CHAPTER 6

Metropolis-Hastings algorithms

6.1 Introduction

In this chapter, Markov chains known under a generic name of Metropolis-Hastings will be presented and discussed. This name stems from papers by Metropolis et al. (1953) and Hastings (1970). These are considered as basic papers for the characterization of the method, although other papers including Barker (1965) and Peskun (1973) have also brought relevant contributions to the method.

The original paper by Metropolis et al. (1953) deals with the calculation of properties of chemical substances and was published in the *Journal of Chemical Physics*. Nevertheless, it later proved itself to have a great impact in Statistics and Simulation.

Consider a substance with d molecules positioned at $\theta = (\theta_1, \ldots, \theta_d)'$. In this case, the component θ_i is formed by the bidimensional vector of positions in the plane of the ith molecule. From Statistical Mechanics, the density of these positions is given by Equation (5.1) where a potential V between molecules can be defined. The potential energy of the substance is then given by $E(\theta) = \Sigma_{i,j} V(\theta_i, \theta_j)/2$.

The calculation of the equilibrium value of any chemical property is given by the expected value of this property with respect to the distribution of the vector of positions. Direct calculation of the expectation is not feasible for d large and is replaced by a Monte Carlo estimate. Metropolis et al. (1953) suggested the following method to deal with the difficult problem of sampling from this density:

1. Start with any initial configuration $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})'$ and set the iteration counter $j = 1$.
2. Move the particles from previous positions $\theta^{(j-1)} = (\theta_1^{(j-1)}, \ldots, \theta_d^{(j-1)})'$ according to a uniform distribution centered at these positions in order to obtain new positions $\phi = (\phi_1, \ldots, \phi_d)'$.
3. Calculate the change ΔE in the potential energy caused by the move. The move in step 2 is accepted with probability $\min\{1, e^{-c\Delta E}\}$, with $c = 1/kT$. If the move is accepted, $\theta^{(j)} = \phi$. Otherwise, $\theta^{(j)} = \theta^{(j-1)}$.
4. Change the counter from j to $j+1$ and return to step 2 until convergence is reached.

After convergence, the vector of positions generated by the method has distribution with density (5.1). It is evident that the above method defines a Markov chain as the transitions depend only on the positions at the previous stage. However, it is not obvious that the method converges to an equilibrium distribution and that this distribution is given by (5.1). Metropolis et al. (1953) present a heuristic proof of this result. The same proof is valid for the case where the moves to ϕ are made according to any symmetric distribution centered at previous positions. This defines a transition kernel q that depends on (θ, ϕ) through $|\phi - \theta|$. Hastings (1970) referred to the above algorithm in this extended form as the Metropolis method. In the next section, a more general version of the algorithm and the proof of its convergence will be presented.

Note that the above algorithm includes an additional step that was not present in the chains previously presented. The transition mechanism now depends on a proposed transition q and a subsequent step of evaluation of this proposal. Note that the proposed positions are completely unrelated from the equilibrium distribution but this is represented in the overall transition through the acceptance probability because

$$\frac{\pi(\phi)}{\pi(\theta^{(j-1)})} = \frac{\exp\{-cE(\phi)\}}{\exp\{-cE(\theta^{(j-1)})\}} = \exp\{-c\Delta E\} .$$

Another important point is that the resulting chain may remain in a low energy (or equivalently, high density) position for many iterations. In this case, it is likely that the proposal will lead to very high energy (very low density) points and $\Delta E \gg 0$ forcing an acceptance probability very close to 0. Computationally, this is not desirable and transition kernels must be carefully chosen to avoid such low acceptance rates.

The next section presents a more common and complete version of the algorithm following the work of Hastings (1970). Important special cases are presented in Section 6.3 and in Section 6.4 variations of the method are discussed. These variations include blocking and the relationship with Gibbs sampling. Finally, application of the algorithm to the context of generalized linear models with hierarchical, dynamic and spatial structure is discussed. A very nice expository introduction to Metropolis-Hastings algorithms is also provided by Chib and Greenberg (1995).

Example 6.1 *It is common in pharmacology studies to specify concentration levels of substances introduced in a system by non-linear equations of the form*

$$f(\psi, x) = \psi_1 + \frac{\psi_2 x}{\psi_3 + x} \tag{6.1}$$

where $\psi = (\psi_1, \psi_2, \psi_3)$ and x is an explanatory variable. Equation (6.1) represents a curve starting at ψ_1 when $x = 0$ and advancing to the asymptotic value $\psi_1 + \psi_2$ when $x \to \infty$. Carlin and Louis (2000, p. 233-234) used this model to explain the effect of the concentration x of Puyromycin (in ppm) on the velocity y of an enzymatic reaction (in counts/min^2). Their model assumes that $y_i = f(\psi, x_i) + \epsilon_i$, $i = 1, \ldots, n$, and that observation errors ϵ_i are normally distributed with zero mean and variance σ^2. Table 6.1 exhibits the $n = 12$ observations. Assume that $(\psi_1, \psi_2, \sigma^2) = (50, 170, 126)$ and set $\psi_3 = \theta$. The prior for θ is $N(0, 100)$. The posterior distribution for θ is

x	*0.02*	*0.02*	*0.06*	*0.06*	*0.11*	*0.11*
	0.22	*0.22*	*0.56*	*0.56*	*1.10*	*1.10*
y	*76*	*47*	*97*	*107*	*123*	*139*
	159	*152*	*191*	*201*	*207*	*200*

Table 6.1 *y: velocity of an enzymatic reaction (in counts/min^2), x: substrate concentration (in ppm).*

$$\pi(\theta) \propto f_N(\theta; 0, 100) \prod_{i=1}^{n} f_N(y_i; 50 + 170x_i/(\theta + x_i), 126). \tag{6.2}$$

Consider the transition kernel $q(\theta, \phi) = f_N(u; 0, 0.01)$, where $u = \theta - \phi$. A typical path is presented in Figure 6.1. The horizontal segments of the trajectory represent iterations where proposed values were repeatedly not accepted. In these cases, the iteration values remained the same until a proposed value is accepted. In total, 188 values were accepted out of the 1000 proposed, thus giving an average acceptance rate for this chain of approximately 18%.

In this setting, inference for θ can be based on the sample obtained from running the chain. After 1000 iterations and discarding the first 100 iterations, the posterior mean and standard deviation of θ are approximately given by 0.132 *and* 0.013. *A 95% credibility interval for θ is given by* $[0.105, 0.156]$.

6.2 Definition and properties

Consider a distribution π from which a sample must be drawn via Markov chains. Again, it is worth stressing that this task will only make sense if the non-iterative generation of π is very complicated or expensive. In this case, a transition kernel $p(\theta, \phi)$ must be constructed in a way such that π is the equilibrium distribution of the chain. A simple way to do this is to consider reversible chains where the kernel p satisfies

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta)\, , \forall\, (\theta, \phi). \tag{6.3}$$

As previously seen in Section 4.6, this is the reversibility condition of the chain. Equation (6.3) is also referred to as the detailed balance equation.

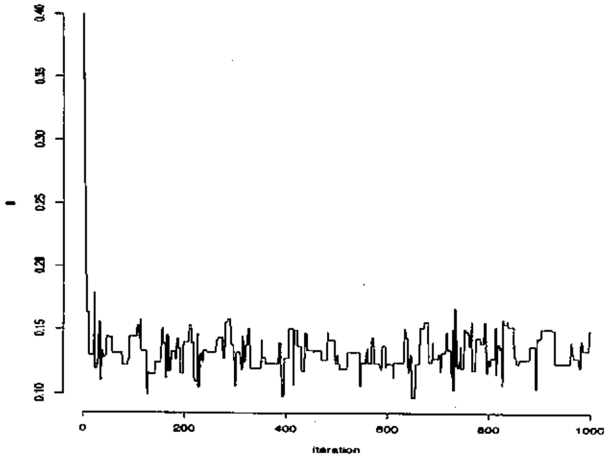

Figure 6.1 *Trajectory of the chain for θ in Example 6.1 with initial value* $\theta^{(0)} = 0.4$.

Even though this is not a necessary condition for convergence, it is a sufficient condition in order that π be the equilibrium distribution of the chain.

The kernel $p(\theta, \phi)$ consists of 2 elements: an arbitrary transition kernel $q(\theta, \phi)$ and a probability $\alpha(\theta, \phi)$ such that

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi)\,, \text{ if } \theta \neq \phi.$$

So, the transition kernel defines a density $p(\theta, \cdot)$ for every possible value of the parameter different from θ. Consequently, there is a positive probability left for the chain to remain at θ given by

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi.$$

These two forms can be grouped in the general expression

$$p(\theta, A) = \int_A q(\theta, \phi)\alpha(\theta, \phi)d\phi + I(\theta \in A)\left[1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi\right] \quad (6.4)$$

for any subset A of the parameter space. So, the transition kernel defines a mixed distribution for the new state ϕ of the chain. For $\phi \neq \theta$, this distribution has a density and for $\phi = \theta$, this distribution has a probability atom.

Hastings (1970) proposed to define the acceptance probability in such a way that when combined with the arbitrary transition kernel, it defines a reversible chain. The expression most commonly cited for the acceptance probability is

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)}\right\}. \quad (6.5)$$

Algorithms based on chains with transition kernel (6.4) and acceptance probability (6.5) will be referred to as Metropolis-Hastings algorithms from now on. This is an acknowledgement of the importance of the contribution from both papers. Hastings (1970) referred to the ratio appearing in (6.5) as the test ratio. A more general expression for α including (6.5) and the acceptance probability used by Barker (1965) as special cases is presented in Hastings (1970). Optimality of these choices can be discussed in terms of minimization of asymptotic variance of moment estimates. Peskun (1973) showed for the discrete case that (6.5) is optimal in a large class of choices. He also showed that for suitable choices of the proposal transition q, Markov chain sampling can be more precise than independent sampling.

The proof that (6.3) is satisfied by p given in (6.4) and hence defines a reversible chain with equilibrium distribution π follows directly from (6.5) and is left as an exercise. Note that (6.3) is satisfied by p but not by q. The proposal transition kernel q has up to now been kept arbitrary and is thus a flexible tool for the construction of the algorithm. Roberts and Smith (1994) showed that if q is irreducible and aperiodic and $\alpha(\theta, \phi) > 0$, for every possible value of (θ, ϕ), then the algorithm defines an irreducible and aperiodic chain with transition kernel p given by (6.4) and limiting distribution π.

In practical terms, simulation of a draw from π using the Markov chain defined by the transition (6.4) can be set up as follows:

1. Initialize the iteration counter $j = 1$ and set an arbitrary initial value $\theta^{(0)}$.
2. Move the chain to a new value ϕ generated from the density $q(\theta^{(j-1)}, \cdot)$.
3. Evaluate the acceptance probability of the move $\alpha(\theta^{(j-1)}, \phi)$ given by (6.5). If the move is accepted, $\theta^{(j)} = \phi$. If it is not accepted, $\theta^{(j)} = \theta^{(j-1)}$ and the chain does not move.
4. Change the counter from j to $j+1$ and return to step 2 until convergence is reached.

Step 3 is performed after the generation of an independent uniform quantity u. If $u \leq \alpha$, the move is accepted and if $u > \alpha$ the move is not allowed. The transition kernel q defines only a possible move that can be confirmed according to the value of α. For that reason, q is generally referred to as the proposal kernel or proposal (conditional) density when looked upon as a (conditional) density $q(\theta, \cdot)$. Other terms sometimes used are probing kernel or density.

In any of the forms of the Metropolis algorithm, q defines a symmetric transition around the previous positions of the molecules. Therefore,

$q(\theta, \phi) = q(\phi, \theta)$, for every (θ, ϕ) and the acceptance probability becomes

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)}{\pi(\theta)}\right\} .$$

depending only on a simplified test ratio $\pi(\phi)/\pi(\theta)$, the ratio of the posterior density values at the proposed and previous positions of the chain.

Note also that the chain may remain in the same state for many iterations. A useful monitoring device of the method is given by the average percentage of iterations for which moves are accepted. Hastings (1970) suggests that this acceptance rate should always be computed in practical applications.

The success of the method depends on not having a very low acceptance rate. A naive approach to the problem is to make the chain move very slowly, i.e., to drive the chain so that its displacements are minute. Assuming for simplicity that $q(\cdot, \cdot)$ and $\pi(\cdot)$ are continuous, similar values for previous and proposed states will lead to a test ratio and hence acceptance probability close to 1. Following this strategy, the chain will have very high acceptance rates and most proposed moves are accepted. The chain however must be capable of traversing the whole parameter space in order to converge to the equilibrium distribution. Very small moves will make it takes many iterations to converge. On the other hand, large displacements may be proposed but they are likely to fall in the tails of the posterior distribution causing a very low value for the test ratio. The chain moves, determined by q, must be paced in such a way as to provide considerable displacements from the current state but with substantial probability, determined by α, of being accepted.

It is also crucial that the proposal kernels are easy to draw from as the method replaces the difficult generation of π by many generations proposed from q. Another less obvious but equally important requirement to be met by q is the correct tuning of the moves it proposes to ensure that moves covering the parameter space can be made and accepted in real computing time.

Optimization studies in this area are not conclusive and are likely never to be. The diversity of models that can be treated and of transitions q that can be proposed make it extremely difficult to allow for general results. Current reasoning, expressed in an applied context in Bennett, Racine-Poon and Wakefield (1995), Besag et al. (1995) and other authors seem to indicate to the direction of acceptance rates between 20% to 50%. In a specific theoretical context, Gelman, Roberts and Gilks (1996) obtained optimal acceptance rates of 24% for high-dimensional problems with normal densities π and q. These values should be looked at only as a generic indication rule and never as a compulsory determination. In the final section, an application with two sampling schemes with very high acceptance rates (larger than 90%) is shown. The performance of the schemes are very different with one of them showing fast convergence whereas the second one has a very slow convergence.

The test ratio can be rewritten as

$$\frac{\pi(\phi)/q(\theta, \phi)}{\pi(\theta)/q(\phi, \theta)} . \tag{6.6}$$

Acceptance of proposed values is based on the ratio of target and proposed density. So, there is a connection here with the resampling schemes described in Section 1.5. There, the proposal density q was to be chosen as similar as possible to π to increase acceptance rates but the methods were not iterative. Also, for the rejection method, the rejection probability depended only on the numerator in (6.6).

The target distribution π enters the algorithm through the test ratio in the form of the ratio $\pi(\phi)/\pi(\theta)$, as in the resampling methods. So again, the complete knowledge of π is not required. In particular, proportionality constants are not needed. When π is a posterior density, even though its functional form is always known, the value of the proportionality constant is rarely known. So, the algorithm is particularly useful for applications to Bayesian inference.

Many of the comments made about Gibbs sampling in the previous chapter are also valid for the Metropolis-Hastings algorithm. So, the discussion about single long against multiple chains is just as relevant here. In using a single long chain, particular attention must be given to spacing between values of the chain taken for the resulting sample. The possible repetition of the same state for many iterations does not hinder convergence properties but makes it more common to produce sequences of repeated values, even after some spacing is allowed. A sample must adequately cover the complete parameter space so it is important that its values are not unnecessarily influenced by their predecessors.

Formal and informal convergence techniques described in Chapter 5 can all be used here. The exception is made up of those based on complete knowledge of conditional densities. Typically, but not necessarily, Metropolis-Hastings algorithms are used when these are not completely known and hence difficult to sample from. When the complete conditional densities are known, Gibbs sampling is generally used. Besag et al. (1995) argued against taking this approach as a general rule. They reasoned that Gibbs sampling does not take into account the previous value of the component being updated and is therefore restrictive. Questions relative to optimization of the algorithm through reparametrization or blocking are deferred to Section 6.4.

6.3 Special cases

As described in the previous section, there is total flexibility for the choice of the proposal transition q apart from a few technical restrictions. Some general considerations have already been made and now we turn to some specific classes. It should be pointed out that although a chain is defined by its transition kernel p and not by a proposal transition q, the names used to categorize the algorithm generally refer to properties of q rather than p.

6.3.1 Symmetric chains

A chain is said to be symmetric if its transition kernel p is symmetric in its arguments, namely $p(\theta, \phi) = p(\phi, \theta)$, for every pair (θ, ϕ) of states. For the Metropolis-Hastings algorithms, the notion of symmetric chain is applied to the proposed transition q. An example of a symmetric chain is the Metropolis version of the algorithm. If q depends on (θ, ϕ) only through $|\phi - \theta|$ then $q(\theta, \phi) = q(\phi, \theta)$. In this case, the acceptance probability reduces to $\min\{1, \pi(\phi)/\pi(\theta)\}$ and does not depend on q. A computational simplification that may well prove to be substantial is thus obtained.

6.3.2 Random walk chains

Again, this characterization refers to the proposal transition q. From Chapter 4, we know that a random walk is a Markov chain with evolution given by $\theta^{(j)} = \theta^{(j-1)} + w_j$ where w_j is a random variable with distribution independent of the chain. In general, the disturbances w_j are independent and identically distributed with density f_w. The chain has proposed moves according to $q(\theta, \phi) = f_w(\phi - \theta)$. If f_w is symmetric around 0, the chain is symmetric and all comments above are valid here. The Metropolis algorithm can then be seen as a special case of a random walk chain.

This is a very common option and most practical implementations of Metropolis-Hastings algorithms use this scheme. The most used choices for f_w are the normal (Muller, 1991b) and Student's t (Geweke, 1992) distributions centered at the origin. Proposed values are then based around the previous values of the chain. An important point still remaining is the choice of the dispersion of f_w. Large values for the variance allow moves that are very distant from previous values but at the likely cost of very small acceptance rates. On the other hand, small values for the variance only allow moves close to the previous values but with high acceptance rates. Tierney (1994) suggested setting the variance matrix of f_w as cV where c is a multiplying scalar playing the role of a tuning constant and V is some form of approximation for the posterior variance (see Chapter 3). This allows the moves along the components of θ to be of the same size relative to the spread of the posterior distribution. The choice of the tuning constant depends on the form of optimization desired (high acceptance rates/large moves). Metropolis et al. (1953) discussed this issue in the context of their application. Tierney (1994) suggested values between 1/2 and 1. Bennett, Racine-Poon and Wakefield (1995) reported the use of a variety of values in the context of non-linear hierarchical models and also obtained the same recommendation based on sample sizes prescribed by the technique of Raftery and Lewis (1992). Gelman, Roberts and Gilks (1996) obtained their optimal acceptance rates for normal random walk proposals with c between 2 and 3.

Example 6.2 *The performance of the random walk Metropolis algorithm is investigated when the target distribution is a two-component mixture of bivariate normal densities*

$$\pi(\theta) = 0.7 f_N(\theta; \mu_1, \Sigma_1) + 0.3 f_N(\theta; \mu_2, \Sigma_2).$$

where

$$\mu_1 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \ \mu_2 = \begin{pmatrix} 0.7 \\ 3.5 \end{pmatrix}, \ \Sigma_1 = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 1.0 & -0.7 \\ -0.7 & 1.0 \end{pmatrix}.$$

Figure 6.2 reveals the bimodal aspect of $\pi(\theta)$. Pretend that direct sampling from $\pi(\theta)$ is not possible and instead sample using a random walk Metropolis with proposal $q(\theta, \phi) = f_N(\phi; \theta, \nu I_2)$. Figure 6.3 shows the effects of the initial value and the tuning parameter ν in the chains.

Notice that the chains perform poorly for both small and large values of the tuning parameter. For small values of the tuning parameter, the chains have difficulty moving across the parameter space. As a result, the trajectories provide inappropriate representations of the target distribution. For large values of the tuning parameter, the chains visit most of the parameter space but have low acceptance rates, leading to a computationally inefficient sampling scheme. For reasonable values of the tuning parameter the chains explore efficiently the parameter space and provide good approximations to the posterior distribution, with an effective sample size of around 200 for a sample of 5000 iterations, for $\nu = 1$ and $t(\theta) = \theta_2$. The disparity between the effective and actual sample sizes is due to the large autocorrelation structure imposed by the random walk nature of the chain. Figure 6.4 illustrates this point. See Exercise 6.9 for a univariate version of this example.

6.3.3 Independence chains

In this case, the proposed transition is formulated independently of the previous position θ of the chain. So, $q(\theta, \phi) = f(\phi)$. It may seem that the independence from the previous state disagrees with the Markovian property of the chain. Once again, it is worth remembering that q is just

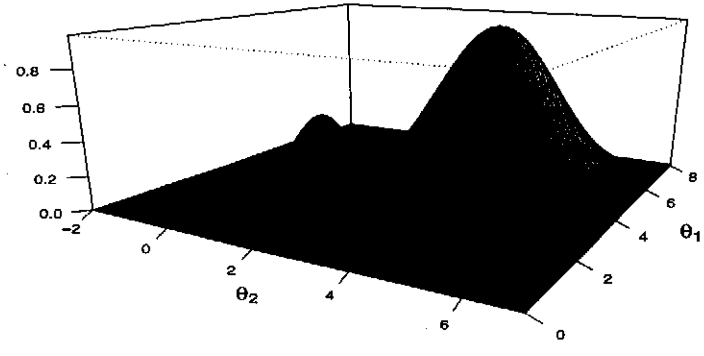

Figure 6.2 *Mixture of normals:* $\pi(\theta) = 0.7f_N(\theta;\mu_1,\Sigma_1) + 0.3f_N(\theta;\mu_2,\Sigma_2)$.

a proposal that is combined with an acceptance probability α to give the transition kernel p of the algorithm. This transition depends on the previous state, preserving the Markovian structure.

Using expression (6.6) for the test ratio, it reduces to $w(\phi)/w(\theta)$ where $w = \pi/f$. The weight function works like the weight function in weighted resampling methods (Section 1.5.2). One popular choice for f is the prior density as in Section 3.5.2 (West, 1996; Knorr-Held, 1997). In this case, $w = l$, the likelihood function and the acceptance probability is $\alpha(\theta,\phi) = \min\{1, l(\phi)/l(\theta)\}$.

The use of the prior distribution as the basis for a resampling scheme was discussed in Section 3.5 with advantages and disadvantages equally relevant here. Computationally, it has the advantages of producing one of the simplest expressions for α. The main disadvantage is when there is conflict between prior and likelihood information. Values more likely to be sampled are those supported by the prior and they will have little posterior support. Values highly supported by the likelihood (and probably by the posterior) will have very little chance of being drawn by such a scheme.

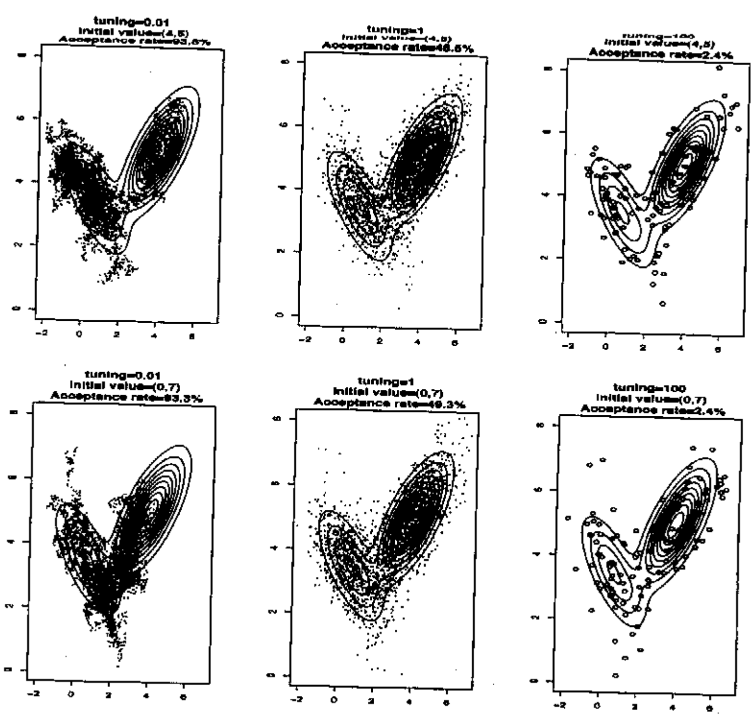

Figure 6.3 *Performance of the random walk Metropolis algorithm. The tuning parameter ν is set at* 0.01, 1 *and* 100 *while two different initial values are used:* (4, 5) *and* (0, 7). *Acceptance rates are based on* 5000 *draws.*

Proposal densities incorporating the likelihood help to avoid this situation. Normal approximations to the likelihood were used by Bennett, Racine-Poon and Wakefield (1995) and Chib and Greenberg (1994) to form normal independence proposals with and without combination with the prior distribution. Jacquier, Polson and Rossi (1994) use moment matching approximation to construct a Gamma proposal in the context of stochastic volatility models.

The general rule for independence chains is to avoid large variation in the weight function as this increases the chances of a chain being retained for many iterations in states with large weights. So, it is recommended that f is chosen in order to make the function w as constant as possible, or at least bounded. As f and π are both densities, this is equivalent to recommending

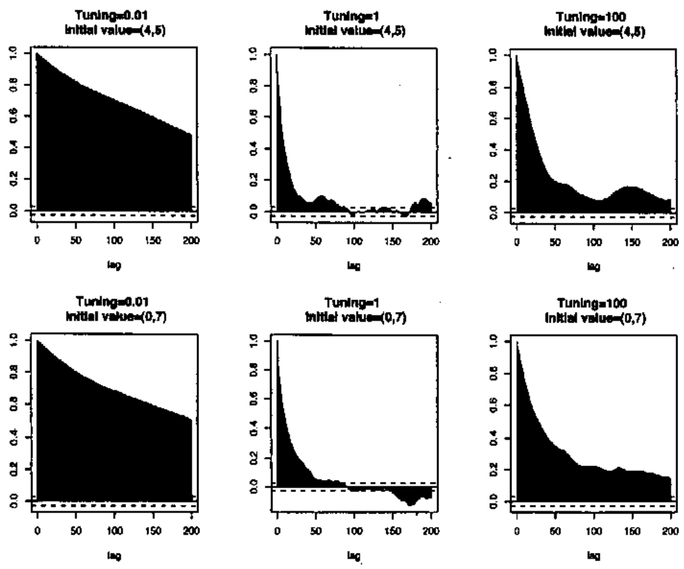

Figure 6.4 *Autocorrelation functions for θ_2 for the random walk Metropolis algorithm.*

that f and π are as similar as possible. That rules out the use of a prior proposal in case of disagreement between prior and likelihood.

Tierney (1994) suggested avoiding densities f with thin tails such as the normal distribution and to use instead t densities with small numbers of degrees of freedom. In this way, the weight function will not be so strongly affected by the tails of f. By doing that, the weight function becomes less likely to have large variations.

The very minimum requirement for such f is to allow sampling from all probable values of π. Otherwise, the resulting chain will almost never visit all likely values of π and resulting sample will be misleading. This recommendation was crucial for simpler Monte Carlo schemes (see Figure 1.4) and remains crucial for MCMC schemes (see example below). Thus, overdispersed proposal distributions should always be preferred over underdispersed proposals to ensure appropriate sampling over appropriate regions of the parameter space.

Example 6.2 *(continued) Consider an independence Metropolis-Hastings algorithm with proposal $q(\theta, \phi) = f_N(\phi; \mu_3, \nu I_2)$, with $\mu_3 = (3.01, 4.55)'$ the mean of the target distribution and ν is a tuning parameter.*

Figure 6.5 shows the effects of the initial value and the tuning parameter in the chains. Again, extremely small and extremely large tuning parameter make the chains move slowly and get trapped. For reasonable values of the tuning parameter the chains explore adequately the parameter space and provide reasonable approximations to the posterior distribution. They do that more efficiently than the random walk proposals because of the low autocorrelation structure implied by the independence form of the proposal (see Figure 6.6). Their acceptance rates are lower but, more importantly, their effective sample sizes, for $\nu = 5$ and $t(\theta) = \theta_2$, are between 600 and 1200, much larger than the effective sample size obtained with the optimized random walk proposal. The independence Metropolis chains outperform the random walk chains in this case. See again Exercise 6.9 for a univariate version of this example.

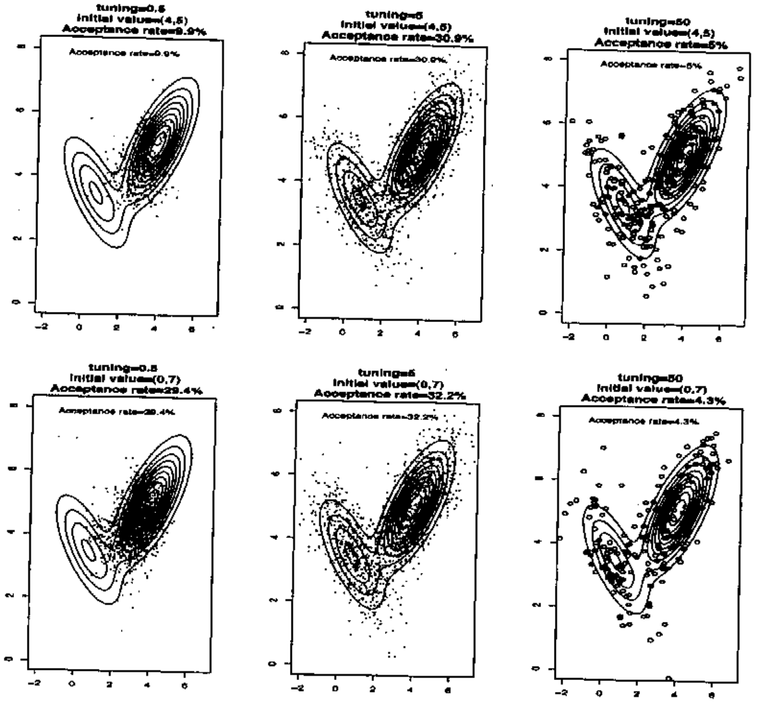

Figure 6.5 *Performance of the independence Metropolis-Hastings algorithm. The tuning parameter ν is set at* 0.5, 5 *and* 50 *while two different initial values are used:* (4, 5) *and* (0, 7). *Acceptance rates are based on* 5000 *draws.*

The message from the above example is that independence proposals are better than random walk proposals. This was due to the fact that in this

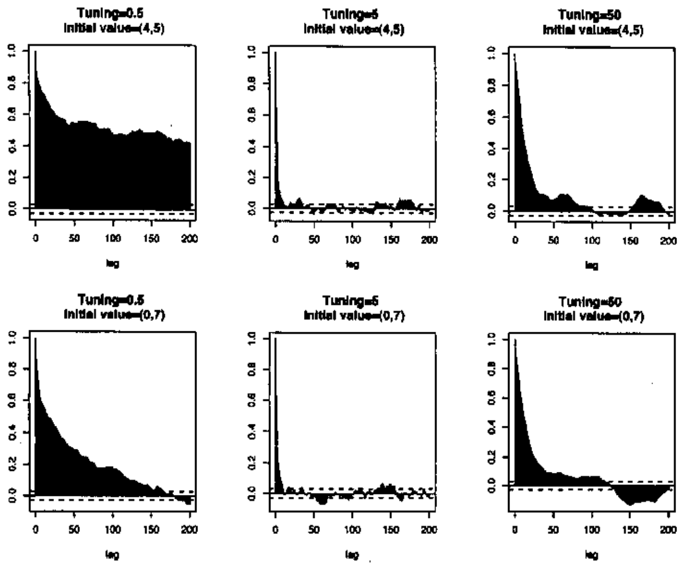

Figure 6.6 *Autocorrelation functions for θ_2 for the independence Metropolis algorithm.*

simple, bivariate context it was rather simple to envision a good approximation to the target distribution. Such feature was successfully applied there. Nonetheless, in more general and higher dimensional problems, it is very difficult to find suitable proposals for the independence sampler. In those cases, it is easier to design and implement suitable random walk proposals.

6.3.4 Other forms

Tierney (1994) presented many other classes of proposal transitions that may be used. In particular, he discussed the use of the rejection method in independence chains. In the case of sampling from π based on rejection sampling from f, f should be an envelope function for π. This requires finding a constant A to achieve this task, which can be difficult due to the complicated form of π. Large values of A ensure a proper envelope at the expense of high rejection rates and small values of A may not ensure a complete envelope. This structure can be put in the context of the Metropolis-Hastings algorithm without having to enforce a (possibly difficult to obtain) specification of the envelope constant A (Exercise 6.6). Chib and Greenberg (1995) gave the details of this application of rejection methods in the Metropolis-Hastings context.

Gilks, Best and Tan (1995) generalized the adaptive rejection sampling scheme to non-log-concave densities using the same idea. The approximating piecewise exponential density does not provide an envelope for these densities. The rejection step may then be replaced by a Metropolis-Hastings step just as outlined above. Limiting results of Metropolis-Hastings algorithms ensure that sampling is still correct.

Another possibility is the extension of random walk chains to autoregressive chains where $\theta^{(j)} = a + b\theta^{(j-1)} + w_j$ (see Section 2.6). Taking the value of $b = 1$ reduces to random walk chains with incorporation of the constant a to the distribution f_w. Taking $b = -1$ produces alternations in the chain forcing negative autocorrelations. These two choices of b are compared for the univariate normal distribution by Hastings (1970) and for a bivariate normal distribution with high correlation by Chib and Greenberg (1995). In both examples, the alternating chain produces better estimates of the target distribution. Barone and Frigessi (1989) provided further theoretical support for this choice of value of b. This alternating effect is not new in simulation and is generally known as antithetic variables with good properties in terms of reducing variance of estimates (Ripley, 1987).

The algorithm of Ritter and Tanner (1992) presented in the previous chapter is, in fact, based on an approximation to the target distribution. This can be improved upon by incorporating it into a Metropolis-Hastings algorithm as the proposal kernel. If the kernel admits a density then this proposal density will appear in the expression of the acceptance probability.

Finally, further proposal densities that do not fit into any of the above schemes will be presented in the applications of Section 6.5. These proposals were motivated by the structure of the model and in similar inferential procedures already available. The diversity of options presented in this section serves the purpose of showing the vast field available for exploration when simulating via Markov chains. Most of them have only started to be explored.

6.4 Hybrid algorithms

In this chapter, a simulation scheme using Markov chains called the Metropolis-Hastings algorithm is being presented. This scheme was introduced in a general form in Section 6.2 and some special cases were presented in Section 6.3. The aim of this section is to present some of the capabilities of the scheme especially in connection with componentwise transition and combinations of different transition schemes.

In the previous chapter, a scheme based on transition by components was presented, the Gibbs sampler. The similarity between the componentwise sampling schemes will be clarified and schemes combining the two algorithms are presented. Also in this section, a discussion about blocking and reparametrization is made.

6.4.1 Componentwise transition

The previous section showed some of the possibilities that are available with Metropolis-Hastings algorithms. In all cases presented, the quantity of interest for sampling θ was updated in a single block. Again here, new transition forms are available when the components $\theta_1,\ldots,\theta_d$ of θ are used separately. Hastings (1970) and Tierney (1994) discussed this possibility. In particular, the components of θ can be updated or changed in the following forms:

a) At each iteration, a single component is updated and the choice of the component is made at random between the d components.

b) At each iteration, a single component is updated and the choice of the component is made in a fixed pre-specified order of the d components. For example, the components are updated in the order $1 \rightarrow 2 \rightarrow \cdots \rightarrow d$.

The above forms are examples of mixtures in case (a) and cycles in case (b) of transitions. In the first case, define transitions p_m with a common equilibrium distribution π and probabilities or weights w_m, $m=1,\ldots,r$, satisfying $w_m \geq 0$ and $\sum_{m=1}^r w_m = 1$. A mixture transition p is formed by taking $p=\sum_{m=1}^r w_m p_m$. Case (a) above is a special case of a mixture with $r=d$, $w_m=1/d$ and each transition p_m moves only the mth component of θ, $m=1,\ldots,d$.

Properties of the transitions p_m are passed on to the mixture transition p. First of all, the mixture kernel p defines a transition kernel of a Markov chain with equilibrium distribution π. Also, if one of the component transition kernels is irreducible and aperiodic then the mixture kernel is irreducible and aperiodic.

In the case of cycle transition kernels with component transition kernels p_c, $c=1,\ldots,r$, an iteration of the new chain is performed after undergoing all moves dictated by the component kernels. For a move from θ to ϕ in a single iteration of a cycle chain, all possible moves to an intermediate state ψ_c through p_c, $c=1,\ldots,r-1$, finally leading to $\psi_r=\phi$ through p_r must be considered. Defining the initial state $\psi_0=\theta$ gives

$$p(\theta,\phi)=\int\cdots\int\prod_{c=1}^{r}p_c(\psi_{c-1},\psi_c)d\psi_1\ldots d\psi_{r-1}$$

which generalizes results from Section 4.6. Case (b) above is the special case of a cycle with $r=d$ and each transition p_m moves only the mth component of θ, $m=1,\ldots,d$.

Many of the properties of the transitions p_c are passed on to the cycle transition p. First, the cycle kernel p defines a transition kernel of a Markov chain with equilibrium distribution π. Unlike mixture kernels, irreducibility and aperiodicity of one of the component transition kernels are not in general sufficient for irreducibility and aperiodicity of the cycle kernel. If all the component kernels are irreducible and aperiodic then the cycle kernel is irreducible and aperiodic (Tierney, 1994).

These forms can be integrated in to the Metropolis-Hastings algorithm. Each of the transition kernels p_i above may be given by a proposal kernel q_i and an acceptance probability α_i. Consider now the cycle scheme with componentwise transitions, namely, the component transition kernel $q_i(\theta,\phi)$ proposes a move of the ith component of θ, $i=1,\ldots,r$. From Chapter 2, $\pi(\theta)=\pi_i(\theta_i)\pi(\theta_{-i})$ where π_i is the full conditional density of θ_i. The move determined by q_i only changes θ_i, so $\theta_{-i}=\phi_{-i}$ and $\pi(\phi)/\pi(\theta)=\pi_i(\phi_i)/\pi_i(\theta_i)$. Note that as far as the transition p_i is concerned, the other components of θ remain fixed and are not affected. Thus, it defines a reducible Markov chain. The proposal transition q_i may then be written in the form $q_i(\theta_i,\phi_i)$ even though it may well depend on the value of θ_{-i}. Consequently, the acceptance probability may also be written as

$$\alpha_i(\theta_i,\phi_i)=\min\left\{1,\frac{\pi_i(\phi_i)q_i(\phi_i,\theta_i)}{\pi_i(\theta_i)q_i(\theta_i,\phi_i)}\right\} . \tag{6.7}$$

Note that each of the component transition kernels above defines a reversible chain with equilibrium distribution $\pi_i(\theta_i)$, $i=1,\ldots,d$. Namely, each component transition kernel satisfies the equation

$$\pi(\phi_i|\theta_{-i})=\int\pi(\theta_i|\theta_{-i})p_i(\theta_i,\phi_i)d\theta_i .$$

Considering only two components, a move from $\theta=(\theta_1,\theta_2)$ to $\phi=(\phi_1,\phi_2)$ is formed after moves for the two components are operated according to their respective kernels. If π is a stationary distribution for this cycle kernel, it must satisfy

$$\pi(\phi)=\int\int\pi(\theta)p_1(\theta_1,\phi_1)p_2(\theta_2,\phi_2)d\theta . \tag{6.8}$$

The right hand side of (6.8) can be rewritten as

$$\begin{aligned}
&\int\int\pi(\theta_1|\theta_2)\pi(\theta_2)p_1(\theta_1,\phi_1)p_2(\theta_2,\phi_2)d\theta_1 d\theta_2\\
=&\int\pi(\theta_2)\left[\int\pi(\theta_1|\theta_2)p_1(\theta_1,\phi_1)d\theta_1\right]p_2(\theta_2,\phi_2)d\theta_2\\
=&\int\pi(\theta_2)\pi(\phi_1|\theta_2)p_2(\theta_2,\phi_2)d\theta_2\\
=&\int\pi(\phi_1)\pi(\theta_2|\phi_1)p_2(\theta_2,\phi_2)d\theta_2\\
=&\ \pi(\phi_1)\int\pi(\theta_2|\phi_1)p_2(\theta_2,\phi_2)d\theta_2
\end{aligned}$$

$$
\begin{aligned}
&= \pi(\phi_1)\pi(\phi_2|\phi_1) \\
&= \pi(\phi),
\end{aligned}
$$

confirming the validity of Equation (6.8). The second and fourth equalities above follow from the stationarity of the conditional densities and the others follow from basic probability and integration operations. The result can be extended to any number of components using induction on the same argument. The derivation above was based on absolutely continuous transition kernels for notational simplicity. All results remain valid with minor technical changes for the mixed kernels of the Metropolis-Hastings algorithm. It can also be shown that despite the reducibility of the component transition kernels, the cycle kernel is irreducible and aperiodic (Tierney, 1994). So, the limiting distribution π is unique.

Once that is done, a new, componentwise version of the Metropolis-Hastings algorithm is given by:

1. Initialize the iteration counter $j = 1$ and set the initial value of the chain $\theta^{(0)}$.
2. Initialize the component counter $i = 1$.
3. Move the ith component of the vector of states of the chain to a new value ϕ_i generated from the density $q_i(\theta_i^{(j-1)}, \phi_i)$.
4. Calculate the acceptance probability of the move $\alpha_i(\theta_i^{(j-1)}, \phi_i)$ given by (6.7). If the move is accepted, $\theta_i^{(j)} = \phi_i$. If the move is not accepted, $\theta_i^{(j)} = \theta_i^{(j-1)}$ and the chain does not move.
5. Change the counter from i to $i + 1$ and return to step 3 until $i = d$. When $i = d$, go to step 6.
6. Change the counter from j to $j+1$ and return to step 2 until convergence is reached.

In fact, this was the form of the algorithm originally proposed by Metropolis et al. (1953). The positions of the molecules were modified one by one according to a symmetric transition with uniform bivariate distribution centered at the previous position of the molecule. The algorithm was introduced in this book with a single global transition to unify the presentation with the work of Hastings (1970).

Example 6.3 *Bennett, Racine-Poon and Wakefield (1996) compared many MCMC schemes in the context of longitudinal data studies with the non-linear mean structure. The response y_{ij} of individual i at time t_{ij} is explained by the non-linear regression model*

$$y_{ij} = f(\psi_i, t_{ij}) + \epsilon_{ij}$$

where $f(\psi, t)$ is given by (6.1), $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\psi_i = (\psi_{1i}, \psi_{2i}, \psi_{3i})$ is the non-linear regression coefficient for individual i, $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. The common structure relating the individuals is given through a hierarchical model where the ψ_i are assumed to be a sample from a $N(\mu, W)$ distribution. The model is completed with a second level where independent prior distributions are specified for the hyperparameters $\mu \sim N(b, B)$, $\sigma^2 \sim IG(n_0/2, n_0 S_0/2)$ and $W \sim IW(n_W/2, n_W S_W/2)$.

Bennett, Racine-Poon and Wakefield (1996) showed that the full conditional distributions of the hyperparameters μ, σ^2 and W are conjugate and easy to sample from but the same is not true for the regression coefficients $\psi_1, \ldots, \psi_m$. For these parameters, they compared sampling directly from the full conditionals with the rejection and ratio-of-uniform methods and indirectly with independence proposals based on a normal approximation to the likelihood and with normal random walk proposals as described in Section 6.2. They found that Gibbs sampling with the rejection method achieves convergence faster but at the expense of many rejections per iteration and conclude that the Metropolis-Hastings schemes are easier to implement and more efficient in terms of computing time. Their findings seem to provided an empirical echo to the intuitive point that whenever the model produces awkward full conditional distributions, one should avoid Gibbs sampling in favor of other MCMC schemes (see also Example 6.5).

The above description of the algorithm concentrated on the scheme where components are updated one by one in the order they are given in the parameter vector θ. The mixture and cycle schemes show that it is also possible to have many other orderings of the components. This also includes schemes where some of the components are systematically updated more often than other components. This may be because the less frequently updated components are more difficult to generate or because they are able to move more freely across the parameter space.

Another possibility worth exploring is to consider mixtures (or cycles) of different transition kernels. These typically apply to different components of θ but nothing in the theory prevents the use of conceptually different sampling schemes to update the same component or group of components in a single Markov chain. This does not generally guarantee faster convergence but Gelfand and Carlin (1995) provide a nice example where considerable improvement in convergence is achieved by mixing different sampling schemes in a single chain.

The different forms of proposal kernels described in the previous section are all relevant here. They may be used to construct each of the proposals q_i considered here. The same can be said about the adequacy of different forms for the situation under study. For some of the components of θ it may be more natural to use independence chains based on the prior distribution if there is enough information available for that component. For other components, the likelihood may suggest more appropriate proposals

(Section 6.5). Failing that, remaining components may be updated by a simple random walk proposal.

Example 6.4 *(Tanner, 1996, p. 67) Times to failure (f) of motorettes were tested at different temperatures (t). The data is presented in Table 6.2. A simple linear regression is fit*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $x_i = 1000/(t_i{+}273.2)$ *and* $y_i = log_{10}(f_i)$. *Without loss of generality, assume that the first* $m = 17$ *observations are uncensored. For simplicity, let* $\sigma^2 = 0.2592$ *and* $p(\beta) \propto 1$, *where* $\beta = (\beta_0, \beta_1)$. *The posterior distribution of* β *is*

$$\pi(\beta) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\frac{\varepsilon_i^2(\beta)}{\sigma^2}\right\}\prod_{i=m+1}^{n}\left[1-\Phi\left(\frac{\varepsilon_i(\beta)}{\sigma}\right)\right]$$

where $\varepsilon_i(\beta) = y_i - \beta_0 - \beta_1 x_i$. *The second term on the righthand side accounts for the censored portion of the data. Analytical posterior inference is unavailable, so the following three variations of the Metropolis-Hastings algorithm are implemented:*

1. *Random walk Metropolis with single move: given the current state of the chain* $\beta^{(j)} = (\beta_0^{(j)}, \beta_1^{(j)})$, β_0 *is sampled from* $N(\beta_0^{(j)}, \tau^2)$ *and then* β_1 *is sampled from* $N(\beta_1^{(j)}, \tau^2)$, *for* $\tau = 0.1$.
2. *Random walk Metropolis with block move:* β *is sampled from* $N(\beta^{(j)}, \tau^2 I_2)$, *for* $\tau = 0.1$.
3. *Independence Metropolis:* β *is sampled from* $N(\hat\beta, \sigma^2(X'X)^{-1})$, *with* $\hat\beta = (X'X)^{-1}X'y$ *where* $y = (y_1, \ldots, y_m)'$ *and* X *is the design matrix with rows* $(1, x_i)$, *for* $i = 1, \ldots, m$. *The correlation between* β_0 *and* β_1 *in the proposal is* -0.999.

Figure 6.7 exhibits the behavior of three chains for each one of the above three schemes for three distinct initial values $(-8, 5.5)$, $(-8, 4)$ *and* $(-5, 3)$. *It can be seen that the independence Metropolis scheme works much better than both single and block random walk Metropolis schemes, the main apparent reason being the careful choice of the proposal density. Effective sample sizes based on either* β_0 *or* β_1 *are roughly around 100 for the random walk Metropolis algorithms (schemes 1 and 2). They are around 600 and 950 for the independence Metropolis algorithm (scheme 3). Posterior summaries are based on the 20000 draws from the independence Metropolis algorithm. Posterior mean, median and 95% credibility interval for* β_0 *are* -6.09, -6.08 *and* $(-7.86, -4.30)$. *Posterior mean, median and 95% credibility interval for* β_1 *are* 4.34, 4.34 *and* $(3.53, 5.15)$. *Posterior correlation between* β_0 *and* β_1 *is* -0.998. *Later, in Section 6.4.4, a reparametrization strategy is implemented to reduce the high correlation. Consequently, mixing of the chains improves considerably for all sampling schemes above.*

Temp.	*time to failures*				
150°	**8064**	**8064**	**8064**	**8064**	**8064**
	8064	**8064**	**8064**	**8064**	**8064**
170°	*1764*	*2772*	*3444*	*3542*	*3780*
	4860	*5196*	**5448**	**5448**	**5448**
190°	*408*	*408*	*1344*	*1344*	*1440*
	1680	**1680**	**1680**	**1680**	**1680**
220°	*408*	*408*	*504*	*504*	*504*
	528	**528**	**528**	**528**	**528**

Table 6.2 *y: time to failure (in hours) of motorettes for four different temperatures (in Celsius). Bold indicates right censoring.*

6.4.2 Metropolis within Gibbs

Transitions are based on the full conditional distributions of the components of θ in the basic componentwise scheme above. The more similar the proposal q_i and the density π, or equivalently the full conditional π_i, the closer to 1 will the acceptance probabilities (6.7) be. This does not necessarily ensure fast convergence but may reduce the computational burden.

In the basic Metropolis algorithm, it was typically not possible to find a sampling kernel q that could approximate π without error. Were that possible, direct sampling from π would be available. Here, the situation is different. It is possible that π has a complicated form that prevents direct sampling from it but (some of) its full conditional distributions π_i can be directly sampled from. In this case, a convenient choice of proposal is to take $q_i = \pi_i$. The proposed value for θ_i is drawn from its full conditional and accepted with probability 1. Hence, the computational burden of the calculation of (6.7) is avoided. If that can be done for all components of θ, their values will be all sampled from the corresponding full conditionals and accepted. This is the Gibbs sampler!

In general, the simplification obtained is very convenient but again no optimality results are available in that direction. Note that the Gibbs sampler only depends on the previous value of the other components, unlike the Metropolis-Hastings scheme that also depends on the previous value of the component being updated. Besag et al. (1995) pointed out that this extra ingredient may allow the construction of a more efficient sampling scheme. The example below shows an instance where it is clearly advantageous to sample from a proposal even when the full conditional is available for sampling.

Example 6.5 *Consider again the Poisson model with change point from Example 5.1. Instead of sampling* m' *from its full conditional* $\pi_m(m)$, *an alternative is to sample* m' *from a proposal kernel* $q(m, m')$ *and accept it*

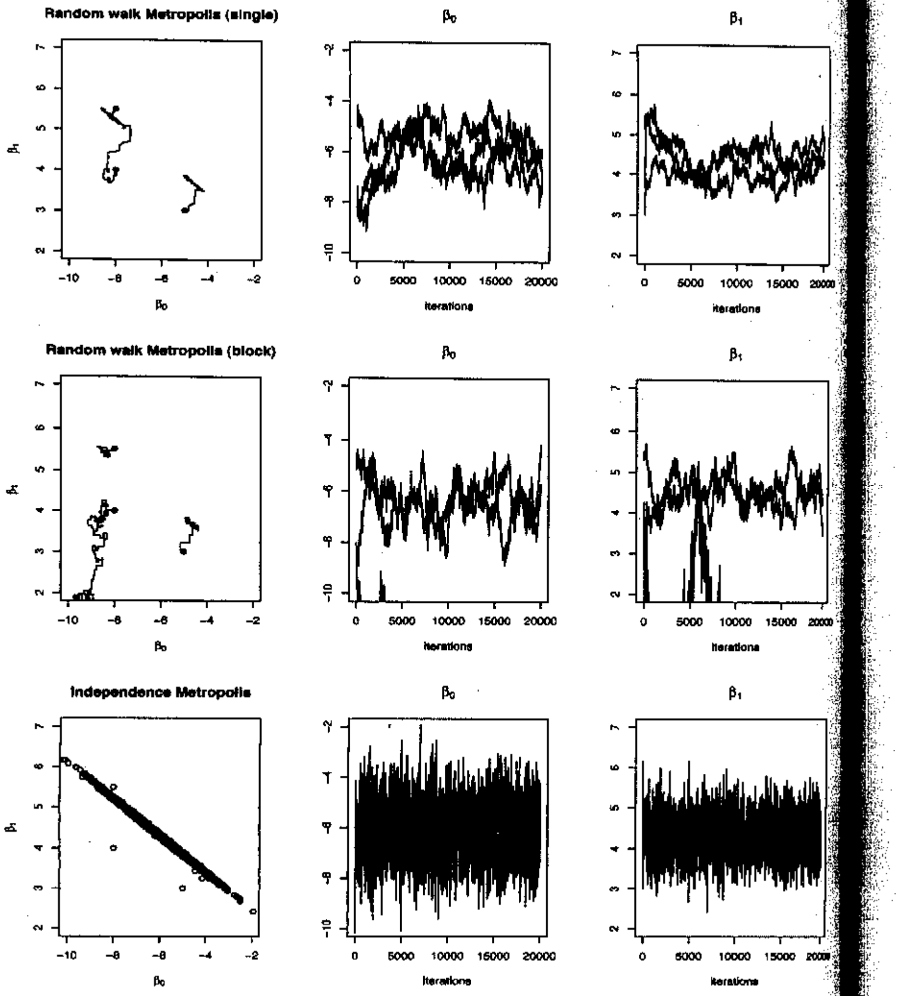

Figure 6.7 *Chain paths: Top frames: random walk Metropolis with single moves. Middle frames: random walk Metropolis with block moves. Bottom frames: independence Metropolis.*

with probability

$$\alpha(m, m') = min\left\{1, \frac{f(y|\lambda; \phi, m')}{f(y|\lambda, \phi, m)} \frac{q(m', m)}{q(m, m')}\right\} . \qquad (6.9)$$

The main advantage of performing a Metropolis-Hastings step as opposed to a Gibbs step is avoiding the evaluation of (5.2) for all possible values of m. This advantage is particularly relevant when the number of evaluations of the full conditional is large and/or when their evaluations are cumbersome.

Assume that m' is drawn from the proposal $q(m, m') \propto exp\{-\tau|m' - m|\}$, where τ is a tuning parameter. The approximate posterior distribution obtained with 5000 draws from the algorithm is very similar to the approximation obtained in Example 5.1 with the Gibbs sampler. Figure 6.8 exhibits average acceptance rates and effective sample sizes for several values of τ. For instance, when $\tau = 0.15$, the average acceptance rate for this chain is approximately 33%, while effective sample sizes based on λ, ϕ and m are 3455, 3330 *and* 900, *respectively. For comparison with the Gibbs sampler, this algorithm was run for a total of 40000 iterations. This task takes only around 40% of the time used to run the 5000 iterations of the Gibbs sampler. Effective sample sizes based on λ, ϕ and m are* 25937, 24837 *and* 5311, *respectively. They are all larger than the effective sample sizes obtained with the 5000 iterations used by the Gibbs sampler. This shows that even though the Metropolis algorithm may generate chains with higher autocorrelation structures, its computational simplicity compensates and produces more efficient approximations in the same amount of time.*

Going back to direct sampling from full conditionals, in complex models it may be possible to establish conditional conjugacy for some but not all components of the parameter. In this case, only some components will be available for direct sampling. For the components that cannot be directly sampled, Muller (1991b) suggested that they are sampled from π_i by a Metropolis-Hasting (sub-)chain inside a Gibbs sampling cycle. So, these components are sampled from a proposal q_i with possible acceptance governed by probabilities (6.7). This process would last for T iterations until convergence and consequent generation from π_i, the limiting distribution of the sub-chain. This sampling scheme is widely known as Metropolis-within-Gibbs. Muller (1991b) used the value $T = 5$ in applications but in fact the construction of the sub-chain is unnecessary as a single iteration is sufficient. Note that the case $T = 1$ reproduces the Metropolis-Hastings cycle scheme that visits each component once and was described in the previous subsection. Nowadays, the version $T = 1$ is almost always used so that the Metropolis-within-Gibbs nomenclature is somewhat misleading. A more appropriate name to describe it could be Gibbs-within-Metropolis as some Gibbs updating schemes are used inside a componentwise Metropolis-Hasting scheme.

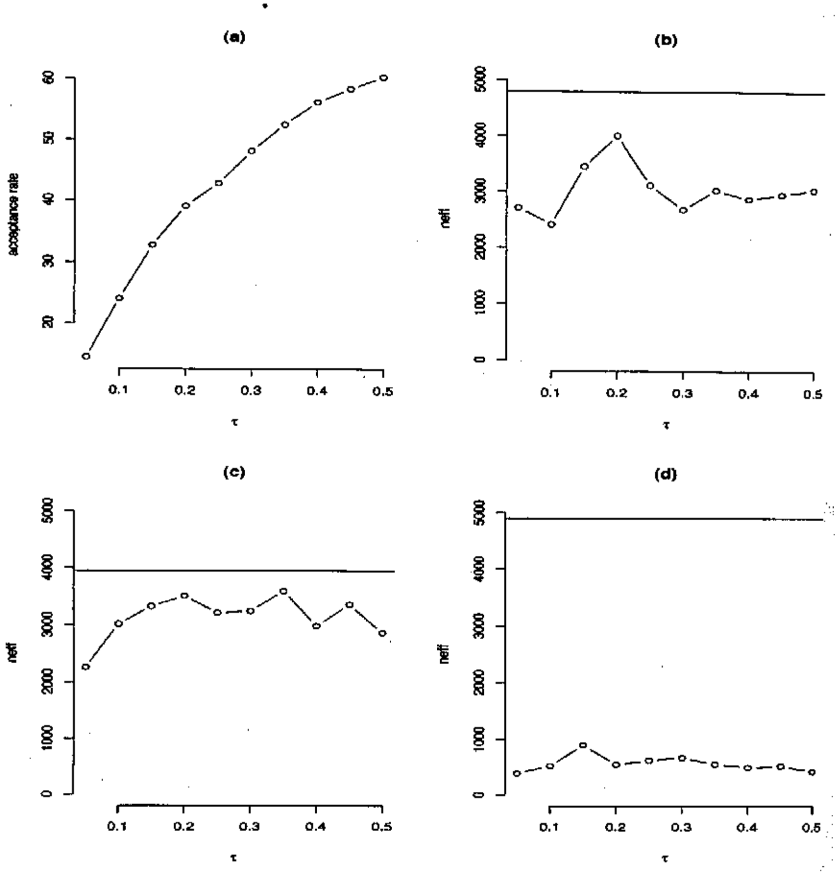

Figure 6.8 *Poisson observations with change point. Results from the Metropolis chains based on 5000 iterations for several values of tuning parameter τ: (a) average acceptance rates, (b),(c) and (d) are effective sample sizes based on λ, ϕ and m, respectively. The horizontal lines in (b), (c) and (d) represent the effective sample sizes based on 5000 iterations of the Gibbs sampler of Example 5.1.*

6.4.3 Blocking

So far, nothing has been said about the choice of blocks $\theta_1, \ldots, \theta_d$. The results obtained for the Gibbs sampler do not reproduce with the same ease here. In particular, it is not true that the larger the block the faster is the convergence. In fact, it is likely that blocking many parameters in a single group in highly multidimensional problems is more detrimental than beneficial.

Example 6.3 *(continued) The parameters of the model are given by $\psi_1, \ldots, \psi_m$, μ, σ^2 and W. Note that each of these parameters is itself a collection of components with the exception of the scalar σ^2. The regression coefficients ψ_i and their mean μ are vector parameters and W is a matrix of parameters. Bennett, Racine-Poon and Wakefield (1996) used this blocking in all their sampling schemes except for the Gibbs sampler with the ratio-of-uniform method where they sample each scalar component of the ψ_i separately.*

Consider the most common case of a random walk chain with proposal q_i for the component θ_i formed by a large number of components. As the number of components gets large, it becomes more likely to have components of θ_i falling well in the tails of π_i. This will produce a proposed value with very low density π_i and hence the test ratio will also be small. As a result, very few proposed values will be accepted and convergence will be very slow. Of course, decreasing the appropriate entries of the variance matrix of the proposal will inhibit such extreme components of θ_i being proposed but that requires tuning of each component of the variance matrix to avoid it. This task is far from easy and can become a time-consuming exercise.

Similar reasoning can be applied to other forms of chains. There is no reason to believe that the proposal q_i will remain close to π_i as the dimension of θ_i increases. This leads to the same end result that increasing the dimension of a block increases the chances of rejection.

Again, there are no theoretical results here but the rule followed in applied work is to form small groups of correlated parameters that belong to the same context in the formulation of the model. A typical example is given in Example 6.3 above. The structure of the model made natural the choice of the blocks $\psi_1, \ldots, \psi_m, \mu, \sigma^2$ and W. With this structure, inference via Gibbs sampling or Metropolis-Hastings sampling may successfully proceed. Gathering all regression coefficients in a single block $\psi = (\psi_1, \ldots, \psi_n)'$ leads to the difficulties described above. Small blocks of ψs may also be formed and would alleviate these difficulties. In the context of this application, it is better to work with each vector of regression coefficient separately because they are all independent conditionally on μ and W. This point will be returned to in more detail below and in the next section.

The important point to make in practice is to block parameters whenever it is possible and needed. It is possible to block whenever the acceptance rate does not fall to very small values, namely single digit percentages. Gelman, Roberts and Gilks (1996) provided some theoretical considerations for aiming at an optimal acceptance rate of around 24% for random walk chains. This rate should be taken as an indication and not as a rule. Efficient chains with higher acceptance rates are also possible.

Blocking of parameters is only needed to break correlations. Parameters that are conditionally independent given other parameters need not be blocked. Remember from Chapter 5 that blocking was used to define more appropriate directions of moves for the chain instead of the unrelated moves along the axes. When components are (conditionally) independent, however, moves will already be made along the individual directions and no improvement is made by changing these directions.

6.4.4 Reparametrization

The same comments made about choice of parametrization in Chapter 5 for the Gibbs sampler are valid here. Good parametrizations are still useful in improving mixing of the chain and accelerating convergence. The apparent freedom of choice of proposals is not unrestricted. Very large displacements are very likely to be rejected as discussed above. Only moves that are compatible with the structure of the model and the form of the posterior distribution will be accepted. So, models with highly correlated parameters will only allow very small moves.

This is the pattern already observed for the Gibbs sampler. For the Metropolis-Hastings algorithm, larger moves may even be proposed but they will very likely fall in the tails of the posterior and will force a small value for the test ratio. The only possible acceptance of these large moves is for the case when they are directed according to the correlation structure of the posterior. In high-dimensional problems, the identification of these directions may be very difficult or time consuming. So, the large moves required for fast convergence will lead to rejected points. If an efficient parametrization can be found, then these large moves will no longer be rejected and fast convergence may be achieved. The application to dynamic models below illustrates this point. Another successful use of reparametrization is the case of centering in generalized linear mixed models as discussed by Gelfand, Sahu and Carlin (1996). They showed that centering in generalized models reduces correlation and improves convergence just as in normal linear mixed models (Section 5.4.3).

Example 6.4 *(continued) Assume the following reparametrization of the simple linear regression, where the x_i is centered around its sample mean*

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$$

for $z_i = x_i - 2.2$. This means that β_0 now corresponds to $\beta_0 - 2.2\beta_1$ in the original model formulation while β_1 remains unchanged (see Example 5.3). Posterior inference is based on samples obtained by performing the same three variations of the Metropolis-Hastings algorithm. Figure 6.9 exhibits the behavior of the chains for three distinct initial values $(3.3, 3)$, $(3.7, 3)$ and $(3.7, 5.5)$ over 20000 draws. Now all algorithms perform reasonably well. Effective sample sizes based on β_0 and β_1, when combining all 60000 draws, appear in Table 6.3. Posterior inference are virtually the same across schemes. Based on the random walk Metropolis scheme, for instance, posterior mean, median and 95% credibility interval for β_0 are 3.47, 3.47 *and* (3.38, 3.57). *Posterior mean, median and 95% credibility interval for β_1 are* 4.33, 4.32 *and* (3.50, 5.19). *Posterior correlation between β_0 and β_1 is* 0.222.

Algorithm	β_0	β_1
RWMS	*7524*	*878*
RWMB	*6311*	*597*
INDM	*2025*	*2476*

Table 6.3 *Effective sample sizes.*

6.5 Applications

This section gives some details of applications of the Metropolis-Hastings methodology for models commonly used in practice. With only a few exceptions, as soon as the model gets away from linearity and normality it loses conditional conjugacy. Chapter 5 showed that conjugacy is a very useful property in conjunction with the Gibbs sampler. It becomes difficult to sample from some of the full conditional distributions without them.

One possibility is to employ some of the resampling techniques, such as the rejection method. A very attractive alternative in terms of simplicity is the use of a Metropolis-Hastings sampling scheme for these awkward sampling distributions. Example 6.3 illustrates this point in the context of non-linearity. The applications below detail uses of the algorithm in the context of non-normal models.

6.5.1 Generalized linear mixed models

Models with random effects have been described in the previous chapter. When they also contain fixed effects, they are called mixed models. A general but not unique form for generalized linear mixed models is given by

$$\begin{aligned} f(y_i|\theta_i) &= a(y_i)\exp\{y_i\theta_i + b(\theta_i)\} \text{ with} \\ E(y_i|\theta_i) &= \mu_i \\ g(\mu_i) &= x_i'\beta + z_i'\gamma_i, \quad i = 1, \ldots, n \end{aligned} \tag{6.10}$$

where the link function g is differentiable, β is the d-dimensional vector of regression coefficients and γ_i is the r-dimensional random effect associated with observation y_i, $i = 1, \ldots, n$. Associated with these parameters, there

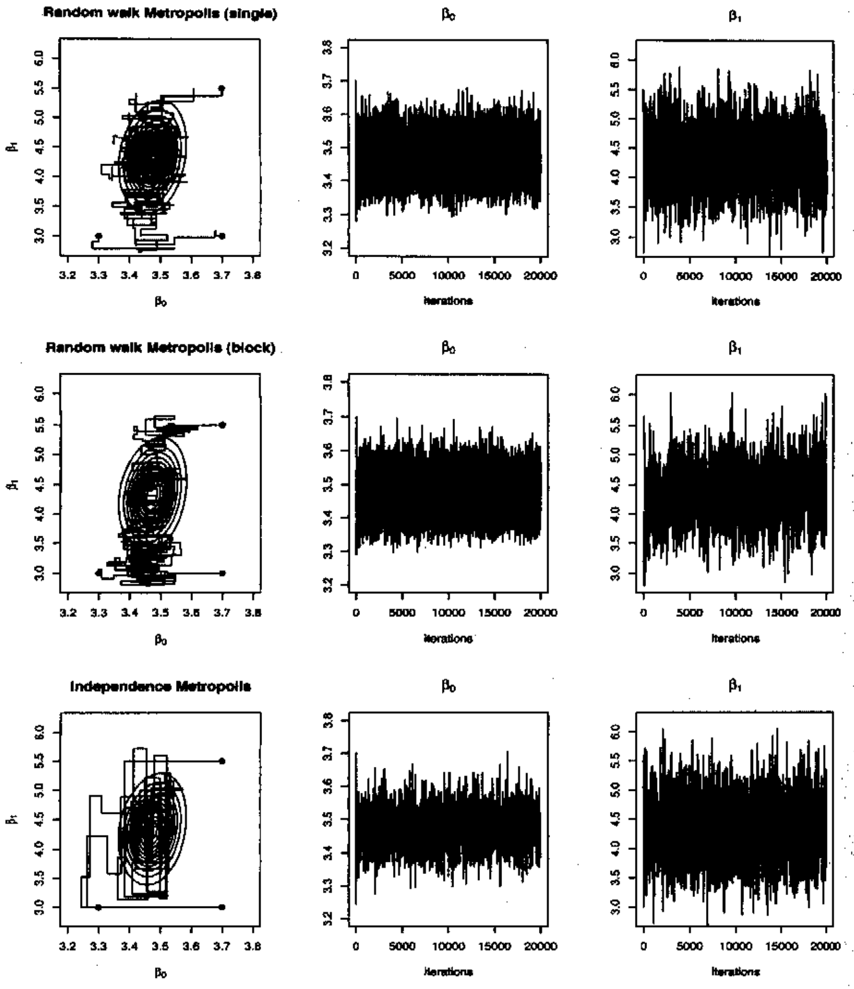

Figure 6.9 *Chain paths based on 20000 draws. Top frames: random walk Metropolis with single moves (RWMS). Middle frames: random walk Metropolis with block moves (RWMB). Bottom frames: independence Metropolis (INDM). Initial values:* (3.3, 3), (3.7, 3) *and* (3.7, 5.5).

are d-dimensional and r-dimensional vectors of covariates x_i and z_i that explain the fixed and random variations in the levels of the observations y_i, $i = 1, \dots, n$. The model is completed with independent prior distributions $\beta \sim N(a, R)$, $\gamma_i | W \sim N(0, W)$, $i = 1, \dots, n$, and $W \sim IW(n/2, nS/2)$. Note that if $x_i = z_i$, $i = 1, \dots, n$, the model becomes a special case of the generalized linear hierarchical model (2.20) with $X_1 = \text{diag}(x_1', \dots, x_n')$, $\beta_1 = (\beta_{11}', \dots, \beta_{1n}')'$, $\beta_{1i} = \beta + \gamma_i$, $i = 1, \dots, n$, $X_2 = I_n$, $C = W$ and $\beta_2 = \beta$.

The natural division of the parameters in blocks is given by $\beta, \gamma_1, \dots, \gamma_n$ and W. Alternative divisions include the block $(\gamma_1, \dots, \gamma_n)$ or even the block $(\beta, \gamma_1, \dots, \gamma_n)$. The former is unnecessary as the random effects are conditionally independent. The latter removes the problems that may be associated with correlation between random and fixed components. However, it is very likely to lead to very low acceptance rates due to its high dimension unless complicated search exercises for an adequate proposal are undertaken. Even then, the solution is likely to be specific to the problem being analyzed and will not be adequate for other data sets.

The only block with full conditional distribution in conjugate form is W. The other blocks have full conditional densities

$$\begin{aligned}
\pi_\beta(\beta) &\propto \exp\left\{\sum_{i=1}^n y_i\theta_i + b(\theta_i)\right\} f_N(\beta; a, R) \\
&\propto \exp\left\{-\frac{1}{2}(\beta - a)' R^{-1}(\beta - a) + \sum_{i=1}^n y_i\theta_i + b(\theta_i)\right\}. \quad (6.11)
\end{aligned}$$

and

$$\begin{aligned}
\pi_{\gamma_i}(\gamma_i) &\propto \exp\{y_i\theta_i + b(\theta_i)\} f_N(\gamma_i; 0, W) \\
&\propto \exp\left\{-\frac{1}{2}\gamma_i' W^{-1}\gamma_i + y_i\theta_i + b(\theta_i)\right\}, \quad i = 1, \dots, n. \quad (6.12)
\end{aligned}$$

None of these distributions is easy to sample from. Zeger and Karim (1991) use rejection sampling with normal envelopes for each of these full conditional distributions. It is very difficult in this situation to tune the constant A to provide a proper envelope over all the parameter space without sacrificing efficiency. Clayton (1996) explored the possibility of sampling each of these parameters componentwise. He used the adaptive rejection method as in most cases of interest the densities above are log-concave.

Other possibilities are likely to involve some use of Metropolis-Hastings methodology. They include sampling from the prior as a proposal in an independence chain, or even a random walk chain with variances given by prior variance or some measure of likelihood uncertainty as the inverse information matrix. As previously discussed, these forms require optimization of the tuning constant, which may be time consuming.

Gamerman (1997) used proposals based on the IRLS algorithm (Section 3.2.2). Consider the start of iteration j of the chain with previous values $\beta^{(j-1)}, \gamma_1^{(j-1)}, \dots, \gamma_n^{(j-1)}$ and $W^{(j-1)}$ and assume that all blocks are updated at every iteration in the order above.

The first step is the construction of the proposal q_β for the block β conditional on all the other parameters being fixed at current values. A vector of *adjusted* observations $\tilde{y} = \tilde{y}(\beta^{(j-1)})$ with corresponding matrix of *adjusted* variances $\tilde{V} = \tilde{V}(\beta^{(j-1)})$ is formed according to the IRLS algorithm and (6.11). An *adjusted* regression model is then formed with

$$\tilde{y}_i \sim N(x_i'\beta + z_i\gamma_i^{(j-1)}, \tilde{V}), i = 1, \ldots, n.$$

Combining with the prior $\beta \sim N(a, R)$ leads to an *adjusted* posterior distribution $\tilde{\pi}_\beta(\beta) = N(m^{(j)}, C^{(j)})$ where $m^{(j)} = C^{(j)}(R^{-1}a + X'\tilde{V}^{-1}\tilde{y}^*)$ and $C^{(j)} = (R^{-1} + X'\tilde{V}^{-1}X)^{-1}$. The vector of *readjusted* observations $\tilde{y}^*$ has components $\tilde{y}_i^* = \tilde{y}_i - z_i'\gamma_i^{(j-1)}$, $i = 1, \ldots, n$. This additional adjustment to the observations is caused by the known displacements $z_i\gamma_i^{(j-1)}$, $i = 1, \ldots, n$, as calculations are conditional on the values of the γ_i. In many cases, $\tilde{\pi}_\beta$ is a good approximation to π_β. Therefore, it is reasonable to take the proposal $q_\beta(\beta^{(j-1)}, \cdot) = \tilde{\pi}_\beta(\cdot)$. A proposed value β^* can be generated and $q_\beta(\beta^{(j-1)}, \beta^*)$ can be calculated. Note that $q_\beta(\beta^{(j-1)}, \cdot)$ depends on $\beta^{(j-1)}$ but in a very intricate way (through the adjusted observations) and this proposal does not fit into any of the categories described in Section 6.3. To calculate the test ratio, the values of $q_\beta(\beta^*, \beta^{(j-1)})$ and $\pi_\beta(\beta^*)/\pi_\beta(\beta^{(j-1)})$ are required. The first one is obtained by repeating the above procedure with β^* replacing $\beta^{(j-1)}$. The second is obtained from (6.11). Depending on the acceptance stage, $\beta^{(j)}$ is taken as β^* or $\beta^{(j-1)}$.

A similar approach is used to construct proposals q_{γ_i}, $i = 1, \ldots, n$. The *adjusted* observations $\tilde{y}_i = \tilde{y}_i(\gamma_i^{(j-1)})$ with *adjusted* variances $\tilde{V}_i = \tilde{V}_i(\gamma_i^{(j-1)})$ form the regression model $\tilde{y}_i \sim N(x_i'\beta^{(j)} + z_i\gamma_i, \tilde{V})$. Combining with the prior $\gamma_i \sim N(0, W)$ leads to the *adjusted* posterior $\tilde{\pi}_{\gamma_i}(\gamma_i) = N(m_i^{(j)}, C_i^{(j)})$ where $m_i^{(j)} = C_i^{(j)} z_i \tilde{V}_i^{-1}\tilde{y}_i^*$ and $C_i^{(j)} = (W^{-1} + z_i\tilde{V}_i^{-1}z_i)^{-1}$. Again, the *readjusted* observation is $\tilde{y}_i^* = \tilde{y}_i - x_i'\beta^{(j)}$. Taking as a proposal $q_{\gamma_i}(\gamma_i^{(j-1)}, \cdot) = \tilde{\pi}_{\gamma_i}(\cdot)$, a new value γ_i^* is proposed and $q_{\gamma_i}(\gamma_i^{(j-1)}, \gamma_i^*)$ can be calculated. Again, $q_{\gamma_i}(\gamma_i^*, \cdot)$ is obtained by repeating the above procedures with γ_i^* replacing $\gamma_i^{(j-1)}$. The value of $\pi_{\gamma_i}(\gamma_i^*)/\pi_{\gamma_i}(\gamma_i^{(j-1)})$ is obtained using (6.12). The test ratio can be calculated and depending on the acceptance stage, $\gamma_i^{(j)}$ is taken as γ_i^* or $\gamma_i^{(j-1)}$. The procedure is repeated for $i = 1, \ldots, n$.

Finally, the value of $W^{(j)}$ is drawn directly from the full conditional distribution of W or $\Phi = W^{-1}$ given by

$$\begin{aligned}
\pi_\Phi(\Phi) &\propto \prod_{i=1}^{n} f_N(\gamma_i; 0, \Phi^{-1})\ f_W(\Phi; n_W/2, n_W S_W/2) \\
&\propto |\Phi|^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \gamma_i'\Phi\gamma_i\right\} |\Phi|^{\frac{n_W-(r+1)}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(n_W S_W \Phi)\right\}
\end{aligned}$$

$$\propto |\Phi|^{[n+n_W-(r+1)]/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(n_W S_W + \sum_{i=1}^{n} \gamma_i\gamma_i'\right)\Phi\right]\right\}.$$

This is the expression of the $W[(n + n_W)/2, (n_W S_W + \sum_{i=1}^n \gamma_i\gamma_i')/2]$ distribution. Techniques for generation of a value of Φ from the Wishart distribution above were described in Chapter 1. A generated value of $W^{(j)}$ is obtained by inversion of the matrix Φ.

An important point is that this solution incorporates the structure of the problem into the construction of the proposal transition. This ensures that the chain moves will have direction and magnitude governed by the model. High acceptance rates are obtained as a result without compromising the amplitude of the chain moves and coverage of the relevant regions of the parameter space. The price paid in this case is the amount of additional calculation required. This may be unnecessary for models with a simpler structure but provides a general framework for analysis of any generalized linear mixed model.

Example 6.6 *Crowder (1978, Table 3) presented a data set with proportions of germinated seeds in $n = 21$ plates. The data set is influenced by the explanatory variables type of seed (s), root extract (r) and an interaction between these covariates. A larger variability than that explained by the binomial model was also noted by Crowder (1978). One possible model for this overdispersion is obtained with random effects. Breslow and Clayton (1993, Section 6.1) proposed to model the germination probabilities p_i associated with the ith plate through the logistic relation*

$$\mathit{logit}(p_i) = x_i'\beta + \gamma_i$$

where $x_i' = (1, s_i, r_i, s_i r_i)$ and $\gamma_i \sim N(0, W)$, $i = 1, \ldots, n$, are the univariate random effects modelling the overdispersion. The model is completed with a non-informative prior distribution $p(\beta, W) \propto 1/W$. Observe that W here is scalar and therefore its full conditional is an $IG(n/2, \sum_{i=1}^n \gamma_i^2/2)$ distribution.. Figure 6.10 presents the marginal histograms for the components of β for an analysis using the Metropolis-Hastings algorithm above. Table 6.4 presents numerical summaries of the posterior distribution along the equivalent ones from the penalized quasi-likelihood (PQL) analysis of Breslow and Clayton (1993). The estimates are very similar but the uncertainty in the Bayesian inference is always larger. This point had previously been noted in the context of generalized linear models by Dellaportas and Smith (1993).

Figure 6.11 presents a graphical summary of the inference for the random effects γ_i. Their posterior mean estimates behave as a sample from a normal distribution. So, they appear to confirm in the posterior the distributional form assumed for their prior. Some of the posterior correlations between the components of β were as large as 0.7 in absolute value but the correlation between β and W was low. This provides some support for the blocking

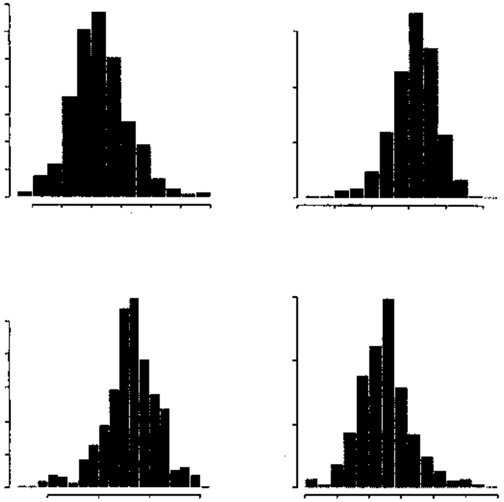

Figure 6.10 *Marginal histograms from the posterior distribution for the regression coefficients for the seeds data based on 600 successive draws from a single chain with a burn-in period of 600 iterations.*

	PQL	*MCMC*
Parameter	*Estimate (SE)*	*Estimate (SE)*
Intercept	*-0.542 (0.190)*	*-0.543 (0.197)*
Seed coef.	*0.077 (0.308)*	*0.074 (0.332)*
Extract coef.	*1.339 (0.270)*	*1.313 (0.274)*
Interaction coef.	*-0.825 (0.430)*	*-0.755 (0.431)*
$\sqrt{W}$	*0.313 (0.121)*	*0.278 (0.167)*

Table 6.4 *Estimation summary for Crowder's seeds data.*

scheme adopted, at least for this data set. The analysis presented was based on a single long chain. Similar results were however obtained by running multiple parallel chains.

This inference procedure can be easily extended to more elaborate forms of random effects. In many observation processes, data is obtained in groups. There may be random effects associated with groups (γ_i) and with individuals within groups (δ_{ij}). These may be described by the linear predictor

$$g(\mu_{ij}) = x'_{ij}\beta + z'_{ij}\gamma_i + t'_{ij}\delta_{ij} \ , j = 1, \ldots, n_i \, , \ i = 1, \ldots, m \ .$$

An application of the above sampling techniques to this model is also presented in Gamerman (1997) and illustrated for a real data set. A very common special case is the so called Laird and Ware (1982) model where there are only random effects associated with groups.

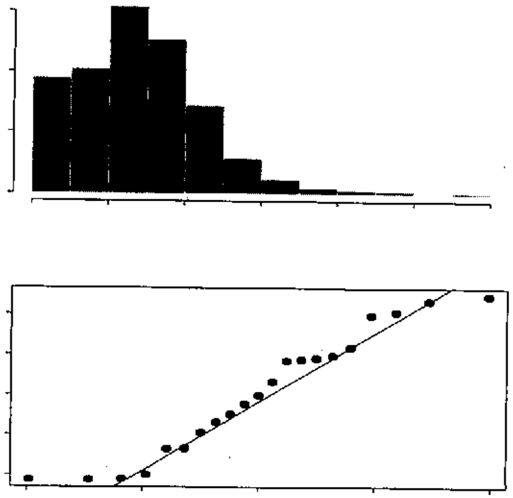

Figure 6.11 *Summary of inference for the random effects in the seeds data: (a) marginal histogram of the posterior distribution of $\sqrt{W}$; (b) QQ plot for normality of random effects. The posterior means for the random effects for each site were estimated from the sample and ordered for the construction of graph (b).*

6.5.2 Dynamic linear models

The dynamic linear model with Equations (2.21) and (2.22) describing the observation and state evolutions, respectively, was introduced in Section 2.5. It was noted in Section 5.5.2 that $\beta_1, \ldots, \beta_n, \sigma^2, W$ could be sampled jointly by first sampling the pair (σ^2, W) from $\pi(\sigma^2, W)$ and then sampling $\beta = (\beta_1, \ldots, \beta_n)$ from $\pi(\beta|\sigma^2, W)$.

From derivations of Section 5.5.2, the joint density of all model parameters can be obtained as $\pi(\beta, \sigma^2, W) \propto f_N(\beta; M, Q^{-1}) f_N(y; FA, FP^{-1}F + \sigma^2 I_n) p(\sigma^2, W)$, where $p(\sigma^2, W)$ is the prior for (σ^2, W) and both M and

Q depend on (σ^2, W). Therefore, this density is completely known up to a proportionality constant. The distribution of $(\beta|\sigma^2, W, y^n)$ was derived in Section 5.5.2 as $N(M, Q^{-1})$. Therefore, the marginal density $\pi(\sigma^2, W) \propto f_N(y; FA, FP^{-1}F + \sigma^2 I_n)p(\sigma^2, W)$ is also known up to a proportionality constant. It does not have a known form and direct sampling becomes difficult, specially if the dimension of the state parameter, and consequently of W, is large. Metropolis-Hastings algorithms can be used instead to propose values for sampling according to some transition kernel q. Gamerman and Moreira (2002) applied this sampling scheme, summarized in the algorithm below.

1. Sample (σ^{2*}, W^*) from $q(\sigma^2, W|\beta^{(j-1)}, \sigma^{2(j-1)}, W^{(j-1)})$.
2. Set $(\sigma^{2(j)}, W^{(j)}) = (\sigma^{2*}, W^*)$, with probability α and $(\sigma^{2(j-1)}, W^{(j-1)})$ with probability $1 - \alpha$, where

$$\alpha = \min\left\{1, \frac{\pi(\sigma^{2*}, W^*)}{\pi(\sigma^{2(j-1)}, W^{2(j-1)})} \frac{q(\sigma^{2(j-1)}, W^{(j-1)}|\beta^{(j-1)}, \sigma^{2*}, W^*)}{q(\sigma^{2*}, W^*|\beta^{(j-1)}, \sigma^{2(j-1)}, W^{(j-1)})}\right\}.$$

Jointly sampling (β, σ^2, W) avoids MCMC convergence problems associated with the posterior correlation between model parameters. The algorithm above uses $\pi(\sigma^2, W)$ only in ratio form. Therefore, the unknown proportionality constant of this density is not required. As shown before, it is important to specify suitably the proposal density q to avoid chains getting trapped or moving slowly. Gamerman and Moreira (2002) used product of log random walk forms where $q(\sigma^2, W|\beta^{(j-1)}, \sigma^{2(j-1)}, W^{(j-1)}) = q_1(\sigma^2|\beta^{(j-1)}, \sigma^{2(j-1)})q_2(W|\beta^{(j-1)}, W^{(j-1)})$, with q_1 given by an inverse Gamma density centered around (ie with means given by) $\sigma^{2(j-1)}$ and q_2 given by an inverse Wishart density centered around $W^{(j-1)}$ with shape parameters tuned for appropriate chain movements.

Example 6.7 *Consider the first order normal dynamic linear model*

$$\begin{aligned} y_t &= \beta_t + \epsilon_t, & \epsilon_t &\sim N(0, \sigma^2) \\ \beta_t &= \beta_{t-1} + \omega_t, & \omega_t &\sim N(0, W), \end{aligned}$$

which is the simplest case of the dynamic model. Regardless of its simplicity, this dynamic model retains the main features of a general dynamic model and is therefore a suitable representative of the general class for the comparison purpose. Fairly vague priors are used: $\beta_1 \sim N(0, 10)$ *and* σ^2 *and* W *have inverse Gamma distributions with means set at their true values and coefficients of variation set at 10. Four combinations of* (n, W) *were entertained,* $(100, 0.01), (100, 0.5), (1000, 0.01)$ *and* $(1000, 0.5)$, *with* $\sigma^2 = 1$ *and 100 simulations per combination (a total of 400 simulations). Signal-to-noise ratio,* $W/\sigma^2 = W$, *are* 0.01 *(small) and* 0.5 *(large). Four different sampling schemes were implemented for each one of the 400 simulations:* Scheme I: *Sampling* $\beta_1, \ldots, \beta_n, \sigma^2$ *and* W *individually from their full conditionals (Carlin, Polson and Stoffer, 1992, and Section 5.5.2),* Scheme II: *Sampling* β *jointly (FFBS algorithm) and* σ^2 *and* W *from their univariate full conditionals (Carter and Kohn, 1994, Frühwirth-Schnatter, 1994, and Section 5.5.2),* Scheme III: *Reparametrization using the fact that the system disturbances* w_t *reproduce in a unique form the state parameters (Gamerman, 1998, and Section 5.5.2), and* Scheme IV: *Sampling* (β, σ^2, W) *(Gamerman and Moreira, 2002, and the above algorithm). For each one of the 400 simulations computation was based on 20000 iterations. Table 6.5 presents the computing times relative to scheme I. When* $n = 100$, *schemes I and III are the fastest and slowest ones, respectively, with scheme IV becoming much slower when* $n = 1000$. *Effective sample sizes are presented in Table 6.6. Figure 6.12 presents effective sample sizes based on* σ^2. *They are smaller for longer time series and/or larger W. Overall, schemes II and III exhibit better performance (see Reis, Salazar and Gamerman, 2006, for more details of this comparison study).*

Scheme	*n=100*	*n=1000*
II	*1.7*	*1.9*
III	*3.5*	*3.8*
IV	*1.9*	*7.2*

Table 6.5 *Computing times relative to scheme I. For instance, when* $n = 100$ *it takes almost 4 times as much to run scheme III and almost 2 times as much to run scheme IV. Statistical summaries are based on a total of 20000 iterations. Scheme III was run with separate updates for* σ^2 *and* W *with tuning parameters set to 35, leading to 62% and 28% average acceptance rates according to whether* $n = 100$ *or* $n = 1000$.

		Scheme			
W	*n*	*I*	*II*	*III*	*IV*
0.01	*1000*	*242*	*8938*	*3028*	*2983*
0.01	*100*	*3283*	*13685*	*2501*	*12263*
0.50	*1000*	*409*	*3043*	*2391*	*963*
0.50	*100*	*1694*	*3404*	*1182*	*923*

Table 6.6 *Effective sample size* n_{eff} *based on* σ^2: *sample mean based on 100 replications.*

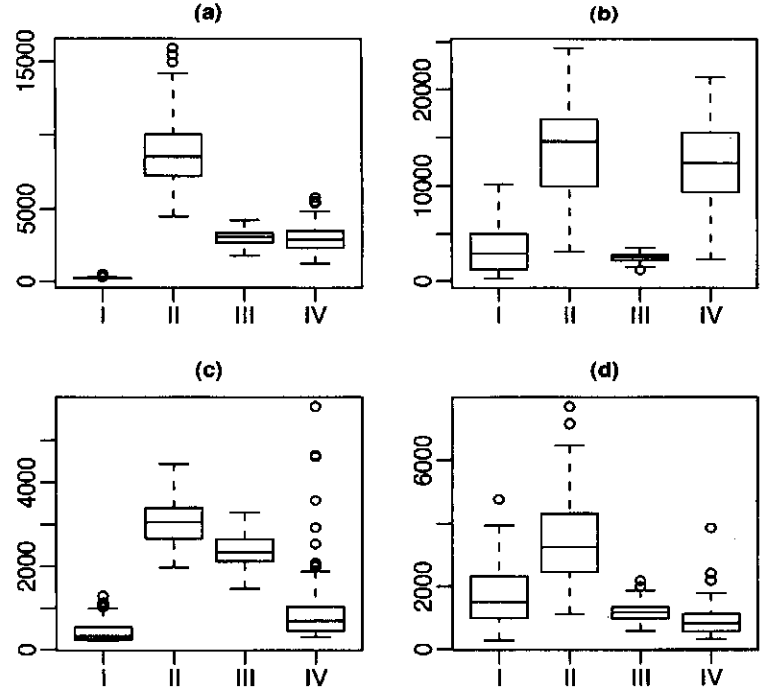

Figure 6.12 *Box plots of the effective sample sizes* n_{eff} *for the four schemes and the four combinations of* W *and* T. *(a)* $W = 0.01, n = 1000$, *(b)* $W = 0.01, n = 100$, *(c)* $W = 0.5, n = 1000$ *and (d)* $W = 0.5, n = 100$.

6.5.3 Dynamic generalized linear models

Dynamic generalized linear models were introduced in Chapter 2. As previously seen, it is not possible to perform exact inference because the relevant marginal distributions cannot be obtained analytically. Assuming that the variances of the system disturbances are constant, the model parameters are given by the state parameters $\beta = (\beta_1, \ldots, \beta_n)'$ and the system variance $W = \Phi^{-1}$. The model is specified with the observation and system equations and completed with the independent prior distributions $\beta_1 \sim N(a, R)$ and $\Phi \sim W(n_W/2, n_W S_W/2)$. The option to work with the precision matrix instead of the variance matrix is made again. The posterior distribution

is given by

$$\pi(\beta, \Phi) \propto \prod_{t=1}^{n} f(y_t|\beta_t) \prod_{i=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\beta_1) p(\Phi) .$$

The full conditional distributions are given by:

a) for block β

$$\begin{aligned}\pi_\beta(\beta) &\propto \prod_{t=1}^{n} f(y_t|\beta_t) \prod_{t=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\beta_1) \\ &\propto \exp\left\{\sum_{t=1}^{n}[y_t\theta_t + b(\theta_t)] - \frac{1}{2}\sum_{t=1}^{n}(\beta_t - G_t\beta_{t-1})'\Phi(\beta_t - G_t\beta_{t-1})\right\} .\end{aligned}$$

b) for block β_t, $t = 2, \ldots, n-1$

$$\begin{aligned}\pi_t(\beta_t) &\propto f(y_t|\beta_t)\, p(\beta_t|\beta_{t-1}, \Phi) p(\beta_{t+1}|\beta_t, \Phi) \\ &\propto \exp\{y_t\theta_t + b(\theta_t)\} \exp\left\{-\frac{1}{2}\left[(\beta_t - G_t\beta_{t-1})'\Phi(\beta_t - G_t\beta_{t-1})\right.\right. \\ &+ \left.\left.(\beta_{t+1} - G_{t+1}\beta_t)'\Phi(\beta_{t+1} - G_{t+1}\beta_t)\ \right]\right\} .\end{aligned}$$

Similar results follow for blocks β_1 and β_n.

c) for block Φ

$$\begin{aligned}\pi_\Phi(\Phi) &\propto \prod_{t=2}^{n} p(\beta_t|\beta_{t-1}, \Phi)\, p(\Phi) \\ &\propto \prod_{t=2}^{n} |\Phi|^{1/2} \exp\left\{-\frac{1}{2}\mathrm{tr}[(\beta_t - G_t\beta_{t-1})(\beta_t - G_t\beta_{t-1})'\Phi]\right\} \\ &\times |\Phi|^{[n_W-(p+1)]/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(n_W S_W \Phi)\right\} \\ &\propto |\Phi|^{[n^*_W-(d+1)]/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[(n^*_W S^*_W)\,\Phi\right]\right\} . \qquad (6.13)\end{aligned}$$

that is the density of the $W(n^*_W/2, n^*_W S^*_W/2)$ distribution with $n^*_W = n_W + n - 1$ and $n^*_W S^*_W = n_W S_W + \sum_{t=2}^{n}(\beta_t - G_t\beta_{t-1})(\beta_t - G_t\beta_{t-1})'$.

The results above show that the full conditional distributions of β and β_t do not belong to any known class of distributions but the full conditional of Φ is a known distribution from which samples can be drawn. As a result, the state parameters cannot be sampled directly. Again, rejection sampling can be used but the same problem of ensuring proper envelopes appears. The natural alternative seems to be the use of a Metropolis-Hastings scheme for this block. Therefore, transition kernels for the block β or for the blocks β_t, $t = 1, \ldots, n$, are required.

Again, many possibilities are available for the construction of the proposed kernels. Knorr-Held (1997) suggested the use of independence chains with prior proposals. He argued that the fast computing time at each iteration partially offsets the slow convergence due to the high correlation between state parameters. Shephard and Pitt (1997) used independence chains with proposals based on both prior and a normal approximation to the likelihood. They also blocked the state parameters β_t in random blocks to speed convergence. Ravinés (2005) used independence normal proposals for the block β with moments given by the approximation of West, Harrison and Migon (1985). Proposal kernels may again be constructed with the IRLS algorithm used for evaluation of the posterior mode.

Singh and Roberts (1982) and Fahrmeir and Wagenpfeil (1997) extended to the dynamic setting the method of mode evaluation described in Section 3.2.2 for static regression. They showed that iterating the posterior mode of β in the *adjusted* normal dynamic linear model given by (2.21) - (2.22) with *adjusted* observations $\tilde{y}_t$ and respective *adjusted* observational variances $\tilde{V}_t$ leads to the posterior mode of β. The expressions of $\tilde{y}_t$ and $\tilde{V}_t$ were given in Section 3.2.2.

The IRLS algorithm provides the *adjusted* full conditional distribution $\tilde{\pi}_\beta$ for block β given by (2.27). From this distribution, *adjusted* full conditional distributions for subsets of β may also be obtained. In particular, the *adjusted* distributions $\tilde{\pi}_t$ for blocks β_t, $t = 1, \dots, n$, are given by (2.26). These distributions may be used as proposal kernels. They define a Markovian process as $\tilde{y}$ and $\tilde{V}$ depend on the value of $\beta^{(j-1)}$. Note that these are all multivariate normal distributions and therefore it is simple to draw from them. The draws are proposed and they may either be accepted or rejected depending on the acceptance probabilities. The complete sampling scheme for β and Φ is given by:

1. Initialize the iteration counter of the chain $j = 1$ and set initial values $\beta^{(0)} = (\beta_1^{(0)}, \dots, \beta_n^{(n)})'$ and $W^{(0)}$.
2. Draw β^* from the density $\tilde{\pi}_\beta(\beta)$.
3. Calculate the acceptance probability $\alpha(\beta^{(j-1)}, \beta^*)$ of the move given by (6.7) with $q(\beta^{(j-1)}, \beta^*) = \tilde{\pi}_\beta(\beta^*)$. If the move is accepted, $\beta^{(j)} = \beta^*$. If the move is not accepted, $\beta^{(j)} = \beta^{(j-1)}$ and the chain does not move.
4. Draw Φ from its full conditional distribution (6.13).
5. Move counter from j to $j + 1$ and return to step 2 until convergence.

When updating with blocks $\beta_1, \dots, \beta_n$ separately, steps 2 and 3 above are replaced by:

2'a. Initialize the component counter $t = 1$.

2'b. Draw β_t^* from the density $\tilde{\pi}_t(\beta_t)$.

2'c. Calculate the acceptance probability $\alpha_t(\beta_t^{(j-1)}, \beta_t^*)$ of the move given by (6.7) with $q_t(\beta_t^{(j-1)}, \beta^*) = \tilde{\pi}_t(\beta_t^*)$. If the move is accepted, $\beta_t^{(j)} = \beta_t^*$. If the move is not accepted, $\beta_t^{(j)} = \beta_t^{(j-1)}$ and the chain does not move.

3'. Move the counter from t to $t + 1$ and return to step 2'b until $t = n$. When $t = n$, go to step 4.

For normal models, it was shown that the Gibbs sampler operated over the block β is superior to the one operated over the blocks β_t. Although it is reasonable to expect the same behavior here, there are important differences. The block β is highly dimensional for time series of large or moderate size. The high correlation between its components forces its complete conditional to be concentrated in a small region of the parameter space. It is very unlikely that proposed values are in this region and therefore, they are likely to fall well into the tails of this distribution. Consequently, their acceptance probability will be very low and the chain virtually does not move as a result.

That does not happen to blocks β_t of much smaller dimension. The proposal gives very good approximations for the full conditional distributions and high acceptance rates result. These high rates are not artificially obtained through small chain moves. They are governed by the structure of the model through the IRLS algorithm. The remaining problem is the high correlation between the components of β, making the chain move slowly towards equilibrium. Similar problems were found by Knorr-Held (1997).

An alternative previously discussed is the reparametrization in terms of the system disturbances w_t. The advantage again is that despite the high prior correlation between the β_t, the disturbances are independent a priori. Again, the components are dealt with separately but as most of the correlation is removed, the scheme is expected to converge at much faster rates. The drawback of the approach is the amount of extra calculations required by the reparametrization (see Table 6.5). Details about the method may be found in Gamerman (1998). Another possibility suggested by Shephard and Pitt (1997) is to form blocks containing small collections of β_t. The groups are divided at random which seems to improve convergence.

Example 6.8 *The data set given in Table 6.7 and Figure 6.13 concerns a study on advertising awareness (Migon and Harrison, 1985). Samples of $n_t = 66$ people were selected at random every week for an opinion poll and asked whether they remembered having seen a given advertising campaign on TV. A weekly cumulative measure of campaign expenditure was constructed and is also depicted in Figure 6.13. The model used for this problem was a dynamic logistic regression*

$$\begin{aligned}
y_t &\sim bin(n_t, \pi_t) \\
\mu_t &= n_t \pi_t \\
logit(\pi_t) &= \beta_{1t} + \beta_{2t} x_t = (1, x_t)\beta_t \\
\beta_t &= \beta_{t-1} + w_t \ , \ w_t \sim N(0, W)
\end{aligned}$$

The main features of this particular data set are a campaign change before

t	x_t	y_t	t	x_t	y_t	t	x_t	y_t
1	490	29	31	66	05	61	501	09
2	450	20	32	60	15	62	454	05
3	406	21	33	54	07	63	483	11
4	365	20	34	48	10	64	522	13
5	331	*	35	43	10	65	559	09
6	315	*	36	39	15	66	519	09
7	376	*	37	35	07	67	467	11
8	441	*	38	32	09	68	420	08
9	506	22	39	50	13	69	378	08
10	502	32	40	116	11	70	340	12
11	544	27	41	196	11	71	306	09
12	489	29	42	268	15	72	276	07
13	440	27	43	325	10	73	248	06
14	396	23	44	367	13	74	232	08
15	357	25	45	386	23	75	201	09
16	321	25	46	397	21	76	181	05
17	289	15	47	413	*	77	163	10
18	260	20	48	423	*	78	146	09
19	234	14	49	420	*	79	132	04
20	211	15	50	490	10	80	119	03
21	190	17	51	539	15	81	107	12
22	171	15	52	581	19	82	96	13
23	154	09	53	603	23	83	86	06
24	138	14	54	580	15	84	78	08
25	124	11	55	524	11	85	70	05
26	112	13	56	499	08	86	63	05
27	100	05	57	552	15	87	57	07
28	91	17	58	597	07	88	51	03
29	82	11	59	611	19	89	46	05
30	73	14	60	557	14	90	41	05

Table 6.7 *Data from Example 6.8. * = missing value.*

week 41 and a few missing points for weeks during which the poll was not made.

The average trajectory of the expenditure coefficient β_{2t} is plotted in Figure 6.14 for the sampling schemes based on the state parameters β_t and the system disturbances w_t. Convergence is clearly faster when sampling the disturbances, as expected. The drawback of higher computational cost is only serious for time series of very long size. The conceptual advantages seem stronger and should prevail in the choice of the method. Estimates for expenditure coefficients are plotted along with confidence limits in Figure 6.15. Two distinct levels are observed with a clear reduction after week 41, showing that the first campaign was more effective in terms of awareness. An increase in the uncertainty levels can be observed for the last weeks of the second campaign.

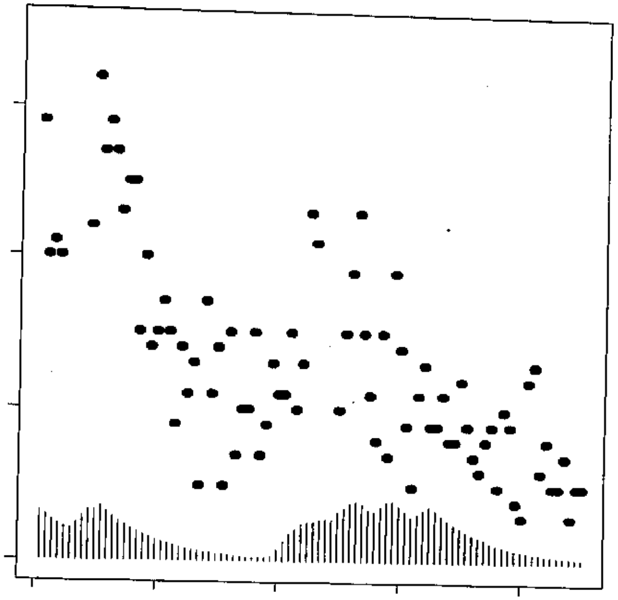

Figure 6.13 *Data on advertising awareness. The dots represent the weekly percentage of people that recalled having watched the advertising campaign of a given product on TV. The vertical bars represent a weekly cumulative measure of advertising expenditure.*

6.5.4 Spatial models

Spatial models were also introduced in Chapter 2 and some MCMC schemes based on Gibbs sampling reported in Section 5.5.3. The presentation here is concentrated on cases where direct sampling from full conditional distributions is not possible or computationally unfeasible.

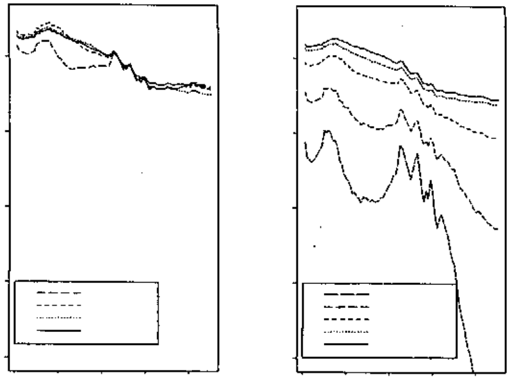

Figure 6.14 *Average trajectory of β_{2t} in 500 parallel chains with number of iterations for sampling from: (a) system disturbances; (b) state parameters.*

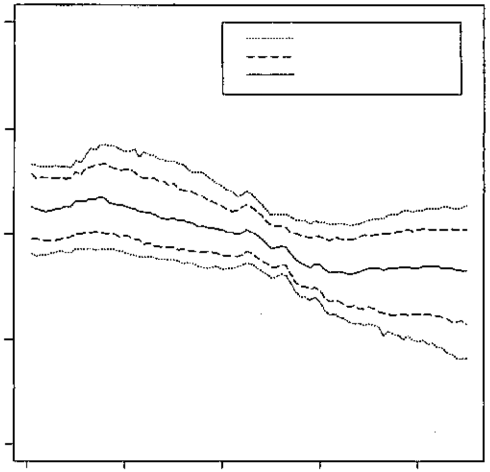

Figure 6.15 *Point estimate with confidence limits of the expenditure coefficient β_{2t}.*

Consider first the GMRF model (2.33) with parameters θ, ϕ and W. These parameters can be sampled separately in a Gibbs cycle. The full conditional distributions of θ_i, θ, ϕ and W were already derived in Section 5.5.3 and shown to be easily sampled from.

An alternative and possibly more efficient sampling scheme is provided by joint sampling all model parameters. Jointly sampling (θ, ϕ, Φ) was also mentioned in that section and is achieved by sampling (ϕ, Φ) from their marginal posterior distribution and then sampling $\theta|(\phi, \Phi)$ from its full conditional posterior distribution. The latter is easily and efficiently sampled from with the methods of Rue and Tjelmeland (2004) and was given in Exercise 2.23. The former is given by

$$\begin{aligned}\pi(\phi, \Phi) \quad &\propto \quad \phi^{(n_\sigma+d)/2}\Phi^{(n_W+d)/2}|C|^{1/2} \\ &\times \quad \exp\left\{-\frac{1}{2}[n_W S_w \Phi + (n_\sigma S_\sigma + y'y)\phi - y'Cy\,\phi^2]\right\} \qquad (6.14)\end{aligned}$$

where $C^{-1} = \phi I_d + \Phi K$.

The density (6.14) cannot easily be sampled directly and Metropolis-Hastings algorithms may be used here. Proposal densities should be specified. In addition to those described earlier in this chapter, other forms can be applied. Gamerman, Moreira and Rue (2003) used Gamma proposals centered around previous values and shape parameters tuned for appropriate acceptance rates. This is equivalent to a random walk proposal in the log scale with log-Gamma proposal densities.

When the observational model is non-Gaussian, then the full conditional of θ is no longer Gaussian and becomes difficult to sample from directly. Once again, Metropolis-Hastings algorithms provide a feasible alternative. The methods used in the previous section to generate normal proposals for state parameters and designed to mimic their posterior distribution could be applied. Alternatively, Rue and Tjelmeland (2004) suggested the use of quadratic approximations directly to the likelihood thus also generating normal proposals.

Analysis of spatially distributed data is an area where simulation methods using Markov chains have been heavily used. Besag, York and Mollié (1991) used the Gibbs sampler with the rejection method to sample from (2.36). Green (1991) suggested using instead proposals in the form $q_i(\theta_i, \cdot) = f_N(\cdot; (1-a)b + a\theta_i, (1-a^2)c)$, $i = 1, \ldots, d$. The constants a, b and c are chosen so as to make the test ratio as constant as possible thus increasing the chances of acceptance of the proposed value. This form of proposal falls into the category of autoregressive chains described in Section 6.2. Note that $a = 1$ gives the random walk with variance c. The guidance provided for the choice of this variance in the above section may be used here for arbitrary values of a.

Once again, the question of blocking arises and different blocking strategies can be applied in this setting. One possibility is to sample θ jointly. In this case, proposals transition kernels $q_\theta(\theta, \cdot; \Phi)$ for θ and $q_\Phi(\Phi, \cdot; \theta)$ for Φ must be specified. Note that both kernels must be conditional on the other parameter as was made explicit in their notation.

Knorr-Held and Rue (2002) discussed and compared blocking schemes. They favored sampling all parameters in a single block based on empirical evidence about the mixing properties of the chains. This means in the setting presented here that new values Φ^* are generated from a proposal q_Φ, then a new value θ^* is proposed from its normal proposal q_θ, conditional on Φ^*. In other words, the (joint) transition kernel is $q((\theta,\Phi),(\theta^*,\Phi^*)) = q_\Phi(\Phi,\Phi^*)q_\theta(\theta,\theta^*;\Phi^*)$. Then, the value of (θ^*,Φ^*) is jointly tested with acceptance probability

$$\min\left\{1, \frac{\pi(\theta^*,\Phi^*)}{\pi(\theta^{(j-1)},\Phi^{(j-1)})} \frac{q_\Phi(\Phi^*,\Phi^{(j-1)})q_\theta(\theta^*,\theta^{(j-1)};\Phi^{(j-1)})}{q_\Phi(\Phi^{(j-1)},\Phi^*)q_\theta(\theta^{(j-1)},\theta^*;\Phi^*)}\right\}.$$

Note that when jointly sampling (θ, Φ), the (marginal) transition kernel for Φ does not depend on θ unlike the (conditional) transition kernel of θ, that depends on Φ.

Similar comments apply to the distance-based GRF models. For normal models, almost all full conditional distributions are easily sampled from and detailed in Section 5.5.3. The only exception is the full conditional of the correlation parameters λ given by (5.9). This distribution has no simple form and typically Metropolis-Hastings proposals are used, specially when λ is a vector of hyperparameters.

Full conditional distributions for θ in non-normal observation models are no longer amenable for easy sampling. Many of the proposals described in this chapter can be applied. Diggle, Tawn and Moyeed (1998) considered these models and proposed MCMC schemes based on single site moves. Ideas of Knorr-Held and Rue (2002) can be adapted for this setting and used to produce sampling schemes with different blocking pattern.

6.6 Exercises

6.1 *Prove that the transition kernel p of the Metropolis-Hastings algorithms satisfies the detailed balance Equation (6.3) and hence has stationary distribution π.*

6.2 *The algorithm proposed by Barker (1965) set*

$$\alpha(\theta,\phi) = \frac{\pi(\phi)}{\pi(\theta)+\pi(\phi)} .$$

Show that this algorithm produces a reversible chain and has stationary distribution π, if q is symmetric.

6.3 *Certify yourself that large moves tend to be rejected and small moves are very slow to converge by considering sampling from a N(0,1) distribution. Consider normal proposal transitions with variances ranging from 0.01 to 100. Compare also the independence chains with an arbitrary, fixed mean with the random walk chain with proposal means given by the previous value of the chain.*

6.4 *Certify yourself by graphical and/or analytical terms that if $q(\theta,\cdot)$ and $\pi(\cdot)$ are continuous, the acceptance probability will be close to 1 when moves proposed by q make the chain move slowly.*

6.5 *Consider random walk chains. Show that if the distribution f_w of the disturbances w_j is symmetric around 0, the chain is symmetric. Show also that the Metropolis algorithm described in Section 6.1 is an example of a random walk chain and specify its distribution f_w.*

6.6 *Consider sampling π by the rejection method with an envelope density q for which complete blanketing is not assured and put it into the context of Metropolis-Hastings.*

(a) *Defining the blanketing region $C = \{\theta | \pi(\theta) < Aq(\theta)\}$, obtain that*

$$\alpha(\theta,\phi) = \begin{cases} 1 & ,\theta \in C \\ Aq(\theta)/\pi(\theta) & ,\theta \notin C\,,\,\phi \in C \\ \min\{1, w(\phi)/w(\theta)\} & ,\theta \notin C\,,\,\phi \notin C \end{cases}$$

where $w = \pi/q$.

(b) *Discuss the computational advantages of the above scheme over independence chains with proposal q and over rejection sampling.*

6.7 *Consider r Markov chains with transition kernels p_i, $i = 1,\ldots,r$, with a common stationary distribution π. Show that the resulting kernel of the mixture of these transitions will also have stationary distribution π. Repeat the exercise for a cycle chain.*

6.8 *Consider a parameter vector $\theta = (\theta_1,\ldots,\theta_d)'$ with posterior density π and the componentwise Metropolis-Hastings algorithm with proposed kernels $q_i(\theta_i,\phi_i)$ and acceptance probabilities α_i given by (6.7), $i = 1,\ldots,d$.*

(a) *Show that the component transition kernels formed define Markov chains with stationary distributions given by the full conditional distributions of θ_i, $i = 1,\ldots,d$.*

(b) *Extend the results of Section 6.4 to prove that π is a stationary distribution of the cycle kernel defined by the componentwise moves through all components of θ.*

6.9 *Consider the following univariate version of Example 6.2*

$$\pi(\theta) = 0.9 f_N(\theta;0,1) + 0.1 f_N(\theta;3.5,0.5)$$

where again the goal is to sample from $\pi(\theta)$ using either the random walk Metropolis or the independence Metropolis algorithm. Assume also that the *initial value is either $\theta^{(0)} = -5$ or $\theta^{(0)} = -7$, the random walk Metropolis proposal is a normal distribution with standard deviation $\tau = 0.1$ or $\tau = 0.5$ and the independent Metropolis proposal is a zero mean normal distribution with standard deviation $\tau = 1$ or $\tau = 3$. Run the algorithms for at least 10000 iterations and compute the respective effective sample sizes (Equation 4.10 from Chapter 4).*

6.10 *Consider the non-linear hierarchical model described in Example 6.3. Obtain the expressions of the full conditional densities for the hyperparameters μ, σ^2 and W and obtain the expressions of the proposed densities for the regression coefficients $\psi_1, \ldots, \psi_n$ that were discussed in the text.*

6.11 *Describe in detail the sampling scheme for the dynamic generalized linear model for blocks $w_1, \ldots, w_n$ and Φ based on the IRLS algorithm. In particular, obtain the expressions of the proposal transition kernels and of the acceptance probabilities.*

6.12 *Consider the dynamic linear model of Section 6.5.2. Show that*

$$\pi(\beta, \sigma^2, W) \propto f_N(\beta; M, Q^{-1}) f_N(y; FA, FP^{-1}F + \sigma^2 I_n) p(\sigma^2, W)$$

and

$$\pi(\sigma^2, W) \propto f_N(y; FA, FP^{-1}F + \sigma^2 I_n) p(\sigma^2, W).$$

6.13 *Show that the posterior density of all model parameters in the linear dynamic model can be written as*

$$\pi(\beta, \sigma^2, W) \propto f_N(\beta; M, Q^{-1}) f_N(y; FA, FP^{-1}F + \sigma^2 I_n) p(\sigma^2, W),$$

where $p(\sigma^2, W)$ is the prior for (σ^2, W).

6.14 *Describe in detail the sampling scheme for the dynamic generalized linear model for blocks $w_1, \ldots, w_n$ and Φ based on the IRLS algorithm. In particular, obtain the expressions of the proposal transition kernels and of the acceptance probabilities.*

CHAPTER 7

Further topics in MCMC

7.1 Introduction

The material presented in the previous chapters covers most of the relevant work on inferential procedures for a given model through Markov chain simulation techniques. Chapter 5 presented the Gibbs sampling technique and Chapter 6 presented the Metropolis-Hastings algorithm. It was assumed there that the adopted model was the true one or at least the most appropriate one throughout the presentation. Therefore, generation of a sample from *the* posterior distribution was all that was required. The techniques presented showed different ways of doing so.

In this final chapter, some points that lie beyond that basic framework will be discussed. Initially, the model will be put under scrutiny. Some techniques for evaluation of the model will be discussed. This evaluation may be divided into two complementary activities. The adequacy of a given model in the light of the observed data is made in Section 7.2. A more encompassing treatment is presented in Section 7.3 where different models are considered simultaneously. Depending on the cardinality and complexity of the set of models considered, alternative methods still based on Markov chains must be considered. Alterations in the structure of a given chain in order to speed up convergence are discussed in Section 7.4. There are many ways of performing these changes, from alterations in the transition kernel to alterations in the target distribution. Other forms of change involve alteration in the generated sample. This chapter presents more advanced ideas, generalizing the material of the previous chapters.

7.2 Model adequacy

Recall from Section 2.7 that a basic ingredient for model assessment is given by the predictive density

$$f(y|M) = \int f(y|\theta, M) p(\theta|M) d\theta, \qquad (7.1)$$

which is the normalizing constant of the posterior distribution (2.3). This predictive density can now be viewed as the likelihood of model M. It is sometimes referred to as predictive likelihood, because it is obtained after marginalization of model parameters.