Markov Chain Monte Carlo Simulation Methods in Econometrics*

Siddhartha Chib
Washington University, St. Louis MO, USA
E-Mail: chib@simon.wustl.edu

Edward Greenberg
Washington University, St. Louis MO, USA
E-Mail: edg@wuecona.wustl.edu

February, 1995

Abstract

We present several Markov chain Monte Carlo simulation methods that have been widely used in recent years in econometrics and statistics. Among these is the Gibbs sampler, which has been of particular interest to econometricians. Although the paper summarizes some of the relevant theoretical literature, its emphasis is on the presentation and explanation of applications to important models that are studied in econometrics. We include a discussion of some implementation issues, the use of the methods in connection with the EM algorithm, and how the methods can be helpful in model specification questions. Many of the applications of these methods are of particular interest to Bayesians, but we also point out ways in which frequentist statisticians may find the techniques useful.

Keywords: Markov chain Monte Carlo, Gibbs sampler, data augmentation, Metropolis-Hastings algorithm, Monte Carlo EM, simulation.

JEL Classification: C11, C15, C20.

*First draft: September 28, 1993; second draft : July 1994. We acknowledge the very helpful comments of three anonymous referees.

1 Introduction

In this paper we explain Markov chain Monte Carlo (MCMC) methods in some detail and illustrate their application to problems in econometrics. These procedures, which enable the simulation of a large set of multivariate density functions, have revolutionized the practice of Bayesian statistics and appear to be applicable to virtually all parametric econometric models regardless of their complexity. Our purpose is to explain how these methods work, both in theory and in practical applications. Since many problems in Bayesian statistics (such as the computation of posterior moments and marginal density functions) can be solved by simulating the posterior distribution, we emphasize Bayesian applications, but these tools are also valuable in frequentist inference, where they can be used to explore the likelihood surface and to find modal estimates or maximum likelihood estimates with diffuse priors.[1]

An MCMC method is a simulation technique that generates a sample (multiple observations) from the *target distribution* in the following way: The transition probability of a Markov process is specified with the property that its limiting invariant distribution is the target distribution. The Markov chain is then iterated a large number of times in a computer-generated Monte Carlo simulation, and the output, after a transient phase and under various sets of conditions, is a sample from the target distribution. The first such method, due to Metropolis et al. (1953) and Hastings (1970), is known as the Metropolis-Hastings (MH) algorithm. In this algorithm, the next value of the Markov chain is generated from a proposal density and then accepted or rejected according to the density at the candidate point relative to the density at the current point. Another MCMC method is the Gibbs sampling algorithm, introduced by Geman and Geman (1984) and extended by Tanner and Wong (1987) and Gelfand and Smith (1990), in which the next draw is obtained by sampling sub-components of a random vector from a sequence of full conditional distributions. Other MCMC methods include hybrid versions of Gibbs sampling and MH sampling [Tierney (1994)] and stochastic versions of the EM algorithm [Celeux and Diebolt (1985)].

[1]Smith and Roberts (1993) and Tanner (1993) contain valuable surveys of some of the same ideas but are addressed to a general statistical audience. We emphasize econometric applications in the present paper.

The generated sample can be used to summarize the target density by graphical means, by exploratory data analysis methods, and by other means.[2] For example, expectations of integrable functions w.r.t. the target density can be estimated by taking a sample average of the function over the simulated draws. Under general conditions the ergodicity of the Markov chain guarantees that this estimate is simulation consistent and satisfies a central limit theorem as the length of the simulation goes to infinity. The MCMC strategy has proved extremely useful in statistical applications, much more so than traditional independent sampling methods, which by and large are difficult to apply in complex, high-dimensional problems. MCMC methods can be applied without knowledge of the normalizing constant of the target density, which is very important in the Bayesian context where the normalizing constant of the target (posterior) density is almost never known. In addition, it is often possible to tailor an MCMC scheme such that models with an intractable likelihood function can be simulated. This is usually achieved, particularly with Gibbs sampling, by the device of "data augmentation" (the strategy of enlarging the parameter space to include missing data or latent variables). Applications of this idea include models with structural breaks at random points [Carlin, Gelfand, and Smith (1992)]; models with censored and discrete data [Chib (1992a) and Albert and Chib (1993a,c)]; models with Markov switching [Albert and Chib (1993b), Chib (1993b), and McCulloch and Tsay (1993)]; models with parameter constraints [Gelfand et al. (1992)], and many others.[3]

The remainder of the paper proceeds as follows. In the second section we review the theory behind generating samples by Markov chain Monte Carlo and discuss implementation issues for the Gibbs and MH algorithms. In the third section these methods are applied to models widely used in econometrics: the seemingly unrelated regression model, the tobit censored regression model, binary and panel probit models, random coefficient model, linear regression with AR(p) errors, and state-space models. In Section 4 we explain how output from an MCMC simulation can be used for statistical inference, and Section 5 contains

[2]This feature is shared by non-MCMC methods (such as those based on rejection sampling) that are designed to sample a density [Rubinstein (1981) and Ripley (1987)].

[3]By contrast, Monte Carlo methods with importance sampling [Kloek and van Dijk (1978), Geweke (1989), Koop (1994)] are difficult to apply in these situations due to the complexity of the likelihood function. In addition, the need to find a suitable importance sampling function is a limitation in high-dimensional problems.

conclusions.

2 Markov chain Monte Carlo sampling methods

We begin the section with an informal presentation of some relevant material from Markov chain theory and then discuss the Gibbs sampling algorithm and the MH algorithm. A much more detailed discussion of Markov theory is provided by Nummelin (1984), Meyn and Tweedie (1993), and Tierney (1994).

2.1 Markov chains

A Markov chain is a collection of random variables (or vectors) $\Phi = \{\Phi_i : i \in T\}$ 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*

$$P(x, A) \equiv \Pr(\Phi_{i+1} \in A | \Phi_i = x, \Phi_j, j < i), \quad x \in \Omega, \quad A \subset \Omega.$$

The assumption that the probability distribution of the next item in the sequence, given the current and the past states, depends only on the current state is the Markov property. Suppose that the transition kernel, for some function $p(x, y) : \Omega \times \Omega \to \Re^+$, is expressed as

$$P(x, dy) = p(x, y)v(dy) + r(x)\delta_x(dy)\,, \tag{1}$$

where $p(x, x) = 0$, $\delta_x(dy) = 1$ if $x \in dy$ and 0 otherwise, $r(x) = 1 - \int_\Omega p(x, y)\, v(dy)$, and ν denote a σ-finite measure on the Borel σ-algebra on Ω, then transitions from x to y occur according to $p(x, y)$ and transitions from x to x occur with probabiltiy $r(x)$. In the case that $r(x) = 0$, the integral of $p(x, y)$ over y is 1 and the function $p(x, y)$ may be referred to as the transition density of the chain. Note that

$$P(x, A) = \int_A P(x, dy)\,. \tag{2}$$

The transition kernel is thus the distribution of Φ_{i+1} given that $\Phi_i = x$. The n-th step ahead transition kernel is given by

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

where $P^{(1)}(x, dy) = P(x, dy)$. Under certain conditions that are discussed below it can be shown that the nth iterate of the transition kernel (as $n \to \infty$) converges to the invariant distribution, π^*. The invariant distribution satisfies

$$\pi^*(dy) = \int_\Omega P(x, dy)\pi(x)v(\,dx) \tag{3}$$

where π is the density of π^* with respect to the measure ν (thus, $\pi^*(dy) = \pi(y)v(dy)$). The invariance condition states that if Φ_i is distributed according to π^*, then so are all subsequent elements of the chain. A chain is said to be *reversible* if the function $p(x, y)$ in (1) satisfies

$$\pi(x)p(x, y) = \pi(y)p(y, x)\,. \tag{4}$$

A reversible chain has π^* as an invariant distribution [see Tierney (1994) or Chib and Greenberg (1994)]. An important notion is π^*-*irreducibility.* A Markov chain is said to be π^*-irreducible if for every $x \in \Omega$, $\pi^*(A) > 0 \Rightarrow P(\Phi_i \in A|\Phi_0 = x) > 0$ for some $i \geq 1$. This condition states that all sets with positive probability under π^* can be reached from any starting point in Ω. 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 $p \geq 2$ such that $P(\Phi^i \in D_{i \bmod(p)}|\Phi_0 \in D_0) = 1$ for all i.

These definitions allow us to state the following (ergodicity) result [see Tierney (1994)], which forms the basis for Markov chain Monte Carlo methods.

Proposition 1 *If $P(\cdot, \cdot)$ is π^*-irreducible and has invariant distribution π^*, then π^* is the unique invariant distribution of $P(\cdot, \cdot)$. If $P(\cdot, \cdot)$ is also aperiodic, then for π^*-almost every $x \in \Omega$, and all sets A*

1. $|P^m(x, A) - \pi^*(A)| \to 0$ *as* $m \to \infty$*;*

2. *for all π^*-integrable real-valued functions h,*

$$\frac{1}{m}\sum_{i=1}^{m} h(\Phi_i) \to \int h(x)\pi(x)\nu(dx) \text{ as } m \to \infty, \ a.s.$$

The first part of this theorem tells us that (under the stated conditions) the probability density of the mth iterate of the Markov chain is, for large m, very close to its unique, invariant density. This means that if drawings are made from $P^m(x, dy)$, then for large m the probability distribution of the drawings is the invariant distribution, regardless of the initial value. The second part states that averages of functions evaluated at sample values (*ergodic averages)* converge (as $m \to \infty$, almost surely) to their expected value under the target density. Sufficient conditions for π^*-irreducibility and aperiodicity are presented below for the Gibbs and MH algorithms.

2.2 Gibbs sampling

As noted above, the objective in MCMC simulation is to find a transition density that has the target density as its invariant distribution. One strategy is the Gibbs sampling algorithm, in which the random vector is partitioned into several blocks and the transition density is defined as the product of the set of full conditional densities (the conditional density of each block given the data and the remaining parameters).[4] The next item in the Markov chain is obtained by successively sampling the full conditional densities, given the most recent values of the conditioning parameters. Casella and George (1992) provide an elementary introduction. The value of this algorithm arises from the fact that in many applications the full conditional densities (perhaps after the parameter space has been augmented by latent data) take convenient forms and can be simulated even though the target density is intractable.

Suppose $\pi(x)$, $x \in \mathcal{S} \subseteq \Re^p$, is the (perhaps unnormalized) target density that we wish to sample. For some decomposition of x into $x_1, \ldots, x_d$, let the full conditional density of the kth block be denoted by $\pi(x_k|x_{-k}) \equiv \pi(x_k|x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$.[5] Then the Gibbs sampling algorithm is defined by the following iterations:

1. Specify starting values $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$ and set $i = 0$.

[4] For frequentist statisticians these distributions can be regarded as proportional to the conditional likelihood functions of each parameter, where the conditioning is on values of all remaining parameters.

[5] Note that the full conditional density $\pi(x_k|\, x_{-k})$ is proportional to the joint density $\pi(x)$. Deriving these is often straightforward.

2. Simulate

$$\begin{array}{lll} x_1^{(i+1)} & \text{from} & \pi(x_1|x_2^{(i)}, x_3^{(i)}, ..., x_d^{(i)}) \\ x_2^{(i+1)} & \text{from} & \pi(x_2|x_1^{(i+1)}, x_3^{(i)}, ..., x_d^{(i)}) \\ x_3^{(i+1)} & \text{from} & \pi(x_3|x_1^{(i+1)}, x_2^{(i+1)}, x_4^{(i)}, ..., x_d^{(i)}) \\ \vdots & & \\ x_d^{(i+1)} & \text{from} & \pi(x_d|x_1^{(i+1)}, x_2^{(i+1)}, ..., x_{d-1}^{(i+1)}). \end{array}$$

3. Set $i = i + 1$ and go to step 2.

This algorithm thus provides the next item of the Markov chain $x^{(i+1)}$ by simulating each of the full conditional densities, where the conditioning elements are revised during a cycle. Since transitions to the same point occur with probability zero, $r(x) = 0$ and transitions of the chain from $x \equiv x^{(i)}$ to $y \equiv x^{(i+1)}$ (two distinct points) take place according to the transition density

$$p_G(x, y) = \prod_{k=1}^{d} \pi(y_k|y_1, \ldots, y_{k-1}, x_{k+1}, \ldots, x_d). \tag{5}$$

It is not difficult to check that this transition density satisfies (3): If ν is Lebesgue measure, $\int p_G(x, y)\, \pi(x)\, d(x)$ is

$$\int \prod_{k=1}^{d} \frac{\pi(y_k|y_1, \ldots, y_{k-1})\pi(x_{k+1}, \ldots, x_d|y_1, \ldots, y_k)}{\pi(x_{k+1}, \ldots, x_d|y_1, \ldots, y_{k-1})} \pi(x_1|x_2, \ldots, x_d)\, \pi(x_2, \ldots, x_d)\, dx$$

by applying Bayes theorem to each term in the transition kernel and writing $\pi(x)$ as $\pi(x_1|x_2, \ldots, x_d)\pi(x_2, \ldots, x_d)$. The calculation is completed by noting that (i) the terms $\pi(y_k|y_1, \ldots, y_{k-1})$ are independent of x, so they factor out as $\prod_{k=1}^{d} \pi(y_k|y_1, \ldots, y_{k-1})$ to give $\pi(y)$; (ii) the integral over x_1 is 1; (iii) the term $\pi(x_2, \ldots, x_d)$ cancels with the denominator for $k = 1$; and (iv) cancellation by telescoping occurs since the numerator element in term $k - 1$ is $p(x_{k+1}, \ldots, x_d|y_1, \ldots, y_{k-1})$, which cancels with the denominator in term k.

We now turn to some issues that arise in implementing the Gibbs sampling algorithm. First, in designing the blocks, highly correlated components should be grouped together; otherwise the Markov chain is likely to display autocorrelations that decay slowly, resulting in slow convergence to the target density [see Liu et al. (1994) and Section 3.4]. Second, a tractable full conditional structure can sometimes be obtained by introducing latent or missing data into the definition of x. The idea of adding variables to the sampler, known as "data augmentation," was introduced by Tanner and Wong (1987) and is illustrated in several of the examples in Section 3.[6] Finally, if some of the full conditional densities are difficult to sample by traditional means (by the method of rejection sampling or by a known generator, for example), that density can be sampled by the MH algorithm [Müller (1991)] or a method that generates independent samples [Gilks and Wild (1992)].

Several sets of sufficient conditions ensure that the Markov chain generated by the Gibbs sampler satisfies the conditions of Proposition 1. A convenient set is due to Roberts and Smith (1994, Theorem 2) [see also Chan (1993)].

Proposition 2 *Suppose that (i)* $\pi(x) > 0$ *implies there exists an open neighborhood* N_x *containing* x *and* $\epsilon > 0$ *such that, for all* $y \in N_x$, $\pi(y) \geq \epsilon > 0$; *(ii)* $\int \pi(x)\, dx_k$ *is bounded for all* k *and all* y *in a open neighborhood of* x; *and (iii) the support of* x *is arc connected. Then* $p_G(x, y)$ *satisfies the conditions of Proposition 1.*

The intuition for these conditions (and their connection to π-irreducibility and aperiodicity) should be noted. The conditions ensure that each full conditional density is well defined and that the support of the density is not separated into disjoint regions so that once the chain moves into one such region it never leaves it. Although these are only sufficient conditions for the convergence of the Gibbs sampler, the conditions are extremely weak and are satisfied in most econometric applications.

2.3 Metropolis-Hastings algorithm

The MH algorithm is another powerful MCMC method that can be used to sample an intractable distribution $\pi^*(.)$. A sequence of draws from that algorithm is obtained as follows: Given that the latest drawing has yielded the value x, the next value in the sequence is generated by drawing a value y from a *candidate generating density* $q(x, y)$ (also called a *proposal density*). The y thus generated is accepted with probability $\alpha(x, y)$, where

$$\alpha(x,y) = \begin{cases} \min\left[\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right] & \text{if } \pi(x)q(x,y) > 0; \\ 1 & \text{otherwise}. \end{cases}$$

[6] The idea of data augmentation also appears in maximum likelihood estimation of missing data models by the EM algorithm [Dempster et al. (1977)].

If the candidate is rejected, the next sampled value is taken to be the current value.

Two important points should be noted. First, the calculation of $\alpha(x, y)$ does not require knowledge of the normalizing constant of $\pi(\cdot)$. Second, if the proposal density is symmetric, i.e., $q(x, y) = q(y, x)$, then the acceptance probability reduces to $\pi(y)/\pi(x)$, which is the original formulation of Metropolis et al. (1953).

To understand the basis for this algorithm first note that the transition kernel of this Markov chain is given by

$$P_{MH}(x, dy) = q(x, y)\alpha(x, y)\, dy + \left[1 - \int_\Omega q(x, y)\alpha(x, y)\, dy\right] \delta_x(dy)\,, \tag{6}$$

which states that transitions from x to y $(y \neq x)$ are made according to

$$p_{MH}(x, y) \equiv q(x, y)\alpha(x, y), \;\; x \neq y,$$

The function $p_{MH}(x, y)$ satisfies the reversibility condition (4). To see this consider the case where $\alpha(x, y) < 1$ (which implies that $\alpha(y, x) = 1$). Then, $\pi(x)\, p_{MH}(x, y) \equiv \pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)$, which is equal to $\pi(y)\, p_{MH}(y, x)$ as was to be checked. Thus π^* is an invariant distribution for $P_{MH}(x, dy)$.

A useful sufficient condition for convergence of chains generated by the MH algorithm can be based on Lemma 1.2 of Mengersen and Tweedie (1993):

Proposition 3 *If $\pi(x)$ and $q(x, y)$ are positive and continuous for all (x, y) then $p_M(x, y)$ satisfies the conditions of Proposition 1.*

Further discussion of sufficient conditions may be found in Smith and Roberts (1993) and Tierney (1994). While Proposition 2 implies convergence, it is not informative about the speed of convergence. This aspect of the theory is under active investigation, the main focus being on geometric ergodicity. Some results may be found in the articles mentioned earlier in this paragraph and in Roberts and Tweedie (1994).

We now turn briefly to the question of specifying the proposal density that drives the MH algorithm. Several generic choices are discussed by Tierney (1994) and Chib and Greenberg (1994). One possibility is to let the proposal density take the form $q(x, y) = q(y - x)$, as, for example, when the candidate is drawn from a multivariate normal density centered at the current value x. This is referred to as the *random walk based MH chain.* Another possibility, suggested by Hastings (1970) and called the *independence MH chain* by Tierney (1994), is specified by letting $q(x, y) = q(y)$, which implies that the density $q(x, y)$ is independent of x. This proposal density can be centered at the posterior mode (or some other suitable value) with the form of q adjusted to ensure that the *acceptance rate* (the proportion of times a candidate value is accepted) is reasonable. What is reasonable depends on the context, but it is important that the proposal density should be chosen so that the chain travels over the support of the target density. This may fail to occur, with a consequent undersampling of low probability regions, if the chain is near the mode and if candidates are drawn too close to the current value.

It is worth emphasizing that once a proposal density is specified, the MH algorithm is a straightforward method of simulating virtually any target density, including an intractable full conditional density that may arise in implementing the Gibbs sampling algorithm. It is easy to show that this combination of Markov chains (Metropolis-within-Gibbs) is itself a Markov chain with the correct invariant distribution. Specifically, consider the case of two blocks and suppose that the full conditional density $\pi(y_1|x_2)$ can be sampled directly but that $\pi(y_2|y_1)$ requires use of the MH algorithm. Under the assumption of Lebesque measure, the transition kernel is then the product of $\pi(y_1|x_2)dy_1$ and the transition kernel of the MH step, which is given by $p_{MH}(x_2, y_2|y_1)dy_2 + r(x_2|y_1)\delta_{x_2}(dy_2)$. Then

$$
\begin{aligned}
& \int\!\!\int \pi(x_1, x_2)\pi(y_1|x_2)\, dy_1[p_{MH}(x_2, y_2|y_1)dy_2 + r(x_2|y_1)\delta_{\,x_2}(dy_2)]\, dx_1\, dx_2 \\
= \quad & \int \pi(x_2)\pi(y_1|x_2)dy_1[p_{MH}(x_2, y_2|y_1)dy_2 + r(x_2|y_1)\delta_{\,x_2}(dy_2)]\, dx_2 \\
= \quad & \pi(y_1)dy_1 \int \pi(x_2|y_1)p_{MH}(x_2, y_2|y_1)dy_2\, dx_2 + \pi(y_2)\pi(y_1|y_2)dy_1 dy_2 r(y_2|y_1) \\
= \quad & \pi(y_1)\pi(y_2|y_1)dy_1 dy_2 \int p_{MH}(y_2, x_2|y_1)\, dx_2 + \pi(y_1, y_2)dy_1 dy_2 r(y_2|y_1) \\
= \quad & \pi(y_1, y_2)dy_1 dy_2(1 - r(y_2|y_1)) + \pi(y_1, y_2)dy_1 dy_2 r(y_2|y_1),
\end{aligned}
$$

and invariance is confirmed. The fourth line above follows from the reversibility of the MH step $\pi(x_2|\, y_1)p_{MH}(x_2, y_2|\, y_1) = \pi(y_2|y_1)p_{MH}(y_2, x_2|y_1)$. It is therefore not necessary to stop the Gibbs sampler to iterate the MH algorithm when an intractable full conditional density is encountered; one value is generated from the MH procedure, followed by the next Gibbs step.

2.4 Implementation issues

Single run vs multiple run sampling: The literature has suggested two methods for generating a sample from an MCMC algorithm—the single-chain and the multiple-chain. In the multiple chain method a starting value is chosen and a sequence is generated from $p(x^{(i-1)}, x^{(i)})$. After a transient phase of N_0 drawings, the $N_0 + 1$ drawing is regarded as a sample from $\pi(\cdot)$. A new starting value is then chosen, and the process is repeated. This method generates an independent sample at the cost of discarding N_0 drawings in each cycle. In the single-run method the sequence $\{x^{(N_0+1)}, x^{(N_0+2)}, \ldots, x^{(N_0+M)}\}$ is regarded as a sample of size M from $\pi(\cdot)$. The resulting sample is correlated because each drawing depends upon the previous draw (the Markov property). The sample is nevertheless useful because the sequence converges to the invariant distribution. The Markov nature of the sample usually introduces strong positive correlation between parameter values at successive iterations, but the correlation often dissipates quickly so that it is close to zero between the iterate at t and $t + n_1$, say, for moderate n_1. In that case an approximately random sample can be found by including in the sample every n_1th item in the sequence after the transient phase has ended.

Detection of convergence: Because the length of the transient phase seems to be model and data dependent, the question of convergence requires considerable care. If the target density being simulated is "well behaved" (as it is in many standard econometric models), then the simulated Markov chain usually mixes rapidly and the serial correlations die out quickly. But with weak identifiability of the parameters and/or multiple modes the chain can be poorly behaved.[7] Many proposals have been made to shed light on these problems. One class of approaches [exemplified by Ritter and Tanner (1992), Gelman and Rubin (1992),

[7]The multiple modes case can be quite deceptive. The chain may appear to mix well but may actually be trapped in a sub-region of the support. This example indicates the importance of understanding by analytical means the target density being simulated and then devising an algorithm to achieve a chain with desirable properties (perhaps by combining MCMC schemes, by abandoning one MCMC algorithm in favor of another, or by using multiple-chain sampling).

Geweke (1992), and Zellner and Min (1993)] attempts to analyze the observed output to determine whether the chain has converged. The Gelman and Rubin approach, which is based on multiple-chain sampling from dispersed starting values, compares the within and between variation in the sampled values. The Ritter and Tanner approach, which requires a single run, monitors the ratio of the target density (up to a normalizing constant) and the current estimate of the target density; stability of the ratio indicates that the chain has converged. Another type of approach [for example, Raftery and Lewis (1992) and Polson (1992)] attempts to produce estimates of the burn-in time *prior to sampling* by analyzing the rate of convergence of the Markov chain to the target density. Considerable work continues to be done in this important area, but no single approach appears to be adequate for all problems.

3 Examples

We now show how the MCMC simulation approach can be applied to a wide variety of econometric models, starting with a simple example in which the Gibbs sampler can be applied without data augmentation and where simulation is from standard distributions only. The later examples require more of the methods described above. Our objectives are to present the logic of the method and to help the reader understand how to apply the method in other situations.

Before presenting the examples, we introduce the assumptions for prior densities that are used throughout this section: The vector β follows a $\mathcal{N}_k(\beta_0, B_0^{-1})$, the variance σ^2 is distributed as inverted gamma $IG(\frac{\nu_0}{2}, \frac{\delta_0}{2})$, and the precision matrix Ω^{-1} follows a Wishart $\mathcal{W}_p(\rho_0, R_0)$ distribution. Hyperparameters of the prior densities, subscripted by a 0, are assumed to be known. A density or distribution function is denoted by $[\cdot]$, a conditional density or distribution by $[\cdot|\cdot]$, and $\overset{d}{=}$ denotes equality in distribution.

3.1 The seemingly unrelated regression model

Our first example is the seemingly unrelated regression (SUR) model, which is widely employed in econometrics. Under the assumption of normally distributed errors, the observed data y_{it} are generated by

$$y_{it} = x_{it}'\beta_i + \epsilon_{it}, \quad \epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{pt})' \sim \text{iid}\mathcal{N}_p(0, \Omega), \quad 1 \le i \le p, \quad 1 \le t \le n,$$

where $\beta_i : k_i \times 1$ and Ω is a positive definite matrix. By stacking observations for each time period, we rewrite the model in vector form as $y_t = X_t\beta + \epsilon_t$, where $y_t = (y_{1t}, \ldots, y_{pt})'$, $X_t = \text{diag}(x_{1t}', \ldots, x_{pt}')$, $\beta = (\beta_1', \ldots, \beta_p')' : k \times 1$, and $k = \sum_i k_i$. We obtain the single equation Gaussian regression model when $p = 1$. It is well known that the maximum likelihood estimators for a sample of data $Y_n = (y_1, \ldots, y_n)$ can be obtained only through an iterative procedure and that the finite sample distribution of these estimators is intractable. In contrast, the Gibbs sampling algorithm provides an exact, small sample Bayesian analysis for this model [Percy (1992) and Chib and Greenberg (1993b)].

Suppose that prior information about (β, Ω^{-1}) is represented by the density $\pi(\beta)\pi(\Omega^{-1})$, where we are assuming that β and Ω^{-1} (the precision matrix) are independent. Then the posterior density of the parameters (proportional to the product of the prior density and the likelihood function) is given by

$$\pi(\beta)\pi(\Omega^{-1}) \times |\Omega^{-1}|^{n/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{n}(y_t - X_t\beta)'\Omega^{-1}(y_t - X_t\beta)\right].$$

This is the target density (with unknown normalizing constant) that must be simulated. Now note that if β and Ω^{-1} are treated as two blocks of parameters, the full conditional densities, $\beta|Y_n, \Omega^{-1}$ and $\Omega^{-1}|Y_n, \beta$ are easy to simulate. In particular, under the priors mentioned above,

$$\beta\;|Y_n, \Omega^{-1} \sim \mathcal{N}_k(\hat{\beta}, B_n^{-1}) \quad \text{and} \quad \Omega^{-1}|Y_n, \beta \sim \mathcal{W}_p(\nu_0 + n, R_n),$$

where $\hat{\beta} = B_n^{-1}(B_0\beta_0 + \sum_{t=1}^{n} X_t'\Omega^{-1}y_t)$, $B_n = (B_0 + \sum_{t=1}^{n} X_t'\Omega^{-1}X_t)$, and $R_n = [R_0^{-1} + \sum_{t=1}^{n}(y_t - X_t\beta)(y_t - X_t\beta)']^{-1}$. It is not difficult to verify the sufficient conditions mentioned in Proposition 2. Therefore, simulating these two distributions by the Gibbs algorithm yields a sample $\{\beta^{(i)}, \Omega^{-1(i)}\}$ such that $\beta^{(i)}$ is distributed according to the marginal density $\pi(\beta|Y_n)$, $\Omega^{-1(i)} \sim \pi(\Omega^{-1}|Y_n)$, and $(\beta^{(i)}, \Omega^{-1(i)})$ is distributed according to the target (joint) density.[8] It should be noted that the sample of draws is obtained without an importance sampling function or the evaluation of the likelihood function.

[8]The Wishart distribution can be simulated by the Bartlett decomposition: If $W \sim \mathcal{W}_p(\nu, G)$, then

3.2 Tobit and probit regression models

In the previous example the Gibbs sampler was applied directly to the parameters of the model. In other situations a tractable set of full conditional distributions can be obtained only by enlarging the parameter space with latent data, as we illustrate next for the tobit and probit models. Interestingly, while the parameter space over which the sampler is defined is extremely large (in the case of the probit model it is larger than the sample size), the number of blocks in the simulation is quite small (three in the tobit model and two in the binary probit model).

Consider the censored regression model of Tobin (1958), in which the observation y_i is generated by

$$z_i \sim \mathcal{N}(x_i'\beta, \sigma^2) \quad \text{and} \quad y_i = \max(0, z_i), \quad 1 \le i \le n.$$

Given a set of n independent observations, the likelihood function for β and σ^2 is

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

where C is the set of censored observations and Φ is the c.d.f. of the standard normal random variable. Clearly, this function (after multiplication by the prior density) is difficult to simplify for use in the Gibbs sampling algorithm. Chib (1992a) shows (in one of the first applications of Gibbs sampling in econometrics) that matters are simplified enormously if the parameter space is augmented by the latent data corresponding to the censored observations.

To see why, suppose we have available the vector $z = (z_i)$, $i \in C$. Let y_z 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 consider applying the Gibbs sampling algorithm with blocks β, σ^2, and z with the respective full conditional densities $[\beta | Y_n, z, \sigma^2]$, $[\sigma^2 | Y_n, z, \beta]$, and $[z | Y_n, \beta, \sigma^2]$. These distributions are all tractable and the Gibbs simulation is readily applied. The first two distributions reduce to

$$\beta | y_z, \sigma^2 \sim \mathcal{N}_k(\hat{\beta}, (B_0 + \sigma^{-2} X'X)^{-1}) \quad \text{and} \quad \sigma^2 | y_z, \beta \sim \mathcal{IG}(\frac{\nu_0 + n}{2}, \frac{\delta_0 + \delta_n}{2}), \tag{7}$$

$W \stackrel{d}{=} LTT'L'$, where $T = (t_{ij})$ is a lower triangular matrix with $t_{ii} \sim \chi^2_{v-i+1}$ and $t_{ij} \sim \mathcal{N}(0,1)$, and L is obtained from the Choleski factorization $LL' = G$.

where $X = (x_1, \ldots, x_n)'$, $\hat{\beta} = (B_0 + \sigma^{-2}X'X)^{-1}(B_0\beta_0 + \sigma^{-2}X'y_z)$, and $\delta_n = (y_z - X\beta)'(y_z - X\beta)$, while the full conditional distribution of the latent data simplifies into the product of n independent distributions, $[z|Y_n, \beta, \sigma^2] = \prod_{i \in C}[z_i|y_i = 0, \beta, \sigma^2]$, where

$$z_i|y_i = 0, \beta, \sigma^2 \sim \mathcal{TN}_{(-\infty,0]}(x_i'\beta, \sigma^2), \quad i \in C,$$

a truncated normal distribution with support $(-\infty, 0]$.[9] The simplification to conditional independence observed in this case (for example, the distributions of β and σ^2 are independent of the censored data given the latent data) usually occurs with data augmentation, which explains why data augmentation is such a useful tool [Morris (1987)].

The value of data augmentation is also clear in the probit model, where we are given n independent observations $Y_n = \{y_i\}$, each y_i being distributed Bernoulli with $\Pr(y_i = 1) = \Phi(x_i'\beta)$. For this model and many others in this class, Albert and Chib (1993a) develop a simple and powerful approach that introduces latent Gaussian data as additional unknown parameters in a Gibbs sampling algorithm. They exploit the fact that the specification

$$z_i = x_i'\beta + u_i, \quad u_i \sim \text{iid}\mathcal{N}(0,1), \quad \text{and} \quad y_i = I[z_i > 0] \tag{8}$$

produces the probit model. The Gibbs sampling algorithm (with data augmentation) is now defined through the full conditional distributions

$$[\beta|Y_n, Z_n] \stackrel{d}{=} [\beta|Z_n] \quad \text{and} \quad [Z_n|Y_n, \beta] \stackrel{d}{=} \prod_{i=1}^{n}[z_i|y_i, \beta].$$

The full conditional distribution of β has the same form as (7) with y_z replaced by Z_n and $\sigma^2 = 1$. The full conditional $[Z_n|Y_n, \beta]$, which factors into the product of independent terms, depends on whether $y_i = 1$ or $y_i = 0$. From (8) we have $z_i \leq 0$ if $y_i = 0$ and $z_i > 0$ if $y_i = 1$. Thus,

$$\begin{aligned} z_i|y_i = 0, \beta &\sim \mathcal{TN}_{(-\infty,0]}(x_i'\beta, 1) \text{ and} \\ z_i|y_i = 1, \beta &\sim \mathcal{TN}_{(0,\infty)}(x_i'\beta, 1), \quad 1 \leq i \leq n. \end{aligned}$$

This MCMC algorithm can be easily modified to estimate a model with an independent student-t link function with ν degrees of freedom [see Albert and Chib (1993a)]. From

[9] To simulate from $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$, we first simulate a uniform random variate U and then obtain the required draw as $\mu + \sigma\Phi^{-1}\{p_1 + U(p_2 - p_1)\}$, where Φ^{-1} is the inverse c.d.f of the normal distribution, $p_1 = \Phi[(a - \mu)/\sigma]$ and $p_2 = \Phi[(b - \mu)/\sigma]$. Alternatively, the method of Geweke (1991) can be used to sample this distribution.

the result that the t-distribution is a scale mixture of normals with mixing distribution $\text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ it is possible to further augment the parameter space by these gamma variables, one for each observation. The full conditionals are again tractable [see also Carlin and Polson (1991) and Geweke (1993a) for use of this idea in linear regression]. Albert and Chib (1993a) also let ν be unknown, which leads to a general robustification of the probit model.

3.3 Random coefficient panel model

We next consider another multiple equation model that is frequently applied to panel data. In this model the data generating equation for the ith observation unit, usually an individual, household, or firm, over the T time periods is given by

$$y_i = X_i b_i + \epsilon_i, \quad \epsilon_i|\sigma^2 \sim \text{iid}\mathcal{N}_T(0, \sigma^2 I_T), \quad 1 \leq i \leq n,$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (x_{i1}, \ldots, x_{iT})'$, and the individual-specific coefficients are assumed to follow the distribution $b_i|\beta, \Omega \sim \mathcal{N}_k(\beta, \Omega)$. A sampling theory discussion of models similar to this may be found in Hsiao (1986).

The first point to note in a Bayesian approach to this model is that a tractable full conditional structure is not available from the likelihood function (obtained by integrating out the random effects). It is, therefore, important to include $\{b_i\}$ as unknown parameters in the Gibbs sampling algorithm [see Wakefield et al. (1994)]. The second point to note is that the parameters β and Ω can also be treated as unknowns and included in the Gibbs sampler without much extra effort.

If the Gibbs sampler is applied to the blocks $\{b_i\}$, σ^2, β, and Ω^{-1}, the hierarchical structure of the model allows us to deduce the following facts: (i) the full conditional distribution $[\{b_i\}|Y_n, \beta, \sigma^2, \Omega^{-1}]$ factors into a product of the distributions $[b_i|\ y_i, \beta, \sigma^2, \Omega^{-1}]$, depending only on the data in the ith cluster; (ii) the full conditional distribution of σ^2 does not depend on β and Ω^{-1}; and (iii) the full conditional distributions of β and Ω^{-1} do not depend on Y_n. Specifically, under our standard prior distributions, the Gibbs sampling algorithm is defined by:

$$b_i|y_i, \beta, \sigma^2, \Omega^{-1} \sim \mathcal{N}_k\left(\hat{b}_i, V_i^{-1}\right), \quad (i \leq n);$$

$$
\begin{aligned}
\beta|\{b_i\},\Omega^{-1} &\sim \mathcal{N}_k\left(\hat{\beta},\left(B_0+n\Omega^{-1}\right)^{-1}\right)\\
\sigma^2|Y_n,\{\,b_i\} &\sim IG\left(\frac{\nu_0+nT}{2},\frac{\delta_0+\sum_{i=1}^n(y_i-X_ib_i)'(y_i-X_ib_i)}{2}\right); \quad \text{and}\\
\Omega^{-1}|\{b_i\},\beta &\sim \mathcal{W}_k\left(\rho_0+n,\left(R_0^{-1}+\sum_{i=1}^n(b_i-\beta)(b_i-\beta)'\right)^{-1}\right),
\end{aligned}
$$

where $\hat{b}_i=(V_i^{-1}\Omega^{-1}\beta+\sigma^{-2}X_i'y_i)$, $V_i=(\Omega^{-1}+\sigma^{-2}X_i'X_i)$, and $\hat{\beta}=(B_0+n\Omega^{-1})^{-1}(B_0\beta_0+\Omega^{-1}\sum_{i=1}^n b_i)$.

A full Bayes analysis of this important model is thus accomplished by simulating the four distributions presented above. It should be noted that an extremely useful by-product of the Gibbs algorithm is the posterior distribution of the random effects. This distribution can be used to study the extent of heterogeneity present in the data [Allenby and Rossi (1993)].

Additional complexity can be introduced into this model without destroying tractability. For example, suppose y_{it} is a binary random variable such that $\Pr(y_{it}=1|b_i)=F(\,x_{it}'b_i)$, where F is a known c.d.f. For the logistic c.d.f., a Gibbs analysis of this model is developed by Zeger and Karim (1991). For the probit c.d.f., introduce latent variables $z_{it}\sim \text{iid}\mathcal{N}(x_{it}'b_i,1)$, $1\le i\le n$, $1\le t\le T$, into the Gibbs sampler, simulating each from the truncated normal distribution $\mathcal{TN}_{(0,\infty)}(x_{it}'b_i,1)$ if $y_{it}=1$ and $\mathcal{TN}_{(-\infty,0]}(x_{it}'b_i,1)$ if $y_{it}=0$ [Albert and Chib (1993c)]. Then, given values of $\{z_{it}\}$, the model reduces to the one presented above.

3.4 State-space model

In the state-space model [Harvey (1981)], the observation vector y_t is generated by

$$
y_t=X_t\theta_t+\epsilon_t,\quad \epsilon_t\sim \text{iid}\mathcal{N}_p(0,\Omega),\quad 1\le t\le n,
$$

and the state vector $\theta_t : m\times 1$ evolves according to the Markov process

$$
\theta_t=G\theta_{t-1}+\eta_t,\quad \eta_t\sim \text{iid}\mathcal{N}_m(0,\Psi). \tag{9}
$$

In the frequentist approach the unknown parameters (Ω,G,Ψ) are estimated by maximum likelihood, and inferences on the states are conducted through the Kalman filter and smoothing recursions, given the estimated parameters. A full Bayes approach for the non-linear version of this model is developed by Carlin, Polson, and Stoffer (1992) and for the present linear case by Carter and Kohn (1992), Chib (1992b), and Chib and Greenberg (1993b). We illustrate the case of known G, but the procedure can be extended to deal with an unknown G.

From the previous examples it is clear that the θ_t should be included in the Gibbs sampler, but this may be done either through the distributions

$$[\theta_t|Y_n,\Omega,\Psi,\theta_s(s\neq t)],\quad [\Omega|Y_n,\{\theta_t\},\Psi],\quad [\Psi|Y_n,\{\theta_t\},\Omega], \tag{10}$$

or through the distributions

$$[\theta_0,\ldots,\theta_n|Y_n,\Omega,\Psi],\quad [\Omega|Y_n,\{\theta_t\},\Psi],\quad [\Psi|Y_n,\{\theta_t\},\Omega]. \tag{11}$$

The two samplers differ in the way they simulate the θ_t's. In (10) the states are simulated from their individual full conditional distributions, while in (11) they are sampled from their joint full conditional distribution. Because the θ_t are correlated (they follow a Markov process), the blocking in (11) will lead to faster convergence to the target distribution and is therefore preferred.

The Gibbs sampler proceeds as follows: If the state vectors are known, the full conditional distributions for Ω^{-1} and Ψ^{-1} are given by

$$\begin{aligned}
\Omega^{-1}|Y_n,\{\theta_t\} &\sim \mathcal{W}_p\left(\rho_0+n,\left[R_0^{-1}+\sum_{t=1}^{n}(y_t-X_t\theta_t)(y_t-X_t\theta_t)'\right]^{-1}\right),\\
\Psi^{-1}|Y_n,\{\theta_t\} &\sim \mathcal{W}_m\left(\delta_0+n,\left[D_0^{-1}+\sum_{t=1}^{n}(\theta_t-G\theta_{t-1})(\theta_t-G\theta_{t-1})'\right]^{-1}\right),
\end{aligned}$$

where δ_0 and $D_0 : m\times m$ are the parameters of the Wishart prior for Ψ^{-1}. These are both standard distributions.

For the simulation of the $\{\theta_t\}$, let $\psi=(\Omega,\Psi)$ and $Y_t=(y_1,\ldots,y_t)$. By writing the joint density of $\{\theta_t\}$ in reverse time order,

$$p(\theta_n|Y_n,\psi)\times p(\theta_{n-1}|Y_n,\theta_n,\psi)\times\ldots\times p(\theta_0|Y_n,\theta_1,\ldots,\theta_n,\psi), \tag{12}$$

we can see how to obtain a draw from the joint distribution: Draw $\tilde{\theta}_n$ from $[\theta_n|Y_n, \psi]$, then draw $\tilde{\theta}_{n-1}$ from $[\theta_{n-1}|Y_n, \tilde{\theta}_n, \psi]$, and so on, until $\tilde{\theta}_0$ is drawn from $[\theta_0|Y_n, \tilde{\theta}_1, \ldots, \tilde{\theta}_n, \psi]$. We now show how to derive the density of the typical term in (12), $p(\theta_t|Y_n, \theta_{t+1}, \ldots, \theta_n, \psi)$.

Let $\theta^s = (\theta_s, \ldots, \theta_n)$ and $Y^s = (y_s, \ldots, y_n)$ for $s \leq n$. Then

$$\begin{aligned} p(\theta_t|Y_n, \theta^{t+1}, \psi) &\propto p(\theta_t|Y_t, \psi)\, p(\theta_{t+1}|Y_t, \theta_t, \psi)\, f(Y^{t+1}, \theta^{t+1}|Y_t, \theta_t, \theta_{t+1}, \psi) \\ &\propto p(\theta_t|Y_t, \psi)\, p(\theta_{t+1}|\theta_t, \psi), \end{aligned} \tag{13}$$

from (9) and the fact that (Y^{t+1}, θ^{t+1}) is independent of θ_t given (θ_{t+1}, ψ). The first density is Gaussian with moments $\hat{\theta}_{t|t}$ and $R_{t|t}$, which are obtained by running the recursions $\hat{\theta}_{t|t} = G\hat{\theta}_{t|t-1} + K_t(\, y_t - X_t\hat{\theta}_{t|t-1})$ and $R_{t|t} = (I - K_tX_t)R_{t|t-1}$, where $\hat{\theta}_{t|t-1} = G\hat{\theta}_{t-1|t-1}$, $F_{t|t-1} = X_tR_{t|t-1}X_t' + \Omega$, $R_{t|t-1} = GR_{t-1|t-1}G' + \Psi$, and $K_t = R_{t|t-1}X_t'F_{t|t-1}^{-1}$. The second density is Gaussian with moments $G\theta_t$ and Ψ. Completing the square in θ_t leads to the following algorithm to sample $\{\theta_t\}$:

1. Run the Kalman filter and save its output $\{\hat{\theta}_{t|t}, R_t, M_t\}$, where $R_t = R_{t|t} - M_tR_{t+1|t}M_t'$ and $M_t = R_{t|t}R_{t+1|t}^{-1}$.

2. Simulate $\tilde{\theta}_n$ from $\mathcal{N}_m(\hat{\theta}_{n|n}, R_{n|n})$; then simulate $\tilde{\theta}_{n-1}$ from $\mathcal{N}_m(\hat{\theta}_{n-1}, R_{n-1})$, and so on, until $\tilde{\theta}_0$ is simulated from $\mathcal{N}_m(\hat{\theta}_0, R_0)$, where $\hat{\theta}_t = \hat{\theta}_{t|t} + M_t\left(\tilde{\theta}_{t+1} - \hat{\theta}_{t|t}\right)$.

3.5 Regression models with AR(p) errors

This subsection illustrates a simulation in which the MH algorithm is combined with the Gibbs sampling algorithm. A detailed analysis of the regression model with ARMA(p, q) errors may be found in Chib and Greenberg (1993a) and Marriott et al. (1993).

Consider the model

$$y_t = x_t'\beta + \epsilon_t, \quad 1 \leq t \leq n, \tag{14}$$

where y_t is a scalar observation. Suppose that the error is generated by the stationary AR(p) process

$$\epsilon_t - \phi_1\epsilon_{t-1} - \ldots - \phi_p\epsilon_{t-p} = u_t \quad \text{or} \quad \phi(L)\epsilon_t = u_t, \tag{15}$$

where $u_t \sim \text{iid}\mathcal{N}(0, \sigma^2)$ and $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ is a polynomial in the lag operator L. The stationarity assumption implies that the roots of $\phi(L)$ lie outside the unit circle;

this constrains $\phi = (\phi_1, \ldots, \phi_p)$ to lie in a subset (say S_ϕ) of $\Re^p$. To conform to this constraint, we take the prior of ϕ to be $\mathcal{N}_p(\phi|\phi_0, \Phi_0^{-1})I_{S_\phi}$, a normal distribution truncated to the stationary region (and assume the standard prior distributions for β and σ^2). The likelihood function for this model can be expressed as

$$f(Y_n|\beta, \phi, \sigma^2) = \Psi(\phi) \times (\sigma^2)^{-(n-p)/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}(y_t^* - x_t^{*\prime}\beta)^2\right],$$

where, for $t \geq p+1$, $y_t^* = \phi(L)y_t$, $x_t^* = \phi(L)x_t$, and

$$\Psi(\phi) = (\sigma^2)^{-p/2}|\Sigma_p|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(Y_p - X_p\beta)\Sigma_p^{-1}(Y_p - X_p\beta)\right] \tag{16}$$

is the (stationary) density of the first p observations. In the above, $Y_p = (y_1, \ldots, y_p)'$, $X_p = (X_1, \ldots, X_p)'$, and $\Sigma_p = \Phi\Sigma_p\Phi' + \ e_1(p)\ e_1(p)'$, with

$$\Phi = \begin{bmatrix} \phi_{-p} & \phi_p \\ I_{p-1} & 0 \end{bmatrix},$$

$e_1(p) = (1, 0, \ldots, 0)'$, and $\phi_{-p} = (\phi_1, \ldots, \phi_{p-1})'$.

How can the posterior density be simulated? The answer lies in recognizing three facts. First, the Gibbs strategy is useful for this problem by taking β, ϕ, and σ^2 as blocks.[10] Second, the full conditional distributions of β and σ^2 can be obtained easily after combining the two exponential terms in the sampling density. Third, the full conditional of ϕ can be simulated with the MH algorithm. We next provide some of the details.

Define $Y_p^* = Q^{-1}\,Y_p$ and $X_p^* = Q^{-1}X_p$, where Q satisfies $QQ' = \Sigma_p$. Let $y^* = (y_1^*, \ldots, y_n^*)'$ and likewise for X^*. Finally let $e = (e_{p+1}, \ldots, e_n)'$ and let E denote the $n-p \times p$ matrix with tth row given by $(e_{t-1}, \ldots, e_{t-p})$, where $e_t = y_t - \ x_t'\beta$, $t \geq p+1$. It is now not difficult to show that the full conditional distributions are

$$\begin{aligned} \beta|Y_n, \phi, \sigma^2 &\sim \mathcal{N}_k\left(\hat\beta, B_n^{-1}\right) \\ \phi|Y_n, \beta, \sigma^2 &\propto \Psi(\phi) \times \mathcal{N}_p\left(\hat\phi, \Phi_n^{-1}\right) I_{S_\phi}, \quad \text{and} \\ \sigma^2|Y_n, \beta, \phi &\sim \mathcal{IG}\left(\tfrac{\nu_0+n}{2}, \tfrac{\delta_0+d_1}{2}\right), \end{aligned} \tag{17}$$

where $\hat\beta = B_n^{-1}(B_0\beta_0 + \sigma^{-2}X^{*\prime}y^*)$, $B_n = (B_0 + \sigma^{-2}X^{*\prime}X^*)$, $d_\beta = ||y^* - X^*\beta||^2$, $\hat\phi = \hat\Phi_n^{-1}(\Phi_0\phi_0 + \sigma^{-2}E'\,e)$, and $\hat\Phi_n = (\Phi_0 + \sigma^{-2}E'E)$.

[10] In the analysis of the AR(p) model conditioned on Y_p, Chib (1993) shows that all full conditional distributions take standard forms.

The full conditionals of β and σ^2 are easily simulated. To simulate ϕ we can employ the MH independence chain with $\mathcal{N}_p\left(\hat{\phi}, \Phi_n^{-1}\right) I_{S_\phi}$ as the candidate generating density. Then the MH step is implemented as follows. At the ith iteration of the Gibbs cycle, draw a candidate $\phi^{(i+1)}$ from a normal density with mean $\hat{\phi}$ and covariance $\sigma^{2(i)}\Phi_n^{-1}$; if it satisfies stationarity, we move to this point with probability

$$\min\left\{\frac{\Psi(\phi^{(i+1)})}{\Psi(\phi^{(i)})}, 1\right\}$$

and otherwise set $\phi^{(i+1)} = \phi^{(i)}$. Chib and Greenberg (1993a) verify the sufficient conditions for the convergence of this algorithm and provide several empirical examples.

3.6 Other models

Other models in addition to those illustrated above lend themselves to MCMC methods and to Gibbs sampling with data augmentation in particular. In the regression framework, missing data can be added to the sampler to generate samples from distributions of the parameters. The important class of multinomial probit models can be analyzed by MCMC simulation (through data augmentation), as discussed by Albert and Chib (1993a), McCulloch and Rossi (1993), and Geweke, Keane, and Runkle (1993). Another important area is that of mixture models, in which each observation in the sample can arise from one of K different populations. Two types of models have been investigated. In the first, the populations are sampled independently from one observation to the next [Diebolt and Robert (1994)]. In the second, the populations are sampled according to a Markov process, which is the Markov switching model [Albert and Chib (1993b), and Chib (1993c)]. New econometric applications that illustrate the versatility of MCMC methods continue to appear: reduced rank regressions [Geweke (1993b)], stochastic volatility models [Jacquier et al. (1994)], cost function models [Koop et al. (1994)], censored autocorrelated data [Zangari and Tsurumi (1994)], and many others.

4 Inference with MCMC methods

We next examine ways in which a sample generated by MCMC methods can be used for statistical inference, including estimation of moments and marginal densities, prediction, sensitivity, model adequacy, and estimation of modes.

4.1 Estimation of moments and numerical standard errors

An implication of Proposition 1 is that output from the MCMC simulation can be used to estimate moments, quantiles, and other summaries of the target density. With ψ denoting parameters and latent data and Y_n denoting the sample data, the target density is the posterior density $\pi(\psi|Y_n)$ and the MCMC sample is the collection $\{\psi^{(i)} : i \leq M\}$. The MCMC estimate of the quantity $\bar{h} = \int h(\psi)\pi(\psi|Y_n)\, d\psi$ (for integrable h) is given by the ergodic average

$$\hat{h} = M^{-1} \sum_{i=1}^{M} h(\psi^{(i)}). \tag{18}$$

This expression can be used to estimate the posterior mean by letting $h(\psi) = \psi$ and the posterior second moment matrix by letting $h(\psi) = \psi\psi'$. It should be noted, however, that the estimate of (18) differs from that of other Monte Carlo methods (importance sampling, for example) because the $\{\psi^{(i)}\}$ are not independent. In particular, the estimate of the Monte Carlo standard error (numerical standard error), which indicates the variation that can be expected in $\hat{h}$ if the simulation were to be repeated, is affected. Following Ripley (1987, Ch. 6), let $Z_i = h(\psi^{(i)})$. Under regularity conditions, since $\{Z_i\}$ is an ergodic time series with autocorrelation sequence $\rho_s = \mathrm{corr}(Z_i, Z_{i-s})$ and variance $\sigma^2 = \mathrm{var}(Z_i)$, we have

$$\begin{aligned} \mathrm{var}(\hat{h}) &= M^{-2} \textstyle\sum_{j,k} \mathrm{cov}(Z_j, Z_k) \\ &= \sigma^2 M^{-2} \textstyle\sum_{j,k=1}^{M} \rho_{|j-k|} \\ &= \sigma^2 M^{-1} \left[1 + 2 \textstyle\sum_{s=1}^{M} (1 - \frac{s}{M})\rho_s\right], \end{aligned}$$

which is larger than σ^2/M (the variance under independent sampling) if all the $\rho_s > 0$, as is frequently the case. The variance equals τ^2/M, where $\tau^2 = 2\pi f(0)$ and $f(\cdot)$ is the spectral density of $\{Z_i\}$. Many methods have been proposed to estimate the variance efficiently; Geweke (1992), for example, estimates the spectral density at frequency zero, while McCulloch and Rossi (1993) use the approach of Newey and West (1987) [see also Geyer (1992)]. An equivalent, more traditional approach is based on the method of "batch means." The data $\{Z_i\}$ are batched or sectioned into k subsamples of length m with means $\{B_i\}$ and the variance of $\hat{h}$ estimated as $[k(k-1)]^{-1} \sum (B_i - \bar{B})^2$. The batch length m is chosen large enough that the first order serial correlation between batch means is less than 0.05.

4.2 Marginal density estimates

Along with moments, the marginal and joint density functions of components of ψ are important summaries of the target density. To obtain such densities, for example the marginal density of ψ_1, it is possible to compute the histogram of the simulated values $\{\psi_1^{(i)}\}$ since these are samples from the marginal posterior $\pi(\psi_1|Y_n)$. More generally, the histogram estimate can be smoothed by standard kernel methods. In the context of the Gibbs sampling algorithm, however, another density estimate is available. Suppose there are d blocks and we wish to compute the marginal density of the first block, $\pi(\psi_1|Y_n) = \int \pi(\psi_1|Y_n, \psi_2, \ldots, \psi_d)\, \pi(\psi_2, \ldots, \psi_d|Y_n)\, d\psi_2 \ldots d\psi_d$. Because $\{\psi_2^{(i)}, \ldots, \psi_d^{(i)}\}$ is a sample from the marginal density $\pi(\psi_2, \ldots, \psi_d|Y_n)$, an estimate of this density at the point ψ_1^* is

$$\hat{\pi}(\psi_1^*|Y_n) = M^{-1} \sum_{i=1}^{M} \pi(\psi_1^*|Y_n, \psi_2^{(i)}, \ldots, \psi_d^{(i)}). \tag{19}$$

Gelfand and Smith (1990) refer to (19) as "Rao-Blackwellization," and Liu et al. (1994) show that this mixture approximation to the marginal density generally produces estimates with smaller numerical standard error than the empirical estimator. They also find that it is preferable to calculate $\int h(\psi_1)\pi(\psi|Y_n)\, d\psi$ by averaging $E(h(\psi_1)|Y_n, \psi_2, \ldots, \psi_d)$ (if the latter is available) over the simulated draws of $(\psi_2, \ldots, \psi_d)$.

4.3 Predictive inference

Consider the question of obtaining the density of a set of future observations y_f given the current model. This is the predictive density $f(y_f|Y_n) = \int f(y_f|Y_n, \psi)\pi(\psi|Y_n)\, d\psi$, where $f(y_f|Y_n, \psi)$ is the conditional density of the future observations given ψ. Even though the integral cannot generally be evaluated it is possible to simulate (by the method of *composition*) a sample of draws from $f(y_f|Y_n)$ given a sample from $\pi(\psi|Y_n)$: For each $\psi^{(i)}$, simulate the vector $y_f^{(i)}$ from the density $f(y_f|Y_n, \psi^{(i)})$. Then $\{y_f^{(i)}\}$ constitutes the desired sample. The simulated forecast values can be summarized in the usual ways. Albert and Chib

(1993b) use this approach to obtain the 4-step ahead prediction density for autoregressive models with Markov switching.

4.4 Sensitivity analysis

It is often of interest to determine the sensitivity of the estimate in (18) to changes in the prior distribution. This can be done by the method of *sampling-importance-resampling* (SIR) without re-running the MCMC simulation [Rubin (1988)]. Specifically, given a sample $\psi^{(1)}, \ldots, \psi^{(M)}$ from $\pi(\psi|Y_n)$, a sample of m draws from a posterior density $p(\psi|Y_n)$ that corresponds to a different prior density can be obtained by resampling the original draws with weights $w(\psi_i) \propto \frac{p(\psi^{(i)}|Y_n)}{\pi(\psi^{(i)}|Y_n)}$, $i = 1, \ldots, M$. The resampled values, which are distributed according to $p(\cdot|\cdot)$ as $M/m \to \infty$, can be used to recompute $\hat{h}$. Other model perturbations can also be similarly analyzed [Gelfand and Smith (1992)].

4.5 Evaluation of model adequacy

The marginal (integrated) likelihood is a central quantity in the comparison of Bayesian models. If the models are defined as $H_k = \{f(Y_n|\psi_k), \pi(\psi_k)\}$, where ψ_k is the parameter vector for the kth model, then the marginal likelihood for model (or hypothesis) H_k is defined as

$$m(Y_n|H_k) = \int f(Y_n|\psi_k)\, \pi(\psi_k)\, d\,\psi_k,$$

which is the integral of the sampling density w.r.t. to the prior density. The evidence in the data for any two models M_k and M_l is summarized by the Bayes factor $B_{kl} = m(Y_n|H_k)/m(Y_n|H_l)$, or by the posterior odds $O_{kl} = B_{kl} \times (p_k/p_l)$ where p_k is the prior probability of M_k [Leamer (1978), Zellner (1984)].

Two distinct methods have been used to compute B_{kl} [see Kass and Raftery (1994) for a comprehensive review]. In the first approach [Newton and Raftery (1994), Chib (1994)], $m(Y_n|H_k)$ is computed directly from the MCMC output corresponding to model M_k. In the second approach [Carlin and Chib (1993)], a model indicator M, $M \in \{1, \ldots, K\}$, is defined, and a Gibbs sampler is constructed from the full conditional distributions $[\psi_1, \ldots, \psi_K|Y_n, M]$ and $[M|Y_n, \psi_1, \ldots, \psi_K]$. The posterior relative frequencies of M are used to compute posterior model probabilities and thence the Bayes factors for any two models. A related approach for models with a common parameter ψ is considered by Carlin and Polson (1991) and George and McCulloch (1993).

4.6 Modal estimates

Markov chain methods can be used to find the modal estimates in models with missing or latent data. This is achieved by sampling the latent or missing data and then evaluating the E step in the EM algorithm using the simulated draws [Celeux and Diebolt (1985), Wei and Tanner (1990), and Ruud (1991)].

Given the current estimate of the maximizer $\theta^{(i)}$, define

$$Q(\theta, \theta^{(i)}) = \int_{Z_n} \log\left(\pi(\theta|Y_n, Z_n)\right)\, d[Z_n|Y_n, \theta^{(i)}],$$

where Y_n is the observed data and Z_n is the latent data. To avoid what is usually an intractable integration, given parameter values we can draw a sample $Z_{n,j}$, $j \leq N$, by MCMC and approximate Q by $\hat{Q}(\theta, \theta^{(i)}) = N^{-1} \sum_j \log\left(\pi(\theta|Y_n, Z_{n,j})\right)$. In the M-step, $\hat{Q}$ is maximized over θ to obtain the new parameter $\theta^{(i+1)}$. These steps are repeated until the difference $\left\|\theta^{(i+1)} - \theta^{(i)}\right\|$ is negligible. When producing the sequence $\{\theta^{(i)}\}$ it is usual to begin with a small value of N and let the number of replications of Z_n increase as the maximizer is approached. This procedure is applied to finite mixture distributions with Markov switching in Chib (1993b) and to partial non-Gaussian state-space models in Shephard (1994).

5 Conclusions

Our survey of developments in the theory and practice of Markov chain Monte Carlo methods, with an emphasis on applications to econometric problems, has shown how Gibbs and Metropolis-Hastings sampling, combined with data augmentation, can be used to organize a systematic approach to Bayesian inference. We have illustrated the ideas in the context of models with censoring, discrete responses, panel data, autoregressive errors, and random and time-varying parameters, but the ideas can be applied to many other econometric models. For frequentist econometricians, we have shown how Monte Carlo versions of the EM algorithm can be used to find the posterior mode.

One of the considerable arguments in favor of MCMC methods (and for simulation-based inference in general) is that they make possible the analysis of models that are difficult to analyze by other means. No longer is analysis in the Bayesian context restricted to tightly specified models and prior distributions. As we have shown, many models, including those with intractable likelihood functions, can be simulated by MCMC methods. Various inference questions, especially those relating to prediction, model and prior perturbations, and model adequacy can be addressed effectively using the output of the simulation.

MCMC methods have already proved extremely useful in econometrics, and more applications continue to appear at a rapid rate. These developments have been enormously aided by significant improvements in computer hardware and software. Great opportunities remain for the work that still needs to be done, especially in the form of applications to new and existing problems and theoretical developments on the speed of convergence, sufficient conditions for validity, and tuning of methods.

References

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

Albert, J. & 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. & S. Chib (1993c), A practical Bayes approach for longitudinal probit regression models with random effects. Unpublished manuscript, Bowling Green State University.

Allenby, G. & P. Rossi (1993), A Bayesian approach to estimating household parameters. *Journal of Marketing Research* 30, 171-182.

Carlin, B. P. & N. G. Polson (1991), Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* 19, 399–405.

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

Carlin, B. P., A. E. Gelfand, & A. F. M. Smith (1992), Hierarchical Bayesian analysis of change point problems. *Journal of the Royal Statistical Society C* 41, 389–405.

Carlin, B. P. and S. Chib (1993), Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, forthcoming.

Carter, C. & R. Kohn (1992), On Gibbs sampling for state space models. *Biometrika,* forthcoming.

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

Celeux, G. and J. Diebolt (1985), The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* 2, 73–82.

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

Chib, S. (1992a), Bayes regression for the tobit censored regression model. *Journal of Econometrics* 51, 79–99.

Chib, S. (1992b), An accelerated Gibbs sampler for state space models and other results. Unpublished manuscript, Washington University.

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

Chib, S. (1993b), Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, forthcoming.

Chib, S. (1994), Marginal likelihood from the Gibbs output. Unpublished manuscript, Washington University.

Chib, S. & E. Greenberg (1993a), Bayes inference for regression models with ARMA(p, q) errors. *Journal of Econometrics*, forthcoming.

Chib, S. & E. Greenberg (1993b), Hierarchical analysis of SUR models with extensions to correlated serial errors and time varying parameter models. *Journal of Econometrics*, forthcoming.

Chib, S. & E. Greenberg (1994), Understanding the Metropolis-Hastings algorithm. Unpublished manuscript, Washington University.

Dempster, A. P., N. Laird, & D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39, 1–38.

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

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

Gelfand, A. E. & A. F. M. Smith (1992), Bayesian statistics without tears: A sampling-resampling perspective. *American Statistician* 46, 84–88.

Gelfand, A. E., A. F. M. Smith, & T. M. Lee (1992), Bayesian analysis of constrained parameter and truncated data problems. *Journal of the American Statistical Association* 87, 523–532.

Gelman, A. & D. B. Rubin (1992), Inference from iterative simulation using multiple sequences. *Statistical Science* 4, 457–472.

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

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

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 & S. Kaufman (eds.), *Computing Science and Statistics: Proceedings of the 23^{rd}* Symposium on the Interface. Pp. 571–578. Fairfax Station VA: Interface Foundation of North America.

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, & A. F. M. Smith (eds.), *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics*. Pp. 169–193. New York: Oxford University Press.

Geweke, J. (1993a), Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics* 8, S19–S40.

Geweke, J. (1993b), A Bayesian analysis of reduced rank regressions. Unpublished manuscript, University of Minnesota.

Geweke, J., M. Keane, & D. Runkle (1993), Alternative computational approaches to inference in the multinomial probit model. Unpublished manuscript, University of Minnesota.

Geyer, C. (1992), Practical Markov chain Monte Carlo. *Statistical Science* 4, 473–482.

Gilks, W. R. & P. Wild (1992), Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 41, 337–348.

Harvey, A. C. (1981), *Time series models*. London: Philip Allan.

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

Hsiao, C. (1986), *Analysis of panel data.* New York: Cambridge University Press.

Jacquier, E., N. G. Polson, & P. E. Rossi (1994), Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, forthcoming.

Kass, R. & A. E. Raftery (1994). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, forthcoming.

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

Koop, G. (1994), Recent progress in applied Bayesian econometrics. *Journal of Economic Surveys* 8, 1–34.

Koop, G., J. Osiewalski, & M. F. J. Steel (1994). Bayesian efficiency analysis with a flexible form: The AIM cost function. *Journal of Business and Economic Statistics* 12, 339–346.

Leamer, E. E. (1978). *Specification searches: Ad-hoc inference with non-experimental data.* John Wiley and Sons: New York.

Liu, J. S., W. W. . Wong, & A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika* 81, 27–40.

Marriott, J., N. Ravishanker, & A. E. Gelfand (1993), Bayesian analysis of ARMA processes: Complete sampling-based inferences under full likelihoods. Unpublished manuscript, University of Connecticut.

McCulloch, R. E. & P. E. Rossi (1993), Exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, forthcoming.

McCulloch, R. E. & R. S. Tsay (1993), Statistical inference of macroeconomic time series via Markov switching models. Unpublished manuscript, University of Chicago.

Mengersen, K. L. & R. L. Tweedie (1993), Rates of convergence of the Hastings and Metropolis algorithms. Unpublished manuscript, Colorado State University.

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

Meyn, S. P. & R. L. Tweedie (1993), *Markov chains and stochastic stability.* London: Springer-Verlag.

Morris, C. N. (1987), Comment: Simulation in hierarchical models. *Journal of the American Statistical Association* 82, 542–543.

Müller, P. (1991), A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Department of Statistics, Purdue University.

Newey, W. K. & and K. D. West (1987), A simple positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Newton, M. A. & A. E. Raftery (1994), Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B* 56, 3–48.

Nummelin, E. (1984), *General irreducible Markov chains and non-negative operators.* Cambridge: Cambridge University Press.

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

Polson, N. G. (1992), Comment. *Statistical Science* 7, 490–491.

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

Ripley, B. (1987), *Stochastic simulation.* New York: John Wiley & Sons.

Ritter, C. & M. A. Tanner (1992), The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association* 87, 861–868.

Roberts, G. O. & A. F. M. Smith (1994), Some convergence theory for Markov chain Monte Carlo. *Stochastic Processes and Applications*, forthcoming.

Roberts, G. O. & R. L. Tweedie (1994), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Unpublished manuscript, University of Cambridge.

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

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo method.* New York: John Wiley & Sons.

Ruud, P. A. (1991), Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* 49, 305–341.

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

Smith, A. F. M. & 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.

Tanner, M. A. (1993), *Tools for Statistical Inference*, 2nd. edition. New York: Springer-Verlag.

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

Tierney, L. (1994), Markov chains for exploring posterior distributions. *Annals of Statistics*, forthcoming.

Tobin, J. (1958), Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36.

Wakefield, J. C., A. E. Gelfand, A. Racine Poon, & A. F. M. Smith (1994), Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics* 43, 201–221.

Wei, G. C. G. & M. A. Tanner (1990), 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.

Zangari, P. J. & H. Tsurumi (1994), A Bayesian analysis of censored autocorrelated data on exports of Japanese passenger cars to the U. S. Unpublished manuscript, Rutgers University.

Zeger, S. L. & 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. (1984), *Basic issues in econometrics.* Chicago: University of Chicago Press.

Zellner, A. & C. Min (1993), Gibbs sampler convergence criteria (GSC^2). Unpublished manuscript, University of Chicago.