Markov Chain Monte Carlo Methods in Biostatistics*

Andrew Gelman
Department of Statistics
Columbia University
New York, NY 10027

Donald B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138

May 21, 1996

1 Introduction

Appropriate models in biostatistics are often quite complicated, reflecting longitudinal data collection, hierarchical structure on units of analysis, multivariate outcomes, and censored and missing data. Although simple, standard, analytically tractable models may sometimes be useful, often special models need to be fit that do not have analytically tractable solutions. It is natural in such cases to turn to Bayesian methods, which can often be implemented using simulation techniques. In fact, as emphasized in Rubin (1984), one of the great scientific advantages of simulation analysis of Bayesian methods is the freedom it gives the researcher to formulate appropriate models rather than be overly interested in analytically neat but scientifically inappropriate models. The basic idea of simulation is simple and important: after collection of data y, uncertainty about a vector of parameters in a statistical model is summarized by a set of random draws of the parameter vector, θ, from a posterior distribution: $p(\theta|y)$. Markov chain Monte Carlo methods are an extremely important set of tools for such simulations.

In this article, we review some important general methods for Markov chain Monte Carlo

*For *Statistical Methods in Medical Research—An International Review Journal*. We thank two referees for helpful comments and the National Science Foundation for partial support through grants SBR-9207456, DMS-9404305, and Young Investigator Award DMS-9457824. We also thank the U.S. Census Bureau for supporting an early version of this article through a contract to the National Opinion Research Center and Datametrics Research, Inc.

simulation of posterior distributions. None of these methods are specific algorithms with automatic computer programs; rather, they are approaches to computation that, at this point, require special programming for each new application. General computer programs for these methods are being developed (see Spiegelhalter et al., 1994), but at this point, individual programming is needed because it is generally necessary to develop a new model for each new statistical problem.

We anticipate that some readers of this article are already experienced in programming for statistical tasks such as computing point estimates and standard errors for classical models in biostatistics. These readers can use this article as an introduction to the ways in which Markov chain Monte Carlo simulation generalizes earlier, deterministic calculations, and as a source of references to more thorough treatments of particular simulation methods. For readers who are not experienced in statistical computation, an important role of this survey is to explain the continuity between the earlier methods of point estimation and the Markov chain Monte Carlo methods that are becoming standard for computing fully Bayesian analyses in complicated models.

1.1 Bayesian models and computation in biostatistics

In Bayesian inference, all unknowns are treated as random variables, which follow the posterior distribution, $p(\theta|y) \propto p(\theta)p(y|\theta)$, after collection of data y. In this notation, θ includes all parameters and uncertain quantities in the model, including (in the terminology of regression) fixed effects, random effects, hierarchical parameters, unobserved indicator variables, and missing data; $p(\theta)$ is the prior or marginal distribution of θ, and $p(y|\theta)$ is the sampling distribution for y, given θ.

Only in a very few simple examples can the posterior distribution be written in a standard analytic form; the most important of these examples are the normal, binomial, Poisson, exponential, and normal linear regression models with conjugate prior distributions. These

examples are important, but there is a much wider variety of models for which exact analytic Bayesian inference is impossible. These include generalized linear models (e.g., Zeger and Karim, 1991, Karim and Zeger, 1992, and Dellaportas and Smith, 1993), hierarchical models (e.g., Longford, 1993), longitudinal models (e.g., Cowles, Carlin, and Connett, 1993), mixture models (e.g., West, 1992), and specific models for problems such as AIDS incidence (e.g., Lange, Carlin, and Gelfand, 1992, and Bacchetti, Segal, and Jewell, 1993), genetic sequencing (e.g., Baldi et al., 1994), epidemiology (e.g., Clayton and Bernardinelli, 1992, and Gilks and Richardson, 1993), and survival analysis (e.g., Kuo and Smith, 1992). Until recently, these problems were handled either in a partially Bayesian manner (which typically meant that some aspects of uncertainty in the models were ignored, as occurs when unknown parameters are replaced by point estimates), or else approximations were used to allow analytic solutions (for example, using a linear approximation to a generalized linear model). Both these approaches can be improved because simplified techniques were used for reasons of computational convenience.

This article surveys methods of iterative simulation, most notably Markov chain Monte Carlo methods, that allow essentially exact Bayesian computation using simulation draws from the posterior distribution. These methods can be applied to a wide range of probability distributions, including those that arise in all of the standard Bayesian models in biostatistics. We discuss the following steps: constructing an approximation to the posterior distribution, constructing a Markov chain Monte Carlo simulation algorithm, and monitoring the convergence of the simulations. After the simulations have essentially converged, the collection of simulated values is used as a discrete approximation to the posterior distribution.

1.2 Posterior simulation

Before delving into any details of Markov chain simulation, we discuss some general points about Bayesian inference using simulation. Given a set of posterior simulation draws, $\theta^1, \theta^2, \ldots, \theta^N$ of a vector parameter θ (where each θ^l represents a draw from the posterior distribution of θ), one can estimate the posterior distribution of any quantity of interest. For example, with $N = 1000$ simulation draws, one can estimate a 95% posterior interval for any function $\phi(\theta, y)$ of parameters and data by the 25th-largest and 975th-largest simulated values of $\phi(\theta^l, y)$, $l = 1, \ldots, 1000$.

Direct simulation. In some simple problems, such as the normal linear regression model, random draws can be obtained from the posterior distribution directly in one step, using standard computer programs (e.g., Gelman et al., 1995, ch. 8). In other somewhat more complicated cases, such as the normal linear regression model with unknown variance, the parameter vector can be partitioned into two sub-vectors, $\theta = (\theta_1, \theta_2)$, such that the posterior distribution of θ_1, $p(\theta_1|y)$, and the conditional posterior distribution of θ_2 given θ_1, $p(\theta_2|\theta_1, y)$, are both standard distributions from which simulations can be easily drawn. Then the simplest and best approach to drawing a posterior simulation is to sample the subvectors in order by performing the following two steps: first draw θ_1 from its marginal posterior density, $p(\theta_1|y)$; then draw θ_2 from its posterior density, $p(\theta_2|\theta_1, y)$, conditional on the drawn value of θ_1. For example, in a normal linear regression with unknown variance (and a noninformative or conjugate prior distribution), one can draw $\sigma^2|y$ from an inverse-χ^2 distribution and then $\beta|\sigma^2, y$ from a normal distribution (see, e.g., Gelman et al., 1995, ch. 8). To obtain N simulated draws, simply repeat the process N times.

Iterative simulation. Unfortunately, for many problems, such as generalized linear models and hierarchical models, direct simulation is not possible, even in two or more steps.

Until recently, these problems have been attacked by approximating the desired posterior distributions by normal or transformed normal distributions, from which direct simulations can be drawn. In recent years, however, iterative simulation methods have been developed to draw from general distributions without any direct need for normal approximations. Markov chain Monte Carlo methods have a long history in computational physics, with the first general presentation in Metropolis et al., 1953, and were more recently introduced for statistical and biostatistical problems by Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990). Recent review articles on the topic include Gelfand et al. (1990), Smith and Roberts (1993), Besag and Green (1993), and Gilks et al. (1993). The recent book edited by Gilks, Richardson, and Spiegelhalter (1996) is a nice practical overview of Markov chain Monte Carlo methods in statistics. More general treatments of Bayesian methods and computation appear in the books by Tanner (1993), Gelman et al. (1995), and Carlin and Louis (1996).

The advantage of these iterative methods is that they can be set up with virtually any model that can be set up in statistics; their disadvantage is that they currently require extensive programming and even more extensive debugging. For this reason and others, the earlier methods of approximation are still important, both for setting up starting points and for providing checks on the answers obtained from the Markov chain methods. Section 2 of this article discusses methods of roughly approximating the posterior distribution of θ as a preliminary to iterative simulation for θ. Section 3 gives a cursory outline of the mathematics of Markov chain simulation, Section 4 discusses implementation, and Section 5 gives an example from an analysis of an experiment involving schizophrenics.

2 What to do before doing Markov chain simulation

2.1 General advice

It is generally a mistake to attempt to run a Markov chain simulation program without knowing roughly where the posterior distribution is located in parameter space. Existing methods and software for parameter estimation are important as starting points for more complicated simulation procedures. For example, suppose one would like to fit a hierarchical generalized linear model in the presence of censoring and missing data. Then it would make sense to use existing computer packages to fit parts of the model (for example, a hierarchical linear model ignoring the missing data with a simple approximation for the censored data; a non-hierarchical generalized linear model using a similar approximation; an off-the-shelf model for analysis with censored data; an off-the-shelf model for imputing the missing data). These separate analyses will not capture all the features of the model and data, but they can be natural, low-effort starting points.

In Sections 2.2–2.4, we describe some basic estimation and approximation strategies; more details appear in Tanner (1993), Gelman and Rubin (1992b), and Gelman et al. (1995, ch. 9–10). These methods will not work for all problems; the point of these section is not to recommend one particular set of algorithms, but rather to explain the principles behind some often-effective methods. We shall see that many of these principles are useful for iterative simulation as well.

2.2 Point estimation and normal or Student-t approximations for unimodal posterior distributions

A point estimate of θ and its associated standard error (or, more generally, its variance-covariance matrix, Σ), are motivated, explicitly or implicitly, by the normal approximation to the posterior distribution, $\theta|y \sim \mathrm{N}(\mu, \Sigma)$. Typically, the mean, μ, of the normal approximation is set equal to the mode (i.e., the maximum likelihood estimate or the posterior

mode), and the inverse variance matrix, Σ^{-1}, is approximated by the negative of the second derivative (with respect to θ) matrix of the log posterior distribution calculated at $\theta = \mu$. Approximating μ and Σ can be difficult in highly multivariate problems. Just finding the mode can require iteration, with Newton's method and EM (Dempster, Laird, and Rubin, 1977) being popular choices for common statistical models. Estimates of Σ can be computed by analytic differentiation, numerical differentiation, or combined methods such as SEM (Meng and Rubin, 1991). Of course, in many problems (for example, generalized linear models), values for μ and Σ can be computed using available software packages.

Because we are creating point estimates only as a way to start iterative simulations, it is usually adequate to be rough in the initial estimation procedure. For example, various methods for approximate EM algorithms in generalized linear models (e.g., Laird and Louis, 1982, and Breslow and Clayton, 1993) often work fine. However, some methods for variance estimation, such as SEM, require an accurate estimate of a local mode.

It can often be useful to replace the normal approximation by a multivariate t, with the same center and scale, but thicker tails corresponding to its degrees of freedom, ν. If z is a draw from a multivariate $\mathrm{N}(0, \Sigma)$ distribution, and x is an independent draw from a χ^2_η distribution, then $\theta = \mu + z\sqrt{\eta/x}$ is a random draw from the multivariate $t_\eta(\mu, \Sigma)$ distribution. Because of its thicker tails (and because it can be easily simulated and its density function is easy to calculate), the multivariate t turns out to be useful as a starting distribution for the iterative simulation methods described below.

2.3 Approximation using a mixture of multivariate normal or Student-t densities for multimodal posterior distributions

When the posterior distribution of θ is multimodal, it is necessary to run an iterative mode-finder several times, starting from different points, in an attempt to find all the modes. This strategy is also sensible and commonly used if the distribution is complicated enough that it *may* be multimodal. Once all K modes are found (possibly a difficult task) and the second

derivative matrix estimated at each mode, the target distribution can be approximated by a mixture of K multivariate normals, each with its own mode μ_k and variance matrix Σ_k; that is, the target density $p(\theta|y)$ can be approximated by

$$p_{\text{approx}}(\theta) = \sum_{k=1}^{K} \frac{\omega_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\theta-\mu_k)^t\Sigma_k^{-1}(\theta-\mu_k)\right),$$

where d is the dimension of θ and ω_k is the mass of the k-th component of the multivariate normal mixture, which can be approximated by setting ω_k proportional to $|\Sigma_k|^{1/2}p(\mu_k|y)$, where $p(\mu_k|y)$ is the posterior density of θ evaluated at $\theta = \mu_k$.

2.4 Nonidentified parameters and informative prior distributions

Bayesian methods can be applied to models in which one or more parameters are poorly identified by the data, so that point estimates (such as maximum likelihood) are difficult or impossible to obtain. In these situations, it is often useful to transform the parameter space to separate the identified and non-identified parts of the model; to handle the uncertainty in the latter, perhaps using straightforward Bayesian methods, it is necessary to assign an informative prior distribution to these parameters.

Even if the parameters in a problem appear to be well identified, one must be careful when using noninformative prior distributions, especially for hierarchical models. For example, assigning an improper uniform prior distribution to the logarithm of a hierarchical variance parameter (such as σ_α^2 in the example of Section 5) yields an improper posterior distribution (Hill, 1965); in this context, "improper" refers to any probability density that does not have a finite integral. An improper posterior distribution is unacceptable because it is cannot be used to create posterior probability statements. In contrast, assigning a uniform prior distribution to the hierarchical variance itself or its square root leads to proper posterior distributions (see, e.g., Exercise 5.8 of Gelman et al., 1995).

3 Methods of iterative simulation

The essential idea of iterative simulation is to draw values of a random variable θ from a sequence of distributions that converge, as iterations continue, to the desired *target distribution* of θ. For inference about θ, iterative simulation is typically less efficient than direct simulation, which is simply drawing from the target distribution, but iterative simulation is applicable across a much wider range of cases, as current statistical literature makes abundantly clear (see, e.g., Smith and Roberts, 1993, Besag and Green, 1993, and Gilks et al., 1993).

3.1 Rejection sampling

A simple way to draw samples from a target distribution $p(\theta|y)$, called *rejection sampling*, uses an approximate starting distribution $p_0(\theta)$, with two requirements. First, one must be able to calculate $p(\theta|y)/p_0(\theta)$, up to a proportionality constant, for all θ; $w(\theta) \propto p(\theta|y)/p_0(\theta)$ is called the *importance ratio* of θ. Second, rejection sampling requires a known constant M that is no less than $\sup w(\theta)$. The algorithm proceeds in two steps:

1. Sample θ at random from $p_0(\theta)$.

2. With probability $\frac{w(\theta)}{M}$, *reject* θ and return to step 1; otherwise, keep θ.

An accepted θ has the correct distribution $p(\theta|y)$; that is, the conditional distribution of drawn θ, given it is accepted, is $p(\theta|y)$.

The above steps can be repeated to obtain additional independent samples from $p = p(\cdot|y)$. Rejection sampling cannot be used if no finite value of M exists, which will happen when $p_0 = p_0(\cdot)$ has lighter tails than p, as when the support of p_0 is smaller than the support of p. (Hence the use of a multivariate t, instead of a normal, for a starting distribution, in Section 2.) In practice, when p_0 is not a good approximation to p, the required M will be so large that almost all samples obtained in step 1 will be rejected in step 2. The virtue

of rejection sampling as an iterative simulation method is that it is self-monitoring—if the simulation is not effective, you will know it, because essentially no simulated draws will be accepted.

A related method is importance resampling (SIR, sampling-importance resampling, see Rubin, 1987, and Gelman et al., 1995, sec. 10.5). Here a large number of draws are made from $p_0(\theta)$ and a small number are redrawn from this initial set without replacement with probability proportional to $w(\theta)$. No value of M need be selected, and the redrawn values are closer than the initial draws to being a sample from $p(\theta|y)$, but the method is only approximate unless such an M exists and the number of initial draws is infinite. Importance resampling is especially useful for creating a few draws from an approximate distribution to be used as starting points for Markov chain simulation.

Markov chain methods are especially desirable when no starting distribution is available that is accurate enough to produce useful importance weights for rejection sampling or related methods such as importance resampling. With any starting distribution that even loosely covers the target distribution, the steps of a Markov chain simulation directly improve the approximate distributions from which samples are drawn. Thus, the distributions used for taking each draw, themselves converge to p. In a wide range of practical cases, it turns out that the iterations of a Markov chain simulation allow accurate inference from starting distributions that are much too vague for useful results from rejection or importance resampling.

3.2 Data augmentation

Data augmentation is an application of iterative simulation to missing data problems, due to Tanner and Wong (1987), that includes an approximation of the target distribution as a mixture that is updated iteratively. The data augmentation algorithm has two steps: the imputation step, drawing values from a mixture of the posterior distributions of the vector of

missing data, y_{mis}, conditional on observed data y and a set of current draws of the vector of model parameters, θ; and the posterior step, obtaining draws from a mixture of the posterior distribution of the model parameters, θ, given the observed data and a set of current draws of imputed data, y_{mis} (a complete data set). This algorithm bears a strong resemblance to the EM algorithm and can be viewed as a stochastic version of it. Obviously, the data augmentation algorithm requires the ability to draw from the two conditional distributions, $p(y_{\text{mis}}|\theta, y)$ and $p(\theta|y_{\text{mis}}, y)$. The draws from data augmentation converge to draws from the target distribution, $p(\theta, y_{\text{mis}}|y)$ as the iterations continue. Data augmentation can also be viewed as a special case of Gibbs sampling, if only one draw of y_{mis} and one draw of θ is made at each iteration. Recent developments in data augmentation include sequential imputation (Kong, Liu, and Wong, 1994). In this context, it is notationally useful to label θ as θ_1 and y_{mis} as θ_2, with $\theta = (\theta_1, \theta_2)$ the random variable whose distribution is sought.

3.3 Gibbs sampling

Geman and Geman (1984) introduced "Gibbs sampling," a procedure for simulating $p(\theta|y)$ by performing a random walk on the vector $\theta = (\theta_1, \ldots, \theta_d)$, altering one component θ_i at a time. Note that each θ_i can itself be a vector, meaning that the parameters can be updated in blocks. At iteration t, an ordering of the d components of θ is chosen and, in turn, each $\theta_i^{(t)}$ is sampled from the conditional distribution given all the other components:

$$p(\theta_i|\theta_{-i}^{(t-1)}, y),$$

where $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$. When $d = 2$, we have the special case of data augmentation where the approximate distributions are not mixtures. The Gibbs sampler too converges to draws from the target distribution, $p(\theta|y)$.

The optimal scenario for the Gibbs sampler is if the components $\theta_1, \ldots, \theta_d$ are independent in the target distribution; in this case, each iteration produces a new independent draw of θ. If the components are highly correlated, the Gibbs sampler can be slow to converge,

and it is often helpful to transform the parameter space so as to draw from conditional distributions that are more approximately independent.

Obviously, as described, the Gibbs sampler requires the ability to draw from the conditional distributions derived from the target distribution; when this is not possible, the more general Metropolis-Hastings algorithm can be used.

3.4 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970) is a general Markov chain Monte Carlo algorithm that includes Gibbs sampling as a special case. The algorithm proceeds as follows:

1. Draw a starting point $\theta^{(0)}$, for which $p(\theta^{(0)}|y) > 0$, from the *starting distribution*, $p_0(\theta)$.

2. For $t = 1, 2, \ldots$:

   (a) At iteration t, take as input the point $\theta^{(t-1)}$.

   (b) Sample a candidate point $\tilde{\theta}$ from a *proposal distribution* at time t, $J_t(\tilde{\theta}|\theta^{(t-1)})$.

   (c) Calculate the ratio of importance ratios,

   $$r = \frac{p(\tilde{\theta}|y)/p(\theta^{(t-1)}|y)}{J_t(\tilde{\theta}|\theta^{(t-1)})/J_t(\theta^{(t-1)}|\tilde{\theta})}.$$

   (r is always defined, because a jump from $\theta^{(t-1)}$ to $\tilde{\theta}$ can only occur if both $p(\theta^{(t-1)}|y)$ and $J_t(\tilde{\theta}|\theta^{(t-1)})$ are nonzero.)

   (d) Set

   $$\theta^{(t)} = \begin{cases} \tilde{\theta} & \text{with probability } \min(r, 1) \\ \theta^{(t-1)} & \text{otherwise.} \end{cases}$$

This method requires the calculation of the relative importance ratios $p(\theta|y)/J_t(\theta|\theta')$ for all θ, θ', and an ability to draw θ from the proposal distribution $J_t(\theta|\theta')$ for all θ' and t.

The proof that the iteration converges to the target distribution has two steps: first, it is shown that the simulated sequence $(\theta^{(t)})$ is a Markov chain with a unique stationary distribution, and second, it is shown that the stationary distribution equals the target distribution. A mathematical discussion of the conditions for convergence appears in Tierney (1995), and a discussion of the relation between the Metropolis-Hastings algorithm and Gibbs sampling appears in Gelman (1992). Each iteration of a d-step Gibbs sampling algorithm can be viewed as d iterations of a Metropolis-Hastings algorithm for which $r = 1$, so that every jump is accepted.

4 Implementing iterative simulation

4.1 Setting up an iterative simulation algorithm

For some relatively simple models such as hierarchical normal regressions, computations can be performed using data augmentation or Gibbs sampling, drawing each parameter or set of parameters conditional on all the others. More generally, some version of the Metropolis-Hastings algorithm can be used; see Gilks, Richardson, and Spiegelhalter (1996) for many examples. In many cases, setting up a reasonable Metropolis-Hastings algorithm takes substantial programming effort.

Although varieties of Metropolis' algorithm, especially the Gibbs sampler, are becoming popular, they can be easily misused relative to direct simulation: in practice, a finite number of iterations must be used to estimate the target distribution, and thus the simulated random variables are, in general, never from the desired target distribution. It is well known (e.g., Gelman and Rubin, 1992a) that inference from a single sequence of a Markov chain simulation can be quite unreliable. Iterative simulation designs using multiple sequences date back at least to Fosdick (1959); Gelman and Rubin (1992b) discuss multiple sequences in a statistical context, which includes incorporating the uncertainty about θ due to the finiteness of the simulation along with the uncertainty about θ in $p(\theta|y)$ due to the finiteness

of the sample data, y.

When applied to a Bayesian posterior distribution, the goal of iterative simulation is typically inference about the target distribution and not merely some moments of the target distribution. The method of Gelman and Rubin (1992b) and later refinements (Liu and Rubin, 1996) use the variances within and between multiple independent sequences of iterative simulations to obtain approximate conservative inference for the target distribution at any point in the simulation. The method is most effective when the simulations are started from an *overdispersed* starting distribution—one that is at least as spread out as the target distribution itself. A critical point for applications is that a crude approximate distribution that is too spread out to be an effective approximation for importance sampling can be acceptable as an overdispersed starting distribution.

We have always found it useful to simulate at least two parallel sequences, typically four or more. If the computations are implemented on a network of workstations or a parallel machine, it makes sense to run as many parallel simulations as there are free workstations or machine processors. The recommendation to always simulate multiple sequences is not new in the iterative simulation literature (e.g., Fosdick, 1959) but is somewhat controversial (see the discussion of Gelman and Rubin, 1992b, and Geyer, 1992). In our experience with Bayesian posterior simulation, however, we have found that the added information obtained from replication in terms of confidence in simulation results and protection from falsely-precise inferences (see, for example, the figures in Gelman and Rubin, 1992a, and Gelman, 1996) outweighs any additional costs in computer time required for multiple rather than single simulations.

It is desirable to choose starting points that are widely dispersed in the target distribution. Overdispersed starting points are an important design feature for two major reasons. First, starting far apart can make lack of convergence apparent. Second, for purposes of inference, starting overdispersed can ensure that all major regions of the target distribution

are represented in the simulations. For many problems, especially those with discrete or bounded parameter spaces, it is possible to pick several starting points that are far apart by inspecting the parameter space and the form of the distribution. For example, the proportions in a two-component mixture model can be started at values of $(0.1, 0.9)$ and $(0.9, 0.1)$ in two parallel sequences.

In more complicated situations, more work may be be needed to find a range of dispersed starting values. In practice, we have found that the additional effort spent on approximating the target density is useful for understanding the problem and for debugging software: after the Markov chain simulations have been completed, the final estimates can be compared to the earlier approximations. In complicated applied statistical problems, it is standard practice to improve models gradually as more information becomes available, and the estimates from each model can be used to obtain starting points for the computation in the next stage.

4.2 Monitoring convergence and debugging

Markov chain simulation is a powerful tool—so easy to apply, in fact, that there is the risk of serious error, including:

1. Inappropriate modeling: the assumed model may not be realistic from a substantive standpoint or may not fit the data.

2. Errors in calculation or programming: the stationary distribution of the simulation process may not be the same as the desired target distribution, or the algorithm, as programmed, may not converge to any proper distribution.

3. Slow convergence: the simulation can remain for many iterations in a region heavily influenced by the starting distribution, so that the iterations do not accurately summarize the target distribution and yield inappropriate inferences.

The first two errors can occur with other statistical methods (such as maximum likelihood), but the combination of the complexity of Markov chain simulation makes mistakes more common. In particular, it is possible to program a method of computation such as the Gibbs sampler or Metropolis' algorithm that only depends on local properties of the model without ever understanding the large-scale features of the joint distribution. For a discussion of this issue in the context of probability models for images, see Besag (1986).

Much has been written about monitoring the convergence of Markov chain simulations in recent years; recent reviews of the topic and many references appear in Cowles and Carlin (1996) and Brooks and Roberts (1995). Our recommended general approach is based on detecting when the Markov chains have "forgotten" their starting points by comparing several sequences drawn from different starting points and checking that they are indistinguishable.

The potential scale reduction factor. For each scalar summary of interest (that is, all parameters and predictions of interest in the model), Gelman and Rubin (1992b) and Gelman (1996) recommend the following strategy: first discarding the first half of the simulated sequences to reduce the influence of the starting points; and then computing the "potential scale reduction factor," labeled $\sqrt{\widehat{R}}$, which is essentially the square root of the variances of the values of the scalar summary for all the simulated sequences mixed together, divided by the average of the variances within the separate sequences. (Minor corrections to the variance ratio are made to account for sampling variability.) In the limit as the number of iterations in the Markov chain simulation approach infinity, the potential scale reductions $\sqrt{\widehat{R}}$ approach 1, but if the sequences are far from convergence, $\sqrt{\widehat{R}}$ can be much larger. It is recommended to continue simulations until $\sqrt{\widehat{R}}$ is close to 1 (below 1.1 or 1.2, say) for all scalar summaries of interest.

As an example, Figure 1 illustrates the convergence of one of the 122 parameters in a hierarchical nonlinear toxicokinetic model (see Gelman, Bois, and Jiang, 1996, for de-

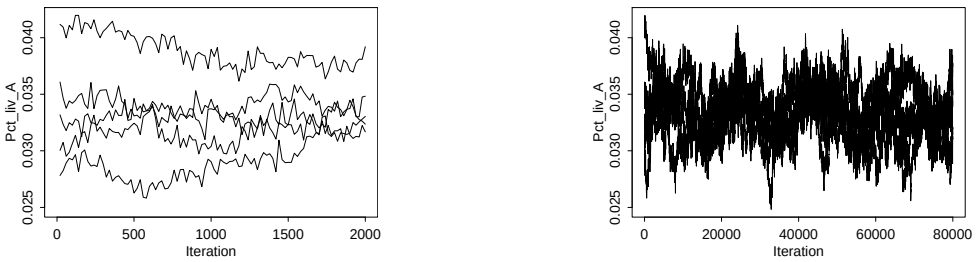

Figure 1: Results of five parallel simulations of a Metropolis algorithm after 2000 iterations and 80,000 iterations for a single parameter of interest (Pct_liv_A, the mass of the liver as a percent of lean body mass for subject A) for a hierarchical nonlinear toxicokinetics model. After 2000 iterations, lack of convergence is apparent; after 80,000, convergence is hard to judge visually. Using the numerical summary given by the potential scale reduction, $\sqrt{\widehat{R}}$: after 2000 iterations, $\sqrt{\widehat{R}}$ (computed from the last half of the simulations; that is, five sequences, each of length 1000) is 1.38; after 80,000 iterations, $\sqrt{\widehat{R}}$ (computed from five sequences, each of length 40,000) decreases to 1.04. (To save memory, only every 20th iteration of the algorithm was recorded.) In practice, the convergence was monitored by running the simulations until $\sqrt{\widehat{R}} < 1.2$ for all 122 parameters in the model. See Gelman, Bois, and Jiang (1996) for details on the model and the simulation.

tails). Five parallel Metropolis-Hastings sequences were simulated (due to the nonlinearity in the model, the conditional posterior distributions did not have closed forms, and so the Gibbs sampler was not possible). Figure 1 displays the results after 2000 and 80,000 iterations, during which the potential scale reduction, $\sqrt{\widehat{R}}$, decreases from 1.38 to 1.04 and the sequences reach approximate convergence.

4.3 Slow convergence

By monitoring convergence of actual simulation runs, it becomes apparent that an MCMC algorithm can be unacceptably slow for many applications, even though it is quite fast and thus acceptable for others. We and others have noticed slowness occurring for three reasons, alone or in combination: (1) the Markov chain moves very slowly through the target distribution, or through bottlenecks of the target distribution (that is, a low "conductance"; see Applegate, Kannan, and Polson, 1990, and Sinclair and Jerrum, 1988); (2) the conditional distributions cannot be directly sampled from, so that each simulation step of the MCMC

algorithm takes substantial computational effort; (3) the function evaluations required to compute the conditional posterior distributions themselves are so slow that an iterative simulation algorithm that is fairly efficient in number of iterations is prohibitively slow in computer time.

A variety of theoretical arguments suggest methods of constructing efficient simulation algorithms or improving the efficiency of existing algorithms. This is an area of much current research; suggested methods in the literature include adaptive rejection sampling (Gilks and Wild, 1992; Gilks, Best, and Tan, 1993), adaptively altering a Metropolis jumping rule (Gelman, Roberts, and Gilks, 1996), reparameterization (Hills and Smith, 1992), adding auxiliary variables and auxiliary distributions to the model (Geyer and Thompson, 1993, Besag et al., 1993), and using early analysis of multiple series to restart the simulations (Liu and Rubin, 1995).

5 Example

We illustrate the methods described here with an application of a mixture model to data from an experiment in psychology. This example is complicated enough that Markov chain simulation methods are the most effective tool for exploring the posterior distribution, but relatively simple in that the model is based on the normal distribution, meaning that all the conditional distributions have simple forms, and computations can be performed using only the Gibbs sampler. The point of this example is not to show the most general variations in computing but rather to illustrate the application of Bayesian computational methods from beginning to end of a problem.

5.1 A study of schizophrenics and non-schizophrenics

In the experiment under study, each of 17 subjects—11 nonschizophrenics and 6 schizophrenics—had their reaction times measured 30 times. We present the data in Figure 2 and briefly

review the basic statistical approach here. More detail on this example appears in Belin and Rubin (1995) and Gelman et al. (1995, ch. 16).

It is clear from Figure 2 that the response times are higher on average for schizophrenics. In addition, the response times for at least some of the schizophrenic individuals are considerably more variable than the response times for the nonschizophrenic individuals. Psychological theory from the last half century and before suggests a model in which schizophrenics suffer from an attentional deficit on some trials, as well as a general motor reflex retardation; both aspects lead to relatively slower responses for the schizophrenics, with motor retardation affecting all trials and attentional deficiency only some.

Finite mixture likelihood model. To address the questions of scientific interest, the following basic model was fit, basic in the sense of minimally addressing the scientific knowledge underlying the data. Response times for nonschizophrenics are described by a normal random-effects model, in which the responses y_{ij} $(i = 1, \ldots, 30)$ of person $j = 1, \ldots, 11$ are normally distributed with distinct person mean α_j and common variance σ_y^2. To reflect the attentional deficiency, the response times for each schizophrenic individual $j = 12, \ldots, 17$ are modeled as a two-component mixture: with probability $(1 - \lambda)$ there is no delay, and the response is normally distributed with mean α_j and variance σ_y^2, and with probability λ responses are delayed, with observations having a mean of $\alpha_j + \tau$ and the same variance, σ_y^2. Because reaction times are all positive and their distributions are positively skewed, even for nonschizophrenics, the above model was fitted to the logarithms of the reaction time measurements.

Hierarchical population model. The comparison of the typical components of $\alpha = (\alpha_1, \ldots, \alpha_{17})$ for schizophrenics versus nonschizophrenics addresses the magnitude of schizophrenics' motor reflex retardation. We include a hierarchical parameter β measuring this

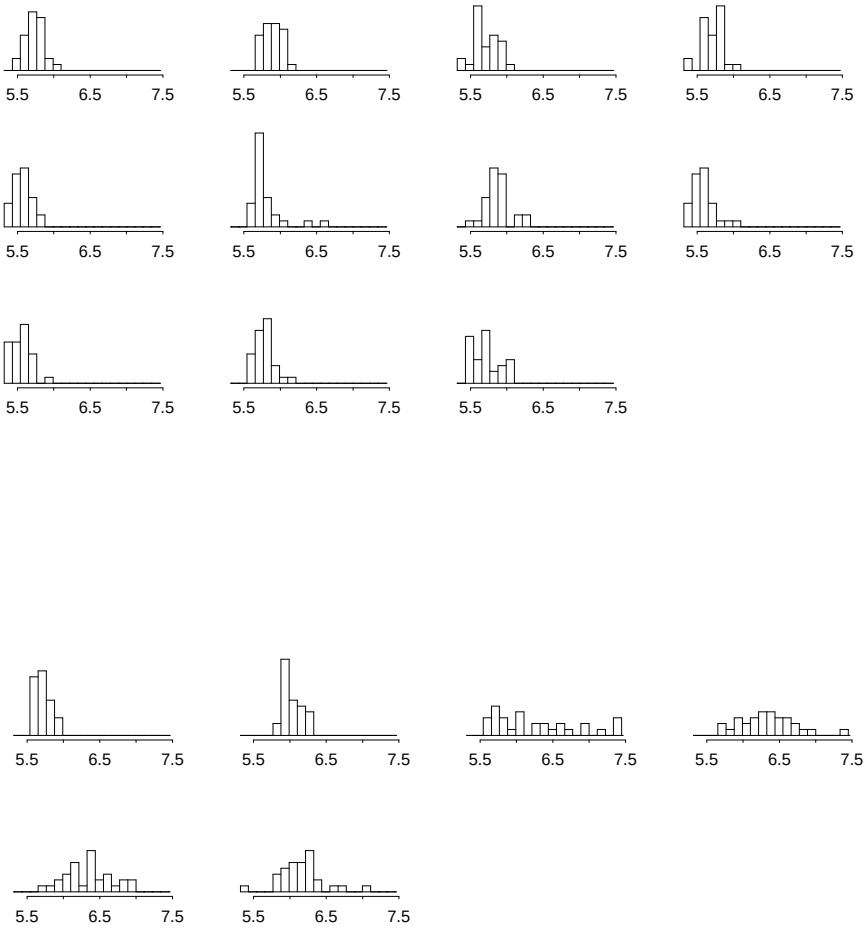

Figure 2: (a) Log response times (in milliseconds) for 11 nonschizophrenic individuals. (b) Log response times for 6 schizophrenic individuals. All histograms are on a common scale, and there are 30 measurements for each individual. From Gelman et al. (1995, ch. 16).

motor retardation. Specifically, variation among individuals is modeled by having the means α_j follow a normal distribution with mean μ for nonschizophrenics and $\mu+\beta$ for schizophrenics, with each distribution having variance σ_α^2. That is, the mean of α_j in the population distribution is $\mu + \beta S_j$, where S_j is an observed indicator variable that is 1 if person j is schizophrenic and 0 otherwise.

We completed the Bayesian model with an improper uniform prior distribution on the hyperparameters $\phi = (\sigma_y^2, \sigma_\alpha^2, \lambda, \mu, \beta, \tau)$. In the experiment at hand, there was adequate information in the data and the hierarchical model to estimate these parameters well enough so that this noninformative prior distribution was acceptable. (To put it another way, posterior inferences would not be sensitive to moderate changes in the prior distribution.)

In probability notation, the full model can be written as:

$$\begin{aligned}
p(y|\alpha,\zeta,\phi) &= \prod_{j=1}^{17}\prod_{i=1}^{30} \mathrm{N}(y_{ij}|\alpha_j, +\tau\zeta_{ij}, \sigma_y^2) \\
p(\alpha|\zeta,\phi) &= \prod_{j=1}^{17} \mathrm{N}(\alpha_j|\mu+\beta S_j, \sigma_\alpha^2) \\
p(\zeta|\phi) &= \prod_{j=12}^{17}\prod_{i=1}^{30} \mathrm{Bernoulli}(\zeta_{ij}|\lambda S_j) \\
p(\phi) &\propto 1,
\end{aligned}$$

where we have introduced ζ, a matrix of indicator variables ζ_{ij} for the schizophrenic observations that take on the value 1 if observation y_{ij} is delayed and 0 otherwise.

The three parameters of primary interest are β, which measures motor reflex retardation, λ, the proportion of schizophrenic responses that are delayed, and τ, the size of the delay when an attentional lapse occurs.

5.2 Approximating the posterior distribution

Crude initial estimate. The first step in the computation is to obtain crude estimates of the model parameters. For this example, each α_j can be roughly estimated by the sample

mean of the observations on subject j, and σ_y^2 can be estimated by the average sample variance within nonschizophrenic subjects. Given the estimates of α_j, we can obtain a quick estimate of the hyperparameters by dividing the α_j's into two groups, nonschizophrenics and schizophrenics. We estimate μ by the average of the estimated α_j's for nonschizophrenics, β by the average difference between the two groups, and σ_α^2 by the variance of the estimated α_j's within groups. We crudely estimate $\hat{\lambda} = 1/3$, and $\hat{\tau} = 1.0$ based on a visual inspection of the histograms of the schizophrenic responses in Figure 2b.

Posterior modes using ECM. We draw 100 points at random from a simplified distribution for ϕ and use each as a starting point for an ECM (expectation conditional maximization) algorithm to search for modes. (ECM is an extension of the EM algorithm; see Meng and Rubin, 1994.) The simplified distribution is obtained by adding some randomness to the crude parameter estimates. Specifically, to obtain a sample from the simplified distribution, we start by setting all the parameters (α, ϕ) at the crude point estimates and then divide each parameter by an independent χ_1^2 random variable in an attempt to ensure that the 100 draws are sufficiently spread out so as to cover the modes of the parameter space.

The ECM algorithm is performed by treating the unknown mixture component corresponding to each schizophrenic observation as "missing data" and then averaging over the resulting vector of 180 missing indicator variables, ζ_{ij}. All steps of the ECM algorithm can then be performed in closed form; see Gelman et al. (1995, ch. 16) for details.

After 100 iterations of ECM from each of 100 starting points, we found three local maxima of (α, ϕ): a major mode and two minor modes. The minor modes are substantively uninteresting, corresponding to near-degenerate models with the mixture parameter λ near zero, and have little support in the data, with posterior density ratios less than e^{-20} with respect to the major mode. We conclude that the minor modes can be ignored and, to the

best of our knowledge, the posterior distribution can be considered unimodal for practical purposes.

Multivariate t approximation. We approximate the posterior distribution by a multivariate t_4, centered at the major mode found by ECM and with covariance matrix set to the inverse of the negative of the numerically-computed second derivative matrix of the log-posterior density. We use the t_4 approximation as a starting distribution for importance resampling (see Rubin, 1987, and Gelman et al., 1995, sec. 10.5) of the parameter vector ϕ. We draw 2000 independent samples of ϕ from the t_4 distribution and importance-resample a subset of 10, which we used as starting points for ten independent Gibbs sampler sequences. This distribution is intended to approximate our ideal starting conditions: for each scalar estimand of interest, the mean is close to the target mean and the variance is greater than the target variance.

5.3 Implementing the Gibbs sampler

The Gibbs sampler is easy to apply for our model once we have performed the "data augmentation" step of including the mixture indicators, ζ_{ij}, in the model. The full conditional posterior distributions have standard forms and can be easily sampled from. The required steps are analogous to the ECM steps used to find the modes of the posterior distribution (once again, details appear in Gelman et al., 1995, ch. 16).

We monitored the convergence of all the parameters in the model for the ten independent sequences of the Gibbs sampler. Table 1 displays posterior inferences and potential scale reduction factors for selected parameters after 20 iterations (still far from convergence, as indicated by the high values of $\sqrt{\widehat{R}}$) and 200 iterations. After 200 iterations, the potential scale reduction factors were below 1.1 for all parameters in the model.

	Inference after 20 iterations				Inference after 200 iterations			
Parameter	2.5%	median	97.5%	$\sqrt{\widehat{R}}$	2.5%	median	97.5%	$\sqrt{\widehat{R}}$
λ	0.05	0.15	0.36	1.9	0.07	0.12	0.18	1.02
τ	0.50	0.78	1.06	1.7	0.74	0.85	0.96	1.02
β	0.13	0.30	0.48	1.2	0.17	0.32	0.47	1.01

Table 1: Posterior quantiles and estimated potential scale reduction factors for some parameters of interest under the old and new mixture models for the reaction time experiment. Ten parallel sequences of the Gibbs sampler were simulated. The table displays inference and convergence monitoring after 20 and then 200 iterations. From Gelman and Rubin (1992b).

5.4 Role of the Markov chain Monte Carlo simulation in the scientific inference process

The Gibbs sampler results allowed us to obtain posterior intervals for all parameters of interest in the model, and also to simulate hypothetical replications of the dataset that could be compared to the observed data. In doing so, we found areas of lack of fit of the model and proceeded to generalize it. It was straightforward to apply the Gibbs sampler to the new model, which had two additional parameters, and then obtain posterior intervals for all the parameters in the expanded model (details appear in Gelman et al., 1995, ch. 16).

The ability to fit increasingly complicated models with little additional programming effort is, in fact, a key advantage of Markov chain Monte Carlo methods. We are no longer limited to those models that we can fit analytically or through elaborate approximations. However, we do not want to understate the effort required in programming these methods for each new problem. As discussed in Sections 2.1 and 2.4, one typically should undertake Markov chain Monte Carlo simulation after a problem has been approximated and explored using simpler methods.

References

Bacchetti, P., Segal, M. R., and Jewell, N. P. (1993). Backcalculation of HIV infection rates (with discussion). *Statistical Science* **8**, 82–119.

Baldi, P., Chauvin, Y., McClure, M., and Hunkapiller, T. (1994). Hidden Markov models of biological primary sequence information, *Proceedings of the National Academy of Science USA*.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B* **48**, 259–302.

Besag, J., and Green, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society B* **55**, 25–102.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall, in preparation.

Clayton, D., and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, ed. P. Elliott, J. Cusick, D. English, and R. Stern, 205–220. Oxford: Oxford University Press.

Cowles, M. K., and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, to appear.

Cowles, M. K., Carlin, B. P. and Connett, J. E. (1993). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data. Research Report 93-007, Division of Biostatistics, University of Minnesota.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. B*, **39**, 1–38.

Fosdick, L. D. (1959). Calculation of order parameters in a binary alloy by the Monte Carlo method. *Physical Review* **116**, 565–573.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.*, **85**, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Stat. Assoc.*, **85**, 398–409.

Gelman, A. (1992). Iterative and non-iterative simulation algorithms. *Computing Science and Statistics* **24**, 433–438.

Gelman, A. (1996). Inference and monitoring convergence. In *Practical Markov Chain Monte Carlo*, ed. W. Gilks, S. Richardson, and D. Spiegelhalter, 131–143. New York: Chapman & Hall.

Gelman, A., Bois, F. Y., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman & Hall.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 599–608. New York: Oxford University Press.

Gelman, A., and Rubin, D. B. (1992a). A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 625–631. New York: Oxford University Press.

Gelman, A., and Rubin, D. B. (1992b). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J., and Thompson, E. A. (1993). Annealing Markov chain Monte Carlo with applications to pedigree analysis. Technical report, School of Statistics, University of Minnesota.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *Journal of the Royal Statistical Society B* **55**, 39–102.

Gilks, W. R. and Richardson, S. (1993). Analysis of disease risks using ancillary risk factors, with application to job-exposure matrices. *Statistics in Medicine* **12**, 1703–1722.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1996). *Practical Markov Chain Monte Carlo.* New York: Chapman & Hall.

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

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association* **60**, 806–825.

Hills, S. E., and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference (with discussion). In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 227–246. New York: Oxford University Press.

Karim, M. R., and Zeger, S. L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **48**, 631–644.

Kong, A., Liu, J. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89**, 278–288.

Kuo, L. and Smith, A. F. M. (1992). Bayesian computation for survival models via the Gibbs sampler. In *Survival Analysis and Related Topics*, ed. J. P. Klien and P. K. Goel. New York: Dekker.

Laird, N. M., and Louis, T. A. (1982) Approximate posterior distributions for incomplete data problems. *Journal of the Royal Statistical Society B* **44**, 190–200.

Lange, N., Carlin, B. P. and Gelfand, A. E. (1992), Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *Journal of the American Statistical Association*, **87**, 615–632.

Liu, C., and Rubin, D. B. (1996). Markov-normal analysis of iterative simulations before their convergence. *Journal of Econometrics*, to appear.

Longford, N. (1993). *Random Coefficient Models.* Oxford: Clarendon Press.

Meng, X. L., and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Meng, X. L., and Rubin, D. B. (1994). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *J. Amer. Stat. Assoc.* **82**, 543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* 395–402, ed. J. Bernardo, Oxford University Press.

Sinclair, A. J., and Jerrum, M. R. (1988). Conductance and the rapid mixing property

of Markov chains: the approximation of the permanent resolved. *Proceedings of the Twentieth Annual Symposium on the Theory of Computing*, 235–244.

Smith, A. F. M., and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society B* **55**, 3–102.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. (1994). BUGS: Bayesian inference using Gibbs sampling, version 0.30. Available from MRC Biostatistics Unit, Cambridge.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, second edition. New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Stat. Assoc.* **82**, 528–550.

Tierney, L. (1995). Markov chains for exploring posterior distributions. *Annals of Statistics.*

West, M. (1992). Modelling with mixtures. In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 503–524. New York: Oxford University Press.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.