MAS3902

Bayesian Inference

Semester 2, 2022–23

Dr. Lee Fawcett
School of Mathematics, Statistics & Physics

Course overview

You were introduced to the Bayesian approach to statistical inference in MAS2903. This module showed statistical analysis in a very different light to the frequentist approach used in other courses. The frequentist approach bases inference on the sampling distribution of (usually unbiased) estimators; as you may recall, the Bayesian framework combines information expressed as expert subjective opinion with experimental data. You have probably realised that the Bayesian approach has many advantages over the frequentist approach. In particular it provides a more natural way of dealing with parameter uncertainty and inference is far more straightforward to interpret.

Much of the work in this module will be concerned with extending the ideas presented in MAS2903 to more realistic models with many parameters that you may encounter in real life situations. These notes are split into four chapters:

- **Chapter 1** reviews some of the key results for Bayesian inference of single parameter problems studied in Stage 2.
- **Chapter 2** studies the case of a random sample from a normal population and determines how to make inferences about the population mean and precision, and about future values from the population. The Group Project is based on this material.
- **Chapter 3** contains some general results for multi-parameter problems. You will encounter familiar concepts, such as how to represent *vague prior information* and the *asymptotic normal posterior distribution*.
- **Chapter 4** introduces *Markov chain Monte Carlo* techniques which have truly revolutionised the use of Bayesian inference in applications. Inference proceeds by simulating realisations from the posterior distribution. The ideas will be demonstrated using an R library specially written for the module. This material is extended in the 4th year Bayesian Statistics module.

Contents

1 Single parameter problems **1**

1.1 Prior and posterior distributions 1

1.2 Different levels of prior knowledge 10

1.2.1 Substantial prior knowledge 10

1.2.2 Limited prior knowledge 10

1.2.3 Vague prior knowledge/prior ignorance 10

1.3 The asymptotic posterior distribution 14

1.4 Bayesian inference 15

1.4.1 Estimation 15

1.4.2 Prediction 19

1.5 Sequential Bayesian updating 23

1.6 Learning objectives 27

2 Inference for a normal population **29**

2.1 Bayes Theorem for many parameters 29

2.2 Prior to posterior analysis 31

2.2.1 Marginal distributions 33

2.3 Confidence intervals and regions 38

2.4 Predictive distribution 41

2.5 Summary 43

2.6 Why do we have so many different distributions? 44

2.7 Learning objectives 45

3 General results for multi-parameter problems **47**

3.1 Different levels of prior knowledge 47

3.2 Asymptotic posterior distribution 52

3.3 Learning objectives 55

4 Non-conjugate multi-parameter problems **57**

4.1 Why is inference not straightforward in non-conjugate problems? 57

4.2 Simulation-based inference 59

4.3 Motivation for MCMC methods 62

4.4 The Gibbs sampler 62

4.4.1 Processing output from a Gibbs sampler 63

4.4.2 Bayesian inference using a Gibbs sampler 69

4.5 Metropolis-Hastings sampling 81

4.5.1 Symmetric chains (Metropolis method) 82

4.5.2 Independence chains 86

4.6 Hybrid methods 86

4.6.1 Componentwise transitions 86

4.6.2 Metropolis within Gibbs 87

4.7 Summary 92

4.8 Learning objectives 93

A Summary of distributions **95**

A.1 Distributions for data 95

A.2 Distributions for prior beliefs 99

Exercises **101**

The R package `nclbayes` **107**

Chapter 1

Single parameter problems

This chapter reviews some of the key results for Bayesian inference of single parameter problems studied in MAS2903.

1.1 Prior and posterior distributions

Suppose we have data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ which we model using the probability (density) function $f(\boldsymbol{x}|\theta)$, which depends on a single parameter θ. Once we have observed the data, $f(\boldsymbol{x}|\theta)$ is the *likelihood function* for θ and is a function of θ (for fixed $\boldsymbol{x}$) rather than of $\boldsymbol{x}$ (for fixed θ).

Also, suppose we have prior beliefs about likely values of θ expressed by a probability (density) function $\pi(\theta)$. We can combine both pieces of information using the following version of Bayes Theorem. The resulting distribution for θ is called the posterior distribution for θ as it expresses our beliefs about θ *after* seeing the data. It summarises all our current knowledge about the parameter θ.

Using Bayes Theorem, the posterior probability (density) function for θ is

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})}$$

where

$$f(\boldsymbol{x}) = \begin{cases} \int_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta)\, d\theta & \text{if } \theta \text{ is continuous,} \\ \\ \sum_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta) & \text{if } \theta \text{ is discrete.} \end{cases}$$

Also, as $f(\boldsymbol{x})$ is not a function of θ, Bayes Theorem can be rewritten as

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta) \times f(\boldsymbol{x}|\theta) \\ i.e.\ \ \text{Posterior} &\propto \text{Prior} \times \text{Likelihood.} \end{aligned}$$

Example 1.1

Table 1.1 shows some data on the number of cases of foodbourne botulism in England and Wales. It is believed that cases occur at random at a constant rate θ in time (a Poisson process) and so can be modelled as a random sample from a Poisson distribution with mean θ.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Cases	2	0	0	0	1	0	2	1

Table 1.1: Number of cases of foodbourne botulism in England and Wales, 1998–2005

An expert in the epidemiology of similar diseases was consulted to construct a suitable prior distribution for the rate θ.

(a) What might be an appropriate probability model for the rate θ? Why is this termed the *prior distribution*?

A gamma distribution might be appropriate here as θ is a rate and must be positive; the gamma distribution has support over the positive real line. This is termed the prior distribution as it represents our beliefs about θ before observing any data.

(b) Suppose the epidemiologist tells you that she expects around two cases of foodbourne botulism per year in England and Wales. After further discussions with the epidemiologist, you are also able to elicit that $SD(\theta) \approx \sqrt{2}$. Using your answer to part (a) find your full prior distribution for θ, and write down its density.

Suppose our prior is $\theta \sim Ga(g, h)$. Then from the information given, we can use the formulae for the gamma mean and standard deviation to write down:

$$E(\theta) = \frac{g}{h} = 2 \quad \text{(using the expert's most likelt value for } \theta\text{)}$$

and

$$SD(\theta) = \frac{\sqrt{g}}{h} = \sqrt{2}.$$

Solving simulataneously for g and h gives $g = 2$ and $h = 1$, and so we have $\theta \sim Ga(2, 1)$. Therefore:

$$\pi(\theta) = \frac{1^2\theta^{2-1}e^{-1\theta}}{\Gamma(2)} = \propto \theta e^{-\theta}, \quad \theta > 0.$$

(c) Determine the posterior distribution for θ.

✎

We know that

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}.$$

We know from the last question that the prior density is $\pi(\theta) \propto \theta e^{-\theta}$. What about the Poisson likelihood? From the data provided, we can write down that

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \frac{e^{-\theta}\theta^2}{2!} \times \frac{e^{-\theta}\theta^0}{0!} \times \cdot \times \frac{e^{-\theta}\theta^1}{1!} \\ &\propto e^{-8\theta}\theta^6. \end{aligned}$$

Therefore

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta) \times f(\boldsymbol{x}|\theta) \\ &\propto \theta e^{-\theta} \times \theta^6 e^{-8\theta} \\ &\propto \theta^7 e^{-9\theta} \quad \theta > 0 \end{aligned}$$

and we recognise this as being proportional to the density of a $Ga(8, 9)$ distribution; that is, $\theta|\boldsymbol{x} \sim Ga(8, 9)$.

Thus the data have updated our beliefs about θ from a $Ga(2, 1)$ distribution to a $Ga(8, 9)$ distribution. Plots of these distributions are given in Figure 1.1, and Table 1.2 gives a summary of the main changes induced by incorporating the data — a $Ga(g, h)$ distribution has mean g/h, variance g/h^2 and mode $(g-1)/h$.

Notice that, as the mode of the likelihood function is close to that of the prior distribution, the information in the data is consistent with that in the prior distribution. Also there is a reduction in variability from the prior to the posterior distributions. The similarity between the prior beliefs and the data has reduced the uncertainty we have about the rate θ at which cases occur.

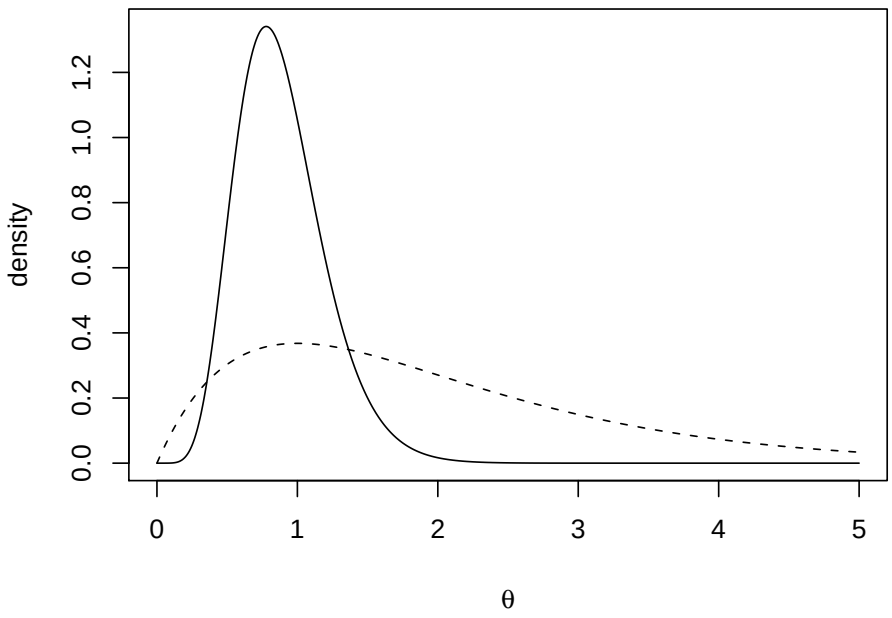


Figure 1.1: Prior (dashed) and posterior (solid) densities for θ

	Prior	Likelihood	Posterior
$Mode(\theta)$	1.00	0.75	0.78
$E(\theta)$	2.00	–	0.89
$SD(\theta)$	1.41	–	0.31

Table 1.2: Changes in beliefs about θ

Example 1.2

Consider now the general case of Example 1.1: suppose $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g\, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{1.1}$$

Determine the posterior distribution for θ.

Solution to Example 1.2 continued...

The likelihood function for θ is

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}, \qquad \theta > 0 \\ &\propto \theta^{n\bar{x}} e^{-n\theta}, \qquad \theta > 0. \end{aligned} \tag{1.2}$$

Using Bayes Theorem, the posterior density function is

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\ &\propto \frac{h^g\, \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^{n\bar{x}} e^{-n\theta}, \qquad \theta > 0 \\ \text{i.e.} \quad \pi(\theta|\boldsymbol{x}) &= k\theta^{g+n\bar{x}-1} e^{-(h+n)\theta}, \qquad \theta > 0 \end{aligned} \tag{1.3}$$

where k is a constant that does not depend on θ. Therefore, the posterior density takes the form $k\theta^{G-1}e^{-H\theta}$, $\theta > 0$ and so the posterior must be a gamma distribution. Thus we have $\theta|\boldsymbol{x} \sim Ga(G = g + n\bar{x}, H = h + n)$.

Summary:

If we have a random sample from a $Po(\theta)$ distribution and our prior beliefs about θ follow a $Ga(g, h)$ distribution then, after incorporating the data, our (posterior) beliefs about θ follow a $Ga(g + n\bar{x}, h + n)$ distribution.

The changes in our beliefs about θ are summarised in Table 1.3, taking $g \geq 1$. Notice

	Prior (1.1)	Likelihood (1.2)	Posterior (1.3)
$Mode(\theta)$	$(g-1)/h$	$\bar{x}$	$(g+n\bar{x}-1)/(h+n)$
$E(\theta)$	g/h	–	$(g+n\bar{x})/(h+n)$
$SD(\theta)$	$\sqrt{g}/h$	–	$\sqrt{g+n\bar{x}}/(h+n)$

Table 1.3: Changes in beliefs about θ

that the posterior mean is greater than the prior mean if and only if the likelihood mode is greater than the prior mean, that is,

$$E(\theta|\boldsymbol{x}) > E(\theta) \quad \Longleftrightarrow \quad Mode_\theta\{f(\boldsymbol{x}|\theta)\} > E(\theta).$$

The standard deviation of the posterior distribution is smaller than that of the prior distribution if and only if the sample mean is not too large, that is

$$SD(\theta|\boldsymbol{x}) < SD(\theta) \quad \Longleftrightarrow \quad Mode_\theta\{f(\boldsymbol{x}|\theta)\} < \left(2 + \frac{n}{h}\right) E(\theta),$$

and this will be true in large samples.

Example 1.3

Suppose we have a random sample from a normal distribution. In Bayesian statistics, when dealing with the normal distribution, the mathematics is more straightforward working with the precision (= 1/variance) of the distribution rather than the variance itself. So we will assume that this population has unknown mean μ but known precision τ: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), where τ is known. Suppose our prior beliefs about μ can be summarised by a $N(b, 1/d)$ distribution, with probability density function

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \tag{1.4}$$

Determine the posterior distribution for μ.

Hint:

$$d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2 = (d + n\tau)\left\{\mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau}\right)\right\}^2 + c$$

where c does not depend on μ.

The likelihood function for μ is

$$\begin{aligned} f(\boldsymbol{x}|\mu) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x_i - \mu)^2\right\} \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}. \end{aligned}$$

Now

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^{n}\{(x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu)\} \\ &= \sum_{i=1}^{n}\{(x_i - \bar{x})^2 + (\bar{x} - \mu)^2\} + 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \end{aligned}$$

Let $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and so

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n\left[s^2 + (\bar{x} - \mu)^2\right].$$

✎ *Solution to Example 1.3 continued...*

Therefore

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\}. \tag{1.5}$$

Using Bayes Theorem, the posterior density function is, for $\mu \in \mathbb{R}$

$$\begin{aligned}
\pi(\mu|\boldsymbol{x}) &\propto \pi(\mu)\, f(\boldsymbol{x}|\mu) \\
&\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu-b)^2\right\} \\
&\qquad\qquad \times \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[d(\mu-b)^2 + n\tau(\bar{x}-\mu)^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(d+n\tau)\left\{\mu - \left(\frac{db+n\tau\bar{x}}{d+n\tau}\right)\right\}^2 + c\right]\right\}
\end{aligned}$$

using the hint

$$\propto \exp\left\{-\frac{1}{2}\left[(d+n\tau)\left\{\mu - \left(\frac{db+n\tau\bar{x}}{d+n\tau}\right)\right\}^2\right]\right\}$$

as c does not depend on μ. Let

$$B = \frac{db+n\tau\bar{x}}{d+n\tau} \qquad \text{and} \qquad D = d + n\tau. \tag{1.6}$$

Then

$$\pi(\mu|\boldsymbol{x}) = k \exp\left\{-\frac{D}{2}(\mu-B)^2\right\}, \tag{1.7}$$

where k is a constant that does not depend on μ. Therefore, the posterior density takes the form $k\exp\{-D(\mu-B)^2/2\}$, $\mu \in \mathbb{R}$ and so the posterior distribution must be a normal distribution: we have $\mu|\boldsymbol{x} \sim N(B, 1/D)$.

Summary:

If we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known) and our prior beliefs about μ follow a $N(b, 1/d)$ distribution then, after incorporating the data, our (posterior) beliefs about μ follow a $N(B, 1/D)$ distribution.

The changes in our beliefs about μ are summarised in Table 1.4. Notice that the posterior

	Prior (1.4)	Likelihood (1.5)	Posterior (1.7)
$Mode(\mu)$	b	$\bar{x}$	$(db + n\tau\bar{x})/(d + n\tau)$
$E(\mu)$	b	–	$(db + n\tau\bar{x})/(d + n\tau)$
$Precision(\mu)$	d	–	$d + n\tau$

Table 1.4: Changes in beliefs about μ

mean is greater than the prior mean if and only if the likelihood mode (sample mean) is greater than the prior mean, that is

$$E(\mu|\boldsymbol{x}) > E(\mu) \quad \Longleftrightarrow \quad Mode_\mu\{f(\boldsymbol{x}|\mu)\} > E(\mu).$$

Also, the standard deviation of the posterior distribution is smaller than that of the prior distribution.

Example 1.4

The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in g/cm^3) are given below.

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.78	5.68	5.85	

Suppose that Cavendish asserts that the error standard deviation of these measurements is $0.2\,g/cm^3$, and assume that they are normally distributed with mean equal to the true earth density μ. Using a normal prior distribution for μ with mean $5.41\,g/cm^3$ and standard deviation $0.4\,g/cm^3$, derive the posterior distribution for μ.

✎

From the data we calculate $\bar{x} = 5.4848$ and $s = 0.1882$. Therefore, the assumed standard deviation $\sigma = 0.2$ is probably okay. We also have $\tau = 1/0.2^2$, $b = 5.41$, $d = 1/0.4^2$ and $n = 23$.

Solution to Example 1.4 continued...

Therefore, using Example 1.3, the posterior distribution is $\mu|\boldsymbol{x} \sim N(B, 1/D)$, where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} = \frac{5.41/0.4^2 + 23 \times 5.4848/0.2^2}{1/0.4^2 + 23/0.2^2} = 5.4840$$

and

$$D = d + n\tau = \frac{1}{0.4^2} + \frac{23}{0.2^2} = \frac{1}{0.0415^2}.$$

Therefore the posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and is shown in Figure 1.2.

The actual mean density of the earth is $5.515\, g/cm^3$ (Wikipedia). We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows. The posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

calculated using the R command `pnorm(5.615,5.484,0.0415)-pnorm(5.415,5.484,0.0415)`.

Without the data, the only basis for determining the earth's density is via the prior distribution. Here the prior distribution is $\mu \sim N(5.4, 0.4^2)$ and so the (prior) probability that the mean density is within 0.2 of the (now known) true value is

$$Pr(5.315 < \mu < 5.715) = 0.1896,$$

calculated using the R command `pnorm(5.615,5.4,0.4)-pnorm(5.415,5.4,0.4)`.

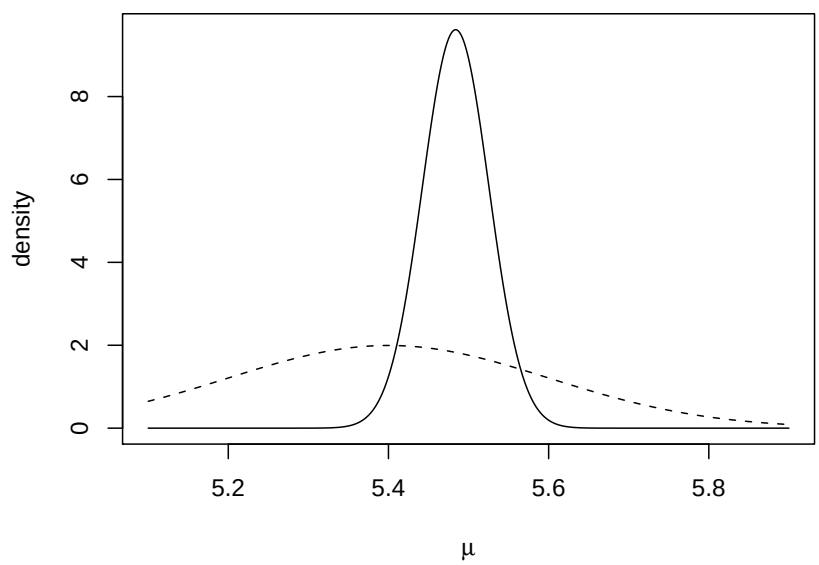


Figure 1.2: Prior (dashed) and posterior (solid) densities for the earth's density

1.2 Different levels of prior knowledge

1.2.1 Substantial prior knowledge

We have *substantial prior information* for θ when the prior distribution *dominates* the posterior distribution, that is $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$.

When we have substantial prior information there can be some difficulties:

1. the intractability of the mathematics in deriving the posterior distribution — though with modern computing facilities this is less of a problem,
2. the practical formulation of the prior distribution — coherently specifying prior beliefs in the form of a probability distribution is far from straightforward.

1.2.2 Limited prior knowledge

When prior information about θ is limited, the pragmatic approach is to choose a distribution which makes the Bayes updating from prior to posterior mathematically straightforward, and use what prior information is available to determine the parameters of this distribution. For example

- Poisson random sample, Gamma prior distribution $\longrightarrow$ Gamma posterior distribution
- Normal random sample (known variance), Normal prior distribution $\longrightarrow$ Normal posterior distribution

In these examples, the prior distribution and the posterior distribution come from the same family. This leads us to the following definition.

Definition 1.1

Suppose that data $\boldsymbol{x}$ are to be observed with distribution $f(\boldsymbol{x}|\theta)$. A family $\mathfrak{F}$ of prior distributions for θ is said to be *conjugate* to $f(\boldsymbol{x}|\theta)$ if for every prior distribution $\pi(\theta) \in \mathfrak{F}$, the posterior distribution $\pi(\theta|\boldsymbol{x})$ is also in $\mathfrak{F}$.

Notice that the conjugate family depends crucially on the model chosen for the data $\boldsymbol{x}$. For example, the only family conjugate to the model "random sample from a Poisson distribution" is the Gamma family.

1.2.3 Vague prior knowledge/prior ignorance

If we have very little or no prior information about the model parameters θ, we must still choose a prior distribution in order to operate Bayes Theorem. Obviously, it would

be sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance. In particular, most of the information about θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$.

We often represent vague prior knowledge by using a prior distribution which is conjugate to the model for $\boldsymbol{x}$ and which is as diffuse as possible, that is, has as large a variance as possible.

We may attempt to represent prior ignorance using a *Jeffreys prior*, specified in terms of Fisher's Information:

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right]. \tag{1.8}$$

Specifically, Jeffreys recommended that we represent prior ignorance by the prior distribution

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

Such a prior distribution is known as a Jeffreys prior distribution.

Example 1.5

In constructing the prior in Example 1.1, did we assume (i) substantial prior knowledge; (ii) limited prior knowledge; or (iii) prior ignorance?

✎

Limited prior knowledge – for mathematical convenience we used a conjugate prior, but we obtained the parameters in this prior through prior information collected from the epidemiologist.

Example 1.6

Suppose the parameter θ represents a probability. What might be the simplest prior to use here, if we are prior ignorant?

✎

Rather than compute a Jeffreys prior, we might simply choose a Uniform $U(0,1)$ prior here; that is, $\theta \sim U(0,1)$, with

$$\pi(\theta) = 1.$$

This would give all outcomes across the permissible range of values for θ equal density.

Example 1.7

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the posterior distribution assuming a vague prior for μ.

✎

The conjugate prior distribution is a normal distribution. We have already seen that if the prior is $\mu \sim N(b, 1/d)$ then the posterior distribution is $\mu|\boldsymbol{x} \sim N(B, 1/D)$ where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad D = d + n\tau.$$

If we now make our prior knowledge vague about μ by letting the prior variance tend to infinity ($d \to 0$), we obtain

$$B \to \bar{x} \qquad \text{and} \qquad D \to n\tau.$$

Therefore, assuming vague prior knowledge for μ results in a $N\{\bar{x}, 1/(n\tau)\}$ posterior distribution. Notice that the posterior mean is the sample mean (the likelihood mode) and that the posterior variance $1/(n\tau) \to 0$ as $n \to \infty$.

Example 1.8

Suppose we have a random sample from a Poisson distribution, that is, $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the posterior distribution assuming a vague prior for θ.

✎

The conjugate prior distribution is a Gamma distribution. Recall that a $Ga(g, h)$ distribution has mean $m = g/h$ and variance $v = g/h^2$. Rearranging these formulae we obtain

$$g = \frac{m^2}{v} \qquad \text{and} \qquad h = \frac{m}{v}.$$

Clearly $g \to 0$ and $h \to 0$ as $v \to \infty$ (for fixed m). We have seen how taking a $Ga(g, h)$ prior distribution results in a $Ga(g + n\bar{x}, h + n)$ posterior distribution. Therefore, taking a vague prior distribution will give a $Ga(n\bar{x}, n)$ posterior distribution. Note that the posterior mean is $\bar{x}$ (the likelihood mode) and that the posterior variance $\bar{x}/n \to 0$ and $n \to \infty$.

Example 1.9

Suppose we have a random sample from an exponential distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the Jeffreys prior for this model.

Recall that $L(\theta|\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \theta^n e^{-n\bar{x}\theta}$, and therefore

$$\begin{aligned}
\log L(\theta|\boldsymbol{x}) &= n \log \theta - n\bar{x}\theta \\
\Rightarrow \quad & \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x}) = \frac{n}{\theta} - n\bar{x} \\
\Rightarrow \quad & \frac{\partial^2}{\partial\theta^2} \log L(\theta|\boldsymbol{x}) = -\frac{n}{\theta^2} \\
\Rightarrow \quad & I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2} \log L(\theta|\boldsymbol{X})\right] = \frac{n}{\theta^2}.
\end{aligned}$$

Hence, the Jeffreys prior for this model is

$$\begin{aligned}
\pi(\theta) &\propto I(\theta)^{1/2} \\
&\propto \frac{\sqrt{n}}{\theta}, \qquad \theta > 0 \\
&\propto \frac{1}{\theta}, \qquad \theta > 0.
\end{aligned}$$

Notice that this distribution is improper since $\int_0^\infty d\theta/\theta$ is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Notice also that this density is, in fact, a limiting form of a $Ga(g, h)$ density (ignoring the integration constant) since

$$\frac{h^g\, \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \propto \theta^{g-1} e^{-h\theta} \to \frac{1}{\theta}, \qquad \text{as } g \to 0,\ h \to 0.$$

Therefore, we obtain the same posterior distribution whether we adopt the Jeffreys prior or vague prior knowledge.

1.3 The asymptotic posterior distribution

If we have a statistical model $f(\boldsymbol{x}|\theta)$ for data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, together with a prior distribution $\pi(\theta)$ for θ then

$$\sqrt{J(\hat{\theta})}\ (\theta - \hat{\theta})|\boldsymbol{x} \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{as } n \to \infty,$$

where $\hat{\theta}$ is the likelihood mode and $J(\theta)$ is the observed information

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x}|\theta).$$

This means that, with increasing amounts of data, the posterior distribution looks more and more like a normal distribution. The result also gives us a useful approximation to the posterior distribution for θ when n is large:

$$\theta|\boldsymbol{x} \sim N\{\hat{\theta}, J(\hat{\theta})^{-1}\} \qquad \text{approximately.}$$

Note that this limiting result is similar to one used in Frequentist statistics for the distribution of the maximum likelihood estimator, namely

$$\sqrt{I(\theta)}\ (\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{as } n \to \infty,$$

where Fisher's information $I(\theta)$ is the expected value of the observed information, where the expectation is taken over the distribution of $\boldsymbol{X}|\theta$, that is, $I(\theta) = E_{\boldsymbol{X}|\theta}[J(\theta)]$. You may also have seen this result written as an approximation to the distribution of the maximum likelihood estimator in large samples, namely

$$\hat{\theta} \sim N\{\theta, I(\theta)^{-1}\} \qquad \text{approximately.}$$

Example 1.10

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the asymptotic posterior distribution for μ.

Recall that

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\},$$

and therefore

$$\begin{aligned}
\log f(\boldsymbol{x}|\mu) &= \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\Rightarrow \quad \frac{\partial}{\partial\mu}\log f(\boldsymbol{x}|\mu) &= -\frac{\tau}{2} \times \sum_{i=1}^{n} -2(x_i - \mu) = \tau\sum_{i=1}^{n}(x_i - \mu) = n\tau(\bar{x} - \mu) \\
\Rightarrow \quad \frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) &= -n\tau \qquad \Rightarrow \quad J(\mu) = -\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) = n\tau.
\end{aligned}$$

✎... *Solution to Example 1.10*

We have

$$\begin{aligned} \frac{\partial}{\partial \mu} \log f(\boldsymbol{x}|\mu) = 0 \qquad &\Longrightarrow \qquad \hat{\mu} = \bar{x} \\ &\Longrightarrow \qquad J(\hat{\mu}) = n\tau \\ &\Longrightarrow \qquad J(\hat{\mu})^{-1} = \frac{1}{n\tau}. \end{aligned}$$

Therefore, for large n, the (approximate) posterior distribution for μ is

$$\mu|\boldsymbol{x} \sim N\left(\bar{x}, \frac{1}{n\tau}\right).$$

Here the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge.

1.4 Bayesian inference

The posterior distribution $\pi(\theta|\boldsymbol{x})$ summarises all our information about θ to date. However, sometimes it is helpful to reduce this distribution to a few key summary measures.

1.4.1 Estimation

Point estimates

There are many useful summaries for a typical value of a random variable with a particular distribution; for example, the mean, mode and median. The mode is used more often as a summary than is the case in frequentist statistics.

Confidence intervals/regions

A more useful summary of the posterior distribution is one which also reflects its variation. For example, a $100(1-\alpha)\%$ *Bayesian confidence interval* for θ is any region C_α that satisfies $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$. If θ is a continuous quantity with posterior probability density function $\pi(\theta|\boldsymbol{x})$ then

$$\int_{C_\alpha} \pi(\theta|\boldsymbol{x})\, d\theta = 1-\alpha.$$

The usual correction is made for discrete θ, that is, we take the largest region C_α such that $Pr(\theta \in C_\alpha|\boldsymbol{x}) \leq 1-\alpha$. Bayesian confidence intervals are sometimes called *credible regions* or *plausible regions*. Clearly these intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution.

A $100(1-\alpha)\%$ *highest density interval* (HDI) for θ is the region

$$C_\alpha = \{\theta : \ \pi(\theta|\boldsymbol{x}) \geq \gamma\}$$

where γ is chosen so that $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$. This region is sometimes called a *most plausible Bayesian confidence interval*. If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions. Also, if the posterior distribution is unimodal (has one mode) and symmetric about its mean then the HDI is an equi-tailed interval, that is, takes the form $C_\alpha = (a, b)$, where $Pr(\theta < a|\boldsymbol{x}) = Pr(\theta > b|\boldsymbol{x}) = \alpha/2$; see Figure 1.3.

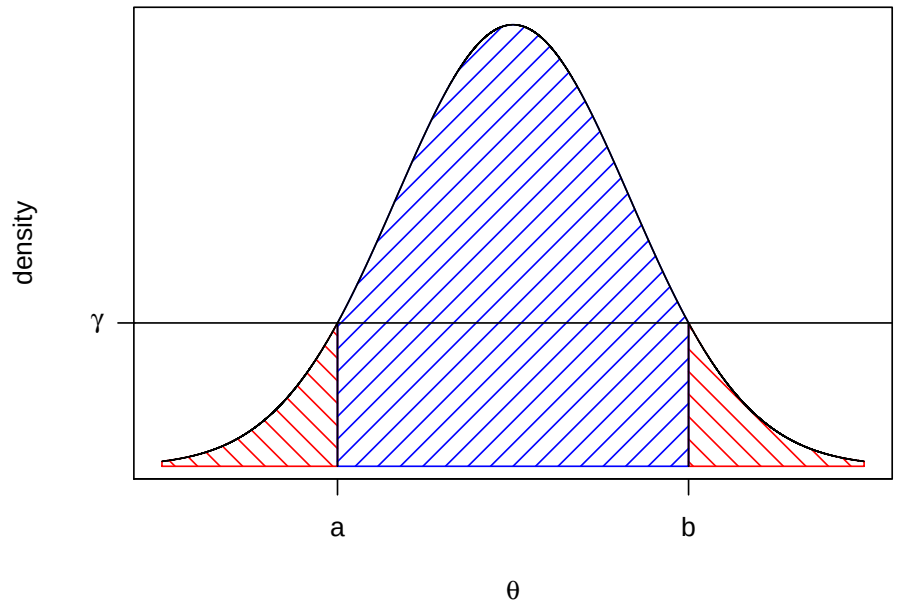


Figure 1.3: Construction of an HDI for a symmetric posterior density

Interpretation of confidence intervals/regions

Suppose C_B is a 95% Bayesian confidence interval for θ and C_F is a 95% frequentist confidence interval for θ. These intervals do not have the same interpretation:

- the probability that C_B contains θ is 0.95;
- the probability that C_F contains θ is either 0 or 1 — since θ does not have a (non-degenerate) probability distribution;
- the interval C_F covers the true value θ on 95% of occasions — in repeated applications of the formula.

Example 1.11

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where τ is known). We have seen that, assuming vague prior knowledge, the posterior distribution is $\mu|\boldsymbol{x} \sim N\{\bar{x}, 1/(n\tau)\}$. Determine the $100(1-\alpha)\%$ HDI for μ.

This distribution has a symmetric bell shape and so the HDI is an equi-tailed interval $C_\alpha = (a, b)$ with $Pr(\mu < a|\boldsymbol{x}) = \alpha/2$ and $Pr(\mu > b|\boldsymbol{x}) = \alpha/2$, that is,

$$a = \bar{x} - \frac{z_{\alpha/2}}{\sqrt{n\tau}} \qquad \text{and} \qquad b = \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n\tau}},$$

where z_α is the upper α-quantile of the $N(0,1)$ distribution. For example, the 95% HDI for μ is

$$\left(\bar{x} - \frac{1.96}{\sqrt{n\tau}}, \bar{x} + \frac{1.96}{\sqrt{n\tau}}\right).$$

Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

Example 1.12

Recall Example 1.1 on the number of cases of foodbourne botulism in England and Wales. The data were modelled as a random sample from a Poisson distribution with mean θ. Using a $Ga(2,1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8,9)$ (see Figure 1.5). Determine the $100(1-\alpha)\%$ HDI for θ.

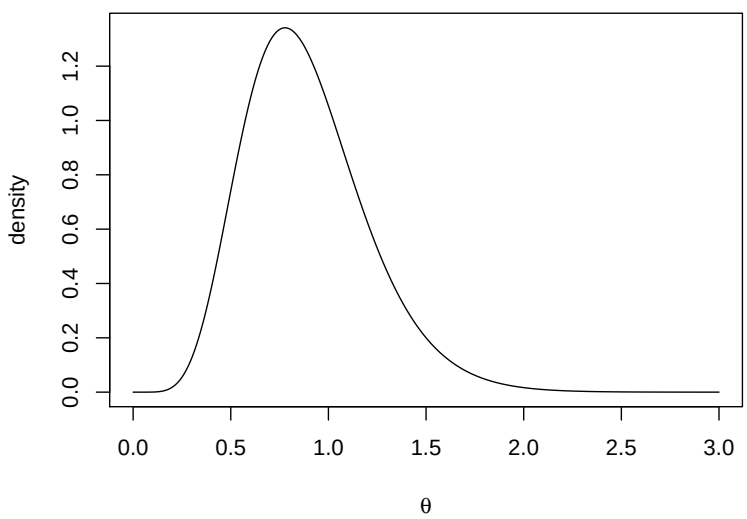


Figure 1.4: Posterior density for θ

✎... *Solution to Example 1.12*

The HDI must take the form $C_\alpha = (a, b)$ if it is to include the values of θ with the highest probability density. Suppose that $F(\cdot)$ and $f(\cdot)$ are the posterior distribution and density functions. Then the end-points a and b must satisfy

$$Pr(a < \theta < b|\boldsymbol{x}) = F(b) - F(a) = 1 - \alpha$$

and

$$\pi(\theta = a|\boldsymbol{x}) - \pi(\theta = b|\boldsymbol{x}) = f(a) - f(b) = 0.$$

Unfortunately, there is no simple analytical solution to these equations and so numerical methods have to be employed to determine a and b. However, if we have the quantile function $F^{-1}(\cdot)$ for $\theta|\boldsymbol{x}$ then we can find the solution by noticing that we can write b in terms of a:

$$b = F^{-1}\{1 - \alpha + F(a)\},$$

for $0 < a < F^{-1}(\alpha)$. Therefore, we can determine the correct choice of a by minimising the function

$$g(a) = \left(f(a) - f[F^{-1}\{1 - \alpha + F(a)\}]\right)^2.$$

within the range $0 < a < F^{-1}(\alpha)$. The values of a and b can be determined using the optimizer function `optimize` in `R`.

The `R` package `nclbayes` contains functions to determine the HDI for several distributions. The function for the Gamma distribution is `hdiGamma` and we can calculate the 95% HDI for the $Ga(8, 9)$ posterior distribution by using the commands

```
install.packages("nclbayes", repos="http://R-Forge.R-project.org")
library(nclbayes)
hdiGamma(p=0.95,a=8,b=9)
```

Taking $1 - \alpha = 0.95$ and using such `R` code gives $a = 0.3304362$ and $b = 1.5146208$. To check this answer, `R` gives $Pr(a < \theta < b|\boldsymbol{x}) = 0.95$, $\pi(\theta = b|\boldsymbol{x}) = 0.1877215$ and $\pi(\theta = a|\boldsymbol{x}) = 0.1877427$. Thus the 95% HDI is $(0.3304362, 1.514621)$.

The package also has functions `hdiBeta` for the Beta distribution and `hdiInvchi` for the Inv-Chi distribution (introduced in Chapter 2).

1.4.2 Prediction

Much of statistical inference (both Frequentist and Bayesian) is aimed towards making statements about a parameter θ. Often the inferences are used as a yardstick for similar future experiments. For example, we may want to predict the outcome when the experiment is performed again.

Clearly there will be uncertainty about the future outcome of an experiment. Suppose this future outcome Y is described by a probability (density) function $f(y|\theta)$. There are several ways we could make inferences about what values of Y are likely. For example, if we have an estimate $\hat{\theta}$ of θ we might base our inferences on $f(y|\theta = \hat{\theta})$. Obviously this is not the best we can do; such inferences ignore the fact that it is unlikely that $\theta = \hat{\theta}$.

Implicit in the Bayesian framework is the concept of the *predictive distribution*. This distribution describes how likely are different outcomes of a future experiment. The predictive probability (density) function is calculated as

$$f(y|\boldsymbol{x}) = \int_{\Theta} f(y|\theta)\, \pi(\theta|\boldsymbol{x})\, d\theta$$

when θ is a continuous quantity. From this equation, we can see that the predictive distribution is formed by weighting the possible values of θ in the future experiment $f(y|\theta)$ by how likely we believe they are to occur $\pi(\theta|\boldsymbol{x})$.

If the true value of θ were known, say θ_0, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$. However, as (generally) θ is unknown, the predictive distribution is used as the next best alternative.

We can use the predictive distribution to provide a useful range of plausible values for the outcome of a future experiment. This *prediction interval* is similar to a HDI interval. A $100(1-\alpha)\%$ *prediction interval* for Y is the region $C_\alpha = \{y : \ f(y|\boldsymbol{x}) \geq \gamma\}$ where γ is chosen so that $Pr(Y \in C_\alpha|\boldsymbol{x}) = 1 - \alpha$.

Example 1.13

Recall Example 1.1 on the number of cases of foodbourne botulism in England and Wales. The data for 1998–2005 were modelled by a Poisson distribution with mean θ. Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8, 9)$. Determine the predictive distribution for the number of cases for the following year (2006).

✎

Suppose the number of cases in 2006 is Y, with $Y|\theta \sim Po(\theta)$. The predictive probability function of Y is, for $y = 0, 1, \ldots$

$$f(y|\boldsymbol{x}) = \int_{\Theta} f(y|\theta)\, \pi(\theta|\boldsymbol{x})\, d\theta = \int_0^\infty \frac{\theta^y e^{-\theta}}{y!} \times \frac{9^8 \theta^7 e^{-9\theta}}{\Gamma(8)}\, d\theta$$

✎... *Solution to Example 1.13*

$$
\begin{aligned}
&= \frac{9^8}{y!\Gamma(8)} \int_0^\infty \theta^{y+7} e^{-10\theta}\, d\theta \\
&= \frac{9^8}{y!\Gamma(8)} \times \frac{\Gamma(y+8)}{10^{y+8}} \\
&= \frac{(y+7)!}{y!7!} \times 0.9^8 \times 0.1^y \\
&= \binom{y+7}{7} \times 0.9^8 \times 0.1^y.
\end{aligned}
$$

You may not recognise this probability function but it is related to that of a negative binomial distribution. Suppose $Z \sim NegBin(r, p)$ with probability function

$$Pr(Z = z) = \binom{z-1}{r-1} p^r (1-p)^{z-r}, \quad z = r, r+1, \ldots.$$

Then $W = Z - r$ has probability function

$$Pr(W = w) = Pr(Z = w + r) = \binom{w+r-1}{r-1} p^r (1-p)^w, \quad w = 0, 1, \ldots.$$

This is the same probability function as our predictive probability function, with $r = 8$ and $p = 0.9$. Therefore $Y|\boldsymbol{x} \sim NegBin(8, 0.9) - 8$. Note that, unfortunately R also calls the distribution of W a negative binomial distribution with parameters r and p. To distinguish between this distribution and the $NegBin(r, p)$ distribution used above, we shall denote the distribution of W as a $NegBin_{\mathrm{R}}(r, p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$. Thus $Y|\boldsymbol{x} \sim NegBin_{\mathrm{R}}(8, 0.9)$.

We can compare this predictive distribution with a naive predictive distribution based on an estimate of θ. Here we shall base our naive predictive distribution on the maximum likelihood estimate $\hat{\theta} = 0.75$, that is, use the distribution $Y|\theta = \hat{\theta} \sim Po(0.75)$. Thus, the naive predictive probability function is

$$f(y|\theta = \hat{\theta}) = \frac{0.75^y\, e^{-0.75}}{y!}, \quad y = 0, 1, \ldots\ .$$

Numerical values for the predictive and naive predictive probability functions are given in Table 1.5.

Again, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses a degenerate posterior distribution $\pi^*(\theta|\boldsymbol{x})$ which essentially allows only one value: $Pr_{\pi^*}(\theta = 0.75|\boldsymbol{x}) = 1$ and standard deviation $SD_{\pi^*}(\theta|\boldsymbol{x}) = 0$. Note that the correct posterior standard deviation of θ is $SD_\pi(\theta|\boldsymbol{x}) =$

y	correct $f(y\|\boldsymbol{x})$	naive $f(y\|\theta=\hat{\theta})$
0	0.430	0.472
1	0.344	0.354
2	0.155	0.133
3	0.052	0.033
4	0.014	0.006
5	0.003	0.001
≥ 6	0.005	0.002

Table 1.5: Predictive and naive predictive probability functions

$\sqrt{8}/9 = 0.314$. Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

$$SD(Y|x=1) = \begin{cases} 0.994 & \text{using the correct } \pi(\theta|\boldsymbol{x}) \\ 0.866 & \text{using the naive } \pi^*(\theta|\boldsymbol{x}), \end{cases}$$

these values being calculated from $NegBin_{\mathbb{R}}(8, 0.9)$ and $Po(0.75)$ distributions.

Using the numerical table of predictive probabilities, we can see that $\{0, 1, 2\}$ is a 92.9% prediction set/interval. This is to be contrasted with the more "optimistic" calculation using the naive predictive distribution which shows that $\{0, 1, 2\}$ is a 95.9% prediction set/interval.

Candidate's formula

In the previous example, a non-trivial integral had to be evaluated. However, when the past data $\boldsymbol{x}$ and future data y are independent (given θ) and we use a conjugate prior distribution, another (easier) method can be used to determine the predictive distribution.

Using Bayes Theorem, the posterior density for θ given $\boldsymbol{x}$ and y is

$$\begin{aligned} \pi(\theta|\boldsymbol{x}, y) &= \frac{\pi(\theta) f(\boldsymbol{x}, y|\theta)}{f(\boldsymbol{x}, y)} \\ &= \frac{\pi(\theta) f(\boldsymbol{x}|\theta) f(y|\theta)}{f(\boldsymbol{x}) f(y|\boldsymbol{x})} \quad \text{since } \boldsymbol{X} \text{ and } Y \text{ are independent given } \theta \\ &= \frac{\pi(\theta|\boldsymbol{x})\, f(y|\theta)}{f(y|\boldsymbol{x})}. \end{aligned}$$

Rearranging, we obtain

$$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)}.$$

This is known as Candidate's formula. The right-hand-side of this equation looks as if it depends on θ but, in fact, any terms in θ will be cancelled between the numerator and denominator.

Example 1.14

Rework Example 1.13 using Candidate's formula to determine the number of cases in 2006.

✎Let Y denote the number of cases in 2006. We know that $\theta|\boldsymbol{x} \sim Ga(8, 9)$ and $Y|\theta \sim Po(\theta)$. Using Example 1.2 we obtain

$$\theta|\boldsymbol{x}, y \sim Ga(8 + y, 10).$$

Therefore the predictive probability function of Y is, for $y = 0, 1, \ldots$

$$\begin{aligned} f(y|\boldsymbol{x}) &= \frac{f(y|\theta)\,\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)} \\ &= \frac{\dfrac{\theta^y\, e^{-\theta}}{y!} \times \dfrac{9^8\theta^7 e^{-9\theta}}{\Gamma(8)}}{\dfrac{10^{8+y}\theta^{7+y} e^{-10\theta}}{\Gamma(8+y)}} \\ &= \frac{\Gamma(8+y)}{y!\Gamma(8)} \times \frac{9^8}{10^{8+y}} \\ &= \frac{(y+7)!}{y!7!} \times 0.9^8 \times 0.1^y \\ &= \binom{y+7}{7} \times 0.9^8 \times 0.1^y. \end{aligned}$$

1.5 Sequential Bayesian updating

Throughout MAS2903, and this review chapter in the current module so far, we have focussed on using Bayes Theorem to perform an update of our beliefs about the parameter θ after having observed a single sample of data $\boldsymbol{x}$.

In many real-world problems in data science, new samples of data are being observed continuously; hence, it might be reasonable to assume our beliefs about θ could also be continuously updated. In such a setting, once the first sample of data $\boldsymbol{x}^{(1)}$ has been used to update our beliefs about θ, the resulting posterior distribution $\pi(\theta|\boldsymbol{x}^{(1)})$ represents our current beliefs about θ; if we were to observe a new sample $\boldsymbol{x}^{(2)}$ we might then use $\pi(\theta|\boldsymbol{x}^{(1)})$ as our "new prior", and update once again to obtain the "new posterior" $\pi(\theta|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$.

In other words, we might replace the usual mantra

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

with

$$\text{Revised} \propto \text{Current} \times \text{New likelihood},$$

with the ability to update our likelihood function with each new sample of data. Why does this work?

Suppose we have data assumed drawn from a probability distribution with parameter θ.

- Before we make any observations at all, our prior beliefs about θ are represented by $\pi(\theta)$. We might refer to this as our "original prior".

- Now we observe a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$; our posterior is then

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \times f(\boldsymbol{x}|\theta) = \pi(\theta) \times \prod_{i=1}^{n} f(x_i|\theta),$$

  assuming our observations are a random sample.

- Now suppose we make a new observation x_{n+1} assumed from the same probability distribution as $\boldsymbol{x}$.

  - Before making this observation, our *current* beliefs about θ are represented by $\pi(\theta|\boldsymbol{x})$; re-label this $\pi_n(\theta)$, our "new prior" after having made the initial n observations.

  - Assuming the new observation is drawn from the same probability model as the sample $\boldsymbol{x}$, then the *new likelihood* has the same functional form as the likelihood for $\boldsymbol{x}$

– Therefore,

$$\begin{aligned}
\text{Revised} &\propto \text{Current} \times \text{New likelihood} \\
\pi_{n+1}(\theta|x_{n+1}) &\propto \pi_n(\theta) \times f(x_{n+1}|\theta) \\
&\propto \pi(\theta|\boldsymbol{x}) \times f(x_{n+1}|\theta) \\
&\propto \left\{ \pi(\theta) \times \prod_{i=1}^{n} f(x_i|\theta) \right\} \times f(x_{n+1}|\theta) \\
&\propto \pi(\theta) \times \prod_{i=1}^{n+1} (x_i|\theta) \\
&\propto \text{original prior} \times \text{full likelihood},
\end{aligned}$$

where the "full likelihood" just represents the likelihood of all our observations together, i.e. the vector $(x_1, x_2, \ldots, x_n, x_{n+1})^T$.

Example 1.15

Table 1.6 shows provides an update to table 1.1, giving the number of cases of foodbourne botulism in England and Wales between 2006 and 2010.

Use your solution to Example 1.1, and sequential Bayesian updating, to obtain your new posterior distribution for θ.

Year	2006	2007	2008	2009	2010
Cases	1	0	0	1	1

Table 1.6: Number of cases of foodbourne botulism in England and Wales, 2006–2010

✎Our current beliefs are represented by the "old posterior" – from Example 1.1, we found that $\theta|\boldsymbol{x} \sim Ga(8, 9)$, and so

$$\text{Current} = \frac{9^8 \theta^7 e^{-9\theta}}{7!} \propto \theta^7 e^{-9\theta}.$$

We also have

$$\begin{aligned}
\text{New likelihood} &= e^{-\theta}\theta^1 \times e^{-\theta}\theta^0 \times \cdots e^{-\theta}\theta^1 \\
&= e^{-5\theta}\theta^{1+0+\cdots+1} \propto e^{-5\theta}\theta^3.
\end{aligned}$$

Therefore

$$\begin{aligned}
\text{Revised} &\propto \theta^7 e^{-9\theta} \times \theta^3 e^{-5\theta} \\
&= \theta^{10} e^{-14\theta},
\end{aligned}$$

and so it must be the case that our revised posterior is $\theta|\boldsymbol{x} \sim Ga(11, 14)$. This gives the further updates as shown in Table 1.7.

From the information in Table 1.7, we see that – in light of this new sample of data – we are more sure about θ, with a further reduction in $SD(\theta)$. We can see this in the updated density plots in Figure 1.5, with a more peaked and concentrated posterior. Before having collected any data, and based on the epidemiologist's prior information only, in a ten year period we might have expected around 20 cases of foodbourne botulism (taking the expected value of around 2 per year); after the first sample of data, this reduces to around 18 cases, and now to just under 16 cases in light of the updated dataset.

	Prior: $Ga(2,1)$ (Original prior)	Posterior: $Ga(8,9)$ (New prior)	Revised: $Ga(11,14)$ (New posterior)
$Mode(\theta)$	1.00	0.78	0.71
$E(\theta)$	2.00	0.89	0.79
$SD(\theta)$	1.41	0.31	0.24

Table 1.7: Changes in beliefs about θ after second update

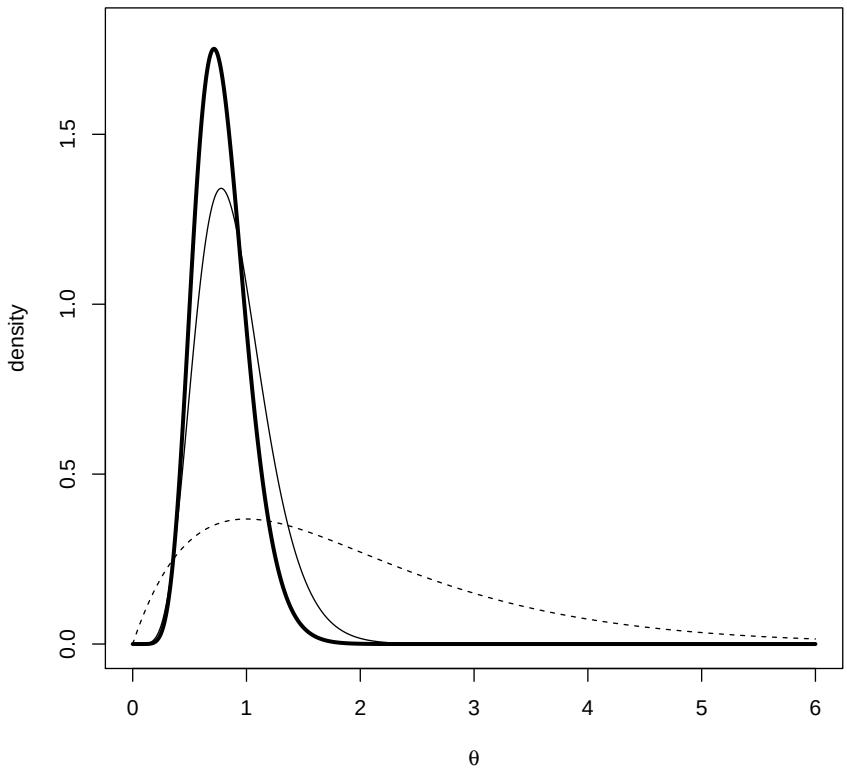


Figure 1.5: Prior (dashed) and posterior (solid) densities for θ, with the heavy line representing the posterior density after the second update

Example 1.16

Devise a generic sequential Bayesian updating plan for successive random samples from a Poisson distribution, assuming a conjugate prior specification.

Suppose our first sample is $\boldsymbol{x}^{(1)}$ with sample size n_1, that is, $\boldsymbol{x}^{(1)} = (x_1, x_2, \ldots, x_{n_1})^T$. From earlier examples, we know that the likelihood function is

$$f(\theta|\boldsymbol{x}^{(1)}) \propto e^{-n_1\theta}\theta^{n_1\bar{x}^{(1)}}.$$

Therefore

$$\pi(\theta|\boldsymbol{x}^{(1)}) \propto \theta^{g-1}e^{-h\theta} \times e^{-n_1\theta}\theta^{n_1\bar{x}^{(1)}} = \theta^{(g+n_1\bar{x}^{(1)})-1}e^{-(h+n_1)\theta},$$

and so $\theta|\boldsymbol{x}^{(1)} \sim Ga(g + n_1\bar{x}^{(1)}, h + n_1)$. Now suppose we have another sample $\boldsymbol{x}^{(2)}$ with sample size n_2; then

$$\begin{aligned}\pi(\theta|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) &\propto \theta^{(g+n_1\bar{x}^{(1)})-1}e^{-(h+n_1)\theta} \times e^{-n_2\theta}\theta^{n_2\bar{x}^{(2)}} \\ &\propto \theta^{(g+n_1\bar{x}^{(1)}+n_2\bar{x}^{(2)})-1}e^{-(h+n_1+n_2)\theta},\end{aligned}$$

giving $\theta|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \sim Ga(g + n_1\bar{x}^{(1)} + n_2\bar{x}^{(2)}, h + n_1 + n_2)$. Continuing this way, we can see that a generic sequential updating scheme would give rise to the posterior:

$$\theta|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots \sim Ga(g + n_1\bar{x}^{(1)} + n_2\bar{x}^{(2)} + \ldots, h + n_1 + n_2 + \ldots);$$

that is, our posterior will always be $Ga(G, H)$, and at each update j G will increase by $n_j\bar{x}^{(j)}$ and H will increase by n_j. The parameters in the posterior can be expressed recursively:

$$\begin{aligned}G_j &= G_{j-1} + n_j\bar{x}^{(j)} \quad \text{and} \\ H_j &= H_{j-1} + n_j.\end{aligned}$$

1.6 Learning objectives

By the end of this chapter, you should be able to:

- Determine the likelihood function using a random sample from **any** distribution
- Combine this likelihood function with **any** prior distribution to obtain the posterior distribution
- Name the posterior distribution if it is a "standard" distribution listed in these notes or on the exam paper – this list may well include distributions that are standard within the subject but which you have not met before. If the posterior distribution is not a "standard" distribution then it is okay just to give its density (or probability function) up to a constant.
- Do all the above for a particular data set or for a general case with random sample $x_1, \ldots, x_n$
- Describe the different levels of prior information; determine and use conjugate priors and vague priors
- Determine the asymptotic posterior distribution
- Determine the predictive distribution, particularly when having a random sample from any distribution and a conjugate prior via Candidate's formula
- Describe and calculate the confidence intervals, HDIs and prediction intervals
- Operate sequential Bayesian updating schemes

Chapter 2

Inference for a normal population

This chapter shows how to make inferences for the mean and variance of a normal population using a conjugate prior distribution. First we need the multi-parameter version of Bayes Theorem.

2.1 Bayes Theorem for many parameters

Suppose that now the probability (density) function we used to describe the data depends on many parameters, that is, $f(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. After observing the data, the likelihood function for $\boldsymbol{\theta}$ is $f(\boldsymbol{x}|\boldsymbol{\theta})$. Prior beliefs about $\boldsymbol{\theta}$ are represented through a probability (density) function $\pi(\boldsymbol{\theta})$. Therefore, using Bayes Theorem, the posterior probability (density) function for $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x})}$$

where

$$f(\boldsymbol{x}) = \begin{cases} \int_\Theta \pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta} & \text{if } \boldsymbol{\theta} \text{ is continuous,} \\ \\ \sum_\Theta \pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta} \text{ is discrete.} \end{cases}$$

As in Chapter 1, this can be rewritten as

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{x}) &\propto \pi(\boldsymbol{\theta}) \times f(\boldsymbol{x}|\boldsymbol{\theta}) \\ i.e.\ \ \text{posterior} &\propto \text{prior} \times \text{likelihood.} \end{aligned}$$

Next we introduce a new distribution which will be useful later on.

Example 2.1

If X has a generalised $t_a(b, c)$ distribution (see formula sheet) then show that $Y = (X - b)/\sqrt{c} \sim t_a \equiv t_a(0, 1)$.

Recall the general result: if X is a random variable with probability density function $f_X(x)$ and g is a bijective (1–1) function then the random variable $Y = g(X)$ has probability density function

$$f_Y(y) = f_X\left\{g^{-1}(y)\right\}\left|\frac{d}{dy}\,g^{-1}(y)\right|. \tag{2.1}$$

Comment

Values for the density function $f_Y(y)$ and the distribution function $F_Y(y)$ can be obtained by using the `R` functions `dgt` and `pgt` in the package `nclbayes`.

It is clear that $t_a(0, 1) \equiv t_a$ by examining their densities. Therefore, it makes sense to think of the t_a distribution as the standard t_a–distribution and make all calculations for the generalised $t_a(b, c)$ distribution from this standard distribution. The relationship between this standard and generalised version of the t-distribution is directly analogous to that between the standard normal $N(0, 1)$ distribution and its more general version: the $N(b, c)$ distribution. In both cases the relationship is one of location and scale:

$$Y \sim N(b, c) \quad \Longrightarrow \quad \frac{Y - b}{\sqrt{c}} \sim N(0, 1)$$

$$Y \sim t_a(b, c) \quad \Longrightarrow \quad \frac{Y - b}{\sqrt{c}} \sim t_a.$$

2.2 Prior to posterior analysis

Suppose we have a random sample from a normal distribution in which both the mean μ and the precision τ are unknown, that is, $X_i|\mu,\tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent). We shall adopt a (joint) prior distribution for μ and τ for which

$$\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right) \qquad \text{and} \qquad \tau \sim Ga(g, h)$$

for known values b, c, g and h. This distribution has density function

$$\begin{aligned}
\pi(\mu,\tau) &= \pi(\mu|\tau)\pi(\tau) \\
&= \left(\frac{c\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{c\tau}{2}(\mu-b)^2\right\} \times \frac{h^g \tau^{g-1} e^{-h\tau}}{\Gamma(g)}, \quad \mu \in \mathbb{R},\ \tau > 0 \\
&\propto \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2 + 2h\right]\right\}, \quad \mu \in \mathbb{R},\ \tau > 0. \qquad (2.2)
\end{aligned}$$

We will use the notation $NGa(b, c, g, h)$ for this distribution. Thus we take the prior distribution

$$\begin{pmatrix}\mu \\ \tau\end{pmatrix} \sim NGa(b, c, g, h).$$

Determine the posterior distribution for $\begin{pmatrix}\mu \\ \tau\end{pmatrix}$.

Hint:

$$c(\mu-b)^2 + n(\bar{x}-\mu)^2 = (c+n)\left\{\mu - \left(\frac{cb+n\bar{x}}{c+n}\right)\right\}^2 + \frac{nc(\bar{x}-b)^2}{c+n}.$$

✎

✎ *...Solution to Example 1.2*

2.2.1 Marginal distributions

Suppose $(\mu, \tau)^T \sim NGa(b, c, g, h)$. From the definition of the NGa distribution we know that $\tau \sim Ga(g, h)$. This also means that $\sigma = 1/\sqrt{\tau} \sim \textit{Inv-Chi}(g, h)$; see formula sheet.

The (marginal) density for μ is, for $\mu \in \mathbb{R}$

$$\begin{aligned}\pi(\mu) &= \int_0^\infty \pi(\mu, \tau)\, d\tau \\ &\propto \int_0^\infty \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2 + 2h\right]\right\} d\tau.\end{aligned}$$

Now, as the integral of a gamma density over its entire range is one, we have

$$\int_0^\infty \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)}\, d\theta = 1 \quad \Longrightarrow \quad \int_0^\infty \theta^{a-1} e^{-b\theta}\, d\theta = \frac{\Gamma(a)}{b^a}.$$

Therefore, for $\mu \in \mathbb{R}$

$$\begin{aligned}\pi(\mu) &\propto \int_0^\infty \tau^{g+\frac{1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[c(\mu-b)^2 + 2h\right]\right\} d\tau \\ &\propto \frac{\Gamma\left(g+\frac{1}{2}\right)}{[\{c(\mu-b)^2 + 2h\}/2\}]^{g+\frac{1}{2}}} \\ &\propto h^{-g-1/2}\left\{1 + \frac{c(\mu-b)^2}{2h}\right\}^{-g-1/2} \\ &\propto \left\{1 + \frac{c(\mu-b)^2}{2h}\right\}^{-\frac{2g+1}{2}}.\end{aligned}$$

Comparing this density with that of the generalised t–distribution (see formula sheet) gives

$$\mu \sim t_{2g}\left(b, \frac{h}{gc}\right). \tag{2.4}$$

Thus, marginally, the prior distribution for μ is a t–distribution.

Similar calculations can be used to determine the (marginal) posterior distributions.

Summary of marginal distributions

The prior $\begin{pmatrix}\mu \\ \tau\end{pmatrix} \sim NGa(b, c, g, h)$ has marginal distributions

- $\mu \sim t_{2g}\left(b, \frac{h}{gc}\right)$
- $\tau \sim Ga(g, h)$; also $\sigma = 1/\sqrt{\tau} \sim \textit{Inv-Chi}(g, h)$.

The posterior $\begin{pmatrix}\mu\\ \tau\end{pmatrix}\Bigg| \boldsymbol{x} \sim NGa(B, C, G, H)$ has marginal distributions

- $\mu|\boldsymbol{x} \sim t_{2G}\left(B, \frac{H}{GC}\right)$
- $\tau|\boldsymbol{x} \sim Ga(G, H)$; also $\sigma|\boldsymbol{x} \sim \textit{Inv-Chi}(G, H)$.

It can be shown that the posterior mean of μ is greater than its prior mean if and only if the sample mean (likelihood mode) is greater than its prior mean, that is,

$$E(\mu|\boldsymbol{x}) > E(\mu) \quad \Longleftrightarrow \quad \bar{x} > b.$$

The relationships between the prior and posterior variance of μ and mean and variance of τ and of σ are rather more complex.

Example 2.2

Recall Example 1.4 on the earth's density. Previously we assumed that the measurements followed a $N(\mu, 0.2^2)$ distribution, that is, the standard deviation of the measurements was known to be $0.2\,g/cm^3$. Now we consider the case where this standard deviation is unknown and determine posterior distributions using the theory in section 2.2.

Before we can proceed, we must specify the parameters in the $NGa(b, c, g, h)$ prior distribution for (μ, τ). In the previous analysis, we assumed that the population measurement precision was $\tau = 1/0.2^2 = 25$ and assumed a $N(5.41, 0.4^2)$ prior distribution for the population mean, that is, $\mu|\tau = 25 \sim N(5.41, 0.4^2)$.

Choice of b and c: the conditional prior distribution for μ is $\mu|\tau \sim N\{b, 1/(c\tau)\}$ and so matching the prior distributions for μ (when $\tau = 25$) gives $b = 5.41$ and $c = 0.25$.

Choice of g and h: the marginal prior distribution for τ is $\tau \sim Ga(g, h)$. Previously, we assumed $\tau = 25$ (with $Var(\tau) = 0$) and so take this value as the prior mean: $E(\tau) = 25$. Suppose we also decide that $Var(\tau) = 250$. These two requirements give $g = 2.5$ and $h = 0.1$. Therefore, we will assume the prior distribution

$$\begin{pmatrix}\mu\\ \tau\end{pmatrix} \sim NGa(5.41, 0.25, 2.5, 0.1).$$

We have seen that if $(\mu, \tau)^T \sim NGa(b, c, g, h)$ then the marginal distribution of μ is $\mu \sim t_{2g}\big\{b, h/(gc)\big\}$. Therefore, with this choice of prior distribution, the marginal prior distribution for μ is

$$\mu \sim t_5(5.41, 0.16).$$

Figure 2.1 shows the close match between the new (marginal) prior distribution for μ and that used previously.

Determine the posterior distribution for $(\mu, \tau)^T$. Also determine the marginal prior distribution for τ and for σ, and the marginal posterior distribution for each of μ, τ and σ.

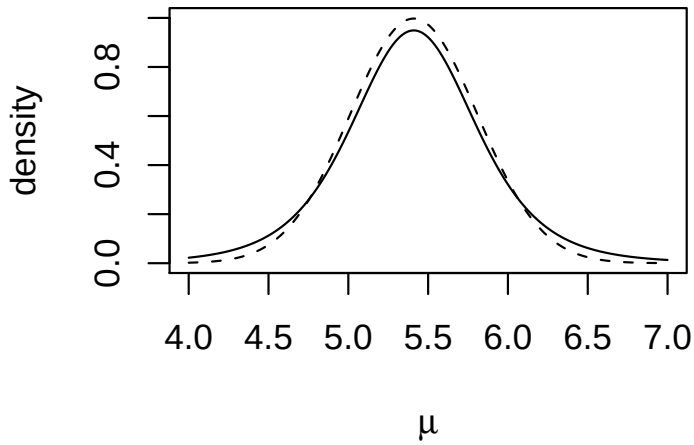


Figure 2.1: Marginal prior density for μ: new version (solid) and previous version (dashed)

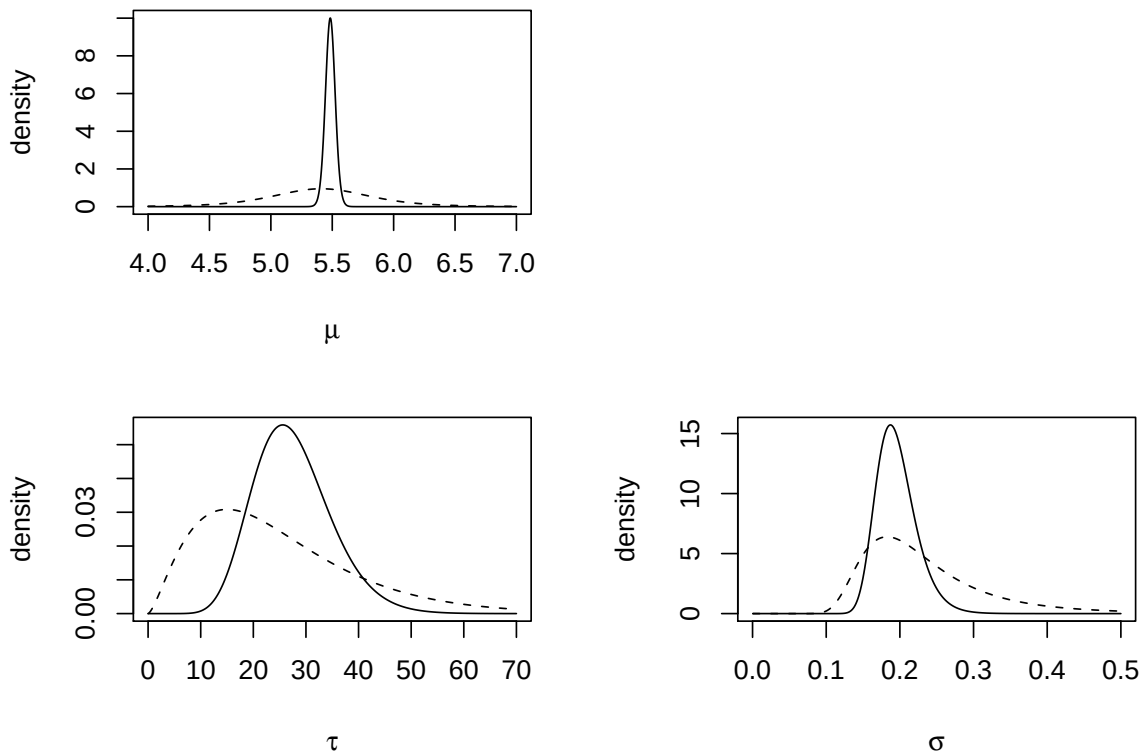


Figure 2.2: Prior (dashed) and posterior (solid) densities for μ, τ and σ

Plots of the (marginal) prior and posterior distributions of μ, τ and σ are given in Figure 2.2. Note that the (marginal) prior and posterior distributions for σ can be determined from that of τ. We can also examine the joint prior and posterior distributions for $(\mu, \tau)^T$ via the contour plots of their densities to see if there is any change in the dependence structure; see Figure 2.3. This figure is produced by using the `R` command `NGacontour` in the `nclbayes` package as follows:

```
mu=seq(4.5,6.5,len=1000)
tau=seq(0,71,len=1000)
NGacontour(mu,tau,b,c,g,h,lty=3)
NGacontour(mu,tau,B,C,G,H,add=TRUE)
```

in which the variables `b,c,g,h,B,C,G,H` have already been set to their prior/posterior values. A careful look at the values of the contour levels plotted shows that the highest contour level plotted for the prior density is 0.024 and the lowest level for the posterior density is 0.05. From this we can conclude that the posterior distribution is far more concentrated than the prior distribution. Also the contours for the posterior distribution are much more elliptical than those for the prior distribution. This indicates a change in the dependence structure. However, the main changes shown by the figure are in the mean and variability of μ and τ.

Wikipedia tells us that the actual mean density of the earth is $5.515\, g/cm^3$. We can determine the (posterior) probability that the mean density is within 0.1 of this value as

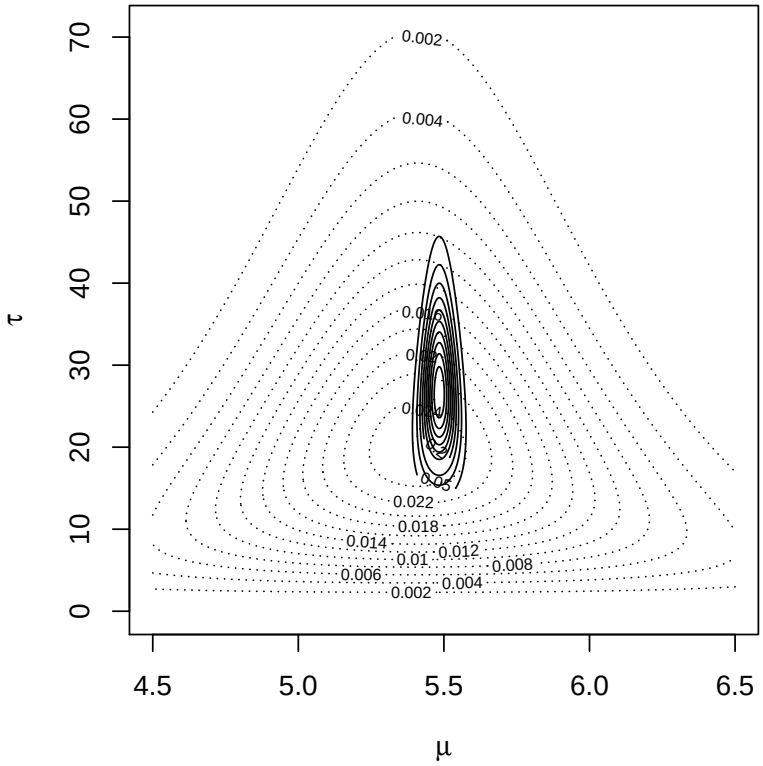


Figure 2.3: Contour plot of the prior (dashed) and posterior (solid) densities for $(\mu, \tau)^T$.

follows. We already know that $\mu|\boldsymbol{x} \sim t_{28}(5.484, 0.001561)$ and so we can calculate

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9529$$

using `pgt(5.615,28,5.484,0.001561)-pgt(5.415,28,5.484,0.001561)`.

Without the data, the only basis for determining the earth's density is via the prior distribution. Here the prior distribution is $\mu \sim t_5(5.41, 0.16)$ and so the (prior) probability that the mean density is within 0.1 of the (now known) true value is

$$Pr(5.415 < \mu < 5.615) = 0.1802,$$

calculated using `pgt(5.615,5,5.41,0.16)-pgt(5.415,5,5.41,0.16)`.

These probability calculations demonstrate that the data have been very informative and changed our beliefs about the earth's density.

2.3 Confidence intervals and regions

Example 2.3

Determine the $100(1-\alpha)\%$ highest density interval (HDI) for the population mean μ in terms of quantiles of the standard t-distribution.

	Prior	Posterior	
μ:	$(4.3818, 6.4382)$	$(5.4031, 5.5649)$	
τ:	$(1.4812, 55.9573)$	$(14.0193, 42.2530)$	$\leftarrow$ HDI
	$(4.1561, 64.1625)$	$(15.0674, 43.7625)$	
σ:	$(0.1062, 0.4246)$	$(0.1466, 0.2505)$	$\leftarrow$ HDI
	$(0.1248, 0.4905)$	$(0.1512, 0.2576)$	

Table 2.1: Prior and posterior 95% intervals for the analysis in Example 2.2

These intervals can be calculated easily using the `R` function `qgt` in the package `nclbayes`. For example, the prior and posterior 95% HDIs for μ can be calculated using

```
c(qgt(0.025,2*g,b,h/(g*c)),qgt(0.975,2*g,b,h/(g*c)))
c(qgt(0.025,2*G,B,H/(G*C)),qgt(0.975,2*G,B,H/(G*C)))
```

Determining a highest density interval (HDI) for the population precision τ or standard deviation σ is more complicated as their posterior distributions are not symmetric. The (marginal) posterior for τ is $\tau|\boldsymbol{x} \sim Ga(G, H)$ and the (marginal) posterior for σ is $\sigma|\boldsymbol{x} \sim$ *Inv-Chi*(G, H). HDIs can be found by using the `R` functions `hdiGamma` and `hdiInvchi` in the package `nclbayes`. More standard equi-tailed confidence intervals can be found using the functions `qgamma` and `qinvchi`.

For example, the prior and posterior 95% HDIs for τ can be calculated using `R` commands `hdiGamma(0.95,g,h)` and `hdiGamma(0.95,G,H)`, and those for σ using commands `hdiInvchi(0.95,g,h)` and `hdiInvchi(0.95,G,H)`. The 95% equi-tailed confidence intervals are calculated in a similar way to the HDIs for μ above. So for τ, the prior and posterior intervals are calculated using

```
c(qgamma(0.025,g,h),qgamma(0.975,g,h))
c(qgamma(0.025,G,H),qgamma(0.975,G,H))
```

and those for σ using

```
c(qinvchi(0.025,g,h),qinvchi(0.975,g,h))
c(qinvchi(0.025,G,H),qinvchi(0.975,G,H))
```

The numerical values for the prior and posterior 95% intervals for the analysis in Example 2.2 are given in Table 2.1. Notice that there is little difference between the posterior HDI and equi-tailed intervals for τ and for σ, whereas the prior intervals are fairly different. This is because the prior distributions are quite skewed but the posterior distributions are fairly symmetric; see Figure 2.2.

In Bayesian inference it can also be useful to determine (joint) confidence regions for several parameters, in this case, for $(\mu, \tau)^T$. In general this is a difficult problem to solve mathematically, and it is in this case.

Example 2.4

Determine a joint confidence region for $(\mu, \tau)^T$.

Using an additional argument in the R function NGacontour produces plots of confidence regions. For example

```
mu=seq(3.5,7.5,len=1000)
tau=seq(0,80,len=1000)
NGacontour(mu,tau,b,c,g,h,p=c(0.95,0.9,0.8),lty=3)
NGacontour(mu,tau,B,C,G,H,p=c(0.95,0.9,0.8),add=TRUE)
```

produces a plot containing the 95%, 90% and 80% prior and posterior confidence regions for $(\mu, \tau)^T$ for the prior and posterior distributions in Example 2.2; see Figure 2.4. The upper plot shows contours of both prior and posterior densities. The numbers within the plot are the contour levels. The largest prior confidence region is the 95% region. The next largest is the 90% prior confidence region and the smallest is the 80% prior confidence region. The same ordering holds for the posterior confidence regions. The posterior contours are so concentrated in the middle of the plot that there is no room to put in the contour levels. However, these can be see on the lower plot which also shows the contours but focuses the parameter range to highlight the contours of the posterior density. The values of the contours in this lower plot show that the posterior density is much more peaked, that is, the posterior has a much reduced variability. The location of the centre of the central contour for both the prior and posterior densities shows that there has been little change in the mean/mode.

2.4 Predictive distribution

Suppose we sample another value y randomly from the population. What values is it likely to take? This is described by its predictive distribution. We can determine this distribution by using the definition of the predictive density

$$f(y|\boldsymbol{x}) = \int f(y|\mu, \tau)\, \pi(\mu, \tau|\boldsymbol{x})\, d\mu\, d\tau$$

or by using Candidate's formula (as this is a conjugate analysis). However, for this model/prior, there is a more straightforward method to determine the predictive distribution in this model.

As Y is a random value from the population, we have that $Y|\mu, \tau \sim N(\mu, 1/\tau)$. We also know that the posterior distribution is $(\mu, \tau)^T|\boldsymbol{x} \sim NGa(B, C, G, H)$. Therefore, we can write

$$Y = \mu + \varepsilon,$$

where

$$\varepsilon|\tau \sim N(0, 1/\tau) \qquad \text{and} \qquad \mu|\boldsymbol{x}, \tau \sim N\left(B, \frac{1}{C\tau}\right).$$

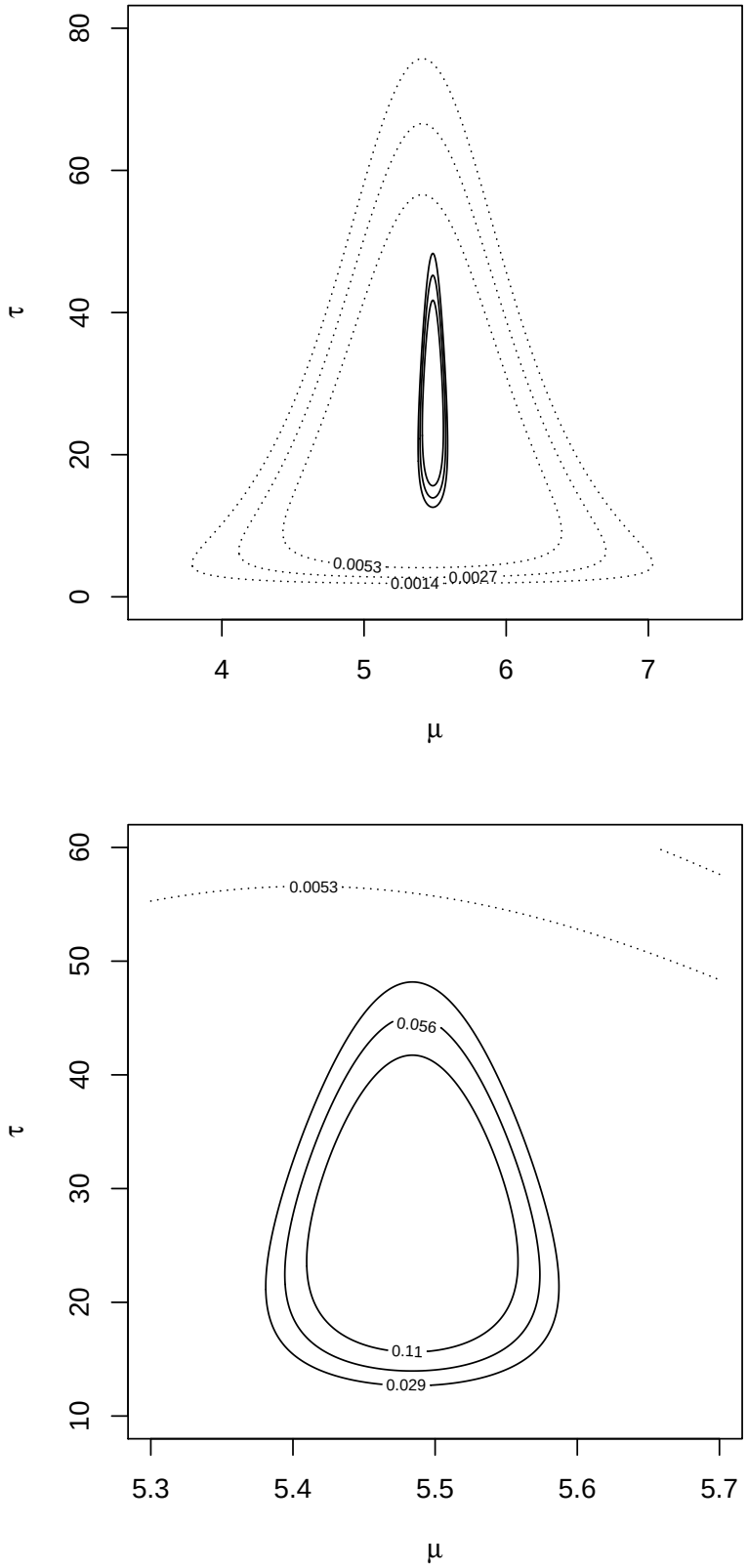


Figure 2.4: 95%, 90% and 80% prior (dashed) and posterior (solid) confidence regions for $(\mu, \tau)^T$

Hence Y is the sum of two independent normal random quantities, and so

$$Y|\boldsymbol{x},\tau \sim N\left(B,\ \frac{1}{\tau}+\frac{1}{C\tau}\right) \equiv N\left(B,\ \frac{C+1}{C\tau}\right).$$

Thus, as $\tau|\boldsymbol{x} \sim Ga(G,H)$

$$\begin{pmatrix} Y \\ \tau \end{pmatrix}\Bigg|\ \boldsymbol{x} \sim NGa\left(B, \frac{C}{C+1}, G, H\right)$$

and so, using (2.4)

$$Y|\boldsymbol{x} \sim t_{2G}\left\{B,\ \frac{H(C+1)}{GC}\right\}.$$

We can determine $100(1-\alpha)\%$ predictive intervals by noting that the predictive distribution is symmetric about its mean and therefore the HDI is

$$\left(B - t_{2G;\alpha/2}\sqrt{\frac{H(C+1)}{GC}},\ B + t_{2G;\alpha/2}\sqrt{\frac{H(C+1)}{GC}}\right).$$

These predictive intervals can be calculated easily using the R function `qgt`. For example, in Example 2.2, the prior and posterior predictive HDIs for a new value Y from the population are $(4.2604, 6.5596)$ and $(5.0855, 5.8825)$ respectively, calculated using

```
c(qgt(0.025,2*g,b,h*(c+1)/(g*c)),qgt(0.975,2*g,b,h*(c+1)/(g*c)))
c(qgt(0.025,2*G,B,H*(C+1)/(G*C)),qgt(0.975,2*G,B,H*(C+1)/(G*C)))
```

2.5 Summary

Suppose we have a normal random sample with $X_i|\mu,\tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent).

(i) $(\mu,\tau)^T \sim NGa(b,c,g,h)$ is a conjugate prior distribution.

(ii) The posterior distribution is $(\mu,\tau)^T|\boldsymbol{x} \sim NGa(B,C,G,H)$ where the posterior parameters are given by (2.3).

(iii) The marginal prior distributions are $\mu \sim t_{2g}\{b, h/(gc)\}$, $\tau \sim Ga(g,h)$, $\sigma = 1/\sqrt{\tau} \sim$ *Inv-Chi*(g,h).

(iv) The marginal posterior distributions are $\mu|\boldsymbol{x} \sim t_{2G}\{B, H/(GC)\}$, $\tau|\boldsymbol{x} \sim Ga(G,H)$, $\sigma|\boldsymbol{x} \sim$ *Inv-Chi*(G,H).

(v) Prior and posterior means and standard deviations for μ, τ and σ can be calculated from the properties of the t, *Gamma* and *Inv-Chi* distributions.

(vi) Prior and posterior probabilities and densities for μ, τ and σ can be calculated using the R functions `pgt, dgt, pgamma, dgamma, pinvchi, dinvchi`.

(vii) HDIs or equi-tailed CIs for μ, τ and σ can be calculated using `qgt`, `hdiGamma`, `hdiInvchi`, `qgamma`, `qinvchi`.

(viii) Contour plots of the prior and posterior densities for $(\mu, \tau)^T$ can be plotted using the `NGacontour` function.

(ix) Prior and posterior confidence regions for $(\mu, \tau)^T$ can be plotted using the `NGacontour` function.

(x) The predictive distribution for a new observation Y from the population is $Y|\boldsymbol{x} \sim t_{2G}\{B, H(C+1)/(GC)\}$ and its HDI can be calculated using the `qgt` function.

2.6 Why do we have so many different distributions?

So far we have used many distributions, some you will have met before and some will be new. After a while the variety and sheer number of different distributions can become overwhelming. Why do we need so many distributions and why do we name so many of them?

Statistics studies the random variation in experiments, samples and processes. The variety of applications leads to their randomness being described by many different distributions. In many applications, bespoke distributions will need to be formulated. However, some distributions come up time and time again for modelling random variation in data and for describing prior beliefs. It is helpful for us to be able to refer to these distributions – and so we give each one a name – and also to be able to quote known results for these distributions such as their mean and variance. In this chapter you have been introduced to a generalisation of the t-distribution and the inverse chi distribution, and we have been able to use results for their mean and variance to study prior and posterior distributions and have been able to plot these distributions using functions in the `R` package.

You will meet several other new distributions in the remainder of the module. You won't be surprised to hear that it is useful to have a working knowledge of each of these distributions but perhaps not vital to remember all their properties listed in these notes. To help in this regard, the exam paper will contain a list of all the distributions used in the exam, together with their density (or probability function) and any useful results such as their mean and variance (as needed for the exam); see the specimen exam paper at the back of this booklet.

2.7 Learning objectives

By the end of this chapter, you should be able to:

- determine the posterior distribution for $(\mu, \tau)^T$
- determine and use the univariate prior and posterior distributions
- determine confidence intervals, HDIs and confidence regions
- determine the predictive distribution of another value from the population, and its predictive interval
- determine the predictive distribution of the mean of another random sample from the population

both in general and for a particular prior and data set. Also you should be able to:

- appreciate the benefit of naming distributions and for having lists of properties for these distributions

Chapter 3

General results for multi-parameter problems

In this chapter we will study some general results for multi-parameter problems.

3.1 Different levels of prior knowledge

We have *substantial prior information* for $\boldsymbol{\theta}$ when the prior distribution *dominates* the posterior distribution, that is $\pi(\boldsymbol{\theta}|\boldsymbol{x}) \sim \pi(\boldsymbol{\theta})$.

When prior information about $\boldsymbol{\theta}$ is limited, this is usually represented through the use of a conjugate prior distribution, with vague prior knowledge represented by making the conjugate distribution as diffuse as possible.

If we represent prior ignorance for a single parameter θ by using uniform or improper priors then we have seen (MAS2903, section 3.4) that, in general, the prior for $g(\theta)$ is not constant and so we are not ignorant about $g(\theta)$. The same problem occurs when we have more than one parameter.

Suppose we represent prior ignorance about $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ using $\pi(\boldsymbol{\theta}) = \textit{constant}$. Let $\phi_i = g_i(\boldsymbol{\theta})$, $i = 1, \ldots, p$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ be a 1–1 transformation. Then, in general, the prior density for $\boldsymbol{\phi}$ is not constant and this suggests that we are not ignorant about $\boldsymbol{\phi}$. However, if we are ignorant about $\boldsymbol{\theta}$ then we must also be ignorant about $g(\boldsymbol{\theta})$.

This contradiction makes it impossible to use this representation of prior ignorance.

Example 3.1

Suppose $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$. If we are ignorant about $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ then show that $\theta_1\theta_2$ does not have a constant prior density.

Example 3.2

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ unknown). Determine the Jeffreys prior for this model.

Hint: We have already seen that the likelihood function can be written as

$$f(\boldsymbol{x}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{n\tau}{2}\left\{s^2 + (\bar{x} - \mu)^2\right\}\right]$$

where

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

✎ *...Solution to Example 3.2*

✎ *...Solution to Example 3.2*

3.2 Asymptotic posterior distribution

Suppose we have a statistical model for data with likelihood function $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, together with a prior distribution with density $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. Then

$$J(\hat{\boldsymbol{\theta}})^{1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})|\boldsymbol{x} \xrightarrow{\mathcal{D}} N_p(0, I_p) \qquad \text{as } n \to \infty,$$

where $\hat{\boldsymbol{\theta}}$ is the likelihood mode, I_p is the $p \times p$ identity matrix and $J(\boldsymbol{\theta})$ is the observed information matrix, with $(i, j)^{th}$ element

$$J_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{x}|\boldsymbol{\theta}),$$

and $A^{1/2}$ denotes the square root matrix of A.

Comments

1. This asymptotic result can give us a useful approximation to the posterior distribution for $\boldsymbol{\theta}$ when n is large:

$$\boldsymbol{\theta}|\boldsymbol{x} \sim N_p\left(\hat{\boldsymbol{\theta}}, J(\hat{\boldsymbol{\theta}})^{-1}\right) \qquad \text{approximately.}$$

2. This limiting result is similar to one for the maximum likelihood estimator in Frequentist Statistics:

$$I(\boldsymbol{\theta})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N_p(0, I_p) \qquad \text{as } n \to \infty,$$

   where $I(\boldsymbol{\theta}) = E_{\boldsymbol{X}|\boldsymbol{\theta}}[J(\boldsymbol{\theta})]$ is Fisher's information matrix. Note that this statement about the distribution of $\hat{\boldsymbol{\theta}}$ for fixed (unknown) $\boldsymbol{\theta}$, whereas the results above is a statement about the distribution of $\boldsymbol{\theta}$ for fixed (known) $\hat{\boldsymbol{\theta}}$.

Example 3.3

Suppose we now have a random sample from a $N(\mu, 1/\tau)$ distribution (with unknown precision). Determine the asymptotic posterior distribution for (μ, τ).

Hint: we have already seen that the likelihood function can be written as

$$f(\boldsymbol{x}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{n\tau}{2}\left\{s^2 + (\bar{x}-\mu)^2\right\}\right]$$

where $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$.

We have

$$\begin{aligned}
\log f(\boldsymbol{x}|\mu,\tau) &= \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{n\tau}{2}\left\{s^2 + (\bar{x}-\mu)^2\right\} \\
\Longrightarrow \frac{\partial}{\partial\mu}\log f(\boldsymbol{x}|\mu,\tau) &= n\tau(\bar{x}-\mu) \\
\frac{\partial}{\partial\tau}\log f(\boldsymbol{x}|\mu,\tau) &= \frac{n}{2\tau} - \frac{n}{2}\left\{s^2 + (\bar{x}-\mu)^2\right\} \\
\Longrightarrow \frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu,\tau) &= -n\tau \\
\frac{\partial^2}{\partial\mu\partial\tau}\log f(\boldsymbol{x}|\mu,\tau) &= n(\bar{x}-\mu) \\
\frac{\partial^2}{\partial\tau^2}\log f(\boldsymbol{x}|\mu,\tau) &= -\frac{n}{2\tau^2}.
\end{aligned}$$

✎ *...Solution to Example 3.3* Now

$$\frac{\partial}{\partial\mu}\log f(\boldsymbol{x}|\mu,\tau)=0 \qquad \Longrightarrow \qquad \hat{\mu}=\bar{x}$$
$$\frac{\partial}{\partial\tau}\log f(\boldsymbol{x}|\mu,\tau)=0 \qquad \Longrightarrow \qquad \hat{\tau}=\frac{1}{s^2}$$

Therefore

$$J_{11}(\hat{\mu},\hat{\tau})=-\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu,\tau)\bigg|_{(\hat{\mu},\hat{\tau})}=n\hat{\tau}=\frac{n}{s^2}$$
$$J_{12}(\hat{\mu},\hat{\tau})=-\frac{\partial^2}{\partial\mu\partial\tau}\log f(\boldsymbol{x}|\mu,\tau)\bigg|_{(\hat{\mu},\hat{\tau})}=-n(\bar{x}-\hat{\mu})=0$$
$$J_{22}(\hat{\mu},\hat{\tau})=-\frac{\partial^2}{\partial\tau^2}\log f(\boldsymbol{x}|\mu,\tau)\bigg|_{(\hat{\mu},\hat{\tau})}=\frac{n}{2\hat{\tau}^2}=\frac{ns^4}{2}$$

and so

$$J(\hat{\mu},\hat{\tau})=\begin{pmatrix}\frac{n}{s^2} & 0\\ 0 & \frac{ns^4}{2}\end{pmatrix},$$

whence

$$J(\hat{\mu},\hat{\tau})^{-1}=\begin{pmatrix}\frac{s^2}{n} & 0\\ 0 & \frac{2}{ns^4}\end{pmatrix}.$$

Therefore, for large n, the (approximate) posterior distribution for (μ,τ) is

$$\begin{pmatrix}\mu\\ \tau\end{pmatrix}\bigg|\,\boldsymbol{x}\sim N_2\left\{\begin{pmatrix}\bar{x}\\ \frac{1}{s^2}\end{pmatrix},\begin{pmatrix}\frac{s^2}{n} & 0\\ 0 & \frac{2}{ns^4}\end{pmatrix}\right\}.$$

Putting this another way, for large n

$$\mu|\boldsymbol{x}\sim N(\bar{x},s^2/n), \qquad \tau|\boldsymbol{x}\sim N\{1/s^2,2/(ns^4)\},$$

independently.

3.3 Learning objectives

By the end of this chapter, you should be able to:

- understand different levels of prior information and have an appreciation for the difficulty of specifying ignorance priors in multi-parameter problems
- determine the asymptotic posterior distribution when the data are a large random sample from **any** distribution
- explain the similarities and differences between the asymptotic posterior distribution and the asymptotic distribution of the maximum likelihood estimator

Chapter 4

Non-conjugate multi-parameter problems

In this chapter we will study some multi-parameter problems in which the prior distribution does not have to be conjugate. Inferences are made by using techniques which simulate realisations from the posterior distribution. These methods are generally referred to as *Markov Chain Monte Carlo* techniques, and often abbreviated to MCMC. There are many different MCMC techniques, but we only have time to look briefly at two of the most fundamental. The first is the *Gibbs sampler*, which was at the forefront of the recent MCMC revolution, and the second is generally known as *Metropolis-Hastings* sampling. In fact, MCMC schemes based on the combination of these two fundamental techniques are still at the forefront of MCMC research.

4.1 Why is inference not straightforward in non-conjugate problems?

Example 4.1

Consider again the problem in section 2.2 in which we have a random sample from a normal distribution where both the mean μ and the precision τ are unknown, that is, $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent). In this section, we showed that a NGa prior for $(\mu, \tau)^T$ was conjugate, that is, if we used a $NGa(b, c, g, h)$ prior distribution for $(\mu, \tau)^T$ then the posterior was a $NGa(B, C, G, H)$ distribution. But what if a $NGa(b, c, g, h)$ prior distribution does not adequately represent our prior beliefs? Suppose instead that our prior beliefs are represented by independent priors for the parameters, with

$$\mu \sim N\left(b, \frac{1}{c}\right) \qquad \text{and} \qquad \tau \sim Ga(g, h)$$

for known values b, c, g and h. What is the posterior distribution for $(\mu, \tau)^T$?

Solution

Previously we have seen that the likelihood function is

$$f(\boldsymbol{x}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{n\tau}{2}\left\{s^2 + (\bar{x}-\mu)^2\right\}\right].$$

Using Bayes Theorem, the posterior density is

$$\begin{aligned}\pi(\mu,\tau|\boldsymbol{x}) &\propto \pi(\mu,\tau)\, f(\boldsymbol{x}|\mu,\tau)\\ &\propto \pi(\mu)\,\pi(\tau)\, f(\boldsymbol{x}|\mu,\tau)\end{aligned}$$

and so, for $\mu \in \mathbb{R}$, $\tau > 0$

$$\begin{aligned}\pi(\mu,\tau|\boldsymbol{x}) &\propto \left(\frac{c}{2\pi}\right)^{1/2} \exp\left\{-\frac{c}{2}(\mu-b)^2\right\} \times \frac{h^g \tau^{g-1} e^{-h\tau}}{\Gamma(g)}\\ &\qquad\qquad \times \tau^{\frac{n}{2}} \exp\left[-\frac{n\tau}{2}\left\{s^2 + (\bar{x}-\mu)^2\right\}\right]\\ &\propto \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\}.\end{aligned}$$

This is not a $NGa(B,C,G,H)$ density as it cannot be written in the form

$$k\,\tau^{G-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[C(\mu-B)^2 + 2H\right]\right\}$$

for any choice of B, C, G and H. In fact $\pi(\mu,\tau|\boldsymbol{x})$ is not the density of any standard distribution.

Can we plot this density? Before we do, we need to know the constant of proportionality, k. As the posterior density must integrate to one

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} k\tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu = 1$$

$$k^{-1} = \int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu$$

Therefore the posterior density is, for $\mu \in \mathbb{R}$, $\tau > 0$

$$\pi(\mu,\tau|\boldsymbol{x}) = \frac{\tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\}}{\displaystyle\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} d\tau\, d\mu}.$$

What is the posterior mean of μ and of τ? What are their marginal distributions? How

can we calculate the moments $E(\mu^{m_1}\tau^{m_2}|\boldsymbol{x})$ of this posterior distribution? Now

$$\begin{aligned}\pi(\mu|\boldsymbol{x}) &= \int_0^\infty \pi(\mu,\tau|\boldsymbol{x})\,d\tau \\ &= \frac{\displaystyle\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\tau}{\displaystyle\int_{-\infty}^\infty\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\tau\,d\mu}\end{aligned}$$

and

$$\begin{aligned}\pi(\tau|\boldsymbol{x}) &= \int_{-\infty}^\infty \pi(\mu,\tau|\boldsymbol{x})\,d\mu \\ &= \frac{\displaystyle\int_{-\infty}^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\mu}{\displaystyle\int_{-\infty}^\infty\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\tau\,d\mu}.\end{aligned}$$

In general, the moments are

$$\begin{aligned}E(\mu^{m_1}\tau^{m_2}|\boldsymbol{x}) &= \int_{-\infty}^\infty\int_0^\infty \mu^{m_1}\tau^{m_2}\pi(\mu,\tau|\boldsymbol{x})\,d\tau\,d\mu \\ &= \frac{\displaystyle\int_{-\infty}^\infty\int_0^\infty \mu^{m_1}\tau^{m_2}\times\tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\tau\,d\mu}{\displaystyle\int_{-\infty}^\infty\int_0^\infty \tau^{g+\frac{n}{2}-1}\exp\left\{-\frac{c}{2}(\mu-b)^2 - h\tau - \frac{n\tau}{2}\left[s^2+(\bar{x}-\mu)^2\right]\right\}\,d\tau\,d\mu}.\end{aligned}$$

These integrals cannot be determined analytically, though it is possible to use numerical integration methods or approximations (for large n). However, in general, the accuracy of the numerical approximation to the integral deteriorates as the dimension of the integral increases.

Comment

The above shows how not using a conjugate prior distribution can cause many basic problems such as plotting the posterior density or determining posterior moments. But having to use conjugate priors is far too restrictive for many real data analyses: (i) our prior beliefs may not be captured using a conjugate prior; (ii) most models for complex data do not have conjugate priors. It was for these reasons that until relatively recently (say mid-1990s), practical Bayesian inference for real complex problems was either not feasible or only undertaken by the dedicated few prepared to develop bespoke computer code to numerically evaluate all the integrals etc.

4.2 Simulation-based inference

One way to get around the problem of having to work out integrals (like those in the previous section) is to base inferences on simulated realisations from the posterior distribution. This is the fundamental idea behind MCMC methods. If we could simulate from

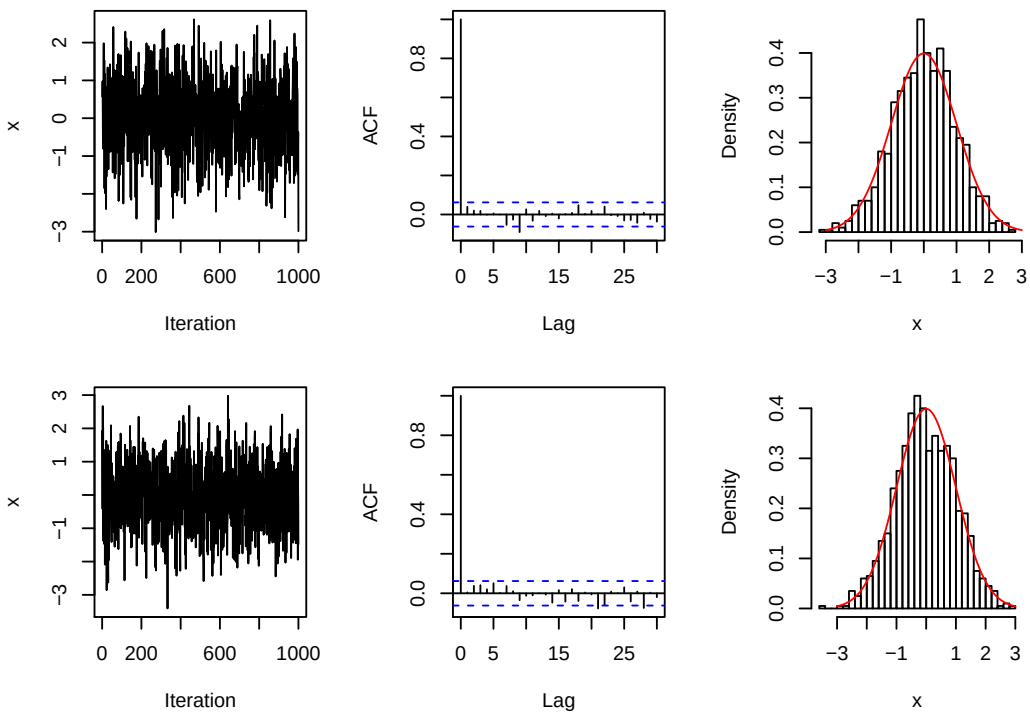


Figure 4.1: Summaries of the 1K realisations from the black box simulator

the posterior distribution then we could use a very large sample of realisations to determine posterior means, standard deviations, correlations, joint densities, marginal densities etc.

As an example, imagine you wanted to know about the standard normal distribution – its shape, its mean, its standard deviation – but didn't know any mathematics so that you couldn't derive say the distribution's zero mean and unit variance. However you've been given a "black box" which can simulate realisations from this distribution. Here we'll use the `R` function `rnorm()` as the black box simulator. If you decide to generate 1K realisations the output might look something like the top row of Figure 4.1. The top left plot shows the trace plot of the output, that is, the realisations from the black box sampler in the order they are produced. The next plot along the top row shows the autocorrelation (ACF) plot. This shows how correlated the realisations are at different lags. We know that the simulator `rnorm()` produces independent realisations and so the (sample) correlation between say consecutive values $corr(x_i, x_{i+1})$ will be almost zero. This is also the case for correlations at all positive lags. Finally the lag 0 autocorrelation $corr(x_i, x_i)$ must be one (by definition). The sample ACF plot is consistent with all of these "claims". Finally the top right plot is a density histogram of the realisations. This too is consistent with the standard normal density (which is also shown). We can also estimate various quantities of the standard normal distribution; for example:

```
 1st Qu.   Median     Mean  3rd Qu.  St.Dev.
-0.65240 -0.00130 -0.00192  0.64810  0.96049
```

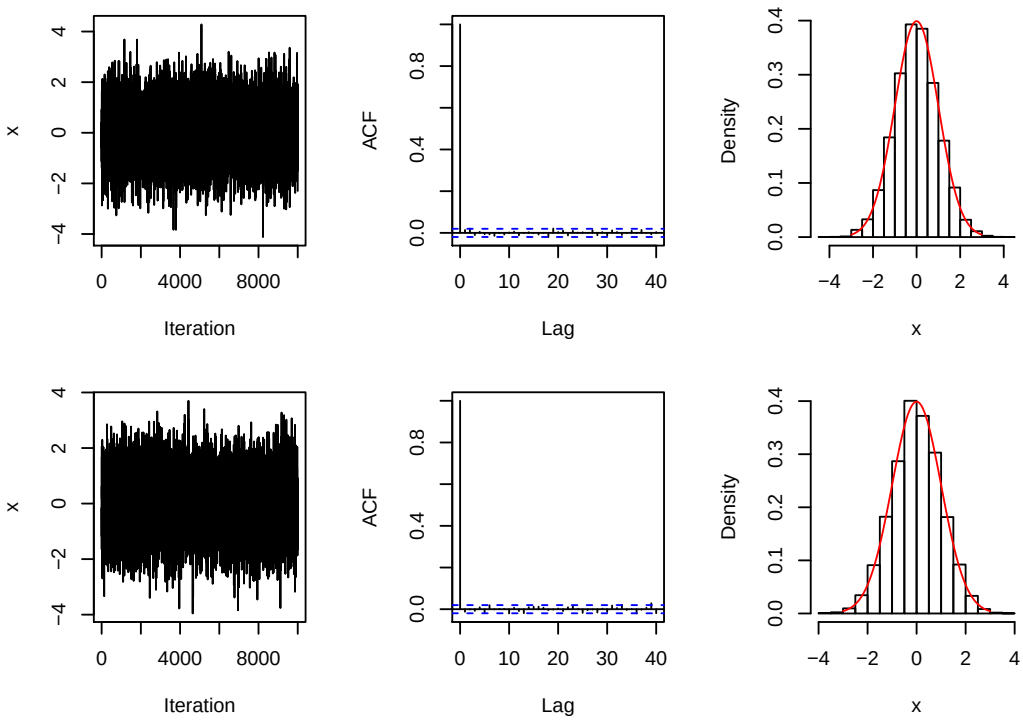


Figure 4.2: Summaries of the 10K realisations from the black box simulator

Here we see that the mean and median are around zero and the standard deviation is around one.

The second row of plots in the figure is another collection of 1K realisations from the black box simulator. These look very similar to those on the top row but are slightly different due to the stochasticity (random nature) of the simulator. This output has the following numerical summaries:

```
 1st Qu.   Median     Mean  3rd Qu.  St.Dev.
-0.69880 -0.09637 -0.03274  0.67330  0.99599
```

Again these numerical summaries are slightly different but essentially tell the same story. In fact we know from previous modules that there is sample variability in estimates of means from random samples. So if we use the simulator again (twice) to obtain 10K realisations, we will get even more similar looking output; see Figure 4.2. The numerical summaries from these outputs are

```
  1st Qu.  Median     Mean  3rd Qu.  St.Dev.
-0.66820 -0.00048  0.00370  0.68070  0.99593
-0.67130  0.01354  0.01008  0.67920  1.00691
```

Here we have much less sampling variability in our estimates due to the larger sample size. In fact we can estimate any "population" quantity to any required accuracy simply by simulating a large enough collection of realisations.

These analyses show how we can make inferences, calculate means, variances, densities etc by using realisations from a distribution. In the rest of this chapter, we will look into how we can construct algorithms for simulating from (complex) posterior distributions, from which we can then make inferences.

4.3 Motivation for MCMC methods

The example in section 4.1 showed that using non-conjugate priors can be problematic. MCMC methods address this problem by providing an algorithm which simulates realisations from the posterior distribution.

We consider a generic case where we want to simulate realisations of two random variables X and Y with joint density $f(x, y)$. This joint density can be factorised as

$$f(x, y) = f(x)f(y|x)$$

and so we can simulate from $f(x, y)$ by first simulating $X = x$ from $f(x)$, and then simulating $Y = y$ from $f(y|x)$. On the other hand, we can write

$$f(x, y) = f(y)f(x|y)$$

and so simulate $Y = y$ from $f(y)$ and then $X = x$ from $f(x|y)$.

We have already seen that dealing with conditional posterior distributions is straightforward when the prior is semi-conjugate, so let's assume that simulating from $f(y|x)$ and $f(x|y)$ is straightforward. The key problem with using either of the above methods is that, in general, we can't simulate from the marginal distribution, $f(x)$ and $f(y)$.

For the moment, suppose we can simulate from the marginal distribution for X, that is, we have an $X = x$ from $f(x)$. We can now simulate a $Y = y$ from $f(y|x)$ to give a pair (x, y) from the bivariate density. Given that this pair is from the bivariate density, the y value must be from the marginal $f(y)$, and so we can simulate an $X = x'$ from $f(x|y)$ to give a new pair (x', y) also from the joint density. But now x' is from the marginal $f(x)$, and so we can simulate a $Y = y'$ from $f(y|X = x')$ to give a new pair (x', y') also from the joint density. And we can keep going.

This alternate sampling from conditional distributions defines a bivariate Markov chain, and the above is an intuitive explanation for why $f(x, y)$ is its stationary distribution. Thus being able to simulate easily from conditional distributions is key to this methodology.

4.4 The Gibbs sampler

Suppose we want to generate realisations from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and that we can simulate from the full conditional distributions (FCDs)

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \boldsymbol{x}) = \pi(\theta_i|\cdot), \qquad i = 1, 2, \ldots, p.$$

The Gibbs sampler follows the following algorithm:

1. Initialise the iteration counter to $j = 1$.
   Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$.
2. Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values
$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{x}) \\
\vdots \qquad & \quad \vdots \qquad\qquad \vdots \\
\theta_p^{(j)} &\sim \pi(\theta_p|\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{p-1}^{(j)}, \boldsymbol{x})
\end{aligned}$$
3. Change counter j to $j + 1$, and return to step 2.

This algorithm defines a homogeneous Markov chain as each simulated value depends only on the previous simulated value and not on any other previous values or the iteration counter j. It can be shown that $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the stationary distribution of this chain and so if we simulate realisations by using a Gibbs sampler, eventually the the Markov chain will converge to the required posterior distribution.

4.4.1 Processing output from a Gibbs sampler

Burn-in period

First we have to determine how many iterations are needed before the Gibbs sampler has reached its stationary distribution. This is known as the *burn-in* period. There are many diagnostic tests available to help determine how long this is but, in general, the most effective method is simply to look at a trace plot of the posterior sample and detect the point after which the realisations look to be from the same distribution. Figure 4.3 illustrates typical output from a Gibbs sampler. Here we see the output when using three different starting points. Initially there is a transient stage and then the distribution of the output becomes the same for each chain (perhaps after iteration 500).

Once the Gibbs sampler has reached its stationary distribution, all subsequent iterates are realisations from the posterior distribution. Suppose that $\boldsymbol{\theta} = (\mu, \tau)^T$ and we have run the Gibbs sampler for N iterations after convergence giving a posterior sample
$$\{(\mu^{(1)}, \tau^{(1)}), (\mu^{(2)}, \tau^{(2)}), \ldots, (\mu^{(N)}, \tau^{(N)})\}.$$
We can use this sample to calculate any features of the posterior distribution. For example, we can estimate the marginal posterior densities $\pi(\mu|\boldsymbol{x})$ and $\pi(\tau|\boldsymbol{x})$ by using histograms of the $\mu^{(j)}$ and of the $\tau^{(j)}$ respectively. Of course, as N is finite we cannot determine these densities exactly. We can also estimate other features of the posterior distribution such as the posterior means, variances and correlation by using their equivalents in the posterior sample: $\bar{\mu}$, $\bar{\tau}$, s_μ^2, s_τ^2 and $r_{\mu\tau}$. Again these estimates will not be exact as N is finite. However, we can make them as accurate as we want by taking a sufficiently large posterior sample, that is, by taking N large enough.

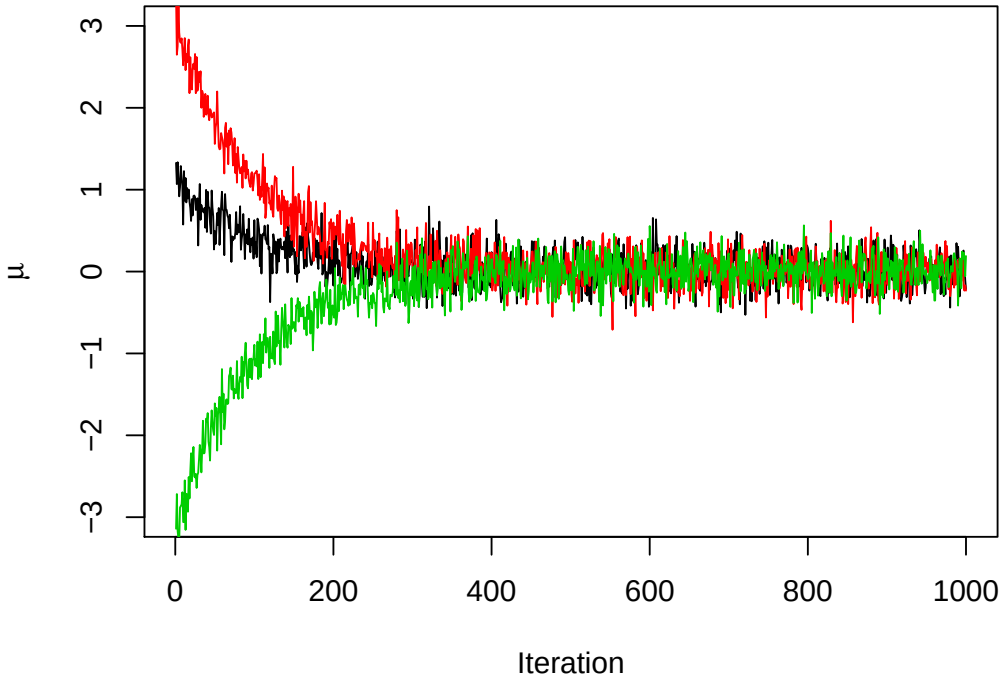


Figure 4.3: Demonstration of burn-in using a Gibbs sampler and three initial values

Dealing with autocorrelation

Another problem with the output from a Gibbs sampler (after convergence) is that it is not a random sample. It should not be surprising that successive realisations from the sampler are autocorrelated, after all the output is a realisation from a Markov chain! To understand the dependence structure, we look at the sample autocorrelation function for each variable. For example, the sample autocorrelation function (ACF) for μ at lag k is

$$r(k) = Corr(\mu^{(j)}, \mu^{(j+k)}).$$

Looking at a plot of this ACF can give an idea as to how much to *thin* the output before it becomes un-autocorrelated (the sample autocorrelations at lags 1,2,3,. . . are small). Thinning here means not taking every realisation by say taking say every mth realisation. In general, an appropriate level of thinning is determined by the largest lag m at which any of the variables have a non-negligible autocorrelation. If doing this leaves a (thinned) posterior sample which is too small then the original Gibbs sampler should be re-run (after convergence) for a sufficiently large number of iterations until the thinned sample is of the required size.

To get a clearer idea of how thinning works, we now look at some output from a moving average $MA(1)$ process. Figure 4.4 shows this output together with its sample autocorrelation function. Theoretically output from this process should have zero autocorrelations at lags greater than one, and this is what we see (with sample noise) in the figure. If we now thin the output by taking every other value then it's clear that the autocorrelations at non-zero lags should be zero. The lower two plots show the thinned output and its

sample autocorrelation function. This ACF plot suggests that the thinned output is not autocorrelated.

We now look at some output from an autoregressive $AR(1)$ process. Figure 4.5 shows this output together with its sample autocorrelation function. As mentioned previously, theoretically output from this process has autocorrelations which decrease geometrically, and this is what we see (with sample noise) in the figure. The smallest lag after which the autocorrelations are negligible is around 7–10. If we now thin the output by taking every 10th value then we should get output which is almost un-autocorrelated. The lower two plots show the thinned output and its sample autocorrelation function and the ACF plot is consistent with the thinned output being un-autocorrelated.

If the MCMC output is un-autocorrelated then the accuracy of $\bar{\mu}$ is roughly $\pm 2s_\mu/\sqrt{N}$. However if the Markov chain followed an autoregressive $AR(1)$ process, its autocorrelations would decrease geometrically, with $r(k) \simeq r(1)^k$, $k \geq 2$. In this case it can be shown that the accuracy of $\bar{\mu}$ is roughly $\pm 2s_\mu/\sqrt{N\{1-r(1)\}^2}$, that is, because the process has autocorrelation, the amount of information in the data is equivalent to a random sample with size $N_{eff} = N\{1-r(1)\}^2$. This effective random sample size calculation gets more complicated for processes with non-zero higher order autocorrelations and this is why we usually adopt the simplistic method of thinning. It's worth noting that, in general, MCMC output with positive autocorrelations has $N_{eff} < N$. Also sometimes MCMC output with some negative autocorrelations can have $N_{eff} > N$.

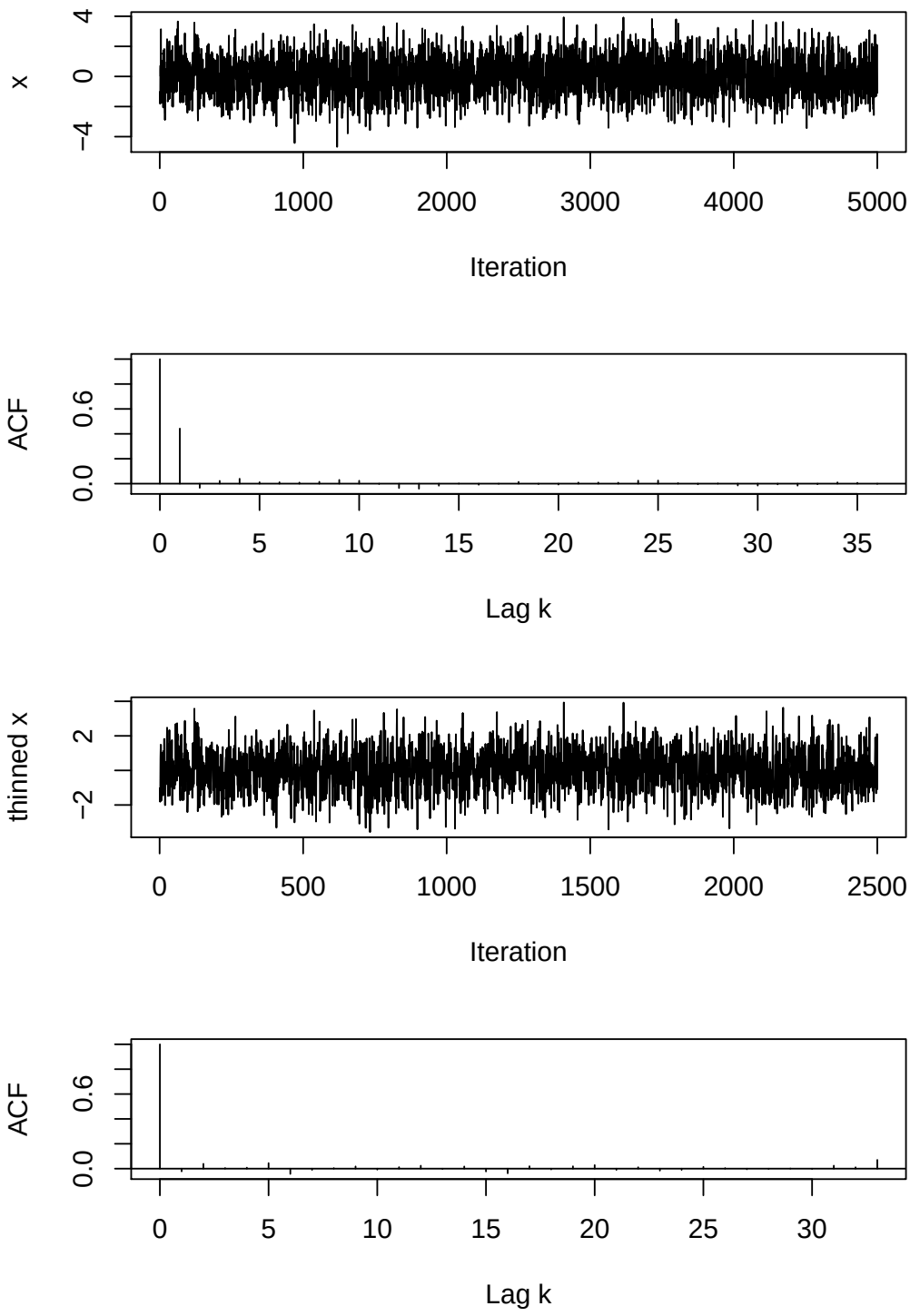


Figure 4.4: Effect of thinning output from a $MA(1)$ process

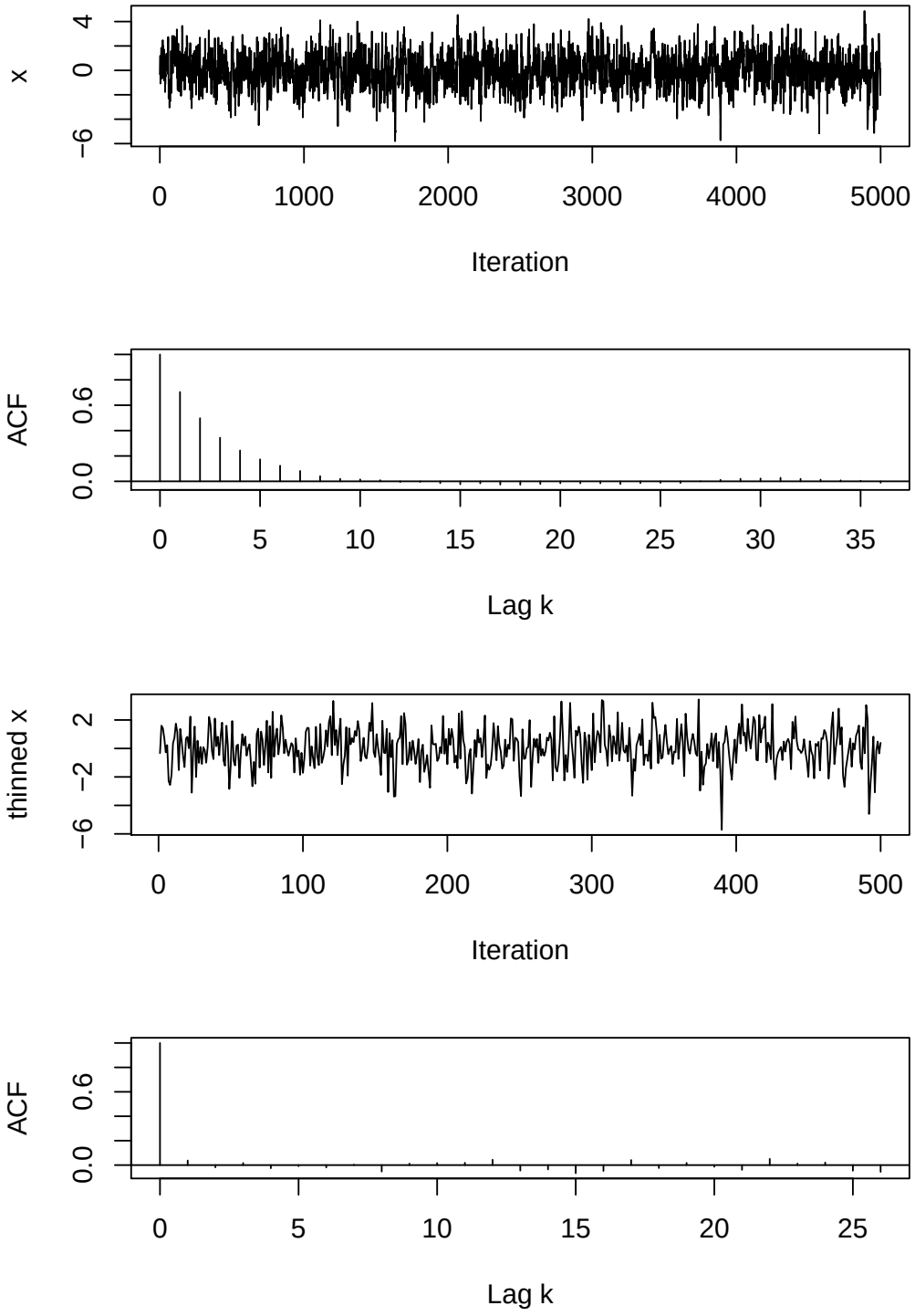


Figure 4.5: Effect of thinning output from a $AR(1)$ process

Strategy

1. Determine the *burn-in* period, after which the Gibbs sampler has reached its stationary distribution. This may involve thinning the posterior sample as slowly snaking trace plots may be due to high autocorrelations rather than a lack of convergence.
2. After this, determine the level of thinning needed to obtain a posterior sample whose autocorrelations are roughly zero.
3. Repeat steps 1 and 2 several times using different initial values to make sure that the sample really is from the stationary distribution of the chain, that is, from the posterior distribution.

Accuracy of posterior summaries

Each time we run an MCMC scheme, we obtain a different sample from the posterior distribution. Suppose that after burn-in and thinning, we have a large sample with N unautocorrelated values, say $\mu_1, \ldots, \mu_N$. In order to determine the accuracy of the sample mean and standard deviation estimates of the posterior mean and standard deviation we need to make some assumption about the posterior distribution. If the data sample size n is large then the posterior distribution will be approximately normal. So we will think of our MCMC output as being a random sample from the posterior distribution.

Suppose the posterior output has sample mean $\bar{\mu}$ and standard deviation s_μ. We need to know the accuracy of these estimates of $M = E(\mu|\boldsymbol{x})$ and $\Sigma = SD(\mu|\boldsymbol{x})$. We saw in Example 3.3 that the asymptotic posterior distribution about the mean and precision $(\mu, \tau)^T$ using a random sample from a normal $N(\mu, 1/\tau)$ distribution was

$$\mu|\boldsymbol{x} \sim N(\bar{x}, s^2/n), \qquad \tau|\boldsymbol{x} \sim N\{1/s^2, 2/(ns^4)\}, \qquad \text{independently.}$$

Rewriting this result in terms of the MCMC sample mean $\bar{\mu}$, standard deviation s_μ and the parameters they estimate gives posterior distributions

$$M \sim N(\bar{\mu}, s_\mu^2/N), \qquad \Sigma^{-2} \sim N\{1/s_\mu^2, 2/(Ns_\mu^4)\}, \qquad \text{independently.}$$

Therefore an approximate 95% HDI for M is

$$\bar{\mu} \pm z_{0.025}\frac{s_\mu}{\sqrt{N}} \simeq \bar{\mu} \pm \frac{2s_\mu}{\sqrt{N}}$$

since $z_{0.025} \simeq 2$.

Also, from the posterior distribution for Σ^{-2}, we have

$$\begin{aligned}
& P\left(\frac{1}{s_\mu^2} - 2\sqrt{\frac{2}{Ns_\mu^4}} < \Sigma^{-2} < \frac{1}{s_\mu^2} + 2\sqrt{\frac{2}{Ns_\mu^4}}\right) \simeq 0.95 \\
\Longrightarrow \quad & P\left(\frac{1 - 2\sqrt{2/N}}{s_\mu^2} < \Sigma^{-2} < \frac{1 + 2\sqrt{2/N}}{s_\mu^2}\right) \simeq 0.95 \\
\Longrightarrow \quad & P\left(\frac{s_\mu}{\sqrt{1 + 2\sqrt{2/N}}} < \Sigma < \frac{s_\mu}{\sqrt{1 - 2\sqrt{2/N}}}\right) \simeq 0.95
\end{aligned}$$

Therefore a 95% confidence interval for Σ is

$$s_\mu \left(1 \pm 2\sqrt{2/N}\right)^{-1/2} \simeq s_\mu \left(1 \pm \frac{1}{2} \times 2\sqrt{2/N}\right) = s_\mu \pm s_\mu \sqrt{\frac{2}{N}}.$$

It can be shown that these accuracy calculations are fairly accurate even when the posterior distribution (from which we have the MCMC sample) is not particularly normal.

4.4.2 Bayesian inference using a Gibbs sampler

Example 4.2

Construct a Gibbs sampler for the posterior distribution in Example 4.1.

Solution

We first need to determine the conditional posterior density for μ and for τ.

The conditional posterior density for μ is, for $\mu \in \mathbb{R}$

$$\begin{aligned}
\pi(\mu|\tau, \boldsymbol{x}) &= \frac{\pi(\mu, \tau|\boldsymbol{x})}{\pi(\tau|\boldsymbol{x})} \\
&\propto \pi(\mu, \tau|\boldsymbol{x}) \\
&\propto \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu - b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\} \\
&\propto \exp\left\{-\frac{c}{2}(\mu - b)^2 - \frac{n\tau}{2}(\bar{x} - \mu)^2\right\}
\end{aligned}$$

after moving all multiplicative constants not involving μ into the proportionality sign, and so after completing the square in the exponent

$$\pi(\mu|\tau, \boldsymbol{x}) \propto \exp\left\{-\frac{c + n\tau}{2}\left(\mu - \frac{bc + n\tau\bar{x}}{c + n\tau}\right)^2\right\}, \mu \in \mathbb{R}$$

that is

$$\mu|\tau, \boldsymbol{x} \sim N\left(\frac{bc + n\tau\bar{x}}{c + n\tau}, \frac{1}{c + n\tau}\right).$$

Note that we could have obtained this result by using Example 1.3 with $d = c$.

The conditional posterior density for τ is, for $\tau > 0$

$$\begin{aligned}
\pi(\tau|\mu, \boldsymbol{x}) &= \frac{\pi(\mu, \tau|\boldsymbol{x})}{\pi(\mu|\boldsymbol{x})} \\
&\propto \pi(\mu, \tau|\boldsymbol{x}) \\
&\propto \tau^{g+\frac{n}{2}-1} \exp\left\{-\frac{c}{2}(\mu - b)^2 - h\tau - \frac{n\tau}{2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\} \\
&\propto \tau^{g+\frac{n}{2}-1} \exp\left\{-\left[h + \frac{n}{2}\left\{s^2 + (\bar{x} - \mu)^2\right\}\right]\tau\right\}
\end{aligned}$$

after moving all multiplicative constants not involving τ into the proportionality sign, and so

$$\tau|\mu, \boldsymbol{x} \sim Ga\left(g + \frac{n}{2}, h + \frac{n}{2}\left\{s^2 + (\bar{x} - \mu)^2\right\}\right).$$

We also need to initialise the algorithm. We might use the prior mean ($\mu^{(0)} = b$, $\tau^{(0)} = g/h$) or the mle ($\mu^{(0)} = \bar{x}$, $\tau^{(0)} = 1/s^2$). Alternatively, if we wanted to compare different runs of the Gibbs sampler, we might simulate $(\mu^{(0)}, \tau^{(0)})$ from the prior distribution.

Therefore the Gibbs sampler is

1. Initialise the iteration counter to $j = 1$.
   Initialise the state of the chain by taking $(\mu^{(0)}, \tau^{(0)})$ as $(b, g/h)$ or $(\bar{x}, 1/s^2)$ or as a random draw from the prior distribution.

2. Obtain new values $\mu^{(j)}$ and $\tau^{(j)}$ from $\mu^{(j-1)}$ and $\tau^{(j-1)}$ via

$$\mu^{(j)} \sim N\left(\frac{bc + n\tau^{(j-1)}\bar{x}}{c + n\tau^{(j-1)}}, \frac{1}{c + n\tau^{(j-1)}}\right)$$
$$\tau^{(j)} \sim Ga\left(g + \frac{n}{2}, h + \frac{n}{2}\left\{s^2 + \left(\bar{x} - \mu^{(j)}\right)^2\right\}\right)$$

3. Change counter j to $j + 1$, and return to step 2.

Comments

Notice that, since μ and τ independent *a priori*, $\mu|\tau \sim N(b, 1/c)$. Therefore, given τ, the normal prior for μ is conjugate. Similarly, $\tau|\mu \sim Ga(g, h)$ and so, given μ, the gamma prior for τ is conjugate. Therefore, both conditional priors (for $\mu|\tau$ and $\tau|\mu$) are conjugate. Such priors are called *semi-conjugate*.

Producing and analysing output from this Gibbs sampler

The `R` function `gibbsNormal` in the library `nclbayes` implements this Gibbs sampling algorithm. The library also contains the functions `mcmcProcess` which can be used to remove the burn–in and thin the output, and `mcmcAnalysis` which analyses the MCMC output. Let us consider the case in which the data have size $n = 100$, mean $\bar{x} = 15$ and standard deviation $s = 4.5$ and the prior distribution has $\mu \sim N(10, 1/100)$ and $\tau \sim Ga(3, 12)$, independently. The following code produces output from this Gibbs sampler, initialising at $(10, 0.25)$ and then analyses the output.

```
library(nclbayes)
posterior=gibbsNormal(N=1000,initial=c(10,0.25),
    priorparam=c(10,1/100,3,12),n=100,xbar=15,s=4.5)
posterior2=mcmcProcess(input=posterior,burnin=10,thin=1)
```

```
op=par(mfrow=c(2,2))
plot(posterior,col=c(1:length(posterior)),main="All realisations")
plot(posterior,type="l",main="All realisations")
plot(posterior2,col=c(1:length(posterior2)),main="After deleting first 10")
plot(posterior2,type="l",main="After deleting first 10")
par(op)

mcmcAnalysis(posterior,rows=2,show=F)
mcmcAnalysis(posterior2,rows=2,show=F)
```

The first block of code runs the function `gibbsNormal`, with initial values `initial` taken as the prior means, to obtain the output. The next block then uses the function `mcmcProcess` to post-process the Gibbs output by deleting an initial `burnin` $= 10$ values and then not thinning by taking `thin` $= 1$. The next block of code produces the plots in Figure 4.6. After this the code analyses the Gibbs sampler output and produces the plots in Figure 4.7.

In Figure 4.6, the top left plot shows the values produced by the Gibbs sampler: notice the initial value $(\mu^{(0)}, \tau^{(0)}) = (b = 10, g/h = 0.25)$ appears at the top left part of the plot and the other values towards the bottom right part (in different colours). The top right plot is another representation of this output but here each consecutive pair $(\mu^{(j)}, \tau^{(j)})$ are joined by a line. This clearly shows that all pairs after the first one remain in the same vicinity (the bottom right part). The lower plots are the equivalent ones to the upper plots but only use the Gibbs sampler output after deleting the first 10 pairs.

In Figure 4.7, the top two rows of plots summarise the Gibbs sampler output using all realisations and the bottom two rows of plots are the equivalent plots but only use the Gibbs sampler output after deleting the first 10 pairs. The first column of plots shows the trace plot of the output, that is, the values for each variable (μ and τ) as the sampler iterates from its first value to its final value. The top two first column plots clearly show the initial value $(\mu^{(0)}, \tau^{(0)}) = (b = 10, g/h = 0.25)$, after which the subsequent values all look to be from the same distribution. In particular, there looks to be no change in the range of values or in the mean value. The bottom two first column plots emphasise these points; here the first 10 values have been deleted as burn-in, though probably we needed only to delete the first 2 or 3 values. Note that the benefit of using `R` code that runs quickly is that adopting a conservative strategy which deletes too many values as burn-in, does not have significant time implications (though this is not a sensible strategy if the code is very slow and takes months to run!).

The second column shows the autocorrelation function for each variable. Note that the spike at lag 0 is due to $r(0) = Corr(\mu^{(j)}, \mu^{(j)}) = 1$. The plots show that the autocorrelations at all other lags are negligible, and so no thinning is needed. The final column shows histograms of the Gibbs sampler output. If using a burn-in of 10 iterations is okay (and it is here!) then the subsequent output can be taken as our posterior sample and therefore the lower two histograms will be good estimates of the marginal densities: good because the output is (almost) uncorrelated and the sample size is quite large.

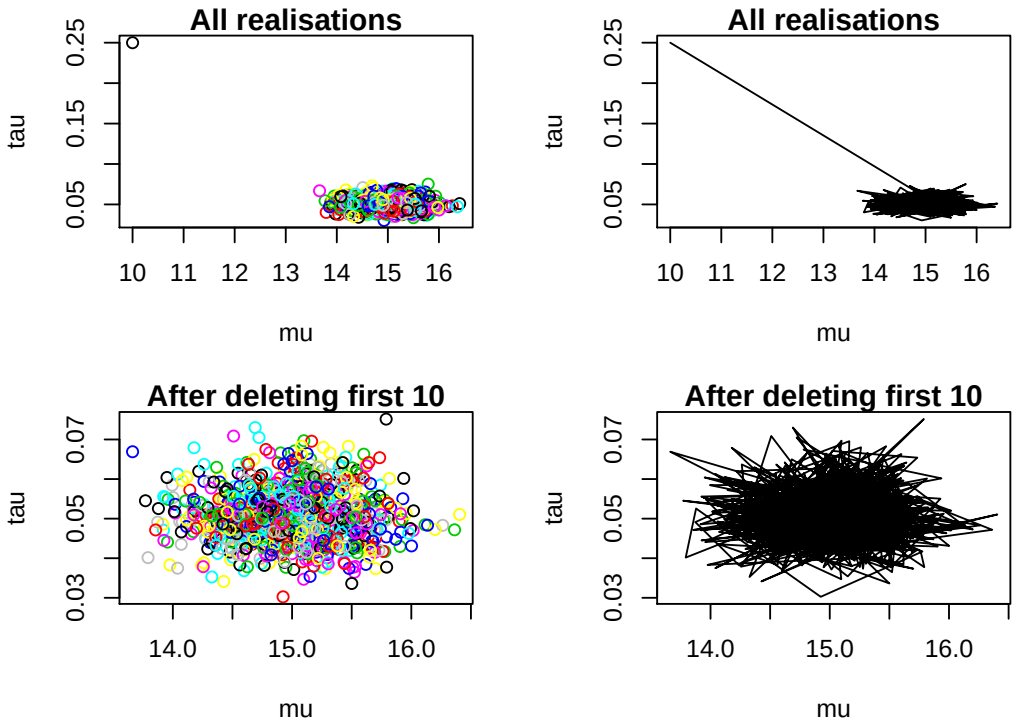


Figure 4.6: Progress of the MCMC scheme

If we use the command

```
mcmcAnalysis(posterior2,rows=2)
```

that is, don't use the `show=F` option, then the function will produce the plots and also various useful numerical summaries. In this run of the Gibbs sampler it gave

```
N = 990 iterations
       mu              tau
 Min.   :13.66   Min.   :0.03025
 1st Qu.:14.68   1st Qu.:0.04663
 Median :15.01   Median :0.05068
 Mean   :14.99   Mean   :0.05110
 3rd Qu.:15.29   3rd Qu.:0.05557
 Max.   :16.41   Max.   :0.07515
Standard deviations:
         mu          tau
0.448562708 0.006743413
```

We can also calculate other features of the joint posterior distribution such as its correlation

$$Corr(\mu, \tau|\boldsymbol{x}) = -0.002706$$

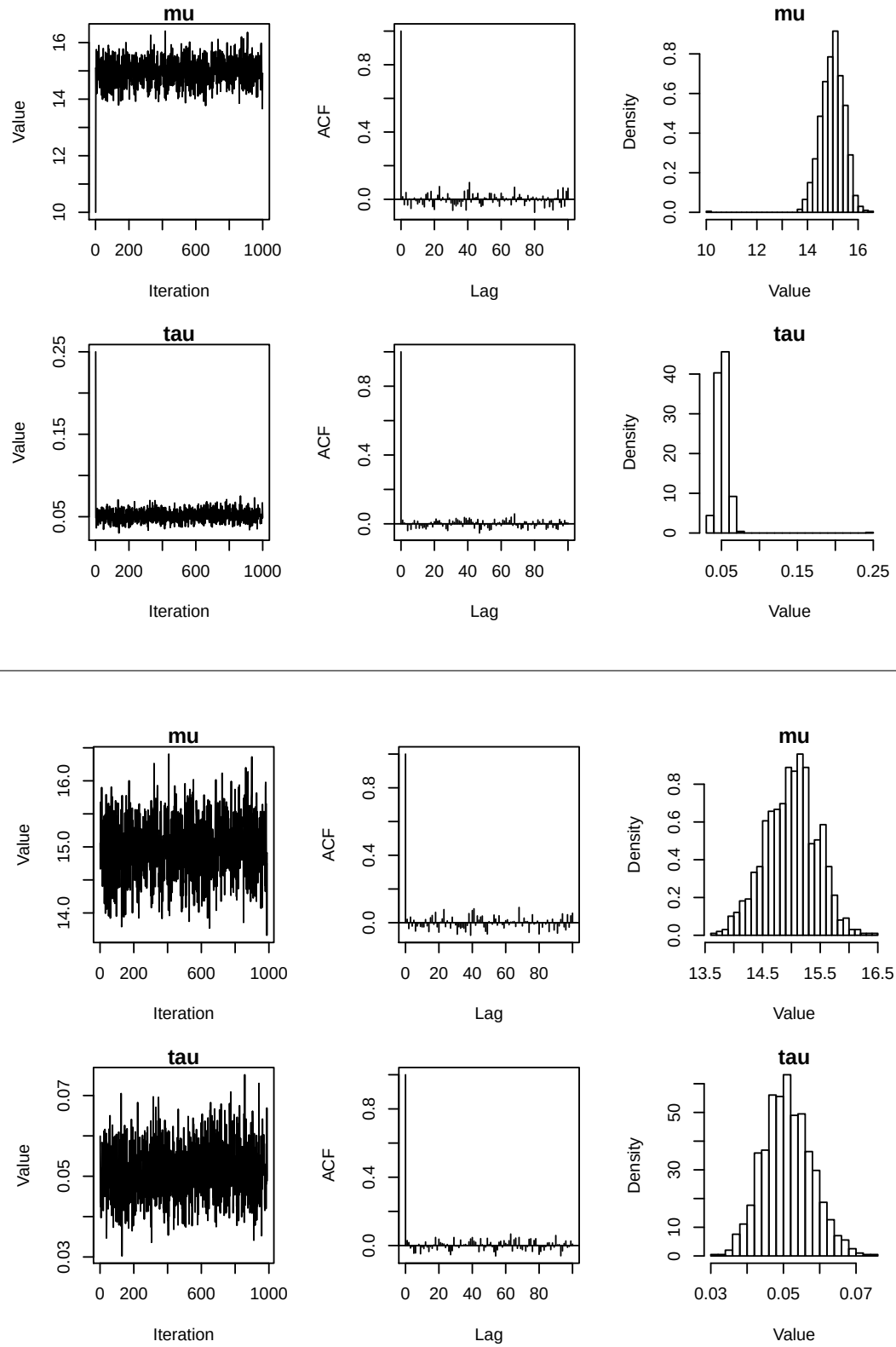


Figure 4.7: Trace plots, autocorrelation plots and histograms of the Gibbs sampler output. Upper plots: all realisations. Lower plots: after deleting the first 10 iterations

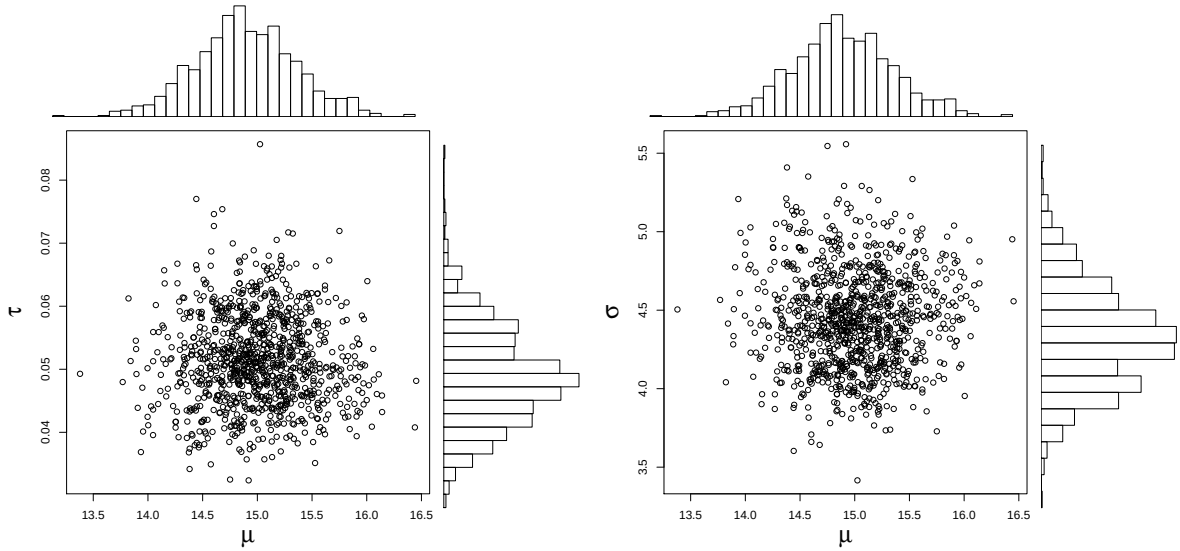


Figure 4.8: Plot of the bivariate posterior sample and their marginal distributions. Left plot: $(\mu, \tau)^T$; right plot: $(\mu, \sigma)^T$.

using the command `cor(posterior2)`, and summarise the posterior sample with a plot of its values and its marginal distributions; see Figure 4.8. We can also determine $100(1-\alpha)\%$ equi-tailed confidence intervals as follows. Suppose we have N realisations from our Gibbs sampler. If we sort the values into increasing order then the confidence interval will have end points which are $N\alpha/2$th and $N(1-\alpha/2)$th values. The `nclbayes` package has a function `mcmcCi` to do this. In this case we would use `mcmcCi(posterior2,level=0.95)` and, for this output, obtain the 95% confidence intervals as

$$\begin{aligned} \mu: &\quad (14.071,\ 15.803) \\ \tau: &\quad (0.03818,\ 0.06499). \end{aligned}$$

Now we have a sample from the posterior distribution, we can determine the posterior distribution for any function of the parameters. For example, if we want the posterior distribution for $\sigma = 1/\sqrt{\tau}$ then we can easily obtain realisations of σ as $\sigma^{(j)} = 1/\sqrt{\tau^{(j)}}$, from which we can produce a plot of its values and its marginal distributions (see Figure 4.8) and also obtain its numerical summaries

```
> sigma=1/sqrt(posterior2[,2])
> summary(sigma)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.648   4.242   4.442   4.453   4.631   5.750
> sd(sigma)
[1] 0.2977992
> quantile(sigma,probs=c(0.025,0.975))
    2.5%    97.5%
3.922480 5.109632
```

Summary

We can use the (converged and thinned) MCMC output to do the following.

- Obtain the posterior distribution for any (joint) functions of the parameters, such as $\sigma = 1/\sqrt{\tau}$ or $(\theta_1 = \mu - \tau, \theta_2 = e^{\mu+\tau/2})^T$
- Look at bivariate posterior distributions via scatter plots
- Look at univariate marginal posterior distributions via histograms or boxplots
- Obtain numerical summaries such as the mean, standard deviation and confidence intervals for single variables and correlations between variables.

Example 4.3

Gibbs sampling can also be used when using a conjugate prior. Construct a Gibbs sampler for the problem in Example 2.2 analysing Cavendish's data on the earth's density. Recall this assumed the data were a random sample from a normal distribution with unknown mean μ and precision τ, that is, $X_i|\mu, \tau \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), and took a conjugate NGa prior distribution for $(\mu, \tau)^T$.

Solution

The joint posterior density is, for $\mu \in \mathbb{R}$, $\tau > 0$

$$\pi(\mu, \tau|\boldsymbol{x}) \propto \tau^{G-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[C(\mu - B)^2 + 2H\right]\right\}$$

where

$$B = \frac{bc + n\bar{x}}{c+n}, \qquad C = c + n,$$
$$G = g + \frac{n}{2}, \qquad H = h + \frac{cn(\bar{x} - b)^2}{2(c+n)} + \frac{ns^2}{2}.$$

Therefore the conditional posterior density for μ is, for $\mu \in \mathbb{R}$

$$\begin{aligned}
\pi(\mu|\tau, \boldsymbol{x}) &= \frac{\pi(\mu, \tau|\boldsymbol{x})}{\pi(\tau|\boldsymbol{x})} \\
&\propto \pi(\mu, \tau|\boldsymbol{x}) \\
&\propto \tau^{G-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[C(\mu - B)^2 + 2H\right]\right\} \\
&\propto \exp\left\{-\frac{C\tau}{2}(\mu - B)^2\right\}
\end{aligned}$$

after moving all multiplicative constants not involving μ into the proportionality sign, and so

$$\mu|\tau, \boldsymbol{x} \sim N\left(B, \frac{1}{C\tau}\right).$$

(Actually we already knew this from the definition of the NGa distribution.)

The conditional posterior density for τ is, for $\tau > 0$

$$\begin{aligned}\pi(\tau|\mu, \boldsymbol{x}) &= \frac{\pi(\mu, \tau|\boldsymbol{x})}{\pi(\mu|\boldsymbol{x})} \\ &\propto \pi(\mu, \tau|\boldsymbol{x}) \\ &\propto \tau^{G-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[C(\mu - B)^2 + 2H\right]\right\}\end{aligned}$$

and so

$$\tau|\mu, \boldsymbol{x} \sim Ga\left(G + \frac{1}{2},\ H + \frac{C}{2}(\mu - B)^2\right).$$

We will initialise the algorithm using the prior means: $\mu^{(0)} = b$ and $\tau^{(0)} = g/h$. Therefore the Gibbs sampler is

1. Initialise the iteration counter to $j = 1$.
   Initialise the state of the chain to $\mu^{(0)} = b$ and $\tau^{(0)} = g/h$.
2. Obtain new values $\mu^{(j)}$ and $\tau^{(j)}$ from $\mu^{(j-1)}$ and $\tau^{(j-1)}$ via

$$\begin{aligned}\mu^{(j)} &\sim N\left(B, \frac{1}{C\tau^{(j-1)}}\right) \\ \tau^{(j)} &\sim Ga\left(G + \frac{1}{2},\ H + \frac{C}{2}\left(\mu^{(j)} - B\right)^2\right)\end{aligned}$$

3. Change counter j to $j + 1$, and return to step 2.

The R function `gibbsNormal2` in the library `nclbayes` implements this Gibbs sampling algorithm. Consider again the analysis of Cavendish's measurements on the earth's density in Example 2.2. These data gave $n = 23$, $\bar{x} = 5.4848$, $s = 0.1882$ and this information was combined with a $NGa(b = 5.41, c = 0.25, g = 2.5, h = 0.1)$ prior distribution to give a $NGa(B = 5.4840, C = 23.25, G = 14, H = 0.5080)$ posterior distribution. Here we analyse the data using a Gibbs sampler and verify that it gives the same results. For example, we know that the marginal posterior distributions are

$$\mu|\boldsymbol{x} \sim t_{2G=28}(B = 5.4840, H/(GC) = 0.001561)$$

and

$$\tau|\boldsymbol{x} \sim Ga(G = 14, H = 0.5080),$$

and so we can compare the Gibbs output with these distributions. The following code runs this Gibbs sampler for this problem.

```
library(nclbayes)
posterior=gibbsNormal2(N=1010,initial=c(5.41,25),
    priorparam=c(5.41,0.25,2.5,0.1),n=23,xbar=5.4848,s=0.1882)
posterior2=mcmcProcess(input=posterior,burnin=10,thin=1)

mcmcAnalysis(posterior,rows=2,show=F)
mcmcAnalysis(posterior2,rows=2,show=F)
```

Figure 4.9 shows the summary of the Gibbs sampler output after deleting the first 1000 iterations as burn-in. The traceplots look like the sampler has converged: they indicate a well mixing chain with similar means and variances in different sections of the chain. Also the autocorrelation plots show that no thinning is needed.

These realisations from the posterior distribution can be summarised using R function `mcmcAnalysis` as

```
N = 1000 iterations
       mu              tau
 Min.   :5.342   Min.   :10.44
 1st Qu.:5.460   1st Qu.:22.81
 Median :5.484   Median :27.70
 Mean   :5.485   Mean   :28.09
 3rd Qu.:5.510   3rd Qu.:32.80
 Max.   :5.619   Max.   :53.66
Standard deviations:
        mu        tau
0.03936649 7.39551702
```

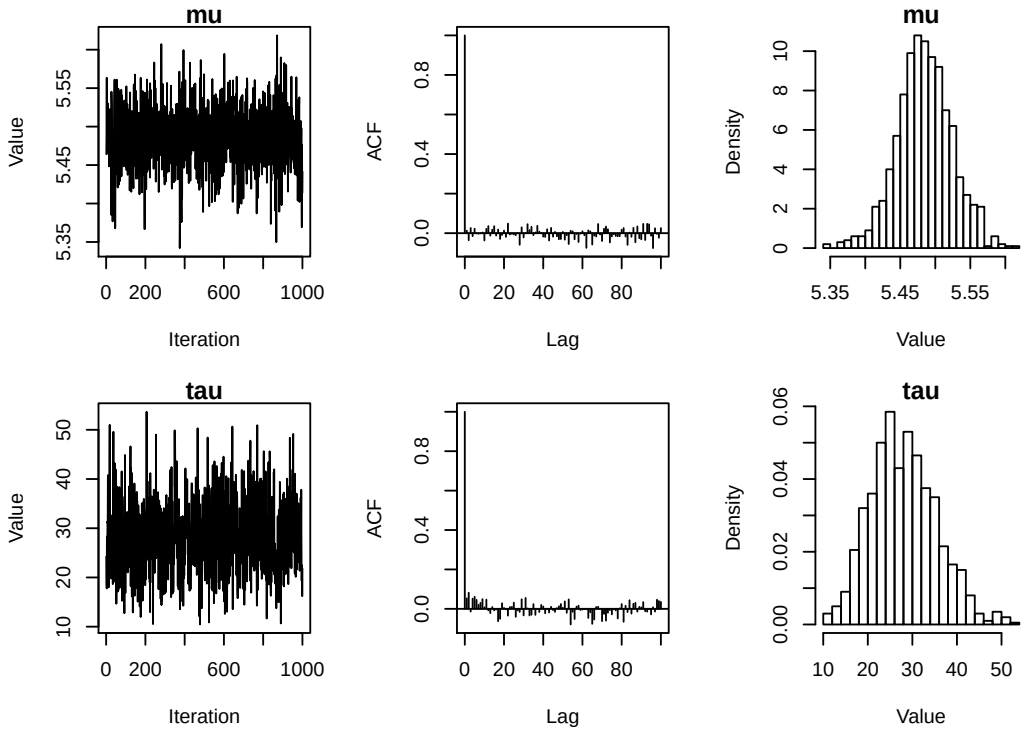


Figure 4.9: Trace plots, autocorrelation plots and histograms of the Gibbs sampler output

These posterior summaries are pretty accurate as we know the correct summaries are

$$E(\mu|\boldsymbol{x}) = B = 5.4840, \qquad SD(\mu|\boldsymbol{x}) = \sqrt{\frac{H}{(G-1)C}} = 0.04100$$

$$E(\tau|\boldsymbol{x}) = \frac{G}{H} = 27.559, \qquad SD(\tau|\boldsymbol{x}) = \frac{\sqrt{G}}{H} = 7.3655.$$

We could obtain even more accurate estimates for these posterior summaries by running the sampler for more iterations. Figure 4.10 shows that the histograms of the Gibbs sampler output are also very close to the (known) marginal posterior densities. These results confirm that our Gibbs sampler is working correctly and does indeed produce realisations from the correct posterior distribution.

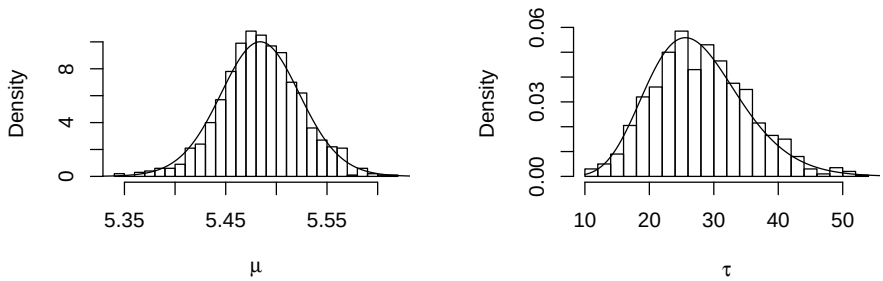


Figure 4.10: Histograms of the Gibbs sampler output and the (known) marginal densities

Example 4.4

Suppose we have a random sample of size n from a gamma $Ga(\alpha, \lambda)$ distribution in which both the index $\alpha > 0$ and scale parameter $\lambda > 0$ are unknown, that is, $X_i|\alpha, \lambda \sim Ga(\alpha, \lambda)$, $i = 1, 2, \ldots, n$ (independent). We shall assume independent prior distributions for these parameters, with $\alpha \sim Ga(a, b)$ and $\lambda \sim Ga(c, d)$ for known values a, b, c and d. Determine the posterior density for $(\alpha, \lambda)^T$ and hence the posterior conditional densities for $\alpha|\lambda$ and $\lambda|\alpha$.

Solution

The likelihood function is

$$\begin{aligned} f(\boldsymbol{x}|\alpha,\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{\alpha} x_i^{\alpha-1} e^{-\lambda x_i}}{\Gamma(\alpha)} \\ &= \frac{\lambda^{n\alpha} \left(\prod x_i\right)^{\alpha-1} e^{-n\bar{x}\lambda}}{\Gamma(\alpha)^n} \\ &= \frac{\lambda^{n\alpha} \bar{x}_g^{n(\alpha-1)} e^{-n\bar{x}\lambda}}{\Gamma(\alpha)^n} \end{aligned}$$

where $\bar{x}_g = \sqrt[n]{\prod x_i}$ is the geometric mean of the data. Using Bayes Theorem, the posterior density is

$$\pi(\alpha,\lambda|\boldsymbol{x}) \propto \pi(\alpha,\lambda)\, f(\boldsymbol{x}|\alpha,\lambda)$$

and so, for $\alpha > 0$, $\lambda > 0$

$$\begin{aligned} \pi(\alpha,\lambda|\boldsymbol{x}) &\propto \alpha^{a-1} e^{-b\alpha} \times \lambda^{c-1} e^{-d\lambda} \times \frac{\lambda^{n\alpha} \bar{x}_g^{n(\alpha-1)} e^{-n\bar{x}\lambda}}{\Gamma(\alpha)^n} \\ &\propto \frac{\alpha^{a-1} \bar{x}_g^{n\alpha} \lambda^{c+n\alpha-1}}{\Gamma(\alpha)^n} \exp\left\{-b\alpha - (d+n\bar{x})\lambda\right\}. \end{aligned}$$

This is not the density of a standard distribution. The conditional posterior density for α is, for $\alpha > 0$

$$\begin{aligned} \pi(\alpha|\lambda,\boldsymbol{x}) &= \frac{\pi(\alpha,\lambda|\boldsymbol{x})}{\pi(\lambda|\boldsymbol{x})} \\ &\propto \pi(\alpha,\lambda|\boldsymbol{x}) \\ &\propto \frac{\alpha^{a-1} \bar{x}_g^{n\alpha} \lambda^{c+n\alpha-1}}{\Gamma(\alpha)^n} \exp\left\{-b\alpha - (d+n\bar{x})\lambda\right\} \\ &\propto \frac{\alpha^{a-1} \bar{x}_g^{n\alpha} \lambda^{n\alpha} e^{-b\alpha}}{\Gamma(\alpha)^n} \\ &\propto \frac{\alpha^{a-1} e^{-(b-n\log\bar{x}_g - n\log\lambda)\alpha}}{\Gamma(\alpha)^n}. \end{aligned}$$

This looks like it might be a gamma density but (i) we would require $b - n\log\bar{x}_g - n\log\lambda > 0$ which can't be guaranteed (ii) the divisor term is wrong: it would have to be $\Gamma(a)$ for this to be a gamma density.

The conditional posterior density for λ is, for $\lambda > 0$

$$\begin{aligned} \pi(\lambda|\alpha,\boldsymbol{x}) &= \frac{\pi(\alpha,\lambda|\boldsymbol{x})}{\pi(\alpha|\boldsymbol{x})} \\ &\propto \pi(\alpha,\lambda|\boldsymbol{x}) \\ &\propto \frac{\alpha^{a-1} \bar{x}_g^{n\alpha} \lambda^{c+n\alpha-1}}{\Gamma(\alpha)^n} \exp\left\{-b\alpha - (d+n\bar{x})\lambda\right\} \\ &\propto \lambda^{c+n\alpha-1} e^{-(d+n\bar{x})\lambda}, \end{aligned}$$

that is, $\lambda|\alpha,\boldsymbol{x} \sim Ga(c+n\alpha,\, d+n\bar{x})$.

The full conditional distribution (FCD) for λ is a standard distribution and so it is straightforward to simulate from this distribution. However, the FCD for α is not a standard distribution and not easy to simulate from. Therefore we cannot use a Gibbs sampler to simulate from the posterior $\pi(\alpha,\lambda|\boldsymbol{x})$. Fortunately, there are other methods available to help us in these situations.

4.5 Metropolis-Hastings sampling

The Gibbs sampler is a very powerful tool but is only useful if the full conditional distributions (FCDs) are standard distributions (which are easy to simulate from). Fortunately there is a class of methods which can be used when the FCDs are non-standard. These methods are known as Metropolis-Hastings schemes.

Suppose we want to simulate realisations from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ and all of the FCDs are non-standard. Suppose further that we have a *proposal distribution* with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, which is easy to simulate from. This distribution gives us a way of proposing new values $\boldsymbol{\theta}^*$ from the current value $\boldsymbol{\theta}$.

Consider the following algorithm:

1. Initialise the iteration counter to $j = 1$, and initialise the chain to $\boldsymbol{\theta}^{(0)}$.
2. Generate a *proposed* value $\boldsymbol{\theta}^*$ using the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})$.
3. Evaluate the *acceptance probability* $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$ of the proposed move, where
$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})\, q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}|\boldsymbol{x})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}.$$
4. Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$, and set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ otherwise.
5. Change the counter from j to $j + 1$ and return to step 2.

In other words, at each stage, a new value is generated from the proposal distribution. This is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is. Whether or not the move is accepted or rejected depends on an acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution. Note that the posterior density $\pi(\cdot|\boldsymbol{x})$ only enters into the acceptance probability as a ratio, and so the method can be used when it is known up to a scaling constant, that is,
$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)\, f(\boldsymbol{x}|\boldsymbol{\theta}^*)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\},$$
since
$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta})\, f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x})}.$$

It can be shown that the above algorithm defines a Markov chain with $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ as its stationary distribution.

Notice that the above description holds for all possible proposal distributions (subject to them generating realisations from the full parameter space). But are some choices better than others? We now discuss some commonly used proposal distributions.

4.5.1 Symmetric chains (Metropolis method)

The simplest case is the Metropolis sampler and uses a symmetric proposal distribution, that is, one with $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$, $\forall\ \boldsymbol{\theta}, \boldsymbol{\theta}^*$. In this case the acceptance probability simplifies to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})}{\pi(\boldsymbol{\theta}|\boldsymbol{x})}\right\},$$

and hence does not involve the proposal density at all. Consequently proposed moves which will take the chain to a region of higher posterior density are always accepted, while moves which take the chain to a region of lower posterior density are accepted with probability proportional to the ratio of the two densities — moves which will take the chain to a region of very low density will be accepted with very low probability. Note that any proposal of the form $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = f(|\boldsymbol{\theta}^* - \boldsymbol{\theta}|)$ is symmetric, where $f(\cdot)$ is some zero mean density function, as $|\boldsymbol{\theta}^* - \boldsymbol{\theta}| = |\boldsymbol{\theta} - \boldsymbol{\theta}^*|$. In this case, the proposal value is a symmetric displacement from the current value. This motivates the following.

Random walk proposals

Consider the random walk proposal in which the proposed value $\boldsymbol{\theta}^*$ depends on the current value $\boldsymbol{\theta}$ via

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{w},$$

where $\boldsymbol{w}$ is a random $p \times 1$ vector from the zero mean density $f(\cdot)$ which is *symmetric* about its mean, and is independent of the state of the chain. We can generate our *proposal* value by first simulating an *innovation* $\boldsymbol{w}$, and then set the *proposal* value to $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{w}$. Clearly $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = f(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = f(\boldsymbol{w})$. Also $\boldsymbol{\theta} = \boldsymbol{\theta}^* - \boldsymbol{w}$ and so $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = f(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = f(-\boldsymbol{w})$. However, as $f(\cdot)$ is a zero mean symmetric density, we have that $f(\boldsymbol{w}) = f(-\boldsymbol{w})$ and so $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$.

But, what distribution should we use for $f(\cdot)$? A distribution which is simple and easy to simulate from would be good, with obvious choices of the uniform or normal distributions, though the normal distribution is generally better, but is a bit more expensive to simulate. What variance should we use for the distribution we choose? This choice will affect the acceptance probability, and hence the overall proportion of accepted moves. If the variance of the innovation is too low, then most proposed values will be accepted, but the chain will move very slowly around the space — the chain is said to be too "cold". On the other hand, if the variance of the innovation is too large, very few proposed values will be accepted, but when they are, they will often correspond to quite large moves — the chain is said to be too "hot". Theoretically it has been shown that the optimal acceptance rate is around 0.234 — this is an asymptotic result (for large samples of data) — but experience suggests that an acceptance rate of around 20–30% is okay. Thus, the variance of the innovation distribution should be "tuned" to get an acceptance rate of around this level.

Normal random walk proposals

A symmetric normal random walk proposal takes the form $\theta^*|\theta \sim N(\theta, k^2)$ for some innovation size $k > 0$. This is a symmetric proposal because $\theta^* = \theta + w$, where $w \sim N(0, k^2)$ has a density which is symmetric about zero. Also the proposal ratio is

$$\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} = \frac{\frac{1}{\sqrt{2\pi k^2}} \exp\left\{-\frac{(\theta-\theta^*)^2}{2k^2}\right\}}{\frac{1}{\sqrt{2\pi k^2}} \exp\left\{-\frac{(\theta^*-\theta)^2}{2k^2}\right\}} = 1.$$

Uniform random walk proposals

A symmetric uniform random walk proposal takes the form $\theta^*|\theta \sim U(\theta - a, \theta + a)$ for some innovation size $a > 0$. This is a symmetric proposal because $\theta^* = \theta + w$, where $w \sim U(-a, a)$ has a density which is symmetric about zero. Also

$$\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} = \frac{1/(2a)}{1/(2a)} = 1.$$

Example 4.5

Suppose the posterior distribution is a standard normal distribution, with density $\phi(\cdot)$. Construct a Metropolis–Hastings algorithm which samples this posterior distribution by using a uniform random walk proposal. Examine how the acceptance rate for this algorithm depends on the width of the uniform distribution.

Solution

We have $\pi(\theta|\boldsymbol{x}) = \phi(\theta)$ and will use a symmetric uniform random walk proposal $\theta^*|\theta \sim U(\theta - a, \theta + a)$ for some choice of $a > 0$. The acceptance probability is

$$\begin{aligned}\alpha(\theta, \theta^*) &= \min\left\{1, \frac{\pi(\theta^*|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x})}\right\} = \min\left\{1, \frac{\phi(\theta^*)}{\phi(\theta)}\right\} \\ &= \min\left\{1, e^{(\theta^2-\theta^{*2})/2}\right\}.\end{aligned}$$

Therefore the algorithm is

1. Initialise the iteration counter to $j = 1$, and initialise the chain to $\theta^{(0)} = 0$ say.
2. Generate a proposed value $\theta^* \sim U(\theta^{(j-1)} - a, \theta^{(j-1)} + a)$.

3. Evaluate the acceptance probability $\alpha(\theta^{(j-1)}, \theta^*) = \min\left\{1, e^{(\theta^{(j-1)^2} - \theta^{*2})/2}\right\}$
4. Set $\theta^{(j)} = \theta^*$ with probability $\alpha(\theta^{(j-1)}, \theta^*)$, and set $\theta^{(j)} = \theta^{(j-1)}$ otherwise.
5. Change the counter from j to $j+1$ and return to step 2.

Of course, in practice we would never simulate from a standard normal distribution using this M–H algorithm as there are much more efficient methods (like the one used in `rnorm`). The purpose here was to illustrate the general method using a very simple choice of posterior distribution.

The `R` function `metropolis` in the library `nclbayes` implements this Metropolis algorithm. The following code runs this algorithm, taking the population mean as its initial value and taking $a = 6$:

```
posterior=metropolis(N=10000,initial=0,a=6)
mcmcAnalysis(posterior,rows=1,show=F)
```

Figure 4.11 shows the output from runs of the algorithm for 10k iterations using different values of a. The top row uses $a = 0.6$ and this chain is too "cold": the innovations are too small and are generally accepted. The acceptance rate for this chain was 0.881. Notice that the autocorrelations are too high and this chain would have to be thinned. Increasing the size of the innovations to $a = 6$ gives the output on the middle row. The autocorrelations are much lower and the acceptance rate was 0.260 (nearer the asymptotic 0.234 M–H acceptance rate). Increasing the size still further to $a = 60$ gives the output on the bottom row. This chain is too "hot" with few proposed values being accepted (acceptance rate 0.027), but when they are, it results in a fairly large move to the chain. This gives fairly high autocorrelations and this chain would have to be thinned.

Normal random walk proposals

Suppose we decide to use a normal random walk with $f(\cdot) = N_p(0, \Sigma)$ and so the proposal distribution is

$$\boldsymbol{\theta}^* | \boldsymbol{\theta} \sim N_p(\boldsymbol{\theta}, \Sigma).$$

Tuning this random walk requires us to choose a value for the covariance matrix Σ. If the posterior distribution is approximately normally distributed (as it is with large data samples) then researchers have shown that the optimal choice is

$$\Sigma = \frac{2.38^2}{p} Var(\boldsymbol{\theta}|\boldsymbol{x}).$$

In practice, of course, we don't know the posterior variance $Var(\boldsymbol{\theta}|\boldsymbol{x})$. However, we could first run the MCMC algorithm substituting in the (generally much larger) prior variance $Var(\boldsymbol{\theta})$. If this chain doesn't converge quickly then we can use its output to get a better idea of $Var(\boldsymbol{\theta}|\boldsymbol{x})$ and run the MCMC code again – this will have more appropriate values for the parameter variances and correlations.

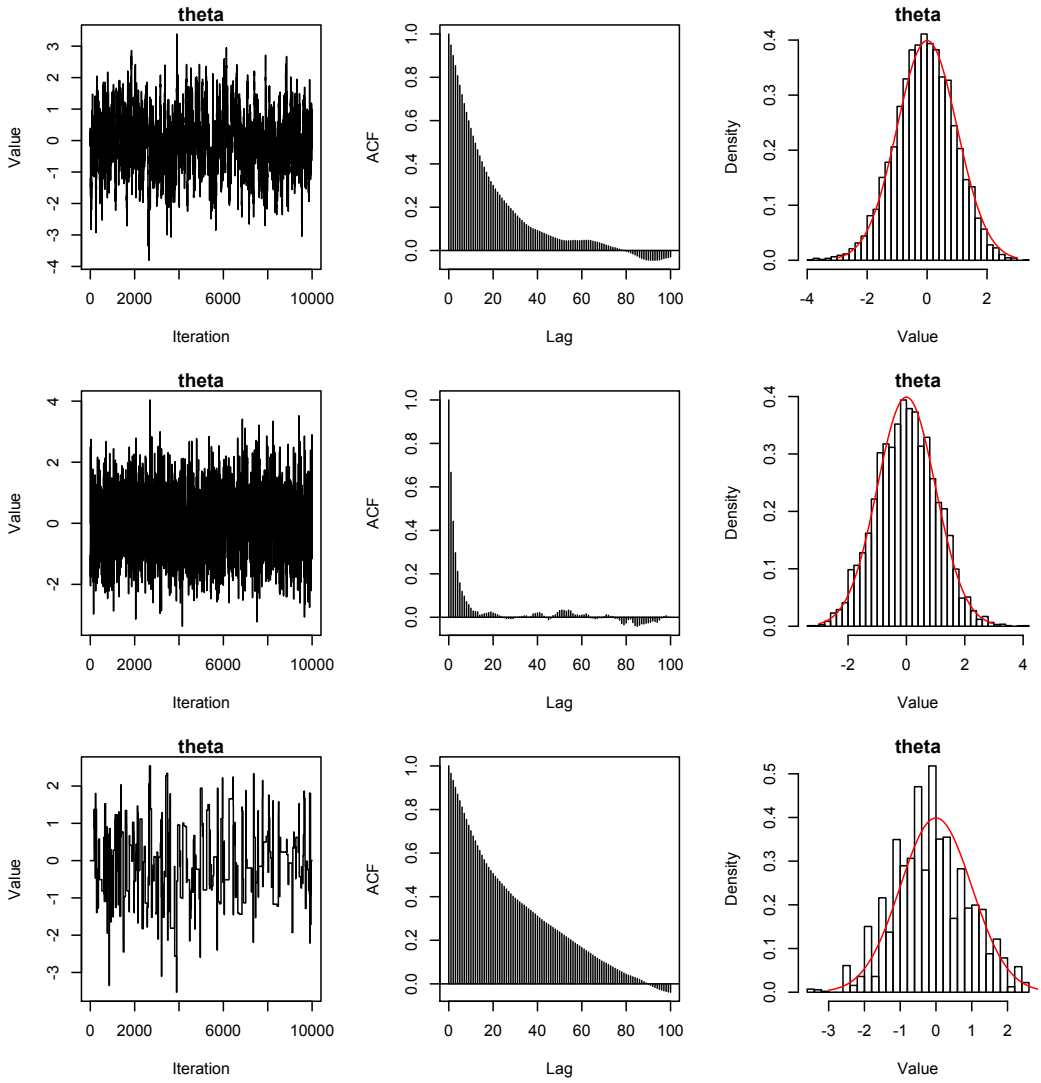


Figure 4.11: Trace plots, autocorrelation plots and histograms of the output from a Metropolis–Hastings sampler using a $U(-a, a)$ random walk proposal. Top row: $a = 0.6$; middle row: $a = 6$; bottom row: $a = 60$

It has been shown from experience that it is not vital to get an extremely accurate value for Σ. Often just getting the correct order of magnitude for its elements will be sufficient, that is, using say 0.1 rather than 0.01 or 1.

4.5.2 Independence chains

In this case, the proposal is formed independently of the position of the chain, and so $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = f(\boldsymbol{\theta}^*)$ for some density $f(\cdot)$. Here the acceptance probability is

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})\, f(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\boldsymbol{x})\, f(\boldsymbol{\theta}^*)}\right\} \\ &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})}{f(\boldsymbol{\theta}^*)} \Big/ \frac{\pi(\boldsymbol{\theta}|\boldsymbol{x})}{f(\boldsymbol{\theta})}\right\},\end{aligned}$$

and we see that the acceptance probability can be increased by making $f(\cdot)$ as similar to $\pi(\cdot|\boldsymbol{x})$ as possible. In this case, the higher the acceptance probability, the better.

Bayes Theorem via independence chains

One possible choice for the proposal density is the prior density. The acceptance probability is then

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{f(\boldsymbol{x}|\boldsymbol{\theta}^*)}{f(\boldsymbol{x}|\boldsymbol{\theta})}\right\},$$

and hence depends only on the likelihood ratio of the proposal and the current value.

4.6 Hybrid methods

We have now seen how we can use the Gibbs sampler to sample from multivariate distributions provided that we can simulate from the full conditional distributions. We have also seen how we can use Metropolis-Hastings methods to sample from awkward FCDs. If we wish, we can combine these in order to form hybrid Markov chains whose stationary distribution is a distribution of interest.

4.6.1 Componentwise transitions

Given a posterior distribution with full conditional distributions that are awkward to sample from directly, we can define a Metropolis-Hastings scheme for each full conditional distribution, and apply them to each component in turn for each iteration. This is like the Gibbs sampler, but each component update is a Metropolis-Hastings update, rather than a direct simulation from the full conditional distribution. Each of these steps will require its own proposal distribution. The algorithm is as follows:

1. Initialise the iteration counter to $j = 1$.
   Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$.

2. Let $\boldsymbol{\theta}_{-i}^{(j)} = \left(\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}\right)^T$, $i = 1, 2, \ldots, p$.
   Obtain a new value $\boldsymbol{\theta}^{(j)}$ from $\boldsymbol{\theta}^{(j-1)}$ by successive generation of values

- $\theta_1^{(j)} \sim \pi(\theta_1|\boldsymbol{\theta}_{-1}^{(j)}, \boldsymbol{x})$ using a Metropolis–Hastings step with proposal distribution $q_1(\theta_1|\theta_1^{(j-1)}, \boldsymbol{\theta}_{-1}^{(j)})$
- $\theta_2^{(j)} \sim \pi(\theta_2|\boldsymbol{\theta}_{-2}^{(j)}, \boldsymbol{x})$ using a Metropolis–Hastings step with proposal distribution $q_2(\theta_2|\theta_2^{(j-1)}, \boldsymbol{\theta}_{-2}^{(j)})$

  $\vdots$
- $\theta_p^{(j)} \sim \pi(\theta_p|\boldsymbol{\theta}_{-p}^{(j)}, \boldsymbol{x})$ using a Metropolis–Hastings step with proposal distribution $q_p(\theta_p|\theta_p^{(j-1)}, \boldsymbol{\theta}_{-p}^{(j)})$

3. Change counter j to $j+1$, and return to step 2.

This is in fact the original form of the Metropolis algorithm. Note that the distributions $\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})$ are just the FCDs.

Suppose we decide to use normal random walks for these M–H steps, that is, take $q_i(\theta_i^*|\theta_i, \boldsymbol{\theta}_{-i}^{(j)})$ is a $N(\theta_i, \Sigma_{ij})$ density. What is the appropriate value for Σ_{ij}? As the proposal in step j is targeting the conditional posterior density $\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})$, the optimal choice of Σ_{ij} is

$$\Sigma_{ij} = \frac{2.38^2}{1} Var(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x}) = 2.38^2 Var(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x}).$$

As these (conditional) posterior variances are not known before running the MCMC code, a sensible strategy might be to replace it with the (probably much larger) prior conditional variance $Var(\theta_i|\boldsymbol{\theta}_{-i}^{(j)})$ or even the prior marginal variance $Var(\theta_i)$. Again, recall that these values are to be used as a guide, generally to get the order of magnitude for the innovation variance.

4.6.2 Metropolis within Gibbs

Given a posterior distribution with full conditional distributions, some of which may be simulated from directly, and others which have Metropolis-Hastings updating schemes, the Metropolis within Gibbs algorithm goes through each in turn, and simulates directly from the full conditional, or carries out a Metropolis-Hastings update as necessary. This algorithm is, in fact, just the above algorithm but uses the full conditional distributions as the proposal distributions when they are easy to simulate from. To see this, suppose that we can simulate from the FCD $\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})$ and use this as the proposal distribution, that is, take $\theta_i^* \sim \pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})$. Then the acceptance probability for this step is

$$\begin{aligned}
\alpha(\theta_i, \theta_i^*) &= \min\left\{1, \frac{\pi(\theta_i^*|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})}{\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})} \frac{q(\theta_i|\theta_i^*, \boldsymbol{\theta}_{-i}^{(j)})}{q(\theta_i^*|\theta_i, \boldsymbol{\theta}_{-i}^{(j)})}\right\} \\
&= \min\left\{1, \frac{\pi(\theta_i^*|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})}{\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})} \frac{\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})}{\pi(\theta_i^*|\boldsymbol{\theta}_{-i}^{(j)}, \boldsymbol{x})}\right\} \\
&= \min(1, 1) \\
&= 1,
\end{aligned}$$

that is, we always accept the proposal from the FCD.

Example 4.6

Construct an MCMC scheme for the problem in Example 4.4 where we had a random sample of size n from a gamma $Ga(\alpha, \lambda)$ distribution and independent gamma $Ga(a, b)$ and $Ga(c, d)$ prior distributions for α and λ respectively. Recall that the FCDs were

$$\pi(\alpha|\lambda, \boldsymbol{x}) \propto \frac{\alpha^{a-1} e^{(-b+n\log\bar{x}_g+n\log\lambda)\alpha}}{\Gamma(\alpha)^n}, \quad \alpha > 0$$

and

$$\pi(\lambda|\alpha, \boldsymbol{x}) \propto \lambda^{c+n\alpha-1}\, e^{-(d+n\bar{x})\lambda}, \quad \lambda > 0.$$

Solution

We have that $\lambda|\alpha, \boldsymbol{x} \sim Ga(c + n\alpha,\ d + n\bar{x})$ but the distribution of $\alpha|\lambda, \boldsymbol{x}$ is non-standard. Therefore we will use a Metropolis within Gibbs algorithm which uses a Gibbs step for λ and a M–H step for α. In the M–H step, we will use a normal random walk proposal distribution, with $\alpha^*|\alpha \sim N(\alpha, \Sigma_\alpha)$, in which the proposal α^* has acceptance probability $\min(1, A)$, where

$$\begin{aligned} A &= \frac{\pi(\alpha^*|\lambda, \boldsymbol{x})}{\pi(\alpha|\lambda, \boldsymbol{x})} \\ &= \frac{\alpha^{*a-1} e^{(-b+n\log\bar{x}_g+n\log\lambda)\alpha^*}}{\Gamma(\alpha^*)^n} \times \frac{\Gamma(\alpha)^n}{\alpha^{a-1} e^{(-b+n\log\bar{x}_g+n\log\lambda)\alpha}}, \quad \alpha^* > 0 \\ &= \left(\frac{\alpha^*}{\alpha}\right)^{a-1} \left\{\frac{\Gamma(\alpha)}{\Gamma(\alpha^*)}\right\}^n e^{(-b+n\log\bar{x}_g+n\log\lambda)(\alpha^*-\alpha)}, \quad \alpha^* > 0, \end{aligned}$$

and zero otherwise.

We need to find a value for α or λ to initialise the MCMC algorithm. We could use a value simulated from the prior or use the prior mean. Alternatively we could use the mle but this is rather complicated to determine. However, the moment estimates are rather more straightforward, and equating first and second population and sample moments gives

$$\text{mean}: \quad \frac{\tilde{\alpha}}{\tilde{\lambda}} = \bar{x} \qquad\qquad \text{variance}: \quad \frac{\tilde{\alpha}}{\tilde{\lambda}^2} = s^2$$

and so

$$\tilde{\lambda} = \frac{\bar{x}}{s^2} \qquad\qquad \tilde{\alpha} = \frac{\bar{x}^2}{s^2}.$$

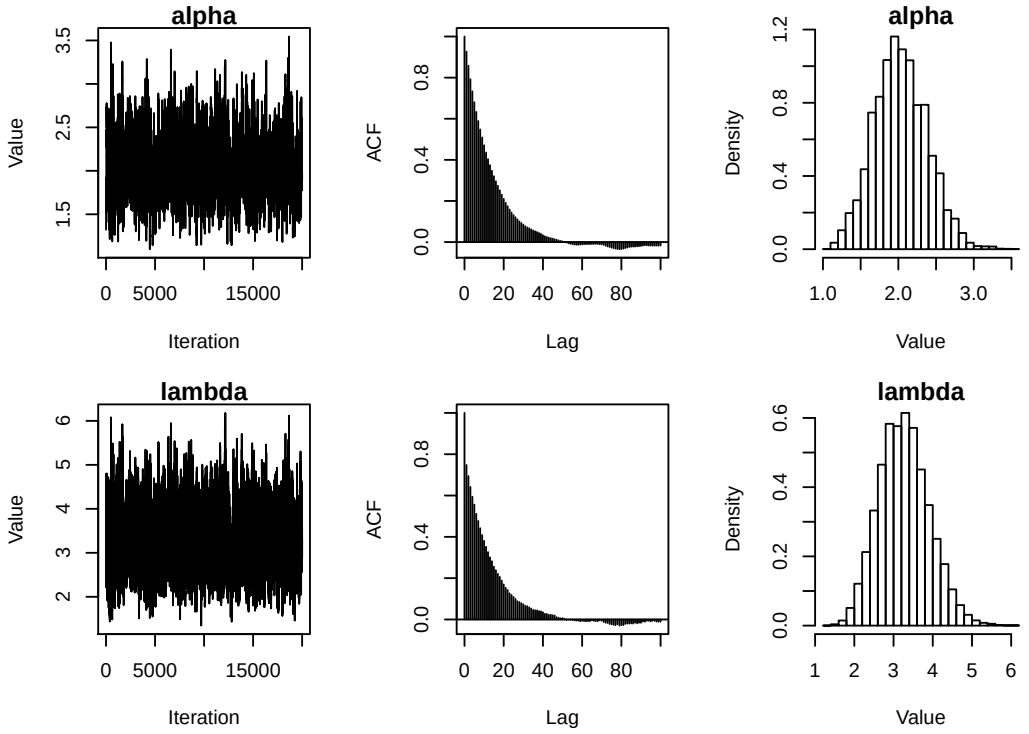
Therefore the MCMC algorithm is

1. Initialise the iteration counter to $j = 1$, and initialise the chain to $\alpha^{(0)} = (\bar{x}/s)^2$.
2. Obtain a new value
$$\lambda^{(j)} \sim Ga(c + n\alpha^{(j-1)},\ d + n\bar{x})$$
3. Generate a proposed value $\alpha^* \sim N(\alpha^{(j-1)}, \Sigma_\alpha)$
4. Evaluate the acceptance probability $\min(1, A)$ at $\alpha^* = \alpha^*$, $\alpha = \alpha^{(j-1)}$ and $\lambda = \lambda^{(j)}$
5. Set $\alpha^{(j)} = \alpha^*$ with probability $\min(1, A)$, and set $\alpha^{(j)} = \alpha^{(j-1)}$ otherwise.
6. Change the counter from j to $j + 1$ and return to step 2.

The R function `mwgGamma` in the library `nclbayes` implements this Metropolis within Gibbs algorithm. The following code produces posterior output from an analysis of a dataset with $n = 50$, $\bar{x} = 0.62$, $\bar{x}_g = 0.46$ and $s = 0.4$, with prior beliefs represented by $a = 2$, $b = 1$, $c = 3$ and $d = 1$, and uses a normal random walk proposal with variance $\Sigma_\alpha = 0.9^2$ as this gives a reasonable acceptance probability of 0.237. The initial value is taken as the moment estimate $\tilde{\alpha} = (\bar{x}/s)^2$.

```
posterior=mwgGamma(N=20000,initial=(0.46/0.4)^2,innov=0.9,
    priorparam=c(2,1,3,1),n=50,xbar=0.62,xgbar=0.46,show=TRUE)
mcmcAnalysis(posterior,rows=2,show=F)
posterior2=mcmcProcess(input=posterior,burnin=10,thin=20)
mcmcAnalysis(posterior2,rows=2)
```

The upper plots in Figure 4.12 show all the output of this MCMC scheme and the lower plots show the output after deleting the first 10 iterations as burn–in and then thinning by only taking every 20th iterate to reduce the autocorrelations.



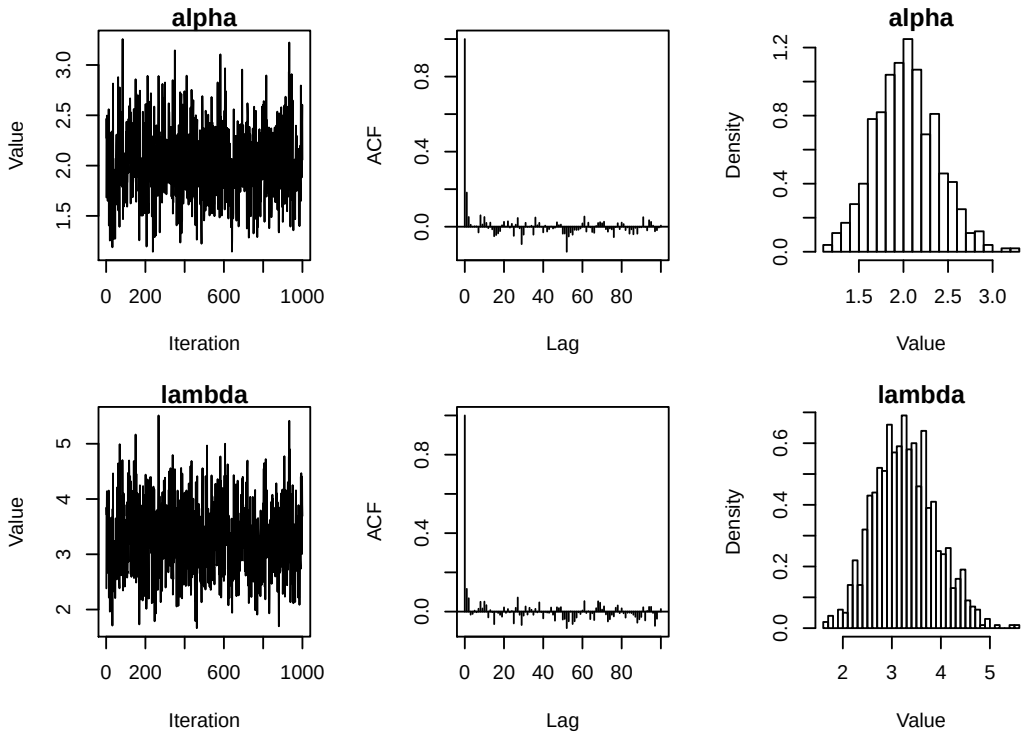


Figure 4.12: Trace plots, autocorrelation plots and histograms of the Metropolis with Gibbs output. Upper plots: all realisations. Lower plots: with burn-in = 10, thin = 20.

Comments

1. If you're unsure whether the proposal distribution is symmetric then it's quite straightforward to examine the proposal ratio. In this last example, we have proposal $\alpha^*|\alpha \sim N(\alpha, \Sigma_\alpha)$ and so

$$\frac{q(\alpha|\alpha^*)}{q(\alpha^*|\alpha)} = \frac{\dfrac{1}{\sqrt{2\pi\Sigma_\alpha}} \exp\left\{-\dfrac{(\alpha-\alpha^*)^2}{2\Sigma_\alpha}\right\}}{\dfrac{1}{\sqrt{2\pi\Sigma_\alpha}} \exp\left\{-\dfrac{(\alpha^*-\alpha)^2}{2\Sigma_\alpha}\right\}} = 1.$$

2. A normal random walk proposal α^* is not accepted if it is negative as, in this case, $A = 0$. This can be wasteful. An alternative is to use a proposal distribution which only generates positive proposal values, such as $\alpha^*|\alpha \sim LN(\log\alpha, \Sigma_\alpha)$. Using this skewed proposal distribution, we have

$$\frac{q(\alpha|\alpha^*)}{q(\alpha^*|\alpha)} = \frac{\dfrac{1}{\alpha\sqrt{2\pi\Sigma_\alpha}} \exp\left\{-\dfrac{(\log\alpha-\log\alpha^*)^2}{2\Sigma_\alpha}\right\}}{\dfrac{1}{\alpha^*\sqrt{2\pi\Sigma_\alpha}} \exp\left\{-\dfrac{(\log\alpha^*-\log\alpha)^2}{2\Sigma_\alpha}\right\}} = \frac{\alpha^*}{\alpha}.$$

   Therefore the acceptance probability for a proposed value α^* is $\min(1, B)$ where

$$\begin{aligned} B &= \frac{\pi(\alpha^*|\lambda, \boldsymbol{x})}{\pi(\alpha|\lambda, \boldsymbol{x})} \times \frac{q(\alpha|\alpha^*)}{q(\alpha^*|\alpha)} \\ &= \frac{\alpha^*\,\pi(\alpha^*|\lambda, \boldsymbol{x})}{\alpha\,\pi(\alpha|\lambda, \boldsymbol{x})}. \end{aligned}$$

   The acceptance probability is still quite straightforward to calculate, and with this proposal distribution we never reject proposal values that are inconsistent with the parameter space (here $\alpha > 0$). Incidentally, $\log X \sim N(\mu, \sigma^2)$ if $X \sim LN(\mu, \sigma^2)$ and so using a log-normal proposal is the same as using a normal random walk on the log scale, that is, a normal random walk for $\log\alpha$. Also log-normal proposals are easy to simulate because if $Y \sim N(\mu, \sigma^2)$ then $e^Y \sim LN(\mu, \sigma^2)$.

3. Dealing with a constraint such as $\alpha > 0$ in optimisation methods or here in MCMC methods, can be solved by re-parameterising the model. Here, for example, we could work with $A = \log\alpha$ and obtain realisations from the posterior distribution for A. Once we have these realisations we can easily obtain realisations from the posterior distribution for $\alpha = e^A$. Working in A rather than α means we have to simulate realisations from the conditional posterior

$$\pi_A(A|\lambda, \boldsymbol{x}) = \pi_\alpha(e^A|\lambda, \boldsymbol{x}) \times \left|\frac{d}{da}\, e^A\right| = e^A\,\pi_\alpha(e^A|\lambda, \boldsymbol{x})$$

   using (2.1). If we also use a normal random walk for proposing new values for A (as

it's unconstrained) then a proposal A^* is accepted with probability $\min(1, C)$ where

$$\begin{aligned} C &= \frac{\pi(A^*|\lambda, \boldsymbol{x})}{\pi(A|\lambda, \boldsymbol{x})} \times \frac{q(A|A^*)}{q(A^*|A)} \\ &= \frac{\pi(A^*|\lambda, \boldsymbol{x})}{\pi(A|\lambda, \boldsymbol{x})} \qquad \text{since the proposal distribution is symmetric about zero} \\ &= \frac{e^{A^*}\,\pi_\alpha(e^{A^*}|\lambda, \boldsymbol{x})}{e^{A}\,\pi_\alpha(e^{A}|\lambda, \boldsymbol{x})} \\ &= \frac{\alpha^*\,\pi_\alpha(\alpha^*|\lambda, \boldsymbol{x})}{\alpha\,\pi_\alpha(\alpha|\lambda, \boldsymbol{x})}. \end{aligned}$$

Notice that the acceptance probabilities B and C are the same, that is, there is no (algorithmic) difference between using a log-normal random walk for a positive parameter or working on the log-scale and using a symmetric normal random walk.

4.7 Summary

(i) Bayesian inference can be complicated when not using a conjugate prior distribution.

(ii) One solution is to use Markov chain Monte Carlo (MCMC) methods.

(iii) These work by producing realisations from the posterior distribution by constructing a Markov chain which has the posterior distribution as its stationary distribution.

(iv) The MCMC methods we have studied are the Gibbs sampler, Metropolis within Gibbs algorithm and the Metropolis–Hastings algorithm.

(v) When obtaining output from these algorithms, we need to assess whether there needs to be a burn-in and whether the output needs to be thinned (by looking at traceplots and autocorrelation plots) using `mcmcAnalysis` and `mcmcProcess`.

(vi) The (converged and thinned) MCMC output are realisations from the posterior distribution. It can be used to

- obtain the posterior distribution for any (joint) functions of the parameters (such as $\sigma = 1/\sqrt{\tau}$ or $(\theta_1 = \mu - \tau, \theta_2 = e^{\mu+\tau/2})^T$);
- look at bivariate posterior distributions via scatter plots;
- look at univariate marginal posterior distributions via histograms or boxplots;
- obtain numerical summaries such as the mean, standard deviation and confidence intervals for single variables and correlations between variables.

(vii) Equi-tailed posterior confidence intervals can be determined from the MCMC output using `mcmcCi`.

4.8 Learning objectives

By the end of this chapter, you should be able to:

- explain why not using a conjugate prior generally causes problems in determining the posterior distribution
- describe the Gibbs sampler, explain why it is a Markov chain and give an outline as to why its stationary distribution is the posterior distribution
- describe the issues of processing MCMC output (burn-in, autocorrelation, thinning etc.) and interpret numerical/graphical output
- derive the full conditional densities for any posterior distribution and name these distributions if they are "standard" distributions given in the notes or on the exam paper
- describe a Metropolis-Hastings algorithm in general terms and when using either symmetric or non-symmetric random walk proposals or independence proposals
- describe the hybrid methods *componentwise transitions* and *Metropolis within Gibbs*
- provide a detailed description of **any** of the MCMC algorithms as they apply to generating realisations from **any** posterior distribution

Appendix A

Summary of distributions

A.1 Distributions for data

Binomial distribution

If $X|\theta \sim Bin(k, \theta)$ then it has probability function

$$f(x|\theta) = \binom{k}{x} \theta^x (1-\theta)^{k-x}, \quad x = 0, 1, \ldots, k,$$

where k is a positive integer and $0 < \theta < 1$. Also, $E(X) = k\theta$ and $Var(X) = k\theta(1-\theta)$.

Exponential distribution

If $X|\lambda \sim Exp(\lambda)$ then it has density

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0$$

where $\lambda > 0$. Also, $E(X) = 1/\lambda$ and $Var(X) = 1/\lambda^2$.

Gamma distribution

If $X|\alpha, \lambda \sim Ga(\alpha, \lambda)$ then it has density

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0,$$

where $\alpha > 0$ and $\lambda > 0$. Also, $E(X) = \alpha/\lambda$ and $Var(X) = \alpha/\lambda^2$.

Inverse Gaussian distribution

If $X|\mu, \lambda \sim IG(\mu, \lambda)$ then it has density

$$f(x|\mu, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0,$$

where $\mu > 0$ and $\lambda > 0$. Also, $E(X) = \mu$ and $Var(X) = \mu^3/\lambda$.

Laplace distribution

If $X|\mu, \sigma \sim La(\mu, \sigma^2)$ then it has density

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2}\,\sigma} \exp\left\{-\frac{\sqrt{2}|x-\mu|}{\sigma}\right\}, \quad x > 0,$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. Also, $E(X) = \mu$ and $Var(X) = \sigma^2$.

Log-normal distribution

If $X|\mu, \sigma \sim LN(\mu, \sigma^2)$ then it has density

$$f(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}, \quad x > 0$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. Also, $E(X) = e^{\mu+\sigma^2/2}$, $Var(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$. Further $E(\log X) = \mu$ and $Var(\log X) = \sigma^2$.

Negative binomial distribution

If $X|\theta \sim NegBin(k, \theta)$ then it has probability function

$$f(x|\theta) = \binom{x-1}{k-1}\theta^k(1-\theta)^{x-k}, \quad x = k, k+1, \ldots ,$$

where k is a positive integer and $0 < \theta < 1$. Also $E(X) = k/\theta$ and $Var(X) = k(1-\theta)/\theta^2$.

Normal distribution

If $X|\mu, \tau \sim N(\mu, 1/\tau)$ then it has density

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\tau > 0$. Also, $E(X) = \mu$ and $Var(X) = 1/\tau$. The distribution has the following quantiles

x	1.2816	1.6449	1.9600	2.3263	2.5758
$\Pr(X < x)$	0.9	0.95	0.975	0.99	0.995

Pareto distribution

If $X|\theta, \alpha \sim Pa(\theta, \alpha)$ then it has density

$$f(x|\alpha, \theta) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, \quad x > \theta$$

where $\alpha > 0$ and $\theta > 0$. Also $E(X) = \alpha\theta/(\alpha - 1)$, $\alpha > 1$ and $Var(X) = \alpha\theta^2/\{(\alpha - 1)^2(\alpha - 2)\}$, $\alpha > 2$.

Poisson distribution

If $X|\theta \sim Po(\theta)$ then it has probability function

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \ldots,$$

where $\theta > 0$. Also, $E(X) = \theta$ and $Var(X) = \theta$.

Rayleigh distribution

If $X|\theta \sim R(\theta)$ then it has density

$$f(x|\theta) = 2x\theta e^{-\theta x^2}, \quad x > 0,$$

where $\theta > 0$. Also $E(X) = \sqrt{\pi/(4\theta)}$ and $Var(X) = (4 - \pi)/(4\theta)$.

Uniform distribution

If $X|\phi, \theta \sim U(\phi, \theta)$ then it has density

$$f(x|\phi, \theta) = \frac{1}{\theta - \phi}, \quad \phi < x < \theta,$$

where $\phi < \theta$. Also, $E(X) = (\phi + \theta)/2$ and $Var(X) = (\theta - \phi)^2/12$.

von Mises distribution

If $X|\mu, \lambda \sim vM(\mu, \lambda)$ then it has density

$$f(x|\mu, \lambda) = \frac{1}{2\pi I_0(\lambda)} \exp\{\lambda \cos(x - \mu)\}, \quad 0 \leq x < 2\pi$$

where $0 \leq \mu < 2\pi$ and $\lambda > 0$. Also $I_0(\cdot)$ is the modified Bessel function of order zero.

Weibull distribution

If $X|\beta, \lambda \sim Wei(\beta, \lambda)$ then it has density

$$f(x|\beta, \lambda) = \beta\lambda^{\beta} x^{\beta-1} e^{-\lambda^{\beta} x^{\beta}}, \quad x > 0,$$

where $\beta > 0$ and $\lambda > 0$. Also, $E(X) = \Gamma(1 + 1/\beta)/\lambda$ and $Var(X) = \Gamma(1 + 2/\beta)/\lambda^2 - E(X)^2$.

A.2 Distributions for prior beliefs

Beta distribution

If $\theta \sim Beta(g, h)$ then it has density

$$\pi(\theta) = \frac{\theta^{g-1}(1-\theta)^{h-1}}{B(g,h)}, \quad 0 < \theta < 1,$$

where $g > 0$ and $h > 0$. Also, $E(\theta) = g/(g+h)$ and $Var(\theta) = gh/\{(g+h)^2(g+h+1)\}$.

Exponential distribution

If $\theta \sim Exp(h)$ then it has density

$$\pi(\theta) = he^{-h\theta}, \quad \theta > 0,$$

where $h > 0$. Also, $E(\theta) = 1/h$ and $Var(\theta) = 1/h^2$.

Gamma distribution

If $\theta \sim Ga(g, h)$ then it has density

$$\pi(\theta) = \frac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0,$$

where $g > 0$ and $h > 0$. Also, $E(\theta) = g/h$ and $Var(\theta) = g/h^2$.

Generalised t distribution

If $\mu \sim t_a(b, c)$ then it has density

$$\pi(\mu) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{ac\pi}\,\Gamma\left(\frac{a}{2}\right)} \left\{1 + \frac{(\mu-b)^2}{ac}\right\}^{-\frac{a+1}{2}}, \quad \mu \in \mathbb{R},$$

where $b \in \mathbb{R}$, $a > 0$ and $c > 0$. Also, $E(\mu) = b$ and $Var(\mu) = ac/(a-2)$ if $a \geq 2$.

Inverse Chi distribution

If $\sigma \sim \textit{Inv-Chi}(a, b)$ then it has density

$$\pi(\sigma|a, b) = \frac{2b^a \sigma^{-2a-1} e^{-b/\sigma^2}}{\Gamma(a)}, \quad \sigma > 0,$$

where $a > 0$, $b > 0$ and $\Gamma(a)$ is the gamma function. Also $E(\sigma) = \sqrt{b}\,\Gamma(a-1/2)/\Gamma(a)$ and $Var(\sigma) = b/(a-1) - E(\sigma)^2$ if $a > 1$. The name of the distribution comes from the fact that $1/\sigma^2 \sim Ga(a, b) \equiv \chi^2_{2a}/(2b)$.

Log-normal distribution

If $\theta \sim LN(b, c^2)$ then it has density

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\, c\, \theta} \exp\left\{-\frac{1}{2c^2}(\log\theta - b)^2\right\}, \quad \theta > 0$$

where $b \in \mathbb{R}$ and $c > 0$. Also, $E(\theta) = e^{b+c^2/2}$, $Var(\theta) = (e^{c^2} - 1)e^{2b+c^2}$. Further $\log\theta \sim N(b, c^2)$ and so $E(\log\theta) = b$ and $Var(\log\theta) = c^2$.

Normal distribution

If $\mu \sim N(b, 1/d)$ then it has density

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}, \quad \mu \in \mathbb{R},$$

where $b \in \mathbb{R}$ and $c > 0$. Also, $E(\mu) = b$ and $Var(\mu) = 1/d$.

Normal-gamma distribution

If $\begin{pmatrix}\mu \\ \tau\end{pmatrix} \sim NGa(b, c, g, h)$ then it has density

$$\pi(\mu, \tau) \propto \tau^{g-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[c(\mu - b)^2 + 2h\right]\right\}, \quad \mu \in \mathbb{R},\ \tau > 0$$

where $b \in \mathbb{R}$ and $c, g, h > 0$. Also, $\mu|\tau \sim N\left(b, \frac{1}{c\tau}\right)$, $\tau \sim Ga(g, h)$ and has marginal distribution $\mu \sim t_{2g}\left(b, \frac{h}{gc}\right)$.

Uniform distribution

If $\theta \sim U(a, b)$ then it has density

$$\pi(\theta) = \frac{1}{b - a}, \quad a < \theta < b,$$

where $a < b$. Also, $E(\theta) = (a + b)/2$ and $Var(\theta) = (b - a)^2/12$.

Exercises

1. Suppose you have a random sample $x_1, x_2, \ldots, x_n$. In the following models, derive the posterior density and name the posterior distribution (and its parameters).

   (i) $f(x|\theta) = \theta^{x-1}(1-\theta)$, $x = 1, 2, \ldots$ with a $Beta(3, 2)$ prior distribution for θ.

   (ii) $f(x|\theta) = \dfrac{e^{-\theta}\theta^x}{x!}$, $x = 0, 1, \ldots$ with a $Exp(2)$ prior distribution for θ.

   (iii) $f(x|\theta) = \dbinom{4}{x}\theta^x(1-\theta)^{4-x}$, $x = 0, 1, 2, 3, 4$ with a $U(0, 1)$ prior distribution for θ.

2. Suppose that a random sample $x_1, x_2, \ldots, x_{10}$ is obtained from a uniform $U(0, \theta)$ distribution. Derive the posterior density for θ assuming a gamma $Ga(20, 1)$ prior distribution θ. Hint: this distribution will depend on the maximum observed x-value x_{max}.

3. Suppose that a random sample $x_1, x_2, \ldots, x_n$ is obtained from a truncated unit exponential distribution, with density

$$f(x|\theta) = \begin{cases} e^{\theta - x}, & x > \theta \\ 0, & \text{otherwise,} \end{cases}$$

   where $\theta \in \mathbb{R}$. Derive the posterior density for θ assuming a normal $N(b, 1/d)$ prior distribution. Hint: this distribution will depend on the minimum observed x-value x_{min}.

4. The dimensions of a component from a long production run vary according to a $N(\mu, 1)$ distribution, and the mean dimension μ varies from production run to production run according to a $N(10, 1/4)$ distribution. From one production run 12 components are drawn at random and their average dimension is found to be $10\frac{1}{3}$. On this information what is the probability that the mean component dimension is at least 10?

5. A trucking company owns a large fleet of well-maintained trucks. Suppose that breakdowns occur at random times. The owner of the company is interested in learning about the daily rate θ at which breakdowns occur. It is known that the number of breakdowns X on a typical day has a Poisson distribution with mean θ. The owner has some knowledge about the rate parameter θ based on the observed number of breakdowns in previous years and expresses these prior beliefs using a

$Ga(4, 2)$ distribution. The daily number of truck breakdowns are obtained on five consecutive days as 3, 2, 3, 1, and 2. Assuming these data are a random sample, determine the posterior distribution of θ.

6. Suppose that the time in minutes required to serve a customer at a certain facility has an exponential distribution for which the parameter θ is unknown, and that the prior distribution of θ is a gamma distribution with mean 0.2 and standard deviation 1. If the average time required to serve a random sample of 20 customers is observed to be 3.8 minutes, determine the posterior distribution of θ.

7. The following data is the time intervals (in minutes) between eruptions of the "Old Faithful" geyser. Note that the average time between eruptions is 67.38 minutes, that is, 1.123 hours. If the eruptions occur randomly in time according to a Poisson process with a rate of θ eruptions per hour then the times between eruptions will form a random sample from an exponential distribution with rate θ. Making this assumption, establish a prior distribution for θ with $E(\theta) = 1$ and $Var(\theta) = 1$. Determine the posterior distribution for θ. Comment on your analysis.

70	64	72	76	80	48	88	53	71	56	69	72	76	54	76
65	54	86	40	87	49	76	51	77	49	71	78	80	51	82
49	80	43	83	49	75	47	78	71	69	63	64	82	68	71
71	63	79	66	75	56	83	67	65	77	72	79	73	53	69
53	78	55	67	68	73	53	70	69	66	79	48	90	49	78
52	79	49	75	75	50	87	40	76	57	71	70	69	72	51
84	43	73	73	70	84	71	79	58	73					

8. The negative binomial distribution is used to model scenarios in which we observe the number of independent (success-fail) trials needed before we see a successful trial; the trials must have the same success probability. Suppose $x_1, x_2, \ldots, x_n$ are a random sample from a negative binomial $NegBin(k, \theta)$ distribution, where k is known.

   (i) Verify that the congugate prior distribution is a Beta distribution.

   (ii) Determine the choice of parameters g and h for the $Beta(g, h)$ distribution that give it maximal variance.
   Hint: reparameterise the distribution in terms of its mean $m = g/(g + h)$ and $s = g + h$; determine the choice of m and s that maximises the variance of the distribution, and hence the choice of g and h.

   Determine the posterior distribution for θ assuming

   (iii) vague prior knowledge;

   (iv) a very large sample.

9. The Rayleigh distribution is often used to measure variability in magnetic resonance imaging (MRI). Suppose $x_1, x_2, \ldots, x_n$ are a random sample from a Rayleigh $R(\theta)$ distribution. Determine the posterior distribution for θ assuming

   (i) vague prior knowledge;

(ii) a very large sample.

10. The Pareto distribution is often used to model data in many areas, ranging from the wealth of individuals to the sizes of meteorites. Suppose $x_1, x_2, \ldots, x_n$ are a random sample from a Pareto $Pa(1, \theta)$ distribution. Determine the posterior distribution for θ assuming

    (i) vague prior knowledge;

    (ii) a very large sample.

11. Suppose that a random sample of size $n = 10$ from an $N(\mu, 1)$ distribution has mean $\bar{x} = 2.5$. The prior distribution for μ is the mixture distribution

$$\mu \sim 0.2\, N(3.3, 0.37^2) + 0.8\, N(1.1, 0.47^2).$$

    (i) Determine the posterior distribution for μ. Note that, for this model

$$f_i(\boldsymbol{x}) = \frac{\pi_i(\mu)\, f(\boldsymbol{x}|\mu)}{\pi_i(\mu|\boldsymbol{x})} \propto \frac{\sqrt{d_i}}{\sqrt{D_i}} \exp\left\{\frac{1}{2}\left[D_i B_i^2 - d_i b_i^2\right]\right\},$$

    where the constant of proportionality doesn't depend on component prior i. In order to get accurate values for the posterior weights, you should calculate posterior component means and standard deviations to at least 4 dp.

    (ii) Calculate the prior and posterior mean and standard deviation.

    (iii) Plot the prior and posterior densities for μ.

    (iv) Calculate the prior and posterior probability that μ exceeds 2.5. Comment on the effect of incorporating the data.

12. Consider the exponential model and gamma mixture prior distribution described in Example **??**. Define your own R functions to calculate the first component weight (p_1^*) and the posterior mean and standard deviation as functions of the sample mean $\bar{x}$. Investigate the behaviour of these functions and give an intuitive explanation of their general properties.

13. Suppose that you have a random sample $x_1, x_2, \ldots, x_n$ from an $N(\mu, \sigma^2)$ distribution and take a $NGa(b, c, g, h)$ prior distribution for $(\mu, \tau)^T$, where $\tau = 1/\sigma^2$. Verify that the posterior mean for μ is greater than the prior mean if and only if the sample mean is greater than the prior mean.

14. Suppose that you have a random sample from an $N(\mu, 1/\tau)$ distribution and believe that the conjugate $NGa(b, c, g, h)$ prior distribution is appropriate for $(\mu, \tau)^T$. Let $t_{\nu,\alpha}$ and $\chi^2_{\nu,\alpha}$ denote the upper α-points of the t_ν and χ^2_ν distributions respectively. Determine

    (i) the equi-tailed 95% Posterior confidence interval for μ;

    (ii) the equi-tailed 95% Posterior confidence interval for τ (hint: if $W \sim Ga(a, b)$ then $2bW \sim \chi^2_{2a}$);

    (iii) a 95% Posterior confidence interval for $\sigma = 1/\sqrt{\tau}$;

(iv) the 95% Posterior HDI for μ.

(v) Why is it not straightforward to determine the 95% HDI for τ? Determine the 95% HDI for τ when the size of the random sample is large.

Compare your answers with the equivalent 95% frequentist confidence intervals:

μ: $(\bar{x} - t_{n-1,0.025}s_u/\sqrt{n}, \bar{x} + t_{n-1,0.025}s_u/\sqrt{n})$
τ: $(\chi^2_{n-1,0.975}/\{(n-1)s_u^2\}, \chi^2_{n-1,0.025}/\{(n-1)s_u^2\})$
σ: $\left(\sqrt{\{(n-1)s_u^2\}/\chi^2_{n-1,0.025}}, \sqrt{\{(n-1)s_u^2\}/\chi^2_{n-1,0.975}}\right)$.

15. Suppose that you have a random sample $x_1, x_2, \ldots, x_n$ from an $N(\mu, 1/\tau)$ distribution and the sample size n is sufficiently large that the posterior distribution for $(\mu, \tau)^T$ is close to its asymptotic form. Determine

(i) the 95% HDI for μ;

(ii) the 95% HDI for τ;

(iii) the 95% HDI for $\sigma = 1/\sqrt{\tau}$.

Compare your answers with the equivalent 95% frequentist confidence intervals:

μ: $(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n})$
τ: $(1/s^2 - 1.96\sqrt{2}/(\sqrt{n}s^2), 1/s^2 + 1.96\sqrt{2}/(\sqrt{n}s^2)$
σ: $(s - 1.96s/\sqrt{2n}, s + 1.96s/\sqrt{2n})$.

16. Suppose that you have a random sample from a $Ga(\alpha, \lambda)$ distribution.

(i) Determine the asymptotic posterior distribution for $\boldsymbol{\theta} = (\alpha, \lambda)^T$. Hint: you should define the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\lambda}$ in terms of the equations they must satisfy, with these involving the sample mean $\bar{x}$, the geometric mean $\bar{x}_g$ and the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$.

(ii) Suppose a sample of size $n = 100$ gives sample mean $\bar{x} = 3.0$ and geometric mean $\bar{x}_g = 2.5$ leading to maximum likelihood estimates $\hat{\alpha} = 2.8983$ and $\hat{\lambda} = 0.9661$. Determine the asymptotic posterior distribution for $\boldsymbol{\theta}$. Also calculate the asymptotic posterior correlation between α and λ. Hint: you will need to use the `R` functions `digamma` for $\psi(\cdot)$ and `trigamma` for $\psi'(\cdot)$.

17. In a calibration experiment, the times between successive emissions from a radioactive source were measured using two techniques: one (X) is very accurate and the other (Y) less precise. Suppose that the observed times $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are a (bivariate) random sample from a distribution in which $X \sim Exp(\lambda)$ and $Y|X = x \sim N(x, \sigma^2)$. Determine the posterior distribution for $(\lambda, \sigma)^T$ when the sample size is large.

18. Consider the Metropolis–Hastings scheme described in Example 4.5 which simulates realisations from the standard normal distribution.

(i) Investigate the relationship between the acceptance rate of the proposals (acc) and the lag 1 autocorrelation (r_1) of the output as the size of the innovation a takes values $1, 2, \ldots, 8$.

(ii) How accurate are your acceptance rates (for each a)? Hint: consider the Bernoulli observations which describe whether a proposal is accepted at each iteration. This model scenario was considered in one of your MAS2903 exercises. Use the asymptotic distribution of the true acceptance probability based on these underlying Bernoulli observations.

(iii) For each value of a, obtain a sample of 1000 realisations which are almost uncorrelated (with $|r_1| < 0.02$) and assess whether each sample is plausibly from the standard normal distribution using a Q–Q plot.

19. Suppose the posterior distribution is a standard normal distribution, with density $\phi(\cdot)$. Construct a Metropolis–Hastings algorithm which samples this posterior distribution by using a normal random walk proposal with standard deviation k.

20. Suppose the posterior distribution is $\theta|\boldsymbol{x} \sim Ga(G, H)$ distribution, with G and H known. Suppose you want to construct a Metropolis–Hastings algorithm which samples this posterior distribution by using a (skewed) $Exp(a/\theta)$ proposal distribution, where $a > 0$, i.e an exponential distribution with its parameter $\lambda = a/\theta$.

    (i) Determine the mean of the proposal distribution.

    (ii) What good feature does this proposal distribution have?

    (iii) Write down the steps in this Metropolis–Hastings algorithm to simulate realisations from the posterior distribution.

21. Suppose the posterior distribution is $\theta|\boldsymbol{x} \sim \textit{Inv-Chi}(G, H)$ distribution, with G and H known. Suppose you want to construct a Metropolis–Hastings algorithm which samples this posterior distribution by using a (skewed) log normal $LN(\log\theta, a^2)$ proposal distribution, where $a > 0$.

    (i) Determine the mean of the proposal distribution.

    (ii) What good feature does this proposal distribution have?

    (iii) Write down the steps in this Metropolis–Hastings algorithm to simulate realisations from the posterior distribution.

22. Rework Qn 21 for the case where interest is about $\beta = \log\theta$: suppose the posterior distribution is $\theta|\boldsymbol{x} \sim \textit{Inv-Chi}(G, H)$ distribution, with G and H known.

    (i) Use equation (2.1) to determine the posterior density of $\beta = \log\theta$.

    Suppose you want to construct a Metropolis–Hastings algorithm which samples the posterior distribution for β by using a normal random walk with variance a^2.

    (ii) Write down the steps in this Metropolis–Hastings algorithm to simulate realisations from the posterior distribution.

    (iii) Show that the acceptance probabilities in these two Metropolis–Hastings algorithms are the same, that is $\alpha(\theta, \theta^*) = \alpha(\beta, \beta^*)$, where $\beta = \log\theta$ and $\beta^* = \log\theta^*$.

23. The Weibull distribution is commonly used to model lifetime data. Suppose we have a random sample $x_1, x_2, \ldots, x_n$ from a Weibull $Wei(\beta, \lambda)$ distribution, with parameters $\beta > 0$ and $\lambda > 0$. Suppose that the prior distribution has $\beta \sim Ga(a, b)$ and $\lambda \sim Ga(c, d)$, independently, for known values a, b, c and d.

    (i) Determine the posterior density for $(\beta, \lambda)^T$ up to a multiplicative constant.

    (ii) Determine the posterior conditional densities for $\beta|\lambda$ and $\lambda|\beta$.

    (iii) Write down the steps in a Metropolis–Hastings algorithm to simulate realisations from the posterior distribution. Your algorithm should have separate steps for each parameter and use normal random walks with variances Σ_β and Σ_λ respectively.

24. The von Mises distribution is commonly used to model circular data, that is, data on the circle such as wind directions. Suppose we have a random sample $x_1, x_2, \ldots, x_n$ from a von Mises $vM(\mu, \lambda)$ distribution with mean direction $\mu \in (0, 2\pi)$ and concentration parameter $\lambda > 0$. Suppose that the prior distribution has $\mu \sim U(0, 2\pi)$ and $\lambda \sim Ga(g, h)$, independently, for known values of g and h.

    (i) Determine the posterior density for $(\mu, \lambda)^T$ up to a multiplicative constant.

    (ii) Determine the posterior densities for $\mu|\lambda$ and $\lambda|\mu$.

    (iii) Write down the steps in Metropolis-Hastings algorithm to simulate realisations from the posterior distribution. Your algorithm should have separate steps for each parameter and use normal random walks with variances Σ_μ and Σ_λ respectively.

    (iv) Using the trigonometric identities $\cos(A - B) = \cos A \cos B + \sin A \sin B$ and $A \cos x + B \sin x = C \cos(x - D)$ where $C = \sqrt{A^2 + B^2}$ and $D = \arctan(B/A)$, show that the update for μ can be achieved using a (more efficient) Gibbs step.

25. A drug company wants to assess the level of side effects from a new drug. A random sample of n people are given the drug and note is taken on the dose level (X) and whether they suffer side effects ($Y = 1$ if yes and $Y = 0$ if no). It is decided that the relationship between dose level and side effects can be described using a linear probit regression model in which

$$Pr(Y = 1|X = x) = \Phi(\beta + \theta x),$$

where $\Phi(\cdot)$ is the standard normal distribution function. Suppose that the prior distribution has $\beta \sim N(a, 1/b^2)$ and $\theta \sim N(c, 1/d^2)$, independently for known values of a, b, c and d.

    (i) Using the data $(y_1, x_1), \ldots, (y_n, x_n)$, determine the likelihood function $f(\boldsymbol{y}|\beta, \theta)$ and hence the posterior density for $(\beta, \theta)^T$ up to a multiplicative constant.

    (ii) Determine the posterior densities for $\beta|\theta$ and $\theta|\beta$.

    (iii) Write down the steps in Metropolis-Hastings algorithm to simulate realisations from the posterior distribution. Your algorithm should have separate steps for each parameter and use normal random walks with variances Σ_β and Σ_θ respectively.

The R package `nclbayes`

The `nclbayes` package is available on all university PC clusters which run R. Please do not install it in your university account.

You can install the package on your own PC by typing (within Rstudio or R)

```
install.packages("nclbayes",repos="http://R-Forge.R-project.org")
```

Note that you only need to install the package once on your own PC – not every time you want to use it.

Distributions

- Inverse Chi distribution [*Inv-Chi(a,b)*]
  density, distribution function, quantile function and random numbers:
  `dinvchi, pinvchi, qinvchi, rinvchi`
- Generalised t distribution [$t_a(b, c)$]
  density, distribution function, quantile function and random numbers:
  `dgt, pgt, qgt, rgt`
- Normal-Gamma distribution [$NGa(b, c, g, h)$]
  density, random numbers, confidence regions and elicitation of parameter values:
  `dnormgamma, rnormgamma, NGacontour, elicitNGa`
- Normal-InvChi distribution
  density and random numbers: `dnorminvchi, rnorminvchi`
- Bivariate Normal distribution [$N_2(\boldsymbol{\mu}, \Sigma)$]
  density: `dbvnorm`
- Bivariate t distribution [$t_a(\boldsymbol{b}, c)$]
  density: `dbvt`

Highest density intervals (HDIs)

- Beta distribution [*Beta(a,b)*]: `hdiBeta`
- Gamma distribution [*Ga(a,b)*]: `hdiGamma`
- Inv-Chi distribution [*Inv-Chi(a,b)*]: `hdiInvchi`

MCMC algorithms

- `gibbsNormal`: Gibbs sampler for a normal random sample with semi-conjugate prior
- `gibbsNormal2`: Gibbs sampler for a normal random sample with conjugate prior
- `gibbsReffects`: Gibbs sampler for a one-way normal random effects model with semi-conjugate prior
- `metropolis`: Metropolis algorithm for simulating from a standard normal distribution
- `mwgGamma`: Metropolis within Gibbs algorithm for a gamma random sample
- `mhReffects`: Metropolis–Hastings algorithm for a one-way normal random effects model using normal and log normal random walk proposals

Other functions

- `mcmcAnalysis`: summarises and plots MCMC output from the above algorithms
- `mcmcProcess`: chops off burnin and then thins MCMC output
- `mcmcCi`: calculates equi-tailed confidence intervals from MCMC output

Demos

- `review`: code used in Chapter 1
- `cavendish`: analyses Cavendish's data on the earth's density
- `gibbs`: code to demonstrate the Gibbs sampler
- `mh`: code to demonstrate the Metropolis-Hastings algorithm
- `sundries`: code used in the notes but not in any other demo

As an example, you can run the demo `cavendish` using `demo(cavendish)` and view the commands in this demo using `demoCommands(cavendish)`.