*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 $\mu$, $\sigma^2$ 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 $\Phi$ 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 $(\sigma^2, W)$.*

**6.14** *Describe in detail the sampling scheme for the dynamic generalized linear model for blocks $w_1, \ldots, w_n$ and $\Phi$ 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.

An important aspect of model evaluation is the calculation of the predictive likelihood. Usually, its expression cannot be analytically obtained due to the complexity of the integrand in (7.1) and approximate methods must be used. Equation (3.4) provided an analytical approximation supported by asymptotic normal theory. In what follows, methods for approximate evaluation of (7.1) using simulation techniques will be presented. Different uses of these approximations for model evaluation will then be shown.

Although evaluation of a model presupposes the existence (or possibility) of other models, the calculations of this section will operate on a single model $M$ at any one time. Therefore, the presence of the model in the conditioning part of probability statements will be suppressed. Approaches that take into consideration different models simultaneously will be considered in the next section. In those cases, the model must be explicitly considered in the conditioning part of the distributions.

### 7.2.1 *Estimates of the predictive likelihood*

For any given model, Equation (7.1) can be written as

$$f(y) = E[f(y|\theta)] \tag{7.2}$$

where expectation is taken with respect to the prior distribution $p(\theta)$. This simple identity and several generalizations can be used to *estimate* $f(y)$. In what follows, the well known Monte Carlo and importance function Monte Carlo identities as well as the bridge and path identities are introduced. Additional estimators derived from normal approximation to the posterior distribution and Chib's identities are also introduced. In what follows, $f(y|\theta)$ and $l(\theta)$ will be used interchangeably.

*Normal approximation*

Normal approximation to the model likelihood has previously been given in Chapter 3. The normal approximation to the posterior gives the estimate (3.4) for $f(y)$, i.e., $p(m)l(m)(2\pi)^{d/2}|V|^{1/2}$, which is based on the evaluation of the values of $m$, the posterior mode, and $V$, an asymptotic approximation for the posterior variance matrix. Sampling-based approximations for $m$ and $V$ can be constructed if a sample $\theta_1, \ldots, \theta_n$ from the posterior is available. The mode $m$ can be estimated as the sample value $\hat{m}$ for which $\pi$ is largest, i.e., $\pi(\hat{m}) = \max_j\{\pi(\theta_j)\}$. Similarly, estimates for the posterior variance matrix may be given in the case of an independent sample by $\hat{V} = \frac{1}{n}\sum_{j=1}^n (\theta_j - \bar{\theta})(\theta_j - \bar{\theta})'$, where $\bar{\theta} = \frac{1}{n}\sum_{j=1}^n \theta_j$. Therefore,

$$\hat{f}_0(y) = p(\hat{m})l(\hat{m})(2\pi)^{d/2}|\hat{V}|^{1/2}$$

is the normal approximation to $f(y)$. Lewis and Raftery (1997) named this estimator the *Laplace-Metropolis* estimator. Kass and Raftery (1995), Raftery (1996) and DiCiccio et al. (1997), among others, discussed alternative calculations of the value of $m$ when computation of $\pi$ is expensive and of the value of $V$ with the use of robust estimators.

*Monte Carlo approximations*

As it was pointed out at the beginning of this section, the Monte Carlo estimate derived from the identity of Equation (7.2) is

$$\hat{f}_1(y) = \frac{1}{n}\sum_{j=1}^n f(y|\theta_j)$$

where $\theta_1, \ldots, \theta_n$ is a sample from the prior distribution $p(\theta)$.

Raftery (1996) argued that this estimator does not work well in cases of disagreement between prior and likelihood, based also on applications by McCulloch and Rossi (1991). Almost all previous chapters contained some discussion of the difficulties associated with approximate inferences based on the prior, especially with sampling-based approaches. In light of this information, it is not surprising to learn that $\hat{f}_1$ does not provide a sensible estimate. It averages likelihood values that are chosen according to the prior. In general, the likelihood is more concentrated than the prior and the majority of $\theta_i$ will be placed in low likelihood regions. Even for large values of $n$, this estimate will be influenced by a few sampled values, making it very unstable. For similar problems, see Figures 1.4 and 3.2 and the penultimate paragraph of Section 6.3.3.

An alternative is to perform importance sampling with the aim of boosting sampled values in regions where the integrand is large. This approach is based on sampling from the importance density $g(\theta) = kg^*(\theta)$ where $g^*$ is the unnormalized form of the density and $k$ is a normalizing constant. When $k$ is known, Equation (7.2) can be rewritten as

$$f(y) = E_g\left[\frac{f(y|\theta)p(\theta)}{g(\theta)}\right] \tag{7.3}$$

where $E_g$ denotes an expectation with respect to the importance distribution $g(\theta)$. This form motivates a new estimate

$$\hat{f}_2(y) = \frac{1}{n}\sum_{j=1}^n \frac{f(y|\theta_j)p(\theta_j)}{g(\theta_j)}$$

where $\theta_1, \ldots, \theta_n$ is a sample from the importance density $g(\theta)$.

In many cases, the value of $k$ is not known and must be estimated. Noting that

$$k = \int kp(\theta)d\theta = \int \frac{p(\theta)}{g^*(\theta)}g(\theta)d\theta$$

leads to the estimator of $k$ given by $\hat{k} = (1/n)\sum_{j=1}^n p(\theta_j)/g^*(\theta_j)$ where, again, the $\theta_i$ are sampled from $g$. Replacing this estimate in $\hat{f}_2$ gives

$$\hat{f}_3(y) = \frac{\sum_{j=1}^{n} f(y|\theta_j)p(\theta_j)/g^*(\theta_j)}{\sum_{j=1}^{n} p(\theta_j)/g^*(\theta_j)} .$$

Special cases of $g$ are given by:

1. *Importance function:* $g(\theta) = \pi(\theta)$

   These results can be applied in the Bayesian context by taking the posterior density $\pi$ as the importance density. After a (Markov chain) simulation process, a sample $\theta_1, \ldots, \theta_n$ from $\pi(\theta) = kl(\theta)p(\theta)$ is available. These values can be used in $\hat{f}_3$ with $g^* = l \times p$. In this case, the estimator simplifies to

   $$\hat{f}_4(y) = \left[\frac{1}{n}\sum_{j=1}^{n}\frac{1}{l(\theta_j)}\right]^{-1} \qquad (7.4)$$

   This estimator is the harmonic mean of likelihood values originally proposed by Newton and Raftery (1994). It is commonly known as the *harmonic mean estimator*. The simplicity of $\hat{f}_4$ make it a very appealing estimator and its use is recommended provided the sample is large enough. Despite its consistency, this estimator is strongly affected by small likelihood values. Raftery (1996) relates this weakness to the occasional divergence of the variance of the terms in (7.4).

2. *Importance function:* $g(\theta) = \delta p(\theta) + (1-\delta)\pi(\theta)$

   Newton and Raftery (1994) introduced an estimator that is a compromise between $\hat{f}_1$, derived from prior draws, and $\hat{f}_4$, derived from posterior draws. More precisely, they suggested $g(\theta) = \delta p(\theta) + (1-\delta)\pi(\theta)$, for $0 < \delta < 1$, as the importance function of identity (7.3). Therefore, when $\theta_1, \ldots, \theta_n$ is a sample from $g(\theta)$, $\hat{f}_2$ is the estimator of $f(y)$. Unfortunately, $f(y)$ needs to be known in order to evaluate $\pi(\theta)$. This dependence suggests the following iterative scheme to estimate $f(y)$:

   $$\hat{f}_5^{(i)}(y) = \frac{\sum_{j=1}^{n} l(\theta_j)\{\delta\hat{f}_5^{(i-1)}(y) + (1-\delta)l(\theta_j)\}^{-1}}{\sum_{j=1}^{n}\{\delta\hat{f}_5^{(i-1)}(y) + (1-\delta)l(\theta_j)\}^{-1}} \qquad (7.5)$$

   for $i = 1, 2, \ldots$ and, say, $\hat{f}_5^{(0)} = \hat{f}_4$. A small number of iterations is usually enough for convergence. Even though $\hat{f}_5$ avoids the instability of $\hat{f}_4$, with the additional cost of also simulating from the prior.

3. *Generalized harmonic mean estimator*

   Another generalization of the harmonic mean estimator was obtained by Gelfand and Dey (1994). For any given density $g(\theta)$,

   $$1 = \int g(\theta)d\theta = \int g(\theta)\frac{f(y)\pi(\theta)}{f(y|\theta)p(\theta)}d\theta$$

where, as before, $p$ is the prior and $\pi$ is the posterior density of $\theta$. So,

$$f(y) = \left[\int \frac{g(\theta)}{f(y|\theta)p(\theta)}\pi(\theta)d\theta\right]^{-1} .$$

Sampling $\theta_1, \ldots, \theta_n$ from $\pi$ leads to the estimate

$$\hat{f}_6(y) = \left[\frac{1}{n}\sum_{j=1}^{n}\frac{g(\theta_j)}{f(y|\theta_j)p(\theta_j)}\right]^{-1} . \qquad (7.6)$$

Even though the method is specified for any density $g$, appropriate choices are very important for a good practical implementation. Gelfand and Dey (1994) suggested using $g$ as an importance density for the posterior and to take a normal or $t$ distribution that approximates $\pi$ with moments based on the sample of $\theta$. Raftery (1996) presented a simple example where $g$ was taken in product forms for each parameter component. The estimates obtained are highly inaccurate, showing that some skill is required in choosing $g$.

### *Annealed importance sampling*

Neal (2001) introduced a modification of the weighted resampling algorithm for sampling from a target distribution (usually the posterior) based on a sequence of proposal densities. The algorithm is particularly useful when the target density and the first density of the sequence have relatively little overlap (vague prior distributions or peaked likelihood functions, for instance). As a by product, the derived weights can be used to compute an estimate of $f(y)$.

More specifically, let $g_0$ be the starting density, $g_k$ be the target density,

$$g_j(\theta) = c_j\{g_0(\theta)\}^{1-\lambda_j}\{g_k(\theta)\}^{\lambda_j}$$

be the intermediate densities, where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_k = 1$ and $c_j$ be the normalizing constants. Starting at $\theta^{(0)}$, a sequence $\theta^{(1)}, \ldots, \theta^{(k)}$ is sampled as follows. First, sample $\theta^*$ from $g_0(\theta)$ and set $\theta^{(1)} = \theta^*$ with probability

$$\alpha(\theta^{(0)}, \theta^*) = \min\left\{1, \frac{g_1(\theta^*)}{g_1(\theta^{(0)})}\frac{g_0(\theta^{(0)})}{g_0(\theta^*)}\right\} .$$

Then, for $i = 2, \ldots, k$, sample $\theta^*$ from $q_{i-1}(\theta^{(i-1)}, \cdot)$ and set $\theta^{(i)} = \theta^*$ with probability

$$\alpha(\theta^{(i-1)}, \theta^*) = \min\left\{1, \frac{g_i(\theta^*)}{g_i(\theta^{(i-1)})}\frac{q_{i-1}(\theta^*, \theta^{(i-1)})}{q_{i-1}(\theta^{(i-1)}, \theta^*)}\right\} .$$

By repeating the previous algorithm $n$ times and keeping the $\theta^{(k)}$s, it can be shown that the sample $\theta_1^{(k)}, \ldots, \theta_n^{(k)}$ and the weights $\omega_1, \ldots, \omega_n$ jointly summarize the target distribution $g_k(\theta)$. Unnormalized weights are computed as (see Exercise 7.2a)

$$\tilde{\omega}_j = \prod_{i=1}^{k} \frac{g_i(\theta_j^{(i)})}{g_{i-1}(\theta_j^{(i)})}, \quad j = 1, \ldots, n. \tag{7.7}$$

Weights are obtained as $\omega_j = \tilde{\omega}_j / \sum_{i=1}^n \tilde{\omega}_i$, $j = 1, \ldots, n$. When $g_0(\theta)$ is the prior $p(\theta)$ and $g_k(\theta)$ is the posterior $\pi(\theta)$, the annealed importance sampling estimator of $f(y)$ is (see Exercise 7.2b)

$$\hat{f}_7(y) = \frac{1}{n} \sum_{j=1}^{n} \tilde{\omega}_j \; . \tag{7.8}$$

As a by product, $E_{g_k}(\theta)$, for instance, can be approximated by the weighted average $\sum_{j=1}^n \omega_j \theta_j^{(k)}$. When $k = 1$, the algorithm degenerates into the simple Monte Carlo estimator $\hat{f}_1$. Neal (2001) argued that the implementation of the annealed importance sampling is commonly straightforward and particularly useful when handling isolated modes. Combining independent sampling with Markov chain sampling is another attractive aspect of the estimator.

### *Bridge and path identities*

Meng and Wong (1996) introduced the *bridge sampling* to estimate ratios of normalizing constants. Notice that $f(y)$ can be rewritten as

$$f(y) = \frac{E_g\{\alpha(\theta)p(\theta)l(\theta)\}}{E_\pi\{\alpha(\theta)g(\theta)\}} \tag{7.9}$$

for any arbitrary *bridge* function $\alpha(\theta)$ with support encompassing both supports of the posterior density $\pi$ and the proposal density $g$ (see Exercise 7.3). If $\alpha(\theta) = 1/g(\theta)$ then the bridge estimator reduces to the simple Monte Carlo estimator $\hat{f}_1$. Similarly, if $\alpha(\theta) = \{p(\theta)l(\theta)g(\theta)\}^{-1}$ then the bridge estimator is a variation of the harmonic mean estimator. If $\theta_1, \ldots, \theta_n$ and $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$ are samples from $\pi$ and $g$, respectively, then

$$\frac{\frac{1}{m}\sum_{j=1}^m \alpha(\tilde{\theta}_j)p(\tilde{\theta}_j)l(\tilde{\theta}_j)}{\frac{1}{n}\sum_{j=1}^n \alpha(\theta_j)g(\theta_j)}$$

is a bridge estimator of $f(y)$. Meng and Wong (1996) showed that the optimal mean square error $\alpha$ function is $\alpha(\theta) = \{g(\theta) + (m/n)\pi(\theta)\}^{-1}$, which depends on $f(y)$ itself. By letting $\omega_j = l(\theta_j)p(\theta_j)/g(\theta_j)$, for $j = 1, \ldots, n$ and $\tilde{\omega}_j = l(\tilde{\theta}_j)p(\tilde{\theta}_j)/g(\tilde{\theta}_j)$, for $j = 1, \ldots, m$, they devised the iterative scheme below to estimate $f(y)$:

$$\hat{f}_8^{(i)}(y) = \frac{\frac{1}{m}\sum_{j=1}^m \tilde{\omega}_j[s_1\tilde{\omega}_j + s_2\hat{f}_8^{(i-1)}(y)]^{-1}}{\frac{1}{n}\sum_{j=1}^n [s_1\omega_j + s_2\hat{f}_8^{(i-1)}(y)]^{-1}} \; ,$$

for $i = 1, 2, \ldots$, $s_1 = n/(m+n)$, $s_2 = m/(m+n)$ and, say, $\hat{f}_8^{(0)} = \hat{f}_4$. A small number of iterations is usually enough for convergence. Other alternatives for $\alpha$ are considered in Meng and Wong (1996).

The bridge sampling efficacy decreases as the distance between $\pi$ and $g$ increases. Metaphorically speaking, it will be costly (and inefficient) to cross the bridge when the length of the bridge is large. Chen and Shao (1997) extend the bridge identity to situations where $g$ and $\pi$ are defined over spaces of different dimensionality. For further details about the bridge sampler, see also Meng and Schilling (1996).

Gelman and Meng (1998) generalized the bridge sampling by replacing one (possibly long) bridge by infinitely many shorter bridges or, as they call it, a *path*. Suppose that instead of using the bridge function $\alpha$ to connect $p(\theta)$ and $p(\theta)l(\theta)$, a path function $h(\theta|\lambda)$ is constructed with $h(\theta|\lambda = 0) = p(\theta)$ and $h(\theta|\lambda = 1) = p(\theta)l(\theta)$. In other words, $h(\theta|\lambda = 0)$ represents the beginning of the path and $h(\theta|\lambda = 1)$ represents the end of the path.

Let $c(\lambda) = \int h(\theta|\lambda)d\theta$ be the normalizing constant of $h(\theta|\lambda)$, so $g(\theta|\lambda) = h(\theta|\lambda)/c(\lambda)$ is a normalized density function and $c(0) = 1$ and $c(1) = f(y)$. Then, it is easy to see that (see Exercise 7.4)

$$f(y) = \exp\left\{\int_0^1 \int_\Theta \frac{H(\theta,\lambda)}{g(\lambda)} g(\theta|\lambda)g(\lambda) \; d\theta \; d\lambda\right\} \tag{7.10}$$

where $H(\theta,\lambda) = \frac{d}{d\lambda}\log h(\theta|\lambda)$ and $g(\lambda)$ is any density for $\lambda$ over the unit interval $[0,1]$. Usually the above integrals cannot be solved analytically and must be approximated, for instance, by Monte Carlo integration leading to the path estimator $\hat{f}_9(y)$.

### *Candidate's estimators*

A very simple estimate, usually called the *candidate's estimator* (Besag, 1989), can be derived from the fact that $f(y) = f(y|\theta)p(\theta)/\pi(\theta)$. Typically, $f(y|\theta)$ and $p(\theta)$ are typically easy to calculate but $\pi(\theta)$ is not. However, if a sample of $\pi$ is available, some form of histogram smoothing can be applied to get an estimate of $\pi$.

Chib (1995) introduced an alternative estimate of $\pi$ when full conditional densities are available in closed form, as in Gibbs sampling, for instance. Note first that, for $\theta = (\theta_1, \ldots, \theta_d)$ and $i = 2, \ldots, d$,

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}) = \int \cdots \int \pi(\theta_i|\theta_{-i})\pi(\theta_{i+1}, \ldots, \theta_d)d\theta_{i+1} \cdots d\theta_d$$

suggests approximating $\pi(\theta_i|\theta_1,\ldots,\theta_{i-1})$ by $\hat{\pi}(\theta_i|\theta_1,\ldots,\theta_{i-1})$ given by

$$\hat{\pi}(\theta_i|\theta_1,\ldots,\theta_{i-1}) = \frac{1}{n}\sum_{j=1}^{n}\pi(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1}^{(j)},\ldots,\theta_d^{(j)})$$

where $(\theta_1^{(j)},\ldots,\theta_d^{(j)})'$, $j = 1,\ldots,n$, is a sample from $\pi(\theta)$.

As $\pi(\theta) = \pi(\theta_1)\prod_{i=1}^{d}\pi(\theta_i|\theta_1,\ldots,\theta_{i-1})$, an approximation $\hat{\pi}$ for the posterior is given by

$$\hat{\pi}(\theta) = \hat{\pi}(\theta_1)\prod_{i=1}^{d}\hat{\pi}(\theta_i|\theta_1,\ldots,\theta_{i-1}) \ .$$

Once an approximation $\hat{\pi}$ for $\pi$ is available, it can be used to give another estimate

$$\hat{f}_{10}(y) = \frac{f(y|\theta)p(\theta)}{\hat{\pi}(\theta)} \ . \qquad (7.11)$$

Note that any value of $\theta$ can be used in the expression of $\hat{f}_{10}$ and if $\pi$ could have been obtained without error, they would all provide the same estimate of $f(y)$. Obviously, $\theta$ should be chosen so that $\hat{\pi}$ has the smallest possible estimation error. This narrows the choice of $\theta$ to the central region of the posterior where $\pi$ is likely to be estimated more accurately. Simple choices are the mode and the mean but any value in that region should be adequate.

Chib and Jeliazkov (2001) extended the above idea for cases where some (or none) of the full conditional densities are of unknown form and difficult to sample from and Metropolis-Hastings output is available. Consider the simple case where a value $\phi$ is sampled in a block from the Metropolis-Hastings proposal $q(\theta,\phi)$ and accepted with probability $\alpha(\theta,\phi)$. For any value of $\phi$, Chib and Jeliazkov (2001) showed that (see Exercise 7.5)

$$\pi(\phi) = \frac{E_{\pi(\theta)}\left\{\alpha(\theta,\phi)q(\theta,\phi)\right\}}{E_{q(\phi,\theta)}\left\{\alpha(\phi,\theta)\right\}}. \qquad (7.12)$$

If $\theta_0^{(1)},\ldots,\theta_0^{(n_0)}$ and $\theta_1^{(1)},\ldots,\theta_1^{(n_1)}$ are samples from $\pi(\cdot)$ and $q(\phi,\cdot)$, respectively, then

$$\hat{\pi}(\phi) = \frac{n_0^{-1}\sum_{j=1}^{n_0}\alpha(\theta_0^{(j)},\phi)q(\theta_0^{(j)},\phi)}{n_1^{-1}\sum_{j=1}^{n_1}\alpha(\phi,\theta_1^{(j)})}$$

is an estimate of $\pi(\phi)$, which can be used to estimate $f(y)$ via

$$\hat{f}_{11}(y) = \frac{f(y|\phi)p(\phi)}{\hat{\pi}(\phi)}.$$

Chib and Jeliazkov (2001) also extended the above idea to multiple parameter blocks, while Chib and Jeliazkov (2005) estimated marginal likelihood based on output from accept-reject Metropolis-Hastings schemes.

By noticing the similarity between Equations (7.9) and (7.12), Mira and Nicholls (2004) showed that Chib and Jeliazkov's estimator is a special case of the bridge sampler with $\alpha(\phi,\theta) = \alpha(\theta)p(\theta)l(\theta)$.

Table 7.1 lists the several estimates of the predictive likelihood introduced in this section, while Examples 7.1 and 7.2 illustrate and compare them.

| Estimate | Proposal density/method |
|---|---|
| $\hat{f}_0$ | normal approximation |
| $\hat{f}_1$ | $p(\theta)$ |
| $\hat{f}_2$ | unnormalized $g^*(\theta)$ |
| $\hat{f}_3$ | unnormalized $g(\theta)$ |
| $\hat{f}_4$ | $\pi(\theta)$ |
| $\hat{f}_5$ | $\delta p(\theta) + (1-\delta)\pi(\theta)$ |
| $\hat{f}_6$ | generalized harmonic mean |
| $\hat{f}_7$ | annealed importance sampling |
| $\hat{f}_8$ | optimal bridge sampling |
| $\hat{f}_9$ | path sampling |
| $\hat{f}_{10}$ | candidate's estimator from Gibbs output |
| $\hat{f}_{11}$ | candidate's estimator from Metropolis output |

Table 7.1 *List of estimates of the predictive likelihood.*

DiCiccio, Kass, Raftery and Wasserman (1997), Han and Carlin (2001) and Lopes and West (2004), among others, compared several of estimators introduced in this section.

**Example 7.1** *Consider Example 2.3 where $y_1,\ldots,y_n$ are a $N(\theta,\sigma^2)$ random sample, with $\sigma^2$ known and Cauchy prior density $p(\theta) = \pi^{-1}(1+\theta^2)^{-1}$. Assume that $\bar{y} = 7$ and $\sigma^2/n = 4.5$. The likelihood function is $l(\theta) = f_N(\theta;\bar{y},\sigma^2/n)$ and the posterior density for $\theta$ is $\pi(\theta) \propto p(\theta)l(\theta)$. A very accurate approximation to the normalizing constant was obtained by $h\sum_{j=1}^{k} l(\theta_j)p(\theta_j) = 0.00963235$, where $\theta_1 = -15, \theta_j = -15+(j-1)h, j = 2,\ldots,k-1, \theta_k = 15$, $k = 10^7$ and $h = 3\times 10^{-6}$.*

*A normal approximation to the posterior density is $g(\theta) = f_N(\theta;m,\hat{V})$ where $m$ is the posterior mode and $\hat{V}$ is a Monte Carlo estimate of the posterior variance of $\theta$. The posterior mode, $m = 5.384$, was obtained by a Newton-Raphson-type algorithm that started at $\theta^{(0)} = 0.05$ and ran for 9 iterations (see Section 3.2.2).*

*Subsequently, 100000 draws from a $N(m,2^2)$ distribution were used in a weighted resampling scheme to generate 100000 draws from the posterior distribution $\pi$ and compute $\hat{V} = 2.49^2$. Figure 7.1(a) exhibits the likelihood function as well as the prior and the normal approximation. The resulting histogram approximation of $\pi$ appears in Figure 7.1(b).*

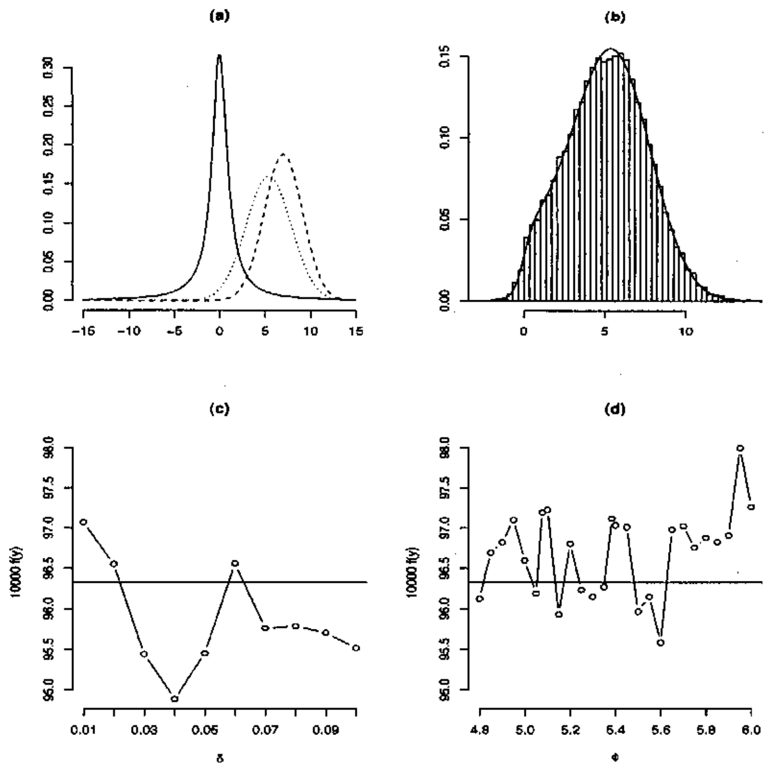

Figure 7.1 *Comparing several estimators of $f(y)$. (a) Prior density $p(\theta)$ (solid line), likelihood function $l(\theta)$ (dashed line) and normal approximation $g(\theta)$ (dotted line). (b) Histogram approximation for $\pi$ along with $\pi$. (c) $\hat{f}_5(y)$ for $\delta \in [0.01, 0.1]$. (d) $\hat{f}_{11}(y)$ for several values of $\phi$ in $[4.8, 6.0]$. Horizontal lines in (c) and (d) correspond to* 0.00963235, *the true value of $f(y)$.*

*For $\hat{f}_5$, $\delta$ was set to* 0.1, *i.e., an average of 10% of prior draws. For $\delta$ in $[0.01, 0.1]$, the estimation error of $\hat{f}_5$ ranged from* 0.02% *to* 2.04% *(see Figure 7.1(c)). For $\hat{f}_6$ and $\hat{f}_8$, the proposal density $g$ is the normal approximation used above. A total of 20000 draws is used to compute $\hat{f}_7$, where $\lambda_j = 0.1j$, for $j = 1, \ldots, 5$ and $\lambda_j = 0.51 + 0.035(j-6)$, for $j = 6, \ldots, 20$ and $q_i(\theta, \phi) = f_N(\phi; \theta, 0.1^2)$, for $i = 1, \ldots, 19$. For $\hat{f}_8$'s iterative algorithm, the initial value was set at $\hat{f}_0$. For $\hat{f}_{11}$, $\phi = m$ and $q(\phi, \theta) = f_N(\theta; \phi, \hat{V})$. Also, $\hat{f}_{11} =$* 0.00970631 *when $\phi = E_\pi(\theta) \approx$* 5.076856. *In fact, the estimation error of $\hat{f}_{11}$ ranged from* 0.06% *to* 1.74%, *for $\phi \in [4.8, 6.0]$ (see Figure 7.1(d)). The results of all estimators are given in Table 7.2.*

| *$f(y)$* | *0.00963235* | |
|---|---|---|
| *Estimator* | *Estimate* | *Error (%)* |
| $\hat{f}_0$ | *0.00932328* | *3.21* |
| $\hat{f}_1$ | *0.00960189* | *0.32* |
| $\hat{f}_4$ | *0.01055301* | *9.56* |
| $\hat{f}_5$ | *0.00957345* | *0.61* |
| $\hat{f}_6$ | *0.00962871* | *0.04* |
| $\hat{f}_7$ | *0.01044794* | *8.47* |
| $\hat{f}_8$ | *0.00963110* | *0.01* |
| $\hat{f}_{11}$ | *0.00969942* | *0.70* |

Table 7.2 *Comparing several estimator of $f(y)$. Error $= 100|\hat{f}(y) - f(y)|/f(y)$. For $\hat{f}_5$, $\delta = 0.1$ and for $\hat{f}_{11}$, $\phi = m$.*

**Example 7.2** *Lopes and West (2004) examined the performance of several estimators of $f(y|k)$ in a traditional $k$-factor model. In one case, a 9-dimensional vector is simulated from a 3-factor model. Subsequently, $k$-factor analysis was entertained for $k = 1, \ldots, 5$, i.e.,*

$$\begin{aligned} y_i | f_{ki}, \theta_k &\sim N(\beta_k f_{ki}, \Sigma_k) \\ f_{ki} &\sim N(0, I_k) \end{aligned}$$

*for $i = 1, \ldots, n$ and $\theta_k = (\beta_k, \Sigma_k)$ and $\Sigma_k = diag(\sigma^2_{k1}, \ldots, \sigma^2_{kk})$. Identifiability constraints impose that $\beta_{kii} > 0$, for $i = 1, \ldots, k$, and $\beta_{kij} = 0$, for $j > i$, which characterizes the lower block triangular shape of $\beta_k$ (Lopes, 2000).*

*The true model parameters are*

$$\beta_3' = \begin{pmatrix} 0.99 & 0.00 & 0.00 & 0.99 & 0.99 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.95 & 0.00 & 0.00 & 0.00 & 0.95 & 0.95 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.00 & 0.00 & 0.00 & 0.00 & 0.90 & 0.90 \end{pmatrix}$$

*and $\Sigma_3 = diag(0.02, 0.19, 0.36, 0.02, 0.02, 0.19, 0.19, 0.36, 0.36)$. The prior distribution of $\beta_{kij}$ is normal, i.e., $\beta_{kij} \sim N(0, 1)$, for $i = 2, \ldots, 9$, $j = 1, \ldots, i-1$ and $k = 1, \ldots, 5$, and the prior distribution of $\sigma^2_{ki}$ is inverse Gamma, i.e., $\sigma^2_{ki} \sim IG(1.1, 0.05)$, so that $E(\sigma^2_{ki}) = 0.5$, for $i = 1, \ldots, k$.*

*Conditional on $k$, a Gibbs sampler is promptly available for inference about $(\beta_k, \Sigma_k)$ and $f_{k1}, \ldots, f_{kn}$ (Geweke and Zhou, 1996; Aguilar and West, 2000; Lopes, 2000; and Lopes and West, 2004). Posterior inference was* *based on* 10000 *iterations as burn-in, followed by a further 10000 iterates that were sampled every ten steps to produce a final MCMC sample of size 1000. For* $\hat{f}_5$*, the control parameter* $\delta$ *was set at* 0.05 *and the iterative scheme run for* 20 *iterations.*

*Uniform prior model probabilities were assumed, i.e.,* $Pr(k) = 0.2$ *for* $k = 1, \ldots, 5$. *Table 7.3 shows the frequencies, out of 1000 simulated sets of data, at which each k-factor model had the highest posterior model probability. Several of the approximation methods reliably identify the true model structure, which gives some indication of their likely utility in real data analysis. Among the approximate Bayesian methods, those based on the harmonic mean estimator* ($\hat{f}_4$)*, the Newton-Raftery estimator* ($\hat{f}_5$) *and the Chib's estimator* ($\hat{f}_{10}$) *exhibited worse performances. Sensitivity to prior distributions was studied by Lopes (2000, 2003).*

| | *Number of factors* | | | | |
|---|---|---|---|---|---|
| *Estimator* | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ |
| $\hat{f}_0$ | *0* | *1* | *999* | *0* | *0* |
| $\hat{f}_4$ | *0* | *0* | *650* | *228* | *122* |
| $\hat{f}_5$ | *0* | *0* | *615* | *258* | *127* |
| $\hat{f}_6$ | *0* | *0* | *998* | *2* | *0* |
| $\hat{f}_8$ | *0* | *11* | *985* | *4* | *0* |
| $\hat{f}_{10}$ | *0* | *12* | *848* | *138* | *2* |

Table 7.3 *Number of times (out of 1000 replications) that a particular k-factor model was chosen as the* best *by each one of several estimators.*

### 7.2.2 *Uses of the predictive likelihood*

When the prior distribution $p(\theta)$ is informative and proper, there is no problem in using $f(y)$ to evaluate different models. When the prior is improper, $f(y)$ cannot be calculated because the integral (7.1) diverges. In cases where the prior is proper but not very informative, the use of $f(y)$ as a tool for model evaluation should be made with care. So, the estimates obtained above for approximating $f(y)$ with weak prior information are likely to be unstable.

Even though the density (7.1) is the canonical form for model evaluation as prescribed by theory, it is not necessarily the only one. Different justifications have led to other approaches by several authors. Consider a sample $y = (y_1, \ldots, y_n)'$ and $y_C$ denotes the subset of $y$ containing the observations with indices in $C$. So, if $C = \{1, n\}$, $y_C = (y_1, y_n)'$ and if $C = \{1, \ldots, i-1, i+1, \ldots, n\}$, $y_C = y_{-i}$. Gelfand and Dey (1994) showed that many densities used for model evaluation can be written in the generic form

$$f(y_{S_1}|y_{S_2}) = \int f(y_{S_1}|\theta)p(\theta|y_{S_2})d\theta.$$

$S_1 = \{1, \ldots, n\}$ and $S_2 = \phi$, the empty set, leads to $f(y)$. If $S_1 = S_2 = \{1, \ldots, n\}$, then the density suggested by Aitkin (1991) for use in the posterior Bayes factor is obtained. The densities appearing in the definitions of the intrinsic Bayes factor of Berger and Pericchi (1996) and the fractional Bayes factor of O'Hagan (1995) also fit into this formulation. The main motivation for these alternative forms is their use with improper priors. An adequate choice of $S_2$ removes the impropriety of $p(\theta|y_{S_2})$ and therefore $f(y_{S_1}|y_{S_2})$ does not diverge and can be calculated.

Another density extensively used by Gelfand (1996) and Gelfand, Dey and Chang (1992) was the cross-validation predictive density $f(y_i|y_{-i})$ (Stone, 1974). Geisser and Eddy (1979) suggested the use of the product $\prod_{i=1}^{n} f(y_i|y_{-i})$ of these densities as a surrogate indicator of the value of the predictive likelihood $f(y)$ through the pseudo Bayes factor

$$\frac{\prod_{i=1}^{n} f(y_i|y_{-i}, M_0)}{\prod_{i=1}^{n} f(y_i|y_{-i}, M_1)}$$

that would approximate the Bayes factor. Gelfand, Dey and Chang (1992) suggested many forms of use of the predictive densities through the expectations of functions $g(y_i)$* under $f(y_i|y_{-i})$. Among them is the prediction error $g_1(y_i) = y_i - y_{i,obs}$ where $y_{i,obs}$ is the observed value of $y_i$. The expectation of $g_1$ is $\gamma_{1i} = E(y_i|y_{-i}) - y_{i,obs}$. These values may be standardized to $\gamma'_{1i} = \gamma_{1i}/\sqrt{Var(y_i|y_{-i})}$. If the model is adequately fitting the data, the values of $\gamma'_{1i}$ should be small. Under approximate normality, 95% of them should roughly lie between -2 and 2. The quantity

$$G_1 = \sum_{i=1}^{n} \gamma'^2_{1i}$$

may be constructed and used as an indicator of model fit. The smaller its value, the better the fit of the model to the data.

Another useful function is $g_2(y_i) = I(y_i \leq y_{i,obs})$ with expectation $\gamma_{2i} = Pr(y_i \leq y_{i,obs}|y_{-i})$. Considering $\gamma_{2i}$ as functions of $y_{i,obs}$ they are $U[0,1]$ distributed but they are not independent. So, the behavior of a sample from a $U[0,1]$ distribution is expected. A large number of $\gamma_{2i}$s close to 0 or 1 indicates observations in the tails of their predictive densities, i.e., poor fit of the model, whereas a large number of $\gamma_{2i}$s close to 1/2 indicates ob-

* The reason why the functions $g$ are considered instead of considering directly their expectations will be made clear below.

servations close to their corresponding predictive medians, i.e., good model fit. A possible summarization of the information from the $\gamma_{2i}$s is obtained with

$$G_2 = \sum_{i=1}^{n} (\gamma_{2i} - 0.5)^2 \ .$$

Again, the smaller its value, the better the fit of the model to the data.

Similarly, the functions $g_3(y_i) = I(y_i \in \{y_i | f(y_i|y_{-i}) \leq f(y_{i,obs}|y_{-i})\}$ and $g_4(y_i) = I(y_i \in [y_{i,obs} - \epsilon, y_{i,obs} + \epsilon])/2\epsilon$ may be used. Their expectations are respectively given by $\gamma_{3i} = Pr(\{y_i|f(y_i|y_{-i}) \leq f(y_{i,obs}|y_{-i})\}|y_{-i})$ and $\gamma_{4i} = f(y_{i,obs}|y_{-i})$, when $\epsilon \to 0$. Again, as functions of $y_{i,obs}$, the $\gamma_{3i}$s are a sample from a $U[0,1]$ distribution and therefore can be summarized by

$$G_3 = \sum_{i=1}^{n} (\gamma_{3i} - 0.5)^2 \ .$$

As for the $\gamma_{4i}$s, also known as the *conditional predictive ordinate* (CPO), the natural summarizing quantity is their product

$$G_4 = \prod_{i=1}^{n} \gamma_{4i}$$

which is also present in the pseudo Bayes factor.

None of the functions of interest $\gamma_{ji}$ can be obtained analytically for most models. However, sampling-based estimates can be obtained as they were written as expectations with respect to a distribution. Assuming the presence of a sample $y_{i1}, \ldots, y_{in}$ from $p(y_i|y_{-i})$, the $\gamma_{ji}$ can be estimated by

$$\hat{\gamma}_{ji} = \frac{1}{n} \sum_{l=1}^{n} g_j(y_{il}) \ , \ j = 1, 2, 3, 4.$$

A sample from $p(y_i|y_{-i})$ can be obtained by noting that

$$\begin{aligned} p(y_i|y_{-i}) &= \int p(y_i, \theta|y_{-i}) d\theta \\ &= \int p(y_i|\theta) p(\theta|y_{-i}) d\theta \end{aligned}$$

where the last equality follows from the usual assumption of conditional independence of the observations given $\theta$. Therefore a draw from $p(y_i|y_{-i})$ is obtained by drawing $\theta_*$ from $p(\theta|y_{-i})$ and subsequently drawing $y_i$ from $f(y_i|\theta_*)$ (Section 1.3).

Only a sampling scheme for $p(\theta|y_{-i})$ remains to be described. Gelfand, Dey and Chang (1992) suggestd the use of a resampling method (Section 1.5) with some approximating density $q(\theta)$. A natural choice for $q$ is the posterior density $\pi(\theta)$, for two reasons. For moderate to large samples, the exclusion of a single observation is not likely to significantly change the posterior. Therefore, $\pi(\theta)$ is likely to approximate $p(\theta|y_{-i})$ well. Also, a sample from $\pi$ is already available from an inferential procedure for a given model.

Bayes' theorem with prior $p(\theta|y_{-i})$ and observation $y_i$ gives $\pi(\theta) \propto f(y_i|\theta)\, p(\theta|y_{-i})$. Hence, following the construction of a sampling-importance resampling scheme in Section 1.5 and assuming the presence of a sample $\theta_1, \ldots, \theta_n$ from $\pi$, the weights

$$w_j \propto \frac{\pi(\theta_j)}{p(\theta_j|y_{-i})} \propto \frac{1}{f(y_i|\theta_j)}$$

can be formed. These weights are normalized to add to 1 and used in the resampling scheme. The resulting sample has approximate distribution $p(\theta|y_{-i})$. A similar scheme can be devised for the rejection method with the disadvantage of finding a constant ensuring complete envelope. Gelfand (1996) also described an alternative form to estimate the $d_{ji}$ directly from a sample of $\pi$. Azevedo (2002), for instance, used CPOs and pseudo Bayes factors when combining related studies through hierarchical models with Dirichlet process priors (Dey, Muller and Sinha, 1998).

**Example 7.3** *Data on the temporal evolution of the dry weight of onion bulbs (y) is presented in Table 7.4 and Figure 7.2 (Ratkowski, 1983).*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *16.08* | *33.83* | *65.80* | *97.20* | *191.55* | *326.20* | *386.87* | *520.53* |
| *590.03* | *651.92* | *724.93* | *699.56* | *689.96* | *637.56* | *717.41* | |

Table 7.4 *Data on the temporal evolution of the dry weight of onion bulbs (y).*

*Gelfand, Dey and Chang (1992) considered two non-linear models*

$$\begin{aligned} \textit{Logistic model} \ &: \ y_t \sim N(\theta_1/(1 + \theta_2\theta_3^t), \theta_4^2) \\ \textit{Gompertz model} \ &: \ y_t \sim N(\phi_1 + e^{\phi_2\phi_3^t}, \phi_4^2) \end{aligned}$$

*where $t = 1, \ldots, n = 15$ and set $\theta = (\theta_1, \ldots, \theta_4)$ and $\phi = (\phi_1, \ldots, \phi_4)$. $\theta_1$ and $\phi_1$ represent asymptotes in both models but the other parameters have different meanings under each model with $\theta_2 > 0$, $0 < \theta_3 < 1$, $\phi_2 > 0$ and $0 < \phi_3 < 1$. Some of the regression parameters in $\theta$ and $\phi$ were transformed so that each of them varies over the real line. $\theta_2$ was transformed to $\log\theta_2$, $\theta_3$ to $\log[\theta_3/(1-\theta_3)]$, $\phi_2$ to $\log\phi_2$, $\phi_3$ to $\log[\phi_3/(1-\phi_3)]$. Standard non-informative priors were then assumed preventing evaluation of the model likelihoods.*

*Gelfand, Dey and Chang (1992) generated a sample of size 2000 of the 15 cross-validation predictive densities. They sampled from these densities using the weighted resampling technique. Rather than getting an initial sample from the posterior distribution, they approximated it by a t distribution* *and sampled from it. The parameters of the t distributions were set in accordance with a previous fit using the non-linear routines from the software SAS for each model. The individual values of the $\gamma_{ji}$ were also provided for each model (Gelfand, Dey and Chang, 1992, Table 2). The summarizing quantities $G_1$, $G_2$ and $G_3$ are reproduced here in Table 7.5. The pseudo Bayes factor was approximated by 1.6863, indicating preference for the logistic model. This preference was confirmed by $G_2$ and $G_3$ as shown in Table 7.5.*

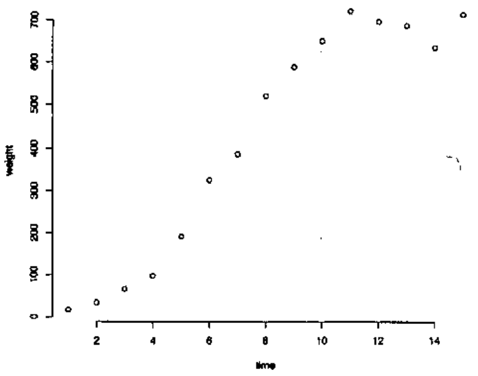

Figure 7.2 *Onion bulb data.*

| *Criteria* | *Logistic* | *Gompertz* |
|---|---|---|
| $G_1$ | *18.27* | *16.82* |
| $G_2$ | *0.92* | *1.32* |
| $G_3$ | *1.57* | *1.95* |

Table 7.5 *Summarization of model evaluation.*

*Posterior predictive criterion*

Gelfand and Ghosh (1998) introduced a posterior predictive criterion that, under squared error loss, favors the model $M_j$ which minimizes

$$D_j^G = P_j^G + G_j^G \tag{7.13}$$

where $P_j^G = \sum_{t=1}^n V(\tilde{y}_t|y, M_j)$, $G_j^G = \sum_{t=1}^n [y_t - E(\tilde{y}_t|y, M_j)]^2$ and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ are predictions/replicates of $y$. The first term, $P_j$, is a penalty term for model complexity with large values for too simple or too complex models. The second term, $G_j$, is a sum of squared residuals and accounts for goodness of fit when $y_t$ is predicted by $\tilde{y}_t$ under model $M_j$. Gelfand and Ghosh (1998) also derived the criteria for more general error loss functions.

Expectations $E(\tilde{y}_t|y, M_j)$ and variances $V(\tilde{y}_t|y, M_j)$ are computed under posterior predictive densities, ie.

$$E[h(\tilde{y}_t)|y, M_j] \;=\; \int\int h(\tilde{y}_t) f(\tilde{y}_t|y, \theta_j, M_j)\pi(\theta_j|M_j)d\theta_j d\tilde{y}_t$$

for $h(\tilde{y}_t) = \tilde{y}_t$ and $h(\tilde{y}_t) = \tilde{y}_t^2$. Therefore, when $\theta_j^{(1)}, \ldots, \theta_j^{(M)}$ is a sample from $\pi(\theta_j|M_j)$ the above integral can be approximated via Monte Carlo (see Section 3.4) by

$$\frac{1}{M}\sum_{i=1}^{M} \int h(\tilde{y}_t) f(\tilde{y}_t|y, \theta_j^{(i)}, M_j) d\tilde{y}_t \,.$$

In general, when $f(\tilde{y}_t|y, \theta_j, M_j) = f(\tilde{y}_t|\theta_j, M_j)$, the above integral can be computed analytically, at least for $h(\tilde{y}_t) = \tilde{y}_t$ and $h(\tilde{y}_t) = \tilde{y}_t^2$.

**Example 7.4** *Considering again Example 2.11, the two components of the minimum posterior predictive loss criteria are (see Exercise 7.6a):*

$$\begin{aligned}
P_1^G &= \frac{n\nu_1\sigma_1^2(1+c_1^{-1})}{\nu_1 - 2} && \textit{and} \quad G_1^G = \sum_{i=1}^n (y_i - \mu_1)^2 \,, \\
P_2^G &= \frac{n\eta_1\psi_1^2}{\eta_1 - 2} && \textit{and} \quad G_2^G = \sum_{i=1}^n (y_i - \alpha)^2 \,, \\
P_3^G &= n(\xi^2 + \delta_1^2) && \textit{and} \quad G_3^G = \sum_{i=1}^n (y_i - \gamma_1)^2 \,.
\end{aligned}$$

*7.2.3 Deviance information criterion*

This section presents different ways to evaluate and use the predictive likelihood, obtained by averaging the likelihood with respect to the prior. It is also conceivable to use the likelihood function averaged with respect to the posterior. Inspired by Dempster's (1997) suggestion to compute the posterior distribution of the log-likelihood, Spiegelhalter et al. (2002) introduced the *deviance information criterion* (DIC)

$$D_j^S = P_j^S + G_j^S \tag{7.14}$$

where $P_j^S = E[D(\theta_j)|y, M_j] - D[E(\theta_j|y, M_j)]$, $G_j^S = E[D(\theta_j)|y, M_j]$, and $D(\theta_j) = -2\log f(y|\theta_j, M_j)$. The DIC is decomposed into two important components: one responsible for goodness of fit ($G_j^S$) and one responsible for model complexity ($P_j^S$), just like the previous criterion. $P_j^S$ is also currently referred to as *the effective number of parameters* of model $M_j$.

If $\theta_j^{(1)}, \ldots, \theta_j^{(M)}$ is a sample from $\pi(\theta_j|M_j)$, then the DIC can be approximated via Monte Carlo by

$$\frac{2}{M}\sum_{i=1}^{M} D(\theta_j^{(i)}) - D\left(\frac{1}{M}\sum_{i=1}^{M}\theta_j^{(i)}\right) .$$

DIC is applied for variable selection in van der Linde (2005). Celeux et al. (2005) compare different DIC constructions in mixture models, random effects models and several missing data models.

This criterion became very popular in the applied Bayesian community due to its computational simplicity and, consequently, its availability in WinBUGS. Further applications appear, amongst many others, in Lopes and Salazar (2006a,b) (nonlinear time series models), Nobre, Schmidt and Lopes (2005) and Zhu and Carlin (2000) (space-time hierarchical models) and Berg, Meyer and Yu (2002) (stochastic volatility models).

**Example 7.4** *(continued) The deviance information criteria are (see Exercise 7.6b)*

$$\begin{aligned}
D_1^S &= -n\log\left(\frac{\nu_1\sigma_1^2}{\nu_1-2}\right) + \frac{(\nu_1+2)\sigma_1^2}{\nu_1}\sum_{i=1}^{n} y_i^2 + \frac{2\mu_1(\nu_1-2)}{\nu_1\sigma_1^2}\sum_{i=1}^{n} y_i \\
&- n\left(c_1^{-1} + \frac{\mu_1^2(\nu_1-2)}{\nu_1\sigma_1^2}\right) + 2nE(\log\sigma^2|y, M_1) \\
&- 2\left(\sum_{i=1}^{n} y_i\right) E\left(\frac{\mu}{\sigma^2}|y, M_1\right) + 2nE\left(\frac{\mu^2}{\sigma^2}|y, M_1\right) , \\
D_2^S &= \frac{\eta_1+2}{\eta_1\psi_1^2}\sum_{i=1}^{n}(y_i-\alpha)^2 - n\log\left(\frac{\eta_1\psi_1^2}{\eta_1-2}\right) + 2nE(\log\tau^2|y, M_2) , \\
D_3^S &= n\log\xi^2 + \frac{1}{\xi^2}\left[2n\delta_1^2 + \sum_{i=1}^{n}(y_i-\gamma_1)^2\right]
\end{aligned}$$

*with $E(\log\sigma^2|y, M_1)$, $E(\mu/\sigma^2|y, M_1)$, $E(\mu^2/\sigma^2|y, M_1)$ and $E(\log\tau^2|y, M_2)$ numerically computed. The term $n\log(2\pi)$ was removed from the DIC since it appears in all $D_j^S$, for $j = 1, 2, 3$.*

**Example 7.5** *Table 7.6 contains $n = 100$ cycles-to-failure times for airplane yarns. For each individual airplane, it has been suggested that an exponential model fits the data well (Leonard and Hsu, 1999; Quesenberry and Kent, 1982). The following Gamma, log-normal and Weibull models*

*are compared:*

$$\begin{aligned}
M_1 &: \; y_i \sim G(\alpha,\beta), && \alpha,\beta > 0 \\
M_2 &: \; y_i \sim LN(\mu,\sigma^2), && \mu \in R, \sigma^2 > 0 \\
M_3 &: \; y_i \sim Weibull(\gamma,\delta) && \gamma,\delta > 0,
\end{aligned}$$

*for $i = 1, \ldots, n$. Under model $M_2$, $\mu$ and $\sigma^2$ are the mean and the variance of $\log\ y_i$, respectively. Under model $M_3$, $p(y_i|\gamma,\delta) = \gamma y_i^{\gamma-1}\delta^{-\gamma}e^{-(y_i/\delta)^\gamma}$. Flat priors were considered for $\theta_1 = (\log\alpha, \log\beta)$, $\theta_2 = (\mu, \log\sigma^2)$ and $\theta_3 = (\log\gamma, \log\delta)$. It is also easy to see that, under model $M_1$, $E(y|\theta_1) = \alpha/\beta$ and $V(y|\theta_1) = \alpha/\beta^2$. Similarly, under model $M_2$, $E(y|\theta_2) = \exp\{\mu+0.5\sigma^2\}$ and $V(y|\theta_2) = \exp\{2\mu+\sigma^2\}(e^{\sigma^2}-1)$, while under model $M_3$, $E(y|\theta_3) = \delta\Gamma(1/\gamma)/\gamma$ and $V(y|\theta_3) = \delta^2\left[2\Gamma(2/\gamma) - \Gamma(1/\gamma)^2/\gamma\right]/\gamma$. These results will be used to compute $D^G$ below.*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 86 | 146 | 251 | 653 | 98 | 249 | 400 | 292 | 131 | 169 | 175 | 176 | 76 |
| 264 | 15 | 364 | 195 | 262 | 88 | 264 | 157 | 220 | 42 | 321 | 180 | 198 |
| 38 | 20 | 61 | 121 | 282 | 224 | 149 | 180 | 325 | 250 | 196 | 90 | 229 |
| 166 | 38 | 337 | 65 | 151 | 341 | 40 | 40 | 135 | 597 | 246 | 211 | 180 |
| 93 | 315 | 353 | 571 | 124 | 279 | 81 | 186 | 497 | 182 | 423 | 185 | 229 |
| 400 | 338 | 290 | 398 | 71 | 246 | 185 | 188 | 568 | 55 | 55 | 61 | 244 |
| 20 | 284 | 393 | 396 | 203 | 829 | 239 | 236 | 286 | 194 | 277 | 143 | 198 |
| 264 | 105 | 203 | 124 | 137 | 135 | 350 | 193 | 188 | | | | |

Table 7.6 *One hundred cycles-to-failure times for samples of yarn airplanes.*

*Weighted resampling schemes, with bivariate normal importance functions, were used to sample from the posterior distributions. For i=1,2,3, the proposals are $q_i(\theta_i) = f_N(\theta_i; \tilde{\theta}_i, V_i)$, where $\tilde{\theta}_1 = (0.15, 0.2)'$, $\tilde{\theta}_2 = (5.16, -0.26)'$, $\tilde{\theta}_3 = (0.47, 5.51)'$, $V_1 = diag\ (0.15, 0.2)$, $V_2 = diag\ (0.087, 0.085)$ and $V_3 = diag\ (0.087, 0.101)$. Posterior draws and contours of the posterior distributions appear in Figure 7.3. Under model $M_1$, the posterior means, posterior standard deviations and 95% posterior credibility intervals for $\alpha$ and $\beta$ are $2.24, 0.21, (1.84, 2.68)$ and $0.01, 0.001, (0.008, 0.012)$, respectively. Under model $M_2$, the posterior means, posterior standard deviations and 95% posterior credibility intervals for $\mu$ and $\sigma^2$ are $5.16, 0.06, (5.05, 5.27)$ and $0.77, 0.04, (0.69, 0.86)$, respectively. Under model $M_3$, posterior means, posterior standard deviations and 95% posterior credibility intervals for $\gamma$ and $\delta$ are $1.60, 0.09, (1.42, 1.79)$ and $248.71, 13.88, (222.47, 276.62)$, respectively.*

*Model comparison criteria appear in Table 7.7. The criteria indicate that both the Gamma and the Weibull models are relatively similar with the Weibull model performing slightly better. Similar conclusions appear in Figure 7.3(d), where the posterior predictive densities for both Gamma and Weibull models are almost identical.*

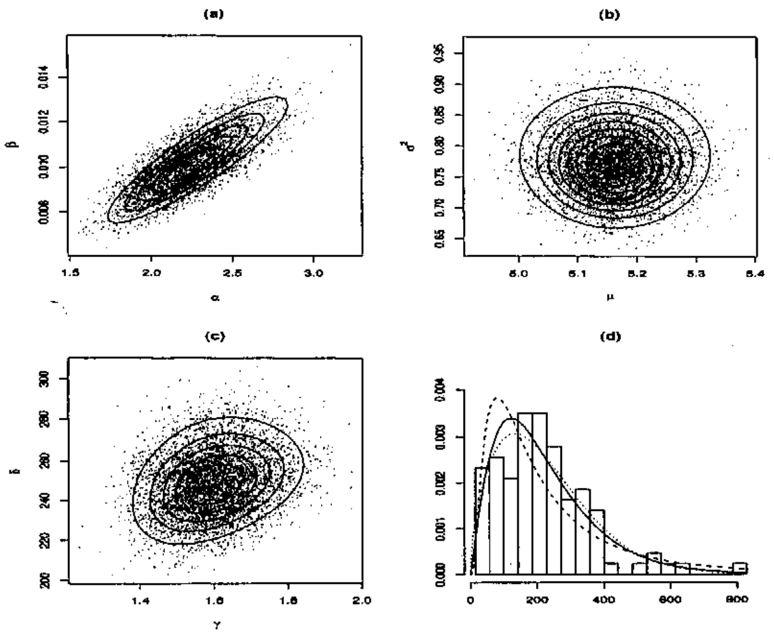

Figure 7.3 *Yarns failure data. Posterior draws and true contours: (a)* $\pi(\alpha, \beta|M_1)$, *(b)* $\pi(\mu, \sigma^2|M_2)$ *and (c)* $\pi(\gamma, \delta|M_3)$. *Panel (d) exhibits the histogram of the data along with approximations for the posterior predictive density* $p(y_{n+1}|y)$ *for the Gamma model (solid line), Lognormal model (dashed line) and Weibull (dotted line).*

| *Model* | *AIC* | *BIC* | $G_4$ | $D^G$ | $D^S$ |
|---|---|---|---|---|---|
| $M_1$ | *1254.489* | *1259.699* | *0.2257* | *1308.370* | *1253.445* |
| $M_2$ | *1267.520* | *1272.731* | *0.2212* | *2268.808* | *1265.842* |
| $M_3$ | *1254.398* | *1259.608* | *0.2232* | *1253.051* | *1253.051* |

Table 7.7 *Model comparison for the Yarns failure data. For* $i = 1, 2, 3$, *Akaike's information criterion (AIC) and Schwarz information criterion (BIC) are* $-2 \log f(y|\hat{\theta}_i) + 2d_i$ *and* $-2 \log f(y|\hat{\theta}_i) + d_i \log n$, *respectively, where* $\hat{\theta}_i$ *is the maximum likelihood estimate of* $\theta_i$. $D^G$*s were normalized so that* $D^G_3 = D^S_3$. *The actual value of* $D^G_3$ *is 4132798.*

## 7.3 Model choice: MCMC over model and parameter spaces

Results from the previous section deal with evaluation of a given model. Comparisons of different models were also contemplated but always within the context where calculations were made separately for each model. In this section, a more formal treatment is given to the problem of choosing between models. This is done by indexing all models under consideration and treating this index as another parameter, to be treated jointly with all other model parameters.

Essentially, two alternative approaches are presented. The first was introduced by Carlin and Chib (1995) and considers all models in a formation, called here a *supermodel*. The Markov chain simulation scheme for this supermodel is presented below. The second approach presents sophisticated simulation techniques using Markov chain with jumps between the different models (Green, 1995).

It will be assumed throughout this section that $y$ is observed and it can be described according to a model $M_j$ with parameter $\theta_j$ of dimension $d_j$ taking values in a parameter space $\Theta_j \subset R^{d_j}$ and $j$ taking values in a countable set $\mathcal{J}$. When $\mathcal{J}$ is finite, it can by identified as $\mathcal{J} = \{1, \ldots, J\}$.

A superparameter $\theta = (\theta_j, j \in \mathcal{J})$ taking values in a parameter space $\Theta = \Theta_1 \times \Theta_2 \cdots$ and a quantity $M$ assuming values in $\mathcal{J}$ can be defined. $M$ serves the purpose of indicating a specific model. For instance, $M = j$ indicates that model $M_j$ is being considered. It is implicitly assumed above that different models do not share any component of their respective parameters. This restrictive assumption is satisfied by many practical applications. Common components can also be included in this formulation with repetition of the components in all $\theta_j$ that contain them.

Assume for the moment that the posterior distribution $\pi(\theta, j)$, the joint distribution of the superparameter and the model indicator, is to be obtained. However, the main interest in inference is to obtain the posterior distributions of $\theta_j|M = j$, $j \in \mathcal{J}$, and of $M$. These distributions respectively provide the posterior inference within each of the models and the posterior probabilities of the models. The joint posterior is useful when dealing with a component of the parameter, say $\phi$, that is shared by all models. Inference about $\phi$ is based on its marginal posterior distribution $\pi(\phi)$. In any case, it can all be obtained from the joint density-probability function $\pi(\theta, j)$. The supermodel approach provides a sample from this more general, perhaps unnecessary posterior distribution whereas the approach with jumps only provides samples from $\theta_j|M = j$, $j \in \mathcal{J}$, and from $M$. The presence of common parameters (for instance, the asymptotes in Example 7.3) does not pose any problem here. Samples from them are naturally obtained from the corresponding samples for all models that include this shared component.

### 7.3.1 *Markov chain for supermodels*

The joint probability model for the superparameter vector $\theta$ and $M$ is defined over $\Theta \times \mathcal{J}$, such that the joint distribution of all random quantities is given by

$$p(y, \theta, j) = f(y|\theta, j) p(\theta|j) p_j$$

where $j$ is the value of $M$ and $p_j = Pr(M = j)$. Godsill (2001) named this the *composite model* and $\Theta \times \mathcal{J}$ the composite model space. Given that $M = j$, the distribution of $y$ depends on $\theta$ only through $\theta_j$, or mathematically,

$$f(y|\theta, j) = f(y|\theta_j, j) \ .$$

Assume also that the $\theta_j$ are conditionally independent given the value of $M$. Hence,

$$p(\theta|j) = \prod_{i \in \mathcal{J}} p(\theta_i|j) \ . \qquad (7.15)$$

Note that the prior distribution $p(\theta_i|j)$, for $i \neq j$, does not make much sense. It specifies the distribution of the parameters of model $i$ conditioned on the fact that this is not the true model. Carlin and Chib (1995) referred to these as *pseudo prior* or linking distributions. Due to the conditional independence (7.15), these priors do not interfere in the expressions of the marginal predictive densities for each model. Nevertheless, they are relevant for the construction of the chain and must be specified.

The full posterior of the supermodel is, therefore,

$$\pi(\theta, j) = \frac{f(y|\theta_j, j) \prod_{i \in \mathcal{J}} p(\theta_i|j) p_j}{f(y)} \ .$$

Assume that $(\theta^{(1)}, j^{(1)}), \ldots, (\theta^{(n)}, j^{(n)})$ is a MCMC sample from the supermodel $\pi(\theta, j)$ obtained by one of the algorithms about to be described. Comparison between models is based on the marginal posterior distribution of $M$, $\pi(j)$, $j \in \mathcal{J}$. These probabilities are estimated by the proportion of values of $M$ equal to $j$ in the sample of size $n$, i.e.,

$$\hat{\pi}(j) = \frac{1}{n} \sum_{l=1}^{n} I(j^{(l)} = j). \qquad (7.16)$$

Inference within each model is based on the conditional posterior distribution $\pi(\theta_j|j)$, $j \in \mathcal{J}$. Note that the sample available for $\theta_j$ is a sample from the marginal posterior $\pi(\theta_j)$. A sample from the conditional posterior is obtained by retaining the $\theta_j^{(l)}$, $l = 1, \ldots, n$, of the posterior sample for which the sampled value for $M$ was $j$ and discarding all other values of $\theta_j^{(l)}$ for which the sampled value for $M$ was not $j$. For instance, $E[g(\theta_j)]$ can be approximated by

$$\tilde{g}_j = \frac{1}{n} \sum_{l=1}^{n} g(\theta_j^{(l)}) I(j^{(l)} = j),$$

where the expectation is taken with respect to $\pi(\theta_j|j)$.

#### *Carlin and Chib sampler*

Carlin and Chib (1995) explored the natural blocking formed by grouping each model parameters and $M$ when $\mathcal{J}$ is a finite set of models. The full conditional distributions for $\theta_1, \ldots, \theta_J$ and $M$ are obtained as follows:

- For block $\theta_j$, $j = 1, \ldots, J$,

$$\pi_j(\theta_j) \propto \begin{cases} f(y|\theta_j, j) p(\theta_j|j) & , \text{ for } M = j \\ p(\theta_j|i) & , \text{ for } M = i \neq j \end{cases} .$$

- For block $M$

$$\pi_M(j) = k^{-1} f(y|\theta_j, j) \prod_{i=1}^{J} p(\theta_i|j) p_j \ , \ j = 1, \ldots, J$$

that is a discrete distribution with proportionality constant

$$k = \sum_{l=1}^{J} f(y|\theta_l, l) \prod_{i=1}^{J} p(\theta_i|l) p_l.$$

$M$ can always be sampled directly because it has a discrete distribution. Direct sampling from blocks $\theta_j$ will depend on the conjugacy structure for model $M = j$ and the form of the pseudo prior distributions. When direct sampling for some of the $\theta_j$s is not possible, Metropolis-Hastings steps described in Chapter 6 may be used.

The above scheme satisfies the conditions of a conventional Markov chain and therefore converges to the target distribution given by the posterior $\pi(\theta, j)$. Random draws from this distribution may be generated by iterative sampling from the full conditional distributions described above. The end result of this process is a sample of size $n$, say $\theta_1^{(l)}, \ldots, \theta_J^{(l)}, j^{(l)}$, $l = 1, \ldots, n$.

The pseudo prior distributions must be carefully chosen as they affect the rate of convergence of the chain. Note that as a modelling device they are meaningless. Therefore, Carlin and Chib (1995) supported the view that total freedom may be given to the specification of these prior distributions. They may even include specifications using the data. After experimenting with a few choices, they recommend setting the pseudo prior distributions $p(\theta_j|i)$, for $i \neq j$ as close as possible to $\pi(\theta_j|j)$. They suggested the use of simple standard approximations based on univariate estimates obtained from pilot (model-specific) chains.

Another difficulty encountered by Carlin and Chib (1995) is connected to the prior model probabilities $p_j$. They observed that for some prior specifications, the chain does not seem to move between models and, as a result, large posterior probabilities are obtained for some of the models. To correct that, they set the prior probabilities in their examples to values that allow movement between the models. Although this may work in practice, it can force practitioners to specify probabilities they may not believe. Their example contained fairly vague prior distributions for models with different dimensions and Bayes factors are known to be very sensitive in these situations. So, this prior setting may need further justification to satisfy potential users.

*Metropolised Carlin and Chib sampler*

The above algorithm is not applicable in the above form when a large or countable number of models is considered. Dellaportas, Forster and Ntzoufras (2002) and Godsill (2001) noticed that its major drawback comes from the calculation of $f(y|\theta_l, l)\prod_{i\in\mathcal{J}} p(\theta_i|l)p_l$ for all $l \in \mathcal{J}$. They suggested replacing the Gibbs sampling step for the model by a Metropolis-type step. Therefore, they *metropolised* the Carlin and Chib algorithm. This idea was previously illustrated in the fixed-dimensional case presented in Example 6.5 where it was shown that replacing a Gibbs step by a Metropolis step proved computationally advantageous.

More precisely, if the chain is currently at $(\theta, j)$, a move to a new pair $(\phi, k)$ is made as follows:

- A new model $k$ is sampled from proposal $q(j,k)$;
- $\phi_k$ is sampled from the pseudo prior $p(\theta_k|j)$. Thus, only the component $k$ of $\theta$ is proposed to move, i.e., $\phi = (\theta_1, \ldots, \theta_{k-1}, \phi_k, \theta_{k+1}, \ldots)$;
- Accept the move to $(\phi, k)$ with probability $\alpha$ where

$$\alpha = \min\left\{1, \frac{f(y|\phi_k, k)\prod_{i=1}^{J} p(\phi_i|k)p_k}{f(y|\theta_j, j)\prod_{i=1}^{J} p(\theta_i|j)p_j} \times \frac{q(k,j)p(\phi_i|k)}{q(j,k)p(\theta_k|j)}\right\}$$

which simplifies to

$$\alpha = \min\left\{1, \frac{\pi(\phi_k|k)p_k}{\pi(\theta_j|j)p_j} \times \frac{q(k,j)p(\phi_j|k)}{q(j,k)p(\theta_k|j)}\right\} \qquad (7.17)$$

since all other pseudo-prior densities cancel out.

**Example 7.2** *(continued) Lopes and West (2004) proposed a sampler for traditional factor analysis that relies upon MCMC-based approximations to the posterior distributions of $(\beta_j, \Sigma_j)$ under the $j$-factor model. In their case $\theta_j = (\tilde{\beta}_j, \sigma_j^2)$ where now $\tilde{\beta}_j$ is the vector obtained by stacking the non-zero elements of the columns of the factor loadings matrix $\beta_j$ and $\sigma_j^2$ is the vector comprising the diagonal of $\Sigma_j$. Therefore, $\bar{\beta}_j$ is a $[9j - j(j-1)/2]$-dimensional vector and $\sigma_j^2$ is a 9-dimensional vector.*

*Their proposal density is $q((\theta_j, j), (\phi_k, k)) = q(j,k)q_k(\phi_k)$, where $q(1,2) = q(5,4) = 1$, $q(l, l+1) = q(l, l-1) = 0.5$, for $l = 2, 3, 4$, and $q_k(\phi_k) = f_N(\tilde{\beta}_k; b_k, \beta V_k)\prod_{l=1}^{k} f_{IG}(\sigma_{kl}^2; \alpha, \alpha\nu_{kl}^2)$. The quantities $b_k$ and $V_k$ are estimates of the posterior mean and variance of $\tilde{\beta}_k$, while $\nu_{kl}^2$ is an estimate of the posterior mean of $\sigma_{kl}^2$, all under the k-factor model. The tuning parameters $\alpha$ and $\beta$ are problem specific and were set at $\alpha = 18$ and $\beta = 2$ in their simulated study. The sampler was based on 1000 iterations, after a burn-in of 10000 draws. 1000 replications of the 3-factor model were simulated. The MCMC algorithms were run and used to identify the model with highest posterior probability for each replication. In 99.3% of the replications, it chose the correct model ($k = 3$), while ($k = 2$) was chosen only in 0.7% of the replications. Results based on other model criteria appeared in Table 7.3. Further discussion and additional simulated and real data analysis appear in Lopes and West (2004).*

In the Metropolised Carlin and Chib algorithm, a new model $k$ is proposed according to the transition kernel $q(j,k)$ and then the parameters of model $k$ are proposed from the pseudo prior $p(\theta_k|j)$. Natural extensions for the above algorithm can be considered. A new model $k$ can be proposed according to a transition depending also on the parameter $\theta_j$ and the parameter $\theta_k$ can be proposed according to a general transition kernel that depends on $j$ and $\theta_j$.

### *7.3.2 Markov chain with jumps*

An alternative route for general construction of Markovian processes with a given stationary distribution is based on the specification of a transition kernel satisfying the detailed balance equation

$$\pi(\theta, j)p((\theta, j), (\phi, k)) = \pi(\phi, k)p((\phi, k), (\theta, j))$$

valid for all points where the move is allowed. This equation ensures reversibility of the chain. This is a sufficient condition and hence imposes more restrictions than necessary for convergence. Nevertheless, it provides a useful basis for specification of appropriate transition kernels. This route was used by Hastings (1970) and described in Chapter 6. It led to many possibilities for the kernel and is again used here in the more general context of a collection of models.

Once again, the transition will be constructed in two stages: a proposal transition and an acceptance probability, correcting the proposal to ensure balance. The main difference here is that many models are simultaneously being considered and therefore many qualitatively different moves are entertained. Green (1995) explored this idea by imposing detailed balance for all possible moves between models. Detailed balance would then be attained globally and convergence to $\pi(\theta, j)$ would result.

Consider for each possible jump move $m$ between models, an arbitrary transition kernel $q_m((\theta, j), (\phi, k))$ and a yet to be specified acceptance probability $\alpha_m((\theta, j), (\phi, k))$. In fact, many different jump moves can be made from $M_j$ to $M_k$ but for simplicity it will be assumed here that each move $m$ uniquely defines a model $M_k$. Many of these moves can be specified and for each of them

$$B_{jm} = \int_\Theta q_m((\theta, j), (\phi, k))d\phi$$

can be defined. $B_{jm}$ gives the probability that the proposed move makes the chain jump from current model $M_j$ to model $M_k$ and may depend on $\theta$. In practice, this is typically not the case and dependence on $\theta$ is dropped for notational simplicity. If only one type of move is allowed from $j$ to $k$ then one can identify $B_{jm} = q(j, k)$ of the previous section. Note now that $q_m((\theta, j), (\cdot, k))/B_{jm}$ defines a proposal transition density for $\theta_k$ associated with jump move $m$.

Each proposed move takes the chain from model $j$ and a corresponding value of $\theta_j$ to model $k$ and a corresponding value of $\phi_k$. Only values of the parameters associated with the models involved in the jump are concerned. If the move is accepted, only the $k$th component of $\theta$ is altered and $\phi_{-k} = \theta_{-k}$. Note also that $k$ may be equal to $j$ and, in this case, the move does not involve a jump between models.

The extension considered by Green (1995) also admits that the chain may not move at every iteration so that $\sum_m B_{jm} = B_j \le 1$. Jump moves $m$ are proposed according to probabilities $B_{jm}$ and there is also a probability $1 - B_j$ that the transition does not propose any moves. Naturally, it is possible to have $B_j = 1$ and in this case a move will always be proposed.

Following (6.4), the transition kernel of the chain is given by

$$\begin{aligned} p((\theta, j), A) &= \sum_m \int_A q_m((\theta, j), (\phi, k))\alpha_m((\theta, j), (\phi, k))d\phi \\ &+ I((\theta, j) \in A)s(\theta, j) \end{aligned} \qquad (7.18)$$

where $A \subset \Theta \times \mathcal{J}$ and

$$\begin{aligned} s(\theta, j) &= 1 - \sum_m \int_\Theta q_m((\theta, j), (\phi, k))\alpha_m((\theta, j), (\phi, k))d\phi \\ &= \sum_m \int_\Theta q_m((\theta, j), (\phi, k))[1 - \alpha_m((\theta, j), (\phi, k))]d\phi \\ &+ 1 - B_j . \end{aligned} \qquad (7.19)$$

As for the Metropolis-Hastings algorithm presented in Chapter 6, the transition kernel (7.18) defines a mixed distribution for the next state of the chain. For points $(\phi, k) \neq (\theta, j)$, the distribution admits a combination of a density for $\phi_k$ and a probability function for $k$ given by $q_m((\theta,j),\cdot)\alpha_m((\theta,j),\cdot)$ where $m$ is the jump move taking the chain from model $M_j$ to model $M_k$. More specifically, there is a probability $B_{jm}$ that the proposed move takes the chain to model $M_k$. Given that the proposed move took the chain to model $M_k$, there is a density $g$ describing the chances of moves to $\phi_k$. Associated with the jump $m$ to model $M_k$, the superparameter can be partitioned as $\theta = (\theta_k, \theta_{-k})$. The density $g$ is given by

$$g(\phi_k) = q_m((\theta_k, \theta_{-k}, j), (\phi_k, \theta_{-k}, k))/B_{jm}. \qquad (7.20)$$

This density governs the proposed value for the parameter of model $k$. This value will then be accepted with probability $\alpha_m((\theta_k, \theta_{-k}, j), (\phi_k, \theta_{-k}, k))$. Note that both the proposal $q_m$ and the acceptance probability $\alpha_m$ depend on $\theta$ only through $\theta_k$; the remaining components are irrelevant once the jump defines model $M_k$ for the next state of the chain.

For $(\phi, k) = (\theta, j)$, this distribution provides a positive probability of no moves given by $s(\theta, j)$. Equation (7.19) informs that this probability is due to two distinct events: either the move took the chain to a state that was not accepted or no move was proposed.

The imposition of general reversibility of the process leads to

$$\alpha_m((\theta, j), (\phi, k)) = \min\left\{1, \frac{\pi(\phi, k)q_m((\phi, k), (\theta, j))}{\pi(\theta, j)q_m((\theta, j), (\phi, k))}\right\} . \qquad (7.21)$$

In practical terms, the simulation of a sample from $\pi$ using the Markov chain defined by transition (7.18) may be summarized as:

1. Initialize the iterations counter $l = 1$ and set as arbitrary an initial value $(\theta^{(0)}, j^{(0)})$.
2. Choose a jump move $m$ accordingly with probabilities $B_{jm}$, thus defining the model $M_k$ considered for move proposals. Hence, the proposed value for $j^{(l)}$ is $k$. If the move chosen was not to move, set $(\theta^{(l)}, j^{(l)}) = (\theta^{(l-1)}, j^{(l-1)})$ and go to step 5. This happens with probability $1 - B_j$.
3. Draw $\phi_k$ from the conditional density (7.20). Hence, the proposed new state for $\theta$ changes its $k$th component to $\phi_k$ while keeping $\phi_{-k} = \theta_{-k}^{(l-1)}$. Hence, all other components of $\theta$ are unchanged.
4. Calculate the acceptance probability $\alpha_m((\theta^{(l-1)}, j^{(l-1)}), (\phi, k))$ of the move given by (7.21). If the move is accepted, $(\theta^{(l)}, j^{(l)}) = (\phi, k)$. If the move is not accepted, $(\theta^{(l)}, j^{(l)}) = (\theta^{(l-1)}, j^{(l-1)})$ and the chain does not move.
5. Change counter from $l$ to $l + 1$ and return to step 2 until convergence is reached.

Steps 2 and 4 are operated after generation of two independent uniform random quantities $u_1$ and $u_2$. The first one is used to determine the model to which the chain will propose a move. This choice is made according to the discrete distribution with probabilities $B_{jm}$. The second random quantity determines the acceptance probability of the proposed move as before: if $u_2 \leq \alpha_m$, the move is accepted and if $u_2 > \alpha_m$ the move is not allowed. The above algorithm is generically known by the acronym RJMCMC, standing for *reversible jump* MCMC.

This process in fact generates a stream of values $\theta^{(l)}_{j^{(l)}}, j^{(l)}$, $l = 1, 2, \ldots$. At each iteration once the value of the model is drawn, only the parameter associated with this model is generated. Therefore, samples from $\theta_j|j$ are automatically provided by restricting attention to the chain values associated with value $j$ for the model. Likewise, samples from $M$ are provided by the marginal samples over all iterations. At each iteration, the chain can be complemented with the current values of the parameters for the other models. This forms a larger stream of values $\theta^{(l)}, j^{(l)}$, $l = 1, 2, \ldots$, as in the previous section. There is no harm, of course, in doing it even though this is in general irrelevant as most questions of interest are already answered by the smaller and variable dimension sequence.

In the case of $J$ models, $J + 1$ moves can be specified: $J - 1$ jumps to the other models, one move within a model and one absence of move. As $J$ may be large, this may imply too many alternative moves. In general, few moves are necessary. The important point is to ensure irreducibility of the chain and freedom of movement throughout the parameter space. Typically, this is achieved by only allowing moves between neighboring models. In the case of a special meaning associated with the ordering $1, 2, \ldots, J$ of the models, Phillips and Smith (1996) and Green (1995) suggested considering only jumps to models $j - 1$ and $j + 1$ from model $j$. Markov chains of this kind were studied in Chapter 4 where they were referred to as birth and death processes. Note also that in the presence of a single model with moves always being proposed, the above structure reduces to the Metropolis-Hastings algorithm described in the previous chapter.

The general algorithm introduced by Green (1995) also takes advantage of possible similarities amongst competing models when proposing model and parameter moves. The supermodel samplers could be thought of as globally proposing new models and model parameters. On the other hand, Green's algorithm focuses on local aspects (associated with a single model), which in turn introduces more flexibility into the search of the model space.

**Example 7.6** *Choosing between two competing models*

*Consider models $M_1$ and $M_2$ with parameters $\theta_1$ and $\theta_2$ of dimensions $d_1$ and $d_2$ satisfying $d_1 + n_1 = d_2$ where $n_1 > 0$. So, model $M_2$ has a parameter space $n_1$ dimensions smaller than model $M_1$. The reversibility condition requires an additional random quantity $u$ of dimension $n_1$ in order to define a bijection $\theta_2 = g(\theta_1, u)$, thus allowing moves in both directions with the same probability. There are many other ways to ensure reversibility discussed by Green (1995) but this is the simplest one. The transition consists of incorporating into the ratio test the information about this bijection and about the distribution $f$ of $u$ that may depend on $\theta_1$. The acceptance probability of the move from model 1 to model 2 is given by*

$$\min\left\{1, \frac{\pi(\theta_2, 2)B_{21}}{\pi(\theta_1, 1)B_{12}q(u|\theta_1)}\left|\frac{\partial g(\theta_1, u)}{\partial(\theta_1, u)}\right|\right\} \qquad (7.22)$$

*where $B_{ij}$ is the probability that a move from model $i$ to model $j$ is proposed, $q(u|\theta_1)$ is the conditional proposal density used to generate the additional quantity $u$ and $\frac{\partial g(\theta_1, u)}{\partial(\theta_1, u)}$ is the matrix of derivatives of the bijection $g(\theta_1, u)$. Likewise, the acceptance probability of the reverse move from model 2 to model 1 is given by*

$$\min\left\{1, \frac{\pi(\theta_1, 1)B_{12}q(u|\theta_1)}{\pi(\theta_2, 2)B_{21}}\left|\frac{\partial(\theta_1, u)}{\partial\theta_2}\right|\right\} .$$

*Details of this derivation are given by Green (1995).*

The extension to more than two models follows the same ideas with specifications of quantities $u_{ij}$ of appropriate dimensions to ensure bijections between the parameters of models $M_i$ and $M_j$. So, if $\theta_i$ ($\theta_j$) is the parameter of model $M_i$ ($M_j$) with dimension $d_i$ ($d_j$) and $d_j - d_i = n_{ij} > 0$, a $n_{ij}$-dimensional random quantity $u_{ij}$ is defined. Then, a deterministic bijection is created between $(\theta_i, u_{ij})$ and $\theta_j$. This simple device ensures that the moves between any two parameter spaces can be reversed. There is no theoretical restriction on the form of the proposal densities used to draw the $u_{ij}$ and of the bijection. Once again, they should lead to moves that are not too small (to render slow convergence) nor too large (to render low acceptance probability). Green (1995) applied this idea to models with multiple change points, image segmentation and partition models. He also discussed some useful forms of bijections and proposals in the context of his applications, as shown in Example 7.7 below.

**Example 7.7** *Non-parametric intensity rate estimation*

*Consider the observation of a Poisson process over $[0, T]$ with unknown intensity rate $\lambda(t)$, $0 \leq t \leq T$. Each of countably many possible models $M_k$, $k = 0, 1, 2, \ldots$, is defined as having a piecewise constant intensity rate taking values $\lambda_j$ at intervals $I_j = [t_j, t_{j+1})$, $j = 0, 1, \ldots, k$, with $t_0 = 0$ and $t_{k+1} = T$. The parameter of $M_k$ is $\theta_k = (\lambda(k), t(k))'$ where $\lambda(k) = (\lambda_0, \lambda_1, \ldots, \lambda_k)$ and $t(k) = (t_1, \ldots, t_k)$. The hierarchical prior used by Green (1995) assumed that, conditional on $k$, $t(k)$ consists of the even-numbered order statistics of a sample of size $2k + 1$ from the $U[0, T]$ distribution and $\lambda(k)$ is a sample from a $G(\alpha, \beta)$ distribution. The prior is completed with a second stage $k \sim Poi(\gamma)$ with probability function $f_P$. Arjas and Heikkinen (1997) replaced the conditional independence of the $\lambda(k)$ by a pairwise difference prior accounting for their spatial interaction as in Section 2.6.*

*Birth and death processes allow for four different types of moves in these circumstances:*

1. *The birth of a new step, thus creating a new interval and moving the chain from model $M_k$ to model $M_{k+1}$. This move has probability $p_k$.*
2. *The death of an existing step, thus deleting an existing interval and moving the chain from model $M_k$ to model $M_{k-1}$. This move has probability $q_k$.*
3. *The change of the intensity rate of a given interval. This move has probability $r_{1k}$.*
4. *The change of the length of a given interval. This move has probability $r_{2k}$.*

*Note that $p_k + q_k + r_{1k} + r_{2k} = 1$ ($q_0 = 0$) and that moves of type 3 and 4 retain model $M_k$. For these moves, assuming randomly chosen intervals for the change, evaluation of acceptance probabilities follows from standard theory and is left as an exercise.*

*A birth move changes model $M_k$ with parameter $\theta_k$ to model $M_{k+1}$ with parameter $\theta_{k+1}$. Green (1995) suggested starting the move by choosing a new endpoint $t^*$ uniformly over $[0,T]$ (and this point will lie in the interval $[t_j, t_{j+1})$, say) and choosing the corresponding new rates $\lambda'_j$ and $\lambda'_{j+1}$ according to $\lambda'_j/\lambda'_{j+1} = u/(1-u)$ where $u \sim U[0,1]$ but preserving the geometric average so that $(t_{j+1}-t^*)log\lambda'_{j+1} + (t^*-t_j)log\lambda'_j = (t_{j+1}-t_j)log\lambda_j$. Note that only two new variables were needed to create the two new parameters. Reversibility automatically defines the new intensity rate of the interval formed by the merging of two adjacent intervals in the death move. The endpoint to be deleted is chosen uniformly at random and the merged intensity rate should also satisfy the same geometric averaging.*

*Following (7.22), the acceptance probability of a birth move is*

$$\min\left\{1, \frac{\pi(\theta_{k+1}, k+1)}{\pi(\theta_k, k)} \frac{B_{k+1,k}}{B_{k,k+1}q(t^*,u)} \left|\frac{\partial\theta_{k+1}}{\partial(\theta_k, t^*, u)}\right|\right\} .$$

*Derivation of the expressions of each of the terms above and of the acceptance probability of a death move are left as exercises.*

Clyde (1999) showed that pre-existing algorithms, such as the MCMC model composition (Raftery, Madigan and Hoeting, 1997) and stochastic search variable selection (George and McCulloch, 1992), can be seen as particular cases of RJMCMC. Other earlier approaches to deal with variable dimension appear in Geyer and Moller (1994) and Ripley (1977) for spatial processes.

In many situations, there is a natural ordering of the component models, usually through the dimensions of their parameter spaces. So, simple chains based on irreducible birth and death processes can be used. The approach has proved to be useful when treating problems with many possible models such as mixtures with an unknown number of components (Richardson and Green, 1997; Lopes, Muller and Rosner, 2003), non-parametric intensity rate estimation (Green, 1995; Arjas and Heikkinen, 1997), non-parametric regression (Dias and Gamerman, 2002), nonlinear classification and regression (Dennison et al., 2002), autoregressive time series models (Huerta and West, 1999a,b; and Huerta and Lopes, 2000), graphical models (Dellaportas and Forster, 1999) and genetic mapping (Waagepetersen and Sorensen, 2001).

### *Proposing from full posterior conditionals*

When $\pi(\phi_k|k)$ is available in closed form and it is easy to sample from, the proposal transition can be written as $q((\theta_j, j), (\phi_k, k)) = \pi(\phi_k|k)q(j,k)$ and the acceptance probability (7.21) can be rewritten as

$$\alpha = \min\left\{1, \frac{\pi(k)}{\pi(j)} \frac{q(k,j)}{q(j,k)}\right\} . \qquad (7.23)$$

In other words, when posterior inference within model is obtained in closed form, the above MCMC scheme searches the space of models with moves essentially driven by posterior odds ratios. So, models with higher posterior model probability will be visited more often with model $M_j$, for instance, visited $100\pi(j)$% of the time.

Inference about $\theta_k$ is made conditional on $k$ using standard within model MCMC algorithms. This result appears, for instance, in MCMC model combination (MC$^3$) for graphical models (Madigan and York, 1995). Even though $\pi(\phi_k|k)$ is only rarely known and/or easy to sample from, the above results suggest that good proposal densities $q$ should mimic the full posterior conditional densities. This recommendation is in line with those previously made in Chapter 6.

### *Partial analytic structure*

Even though direct evaluation and/or sampling from $\pi(\theta_j|j)$ may be unrealistic in most situations, there are several instances where $\pi(\theta_{j2}|\theta_{j1}, j)$ is available in closed form, for some partition $(\theta_{j1}, \theta_{j2})$ of $\theta_j$ under model $M_j$. Suppose that, in addition, there exists a similar partition of $\phi_k$, $(\phi_{k1}, \phi_{k2})$, such that the dimensions of $\theta_{j1}$ and $\phi_{k1}$ are identical. When current and proposed models share common parameters, Godsill (2001) suggests an MCMC algorithm that proposes a move from $(\theta, j)$ to $(\phi, k)$ by setting $\phi_{k1} = \theta_{j1}$, sampling $\phi_{k2}$ from its full conditional $\pi(\phi_{k2}|\phi_{k1}, k)$ and accepting the move with probability

$$\alpha = \min\left\{1, \frac{\pi_k(k|\phi_{k2})q(j,k)}{\pi_j(j|\theta_{j2})q(k,j)}\right\},$$

where $\pi_l(l|\theta_{l2}) = \int \pi(l, \theta_{l1}|\theta_{l2}) d\theta_{l1}$, for $l = j, k$. After that, the common component $\theta_{j1}$ say, is sampled from its full conditional $\pi(\theta_{j1}|\theta_{j2}, j)$.

**Example 7.3** *(continued) $\theta_1 = \phi_1$ represents the asymptote in both models but $\theta_{-1}$ and $\phi_{-1}$ have different meanings under each model. Assume now that proper prior distributions are considered. Additionally, assume that the prior distribution of $\theta_1$ and $\phi_1$ are both $N(m_0, C_0)$ and independent of the other parameters. It is easy to verify that the full conditional distributions of $\theta_1$ and $\phi_1$ are normal. Therefore, the implementation of the above partial analytic structure is straightforward. Suppose that the current model is the logistic model, for instance, with parameter value $\theta$ and the proposed model is the Gompertz model, then $(\phi_2, \phi_3, \phi_4)$ is sampled from its full conditional distribution. The move is accepted with probability*

$$\alpha = min\left\{1, \frac{f(y|\phi_2, \phi_3, \phi_4)p_2}{f(y|\theta_2, \theta_3, \theta_4)p_1} \frac{q(2,1)}{q(1,2)}\right\}$$

*where $f(y|\theta_2, \theta_3, \theta_4) = f_N(y; xm_0, \theta_4^2 I_n + xx'C_0)$ and $f(y|\phi_2, \phi_3, \phi_4) = f_N(y; 1_n m_0 + z, \phi_4^2 + 1_n 1_n' C_0)$, for $x = (x_1, \ldots, x_n)'$, $z = (z_1, \ldots, z_n)'$, $x_t = (1 + \theta_2 \theta_3^t)^{-1}$ and $z_t = \exp\{\phi_2 \phi_3^t\}$, for $t = 1, \ldots, n$. If the move is (not) accepted then $\phi_1$ ($\theta_1$) is sampled from a standard MCMC algorithm under the Gompertz (logistic) model.*

**Example 7.8** *Lopes and Salazar (2006a,b) revisited the Canadian lynx data (Figure 7.4). They designed RJMCMC algorithms to account for model uncertainty in logistic smooth transition autoregressive models of order $j$ ($M_j$)*

$$y_t|\theta_j \sim N(x_t'\beta_{j1} + g(\gamma_j, c_j, s_t)x_t'\beta_{j2}, \sigma_j^2)$$

*where $x_t = (y_{t-1}, \ldots, y_{t-j})'$ and $\beta_{j1}$ and $\beta_{j2}$ are $j$-dimensional vectors of regression coefficients, so that $\theta_j = (\theta_{j1}, \theta_{j2})$, $\theta_{j1} = (\gamma_j, c_j, \sigma_j^2)$ and $\theta_{j2} = (\beta_{j1}, \beta_{j2})$. The function $g_j = g(\gamma_j, c_j, s_t) = (1 + e^{-\gamma(s_t - c_j)})^{-1}$ plays the role of a smooth transition continuous function bounded between* 0 *and* 1*. When $g_j = 0$, the above model reduces to a simple linear autoregressive model. $s_t$ is called the transition variable, with $s_t = y_{t-d}$ a common choice. One of the prior specification they used is*

$$p(\theta_j) \propto f_N(\beta_{j2}; 0, \sigma_j^2 e^\gamma I_{j+1})(1 + \gamma_j^2)^{-1} \sigma_j^{-2}$$

*for $c_j \in [\hat{F}^{-1}(0.15), \hat{F}^{-1}(0.85)]$ and $\hat{F}$ the empirical cumulative distribution function of the data. They showed that the model acceptance probability becomes*

$$\alpha = \min\left\{1, \frac{f(y|k, \theta_{k1})}{f(y|j, \theta_{j1})} \frac{q(k,j)}{q(j,k)}\right\}$$

*where*

$$f(y|l, \theta_{l1}) \propto \left[\frac{2\pi}{\sigma_l^2 \exp(\gamma_l)}\right]^{\frac{l+1}{2}} |C_l|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\sigma_l^{-2} y'y - m_l' C_l m_l)\right\}$$

*and*

$$\begin{aligned} m_l &= C_l \sigma_l^{-2} \sum z_t y_t \ , \ \ C_l = (\sigma_l^{-2} \sum z_t z_t' + \Sigma^{-1}) \\ z_t' &= (x_t', g_l x_t') \ , \ \ \Sigma^{-1} = diag(0, \sigma_l^{-2} e^{-\gamma_l} I_{l+1}), \end{aligned}$$

*for $l = j, k$. The MCMC algorithm is completed by sampling $\theta_{k2}$ conditional on $\theta_{k1}$, $k$ and $y$. Table 7.8 presents the posterior model probabilities obtained by their MCMC algorithm. The modal model is an AR(11) model with $Pr(M_{11}|y) = 0.56$. Nonetheless, AR(3), AR(4), AR(12) and AR(13) models also exhibit non-negligible posterior model probabilities.*

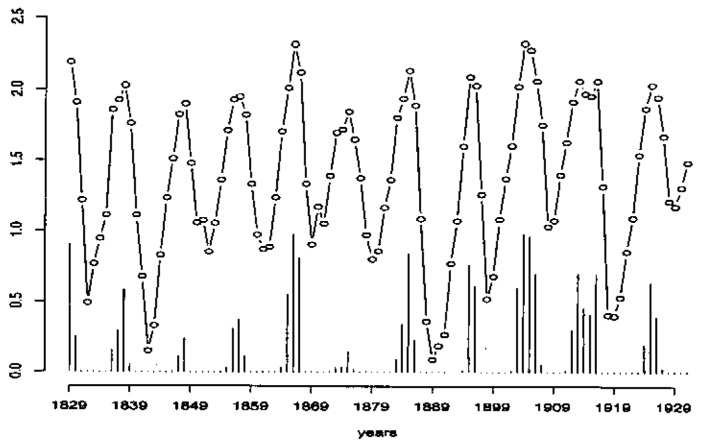

Figure 7.4 *Logarithmic transformation of the number of Canadian lynx trapped in the Mackenzie River district of Northwest Canada over the period from* 1821 *to* 1934*. The vertical bars represent the posterior mean of $g(\gamma, c, y_{t-d})$.*

| *k* | *1* | *2* | *3* | *4* | *5* | *6* | *7* |
|---|---|---|---|---|---|---|---|
| *Pr(k\|y)* | *0.000* | *0.001* | *0.066* | *0.073* | *0.013* | *0.001* | *0.002* |
| *k* | *8* | *9* | *10* | *11* | *12* | *13* | |
| *Pr(k\|y)* | *0.000* | *0.001* | *0.000* | *0.559* | *0.162* | *0.121* | |

Table 7.8 *Canadian Lynx: posterior model probabilities.*

### 7.3.3 Further issues related to RJMCMC algorithms

*Convergence diagnostics*

Despite the broad use of reversible jump MCMC algorithms since Green's (1995) seminal paper, literature on assessing convergence is still scarce. Brooks and Giudici (1999, 2000) generalized the convergence diagnostics of Gelman and Rubin (1992) and Brooks and Gelman (1998) by splitting the total variation of parallel chains in two major components: (i) variation between chains and (ii) variation between models. In order to adapt Gelman and Rubin's convergence diagnostic, a subset of parameters, say $\theta_0$, must keep the same interpretation across models. In such case, which appears, for instance, when models are nested, they introduced a two-way ANOVA decomposition of the variance of $\theta_0$. See also Castelloe and Zimmerman (2002) who proposed unbalanced two-way ANOVA and multivariate ANOVA versions of the diagnostics of Brooks and Giudici (1999,2000).

Another idea, suggested by Brooks, Giudici and Philippe (2003), is to monitor the model indicator $M_k$ through straightforward implementation of nonparametric hypothesis tests. They applied their convergence diagnostic to mixture and graphical Gaussian models. In principle two drawbacks are apparent. First, care must be taken when assessing convergence based on marginal convergence diagnostics, as opposed to assessing for $(\theta, j)$ jointly. Second, it may be restrictive to use univariate test statistics to assess convergence of complex chains over highly dimensional parameter and model spaces.

Regardless of the scarcity of research in this area and the limitations of the proposed diagnostics, RJMCMC algorithms are being routinely used and they heavily depend on reliable convergence assessment. This area will benefit from further research.

*Choice of the proposal*

Green (2003) introduced a transdimensional version of the random walk Metropolis algorithm. He suggested running pilot chains within each model separately and use the draws to estimate the posterior moments required for his extension. An independence Metropolis version of this algorithm was presented by Lopes (2000) and Lopes and West (2004) (see Example 7.2 for details). This line of approach is practically limited to small or moderately small model sets $\mathcal{J}$. Further ideas are presented and discussed in Brooks, Giudici and Roberts (2003). See also Ehlers and Brooks (2002) for efficient construction of RJMCMC proposal densities for autoregressive models.

*RJMCMC and other transdimensional algorithms*

A similar treatment to the problem of inference about parameters in different models and model choice was given by Phillips and Smith (1996) based on jump diffusion processes (Grenander and Miller, 1994). In these processes, the chain moves within a model following a diffusion process, namely a Markovian process at continuous time. Sampling from this process requires a discretization of time and consequent approximation of the stochastic differential equation governing the process by a difference equation similar to the one used to define a system equation in dynamic models. A jump process is superimposed onto this diffusion to allow for moves between models. The jumps occur according to the marginal jump intensity. This intensity depends on the state $(\theta, j)$ of the chain and is obtained by the integration of the jump intensity $q((\theta, j), (\phi, k))$ over all possible jump points $(\phi, k)$. As before, this intensity $q$ is constructed so as to ensure reversibility. Many possibilities for $q$ based on Gibbs samplers and Metropolis-Hastings algorithms are presented by Phillips and Smith (1996) and illustrated in the context of identification of mixture components, object recognition, variable selection and identification of change points.

Stephens (2000) proposed and applied a discrete birth and death process for model search in mixture models. Later, Cappé, Robert and Rydén (2003) showed that Stephen's processes and Green's RJMCMC are quite similar when continuous-time processes are considered.

Sisson (2005) presented a comprehensive review of RJMCMC and lists up to date references and URL for several freely available software implementing RJMCMC and other transdimensional algorithms. See also Waagepetersen and Sorensen (2001) and Green (2003) for further review on RJMCMC.

## 7.4 Convergence acceleration

Previous sections dealing with techniques for improving the convergence of the chain basically considered them in the context of blocking parameters and reparametrization. There are many other suggestions for improving the convergence of the chain to the equilibrium distribution. Some of these techniques are described in this section. For presentation purposes they are divided in two large groups: alterations in the chain and alterations in the equilibrium distribution.

The chain may be altered by specification of alternative transition kernels with improved convergence properties or by manipulation of the draws generated from the chain. In both cases, the goal is the same: to make the chain get to the equilibrium faster.

### 7.4.1 Alterations to the chain

One of the greatest problems for the convergence of the chain is the fact that it moves the components along the directions determined by the components of $\theta$. One alteration of this scheme is reparametrization, which essentially promotes a transformation of the parameter space and can be seen as a redefinition of the sampling axes.

Rather than seeking useful reparametrization that will then define more appropriate directions for sampling, new directions can be suggested directly. Schmeiser and Chen (1991) proposed a chain where the direction $e$ of the moves is chosen at random from the unit vectors in $R^d$ where $d$ is the dimension of $\theta$. Once a direction is chosen, the next state of the chain is chosen in that direction according to the probabilities given by $\pi$. The complete scheme is:

1. Initialize the iteration counter of the chain $j = 1$ and set an initial value $\theta^{(0)}$.
2. Choose a direction $e^{(j)}$ in $R^d$ uniformly on $\{(x_1, \ldots, x_d) : \sum_{i=1}^d x_i^2 = 1\}$.
3. Choose a scalar $c^{(j)}$ generated from the density $g(c) \propto \pi(\theta^{(j-1)} + ce^{(j)})$ and set $\theta^{(j)} = \theta^{(j-1)} + c^{(j)}e^{(j)}$.
4. Change counter from $j$ to $j+1$ and return to step 2 until convergence is reached.

The generation of $e$ can be made univariate by drawing $r_i$ from some distribution symmetric around 0 and taking $e_i = r_i / \sum_j r_j^2$, $i = 1, \ldots, d$. Once the direction $e$ is chosen, the amount $c$ of movement along $e$ is chosen according to the posterior distribution. No matter how complicated is the form of the posterior, the generation of $c$ is univariate.

**Example 7.9** *Assume that the posterior distribution of interest is a bivariate mixture of normals $wN(\mu_1, \Sigma) + (1-w)N(\mu_2, \Sigma)$ where $0 < w < 1$ and*

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \; i = 1, 2, \; \textit{and } \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

*At iteration $j$ and given direction $e^{(j)} = (e_1, e_2)'$, the generating density for $c^{(j)}$ is*

$$\begin{aligned} g(c) \propto\ & w \exp\left\{-\frac{1}{2}(\theta^{(j-1)} + ce^{(j)} - \mu_1)'\Sigma^{-1}(\theta^{(j-1)} + ce^{(j)} - \mu_1)\right\} \\ & + (1-w) \exp\left\{-\frac{1}{2}(\theta^{(j-1)} + ce^{(j)} - \mu_2)'\Sigma^{-1}(\theta^{(j-1)} + ce^{(j)} - \mu_2)\right\} \\ \propto\ & w \exp\left\{-\frac{1}{2}(rc^2 - 2cs_1 + t_1)\right\} + (1-w) \exp\left\{-\frac{1}{2}(rc^2 - 2cs_2 + t_2)\right\} \end{aligned}$$

*where $r = e^{(j)'}\Sigma^{-1}e^{(j)}$, $s_i = e^{(j)'}\Sigma^{-1}(\theta^{(j-1)} - \mu_1)$ and $t_i = (\theta^{(j-1)} - \mu_i)'\Sigma^{-1}(\theta^{(j-1)} - \mu_i)$, $i = 1, 2$. It is clear that $g$ is the density of a mixture of normal distributions with means $s_i/r$, $i = 1, 2$, weights $w_1$ and $1 - w_1$ where $w_1 = we^{(s_1^2/2r)-(t_1/2)}/[we^{(s_1^2/2r)-(t_1/2)} + (1-w)e^{(s_2^2/2r)-(t_2/2)}]$ and common variance $r^{-1}$. So, draws of $c$ are easily obtained.*

A generalization of this algorithm was proposed by Phillips and Smith (1993). Each iteration in their sampling scheme comprises a set of $d$ directions $e_1, \ldots, e_d$. When the $e_i$ form the canonical basis of $R^d$ indicating the $d$ axes, they reproduce the standard algorithms presented so far. When the $e_i$ form another basis of $R^d$, the algorithm corresponds to sampling through components after a linear reparametrization. Other choices of direction are available. In particular, random choices can be made. Completely random choices correspond to the algorithm of Schmeiser and Chen (1991) observed at every $d$ iterations. Phillips and Smith (1993) suggested only choosing $e_1$ at random and then choosing $e_2, \ldots, e_d$ to be mutually orthogonal and orthogonal to $e_1$. The directions can also be tuned to improve sampling in every specific setting. This will generally destroy the convergence properties of the chain so this tuning can only be operated at a transient phase of the chain.

Phillips and Smith (1993) also suggested a generalization in the choice of $c$. When sampling $c$ from a general density $g(c|\theta^{(j-1)}, e^{(j)})$, the proposed value of the chain $\theta^{(j)} = \theta^{(j-1)} + c^{(j)}e^{(j)}$ is only accepted with probability

$$\min\left\{1, \frac{\pi(\theta^{(j-1)} + c^{(j)}e^{(j)})g(-c^{(j)}|\theta^{(j-1)} + c^{(j)}e^{(j)}, e^{(j)})}{\pi(\theta^{(j-1)})g(c^{(j)}|\theta^{(j-1)}, e^{(j)})}\right\}.$$

The Schmeiser and Chen (1991) algorithm corresponds to $g(c|\theta^{(j-1)}, e^{(j)}) \propto \pi(\theta^{(j-1)} + ce^{(j)})$ and accepting the proposed value with probability 1.

These algorithms work well for the possibility of large moves when appropriate directions are chosen. This is generally not possible when moving along the direction of the components. The performance of these algorithms in highly dimensional models with correlated parameters or with concentrated modes may be very poor. The empirical evidence obtained by Phillips and Smith (1993) suggested that the orthogonalization of random directions improves convergence over sampling from entirely random directions which also improves convergence over Gibbs sampling. The improvement increases substantially when directions are chosen according to a principal component analysis based on (an approximation to) the posterior variance matrix. The comparisons mentioned above were based on the convergence prescription of Raftery and Lewis (1992) and on the integrated autocorrelation time (Green and Han, 1992).

Another approach to the appropriate selection of the sampling directions is given in Gilks, Roberts and George (1994). They proposed adaptive methods for choosing the direction of sampling. The method is based on a current sample $\theta_1^{(j)}, \ldots, \theta_n^{(j)}$ at each iteration of the chain and allows the choice of the next direction $e^{(j+1)}$ to be based on the current sample. Dependence on a succession of values is not a problem for convergence as the Markov chain may be enlarged to contemplate $n$ of the original steps. At each iteration, an element from the sample is randomly chosen and replaced by a point chosen according to a specified direction. This direction may depend on the other points in the current sample. When $n = 1$, the Schmeiser and Chen (1991) algorithm is obtained. A variation is to allow only directions connecting the chosen point to the other points in the sample. When these choices are made with the posterior controlling the displacements along the directions, the points are automatically chosen. When the displacements are chosen according to some other distribution, the points are only proposed and an acceptance probability must be evaluated.

The motivation for the adaptive methods is that the sample points will cover adequately regions of high probability of $\pi$ near the equilibrium and subsequent points will tend to concentrate in these areas. For slow mixing chains, however, the sample from initial iterations may be far from these regions. Movements of the sample toward high probability regions may be slow. Hence, these difficulties and the extra computations required prevent a straightforward recommendation of these techniques. Rather, they should be used as auxiliary devices.

Tierney and Mira (1999) and Mira and Tierney (2002) introduced the *delayed rejection Metropolis* algorithm. In its simplest version a rejected draw is still used to help proposing a new candidate. Suppose that $\theta$ is the current state of the Markov chain, $\theta_1$ is a candidate draw from $q_1(\theta, \cdot)$ and $\alpha(\theta, \theta_1)$ is the acceptance probability. Instead of repeating the current value $\theta$ when $\theta_1$ is rejected, the algorithm suggests delaying the rejection and instead sample a second candidate draws $\theta_2$ from $q_2(\theta, \theta_1, \cdot)$, which is accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(\theta_2)}{\pi(\theta)} \frac{q_1(\theta_2, \theta_1)}{q_1(\theta, \theta_1)} \frac{[1 - \alpha(\theta_2, \theta_1)]q_2(\theta_2, \theta_1, \theta)}{[1 - \alpha(\theta, \theta_1)]q_2(\theta, \theta_1, \theta_2)}\right\}.$$

Green and Mira (2001) generalized the delayed rejection algorithm by noticing that $\alpha$ is derived under the unnecessary assumption that the return path has to visit $\theta_1$. They pointed out that this restriction may limit the formulation of delayed rejection algorithms and completely prevents its natural extension to transdimensional situations. See also Al-Awadhi, Hurn and Jennison (2004) for a related algorithm.

**Example 7.10** *Assume that the target distribution is the following mixture of two univariate normal densities:*

$$\pi(\theta) = 0.9 f_N(\theta; 0, 1) + 0.1 f_N(\theta; 10, 1)$$

*such that $\pi$ is bimodal and the modes are relatively far away from each other. The performance of the above delayed rejection Metropolis algorithm is compared to the performance of a simple random walk Metropolis algorithm. For both algorithms the proposal density is $q(\theta, \theta_1) = f_N(\theta_1; \theta, \tau_1^2)$ where $\tau_1 = 2$. For the delayed rejection algorithm $q_2(\theta, \theta_1, \theta_2) = f_N(\theta_2; \theta_1, \tau_2^2)$, where $\tau_2 = 4$. Both chains started at $\theta^{(0)} = 10.5$ and were run for 15000 iterations. As it can be seen in Figure 7.5, the delayed rejection algorithm induces the appropriate number visits to the lower mode of the target distribution and better mixing of the chain to lower chain autocorrelation.*

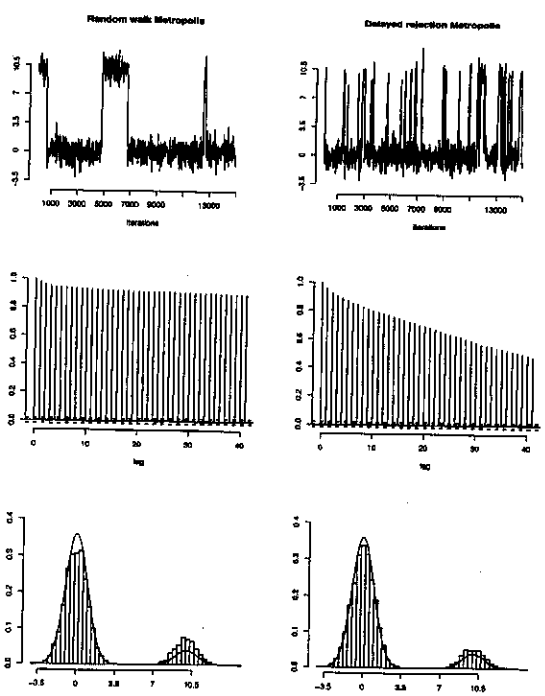

Figure 7.5 *Comparison between random walk Metropolis (left panels) and delayed rejection Metropolis (right panels): chain paths (top panels); autocorrelation functions (middle panels); and target density versus histogram approximation (bottom panels).*

Convergence acceleration can also be achieved by directly changing the chain output with resampling techniques. Assume $m$ parallel chains are used for sampling and a sample $\theta_1^{(j)}, \ldots, \theta_m^{(j)}$ is available at iteration $j$. This is a sample from the marginal distribution of the chain $\pi^{(j)}$ at iteration $j$ and the objective is to use this sample to obtain a sample from $\pi$. Successively drawing samples through the chain will attain that but Gelfand and Sahu (1994) considered the possibility of making this sample an approximate sample from $\pi$ in a single iteration. Methods to achieve this goal were described in Section 1.5 where it was shown how an arbitrary approximating density $q$ can be used to generate a sample from $\pi$. In the present context, the approximating density $\pi^{(j)}$ is not directly available and must be estimated from its sample. Methods for doing it were discussed in Chapter 5. Once this density estimate $\hat{\pi}^{(j)}$ is obtained, resampling techniques can be used. Rejection methods are preferred to weighted resampling in principle when calculation of the enveloping constant is feasible. This is rarely the case and Gelfand and Sahu (1995) concentrated on weighted resampling.

An approximate sample from $\pi$ is obtained by forming weights $w_i \propto \pi(\theta_i^{(j)})/\hat{\pi}^{(j)}(\theta_i^{(j)})$, $i = 1, \ldots, m$, and resampling from the discrete distribution with probabilities $w_1, \ldots, w_m$. If $\hat{\pi}^{(j)}$ is a reasonable approximation for $\pi$ and $m$ is large, the resulting resample will be a good approximation to a sample from $\pi$. This resample substitutes the current sample with consequent approximation to the limiting distribution. Even when the approximation is not very accurate, the resulting resample should be closer to a sample from $\pi$ with convergence improvements.

Another suggestion made by Gelfand and Sahu (1995) is the adaptation of the transition kernels to forms that speed up the convergence. They work in the discrete case with an adaptive initial phase in the chain designed to identify better transition matrices among those with the same limiting distribution. This procedure does not affect the convergence of the chain but allows identification of better transitions. The changes between transitions in the initial phase are deterministic and do not violate convergence results. They apply these results to the choice of tuning parameters in specific proposal transition kernels, thus providing further theoretical justification for their choice (see also Section 6.3).

Liu, Liang and Wong (2000) introduced the *multiple-try Metropolis* algorithm. They claim that the algorithm improves exploration of *neighboring regions* defined by the transition kernel $q(\theta, \phi)$. They define weights

$$w(\theta, \phi) = \pi(\phi) q(\theta, \phi) \lambda(\theta, \phi),$$

where $\lambda(\theta, \phi)$ is a nonnegative symmetric function in $\theta$ and $\phi$ with $\lambda(\theta, \phi) > 0$ whenever $q(\theta, \phi) > 0$. Suppose that $\theta$ is the current state of the Markov chain, then one iteration of the algorithm proceeds according to the following steps:

1. Draw a random sample $\phi_1^*, \ldots, \phi_k^*$ from $q(\theta, \cdot)$;
2. Select $\phi_k$ from $\{\phi_1^*, \ldots, \phi_k^*\}$ with probability proportional to $w(\phi_j^*, \theta)$;
3. Draw $\phi_1, \ldots, \phi_{k-1}$ from $q(\phi_k, \cdot)$;
4. Calculate the acceptance probability

$$\alpha = \min\left\{1, \frac{w(\phi_1^*, \theta) + \cdots + w(\phi_k^*, \theta)}{w(\phi_1, \phi_k) + \cdots + w(\phi_k, \phi_k)}\right\}.$$

If the move is accepted, set $\phi = \phi_k$. If the move is not accepted, set $\phi = \theta$.

Liu, Liang and Wong (2000) pointed out that the multiple try Metropolis

algorithm corresponds to Frenkel and Smit's (1996) *orientational-biased Monte Carlo* when $q(\theta, \phi)$ is symmetric . Two of the $\lambda$ functions introduced by Liu, Liang and Wong (2000) are $\lambda_1(\theta, \phi) = 2\{q(\theta, \phi) + q(\phi, \theta)\}^{-1}$ and $\lambda_2(\theta, \phi) = \{q(\theta, \phi) q(\phi, \theta)\}^{-\alpha}$. When $\alpha = 1$, $w(\theta, \phi)$ can be thought of as weights obtained in a weighted resampling algorithm with proposal density $q(\theta, \phi)$ and target $\pi$. Liu, Liang and Wong (2000) recommend using $\alpha$ close to one.

**Example 7.11** *Consider Example 3.7 again, but replacing the independent uniform prior distributions for $\beta_1$ and $\beta_2$ by independent normal prior distributions $\beta_1 \sim N(\beta_1; 20, 20^2)$ and $\beta_2 \sim N(\beta_2; 0, 1.5^2)$. This change leads to the following posterior distribution (after integrating $\sigma^2$ out)*

$$\pi(\beta) \propto \left( \sum_{i=1}^{6} [y_i - \beta_1 - \beta_1 e^{-\beta_2 x_i}]^2 \right)^{-2} f_N(\beta_1; 20, 20^2) f_N(\beta_2; 0, 1.5^2) .$$

*Both random walk Metropolis and random walk multiple try Metropolis algorithms are implemented. The random walk was run for 10000 iterations, while a computationally comparable multiple-try was run for 2000 iterations with $k = 5$. The proposal distributions for both algorithms are $q(\beta_1, \tilde{\beta}_1) = f_N(\tilde{\beta}_1; \beta_1, 25^2)$ and $q(\beta_2, \tilde{\beta}_2) = f_N(\tilde{\beta}_2; \beta_2, 3^2)$. Using the random walk algorithm, approximations for the posterior mean, standard deviation and 95% credibility interval of $\beta_1$ are* 20.066, 6.506 *and* (12.300, 38.859)*, respectively. Similarly, approximations for the posterior mean, standard deviation and 95% credibility interval of $\beta_2$ are* 0.808, 0.621 *and* (0.109, 2.569)*, respectively. Using the multiple-try algorithm, approximations for the posterior mean, standard deviation and 95% credibility interval of* 18.190, 5.951 *and* (9.527, 32.912)*, respectively. Similarly, approximations for the posterior mean, standard deviation and 95% credibility interval of $\beta_2$ are* 1.073, 0.718 *and* (0.171, 2.766)*, respectively. Results appear in Figure 7.6 showing a good agreement between both sampling approaches.*

*Acceptance rates for $\beta_1$ and $\beta_2$ are* 0.1310 *and* 0.1185 *for the random walk Metropolis algorithm and* 0.5910 *and* 0.6520 *for the multiple try Metropolis algorithm, suggesting an efficiency gain with the use of multiple-try algorithms. Effective sample sizes based on $\beta_1$ and $\beta_2$ are* 132 *and* 29 *for the random walk Metropolis algorithm and* 1365 *and* 577 *for the multiple try Metropolis algorithm. These results shows a clear superiority of the use of multiple try over standard Metropolis algorithms.*

Finally, another important development in the area should be mentioned. Propp and Wilson (1996) suggested a Markov chain sampling scheme that would enable exact draws from the target distribution. For this reason, their approach was named *perfect sampling*. Their work attracted interest from the statistical community and has been attracting a considerable amount of research effort. Nevertheless, most of the work to date is restricted to theoretical aspects and applications only in a limited set of situations.

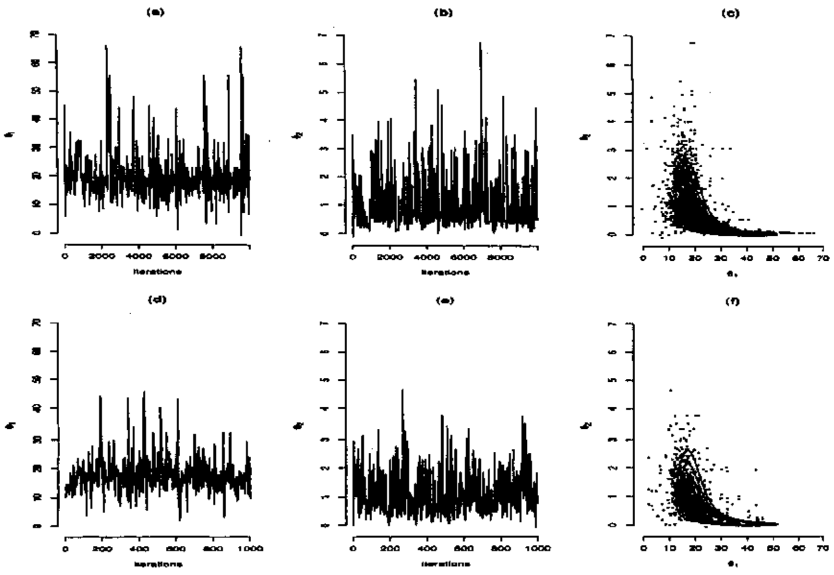

Figure 7.6 *Random walk Metropolis and multiple try random walk Metropolis with proposal $q(\beta_1, \tilde\beta_1) = f_N(\tilde\beta_1; \beta_1, 25^2)$ and $q(\beta_2, \tilde\beta_2) = f_N(\tilde\beta_2; \beta_2, 3^2)$. Random walk Metropolis: (a) trace plot of $\beta_1$, (b) trace plot of $\beta_2$ and (c) contours of $\pi(\beta)$ against draws from the posterior distribution. Random walk multiple try Metropolis: (d) trace plot of $\beta_1$, (e) trace plot of $\beta_2$ and (f) contours of $\pi(\beta)$ against draws from the posterior distribution.*

According to Sisson (2005), there are indications that this area would not be as fruitful as it initially seemed. Future research will indicate if this is the case.

### *7.4.2 Alterations to the equilibrium distribution*

The use of resampling described above was applied to intermediate iterations of the chain, before it had reached equilibrium. The same ideas could be applied directly to the equilibrium stage. Assume it is possible to construct a chain with fast convergence to a limiting distribution $q$ that approximates $\pi$. A sample from $q$ is available after convergence of the chain is established but the objective is to obtain a sample from $\pi$. Once again, resampling methods can be used. Unlike the previous case, the expression for $q$ will typically be available here but the enveloping constant will still be difficult to obtain. Weighted resampling seems a better option although it provides a sample from an approximation to $\pi$. Weights $w_i$ based on the

ratio $\pi/q$ at sampled values can be calculated and the resample drawn from the discrete distribution concentrated at the sample from $q$ with respective weights $w_1, \ldots, w_m$.

One possibility for $q$ is obtained by *heating* the target distribution according to $q(\theta) \propto \pi(\theta)^{1/T}$ where the constant $T > 1$ receives the physical interpretation of system temperature, hence the nomenclature used. This mechanism was suggested by Jennison (1993). It is based on simulated annealing (Kirkpatrick, Gelatt and Vecchi, 1983; Ripley, 1987), an optimization technique designed to find maxima of functions. The equilibrium distribution $\pi$ corresponds to the basal temperature $T = 1$. The *heated* distribution $q$ is flattened with respect to $\pi$ and its density gets closer to the uniform distribution. It becomes easier to design a Markov chain that converges faster to the equilibrium distribution. This alteration is particularly relevant for the case of a distribution with distant modes. It is generally difficult to construct chains allowing for frequent moves between regions around the modes. Gibbs sampling will make these movements very slowly and Metropolis-Hastings algorithms will tend to reject most of these moves. By flattening the modes, the moves required to cover adequately the parameter space become more likely. Gilks and Roberts (1996) suggested modifying the distribution by the inclusion of what they call *stepping stones.* These are lumps of probability redistributed to regions to ease the moves between modes. This is in a way a discrete version of the *heating* procedure that redistributes weights continuously over the parameter space. Besag and Green (1993) also discussed approaches to multimodality in more qualitative terms.

The idea of *heated* equilibrium distributions was also used by Geyer (1991). The difference here is to use $m$ parallel chains, each having equilibrium distribution

$$\pi_i(\theta) \propto \pi(\theta)^{1/T_i} \tag{7.24}$$

gradually *heated* according to the rule $T_i = 1 + \lambda(i-1)$, $i = 1, \ldots, m$, for some $\lambda > 0$. The target distribution is simply one of these chains, corresponding to the case $i = 1$. Chains with higher temperatures will have more movement than *cooler* chains, including the chain of interest $i = 1$. The objective here is to make the low temperature chain benefit from the moves from the *heated* chains. So, jumps between the chains are proposed in addition to regular moves within each chain. At iteration $j$, an exchange between the states of chains $i$ and $k$ is proposed. The acceptance probability of this swap is

$$\alpha_{ik}(\theta_i^{(j)}, \theta_k^{(j)}) = \min\left\{1, \frac{\pi_i(\theta_k^{(j)})\pi_k(\theta_i^{(j)})}{\pi_i(\theta_i^{(j)})\pi_k(\theta_k^{(j)})}\right\}.$$

After convergence is reached at the $m$ chains, a sample from chain $i = 1$

is drawn. The remaining $m-1$ are only used to speed the convergence of the chain of interest and once that is achieved, they are discarded. This is an obvious computational disadvantage of the method. Another disadvantage is that sampling from the heated distributions will generally be more difficult.

**Example 7.12** *The number of positive responses $y_i$ out of $n_i$ trials at level $x_i$ is modelled by the binomial model*

$$y_i|\beta \sim Bin\left(n_i, \frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}}\right), \quad i=1,2,3,4.$$

*where $\beta=(\beta_1,\beta_2)$. Let $x=(-0.863,-0.296,-0.053,0.727)$, $n=(5,5,5,5)$ and $y=(0,1,3,5)$. When a flat prior is adopted for $\beta$, the posterior distribution becomes*

$$\pi(\beta) \propto \prod_{i=1}^{4} e^{(\beta_1+\beta_2 x_i)y_i} \prod_{i=1}^{4} \left\{1+e^{\beta_1+\beta_2 x_i}\right\}^{-n_i} .$$

*The objective here is to find the posterior mode. The above simulated annealing algorithm is implemented for combinations of four initial values and two cooling schedules. The proposal distribution is $q(\beta|\beta^{(i)}) = f_N(\beta;\beta^{(i)}, 0.05^2 I_2)$. Figure 7.7 exhibits the behavior of the algorithm. It can be seen that the convergence is achieved faster when $T_i=1/i$ for both $\beta_1$ and $\beta_2$. For instance, the approximate mode is $(0.88, 7.99)$ when $(\beta_1^{(0)}, \beta_2^{(0)}) = (5,30)$ was the initial value. The actual mode of $\pi$ is $(0.87, 7.91)$, which was obtained by a standard Newton-Raphson-like algorithm.*

A similar idea was proposed by Marinari and Parisi (1992) under the name of simulated tempering. Again, $m$ chains with respective equilibrium distributions $\pi_i$, $i=1,\ldots,m$, are used, but in series and not in parallel as above. One possibility for the $\pi_i$ is given by (7.24). The chain thus formed alternates between equilibrium distributions and is formally extended to a chain of $(\theta, i)$ where $i$ denotes the sampling scheme (with respective equilibrium distribution $\pi_i$) considered.

This structure is very similar to that used for model choice based on chains with jumps. The only changes are the fact that the parameter is the same through all components and the interpretation given to the additional component introduced. For model choice, it represents the model considered. Here, it only specifies the auxiliary chain constructed to accelerate the convergence of the chain of interest. Geyer and Thompson (1995) suggested that only jumps between neighboring models are allowed and provide further discussion on implementation issues. Once again, after convergence is reached, only samples of $\theta$ corresponding to $i=1$ are retained.

Simulated annealing algorithms can, at least in principle, be used to search for modal models or classes of modal models when the number of entertained models is large, or even uncountable. Andrieu, de Freitas and

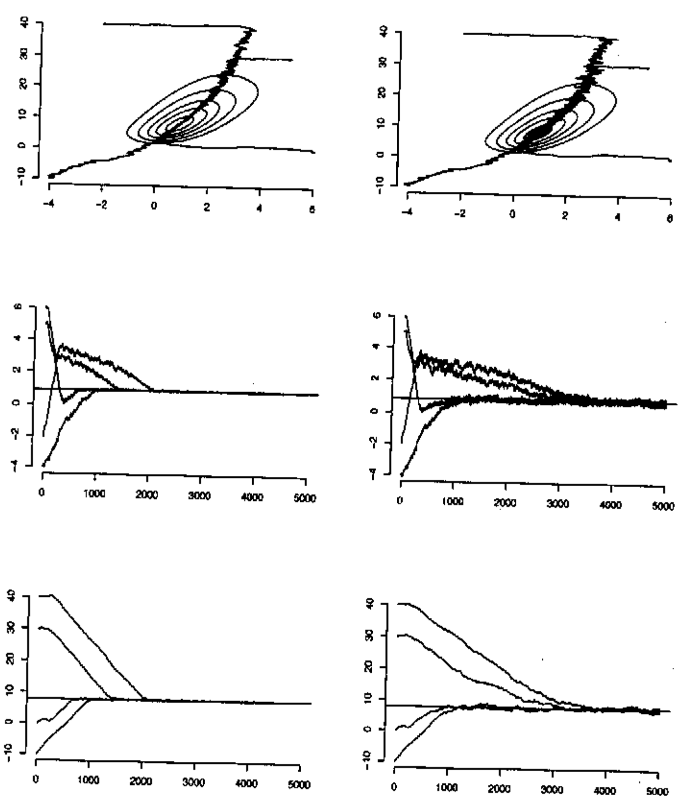

Figure 7.7. *Simulated annealing paths for four distinct initial values $(5,30)$, $(-2,40)$, $(-4,-10)$ and $(6,0)$, two cooling schedules $T_i = 1/i$ (left column) and $T_i = 1/[10\log(1+i)]$ (right column) and proposal distribution $q(\beta|\beta^{(i)}) = f_N(\beta;\beta^{(i)}, 0.05^2 I_2)$. Joint trajectories for $(\beta_1,\beta_2)$ appear on the top row along with contours of the posterior density $\pi(\beta)$, while individual trajectories for $\beta_1$ and $\beta_2$ appear on the second and third rows, respectively.*

Doucet (2000) and Brooks, Friel and King (2003) extended standard simulated annealing to transverse the model space with limiting distribution

$$f_T(\theta_j, j) \propto \exp\left\{-\frac{E(\theta_j, j)}{T}\right\}$$

where $E(\theta_j, j)$ is a model-ranking function to be minimized. In their capture-recapture applications, King and Brooks (2004) and Sisson and Fan (2004) adopted AIC-type criteria where $E(\theta_j, j) = -2f(y|\theta_j, j) + 2d_j$, for $d_j$ the dimension of $\theta_j$. Unfortunately, models that are chosen based on arbitrary rankings are not guaranteed to be models with high posterior model probabilities (Sisson, 2005). Nonetheless, from a decision-theoretic perspective, the above ideas can be thought of as strategies for selecting models (decisions) that maximize utility functions based on $E(\theta_j, j)$.

### *7.4.3 Auxiliary variables*

Simulated tempering is a scheme where an additional variable was introduced with the aim of reducing slow mixing of the original chain. Slow mixing of the chain leads to slow convergence and is normally due to particularities of the target distribution such as multimodality or high correlation between some of the components of $\theta$. Whatever the reason, convergence is slow due to high autocorrelation in the chain.

The introduction of auxiliary variables attempts to remove these sources of correlation and to hasten convergence of the extended chain. Auxiliary variables are also introduced in order to make complicated posterior distributions more manageable (see Example 7.13 below). Consider the introduction of variables $\phi$ with known (and preferably easy to sample from) conditional distribution $\pi(\phi|\theta)$. The equilibrium distribution now becomes $\pi(\theta, \phi) = \pi(\theta)\pi(\phi|\theta)$ where the first term to the right hand side is the target distribution. The extended chain alternates generations from the full conditional distributions $\pi_\theta(\theta) \propto \pi(\theta)\pi(\phi|\theta)$ and $\pi_\phi(\phi) = \pi(\phi|\theta)$ at every iteration. If these generations can be made directly, Gibbs sampling can be applied to blocks $\theta$ and $\phi$ and discarding the $\phi$s once the MCMC sample is obtained. Besag and Green (1993) considered chains with more general transition kernels with special attention to spatial statistics problems.

**Example 7.13** *Assume that $\theta$ has a posterior distribution that can be written as*

$$\pi(\theta) = q(\theta) \prod_{i=1}^{I} b_i(\theta)$$

*where $q$ has an easy to sample distribution and the functions $b_i$ are complicated terms involving interactions between the components of the vector $\theta$ (Edwards and Sokal, 1988). Examples include spatial and temporal correlation. A vector $\phi = (\phi_1, \ldots, \phi_I)'$ of components conditionally independent given $\theta$ can be defined with distributions $\phi_i|\theta \sim U[0, b_i(\theta)]$, $i = 1, \ldots, I$. Generation from $\phi|\theta$ is simple and*

$$\begin{aligned} \pi(\theta, \phi) &= \pi(\theta)\,\pi(\phi|\theta) \\ &= q(\theta) \prod_{i=1}^{I} b_i(\theta) \times \prod_{i=1}^{I} \frac{I(0 \le \phi_i \le b_i(\theta))}{b_i(\theta)} \\ &= q(\theta) \prod_{i=1}^{I} I(0 \le \phi_i \le b_i(\theta)) \; . \end{aligned}$$

*Generation of $\theta$ from $\pi_\theta$ involves a generation from $q$ followed by verification of conditions $b_i(\theta) \ge \phi_i$. By the rejection method, the generated value is retained if they are all satisfied. Otherwise, a new value is generated from $q$ until all conditions are satisfied.*

**Example 7.14** *The vector $\theta = (\theta_1, \ldots, \theta_d)'$ represents the colors in the pixels of a given image. Each position $i$ has color $\theta_i$ varying in a finite set of possibilities $\{1, \ldots, L\}$. A frequently adopted distribution for $\theta$ is given by the Potts model that seeks to reflect similarities in colors of neighboring pixels. Its probability function is given by the Gibbs distribution with energy $E(\theta)$ being the number of neighboring pairs of the same color. So, $E(\theta) = \sum_{j \sim k} I(\theta_j \ne \theta_k)$, where $j \sim k$ denotes that the pair $i = (j, k)$ consists of neighboring positions. The probability function can be written as*

$$\pi(\theta) \propto \prod_i b_i(\theta) \text{ where } b_i(\theta) = e^{-\beta I(\theta_j \ne \theta_k)} \text{ and } \beta = 1/(kT) \; .$$

*If $\beta$ is large, there is high correlation between the components of $\theta$ and the use of auxiliary variables is recommended.*

*Define an auxiliary vector $\phi = (\phi_1, \ldots, \phi_I)$ where $I$ is the number of pairs of neighbors and the auxiliary variable $\phi_i$ associated with the pair $i = (j, k)$ has independent bern$(b_i(\theta))$ distributions, $i = 1, \ldots, I$. If $\theta_j = \theta_k$, $\phi_i = 1$ and if $\theta_j \ne \theta_k$, $\phi_i \sim$ bern$(e^{-\beta})$. Generation from $\phi|\theta$ is therefore trivial. Generation of $\theta|\phi$ is based on a uniform distribution over $\{1, \ldots, L\}^d$. The generated value will be accepted if all pairs of neighbors satisfy the configuration given by $\phi$. This generation mechanism was proposed by Swendsen and Wang (1987). It is discussed in the statistical context by Besag and Green (1993) and Green (1996).*

Gilks and Roberts (1996) considered other possible uses of auxiliary variables. They include the important case of missing observations and a version of the rejection method where the auxiliary variable is again an indicator variable controlling the acceptance probability. Gilks, Best and Tan (1995) suggested a generalization of the adaptive rejection method where the acceptance probability is replaced by a Metropolis step. They show that their method can also be seen as another use of auxiliary variables. Finally, Besag and Green (1993) discuss other possibilities still preserving reversibility of the chain. See Higdon (1998) and Damien, Wakefield and

Walker (1999) for further developments and discussions about auxiliary variables.

*The slice sampler*

Perhaps one of the most popular auxiliary variable MCMC algorithm to date, the *slice sampler* is derived from the Example 7.13 by setting $I = 1$, $q(\theta) = 1$ and $\pi(\phi|\theta) = \pi^{-1}(\theta)I(0 \leq \phi \leq \pi(\theta))$. In this case,

$$\pi(\theta,\phi) = \pi(\theta)\pi(\phi|\theta) = \pi(\theta)\frac{1}{\pi(\theta)}I(0 \leq \phi \leq \pi(\theta)),$$

suggesting that standard MCMC schemes can be applied to sample from $\pi(\phi|\theta)$ and $\pi(\theta|\phi)$ iteratively.

Sampling $\phi$ from $\pi(\phi|\theta)$ usually imposes no difficulty. $\theta$ is sampled from a uniform distribution over the region $\{\theta, \pi(\theta) \geq \phi\}$ (see Figure 7.8(a)). The name of the sampler derives from the fact that the region is defined by *slicing* the density $\pi(\theta)$ horizontally at the contour level $\phi$. Sampling $\theta$ may not be straightforward because defining the region requires solving the inequality $\pi(\theta) \geq \phi$ for $\theta$. This can be a computationally challenging task in higher dimensional parameter spaces. Neal (2003) proposed an adaptive rejection-like algorithm to overcome this difficulty. Silva, Lopes and Migon (2006) uses Neal's adaptive algorithm in the context of generalized inverse Gaussian models. Roberts and Rosenthal (1999) and Mira and Tierney (2002) examined theoretical properties of the slice sampler.

**Example 7.15** *Consider Example 3.6 again, but now suppose that the total number of animals is 22 and that counts are* $y = (14, 3, 5)$. *Adopting an uniform prior for* $\theta$ *leads to* $\pi(\theta) \propto (2+\theta)^{14}(1-\theta)^3\theta^5$. *Posterior simulation is now performed through the above slice sampling algorithm. Based on a sample of size* $M = 5000$ *from* $\pi(\theta)$, *approximations for the posterior mean, standard deviation and 95% credibility interval of* $\theta$ *are* 0.698, 0.123 *and* (0.417, 0.908), *respectively. Figure 7.8 illustrates the sampler.*

## 7.5 Exercises

**7.1** *Show that all the estimators of the model likelihood presented in Section 7.2 are consistent and obtain their asymptotic variance. Also, discuss whether variance of* $\{\hat{f}_4(y)\}^{-1}$ *diverges or not.*

**7.2** *Consider the annealed importance sampling algorithm.*

*(a) Derive the annealed importance weights (7.7).*

*(b) Derive the annealed importance sampling estimator of* $f(y)$ *of Equation (7.8).*

**7.3** *Derive the bridge identity (7.9).*

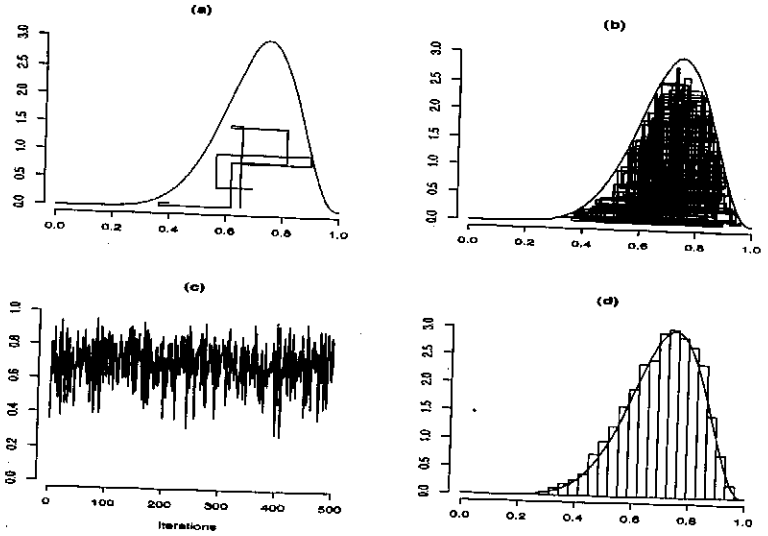

Figure 7.8 *Slice sampling performance after (a) 20 and (b) 500 iterations. The first 500 draws appear in (c) and the resulting histogram approximation of* $\pi$ *appears in (d).*

**7.4** *Consider the path algorithm.*

*(a) Derive Equation (7.10).*

*(b) (Friel and Pettitt, 2005) Consider the power posterior density* $g(\theta|\lambda) = p(\theta)l^\lambda(\theta)/c(\lambda)$ *and assume that* $\lambda \sim U(0,1)$ *and* $(\theta_1,\lambda_1),\ldots,(\theta_n,\lambda_n)$ *is a sample from the joint density* $g(\theta|\lambda)g(\lambda)$. *Show that* $H(\theta,\lambda) = p(\theta)l(\theta)^\lambda \log l(\theta)$ *and the path estimator of* $f(y)$ *is*

$$\exp\left\{\frac{1}{n}\sum_{j=1}^{n} p(\theta_j)l^{\lambda_j}(\theta_j)\log l(\theta_j)\right\}.$$

**7.5** *Derive the identity from Equation (7.12).*

**7.6** *Consider Example 7.4.*

*(a) Derive the minimum posterior predictive loss criteria* $D_1^G$, $D_2^G$ *and* $D_3^G$.

*(b) Derive the deviance information criteria* $D_1^S$, $D_2^S$ *and* $D_3^S$.

**7.7** *Show that the densities used in the intrinsic Bayes factor of Berger and Pericchi (1996) and the fractional Bayes factor of O'Hagan (1995) can be written in the form* $p(y_{S_1}|y_{S_2})$, *identifying the sets* $S_1$ *and* $S_2$ *in each case.*

**7.8** *(Gelfand, 1996) Table 7.9 presents measurements of tree trunk circumferences (Draper and Smith, 1981). Let $y_{ij}$ be the time $t_j$ measurement of the $i^{th}$ tree, for $i = 1, \ldots, I = 5$ and $j = 1, \ldots, J = 7$. Consider the fit of the following four different random effects (or hierarchical) models:*

$$\begin{aligned}
M_1 : &\quad y_{ij} = \beta_0 + b_i + \varepsilon_{ij} \\
M_2 : &\quad y_{ij} = \beta_0 + \beta_1 t_j + b_i + \varepsilon_{ij} \\
M_3 : &\quad y_{ij} = \beta_0 (1 + \beta_1 e^{\beta_2 t_j})^{-1} + \varepsilon_{ij} \\
M_4 : &\quad y_{ij} = (\beta_0 + b_i)(1 + \beta_1 e^{\beta_2 t_j})^{-1} + \varepsilon_{ij}
\end{aligned}$$

*where $\varepsilon_{ij}$ are a $N(0, \sigma^2)$ random sample and $b_i$ are assumed $N(0, \tau^2)$. Let $\theta_i$ be the vector of parameters of model $M_i$, for $i = 1, \ldots, 4$. Assume that the joint priors are $p(\theta_1|M_1) \propto \sigma^{-2}\tau^{-3}\exp\{-2(\sigma^{-2} + \tau^{-2}\}$, $p(\theta_2|M_2) \propto \sigma^{-2}\tau^{-3}\exp\{-2(\sigma^{-2} + \tau^{-2}\}$, $p(\theta_3|M_3) \propto \sigma^{-2}\exp\{-2\sigma^{-2}\}$ and $p(\theta_4|M_4) \propto \sigma^{-2}\tau^{-3}\exp\{-2(\sigma^{-2} + \tau^{-2}\}$ . Compute the deviance information criteria ($D^S$) and the pseudo-Bayes factors ($G_4$) and compare them.*

| *Trees* | | | | | |
|---|---|---|---|---|---|
| *1* | *2* | *3* | *4* | *5* | *Days* |
| *30* | *33* | *30* | *32* | *30* | *118* |
| *58* | *69* | *51* | *62* | *49* | *484* |
| *87* | *111* | *75* | *112* | *81* | *664* |
| *115* | *156* | *108* | *167* | *125* | *1004* |
| *120* | *172* | *115* | *179* | *142* | *1231* |
| *142* | *203* | *139* | *209* | *174* | *1372* |
| *145* | *203* | *140* | *214* | *177* | *1582* |

Table 7.9 *Trunk circumference (in milimeters) of 5 orange trees measured at 7 different times (in days) (Draper and Smith, 1981).*

**7.9** *Consider the model choice setting of Section 7.3 with a superparameter $\theta = (\theta_1, \ldots, \theta_J)$ and a quantity $M$ indicating a model with joint posterior $\pi(\theta, j)$.*

*(a) Obtain the posterior distributions of $\theta_j|M = j$, $j = 1, \ldots, J$, and $M$. Show that samples from $\theta_j|M = j$ are obtained by retaining the draws of $\theta_j$ associated with a value $j$ for $M$.*

*(b) Consider a component of the parameter, say $\phi$, that is shared by all models. Obtain its marginal posterior distribution $\pi(\phi)$.*

**7.10** *Show that the supermodel approach of Carlin and Chib (1995) cannot be applied when the number of models considered is not finite.*

**7.11** *Show that the reversible jump Markov chain approach for inference with a collection of models reduces to the Metropolis-Hastings algorithm if only one model is considered and moves are always proposed. What happens if there is a positive probability of not proposing a move?*

**7.12** *Consider the reversible jump Markov chain approach with transition kernel $q_m((\theta, j), (\phi, k))$ for each proposed move $m$. Show that reversibility of chain is ensured if the acceptance probability associated with move $m$ is given by*

$$\alpha_m((\theta, j), (\phi, k)) = \min\left\{1, \frac{\pi(\phi, k) q_m((\phi, k), (\theta, j))}{\pi(\theta, j) q_m((\theta, j), (\phi, k))}\right\} .$$

**7.13** *(Green, 1995) Consider again the conditions of Example 7.7 with irreducible birth and death chains moving across models $M_k$ with piecewise constant intensity rates having $k$ steps and model parameter $\theta_k = (\lambda(k), t(k))'$ where $\lambda(k) = (\lambda_0, \lambda_1, \ldots, \lambda_k)$ and $t(k) = (t_1, \ldots, t_k)$, $k = 0, 1, 2, \ldots$*

*(a) Show that the likelihood for model $M_k$ is given by*

$$l(\theta_k, k) = \prod_{j=0}^{k} \lambda_j^{d_j} e^{-\lambda_j (t_{j+1} - t_j)}$$

*where $d_j$ is the number of occurrences in interval $I_j$, $j = 0, 1, \ldots, k$.*

*(b) Show that the likelihood ratio between model $M_k$ with parameters $\theta'_k$ and $\theta_k$ is given by*

$$lr_\lambda = \prod_{j=0}^{k} \left(\frac{\lambda'_j}{\lambda_j}\right)^{d_j} e^{(\lambda_j - \lambda'_j)(t_{j+1} - t_j)}$$

*if they differ only on the values of $\lambda(k)$ and*

$$lr_t = \prod_{j=0}^{k} \lambda_j^{d'_j - d_j} e^{\lambda_j [(t_{j+1} - t_j) - (t'_{j+1} - t'_j)]}$$

*if they differ only on the values of $t(k)$.*

*(c) Consider a move of type 3 which proposes a change of the intensity rate of a randomly chosen interval, say $j$, from $\lambda_j$ to $\lambda'_j$ according to a random walk log $\lambda'_j \sim U[\log \lambda_j - 1/2, \log \lambda_j + 1/2]$. Show that the acceptance probability of this move is*

$$\min\left\{1, lr_\lambda \left(\frac{\lambda'_j}{\lambda_j}\right)^{\alpha} \exp[-\beta(\lambda'_j - \lambda_j)]\right\} .$$

*(d) Consider a move of type 4 which proposes a change of the endpoint of a randomly chosen interval, say $j$, from $t_j$ to $t'_j$ according to a $U[t_{j-1}, t_{j+1}]$*

*distribution. Show that the acceptance probability of this move is*

$$\min\left\{1, lr_t \frac{(t_{j+1}-t'_j)(t'_j-t_{j-1})}{(t_{j+1}-t_j)(t_j-t_{j-1})}\right\} .$$

*(e) Consider a move of type 1 which proposes the birth of a new endpoint at a point uniformly chosen on* $[0,T]$. *Show that the acceptance probability of this move is*

$$\min\left\{1, \frac{\pi(\theta_{k+1},k+1)}{\pi(\theta_k,k)} \times \frac{B_{k+1,k}}{B_{k,k+1}q(t^*,u)} \times \left|\frac{\partial\theta_{k+1}}{\partial(\theta_k,t^*,u)}\right|\right\} .$$

*(f) Show that in the above expression*

$$\frac{\pi(\theta_{k+1},k+1)}{\pi(\theta_k,k)} = \frac{l(\theta_{k+1},k+1)}{l(\theta_k,k)} \times \frac{p(\lambda(k+1),t(k+1),k+1)}{p(\lambda(k),t(k),k)}$$

*where*

$$\begin{aligned} p(\lambda(l),t(l),l) &= p(\lambda(l)|l)\,p(t(l)|l)\,f_P(l)\,,\; l=1,2,\ldots, \\ \frac{p(\lambda(k+1)|k+1)}{p(\lambda(k)|k)} &= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{\lambda'_j\lambda'_{j+1}}{\lambda_j}\right)^{\alpha-1} e^{-\beta(\lambda'_j+\lambda'_{j+1}-\lambda_j)} \text{ and} \\ \frac{p(t(k+1)|k+1)}{p(t(k)|k)} &= \frac{2(k+1)(2k+3)}{T^2}\frac{(t^*-t_j)(t_{j+1}-t^*)}{t_{j+1}-t_j} . \end{aligned}$$

*(g) Show that* $B_{k,k+1}=p_k$, $B_{k+1,k}=q_{k+1}/(k+1)$ *and* $q(t^*,u)=1/T$ *and that the Jacobian is given by* $(\lambda'_j+\lambda'_{j+1})^2/\lambda_j$.

*(h) Show that the acceptance probability of a death move is*

$$\min\left\{1, \frac{\pi(\theta_k,k)}{\pi(\theta_{k+1},k+1)} \times \frac{B_{k,k+1}q(t^*,u)}{B_{k+1,k}} \times \left|\frac{\partial(\theta_k,t^*,u)}{\partial\theta_{k+1}}\right|\right\} .$$

**7.14** *Show that the Schmeiser and Chen (1991) algorithm corresponds to taking the proposal density g in the Phillips and Smith (1993) algorithm as* $g(c|\theta^{(j-1)},e^{(j)}) \propto \pi(\theta^{(j-1)}+ce^{(j)})$. *Discuss other possible forms for g, commenting on their advantages/disadvantages with respect to the above choices.*

**7.15** *Generate samples from the Potts model described in Example 7.13 varying the values of* $\beta$ *and using both a componentwise Markov chain sampler and the method of auxiliary variables. Compare the convergence of the generated samples from both approaches.*

# References

Abanto, C. A., Lopes, H. F. and Migon, H. S. (2005) Simulation-based sequential analysis for the bivariate stochastic volatility-volume model. Technical Report, Graduate School of Business, University of Chicago.

Abramowitz, M. and Stegun, I. A. (eds) (1965) *Handbook of Mathematical Functions*, National Bureau of Standards, Washington.

Achcar, J. A. and Smith, A. F. M. (1989) Aspects of reparametrisation in approximate Bayesian inference. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honour of George A. Barnard* (eds S. Geisser et al.), North Holland, Amsterdam, 439-52.

Aguilar, O. and West, M. (2000) Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, **18**, 338-57.

Ahrens, J. H. and Dieter, U. (1974) Computer methods for sampling gamma, beta, Poisson and binomial distributions. *Computing*, **12**, 223-46.

Aitchinson, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis*, Cambridge University Press, Cambridge.

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 111-42.

Al-Awadhi, F., Hurn, M. and Jennison, C. (2004) Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters*, **69**, 189-98.

Albert, J. H. (1988) Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, **83**, 1037-45.

Albert, J. H. (1996) A MCMC algorithm to fit a general exchangeable model. *Communications in Statistics - Simulation and Computation*, **25**, 573-92.

Anderson, T. W. (1958) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normality. *Journal of the Royal Statistical Society, Series B*, **36**, 99-102.

Andrieu, C., de Freitas, J. and Doucet, A. (2000) Reversible jump MCMC simulated annealing for neural networks. In *Uncertainty in Artificial Inteligence* (eds C. Boutilier and M. Goldszmidt), Morgan Kaufmann, San Francisco, 11-8.