the marginal posterior distributions for the three parameters as well as a scatterplot of the samples of β_0 and β_1. The scatterplot shows that there is a negative relationship between β_0 and β_1.

Figure 5: Marginal posterior distributions for β_0, β_1, σ^2, and the joint distribution of β_0 and β_1



A R Code

A.1 Example: Posterior μ with known σ^2

```
#-- Posterior mu given known sigma
sigma <- 50; m <- 0.6744898^2; n <- 20; ybar <- 165; mu.0<-200
w <- n/(m+n)
post.E <- w*ybar+ (1-w)*mu.0;
post.V <- sigma^2/(m+n)
cat("w=",w,"1-w=",1-w,"post.E=",post.E, "post.V=",post.V,"post.sd=",sqrt(post.V),"\n")

#--plot prior and posterior
x.seq <- 10:450
y.prior <- dnorm(x.seq,mean=mu.0,sd=sigma/sqrt(m))
y.post  <- dnorm(x.seq,mean=post.E,sd=sqrt(post.V))
my.ylim <- range(c(y.prior,y.post))
plot(x.seq,y.prior,xlab=expression(mu),ylab="",ylim=my.ylim,type="l")
lines(x.seq,y.post,lty=2,col=2)
legend("topright",legend=c("Prior","Posterior"),col=1:2,lty=1:2)
```

A.2 Example: Posterior σ^2 with known μ

```
library(MCMCpack)  #this has dinvgamma() function

inv.gamma.param.calc <- function(mu,V) {
  alpha <- 2+mu^2/V
  beta  <- mu*(alpha-1)
  out <- list(alpha=alpha,beta=beta)
  return(out)
}
set.seed(742)
n <- 10
y <- rnorm(n=n,mean=3.5,sd=sqrt(0.01))
y <- round(y,3)
sse <- sum((y-3.5)^2)
cat("sse=",sse,"\n")

temp <- inv.gamma.param.calc(0.05,0.02)
prior.alpha <- temp$alpha
prior.beta  <- temp$beta

post.alpha <- prior.alpha+n/2
post.beta <- prior.beta + sse/2

post.mean <- post.beta/(post.alpha-1)
post.var  <- post.beta^2/((post.alpha-1)^2*(post.alpha-2))

cat("Prior alpha and beta=",prior.alpha,prior.beta,"\n")
cat("Post alpha and beta=", post.alpha,post.beta,"\n")
cat("Post mean=",post.mean,"var=",post.var,"\n")

theta.seq <- seq(0.01,0.4,length=100)
prior.density <- dinvgamma(x=theta.seq, shape=prior.alpha, scale = prior.beta)
post.density <- dinvgamma(x=theta.seq, shape=post.alpha, scale = post.beta)
my.ylim <- range(c(prior.density,post.density))
plot(theta.seq,prior.density,type="l",xlab=expression(sigma^2),ylab="",ylim=my.ylim)
lines(theta.seq,post.density,col=2,lty=2)
legend("topright",legend=c("Prior","Posterior"),lty=1:2,col=1:2)

# 95% credible interval
x <- qgamma(c(0.025,0.975),shape=7.125, rate=0.1152385)
print(x)
print(1/x)
```

A.3 Example: Linear regression with Dugong data

The R code for fitting the Dugong data and the call to JAGS are shown below.

```
library(rjags)

dugong.data <-
  list(age = c( 1.0, 1.5, 1.5, 1.5, 2.5, 4.0, 5.0, 5.0, 7.0,
                8.0, 8.5, 9.0, 9.5, 9.5, 10.0, 12.0, 12.0, 13.0,
                13.0, 14.5, 15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5),
       length = c(1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47,
                  2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43,
                  2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, 2.72, 2.57), n = 27)

log.age <- log(dugong.data$age)
plot(dugong.data$length ~ log.age,xlab="Log(Age)",ylab="",main="Dugong length vs log(age)")

# Initial values for running 3 MCMC chains in JAGS
num.chains   <- 3
beta0.set    <- c(-1,0,1)
beta1.set    <- c(-1,0,1)
sigma2.set   <- c(3,10,15)
dugong.inits <- list()
for(i in 1:num.chains) {
  dugong.inits[[i]] <- list(beta0=beta0.set[i],beta1=beta1.set[i],
                            sigma2=sigma2.set[i])
}


dugong.model <-   "model {
 # data that will be read in are age and length and n
 #Hyperparameters
 beta.low    <- -50
 tau   <- 1/sigma2

 #priors
 beta0 ~ dunif(-50,50)
 beta1 ~ dunif(-50,50)
 sigma2 ~ dunif(0.01,20)

#Likelihood
for(i in 1:n) {
 logage[i] <- log(age[i])
 mu[i] <- beta0 + beta1 * logage[i]
 length[i] ~ dnorm(mu[i], tau)
 }
}"

set.seed(742)
burnin <- 2000
inference.length <- 10000
dugong.results.initial <- jags.model(file=textConnection(dugong.model),
                                     data=dugong.data, inits=dugong.inits,
                                     n.chains=num.chains)
update(dugong.results.initial, n.iter=burnin)
dugong.results.final <- coda.samples(model=dugong.results.initial,
          variable.names=c("beta0","beta1","sigma2"),
                       n.iter=inference.length,thin=10)
summary(dugong.results.final)

#--- for looking at the entire combined results convert the mcmc.list object to a data frame
dugong.results.df <- as.data.frame(as.matrix(dugong.results.final))
head(dugong.results.df)
par(mfrow=c(2,2),oma=c(0,0,3,0))
plot(density(dugong.results.df$beta0),xlab=expression(beta[0]),ylab="",
     main=expression(beta[0]))
```

```
plot(density(dugong.results.df$beta1),xlab=expression(beta[1]),ylab="",
     main=expression(beta[1]))
plot(density(dugong.results.df$sigma2),xlab=expression(sigma^2),ylab="",
     main=expression(sigma^2))
plot( dugong.results.df$beta0,dugong.results.df$beta1,xlab=expression(beta[0]),ylab="",
      main= expression(paste("Joint dist ",beta[0]," ",beta[1])))
par(mfrow=c(1,1))
#if(plot.it) dev.copy2pdf(file=paste0(output,"L5_F_dugong_posterior_plots.pdf"))

#Frequentist results
dugong.freq.lm <- lm(length ~ log.age,data=dugong.data)
summary(dugong.freq.lm)
```