Christensen, Ronald, Wesley Johnson, Adam Branscum, and Timothy E. Hanson.
Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC press, 2011.

Chapter 6

Simulations

Modern Bayesian analysis is typically performed by simulating the posterior distribution using Markov chain Monte Carlo (MCMC) methods. Monte Carlo methods are a traditional name for simulation methods.

Historically, Bayesian methods were restricted by the need to perform integrations analytically. More recently, approximate Bayesian analysis has been performed by using numerical integrations (Naylor and Smith, 1982; Smith et al., 1985), by using the analytic Laplace approximation (Leonard,1982; Tierney and Kadane, 1986; Kass et al., 1988), and by using Monte Carlo methods (Zellner and Rossi, 1984; Gelfand and Smith, 1990; Dellaportas and Smith, 1993). See Gelman et al. (2004, Chaps. 9–11) for a nice summary of these methods. We prefer Monte Carlo methods to Laplace approximations in regression problems because when performing many predictions, only a single Monte Carlo sample is necessary to perform all predictions, whereas the Laplace method requires a separate analytic approximation for each prediction. We prefer Monte Carlo methods to numerical integration because of their potential to deal with high-dimensional problems.

This chapter provides a short introduction to simulation, specifically, traditional simulation methods such as rejection sampling and importance sampling, as well as Markov chain theory and MCMC methods. Section 1 presents the basics of simulation and Section 2 presents the traditional methods of Acceptance-Rejection, and Importance Sampling. Section 3 presents a short version of the theory of Markov chains and its application to the methods of Gibbs Sampling, Metropolis Algorithm, and Slice Sampling, which are all used in WinBUGS. Section 3 also discusses adaptive rejection sampling and methods of assessing convergence. All readers should study Sections 6.3.0 and 6.3.5. Additional information on these topics can be found in Robert and Casella (2004), Chen, Shao, and Ibrahim (2000), Gilks, Richardson, and Spiegelhalter (1996), and Gamerman and Lopes (2006).

6.1 Generating Random Samples

Let Y be a random variable with strictly monotone cdf $F(y)$ that has an inverse function $F^{-1}(u)$. By definition

$$\Pr(Y \le y_0) = F(y_0).$$

Let U have a $U(0,1)$ distribution so that $\Pr[U \le u_0] = u_0$ and consider the random variable $F^{-1}(U)$.

$$\Pr[F^{-1}(U) \le y_0] = \Pr[U \le F(y_0)] = F(y_0),$$

so Y and $F^{-1}(U)$ have the same distribution. In particular, if you have a way of generating random samples $U_1, \ldots, U_m$ from a Uniform(0,1) distribution, the independent random observations $F^{-1}(U_1), \ldots, F^{-1}(U_m)$ will all have the same distribution as Y.

EXAMPLE 6.1.1. *Exponentials, Gammas, and Betas*. If $Y \sim \text{Exp}(\theta)$ then $F(y) = 1 - e^{-\theta y}$ and $F^{-1}(u) = -\log(1-u)/\theta$. If we generate $U \sim U(0,1)$, $-\log(U)/\theta \sim \text{Exp}(\theta)$, since $U \sim 1 - U$. From inspecting the densities in Table 2.1, it is not difficult to see that $\text{Exp}(\theta) = \text{Gamma}(1,\theta)$.

A well-known property of Gamma distributions is that if $Y_1,\ldots,Y_n$ are independent Gamma(a_i,b), then $\sum_{i=1}^n Y_i \sim$ Gamma$(\sum_{i=1}^n a_i, b)$. Thus if $U_1,\ldots,U_n$ are iid $U(0,1)$, $-\sum_{i=1}^n \log(U_i)/\theta \sim$ Gamma(n,θ).

Another well-known property of Gamma distributions is that if Y_1, Y_2 are independent Gamma(a_i,b) distributions, $Y_1/(Y_1+Y_2) \sim$ Beta(a_1,a_2). As such, it is a simple matter for us to simulate Beta distributions with integer parameters.

EXERCISE 6.1. Suppose $y \sim$ Beta$(a,1)$, $a>0$. Explain how to simulate y using its cdf.

EXERCISE 6.2. Suppose y is a random variable with cdf $F(y) = 1-e^{-\lambda y^\alpha}$ for $y>0$, $\alpha>0$. We say $y \sim$ Weib(α,λ). Explain how to simulate y. Note that y is a simple transformation of an exponential random variable. What is the transformation?

EXAMPLE 6.1.2. *Normals*. We now consider a convenient way to generate normal random variables. Take Z_1, Z_2 iid $N(0,1)$. Their distribution is rotationally symmetric about the origin in two-dimensional space. In other words, the density is constant on every circle. To specify the joint distribution of Z_1 and Z_2, we need only specify how much density needs to be associated with every circle centered at the origin. In particular, all the points on a circle of radius $\sqrt{Y}$ are all the Z_1, Z_2 values satisfying $Z_1^2+Z_2^2 = Y$. Knowing the distribution of Y and the rotational symmetry is enough to get us the entire distribution. By the definition of a chi-squared random variable, $Z_1^2+Z_2^2 \sim \chi^2(2) =$ Gamma$(2/2,1/2)$, something we know how to simulate, cf. Example 6.1.1.

Take $U_1 \sim U(0,1) \perp\!\!\!\perp U_2 \sim U(0,2\pi)$. From Example 6.1.1

$$-2\log(U_1) \sim Z_1^2+Z_2^2.$$

Intuitively, if U_2's distribution is uniform on a circle of radius 1 about the origin, the corresponding points are $(\cos(U_2),\sin(U_2))'$, so the distribution of

$$[Z_1,Z_2]' \equiv [\sqrt{-2\log(U_1)}\cos(U_2), \sqrt{-2\log(U_1)}\sin(U_2)]'$$

is rotationally symmetric about the origin and puts the correct density on each circle. In other words,

$$Z_1, Z_2 \overset{iid}{\sim} N(0,1). \tag{1}$$

This gives two $N(0,1)$ samples by sampling two uniforms. To sample $Y_i \sim N(\mu,\sigma^2)$, just take $Y_i = \mu + \sigma Z_i$.

EXERCISE 6.3. Using Proposition B.4, show that if $r^2 \sim \chi^2(2) \perp\!\!\!\perp \phi \sim U(0,2\pi)$, then $Z_1 = r\cos(\phi)$ and $Z_2 = r\sin(\phi)$ are iid $N(0,1)$. This result establishes (1). Remember that $r^2 \equiv q$ is the original random variable and $r = \sqrt{r^2} = \sqrt{q}$ is the transformation, so you need to find $d\sqrt{r^2}/d(r^2) \equiv \sqrt{q}/dq$.

EXAMPLE 6.1.3. *Multivariate Normals*. To sample a multivariate normal random vector $y = (y_1,\ldots,y_p)'$ with mean $\mu = (\mu_1,\ldots,\mu_p)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix},$$

write $\Sigma = AA'$, generate $z_1,\ldots,z_p$ iid $N(0,1)$, and compute

$$y = AZ + \mu$$

where $Z = (z_1, \ldots, z_p)'$. One way to find an appropriate matrix A is to compute the spectral decomposition of Σ, that is,

$$\Sigma = PD(\lambda_i)P'$$

where $D(\lambda_i)$ is a diagonal matrix with nonnegative diagonal elements λ_i. The λ_is are eigenvalues of Σ and the columns of P are corresponding orthonormal eigenvectors. Many computer packages have the ability to compute these. Take $A = PD(\sqrt{\lambda_i})$.

EXAMPLE 6.1.4. *Multivariate Extension of Beta*. Beta distributions are useful for modeling prior beliefs about a single probability associated with whether an event occurs or does not occur. However, we often need to deal with multiple events. For example, we might wonder if a person's hair is blond, black, or brown. To examine this, we need probabilities for each of the outcomes. The Dirichlet distribution is a useful way of simultaneously specifying probabilities for multiple events.

Another well-known property of Gamma distributions is that if Y_1, Y_2, and Y_3 are independent Gamma(a_i, b) distributions,

$$\left(\frac{Y_1}{Y_1+Y_2+Y_3}, \frac{Y_2}{Y_1+Y_2+Y_3}\right) \sim \text{Dirichlet}(a_1, a_2, a_3).$$

For integer values of the a_is, from Example 6.1.1 we know how to simulate the Gamma variables. Often we also use the notation

$$\left(\frac{Y_1}{Y_1+Y_2+Y_3}, \frac{Y_2}{Y_1+Y_2+Y_3}, \frac{Y_3}{Y_1+Y_2+Y_3}\right) \sim \text{Dirichlet}(a_1, a_2, a_3)$$

but in 3 dimensions this distribution has no density because the random variables are redundant. (They always add up to 1.) Also note that clearly

$$\frac{Y_1}{Y_1+Y_2+Y_3} \sim \text{Beta}(a_1, a_2+a_3).$$

It is a simple matter to extend the Dirichlet distribution to allow for an arbitrary number of categories. If $Y_1, Y_2, \ldots, Y_k$ are independent Gamma(a_i, b) distributions, then define $S = \sum_{i=1}^{k} Y_i$ and write

$$Z \equiv (Z_1, \ldots, Z_k) \equiv \left(\frac{Y_1}{S}, \frac{Y_2}{S}, \ldots, \frac{Y_k}{S}\right) \sim \text{Dirichlet}(a_1, \ldots, a_k).$$

Since $\sum_i Z_i = 1$, there are really only $k-1$ free components of Z. From Example 6.1.1,

$$Z_i \sim \text{Beta}\left(a_i, \sum_{j \neq i} a_j\right).$$

EXERCISE 6.4. Use Proposition B.4 to transform $Y_1, \ldots, Y_k$ that are independent Gamma(a_i, b)s into the vector $(Z_1, \ldots, Z_{k-1}, S)'$. Show that S is independent of the other variables by showing that the joint density for $(Z_1, \ldots, Z_{k-1}, S)$ is the product of a Dirichlet$(a_1, \ldots, a_k)$ density and a Gamma$(\sum_i a_i, b)$ density. The Dirichlet density is an obvious extension of that given in Table 2.1.

EXERCISE 6.5. Place a prior on the vector of probabilities (π_1, π_2, π_3). Here π_1 is the prevalence of infection among individuals showing clinical symptoms while π_2 is the prevalence among those who do not show clinical symptoms, so $\pi \equiv \pi_1 + \pi_2$ is the prevalence of infection in a population of interest. For example, if an individual has the human immunodeficiency virus (HIV), they may or may not have acquired immunodeficiency syndrome (AIDS). (a) Construct a Dirichlet prior that has mean vector $(2/10, 2/10, 6/10)$ and that has 95th percentile of 0.3 for π_1. Find a_1 and $a_2 + a_3$ by trial and error using BetaBuster. (b) What is the marginal prior for π_2 and for π_3? (c) Is it possible to find a Dirichlet prior with the given means and percentile that also has a 95th percentile for π_2 of 0.5? Explain.

6.2 Traditional Monte Carlo Methods

In this section, we examine two methods that traditionally have been used for Monte Carlo computations: Acceptance-rejection sampling and importance sampling. In applications, these methods have largely been supplanted by the Markov chain Monte Carlo methods discussed in the next section. Their primary importance now is as a supplement to MCMC methods when individual distributions in a larger Markov chain are difficult to sample. Smith and Gelfand (1992) presented an introduction to the use of importance sampling and the rejection method for Bayesian analysis.

6.2.1 Acceptance-Rejection Sampling

We want to sample from a univariate distribution with density $p(\theta)$ that is not recognizable as any easily sampled distribution. In fact, we may only know the kernel of the density, say $p_*(\theta)$, with $p_*(\theta) \propto p(\theta)$. (The posterior is often unrecognizable but it is proportional to the prior times the likelihood, two things we know.) To sample from $p(\theta)$, acceptance-rejection sampling uses another density that is "easy" to sample, say, $q(\theta)$, for which we can find a constant M that has $p_*(\theta) \le Mq(\theta)$ for all θ. The function $Mq(\theta)$ is called the upper envelope of the kernel. Ideally, it is as close as possible to $p_*(\theta)$. We return to the construction of $q(\theta)$ and the determination of M shortly, but first assume that they are already known.

To sample from $p(\cdot)$, start by sampling $\theta \sim q(\cdot)$ and independently $U \sim U[0,1]$. The idea is that either θ is acceptable as a sample from $p(\cdot)$ or it is rejected and we start over. U determines whether θ is acceptable. For given θ, $Mq(\theta)U \sim U[0, Mq(\theta)]$, so it is uniformly distributed below the envelope. We accept the sampled θ if $Mq(\theta)U < p_*(\theta)$ and reject it otherwise. Thus values of $Mq(\theta)U$ between the upper envelope and the kernel are rejected. The probability of rejection is small if the area between the envelope and $p_*(\cdot)$ is small.

Before showing that this method actually provides samples from $p(\theta)$, consider one particular method for constructing $Mq(\theta)$. Suppose that $\ell(\theta) \equiv \log[p_*(\theta)]$ is concave. To pick $Mq(\theta)$ find the mode of $\ell(\theta)$, pick points on either side of the mode, say, $\tilde{\theta}_1$ and $\tilde{\theta}_2$, and envelop $\ell(\theta)$ using the tangent lines at $\tilde{\theta}_1$ and $\tilde{\theta}_2$. The tangent lines are

$$\gamma_i(\theta) \equiv \ell(\tilde{\theta}_i) + \dot{\ell}(\tilde{\theta}_i)(\theta - \tilde{\theta}_i)$$

where $\dot{\ell}(\theta)$ is the derivative. Because $\ell(\theta)$ is concave, for $i = 1, 2$, $\ell(\theta) \le \gamma_i(\theta)$. Exponentiating, we get

$$p_*(\theta) = e^{\ell(\theta)} \le e^{\gamma_i(\theta)}.$$

This allows us to pick our envelope function as

$$Mq(\theta) = \min\left\{e^{\gamma_1(\theta)}, e^{\gamma_2(\theta)}\right\}.$$

We have defined the product $Mq(\theta)$ but for acceptance-rejection sampling, we need to actually know the density $q(\theta)$. As seen in Exercise 6.6, it is easy to find where the two tangent lines intersect, say at θ_*, so

$$Mq(\theta) = \begin{cases} e^{\gamma_1(\theta)} & \text{if } \theta \le \theta_* \\ e^{\gamma_2(\theta)} & \text{if } \theta \ge \theta_* \end{cases}.$$

Also as in Exercise 6.6, it is not difficult to integrate under this curve to find M and thus $q(\theta)$. Finally, Exercise 6.6 establishes that it is easy to sample from $q(\theta)$.

EXERCISE 6.6. Let $\gamma_i(\theta) = a_i + b_i\theta$. We know by construction that $b_1 > 0$ and $b_2 < 0$. (a) Find θ_* by setting $\gamma_1(\theta_*) = \gamma_2(\theta_*)$ and solving. (b) Integrate $Mq(\theta)$ over $(-\infty, \infty)$ to determine M as a function of, say, θ_* and the a_is and b_is. (c) Obtain the cdf based on the density $q(\cdot)$, say $Q(v) \equiv \int_{-\infty}^{v} q(\theta)d\theta$. Do this first for $v \le \theta_*$, and then for $v > \theta_*$. Calculate the latter as $\int_{-\infty}^{\theta_*} q(\theta)d\theta +$

$\int_{\theta_*}^{v} q(\theta)d\theta$. (d) Finally, solve $Q(v) = u$ for v so that $v = Q^{-1}(u)$. Thus if we sample $U \sim U[0,1]$, we have $Q^{-1}(U) \sim q(\cdot)$. Carefully organize your presentation of the numerous steps involved.

EXERCISE 6.7. Suppose $y|\theta \sim \text{Pois}(\theta)$ and $\theta \sim \text{Gamma}(10, 0.5)$. If $y = 15$ is observed, develop a method for sampling from the posterior. Find the explicit form of the envelope function $Mq(\theta)$, including the point of intersection of the two tangent lines, θ^*, corresponding to the tangent lines at $\tilde{\theta}_1 = 10$ and $\tilde{\theta}_2 = 22$. Explain how you would find M and give an explicit algorithm for the accept-reject algorithm applied to this problem. You need not explicitly find M.

Finally, we establish the validity of the acceptance-rejection procedure. Define $K(\theta) = p_*(\theta)/Mq(\theta)$. Let A denote the event that θ is accepted. Also let c_* be the constant that satisfies $p(\theta) = c_* p_*(\theta)$. We show that $\Pr(\theta \le v|A)$ equals the cdf corresponding to the density $p(\cdot)$.

$$\begin{aligned}
\Pr(\theta \le v|A) &= \frac{\Pr(\theta \le v \text{ and } A)}{\Pr(A)} \\
&= \frac{\Pr[\theta \le v \text{ and } Mq(\theta)U \le p_*(\theta)]}{\Pr[Mq(\theta)U \le p_*(\theta)]} \\
&= \frac{\Pr[\theta \le v \text{ and } U \le p_*(\theta)/Mq(\theta)]}{\Pr[U \le p_*(\theta)/Mq(\theta)]} \\
&= \frac{\Pr[\theta \le v \text{ and } U \le K(\theta)]}{\Pr[U \le K(\theta)]} \\
&= \frac{\int_{-\infty}^{v} \left[\int_0^{K(\theta)} 1du\right] q(\theta)d\theta}{\int_{-\infty}^{\infty} \left[\int_0^{K(\theta)} 1du\right] q(\theta)d\theta} \\
&= \frac{\int_{-\infty}^{v} K(\theta)q(\theta)d\theta}{\int_{-\infty}^{\infty} K(\theta)q(\theta)d\theta} \\
&= \frac{\int_{-\infty}^{v} [p_*(\theta)/Mq(\theta)]\, q(\theta)d\theta}{\int_{-\infty}^{\infty} [p_*(\theta)/Mq(\theta)]\, q(\theta)d\theta} \\
&= \frac{\int_{-\infty}^{v} c_* p_*(\theta)d\theta}{\int_{-\infty}^{\infty} c_* p_*(\theta)d\theta} \\
&= \int_{-\infty}^{v} p(\theta)d\theta.
\end{aligned}$$

The last equality holds because the integral of $p(\cdot)$ in the denominator is 1.

6.2.2 *Importance Sampling*

A random sample $\theta^1, \ldots, \theta^s$ from the posterior distribution can be viewed as an approximation to the posterior that, for $k = 1, \ldots, s$, takes the value θ^k with probability $1/s$. Importance sampling also provides an approximation to the posterior distribution that is discrete but one with unequal probabilities. It takes on values θ^k with probability w_k. More formally, importance sampling provides numerical approximations to posterior integrals of the form

$$\int h(\theta)p(\theta|y)d\theta = \frac{\int h(\theta)L(\theta|y)p(\theta)d\theta}{\int L(\theta|y)p(\theta)d\theta}. \tag{1}$$

for some function $h(\cdot)$.

In importance sampling, one chooses a *known* density function $q(\theta)$ that is easy to sample. The procedure works best if $q(\theta)$ is similar in shape to the known kernel of the posterior $L(\theta|y)p(\theta)$

with tails that do not decay more rapidly than the tails of the posterior. Sample $\theta^1, \ldots, \theta^s$ from the distribution with density $q(\theta)$. Define

$$\tilde{w}(\theta) \equiv \frac{L(\theta|y)p(\theta)}{q(\theta)}$$

and for $k = 1, \ldots, s$,

$$w_k \equiv \tilde{w}(\theta^k) \Big/ \sum_{j=1}^{s} \tilde{w}(\theta^j).$$

By design, $\sum_k w_k = 1$. The discrete approximation to the posterior takes the value θ^k with probability w_k.

To see that this discrete approximation to the posterior provides accurate estimates of integrals like (1), rewrite (1) as

$$\int h(\theta)p(\theta|y)d\theta = \frac{\int h(\theta)\,[L(\theta|y)p(\theta)/q(\theta)]\,q(\theta)d\theta}{\int [L(\theta|y)p(\theta)/q(\theta)]\,q(\theta)d\theta} = \frac{\int h(\theta)\tilde{w}(\theta)q(\theta)d\theta}{\int \tilde{w}(\theta)q(\theta)d\theta}.$$

Applying the Law of Large Numbers, for large s

$$\sum_{j=1}^{s} h(\theta^j)\tilde{w}(\theta^j)/s \doteq \int h(\theta)\tilde{w}(\theta)q(\theta)d\theta$$

and

$$\sum_{j=1}^{s} \tilde{w}(\theta^j)/s \doteq \int \tilde{w}(\theta)q(\theta)d\theta,$$

so applying the discrete approximation to $\mathrm{E}[h(\theta)|y]$ gives

$$\hat{\theta}_h \equiv \sum_{j=1}^{s} h(\theta^j)w_j = \frac{\sum_{j=1}^{s} h(\theta^j)\tilde{w}(\theta^j)/s}{\sum_{j=1}^{s} \tilde{w}(\theta^j)/s} \doteq \frac{\int h(\theta)\tilde{w}(\theta)q(\theta)d\theta}{\int \tilde{w}(\theta)q(\theta)d\theta} = \int h(\theta)p(\theta|y)d\theta.$$

To avoid any difficulties with the unequal weights in this discrete approximation to the posterior, we can obtain an approximate random sample from the posterior by taking a new Monte Carlo sample from this discrete distribution. The process is called *Sampling Importance Resampling* or *SIR*. Thus if a SIR sample is represented as $(\theta^{*1}, \ldots, \theta^{*m})$, we can use this sample to make full inferences about $h(\theta)$ by calculating $h(\theta^{*1}), \ldots, h(\theta^{*m})$ and obtaining a smoothed histogram, the mean, standard deviation, and quantiles in the usual way. (SIR is also the acronym for *sliced inverse regression* and we have used it as an acronym for the *standard improper reference* prior for normal data.)

A standard importance distribution $q(\cdot)$ is a multivariate normal or multivariate Student distribution (see Exercise 9.3) with mean (location) equal to the posterior mode or MLE of θ and covariance matrix (dispersion) equal to a scaled version of the Fisher Observed Information matrix's inverse (see Section 4.10). The importance distribution is taken to have heavier tails than the actual posterior because otherwise the weights w_k may get large when θ^k is observed in an extreme tail of $q(\theta)$. Suppose an unusual value θ^k occurs with very small $q(\theta^k)$. If the tails of q are much lighter than the kernel of the posterior, $\tilde{w}(\theta^k) = p(\theta^k)L(\theta^k|y)/q(\theta^k)$ will be large and the weight given to θ^k will be large. This results in unstable approximations to (1). We could have, say, $w_k = 0.6$, which would obviously result in a very poor discrete approximation to a continuous posterior. In theory, these issues take care of themselves, but in practice, we want to eliminate the possibility that outliers in the simulation sample, i.e., unusual values θ^k, are given high weight. See Christensen (1997, Sec. 13.4) for details on applying importance sampling to logistic regression. Smith and Gelfand (1992) used the prior distribution as an importance function. This is often easier to sample, but it is unlikely to mimic the posterior as well.

EXERCISE 6.8. Suppose that $y_1, y_2, \ldots, y_n$ are iid Pois(θ) and assume a Jeffreys' prior. Explain in detail how you would implement an importance sampling algorithm for obtaining the posterior probability that $\theta > 20$.

EXERCISE 6.9. Suppose that $y_1, y_2, \ldots, y_n$ are iid Gamma(α, β) and assume $p(\alpha, \beta) = p(\alpha)p(\beta)$, where $\beta \sim$ Gamma(a, b) and $\alpha \sim$ Gamma$(1, 1)$. The prior mean for α is one, so we are in some sense centering the prior on the exponential distribution but allowing departures from it. Explain in detail how you would implement an importance sampling algorithm for obtaining full inferences about the mean of the ys, α/β.

6.3 Markov Chain Monte Carlo

The idea of Markov chain Monte Carlo is to define a sequence of random vectors $\theta^1, \theta^2, \theta^3, \ldots$ in which the distribution of θ^k near the beginning of the sequence can be just about anything but in which the distributions eventually settle down to the posterior distribution. Thus, if θ^k has a marginal density $q_k(\theta)$, as k gets large, these densities approach the *posterior density $p(\theta|y)$, which for this section we will abbreviate as $p(\theta)$*. Specifically, the sequence of θ^ks is a Markov chain as defined in Subsection 6.3.1. Markov chains have other useful applications but our interest is restricted to sampling from the joint posterior. Under mild conditions, Markov chain theory indicates that if k is large, the θ^ks are (approximately) identically distributed with density $p(\theta)$. However, the θ^ks are not typically independent, so the θ^ks do not constitute an approximate random sample from the posterior. Nonetheless, a version of the Law of Large Numbers, sometimes called an *Ergodic Theorem*, applies to these sequences. Under some conditions, if $\theta^1, \ldots, \theta^s$ are sampled from a Markov chain and h is a function with finite expectation under the posterior distribution, then with probability one

$$\lim_{s\to\infty} \sum_{j=1}^{s} h(\theta^j)/s = \int h(\theta)p(\theta)d\theta.$$

Thus we can approximate probabilities and expected values relative to the posterior distribution just by taking the sample mean of appropriate functions of the θ^ks.

Burning-in can improve the approximations. Intuitively, observations obtained after the chain has settled down to the posterior will be more useful in estimating probabilities and expectations for $p(\theta)$. If we throw out the early observations, taken while the process was settling down, the remainder of the process should be a very close approximation to one in which every observation is sampled from the posterior. Dropping the early observations is referred to as using a *burn-in* period. With simple statistical models, we might run the chain for 6,000 or 11,000 observations with a burn in of 1,000 observations, thus using the last 5,000 or 10,000 samples to estimate probability integrals associated with the posterior distribution. More complicated probability models typically require longer chains and burn-ins. Subsection 6.3.5 discusses checking on whether the chain has settled down. Without getting technical, we refer to the process of settling down as achieving stationarity, cf. Christensen (2001a, Section 4.1).

Given an observed sequence, a plot of the pairs (k, θ^k) is known as a history. Figure 6.1 shows histories of four pairs of chains with each pair started at distinct initial values. Figure 6.1(a) shows the behavior we like to see. After a burn-in of 5,000 iterations, each chain has settled down nicely. Figure 6.1(b) shows typical behavior during the burn-in period. The two chains differ markedly in the initial stages but by 750 iterations we see that they are settling down. Figure 6.1(c) shows chains that are nowhere near settling down between 501 and 1,500 iterations, and Figure 6.1(d) shows the same chains having settled down nicely between 50,000 and 100,000 iterations. Although they have certainly converged by 50,000 iterations, observe the "waviness" of the histories compared with those in Figure 6.1(a). This is due to autocorrelation, which is discussed later.

Usually, even after the burn-in phase, the iterates are correlated. They are eventually identically distributed but not independent. If $s = 10{,}000$ and we have correlation 1 between all pairs through

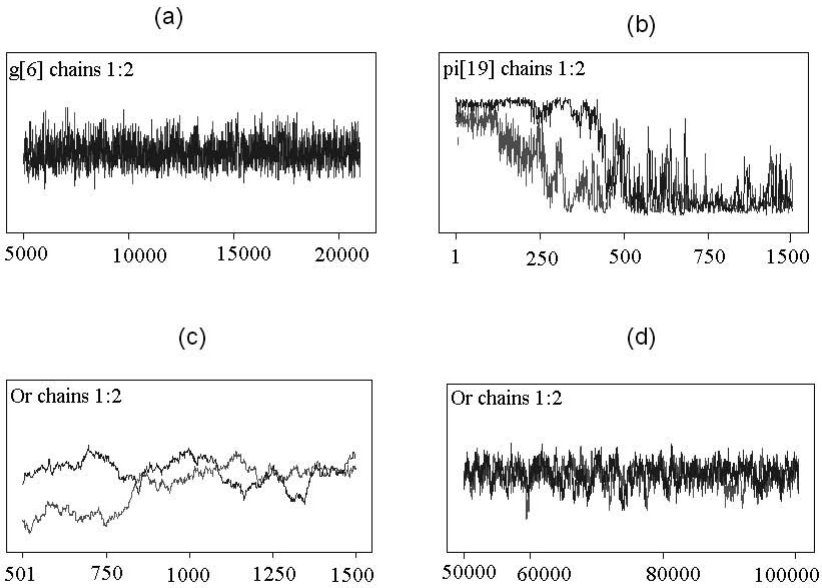

Figure 6.1: *Histories of Markov chains with two starting values: (a) Chains that have converged after 5,000 iterations, (b) chains that diverge and then converge by 1,000 iterations, (c) chains not converging and exhibiting high autocorrelation, and (d) chains from (c) that ultimately converge sometime before 50,000 iterations, but that exhibit high autocorrelation.*

time, then the effective Monte Carlo sample size is 1. This extreme case is unlikely ever to happen, but it illustrates a point. With iid sampling under typical circumstances, one can probably get reasonable approximations from a few thousand samples. If the samples are identically distributed but not independent, to attain sufficient accuracy in our numerical approximation to the posterior, we may need to take much larger sample sizes. Fortunately, computers are well equipped for that.

Thinning is a process used to make the observations more nearly independent, hence more nearly a random sample from the posterior distribution. Frankly, after a burn-in, there is not much point in thinning unless the correlations are extremely large. If there is a lot of correlation between adjacent observations, a larger overall MC sample size is needed to achieve reasonable numerical accuracy, in addition to needing a much longer burn-in. To check the level of dependence, look at the estimated *autocorrelation function (ACF)*. This is a function of the integers j that gives the estimated correlation between θ^k and θ^{k+j}. After a burn-in, this correlation should depend on the *lag* j, but not on k. It is computed as the sample correlation between the pairs (θ^k, θ^{k+j}), $k = 1, \ldots, s-j$. If the autocorrelations are near zero except for, say, the first two, $j = 1, 2$, then we could thin by taking every third θ^k. That is, our sample could be θ_{3k}, $k = 1, \ldots, s$, after the burn-in. This sample should be nearly uncorrelated but it throws away information. Unless there is severe autocorrelation, e.g., high correlation even with, say $j = 30$, we don't believe that thinning is worthwhile.

Figure 6.2 gives histories of two chains on the left and their corresponding autocorrelations on the right. The history and ACF on the top both look good. The ACF rapidly approaches zero and stays there. The autocorrelation function in Figure 6.2(d) dies out very slowly with autocorrelation of at least 0.5 even at a lag of 50 iterations. In time domain analysis of time series, such autocorrelation functions are taken as evidence of non-stationarity, cf. Christensen (2001a, Section 5.6). The waviness of the history plot indicates that iterations near one another are nearly the same. This waviness and the huge autocorrelations indicate the need for a larger Monte Carlo sample size.

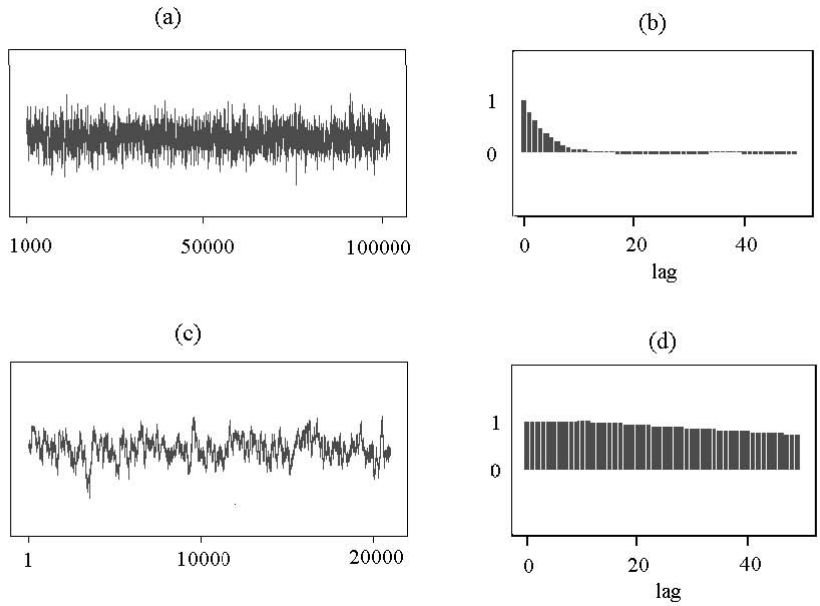

Figure 6.2: *Histories of Markov chains (left panels) and corresponding autocorrelations by lag (right panels): Panel (a) shows the history of a chain with strong autocorrelation as exhibited by the corresponding ACF in panel (b), and panel (c) shows the history of a chain with some autocorrelation as exhibited in panel (d).*

As discussed in the next section, the sequence (chain) should eventually settle down to be approximately stationary but these plots suggest that the convergence to stationarity is occurring very slowly or not at all. (The conditions that imply convergence may not hold.) We ran the chain from panel (c) out to 100,000 iterations and thinned by 20, and the ACF for the thinned chain looked very similar to the one in panel (b), and the history looked similar to the one in panel (a).

As a diagnostic to check on whether we have approximately identical distributions after burn-in, we might run several chains with θ^1 values chosen independently from an initial distribution. Such samples are independent and should give similar estimates for the posterior distribution. More on this is in Subsection 6.3.5.

The next subsection introduces some theory for Markov chains. The subsequent three subsections are devoted to specific methods for generating MCs that result in samples from the joint posterior after the burn-in phase, namely Gibbs sampling, the Metropolis algorithm, and Slice sampling. Casella and George (1992) presented an early introduction to the Gibbs sampler. When Gibbs sampling is applicable, it is almost always used. The Metropolis and Slice algorithms are more generally applicable, and are also used to augment the Gibbs sampler.

6.3.1 Markov Chains

Consider a sequence of random vectors $\theta^1, \theta^2, \theta^3, \ldots$. This is a *Markov chain (MC)* if for any set A,

$$\Pr(\theta^k \in A | \theta^1, \ldots, \theta^{k-1}) = \Pr(\theta^k \in A | \theta^{k-1}). \tag{1}$$

In other words, what happens at step k depends only on what happened at step $k-1$. Another way of thinking about it is that when you are at step $k-1$, where you go depends only on where you are — it does not depend on how you got to where you are. This dependence of each new observation on only the previous observation is known as the *Markov property*.

Let $q_1(\theta^1)$ be the initial density for θ^1, let $q_{k|\cdot}(\theta^k|\theta^1,\ldots,\theta^{k-1})$ be the obvious conditional density, and as mentioned earlier, $q_k(\theta^k)$ is the marginal density of θ^k. From standard probability

$$\begin{aligned}\Pr(\theta^k \in A) &= \int_A q_k(\theta)d\theta \\ &= \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q_{k|\cdot}(\theta^k|\theta^1,\ldots,\theta^{k-1}) \cdots q_{2|\cdot}(\theta^2|\theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.\end{aligned}$$

If we have the Markov property (1), then we must have $q_{j|\cdot}(\theta^j|\theta^1,\ldots,\theta^{j-1}) = q_{j|j-1}(\theta^j|\theta^{j-1})$ for $j = 2,3,\ldots$, so a Markov chain has

$$\Pr(\theta^k \in A) = \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q_{k|k-1}(\theta^k|\theta^{k-1}) \cdots q_{2|1}(\theta^2|\theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.$$

Constructing a Markov chain is simple. All you have to do is specify the initial distribution $q_1(\theta^1)$ and the conditional distributions $q_{j|j-1}(\theta^j|\theta^{j-1})$ for $j = 2,3,\ldots$. To show that something is a Markov chain, all you have to do is show that the conditional distributions are of the form $q_{j|j-1}(\theta^j|\theta^{j-1})$ for $j = 2,3,\ldots$. Moreover, it is simple to sample from an MC if you know $q_1(\cdot)$ and all of the conditional densities. Generate θ^1 from q_1, then, since you know θ^1, generate θ^2 from the appropriate conditional distribution, and continue.

A simplifying assumption that we make *henceforth* is that of *stationary transition probabilities*. This assumption states that the conditional distribution of going from step $j-1$ to step j is the same regardless of the value of j. If we write the densities more carefully as in Appendix B, $q_{j|j-1}(\theta^j|\theta^{j-1}) \equiv q_{\theta^j|\theta^{j-1}}(u|v)$. For this function not to depend on the steps $j-1$ and j, it must be some function $q(u|v)$, which we commonly write as $q(\theta^j|\theta^{j-1})$ to help us keep track of where it is being applied. (Many discussions use the notation $q(u|v) \equiv k(v,u)$, which is called a *transition kernel*.) A Markov chain with stationary transition probabilities now has

$$\Pr(\theta^k \in A) = \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(\theta^k|\theta^{k-1}) \cdots q(\theta^2|\theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.$$

Although the transition probabilities do not depend on the step of the chain, the marginal distribution of a θ^k typically depends on k, hence $q_k(\theta^k)$ still indicates the marginal density at step k.

Historically, an interesting issue for Markov chains with stationary transition probabilities is studying the effect that the choice of an initial distribution $q_1(\theta^1)$ has on the marginal distributions $q_k(\theta^k)$. It is the solution of this problem that makes Markov chains useful in statistical simulation.

One aspect of the solution involves the idea of a *stationary distribution*. A stationary distribution, say $p(\cdot)$, has the property for all k that

$$\Pr(\theta^k \in A) = \int_A p(\theta)d\theta. \tag{2}$$

In particular, $q_k(\theta) = p(\theta)$ for all θ and for all k, including $k = 1$ which means we start the chain with a value taken from the stationary distribution itself. Under the assumption of a stationary distribution and using the law of total probability, we find

$$\begin{aligned}\int_A p(\theta)d\theta &= \Pr(\theta^k \in A) \\ &= \int \Pr(\theta^k \in A|\theta^{k-1}) p(\theta^{k-1}) d\theta^{k-1} \\ &= \int \left[\int_A q(\theta^k|\theta^{k-1})d\theta^k\right] p(\theta^{k-1})d\theta^{k-1} \\ &= \int_A \left[\int q(\theta^k|\theta^{k-1}) p(\theta^{k-1})d\theta^{k-1}\right] d\theta^k.\end{aligned}$$

Since this holds for any set A, and since θ^k and θ^{k-1} are just dummy variables in the integrals, we must have

$$p(\theta) = \int q(\theta|\theta^{k-1})p(\theta^{k-1})d\theta^{k-1} = \int q(\theta|\theta^*)p(\theta^*)d\theta^*. \tag{3}$$

Thus a chain that satisfies (2) for all k must also satisfy (3).

The result that we really want is the converse. If we start our chain with $q_1(\theta) = p(\theta)$, and if our transition density satisfies (3), then (2) will be satisfied, namely, the chain will have stationary density $p(\theta)$. This is easily established by induction. Assume that the first iterate is taken from $p(\theta)$ and that (3) holds. It follows from the law of total probability that

$$q_2(\theta^2) = \int q(\theta^2|\theta^1)p(\theta^1)d\theta^1 = p(\theta^2).$$

Then assuming that $q_k(\theta) = p(\theta)$ as the induction hypothesis, exactly the same argument leads to the conclusion that $q_{k+1}(\theta) = p(\theta)$ and we are done.

Henceforth we assume the existence of a stationary distribution. However, it is possible that there could be more than one density p that satisfies (3), hence more than one stationary distribution. Under relatively weak conditions, that cannot happen.

So far, we have not accomplished very much. If we construct a Markov chain that has the posterior $p(\theta)$ as a stationary distribution, then if we begin the MC by sampling from $p(\theta)$, every subsequent observation will also be an observation with density $p(\theta)$. But if we knew how to sample from $p(\theta)$, we would not need any of this Markov chain machinery. The point is to run a Markov chain that starts arbitrarily but eventually gives us samples from the posterior. We need some powerful results to achieve this goal.

Under conditions discussed later, when a proper stationary distribution exists, it is unique, also, as k gets large, regardless of the initial distribution $q_1(\theta^1)$, the marginal distribution of the θ^ks settles down to the stationary distribution, and, finally, the Markov chain satisfies a version of the Law of Large Numbers. Rephrasing the middle result, as k gets large, the $q_k(\theta^k)$ distribution approaches the $p(\theta^k)$ distribution, that is, for large k,

$$\Pr(\theta^k \in A) \doteq \int_A p(\theta)d\theta,$$

or more technically,

$$\lim_{k\to\infty} \Pr(\theta^k \in A) = \int_A p(\theta)d\theta. \tag{4}$$

In particular, by picking q_1 to give probability one to the value θ_*,

$$\lim_{k\to\infty} \Pr(\theta^k \in A|\theta^1 = \theta_*) = \int_A p(\theta)d\theta. \tag{5}$$

This suggests that, rather than randomly picking where to start the chain, we can start it anywhere we want. Nonetheless, it seems obvious that the convergence should be faster if we can pick an initial distribution q_1 that is somehow close to p, or an initial value θ^1 that is characteristic of p.

The key result is the Markov chain version of the Law of Large Numbers, sometimes called an *Ergodic Theorem*. It states that if $\theta^1, \ldots, \theta^s$ are sampled from the Markov chain and h is a function with finite expectation under the stationary distribution, then with probability one

$$\lim_{s\to\infty} \sum_{j=1}^{s} h(\theta^j)/s = \int h(\theta)p(\theta)d\theta. \tag{6}$$

Thus we can approximate probabilities and expected values relative to the stationary (in practice, posterior) distribution. This makes sense because for large k, the θ^ks are approximately identically distributed with density $p(\theta)$, although they are typically not independent.

The usefulness of all of this is that we can construct Markov chains with particular stationary transition probabilities that have the posterior distribution as their stationary distribution. Then, regardless of the initial distribution we choose, if we sample from the MC long enough, the samples will, to a good approximation, be identically distributed from the posterior and sample means converge with probability one to their corresponding posterior expectations. To construct such an MC, we need to establish that it has stationary transition probabilities, that the posterior satisfies equation (3), and a little more.

We now make some definitions and present formal conditions for (4), (5), and (6) to hold. We take most of our results from Tierney (1994). An MC with stationary distribution $p(\theta)$ is *p-irreducible* if, for any initial value, there is positive probability of eventually reaching any set A for which $\int_A p(\theta)d\theta > 0$. A chain is *periodic* if it can only return to an initial set at regularly spaced times, e.g., on the 2nd, 4th, 6th, etc. iterations. Otherwise, it is *aperiodic*. A sufficient condition for aperiodicity is that $\int_A q(\theta|\theta^1)d\theta > 0$ for all θ^1, provided $\int_A p(\theta)d\theta > 0$, which means that it is possible to get to any set of interest in one transition, regardless of where the chain was started. This condition is also sufficient for p-irreducibility.

If a chain has a proper stationary distribution p, is p-irreducible, and is aperiodic, then not only is the stationary distribution unique, that is, no other choice for p can satisfy (3) for the given transition distribution $q(\cdot|\cdot)$, but (5) is satisfied.

Convergence theory for MCs involves assessing whether a chain will repeatedly return to any specified set with positive probability under $p(\theta)$. This is called *recurrence*. The form of recurrence we need is *Harris recurrence*. A chain is Harris recurrent if, for every starting value $\theta^1 = \theta_*$ and any set A with positive probability under $p(\theta)$, the probability that A is revisited by the chain infinitely often is one. This guarantees, in theory if not in practice, *good mixing* of the chain. We want the MC to explore the entire support of the posterior (stationary) distribution, so it is important that every region of the support that has positive probability be visited infinitely often by the chain. A sufficient condition for Harris recurrence in an MC with stationary p is that it be p-irreducible and that $\int_A p(\theta)d\theta = 0$ implies $\int_A q(\theta \mid \theta^1)d\theta = 0$, for all initial values θ^1. This latter condition is referred to as the transition distribution being absolutely continuous with respect to the stationary distribution. The absolute continuity condition reduces to checking whether sets with posterior probability zero also have transition probability zero, regardless of the initial value.

A Markov chain is called ergodic if it is Harris recurrent and aperiodic. If we have an ergodic chain with stationary distribution $p(\theta)$, (4) holds, and if $h(\cdot)$ is integrable with respect to $p(\theta)$, then (6) holds.

In practice, we check that $\int_A p(\theta)d\theta = 0$ if and only if $\int_A q(\theta \mid \theta^1)d\theta = 0$ for all θ^1. If this is true, the MC is aperiodic, p-irreducible, and Harris recurrent, so all of (4), (5), and (6) hold.

MCs generated by Gibbs samplers, the Metropolis algorithm, and slice sampling all have the posterior satisfying (3) and are usually ergodic. We establish (3) for Gibbs sampling and the Metropolis algorithm in the next two subsections. Slice sampling is a special case of the Gibbs sampler. Details for establishing ergodicity can be found in Tierney (1994) and Robert and Casella (2004).

6.3.2 Gibbs Sampling

Gibbs sampling is a method for constructing a Markov chain that is extremely useful when one can isolate the conditional distribution of each parameter given all of the other parameters. The process involves obtaining samples from each conditional distribution in turn. More generally, it can be applied to sets of parameters. Temporarily treating vectors as row vectors, we illustrate the ideas for three blocks or subvectors, that is, $\theta^k = (\theta_1^k, \theta_2^k, \theta_3^k)$. The dimensions of each block are arbitrary. We construct the chain so that the posterior $p(\theta) = p(\theta_1, \theta_2, \theta_3)$ is the stationary distribution. Here *we have again dropped the explicit dependence on the data for brevity.* Gibbs sampling is based on

sampling from the *full conditional distributions* determined by the posterior, i.e.,

$$p_{1|23}(\theta_1 \,|\, \theta_2, \theta_3), \quad p_{2|13}(\theta_2 \,|\, \theta_1, \theta_3), \quad p_{3|12}(\theta_3 \,|\, \theta_1, \theta_2).$$

To define the MC, first sample θ^1 from the initial distribution $q(\theta_1, \theta_2, \theta_3) \equiv q_1(\theta)$. This can be a one-point distribution, i.e., just pick a starting value, or, if making a random selection, it may be convenient to have the three blocks independent, thus sampling $\theta^1 = (\theta_1^1, \theta_2^1, \theta_3^1)$ as

$$\theta_1^1 \sim q_1(\theta_1), \qquad \theta_2^1 \sim q_2(\theta_2), \qquad \theta_3^1 \sim q_3(\theta_3)$$

where we have implicitly redefined the q_1 notation. The key to Gibbs sampling is that the transition probabilities are defined in terms of the full conditional distributions. The second complete step of the chain defines θ^2 in three phases. First,

$$\theta_1^2 \,|\, \theta_2^1, \theta_3^1 \quad \sim \quad p_{1|23}(\theta_1 \,|\, \theta_2^1, \theta_3^1),$$

then

$$\theta_2^2 \,|\, \theta_1^2, \theta_3^1 \quad \sim \quad p_{2|13}(\theta_2 \,|\, \theta_1^2, \theta_3^1),$$

and finally

$$\theta_3^2 \,|\, \theta_1^2, \theta_2^2 \quad \sim \quad p_{3|12}(\theta_3 \,|\, \theta_1^2, \theta_2^2).$$

In general, we sample $\theta^k = (\theta_1^k, \theta_2^k, \theta_3^k)$ as

$$\begin{aligned} \theta_1^k \,|\, \theta_2^{k-1}, \theta_3^{k-1} \quad &\sim \quad p_{1|23}(\theta_1 \,|\, \theta_2^{k-1}, \theta_3^{k-1}), \\ \theta_2^k \,|\, \theta_1^k, \theta_3^{k-1} \quad &\sim \quad p_{2|13}(\theta_2 \,|\, \theta_1^k, \theta_3^{k-1}), \end{aligned}$$

and

$$\theta_3^k \,|\, \theta_1^k, \theta_2^k \quad \sim \quad p_{3|12}(\theta_3 \,|\, \theta_1^k, \theta_2^k).$$

By construction, this defines a valid conditional distribution for transitioning from θ^{k-1} to θ^k that does not depend on k. In particular, the stationary transition distribution is

$$\begin{aligned} q(\theta^k \,|\, \theta^{k-1}) \quad &\equiv \quad q(\theta_1^k, \theta_2^k, \theta_3^k \mid \theta_1^{k-1}, \theta_2^{k-1}, \theta_3^{k-1}) \\ &\equiv \quad p_{1|23}(\theta_1^k \mid \theta_2^{k-1}, \theta_3^{k-1}) p_{2|13}(\theta_2^k \mid \theta_1^k, \theta_3^{k-1}) p_{3|12}(\theta_3^k \mid \theta_1^k, \theta_2^k). \end{aligned} \tag{7}$$

With these transition distributions, the posterior is the stationary distribution, a fact that we will later illustrate for a two-block Gibbs sampler.

EXAMPLE 6.3.1. *Normal Data with Independence Prior.* Suppose we have observations

$$y_1, \ldots, y_n | \mu, \tau \overset{iid}{\sim} N(\mu, 1/\tau)$$

with prior

$$\mu \sim N(a, 1/b) \quad \perp\!\!\!\perp \quad \tau \sim \text{Gamma}(c, d).$$

As discussed in Subsection 5.2.4, the posterior density $p(\mu, \tau | y)$ is not recognizable as any parametric form but we saw that

$$\mu | \, \tau, y \sim N[\hat{\mu}(\tau), 1/(n\tau + b)]$$

and

$$\tau | \, \mu, y \sim \text{Gamma}\left(c + \frac{n}{2}, d + \frac{1}{2}\left[n(\bar{y}_\cdot - \mu)^2 + (n-1)s^2\right]\right).$$

To sample these distributions iteratively, we start with initial values (μ^1, τ^1). To get off to a good start, these are often picked to be the modes of informative prior distributions. We then sample a new value μ^2 from the $N[\hat{\mu}(\tau^1), 1/(n\tau^1 + b)]$ distribution followed by a sample τ^2 taken from the Gamma$(c + n/2, d + [n(\bar{y}_\cdot - \mu^2)^2 + (n-1)s^2]/2)$ distribution. We continue in this fashion to obtain s iterates, $\{(\mu^k, \tau^k) : k = 1, \ldots, s\}$. Moreover, as discussed in Chapter 3, for any function $\gamma = g(\mu, \tau)$, we can approximate the posterior $p(\gamma|y)$ numerically by using the sample $\{\gamma^k \equiv g(\mu^k, \tau^k) : k = 1, \ldots, s\}$.

The Gibbs sampler presupposes that we actually know how to sample from the full conditional distributions. Sometimes these distributions are recognizable, such as normal, beta, or gamma distributions, in which case sampling from them is easy. In any case, we can find the kernels of the full conditionals which lets us sample from the distribution using something like an *adaptive-rejection method,* a Metropolis method, or a Slice sampler. Adaptive-rejection sampling is a modification of the acceptance-rejection method of Section 2 but is more efficient as applied to Gibbs sampling and is implemented in WinBUGS when the full conditional is log concave. The other methods are discussed in the next two subsections and are also built into WinBUGS to handle full conditionals that are not log concave.

Adaptive-rejection sampling was developed by Gilks and Wild (1992). A simpler algorithm was used by Dellaportas and Smith (1993) for Bayesian generalized linear models, which generally have log concave posteriors provided the priors are log concave. They combined the Gibbs sampler with the envelope rejection method for sampling unrecognizable full conditionals. To sample an observation, say θ^k, using the envelope method requires choosing two values, say $\tilde{\theta}_j$, $j = 1, 2$, one on either side of the mode. Taking the $\tilde{\theta}_j$s to be the most recent values of $\theta^{k-1}, \ldots, \theta^1$ to fall on either side of the mode seems to work well.

EXAMPLE 6.3.2. *Weibull Data.* Feigl and Zelen (1965) present data on the survival times, measured in weeks, of patients who were diagnosed with leukemia. The patients were classified according to one of two characteristics of white blood cells. We only examine those whose blood is AG+. The sample consists of $n = 17$ times in weeks from diagnosis to death: 65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65. Survival times are dealt with in detail in Chapters 12 and 13.

The data must be strictly positive, so we need models that are concentrated on $(0, \infty)$. One such model is the Weibull distribution. We denote

$$y_i \sim \text{Weib}(\alpha, \lambda)$$

if the density and cdf are

$$f_*(y_i|\,\alpha, \lambda) = \lambda\,\alpha y_i^{\alpha-1} e^{-\lambda y_i^\alpha}; \qquad F_*(y_i|\,\alpha, \lambda) = 1 - e^{-\lambda y_i^\alpha}, \qquad y_i > 0.$$

With data

$$y_1, \ldots, y_n|\alpha, \lambda \overset{iid}{\sim} \text{Weib}(\alpha, \lambda),$$

the likelihood is

$$L(\alpha, \lambda) \propto \prod_{i=1}^n \lambda\,\alpha y_i^{\alpha-1} e^{-\lambda y_i^\alpha} = \lambda^n \alpha^n \left(\prod_{i=1}^n y_i\right)^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^n y_i^\alpha\right).$$

Discussion of placing an informative prior on (α, λ) is deferred to Chapters 12 and 13 but we select a prior with $\lambda \sim \text{Gamma}(a, b)$, α independent of λ, and α with a prior density $p_0(\alpha)$ having support $(0, \infty)$. The joint prior density has the form

$$p(\alpha, \lambda) \propto p_0(\alpha) \lambda^{a-1} e^{-\lambda b}.$$

The posterior density has the form

$$\begin{aligned} p(\alpha,\lambda|y) &\propto \lambda^n \alpha^n \left(\prod_{i=1}^n y_i\right)^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^n y_i^\alpha\right) p_0(\alpha)\lambda^{a-1}e^{-\lambda b} \\ &\propto \lambda^{a+n-1}\exp\left[-\lambda\left(b+\sum_{i=1}^n y_i^\alpha\right)\right] \alpha^n \left(\prod_{i=1}^n y_i\right)^\alpha p_0(\alpha). \end{aligned}$$

We know of no choice for $p_0(\alpha)$ that makes the joint posterior recognizable.

Gibbs sampling for this problem involves an additional wrinkle. The conditional density for $\lambda \,|\, \alpha, y$ is easily seen to be

$$p(\lambda \,|\, \alpha, y) \propto \lambda^{a+n-1}\exp\left[-\lambda\left(b+\sum_{i=1}^n y_i^\alpha\right)\right],$$

so

$$\lambda \,|\, \alpha, y \sim \text{Gamma}\left(a+n, b+\sum_{i=1}^n y_i^\alpha\right).$$

However, the conditional density for $\alpha \,|\, \lambda, y$ is

$$p(\alpha \,|\, \lambda, y) \propto \alpha^n \left(\prod_{i=1}^n y_i\right)^\alpha \exp\left[-\lambda \sum_{i=1}^n y_i^\alpha\right] p_0(\alpha).$$

Again, we know of no choice for $p_0(\alpha)$ that makes the conditional recognizable. However, we have discussed methods of sampling from an unknown distribution, so we would simply use one of them. The full conditional is log concave provided $p_0(\alpha)$ is log concave, and so it is possible to sample from it using adaptive-rejection sampling. In fact, that is precisely what WinBUGS does for this problem.

Gibbs sampling again begins with an initial value for (α,λ), say (α^1,λ^1). It is easy to sample a new λ for given α^1 by sampling a Gamma$(a+n, b+\sum_{i=1}^n y_i^{\alpha^1})$ variate. Call it λ^2. Next, sample a new value α^2 from the conditional distribution $\alpha \,|\, \lambda^2, y$, and continue until a sufficient number of samples have been taken. We thus obtain $\{(\alpha^k,\lambda^k) : k = 1,\ldots,s\}$, which after a burn-in phase constitutes a sample from the joint posterior distribution.

EXERCISE 6.10. The following code provides an analysis for the data of Example 6.3.2 using independent gamma priors on the parameters and allows inferences about the median time to death and the 24-week survival rate, i.e., $1 - F_*(24|\alpha,\lambda)$ where F_* is the cdf of the Weibull.

```
model{
 for(i in 1:n){ y[i] ~ dweib(alpha,lambda) }
 lambda ~ dgamma(1.53,26.3)
 alpha ~ dgamma(1,1)
 median <- log(2)/lambda
 S24 <- exp(-lambda*pow(24,alpha))
}
list(n=17,
     y=c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65))
list(lambda=0.05, alpha =1)
```

Run the code and make inferences about the median and 24-week survival times. Also check to see if the data might suggest that α is near 1, which would imply that an exponential distribution might suffice as a model for the data.

EXERCISE 6.11. Show that the full conditional for α in Example 6.3.2 is log concave provided $p_0(\alpha)$ is log concave.

EXERCISE 6.12. Argue that the chain generated in Example 6.3.2 by sampling the full conditional for λ from the appropriate Gamma distribution, and the full conditional for α using acceptance-rejection sampling, as described in Subsection 6.2.1, will result in samples from the joint posterior.

*6.3.2.1 Proof that $p(\theta)$ is the Stationary Distribution in the Two-Block Case**

We need to show that (3) holds for the two block case. In the two block case, using the same argument as in (7), the stationary transition density for Gibbs sampling is

$$\begin{aligned} q(\theta^k \mid \theta^{k-1}) &\equiv q(\theta_1^k, \theta_2^k \mid \theta_1^{k-1}, \theta_2^{k-1}) \\ &= p_{1|2}(\theta_1^k \mid \theta_2^{k-1}) p_{2|1}(\theta_2^k \mid \theta_1^k). \end{aligned}$$

In order to establish (3), we must show that

$$\int p(\theta^{k-1}) q(\theta^k | \theta^{k-1}) d\theta^{k-1} = p(\theta^k).$$

From the definition of the Gibbs sampler and using its transition probability density, the left hand side above is

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\theta_1^{k-1}, \theta_2^{k-1}) p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_1^{k-1} d\theta_2^{k-1} \\ = & \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} p(\theta_1^{k-1}, \theta_2^{k-1}) d\theta_1^{k-1} \right] p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ = & \int_{-\infty}^{\infty} p_2(\theta_2^{k-1}) p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ = & \int_{-\infty}^{\infty} p(\theta_1^k, \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ = & \; p_1(\theta_1^k) p_{2|1}(\theta_2^k | \theta_1^k) \\ = & \; p(\theta^k), \end{aligned}$$

as we intended to show. The proof in the multiblock case can be found in Robert and Casella (2004).

6.3.3 Metropolis Algorithm

A very general method of defining an MC is the Metropolis algorithm (Metropolis et al., 1953) as extended by Hastings (1970). This method differs from Gibbs sampling in that it can be used to simulate the entire vector θ at each iteration of the algorithm. It is also used to sample full conditionals of a Gibbs sampler when they are unrecognizable. It is a remarkable algorithm that we show, for a special case, generates a Markov chain with the joint posterior as stationary distribution. Moreover, when used as a method of sampling unrecognizable full conditionals in Gibbs sampling, it can also be shown that a hybrid sampler that replaces a sample from a full conditional with one step of the Metropolis algorithm also has the joint posterior as its stationary distribution (see Robert and Casella, 2004). This algorithm is sometimes referred to as "Metropolis within Gibbs."

The Metropolis algorithm is another type of accept-reject algorithm. It requires a *candidate generating distribution*; sometimes referred to as the *proposal distribution*. The algorithm begins with an initial value θ^1. At the kth iteration, we have $(\theta^1, \theta^2, \ldots, \theta^k)$. The $k+1$st iteration first generates θ^* from a proposal density $h(\theta^* | \theta^k)$. This density should mimic the actual posterior

distribution in some sense, but in theory, it can be any distribution with the same support as the posterior. Define

$$\alpha(\theta^*, \theta^k) = \min\left\{1, \frac{p(\theta^*)h(\theta^k|\theta^*)}{p(\theta^k)h(\theta^*|\theta^k)}\right\} \equiv \alpha.$$

We then simulate $U \sim U[0,1]$ and we select $\theta^{k+1} = \theta^*$ if $U \leq \alpha$ and otherwise take $\theta^{k+1} = \theta^k$. Thus

$$\theta^{k+1} = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*, \theta^k) \\ \theta^k & \text{with probability } 1 - \alpha(\theta^*, \theta^k) \end{cases}.$$

Note that α only uses the ratio of two values of $p(\cdot)$, so it is enough to know the kernel of the posterior density.

EXAMPLE 6.3.3. *Proposal Distribution for Poisson.* Suppose we have iid Pois(θ) data with a conjugate gamma prior as in Subsection 5.3.1. A normal approximation with matching mean and variance is often a reasonable approximation to the Gamma posterior. If we didn't know how to sample from a Gamma(a,b) distribution, we could use the Metropolis algorithm with, say, a $N(\theta^k, a/b^2)$ proposal distribution, namely,

$$h(\theta^*|\theta^k) \propto \exp\{-0.5(\theta^* - \theta^k)^2/(a/b^2)\}.$$

Note that $h(\theta^*|\theta^k) = h(\theta^k|\theta^*)$, so α simplifies to

$$\alpha(\theta^*, \theta^k) = \min\{1, [\theta^*/\theta^k]^{a-1}\exp(b(\theta^k - \theta^*))\},$$

which is easy to compute.

EXERCISE 6.13. Write code in R to sample from a Gamma(a,b) distribution using the Metropolis algorithm with a normal proposal. You might scale the variance of the proposal to see if you can get the acceptance rate in the range of 20-40%. Plot your history of the chain and see if you can identify an appropriate burn-in value. Also try different initial values and compare histories. Try different values of (a,b). For example, the normal candidate should work better if a is moderate to large.

The original Metropolis algorithm assumed that $h(\theta^k|\theta^*) = h(\theta^*|\theta^k)$ so that $\alpha(\theta^*, \theta^k) = \min\{1, p(\theta^*)/p(\theta^k)\}$. This is called the *random walk*. In that case, it is easy to see that we use θ^* if its density $p(\theta^*)$ is larger than $p(\theta^k)$, the density for θ^k. If the density $p(\theta^*)$ for θ^* is smaller than $p(\theta^k)$, we use θ^* with probability $\alpha(\theta^*, \theta^k)$, which will be large when the density of θ^* is nearly as big as the density for θ^k, and will be small when the density of θ^* is much smaller than the density for θ^k. Later we will show that this gives the correct stationary distribution but now we discuss how to implement Metropolis.

Various suggestions have been made about how to choose $h(\theta^*|\theta^k)$. Often, it is taken as a $N(\theta^k, \Sigma^k)$ distribution with various suggestions for Σ^k.

I. Often one can find the posterior mode θ_M of $p(\theta)$ and, up to a constant multiple, the second derivative of $\log[p(\theta)]$, say, $\ddot{\ell}(\theta)$. Define

$$\tilde{\Sigma} = \left[-\ddot{\ell}(\theta_M)\right]^{-1}.$$

It is generally recommended to take

$$\Sigma^k = c\tilde{\Sigma},$$

where c is a "tuning" parameter chosen so that θ^* values are accepted between 20% and 40% of the time.

II. Another proposal is to pick any old Σ_0, use this on every iteration of the MC for a few thousand iterations, compute the sample covariance matrix of the θ^ks from this initial run, say, $\tilde{\Sigma}$, and use $\Sigma^k = c\tilde{\Sigma}$ on all subsequent iterations, where again, c is a tuning parameter chosen to give an acceptance rate between 20% and 40%.

III. A third proposal is similar to the second but uses only the diagonal elements of the sample covariance matrix.

IV. Another proposal, called the *independence* proposal, simply samples from a fixed distribution with density, say $g(\theta)$, possibly a multivariate normal with mean vector equal to the posterior mode and with covariance selected in one of the above ways. Thus all of the proposed samples are iid from this density. However, since the acceptance probability depends on the last value in the chain, the resulting chain is indeed a Markov chain with dependence from one iteration to the next.

The actual Metropolis sampler is a mixture of continuous and discrete distributions since at each iteration, there is often positive probability of the new iterate being identically equal to the last iterate, and also positive probability that it will be the value that was taken from the (continuous) candidate generating distribution. If we use the Metropolis-within-Gibbs hybrid sampler, it is not immediately obvious that the overall chain has the appropriate stationary distribution. While it is not particularly difficult to establish this result, we refer the reader to Tierney (1994).

EXERCISE 6.14. (a) Write an algorithm to sample from the joint posterior for the Weib(α,λ) model assuming independent Gamma priors for α and λ, and using the Metropolis algorithm with a bivariate normal random walk proposal distribution. [This is a random walk in the sense that when at θ^k the next proposed step to θ^* consists of adding an independent $N(0,\Sigma)$ variate to θ^k.] (b) Write an algorithm to perform Gibbs sampling where the full conditional for λ is sampled directly and where one step of the Metropolis algorithm is used to sample the full conditional for α. You must obtain an explicit formula for $\alpha(\theta^*,\theta^k)$ in each case with appropriately defined θ. Don't confuse the α in the Weibull parametrization and the acceptance probability α of the Metropolis algorithm! (c) Write R code to sample from these distributions. (d) Run your R code using the leukemia data of Example 6.3.2. Compare your results with the results from using WinBUGS.

*6.3.3.1 Proof that $p(\theta)$ is the Stationary Distribution**

Denote the MC as θ^r, $r = 1,2,\ldots$. This proof is for $p(\theta)$ discrete. By changing (i,j,k) to $(\theta^*,\theta^r,\theta^{r+1})$ and sums to integrals, the proof works in the "continuous" case. Unfortunately, there really isn't a continuous case. The distributions $p(\cdot)$ and $h(\theta^*|\theta^r)$ may be continuous, but from the definition of the MC, the distribution of θ^{r+1} given θ^r is a mixture of a continuous distribution and a discrete distribution. This causes some technical (measure theoretic) difficulties. Simple justifications for the Metropolis algorithm gloss over this fact.

Given $\theta^r = j$, generate θ^* from density $h(i|j)$. Define

$$\alpha(i,j) = \min\left\{1, \frac{p(i)h(j|i)}{p(j)h(i|j)}\right\}$$

and

$$\theta^{r+1} = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*,j) \\ j & \text{with probability } 1-\alpha(\theta^*,j) \end{cases}.$$

We need to establish (3). Let $g(k)$ be the density for θ^{r+1}. We need to show that if the density for θ^r is $p(j)$ then $g(k) = p(k)$. Let

$$\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}.$$

From the definition of the Markov process, considering all the possibilities of $(\theta^*, \theta^r) = (i, j)$

$$\begin{aligned} g(k) &= \sum_i \sum_j \delta_{ik}\alpha(i,j)h(i|j)p(j) + \sum_i \sum_j \delta_{jk}[1-\alpha(i,j)]h(i|j)p(j) \\ &= \sum_j \alpha(k,j)h(k|j)p(j) + \sum_j \delta_{jk} \sum_i [1-\alpha(i,j)]h(i|j)p(j). \end{aligned} \tag{8}$$

We look at each term individually, starting with the second

$$\begin{aligned} &\sum_j \delta_{jk} \sum_i [1-\alpha(i,j)]h(i|j)p(j) \\ &= \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)=1\}} [1-\alpha(i,j)]h(i|j)p(j) \\ &\quad + \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} [1-\alpha(i,j)]h(i|j)p(j) \\ &= 0 + \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} \left[1 - \frac{p(i)h(j|i)}{p(j)h(i|j)}\right] h(i|j)p(j) \\ &= \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} 1h(i|j)p(j) - \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} p(i)h(j|i) \\ &= \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} [h(i|j)p(j) - p(i)h(j|i)] \\ &= \sum_{\{i:\alpha(i,k)\neq 1\}} [h(i|k)p(k) - p(i)h(k|i)] \\ &= \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \end{aligned}$$

Thus we have all together

$$g(k) = \sum_j \alpha(k,j)h(k|j)p(j) + \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)].$$

Now examine the first term:

$$\begin{aligned} &\sum_j \alpha(k,j)h(k|j)p(j) \\ &= \sum_{\{j:\alpha(k,j)=1\}} \alpha(k,j)h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} \alpha(k,j)h(k|j)p(j) \\ &= \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} \frac{p(k)h(j|k)}{p(j)h(k|j)} h(k|j)p(j) \\ &= \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k). \end{aligned}$$

So all together we have

$$\begin{aligned} g(k) &= \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) \\ &\quad + \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \end{aligned}$$

Finally we note that by the definition of $\alpha(k,j)$

$$\{j : \alpha(k,j) = 1\} = \{j : \alpha(j,k) \neq 1\} \cup \{j : p(k)h(j|k) = p(j)h(k|j)\} \tag{9}$$

so, using this decomposition and recalling that $h(j|k)$ is a well-defined conditional density so that $1 = \sum_j h(j|k)$, we get

$$\begin{aligned} g(k) &= \sum_{\{j:\alpha(j,k)\neq 1\}} h(k|j)p(j) + \sum_{\{j:p(k)h(j|k)=p(j)h(k|j)\}} h(k|j)p(j) \\ &\quad + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) + \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)] \\ &= \sum_{\{j:\alpha(j,k)\neq 1\}} h(j|k)p(k) + \sum_{\{j:p(k)h(j|k)=p(j)h(k|j)\}} p(k)h(j|k) \\ &\quad + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) \\ &= \sum_j h(j|k)p(k) \qquad (10) \\ &= p(k). \end{aligned}$$

EXERCISE 6.15. Use the law of total probability and invent any necessary notation to establish why (8) holds. Also establish why (9) and (10) hold.

6.3.4 Slice Sampling

Slice sampling uses a Markov chain to sample from a *univariate* distribution with density $p(\theta)$. Since we generally won't know the constant of integration, we work with the kernel of the posterior, $p_*(\theta)$. Imagine the two-dimensional graph of $p_*(\theta)$. Slice sampling takes a random walk through the area under the graph of $p_*(\theta)$, alternating steps along the orthogonal axes. To sample a slice of this method

I. Choose an arbitrary θ^1 in the support of $p_*(\theta)$ to initialize the chain. The random walk starts at $(\theta^1, 0)$.

II. Given θ^1, step up towards the density by simulating $u^1 \sim U[0, p_*(\theta^1)]$ and move to (θ^1, u^1).

III. From a point (θ^k, u^k) with a *unimodal* $p_*(\theta)$, step along the θ axis by finding the two points $\theta_{k1} < \theta_{k2}$ with $u^k = p_*(\theta_{k1}) = p_*(\theta_{k2})$ and simulate $\theta^{k+1} \sim U[\theta_{k1}, \theta_{k2}]$. Move to (θ^{k+1}, u^k). For non-unimodal densities, find $\{\theta : p_*(\theta) \geq u^k\}$ and sample from a uniform distribution on that set, say $U[\theta : p_*(\theta) \geq u^k]$. The main difficulty in slice sampling is finding the set $\{\theta : p_*(\theta) \geq u\}$ when $p_*(\theta)$ is not unimodal.

IV. Simulate $u^{k+1} \sim U[0, p_*(\theta^{k+1})]$ and move to (θ^{k+1}, u^{k+1}).

The random walk used for slice sampling is a special case of Gibbs sampling since we alternate between sampling $U|\,\theta = \theta^* \sim U[0, p_*(\theta^*)]$ and $\theta|U = u \sim U[\theta' : p_*(\theta') \geq u]$.

It is not difficult to prove that p is the stationary distribution of the subchain θ^k, and that it is also aperiodic. See Robert and Casella (2004) for details.

EXERCISE 6.16. Let $y \sim \text{Exp}(\theta)$. Give an explicit algorithm for sampling from this distribution using the slice sampler.

EXERCISE 6.17. Assume that there exists a joint density that gives rise to the full conditionals for the slice sampler, i.e., assume that $p(\theta, u)$ gives rise to $p_{\theta|u}(\theta) = p(\theta, u)/p_u(u)$ and $p_{u|\theta}(u) = p(\theta, u)/p_\theta(\theta)$. It immediately follows that $p_{\theta|u}(\theta)/p_{u|\theta}(u) = p_\theta(\theta)/p_u(u) \propto p_\theta(\theta)$. Thus under this assumption, which can be verified for this problem, obtain the kernel of the pdf for θ.

EXERCISE 6.18. We say that the full conditionals for the slice sampler are *compatible* if they uniquely determine a joint distribution. A necessary and sufficient condition for the existence of a

joint density that corresponds to two given conditional densities is (i) that the support of the two full conditionals are identical, i.e., $\{(\theta,u) : p(\theta|u) > 0\} = \{(\theta,u) : p(u|\theta) > 0\}$ and (ii) the ratio of kernels for the full conditionals, say the kernel for $\theta|u$ divided by the kernel for $u|\theta$, is integrable as a function of θ (Arnold and Press, 1989). (a) Using this result, establish that the joint density for (θ,u) for the slice sampler exists. (b) Consider the two conditional distributions: $x|y \sim \text{Exp}(y)$ and $y|x \sim \text{Exp}(x)$. Applying the Arnold and Press result, establish that there does not exist a joint density that is compatible with these two full conditionals.

6.3.5 Checking MCMC Samples

Our first check on convergence of a Markov chain to the stationary distribution $p(\theta)$ is to plot histories (k, θ_j^k) for each component of the parameter vector $j = 1,\ldots,r$, as shown in Figure 6.1. After a burn-in, the chain should not show any trends or patterns. An essentially ideal plot looks like Figure 6.1(a) and the worst plots look like Figure 6.1(c).

A second diagnostic is the autocorrelation function. If there is a lot of autocorrelation, we may have to sample a huge number of values to get reasonable numerical accuracy for our inferences. See Figure 6.2.

WinBUGS provides an estimate of Monte Carlo error that is analogous to a frequentist standard error. For example, if $\theta^1 \ldots, \theta^m$ are iid from $p(\theta)$, a component of the vector θ, say θ_j, has a posterior mean estimated by the sample mean $\bar{\theta}_j$ and a posterior variance estimated by the sample variance s_j^2. The estimated standard deviation of $\bar{\theta}_j$ is $[s_j^2/m]^{1/2}$ and an approximate 95% confidence interval for the posterior mean has endpoints $\bar{\theta}_j \pm 2[s_j^2/m]^{1/2}$. If the sample size m is large enough so that the confidence interval is, say, 10 ± 0.0001, we will be quite happy with the Monte Carlo approximation. On the other hand if our interval is 10 ± 0.5 or 0.37 ± 0.1, we might want to increase m. WinBUGS provides an estimated standard deviation for $\bar{\theta}_j$ that is appropriate for the dependent samples obtained from the Markov chain.

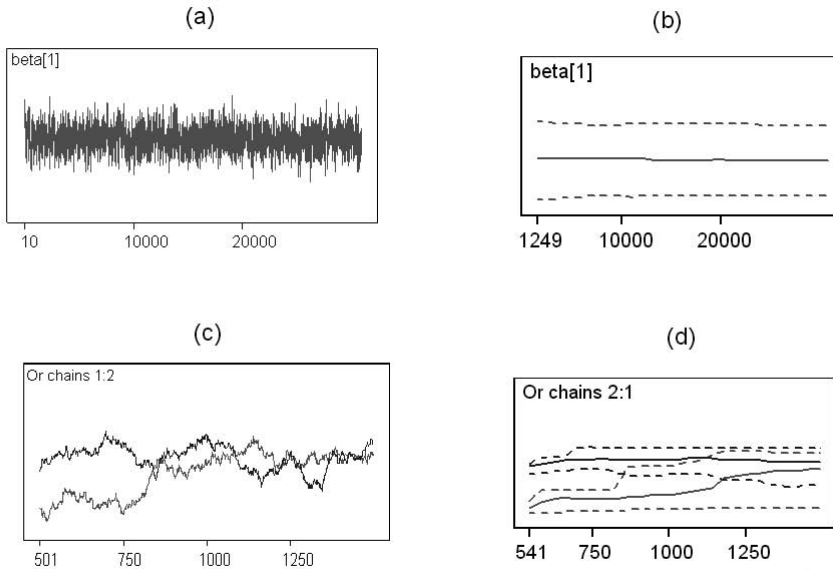

Figure 6.3: *History and quantile plots for chains that: (a–b) have converged after 10 iterations, and (c–d) chains that are not converging.*

Another plot of interest looks at running quantiles as the Markov chain is sampled. It is not easy to estimate the 0.025 and 0.975 quantiles of any sample, since there are fewer data in the tails of the distribution. What we look for in quantile plots is stability after a burn-in phase. Ideally, plots of quantiles are constant through time. Figure 6.3 gives two quantile plots. Figure 6.3(a) shows a Markov chain history that has converged with a burn-in of only 10 iterations and the corresponding quantile plots are in (b). The quantiles plotted are the median, and the 0.025 and 0.975 quantiles, calculated using samples up to the current iteration, across all iterations. Figure 6.3(c) shows the same chains (with two distinct starting values) that were given in Figure 6.1(c); the corresponding quantile plots are in Figure 6.3(d). The quantile plots show two running medians, and two running 0.025 and 0.975 quantiles, which are obviously running amok, indicating serious lack of convergence.

Given a burned-in, thinned chain, one could perform control charts on the data. Control charts for means are specifically designed to validate the assumption that a sample consists of iid observations, cf. Christensen (2001b).

Perhaps the best check is to see whether multiple chains that are initialized at points spread out in the parameter space ultimately converge to the same distribution. Figure 6.1 illustrated convergence and nonconvergence of chains.

A more formal approach takes the following course. Sample independent initial values, $\theta_1^1, \ldots, \theta_m^1$ from a highly dispersed initial distribution $q(\theta)$. For each $i = 1, \ldots, m$, independently generate chains $\theta_i^1, \theta_i^2, \ldots, \theta_i^m$. After a burn-in of BI observations and thinning so that the observations are approximately independent, there should remain, say, s observations on each of m groups. This constitutes data for a balanced one-way ANOVA. Moreover, if the chains have converged to the same distribution, the means should be the same in each group, so the null hypothesis for the standard analysis of variance F test should hold. We can also compare the sample variances and other characteristics of the multiple chains. Gelman and Rubin (1992) and Brooks and Gelman (1998) provide related diagnostics.

EXERCISE 6.19. Using the Weibull model, the data in Example 6.3.2, and the prior specified in Exercise 6.14, run the model with three distinct initial values for both parameters and look at the quantiles plot, history plots, and autocorrelation plots. Determine an appropriate value of the burn-in. Try thinning the chains to see if you can reduce any autocorrelation. Pay particular attention to the Monte Carlo error that is reported and make sure that you ultimately have a large enough MC sample size to achieve reasonable accuracy for all of the posterior means. Comment on all of these issues in a short report.