CHAPTER 5

Gibbs sampling

5.1 Introduction

This chapter introduces the first widely used class of schemes for stochastic simulation using Markov chains. It is generically known as Gibbs sampling because it originated in the context of image processing. In this context, the posterior of interest for sampling is a Gibbs distribution. Borrowing concepts from Mechanical Statistics, the density of the Gibbs distribution can be written as

$$f(x_1, \ldots, x_d) \propto \exp\left[-\frac{1}{kT}E(x_1, \ldots, x_d)\right] \tag{5.1}$$

where k is a positive constant, T is the temperature of the system, E is the energy of the system, a positive function, and x_i is the characteristic of interest for the ith component of the system, $i = 1, \ldots d$. In Mechanical Statistics, x_i is the position or perhaps the velocity and position of the ith particle and in image processing it is (an indicator of) the colour of the ith pixel of an image.

The energy function E is commonly given by a sum of potential functions V. These sums operate over collections of subgroups of components over which each potential function is evaluated. The subgroups generally obey some neighboring relationship in their definition. This leads to a probability specification based on local properties, useful for modelling spatial interaction between components. The main drawback is the difficulty in the determination of the global properties, such as the normalizing constant.

Geman and Geman (1984) discuss this modelling problem extensively with special regard for sampling schemes and comparison with Markov random fields. Their sampling scheme explored the conditional structure implied by the local specification. Even though it was a well known and influential paper in the area, their paper was not published in a mainstream statistical journal. This is one of the few possible explanations for the delay in the introduction of their powerful results for the solution of Bayesian problems in general. Gelfand and Smith (1990) were the first authors to successfully point out to the statistical community at large that the sampling scheme devised by Geman and Geman (1984) for Gibbs distributions could in fact be used for a host of other posterior distributions. In that sense, it is somewhat misleading that the scheme retained the name

Gibbs sampling and Robert (2001) proposed to change it to Bayesian sampling. The paper by Gelfand and Smith (1990) also compared the Gibbs sampling scheme with the data augmentation algorithm (Section 4.9) and sampling-importance resampling (Section 3.5).

This chapter tries to describe the development of the area up to now. Some questions are still not entirely settled and there is a risk of obsolescence involved. The Gibbs sampling algorithm is described in the next section and some of its main properties presented. Section 5.3 deals with the description of implementation and convergence acceleration techniques. As previously discussed, one of the main difficulties when sampling via Markov chains is the verification of convergence of the chain. This problem is addressed in Section 5.4 where statistical techniques for convergence monitoring and identification are introduced. Most of the material from these two sections can be applied to any MCMC scheme, not just the Gibbs sampler. Section 5.5 applies Gibbs sampling to hierarchical, dynamic and spatial models. Finally, the chapter provides a brief description of the main software available for Bayesian inference using Gibbs sampling.

5.2 Definition and properties

Gibbs sampling is a MCMC scheme where the transition kernel is formed by the full conditional distributions. Assume as before that the distribution of interest is $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)'$. Each one of the components θ_i can be a scalar, a vector or a matrix.* Consider also that the full conditional distributions $\pi_i(\theta_i) = \pi(\theta_i|\theta_{-i})$, $i = 1, \ldots, d$ are available. This means that they are completely known and can be sampled from.

The problem to be solved is to draw from π when direct generation schemes are costly, complicated or simply unavailable but when generations from the π_i are possible. Gibbs sampling provides an alternative generation scheme based on successive generations from the full conditional distributions. It can be described in the following way:

1. Initialize the iteration counter of the chain $j = 1$ and set initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})'$;
2. Obtain a new value $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_d^{(j)})'$ from $\theta^{(j-1)}$ through successive generation of values

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}) , \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) , \\
&\vdots \\
\theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \ldots, \theta_{d-1}^{(j)}) ;
\end{aligned}$$

* The reader may prefer to think about them as scalars if that helps. This point is readdressed in Section 5.3.4.

3. Change counter j to $j + 1$ and return to step 2 until convergence is reached.

When convergence is reached, the resulting value $\theta^{(j)}$ is a draw from π. As the number of iterations increases, the chain approaches its equilibrium condition. Convergence is then assumed to hold approximately.

This presentation where each iteration consists of a single change to all components is favored by Gelfand and Smith (1990). The original work of Geman and Geman (1984) presented a chain with iterations formed by a change to a given component. Step 2 is obtained in the special case where the components are changed in a fixed and constant order.

The obvious form to obtain a sample of size n from π is to replicate n chains until convergence. Alternatively, after convergence all draws from a chain come from the stationary distribution. Therefore n successive values from this chain after the burn-in period will also provide a sample from π. The issue of how to form a sample is readdressed in more detail in the next section.

A typical trajectory of a Gibbs sampling chain is illustrated in Figure 5.1. An iteration is completed after d moves along the coordinate axes of the components of θ. Convergence diagnostics are complex as d can be very large and will be left for Section 5.4. Note that the convergence must be in distribution which means that the joint distribution of all parameter components must converge to the joint posterior for all values of θ. This exhaustive verification is far from trivial.

Example 5.1 *(Carlin, Gelfand and Smith, 1992) Let $y_1, \ldots, y_n$ be a sample from a Poisson distribution for which there is a suspicion of a change point m along the observation process where the means change, $m = 1, \ldots, n$. Given m, the observation distributions are $y_i|\lambda \sim Poi(\lambda)$, $i = 1, \ldots, m$ and $y_i|\phi \sim Poi(\phi)$, $i = m+1, \ldots, n$. The model is completed with independent prior distributions $\lambda \sim G(\alpha, \beta)$, $\phi \sim G(\gamma, \delta)$ and m uniformly distributed over $\{1, \ldots, n\}$ where α, β, γ and δ are known constants. The posterior density is*

$$\begin{aligned}
\pi(\lambda, \phi, m) &\propto f(y_1, \ldots, y_n|\lambda, \phi, m)p(\lambda, \phi, m) \\
&= \prod_{i=1}^{m} f_P(y_i; \lambda) \prod_{i=m+1}^{n} f_P(y_i; \phi) \quad f_G(\lambda; \alpha, \beta) f_G(\phi; \gamma, \delta) \frac{1}{n} \\
&\propto \prod_{i=1}^{m} e^{-\lambda} \lambda^{y_i} \prod_{i=m+1}^{n} e^{-\phi} \phi^{y_i} \lambda^{\alpha-1} e^{-\beta\lambda} \phi^{\gamma-1} e^{-\delta\phi} \\
&\propto \lambda^{\alpha+s_m-1} e^{-(\beta+m)\lambda} \phi^{\gamma+s_n-s_m-1} e^{-(\delta+n-m)\phi}
\end{aligned}$$

where $s_l = \sum_{i=1}^{l} y_i$ for $l = 1, \ldots, n$. It becomes simple to obtain the full conditional densities

$$\pi_\lambda(\lambda) = G(\alpha + s_m, \beta + m) ,$$

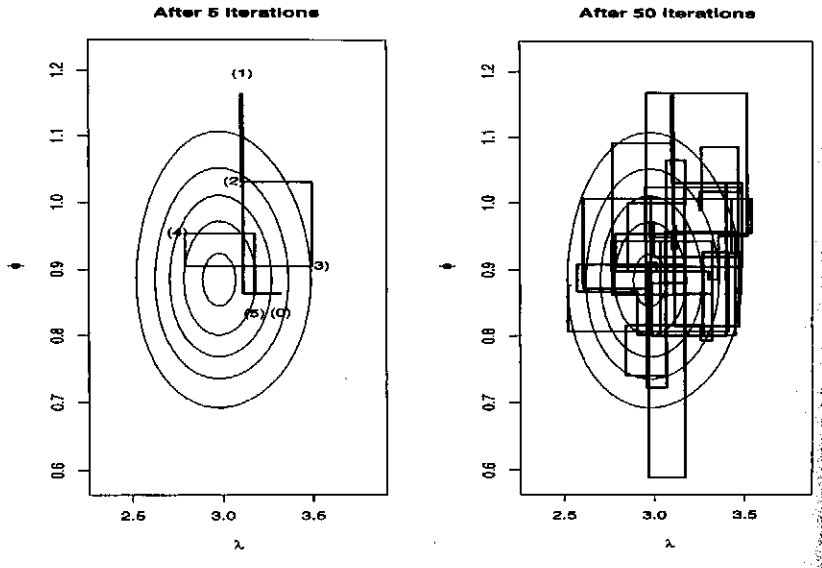

Figure 5.1 *Typical trajectories of the Gibbs sampler in a bidimensional parametric space, $d = 2$, after 5 and 50 iterations. The concentric curves represent the contour lines of the posterior density.*

$$\begin{aligned} \pi_\phi(\phi) &= G(\gamma + s_n - s_m, \delta + n - m), \\ \pi_m(m) &= \frac{\lambda^{\alpha+s_m-1}e^{-(\beta+m)\lambda}\phi^{\gamma+s_n-s_m-1}e^{-(\delta+n-m)\phi}}{\sum_{l=1}^{n}\lambda^{\alpha+s_l-1}e^{-(\beta+l)\lambda}\phi^{\gamma+s_n-s_l-1}e^{-(\delta+n-l)\phi}}, \end{aligned} \tag{5.2}$$

for $m = 1, \ldots, n$. It can be seen that λ and ϕ are conditionally independent given m, a posteriori. Thus, $\pi(\lambda|m) = \pi_\lambda(\lambda)$ and $\pi(\phi|m) = \pi_\phi(\phi)$. All these distributions are easily sampled from (see Chapter 1) and the iterative scheme repeating steps 1-3 can be operated without difficulty.

In this specific setting, it is possible to obtain the marginal posterior distributions analytically. It can be shown that (see Exercise 5.1)

$$\pi(m) \propto \frac{\Gamma(\alpha + s_m)\Gamma(\gamma + s_n - s_m)}{(m+\beta)^{\alpha+s_m}(n-m+\delta)^{\gamma+s_n-s_m}}, \tag{5.3}$$

for $m = 1, \ldots, n$. The normalizing constant is obtained by summing up the posterior probabilities and ensuring sum 1. Therefore $\pi(\lambda)$ and $\pi(\phi)$ can be analytically obtained as $\pi(\lambda) = \sum_{m=1}^{n}\pi(\lambda|m)\pi(m)$ and $\pi(\phi) = \sum_{m=1}^{n}\pi(\phi|m)\pi(m)$. For instance, $E_\pi(\lambda) = \sum_{m=1}^{n}E(\lambda|m,y)\pi(m)$, while $Var_\pi(\phi) = \sum_{m=1}^{n}Var(\phi|m,y)\pi(m)$.

This model can be applied to the $n = 112$ observations in Table 5.1 and

4	5	4	1	0	4	3	4	0	6	3	3	4	0	2	6	3	3	5	4
5	3	1	4	4	1	5	5	3	4	2	5	2	2	3	4	2	1	3	2
2	1	1	1	1	3	0	0	1	0	1	1	0	0	3	1	0	3	2	2
0	1	1	1	0	1	0	1	0	0	0	2	1	0	0	0	1	1	0	2
3	3	1	1	2	1	1	1	1	2	4	2	0	0	0	1	4	0	0	0
1	0	0	0	0	0	1	0	0	1	0	1								

Table 5.1 *Counts of coal mining disasters in Great Britain by year from 1851 to 1962 (Jarret, 1979).*

	True			*Gibbs*		
Par.	*Mean*	*Var*	*95% C.I.*	*Mean*	*Var*	*95% C.I.*
λ	*3.120*	*0.280*	*(2.571,3.719)*	*3.131*	*0.290*	*(2.582,3.733)*
ϕ	*0.923*	*0.113*	*(0.684,0.963)*	*0.922*	*0.118*	*(0.703,1.167)*
m	*1890*	*2.423*	*(1886,1895)*	*1890*	*2.447*	*(1886,1896)*

Table 5.2 *Exact and approximate posterior quantities. Gibbs results are based on 5000 draws starting at $m^{(0)} = 1891$. C.I. stands for credibility interval.*

Figure 5.2(a). Figure 5.2(b) shows some evidence of a changing pattern around 1890 more clearly.

The Gibbs sampler was implemented and run for 5000 iterations, starting at $m^{(0)} = 1891$ and hyperparameters $\alpha = \beta = \gamma = \delta = 0.001$. Figure 5.1 exhibits the trajectories at selected iterations.

Inference can be approximately performed via MCMC by using all the 5000 iterations of the chain. The MCMC approximation to the marginal posterior distributions of λ, ϕ and m appear in Figures 5.2(c) and 5.2(d). It can be seen that with high probability a change point occurs around 1891 and that the Poisson rates are quite distinct before and after this change point. True values and MCMC approximations to the posterior means, posterior variances and posterior 95% credibility intervals of λ, ϕ and m appear in Table 5.2.

Effective sample sizes based on λ, ϕ and m are 4800, 3950 and 4900, respectively. They are very close to the actual number of iterations of the chain, reflecting the very low autocorrelation structure of this Gibbs sampler. They are calculated with theoretical chain autocorrelations ρ_k used in (4.10) approximated by their moment estimates $\hat{\rho}_k$ (see Equation 4.16).

A few basic facts must be established beforehand. First, the Gibbs sampler does define a Markov chain. This is clearly the case as the probabilistic change at iteration j depends only on chain values at step $j-1$. Also, the chain is homogeneous as the transitions are only affected by the iteration

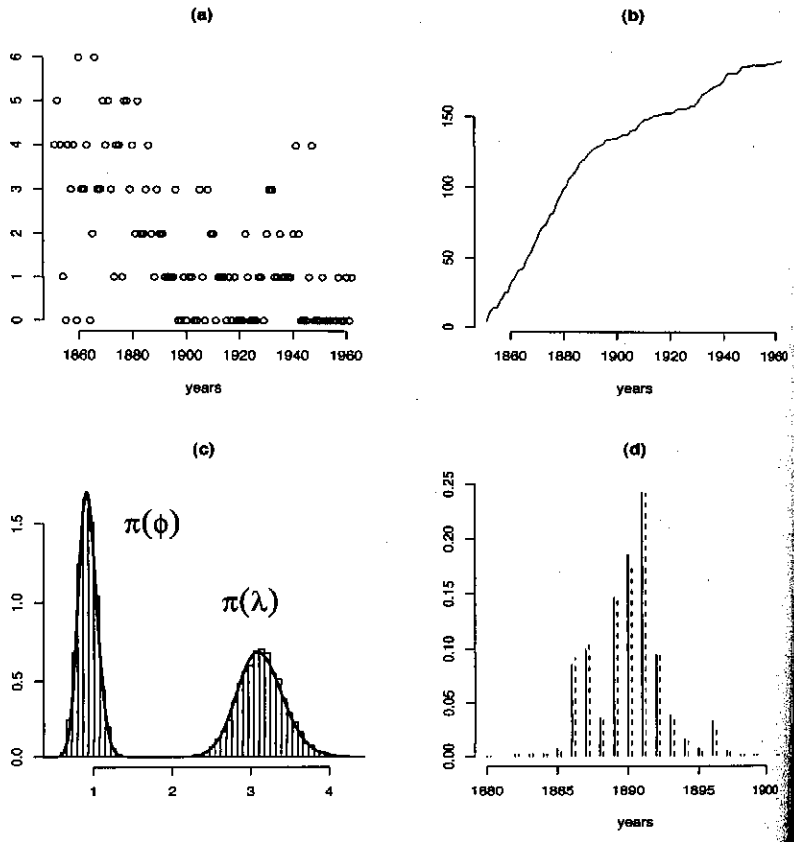

Figure 5.2 *Poisson observations with change point. (a) counts of coal mining disasters in Great Britain by year from 1851 to 1962, (b) cumulative counts, (c) true (solid lines) and histogram approximations of the marginal posterior distributions of λ and ϕ, (d) true (solid lines) and histogram approximations (dashed lines) of the marginal posterior distribution of m.*

through the chain values. It is not difficult to obtain the transition kernel as

$$p(\theta, \phi) = \prod_{i=1}^{d} \pi(\phi_i | \phi_1, \ldots, \phi_{i-1}, \theta_{i+1}, \ldots, \theta_d) \tag{5.4}$$

which clearly depends on the iterations only through the chain values θ and ϕ. This chain with a complete scan over all components is not reversible

although each individual change is reversible. Green (1995) pointed out that the chain can be made reversible by taking each iteration of the chain to consist of the complete scan through the components followed by another scan through the components in reversed order (Besag, 1986).

Another important result is the derivation that the equilibrium distribution of a chain with transition kernel (5.4) is π. This result was derived in a special case of $d = 2$ discrete components in Example 4.7. In the continuous case, the same argument cannot be used but the mechanics of the algorithm is the same. If a Markov chain with transition kernel $p(\theta, \phi)$ has limiting distribution π^∞, then the stationarity condition (4.14) must be satisfied by p and π^∞ and the chain must be irreducible. Irreducibility is easy to verify for each application by checking that $P(x, A) > 0$ for all sets A with positive posterior probability. For statistical applications, it is generally satisfied but see Example 5.2 below for a counterexample.

To check stationarity, let $\theta = (\theta_1, \theta_2)$ with marginal limiting densities $\pi^\infty(\theta_1)$ and $\pi^\infty(\theta_2)$. The limiting full conditional distribution of θ_1 is $\pi_1^\infty(\theta_1)$. The transition kernel (5.4) simplifies to

$$p(\theta, \phi) = \pi(\phi_1 | \theta_2) \pi(\phi_2 | \phi_1)$$

where $\phi = (\phi_1, \phi_2)$ has components with the same dimensions as those of θ. As $\int\int \pi(\phi_2|\phi_1)\pi^\infty(\theta_1|\theta_2)d\theta_1 d\phi_2 = \int \pi(\phi_2|\phi_1)d\phi_2 \times \int \pi^\infty(\theta_1|\theta_2)d\theta_1 = 1$,

$$\begin{aligned} \pi^\infty(\theta_2) &= \int\int \pi(\phi_2|\phi_1)\pi^\infty(\theta_1|\theta_2)\pi^\infty(\theta_2)d\theta_1 d\phi_2 \\ &= \int\int \pi(\phi_2|\phi_1)\pi^\infty(\theta)d\theta_1 d\phi_2 \,. \end{aligned} \tag{5.5}$$

Integrating (4.14) with respect to ϕ_2 gives the marginal limiting density of ϕ_1 as

$$\begin{aligned} \pi^\infty(\phi_1) &= \int\int p(\theta, \phi)\pi^\infty(\theta)d\theta d\phi_2 \\ &= \int\int\int \pi(\phi_1|\theta_2)\,\pi(\phi_2|\phi_1)\pi^\infty(\theta)d\theta_1 d\theta_2 d\phi_2 \\ &= \int \pi(\phi_1|\theta_2)\pi^\infty(\theta_2)d\theta_2 \end{aligned} \tag{5.6}$$

where the last equality follows from (5.5). The only distribution satisfying (5.6) must have $\pi(\phi_1|\phi_2) = \pi^\infty(\phi_1|\phi_2)$. The same argument could be used to give $\pi(\phi_2|\phi_1) = \pi^\infty(\phi_2|\phi_1)$ and the limiting distribution must have the same full conditionals as the posterior. The same argument follows for θ divided into d blocks of components as these can always be rearranged in two blocks θ_i and θ_{-i}. This means that all limiting full conditionals are given by the posterior full conditionals. This does not in general guarantee that $\pi^\infty = \pi$ (see Exercise 5.3 for an example where it fails). Nevertheless, Besag (1974) showed that under very mild conditions, the set of all full

conditional distributions determine the joint distribution. Therefore, the Markov chain with transition kernel (5.4) converges to the distribution of interest π and the iterative sampling scheme with steps 1-3 above draws in the limit a value from this distribution.

Formal convergence conditions for the Gibbs sampler were established by Roberts and Smith (1994) and Tierney (1994). The results are presented in terms of continuous parameter spaces but can be extended for combinations of continuous and discrete parameters (Example 5.1). A simple example of a reducible chain where convergence fails is given below.

Example 5.2 *(O'Hagan and Forster, 2004; Roberts, 1996) Consider* $\theta = (\theta_1, \theta_2)$ *uniformly distributed over two disjoint regions* $A = A_1 \times A_2$ *and* $B = B_1 \times B_2$ *of the plane with probabilities* p_A *and* p_B *adding up to 1. Consider also that* A_1 *and* B_1 *are disjoint regions on the* θ_1 *axis and* A_2 *and* B_2 *are disjoint regions on the* θ_2 *axis. This implies that full conditionals are also uniform but over regions that depend on the starting point. Chains that start in* A_1 *will lead to sampling* θ_2 *uniformly over* A_2 *which implies sampling* θ_1 *uniformly over* A_1 *and this situation perpetuates itself. Points from B will never be reached. Analogously, chains starting in* B_1 *will never reach points in A. This chain is clearly reducible and will have uniform limiting distribution over A or B depending on the starting point.*

5.3 Implementation and optimization

Despite the theoretical results ensuring the convergence of the Gibbs sampler, its practical implementation may be complicated by the potential complexity of the models considered. Convergence of the sampler becomes difficult to characterize. Given that it is a numeric and iterative method, practical strategies to improve the efficiency of the method may have a considerable impact on its computational cost. Efficiency broadly consists of reducing the number of burn-in iterations and the amount of arithmetic operations required at each iteration. The techniques presented are related to the basic MCMC methods as described in the previous chapter and represented by the Gibbs sampler. More general techniques using other forms of chains will be presented in Chapter 7.

5.3.1 Forming the sample

The previous section presented two forms to obtain a sample of size n from the posterior distribution π. The obvious one is to process n chains in parallel until convergence, say after m iterations, and take as sample elements the mth chain value from each of the n chains. The generation procedure will then require mn generations from the chain. If chains are initialized independently, the sample consists of independent values from

π. Independence is easier to establish if the initial values are all different and preferably with larger dispersion than in the posterior (Section 5.4.3).

Another form is to consider a single chain and explore ergodic results. After convergence, all chain values have marginal distribution given by the equilibrium distribution π. So, a sample of size n may be formed by n successive values from this chain. This generation will require $m+n$ generations from the chain. This is substantially less than independent sampling. The difficulty here is that the sample elements are no longer independent due to chain dependence. Ergodic theorems ensure that inference based on this sample is still valid. From a practical point of view, there may be problems if the chain autocorrelation is too high and the sample is not large enough to acknowledge it. In these cases, chains may take too long to adequately cover the entire parameter space appropriately. As a result, some relevant regions may be underrepresented in the sample.

An alternative approach accommodating independence is to take for the sample chain values at every kth iteration after the burn-in period. Markovian processes only have first order dependence. As the lag between iterations increases, chain values become less and less correlated and are virtually independent for a large enough value of the lag k. A sample of size n with *quasi*-independent elements thus requires $m + kn$ generations from the chain. The value of k is typically smaller than m and again an improvement over independent sampling is obtained. There is no gain in efficiency, however, by this approach and estimation is shown below to be always less precise than retaining all chain values. This procedure is advantageous if computer storage of values is limited. Useful indicators of dependence are given by the chain autocorrelations (Section 4.8). A formal approach for selection of k is given in Section 5.4.4.

Another compromise is to take a small number l, say less than 10, of independent chains, run them until convergence and then retain from each of them n/l successive values from the sample. This will lead to a sample of size n obtained after $l[m + (n/l)] = lm + n$ generations from the chains. Yet another variant is obtained by retaining every kth chain value after convergence with a total of $l[m + (n/l)k] = lm + kn$ generations. In computational terms, there are efficiency losses with respect to using a single chain and gains with respect to independent sampling.

The independent sampling approach was suggested by Gelfand and Smith (1990) and used by some authors shortly afterwards. The single chain approach was emphatically advocated by Geyer (1992) backed by ergodic theorems. Sampling every kth iteration was discussed by Raftery and Lewis (1992). Gelman and Rubin (1992a) recommended the use of a small number of independent chains backed by an example from Gelman and Rubin (1992b) where single chains provide indication of convergence of ergodic averages to different limits. Gelman et al. (2004) argue that somehow the

benefits from *quasi*-independent sampling are diluted when running few chains.

There is no general agreement on the subject although it is generally agreed that running n parallel chains in practice is computationally inefficient and unnecessary. The main debate is whether a few parallel chains are needed. If the convergence properties of the chain are well understood then clearly a single chain suffices. As these characteristics are hard to obtain, prudence suggests that a few pilot parallel chains should be run. If they quickly settle around common values then a single chain can be safely used to extract a large sample for inference. Otherwise, there may be minor characteristics of the posterior distribution such as secondary modes far from the mode that require very large samples to be noticed. In this case, these parallel chains should be run longer and their values should be retained for the sample. Convergence diagnostics are the subject of the next section and these points will be returned to in more detail there.

5.3.2 Scanning strategies

The Gibbs sampler described in the previous section involved a complete scan over the components. All iterations consisted of visits to update the components in the same deterministic order, typically $1 \to 2 \to \cdots \to d$. There are many other possible scanning or updating strategies for visiting the components of θ.

Geman and Geman (1984) proved convergence to the joint distribution in a discrete setting for all visiting schemes that guarantee that all components are visited i.o. when the chain is run indefinitely. The reversible Gibbs sampler where at each iteration each component is visited in a fixed order and then visited again in reversed order satisfies this property. In this case, each iteration consists of $2d$ updates and comparisons between strategies should bear that in mind.

Another scheme where an i.o. schedule is guaranteed draws a number i from $\{1, \ldots, d\}$ with fixed positive probabilities at each iteration and only updates the θ_i at that iteration. To make it more comparable with the deterministic scan, an iteration of these random scans can be defined by a collection of d such updates.

Roberts and Sahu (1997) consider a random permutation scan where at each iteration a permutation of $\{1, \ldots, d\}$ is selected and components are visited in that order. Zeger and Karim (1991) describe a Gibbs sampling scheme where some components were visited only every kth iteration. This also guarantees an i.o. visiting schedule for fixed, finite k.

Assume now that π is a multivariate normal distribution with precision matrix $\Phi = (\phi_{ij})$. For this setting, Roberts and Sahu (1997) showed that convergence for the deterministic scan is faster than for the random scan if Φ is tridiagonal ($\pi(\theta_i|\theta_{-i}) = \pi(\theta_i|\theta_{i-1}, \theta_{i+1})$, for all i) or if Φ has non-

negative partial correlations ($\phi_{ij} \leq 0$). This result is particularly important because both dynamic and hierarchical models lead to tridiagonal matrices if variances are known. Their results also indicate that more precise distributions lead to faster convergence both for the deterministic and random scans.

5.3.3 Using the sample

Whatever the scheme chosen for forming the sample, after it is used a sample of vectors $\theta_1, \ldots, \theta_n$ generated from the posterior distribution π is available. Assume also the more general case where these are successive values from a single Markov chain. A sample from the ith component of θ is given by $\theta_{1i}, \ldots, \theta_{ni}$. Marginal point or interval summaries of any real function $\psi = t(\theta)$ are estimated by their corresponding estimators based on the sample. This is always a consistent estimator by the ergodic theorem (4.6). The quality of this estimator can be judged by the central limit theorem (4.11) from where approximate confidence intervals about the MCMC estimates may be formed.

So, the posterior mean of ψ is estimated by $\hat{E}(\psi) = \hat{\psi} = (1/n)\sum_{j=1}^{n} \psi_j$ where $\psi_j = t(\theta_j)$, $j = 1, \ldots, n$. The posterior variance of ψ is similarly estimated by noting that $\sigma_\psi^2 = Var(\psi) = E(\psi^2) - [E(\psi)]^2$. Each expectation is estimated by an application of (4.6) and σ_ψ^2 is estimated by $\hat{\sigma}_\psi^2$ where

$$\hat{\sigma}_\psi^2 = \hat{E}(\psi^2) - [\hat{E}(\psi)]^2 = \frac{1}{n}\sum_{j=1}^{n}(\psi_j - \hat{\psi})^2 ,$$

the sample variance. The denominator n may be replaced by $n-1$ but this change is irrelevant for two reasons. First, typically n is large which makes the change immaterial. Second, it does not remove the bias of the estimator as usually happens when independent sampling is performed.

Consider again the problem of choosing between a sample of n successive values and a sample of $m = n/k$ values obtained by skipping every kth iteration. Note that k such sub-samples with *quasi*-independent draws are formed. Denote by $\hat{\psi}_1, \ldots, \hat{\psi}_k$ the averages of the sub-samples and $\hat{\psi} = (1/k)\sum_{j=1}^{k} \hat{\psi}_j$ the average over the complete sample. There are $k+1$ estimators of $E(\psi)$ and they are all consistent estimators by the ergodic theorem. It can be shown that $Var(\hat{\psi}) \leq Var(\hat{\psi}_j)$, for all j (O'Hagan and Forster, 2004; MacEachern and Berliner, 1994). This means that independence sampling comes at the expense of reduced efficiency.

Credibility intervals are similarly obtained by estimating the interval limits by the respective sample quantiles. As in Section 3.5, if $n = 1000$, and an equal tails 95% probability interval for ψ is required, it can be estimated by the interval with limits given by the 25th and 975th largest

sample values of ψ. Again, these values are consistent estimators of the 0.025 and 0.975 quantiles of ψ by (4.6).

All above estimators have a sampling distribution that is approximated by (4.11). The asymptotic variance of these estimators, which is different from the posterior variance, can be estimated by the methods described in Section 4.8. The central limit theorem ensures that estimation errors are $O(n^{-1/2})$.

The marginal densities $\pi(\theta_i)$ can be estimated by (a smoothed version of) the histogram of sampled values of θ_i (Section 3.5). Better estimators can be obtained by using conditional distributions. Recalling that $\pi(\theta_i) = \int \pi(\theta_i|\theta_{-i})\pi(\theta_{-i})d\theta_{-i}$, a Monte Carlo estimator is given by

$$\hat{\pi}(\theta_i) = \frac{1}{n}\sum_{j=1}^{n} \pi(\theta_i|\theta_{j,-i}) \tag{5.7}$$

where the $\theta_{j,-i}$, $j = 1, \ldots, n$ are a sample from the marginal $\pi(\theta_{-i})$. Notice that Equation (5.7) is a generalization of the results from Section 3.4 for dependent samples. Again, the ergodic theorem ensures that $\hat{\pi}$ is a consistent estimator and it obeys a central limit theorem for every value of θ_i. These estimators are always continuous for continuous parameters. More importantly, they are based on information about the form of the posterior. For that reason, Gelfand and Smith (1990) call it a Rao-Blackwellized density estimator. This is a reference to the Rao-Blackwell theorem that states that estimators are always improved (in the sense of reducing sampling variance) by conditioning on sufficient statistics. They proved the result for density estimation in the context of independent sampling. The general proof of the result for Markov chain sampling is given by Liu, Wong and Kong (1994). The same idea can be used to obtain better estimates of moments of $t(\theta_i)$ through

$$\hat{E}[t(\theta_i)] = \frac{1}{n}\sum_{j=1}^{n} E[t(\theta_i)|\theta_{j,-i}]$$

although the gains are not as large here.

5.3.4 Reparametrization

Going back to Figure 5.1, an iteration is formed by moves along the coordinate axes of the components of θ. If there is weak dependence between the components, the moves will be ample. The chain will then move freely through the parametric space and convergence will be fast. An extreme case is posterior independence between the components. The full conditionals are equal to the marginals and convergence is immediate.

Often, the posterior structure leads to high correlation between some of the components of θ (Section 5.5.2). Figure 5.3 illustrates this point

for a bidimensional parameter. The contours of the posterior show strong dependence between the components of θ and chain moves, governed by the conditional densities, will be small. The chain will take many iterations to adequately cover the parametric space and as a result convergence is slow. In this case, the Gibbs sampler will be inefficient. Examples can be constructed in larger dimension models where convergence can be slowed to any arbitrary amount of iterations (Shephard, 1994).

A simple and sometimes effective way to reduce convergence time is to use reparametrizations. This point was discussed in Chapter 3 in the context of improving approximations. Adequate transformations in the parameter space may produce situations of near independence that are ideal for fast convergence of the chain. Unfortunately, there are no rules to determine suitable transformations but frequently linear transformations that produce a diagonal variance matrix provide good results. Two important classes of models where these transformations can be found are presented below.

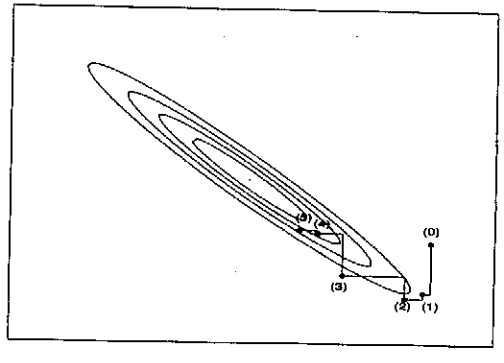

Figure 5.3 *Contour lines of a bivariate posterior density with components highly correlated. A possible chain trajectory is also depicted to illustrate slow convergence, with iterations in parentheses. The contours are from a bivariate normal distribution with marginal distributions* $\theta_1 \sim N(2,1)$ *and* $\theta_2 \sim N(3,1)$ *and correlation* -0.97. *The trajectory is obtained by sampling from the full conditional distributions* $\theta_2|\theta_1$ *and* $\theta_1|\theta_2$.

Example 5.3 *For the regression model described in Section 2.3, the conditional posterior was* $\pi(\beta|\phi) = N(b_\phi, B_\phi)$ *and therefore posterior correlations depend on the posterior variance* B_ϕ. *Numerical techniques described in Section 1.4 can be applied to obtain the square root matrix* A_ϕ *such that*

$A_\phi A'_\phi = B_\phi^{-1}$. So, $\alpha = A_\phi \beta \sim N(A_\phi b_\phi, I_d)$ given ϕ and the components of vector α are independent a posteriori given ϕ.

In the case of simple linear regression $y_i = \beta_1 + \beta_2 x_i + e_i$ with non-informative prior $p(\beta) \propto k$, α has components $\alpha_1 = \beta_1 + \beta_2 \bar{x}$ and $\alpha_2 = \beta_2$. This is equivalent to centering covariates and working with the model $y_i = \alpha_1 + \alpha_2(x_i - \bar{x}) + e_i$. Depending on how close the values of x_i are, the plot of the posterior conditional density of $\beta|\phi$ will present contour lines that are as concentrated as those exhibited in Figure 5.3. In multiple linear regression, centering covariates is usually enough.

The conditional dependence between components of α is removed and the only remaining posterior correlation is between α and ϕ. Sampling components of α is as simple as sampling components of β but the chain will converge faster. Once α is sampled, a sample value of β is obtained by $\beta = A_\phi^{-1}\alpha$.

Example 5.4 *Consider the hierarchical (or random effects) model*

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2)$$

where the random effects α_i associated with the observation groups have distribution $\alpha_i \sim N(0, \tau^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. Note that by writing $\beta_i = \mu + \alpha_i$ and $\beta = \mu$, the one-way classification model of Example 2.8 is recovered. Assume also a non-informative prior $p(\mu) \propto k$. It can be shown that given variance components σ^2 and τ^2, posterior correlations between model parameters are given by

$$Cor(\mu, \alpha_i) = -\left(1 + \frac{\sigma^2/n_i}{\tau^2/m}\right)^{-1/2} \quad \text{and} \quad Cor(\alpha_i, \alpha_j) = \left(1 + \frac{\sigma^2/n_i}{\tau^2/m}\right)^{-1}.$$

High posterior correlations occur if $\sigma^2/n_i \ll \tau^2/m$. Roberts and Sahu (1997) showed that asymptotically (as $m \to \infty$) the deterministic scan over the parameters has faster convergence than random scans.

Gelfand, Sahu and Carlin (1995) suggest the hierarchical parametrization with the β_i replacing the α_i and show that given the variance components σ^2 and τ^2, posterior correlations between model parameters are given by

$$Cor(\mu, \beta_i) = -\left(1 + \frac{m\tau^2}{\sigma^2/n_i}\right)^{-1/2} \quad \text{and} \quad Cor(\beta_i, \beta_j) = \left(1 + \frac{m\tau^2}{\sigma^2/n_i}\right)^{-1}.$$

Now, low correlations are obtained for the conditions of high correlations in the original parametrization. So, depending on the data structure, it is more appropriate to work on another parametrization. Theoretical support is provided by Roberts and Sahu (1997). Gelfand, Sahu and Carlin (1995) argue that in practice the presence of random effects implies excess randomness and therefore it is expected that σ^2/n_i will be smaller than τ^2/m, which would justify the use of the β_i. They extended their approach to nested random effects models. Generalized linear models with random effects were studied by Gelfand, Sahu and Carlin (1996) and, even though analytic results are no longer available, they arrived at the same qualitative recommendations.

Vines, Gilks and Wild (1996) suggest the reparametrization $\nu = \mu - \bar{\alpha}$ and $\xi_i = \alpha_i - \bar{\alpha}$ and show that given the variance components σ^2 and τ^2, posterior correlations between model parameters are given by

$$Cor(\nu, \xi_i) = 0 \quad \text{and} \quad Cor(\xi_i, \xi_j) = -\frac{1}{m}.$$

Now, correlations have the advantage of not depending on the variance components. Again, the idea can be extended to more general models. For this parametrization, convergence with the random permutation scans is faster than for all other scanning strategies.

These examples suggest a general strategy based on approximate posterior normality (Section 3.2). The approximate variance V is a first order approximation for the posterior variance. Its square root matrix A can be calculated and a linear transformation $\alpha = A^{-1}\theta$ operated. This will provide approximate posterior independence to the first order. In more general models, in addition to the computational cost of finding A, there is also the added cost of sampling α instead of θ. Other approximations to the posterior variance may be sought. Hills and Smith (1992) suggest using the sample variance obtained from a pilot chain.

Another simple but important point is to observe the structure of the model and parameters. For example, these transformations will provide a better result if the posterior for each parameter behaves like the normal distribution. Variance parameters will not have this behavior unless a large number of observations is collected. Otherwise, a logarithmic transformation is recommended before application of the orthogonalization procedures above. Optimal strategies for some parameters should not be blindly applied for other parameters.

5.3.5 Blocking

So far, nothing has been said about the choice of components that form the parameter vector θ. In principle, the way the components are arranged in blocks of parameters is completely arbitrary and includes as a special case blocks formed by scalar components. The structure of Gibbs sampling, also illustrated in Figure 5.1, makes moves according to the coordinate axes of the blocks. Scalar blocks lead to moves along each component of θ. Larger blocks allow moves in more general directions. This can be very beneficial computationally when there is high correlation between components. The slow, componentwise moves may be replaced by fast moves incorporating the information about dependence between components. These moves are dictated by the joint full conditional for the block of parameters considered,

which incorporates the correlation structure. This intuitive consideration is confirmed by the theoretic results of Liu, Wong and Kong (1994). They showed that estimates obtained by blocking components are generally more precise than those obtained by treating each component separately.

Derivations of Roberts and Sahu (1997) show that, for random scans, convergence improves as the number of blocks decreases. Thus, blocking is beneficial. They also proved that blocking is beneficial for non-negative partial correlation distributions and more so as the partial correlation of the components in the block gets larger. These results were obtained only for a multivariate normal π and extrapolations should be made with caution. They also provided an example where blocking worsens convergence.

Although it is hard to determine optimal blocking strategies, some basic rules should be followed. When a parametric vector or matrix is specified in block, it generally has joint full conditionals that are easy to sample from. The important message is to block as much as possible for sampling. Of course, if the complete parameter vector forming a single block could be sampled, there would be no need for Gibbs sampling! Therefore, the only restriction is the ability to sample from the full conditional distributions formed.

5.3.6 Sampling from the full conditional distributions

In some cases, the form of the full conditional distribution is not recognizable which prevents sampling via the conventional algorithms. Chapter 1 presented a host of other general-purpose options such as (adaptive) rejection and reweighted sampling methods. According to Carlin and Louis (2000), these situations are an indication that an altogether different approach should be applied instead of insisting on Gibbs sampling.

Ritter and Tanner (1992) developed yet another sampling scheme from difficult full conditionals. Their approach is similar to adaptive rejection by being based on the evaluation of the full conditional at a few selected points. For that reason, they called it the griddy Gibbs sampler. Let $\pi_i(\theta_i)$ be a difficult full conditional distribution. Then, sampling from π_i can be approximately performed as follows:

1. Take a grid of points $\theta_{i1}, \ldots, \theta_{im}$, evaluate $\pi_i(\theta_{ij})$, $j = 1, \ldots, m$, and normalize them to obtain weights $w_1, \ldots, w_m$.
2. Use the weights $w_1, \ldots, w_m$ to construct a simple approximation to the distribution function of π_i.
3. Draw a value from π_i by the probability integral transform method (Section 1.3).

There are many possibilities for the construction in step 2. The simplest one is to use piecewise constant functions (discrete distribution). Piecewise linear functions are also easy to sample from and allow for continuous

sampling. Higher order polynomials and even splines may be used but it is important to keep it simple to sample from. Tanner (1996) suggests that the number of points m should be kept small for the burn-in period of the chain and doubled after convergence for good approximations only when it matters. He also suggests an adaptive scheme to revise the selected grid to include more points in higher density regions.

5.4 Convergence diagnostics

As previously discussed, a value from the distribution of interest π is only obtained when the number of iterations of the chain approaches infinity. In practice this is not attainable and a value obtained at a sufficiently large iteration is taken instead of being drawn from π. The difficulty is the determination of how large this iteration should be. There is no simple answer to this question and most efforts have been directed at studying as close as possible the convergence characteristics of the chain. Most results below can be applied to any MCMC method although for a few of them the use of Gibbs sampling is required.

There are two main ways to approach the study of convergence. The first one is more theoretical and tries to measure distances and establish bounds on distribution functions generated from a chain. In particular, one can study the total variation distance between the distribution of the chain at iteration j and the limiting distribution π. Special aspects derived from the probabilistic structure of the chain can also be studied. This approach was pursued by Meyn and Tweedie (1994), Polson (1996), Roberts and Polson (1994), Roberts and Tweedie (1994) and Rosenthal (1993) to cite just a few papers (see also the references in those papers). This is an area that is certainly going to grow as we increase our understanding of the subject. At the moment, however, the results have had little impact on practical work (Cowles and Carlin, 1996).

The study of convergence of the chain can also be approached from a statistical perspective, i.e., by analyzing the properties of the observed output from the chain. This is an empirical as opposed to a theoretical treatment of the problem and is obviously more practical. The difficulty with this approach is that it can never guarantee convergence because it is only based on observations from the chain (see Example 5.5 below).

Although the two approaches to the study of convergence are valid and complement each other, theoretical results have proved to be more difficult to obtain and apply to practical problems. This book will provide a more detailed description of the convergence diagnostics based on the statistical properties of the observed chain. Cowles and Carlin (1996) and Brooks and Roberts (1998) provide comparative and illustrative reviews of many of these methods. Robert (1995) reviews some possibilities involving the two approaches.

5.4.1 Rate of convergence

Geman and Geman (1984) showed for the discrete case that the Gibbs sampler is a uniformly ergodic Markov chain. Uniform ergodicity determines an exponential rate of convergence of the chain to the limiting distribution and could be taken as an indication of fast convergence. However, there is no indication of control over the rate of convergence and the Gibbs sampler can have an extremely slow convergence in some cases. The example below illustrates this point in a very simple context.

Example 5.5 *(O'Hagan and Forster, 2004) Consider again the situation of Example 4.7 where $\theta = (\theta_1, \theta_2)'$ is bivariate and $\pi(\theta)$ is given by the table of probabilities below*

	θ_2	
θ_1	*0*	*1*
0	$p/2$	$(1-p)/2$
1	$(1-p)/2$	$p/2$

Observe that $\pi(\theta_i) = bern(1/2)$, $i = 1, 2$, and the posterior correlation between θ_1 and θ_2 is $\rho = 2p - 1$. Using properties of the Gibbs sampler, it is easy to obtain that $Pr(\theta_1^{(j)} = 1|\theta_1^{(j-1)} = 1) = Pr(\theta_1^{(j)} = 0|\theta_1^{(j-1)} = 0) = p^2 + (1-p)^2$ and, consequently, $Pr(\theta_1^{(j)} = 1|\theta_1^{(j-1)} = 0) = Pr(\theta_1^{(j)} = 0|\theta_1^{(j-1)} = 1) = 2p(1-p)$. Taking $p_j = Pr(\theta_1^{(j)} = 1)$ gives $p_j = \rho^2 p_{j-1} + b$ where $b = 2p(1-p)$. The solution for p_j is $p_j = \rho^{2(j+1)} p_0 + b(1 - \rho^{2(j+1)})/(1-\rho^2)$.

The transition matrix formed by the marginal chain $(\theta_1^{(j)})_{j \geq 0}$ has eigenvalues 1 and ρ and therefore the rate of convergence is $|\rho|$. Ergodicity of the chain is ensured if $p > 0$ but if p is close to 1 or 0, this rate will be close to 1, the chain will tend not to move and convergence to the limiting distribution is very slow. In the limit, $p_j \to b/(1-\rho^2) = 1/2$ as expected. However, if $p = 0.999$, $\rho = 0.998$ and, after 100 iterations, $p_{100} = 0.667 p_0 + 0.165$ which is still far from the appropriate limit.

The point raised in this example is far from rare in many applications and although correlations as high as 0.998 are not common, a similar effect is obtained with high dimensional parameter spaces with much smaller correlations. Once again, it seems sensible in these cases to reparametrize the model. This point was already discussed in the previous sections and will be returned to in Section 5.5.

5.4.2 Informal convergence monitors

Gelfand and Smith (1990) suggested a few informal checks of convergence based on graphical techniques. After m iterations in n parallel chains, a histogram of the n values of the mth iterates of a given function of θ can be plotted. This function can be one of the components of θ and the histogram may be smoothed if desired. The procedure is repeated after a further k iterates are obtained in the chains. The value of k does not need to be large if one suspects convergence after m iterations. It cannot be low as the chain correlation will still be affecting possible similarities of the histograms. Typically, values between 10 and 50 are reasonable. Convergence is accepted if the histograms cannot be distinguished.

Same ideas can be used with a single chain. A trajectory of the chain exhibiting the same qualitative behavior through iterations after a transient initial period is an indication of convergence. Similarly, the trajectory of the ergodic averages can be evaluated and plotted. An asymptotic behavior over many successive iterations indicates convergence. Figure 5.4 shows the ergodic averages of variance components in a nested random effects model (Gamerman, 1997). The indication of convergence for both components seems to be very clear.

Similar ideas can be used with graphical representations of the simulated values of a few chosen (transformations of) parameters. The resulting plots provide a rough indication of stationarity behavior when the sequence of values tends to concentrate around the same pattern. This visual impression can be reinforced when chains started at different values oscillate in the same region.

Example 5.6 *Souza (1999) considers a number of hierarchical and dynamic models to describe the nutritional pattern of pregnant women. The data depicted in Figure 5.5 consist of the weight gains of $I = 68$ pregnant women at 5 to 7 visits to the Instituto de Puericultura e Pediatria Martagão Gesteira from the Universidade Federal do Rio de Janeiro. One of the simplest models she adopted was the simple hierarchical regression on time where*

$$\begin{aligned} y_{ij}|\alpha_i, \beta_i, \phi &\sim N(\alpha_i + \beta_i t_{ij}, \sigma^2), \\ (\alpha_i, \beta_i)'|\alpha, \beta &\sim N((\alpha, \beta)', diag(\tau_\alpha^{-1}, \tau_\beta^{-1})), \\ (\alpha, \beta)' &\sim N((0,0)', diag(P_\alpha^{-1}, P_\beta^{-1})), \end{aligned}$$

prior independent scale parameters σ^{-2}, τ_α and $\tau_\beta \sim G(a, b)$ and y_{ij} and t_{ij} are the jth weight measurement and visit time of the ith women, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$. Here, $n = \sum_{i=1}^{N} n_i = 427$, $P_\alpha = P_\beta = 1/1000$ and $a = b = 0.001$.

The Gibbs sampler can be applied after calculation of the full conditional distributions. These are all normal and Gamma distributions which can

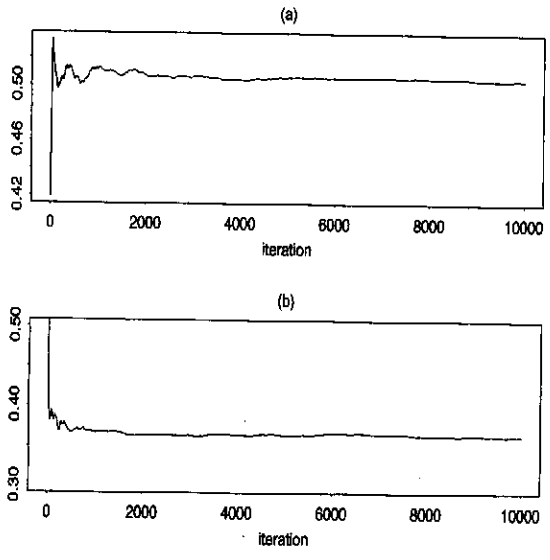

Figure 5.4 *Ergodic averages of two parameters with number of iterations of the chain. The parameters are standard deviations of random effects at: (a) individual; (b) unit level in a longitudinal study of epilepsy treatment (Gamerman, 1997).*

be easily sampled from (see Exercise 5.9). Figure 5.6 shows traces of some model parameters for two parallel chains started at different points. It seems to indicate convergence after around 1500 iterations.

These techniques must be used with caution and should always be accompanied by some theoretical reasoning. Graphical techniques may be deceptive indicating constancy that may not be so evident under a different scale. More importantly, there are many chains that exhibit every indication of convergence without actually achieving it. They are called metastable chains and are the subject of much research in probability theory (Capocaccia, Cassandro and Olivieri, 1977; Cassandro et al., 1984).

Example 5.5 *(continued) It was established that if p is close to 1 or 0, the chain will tend not to move. Therefore, the observed trajectory will have long constant stretches indicating a metastable behavior. This is known not to be an indication of convergence because $\pi(\theta_i = j) = 1/2$, $i = 1, 2$; $j = 0, 1$, and convergence will only be achieved when the chain begins to alternate values 0 and 1 with very similar frequencies.*

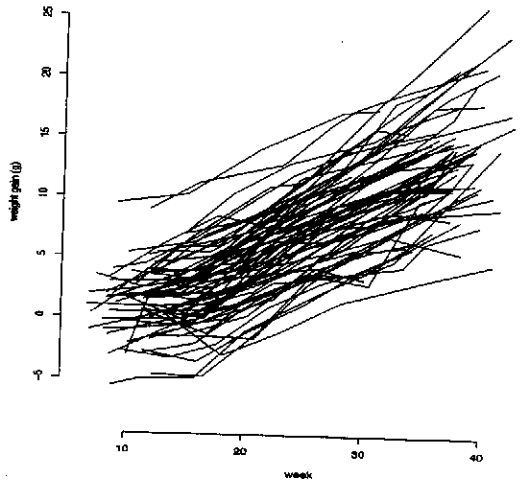

Figure 5.5 *Data on the weight gain of pregnant women (Souza, 1999).*

5.4.3 Convergence prescription

Raftery and Lewis (1992) proposed a method to establish the length of a chain required for a MCMC run. More specifically, the methodology suggests values of m, the number of burn-in iterations, k, the number of iterations to be skipped between stored chain values and n, the size of the sample values that must be stored to achieve a given Monte Carlo precision of estimates.

The setting for these choices is the estimation of u, the q quantile of a given function $\psi = t(\theta)$, i.e. $q = Pr_\pi(\psi \leq u)$. The method requires that the Monte Carlo estimate $\hat{q}$ satisfies $Pr(|\hat{q} - q| \leq r) = s$. A common choice is the tail probability with $q = 0.025$ in which case u is the lower limit of the equal tail 95% posterior credibility interval for ψ. One may require that the value of this probability be estimated in a MCMC run with error smaller than $r = 0.01$ with confidence $s = 0.99$. So, 95% posterior intervals would be given by intervals with posterior probabilities between 93% and 97% with 99% confidence. This confidence level is due to the estimation of q by MCMC and should not be confused with posterior uncertainty about ψ, governed by π.

This problem is tackled at the simpler level of a binary chain $Z^{(j)} = I(\psi^{(j)} \leq u)$ where $\psi^{(j)}$ is the value of ψ at the jth iteration of the MCMC

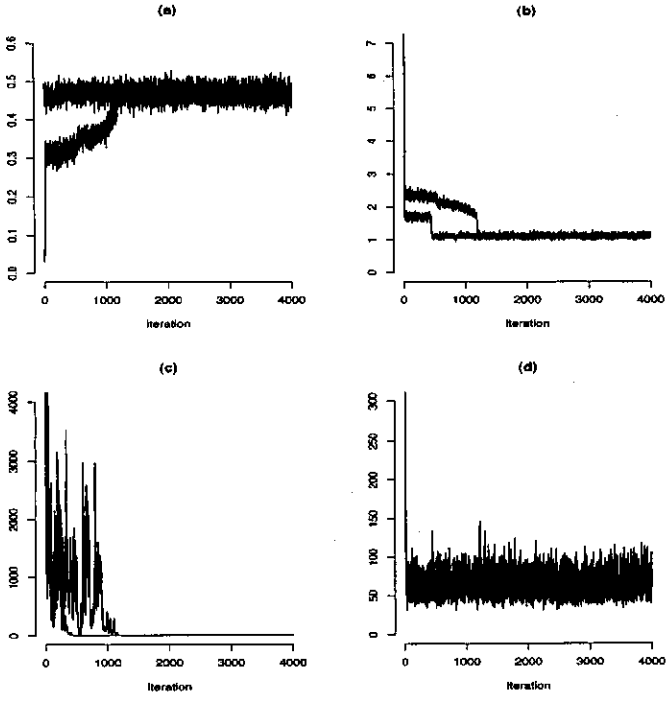

Figure 5.6 *Traces of model parameters with number of iterations of the two chains. The parameters are: (a) β, the populational growth; (b) σ, the observational standard deviation; (c) τ_α, the precision of the population of intercepts; (d) τ_β, the precision of the population of regression coefficients. First set of initial values: $\alpha_i = 0$ and $\beta_i = 0$ for $i = 1, \ldots, I$, $\alpha = \beta = 0$, $\tau = \tau_\alpha = \tau_\beta = 1.0$. Second set of initial values: $\alpha_i = -4$ and $\beta_i = 0.5$ for $i = 1, \ldots, I$, $\alpha = -4$, $\beta = 0.5$, $\tau = \tau_\alpha = \tau_\beta = 1.0$.*

for θ. $Z^{(j)}$ is derived from a Markov chain but is not a Markov chain. It is reasonable to assume however that dependencies between iterations fall off quickly with lag and new chains $Z_k^{(j)}$ may be formed by taking the values of $Z^{(j)}$ at every kth iteration. The value of k is chosen as the smallest lag to make the first order Markov chain preferable to the second order Markov chain for the chain $Z_k^{(j)}$. A test using the BIC (Schwarz, 1978) is used to choose between the two models for every value of k (Raftery and Lewis, 1992).

Once the value of k is chosen, the next step is to determine the number $m = km^*$ of iterations to be discarded. This is done by choosing m^*

such that at iteration m^*, the chain $Z_k^{(j)}$ has marginal distribution arbitrarily close to the limiting distribution implied by π. This is equivalent to requiring that $|Pr(Z_k^{(m^*)} = 1 | Z_k^{(0)} = j) - Pr_\pi(\psi \leq u)| < \epsilon$ and $|Pr(Z_k^{(m^*)} = 0 | Z_k^{(0)} = j) - Pr_\pi(\psi > u)| < \epsilon$, $j = 0, 1$. This is obtained using the results from Example 4.5 as

$$m^* = \frac{\log\left(\frac{\epsilon(\alpha+\beta)}{\max\{\alpha,\beta\}}\right)}{\log|1-\alpha-\beta|}$$

where α and β are the $(0,1)$ and $(1,0)$ elements of P_k and P_k is the transition matrix of the chain $Z_k^{(j)}$ (Exercise 5.10).

The value of $n = kn^*$ is chosen using a central limit theorem for the chain $Z_k^{(j)}$. Note that q is estimated by the ergodic average $\hat{q} = \bar{Z}_{k,n} = (1/n)\sum_{j=1}^{n} Z_k^{(m+j)}$, that has asymptotic variance given by $\tau^2 = \alpha\beta(2-\alpha-\beta)/(\alpha+\beta)^3$. The approximating distribution for $\hat{q}$ gives

$$n^* = \left(\frac{\tau z_{(1+s)/2}}{r}\right)^2$$

where z_γ is the γ quantile of the $N(0,1)$ distribution. As an assessment of the magnitudes involved, in the most favorable case of independent sampling with $k = 1$, estimation of the quantile 0.025 with largest error 0.0125 with 95% confidence level requires $n = 600$ iterations. When the error is reduced to 0.005, the number of iterations required increases to 3746.

So, one must specify the quantile of interest q, the convergence tolerance ϵ, the estimation tolerance r and confidence level s in advance. In addition, a pilot run must be observed to estimate the values of α and β. An appropriate run length will then be prescribed. This run can be used to refine the estimates of α and β, suggesting an iterative procedure.

This diagnostic obviously depends on the chosen ψ andquantile. Raftery and Lewis (1996) recommended using it for all quantities of interest with $q = 0.025$ and $q = 0.975$ as the tails are harder to estimate in general. These provide a collection of values of k, m and n and the largest of each is chosen. If l chains are to be used, then each should be run for $m + n/l$ iterations to ensure convergence. Note that the procedure does not require any information about the chain itself, just its output. Brooks and Roberts (1998) provided an example of a slow convergence chain where severe under- and over-estimation of the required length are observed. They pointed out the use of marginal indicators and estimation of α and β as the weak points of the method and suggested its use alongside other convergence diagnostics.

Example 5.6 *(continued) Two chains were run for 4000 iterations to obtain the prescribed values of the spacing k, the burn-in period m and the sample size n retained for inference. Based on the observed chain values*

Parameter	k	m	n
β	1 (3)	2(18)	3866 (15759)
σ	1 (1)	3 (2)	4112 (3946)
τ_α	1 (1)	3 (3)	4285 (4112)
τ_β	2 (1)	8 (2)	8128 (3787)

Table 5.3 *Convergence prescription summary for data on pregnant women. Values in parentheses refer to the second chain.*

these are given for the two chains according to values in Table 5.3 with $q = 0.025$, $r = 0.005$ and $s = 0.95$. They seem to indicate that the chain lengths used are generally appropriate. Figure 5.6 suggests that the burn-in periods have been mostly underestimated. Note also that different starting values may provide large variation in the prescribed values.

5.4.4 Formal convergence methods

As in the previous subsection, the methods presented here diagnose convergence based on exploration of the statistical properties of the observed chain. The methods here attempt to decide whether convergence can be safely assumed to hold rather that prescribing the run length to achieve convergence. There have been many methods presented in the literature. Most of them are covered by the review papers of Brooks and Roberts (1998) and Cowles and Carlin (1996). Only a few of the most cited and used in the literature are presented here.

Time series analysis

Consider a real function $\psi = t(\theta)$ and its trajectory $\psi^{(1)}, \psi^{(2)}, \ldots$ obtained from $\psi^{(j)} = t(\theta^{(j)})$, $j = 1, 2, \ldots$ This trajectory defines a time series and ergodic averages of this series can be evaluated. Geweke (1992) suggested the use of tests on ergodic averages to verify convergence of the chain based on the series $\psi^{(j)}$.

Assume observation of the chain for $m + n$ iterations and form averages

$$\bar{\psi}_b = \frac{1}{n_b} \sum_{j=m+1}^{m+n_b} \psi^{(j)} \quad \text{and} \quad \bar{\psi}_a = \frac{1}{n_a} \sum_{j=m+n-n_a+1}^{m+n} \psi^{(j)}$$

where $n_b + n_a < n$. If m is the length of the burn-in period, then $\bar{\psi}_a$ and $\bar{\psi}_b$ are the ergodic averages at the end and beginning of the convergence period and should behave similarly. As n gets large and the ratios n_a/n and n_b/n remain fixed then

$$z_G = \frac{\bar{\psi}_a - \bar{\psi}_b}{\sqrt{\hat{V}ar(\psi_a) + \hat{V}ar(\psi_b)}} \xrightarrow{d} N(0,1).$$

So, the standardized difference z_G between the ergodic averages at the beginning and at the end of the convergence period should not be large if convergence has been achieved. Large differences indicate lack of convergence but small differences do not imply convergence. Geweke (1992) suggested the use of values $n_b = 0.1n$ and $n_a = 0.5n$ and used spectral density estimators for the variances. This is a univariate technique but can be applied to posterior density by taking $t(\theta) = -2 \log \pi(\theta)$. As with the Raftery and Lewis (1992) diagnostic, it requires only the output from the chain and can be used with any MCMC scheme.

	All 4000 draws		*Last 2500 draws*	
Parameter	*1st chain*	*2nd chain*	*1st chain*	*2nd chain*
β	*0.241*	*-2.936*	*-0.508*	*-0.068*
σ	*3.212*	*-1.464*	*-0.552*	*-0.087*
τ_α	*-0.001*	*-0.043*	*-0.063*	*-0.031*
τ_β	*0.069*	*0.145*	*0.047*	*0.044*

Table 5.4 *Geweke diagnostic z_G summary for data on pregnant women.*

Example 5.6 *(continued) The values of z_G for both chains based on all 4000 iterations and on the last 2500 iterations are given in Table 5.4. Apart from σ (both chains) and β (second chain), all the other parameters exhibit convergence. When the first 1500 iterations are discarded, all parameters appear to have converged. Further insight into these values is provided in Figure 5.7, displaying the values of z_G for each of the four parameters after removal of a given number of iterations. The figure seems to indicate convergence after around 1200 iterations, confirming results from Figure 5.6.*

Further exploration of the time series structure of the chain to study convergence of the chain has been the subject of research of a number of authors in the area of Operational Research (Heidelberger and Welch, 1983; Schruben, Singh and Tierney, 1983; and references therein). These techniques investigate similarities between the observed series and their expected behavior under stationarity.

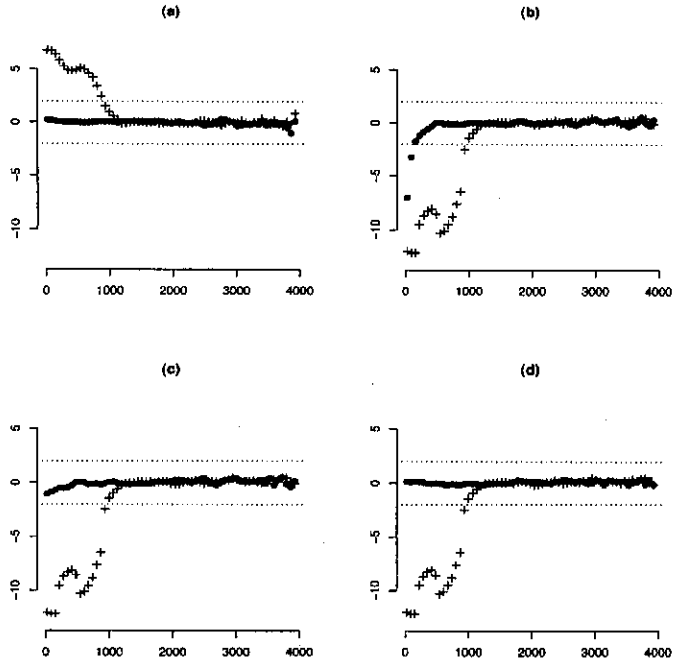

Figure 5.7 *Geweke convergence diagnostic z_G for the two chains (1st chain: •, 2nd chain: +) and the four chosen parameters for the data on pregnant women plotted against the number of discarded iterations: (a) β; (b) σ; (c) τ_α; (d) τ_β.*

Multiple chains

Another simple method to check convergence is to use parallel chains started at different points. This technique explores the same ideas used in iterative optimization to avoid convergence to local maxima. Use of multiple chains would then prevent chains getting trapped in regions around local modes. Also, slow convergence may give rise to metastable behavior of the chains and this can be easily detected through parallel chains. Examples of this behavior are provided by Gelman and Rubin (1992b) and Gelman (1996). After convergence, all chains will have the same quantitative and qualitative behavior.

Gelman and Rubin (1992a) elaborated on the idea that the chain trajectories should be the same after convergence using analysis of variance techniques. The overall idea is to test whether dispersion within chains is

larger than dispersion between chains. This is equivalent to the histogram of all chains being similar to all the histograms of individual chains.

The procedure starts by initializing the chains at points that are overdispersed with respect to the posterior distribution. The number of chains does not need to be too large to avoid computational waste and is typically given by single digit numbers. For components of θ restricted to an interval, two chains initialized close to the limits of the interval is an adequate choice. For continuous components, a search for the mode(s) and respective curvature(s) (Section 3.2) can be set and initial states of the chains drawn from (mixtures of) Student's t distribution(s) with moment(s) matching mode(s) and curvature(s). When there is indication of posterior multimodality, it is advisable to start at least one chain from each mode. Gelman (1996) pointed out that it is easy to adapt programs for calculation of Gibbs samplers to maximization of the posterior. All that is required is the substitution of random moves by deterministic moves in the direction of higher posterior density.

Considering m parallel chains and a real function $\psi = t(\theta)$, there are m trajectories $\{\psi_i^{(1)}, \psi_i^{(2)}, \ldots, \psi_i^{(n)}\}$, $i = 1, \ldots, m$, for ψ. The variances between chains B and within chains W are given by

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\psi}_i - \bar{\psi})^2 \text{ and } W = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(\psi_i^{(j)} - \bar{\psi}_i)^2$$

where $\bar{\psi}_i$ is the average of observations of chain i, $i = 1, \ldots, m$, and $\bar{\psi}$ is the average of these averages. Under convergence, all these mn values are drawn from the posterior and σ_ψ^2, the variance of ψ, can be consistently estimated by W, B and the weighted average $\hat{\sigma}_\psi^2 = (1 - 1/n)W + (1/n)B$.

If, however, the chains have not yet converged, then initial values will still be influencing the trajectories. Due to their overdispersion, they will force $\hat{\sigma}_\psi^2$ to overestimate σ_ψ^2 until stationarity is reached. On the other hand, before convergence, W will tend to underestimate σ_ψ^2 because each chain will not have adequately traversed the complete state space. Following this reasoning, an indicator of convergence can be formed by the estimator of potential scale reduction given by $\hat{R} = \sqrt{\hat{\sigma}_\psi^2/W}$, that is always larger than 1. As $n \to \infty$, both estimators converge to σ_ψ^2 by the ergodic theorem and $\hat{R} \to 1$. Convergence can be evaluated by the proximity of $\hat{R}$ to 1. Gelman (1996) suggested accepting convergence when the value of $\hat{R}$ is below 1.2. The original estimator proposed by Gelman and Rubin (1992a) is far more elaborate and its derivation is left as an exercise. It seems that the elaboration brings unnecessary complication as there are no formal tests applied to the statistic $\hat{R}$.

Example 5.6 *(continued) The values of $\hat{R}$ were evaluated for the four pa-*

rameters previously chosen based on the two chains with 4000 iterations. For the scale parameters, a logarithmic transformation was used to improve the normality pattern of the posterior sample. They all lie below 1.05 providing further indication of convergence. At this stage, one can safely assume that after 2000 iterations all draws arise from the posterior distribution.

The potential scale reduction should be evaluated for all quantities of interest to provide reasonable information about convergence of the chain. Note that this is a univariate technique but, again, can be applied to the complete posterior density by taking $t(\theta) = -2 \log \pi(\theta)$.

A problem of this method is the dependence on normal theory present in the choice of initial states of the chains and formulation of variance estimators. Alternatively, non-parametric estimators of variance can be used. Also, reparametrizations may be applied to components expected to have non-normal behavior but this increases the complexity of the verification. Another problem is the inefficiency associated with multiple chains (Section 5.3.1) which should lead to very parsimonious choices of the number of chains.

Methods based on conditional distributions

Assume that θ can be divided in two blocks θ_1 and θ_2. Then, $\pi(\theta) = \pi(\theta_1|\theta_2)\pi(\theta_2) = \pi(\theta_2|\theta_1)\pi(\theta_1)$, for all θ. In the applications where Gibbs sampling can be used, full conditionals are easy to obtain but the marginal distributions are not. However, they can be estimated by (5.7) so let $\hat{\pi}(\theta_i)$ be the estimate of $\pi(\theta_i)$, $i = 1, 2$.

Zellner and Min (1995) proposed two criteria for verification of convergence of the Gibbs sampler. The difference criterion is based on the statistic

$$\hat{\eta} = \pi(\theta_1|\theta_2)\hat{\pi}(\theta_2) - \pi(\theta_2|\theta_1)\hat{\pi}(\theta_1).$$

If the chain has converged, then $\hat{\eta}$ will be close to $\eta = 0$ for all θ. The ratio criterium is based on the statistics

$$\hat{\xi}_1 = \frac{\pi(\theta_2|\theta_1)\hat{\pi}(\theta_1)}{\pi(\theta_2^*|\theta_1^*)\hat{\pi}(\theta_1^*)} \quad \text{and} \quad \hat{\xi}_2 = \frac{\pi(\theta_1|\theta_2)\hat{\pi}(\theta_2)}{\pi(\theta_1^*|\theta_2^*)\hat{\pi}(\theta_2^*)}$$

where $\theta^* = (\theta_1^*, \theta_2^*)'$ is another value from the state space. Both $\hat{\xi}_1$ and $\hat{\xi}_2$ are estimates of $\xi = \pi(\theta)/\pi(\theta^*)$. If the chain has converged, then $\hat{\xi}_1$ and $\hat{\xi}_2$ will be close. In addition, if they are close to ξ then the chain has converged to the correct equilibrium distribution. Zellner and Min (1995) formalized their approach by assuming a normal sampling distribution for the estimates based on (3.8). They proceeded with a Bayesian analysis by assuming a vague prior distribution for the estimand, evaluating the criteria at a sample of θ values, constructing credibility intervals and testing the hypotheses of interest.

Ritter and Tanner (1992) also proposed to assess convergence of the chain by looking at ratio statistics such as $\hat{\xi}_1$ and $\hat{\xi}_2$. They suggested evaluating the ratios at the chain values $\theta^{(n)}$ and plotting the histograms of the ratios. As $n \to \infty$, these histograms should become closer to a degenerate distribution at the value of 1. See also Gelfand (1992) for further discussion of analysis of histograms and Roberts (1992) for expressions of moments of ratio statistics for reversible Gibbs samplers. Again, metastable behavior may be a problem and use of multiple chains should remedy the situation. Another problem with these methods is the need for the expression of the full conditionals which restrict their application to Gibbs samplers. It is also not clear how to split the parameter into two blocks. When the parameter dimension d is large, there are too many ways of splitting them for it to be feasible to perform convergence checks in all of them.

Other methods

There are many other methods of convergence diagnostics proposed in the literature (see Mengersen, Robert and Cuihenneuc-Jouyaux (1999) for a review). Liu, Liu and Rubin (1992) proposed a method based on control variables, Garren and Smith (1993) estimated the rate of convergence of chains formed by indicator variables and Johnson (1996) used coupled chains with overdispersed starting points.

Some of the advantages and disadvantages of the approaches have been discussed. The main points to consider are ease of implementation, applicability to MCMC schemes, interpretability, dependence on chain structure and availability of software (Section 5.6). As previously mentioned, none of the schemes can guarantee convergence. So it is advisable that as many as possible are used in any given problem.

5.5 Applications

Applications of Gibbs sampling have been restricted so far to simple models or separate derivations of full conditional distributions. This section will provide a more complete treatment for a few special models. Inference via Gibbs sampling will be detailed for hierarchical models (Section 2.4), dynamic models (Section 2.5) and spatial models (Section 2.6).

5.5.1 Hierarchical models

Consider initially the 2-stage normal hierarchical model described at the beginning of Section 2.4 with

$$\begin{aligned} y|\beta_1, \phi &\sim N(X_1\beta_1, \phi^{-1}I_n) \\ \beta_1|\beta_2, C &\sim N(X_2\beta_2, C^{-1}) \\ \beta_2 &\sim N(b, B) \end{aligned}$$

$$\phi \sim G\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right) \quad \text{independent of} \quad C \sim W\left(\frac{n_W}{2}, \frac{n_W S_W}{2}\right)$$

where n_0, n_W and S_0 are positive constants, b is an r-dimensional vector of constants and B and S_W are $r \times r$ and $d \times d$ positive definite matrices of constants. The parameters of the model are the d- and r-dimensional vectors β_1 and β_2 respectively, the scalar ϕ and the dispersion matrix C. Typically $r \leq d$ although this is not mathematically necessary.

The model includes as special cases the one-way classification model (Example 2.8), the random effects model (Example 5.4) and the exchangeable regression model

$$\begin{aligned} y_i|\beta_i, \phi &\sim N(X_i\beta_i, \phi^{-1}I_{n_i}), \quad i = 1, \ldots, m \\ \beta_i|\beta_2, C &\sim N(\beta_2, C^{-1}), \quad i = 1, \ldots, m \\ \beta_2 &\sim N(b, B) \\ \phi \sim G\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right) &\quad \text{independent of} \quad C \sim W\left(\frac{n_W}{2}, \frac{n_W S_W}{2}\right) \end{aligned}$$

where $y_i = (y_{i1}, \ldots, y_{in_i})'$, $i = 1, \ldots, m$. The analysis for this model using Gibbs sampling is described and illustrated in Gelfand et al. (1990).

Other versions of this model are possible having, for example, an unknown observational dispersion matrix or the dispersion matrix C premultiplied by ϕ. Also, the prior distributions can be changed to other nonconjugate forms or more stages can be included.

The full conditional distributions for the blocks β_1, β_2 and ϕ were obtained in Section 2.4 as:

1. $\beta_1|\beta_2, \phi, C \sim N(b_\phi, B_\phi)$;
2. $\beta_2|\beta_1, \phi, C \sim N(\mu^*, C_2^*)$;
3. $\phi|\beta_1, \beta_2, C \sim G(n_1/2, n_1 S_1/2)$;

where the expressions of b_ϕ, B_ϕ, b^*, B^*, n_1 and S_1 were given there.

The novelty here is the assumption that C is unknown which requires the evaluation of its full conditional distribution. The density is given by

$$\begin{aligned} \pi(C|\beta_1, \beta_2, \phi) &\propto f_N(\beta_1; X_2\beta_2, C^{-1})\, f_W(C; n_W/2, n_W S_W/2) \\ &\propto |C|^{1/2} \exp\left\{-\frac{1}{2}\mathrm{tr}[(\beta_1 - X_2\beta_2)(\beta_1 - X_2\beta_2)'C]\right\} \\ &\times |C|^{(n_W - r + 1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}[n_W S_W C]\right\} \\ &\propto |C|^{\frac{n_W - r}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}[(n_W S_W + (\beta_1 - X_2\beta_2)(\beta_1 - X_2\beta_2)')C]\right\} \end{aligned}$$

which leads to

4. $C|\beta_1, \beta_2, \phi \sim W(n_W^*/2, n_W^* S_W^*/2)$ where $n_W^* = n_W + 1$ and $n_W^* S_W^* = n_W S_W + (\beta_1 - X_2\beta_2)(\beta_1 - X_2\beta_2)'$.

A complete cycle of the Gibbs sampler involves successive sampling from the distributions given in steps 1 – 4. Generation from all these distributions is described in Chapter 1. Note that the blocks were naturally determined by the structure of the model.

Example 5.6 *(continued) The model used is a special case of the above models and draws from the posterior distribution can be obtained. The application indicates convergence after 1500 iterations. Therefore, the remaining 2500 values from the two chains can be taken to form a sample of size 5000 from the posterior. The resulting histogram can be smoothed and the resulting marginal posteriors appear in Figure 5.8. They show a normal-like form for the populational growth and gamma-like form for the scale parameters with more variation in the population of the αs than for the βs.*

Another important aspect is the assumed normality of errors at all levels. This is an unnecessary restriction now and in particular thicker-tailed distributions as the Student's t may be used. If the error distribution can be written as a (discrete or scale) mixture of normals, all full conditionals can be easily sampled from. In the context of exchangeable regression models, one may replace the first level equation by

$$\beta_i|\mu, \lambda_i, C \sim N(\mu, \lambda_i^{-1}C^{-1}) \text{ and } \lambda_i \sim F_\lambda \quad , \quad i = 1, \ldots, m.$$

If F_λ is a Gamma distribution, the regression coefficients are t distributed. In this case, the full conditional distributions of $\beta = (\beta_1, \ldots, \beta_m)'$ and μ alter only by the substitutions of C by $\lambda_i C$. The full conditional distribution of C now has $n_W^* S_W^* = n_W S_W + \Sigma_i \lambda_i (\beta_i - \beta_2)(\beta_i - \beta_2)'$. If $F_\lambda = G(n_\lambda/2, n_\lambda S_\lambda/2)$, the full conditional distribution of $\lambda = (\lambda_1, \ldots, \lambda_m)'$ is given by

$$\begin{aligned} \pi_\lambda(\lambda) &\propto \prod_{i=1}^m f_N(\beta_i; \mu, \lambda_i^{-1}C^{-1}) \prod_{i=1}^m f_G(\lambda_i; n_\lambda/2, n_\lambda S_\lambda/2) \\ &\propto \prod_{i=1}^m \lambda_i^{1/2} \exp\left\{-\frac{\lambda_i}{2}(\beta_i - \beta_2)'C(\beta_i - \beta_2)\right\} \lambda_i^{n_\lambda/2} \exp\left\{-\frac{\lambda_i}{2} n_\lambda S_\lambda\right\} \\ &\propto \prod_{i=1}^m \lambda_i^{(n_\lambda+1)/2} \exp\left\{-\frac{\lambda_i}{2}[n_\lambda S_\lambda + (\beta_i - \beta_2)'C(\beta_i - \beta_2)]\right\} \end{aligned}$$

and, a posteriori, the λ_is remain conditionally independent with distributions $G\{(n_\lambda + 1)/2, [n_\lambda S_\lambda + (\beta_i - \beta_2)'C(\beta_i - \beta_2)]/2\}$, $i = 1, \ldots, m$. So, minor modifications in the already existing steps and the introduction of

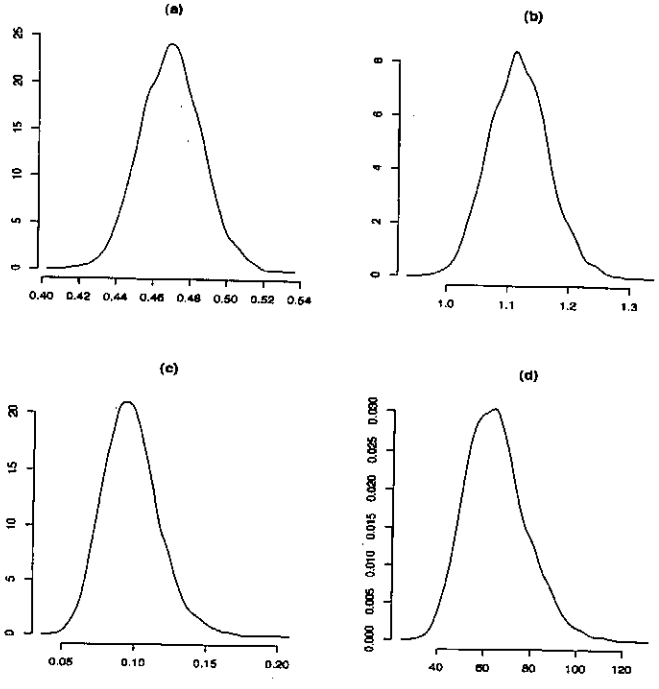

Figure 5.8 *Smoothed marginal density estimates for data on pregnant women, (a) β; (b) σ; (c) τ_α; (d) τ_β.*

an additional step with independent Gamma draws are the only changes required to robustify the model.

5.5.2 Dynamic models

The dynamic models considered here are given by (2.21) – (2.22) with constant variances of the observation and system disturbances, i.e., $\sigma_t^2 = \sigma^2$ and $W_t = W$, for all t. This restriction is aimed mainly at presentation clarity. It also provides for a more parsimonious model. The extension to the general case of unequal variances is not difficult and is left as Exercise 5.15.

As in hierarchical models, the specification of the model can be taken as a basis for blocking parameters. So, the natural choice is to form blocks $\beta_1, \ldots, \beta_n, \sigma^2$ and W.

The full conditional distributions of the β_t were obtained in Section 2.5.2

and are given by $\pi(\beta_t|\beta_{-t}, \sigma^2, W) = N(b_t, B_t)$ where

$$b_t = \begin{cases} B_t(\sigma^{-2}F_t y_t + G'_{t+1}W^{-1}\beta_{t+1} + R^{-1}a) & , t = 1 \\ B_t(\sigma^{-2}F_t y_t + G'_{t+1}W^{-1}\beta_{t+1} + W^{-1}G_t\beta_{t-1}) & , t = 2, \ldots, n-1 \\ B_t(\sigma^{-2}F_t y_t + W^{-1}G_t\beta_{t-1}) & , t = n \end{cases}$$

and

$$B_t = \begin{cases} (\sigma^{-2}F_t F'_t + G'_{t+1}W^{-1}G_{t+1} + R^{-1})^{-1} & , t = 1 \\ (\sigma^{-2}F_t F'_t + G'_{t+1}W^{-1}G_{t+1} + W^{-1})^{-1} & , t = 2, \ldots, n-1. \\ (\sigma^{-2}F_t F'_t + W^{-1})^{-1} & , t = n \end{cases}$$

Assuming now independent priors $\phi = \sigma^{-2} \sim G(n_\sigma/2, n_\sigma S_\sigma/2)$ and $\Phi = W^{-1} \sim W(n_W/2, n_W S_W/2)$, their full conditional posterior distributions were also obtained in Section 2.5.2 as $G(n^*_\sigma/2, n^*_\sigma S^*_\sigma/2)$ and $W(n^*_W/2, n^*_W S^*_W/2)$, respectively. Once again, both parameters are conditionally conjugate.

These full conditional distributions complete a cycle of the Gibbs sampler. They are all easy to sample from (see Chapter 1) and one can proceed with a sampling-based Bayesian inference. This approach was introduced by Carlin, Polson and Stoffer (1992). They also extended the analysis to models with non-normal disturbances ϵ_t and w_t and non-linear models. Non-normality was introduced through scale mixtures of normals and therefore the same methods used in Section 5.5.1 can be applied here. Non-linearity was introduced with the replacement of the linear forms $F'_t\beta_t$ and $G_t\beta_{t-1}$ by arbitrary functions $F_t(\beta_t)$ and $G_t(\beta_{t-1})$. In these cases, the authors suggested the use of rejection methods with a normal density q based on the linear parts of the model.

Unfortunately, this approach may be very inefficient. The system equation introduces prior correlation between system parameters $\beta=(\beta_1, \ldots, \beta_n)'$. This correlation is controlled by the system variance matrix W. The smaller their elements, the larger in absolute value the correlation between the β_t will be. This correlation is partially preserved in the posterior although its quantification is more complicated. In the limit, when $W = 0$, the prior and consequently the posterior correlation is one. This is a highly-dimensional version of the same phenomenon depicted in Figure 5.3. The high dimensionality of the state space brings convergence problems to Gibbs sampling.

Typically, the values of W are much smaller than the values of σ^2. In this case, posterior correlation between state parameters will be high and the chain will tend to move slowly across the state space. As a result, a large number of iterations is required both for the burn-in period and for collecting the sample from the limiting distribution. In the latter case, it is advisable to retain draws from every kth iteration for a final sample with fixed size. The high chain autocorrelation will tend to force similar values at successive iterations and appropriate coverage of the parameter space will only be achieved with a large sample or with spacing between draws.

However, if the entries of the system variance matrix W are large then the correlation between the components of β will be low and the Gibbs sampler will work well. In this case, the system parameters experience large variation and the very use of dynamic models becomes questionable; little information will be passed through model parameters. Dynamic models should be used when there is relevant passage of information through the system and in this case the correlation between model parameters will be high.

There are two alternative approaches to high correlation, both described in Section 5.3: block sampling and reparametrization. As previously discussed, it is generally preferable to sample correlated parameters in blocks when using Gibbs sampling. This is also possible for dynamic models by using Equation (2.27). It shows that the full conditional distribution of the block β is normal and can be decomposed in tractable densities that can be obtained and sampled from the updating equations. Incorporating explicitly the conditional on σ^2 and W, each term in (2.27) is given by Bayes' theorem as

$$\begin{aligned} p(\beta_t|\beta_{t+1}, \sigma^2, W, y^t) &\propto p(\beta_{t+1}|\beta_t, \sigma^2, W, y^t) p(\beta_t|\sigma^2, W, y^t) \\ &\propto f_N(\beta_{t+1}; G_t\beta_t, W) f_N(\beta_t; m_t, C_t) \end{aligned}$$

where the first term on the right hand side plays the role of the likelihood and the second term plays the role of the prior. From this perspective, the *observations* β_{t+1} form a regression model with design matrix G_t and parameters β_t whose *prior* is given by the updating equations. It becomes easy to obtain that

$$\begin{aligned} (\beta_t|\beta_{t+1}, \sigma^2, W, y^t) \sim\ & N[(G_t'W^{-1}G_t + C_t^{-1})^{-1}(G_t'W^{-1}\beta_{t+1} + C_t^{-1}m_t), \\ & (G_t'W^{-1}G_t + C_t^{-1})^{-1}] \end{aligned} \tag{5.8}$$

for $t = 1, \ldots, n-1$ using results from Section 2.5 (Exercise 2.20). So, a scheme for sampling from the full conditional of the block β is given by:

1. Sample β_n from its updated distribution (2.25) and set $t = n-1$.
2. Sample β_t from the distribution (5.8).
3. Decrease t to $t-1$ and return to step 2 until $t = 1$.

Step 1 is obtained by running the Kalman filter from $t = 1$ to $t = n$ with given values of σ^2 and W. When running the filter, the updated means m_t and variances C_t, $t = 1, \ldots, n$, are stored for use in step 2.

The above sampling scheme draws a value from the full conditional $\pi(\beta|\sigma^2, W)$. It was independently proposed by Carter and Kohn (1994) and Frühwirth-Schnatter (1994) and is widely known as the *forward filtering-backward sampling* (FFBS) algorithm, with m_t and C_t computed in a forward filtering step and $\beta_1, \ldots, \beta_n$ sampled backwards from Equation (5.8). Examples in these papers and in Shephard (1994) show that convergence

becomes orders of magnitude faster than sampling each β_t at a time. Also, the computational cost of each iteration is higher but comparable in magnitude to the cost of obtaining and sampling from the full conditionals of the β_t.

A more numerically advantageous way of obtaining a sample from β is by directly evaluating the prior full conditional of β, which turns out to be $N(A, P^{-1})$, for P block tridiagonal (see Exercise 5.17). The likelihood is now given in matrix form by $y|\beta, \sigma^2 \sim N(F\beta, \sigma^2 I_n)$, where $y = (y_1, \ldots, y_n)$ and $F = diag(F_1', \ldots, F_n')$. Combining prior and likelihood leads to the posterior $\theta|\sigma^2, W, y^n \sim N(M, Q^{-1})$ where Q is a block tridiagonal matrix (see Exercise 5.18). Great computational advantages can be obtained from the sparseness of Q. In particular, fast inversion algorithms can be used to ensure that samples from β are efficiently drawn (Rue, 2001). See Migon et al. (2005) for further details.

In many situations, it is computationally feasible to sample σ^2 and W from their joint marginal posterior $\pi(\sigma^2, W)$. In such cases, block sampling reduces to sampling the whole parameter vector jointly by noticing that $\pi(\beta, \sigma^2, W) = \pi(\beta, \sigma^2, W)\pi(\sigma^2, W)$. Thus, a sample from β, σ^2, W is obtained by: (1) sampling (σ^2, W) from $\pi(\sigma^2, W)$, and (2) conditional on sampled values of (σ^2, W), sampling β using the FFBS algorithm. Details about this sampling scheme are presented in Section 6.5.2.

The other alternative is reparametrization and was explored by Gamerman (1998). The main source of correlation between the β_t is induced by the system equation. On the other hand, the β_t can be completely determined by the values of β_1 and disturbances w_t. The system equation can be rewritten as $w_t = \beta_t - G_t\beta_{t-1}$, $t = 2, \ldots, n$. Setting $w_1 = \beta_1$ and applying recursively the system equation leads to the inverse relation

$$\beta_t = \sum_{l=1}^{t} \left(\prod_{k=1}^{t-l} G_{t-k+1} \right) w_l$$

for $t = 2, \ldots, n$ and $\beta_1 = w_1$. For the majority of dynamic models of interest, $G_t = G$, for all t. In this case, the above expression simplifies to

$$\beta_t = \sum_{l=1}^{t} G^{t-l} w_l \ .$$

The model can be written in terms of the new parameters w_t and their full conditionals can be obtained (Exercise 5.16). The disturbances are prior independent by construction but are not posterior independent. So, this scheme is not as efficient as block sampling β. Nevertheless, it removes the main source of correlation, the system equation, from the sampling scheme. This ensures fast convergence of the sampling algorithm although in this case the computational cost of each iteration is higher. This reparametriza-

tion will be explored in the next chapter in connection with non-normal dynamic models.

Reis, Salazar and Gamerman (2006) compare the performance of these sampling schemes in the context of the first order normal dynamic linear model. Section 6.5.2 gives some details about the comparisons.

5.5.3 Spatial models

The spatial models considered here are the simpler members of the GMRF and distance-based GRF families with no covariates and only the intercept θ is varying in space. Derivation for the extensions where covariates are present and their effects are space-varying are left as Exercises 5.19 and 5.20. Consider initially the GMRF model (2.33) and, as before, set $\phi = 1/\sigma^2$. The full set of parameters is $\theta_1, \ldots, \theta_d, \phi$ and $\Phi = W^{-1}$. The full conditional posterior for the θ_is and for θ were obtained in Exercise 2.23 as:

1. $\theta_i|\theta_{-i}, \sigma^2, W \sim N(m_i, C_i)$, for $i = 1, \ldots, d$;
2. $\theta|\sigma^2, W \sim N(m, C)$;

where values of m_is, C_is, m and C are given in the exercise. In practical applications the value of d is usually large, making direct sampling of θ very slow. Rue (2001) suggested numerical techniques to take advantage of the sparseness (presence of many 0s) of C^{-1} and hence improve sampling considerably. These techniques basically involve appropriate reordering of the pixels to ensure a minimal diagonal band structure to C^{-1} followed by a numerically efficient Cholesky decomposition that takes advantage of this band diagonalization.

Assuming independent Gamma priors

$$\phi \sim G\left(\frac{n_\sigma}{2}, \frac{n_\sigma S_\sigma}{2}\right) \text{ and } \Phi \sim G\left(\frac{n_W}{2}, \frac{n_W S_W}{2}\right),$$

the full conditional posterior distributions of ϕ and Φ are given by

$$\begin{aligned}
\pi(\phi|\theta, W) &\propto \prod_{i=1}^{d} p(y_i|\theta_i, \phi)\; p(\phi) \\
&\propto \prod_{i=1}^{d} f_N(y_i; \theta_i, \phi^{-1})\; f_G(\phi; n_\sigma/2, n_\sigma S_\sigma/2) \\
&\propto f_G(\phi; n^*_\sigma/2, n^*_\sigma S^*_\sigma/2) \text{ and} \\
\pi(\Phi|\theta, \phi) &\propto p(\theta_1, \ldots, \theta_d|\Phi)\; p(\Phi) \\
&\propto \Phi^{d/2} \exp\left\{-\frac{\Phi}{2}\sum_{i<j} w_{ij}(\theta_i - \theta_j)^2\right\}\; f_G\left(\Phi; \frac{n_W}{2}, \frac{n_W S_W}{2}\right) \\
&\propto f_G\left(\Phi; n^*_W/2, n^*_W S^*_W/2\right)
\end{aligned}$$

where $n^*_\sigma = n_\sigma + d$, $n^*_\sigma S^*_\sigma = n_\sigma S_\sigma + \Sigma_i(y_i - \theta_i)^2$, $n^*_W = n_W + d$ and $n^*_W S^*_W = n_W S_W + \sum_{i<j} w_{ij}(\theta_i - \theta_j)^2$. Therefore, a posteriori, $\phi \sim G(n^*_\sigma/2, n^*_\sigma S^*_\sigma/2)$ and $\Phi \sim G(n^*_W/2, n^*_W S^*_W/2)$ and all parameters of this model exhibit conditional conjugacy.

Thus, one can sample each scalar parameter individually or sample jointly all components of $\theta = (\theta_1, \ldots, \theta_d)$. It is also computationally feasible to sample all parameters (θ, ϕ, W) jointly by noting that $\pi(\theta, \phi, W) = \pi(\theta|\phi, W)\pi(\phi, W)$. Since $\pi(\theta|\phi, W)$ is available in closed form and $\pi(\theta, \phi, W)$ is known up to a proportionality constant, $\pi(\phi, W)$ can also be obtained up to a proportionality constant. This sampling strategy was also presented for dynamic models. Details of this sampling scheme are left for Section 6.5.4.

Blocking is still beneficial in the spatial context but to an extent smaller than for dynamic models. Empirical evidence from Gamerman, Moreira and Rue (2003) did not present large differences between these sampling schemes but ranked them in the increasing order of blocking. Knorr-Held and Rue (2002) also advocated the importance of blocking as much as possible in the spatial context based on their empirical evidence.

Consider now the distance-based GRF model (2.38) and, as before, set $\phi = 1/\sigma^2$. The full set of parameters is $\theta_1, \ldots, \theta_d, \Psi$, with $\Psi = (\mu, \phi, \tau^2, \lambda)$. The full conditional posterior for the model parameters are obtained as

1. $\theta|\Psi \sim N(m_\theta, C_\theta)$, with $m_\theta = C_\theta(\tau^{-2}R^{-1}_\lambda 1_d\mu + \phi y)$ and $C^{-1}_\theta = \tau^{-2}R^{-1}_\lambda + \phi I_d$ and R_λ as defined in (2.37) implied by the assumed correlation function;
2. $\mu|\theta, \phi, \tau^2, \lambda \sim N(m_\mu, C_\mu)$, with $m_\mu = C_\mu(\tau^{-2}1'_d R^{-1}_\lambda \theta + R^{-1}a)$ and $C_\mu = (\tau^{-2}1'_d R^{-1}_\lambda 1_d + R^{-1})$;
3. $\phi|\theta, \mu, \tau^2, \lambda \sim G(n^*_\sigma/2, n^*_\sigma S^*_\sigma/2)$, with $n^*_\sigma = n_\sigma + d$, $n^*_\sigma S^*_\sigma = n_\sigma S_\sigma + \Sigma_i[y_i - \theta(s_i)]^2$;
4. $\tau^2|\theta, \mu, \phi, \lambda \sim IG(n^*_\tau/2, n^*_\tau S^*_\tau/2)$, with $n^*_\tau = n_\tau + d$, $n^*_\tau S^*_\tau = n_\tau S_\tau + Q_\lambda$ with $Q_\lambda = (\theta - 1_d\mu)'R^{-1}_\lambda(\theta - 1_d\mu)$;
5. $\lambda|\theta, \mu, \phi, \tau^2 \sim F_\lambda$ with density

$$\pi(\lambda|\theta, \mu, \phi, \tau^2) \propto p(\lambda)\,|R_\lambda|^{-1/2} \exp\{-\frac{1}{2\tau^2}Q_\lambda\}, \tag{5.9}$$

which has no amenable form for direct sampling;

The derivation of the full conditional posterior distributions above is left as Exercise 5.20.

A major computational problem arises here when d is large. The inversion of the square matrix R_λ of order d is required at every iteration, slowing down the computations. Nevertheless, it is expected in the case of large d to have many (possibly most) entries in R_λ with negligible values due to many small correlation values. This fact can be used to design approximation

methods to band diagonalize this matrix and ensure application of fast inversion algorithms (see Rue and Tjelmeland (2004) for more details).

5.6 MCMC-based software for Bayesian modeling

One of the greatest impediments on the development of Bayesian inference was the difficulty of its implementation in practical situations. This difficulty was due to a host of possibilities for the specification of the prior distribution and to the difficulty of summarization of the resulting posterior distribution. The first source of difficulty is being eliminated by the introduction of symbolic languages that accommodate many specifications in a computational system. The second source of difficulty was greatly eliminated by the introduction of MCMC methods such as the Gibbs sampler that allow the analysis of complex models through decomposition and sampling from full conditional distributions.

Any system capable of specifying a variety of prior distributions for any given model and sampling from the resulting full conditionals would solve a great number of Bayesian problems. One such system is BUGS (Spiegelhalter et al., 1995), which stands as an acronym for Bayesian inference Using Gibbs Sampling. BUGS is a system developed at the Biostatistics Unit of the Medical Research Council, United Kingdom. Since its advent, BUGS has been expanded in several fronts with the implementation of Metropolis-Hastings steps and its Windows version, WinBUGS. WinBUGS was developed jointly with the Imperial College School of Medicine. BUGS consists of a set of functions that allows specification of models and probability distributions for all its random components (observations and parameters). Model specification is surprisingly simple given the complexity of models that it can tackle. Among those models already analyzed with WinBUGS and described in its manual (Spiegelhalter et al., 2003) are generalized linear models with random effects, regression analysis of survival data, models for spatially dependent data and non-parametric smoothing models.

For each combination of data set and model, BUGS outputs samples of model parameters at every $k \geq 1$ iterations after m iterations. The values of k and m as well as sampled parameters to be stored are chosen by the user. In addition, it provides sample-based estimates of posterior mean and credibility interval for the parameters. Both the system language as well as data input and output follow the syntax of the S-plus and R languages thus providing a useful interface for other data manipulations the user may wish to entertain. The system is freely available at the BUGS Project official web site www.mrc-bsu.cam.ac.uk/bugs/.

The system recognizes conditional conjugacy and uses it to sample efficiently. Failing that, it uses rejection and adaptive rejection methods (Section 1.5) or the Metropolis-Hastings algorithm (Chapter 6). The latest release of WinBUGS at the time of writing, version 1.4, already allows for

specification of multivariate normal and Wishart distributions, block sampling and graphical model specification. WinBUGS can be run in batch mode by using, for example, the R2WinBUGS package (Sturtz, Ligges and Gelman, 2005). R2WinBUGS is one of the R contributed packages and can be found at cran.r-project.org/src/contrib/Descriptions/R2WinBUGS.html.

Example 5.7 *The data in Figure 5.9 describe the evolution in the height of teeth of children through time and was introduced by Elston and Grizzle (1962). In addition to the dynamic component of the trajectories through time, there is also the hierarchical component of similarities between trajectories. Gamerman and Smith (1996) proposed a mixture model that incorporates both aspects above while still preserving the individuality of the series. Their model was given by*

$$\begin{aligned} y_{ti} &\sim N(\theta_{ti}, \sigma^2) \\ \theta_{ti} &\sim (1-p)N(\theta_{t-1,i} + \lambda_{ti}, W_1) + pN(\mu_t, V_1) \\ \lambda_{ti} &\sim (1-p)N(\lambda_{t-1,i}, W_2) + pN(\gamma_t, V_2) \\ \mu_t &\sim N(\mu_{t-1} + \gamma_t, W_1) \\ \gamma_t &\sim N(\gamma_{t-1}, W_2) \end{aligned}$$

and was completed with independent vague priors μ_1 and $\gamma_1 \sim N(0, 10^3)$, σ^2, V_1, V_2, W_1 and $W_2 \sim IG(0.1, 0.1)$.

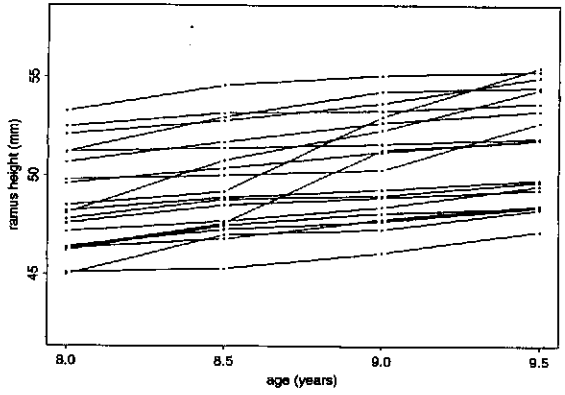

Figure 5.9 *Data on the height of the ramus tooth in boys. Each trajectory represents the evolution of the height for one boy.*

The analysis was performed using BUGS. *Appendix 5.A below illustrates how the model is described in* BUGS. *It is written as in* S *(or* R*) language with the added bonus of probabilistic attributions. Note that discrete mixture*

of normals is not a common distribution and is not directly available. It is reproduced in BUGS *using the same device of indicator variables described in Section 3.2. For this data set, despite the dynamic nature of the model, a single long chain was used. The very short time length of the series reduces the problems caused by the correlation of successive state parameters. After discarding the first 10000 iterations the next 1000 iterations constituted the sample used for inference. Part of the results is exhibited in Figure 5.10 where the mixture character of the model is evidenced.*

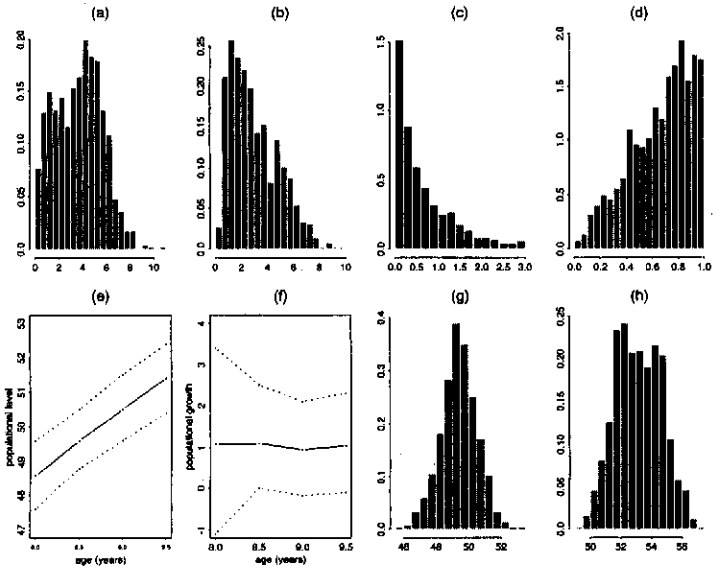

Figure 5.10 *Summary of estimation - posterior histograms of: (a) σ; (b) V_1; (c) W_2; (d) p; (g) $\theta_{2,12}$; (h) $\theta_{4,18}$. Point estimates (full line) and 90% credibility limits (broken lines) for populational parameters: (e) μ_t; (f) γ_t. Note the difference between the histograms of individual levels in agreement ($\theta_{2,12}$) and in disagreement ($\theta_{4,18}$) with the populational levels.*

Example 5.8 *Nobre, Schmidt and Lopes (2005) proposed a Poisson time series version of the spatial model (2.33) when investigating the effects of rainfall (x) on malaria incidence (y) in Pará, one of Brazil's largest states, for a number of years. They used a Poisson-normal model where malaria counts for any two counties (areal data) are conditionally independent and Poisson distributed. For counties $i = 1, \ldots, d$ and years $t = 1, \ldots, n$,*

$$\begin{aligned} y_{it} &\sim Poi(e^{\theta_{it}}) \\ \theta_{it} &= \alpha_t + \beta_t x_{it} + b_{it} \end{aligned}$$

with $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \tau_\alpha^2)$, $\beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \tau_\beta^2)$, $\alpha_0 \sim N(0.001, 1000)$ and $\beta_0 \sim N(0, \tau_\beta^2)$. Four model structures were considered for spatio-temporal variation of parameters b_{it}:

$$\begin{aligned} M_1: &\quad b_{it} \sim N(0, \sigma_t^2) &&\text{and}\quad \sigma_t^2 \sim IG(1,1)\ ; \\ M_2: &\quad b_{it} \sim CAR(w, \sigma_t^2) &&\text{and}\quad \sigma_t^2 \sim IG(1,1)\ ; \\ M_3: &\quad b_{it} \sim N(0, \sigma_t^2) &&\text{and}\quad \sigma_t^2 \sim LN(\log(\sigma_{t-1}^2), \tau_\sigma^2)\ ; \\ M_4: &\quad b_{it} \sim CAR(w, \sigma_t^2) &&\text{and}\quad \sigma_t^2 \sim LN(\log(\sigma_{t-1}^2), \tau_\sigma^2), \end{aligned}$$

where w denotes the weights given by neighbor indicators and $LN(\mu, \sigma^2)$ is the lognormal distribution such that $X \sim LN(\mu, \sigma^2)$ if and only if $log(X) \sim N(\mu, \sigma^2)$. For models M_3 and M_4, $\log(\sigma_1^2) \sim N(0, \tau_\sigma^2)$. The prior distributions for the hyperparameters were $\tau_\alpha^{-2} \sim G(1,1)$, $\tau_\beta^{-2} \sim G(1,1)$ and $\tau_\sigma^2 \sim G(1,1)$.

The analysis was performed using BUGS. *Appendix 5.B below illustrates how the model is described in* BUGS. *Figure 5.11 shows the posterior median and associated 95% credibility interval for the rainfall coefficient β_t estimated under each of the four models. It is noticeable that the effect of rainfall:*
(a) is significant for models M_1 and M_3 and not significant for models M_2 and M_4;
(b) is negative for all models;
(c) drops quite abruptly for models M_1 and M_3 around December of 1997, probably due to the amount of rainfall predicted in October of 1997.

Convergence diagnotics of a BUGS output can be performed through CODA (Best, Cowles and Vines, 1995). The latest version of CODA is 0.5-3 and is freely available from `www-fis.iarc.fr/coda/`. A similar software is BOA. Its latest version is 1.1.5 and is also freely available from `www.public-health.uiowa.edu/boa/`. CODA stands for Convergence Diagnostics and Output Analysis and BOA stands for Bayesian Output Analysis. CODA and BOA are systems that may but do not need to be used in conjunction with WinBUGS and have versions available for use in R/S-PLUS. CODA is being maintained and distributed by the same research group responsible for BUGS. Both CODA and BOA contain many summarizing statistics and the convergence diagnostics of Gelman and Rubin (1992a), Geweke (1992) (also available in BUGS), Raftery and Lewis (1992) and Heidelberger and Welch (1983). All simulations of Example 5.6 were made with BUGS and the convergence diagnostics of Table 5.3 were made with CODA.

Additional software for Bayesian analysis is an increasingly large list. For example, BayesX is a software for Bayesian semiparametric regression and can be freely downloaded from `www.stat.uni-muenchen.de/~bayesx`. As of now, R contains more than a dozen modules for Bayesian inference via MCMC to a broad range of statistical models.

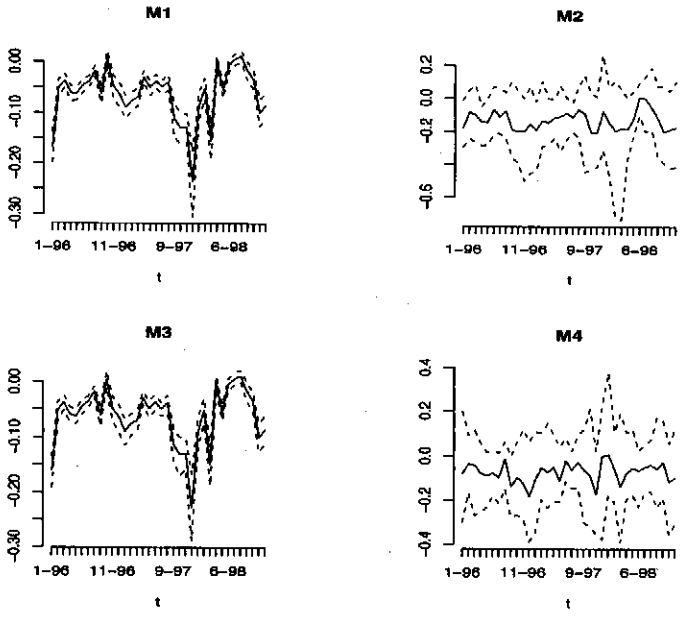

Figure 5.11 *Posterior median (solid line) and 95% credibility interval (dashed line) for the rainfall coefficient β_t under models: M_1, M_2, M_3 and M_4.*

Appendix 5.A: BUGS code for Example 5.7

```
model mixture;
const
  m = 20, # number of series
  n = 4;  # number of time periods
var
   y[n,m],theta[n,m],lambda[n,m],k,k1,mi1[n,m,2],mi2[n,m,2],
   vi1[n,2],vi2[n,2],mu[n],gamma[n],eta[n],p,sigma,tau1,
   tau2,w1,w2,w0;
data in "ramus.dat";
inits in "ramus.in";
{
  # Prior specification for hyperparameters
  sigma ~ dgamma(0.1,0.1);
  tau1  ~ dgamma(0.1,0.1);
  tau2  ~ dgamma(0.1,0.1);
  w1    ~ dgamma(0.1,0.1);
```



```
  w2    ~ dgamma(0.1,0.1);
  p     ~ dbeta(1,1);
  # Prior specification for 2nd. level parameters
  w0       <- 0.001 ;
  mu[1]    ~ dnorm(0,w0);
  gamma[1] ~ dnorm(0,w0);
  # Prior specification for 1st. level parameters
  k        ~ dbern(p);
  k1       <- 1 + k;
  vi1[1,1] <- w0;
  vi1[1,2] <- tau1;
  vi2[1,1] <- w0;
  vi2[1,2] <- tau2;
  for(i in 1:m){
    mi1[1,i,1] <- 0;
    mi1[1,i,2] <- mu[1];
    mi2[1,i,1] <- 0 ;
    mi2[1,i,2] <- gamma[1];
    theta[1,i] ~ dnorm(mi1[1,i,k1],vi1[1,k1]);
    lambda[1,i] ~ dnorm(mi2[1,i,k1],vi2[1,k1]);
    # Observation equation
    y[1,i] ~ dnorm(theta[1,i],sigma)
  }
  # Model specification for t=2,...,n
  # Evolution for 2nd. level parameters
  for(t in 2:n){
    gamma[t] ~ dnorm(gamma[t-1],w2);
    eta[t]  <- mu[t-1]+gamma[t];
    mu[t]    ~ dnorm(eta[t], w1);
    # Evolution for 1st. level parameters
    vi1[t,1] <- w1; vi1[t,2] <- tau1;
    vi2[t,1] <- w2; vi2[t,2] <- tau2;
    for(i in 1:m){
      mi2[t,i,1] <- lambda[t-1,i];
      mi2[t,i,2] <- gamma[t];
      lambda[t,i] ~ dnorm(mi2[t,i,k1],vi2[t,k1]);
      mi1[t,i,1] <- beta[t-1,i]+lambda[t,i];
      mi1[t,i,2] <- mu[t];
      theta[t,i]   ~ dnorm(mi1[t,i,k1],vi1[t,k1]);
      # Observation equation
      y[t,i]  ~ dnorm(theta[t,i],sigma)
    }
  }
}
```

Appendix 5.B BUGS code for Example 5.8

```
# BUGS works with precisions as opposed to variances.
# The prefix "i" in isigma2, itau2.alpha, itau2.beta
# and itau2.sigma stands for inverse.
# For instance isigma2 is the inverse of sigma2.
model{
  for (t in 1:n){
    b[1:d,t] ~ car.normal(adj[],weights[],num[],isigma2[t])
    for (i in 1:n){
      log(mu[i,t]) <- alpha[t]+beta[t]*chuva[i,t]+b[i,t]
      Y[i,t]       ~ dpois(mu[i,t])
    }
  }
  isigma2[1]     <- 1/exp(logsigma2[1])
  alpha[1]       ~ dnorm(alpha0,itau2.alpha)
  beta[1]        ~ dnorm(0,itau2.beta)
  logsigma2[1]   ~ dnorm(0,itau2.sigma)
  for (t in 2:n){
    isigma2[t]    <- 1/exp(logsigma2[t])
    alpha[t]      ~ dnorm(alpha[t-1],itau2.alpha)
    beta[t]       ~ dnorm(beta[t-1],itau2.beta)
    logsigma2[t]  ~ dnorm(logsigma2[t-1],itau2.sigma)
  }
  alpha0       ~ dflat()
  itau2.beta   ~ dgamma(1,1)
  itau2.alpha  ~ dgamma(1,1)
  itau2.sigma  ~ dgamma(1,1)
}
```

5.7 Exercises

5.1 *Consider Example 5.1. Prove Equation (5.3).*

5.2 *Consider the Gibbs sampler with transition kernel (5.4).*

(a) Show that the chain with a complete scan over all components is not reversible.

(b) Under what conditions is the chain with transitions formed by individual changes on a single component reversible?

(c) Show that the chain that takes each iteration to consist of the complete scan through the components followed by another scan through the components in reversed order is reversible (Besag, 1986).

5.3 *(Casella and George, 1992) Let x and y be random quantities with conditional densities $f(x|y) = ye^{-yx}$, $x > 0$ and $f(y|x) = xe^{-xy}$, $y > 0$.*

Show that the only possible solution for $f(x)$ is $f(x) = 1/x$, which is not a proper density, and that Gibbs sampling cannot be applied in this case.

5.4 *Consider Example 5.1 but assume now that the regions A and B have non-null intersection for only one of the axes, for example, the θ_1 axis. Discuss whether conditions for convergence to the joint distribution are satisfied.*

5.5 *(MacEachern and Berliner, 1994; O'Hagan and Forster 2004) Consider the estimation of a real function $\psi = t(\theta)$ from a stream of n successive values from a chain. Form k sub-samples of size $m = n/k$ by skipping every k iterations and assume that k is large enough to ensure approximate independence between sample values. Denote by $\hat{\psi}_1, \ldots, \hat{\psi}_k$ the averages of the sub-samples and $\hat{\psi} = (1/k)\sum_{j=1}^{k} \hat{\psi}_j$ the average over the complete sample.*

(a) Show that $Var(\hat{\psi}_j) = Var(\psi)/m$, for all j.

(b) Use the Cauchy-Schwarz (or correlation) inequality $Cov(x, y) \leq \sqrt{Var(x)Var(y)}$ to show that $Var(\hat{\psi}) \leq Var(\hat{\psi}_j)$.

5.6 *Consider the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where $e_i \sim N(0, \sigma^2)$ are independent, $i = 1, \ldots, n$, with non-informative marginal $p(\beta) \propto k$.*

(a) Show that if α has components $\alpha_1 = \beta_1 + \beta_2 \bar{x}$ and $\alpha_2 = \beta_2$ then α_1 and α_2 are conditionally independent a posteriori given σ^2.

(b) Show that using α is equivalent to centering the covariate and using the model $y_i = \alpha_1 + \alpha_2(x_i - \bar{x}) + e_i$.

(c) Generalize the result to multiple linear regression.

5.7 *Show that for the random effects model of Example 5.4, posterior correlations are given by*

$$Cor(\mu, \alpha_i) = -\left(1 + \frac{\sigma^2/n_i}{\tau^2/m}\right)^{-1/2} \quad \text{and} \quad Cor(\alpha_i, \alpha_j) = \left(1 + \frac{\sigma^2/n_i}{\tau^2/m}\right)^{-1}$$

in the original parametrization,

$$Cor(\mu, \beta_i) = -\left(1 + \frac{m\tau^2}{\sigma^2/n_i}\right)^{-1/2} \quad \text{and} \quad Cor(\beta_i, \beta_j) = \left(1 + \frac{m\tau^2}{\sigma^2/n_i}\right)^{-1}$$

in the centered parametrization and

$$Cor(\nu, \xi_i) = 0 \quad \text{and} \quad Cor(\xi_i, \xi_j) = -\frac{1}{m}$$

in the parametrization of Vines, Gilks and Wild (1996).

5.8 *(O'Hagan and Forster, 2004) Consider the situation of Example 5.5 where $\theta = (\theta_1, \theta_2)$ is bivariate and $\pi(\theta)$ is given by the table of probabilities below*

θ_1	θ_2	
	0	*1*
0	$p/2$	$(1-p)/2$
1	$(1-p)/2$	$p/2$

(a) Obtain that $\pi(\theta_i) = bern(1/2)$, $i = 1, 2$, *and that the posterior correlation between* θ_1 *and* θ_2 *is* $\rho = 2p - 1$.

(b) Show that $Pr(\theta_1^{(j)} = 1|\theta_1^{(j-1)} = 1) = Pr(\theta_1^{(j)} = 0|\theta_1^{(j-1)} = 0) = p^2 + (1-p)^2$ *and, consequently,* $Pr(\theta_1^{(j)} = 1|\theta_1^{(j-1)} = 0) = Pr(\theta_1^{(j)} = 0|\theta_1^{(j-1)} = 1) = 2p(1-p)$.

(c) Show that if $p_j = Pr(\theta_1^{(j)} = 1)$, $j = 0, 1, \ldots$, *then* $p_j = \rho^2 p_{j-1} + b$ *where* $b = 2p(1-p)$ *and derive that* $p_j = \rho^{2(j+1)} p_0 + b(1 - \rho^{2(j+1)})/(1-\rho^2)$.

(d) Show that the transition matrix formed by the marginal chain $(\theta_1^{(j)})_{j \geq 0}$ *has eigenvalues 1 and* ρ.

(e) Show that in the limit, $p_j \to b/(1-\rho^2) = 1/2$.

(f) Plot $p_j \times j$ *for* $p = 0.999$ *for a given value of* p_0.

5.9 *Consider Example 5.6. Show that the full conditional distributions of* $\alpha, \beta, \tau_\alpha, \tau_\beta, \sigma^{-2}, \alpha_i, \beta_i$ *and* (α_i, β_i) *are:*

$$\begin{aligned}
\alpha &\sim N\left(\frac{\tau_\alpha \sum_{i=1}^I \alpha_i}{\tau_\alpha I + P_\alpha}\ ,\ \frac{1}{\tau_\alpha I + P_\alpha}\right),\\
\beta &\sim N\left(\frac{\tau_\beta \sum_{i=1}^I \beta_i}{\tau_\beta I + P_\beta}\ ,\ \frac{1}{\tau_\beta I + P_\beta}\right),\\
\tau_\alpha &\sim G\left(a + \frac{I}{2}\ ,\ b + \frac{\sum_{i=1}^I (\alpha_i - \alpha)^2}{2}\right),\\
\tau_\beta &\sim G\left(a + \frac{I}{2}\ ,\ b + \frac{\sum_{i=1}^I (\beta_i - \beta)^2}{2}\right),\\
\sigma^{-2} &\sim G\left(a + \frac{n}{2}\ ,\ b + \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \alpha_i - \beta_i t_{ij})^2}{2}\right),\\
\alpha_i &\sim N\left(\frac{\sigma^{-2} \sum_{j=1}^{n_i} (y_{ij} - \beta_i t_{ij}) + \alpha\tau_\alpha}{n_i\sigma^{-2} + \tau_\alpha}\ ,\ \frac{1}{n_i\sigma^{-2} + \tau_\alpha}\right),\\
\beta_i &\sim N\left(\frac{\sigma^{-2} \sum_{j=1}^{n_i} t_{ij}(y_{ij} - \alpha_i) + \beta\tau_\beta}{\sigma^{-2} \sum_{j=1}^{n_i} t_{ij}^2 + \tau_\beta}\ ,\ \frac{1}{\sigma^{-2} \sum_{j=1}^{n_i} t_{ij}^2 + \tau_\beta}\right) \text{ and}
\end{aligned}$$

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left[\Sigma_i(A^{-1}\theta + \sigma^{-2} x_i' y_i)\ ,\ \Sigma_i\right]$$

where $\theta = (\alpha, \beta)'$, $y_i = (y_{i1}, \ldots, y_{in_i})'$, $t_i = (t_{i1}, \ldots, t_{in_i})'$, $x_i = (1_{n_i}, t_i)$, $A^{-1} = diag(\tau_\alpha, \tau_\beta)$ *and* $\Sigma_i = (A^{-1} + \sigma^{-2} x_i' x_i)^{-1}$.

5.10 *Consider the chain* $Z_k^{(j)}$ *described in Section 5.4 and its transition matrix*

$$P_k = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

(a) Show that the equilibrium distribution is $(\pi_0, \pi_1) = (\beta, \alpha)/(\alpha+\beta)$ *and that*

$$P_k^l = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{\lambda^l}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}$$

where $\lambda = 1 - \alpha - \beta$.

(b) Show that if $Pr(Z_k^{(m)} = i|Z_k^{(0)} = j)$ *is required to be within* ϵ *from its limiting value* π_i, *i.e.,* $|Pr(Z_k^{(m)} = i|Z_k^{(0)} = j) - \pi_i| \leq \epsilon$, $i, j = 0, 1$, *then*

$$m \geq m^* = \frac{log\left[\frac{(\alpha+\beta)\epsilon}{max(\alpha,\beta)}\right]}{log\,|\lambda|}$$

and therefore $m = m^* k$ *should be taken as the burn-in period.*

5.11 *Still in the conditions of Section 5.4.2, show*

(a) using (4.11) that for large n

$$\bar{Z}_{k,n} \dot{\sim} N\left[q, \frac{1}{n}\frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}\right].$$

(b) that the smallest n *satisfying* $Pr(|\bar{Z}_{k,n} - q| < r) = s$ *is*

$$n^* = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}\left\{\frac{z_{(1+s)/2}}{r}\right\}^2$$

where z_γ *is the* γ *quantile of the* $N(0,1)$ *distribution.*

(c) that n^* *is minimized in the case of independent sampling in which case* $1 - \alpha = \beta = q$ *and is given by*

$$n^* = \frac{q(1-q)z_{(1+s)/2}^2}{r^2}.$$

5.12 *Consider* m *parallel chains and a real function* $\psi = t(\theta)$. *There are* m *trajectories* $\{\psi_i^{(1)}, \psi_i^{(2)}, \ldots, \psi_i^{(n)}\}$, $i = 1, \ldots, m$, *for* ψ. *The variances between chains* B *and within chains* W *are given by*

$$B = \frac{n}{m-1}\sum_{i=1}^m (\bar{\psi}_i - \bar{\psi})^2 \quad \text{and} \quad W = \frac{1}{m(n-1)}\sum_{i=1}^m \sum_{j=1}^n (\psi_i^{(j)} - \bar{\psi}_i)^2$$

where $\bar{\psi}_i$ is the average of observations of chain i, $i = 1, \ldots, m$, and $\bar{\psi}$ is the average of these averages.

(a) Use the ergodic theorem to show that W, B and $\hat{\sigma}^2_\phi = (1 - 1/n)W + (1/n)B \to \sigma^2_\psi$, the variance of ψ, when $n \to \infty$.

(b) (Gelman and Rubin, 1992a) Let $V = \sigma^2_\psi + Var(\bar{\psi})$ and $\hat{V} = \hat{\sigma}^2_\psi + B/mn$ be an estimator of V. Show that $E(\hat{V}) = V$ and if the distribution of $\hat{V}/V$ is approximated by a χ^2_ν then ν is estimated by the method of moments as

$$\hat{\nu} = 2\frac{\hat{V}^2}{\hat{Var}(\hat{V})} .$$

Note: the potential scale reduction estimator originally proposed was given by $\hat{R} = \sqrt{(\hat{V}/W)(\hat{\nu}/(\hat{\nu} - 2))}$.

5.13 *(Zellner and Min, 1995) Derive the central limit theorem for the Rao-Blackwellized estimator of marginal densities given by (5.7).*

(a) Derive confidence intervals and a test for convergence based on the result and using the difference criterium statistic $\hat{\eta}$ evaluated at m different states $\theta_1, \ldots, \theta_m$.

(b) Assuming a vague prior for η, obtain its posterior distribution and construct credibility intervals for η.

(c) Repeat items (a) and (b) to derive tests and confidence and credibility intervals for convergence and for correct convergence based on the ratio criterium statistics.

5.14 *Specify a version of the 3-stage hierarchical model and obtain the full conditional distributions required for implementation of the Gibbs sampler.*

5.15 *Obtain the full conditional distributions required for implementation of the Gibbs sampler for the dynamic model with observational and system variances having independent prior distributions $IG(n_{\sigma,t}/2, n_{\sigma,t}S_{\sigma,t}/2)$ and $IW(n_{W,t}/2, n_{W,t}S_{W,t}/2)$ respectively, $t = 1, \ldots, n$. Compare and discuss in this context the relative efficiency of the methods based on separate sampling of the β_t, block sampling of β and sampling via the reparametrizations w_t.*

5.16 *Show that*

$$\beta_t = \sum_{l=1}^{t} \left(\prod_{k=1}^{t-l} G_{t-k+1} \right) w_l$$

for $t = 2, \ldots, n$ and $\beta_1 = w_1$ and, if $G_t = G$, for all t, the above expression simplifies to

$$\beta_t = \sum_{l=1}^{t} G^{t-l} w_l .$$

Show also that the full conditional distribution of w_t is $N(b_t, B_t)$ where

$$b_t = B_t \sum_{l=t}^{n} \sigma^{-2} H_{tl}(y_l - k_{tl}) \quad \text{and} \quad B_t^{-1} = W^{-1} + \sum_{l=t}^{n} \sigma^{-2} H_{tl} H'_{tl}$$

for $t = 2, \ldots, n$ and

$$b_1 = B_1 \left[R^{-1} a + \sum_{l=1}^{n} \sigma^{-2} H_{1l}(y_l - k_{1l}) \right] \quad \text{and} \quad B_1^{-1} = R^{-1} + \sum_{l=1}^{n} \sigma^{-2} H_{1l} H'_{1l}$$

where $H'_{tl} = F'_t G^{l-t}$ and $k_{tl} = F'_t \sum_{i=1, i \neq t}^{j} G^{l-i} w_i$, $l \geq t$, $t = 1, \ldots, n$.

5.17 *Based on the dynamic models developed on Section 5.5.2, show that $(\beta|\sigma^2, W) \sim N(A, P^{-1})$ with $A' = \left(I, G'_2, G'_3 G'_2, \cdots, \prod_{t=2}^{n} G'_{n-t+2}\right) a$ and symmetric, block tridiagonal precision matrix P, with main diagonal blocks $P_{11} = R^{-1} + G'_2 W^{-1} G_2$, $P_{nn} = W^{-1}$, $P_{tt} = W^{-1} + G'_{t+1} W^{-1} G_{t+1}$, for $t = 2, \ldots, n-1$, and secondary diagonal blocks given by $P'_{t,t+1} = P_{t+1,t} = W^{-1} G_t$, for $t = 1, \ldots, n-1$, where $\beta_1 \sim N(a, R)$.*

5.18 *Show that combination of the likelihood $y|\beta, \sigma^2 \sim N(F\beta, \sigma^2 I_n)$, where $y = (y_1, \ldots, y_n)$ and $F = diag(F'_1, \ldots, F'_n)$, with the prior from Exercise 5.17 leads to the full conditional posterior $\beta|\sigma^2, W, y^n \sim N(M, Q^{-1})$ where $M = Q^{-1}(\sigma^{-2} Fy + PA)$ and $Q = P + \sigma^{-2} F'F$.*

5.19 *Derive the expressions for the full conditional posterior distributions of parameters of model (2.33) with the presence of a regression term $X\beta$. Repeat the derivations considering now that the regression coefficients β are space-varying.*

5.20 *Derive the expressions for the full conditional posterior distributions of parameters of model (2.38). Repeat the derivations for the model with the presence of a regression term $X\beta$. Repeat the derivations considering now that the regression coefficients β are space-varying according to independent GRF for each covariate effect.*