Chapter 57

MARKOV CHAIN MONTE CARLO METHODS: COMPUTATION AND INFERENCE

SIDDHARTHA CHIB[*]

John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Dr.,
St. Louis, MO 63130, USA

Contents

[*] email: chib@olin.wustl.edu

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

Abstract

This chapter reviews the recent developments in Markov chain Monte Carlo simulation methods. These methods, which are concerned with the simulation of high dimensional probability distributions, have gained enormous prominence and revolutionized Bayesian statistics. The chapter provides background on the relevant Markov chain theory and provides detailed information on the theory and practice of Markov chain sampling based on the Metropolis–Hastings and Gibbs sampling algorithms. Convergence diagnostics and strategies for implementation are also discussed. A number of examples drawn from Bayesian statistics are used to illustrate the ideas. The chapter also covers in detail the application of MCMC methods to the problems of prediction and model choice.

Keywords

JEL classification: C1, C4

1. Introduction

This chapter is concerned with the theory and practice of Markov chain Monte Carlo (MCMC) simulation methods. These methods which deal with the simulation of high dimensional probability distributions, have over the last decade gained enormous prominence, sparked intense research interest, and energized Bayesian statistics [Tanner and Wong (1987), Casella and George (1992), Gelfand and Smith (1990, 1992), Smith and Roberts (1993), Tierney (1994), Chib and Greenberg (1995a, 1996), Besag, Green, Higdon and Mengersen (1995), Albert and Chib (1996), Tanner (1996), Gilks, Richardson and Spiegelhalter (1996), Carlin and Louis (2000), Geweke (1997), Gammerman (1997), Brooks (1998), Robert and Casella (1999)]. The idea behind these methods is simple and extremely general. In order to sample a given probability distribution that is referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution is the target distribution. Depending on the specifics of the problem, the Markov chain can be constructed by the Metropolis–Hastings algorithm, the Gibbs sampling method, a special case of the Metropolis method, or hybrid mixtures of these two algorithms. Once the Markov chain has been constructed, a sample of (correlated) draws from the target distribution can be obtained by simulating the Markov chain a large number of times and recording its values. In many situations, Markov chain Monte Carlo simulation provides the only practical way of obtaining samples from high dimensional probability distributions.

Markov chain sampling methods originated with the work of Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) who proposed an algorithm to simulate a high dimensional discrete distribution. This algorithm found wide application in statistical physics but was mostly unknown to statisticians until the paper of Hastings (1970). Hastings generalized the Metropolis algorithm and applied it to the simulation of discrete and continuous probability distributions such as the normal and Poisson. Outside of statistical physics, Markov chain methods first found applications in spatial statistics and image analysis [Besag (1974)]. The more recent interest in MCMC methods can be traced to the papers of Geman and Geman (1984), who developed an algorithm that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987), who proposed a MCMC scheme involving "data augmentation" to sample posterior distributions in missing data problems, and Gelfand and Smith (1990), where the value of the Gibbs sampler was demonstrated for general Bayesian inference with continuous parameter spaces.

In Bayesian applications, the target distribution is typically the posterior distribution of the parameters, given the data. If $\mathcal{M}$ denotes a particular model, $p(\boldsymbol{\psi}|\mathcal{M})$, $\boldsymbol{\psi} \in \Re^d$, the prior density of the parameters in that model and $f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M})$ the assumed sampling density (likelihood function) for a vector of observations $\boldsymbol{y}$, then the posterior density is given by

$$\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}) \propto p(\boldsymbol{\psi}|\mathcal{M}) f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M}), \tag{1}$$

where the normalizing constant of the density, called the marginal likelihood,

$$m(\boldsymbol{y}|\mathcal{M}) = \int_{\Re^d} p(\boldsymbol{\psi}|\mathcal{M}) f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi},$$

is almost never known in analytic form. As may be expected, an important goal of the Bayesian analysis is to summarize the posterior density. Particular summaries, such as the posterior mean and posterior covariance matrix, are especially important as are interval estimates (called credible intervals) with specified posterior probabilities. The calculation of these quantities reduces to the evaluation of the following integral

$$\int_{\Re^d} h(\boldsymbol{\psi})\,\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi},$$

under various choices of the function h. For example, to get the posterior mean, one lets $h(\boldsymbol{\psi}) = \boldsymbol{\psi}$ and for the second moment matrix one lets $h(\boldsymbol{\psi}) = \boldsymbol{\psi}\boldsymbol{\psi}'$, from which the posterior covariance matrix and posterior standard deviations may be computed.

In the pre MCMC era, posterior summaries were usually obtained either by analytic approximations, such as the method of Laplace for integrals [Tierney and Kadane (1986)], or by the method of importance sampling [Kloek and van Dijk (1978), Geweke (1989)]. Although both techniques continue to have uses (for example, the former in theoretical, asymptotic calculations), neither method is sufficiently flexible to be used routinely for the kinds of high-dimensional problems that arise in practice. A shift in thinking was made possible by the advent of MCMC methods. Instead of focusing on the question of moment calculation directly one may consider the more general question of drawing sample variates from the distribution whose summaries are sought. For example, to summarize the posterior density $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})$ one can produce a simulated sample $\{\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}\}$ from this posterior density, and from this simulated sample, the posterior expectation of $h(\boldsymbol{\psi})$ can be estimated by the average

$$M^{-1} \sum_{j=1}^{M} h(\boldsymbol{\psi}^{(j)}). \tag{2}$$

Under independent sampling from the posterior, which is rarely feasible, this calculation would be justified by classical laws of large numbers. In the context of MCMC sampling the draws are correlated but, nonetheless, a suitable law of large numbers for Markov chains that is presented below can be used establish the fact that

$$M^{-1} \sum_{j=1}^{M} h(\boldsymbol{\psi}^{(j)}) \to \int_{\Re^d} h(\boldsymbol{\psi})\,\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi}, \quad M \to \infty.$$

It is important to bear in mind that the convergence specified here is in terms of the simulation sample size M and not in terms of the data sample size n which is fixed.

This means that one can achieve any desired precision by taking M to be as large as required, subject to the constraint on computing time.

The Monte Carlo approach to inference also provides elegant solutions to the Bayesian problems of prediction and model choice. For the latter, algorithms are available that proceed to sample over both model space and parameter space, such as in the methods of Carlin and Chib (1995) and Green (1995), or those that directly compute the evidential quantities that are required for Bayesian model comparisons, namely marginal likelihoods and their ratios, Bayes factors [Jeffreys (1961)]; these approaches are developed by Gelfand and Dey (1994), Chib (1995), Verdinelli and Wasserman (1995), Meng and Wong (1996), DiCiccio, Kass, Raftery and Wasserman (1997), Chib and Jeliazkov (2001), amongst others. Discussion of these techniques is provided in detail below.

1.1. Organization

The rest of the chapter is organized as follows. Section 2 provides a brief review of three classical sampling methods that are discussed or used in the sequel. Section 3 summarizes the relevant Markov chain theory that justifies simulation by MCMC methods. In particular, we provide the conditions under which discrete-time and continuous state space Markov chains satisfy a law of large numbers and a central limit theorem. The Metropolis–Hastings algorithm is discussed in Section 4 followed by the Gibbs sampling algorithm in Section 5. Methods for diagnosing convergence are considered in Section 6 and strategies for improving the mixing of the Markov chains in Section 7. In Section 8 we discuss how MCMC methods can be applied to simulate the posterior distributions that arise in various canonical statistical models. Bayesian prediction and model choice problems are presented in Sections 9 and 10, respectively, and the MCMC-based EM algorithm is considered in Section 11. Section 12 concludes with brief comments about new and emerging directions in MCMC methods.

2. Classical sampling methods

We now briefly review three sampling methods, that we refer to as classical methods, that deliver independent and identically distributed draws from the target density. Authoritative surveys of these and other such methods are provided by Devroye (1985), Ripley (1987) and Gentle (1998). Although these methods are technically outside the scope of this chapter, the separation is somewhat artificial because, in practice, all MCMC methods in one way or another make some use of classical simulation methods. The ones we have chosen to discuss here are those that are mentioned or used explicitly in the sequel.

2.1. Inverse transform method

This method is particularly useful in the context of discrete distribution functions and is based on taking the inverse transform of the cumulative distribution function (hence its name). Suppose we want to generate the value of a discrete random variable with mass function

$$\Pr(\psi = \psi_j) = p_j,\ j = 1, 2, \ldots, \sum_j p_j = 1,$$

and cumulative mass function

$$\Pr(\psi \leqslant \psi_j) \equiv F(\psi_j) = p_1 + p_2 + \cdots + p_j.$$

The function F is a right-continuous stair function that has jumps at the point ψ_j equal to p_j and is constant otherwise. It is not difficult to see that its inverse takes the form

$$F^{-1}(u) = \psi_j \quad \text{if} \quad p_1 + \cdots + p_{j-1} \leqslant u \leqslant p_1 + \cdots + p_j. \tag{3}$$

A random variate from this distribution is obtained by generating U uniform on $(0, 1)$ and computing $F^{-1}(U)$ where F^{-1} is the inverse function in Equation (3). This method samples ψ_j with probability p_j because

$$\begin{aligned}\Pr(F^{-1}(U) = \psi_j) &= \Pr(p_1 + \cdots + p_{j-1} \leqslant U \leqslant p_1 + \cdots + p_j)\\ &= p_j.\end{aligned}$$

An equivalent version is available for continuous random variables. An important application is to the sampling of a truncated normal distribution. Suppose, for example, that

$$\psi \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2),$$

a univariate truncated normal distribution truncated to the interval (a, b), with distribution function

$$F(t) = \begin{cases} 0 & \text{if } \psi < a \\ \frac{1}{p_2 - p_1}\left(\Phi(\frac{t-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right) & \text{if } a < \psi < b \\ 1 & \text{if } b < \psi \end{cases}, \tag{4}$$

where

$$p_1 = \Phi\left(\frac{a-\mu}{\sigma}\right); \quad p_2 = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

To generate a sample variate from this distribution one must solve the equation $F(t) = U$, where U is uniform on $(0, 1)$. Algebra yields

$$t = \mu + \sigma\Phi^{-1}\left(p_1 + U(p_2 - p_1)\right). \tag{5}$$

Although the inverse distribution method is useful it is rather difficult to apply in the setting of multi-dimensional distributions.

2.2. Accept–reject algorithm

The accept–reject method is the basis for many of the well known univariate random number generators that are provided in software programs. This method is characterized by a source density $h(\boldsymbol{\psi})$ which is used to supply candidate values and a constant c, that is determined by analysis, such that for all $\boldsymbol{\psi}$

$$\pi(\boldsymbol{\psi}) \leqslant ch(\boldsymbol{\psi}).$$

Note that the accept–reject method does not require knowledge of the normalizing constant of π because that constant can be absorbed in c. Then, in the accept–reject method, one draws a variate from h, accepting it with probability $\pi(\boldsymbol{\psi})/\{ch(\boldsymbol{\psi})\}$. If the particular proposal is rejected, a new one is drawn and the process continued until one is accepted. The accepted draws constitute an independent and identically distributed (i.i.d.) sample from π.

In algorithmic form, the accept–reject method can be described as follows.

Algorithm 1: Accept–reject

```
(1) Repeat for $j = 1, 2, \ldots, M$.
    (a) Generate
            $\boldsymbol{\psi}' \sim h(\boldsymbol{\psi});\quad U \sim \mathrm{Unif}(0, 1)$.
    (b) Let $\boldsymbol{\psi}^{(j)} = \boldsymbol{\psi}'$ if
            $U \leqslant \frac{\pi(\boldsymbol{\psi}')}{ch(\boldsymbol{\psi}')}$,
        otherwise go to step 1(a).
(2) Return the values $\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\}$.
```

The idea behind this algorithm may be explained quite simply using Figure 1. Imagine drawing random bivariate points in the region bounded above by the function $ch(\boldsymbol{\psi})$ and below by the x-axis. A point in this region may be drawn by first drawing $\boldsymbol{\psi}'$ from $h(\boldsymbol{\psi})$, which fixes the x-coordinate of the point, and then drawing the y-coordinate of the point as $Uch(\boldsymbol{\psi}')$. Now, if $Uch(\boldsymbol{\psi}') \leqslant \pi(\boldsymbol{\psi}')$, the point lies below π and is accepted; but the latter is simply the acceptance condition of the AR method, which completes the justification.

Below we shall discuss a Markov chain Monte Carlo version of the accept–reject method that can be used when the condition $\pi(\boldsymbol{\psi}) \leqslant ch(\boldsymbol{\psi})$ does not hold for all values of $\boldsymbol{\psi}$.

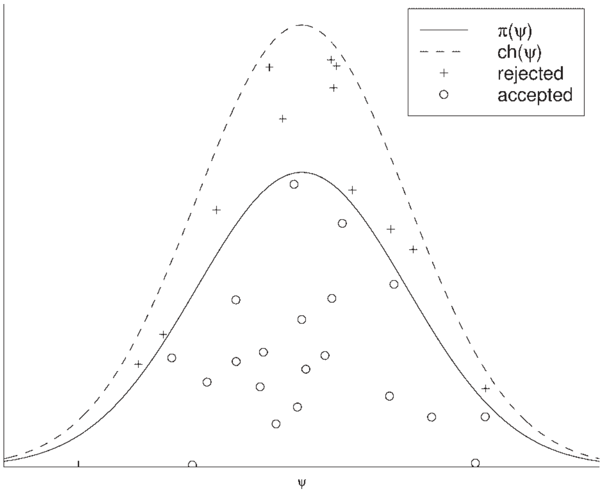

Fig. 1. Graphical illustration of the accept–reject method.

2.3. Method of composition

This method is based on the observation that if the joint density $\pi(\psi_1, \psi_2)$ is expressed as

$$\pi(\psi_1, \psi_2) = \pi(\psi_1)\,\pi(\psi_2|\psi_1),$$

and each density on the right hand side is easily sampled, then a draw from the joint distribution may be obtained by
(1) drawing $\psi_1^{(j)}$ from $\pi(\psi_1)$ and then
(2) drawing $\psi_2^{(j)}$ from $\pi(\psi_2|\psi_1^{(j)})$.

Because $(\psi_1^{(j)}, \psi_2^{(j)})$ is a draw from the joint distribution it follows that the second component of the simulated vector is a draw from the marginal distribution of ψ_2:

$$\psi_2^{(j)} \sim \pi(\psi_2) = \int \pi(\psi_2|\psi_1)\,\pi(\psi_1)\,\mathrm{d}\psi_1.$$

Thus, to obtain a draw $\psi_2^{(j)}$ from $\pi(\psi_2)$, it is sufficient to produce a sample from the joint distribution and retain the second component. This method is quite important and arises frequently in the setting of MCMC methods.

3. Markov chains

Markov chain Monte Carlo is a method to sample a given multivariate distribution π^* by constructing a suitable Markov chain with the property that its limiting, invariant distribution, is the target distribution π^*. In most problems of interest, the distribution π^* is absolutely continuous and, as a result, the theory of MCMC methods is based on that of Markov chains on continuous state spaces outlined, for example, in Nummelin (1984) and Meyn and Tweedie (1993). Tierney (1994) is the fundamental reference for drawing the connections between this elaborate Markov chain theory and MCMC methods. Basically, the goal of the analysis is to specify conditions under which the constructed Markov chain converges to the invariant distribution, and conditions under which sample path averages based on the output of the Markov chain satisfy a law of large numbers and a central limit theorem.

3.1. *Definitions and results*

A Markov chain is a collection of random variables (or vectors) $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_i : i \in T\}$ where $T = \{0, 1, 2, \ldots\}$. The evolution of the Markov chain on a space $\Omega \subseteq \Re^p$ is governed by the *transition kernel*

$$\begin{aligned} P(\boldsymbol{x}, A) &\equiv \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}, \boldsymbol{\Phi}_j, j < i) \\ &= \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad A \subset \Omega, \end{aligned}$$

which embodies the Markov assumption that the distribution of each succeeding state in the sequence, given the current and the past states, depends only on the current state.

In general, in the context of Markov chain simulations, the transition kernel has both a continuous and a discrete component. For some function $p(\boldsymbol{x}, \boldsymbol{y}) : \Omega \times \Omega \to \Re^+$, the kernel can be expressed as

$$P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} + r(\boldsymbol{x})\, \delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}), \tag{6}$$

where $p(\boldsymbol{x}, \boldsymbol{x}) = 0$, $\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}) = 1$ if $\boldsymbol{x} \in \mathrm{d}\boldsymbol{y}$ and 0 otherwise, $r(\boldsymbol{x}) = 1 - \int_\Omega p(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$. This transition kernel specifies that transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ occur according to $p(\boldsymbol{x}, \boldsymbol{y})$ and transitions from $\boldsymbol{x}$ to $\boldsymbol{x}$ occur with probability $r(\boldsymbol{x})$.

The transition kernel is thus the distribution of $\boldsymbol{\Phi}_{i+1}$ given that $\boldsymbol{\Phi}_i = \boldsymbol{x}$. The nth-step-ahead transition kernel is given by

$$P^{(n)}(\boldsymbol{x}, A) = \int_\Omega P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\, P^{(n-1)}(\boldsymbol{y}, A),$$

where $P^{(1)}(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ and

$$P(\boldsymbol{x}, A) = \int_A P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}). \tag{7}$$

The objective is to elucidate the conditions under which the nth iterate of the transition kernel converges to the invariant distribution π^* as $n \to \infty$. The invariant distribution satisfies

$$\pi^*(\mathrm{d}\boldsymbol{y}) = \int_\Omega P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{8}$$

where π is the density of π^* with respect to the Lebesgue measure (thus, $\pi^*(\mathrm{d}\boldsymbol{y}) = \pi(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$). The invariance condition states that if $\boldsymbol{\Phi}_i$ is distributed according to π^*, then all subsequent elements of the chain are also distributed as π^*. It should be noted that Markov chain samplers are invariant by construction and therefore the existence of the invariant distribution does not have to be checked in any particular application of MCMC methods.

A Markov chain is said to be *reversible* if the function $p(\boldsymbol{x},\boldsymbol{y})$ in Equation (6) satisfies

$$f(\boldsymbol{x})\,p(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{y})\,p(\boldsymbol{y},\boldsymbol{x}), \tag{9}$$

for a density $f(\cdot)$. If this condition holds, it can be shown that $f(\cdot) = \pi(\cdot)$. A reversible chain has π^* as an invariant distribution [see Tierney (1994)]. To verify this we evaluate the right hand side of Equation (8):

$$\begin{aligned}\int P(\boldsymbol{x},A)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int \left\{ \int_A p(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \right\} \pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int r(\boldsymbol{x})\,\delta_{\boldsymbol{x}}(A)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A \left\{ \int p(\boldsymbol{x},\boldsymbol{y})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right\} \mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A \left\{ \int p(\boldsymbol{y},\boldsymbol{x})\,\pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} \right\} \mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A (1 - r(\boldsymbol{y}))\,\pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A \pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}.\end{aligned} \tag{10}$$

A minimal requirement to ensure that the Markov chain satisfies a law of large numbers is that of π^**-irreducibility*. This is the requirement that the chain is able to visit all sets with positive probability under π^* from any starting point in Ω. Formally, a Markov chain is said to be π^*-irreducible if for every $\boldsymbol{x} \in \Omega$,

$$\pi^*(A) > 0 \Rightarrow P(\boldsymbol{\Phi}_i \in A|\boldsymbol{\Phi}_0 = \boldsymbol{x}) > 0,$$

for some $i \geqslant 1$. If the space Ω is connected and the function $p(\boldsymbol{x},\boldsymbol{y})$ is positive and continuous, then the Markov chain with transition kernel given by Equation (7) and invariant distribution π^* is π^*-irreducible.

Another important property of a chain is *aperiodicity*, which ensures that the chain does not cycle through a finite number of sets. A Markov chain is aperiodic if there exists no partition of $\Omega = (D_0, D_1, \ldots, D_{p-1})$ for some $p \geqslant 2$ such that $P(\boldsymbol{\Phi}^i \in D_{i \bmod (p)}|\boldsymbol{\Phi}_0 \in D_0) = 1$ for all i.

These definitions allow us to state the following results [see Tierney (1994)], which form the basis for Markov chain Monte Carlo methods. The first of these results gives conditions under which a strong law of large numbers holds and the second gives conditions under which the probability density of the Mth iterate of the Markov chain converges to its unique, invariant density.

Theorem 1. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible Markov chain with transition kernel $P(\cdot,\cdot)$ and invariant distribution π^*, then π^* is the unique invariant distribution of $P(\cdot,\cdot)$ and for all π^*-integrable real-valued functions h,*

$$\frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i) \to \int h(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{as} \quad M \to \infty, \ \textit{a.s.}$$

Theorem 2. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible, aperiodic Markov chain with transition kernel $P(\cdot,\cdot)$ and invariant distribution π^*. Then for π^*-almost every $\boldsymbol{x} \in \Omega$, and all sets A*

$$\| P^M(\boldsymbol{x},A) - \pi^*(A) \| \to 0 \quad \text{as} \quad M \to \infty,$$

where $\| \cdot \|$ denotes the total variation distance.

A further strengthening of the conditions is required to obtain a central limit theorem for sample-path averages. A key requirement is that of an ergodic chain, i.e., chains that are irreducible, aperiodic and positive Harris-recurrent [for a definition of the latter, see Tierney (1994)]. In addition, one needs the notion of geometric ergodicity. An ergodic Markov chain with invariant distribution π^* is a geometrically ergodic if there exists a non-negative real-valued function (bounded in expectation under π^*) and a positive constant $r < 1$ such that

$$\| P^M(\boldsymbol{x},A) - \pi^*(A) \| \leqslant C(\boldsymbol{x})\, r^n,$$

for all $\boldsymbol{x}$ and all n and sets A. Chan and Geyer (1994) show that if the Markov chain is ergodic, has invariant distribution π^*, and is geometrically ergodic, then for all L^2 measurable functions h, taken to be scalar-valued for simplicity, and any initial distribution, the distribution of $\sqrt{M}(\hat{h}_M - \mathrm{E}h)$ converges weakly to a normal distribution with mean zero and variance $\sigma_h^2 \geqslant 0$, where

$$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i)$$

$$\mathrm{E}h = \int h(\boldsymbol{\Phi})\,\pi(\boldsymbol{\Phi})\,\mathrm{d}\boldsymbol{\Phi},$$

and

$$\sigma_h^2 = \mathrm{Var}\, h(\boldsymbol{\Phi}_0) + 2\sum_{k=1}^{\infty} \mathrm{Cov}\,\{h(\boldsymbol{\Phi}_0), h(\boldsymbol{\Phi}_k)\}\,. \tag{11}$$

3.2. Computation of numerical accuracy and inefficiency factor

Let $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_M$ denote the output from a Markov chain, possibly collected after discarding the iterates from an initial burn-in period, and suppose that, as above,

$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i)$ denotes the sample average of the scalar function h. Then, in this context, the variance of $\hat{h}_M$ based on $\{h(\boldsymbol{\Phi}_1), \ldots, h(\boldsymbol{\Phi}_M)\}$ is an estimate of σ_h^2 where the square root of the variance of $\hat{h}_M$ is referred to as the *numerical standard error*.

To describe consistent in M estimators of σ_h^2, let $Z_i = h(\boldsymbol{\Phi}_i)$ $(i \leqslant M)$. Then, due to the fact that $\{Z_i\}$ is a dependent sequence

$$\begin{aligned}\operatorname{Var}(\hat{h}_M) &= M^{-2}\sum_{j,k}\operatorname{Cov}(Z_j, Z_k)\\ &= s^2 M^{-2}\sum_{j,k=1}^{M}\rho_{|j-k|}\\ &= s^2 M^{-1}\left\{1 + 2\sum_{s=1}^{M}(1 - \frac{s}{M})\rho_s\right\},\end{aligned}$$

where s^2 is the sample variance of $\{Z_i\}$ and ρ_s is the estimated autocorrelation at lag s [see Ripley (1987, Ch. 6)]. If $\rho_s > 0$ for each s, then this variance is larger than s^2/M which is the variance under independence. Another estimate of the variance can be found by consistently estimating the spectral density f of $\{Z_i\}$ at frequency zero and using the fact that $\operatorname{Var}(\hat{h}_M) = \tau^2/M$, where $\tau^2 = 2\pi f(0)$. Finally, a traditional approach to finding the variance is by the method of "batch means." In this approach, the data $(Z_1, \ldots, Z_M)$ is divided into k batches of length m with means $B_i = m^{-1}[Z_{(i-1)m+1} + \cdots + Z_{im}]$ and the variance of $\hat{h}_M$ estimated as

$$\operatorname{Var}(\hat{h}_M) = \frac{1}{k(k-1)}\sum_{i=1}^{k}(B_i - \bar{B})^2, \tag{12}$$

where the batch size m is chosen to ensure that the first order serial correlation of the batch means is less than 0.05.

Given the numerical variance it is common to calculate the *inefficiency factor,* which is also called the *autocorrelation time,* defined as

$$\kappa_{\hat{h}} = \frac{\operatorname{Var}(\hat{h}_M)}{s^2/M}. \tag{13}$$

This quantity is interpreted as the ratio of the numerical variance of $\hat{h}_M$ to the variance of $\hat{h}_M$ based on independent draws, and its inverse is the relative numerical efficiency defined in Geweke (1992). The inefficiency factor serves to quantify the relative efficiency loss in the computation of $\hat{h}_M$ from correlated versus independent samples.

4. Metropolis–Hastings algorithm

The Metropolis–Hastings (M–H) method is a general MCMC method to produce sample variates from a given multivariate density [Tierney (1994), Chib and Greenberg

(1995a)]. It is based on a candidate generating density that is used to supply a proposal value and a probability of move that is used to determine if the proposal value should be taken as the next item of the chain. The probability of move is based on the ratio of the target density (evaluated at the proposal value in the numerator and the current value in the denominator) times the ratio of the proposal density (at the current value in the numerator and the proposal value in the denominator). Because ratios of the target density are involved, knowledge of the normalizing constant of the target density is not required. There are a number of special cases of this method, each defined either by the form of the proposal density or by the form in which the components of $\boldsymbol{\psi}$ are revised, say in one block or in several blocks. The method is extremely general and powerful, it being possible in principle to view almost any MCMC algorithm, in one way or another, as a variant of the M–H algorithm.

4.1. The algorithm

The goal is to simulate the d-dimensional distribution $\pi^*(\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \Psi \subseteq \Re^d$ that has density $\pi(\boldsymbol{\psi})$ with respect to some dominating measure. To define the algorithm, let $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ denote the *candidate generating density,* also called a proposal density, that is used to supply a candidate value $\boldsymbol{\psi}'$ given the current value $\boldsymbol{\psi}$, and let $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ denote the function

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \begin{cases} \min\left[\frac{\pi(\boldsymbol{\psi}')\,q(\boldsymbol{\psi}',\boldsymbol{\psi})}{\pi(\boldsymbol{\psi})\,q(\boldsymbol{\psi},\boldsymbol{\psi}')}, 1\right] & \text{if} \quad \pi(\boldsymbol{\psi})\,q(\boldsymbol{\psi}, \boldsymbol{\psi}') > 0; \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

Then, in the M–H algorithm, a candidate value $\boldsymbol{\psi}'$ is drawn from the proposal density and taken to be the next item of the chain with probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$. If the proposal value is rejected, then the next sampled value is taken to be the current value. In algorithmic form, the simulated values are obtained by the following recursive procedure.

Algorithm 2: Metropolis–Hastings

```
(1) Specify an initial value $\boldsymbol{\psi}^{(0)}$:
(2) Repeat for $j = 1, 2, \ldots, M$.
    (a) Propose
            $\boldsymbol{\psi}' \sim q(\boldsymbol{\psi}^{(j)})$.
    (b) Let
            $\boldsymbol{\psi}^{(j+1)} = \begin{cases} \boldsymbol{\psi}' & \text{if} \quad \text{Unif}(0,1) \leqslant \alpha(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'); \\ \boldsymbol{\psi}^{(j)} & \text{otherwise.} \end{cases}$
(3) Return the values $\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\}$.
```

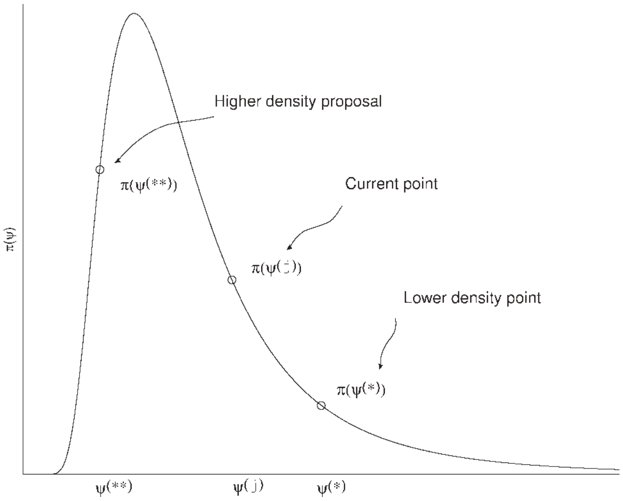

Fig. 2. Original Metropolis algorithm: higher density proposal is accepted with probabability one and the lower density proposal with probability α.

The M–H algorithm delivers variates from π under general conditions. Of course, the variates are from π only in the limit as the number of iterations becomes large but, in practice, after an initial burn-in phase consisting of (say) n_0 iterations, the chain is assumed to have converged and subsequent values are taken as approximate draws from π. Because theoretical calculation of the burn-in is not easy it is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the invariant distribution without staying at one place for many iterations. Generally, the empirical behavior of the M–H output is monitored by the autocorrelation time of each component of $\boldsymbol{\psi}$ and by the *acceptance rate,* which is the proportion of times a move is made as the sampling proceeds.

One should observe that the target density appears as a ratio in the probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and therefore the algorithm can be implemented without knowledge of the normalizing constant of $\pi(\cdot)$. Furthermore, if the candidate-generating density is symmetric, i.e., $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$, the acceptance probability only contains the ratio $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$; hence, if $\pi(\boldsymbol{\psi}') \geqslant \pi(\boldsymbol{\psi})$, the chain moves to $\boldsymbol{\psi}'$, otherwise it moves with probability given by $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$. The latter is the algorithm originally proposed by Metropolis et al. (1953). This version of the algorithm is illustrated in Figure 2.

Different proposal densities give rise to specific versions of the M–H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}' - \boldsymbol{\psi})$. The candidate $\boldsymbol{\psi}'$ is thus drawn according to the process $\boldsymbol{\psi}' = \boldsymbol{\psi} + \boldsymbol{z}$, where $\boldsymbol{z}$ follows the distribution q. Since the candidate is equal to the current value plus noise, this case is called a *random walk M–H* chain. Possible choices for q include the multivariate normal density and the multivariate-t. The random walk M–H chain is perhaps the simplest version of the M–H algorithm [and was the one used by Metropolis et al. (1953)] and quite popular in applications. One has to be careful, however, in setting the variance of z; if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. Both circumstances will tend to generate draws that are highly serially correlated. Note that when q is symmetric, the usual circumstance, $q(z) = q(-z)$ and the probability of move only contains the ratio $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$. As mentioned earlier, the same reduction occurs if $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$.

Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}')$. Tierney (1994) refers to this as an *independence M–H chain* because, in contrast to the random walk chain, the candidates are drawn independently of the current location $\boldsymbol{\psi}$. In this case, the probability of move becomes

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \min\left\{\frac{w(\boldsymbol{\psi}')}{w(\boldsymbol{\psi})}, 1\right\},$$

where $w(\boldsymbol{\psi}) = \pi(\boldsymbol{\psi})/q(\boldsymbol{\psi})$ is the ratio of the target and proposal densities. For this method to work and not get stuck in the tails of π, it is important that the proposal density have thicker tails than π. A similar requirement is placed on the importance sampling function in the method of importance sampling [Geweke (1989)]. In fact, Mengersen and Tweedie (1996) show that if $w(\boldsymbol{\psi})$ is uniformly bounded then the resulting Markov chain is ergodic.

Chib and Greenberg (1994) discuss a way of formulating proposal densities in the context of time series autoregressive-moving average models that has a bearing on the choice of proposal density for the independence M–H chain. They suggest matching the proposal density to the target at the mode by a multivariate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode. Specifically, the parameters of the proposal density are taken to be

$$\begin{aligned} &\boldsymbol{m} = \arg\max \log \pi(\boldsymbol{\psi}) \quad \text{and} \\ &\boldsymbol{V} = \tau\left\{-\frac{\partial^2 \log \pi(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right\}^{-1}_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}, \end{aligned} \tag{15}$$

where τ is a tuning parameter that is adjusted to control the acceptance rate. The proposal density is then specified as $q(\boldsymbol{\psi}') = f(\boldsymbol{\psi}'|\boldsymbol{m}, \boldsymbol{V})$, where f is some multivariate density. This may be called a *tailored M–H* chain.

Another way to generate proposal values is through a Markov chain version of the accept–reject method. In this version, due to Tierney (1994), a pseudo accept–reject step is used to generate candidates for an M–H algorithm. Suppose $c > 0$ is a known constant and $h(\boldsymbol{\psi})$ a source density. Let $C = \{\boldsymbol{\psi} : \pi(\boldsymbol{\psi}) \leqslant ch(\boldsymbol{\psi})\}$ denote the set of value for which $ch(\boldsymbol{\psi})$ dominates the target density and assume that this set has high probability under π^*. Now given $\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}$, the next value $\boldsymbol{\psi}^{(n+1)}$

is obtained as follows: First, a candidate value $\boldsymbol{\psi}'$ is obtained, *independent of the current value* $\boldsymbol{\psi}$, by applying the accept–reject algorithm with $ch(\cdot)$ as the "pseudo dominating" density. The candidates $\boldsymbol{\psi}'$ that are produced under this scheme have density $q(\boldsymbol{\psi}') \propto \min\{\pi(\boldsymbol{\psi}'), ch(\boldsymbol{\psi}')\}$. If we let $w(\boldsymbol{\psi}) = c^{-1}\pi(\boldsymbol{\psi})/h(\boldsymbol{\psi})$ then it can be shown that the M–H probability of move is given by

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \begin{cases} 1 & \text{if } \boldsymbol{\psi} \in C, \\ 1/w(\boldsymbol{\psi}) & \text{if } \boldsymbol{\psi} \notin C, \boldsymbol{\psi}' \in C, \\ \min\{w(\boldsymbol{\psi}')/w(\boldsymbol{\psi}), 1\} & \text{if } \boldsymbol{\psi} \notin C, \boldsymbol{\psi}' \notin C. \end{cases} \tag{16}$$

The choices mentioned above are not exhaustive. Other proposal densities can be generated by mixing over a set of proposal densities, using one proposal density for a certain number of iterations before switching to another.

4.2. Convergence results

In the M–H algorithm the transition kernel of the chain is given by

$$P(\boldsymbol{\psi}, \mathrm{d}\boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \mathrm{d}\boldsymbol{\psi}' + r(\boldsymbol{\psi})\, \delta_{\psi}(\mathrm{d}\boldsymbol{\psi}'), \tag{17}$$

where $\delta_{\psi}(\mathrm{d}\boldsymbol{\psi}') = 1$ if $\boldsymbol{\psi} \in \mathrm{d}\boldsymbol{\psi}'$ and 0 otherwise and

$$r(\boldsymbol{\psi}) = 1 - \int_{\Omega} q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \mathrm{d}\boldsymbol{\psi}'.$$

Thus, transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ ($\boldsymbol{\psi}' \neq \boldsymbol{\psi}$) are made according to the density

$$p(\boldsymbol{\psi}, \boldsymbol{\psi}') \equiv q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}'), \quad \boldsymbol{\psi} \neq \boldsymbol{\psi}',$$

while transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}$ occur with probability $r(\boldsymbol{\psi})$. In other words, the density function implied by this transition kernel is of mixed type,

$$K(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') + r(\boldsymbol{\psi})\, \delta_{\psi}(\boldsymbol{\psi}'), \tag{18}$$

having both a continuous and discrete component where now, with change of notation, $\delta_{\psi}(\boldsymbol{\psi}')$ is the Dirac delta function defined as $\delta_{\psi}(\boldsymbol{\psi}') = 0$ for $\boldsymbol{\psi}' \neq \boldsymbol{\psi}$ and $\int_{\Omega} \delta_{\psi}(\boldsymbol{\psi}')\, \mathrm{d}\boldsymbol{\psi}' = 1$.

Chib and Greenberg (1995a) provide a way to derive and interpret the probability of move $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$. Consider the proposal density $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$. This proposal density q is not likely to be reversible for π (if it were then we would be done and M–H sampling would not be necessary). Without loss of generality, suppose that $\pi(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\psi}') > \pi(\boldsymbol{\psi}')\, q(\boldsymbol{\psi}', \boldsymbol{\psi})$ implying that the rate of transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ exceed those in the reverse direction. To reduce the transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ one can introduce a function $0 \leqslant \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') \leqslant 1$ such that

$\pi(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \pi(\boldsymbol{\psi}')\, q(\boldsymbol{\psi}', \boldsymbol{\psi})$. Solving for $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ yields the probability of move in the M–H algorithm. This calculation reveals the important point that the function $p(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is reversible by construction, i.e., it satisfies the condition

$$q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \pi(\boldsymbol{\psi}) = q(\boldsymbol{\psi}', \boldsymbol{\psi})\, \alpha(\boldsymbol{\psi}', \boldsymbol{\psi})\, \pi(\boldsymbol{\psi}'). \tag{19}$$

It immediately follows, therefore, from the argument in Equation (10) that the M–H kernel has $\pi(\boldsymbol{\psi})$ as its invariant density.

It is not difficult to provide conditions under which the Markov chain generated by the M–H algorithm satisfies the conditions of Propositions 1–2. The conditions of Proposition 1 are satisfied by the Metropolis–Hastings chain if $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is positive for $(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and continuous and the set $\boldsymbol{\psi}$ is connected. In addition, the conditions of Proposition 2 are satisfied if q is not reversible (which is the usual situation) which leads to a chain that is aperiodic. Conditions for ergodicity, required for use of the central limit theorem, are satisfied if in addition π is bounded. Other similar conditions are provided by Robert and Casella (1999).

4.3. Example

To illustrate the M–H algorithm consider count data taken from Hand et al. (1994) on the number of seizures for 58 epilepsy patients measured first over a eight week baseline period and then over four subsequent two week intervals. At the end of the baseline, each patient is randomly assigned to either a treatment group, which is given the drug Progabide, or a control group which is given a placebo. The model for these data on the ith patient at the jth occasion is taken to be

$$\begin{aligned} y_{ij} | \mathcal{M}, \boldsymbol{\beta} &\sim \text{Poisson}(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \ln t_{ij}, \\ \boldsymbol{\beta} &\sim \mathcal{N}_4(0, 10\, \boldsymbol{I}_4), \end{aligned}$$

where x_1 is an indicator for treatment status, x_2 is an indicator of period, equal to zero for the baseline and one otherwise, $x_3 = x_1 x_2$ and t_{ij} is the offset that is equal to eight in the baseline period and two otherwise. Because the purpose of this example is illustrative, the model does not incorporate the obvious intra-cluster dependence that is likely to be present in the counts.

The target density in this case is the Bayesian posterior density

$$\pi(\boldsymbol{\beta} | \boldsymbol{y}, \mathcal{M}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{58} \prod_{j=0}^{4} \exp(-\lambda_{ij})\, \lambda_{ij}^{y_{ij}},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\pi(\boldsymbol{\beta})$ is the density of the $\mathcal{N}(0, 10\, \boldsymbol{I}_4)$ distribution. To draw sample variates on $\boldsymbol{\beta}$ from this density we apply the AR–M–H chain.

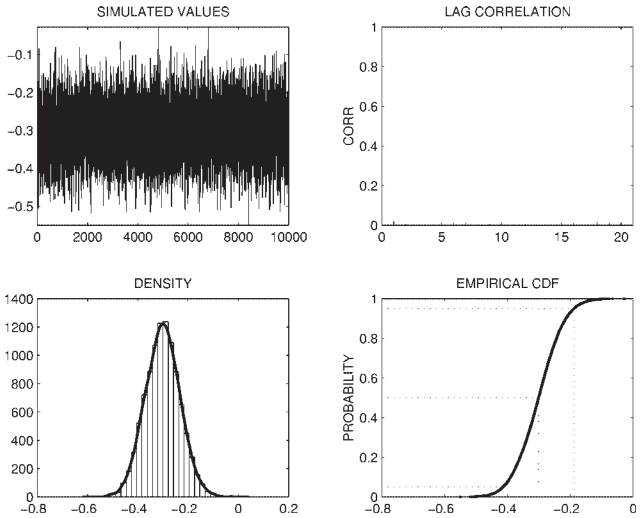

Fig. 3. Marginal posterior distribution of β_1 in Poisson count example. Top left, simulated values by iteration; top right, autocorrelation function of simulated values; bottom left, histogram and superimposed kernel density estimate of marginal density; bottom right, empirical cdf with .05 percentile, 50th percentile and 97.5th percentile marked.

Let $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{V}$ denote the maximum likelihood estimate and inverse of observed information matrix, respectively. Then, the source density $h(\boldsymbol{\beta})$ for the accept–reject method is specified as $f_T(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{V}, 15)$, a multivariate-*t* density with fifteen degrees of freedom. The constant c is set equal to 1.5 which implies that the probability of move in Equation (16) is defined in terms of the weight

$$w(\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{58} \prod_{j=0}^{4} \exp(-\lambda_{ij})\lambda_{ij}^{y_{ij}}}{1.5 f_T(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{V}, 15)}.$$

The MCMC sampler is now run for 10 000 iterations beyond a burn-in of 200 iterations. Of interest in this case is the marginal posterior of β_1 which is summarized in Figure 3.

The figure includes a time series plot of the sampled values, against iteration, and the associated autocorrelation function. These indicate that there is no sign of serial correlation in the sampled values. Although mixing of this kind is often not achieved, this example shows that it is sometimes possible to have a MCMC algorithm produce virtually i.i.d. draws from the target distribution. We also summarize the marginal posterior distribution by a histogram/kernel smoothed plot and the empirical cumulative distribution function. Because the entire distribution is concentrated on negative values it appears that the drug Progabide tends to lower the seizure counts, conditional on the specified model.

4.4. Multiple-block M–H algorithm

In applications when the dimension of $\boldsymbol{\psi}$ is quite large it is preferable to construct the Markov chain simulation by first grouping the variables $\boldsymbol{\psi}$ into p blocks $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)$, with $\boldsymbol{\psi}_k \in \Omega_k \subseteq \Re^{d_k}$, and sampling each block, conditioned on the rest, by the M–H algorithm. Hastings (1970) considers this general situation and mentions different possibilities for constructing a Markov chain on the product space $\Omega = \Omega_1 \times \cdots \times \Omega_p$.

Let $\boldsymbol{\psi}_{-k} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-1}, \boldsymbol{\psi}_{k+1}, \ldots, \boldsymbol{\psi}_p)$ denote the variables (blocks) excluding $\boldsymbol{\psi}_k$, in order to describe the multiple-block M–H algorithm. Also let $\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})$ denote the joint density of $\boldsymbol{\psi}$, regardless of where $\boldsymbol{\psi}_k$ appears in the list $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)$. Furthermore, let $\{q_k(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k}), \; k \leqslant p\}$ denote a collection of proposal densities, one for each block $\boldsymbol{\psi}_k$, where the proposal density q_k may depend on the current value of the remaining blocks and is specified along the lines mentioned in connection with the single-block M–H algorithm. Finally, define

$$\alpha_k(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k}) = \min\left\{ \frac{\pi(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_{-k})\, q_k(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k | \boldsymbol{\psi}_{-k})}{\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})\, q_k(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k})}, 1 \right\}, \tag{20}$$

as the probability of move for block $\boldsymbol{\psi}_k$ conditioned on $\boldsymbol{\psi}_{-k}$. Then, in the multiple-block M–H algorithm, one cycle of the algorithm is completed by updating each block, say sequentially in fixed order, using a M–H step with the above probability of move, given the most current value of the remaining blocks. The algorithm may be summarized as follows.

Algorithm 3: Multiple-block Metropolis–Hastings

```
(1) Specify an initial value $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\psi}_1^{(0)}, \ldots, \boldsymbol{\psi}_p^{(0)})$
(2) Repeat for $j = 1, 2, \ldots, M$
    (a) Repeat for $k = 1, 2, \ldots, p$
        (i) Propose
                $\boldsymbol{\psi}'_k \sim q(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k})$.
        (ii) Calculate
                $\alpha_k(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k}) = \min\left\{ \frac{\pi(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_{-k})\, q_k(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k^{(j)} | \boldsymbol{\psi}_{-k})}{\pi(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}_{-k})\, q_k(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k})}, 1 \right\}$.
        (iii) Set
                $\boldsymbol{\psi}_k^{(j+1)} = \begin{cases} \boldsymbol{\psi}'_k & \text{if } \operatorname{Unif}(0,1) \leqslant \alpha_k(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k}) \\ \boldsymbol{\psi}_k^{(j)} & \text{otherwise.} \end{cases}$
(3) Return the values $\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\}$.
```

Before we examine this algorithm, some features of this method should be noted. First, the version of the algorithm presented above assumes that the blocks are revised sequentially in fixed order. This is not necessary and the blocks may be updated in random order. Second, at the moment block k is updated in this algorithm, the blocks $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-1})$ have already been revised while the blocks $(\boldsymbol{\psi}_{k+1}, \ldots, \boldsymbol{\psi}_p)$ have not. Thus, at each step of the algorithm one must be sure to condition on the *most current value of the blocks in* $\boldsymbol{\psi}_{-k}$. Finally, if the proposal density q_k is determined by tailoring to $\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})$, as in Chib and Greenberg (1994), then this implies that the proposal density is not fixed but varies across iterations.

To understand the multiple-block M–H algorithm, first note that the transition kernel of the kth block, conditioned on $\boldsymbol{\psi}_{-k}$, may be expressed as

$$P_k(\boldsymbol{\psi}_k, \mathrm{d}\boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k}) = q(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k})\,\alpha(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k})\,\mathrm{d}\boldsymbol{\psi}'_k + r(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})\,\delta_{\boldsymbol{\psi}_k}(\mathrm{d}\boldsymbol{\psi}'_k), \tag{21}$$

where the notation is similar to that of Equation (17). It can be readily shown that, for a given $\boldsymbol{\psi}_{-k}$, this kernel satisfies what may be called the *local reversibility condition*

$$\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})\,q(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k})\,\alpha(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k}) = \pi(\boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k})\,q(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})\,\alpha(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}). \tag{22}$$

As a consequence, the transition kernel of the move from $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_k)$ to $\boldsymbol{\psi}' = (\boldsymbol{\psi}'_1, \boldsymbol{\psi}'_2, \ldots, \boldsymbol{\psi}'_k)$, under the assumption that the blocks are revised sequentially in fixed order, is given by the product of transition kernels

$$P(\boldsymbol{\psi}, \mathrm{d}\boldsymbol{\psi}') = \prod_{k=1}^{p} P_k(\boldsymbol{\psi}_k, \mathrm{d}\boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k}). \tag{23}$$

This transition kernel is not reversible, as can be easily checked, because under fixed sequential updating of the blocks updating in the reverse order never occurs. The multiple-block M–H algorithm, however, satisfies the weaker condition of invariance. To show this, we follow Chib and Greenberg (1995a). Consider for notational simplicity the case of two blocks, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$, where $\boldsymbol{\psi}_k\colon d_k \times 1$. Now, due to the fact that the local moves satisfy the local reversibility condition (22), the transition kernel $P_1(\boldsymbol{\psi}_1, d\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)$ has $\pi^*_{1|2}(\cdot|\boldsymbol{\psi}_2)$ as its local invariant distribution (with density $\pi_{1|2}(\cdot|\boldsymbol{\psi}_2)$), i.e.,

$$\pi^*_{1|2}(\mathrm{d}\boldsymbol{\psi}_1|\boldsymbol{\psi}_2) = \int P_1(\boldsymbol{\psi}_1, \mathrm{d}\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\,\pi_{1|2}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\,\mathrm{d}\boldsymbol{\psi}_1. \tag{24}$$

Similarly, the conditional transition kernel $P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2|\boldsymbol{\psi}_1)$ has $\pi^*_{2\,|\,1}(\cdot|\boldsymbol{\psi}_1)$ as its invariant distribution, for a given value of $\boldsymbol{\psi}_1$. Then, the kernel formed by multiplying the conditional kernels is invariant for $\pi^*(\cdot,\cdot)$:

$$\begin{aligned}
&\int\int P_1(\boldsymbol{\psi}_1, \mathrm{d}\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\, P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')\, \pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\, \mathrm{d}\boldsymbol{\psi}_1\, \mathrm{d}\boldsymbol{\psi}_2 \\
&\quad = \int P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1') \left[\int P_1(\boldsymbol{\psi}_1, \mathrm{d}\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\, \pi_{1|2}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\, \mathrm{d}\boldsymbol{\psi}_1\right] \pi_2(\boldsymbol{\psi}_2)\, \mathrm{d}\boldsymbol{\psi}_2 \\
&\quad = \int P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')\, \pi^*_{1\,|\,2}(\mathrm{d}\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\, \pi_2(\boldsymbol{\psi}_2)\, \mathrm{d}\boldsymbol{\psi}_2 \\
&\quad = \int P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1') \frac{\pi_{2\,|\,1}(\boldsymbol{\psi}_2|\boldsymbol{\psi}_1')\, \pi^*_1(\mathrm{d}\boldsymbol{\psi}_1')}{\pi_2(\boldsymbol{\psi}_2)} \pi_2(\boldsymbol{\psi}_2)\, \mathrm{d}\boldsymbol{\psi}_2 \\
&\quad = \pi^*_1(\mathrm{d}\boldsymbol{\psi}_1') \int P_2(\boldsymbol{\psi}_2, \mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')\, \pi_{2\,|\,1}(\boldsymbol{\psi}_2|\boldsymbol{\psi}_1')\, \mathrm{d}\boldsymbol{\psi}_2 \\
&\quad = \pi^*_1(\mathrm{d}\boldsymbol{\psi}_1')\, \pi^*_{2\,|\,1}(\mathrm{d}\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1') \\
&\quad = \pi^*(\mathrm{d}\boldsymbol{\psi}_1', \mathrm{d}\boldsymbol{\psi}_2'),
\end{aligned}$$

where the third line follows from Equation (24), the fourth from Bayes theorem, the sixth from assumed invariance of P_2, and the last from the law of total probability.

The implication of this "product of kernels" result is that it allows us to take draws in succession from each of the kernels, instead of having to run each to convergence for every value of the conditioning variable.

5. The Gibbs sampling algorithm

Another MCMC method, which is a special case of the multiple-block Metropolis–Hastings method, is called the Gibbs sampling method and was brought into statistical prominence by Gelfand and Smith (1990). An elementary introduction to Gibbs sampling is provided by Casella and George (1992). In this algorithm the parameters are grouped into p blocks $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)$ and each block is sampled according to the *full conditional distribution* of block $\boldsymbol{\psi}_k$, defined as the conditional distribution under π of $\boldsymbol{\psi}_k$ given all the other blocks $\boldsymbol{\psi}_{-k}$ and denoted as $\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})$. In parallel with the multiple-block M–H algorithm, the most current value of the remaining blocks is used in deriving the full conditional distribution of each block. Derivation of the full conditional distributions is usually quite simple since, by Bayes theorem, $\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}) \propto \pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})$, the joint distribution of all the blocks. In addition, the powerful device of data augmentation, due to Tanner and Wong (1987), in which latent or auxiliary variables are artificially introduced into the sampling, is often used to simplify the derivation and sampling of the full conditional distributions.

5.1. The algorithm

To define the Gibbs sampling algorithm, let the set of full conditional distributions be

$$\{\pi(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_p); \pi(\boldsymbol{\psi}_2|\boldsymbol{\psi}_1, \boldsymbol{\psi}_3, \ldots, \boldsymbol{\psi}_p); \ldots, \pi(\boldsymbol{\psi}_p|\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{d-1})\}.$$

Now one cycle of the Gibbs sampling algorithm is completed by simulating $\{\boldsymbol{\psi}_k\}_{k=1}^p$ from these distributions, recursively updating the conditioning variables as one moves through each distribution. When $d = 2$ one obtains the two block Gibbs sampler that is featured in the work of Tanner and Wong (1987). The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm 4: Gibbs sampling

```
(1) Specify an initial value $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\psi}_1^{(0)}, \ldots, \boldsymbol{\psi}_p^{(0)})$
(2) Repeat for $j = 1, 2, \ldots, M$
      Generate $\boldsymbol{\psi}_1^{(j+1)}$ from $\pi(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2^{(j)}, \boldsymbol{\psi}_3^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)})$.
      Generate $\boldsymbol{\psi}_2^{(j+1)}$ from $\pi(\boldsymbol{\psi}_2|\boldsymbol{\psi}_1^{(j+1)}, \boldsymbol{\psi}_3^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)})$.
      ⋮
      Generate $\boldsymbol{\psi}_p^{(j+1)}$ from $\pi(\boldsymbol{\psi}_p|\boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{p-1}^{(j+1)})$.
(3) Return the values $\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\}$.
```

Thus, the transition of $\boldsymbol{\psi}_k$ from $\boldsymbol{\psi}_k^{(j)}$ to $\boldsymbol{\psi}_k^{(j+1)}$ is effected by taking a draw from the conditional distribution

$$\pi\left(\boldsymbol{\psi}_k|\boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{k-1}^{(j+1)}, \boldsymbol{\psi}_{k+1}^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right),$$

where the conditioning elements reflect the fact that when the kth block is reached, the previous $(k - 1)$ blocks have already been updated. The transition density of the chain, again under the maintained assumption that π is absolutely continuous, is therefore given by the product of transition kernels for each block:

$$K\left(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}^{(j+1)}\right) = \prod_{k=1}^{p} \pi\left(\boldsymbol{\psi}_k|\boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{k-1}^{(j+1)}, \boldsymbol{\psi}_{k+1}^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right). \tag{25}$$

To illustrate the manner in which the blocks are revised, we consider a two block case, each with a single component, and trace out in Figure 4 a possible trajectory of the sampling algorithm. The contours in the plot represent the joint distribution of $\boldsymbol{\psi}$ and the labels "(0)", "(1)", etc., denote the simulated values. Note that one iteration of the algorithm is completed after both components are revised. Also notice that each component is revised along the direction of the coordinate axes. This feature can be a source of problems if the two components are highly correlated because then

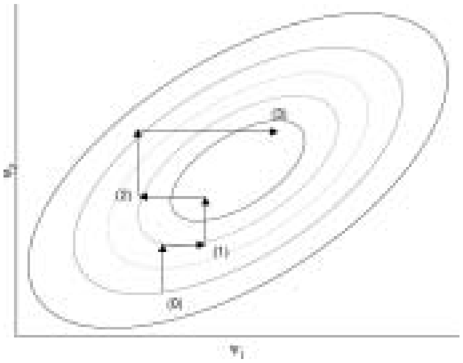

Fig. 4. Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations.

the contours become compressed and movements along the coordinate axes tend to produce only small moves. We return to this issue below.

5.2. Connection with the multiple-block M–H algorithm

A connection with the M–H algorithm can be drawn by noting that the full conditional distribution by Bayes theorem is proportional to the joint distribution, i.e.,

$$\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}) \propto \pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k}).$$

Now recall that the probability of move in the multiple-block M–H algorithm from Equation (20) is

$$\alpha_k(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k}) = \min\left\{\frac{\pi(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_{-k})\, q(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})}{\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})\, q(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k})}, 1\right\},$$

so if one substitutes

$$\begin{aligned} &q(\boldsymbol{\psi}_k, \boldsymbol{\psi}'_k|\boldsymbol{\psi}_{-k}) = \pi(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_{-k}),\\ &q(\boldsymbol{\psi}'_k, \boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}) = \pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k}),\end{aligned}$$

in this expression all the terms cancel implying that the probability of accepting the proposal is one. Thus, the Gibbs sampling algorithm is a special case of the multiple-block M–H algorithm.

It should be noted that a multiple-block M–H algorithm in which only some of the blocks are sampled using the full conditional distributions are sometimes called *hybrid samplers* or Metropolis-within-Gibbs samplers. These names are not very informative or precise and it is preferable to continue to refer to such algorithms as multiple-block M–H algorithms. The only algorithm that should properly be referred to as the Gibbs algorithm is the one in which each block is sampled directly from its full conditional distribution.

5.3. Invariance of the Gibbs Markov chain

The Gibbs transition kernel is invariant by construction. This is a consequence of the fact that the Gibbs algorithm is a special case of the multiple-block M–H algorithm which is invariant as was established in the last section. A direct calculation also reveals the same result. Consider for simplicity the situation of two blocks when the transition kernel density is

$$K(\boldsymbol{\psi}, \boldsymbol{\psi}') = \pi(\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\,\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1').$$

To check invariance we need to show that

$$\int K(\boldsymbol{\psi}, \boldsymbol{\psi}')\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\,\mathrm{d}\boldsymbol{\psi}_1\mathrm{d}\boldsymbol{\psi}_2 = \int \pi(\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\,\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')\,\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\,\mathrm{d}\boldsymbol{\psi}_1\mathrm{d}\boldsymbol{\psi}_2,$$

is equal to $\pi(\boldsymbol{\psi}_1', \boldsymbol{\psi}_2')$. This is easily verified because $\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')$ comes out of the integral, and the integral over $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ produces $\pi(\boldsymbol{\psi}_1')$. This calculation can be extended to any number of blocks in the same way. In addition, the Gibbs Markov chain is not reversible. Reversible Gibbs samplers are discussed by Liu, Wong and Kong (1995).

5.4. Sufficient conditions for convergence

Under rather general conditions, which are easy to verify, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of iterations become large. Formally, if we let $K(\boldsymbol{\psi}, \boldsymbol{\psi}')$ represent the transition density of the Gibbs algorithm and let $K^{(M)}(\boldsymbol{\psi}_0, \boldsymbol{\psi}')$ be the density of the draw $\boldsymbol{\psi}'$ after M iterations given the starting value $\boldsymbol{\psi}_0$, then

$$\| K^{(M)}\left(\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}'\right) - \pi(\boldsymbol{\psi}') \| \to 0 \quad \text{as} \quad M \to \infty. \tag{26}$$

Roberts and Smith (1994) [see also Chan (1993)] have shown that the conditions of Proposition 2 are satisfied under the following conditions: (i) $\pi(\boldsymbol{\psi}) > 0$ implies there exists an open neighborhood $N_{\boldsymbol{\psi}}$ containing $\boldsymbol{\psi}$ and $\epsilon > 0$ such that, for all $\boldsymbol{\psi}' \in N_{\boldsymbol{\psi}}$, $\pi(\boldsymbol{\psi}') \geqslant \epsilon > 0$; (ii) $\int f(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi}_k$ is locally bounded for all k, where $\boldsymbol{\psi}_k$ is the kth block of parameters; and (iii) the support of $\boldsymbol{\psi}$ is arc connected.

It is difficult to find non-pathological problems where these conditions are not satisfied.

5.5. Estimation of density ordinates

We mention that if the full conditional densities are available, whether in the context of the multiple-block M–H algorithm or that of the Gibbs sampler, then the MCMC output can be used to estimate posterior marginal density functions Tanner and Wong (1987) and Gelfand and Smith (1990). One possibility is to use a non-parametric kernel smoothing method which, however, suffers from the curse of dimensionality problem. A more efficient possibility is to exploit the fact that the marginal density of $\boldsymbol{\psi}_k$ at the point $\boldsymbol{\psi}_k^*$ is

$$\pi(\boldsymbol{\psi}_k^*) = \int \pi(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k})\, \pi(\boldsymbol{\psi}_{-k})\mathrm{d}\boldsymbol{\psi}_{-k},$$

where as before $\boldsymbol{\psi}_{-k} = \boldsymbol{\psi} \backslash \boldsymbol{\psi}_k$. Provided the normalizing constant of $\pi(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k})$ is known, we can estimate the marginal density as an average of the full conditional density over the simulated values of $\boldsymbol{\psi}_{-k}$:

$$\hat{\pi}(\boldsymbol{\psi}_k^*) = M^{-1} \sum_{j=1}^{M} \pi(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k}^{(j)}).$$

Then, under the assumptions of Proposition 1,

$$M^{-1} \sum_{j=1}^{M} \pi(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k}^{(j)}) \to \pi(\boldsymbol{\psi}_k^*), \quad \text{as} \quad M \to \infty.$$

Gelfand and Smith (1990) refer to this approach as "Rao–Blackwellization" because of the connections with the Rao–Blackwell theorem in classical statistics. That connection is more clearly seen in the context of estimating (say) the mean of $\boldsymbol{\psi}_k$, $E(\boldsymbol{\psi}_k) = \int \boldsymbol{\psi}_k \pi(\boldsymbol{\psi}_k)\, \mathrm{d}\boldsymbol{\psi}_k$. By the law of the iterated expectation,

$$E(\boldsymbol{\psi}_k) = E\{E(\boldsymbol{\psi}_k | \boldsymbol{\psi}_{-k})\},$$

and therefore the estimates

$$M^{-1} \sum_{j=1}^{M} \boldsymbol{\psi}_k^j,$$

and

$$M^{-1} \sum_{j=1}^{M} E(\boldsymbol{\psi}_k | \boldsymbol{\psi}_{-k}^{(j)}),$$

both converge to $E(\boldsymbol{\psi}_k)$ as $M \to \infty$. Under i.i.d. sampling, and under Markov sampling provided some conditions are satisfied [see Liu, Wong and Kong (1994), Geyer (1995), Casella and Robert (1996) and Robert and Casella (1999)], it can be shown that the variance of the latter estimate is smaller than that of the former. Thus, it can help to average the conditional mean $E(\boldsymbol{\psi}_k | \boldsymbol{\psi}_{-k})$, if that were available, rather than average the draws directly. Gelfand and Smith appeal to this analogy to argue that the Rao–Blackwellized estimate of the density is preferable to that based on the method of kernel smoothing. Chib (1995) extends the Rao–Blackwellization approach to estimate "reduced conditional ordinates" defined as the density of $\boldsymbol{\psi}_k$ conditioned on one or more of the remaining blocks. More discussion of this is provided below in Section 10 on Bayesian model choice. Finally, Chen (1994) provides an importance weighted estimate of the marginal density for cases where the conditional posterior density does not have a known normalizing constant. Chen's estimator is based on the identity

$$\pi(\boldsymbol{\psi}_k^*) = \int w(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}) \frac{\pi(\boldsymbol{\psi}_k^*, \boldsymbol{\psi}_{-k})}{\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})} \pi(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi},$$

where $w(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})$ is a completely known conditional density whose support is equal to the support of the full conditional density $\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})$. In this form, the normalizing constant of the full conditional density is not required and given a sample of draws $\{\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}\}$ from $\pi(\boldsymbol{\psi})$, a Monte Carlo estimate of the marginal density is given by

$$\hat{\pi}(\boldsymbol{\psi}_k^*) = M^{-1} \sum_{j=1}^{M} w(\boldsymbol{\psi}_k^{(j)}|\boldsymbol{\psi}_{-k}^{(j)}) \frac{\pi(\boldsymbol{\psi}_k^*, \boldsymbol{\psi}_{-k}^{(j)})}{\pi(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}_{-k}^{(j)})}.$$

Chen (1994) discusses the choice of the conditional density w. Since it depends on $\boldsymbol{\psi}_{-k}$, the choice of w will vary from one sampled draw to the next.

5.6. *Example: simulating a truncated multivariate normal*

To illustrate the Gibbs sampling algorithm consider the question of sampling a trivariate normal distribution truncated to the positive orthant. In particular, let the target distribution be

$$\pi(\boldsymbol{\psi}) = \frac{1}{\Pr(\boldsymbol{\psi} \in A)} f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, I(\boldsymbol{\psi} \in A) \propto f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, I(\boldsymbol{\psi} \in A),$$

where $\boldsymbol{\mu} = (.5, 1, 1.5)'$, $\boldsymbol{\Sigma}$ is in equi-correlated form with units on the diagonal and 0.7 on the off-diagonal, $A = (0, \infty) \times (0, \infty) \times (0, \infty)$ and $\Pr(\boldsymbol{\psi} \in A)$ is the normalizing constant which is difficult to compute. Following Geweke (1991), one may define the Gibbs sampler with the blocks ψ_1, ψ_2, ψ_3 and the full conditional distributions

$$\pi(\psi_1|\psi_2, \psi_3);\ \pi(\psi_2|\psi_1, \psi_3);\ \pi(\psi_3|\psi_1, \psi_2),$$

where each of the these full conditional distributions is univariate truncated normal restricted to the interval $(0, \infty)$:

$$\pi(\psi_k|\boldsymbol{\psi}_{-k}) \propto f_N\left(\psi_k|\mu_k + \boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}(\boldsymbol{\psi}_{-k} - \boldsymbol{\mu}_{-k}), \Sigma_k - \boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}\boldsymbol{C}_k\right) I(\psi_k \in (0, \infty)). \tag{27}$$

In this expression we have utilized the well known result about conditional normal distributions and have let $\boldsymbol{C}_k = \mathrm{Cov}(\psi_k, \boldsymbol{\psi}_{-k})$, $\boldsymbol{\Sigma}_{-k} = \mathrm{Var}(\boldsymbol{\psi}_{-k})$ and $\boldsymbol{\mu}_{-k} = E(\boldsymbol{\psi}_{-k})$. Note

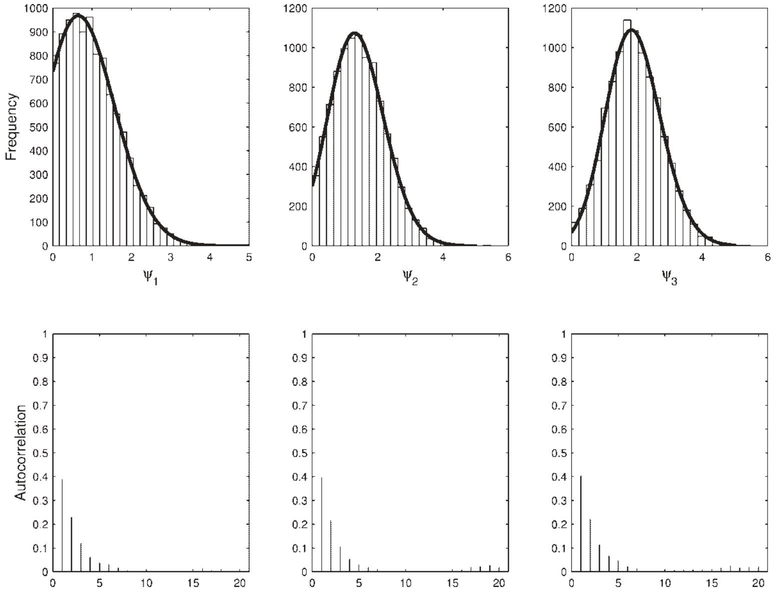

Fig. 5. Marginal distributions of ψ in truncated multivariate normal example (top panel). Histograms of the sampled values and Rao–Blackwellized estimates of the densities are shown. Autocorrelation plots of the Gibbs MCMC chain are in the bottom panel. Graphs are based on 10 000 iterations following a burn-in of 500 cycles.

that, unfortunately, the use of singleton block sizes is unavoidable in this problem because the conditional distribution of any two components given the third is not easy to simulate.

Figure 5 gives the marginal distribution of each component of ψ_k from a Gibbs sampling run of $M = 10\,000$ iterations with a burn-in of 100 cycles. The figure includes both the histograms of the sampled values and the Rao–Blackwellized estimates of the marginal densities based on the averaging of Equation (27) over the simulated values of $\boldsymbol{\psi}_{-k}$. The agreement between the two density estimates is close. In the bottom panel of Figure 5 we plot the autocorrelation function of the sampled draws. The rapid decline in the autocorrelations for higher lags indicates that the sampler is mixing well.

6. Sampler performance and diagnostics

In implementing a MCMC method it is important to assess the performance of the sampling algorithm to determine the rate of mixing and the size of the burn-in, both having implications for the number of iterations required to get reliable answers. A large literature has now emerged on these issues, for example, Robert (1995), Tanner (1996, Section 6.3), Cowles and Carlin (1996), Gammerman (1997, Section 5.4), Brooks, Dellaportas and Roberts (1997) and Robert and Casella (1999), but the ideas, although related in many ways, have not coalesced into a single prescription.

One approach for determining sampler performance and the size of the burn-in time is to employ analytical methods to the specified Markov chain, prior to sampling. This approach is exemplified in the work of, for example, Meyn and Tweedie (1994), Polson (1996), Roberts and Tweedie (1996) and Rosenthal (1995). Two factors have inhibited the growth and application of these methods. The first is that the calculations are difficult and problem-specific, and second, the upper bounds for the burn-in that emerge from such calculations are usually highly conservative.

At this time the more popular approach is to utilize the sampled draws to assess both the performance of the algorithm and its approach to the stationary, invariant distribution. Several such relatively informal methods are now available. Gelfand and Smith (1990) recommend monitoring the evolution of the quantiles as the sampling proceeds. Another quite useful diagnostic, one that is perhaps the simplest and most direct, are autocorrelation plots (and autocorrelation times) of the sampled output. Slowly decaying correlations indicate problems with the mixing of the chain. It is also useful in connection with M–H Markov chains to monitor the acceptance rate of the proposal values with low rates implying "stickiness" in the sampled values and thus a slower approach to the invariant distribution.

Somewhat more formal sample-based diagnostics are also available in the literature, as summarized in the CODA routines provided by Best, Cowles and Vines (1995). Although these diagnostics often go under the name "convergence diagnostics" they are in principle approaches that detect *lack* of convergence. Detection of convergence based entirely on the sampled output, without analysis of the target distribution, is extremely difficult and perhaps impossible. Cowles and Carlin (1996) discuss and evaluate thirteen such diagnostics [for example, those proposed by Geweke (1992), Raftery and Lewis (1992), Ritter and Tanner (1992), Gelman and Rubin (1992), Zellner and Min (1995), amongst others] without arriving at a consensus. Difficulties in evaluating these methods stem from the fact that some of these methods apply only to Gibbs Markov chains [for example, those of Ritter and Tanner (1992) and Zellner and Min (1995)] while others are based on the output not just of a single chain but on that of multiple chains specifically run from "disparate starting values" as in the method of Gelman and Rubin (1992). Finally, some methods assess the behavior of univariate moment estimates [as in the approach of Geweke (1992) and Gelman and Rubin (1992)] while others are concerned with the behavior of the entire transition kernel [as in Ritter and Tanner (1992) and Zellner and Min (1995)]. Further developments in this area are ongoing.

7. Strategies for improving mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation time, because such samplers can be expected to converge more quickly to the invariant distribution. Over the years a number of different recipes for designing samplers with low autocorrelation times have been proposed although it may sometimes be difficult, because of the complexity of the problem, to apply any of these recipes.

7.1. Choice of blocking

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution and the sampled draws would tend to display autocorrelations that decay slowly. To get a sense of the problem, it may be worthwhile for the reader to use the Gibbs sampler to simulate a bivariate normal distribution with unit variances and covariance (correlation) of 0.95.

The importance of coarse, or highly grouped, blocking has been highlighted in a number of different problems for example, the state space model, hidden Markov model and longitudinal data models with random effects. In each of these situations, which are further discussed below in detail, the parameter space is quite large on account of the fact that auxiliary variables are included in the sampling (the latent states in the case of the state space model and the random effects in the case of the longitudinal data model). These latent variables tend to be highly correlated either amongst themselves, as in the case of the state space model, or with a different set of variables as in the case of the panel model.

Blocks can be combined by the method of composition. For example, suppose that $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ denote three blocks and that the distribution $\boldsymbol{\psi}_1|\boldsymbol{\psi}_3$ is tractable (i.e., can be sampled directly). Then, the blocks $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ can be collapsed by first sampling $\boldsymbol{\psi}_1$ from $\boldsymbol{\psi}_1|\boldsymbol{\psi}_3$ followed by $\boldsymbol{\psi}_2$ from $\boldsymbol{\psi}_2|\boldsymbol{\psi}_1, \boldsymbol{\psi}_3$. This amounts to a two block MCMC algorithm. In addition, if it is possible to sample $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ marginalized over $\boldsymbol{\psi}_3$ then the number of blocks is reduced to one. Liu (1994) and Liu, Wong and Kong (1994) discuss the value of these strategies in the context of a three-block Gibbs MCMC chains. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

7.2. Tuning the proposal density

As mentioned above, the proposal density in a M–H algorithm has an important bearing on the mixing of the MCMC chain. Fortunately, one has great flexibility in the choice of candidate generating density and it is possible to adapt the choice to the specific context of a given problem. For example, Chib, Greenberg and Winkelmann (1998) develop and compare four different choices in the context of longitudinal random effects for count data. In this problem, each cluster (or individual) has its own random effects and each of these has to be sampled from an intractable target distribution.

If one lets n denote the number of clusters, where n is typically large, say in excess of a thousand, then the number of blocks in the MCMC implementation is $n + 3$ (n for each of the random effect distributions, two for the fixed effects and one for the variance components matrix). For this problem, the multiple-block M–H algorithm requires $n + 1$ M–H steps within one iteration of the algorithm. Tailored proposal densities are therefore computationally quite expensive but one can use a mixture of proposal densities where a less demanding proposal, for example a random walk proposal, is combined with the tailored proposal to sample each of the n random effect target distributions. Further discussion of mixture proposal densities for the purpose of improving mixing is contained in Tierney (1994).

7.3. Other strategies

In some problems it is possible to reparameterize the variables to make the blocks less correlated. See Hills and Smith (1992) and Gelfand, Sahu and Carlin (1995) where under certain circumstances reparameterization is shown to be beneficial for simple one-way analysis of variance models, and for general hierarchical normal linear models.

Another strategy that can prove useful is importance resampling in which the MCMC sampler is applied not to the target distribution π but to a modified distribution π^*, for which a well mixing sampler can be designed, and which is close to π. Now suppose $\{\boldsymbol{\psi}^{(1)}, \dots, \boldsymbol{\psi}^{(M)}\}$ are draws from the target distribution π^*. These can be made to correspond to the target distribution π by attaching the weight $w_j = \pi(\boldsymbol{\psi}^{(j)})/\pi^*(\boldsymbol{\psi}^{(j)})$ to each draw and then re-sampling the sampled values with probability given by $\{w_j / \sum_{g=1}^{M} w_g\}$. This strategy was introduced for a different purpose by Rubin (1988) and then employed by Gelfand and Smith (1992) and Albert (1993) to study the sensitivity of the posterior distribution to small changes in the prior without involving a new MCMC calculation. Its use for improving mixing in the MCMC context is illustrated by Kim, Shephard and Chib (1998) where a nonlinear state space model of stochastic volatility is approximated accurately by a mixture of state space models; an efficient MCMC algorithm is then developed for the latter target distribution and the draws are finally re-sampled to correspond to the original nonlinear model.

Other approaches have also been discussed in the literature. Marinari and Parisi (1992) develop the simulated tempering method whereas Geyer and Thompson (1995) develop a related technique that they call the Metropolis-coupled MCMC method. Both these approaches rely on a series of transition kernels $\{K_1, \dots, K_m\}$ where only K_1 has π^* as the stationary distribution. The other kernels have equilibrium distributions π_i, which Geyer and Thompson take to be $\pi_i(\boldsymbol{\psi}) = \pi(\boldsymbol{\psi})^{1/i}$, $i = 2, \dots, m$. This specification produces a set of target distributions that have higher variance than π^*. Once the transition kernels and equilibrium distributions are specified then the Metropolis-coupled MCMC method requires that each of the m kernels be used in parallel. At each iteration, after the m draws have been obtained, one randomly selects two chains to see if the states should be swapped. The probability of swap is based on the M–H acceptance condition. At the conclusion of the sampling, inference is based on the sequence of draws that correspond to the distribution π^*. These methods promote rapid mixing because draws from the various "flatter" target densities have a chance of being swapped with the draws from the base kernel K_1. Thus, variates that are unlikely under the transition K_1 have a chance of being included in the chain, leading to more rapid exploration of the parameter space.

8. MCMC algorithms in Bayesian estimation

8.1. Overview

Markov chain Monte Carlo methods have proved enormously popular in Bayesian statistics [for wide-ranging discussions of the Bayesian paradigm see, for example, Zellner (1971), Leamer (1978), Berger (1985), O'Hagan (1994), Bernardo and Smith (1994), Poirier (1995), Gelman, Meng, Stern and Rubin (1995)], where these methods have opened up vistas that were unimaginable fifteen years ago. Within the Bayesian framework, where both parameters and data are treated as random variables and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods. Sometimes the target distribution is the posterior distribution of the parameters augmented by latent data, in which case the MCMC scheme operates on a space that is considerably larger than the parameter space. This strategy, which goes under the name of data augmentation, is illustrated in several models below and its main virtue is that it allows one to conduct the MCMC simulation without having to evaluate the likelihood function of the parameters. The latter feature is of considerable importance especially when the model of interest has a complicated likelihood function and likelihood based inference is difficult. Admittedly, in standard problems such as the linear regression model, there may be little to be gained by utilizing MCMC methods or in fact by adopting the Bayesian approach, but the important point is that MCMC methods provide a complete computational toolkit for conducting Bayesian inference in models that are both simple and complicated. This is the central reason for the current growing appeal of Bayesian methods in theoretical and practical work and this appeal is likely to increase once MCMC Bayesian software, presently under development at various sites, becomes readily available.

Papers that develop some of the important general MCMC ideas for Bayesian inference appeared early in the 1990's. Categorized by topics, these include, normal and student-*t* data models [Gelfand et al. (1990), Carlin and Polson (1991)]; binary and ordinal response models [Albert and Chib (1993a, 1995)]; tobit censored regression models [Chib (1992)]; generalized linear models [Dellaportas and Smith (1993), Mallick and Gelfand (1994)]; change point models [Carlin et al. (1992), Stephens (1994)]; autoregressive models [Chib (1993), McCulloch and Tsay (1994)];

autoregressive-moving average models [Chib and Greenberg (1994)]; hidden Markov models [Albert and Chib (1993b), Robert et al. (1993), McCulloch and Tsay (1994), Chib (1996)]; state space models [Carlin, Polson and Stoffer (1992), Carter and Kohn (1994, 1996), Chib and Greenberg (1995b), de Jong and Shephard (1995)]; measurement error models [Mallick and Gelfand (1996)]; mixture models [Diebolt and Robert (1994), Escobar and West (1995), Muller, Erkanli and West (1996)]; longitudinal data models [Zeger and Karim (1991), Wakefield et al. (1994)].

More recently, other model and inference situations have also come under scrutiny. Examples include, ARMA models with switching [Billio, Monfort and Robert (1999)]; CART models [Chipman, George and McCulloch (1998), Denison, Mallick and Smith (1998)]; conditionally independent hierarchical models [Albert and Chib (1997)]; estimation of HPD intervals [Chen and Shao (1999)]; item response models [Patz and Junker (1999)]; selection models [Chib and Hamilton (2000)]; partially linear and additive regression models [Lenk (1999), Shively, Kohn and Wood (1999)]; sequential Monte Carlo for state space models [Liu and Chen (1998), Pitt and Shephard (1999)]; stochastic differential equation models [Elerian, Chib and Shephard (1999)]; models with symmetric stable distributions [Tsionas (1999)]; neural network models [Muller and Insua (1998)]; spatial models [Waller, Carlin, Xia and Gelfand (1997)].

MCMC methods have also been extended to the realm of Bayesian model choice. Problems related to variable selection in regression models, hypothesis testing in nested models and the general problem of model choice are now all amenable to analysis by MCMC methods. The basic strategies are developed in the following papers: variable selection in regression [George and McCulloch (1993)]; hypothesis testing in nested models [Verdinelli and Wasserman (1995)]; predictive model comparison [Gelfand and Dey (1994)]; marginal likelihood and Bayes factor computation [Chib (1995)]; composite model space and parameter space MCMC [Carlin and Chib (1995), Green (1995)]. These developments are discussed in Section 10.

We now provide a set of applications of MCMC methods to models largely drawn from the list above. These models serve to illustrate a number of general techniques, for example, derivations of full conditional distributions, use of latent variables in the sampling (data augmentation) to avoid computation of the likelihood function, and issues related to blocking. Because of the modular nature of MCMC methods, the algorithms presented below can serve as the building blocks for other models not considered here. In some instances one would only need to combine different pieces of these algorithms to fit a new model.

8.2. Notation and assumptions

To streamline the discussion we collect some of the notation that is used in the rest of the paper.

The d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$ is denoted by $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Its density at the point $\boldsymbol{t} \in \Re^d$ is denoted by $\phi_d(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Omega})$. The univariate normal density truncated to the interval (a, b) is denoted by $\mathcal{TN}_{[a,b]}(\mu, \sigma^2)$

with density at the point $t \in (a, b)$ given by $\phi(t|\mu, \sigma^2)/[\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)]$, where ϕ is the univariate normal density and $\Phi(\cdot)$ is the c.d.f. of the standard normal random variable.

A d-variate random vector distributed according to the multivariate-t distribution with mean vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$ and ξ degrees of freedom has density $f_T(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \xi)$ given by

$$\frac{\Gamma((\xi+1)/2)\Gamma(\xi/2)}{(\xi\pi)^{1/2}|\boldsymbol{\Omega}|^{1/2}} \left\{ 1 + \frac{1}{\xi}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{t}-\boldsymbol{\mu}) \right\}^{-(\xi+d)/2}.$$

The gamma distribution is denoted by $\Gamma(a, b)$ with density at the point t by $f_G(t|a, b) \propto t^{a-1}\exp(-bt)\, I[t > 0]$, where $I[A]$ is the indicator function of the event A. The inverse gamma distribution is the distribution of the inverse of a gamma variate.

A random symmetric positive definite matrix $\boldsymbol{W}\colon p \times p$ is said to follow a Wishart distribution $\mathcal{W}_p(\boldsymbol{W}|\upsilon, \boldsymbol{R})$ if the density of $\boldsymbol{W}$ is given by

$$c\frac{|\boldsymbol{W}|^{(\nu-p-1)/2}}{|\boldsymbol{R}|^{\nu/2}} \exp\left\{-\tfrac{1}{2}\operatorname{tr}(\boldsymbol{R}^{-1}\boldsymbol{W})\right\}, \quad |\boldsymbol{W}| > 0,$$

where c is a normalizing constant, $\boldsymbol{R}$ is a hyperparameter matrix and "tr" is the trace function. To simulate the Wishart distribution, one utilizes the expression $\boldsymbol{W} = \boldsymbol{LTT'L'}$, where $\boldsymbol{R} = \boldsymbol{LL'}$ and $\boldsymbol{T} = (t_{ij})$ is a lower triangular matrix with $t_{ii} \sim \sqrt{\chi^2_{\upsilon-i+1}}$ and $t_{ij} \sim \mathcal{N}(0, 1)$.

In connection with the sampling design of the observations and the error terms we use "ind" to denote independent and "i.i.d." to denote independent and identically distributed. The response variable (or vector) of the model is denoted by either y_i or y_t, the sample size by n and the entire collection of sample data by $\boldsymbol{y} = (y_1, \ldots, y_n)$. In some instances, we let $\boldsymbol{Y}_t = (y_1, \ldots, y_t)$ denote the data upto time t and $\boldsymbol{Y}^t = (y_t, \ldots, y_n)$ to denote the values from t to the end of the sample. The covariates are denoted as x_i if the corresponding response is a scalar and as $\boldsymbol{X}_i$ or $\boldsymbol{X}_t$ if the response is a vector. The regression coefficients are denoted by $\boldsymbol{\beta}$ and the error variance (if y_i is a scalar) by σ^2 and the error covariance by $\boldsymbol{\Omega}$ if y_i is a vector. The parameters of the model are denoted by $\boldsymbol{\theta}$ and the variables used in the MCMC simulation by $\boldsymbol{\psi}$ (consisting of $\boldsymbol{\theta}$ and other quantities).

When denoting conditional distributions only dependence on random quantities, such as parameters and random effects, is included in the conditioning set. Covariates are never included in the conditioning. The symbol p is used to denote the prior density if general notation is required.

It is always assumed that each distinct set of parameters, for example, regression coefficients and covariance elements, are a priori independent. The joint prior distribution is therefore specified through the marginal distribution of each distinct set of parameters. Distributions for the parameters are chosen from the class of conditionally conjugate distributions in keeping with the existing literature on these models. The parameters of the prior distributions, called hyperparameters, are assumed known. These will be indicated by the subscript “0.” In some cases, when the hyperparameters are unknown, hierarchical priors, defined by placing prior distributions on the prior hyperparameters, are used.

8.3. Normal and student-t regression models

Consider the univariate regression model defined by the specification

$$\begin{aligned} y_i|\mathcal{M}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \quad i \leqslant n,\\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0),\\ \sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2}, \frac{\delta_0}{2}\right). \end{aligned}$$

The target distribution is

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathcal{M}, \boldsymbol{y}) \propto p(\boldsymbol{\beta})p(\sigma^2)\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2),$$

and MCMC simulation proceeds by a Gibbs chain defined through the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \sigma^2;\ \sigma^2|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}.$$

Each of these distributions is straightforward to derive because conditioned on σ^2 both the prior and the likelihood have Gaussian forms (and hence the updated distribution is Gaussian with moments found by completing the square for the terms in the exponential function) while conditioned on $\boldsymbol{\beta}$, the updated distribution of σ^2 is inverse gamma with parameters found by adding the exponents of the prior and the likelihood.

Algorithm 5: Gaussian multiple regression

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

(2) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right\}$$

(3) `Goto 1`.

This algorithm can be easily modified to permit the observations y_i to follow a Student-*t* distribution. The modification, proposed by Carlin and Polson (1991), utilizes the fact that if

$$\lambda_i \sim \mathcal{G}\left(\frac{\xi}{2}, \frac{\xi}{2}\right),$$

and

$$y_i | \mathcal{M}, \boldsymbol{\beta}, \sigma^2, \lambda_i \sim \mathcal{N}(\boldsymbol{x}_i' \boldsymbol{\beta}, \lambda_i^{-1} \sigma^2),$$

then

$$y_i | \mathcal{M}, \boldsymbol{\beta}, \sigma^2 \sim f_T(y_i | \boldsymbol{x}_i' \boldsymbol{\beta}, \sigma^2, \xi), \quad i \leqslant n.$$

Hence, if one defines $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2, \{\lambda_i\})$ then, conditioned on $\{\lambda_i\}$, the model is Gaussian and a variant of Algorithm 5 can be used. Furthermore, conditioned on $(\boldsymbol{\beta}, \sigma^2)$, the full conditional distribution of $\{\lambda_i\}$ factors into a product of independent Gamma distributions.

Algorithm 6: Student-*t* multiple regression

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_{n,\lambda}\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i y_i\right),\ \boldsymbol{B}_{n,\lambda} = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}\right).$$

(2) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n} \lambda_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right\}.$$

(3) `Sample`

$$\lambda_i \sim \mathcal{G}\left[\frac{\xi + 1}{2}, \frac{\xi + \sigma^{-2}(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{2}\right], \quad i \leqslant n.$$

(4) `Goto 1`.

Another modification of Algorithm 5 is to Zellner's seemingly unrelated regression model (SUR). In this case a vector of p observations are generated from the model

$$\begin{aligned} y_t | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Omega} &\sim \mathcal{N}(\boldsymbol{X}_t \boldsymbol{\beta}, \boldsymbol{\Omega}), \quad t \leqslant n, \\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0), \\ \boldsymbol{\Omega}^{-1} &\sim \mathcal{W}_p(\nu_0, \boldsymbol{R}_0), \end{aligned}$$

where $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})'$, $\boldsymbol{X}_t = \mathrm{diag}(\boldsymbol{x}_{1t}', \ldots, \boldsymbol{x}_{pt}')$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')' : k \times 1$, and $k = \sum_i k_i$.

To deal with this model, a two block MCMC approach can be used as proposed by Blattberg and George (1991) and Percy (1992). Chib and Greenberg (1995b) extend that algorithm to SUR models with hierarchical priors and time-varying parameters of the type considered by Gammerman and Migon (1993).

For the SUR model, the posterior density of the parameters is proportional to

$$\pi(\boldsymbol{\beta})\pi(\boldsymbol{\Omega}^{-1}) \times \left|\boldsymbol{\Omega}^{-1}\right|^{n/2} \exp\left\{-\tfrac{1}{2}\sum_{t=1}^{n}(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta})\right\},$$

and the MCMC algorithm is defined by the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\Omega}^{-1};\ \boldsymbol{\Omega}^{-1}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}.$$

These are both tractable, with the former a normal distribution and the latter a Wishart distribution.

Algorithm 7: Gaussian SUR

```
(1) Sample
        $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{t=1}^{n}\boldsymbol{X}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{y}_t\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{t=1}^{n}\boldsymbol{X}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{X}_t\right)^{-1}\right).$
(2) Sample
        $\boldsymbol{\Omega}^{-1} \sim \mathcal{W}_p\left[\nu_0 + n, \left\{\boldsymbol{R}_0^{-1} + \sum_{t=1}^{n}(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta})(\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta})'\right\}^{-1}\right].$
(3) Goto 1.
```

8.4. Binary and ordinal probit

Suppose that each y_i is binary and the model of interest is

$$y_i|\mathcal{M}, \boldsymbol{\beta} \sim \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}), i \leqslant n; \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0).$$

The posterior distribution does not belong to a named family of distributions. To deal with the problem, Albert and Chib (1993a) introduce a technique that has formed the basis for a unified methodology for univariate and multivariate binary and ordinal response models and led to many applications. The Albert–Chib algorithm capitalizes on the simplifications afforded by introducing latent or auxiliary data into the sampling.

Instead of the specification above, the model of interest is specified in equivalent form as

$$z_i|\mathcal{M}, \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1), \quad y_i = I[z_i > 0], i \leqslant n, \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0).$$

Now the MCMC Gibbs algorithm proceeds with the sampling of the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \{z_i\}; \quad \{z_i\}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta},$$

where

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \{z_i\} \overset{d}{=} \boldsymbol{\beta}|\mathcal{M}, \{z_i\},$$

has the same form as in the linear regression model with σ^2 set equal to one and y_i replaced by z_i and

$$\{z_i\}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta} \overset{d}{=} \prod_{i=1}^{n} z_i|y_i, \mathcal{M}, \boldsymbol{\beta},$$

factor into a set of n independent distributions with each depending on the data only through y_i. The distributions $z_i|y_i, \mathcal{M}, \boldsymbol{\beta}$ are obtained by reasoning as follows. Suppose that $y_i = 0$, then from Bayes theorem

$$\begin{aligned} f(z_i|y_i = 0, \mathcal{M}, \boldsymbol{\beta}) &\propto f_N(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\, f(y_i = 0|z_i, \mathcal{M}, \boldsymbol{\beta}) \\ &\propto f_N(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\, I[z_i \leqslant 0], \end{aligned}$$

because $f(y_i = 0|z_i, \mathcal{M}, \boldsymbol{\beta})$ is equal to one if z_i is negative and equal to zero otherwise, which is the definition of $I[z_i \leqslant 0]$. Hence, the information $y_i = 0$ simply serves to truncate the support of z_i. By a similar argument it is shown that the support of z_i is $(0, \infty)$ when conditioned on the event $y_i = 1$. Each of these truncated distributions is simulated by the formula given in Equation (5). This leads to the following algorithm.

Algorithm 8: Binary probit

```
(1) Sample
        $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i z_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$
(2) Sample
        $z_i \sim \begin{cases} \mathcal{TN}_{(-\infty,0]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) & \text{if } y_i = 0, \\ \mathcal{TN}_{(0,\infty)}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) & \text{if } y_i = 1, \end{cases} \quad i \leqslant n.$
(3) Goto 1.
```

Albert and Chib (1993a) also extend this algorithm to the ordinal categorical data case where y_i can take one of the values $\{0, 1, \ldots, J\}$ according to the probabilities

$$\Pr(y_i \leqslant j|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi(\gamma_j - \boldsymbol{x}_i'\boldsymbol{\beta}), \quad j = 0, 1, \ldots, J. \tag{28}$$

In this model the $\{\gamma_j\}$ are category specific cut-points with γ_0 normalized to zero and γ_J to infinity. The remaining cut-points $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})$ are assumed to satisfy the order restriction $\gamma_1 \leqslant \cdots \leqslant \gamma_{J-1}$ which ensures that the cumulative probabilities are non-decreasing. For given data $y_1, \ldots, y_n$ from this model, the likelihood function is given by

$$f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{j=0}^{J} \prod_{i:y_i=j} \left[\Phi(\gamma_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})\right], \tag{29}$$

and the posterior density, under the prior $p(\boldsymbol{\beta}, \boldsymbol{\gamma})$, is proportional to $p(\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$. Posterior simulation is again feasible with the the introduction of latent variables $z_1, \ldots, z_n$, where $z_i|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, 1)$. A priori, we observe $y_i = j$ if the latent variable z_i falls in the interval $[\gamma_{j-1}, \gamma_j)$. Now the basic Albert and Chib MCMC scheme draws the latent data, regression parameters and cut-points in sequence. Given $y_i = j$, the sampling of the latent data z_i is from $\mathcal{TN}_{[\gamma_{j-1}, \gamma_j]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ and the sampling of the parameters $\boldsymbol{\beta}$ is as in Algorithm 8. For the cut-points, Cowles (1996) and Nandram and Chen (1996) proposed that the cut-points be generated by the M–H algorithm, marginalized over $\boldsymbol{z}$. Subsequently, Albert and Chib (1998) simplified the latter step by transforming the cut-points $\boldsymbol{\gamma}$ so as to remove the ordering constraint. The transformation is defined by the one-to-one map

$$\delta_1 = \log \gamma_1; \, \delta_j = \log(\gamma_j - \gamma_{j-1}), \quad 2 \leqslant j \leqslant J - 1. \tag{30}$$

The advantage of working with $\boldsymbol{\delta}$ instead of $\boldsymbol{\gamma}$ is that the parameters of the tailored proposal density in the M–H step for $\boldsymbol{\delta}$ can be obtained by an unconstrained optimization and the prior $p(\boldsymbol{\delta})$ on $\boldsymbol{\delta}$ can be an unrestricted multivariate normal. The algorithm is defined as follows.

Algorithm 9: Ordinal probit

```
(1) M-H
    (a) Calculate
            $\boldsymbol{m} = \arg\max_{\boldsymbol{\delta}} \log f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta})$,
        and $\boldsymbol{V} = \{-\partial \log f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta})/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'\}^{-1}$, the negative inverse of
        the hessian at $\boldsymbol{m}$.
    (b) Propose
            $\boldsymbol{\delta}' \sim f_T(\boldsymbol{\delta}|\boldsymbol{m}, \boldsymbol{V}, \xi).$
    (c) Calculate
            $\alpha = \min\left\{\frac{p(\boldsymbol{\delta}')f(\boldsymbol{y}|\mathcal{M},\boldsymbol{\beta},\boldsymbol{\delta}')}{p(\boldsymbol{\delta})f(\boldsymbol{y}|\mathcal{M},\boldsymbol{\beta},\boldsymbol{\delta})}\frac{f_T(\boldsymbol{\delta}|\boldsymbol{m},\boldsymbol{V},\xi)}{f_T(\boldsymbol{\delta}'|\boldsymbol{m},\boldsymbol{V},\xi)}, 1\right\}.$
    (d) Move to $\boldsymbol{\delta}'$ with probability $\alpha$. Transform the new $\boldsymbol{\delta}$ to $\boldsymbol{\gamma}$ via
        the inverse map $\gamma_j = \sum_{i=1}^{j} \exp(\delta_i),\ 1 \leqslant j \leqslant J-1.$
(2) Sample
        $z_i \sim \mathcal{TN}_{[\gamma_{j-1},\gamma_j]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) \quad \text{if} \quad y_i = j, \quad i \leqslant n.$
(3) Sample
        $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i z_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right).$
(4) Goto 1.
```

8.5. Tobit censored regression

Consider now a model in the class of the Tobit family in which the data y_i is generated by

$$z_i|\mathcal{M}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

$$y_i = \max(0, z_i), \quad 1 \leqslant i \leqslant n,$$

indicating that the observation z_i is observed only when z_i is positive. This model gives rise to a mixed discrete-continuous distribution with a point mass of $[1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)]$ at zero and a density $f_N(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$ on $(0, \infty)$. The likelihood function is given by

$$\prod_{i \in C}\{1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)\}\prod_{i \in C'}(\sigma^{-2})\exp\left\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right\},$$

where C is the set of censored observations and Φ is the c.d.f. of the standard normal random variable.

A MCMC procedure for this model is developed by Chib (1992) while Wei and Tanner (1990a) discuss a related approach for a model that arises in survival analysis. A set of tractable full conditional distributions is obtained by including the vector $\boldsymbol{z} = (z_i)$, $i \in C$ in the sampling. Let $\boldsymbol{y}_z = (y_{zi})$ be a $n \times 1$ vector with ith component y_i if the ith observation is not censored and z_i if it is censored. Now apply the Gibbs sampling algorithm with blocks $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z})$ and associated full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}, \sigma^2; \quad \sigma^2|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2.$$

The first two of these distributions follow from the results for linear regression with Gaussian errors (with y_{zi} used in place of y_i) and the third distribution, analogous to the probit case, is truncated normal on the interval $(-\infty, 0]$.

Algorithm 10: Tobit censored regression

```
(1) Sample
        $$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t' y_{zi}\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\right).$$
(2) Sample
        $$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_{zi} - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right\}.$$
(3) Sample
        $$z_i \sim \mathcal{TN}_{(-\infty,0]}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \quad i \in C.$$
(4) Goto 1.
```

8.6. Regression with change point

Suppose that $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ is a time series such that the density of y_t given $\boldsymbol{Y}_{t-1} = (y_1, \ldots, y_{t-1})$ is specified as

$$y_t|\mathcal{M}, \boldsymbol{Y}_{t-1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \tau \sim \begin{cases} \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_1, \sigma_1^2) & \text{if } t \leqslant \tau, \\ \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_2, \sigma_2^2) & \text{if } \tau < t, \end{cases}$$

where τ is an unknown change point. The objective is to estimate the parameter vectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, the regression variances $\sigma^2 = (\sigma_1^2, \sigma_2^2)$ and the change point τ.

An analysis of such models from a MCMC perspective was initiated by Carlin, Gelfand and Smith (1992). It is based on the inclusion of the change point τ in the MCMC sampling. Stephens (1994) generalized the approach of Carlin, Gelfand and Smith for models with multiple change points by including each of the unobserved change points in the sampling. In this generalization, however, the step that involves the simulation of the change points conditioned on the parameters and the data can be computationally very demanding when the sample size n is large. A different approach to multiple change point problems which is computationally simpler is developed by Chib (1998). An important aspect of the MCMC approach for change point problems is that it can be easily adapted for binary and count data.

Assume that

$$\boldsymbol{\beta}_j \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0); \quad \sigma_j^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right); \quad \tau \sim \text{Unif}\{a_0, a_0+1, \ldots, b_0\},$$

where τ follows a discrete uniform distribution on the integers $\{a_0, b_0\}$. Then the posterior density is

$$\pi(\boldsymbol{\beta}, \sigma^2, \tau | \boldsymbol{y}, \mathcal{M}) \propto p(\boldsymbol{\beta}) p(\sigma^2) p(\tau) \prod_{t \leqslant \tau} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{\tau < t} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

Conditional on τ the data splits into two parts and the conditional distributions of the regression parameters are obtained from the regression updates of Algorithm 5. On the other hand, given the regression parameters, the full conditional distribution of τ is concentrated on $\{a_0, b_0\}$ with mass function

$$\Pr(\tau = k | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2) \propto \prod_{t \leqslant k} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{k < t} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

The normalizing constant of this mass function is the sum of the right hand side over k.

Algorithm 11: Regression with change point

(1) `Sample for` $j = 1, 2$

$$\boldsymbol{\beta}_j \sim \mathcal{N}_k(\hat{\boldsymbol{\beta}}_j, \boldsymbol{B}_j),$$

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{B}_j \left(\boldsymbol{B}_0^{-1} \boldsymbol{\beta}_0 + \sigma_j^{-2} \sum_{t=l_j}^{u_j} \boldsymbol{x}_t' y_t \right),$$

$$\boldsymbol{B}_j = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2} \sum_{t=l_j}^{u_j} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1},$$

$$l_j = 1 + (j-1)\tau; \; u_j = \tau + (j-1)(n-\tau).$$

(2) `Sample for` $j = 1, 2$

$$\sigma_j^2 \sim \mathcal{IG}\left\{ \frac{\upsilon_0 + n_j}{2}, \frac{\delta_0 + \sum_{t=l_j}^{u_j} (y_t - \boldsymbol{x}_t' \boldsymbol{\beta}_j)^2}{2} \right\},$$

$$n_j = \tau + (j-1)(n - 2\tau).$$

(3) `Calculate for` $k = a_0, a_0 + 1, \ldots, b_0$

$$p_k \propto \prod_{t \leqslant k} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{k < t} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

(4) `Sample`

$$\tau \sim \{p_{a_0}, p_{a_0+1}, \ldots, p_{b_0}\}.$$

(5) `Goto 1`.

8.7. Autoregressive time series

Consider the model

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \epsilon_t, \quad 1 \leqslant t \leqslant n,$$

where the error is generated by the stationary AR(p) process

$$\epsilon_t - \phi_1 \epsilon_{t-1} - \cdots - \phi_p \epsilon_{t-p} = u_t \quad \text{or} \quad \phi(L)\, \epsilon_t = u_t,$$

where $u_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ is a polynomial in the lag operator L. One interesting complication in this model is that the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, due to the stationarity assumption, are restricted to lie in the region $S_{\boldsymbol{\phi}}$ of $\Re^p$ where the roots of $\phi(L)$ are all outside the unit circle. Chib and Greenberg (1994), based on Chib (1993), derive a multiple-block Metropolis–Hastings MCMC algorithm for this model in which the proposal densities for $\boldsymbol{\beta}$ and σ^2 are the respective full conditional densities while that of $\boldsymbol{\phi}$ is a normal density constructed from the observations y_t, $t \geqslant p + 1$.

Denote the first p observations as $\boldsymbol{Y}_p = (y_1, \ldots, y_p)'$ and $\boldsymbol{X}_p = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)'$ and let $y_t^* = \phi(L) y_t$ and $\boldsymbol{x}_t^* = \phi(L) \boldsymbol{x}_t$, $t \geqslant p + 1$. Also define the p dimensional matrix $\boldsymbol{\Sigma}_p$ through the matrix equation

$$\boldsymbol{\Sigma}_p = \boldsymbol{\Phi} \boldsymbol{\Sigma}_p \boldsymbol{\Phi}' + \boldsymbol{e}_1(p)\, \boldsymbol{e}_1(p)',$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}_{-p}' & \phi_p \\ \boldsymbol{I}_{p-1} & \mathbf{0} \end{pmatrix},$$

$\boldsymbol{e}_1(p) = (1, 0, \ldots, 0)'$ and $\boldsymbol{\phi}_{-p} = (\phi_1, \ldots, \phi_{p-1})'$. Let the cholesky factorization of $\boldsymbol{\Sigma}_p$ be $\boldsymbol{Q}\boldsymbol{Q}'$ and define $\boldsymbol{Y}_p^* = \boldsymbol{Q}^{-1} \boldsymbol{Y}_p$ and $\boldsymbol{X}_p^* = \boldsymbol{Q}^{-1} \boldsymbol{X}_p$ which are functions of $\boldsymbol{\phi}$. Finally define $e_t = y_t - \boldsymbol{x}_t' \boldsymbol{\beta}$, $t \geqslant p + 1$.

One can now proceed by noting that given $\boldsymbol{\phi}$, updates of $\boldsymbol{\beta}$ and σ^2 follow from the model

$$\begin{aligned} y_t^*|\mathcal{M},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_t^{*\prime}\boldsymbol{\beta},\sigma^2), \quad t \geqslant 1,\\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0,\boldsymbol{B}_0),\\ \sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2},\frac{\delta_0}{2}\right), \end{aligned}$$

while conditioned on $(\boldsymbol{\beta},\sigma^2)$, and the assumption that the prior density of $\boldsymbol{\phi}$ is $\mathcal{N}(\boldsymbol{\phi}_0,G_0)$ truncated to the region $S_{\boldsymbol{\phi}}$, the full conditional of $\boldsymbol{\phi}$ is

$$\pi(\boldsymbol{\phi}|\boldsymbol{y},\mathcal{M},\boldsymbol{\beta},\sigma^2) \propto \Psi(\boldsymbol{\phi}) \times \mathcal{N}_p(\hat{\boldsymbol{\phi}},\boldsymbol{V})\, I_{S_{\phi}},$$

where

$$\Psi(\boldsymbol{\phi}) = |\boldsymbol{\Sigma}_p|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}_p - \boldsymbol{X}_p\boldsymbol{\beta})\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{Y}_p - \boldsymbol{X}_p\boldsymbol{\beta})\right\},$$

$\hat{\boldsymbol{\phi}} = \boldsymbol{V}(\boldsymbol{G}_0^{-1}\boldsymbol{\phi}_0 + \sum_{t=p+1}^{n}\boldsymbol{E}_t\boldsymbol{e}_t)$, $\boldsymbol{V} = (\boldsymbol{G}_0^{-1} + \sigma^{-2}\sum_{t=p+1}^{n}\boldsymbol{E}_t\boldsymbol{E}_t')^{-1}$, $\boldsymbol{E}_t = (e_{t-1},\ldots,e_{t-p})'$. To sample this density the proposal density is specified as

$$q(\boldsymbol{\phi}|\boldsymbol{y},\mathcal{M},\boldsymbol{\beta},\sigma^2) = \mathcal{N}_p(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}},\boldsymbol{V}).$$

With this tailored proposal density the probability of move just involves $\Psi(\boldsymbol{\phi})$, leading to a M–H step that is both fast (because it entails the calculation of a function based on the first p observations and not the entire sample) and highly efficient (because the proposal density is matched to the target).

Algorithm 12: Regression with autoregressive errors

(1) Calculate $(y_t^*,\boldsymbol{x}_t^*)$, $t \leqslant n$.

(2) Sample

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t^{*\prime}y_t^*), \boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t^*\boldsymbol{x}_t^{*\prime})^{-1}\right).$$

(3) Sample

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_t^* - \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta})^2}{2}\right\}.$$

(4) M-H

(a) Calculate

$$\hat{\phi} = \boldsymbol{V}(\boldsymbol{G}_0^{-1}\boldsymbol{\phi}_0 + \sigma^{-2}\sum_{t=p+1}^{n}\boldsymbol{E}_t'e_t); \quad \boldsymbol{V} = (\boldsymbol{G}_0^{-1} + \sigma^{-2}\sum_{t=p+1}^{n}\boldsymbol{E}_t\boldsymbol{E}_t')^{-1}.$$

```
    (b) Propose
            $\phi' \sim \mathcal{N}_p(\hat{\phi}, V)$.
    (c) Calculate
            $\alpha = \min\left\{1, \frac{\Psi(\phi') I_{S_{\phi'}}}{\Psi(\phi)}\right\}$.
    (d) Move to $\phi'$ with probability $\alpha$.
(5) Goto 1.
```

8.8. Hidden Markov models

In this subsection we consider the MCMC-based analysis of hidden Markov models (or Markov mixture models or Markov switching models). The general model is described as

$$\begin{aligned} y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, s_t = k, \boldsymbol{\theta} &\sim f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_k), \quad k = 1, \ldots, m, \\ s_t|s_{t-1}, \boldsymbol{P} &\sim \text{Markov}(\boldsymbol{P}, \pi_1), \\ \boldsymbol{\theta} &\sim \pi, \\ \boldsymbol{p}_i &\sim \text{Dirichlet}(\alpha_{i1}, \ldots, \alpha_{im}), \quad i \leqslant m, \end{aligned}$$

where $s_t \in \{1, \ldots, m\}$ is an *unobservable* random variable which evolves according to a Markov process with transition matrix $\boldsymbol{P} = \{p_{ij}\}$, with $p_{ij} = \Pr(s_t = j|s_{t-1} = i)$, and initial distribution π_1 at $t = 1$, f is a density or mass function, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ are the parameters of f under each possible value of s_t, and $\boldsymbol{p}_i$ is the ith row of $\boldsymbol{P}$ that is assumed to have a Dirichlet prior distribution with parameters $(\alpha_{i1}, \ldots, \alpha_{im})$. For identifiability reasons, the Markov chain of s_t is assumed to be time-homogeneous, irreducible, and aperiodic.

The MCMC analysis of such models was initiated by Albert and Chib (1993b) in the context of a more general model than the one above where the conditional density of the data depends not just on s_t but also on the previous values $\{s_{t-1}, \ldots, s_{t-r}\}$, as in the model of Hamilton (1989). The approach relies on augmenting the parameter space to include the unobserved states and simulating $\pi(\boldsymbol{S}_n, \boldsymbol{\theta}, \boldsymbol{P}|\boldsymbol{y}, \mathcal{M})$ via the conditional distributions

$$s_t|\boldsymbol{y}, \mathcal{M}, \boldsymbol{S}_{(-t)}, \boldsymbol{\theta}, \boldsymbol{P} (t \leqslant n); \quad \boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{S}_n, \boldsymbol{P}; \{\boldsymbol{p}_i\}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{S}_n,$$

where $\boldsymbol{S}_n = (s_1, \ldots, s_n)$ denotes the entire collection of states. Robert, Celeux and Diebolt (1993) and McCulloch and Tsay (1994) developed a similar approach for the simpler model in which only the current state s_t appears in the density of y_t while Billio, Monfort and Robert (1999) consider ARMA models with Markov switching.

Chib (1996), whose approach we now follow, modifies the first set of blocks of the above scheme to sample the states jointly from

$$\boldsymbol{S}_n | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P},$$

in one block. This leads to a more efficient MCMC algorithm. The sampling of $\boldsymbol{S}_n$ is achieved by *one* forward and backward pass through the data. In the forward pass, one recursively produces the sequence of mass functions $\{p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})\}$ $(t \leqslant n)$ as follows: assume that the function $p(s_{t-1}|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ is available. Then, one obtains $p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ by calculating

$$p(s_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) = \sum_{l=1}^{m} p(s_t|\, s_{t-1} = l, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t-1} = l|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}),$$

followed by

$$p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) = \frac{p(s_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_{s_t}, \boldsymbol{P})}{\sum_{l=1}^{m} p(s_t = l|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_l, \boldsymbol{P})}.$$

These forward recursions can be initialized at $t = 1$ by setting $p(s_1|\boldsymbol{Y}_0, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ to be the stationary distribution of the chain (the left eigenvector corresponding to the eigenvalue of one).

Then, in the backward pass one simulates $\boldsymbol{S}_n$ by the method of composition, first simulating s_n from $s_n|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}$ and then the s_t's using the probability mass functions

$$p(s_t = k|\boldsymbol{y}, \mathcal{M}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P}) = \frac{p(s_t = k|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t+1}|s_t = k, \boldsymbol{P})}{\sum_{l=1}^{m} p(s_t = l|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t+1}|s_t = l, \boldsymbol{P})},$$
$$k \leqslant m, \quad t \leqslant n-1,$$

where $\boldsymbol{S}^{t+1} = (s_{t+1}, \ldots, s_n)$ consists of the simulated values from earlier steps and the second term of the numerator is the Markov transition probability, which is picked off from the column of $\boldsymbol{P}$ determined by the simulated value of s_{t+1}.

Given the simulated vector $\boldsymbol{S}_n$, the data separates into m non-contiguous pieces and the simulation of $\boldsymbol{\theta}_k$ is from the full conditional distribution

$$\pi(\boldsymbol{\theta}_k) \prod_{t:s_t=k} f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}).$$

Depending on the form of f and p this may belong to a named distribution. Otherwise, this distribution is sampled by a M–H step. Finally, the last distribution depends simply on $\boldsymbol{S}_n$ with each row $\boldsymbol{p}_i$ of $\boldsymbol{P}$ independently an updated Dirichlet distribution:

$$\boldsymbol{p}_i|\boldsymbol{S}_n \sim \mathcal{D}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{i1} + n_{im}), \quad (i \leqslant m),$$

where n_{ik} is the total number of *one-step* transitions from state i to state k in the vector $\boldsymbol{S}_n$.

Algorithm 13: Hidden Markov model

```
(1) Calculate and store for $t = 1, 2, \ldots, n$
        $p(s_t | \boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(2) Sample
        $s_n \sim p(s_n | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(3) Sample for $t = n-1, n-2, \ldots, 1$
        $s_t \sim p(s_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P})$.
(4) Sample for $k = 1, \ldots, m$
        $\boldsymbol{\theta}_k \propto \pi(\boldsymbol{\theta}_k) \prod_{t : s_t = k} f(y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(5) Sample for $i = 1, 2, \ldots, m$
        $\boldsymbol{p}_i \sim \mathcal{D}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{i1} + n_{im})$.
(6) Goto 1.
```

8.9. State space models

Consider next a linear state space model in which a scalar observation y_t is generated as

$$\begin{aligned} y_t | \mathcal{M}, \boldsymbol{\theta}_t &\sim \mathcal{N}(\boldsymbol{x}_t' \boldsymbol{\theta}_t, \sigma^2), \\ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1} &\sim \mathcal{N}_m(\boldsymbol{G}\boldsymbol{\theta}_{t-1}, \boldsymbol{\Psi}), \quad 1 \leqslant t \leqslant n, \\ \sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2}, \frac{\delta_0}{2}\right), \\ \boldsymbol{\Psi}^{-1} &\sim \mathcal{W}_m(\rho_0, \boldsymbol{R}_0), \end{aligned}$$

where $\boldsymbol{\theta}_t$ is an $m \times 1$ state vector and $\boldsymbol{G}$ is assumed known. For nonlinear versions of this model, a MCMC fitting approach is provided by Carlin, Polson and Stoffer (1992). It is based on the inclusion of the variables $\{\boldsymbol{\theta}_t\}$ in the sampling followed by one-at-a-time sampling of $\boldsymbol{\theta}_t$ given $\boldsymbol{\theta}_{-t}$ (the remaining $\boldsymbol{\theta}_t$'s) and $(\sigma^2, \boldsymbol{\Psi})$. For the linear version presented above, Carter and Kohn (1994) and Fruhwirth-Schnatter (1994) show that a reduced blocking scheme involving the joint simulation of $\{\boldsymbol{\theta}_t\}$ is possible and desirable, because the $\boldsymbol{\theta}_t$'s are correlated by construction, while de Jong and Shephard (1995) provide an important alternative procedure called the simulation smoother that is particularly useful if $\boldsymbol{\Psi}$ is not positive definite or if the dimension m of the state vector is large. Carter and Kohn (1996) and Shephard (1994) also consider models, called conditionally Gaussian state space models, that have Gaussian observation densities conditioned on a discrete or continuous variable s_t. An example of this is provided below in Section 8.10. Chib and Greenberg (1995b) consider hierarchical and vector versions of the above model while additional issues related to the fitting and parameterization of state space models are considered by Pitt and Shephard (1997).

The MCMC implementation for this model is based on the distributions

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}; \quad \sigma^2 | \boldsymbol{y}, \mathcal{M}, \{\boldsymbol{\theta}_t\}, \boldsymbol{\Psi}; \quad \boldsymbol{\Psi}^{-1} | \boldsymbol{y}, \mathcal{M}, \{\boldsymbol{\theta}_t\}.$$

To see how the $\boldsymbol{\theta}_t$'s are sampled, write the joint distribution as

$$p(\boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}) \times p(\boldsymbol{\theta}_{n-1} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}_n, \sigma^2, \boldsymbol{\Psi}) \\ \times \cdots \times p(\boldsymbol{\theta}_1 | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n, \sigma^2, \boldsymbol{\Psi}),$$

where, on letting $\boldsymbol{\theta}^s = (\boldsymbol{\theta}_s, \ldots, \boldsymbol{\theta}_n)$, $\boldsymbol{Y}_s = (y_1, \ldots, y_s)$ and $\boldsymbol{Y}^s = (y_s, \ldots, y_n)$ for $s \leqslant n$, the typical term is

$$p(\boldsymbol{\theta}_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}^{t+1}, \sigma^2, \boldsymbol{\Psi}) \propto p(\boldsymbol{\theta}_t | \boldsymbol{Y}_t, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}) p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}),$$

due to the fact that $(\boldsymbol{Y}^{t+1}, \boldsymbol{\theta}^{t+1})$ is independent of $\boldsymbol{\theta}_t$ given $(\boldsymbol{\theta}_{t+1}, \sigma^2, \boldsymbol{\Psi})$. The first density on the right hand side is Gaussian with moments given by the Kalman filter recursions. The second density is Gaussian with moments $\boldsymbol{G}\boldsymbol{\theta}_t$ and $\boldsymbol{\Psi}$. By completing the square in $\boldsymbol{\theta}_t$ the moments of $p(\boldsymbol{\theta}_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}^{t+1}, \sigma^2, \boldsymbol{\Psi})$ can be derived. Then, the joint distribution $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}$ can be sampled by the method of composition.

Algorithm 14: Gaussian state space

```
(1) Kalman filter
    (a) Calculate for $t = 1, 2, \ldots, n$

            $\hat{\boldsymbol{\theta}}_{t|t-1} = \boldsymbol{G}\hat{\boldsymbol{\theta}}_{t-1|t-1}$,    $\boldsymbol{R}_{t|t-1} = \boldsymbol{G}\boldsymbol{R}_{t-1|t-1}\boldsymbol{G}' + \boldsymbol{\Psi}$,
            $f_{t|t-1} = \boldsymbol{x}_t'\boldsymbol{R}_{t|t-1}\boldsymbol{x}_t + \sigma^2$,    $\boldsymbol{K}_t = \boldsymbol{R}_{t|t-1}\boldsymbol{x}_t f_{t|t-1}^{-1}$,
            $\hat{\boldsymbol{\theta}}_{t|t} = \hat{\boldsymbol{\theta}}_{t|t-1} + \boldsymbol{K}_t(y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\theta}}_{t|t-1})$,    $\boldsymbol{R}_{t|t} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{x}_t')\boldsymbol{R}_{t|t-1}$,
            $\boldsymbol{M}_t = \boldsymbol{R}_{t|t}\boldsymbol{G}'\boldsymbol{R}_{t+1|t}^{-1}$,    $\boldsymbol{R}_t = \boldsymbol{R}_{t|t} - \boldsymbol{M}_t\boldsymbol{R}_{t+1|t}\boldsymbol{M}_t'$.

    (b) Store

            $\hat{\boldsymbol{\theta}}_{t|t}; \boldsymbol{M}_t; \boldsymbol{R}_t$.

(2) Simulation step
    (a) Sample

            $\boldsymbol{\theta}_n \sim \mathcal{N}_m(\hat{\boldsymbol{\theta}}_{n|n}, \boldsymbol{R}_{n|n})$.
```

(b) Sample for $t = n-1, n-2, \ldots, 1$

$$\boldsymbol{\theta}_t \sim \mathcal{N}_m(\hat{\boldsymbol{\theta}}_t, \boldsymbol{R}_t), \quad \hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t\,|\,t} + \boldsymbol{M}_t \left(\boldsymbol{\theta}_{t+1} - \boldsymbol{G}\hat{\boldsymbol{\theta}}_{t\,|\,t}\right).$$

(3) Sample

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_t - \boldsymbol{x}_i'\boldsymbol{\theta}_t)^2}{2}\right\}.$$

(4) Sample

$$\boldsymbol{\Psi}^{-1} \sim \mathcal{W}_m\left[\rho_0 + n, \left\{\boldsymbol{R}_0^{-1} + \sum_{t=1}^{n}(\boldsymbol{\theta}_t - \boldsymbol{G}\boldsymbol{\theta}_{t-1})(\boldsymbol{\theta}_t - \boldsymbol{G}\boldsymbol{\theta}_{t-1})'\right\}^{-1}\right].$$

(5) Goto 1.

8.10. Stochastic volatility model

Suppose that time series observations $\{y_t\}$ are generated by the stochastic volatility (SV) model [see, for example, Taylor (1994), Shephard (1996), and Ghysels, Harvey and Renault (1996)]

$$y_t = \exp(h_t/2)u_t, \quad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t, \quad t \leqslant n,$$

where $\{h_t\}$ is the latent log-volatility of y_t and $\{u_t\}$ and $\{\eta_t\}$ are white noise standard normal random variables. This is an example of a state space model in which the state variable h_t appears non-linearly in the observation equation. The model can be extended to include covariates in the observation and evolution equations and to include a heavy-tailed, non-Gaussian distribution for u_t. The MCMC analysis of this model was initiated by Jacquier, Polson and Rossi (1994) based on the general approach of Carlin, Polson and Stoffer (1992). If we let $\boldsymbol{\theta} = (\phi, \mu, \sigma^2)$, then the algorithm of Jacquier, Polson and Rossi (1994) is based on the $(n+3)$ full conditional distributions

$$h_t|\boldsymbol{y}, \mathcal{M}, h_{-t}, \boldsymbol{\theta}, \quad t = 1, 2, \ldots, n,$$
$$\phi|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \mu, \sigma^2; \quad \mu|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \phi, \sigma^2; \; \sigma^2|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \phi, \mu,$$

where the latent variables h_t are sampled by a sequence of Metropolis–Hastings steps. Subsequently, Kim, Shephard and Chib (1998) discussed an alternative approach that leads to considerable improvements in the mixing of the Markov chain. The latter approach has been further refined by Chib, Nardari and Shephard (1998, 1999).

The idea behind the Kim, Shepard and Chib approach is to approximate the SV model by a conditionally Gaussian state space model with the introduction of

Table 1
Parameters of seven-component Gaussian mixture to approximate the distribution of $\log \chi_1^2$

s_t	q	m_{s_t}	$\upsilon_{s_t}^2$
1	0.00730	−11.40039	5.79596
2	0.10556	−5.24321	2.61369
3	0.00002	−9.83726	5.17950
4	0.04395	1.50746	0.16735
5	0.34001	−0.65098	0.64009
6	0.24566	0.52478	0.34023
7	0.25750	−2.35859	1.26261

multinomial random variables $\{s_t\}$ that follow a seven-point discrete distribution. Conditioned on $\{s_t\}$, the model is Gaussian and the variables h_t appear linearly in the observation equation. Then, the entire set of $\{h_t\}$ are sampled jointly conditioned on $\boldsymbol{\theta}$ and $\{s_t\}$ by either the simulation smoother of de Jong and Shephard (1995) or by the algorithm for simulating states given in Algorithm 14. Once the MCMC simulation is concluded the parameter draws are reweighted to correspond to the original non-linear model.

To begin with, reexpress the SV model as

$$y_t^* = h_t + z_t, \quad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t,$$

where $y_t^* = \ln(y_t^2)$ and $z_t = \log \varepsilon_t^2$ is distributed as a log of chi-squared random variable with one degrees of freedom. Now approximate the distribution of $y_t^*|h_t$ by a mixture of normal distributions. A very accurate representation is given by the mixture distribution

$$y_t^*|h_t, s_t \sim \mathcal{N}(m_{s_t} + h_t, \upsilon_{s_t}^2), \quad \Pr(s_t = i) = q_i, \quad i \leqslant 7, \quad t \leqslant n,$$

where $s_t \in (1, 2, \ldots, 7)$ is an unobserved component indicator with probability mass function $q = \{q_i\}$ and the parameters $\{q, m_{s_t}, \upsilon_{s_t}^2\}$ are as reported in Table 1. Now the parameters and the latent variables can be simulated by a *two block* MCMC algorithm defined by the distributions

$$(\boldsymbol{\theta}, h_1, \ldots, h_n)|\{y_t^*\}, \{s_t\},$$
$$\{s_t\}|\{y_t^*\}, \{h_t\}, \boldsymbol{\theta}.$$

where the first block is sampled by the method of composition by first drawing $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta}|\{y_t^*\}, \{s_t\})$ by a M–H step followed by a draw of $\{h_t\}$ by the simulation smoother. In the former step the target distribution is

$$\begin{aligned}\pi(\boldsymbol{\theta}|\{y_t^*\}, \{s_t\}) &\propto p(\boldsymbol{\theta}) f(y_1^*, \ldots, y_n^*|\{s_t\}, \boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}) \prod_{t=1}^{n} f(y_t^*|\mathcal{F}_{t-1}^*, \{s_t\}, \boldsymbol{\theta}),\end{aligned}$$

where each one-step ahead density $f(y_t^*|\mathcal{F}_{t-1}^*, \{s_t\}, \boldsymbol{\theta})$ can be derived from the output of the Kalman filter recursions, adapted to the differing components, as indicated by the component vector $\{s_t\}$, and $p(\boldsymbol{\theta})$ is the prior density. For ϕ the prior can be taken to be the scaled beta density

$$p(\phi) = c\,(0.5(1+\phi))^{\phi^{(1)}-1}\,(0.5(1-\phi))^{\phi^{(2)}-1}, \quad \phi^{(1)}, \phi^{(2)} > 0.5, \tag{31}$$

where

$$c = 0.5\frac{\Gamma(\phi^{(1)}+\phi^{(2)})}{\Gamma(\phi^{(1)})\Gamma(\phi^{(2)})},$$

with prior mean of $2\phi^{(1)}/(\phi^{(1)}+\phi^{(2)}-1)$, while those on μ and σ^2 can be normal and inverse gamma densities, respectively.

Algorithm 15: Stochastic volatility

```
(1) Initialize $\{s_t\}$
(2) M-H
    (a) Calculate $\boldsymbol{m} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ where

            $l(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{n} \ln f_{t|t-1} - \frac{1}{2}\sum_{t=1}^{n} \frac{(y_t^* - m_{s_t} - \hat{h}_{t|t-1})^2}{f_{t|t-1}}$

        and $\boldsymbol{V} = \left\{-\partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right\}^{-1}$, the negative inverse of the
        hessian at $\boldsymbol{m}$, where $f_{t|t-1}$ and $\hat{h}_{t|t-1}$ are computed from the
        Kalman filter recursions

            $\hat{h}_{t|t-1} = \mu + \phi(\hat{h}_{t-1|t-1} - \mu)$,        $\boldsymbol{R}_{t|t-1} = \phi^2\boldsymbol{R}_{t-1|t-1} + \sigma^2$,
            $f_{t|t-1} = \boldsymbol{R}_{t|t-1} + \upsilon_{s_t}^2$,                $\boldsymbol{K}_t = \boldsymbol{R}_{t|t-1} f_{t|t-1}^{-1}$,
            $\hat{h}_{t|t} = \hat{h}_{t|t-1} + \boldsymbol{K}_t(y_t^* - m_{s_t} - \hat{h}_{t|t-1})$, $\boldsymbol{R}_{t|t} = (1 - \boldsymbol{K}_t)\boldsymbol{R}_{t|t-1}$.

    (b) Propose

            $\boldsymbol{\theta}' \sim f_T(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{V}, \xi)$.

    (c) Calculate

            $\alpha = \min\left\{\frac{p(\boldsymbol{\theta}')\,l(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})\,l(\boldsymbol{\theta})}\frac{f_T(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{V}, \xi)}{f_T(\boldsymbol{\theta}'|\boldsymbol{m}, \boldsymbol{V}, \xi)}, 1\right\}$.

    (d) Move to $\boldsymbol{\theta}'$ with probability $\alpha$.
(3) Sample $\{h_t\}$ using algorithm 13, or the simulation smoother
    algorithm, modified to include the components of the
    mixture selected by $\{s_t\}$.
```

(4) `Sample`

$$s_t \sim \Pr(s_t | y_t^*, h_t, \boldsymbol{\psi}) \propto \Pr(s_t) f_N(y_t^* | \mu_{s_t} + h_t, \upsilon_{s_t}^2).$$

(5) `Goto 2`.

8.11. Gaussian panel data models

For continuous clustered or panel data a common model formulation is that of Laird and Ware (1982)

$$\boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma^2 \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{b}_i, \sigma^2 \boldsymbol{I}_{n_i}), \quad \boldsymbol{b}_i | \boldsymbol{D} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D}),$$

$$\boldsymbol{D}^{-1} \sim \mathcal{W}_p(\rho_0, \boldsymbol{R}_0), \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0), \quad \sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

where $\boldsymbol{y}_i$ is a n_i vector of observations and the matrix $\boldsymbol{W}_i$ is a subset of $\boldsymbol{X}_i$. If $\boldsymbol{W}_i$ is a vector of units, then the model reduces to a panel model with intercept heterogeneity. If $\boldsymbol{W}_i = \boldsymbol{X}_i$, then the model becomes the random coefficient panel model.

Zeger and Karim (1991) and Wakefield et al. (1994) propose a Gibbs MCMC approach for this model that is based on including $\{\boldsymbol{b}_i\}$ in the sampling in conjunction with full blocking. This blocking scheme is not very desirable because the random effects and the fixed effects $\boldsymbol{\beta}$ tend to be highly correlated and treating them as separate blocks creates problems with mixing Gelfand, Sahu and Carlin (1995). To deal with this problem, Chib and Carlin (1999) suggest a number of reduced blocking schemes. One of the simplest proceeds by sampling $\boldsymbol{\beta}$ and $\{\boldsymbol{b}_i\}$ in one block by the method of composition: first sampling $\boldsymbol{\beta}$ marginalized over $\{\boldsymbol{b}_i\}$ and then sampling $\{\boldsymbol{b}_i\}$ conditioned on $\boldsymbol{\beta}$. What makes reduced blocking possible is the fact that the distribution of $\boldsymbol{y}_i$ marginalized over $\boldsymbol{b}_i$ is also Gaussian:

$$\boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad \boldsymbol{V}_i = \sigma^2 \boldsymbol{I}_{n_i} + \boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'.$$

The updated distribution of $\boldsymbol{\beta}$, marginalized over $\{\boldsymbol{b}_i\}$ is, therefore, easy to derive. The rest of the algorithm follows the steps of Wakefield et al. (1994). In particular, the sampling of the random effects is from independent normal distributions that are derived by treating $(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ as the "data," $\boldsymbol{b}_i$ as the regression coefficient and $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$ as the prior. The sampling of $\boldsymbol{D}^{-1}$ is from an Wishart distribution and that of σ^2 from an inverse gamma distribution.

Algorithm 16: Gaussian Panel

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{V}_i^{-1}\boldsymbol{y}_i), \boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\right).$$

(2) `Sample`

$$\boldsymbol{b}_i \sim \mathcal{N}_q\left(\boldsymbol{D}_i \boldsymbol{W}_i' \sigma^{-2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{D}_i = (\boldsymbol{D} + \sigma^{-2}\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right), \quad i \leqslant n.$$

(3) `Sample`

$$\boldsymbol{D}^{-1} \sim \mathcal{W}_p\left\{\rho_0 + n, \left(\boldsymbol{R}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{b}_i \boldsymbol{b}_i'\right)^{-1}\right\}.$$

(4) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \sum_{i=1}^{n} \| \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{W}_i\boldsymbol{b}_i \|^2}{2}\right).$$

(5) `Goto 1`.

8.12. *Multivariate binary data models*

To model correlated binary data a canonical model is the multivariate probit (MVP). Let y_{ij} denote the binary response on the ith observation unit and jth variable, and let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$, $1 \leqslant i \leqslant n$, denote the collection of responses on all J variables. Then, under the MVP model the marginal probability of $y_{ij} = 1$ is

$$\Pr(y_{ij} = 1 | \mathcal{M}, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j),$$

and the joint probability that $\boldsymbol{Y}_i = \boldsymbol{y}_i$ conditioned on the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \equiv \Pr(\boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{A_{iJ}} \cdots \int_{A_{i1}} \phi_J(\boldsymbol{t} | \boldsymbol{0}, \boldsymbol{\Sigma})\, \mathrm{dt},$$

where as in the SUR model, $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J') \in \Re^k$, $k = \sum k_j$, but unlike the SUR model, the $J-$ matrix $\boldsymbol{\Sigma} = \{\sigma_{jk}\}$ is in correlation form (with units on the diagonal), and A_{ij} is the interval

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x}_{ij}'\boldsymbol{\beta}_j) & \text{if } y_{ij} = 1, \\ \left[\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j, \infty\right) & \text{if } y_{ij} = 0. \end{cases}$$

To simplify the MCMC implementation for this model Chib and Greenberg (1998) follow the general approach of Albert and Chib (1993a) and employ latent variables. Let

$$\boldsymbol{z}_i \sim \mathcal{N}_J(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

with the observed data given by the sign of z_{ij}:

$$y_{ij} = I(z_{ij} > 0), \quad j = 1, \ldots, J,$$

where $I(A)$ is the indicator function of the event A. If we let $\boldsymbol{\sigma} = (\sigma_{21}, \sigma_{31}, \sigma_{32}, \ldots, \sigma_{JJ})$ denote the $J(J-1)/2$ distinct elements of $\boldsymbol{\Sigma}$, and let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ denote the latent values corresponding to the observed data $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{n}$, then the algorithm proceeds with the sampling of the augmented posterior density

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{z} | \boldsymbol{y}, \mathcal{M}) &\propto p(\boldsymbol{\beta}) p(\boldsymbol{\sigma}) f(\boldsymbol{z} | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \Pr(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &\propto p(\boldsymbol{\beta}) p(\boldsymbol{\sigma}) \prod_{i=1}^{n} \left\{ \phi_J(\boldsymbol{z}_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}) \Pr(\boldsymbol{y}_i | \boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \right\}, \boldsymbol{\beta} \in \Re^k, \boldsymbol{\sigma} \in C,\end{aligned}$$

where

$$\Pr(\boldsymbol{y}_i | \boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J} \left\{ I(z_{ij} > 0) I(y_{ij} = 1) + I(z_{ij} \leqslant 0) I(y_{ij} = 0) \right\},$$

$p(\boldsymbol{\sigma})$ is a normal density truncated to the region C, and C is the set of values of $\boldsymbol{\sigma}$ that produce a positive definite correlation matrix $\boldsymbol{\Sigma}$.

Conditioned on $\{\boldsymbol{z}_i\}$ and $\boldsymbol{\Sigma}$, the update for $\boldsymbol{\beta}$ is as in the SUR model, while conditioned on $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, z_{ij} can be sampled one at a time conditioned on the other latent values from truncated normal distributions, where the region of truncation is either $(0, \infty)$ or $(-\infty, 0)$ depending on whether the corresponding y_{ij} is one or zero. The key step in the algorithm is the sampling of $\boldsymbol{\sigma}$, the unrestricted elements of $\boldsymbol{\Sigma}$, from the full conditional density $\pi(\boldsymbol{\sigma} | \mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta}) \propto p(\boldsymbol{\sigma}) \prod_{i=1}^{n} \phi_J(\boldsymbol{z}_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma})$. This density, which is truncated to the complicated region C, is sampled by a M–H step with tailored proposal density $q(\boldsymbol{\sigma} | \mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta}) = f_T(\boldsymbol{\sigma} | \boldsymbol{m}, \boldsymbol{V}, \xi)$ where

$$\begin{aligned}\boldsymbol{m} &= \arg\max_{\boldsymbol{\sigma} \in C} \sum_{i=1}^{n} \ln \phi_J(\boldsymbol{z}_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}), \\ \boldsymbol{V} &= -\left\{ \frac{\partial^2 \sum_{i=1}^{n} \ln \phi_J(\boldsymbol{z}_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'} \right\}_{\boldsymbol{\sigma} = \boldsymbol{m}}^{-1},\end{aligned}$$

are the mode and curvature of the target distribution, given the current values of the conditioning variables. Note that, as in Algorithm 12, no truncation is enforced on the proposal density.

Algorithm 17: Multivariate probit

```
(1) Sample for $i \leqslant n, j \leqslant J$
        $z_{ij} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{ij}, \upsilon_{ij}) & \text{if } y_{ij} = 1, \\ \mathcal{TN}_{(-\infty,0])}(\mu_{ij}, \upsilon_{ij}) & \text{if } y_{ij} = 0, \end{cases}$
        $\mu_{ij} = \mathrm{E}(z_{ij} | Z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}),$
        $\upsilon_{ij} = \mathrm{Var}(z_{ij} | Z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma}).$
```

```
(2) Sample

        $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i\right), \boldsymbol{B}_n=\left(\boldsymbol{B}_0^{-1}+\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i^{-1}\right)^{-1}\right).$

(3) M-H
    (a) Calculate the parameters $(\boldsymbol{m}, \boldsymbol{V})$.
    (b) Propose

            $\boldsymbol{\sigma}' \sim f_T(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi).$

    (c) Calculate

            $\alpha = \min\left\{\frac{p(\boldsymbol{\sigma}')\prod_{i=1}^{n}\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}')\, I[\boldsymbol{\sigma}' \in C]}{p(\boldsymbol{\sigma})\prod_{i=1}^{n}\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})}\frac{f_T(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)}{f_T(\boldsymbol{\sigma}'|\boldsymbol{m}, \boldsymbol{V}, \xi)}, 1\right\}.$

    (d) Move to $\boldsymbol{\sigma}'$ with probability $\alpha$.
(4) Goto 1.
```

As an application of this algorithm consider a data set in which the multivariate binary responses are generated by a panel strucure. The data is concerned with the health effects of pollution on 537 children in Stuebenville, Ohio, each observed at ages 7, 8, 9 and 10 years, and the response variable is an indicator of wheezing status [Diggle, Liang and Zeger (1995)]. Suppose that the marginal probability of wheeze status of the ith child at the jth time point is specified as

$$\Pr(y_{ij}=1|\boldsymbol{\beta}) = \Phi(\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\beta_3 x_{3ij}), \quad i \leqslant 537, j \leqslant 4,$$

where $\boldsymbol{\beta}$ is constant across categories, x_1 is the age of the child centered at nine years, x_2 is a binary indicator variable representing the mother's smoking habit during the first year of the study, and $x_3 = x_1 x_2$. Suppose that the Gaussian prior on $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ is centered at zero with a variance of $10\boldsymbol{I}_k$ and let $p(\boldsymbol{\sigma})$ be the density of a normal distribution, with mean zero and variance $\boldsymbol{I}_6$, *restricted* to region that leads to a positive-definite correlation matrix, where $(\sigma_{21}, \sigma_{31}, \sigma_{32}, \sigma_{41}, \sigma_{42}, \sigma_{43})$. From 10 000 cycles of Algorithm 17 one obtains the following covariate effects and posterior distributions of the correlations.

Notice that the summary tabular output in Table 2 contains not only the posterior means and standard deviations of the parameters but also the 95% credibility intervals, all computed from the sampled draws. It may be seen from Figure 6 that the posterior distributions of the correlations are similar suggesting that an equicorrelated correlation structure might be appropriate for these data. This issue is considered more formally in Section 10.2 below.

Table 2
Covariate effects in the Ohio wheeze data: MVP model with unrestricted correlations[1]

β	Prior		Posterior[2]				
	Mean	Std. dev.	Mean	NSE	Std. dev.	Lower	Upper
$\boldsymbol{\beta}_1$	0.000	3.162	–1.108	0.001	0.062	–1.231	–0.985
$\boldsymbol{\beta}_2$	0.000	3.162	–0.077	0.001	0.030	–0.136	–0.017
$\boldsymbol{\beta}_3$	0.000	3.162	0.155	0.002	0.101	–0.043	0.352
$\boldsymbol{\beta}_4$	0.000	3.162	0.036	0.001	0.049	–0.058	0.131

[1] The results are based on 10 000 draws from Algorithm 17.

[2] NSE denotes the numerical standard error, lower is the 2.5th percentile and upper is the 97.5th percentile of the simulated draws.

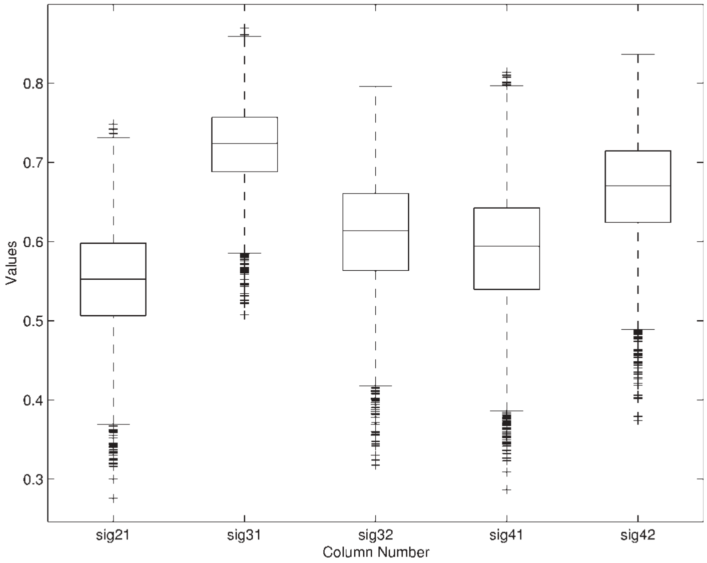

Fig. 6. Posterior boxplots of the correlations in the Ohio wheeze data: MVP model.

9. Sampling the predictive density

A fundamental goal of any statistical analysis is to predict a set of future or unobserved observations $\boldsymbol{y}_f$ given the current data $\boldsymbol{y}$ and the assumed model $\mathcal{M}$. In the Bayesian context this problem is solved by the calculation of the Bayesian prediction density which is defined as the distribution of $\boldsymbol{y}_f$ conditioned on $(\boldsymbol{y}, \mathcal{M})$ but marginalized over the parameters $\boldsymbol{\theta}$. More formally, the predictive density is defined as

$$f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}) = \int f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M})\, \mathrm{d}\boldsymbol{\theta}, \tag{32}$$

where $f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta})$ is the conditional density of $\boldsymbol{y}_f$ given $(\boldsymbol{y},\mathcal{M},\boldsymbol{\theta})$ and the marginalization is with respect to the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$ of $\boldsymbol{\theta}$. In general, the predictive density is not available in closed form. However, in the context of MCMC problems that deliver a sample of (correlated) draws

$$\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(M)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M}),$$

this is hardly a problem. One can utilize the posterior draws in conjunction with the method of composition to produce a sample of draws from the predictive density. This is done by appending a step at the end of the MCMC iterations where for each value $\boldsymbol{\theta}^{(j)}$ one simulates

$$\boldsymbol{y}_f^{(j)} \sim f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta}^{(j)}),\quad j \leqslant M, \tag{33}$$

from the density of the observations, conditioned on $\boldsymbol{\theta}^{(j)}$. The collection of simulated values $\{\boldsymbol{y}_f^{(1)},\ldots,\boldsymbol{y}_f^{(M)}\}$ is a sample from the Bayes prediction density $f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M})$. The simulated sample can be summarized in the usual way by the computation of sample averages and quantiles. Thus, to sample the prediction density one simply has to simulate the data generating process for each simulated value of the parameters.

In some problems, that have a latent data structure, a modified procedure to sample the predictive density may be necessary. Suppose that $\boldsymbol{z}_f$ denotes the latent data in the prediction period and $\boldsymbol{z}$ denote the latent data in the sample period. Let $\boldsymbol{\psi} = (\boldsymbol{\theta},\boldsymbol{z})$ and suppose that the MCMC sampler produces the draws

$$\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(M)} \sim \pi(\boldsymbol{\psi}|\boldsymbol{y},\mathcal{M}).$$

In this situation, the predictive density can be expressed as

$$f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M}) = \int f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{z}_f,\boldsymbol{\psi})\,\pi(\boldsymbol{z}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi})\,\pi(\boldsymbol{\psi}|\boldsymbol{y},\mathcal{M})\,\mathrm{d}\boldsymbol{z}_f\,\mathrm{d}\boldsymbol{\psi}, \tag{34}$$

which may again be sampled by the method of composition where for each value $\boldsymbol{\psi}^{(j)} \sim \pi(\boldsymbol{\psi}|\boldsymbol{y},\mathcal{M})$ one simulates

$$\boldsymbol{z}_f^{(j)} \sim \pi(\boldsymbol{z}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^{(j)}),\quad \boldsymbol{y}_f^{(j)} \sim f(\boldsymbol{y}_f|\boldsymbol{y},\mathcal{M},\boldsymbol{z}_f^{(j)},\boldsymbol{\psi}^{(j)}).$$

The simulated values of $\boldsymbol{y}_f$ from this two step process are again from the predictive density.

To illustrate the one step procedure, suppose that one is interested in predicting $\boldsymbol{y}_f = (y_{n+1}, y_{n+2})$ from a regression model with autoregressive errors of order two where

$$y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\beta}, \phi, \sigma^2 \sim \mathcal{N}(\phi_1 y_{t-1} + \phi_2 y_{t-2} + (\boldsymbol{x}_t - \phi_1 \boldsymbol{x}_{t-1} - \phi_2 \boldsymbol{x}_{t-2})' \boldsymbol{\beta}, \sigma^2).$$

Then, for each draw $(\boldsymbol{\beta}^{(j)}, \phi^{(j)}, \sigma^{2(j)})$ from Algorithm 12, one simulates $\boldsymbol{y}_f$ by sampling

$$y_{n+1}^{(j)} \sim \mathcal{N}(\phi_1^{(j)} y_n + \phi_2^{(j)} y_{n-1} + (\boldsymbol{x}_{n+1} - \phi_1^{(j)} \boldsymbol{x}_n - \phi_2^{(j)} \boldsymbol{x}_{n-1})' \boldsymbol{\beta}^{(j)}, \sigma^{2(j)})$$

and

$$y_{n+2}^{(j)} \sim \mathcal{N}(\phi_1^{(j)} y_{n+1}^{(j)} + \phi_2^{(j)} y_n + (\boldsymbol{x}_{n+2} - \phi_1^{(j)} \boldsymbol{x}_{n+1} - \phi_2^{(j)} \boldsymbol{x}_n)' \boldsymbol{\beta}^{(j)}, \sigma^{2(j)}).$$

The sample of simulated values $\{y_{n+1}^{(j)}, y_{n+2}^{(j)}\}$ from repeating this process is a sample from the (joint) predictive density.

As an example of the two step procedure consider a specific hidden Markov model in which

$$y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \beta_0, \gamma, \sigma^2 \sim \mathcal{N}(\beta_0 + \gamma s_t, \sigma^2),$$

where $s_t \in \{0, 1\}$ is a unobserved state variable that follows a two-state Markov process with unknown transition probabilities

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

In this case, suppose that Algorithm 13 has been used to deliver draws on $\boldsymbol{\psi} = (\beta_0, \gamma, \sigma^2, a, b, S_n)$. As described by Albert and Chib (1993b), to predict y_{n+1} we take each draw of $\boldsymbol{\psi}^{(j)}$ and sample

$$s_{n+1}^{(j)} \sim p(s_{n+1} | s_n^{(j)}, p_{11}^{(j)}, p_{22}^{(j)})$$

from the Markov chain (this is just a two point discrete distribution), and then sample

$$y_{n+1}^{(j)} \sim \mathcal{N}(\beta_0^{(j)} + \gamma^{(j)} s_{n+1}^{(j)}, \sigma^{2(j)}).$$

The next value $y_{n+2}^{(j)}$ is drawn in the same way after $s_{n+2}^{(j)}$ is simulated from the Markov chain $p(s_{n+2} | s_{n+1}^{(j)}, p_{11}^{(j)}, p_{22}^{(j)})$. These two steps can be iterated for any number of periods into the future and the whole process repeated for each simulated value of $\boldsymbol{\psi}$.

10. MCMC methods in model choice problems

10.1. Background

Consider the situation in which there are K possible models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ for the observed data defined by the sampling densities $\{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\}$ and proper prior densities $\{p(\boldsymbol{\theta}_k|\mathcal{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1|\boldsymbol{y}), \ldots, \Pr(\mathcal{M}_K|\boldsymbol{y})\}$ conditioned on the data but marginalized over the unknowns $\boldsymbol{\theta}_k$. Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$\begin{aligned}\Pr(\mathcal{M}_k|\boldsymbol{y}) &= \frac{\Pr(\mathcal{M}_k)\, m(\boldsymbol{y}|\mathcal{M}_k)}{\sum_{l=1}^{K} \Pr(\mathcal{M}_l)\, m(\boldsymbol{y}|\mathcal{M}_l)}\\ &\propto \Pr(\mathcal{M}_k)\, m(\boldsymbol{y}|\mathcal{M}_k), \quad (k \leqslant K),\end{aligned}$$

where

$$m(\boldsymbol{y}|\mathcal{M}_k) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k, \tag{35}$$

is the marginal density of the data and is called the marginal likelihood of $\mathcal{M}_k$. In words, the posterior probability of $\mathcal{M}_k$ is proportional to the prior probability of $\mathcal{M}_k$ times the marginal likelihood of $\mathcal{M}_k$. The evidence provided by the data about the models under consideration is summarized by the posterior probability of each model.

Often the posterior probabilities are summarized in terms of the posterior odds

$$\frac{\Pr(\mathcal{M}_i|\boldsymbol{y})}{\Pr(\mathcal{M}_j|\boldsymbol{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \frac{m(\boldsymbol{y}|\mathcal{M}_i)}{m(\boldsymbol{y}|\mathcal{M}_j)},$$

which provides the relative support for the two models. The ratio of marginal likelihoods in this expression is the Bayes factor of $\mathcal{M}_i$ vs $\mathcal{M}_j$.

If interest centers on the prediction of observables then it is possible to mix over the alternative predictive densities by utilizing the posterior probabilities as weights. More formally, the prediction density of a set of observations $\boldsymbol{y}_f$ marginalized over both $\{\boldsymbol{\theta}_k\}$ and $\{\mathcal{M}_k\}$ is given by

$$f(\boldsymbol{y}_f|\boldsymbol{y}) = \sum_{j=1}^{K} \Pr(\mathcal{M}_k|\boldsymbol{y}) f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}_k),$$

where $f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}_k)$ is the prediction density in Equation (34).

10.2. Marginal likelihood computation

A central problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of $\boldsymbol{\theta}_k$. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density because that estimate would converge to

$$\int f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k,$$

which is not the marginal likelihood. In addition, taking draws from the prior density to do the averaging produces an estimate that is simulation-consistent but highly inefficient because draws from the prior density are not likely to be in high density regions of the sampling density $f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)$. A natural way to correct this problem is by the method of importance sampling. If we let $h(\boldsymbol{\theta}_k|\mathcal{M}_k)$ denote a suitable importance sampling function, then the marginal likelihood can be estimated as

$$\hat{m}_I(\boldsymbol{y}|\mathcal{M}_k) = M^{-1} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)}{h(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)},$$
$$\boldsymbol{\theta}_k^{(j)} \sim h(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k) \quad (j \leqslant M).$$

This method is useful when it can be shown that the ratio is bounded, which can be difficult to check in practice, and when the sampling density is not expensive to compute which, unfortunately, is often not true. We mention that if the importance sampling function is taken to be the unnormalized posterior density then that leads to

$$\hat{m}_{\mathrm{NR}} = \left[\frac{1}{M} \sum_{j=1}^{M} \left\{ \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)} \right\} \right]^{-1},$$

the harmonic mean of the likelihood values. This estimate, proposed by Newton and Raftery (1994), can be unstable because the inverse likelihood does not have finite variance. Gelfand and Dey (1994) propose a modified stable estimator

$$\hat{m}_{\mathrm{GD}} = \left[\frac{1}{M} \sum_{j=1}^{M} \left\{ \frac{h(\boldsymbol{\theta}^{(j)})}{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)} \right\} \right]^{-1},$$

where $h(\boldsymbol{\theta})$ is a density with tails thinner than the product of the prior and the likelihood. Unfortunately, this estimator is difficult to apply in models with latent or missing data.

The Laplace method for integrals can be used to provide a non-simulation based estimate of the marginal likelihood. Let d_k denote the dimension of $\boldsymbol{\theta}_k$ and let $\hat{\boldsymbol{\theta}}_k$

denote the posterior mode of $\boldsymbol{\theta}_k$, and $\boldsymbol{\Sigma}_k$ the inverse of the negative Hessian of $\ln\{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)\}$ evaluated at $\hat{\boldsymbol{\theta}}_k$. Then the Laplace estimate of marginal likelihood, on the customary log base ten scale, is given by

$$\log \hat{m}_L(\boldsymbol{y}|\mathcal{M}_k) = (d_k/2)\log(2\pi) + (1/2)\log\det(\boldsymbol{\Sigma}_k) + \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) + \log p(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_k).$$

The Laplace estimate has a large sample justification and can be shown to equal the true value upto an error that goes to zero in probability at the rate n^{-1}.

Both the importance method and the Laplace estimate may be considered as the traditional methods for computing the marginal likelihood. More recent methods exploit two additional facts about the marginal likelihood. The first that the marginal likelihood is the normalizing constant of the posterior density and therefore under this view the Bayes factor can be interpreted as the ratio of two normalizing constants. There is a large literature in physics (in a quite different context, however) on precisely the latter problem stemming from Bennett (1976). This literature was adapted in the mid 1990's for statistical problems by Meng and Wong (1996) utilizing the bridge sampling method and by Chen and Shao (1997) based on umbrella sampling. The techniques presented in these papers, although based on the work in physics, contain modifications of the ideas to handle problems such as models with differing dimensions. DiCiccio, Kass, Raftery and Wasserman (1997) present a comparative analysis of the bridge sampling method in relation to other competing methods of computing the marginal likelihood. At this time, however, the bridge sampling method and its refinements have not found significant use in applications perhaps because the methods are quite involved and because simpler methods are available.

Another approach that deals with the estimation of Bayes factors, again in the context of nested models, is due to Verdinelli and Wasserman (1995) and is called the Savage–Dickey density ratio method. Suppose a model is defined by a parameter $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\psi})$ and the first model $\mathcal{M}_1$ is defined by the restriction $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and the second model $\mathcal{M}_2$ by letting $\boldsymbol{\omega}$ be unrestricted. Then, it can be shown that the Bayes factor is given by

$$B_{12} = \frac{\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)}{\pi(\boldsymbol{\omega}_0|\mathcal{M}_2)} E\left\{\frac{p(\boldsymbol{\psi}|\mathcal{M}_1)}{p(\boldsymbol{\psi}|\mathcal{M}_1, \boldsymbol{\omega}_0)}\right\},$$

where the expectation is with respect to $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\omega}_0)$. If $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\psi})$ is available in closed form then $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)$ can be estimated by the Rao–Blackwell method and the second expectation by taking draws from the posterior $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\omega}_0)$, which can be obtained by the method of reduced runs discussed below, and averaging the ratio of prior densities. This method provides a simple approach for nested models but the method is not efficient if the dimensions of the two models are substantially different because then the ordinate $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)$ tends to be small and the simulated values used to average the ratio tend to be in low density regions.

The second fact about marginal likelihoods, highlighted in a paper by Chib (1995), is that the marginal likelihood by virtue of being the normalizing constant of the posterior density can be expressed as

$$m(\boldsymbol{y}|\mathcal{M}_k) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{\pi(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M}_k)}. \tag{36}$$

This expression is an identity in $\boldsymbol{\theta}_k$ because the left hand side is free of $\boldsymbol{\theta}_k$. Chib (1995) refers to it as the basic marginal likelihood identity (BMI). Based on this expression an estimate of the marginal likelihood on the log-scale is given by

$$\log \hat{m}(\boldsymbol{y}|\mathcal{M}_k) = \log f(\boldsymbol{y}|\boldsymbol{\theta}_k^*, \mathcal{M}_k) + \log p(\boldsymbol{\theta}_k^*|\mathcal{M}_k) - \log \hat{\pi}(\boldsymbol{\theta}_k^*|\boldsymbol{y}, \mathcal{M}_k), \tag{37}$$

where $\boldsymbol{\theta}_k^*$ denotes an arbitrarily chosen point and $\hat{\pi}(\boldsymbol{\theta}_k^*|\boldsymbol{y}, \mathcal{M}_k)$ is the estimate of the posterior density at that single point. Two points should be noted. First, this estimate requires only one evaluation of the likelihood function. This is particularly useful in situations where repeated evaluation of the likelihood function is computationally expensive. Second, to increase the computational efficiency, the point $\boldsymbol{\theta}_k^*$ should be taken to be a high density point under the posterior.

To estimate the posterior ordinate one utilizes the MCMC output in conjunction with a marginal/conditional decomposition. To simplify notation, drop the model subscript k and suppose that the parameter vector is blocked into B blocks as $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B$. In addition, let $\boldsymbol{z}$ denote additional variables (latent or missing data) that may be included in the simulation to clarify the structure of the full conditional distributions. Also let $\boldsymbol{\psi}_i = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i)$ and $\boldsymbol{\psi}^i = (\boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_B)$ denote the list of blocks upto i and the set of blocks from i to B, respectively. Now write the posterior ordinate at the point $\boldsymbol{\theta}^*$ by the law of total probability as

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y}, \mathcal{M}) = \pi(\boldsymbol{\theta}_1^*|\boldsymbol{y}, \mathcal{M}) \times \cdots \times \pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*) \times \cdots \times \pi(\boldsymbol{\theta}_B^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{B-1}^*), \tag{38}$$

where the first term in this expression is the marginal density of $\boldsymbol{\theta}_1$ evaluated at $\boldsymbol{\theta}_1^*$, and the typical term is of the form

$$\pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*) = \int \pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1}, \boldsymbol{z})\, \pi(\boldsymbol{\psi}^{i+1}, \boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*)\, \mathrm{d}\boldsymbol{\psi}^{i+1}\, \mathrm{d}\boldsymbol{z}.$$

This may be called a *reduced conditional ordinate.* It is important to bear in mind that in finding the reduced conditional ordinate one must integrate only over $(\boldsymbol{\psi}^{i+1}, \boldsymbol{z})$ and that the integrating measure is conditioned on $\boldsymbol{\psi}_{i-1}^*$.

Assume that the normalizing constants of each full conditional density is known, an assumption that is relaxed below. Then, the first term of Equation (38) can be estimated by the Rao–Blackwell method. To estimate the typical reduced conditional ordinate, Chib (1995) defines a *reduced MCMC* run consisting of the full conditional distributions

$$\begin{aligned}&\big\{\pi(\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1},\boldsymbol{z});\ \cdots;\ \pi(\boldsymbol{\theta}_B|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_{B-1},\boldsymbol{z});\\&\quad\pi(\boldsymbol{z}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^i)\big\},\end{aligned} \tag{39}$$

where the blocks in $\boldsymbol{\psi}_{i-1}$ are set equal to $\boldsymbol{\psi}^*_{i-1}$. By MCMC theory, the draws on $(\boldsymbol{\psi}^{i+1},\boldsymbol{z})$ from this run are from the distribution $\pi(\boldsymbol{\psi}^{i+1},\boldsymbol{z}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1})$ and so the reduced conditional ordinate can be estimated as the average

$$\hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1}) = M^{-1}\sum_{j=1}^{M}\pi(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1,(j)},\boldsymbol{z}^{(j)}),$$

over the simulated values of $\boldsymbol{\psi}^{i+1}$ and $\boldsymbol{z}$ from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (38) can be estimated in the same way though, conveniently, with fewer and fewer distributions appearing in the reduced runs. Given the marginal and reduced conditional ordinates, the Chib estimate of the marginal likelihood on the log scale is defined as

$$\log\hat{m}(\boldsymbol{y}|\mathcal{M}) = \log f(\boldsymbol{y}|\boldsymbol{\theta}^*,\mathcal{M}) + \log p(\boldsymbol{\theta}^*) - \sum_{i=1}^{B}\log\hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1}), \tag{40}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta}^*,\mathcal{M})$ is the density of the data marginalized over the latent data $\boldsymbol{z}$.

It is worth noting that an alternative approach to estimate the posterior ordinate is developed by Ritter and Tanner (1992) in the context of Gibbs MCMC chains with fully known full conditional distributions. If one lets

$$K_G(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \prod_{i=1}^{B}\pi(\boldsymbol{\theta}^*_k|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta}^*_1,\ldots,\boldsymbol{\theta}^*_{k-1},\boldsymbol{\theta}_{k+1},\ldots,\boldsymbol{\theta}_B),$$

denote the Gibbs transition kernel, then by virtue of the fact that the Gibbs chain satisfies the invariance condition $\pi(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \int K_G(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,\pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})\,\mathrm{d}\boldsymbol{\theta}$, one can obtain the posterior ordinate by averaging the transition kernel over draws from the posterior distribution:

$$\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = M^{-1}\sum_{g=1}^{M}K_G(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}).$$

This estimate only requires draws from the full Gibbs run but when $\boldsymbol{\theta}$ is high dimensional and the model contains latent variables, this estimate is less accurate than Chib's posterior density decomposition method.

It should be observed that the above methods of estimating the posterior ordinate require knowledge of the normalizing constants of each full conditional density. What can be done when this condition does not hold? DiCiccio, Kass, Raftery and Wasserman (1997) and Chib and Greenberg (1998) suggest the use of kernel smoothing in this case. Suppose, for example, that the problem occurs in the distribution of the ith block. Then, the draws on $\boldsymbol{\theta}_i$ from the reduced MCMC run in Equation (39) can be smoothed by kernel methods to find the ordinate at $\boldsymbol{\theta}_i^*$. This approach should only be used when the dimension of the recalcitrant block is not large. A more general technique has recently been developed by Chib and Jeliazkov (2001). The first main result of the paper is that if sampling is done in one block by the M–H algorithm then the posterior ordinate can be written as

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \frac{E_1\left\{\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,q(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})\right\}},$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$ and the denominator expectation E_2 is with respect to the proposal density of $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\theta}^*$, $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$, and $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})$ is the probability of move in the M–H step. This expression implies that a simulation consistent estimate of the posterior ordinate can be defined as

$$\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,q(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})}{J^{-1}\sum_{j=1}^{M}\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)}|\boldsymbol{y},\mathcal{M})}, \tag{41}$$

where $\{\boldsymbol{\theta}^{(g)}\}$ are the given draws from the posterior distribution while the draws $\boldsymbol{\theta}^{(j)}$ in the denominator are from $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$, given the fixed value $\boldsymbol{\theta}^*$. The second main result of the paper is that in the context of the multiple block M–H algorithm the reduced conditional ordinate can be expressed as

$$\begin{aligned}&\pi(\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_{i-1}^*)\\&\quad=\frac{E_1\left\{\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})q_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})\right\}},\end{aligned} \tag{42}$$

where E_1 is the expectation with respect to $\pi(\boldsymbol{\theta}_i,\boldsymbol{\psi}^{i+1}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*)$ and E_2 that with respect to the product measure $\pi(\boldsymbol{\psi}^{i+1}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_i^*)\,q_i(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})$. The quantity $\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})$ is the usual *conditional* M–H probability of move. The two expectations can be estimated from the output of the reduced runs in an obvious way. An example of this technique in action is provided next.

Consider the data set that was introduced in Section 8 in connection with the multivariate probit model. In this setting, the full conditonal density of the correlations is not in tractable form. Assume as before that the marginal probability of wheeze is given by

$$\Pr(y_{ij}=1|\mathcal{M}_k,\boldsymbol{\beta}) = \Phi(\beta_0+\beta_1x_{1ij}+\beta_2x_{2ij}+\beta_3x_{3ij}), \quad i \leqslant 537,\ j \leqslant 4,$$

where, as before, the dependence of $\boldsymbol{\beta}$ on the model is suppressed for convenience, x_1 is the age of the child centered at nine years, x_2 is a binary indicator variable representing the mother's smoking habit during the first year of the study, and $x_3 = x_1 x_2$. Now suppose that interest centers on three alternative models generated by three alternative correlation matrices. Let these models be defined as

- $\mathcal{M}_1$: Unrestricted $\boldsymbol{\Sigma}$ except for the unit constraints on the diagonal. In this case $\boldsymbol{\sigma}$ consists of six unknown elements.
- $\mathcal{M}_2$: Equicorrelated $\boldsymbol{\Sigma}$ where the correlations are all equal and described by a single parameter ρ.
- $\mathcal{M}_3$: Toeplitz $\boldsymbol{\Sigma}$ wherein the correlations depend on a single parameter ω but under the restriction that $\mathrm{Corr}(Z_{ik}, Z_{il}) = \omega^{|k-l|}$.

Assume that the prior on $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is independent Gaussian with a mean of zero and a variance of ten. Also let the prior on the correlations $\boldsymbol{\sigma}$ be normal with mean of zero and covariance equal to the identity matrix (truncated to the region C) and that on ρ and ω be normal truncated to the interval $(-1, 1)$.

For each model, 10 000 iterations of Algorithm 17 are used to obtain the posterior sample and the posterior ordinate, using $\mathcal{M}_1$ for illustration, is computed as

$$\pi(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1) = \pi(\boldsymbol{\sigma}^* | \boldsymbol{y}, \mathcal{M}_1)\, \pi(\boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*).$$

To estimate the marginal ordinate one can apply Equation (42) leading to the estimate

$$\hat{\pi}(\boldsymbol{\sigma}^* | \boldsymbol{y}, \mathcal{M}_1) = \frac{M^{-1} \sum_{g=1}^{M} \alpha(\boldsymbol{\sigma}^{(g)}, \boldsymbol{\sigma}^* | \boldsymbol{y}, \boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\})\, q(\boldsymbol{\sigma}^* | \boldsymbol{y}, \boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\})}{J^{-1} \sum_{j=1}^{J} \alpha(\boldsymbol{\sigma}^{(j)} | \boldsymbol{y}, \boldsymbol{\beta}^{(j)}, \{\boldsymbol{z}_i^{(j)}\})}, \tag{43}$$

where α is the probability of move defined in Algorithm 17, $\{\boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\}, \boldsymbol{\sigma}^{(g)}\}$ are values drawn from the full MCMC run and the values $\{\boldsymbol{\beta}^{(j)}, \{\boldsymbol{z}_i^{(j)}\}, \boldsymbol{\sigma}^{(j)}\}$ in the denominator are from a reduced run consisting of the densities

$$\pi(\boldsymbol{\beta} | \boldsymbol{y}, \mathcal{M}_1, \{\boldsymbol{z}_i\}, \boldsymbol{\Sigma}^*); \quad \pi(\{\boldsymbol{z}_i\} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*), \tag{44}$$

after $\boldsymbol{\Sigma}$ is fixed at $\boldsymbol{\Sigma}^*$. In particular, the draws for the denominator are from the distributions

$$\begin{aligned} &\boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)} \sim \pi(\boldsymbol{\beta}, \boldsymbol{z} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*), \\ &\quad \boldsymbol{\sigma}^{(j)} \sim q(\boldsymbol{\sigma}^*, \boldsymbol{\sigma} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)}), \quad j \leqslant J. \end{aligned}$$

The sampled variates $\{\boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)}\}$ from this reduced run are also used to estimate the second ordinate as

$$\hat{\pi}(\boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*) = M^{-1} \sum_{j=1}^{M} \phi_J(\boldsymbol{\beta}^* | \hat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{B}_n^*), \tag{45}$$

where $\hat{\boldsymbol{\beta}}^{(j)} = \boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{*-1} \boldsymbol{z}_i^{(j)})$ and $\boldsymbol{B}_n^* = (\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{*-1} \boldsymbol{X}_i)^{-1}$. It should be noted that estimates of *both* ordinates are available at the conclusion of the single reduced run.

Table 3
Log-likelihood and log marginal likelihood by the Chib method of three models fit to the Ohio wheeze data [1]

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$\ln f(y \mid \mathcal{M}, \theta^*)$	−795.1869	−798.5567	−804.4102
$\ln m(y \mid \mathcal{M})$	−823.9188	−818.009	−824.0001

[1] $\mathcal{M}_1$, MVP with unrestricted correlations; $\mathcal{M}_2$, MVP with an equicorrelated correlation; $\mathcal{M}_3$, MVP with Toeplitz correlation structure.

The marginal likelihood computation is completed by evaluating the likelihood function at the point $(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)$ by the Geweke–Hajivassiliou–Keane method. The resulting marginal likelihoods of the three alternative models are reported in Table 3. On the basis of these marginal likelihoods we conclude that the data tend to support the MVP model with equicorrelated correlations.

10.3. Model space-parameter space MCMC algorithms

When one is presented with a large collection of candidate models $\{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$, each with parameters $\boldsymbol{\theta}_k \in B_k \subseteq \Re^{d_k}$, direct fitting of each model to find the marginal likelihood can be computationally expensive. In such cases it may be more fruitful to utilize model space-parameter space MCMC algorithms that eschew direct fitting of each model for an alternative simulation of a "mega model" where a model index random variable, denoted as $\mathcal{M}$, taking values on the integers from 1 to K, is sampled in tandem with the parameters. The posterior distribution of $\mathcal{M}$ is then computed as the frequency of times each model is visited.

In this section we discuss two general model space-parameter space algorithms that have been proposed in the literature. These are the algorithms of Carlin and Chib (1995) and the reversible jump method of Green (1995).

To explain the Carlin and Chib (1995) algorithm, write $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ and assume that each model is defined by the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)$ and (proper) priors $p(\boldsymbol{\theta}_k|\mathcal{M} = k)$. Note that each model is non-nested. Now by the law of total probability the joint distribution of the data, the parameters and the model index is given by

$$f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M} = k) = f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\, p(\boldsymbol{\theta}_k|\mathcal{M} = k)\, p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k) \Pr(\mathcal{M} = k). \tag{46}$$

Thus, in addition to the usual inputs, the joint probability model requires the specification of the densities $\{p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k),\ k \leqslant K\}$. These are called *pseudo priors* or *linking densities* and are necessary to complete the probability model but play no role in determining the marginal likelihood of $\mathcal{M} = k$ since

$$m(\boldsymbol{y}, \mathcal{M} = k) = \int f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M} = k)\, \mathrm{d}\boldsymbol{\theta},$$

regardless of what pseudo priors are chosen. Hence, the linking densities may be chosen in any convenient way that promotes the working of the MCMC sampling procedure. The goal now is to sample the posterior distribution on model space and parameter space

$$\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathcal{M}|\boldsymbol{y}) \propto f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M}),$$

by MCMC methods.

Algorithm 18: Model space MCMC

```
(1) Sample
        $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M} = k) \propto f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\,\pi(\boldsymbol{\theta}_k|\mathcal{M} = k), \quad \mathcal{M} = k,$
        $\boldsymbol{\theta}_{-k} \sim p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k), \quad \mathcal{M} \neq k.$
(2) Model jump
    (a) Calculate
        $p_k = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)p(\boldsymbol{\theta}_k|\mathcal{M} = k)p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k)\Pr(\mathcal{M} = k)}{\sum_{l=1}^{K} f(\boldsymbol{y}|\boldsymbol{\theta}_l, \mathcal{M} = l)p(\boldsymbol{\theta}_l|\mathcal{M} = l)p(\boldsymbol{\theta}_{-l}|\boldsymbol{\theta}_l, \mathcal{M} = l)\Pr(M = l)}, k \leqslant K.$
    (b) Sample
        $\mathcal{M} \sim \{p_1, \ldots, p_K\}.$
(3) Goto 1.
```

Thus, when $\mathcal{M} = k$, we sample $\boldsymbol{\theta}_k$ from its full conditional distribution and the remaining parameters from their pseudo priors and the model index is sampled from the a discrete point distribution with probabilities $\{p_k\}$.

Algorithm 18 is conceptually quite simple and can be used without any difficulties when the number of models under consideration is small. When K is large, however, the specification of the pseudo priors and the requisite generation of each $\boldsymbol{\theta}_k$ within each cycle of the MCMC algorithm can be a computational burden. We also mention that the pseudo priors should be chosen to be close to the model specific posterior distributions. To understand the rationale for this recommendation suppose that the pseudo priors can be set exactly equal to the model specific posterior distributions as

$$p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k) = \prod_{l \neq k} \pi(\boldsymbol{\theta}_l|\boldsymbol{y}, \mathcal{M} = l).$$

Substituting this choice into the equation of p_k and simplifying we get

$$p_k = \frac{m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k)}{\sum_{l=1}^{K} m(\boldsymbol{y}|\mathcal{M} = l)\Pr(\mathcal{M} = l)}, \tag{47}$$

which is $\Pr(\mathcal{M} = k|\boldsymbol{y})$. Therefore, under this choice of pseudo priors, the Carlin–Chib algorithm generates the model move at each iteration of the sampling according to their posterior probabilities, without any required burn-in. Thus, by utilizing pseudo priors that are close to the model specific posterior distributions one promotes mixing on model space and more rapid convergence to the invariant target distribution $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K, \mathcal{M}|\boldsymbol{y})$.

Another point in connection with the above algorithm is that the joint distribution over parameter space and model space can be sampled by the M–H algorithm. For example, Dellaportas, Forster and Ntzoufras (1998) suggest that the discrete conditional distribution on the models be sampled by M–H algorithm in order to avoid the calculation of the denominator of p_k. Godsill (1998) considers the sampling of the entire joint distribution in Equation (46) by the M–H algorithm. Suppose that the proposal density on the joint space is specified as

$$q\left\{(\mathcal{M}=k, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k}), (\mathcal{M}=k', \boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}'_{-k'})\right\} = q_1(k,k')\, q_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_{k'}|k,k')\, p(\boldsymbol{\theta}'_{-k'}|\boldsymbol{\theta}'_{k'}, \mathcal{M}=k'), \tag{48}$$

where the pseudo prior is the proposal density of the parameters $\boldsymbol{\theta}_{-k'}$ not in the proposed model k'. It is important that q_1 not depend on the current value $(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k})$ and that q_2 not depend on the current value of $\boldsymbol{\theta}_{k'}$ in the model being proposed. Then, the probability of move from $(\mathcal{M}=k, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k})$ to $(\mathcal{M}=k', \boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}'_{-k'})$ in the M–H step, after substitutions and cancellations, reduces to

$$\min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}'_{k'}, \mathcal{M}=k')\, p(\boldsymbol{\theta}'_{k'}|\mathcal{M}=k')\, \Pr(\mathcal{M}=k')}{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}=k)\, p(\boldsymbol{\theta}_k|\mathcal{M}=k)\, \Pr(\mathcal{M}=k)} \frac{q_1(k',k)\, q_2(\boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}_k|k,k')}{q_1(k,k')\, q_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_{k'}|k,k')}\right\}, \tag{49}$$

which is completely independent of the pseudo priors. Thus, the sampling, or specification, of pseudo priors is not required in this version of the algorithm but the requirement that the parameters of each model be proposed in one block rules out many important problems.

We now turn to the reversible jump algorithm of Green (1995) which is designed primarily for *nested* models. In this algorithm, model space and parameter space moves from the current point $(\mathcal{M} = k, \boldsymbol{\theta}_k)$ to a new point $(\mathcal{M} = k', \boldsymbol{\theta}'_{k'})$ are made by a Metropolis–Hastings step in conjunction with a dimension matching condition to ensure that the resulting Markov chain is reversible. An application of the reversible jump method to choosing the number of components in a finite mixture of distribution model is provided by Richardson and Green (1997). The parameter space in this method is based on the *union* of the parameter spaces B_k. To describe the algorithm we let q denote a discrete mass function that gives the probability of each possible model given the current model and we let u' denote an increment/decrement random variable that takes one from the current point $\boldsymbol{\theta}_k$ to the new point $\boldsymbol{\theta}'_{k'}$.

Algorithm 19: Reversible jump model space MCMC

```
(1) Propose a new model $k'$

        $k' \sim q_1(k, k')$.
```

```
(2) Dimension matching
    (a) Propose
            $u' \sim q_2(u'|\boldsymbol{\theta}_k, k, k')$.
    (b) Set
            $(\boldsymbol{\theta}'_{k'}, u) = g_{k,k'}(\boldsymbol{\theta}_k, u')$,
        where $g_{k,k'}$ is a bijection between $(\boldsymbol{\theta}'_{k'}, u)$ and $(\boldsymbol{\theta}_k, u')$ and $\dim(\boldsymbol{\theta}_k)$
        $+ \dim(u') = \dim(\boldsymbol{\theta}'_{k'}) + \dim(u)$.
(3) M-H
    (a) Calculate
            $\alpha = \min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}'_{k'}, \mathcal{M} = k')\, p(\boldsymbol{\theta}'_{k'}|\mathcal{M} = k')\, \Pr(\mathcal{M} = k')}{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\, p(\boldsymbol{\theta}_k|\mathcal{M} = k)\, \Pr(\mathcal{M} = k)} \frac{q_1(k', k)\, q_2(u|\boldsymbol{\theta}_k, k, k')}{q_2(k, k')\, q_2(u'|\boldsymbol{\theta}_k, k, k')} \cdot J\right\}$,
        where
            $J = \left|\frac{\partial g_{k,k'}(\boldsymbol{\theta}_k, u')}{\partial(\boldsymbol{\theta}_k, u')}\right|$.
    (b) Move to $(k'; \boldsymbol{\theta}'_{k'}, u')$ with probability $\alpha$.
(4) Goto 1.
```

In the reversible jump method most of the tuning is in the specification of the proposal distribution q_2; a different proposal distribution is required if k' is a model with more parameters than model k than for the case when model k' has fewer parameters. This is the reason for the dependence of q_2 on not just $\boldsymbol{\theta}_k$ but also on (k, k'). In addition, the algorithm as stated by Green (1995) is designed for the situation where the competing models are nested and obtained by the removal or addition of different parameters, as for example in a variable selection problem.

10.4. Variable selection

Model space MCMC methods described above can be specialized to the problem of variable selection in regression. We first focus on this problem in the context of linear regression models with conjugate priors before discussing a more general situation.

Consider then the question of building a multiple regression model for a vector of n observations $\boldsymbol{y}$ in terms of a given set of covariates $\boldsymbol{X} = \{x_1, \ldots, x_p\}$. The goal is to find the "best" model of the form

$$\mathcal{M}_k\colon \boldsymbol{y} = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sigma\varepsilon,$$

where $\boldsymbol{X}_k$ is a $n \times d_k$ matrix composed of some or all variables from $\boldsymbol{X}$, σ^2 is a variance parameter and ε is $\mathcal{N}(0, \boldsymbol{I}_n)$. Under the assumption that any subset of the variables in

$\boldsymbol{X}$ can be used to form $\boldsymbol{X}_k$ it follows that the number of possible models is given by $K = 2^p$, which is a large number even if p is as small as fifteen. Thus, unless p is small, when the marginal likelihoods can be computed for each possible $\boldsymbol{X}_k$, it is helpful to use simulation-based methods that traverse the space of possible models to determine the subsets that are most supported by the data.

Raftery, Madigan and Hoeting (1997) develop one approach that is based on the use of conjugate priors. Let the parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma^2)$ of model $\mathcal{M}_k$ follow the conjugate prior distributions

$$\boldsymbol{\beta}_k|\mathcal{M} = k, \sigma^2 \sim \mathcal{N}_{d_k}(\mathbf{0}, \sigma^2\boldsymbol{B}_{0k}); \, \sigma^2|\mathcal{M} = k \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right), \tag{50}$$

which implies after some algebra that the marginal likelihood of $\mathcal{M}_k$ is

$$m(\boldsymbol{y}|\mathcal{M} = k) = \frac{\Gamma\{(\nu_0 + n)/2\}}{\Gamma(\nu_0/2)(\delta_0\pi)^{n/2}} |\boldsymbol{B}_k|^{1/2} \times \left(1 + \frac{1}{\delta_0}\boldsymbol{y}'\boldsymbol{B}_k\boldsymbol{y}\right)^{-(n+\nu_0)/2},$$

where

$$\boldsymbol{B}_k = \boldsymbol{I}_n - \boldsymbol{X}_k(\boldsymbol{B}_{0k}^{-1} + \boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'.$$

Raftery, Madigan and Hoeting (1997) specify a MCMC chain to sample model space in which the target distribution is the *univariate* discrete distribution with probabilities

$$\Pr(\mathcal{M} = k|\boldsymbol{y}) = p_k \propto m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k), \quad k \leqslant K. \tag{51}$$

Although this distribution can in principle be normalized, the normalization constant is computationally expensive to calculate when K is large (but one can argue that expending the necessary computational effort is always desirable). This motivates the sampling of Equation (51) by the Metropolis–Hastings algorithm. For each model $\mathcal{M} = k$ define a neighborhood nbd($\mathcal{M} = k$) which consists of the model $\mathcal{M} = k$ and models with either one more variable or one fewer variable than $\mathcal{M} = k$. Define a transition matrix $q_1(k, k')$ which puts uniform probability over models k' that are in nbd($\mathcal{M} = k$) and zero probability for all other models. Given that the chain is currently at the point ($\mathcal{M} = k$) a move to the proposed model k' is made with probability

$$\min\left\{\frac{m(\boldsymbol{y}|\mathcal{M} = k')\Pr(\mathcal{M} = k')}{m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k)}\frac{q_1(k', k)}{q_1(k, k')}, 1\right\}. \tag{52}$$

If the proposed move is rejected the chain stays at $\mathcal{M} = k$.

When conjugate priors are not assumed for $\boldsymbol{\theta}_k$, or when the model is more complicated than multiple regression, it is not possible to find the marginal likelihood of each model analytically. It then becomes necessary to sample both the parameters and the model index jointly as in the general model space-parameter space algorithms mentioned above. The approaches that have been developed for this case, however, treat the various models as nested.

Suppose that the coefficients attached to the p possible covariates in the model are denoted by $\eta = \{\eta_1, \ldots, \eta_p\}$, where any common noise variances or other common parameters are suppressed from the notation and the discussion. Now associate with each coefficient η_j an indicator variable δ_j which takes the value one if the coefficient is in the model and the value zero otherwise and let η_δ denote the set of active η_j's given a configuration $\boldsymbol{\delta}$ and let $\eta_{-\boldsymbol{\delta}}$ denote the complementary η_j's. For example, if $p = 5$ and $\boldsymbol{\delta} = \{1, 0, 0, 1, 1\}$, then $\eta_{\boldsymbol{\delta}} = \{\eta_1, \eta_4, \eta_5\}$ and $\eta_{-\boldsymbol{\delta}} = \{\eta_2, \eta_3\}$. A variable selection MCMC algorithm can now be developed by sampling the joint posterior distribution $\pi(\delta_1, \eta_1, \ldots, \delta_p, \eta_p|\boldsymbol{y})$. Particular implementations representing different blocking schemes to sample this joint distribution are discussed by Kuo and Mallick (1998), Geweke (1996) and Smith and Kohn (1996). For example, in the algorithm of Kuo and Mallick (1998), the posterior distribution is sampled by recursively simulating the $\{\eta_1, \ldots, \eta_p\}$ from the distributions

$$\eta_j \sim \pi(\eta_j|\boldsymbol{y}, \eta_{-j}, \boldsymbol{\delta}) \propto \begin{cases} f(\boldsymbol{y}|\eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})p(\eta_{\boldsymbol{\delta}}|\boldsymbol{\delta}) & \text{if } \delta_j = 1, \\ p(\eta_j|\eta_{-j}, \boldsymbol{\delta}) & \text{if } \delta_j = 0, \end{cases}$$

where $p(\eta_j|\eta_{-j}, \boldsymbol{\delta})$ is a pseudo prior because it represents the distribution of η_j when η_j is not in the current configuration. Next, the variable indicators $\{\delta_1, \ldots, \delta_p\}$ are sampled one at a time from the two point mass function

$$\delta_j \sim \Pr(\delta_j|\boldsymbol{y}, \eta_{-j}, \boldsymbol{\delta}_{-j}) \propto f(\boldsymbol{y}|\eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})\, p(\eta_{\boldsymbol{\delta}}|\boldsymbol{\delta})\, p(\eta_{-\boldsymbol{\delta}}|\eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})\, p(\delta_j),$$

where $p(\eta_{-\boldsymbol{\delta}}|\eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})$ is the pseudo prior. These two steps are iterated. Procedures to sample (δ_j, η_j) in one block given all the other blocks are presented by Geweke (1996) and Smith and Kohn (1996).

George and McCulloch (1993, 1997) develop an important alternative simulation-based approach for the variable selection problem that has been extensively studied and refined. In their approach, the variable selection problem is cast in terms of a hierarchical model of the type

$$\begin{aligned} &\boldsymbol{y} \sim \boldsymbol{X}\boldsymbol{\beta} + \sigma\varepsilon, \quad \boldsymbol{\beta}_j|\gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2), \\ &\quad \Pr(\gamma_j = 1) = 1 - \Pr(\gamma_j = 0) = p_j, \end{aligned}$$

where τ_j^2 is a small positive number and c_j a large positive number. In this specification each component of $\boldsymbol{\beta}$ is assumed to come from a mixture of two normal distributions such that $\gamma_j = 0$ corresponds to the case where $\boldsymbol{\beta}_j$ can be assumed to be zero. It should be noted that in this framework a particular covariate is never strictly removed from the model; exclusion from the model corresponds to a high posterior probability of the event that $\gamma_j = 0$. George and McCulloch (1993) sample the posterior distribution of $(\boldsymbol{\beta}, \{\gamma_j\})$ by the Gibbs sampling algorithm.

10.5. Remark

We conclude this discussion by pointing out that convergence checks of the Markov chain in model space algorithms is quite difficult and has not been satisfactorily addressed in the literature. When the model space is large, as for example in the variable selection problem, one cannot be sure that all models supported by the data have been visited according to their posterior probabilities. Of course if the model space is diminished to ensure better coverage of the various models it may happen that direct computation of the marginal likelihood becomes feasible, thereby removing any justification for considering a model space algorithm in the first place. This tension in the choice between direct computation and model space algorithms is real and cannot be adjudicated in the absence of a concrete problem.

11. MCMC methods in optimization problems

Suppose that we are given a particular function $h(\boldsymbol{\theta})$, say the log likelihood of a given model, and interest lies in the value of $\boldsymbol{\theta}$ that maximizes this function. In some cases, this optimization problem can be quite effectively solved by MCMC methods. One somewhat coarse possibility is to obtain draws $\{\boldsymbol{\theta}^{(j)}\}$ from a density proportional to $h(\boldsymbol{\theta})$ and to find the value of $\boldsymbol{\theta}$ that corresponds to the maximum of $\{h(\boldsymbol{\theta}^{(j)}\}$. Another more precise technique goes by the name of simulated annealing which appears in Metropolis et al. (1953) and is closely related to the Metropolis simulation method. In the simulated annealing method, which is most typically used to maximize a function on a finite but large set, one uses the Metropolis method to sample the distribution

$$\pi(\boldsymbol{\theta}) \propto \exp\{h(\boldsymbol{\theta})/T\},$$

where T is referred to as the temperature. The temperature variable is gradually reduced as the sampling proceeds [for example, see Geman and Geman (1984)]. It can be shown that in the finite case, the values of $\boldsymbol{\theta}$ produced by the simulated annealing method concentrate around the local maximum of the function $h(\boldsymbol{\theta})$.

Another method of interest is a MCMC version of the EM algorithm which can be used to find the maximum likelihood estimate in certain situations. Suppose that $\boldsymbol{z}$ represents missing data and $f(\boldsymbol{y}|\mathcal{M},\boldsymbol{\theta})$ denotes the likelihood function. Also suppose that

$$f(\boldsymbol{y}|\mathcal{M},\boldsymbol{\theta}) = \int f(\boldsymbol{y},\boldsymbol{z}|\mathcal{M},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z},$$

is difficult to compute but that the complete data likelihood $f(\boldsymbol{y},\boldsymbol{z}|\mathcal{M},\boldsymbol{\theta})$ is available, as in the models with a missing data structure in Section 8. For this problem, the standard EM algorithm [Dempster, Laird and Rubin (1977)] requires the recursive implementation of two steps: the expectation or E-step and the maximization or M-step. In the E-step, given the current guess of the maximizer $\boldsymbol{\theta}^{(j)}$, one computes

$$Q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}) = \int \ln f(\boldsymbol{y}, \boldsymbol{z}|\mathcal{M}, \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z},$$

while in the M-step the Q function is maximized to obtain a revised guess of the maximizer, i.e.,

$$\boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}).$$

Wu (1983) has shown that under regularity conditions the sequence of values $\{\boldsymbol{\theta}^{(j)}\}$ generated by these steps converges to the maximizer of the function $f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\theta})$.

The MCEM algorithm is a variant of the EM algorithm, proposed by Wei and Tanner (1990b), in which the E-step, which is often intractable, is computed by Monte Carlo averaging over values of $\boldsymbol{z}$ drawn from $f(\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})$, which in the MCMC context is the full conditional distribution of the latent data. Then, the revised value of $\boldsymbol{\theta}$ is obtained by maximizing the Monte Carlo estimate of the Q function. Specifically, the MCEM algorithm is defined by iterating on the following steps:

$$\hat{Q}_M(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}) = M^{-1} \sum_{j=1}^{M} \ln f(\boldsymbol{y}, \boldsymbol{z}^{(j)}|\mathcal{M}, \boldsymbol{\theta}),$$
$$\boldsymbol{z}^{(j)} \sim f(\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}), \quad \boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} \hat{Q}(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}).$$

As suggested by Wei and Tanner (1990b), these iterations are started with a small value of M that is increased as the maximizer is approached. One point to note is that in general, the MCEM algorithm, similar to the EM algorithm, can be slow to converge to the mode but it should be possible to adapt the ideas described in Liu, Rubin and Wu (1998) to address this problem. Another point to note is that the computation of the $\hat{Q}_M$ function can be expensive when M is large. Despite these potential difficulties, a number of applications of the MCEM algorithm have now appeared in the literature. These include Chan and Ledolter (1995), Chib (1996, 1998), Chib and Greenberg (1998), Chib, Greenberg and Winkelmann (1998) and Booth and Hobert (1999).

Given the modal value $\hat{\theta}$, the standard errors of the MLE are obtained by the formula of Louis (1982). In particular, the observed information matrix is given by

$$-E\left\{\frac{\partial^2 \ln f(\boldsymbol{y}, \boldsymbol{z}|\mathcal{M}, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} - \mathrm{Var}\left\{\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}|\mathcal{M}, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\},$$

where the expectation and variance are with respect to the distribution $\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \hat{\theta}$. This expression is estimated by taking an additional J draws $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(J)}\}$ from $\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \hat{\theta}$ and computing

$$-J^{-1}\sum_{k=1}^{J}\frac{\partial^2 \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$
$$-J^{-1}\sum_{k=1}^{J}\left(\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}}-\boldsymbol{m}\right)\left(\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}}-\boldsymbol{m}\right)',$$

where

$$\boldsymbol{m} = J^{-1}\sum_{k=1}^{J}\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \hat{\theta})}{\partial\boldsymbol{\theta}}.$$

Standard errors are equal to the square roots of the diagonal elements of the inverse of the estimated information matrix.

12. Concluding remarks

In this survey we have provided an outline of Markov chain Monte Carlo methods with emphasis on techniques that prove useful in Bayesian statistical inference. Further developments of these methods continue to occur but the ideas and details presented in this survey should provide a reasonable starting point to understand the current and emerging literature. Two recent developments are the slice sampling method discussed by Mira and Tierney (1998), Damien et al. (1999) and Roberts and Rosenthal (1999) and the perfect sampling method proposed by Propp and Wilson (1996). The slice sampling method is based on the introduction of auxiliary uniform random variables to simplify the sampling and improve mixing while the perfect sampling method uses Markov chain coupling to generate an exact draw from the target distribution. These methods are in their infancy and can be currently applied only under rather restrictive assumptions on the target distribution but it is possible that more general versions of these methods will eventually become available.

Other interesting developments are now occurring in the field of applied Bayesian inference as practical problems are being addressed by the methods summarized in this survey. These applications are appearing at a steady rate in various areas. For example, a partial list of fields and papers within fields include: biostatistical time series analysis [West, Prado and Krystal (1999)]; economics [Chamberlain and Hirano (1997), Filardo and Gordon (1998), Gawande (1998), Lancaster (1997), Li (1998), Kiefer and Steel (1998), Kim and Nelson (1999), Koop and Potter (1999), Martin (1999), Paap and van Dijk (1999), So, Lam and Li (1998)]; finance [Jones (1999), Pastor and Stambaugh

(1999)]; marketing [Allenby, Leone and Jen (1999), Bradlow and Zaslavsky (1999), Manchanda, Ansari and Gupta (1999), Montgomery and Rossi (1999), Young, DeSarbo and Morwitz (1998)]; political science [King, Rosen and Tanner (1999), Quinn, Martin and Whitford (1999), Smith (1999)]; and many others.

One can claim that with the ever increasing power of computing hardware, and the experience of the past ten years, the future of simulation-based inference using MCMC methods is secure.

References

Albert, J. (1993), “Teaching Bayesian statistics using sampling methods and MINITAB”, American Statistician 47:182–191.

Albert, J., and S. Chib (1993a), “Bayesian analysis of binary and polychotomous response data”, Journal of the American Statistical Association 88:669–679.

Albert, J., and S. Chib (1993b), “Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts”, Journal of Business and Economic Statistics 11:1–15.

Albert, J., and S. Chib (1995), “Bayesian residual analysis for binary response models”, Biometrika 82:747–759.

Albert, J., and S. Chib (1996), “Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo”, in: T. Fomby and R.C. Hill, eds., Advances in Econometrics, Vol. 11A (Jai Press, Greenwich, CT) 3–24.

Albert, J., and S. Chib (1997), “Bayesian tests and model diagnostics in conditionally independent hierarchical models”, Journal of the American Statistical Association 92:916–925.

Albert, J., and S. Chib (1998), “Sequential Ordinal Modeling with Applications to Survival Data”. Biometrics, in press.

Allenby, G.M., R.P. Leone and L. Jen (1999), “A dynamic model of purchase timing with application to direct marketing”, Journal of the American Statistical Association 94:365–374.

Bennett, C.H. (1976), “Efficient estimation of free energy differences from Monte Carlo data”, Journal of Computational Physics 22:245–268.

Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis, 2nd edition (Springer, New York).

Bernardo, J.M., and A.F.M. Smith (1994), Bayesian Theory (Wiley, New York).

Besag, J. (1974), “Spatial interaction and the statistical analysis of lattice systems (with discussion)”, Journal of the Royal Statistical Society B 36:192–236.

Besag, J., E. Green, D. Higdon and K.L. Mengersen (1995), “Bayesian computation and stochastic systems (with discussion)”, Statistical Science 10:3–66.

Best, N.G., M.K. Cowles and S.K. Vines (1995), “CODA: convergence diagnostics and output analysis software for Gibbs sampling”, Technical report (Cambridge MRC Biostatistics Unit).

Billio, M., A. Monfort and C.P. Robert (1999), “Bayesian estimation of switching ARMA models”, Journal of Econometrics 93:229–255.

Blattberg, R.C., and E.I. George (1991), “Shrinkage estimation of price and promotional elasticities: seemingly unrelated equations”, Journal of the American Statistical Association 86:304–315.

Booth, J.G., and J.P. Hobert (1999), “Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm”, Journal of the Royal Statistical Society B 61:265–285.

Bradlow, E., and A.M. Zaslavsky (1999), “A hierarchical latent variable model for ordinal data from a customer satisfaction survey with “no answer” responses”, Journal of the American Statistical Association 94:43–52.

Brooks, S.P. (1998), “Markov chain Monte Carlo and its application”, Statistician 47:69–100.

Brooks, S.P., P. Dellaportas and G.O. Roberts (1997), "A total variation method for diagnosing convergence of MCMC algorithms", Journal of Computational and Graphical Statistics 6:251–265.

Carlin, B., A.E. Gelfand and A.F.M. Smith (1992), "Hierarchical Bayesian analysis of changepoint problems", Applied Statistics 41:389–405.

Carlin, B.P., and S. Chib (1995), "Bayesian model choice via Markov Chain Monte Carlo methods", Journal of the Royal Statistical Society B 57:473–484.

Carlin, B.P., and T.A. Louis (2000), Bayes and Empirical Bayes Methods for Data Analysis, 2nd Edition (Chapman and Hall, London).

Carlin, B.P., and N.G. Polson (1991), "Inference for non-conjugate Bayesian models using the Gibbs sampler", Canadian Journal of Statistics 19:399–405.

Carlin, B.P., N.G. Polson and D.S. Stoffer (1992), "A Monte Carlo approach to nonnormal and nonlinear state-space modeling", Journal of the American Statistical Association 87:493–500.

Carter, C., and R. Kohn (1994), "On Gibbs sampling for state space models", Biometrika 81:541–553.

Carter, C., and R. Kohn (1996), "Markov chain Monte Carlo for conditionally Gaussian state space models", Biometrika 83:589–601.

Casella, G., and E.I. George (1992), "Explaining the Gibbs sampler", American Statistician 46:167–174.

Casella, G., and C.P. Robert (1996), "Rao-Blackwellization of sampling schemes", Biometrika 83:81–94.

Chamberlain, G., and K. Hirano (1997), "Predictive distributions based on longitudinal earnings data", Manuscript (Department of Economics, Harvard University).

Chan, K.S. (1993), "Asymptotic behavior of the Gibbs sampler", Journal of the American Statistical Association 88:320–326.

Chan, K.S., and C.J. Geyer (1994), "Discussion of Markov chains for exploring posterior distributions", Annals of Statistics 22:1747–1758.

Chan, K.S., and J. Ledolter (1995), "Monte Carlo EM estimation for time series models involving counts", Journal of the American Statistical Association 90:242–252.

Chen, M.-H. (1994), "Importance-weighted marginal Bayesian posterior density estimation", Journal of the American Statistical Association 89:818–824.

Chen, M.-H., and Q.-M. Shao (1997), "On Monte Carlo methods for estimating ratios of normalizing constants", Annals of Statistics 25:1563–1594.

Chen, M.-H., and Q.-M. Shao (1999), "Monte Carlo estimation of Bayesian credible and HPD intervals", Journal of Computational and Graphical Statistics 8:69–92.

Chib, S. (1992), "Bayes regression for the Tobit censored regression model", Journal of Econometrics 51:79–99.

Chib, S. (1993), "Bayes regression with autocorrelated errors: a Gibbs sampling approach", Journal of Econometrics 58:275–294.

Chib, S. (1995), "Marginal likelihood from the Gibbs output", Journal of the American Statistical Association 90:1313–1321.

Chib, S. (1996), "Calculating posterior distributions and modal estimates in Markov mixture models", Journal of Econometrics 75:79–97.

Chib, S. (1998), "Estimation and comparison of multiple change point models", Journal of Econometrics 86:221–241.

Chib, S., and B.P. Carlin (1999), "On MCMC sampling in hierarchical longitudinal models", Statistics and Computing 9:17–26.

Chib, S., and E. Greenberg (1994), "Bayes inference for regression models with ARMA(p,q) errors", Journal of Econometrics 64:183–206.

Chib, S., and E. Greenberg (1995a), "Understanding the Metropolis–Hastings algorithm", American Statistician 49:327–335.

Chib, S., and E. Greenberg (1995b), "Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models", Journal of Econometrics 68:339–360.

Chib, S., and E. Greenberg (1996), "Markov chain Monte Carlo simulation methods in econometrics", Econometric Theory 12:409–431.

Chib, S., and E. Greenberg (1998), “Analysis of multivariate probit models”, Biometrika 85:347–361.

Chib, S., and B. Hamilton (2000), “Bayesian analysis of cross section and clustered data treatment models”, Journal of Econometrics 97:25–50.

Chib, S., and I. Jeliazkov (2001), “Marginal likelihood from the Metropolis–Hastings output”, Journal of the American Statistical Association 96:270–281.

Chib, S., E. Greenberg and R. Winkelmann (1998), “Posterior simulation and Bayes factors in panel count data models”, Journal of Econometrics 86:33–54.

Chib, S., F. Nardari and N. Shephard (1998), “Markov Chain Monte Carlo analysis of generalized stochastic volatility models”, Journal of Econometrics, under review.

Chib, S., F. Nardari and N. Shephard (1999), “Analysis of high dimensional multivariate stochastic volatility models”, Technical report (John M. Olin School of Business, Washington University, St. Louis).

Chipman, H.A., E.I. George and R.E. McCulloch (1998), “Bayesian CART model search (with discussion)”, Journal of the American Statistical Association 93:935–948.

Cowles, M.K. (1996), “Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models”, Statistics and Computing 6:101–111.

Cowles, M.K., and B. Carlin (1996), “Markov chain Monte Carlo convergence diagnostics: a comparative review”, Journal of the American Statistical Association 91:883–904.

Damien, P., J. Wakefield and S. Walker (1999), “Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables”, Journal of the Royal Statistical Society B 61:331–344.

de Jong, P., and N. Shephard (1995), “The simulation smoother for time series models”, Biometrika 82:339–350.

Dellaportas, P., and A.F.M. Smith (1993), “Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling”, Applied Statistics 42:443–459.

Dellaportas, P., J.J. Forster and I. Ntzoufras (1998), “On Bayesian model and variable selection using MCMC”, Technical report (University of Economics and Business, Greece).

Dempster, A.P., N.M. Laird and D.B. Rubin (1977), “Maximum likelihood estimation from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society B 39:1–38.

Denison, D.G.T., B.K. Mallick and A.F.M. Smith (1998), “A Bayesian CART algorithm”, Biometrika 85:363–377.

Devroye, L. (1985), Non-Uniform Random Variate Generation (Springer, New York).

DiCiccio, T.J., R.E. Kass, A.E. Raftery and L. Wasserman (1997), “Computing Bayes factors by combining simulation and asymptotic approximations”, Journal of the American Statistical Association 92:903–915.

Diebolt, J., and C.P. Robert (1994), “Estimation of finite mixture distributions through Bayesian sampling”, Journal of the Royal Statistical Society B 56:363–375.

Diggle, P., K.-Y. Liang and S.L. Zeger (1995), Analysis of Longitudinal Data (Oxford University Press, Oxford).

Elerian, O., S. Chib and N. Shephard (1999), “Likelihood inference for discretely observed nonlinear diffusions”, Econometrica, in press.

Escobar, M.D., and M. West (1995), “Bayesian prediction and density estimation”, Journal of the American Statistical Association 90:577–588.

Filardo, A.J., and S.F. Gordon (1998), “Business cycle durations”, Journal of Econometrics 85:99–123.

Fruhwirth-Schnatter, S. (1994), “Data augmentation and dynamic linear models”, Journal of Time Series Analysis 15:183–202.

Gammerman, D. (1997), Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference (Chapman and Hall, London).

Gammerman, D., and H.S. Migon (1993), “Dynamic hierarchical models”, Journal of the Royal Statistical Society B 55:629–642.

Gawande, K. (1998), “Comparing theories of endogenous protection: Bayesian comparison of Tobit models using Gibbs sampling output”, Review of Economics and Statistics 80:128–140.

Gelfand, A.E., and D. Dey (1994), “Bayesian model choice: asymptotics and exact calculations”, Journal of the Royal Statistical Society B 56:501–514.

Gelfand, A.E., and A.F.M. Smith (1990), “Sampling-based approaches to calculating marginal densities”, Journal of the American Statistical Association 85:398–409.

Gelfand, A.E., and A.F.M. Smith (1992), “Bayesian statistics without tears: a sampling–resampling perspective”, American Statistician 46:84–88.

Gelfand, A.E., S. Hills, A. Racine-Poon and A.F.M. Smith (1990), “Illustration of Bayesian inference in normal data models using Gibbs sampling”, Journal of the American Statistical Association 85:972–982.

Gelfand, A.E., S.K. Sahu and B.P. Carlin (1995), “Efficient parameterizations for normal linear mixed models”, Biometrika 82:479–488.

Gelman, A., and D.B. Rubin (1992), “Inference from iterative simulation using multiple sequences”, Statistical Science 4:457–472.

Gelman, A., X.L. Meng, H.S. Stern and D.B. Rubin (1995), Bayesian Data Analysis (Chapman and Hall, London).

Geman, S., and D. Geman (1984), “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”, IEEE Transactions on Pattern Analysis and Machine Intelligence 12:609–628.

Gentle, J.E. (1998), Random Number Generation and Monte Carlo Methods (Springer, New York).

George, E.I., and R.E. McCulloch (1993), “Variable selection via Gibbs sampling”, Journal of the American Statistical Association 88:881–889.

George, E.I., and R.E. McCulloch (1997), “Approaches to Bayesian variable selection”, Statistica Sinica 7:339–373.

Geweke, J. (1989), “Bayesian inference in econometric models using Monte Carlo integration”, Econometrica 57:1317–1340.

Geweke, J. (1991), “Efficient simulation from the multivariate normal and student-*t* distributions subject to linear constraints”, in: E. Keramidas and S. Kaufman, eds., Computing Science and Statistics: Proceedings of the 23rd Symposium (Interface Foundation of North America) 571–578.

Geweke, J. (1992), “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments”, in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Bayesian Statistics (Oxford University Press, New York) 169–193.

Geweke, J. (1996), “Variable selection and model comparison in regression”, in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Bayesian Statistics (Oxford University Press, New York) 609–620.

Geweke, J. (1997), “Posterior simulators in econometrics”, in: D.M. Kreps and K.F. Wallis, eds., Advances in Economics and Econometrics: Theory and Applications, 7th World Congress (Cambridge University Press, Cambridge) 128–165.

Geyer, C. (1995), “Conditioning in Markov chain Monte Carlo”, Journal of Computational and Graphical Statistics 4:148–154.

Geyer, C.J., and E.A. Thompson (1995), “Annealing Markov chain Monte Carlo with applications to ancestral inference”, Journal of the American Statistical Association 90:909–920.

Ghysels, E., A.C. Harvey and E. Renault (1996), “Stochastic volatility”, in: C.R. Rao and G.S. Maddala, eds., Statistical Methods in Finance (North-Holland, Amsterdam) 119–191.

Gilks, W.R., S. Richardson and D.J. Spiegelhalter (1996), Markov Chain Monte Carlo in Practice (Chapman and Hall, London).

Godsill, S.J. (1998), “On the relationship between model uncertainty methods”, Technical report (Signal Processing Group, Cambridge University).

Green, P.E. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”, Biometrika 82:711–732.

Hamilton, J.D. (1989), “A new approach to the economic analysis of nonstationary time series subject to changes in regime”, Econometrica 57:357–384.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway and E. Ostrowski (1994), A Handbook of Small Data Sets (Chapman and Hall, London).

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", Biometrika 57:97–109.

Hills, S.E., and A.F.M. Smith (1992), "Parameterization issues in Bayesian inference", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 641–649.

Jacquier, E., N.G. Polson and P.E. Rossi (1994), "Bayesian analysis of stochastic volatility models (with discussion)", Journal of Business and Economic Statistics 12:371–417.

Jeffreys, H. (1961), Theory of Probability, 3rd edition (Oxford University Press, New York).

Jones, C.S. (1999), "The dynamics of stochastic volatility", Manuscript (University of Rochester).

Kiefer, N.M., and M.F.J. Steel (1998), "Bayesian analysis of the prototypal search model", Journal of Business and Economic Statistics 16:178–186.

Kim, C.-J., and C.R. Nelson (1999), "Has the US become more stable? A Bayesian approach based on a Markov-switching model of business cycle", The Review of Economics and Statistics 81:608–616.

Kim, S., N. Shephard and S. Chib (1998), "Stochastic volatility: likelihood inference and comparison with ARCH models", Review of Economic Studies 65:361–393.

King, G., O. Rosen and M.A. Tanner (1999), "Binomial-beta hierarchical models for ecological inference", Sociological Method Research 28:61–90.

Kloek, T., and H.K. van Dijk (1978), "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo", Econometrica 46:1–20.

Koop, G., and S.M. Potter (1999), "Bayes factors and nonlinearity: evidence from economic time series", Journal of Econometrics 88:251–281.

Kuo, L., and B. Mallick (1998), "Variable selection for regression models", Sankhya B 60:65–81.

Laird, N.M., and J.H. Ware (1982), "Random-effects models for longitudinal data", Biometrics 38: 963–974.

Lancaster, T. (1997), "Exact structural inference in optimal job-search models", Journal of Business and Economic Statistics 15:165–179.

Leamer, E.E. (1978), Specification Searches: Ad Hoc Inference with Experimental Data (Wiley, New York).

Lenk, P.J. (1999), "Bayesian inference for semiparametric regression using a Fourier representation", Journal of the Royal Statistical Society B 61:863–879.

Li, K. (1998), "Bayesian inference in a simultaneous equation model with limited dependent variables", Journal of Econometrics 85:387–400.

Liu, C., D.B. Rubin and Y.N. Wu (1998), "Parameter expansion to accelerate EM: the PX-EM algorithm", Biometrika 85:755–770.

Liu, J.S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem", Journal of the American Statistical Association 89:958–966.

Liu, J.S., and R. Chen (1998), "Sequential Monte Carlo methods for dynamic systems", Journal of the American Statistical Association 93:1032–1044.

Liu, J.S., W.H. Wong and A. Kong (1994), "Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and data augmentation schemes", Biometrika 81:27–40.

Liu, J.S., W.H. Wong and A. Kong (1995), "Covariance structure and convergence rate of the Gibbs sampler with various scans", Journal of the Royal Statistical Society B 57:157–169.

Louis, T.A. (1982), "Finding the observed information matric when using the EM algorithm", Journal of the Royal Statistical Society B 44:226–232.

Mallick, B., and A.E. Gelfand (1994), "Generalized linear models with unknown link function", Biometrika 81:237–246.

Mallick, B., and A.E. Gelfand (1996), "Semiparametric errors-in-variables models: a Bayesian approach", Journal of Statistical Planning and Inference 52:307–321.

Manchanda, P., A. Ansari and S. Gupta (1999), "The "shopping basket": a model for multicategory purchase incidence decisions", Marketing Science 18:95–114.

Marinari, E., and G. Parisi (1992), “Simulated tempering: a new Monte Carlo scheme”, Europhysics Letters 19:451–458.

Martin, G. (1999), “US deficit sustainability: a new approach based on multiple endogenous breaks”, Journal of Applied Econometrics, in press.

McCulloch, R.E., and R. Tsay (1994), “Statistical analysis of macroeconomic time series via Markov switching models”, Journal of Time Series Analysis 15:523–539.

Meng, X.-L., and W.H. Wong (1996), “Simulating ratios of normalizing constants via a simple identity: a theoretical exploration”, Statistica Sinica 6:831–860.

Mengersen, K.L., and R.L. Tweedie (1996), “Rates of convergence of the Hastings and Metropolis algorithms”, Annals of Statistics 24:101–121.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), “Equations of state calculations by fast computing machines”, Journal of Chemical Physics 21:1087–1092.

Meyn, S.P., and R.L. Tweedie (1993), Markov chains and stochastic stability (Springer, London).

Meyn, S.P., and R.L. Tweedie (1994), “Computable bounds for convergence rates of Markov chains”, Annals of Applied Probability 4:981–1011.

Mira, A., and L. Tierney (1998), “On the use of auxiliary variables in Markov chain Monte Carlo methods”, Technical Report (University of Minnesota).

Montgomery, A.L., and P.E. Rossi (1999), “Estimating price elasticities with theory-based priors”, Journal of Marketing Research 36:413–423.

Muller, P., and D.R. Insua (1998), “Issues in Bayesian analysis of neural network models”, Neural Computation 10:749–770.

Muller, P., A. Erkanli and M. West (1996), “Curve fitting using multivariate normal mixtures”, Biometrika 83:63–79.

Nandram, B., and M.-H. Chen (1996), “Accelerating Gibbs sampler convergence in the generalized linear models via a reparameterization”, Journal of Statistical Computation and Simulation 54:129–144.

Newton, M.A., and A.E. Raftery (1994), “Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion)”, Journal of the Royal Statistical Society B 56:1–48.

Nummelin, E. (1984), General Irreducible Markov Chains and Non-Negative Operators (Cambridge University Press, Cambridge).

O’Hagan, A. (1994), Kendall’s Advanced Theory of Statistics, Vol. 2B, Bayesian Inference (Halsted Press, New York).

Paap, R., and H.K. van Dijk (1999), “Bayes estimates of Markov trends in possibly cointegrated series: An application to US consumption and income”, Manuscript (RIBES, Erasmus University).

Pastor, L., and R.F. Stambaugh (1999), “Costs of equity capital and model mispricing”, Journal of Finance 54:67–121.

Patz, R.J., and B.W. Junker (1999), “A straightforward approach to Markov chain Monte Carlo methods for item response models”, Journal of Education and Behavioral Statistics 24:146–178.

Percy, D.F. (1992), “Prediction for seemingly unrelated regressions”, Journal of the Royal Statistical Society B 54:243–252.

Pitt, M.K., and N. Shephard (1997), “Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models”, Journal of Time Series Analysis 20:63–85.

Pitt, M.K., and N. Shephard (1999), “Filtering via simulation: auxiliary particle filters”, Journal of the American Statistical Association 94:590–599.

Poirier, D.J. (1995), Intermediate Statistics and Econometrics: A Comparative Approach (MIT Press, Cambridge).

Polson, N.G. (1996), “Convergence of Markov chain Monte Carlo algorithms”, in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fifth Valencia International Conference on Bayesian Statistics (Oxford University Press, Oxford) 297–323,.

Propp, J.G., and D.B. Wilson (1996), “Exact sampling with coupled Markov chains and applications to statistical mechanics”, Random Structures and Algorithms 9:223–252.

Quinn, K.M., A.D. Martin and A.B. Whitford (1999), "Voter choice in multi-party democracies: a test of competing theories and models", American Journal of Political Science 43:1231–1247.

Raftery, A.E., and S.M. Lewis (1992), "How many iterations in the Gibbs sampler?" in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 763–774.

Raftery, A.E., A.D. Madigan and J.A. Hoeting (1997), "Bayesian model averaging for linear regression models", Journal of the American Statistical Association 92:179–191.

Richardson, S., and P.J. Green (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)", Journal of the Royal Statistical Society B 59:731–792.

Ripley, B. (1987), Stochastic Simulation (Wiley, New York).

Ritter, C., and M.A. Tanner (1992), "Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy-Gibbs Sampler", Journal of the American Statistical Association 87:861–868.

Robert, C.P. (1995), "Convergence control methods for Markov chain Monte Carlo algorithms", Statistical Science 10:231–253.

Robert, C.P., and G. Casella (1999), Monte Carlo Statistical Methods (Springer, New York).

Robert, C.P., G. Celeux and J. Diebolt (1993), "Bayesian estimation of hidden Markov models: a stochastic implementation", Statistics and Probability Letters 16:77–83.

Roberts, G.O., and J.S. Rosenthal (1999), "Convergence of slice sampler Markov chains", Journal of the Royal Statistical Society B 61:643–660.

Roberts, G.O., and S.K. Sahu (1997), "Updating schemes, correlation structure, blocking, and parametization for the Gibbs sampler", Journal of the Royal Statististical Society B 59:291–317.

Roberts, G.O., and A.F.M. Smith (1994), "Some simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms", Stochastic Processes and its Applications 49:207–216.

Roberts, G.O., and R.L. Tweedie (1996), "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms", Biometrika 83:95–110.

Rosenthal, J.S. (1995), "Minorization conditions and convergence rates for Markov chain Monte Carlo", Journal of the American Statistical Association 90:558–566.

Rubin, D.B. (1988), "Using the SIR algorithm to simulate posterior distributions", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 395–402.

Shephard, N. (1994), "Partial non-Gaussian state space", Biometrika 81:115–131.

Shephard, N. (1996), "Statistical aspects of ARCH and stochastic volatility", in: D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielson, eds., Time Series Models with Econometric, Finance and Other Applications (Chapman and Hall, London) 1–67.

Shively, T.S., R. Kohn and S. Wood (1999), "Variable selection and function estimation in additive nonparametric regression using a data-based prior", Journal of the American Statistical Association 94:777–794.

Smith, A. (1999), "Testing theories of strategic choice: the example of crisis escalation", American Journal of Political Science 43:1254–1283.

Smith, A.F.M., and G.O. Roberts (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", Journal of the Royal Statistical Society B 55:3–24.

Smith, M., and R. Kohn (1996), "Nonparametric regression using Bayesian variable selection", Journal of Econometrics 75:317–343.

So, M.K.P., K. Lam and W.K. Li (1998), "A stochastic volatility model with Markov switching", Journal of Business and Economic Statistics 16:244–253.

Stephens, D.A. (1994), "Bayesian retrospective multiple-changepoint identification", Applied Statistics 43:159–178.

Tanner, M.A. (1996), Tools for Statistical Inference, 3rd. edition (Springer, New York).

Tanner, M.A., and W.H. Wong (1987), "The calculation of posterior distributions by data augmentation", Journal of the American Statistical Association 82:528–549.

Taylor, S.J. (1994), "Modelling stochastic volatility", Mathematical Finance 4:183–204.

Tierney, L. (1994), “Markov chains for exploring posterior distributions (with discussion)”, Annals of Statistics 22:1701–1762.

Tierney, L., and J. Kadane (1986), “Accurate approximations for posterior moments and marginal densities”, Journal of the American Statistical Association 81:82–86.

Tsionas, E.G. (1999), “Monte Carlo inference in econometric models with symmetric stable disturbances”, Journal of Econometrics 88:365–401.

Verdinelli, I., and L. Wasserman (1995), “Computing Bayes factors using a generalization of the Savge–Dickey density ratio”, Journal of the American Statistical Association 90:614–618.

Wakefield, J.C., A.F.M. Smith, A. Racine-Poon and A.E. Gelfand (1994), “Bayesian analysis of linear and non-linear population models by using the Gibbs sampler”, Applied Statistics 43:201–221.

Waller, L.A., B.P. Carlin, H. Xia and A.E. Gelfand (1997), “Hierarchical spatio-temporal mapping of disease rates”, Journal of the American Statistical Association 92:607–617.

Wei, G.C.G., and M.A. Tanner (1990a), “Posterior computations for censored regression data”, Journal of the American Statistical Association 85:829–839.

Wei, G.C.G., and M.A. Tanner (1990b), “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithm”, Journal of the American Statistical Association 85:699–704.

West, M., R. Prado and A.D. Krystal (1999), “Evaluation and comparison of EEG traces: latent structure in nonstationary time series”, Journal of the American Statistical Association 94:375–387.

Wu, C.F.J. (1983), “On the convergence properties of the EM algorithm”, Annals of Statistics 11:95–103.

Young, M.R., W.S. DeSarbo and V.G. Morwitz (1998), “The stochastic modeling of purchase intentions and behavior”, Management Science 44:188–202.

Zeger, S.L., and M.R. Karim (1991), “Generalized linear models with random effects: a Gibbs sampling approach”, Journal of the American Statistical Association 86:79–86.

Zellner, A. (1971), Introduction to Bayesian Inference in Econometrics (Wiley, New York).

Zellner, A., and C. Min (1995), “Gibbs sampler convergence criteria”, Journal of the American Statistical Association 90:921–927.