Inference from Iterative Simulation Using Multiple Sequences

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

and

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

Technical Report No. 307
June 1992

Department of Statistics
University of California
Berkeley, California 94720

Inference from Iterative Simulation Using Multiple Sequences*

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

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

July 21, 1992

Abstract

The Gibbs sampler, the algorithm of Metropolis, and similar iterative simulation methods are potentially very helpful for summarizing multivariate distributions. Used naively, however, iterative simulation can give misleading answers. Our methods are simple and generally applicable to the output of any iterative simulation; they are designed for researchers primarily interested in the science underlying the data and models they are analyzing, rather than for researchers interested in the probability theory underlying the iterative simulations themselves. Our recommended strategy is to use several independent sequences, with starting points sampled from an overdispersed distribution. At each step of the iterative simulation, we obtain, for each univariate estimand of interest, a distributional estimate and an estimate of how much sharper the distributional estimate might become if the simulations were continued indefinitely. Because our focus is on applied inference for Bayesian posterior distributions in real problems, which often tend toward normality after transformations and marginalization, we derive our results as normal-theory approximations to exact Bayesian inference, conditional on the observed simulations. The methods are illustrated on a random-effects mixture model applied to experimental measurements of reaction times of normal and schizophrenic patients.

*To appear in *Statistical Science*, with discussion. Thanks to John Carlin, Brad Carlin, Tom Belin, Xiao-Li Meng, the editors, and the referees for useful comments, NSF for grant SES-88-05433 and a mathematical sciences postdoctoral fellowship, and NIMH for grants MH-31-154 and MH-31-340. In addition, some of this work was done at AT&T Bell Laboratories.

Keywords: Bayesian inference, convergence of stochastic processes, EM, ECM, Gibbs sampler, importance sampling, Metropolis algorithm, multiple imputation, random effects model, SIR.

1 Introduction

Currently, one of the most active topics in statistical computation is inference from iterative simulation, especially the Metropolis algorithm and the Gibbs sampler (Metropolis and Ulam, 1949; Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; and Gelfand et al., 1990). The essential idea of iterative simulation is to draw values of a random variable x from a sequence of distributions that converge, as iterations continue, to the desired *target distribution* of x. For inference about x, iterative simulation is typically less efficient than direct simulation, which is simply drawing from the target distribution, but iterative simulation is applicable in a much wider range of cases, as current statistical literature makes abundantly clear.

1.1 Objective: applied Bayesian inference

Iterative simulation has tremendous potential for aiding applied Bayesian inference by summarizing awkward posterior distributions, but it has its pitfalls; although we and our colleagues have successfully applied iterative simulation to previously intractable posterior distributions, we have also encountered numerous difficulties, ranging from detecting coding errors to assessing uncertainty in how close a presumably correctly coded simulation is to convergence. In response to these difficulties, we have developed a set of tools that can be applied easily and can lead to honest inferences across a broad range of problems. In particular, our methods apply even when the iterative simulations are not generated from a Markov process. Consequently, we can monitor the convergence of, for example, low-dimensional summaries of Gibbs sampler sequences. We do not pretend to solve all

problems of iterative simulation.

Our focus is on Bayesian posterior distributions arising from relatively complicated practical models, often with a hierarchical structure and many parameters. Many such examples are currently being investigated; for instance, Zeger and Karim (1991) and McCullogh and Rossi (1992) apply the Gibbs sampler to generalized linear models and the multinomial probit model, respectively, and Gilks et al. (1993) review some recent applications of the Gibbs sampler to Bayesian models in medicine. Best results will be obtained for distributions whose marginals are approximately normal, and preliminary transformations to improve normality should be employed, just as with standard asymptotic approximations (e.g., take logarithms of all-positive quantities and logits of quantities that lie between 0 and 1).

1.2 What is difficult about inference from iterative simulation?

Many authors have addressed the problem of drawing inference from iterative simulation, including Ripley (1987), Gelfand and Smith (1990), Geweke (1992), and Raftery and Lewis (1992) in the recent statistical literature. Practical use of iterative simulation methods can be tricky because after any finite number of iterations, the intermediate distribution being used to draw x lies between the starting and target distributions. As Gelman and Rubin (1992) demonstrate for the Ising lattice model, which is a standard application of iterative simulation (Kinderman and Snell, 1980), it is not generally possible to monitor convergence of an iterative simulation from a single sequence (i.e., one random walk). The basic difficulty is that the random walk can remain for many iterations in a region heavily influenced by the starting distribution. This problem can be especially acute when examining a lower-dimensional summary of the multi-dimensional random variable that is being simulated, and can happen even when the summary's

target distribution is univariate and unimodal, as in the Gelman and Rubin (1992) example.

Neither the problem nor the solution is entirely new. Iterative simulation is like iterative maximization; with maximization, one cannot use a single run to find all maxima, and so general practice is to take dispersed starting values and run multiple iterative maximizations. The same idea holds with iterative simulation in the real world; multiple starting points are needed with *finite-length* sequences to avoid inferences being unduly influenced by slow-moving realizations of the iterative simulation. If the parameter space of the simulation has disjoint regions, multiple starting points are needed even with theoretical sequences of infinite length. In general, one should look for all modes and create simple approximations before doing iterative simulation, because by comparing stochastic (i.e., simulation-based) results to modal approximations, we are more likely to discover limitations of both approaches, including programming errors and other mistakes.

1.3 Our approach

Our method is composed of two major steps. First, an estimate of the target distribution is created, centered about its mode (or modes, which are typically found by an optimization algorithm), and "overdispersed" in the sense of being more variable than the target distribution. The approximate distribution is then used to start several independent sequences of the iterative simuation. The second major step is to analyze the multiple sequences to form a distributional estimate of what is known about the target random variable, given the simulations thus far. This distributional estimate, which is in the form of a Student's t distribution for each scalar estimand, is somewhere between its starting and target distributions, and provides the basis for an estimate of how close the simulation process is to covergence—that is, how much sharper the distributional estimate might become if the

simulations were run longer.

With multiple sequences, the target distribution of each estimand can be estimated in two ways. First, a basic distributional estimate is formed, using between-sequence as well as within-sequence information, which is more variable than the target distribution, due to the use of overdispersed starting values. Second, a pooled within-sequence estimate is formed and used to monitor the convergence of the simulation process. Early on, when the simulations are far from convergence, the individual sequences will be less variable than the target distribution or the basic distributional estimate, but as the individual sequences converge to the target distribution, the variability within each sequence will grow to be as large as the variability of the basic distributional estimate.

Multiple sequences help us in two ways. First, by having several sequences, we are able to use the variability present in the starting distribution. In contrast, inference from a finite sample of a single sequence requires extrapolation to estimate the variability that has not been seen. Second, having several independent replications allows easy estimation of the sampling variability of our estimators, without requiring inference about the time-series structure of the simulations. Our use of Student's t reference distributions is analogous to their use in the analysis of linear models, where, in practice, they are generally conditional on a set of simple but insufficient summary statistics; see Pratt (1965) for a discussion of these ideas from a formal Bayesian perspective.

1.4 Remarks

We believe that for any iterative simulation of finite length, valid inference for the target distribution must include a distributional estimate, which reflects uncertainty in the simulation, and also an estimate of how much this distributional estimate may improve as iterations continue. Multiple

independent sequences are essential to routinely obtaining valid inferences from iterative simulations of finite length, and moreover are ideally suited to parallel computing environments. In such environments, running many parallel sequences can be essentially as cheap in computing time as running a single sequence of the same length. Certainly, given the effort needed to formulate scientific models, design real experiments and studies, collect data, conduct exploratory data analyses, reformulate models, and set up one run of an iterative simulation, the extra cost of running independent replications of the same length is trivial in most scientific contexts.

Although we briefly review the problems of *inference* from iterative simulation, we do not discuss the vast and expanding variety of simulation methods themselves, which have been described in several recent review articles; e.g., Tierney (1991), Smith and Roberts (1993), Besag and Green (1993), and Gelman (1992). In fact, one point of our presentation is that the problems of creating simulations and obtaining inferences from simulations are separate to some extent, so our routine methods of inference can be useful for a wide range of simulation methods.

After presenting our suggestions in Section 2 and deriving our inferential methods in Section 3, we apply them in Section 4 to an analysis of a random-effects mixture model, fit to data from a psychological experiment. Our final comments will appear after the discussants to this article have presented their views.

2 An approach to inference from iterative simulation

Our approach to iterative simulation has two major parts: creating an overdispersed approximate distribution from which to obtain multiple starting values, and using multiple sequences to obtain inferences about the target distribution.

2.1 Creating a starting distribution

We seek to begin an iterative simulation with an approximation to the target distribution from which to draw starting values for multiple iterative sequences. Ideally, the starting distribution should be overdispersed but not wildly inaccurate. We find such a distribution in three steps. First, we locate the high-density regions of the (multivariate) target distribution of x, to ensure that our initial values for the iterative simulation do not entirely miss important regions of the target distribution. Second, we create an overdispersed approximation, so that the starting distribution covers the target distribution in the same sense that an approximate distribution for rejection sampling should cover the exact distribution. Third, we downweight draws from the approximate distribution that have relatively low density under the target distribution. These three steps are useful, not only for improving the iterative simulation, but also for better understanding the object of the simulation—the target distribution itself. A variety of methods exist for attacking each of these objectives; here, we present an approach that can often be useful in most statistical problems where the posterior distribution has one or more modes.

First, we find the modes using either an optimization program or a statistical method such as EM (Dempster, Laird, and Rubin, 1977). When a distribution is multimodal, it is necessary to run an iterative mode-finder several times, starting from different points, in an attempt to find all modes. Often, starting values for the mode finder can be found by first discarding information in the data set to obtain inefficient but simpler distributions for the parameters, from which starting values are drawn; this process is illustrated for our example in Section 4.2. Searching for modes is also sensible and commonly done if the distribution is complicated enough that it *may* be multimodal.

Once K modes are found, a second derivative matrix should be estimated

at each mode. Then the high-density regions of 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, fit to the second derivative matrix at each mode. That is, the target density $P(x)$ can be approximated by

$$\hat{P}(x) = \sum_{k=1}^{K} \omega_k (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right), \qquad (1)$$

where d is the dimension of x, and ω_k is the mass of the k-th component of the multivariate normal mixture. The masses ω_k can be calculated by equating the approximate density $\hat{P}$ to the exact density P at the k modes, so that $\hat{P}(\mu_k) = P(\mu_k)$, for $k = 1, \ldots, K$. Assuming the modes are well-separated, this implies that for each k, the mass ω_k is roughly proportional to $|\Sigma_k|^{1/2} P(\mu_k)$.

Second, we obtain samples from an overdispersed distribution by first drawing from the normal mixture and then dividing each sample vector by a positive scalar random variable, an obvious choice being a χ^2_η random deviate divided by η. Making this choice, the new distribution is then a mixture of multivariate t distributions, with the following density function:

$$\tilde{P}(x) \propto \sum_{k=1}^{K} \omega_k |\Sigma_k|^{-1/2} (\eta + (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k))^{-(d+\eta)/2}. \qquad (2)$$

A Cauchy mixture (i.e., $\eta = 1$) is a conservative choice to ensure overdispersion, but if the parameter space is high-dimensional, most draws from a multivariate Cauchy might be too far from the mode to reasonably approximate the target distribution. For most posterior distributions arising in practice, especially those without long-tailed underlying models, a value such as $\eta = 4$, which has three finite moments for $\tilde{P}(x)$, is probably dispersed enough. Further improvements in the approximate distribution can sometimes be obtained by analytically or numerically integrating out nuisance parameters, or by bounding the range of parameter values. Special

efforts may be needed for difficult problems such as banana-shaped posterior distributions in many dimensions, which can arise in practice (e.g., logistic regressions with sparse data in Clogg, Rubin, Schenker, Schultz, and Wideman, 1991).

Third, we sharpen the overdispersed approximation, while keeping it overdispersed, by downweighting regions that have relatively low density under the target distribution. One way of improving the approximation in this way is to use *importance resampling*, also known as SIR (Rubin, 1987b, 1988), which proceeds as follows. First draw N independant samples from the multivariate t mixture (2). For each drawn x, calculate the *importance ratio*, $P(x)/\tilde{P}(x)$, which only needs to be known up to an arbitrary multiplicative constant. Now draw a sample of size m, *without replacement*, from the set of N, as follows. First draw one point from the set of N, with the probability of sampling each x proportional to its importance weight, $P(x)/\tilde{P}(x)$. Then draw a second sample using the same procedure, but from the set of $N-1$ remaining values. Repeatedly sample without replacement $m-2$ more times. Sampling without replacement proportional to the importance weights yields draws from a distribution that lies between $\tilde{P}$ and P. (In the limit as $N \to \infty$, the m importance-resampled draws follow the target distribution, P, under mild regularity conditions.)

We typically sample about $N = 1000$ points from the overdispersed approximate distribution and then draw about $m = 10$ importance-weighted resamples, using larger samples when more than one major mode exists. Although it would be nice if the resultant draws, which we use to start the iterative simulations, were close to draws from the target distribution, this is often not necessary. In contrast to some simulation methods, such as rejection or importance sampling, practical use of Markov chain simulation does not require the starting distribution to be *close* to the target distribution, because in Markov chain simulation, the approximate distribution,

P_t, used for taking draws at time (iteration) t, itself converges to the target distribution, P.

The three steps presented here for creating a starting distribution help avoid pitfalls in a wide variety of problems. Finding the modes is helpful in enabling the iterative simulations to begin roughly centered at the high-density regions of the target distribution. An overdispersed starting distribution allows us to make conservative inferences from multiple sequences of finite length, and is also the key to our method of monitoring convergence. Finally, adjusting the simulations from the starting distribution using importance sampling to make them more typical of the target distribution will generally speed the convergence of the iterative simulation.

Of course, in easy problems, not all of these three preliminary steps will be necessary. In some other cases, far better starting distributions may be found using other methods, for instance, ones that capitalize on analytic and numerical integration methods, such as presented by Tierney and Kadane (1986) and Morris (1988), which are a substantial topic in their own right. In addition, mode-based distributions will not work for every problem; for example, in the Ising distribution discussed in Gelman and Rubin (1992), the multivariate mode of the parent random variable being simulated projects to an extremum in the distribution of the scalar estimand of interest. For particularly difficult problems, the creation of a starting distribution may itself be an iterative process. Any such starting distribution should, however, err on the side of over- rather than under-dispersion. In any case, we believe that the multiple sequences provide critical information beyond that available in one sequence, and that our methods, described in 2.2, will reveal much of this information.

2.2 Prescriptive summary of using components of variance from multiple sequences

Our approach to inference from multiple iteratively simulated sequences examines each scalar estimand of interest separately, and so henceforth we use the notation x to represent a scalar estimand, rather than the whole multicomponent parent random variable being simulated.

We proceed in seven steps.

First, independently simulate $m \geq 2$ sequences, each of length $2n$, with starting points drawn from an overdispersed distribution. To diminish the effect of the starting distribution, discard the first n iterations of each sequence and focus attention on the last n.

Second, for each scalar parameter of interest, calculate

B/n = the variance between the m sequence means, $\overline{x}_{i.}$, each based on n values of x, $B/n = \sum_{i=1}^{m}(\overline{x}_{i.} - \overline{x}_{..})^2/(m-1)$, and

W = the average of the m within-sequence variances, s_i^2, each based on $n-1$ degrees of freedom, $W = \sum_{i=1}^{m} s_i^2/m$.

If only one sequence is simulated, B cannot be calculated.

Third, estimate the target mean, $\mu = \int x P(x) dx$, by $\hat{\mu}$, the sample mean of the mn simulated values of x, $\hat{\mu} = \overline{x}_{..}$.

Fourth, estimate the target variance, $\sigma^2 = \int (x-\mu)^2 P(x) dx$, by a weighted average of W and B, namely

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{1}{n} B, \tag{3}$$

which overestimates σ^2 assuming the starting distribution is appropriately overdispersed, but is unbiased for σ^2 under stationarity (i.e., if the starting distribution equals the target distribution), or in the limit $n \to \infty$. Meanwhile, for any finite n, W should be less than σ^2 because the individual sequences have not had time to range over all of the target distribution and,

as a result, will have less variablility; in the limit as $n \to \infty$, the expectation of W approaches σ^2.

Fifth, estimate what is now known about x. We can improve upon the optimistic (i.e., overly-precise) $N(\hat{\mu}, \hat{\sigma}^2)$ estimate of the target distribution by allowing for the sampling variability of the estimates, $\hat{\mu}$ and $\hat{\sigma^2}$. The result is an approximate Student's t distribution for x with center $\hat{\mu}$, scale $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + B/mn}$, and degrees of freedom $\mathrm{df} = 2\hat{V}^2/\widehat{\mathrm{var}}(\hat{V})$, where

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{V}) &= \left(\frac{n-1}{n}\right)^2 \frac{1}{m}\widehat{\mathrm{var}}(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1}B^2 + \\ &\quad + 2\frac{(m+1)(n-1)}{mn^2} \cdot \frac{n}{m}\left[\widehat{\mathrm{cov}}(s_i^2, \overline{x}_{i\cdot}^2) - 2\overline{x}_{\cdot\cdot}\widehat{\mathrm{cov}}(s_i^2, \overline{x}_{i\cdot})\right], \quad (4)\end{aligned}$$

and the estimated variances and covariances are obtained from the m sample values of $\overline{x}_{i\cdot}$ and s_i^2; $\mathrm{df} \to \infty$ as $n \to \infty$.

Sixth, monitor convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for x might be reduced if the simulations were continued in the limit $n \to \infty$. This potential scale reduction is estimated by $\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}\frac{\mathrm{df}}{\mathrm{df}-2}}$, which declines to 1 as $n \to \infty$. $\hat{R}$ is the ratio of the current variance estimate, $\hat{V}$, to the within-sequence variance, W, with a factor to account for the extra variance of the Student's t distribution. If the potential scale reduction is high, then we have reason to believe that proceeding with further simulations may improve our inference about the target distribution.

Seventh, once $\hat{R}$ is near 1 for all scalar estimands of interest, it is typically desirable to summarize the target distribution by a set of simulations, rather than normal-theory approximations, in order to detect nonnormal features of the target distribution. The simulated values from the last halves of the simulated sequences provide such draws.

A fifty-line program is available in Statlib that implements the above steps in the S computer language.

2.3 Previous methods for monitoring convergence using multiple sequences

Our approach can be viewed as combining and formalizing some ideas from previous uses of multiple sequences to monitor convergence of iterative simulation procedures. Fosdick (1959) simulated multiple sequences, stopping when the difference between sequence means was less than a prechosen error bound, thus basically using B but without comparing it to W. Similarly, Ripley (1987) suggested examining at least three sequences as a check on relatively complicated single-sequence methods involving graphics and time-series analysis, thereby essentially estimating W quantitatively and B qualitatively. Tanner and Wong (1987) and Gelfand and Smith (1990) simulated multiple sequences, monitoring convergence by qualitatively comparing the set of m simulated values at time t to the corresponding set at a later time t'; this approach can be thought of as a qualitative comparison of values of B at two time points in the sequences, without using W as a comparison.

Our approach differs from previous multiple-sequence methods by being fully quantitative in monitoring convergence (i.e., our method is not based on visual inspection of the simulations) and incorporating the uncertainty due to finite-length sequences into the distributional estimates. This reflection of extra uncertainty is analogous to the correction for a finite number of imputations when using multiple imputation to summarize a distribution (Rubin, 1987a).

2.4 Limitations of our method

Multimodal target distributions can give iterative simulation algorithms serious problems because the random walks may take a long time to move from the region of one mode to another. Our analysis should reveal (but not solve) this problem when it occurs: the estimated potential scale reduction will not decline to 1 when different sequences remain in the neighborhoods

of the different modes in which they started. In such cases, the iterative simulation algorithm itself may have to be altered in order to speed convergence, for example, by reparameterizing (Hills and Smith, 1992), adding auxiliary variables (Besag and Green, 1993), or improving the jumping step in the generalized Metropolis algorithm (Green and Han, 1991).

In addition, with better analysis or understanding of the time-series structure of the simulations, the details of our inferential method could be improved upon in a variety of ways; see, Liu (1991) for example. Here, we merely work out the details of the simplest approach that we believe can yield reliable inferences.

3 Derivations for multiple sequence analysis

3.1 Definitions

For our derivations, the previous notation, which suppressed the dependence of the target distribution on parameters (e.g., μ and σ^2 in Section 2.2), will be modified slightly. In particular, let θ denote the parameters of the target distribution, of which μ and σ^2 are functions, so that the target distribution is $P(x|\theta)$, for scalar estimand x.

The target distribution, $P(x|\theta)$, is to be estimated from m independent replications of the following process:

1. A starting point x_0 is drawn at random from a known distribution P_0.

2. A sequence $(x_1, \ldots, x_n)$ is created stochastically by an iterative simulation algorithm, applied to the multivariate random variable of which x is a scalar function. We assume the algorithm eventually converges: $\mathcal{L}(x_n) \to P(x|\theta)$.

The simulations create m sequences:

$$(x_{10}, x_{11}, \ldots, x_{1n})$$

$$\ldots \quad \ldots$$

$$(x_{m0}, x_{m1}, \ldots, x_{mn}).$$

We use the notation (x_{it}) for the matrix of mn simulated values of x. Each sequence is an independent observation of $(x_0, x_1, \ldots, x_n)$, the random vector produced by the iterative simulation algorithm that starts at the random variable x_0. For each $t = 0, 1, \ldots$, we label the mean and variance of the iterate x_t as μ_t and σ_t^2. As $t \to \infty$, μ_t and σ_t^2 approach μ and σ^2, the mean and variance of the target distribution. For notational convenience, we denote the set of all parameters governing the iterative simulation—i.e., the joint distribution of $(x_0, x_1, \ldots)$—by (θ, ξ).

3.2 Approach to inference

We seek a conditional distribution for x, given the mn simulated values (x_{it}),

$$P_{mn}(x) \equiv P(x|(x_{it})),$$

that is valid in the sense of reflecting uncertainty about x due to finite m and n, as well as the uncertainty present in the target distribution itself reflected by nonzero variance σ^2. That is, we want to reflect the fact that $P_{mn}(x) \neq P(x|\theta)$ even though we do assume that $\lim_{n\to\infty} P_{mn}(x) = P(x|\theta)$ for any $m \geq 1$. Intervals for x derived from $P_{mn}(x)$ should ideally have approximately their nominal coverage over the target distribution, $P(x)$. That is, if $\int_I P_{mn}(x)dx = \alpha$, then we ideally want $\int_I P(x)dx \approx \alpha$, where I is an interval about the posterior mean of x, and α is a nominal coverage probability that is typically at least 50%. Usually, errors on the side of conservatism, $\int_I P(x)dx > \alpha$, are more acceptable than liberal errors, and so our approach tends to be conservative, preferring overly-wide to overly-narrow intervals.

For the primary analysis, the distributional estimate $P_{mn}(x)$ is derived

as

$$P_{mn}(x) = \int P(x|\theta,\xi,(x_{it}))P(\theta,\xi|(x_{it}))d(\theta,\xi), \tag{5}$$

where the simulations (x_{it}) are treated as "data" and, by definition,

$$P(x|\theta,\xi,(x_{it})) = P(x|\theta). \tag{6}$$

The logic underlying this analysis is analogous to inference from multiple imputation (Rubin, 1987a); in that framework, "data" refers to the m data sets completed by imputation. The primary analysis makes two assumptions. First,

$$P(x|\theta) = P(x|\mu,\sigma^2), \tag{7}$$

so that the target distribution is known up to a location-scale family, which for our derivations is $N(x|\mu,\sigma^2)$. Second, our primary analysis assumes that that $P(x|\theta) = P_0(x) = P_t(x)$ for all t, an assumption we call *strong stationarity*. The assumption that the starting distribution is equal to the target distribution is certainly unrealistic in practice: if we could start by sampling from the target distribution, we would not need iterative simulation at all. Nevertheless, the ideal case suggests an approach to inference that proves useful in practice.

Once the primary normal-theory analysis suggests that $P_{mn}(x)$ is essentially equal to $P_{m\infty}(x) = P(x|\theta)$, the target distribution $P(x|\theta)$ can be represented by random draws from the last half of the simulated sequences to reflect possible deviations from the assumed normality.

Before deriving $P_{mn}(x)$, we motivate in Section 3.3 the use of the simple statistics $\hat{\mu}$ and $\hat{\sigma}^2$. The remainder of Section 3 derives the Student's t approximation to $P_{mn}(x)$ and the related estimate of potential scale reduction, $\sqrt{\hat{R}}$, which assesses how close $P_{mn}(x)$ is to $P(x|\theta)$ under normality.

3.3 Unbiased estimation of μ and σ^2 under strong normal stationarity

The estimate $\hat{\mu} = \overline{x}_{..}$, the average of the mn simulated values, is unbiased for μ given strong stationarity, $E(\hat{\mu}|\theta,\xi) = \mu$, and would be efficent if the n values in each sequence were exchangeable. (Although not necessary here, we retain ξ because of its later use when we drop the assumption of strong normal stationarity.) Since $\hat{\mu}$ can be expressed as the average of m independent sequence means, the sampling variance of $\hat{\mu}$ is unbiasedly estimated by

$$\begin{aligned} \widehat{\text{var}}(\hat{\mu}) &= \frac{1}{m} \cdot (\text{sample variance of the } m \text{ sequence means } \overline{x}_{i.}) \qquad (8) \\ &= \frac{B}{mn}. \end{aligned}$$

Unbiasedness here means that, over repeated simulations, $E(\frac{B}{mn}|\theta,\xi) = \text{var}(\hat{\mu}|\theta,\xi)$.

Viewing the "data" (x_{it}) as a one-way layout layout with m blocks and n observations per block, B is the usual between-sequence mean square and W is the pooled within-sequence mean square from the analysis of variance. Under strong normal stationarity, the variance components B and W yield an unbiased estimate of σ^2, based on applying the following identity to each sequence $(x_{i1},\ldots,x_{in})$:

$$\frac{1}{n}\sum_{t=1}^{n}(x_{it}-\mu)^2 = \frac{1}{n}\sum_{t}(x_{it}-\overline{x}_{i.})^2 + (\overline{x}_{i.}-\mu)^2. \qquad (9)$$

Taking expectations of both sides under strong stationarity yields

$$\sigma^2 = E(\frac{n-1}{n}s_i^2|\theta,\xi) + \text{var}(\overline{x}_{i.}|\theta,\xi), \qquad (10)$$

where $s_i^2 = \frac{1}{n-1}\sum_t(x_{it}-\overline{x}_{i.})^2$. Unbiased estimates for the two terms $E(s_i^2|\theta,\xi)$ and $\text{var}(\overline{x}_{i.}|\theta,\xi)$ can be expressed in terms of the ANOVA mean squares, B and W.

$$E(\frac{n-1}{n}W|\theta,\xi) = E(\frac{n-1}{n}s_i^2|\theta,\xi) \quad \text{and} \quad E(\frac{1}{n}B|\theta,\xi) = \text{var}(\overline{x}_{i.}|\theta,\xi),$$

whence an unbiased estimate for σ^2 is $\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}B$, from equation (3); under strong stationarity, $E(\hat{\sigma}^2|\theta, \xi) = \sigma^2$.

3.4 Approximate conservative posterior distribution for x

For the purposes of monitoring convergence and obtaining standard inference statements, we approximate the posterior distribution $P_{mn}(x)$ under strong normal stationarity by a Student's t distribution,

$$P_{mn}(x) \approx t_{\mathrm{df}}(x|\hat{\mu}, \widehat{V}), \tag{11}$$

where the squared scale,

$$\widehat{V} = \hat{\sigma}^2 + B/mn, \tag{12}$$

incorporates the predictive variance, $\sigma^2 = \mathrm{var}(x|\theta)$, and also the uncertainty about μ, B/mn.

The distribution (11) is intended to be "conservative" in four senses. First, the estimated distribution is conservative in that it is a Bayesian estimate conditional on insufficient statistics; the estimates of $\hat{\mu}$, $\hat{V}$, and df are based only on the means and variances of the simulations, ignoring any other moments and their time-series structure. Second, the dispersion, $\hat{V}$, is an overestimate, assuming the simulation distributions are overdispersed, and converges to σ^2 as $n \to \infty$. Third, even though we are assuming the (marginal) target distribution of x is normal, we are summarizing what is known about x by a Student's t due to the finite number of simulations used in the estimated scale, $\hat{V}^{1/2}$. The fourth sense in which (11) should be conservative is a consequence of the first three: central confidence intervals, I, derived from $P_{mn}(x)$ should have at least their nominal coverage over the target distribution, $P(x|\theta)$, at least in expectation over repeated iteratively simulated "data" matrices (x_{it}): if I is an interval centered at $\hat{\mu}$, and $\int_I P_{mn}(x)dx = \alpha$, then we should have $E[\int_I P(x|\theta)dx|\theta, \xi] \geq \alpha$.

The t distribution for x is derived in several straightforward steps. Equations (5)–(7) and strong normal stationarity imply,

$$P_{mn}(x) = \int N(x|\mu,\sigma^2)\Pr(\mu,\sigma^2,\xi|(x_{it}))d(\mu,\sigma^2,\xi). \tag{13}$$

Two approximations are used to obtain the t distribution (11) from (13).

First, we conservatively approximate $\Pr(\mu|\sigma^2,\xi,(x_{it}))$ by discarding all information in (x_{it}) except $\hat{\mu}$. The sampling distribution of $\hat{\mu}$ is essentially

$$(\hat{\mu}|\mu,\sigma^2,\xi) \sim N(\mu, \mathrm{var}(\hat{\mu}|\mu,\sigma^2,\xi)), \tag{14}$$

where $\mathrm{var}(\hat{\mu}|\mu,\sigma^2,\xi) = \mathrm{var}(\hat{\mu}|\sigma^2,\xi)$. Combining the sampling distribution (14) with a uniform prior density for μ yields the conditional distribution,

$$(\mu|\sigma^2,\xi,\hat{\mu}) \sim N(\hat{\mu}, \mathrm{var}(\hat{\mu}|\sigma^2,\xi)).$$

Thus, accepting $\Pr(\mu|\sigma^2,\xi,(x_{it})) = \Pr(\mu|\sigma^2,\xi,\hat{\mu})$ gives,

$$P_{mn}(x) = \int N(x \mid \hat{\mu}, \sigma^2 + \mathrm{var}(\hat{\mu}|\sigma^2,\xi))\Pr(\sigma^2,\xi|(x_{it}))d(\sigma^2,\xi).$$

That is, $P_{mn}(x)$ is, conservatively, a mixture of normals with common mean $\hat{\mu}$ and therefore is symmetric and unimodal. Labeling

$$V = \sigma^2 + \mathrm{var}(\hat{\mu}|\sigma^2,\xi),$$

the density $P_{mn}(x)$ may be written as,

$$P_{mn}(x) = \int N(x|\hat{\mu},V)\Pr(V|(x_{it}))dV. \tag{15}$$

In our second approximation, we adopt a Student's t distribution for $P_{mn}(x)$, and determine its degrees of freedom, df, by setting the mixing distribution for the variance, $\Pr(V|(x_{it}))$, equal to $\widehat{V}\cdot \mathrm{df}/\chi^2_{\mathrm{df}}$, where the statistic $\widehat{V} = \hat{\sigma}^2 + B/mn$ in (12) is unbiased for V,

$$E(\widehat{V}|\sigma^2,\xi) = V. \tag{16}$$

Then df is determined by matching the second moment of $\hat{V}$ to the χ^2 mixing distribution, as described in Section 3.5. The location and squared scale of the t distribution are just $\hat{\mu}$ and $\hat{V}$, respectively. As $n \to \infty$, df$\to \infty$, and the approximation to converges to $N(\mu, \sigma^2)$.

3.5 The degrees of freedom for the Student's t distribution

We approximate the sampling distribution of $\hat{V}/V$, conditional on (σ^2, ξ), as

$$\hat{V}/V \sim \chi^2_{\mathrm{df}}/\mathrm{df},$$

with degrees of freedom estimated by the method of moments,

$$\mathrm{df} = 2\frac{\hat{V}^2}{\widehat{\mathrm{var}}(\hat{V}|\sigma^2, \xi)}.$$

Many similar competing estimates for the degrees of freedom can be devised, either by matching a different set of moments or by a more sophisticated approach. No particular choice seemed to be dominant in our initial investigations, so we chose this particular estimator for its simplicity, and because of its previous use in similar problems (Satterthwaite, 1946; Rubin, 1987a).

Each of the terms of

$$\begin{aligned}\mathrm{var}(\hat{V}|\sigma^2, \xi) &= \left(\frac{n-1}{n}\right)^2 \mathrm{var}(W|\sigma^2, \xi) + \left(\frac{m+1}{mn}\right)^2 \mathrm{var}(B|\sigma^2, \xi) + \\ &\quad + 2\frac{(m-1)(n-1)}{mn^2}\mathrm{cov}(W, B\,|\,\sigma^2, \xi) \qquad (17)\end{aligned}$$

can be estimated unbiasedly using multiple independent sequences, assuming strong stationarity. The within mean square, W, is just the average of m iid within-sequence variances s_i^2, and so we estimate its sampling variance by the sample variance of the s_i^2 values, divided by m. The between mean square, B, is the variance of m iid components, and so its sampling distribution is approximately proportional to a χ^2_{m-1} distribution. Matching moments, we estimate the sampling variance of B by $\frac{2}{m-1}B^2$.

Finally, $\text{cov}(W, B \mid \sigma^2, \xi)$ may be estimated using the following expression, which derives from the independence of the m simulated sequences.

$$\begin{aligned}
\text{cov}(W, B|\sigma^2, \xi) &= \text{cov}\left(\frac{1}{m}\sum_i s_i^2, \frac{n}{m-1}\sum_i (\overline{x}_{i.} - \overline{x}_{..})^2 \middle| \sigma^2, \xi \right) \\
&= \frac{n}{m(m-1)} \text{cov}\left(\frac{1}{m}\sum_i s_i^2, \sum_{i,j;i<j}\sum (\overline{x}_{i.} - \overline{x}_{j.})^2 \middle| \sigma^2, \xi \right) \\
&= \frac{n}{m(m-1)}[(m-1)\text{cov}(s_i^2, \overline{x}_{i.}^2 \mid \sigma^2, \xi) - 2(m-1)\text{cov}(s_i^2, \overline{x}_{i.}\overline{x}_{j.} \mid \sigma^2, \xi)] \\
&= \frac{n}{m}[\text{cov}(s_i^2, \overline{x}_{i.}^2 \mid \sigma^2, \xi) - 2E(\overline{x}_{j.} \mid \sigma^2, \xi)\text{cov}(s_i^2, \overline{x}_{i.} \mid \sigma^2, \xi)]. \qquad (18)
\end{aligned}$$

We estimate $\text{cov}(s_i^2, \overline{x}_{i.}^2 \mid \sigma^2, \xi)$ and $\text{cov}(s_i^2, \overline{x}_{i.} \mid \sigma^2, \xi)$ from the corresponding sample covariances, and $E(\overline{x}_{j.})$ is of course estimated by $\hat{\mu} = \overline{x}_{..}$.

Inserting these estimates for $\text{var}(W)$, $\text{var}(B)$, and $\text{cov}(W, B)$ into equation (17) yields the estimate of $\text{var}(\hat{V})$ in equation (4) in Section 2.2. Then, assuming an approximate uniform prior distribution on V yields the conservative posterior distribution: $V/\widehat{V} \sim \text{df}/\chi^2_{\text{df}}$.

3.6 Conservative estimation given overdispersion

The derivation in Section 3.3 of the expectation of the variance estimate $\hat{\sigma}^2$ can be generalized to show that given overdispersion, $\hat{\sigma}^2$ overestimates σ^2. We first use equations (10) and (3), which hold without stationarity or any distributional assumptions, to express $E(\hat{\sigma}^2)$ in terms of the statistics of the sample series,

$$E(\hat{\sigma}^2|\theta, \xi) = E(\frac{n-1}{n}s_i^2|\theta, \xi) + \text{var}(\overline{x}_{i.}|\theta, \xi).$$

We then use the algebraic identity of the right-hand sides of equations (9) and (10) to obtain

$$E(\hat{\sigma}^2|\theta, \xi) = E\left[\frac{1}{n}\sum_t (x_{it} - E(\overline{x}_{i.}|\theta, \xi))^2 \middle| \theta, \xi \right],$$

and finally express the result in terms of the mean and variance of the distribution at iteration t:

$$E(\hat{\sigma}^2|\theta,\xi) = \frac{1}{n}[\sum_t \sigma_t^2 + (\mu_t - E(\overline{x}_{i.}|\theta,\xi))^2] \quad (19)$$
$$\geq \sigma^2 \text{ if } \sigma_t^2 \geq \sigma^2 \text{ for all } t.$$

Because $\hat{\sigma}^2$ is an overestimate given overdispersion, the Student's t distribution (11) is wider than the distribution that would have been obtained under strong stationarity.

3.7 Monitoring convergence

Rather than test the generally false hypothesis that the iterative simulation has "converged," we monitor convergence by estimating the factor by which the scale of the conservative posterior distribution $P_{mn}(x)$ will shrink as $n \to \infty$. The ratio of the information about x in the limiting normal distribution to that in the $t_{\text{df}}(x|\hat{\mu},\widehat{V})$ distribution is given by Fisher (1935, §74):

$$R = \frac{\widehat{V}}{\sigma^2} \cdot \frac{\text{df}}{\text{df}-2}.$$

Equivalently, the potential scale reduction is given by $\sqrt{R}$, which we estimate from our finite simulated sequences. In general (without assuming strong stationarity), no unbiased estimate of σ^2 is possible; however, we can overestimate R by inserting an underestimate of σ^2. Fortunately, many downwardly biased estimates of the target variance σ^2 are available by applying finite-length time-series methods to the m simulated sequences individually. For convenience, we simply use the within-sequence variance W, which seems to work fine in practice. (Under strong stationarity and positively-correlated simulations, $E(W|\theta,\xi) < \sigma^2$ for finite n. Under overdispersion, it is possible that $E(W|\theta,\xi) > \sigma^2$, but in typical examples, $E(\widehat{V}|\theta,\xi)$ exceeds σ^2 by an even greater factor, so that the ratio $E(\frac{\widehat{V}}{W}\,|\,\theta,\xi)$ exceeds 1.)

Thus, we obtain for our estimated scale reduction,

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}\frac{\mathrm{df}}{\mathrm{df}-2}} = \sqrt{\left(\frac{n-1}{n}+\frac{m+1}{mn}\frac{B}{W}\right)\frac{\mathrm{df}}{\mathrm{df}-2}}, \quad (20)$$

which is itself subject to sampling variability. To reflect that uncertainty, we approximate $\frac{B}{W}$ by an F distribution with $m-1$ degrees of freedom for the numerator and $2W^2/\widehat{\mathrm{var}}(W|\theta,\xi)$ for the denominator. (The estimated sampling variance, $\widehat{\mathrm{var}}(W|\theta,\xi) = \frac{1}{m}\mathrm{var}(s_i^2)$, is derived in Section 3.5.) We ignore the minor contribution to variability in the factor $\frac{\mathrm{df}}{\mathrm{df}-2}$. The resulting distribution for $\frac{B}{W}$ overestimates variability, because the true joint sampling distribution of B and W will generally exhibit positive correlation. In practice, we are concerned if the scale reduction is large, but not if it is small, so we report only the estimated potential scale reduction and its upper 97.5% confidence limit (see the center columns of Figures 2 and 3), both of which we regard as conservative (i.e., overestimates of $\sqrt{R}$).

When the potential scale reduction is large, one (or both) of the following statements must be true:

1. Further simulation will substantially decrease $\hat{\sigma}$; i.e., inference about x can be made more precise by allowing the simulations to converge to the target distribution, so that $\hat{\sigma} \to \sigma$.

2. Further simulation will substantially increase W; i.e., the simulated sequences have not each made a complete tour of the target distribution.

In either case, the stochastic process of iterative simulation, at least as represented by the last half of the sequences, is far from convergence to the stationary distribution.

When the potential scale reduction is near 1, we conclude that each set of the m sets of n simulated values is close to the target distribution. The discussion of Tables 2, 3, and 4 from our example in Section 4 illustrates how the estimated ineffiency can be used to monitor convergence in practice.

Once the simulation is done, and approximate convergence is assessed, it is important to check that the key assumption of overdispersion has been satisfied. A direct way of assessing overdispersion is to see whether the early intervals have conservative coverage over later distributions, as illustrated by Figure 3 and Table 3 in our example in Section 4.

4 Example

4.1 The data and model

Psychologists at Harvard University (P. Holtzman, H. Gale, and S. Levin) performed an experiment measuring thirty reaction times for each of seventeen subjects: eleven non-schizophrenics and six schizophrenics. Belin and Rubin (1990, 1992) describe the data and several probability models in detail; we present the data in Figure 1 and briefly review their basic approach here.

It is clear 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 non-schizophrenic individuals. Current psychological theory 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 a delay in the schizophrenics' responses, with motor retardation affecting all trials, and attentional deficiency only some.

To address the questions of scientific interest, the following basic model was fit. Response times for non-schizophrenics are thought of as arising from a normal random effects model, in which the responses of person $i =$

$1, \ldots, 11$ are normally distributed with distinct person mean α_i and common variance σ^2_{obs}. To reflect the attentional deficiency, the response times for each schizophrenic individual $i = 12, \ldots, 17$ are fit to a two-component mixture: with probability $(1 - \lambda)$, there is no delay and the response is normally distributed with mean α_i and variance σ^2_{obs}, and with probability λ, responses are delayed, with observations having a mean of $\alpha_i + \tau$ and the

Figure 1a: Log response times for 11 non-schizophrenic individuals

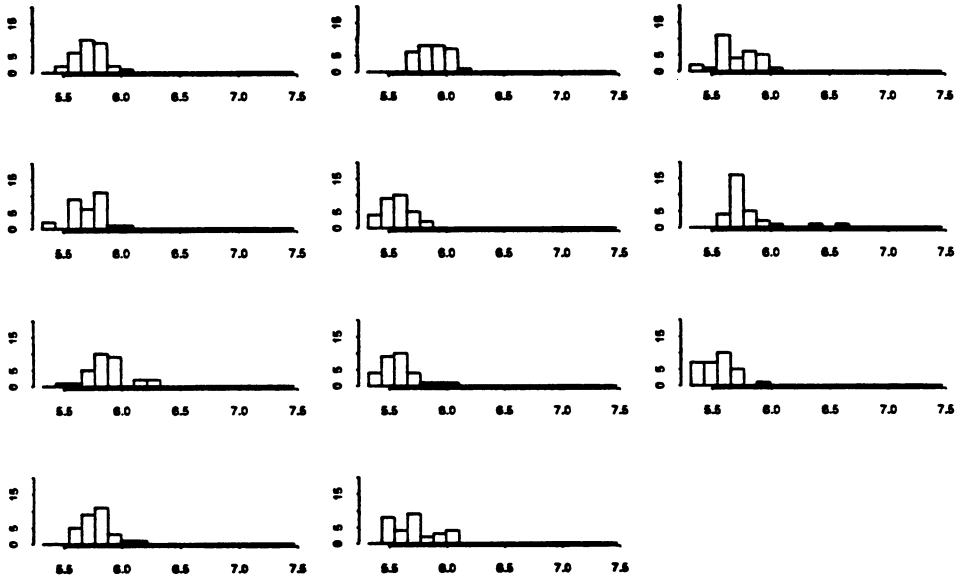
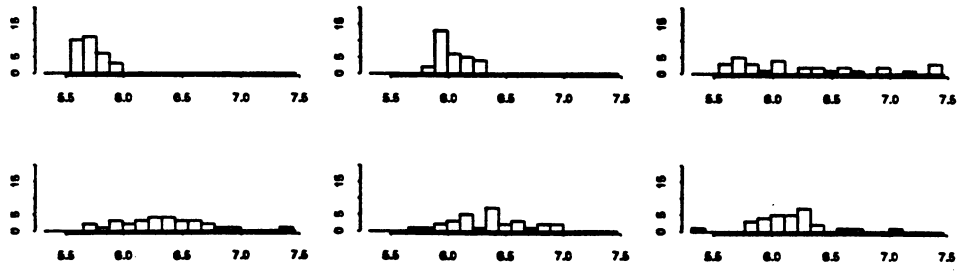


Figure 1b: Log response times for 6 schizophrenic individuals



same variance. Because the reaction times are all positive and their distributions are positively skewed, even for non-schizophrenics, the model was fit to the logarithms of the reaction time measurements.

The comparison of the components of $\alpha = (\alpha_1, \ldots, \alpha_{17})$ for schizophrenics versus non-schizophrenics addresses the magnitude of schizophrenics' modal retardation. We modify the basic model of Belin and Rubin (1992) to incorporate a hierarchical parameter β measuring motor retardation. Specifically, variation among individuals is modeled by having the person means α_i follow a normal distribution with mean $\nu + \beta S_i$ and variance σ_α^2, where ν is the overall mean response time of non-schizophrenics, and the observed indicator S_i is 1 if person i is schizophrenic and 0 otherwise.

Letting y_{ij} be the jth response of individual i, the model can be written in the following hierarchical form.

$$\begin{aligned} y_{ij}|\alpha_i, z_{ij}, \phi &\sim N(\alpha_i + \tau z_{ij}, \sigma_{obs}^2), \\ \alpha_i|z, \phi &\sim N(\nu + \beta S_i, \sigma_\alpha^2), \\ z_{ij}|\phi &\sim \text{Bernoulli}(\lambda S_i), \end{aligned}$$

where $\phi = (\log(\sigma_\alpha^2), \beta, \text{logit}(\lambda), \tau, \nu, \log(\sigma_{obs}^2))$, and z_{ij} is an unobserved indicator variable that is 1 if measurement j on person i arose from the delayed component and 0 if it arose from the undelayed component. The indicator random variables z_{ij} are not necessary to formulate the model, but allow convenient computation of the modes of (α, ϕ) using the iterative ECM algorithm (Meng and Rubin, 1992) and simulation using the Gibbs sampler. For the Bayesian analysis, the parameters σ_α^2, β, λ, τ, ν, and σ_{obs}^2 are assigned a joint uniform prior distribution, except that λ is restricted to the range $[.001, .999]$, τ is restricted to be positive to identify the model, and σ^2 and σ_{obs}^2 are of course restricted to be positive.

The three parameters of primary interest are: β, which measures motor reflex retardation; λ, the proportion of schizophrenic responses that

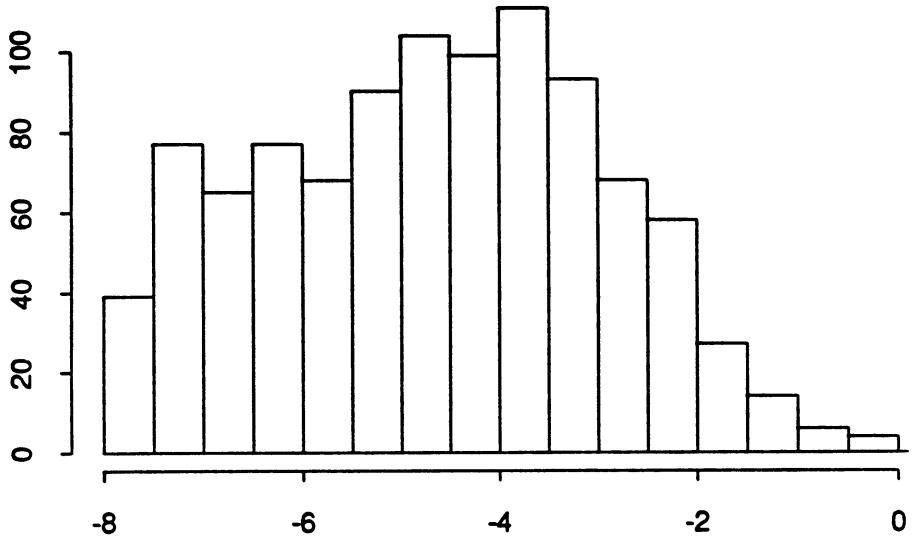
are delayed; and τ, the size of the delay when an attentional lapse occurs. Substantive analyses of this basic model are presented in Belin and Rubin (1992), who show that this model, chosen as the simplest model to include all the substantive parameters of interest and reasonably fit the data, does not provide an entirely satisfactory fit to the data, and suggest some extensions. Here, we focus on computational issues and use the data and the basic model for this purpose.

4.2 Creating an approximate distribution

Our model cannot be fit to the data in closed form. We began our exploration of the posterior distribution of (α, ϕ) by drawing 50 points at random from a simplified distribution for (α, ϕ) and using each as a starting point for the ECM maximizing routine, to search for modes. In the simplified distribution, there were no random effects and no mixture components; the parameters τ and logit(λ) were both set to zero, and each of the 17 subject effects, α_i, was independently sampled from a normal distribution based on the mean and variance of the 30 observations from that subject. The hyperparameters $\log(\sigma_\alpha^2)$, β, and ν were estimated by (closed-form) maximum likelihood conditional on the drawn subject effects, and each was then then perturbed by dividing by independent χ_1^2 random deviates in an attempt to cover the modes of the parameter space with marginal Cauchy distributions.

Three local maxima of (α, ϕ) were found: a major mode and two minor modes. The minor modes were substantively uninteresting, corresponding to near-degenerate models with the mixture parameter λ near zero, and had little support in the data, with posterior density ratios with respect to the major mode below e^{-20}. We concluded that the minor modes could be ignored and, to the best of our knowledge, the target distribution could be considered unimodal for practical purposes. To put it another way, the importance ratios at the minor modes were so low, we simply discarded

Figure 2: Logarithms of the largest importance ratios from the multivariate t approximation



them in our approximation, before going to the work of estimating associated second derivatives and forming the mixture approximation. Had we included the minor modes, any draws from them would have had essentially zero importance weights and would almost certainly have not appeared in the importance-weighted resamples.

Random samples of (α, ϕ) were drawn from $\tilde{P}$, the multivariate t approximation, with $\eta = 4$ degrees of freedom, centered at the major mode with scale determined by the second derivative matrix at the mode, which was computed by numerical differentiation. An alternative would have been to use the SECM algorithm (Meng and Rubin, 1991). The corresponding exact posterior distribution of (α, ϕ), P, has the indicators z_{ij} integrated out, and is thus an easily computed product of mixture forms. The number of independent samples drawn was $N = 2000$, and a histogram of the

relative values of 1000 largest log importance weights, presented in Figure 2, shows little variation—an indication of the adequacy of the overdispersed approximation as a basis for taking draws to be resampled to create the starting distribution.

4.3 Applying the Gibbs sampler

A set of $m = 10$ starting points was drawn by importance-weighted resampling (SIR), as described in Section 2.1, and represents the starting distribution, P_0. 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.

The Gibbs sampler is easy to apply for our model because the full conditional posterior distributions—$P(\phi|\alpha, z)$, $P(\alpha|\phi, z)$, and $P(z|\alpha, \phi)$—have closed form and can be easily sampled from. We simulated ten independent sequences of 200 iterations each (all the variables were updated in each iteration), and we then examined the results using the method described in Section 2.2.

4.4 Detailed results for a single parameter

For exposition, we first display detailed results for a single scalar estimand—τ, the increase in log reaction time of schizophrenics due to lapse of attention. In fact, this parameter was one of the slower estimands to converge in the iterative simulation.

Figure 3 shows the paths of three simulated sequences (randomly chosen from the ten used for the computation) for the first 50 iterates of τ. The vertical bars at 10, 20, 30, 40, and 50 are the conservative 95% posterior intervals for τ based on iterations 6–10, 11–20, 16–30, 21–40, and 26–50, respectively, of the ten simulated sequences (i.e., intervals from our procedure with $m = 10$ and $n = 5, 10, 15, 20$, and 25, respectively). The final vertical

Figure 3: Multiple simulations and 95% posterior intervals for τ

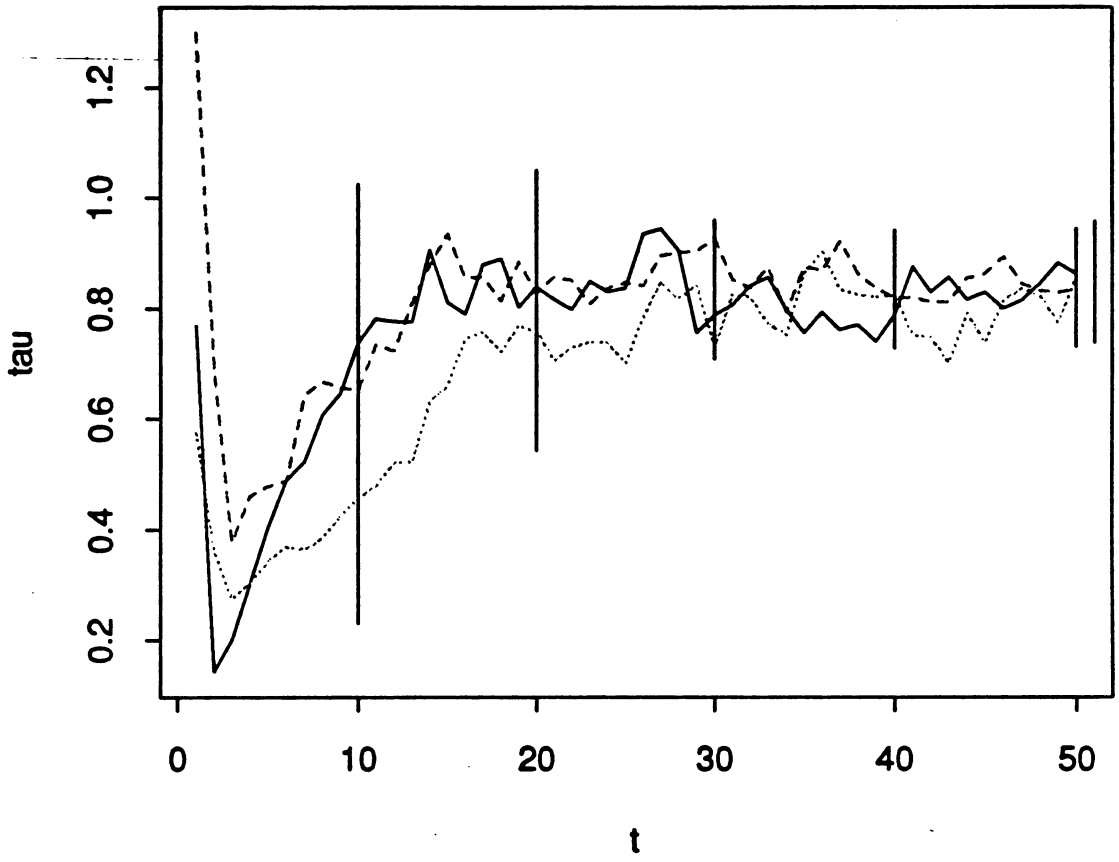


Table 1: Relative eventual reductions in posterior interval widths for τ

$2n$ = number of iterations	10	20	30	40	50
Estimated potential reduction factor	3.4	1.7	1.3	1.1	1.1
and its 97.5% quantile	5.3	2.4	1.6	1.3	1.1
Actual reduction factor, relative to $2n = 200$	3.9	2.5	1.3	1.0	1.0

bar just to the right of 50 is the interval based on iterations 101–200 (i.e., $n = 100$). The intervals are Student's t intervals centered at $\hat{\mu}$, with scale $\sqrt{\widehat{V}}$ and degrees of freedom df as discussed in Section 2.2.

Clearly, the early intervals include the later intervals, and each is "conservative" in that it includes more than its nominal 95% coverage of the probability under the target distribution. The greater width of the early intervals partly reflects sampling variability due to finite m and n, but the main cause is that, when the starting distribution is overdispersed, $\hat{\sigma}^2$ overestimates the target variance, σ^2. The later intervals rapidly approach stability, and are approximately correct because as n increases, the bias of $\hat{\sigma}^2$ decreases to zero.

Watching the posterior intervals for τ approach stability is comforting, but to measure their convergence, we need a standard of comparison. The first two rows of Table 1 show the estimated potential scale reduction, $\sqrt{\widehat{R}}$, and its 97.5% quantile, which together give a rough indication of the factor by which we expect the posterior intervals may shrink under continued simulation. (The upper quantile comes from the approximate sampling distribution of $\sqrt{\widehat{R}}$, derived in Section 3.7.) The estimated improvement factors displayed are based on only the first 10, 20, 30, 40, and 50 iterations of the ten sequences. The last row of the table shows the actual eventual reductions in interval width, using the results based on iterations 101–200 as a standard of comparison. Even the early estimates of potential reductions in interval width appear to have been reasonably accurate.

4.5 Results for other parameters of interest

Section 4.4 showed every step of our analysis for a single scalar estimand, τ; in practice, one will typically examine several estimand, but in less detail. In fact, as we now show, the iterative simulation process can be monitored reliably without examining any time-series graphs like Figure 3.

Table 2: Inference for scalar estimands based on 10 sequences, iterations 101–200

	Normal-theory posterior interval			Potential scale reduction		Simulated quantiles				
	2.5%	$\hat{\mu}$	97.5%	est.	97.5%	2.5%	25%	median	75%	97.5%
α_1	5.66	5.73	5.80	1.00	1.00	5.66	5.71	5.73	5.76	5.80
α_2	5.82	5.89	5.95	1.00	1.00	5.82	5.86	5.89	5.91	5.95
α_3	5.64	5.71	5.78	1.00	1.01	5.65	5.69	5.71	5.73	5.78
α_4	5.64	5.71	5.77	1.00	1.02	5.64	5.68	5.71	5.73	5.77
α_5	5.51	5.58	5.65	1.00	1.01	5.51	5.56	5.58	5.60	5.65
α_6	5.73	5.80	5.86	1.00	1.00	5.73	5.77	5.80	5.82	5.86
α_7	5.79	5.86	5.92	1.00	1.00	5.79	5.83	5.86	5.88	5.92
α_8	5.52	5.59	5.66	1.00	1.00	5.52	5.56	5.59	5.61	5.65
α_9	5.48	5.55	5.62	1.00	1.00	5.49	5.53	5.55	5.57	5.62
α_{10}	5.71	5.77	5.84	1.00	1.01	5.71	5.75	5.77	5.80	5.84
α_{11}	5.65	5.72	5.78	1.00	1.01	5.65	5.69	5.72	5.74	5.78
α_{12}	5.66	5.73	5.80	1.00	1.00	5.66	5.71	5.73	5.75	5.80
α_{13}	5.97	6.03	6.10	1.00	1.00	5.96	6.01	6.03	6.05	6.10
α_{14}	5.93	6.01	6.09	1.00	1.01	5.93	5.98	6.01	6.04	6.09
α_{15}	6.08	6.19	6.29	1.03	1.07	6.08	6.15	6.19	6.22	6.29
α_{16}	6.11	6.19	6.27	1.01	1.03	6.10	6.16	6.19	6.22	6.26
α_{17}	6.00	6.07	6.14	1.01	1.02	5.99	6.04	6.07	6.09	6.14
σ_α	0.09	0.14	0.21	1.00	1.00	0.10	0.12	0.14	0.16	0.21
β	0.17	0.32	0.47	1.01	1.02	0.17	0.27	0.32	0.37	0.48
λ	0.07	0.12	0.19	1.02	1.04	0.07	0.10	0.12	0.14	0.18
τ	0.74	0.85	0.96	1.02	1.05	0.74	0.81	0.85	0.88	0.96
σ_{obs}	0.18	0.19	0.20	1.01	1.02	0.18	0.18	0.19	0.19	0.20
$\sigma_\alpha/\sigma_{obs}$	0.50	0.74	1.10	1.00	1.00	0.51	0.64	0.73	0.85	1.11
$-2\log(\text{density})$	727.81	747.33	766.86	1.01	1.01	729.98	739.92	746.88	753.84	768.35

We computed several univariate estimands: 17 random effects α_i and their standard deviation σ_α, the shift parameters τ and β, the standard deviation of observations σ_{obs}, the mixture parameter λ, the ratio of standard deviations $\sigma_\alpha/\sigma_{obs}$, and $-2\log(\text{posterior density})$. Table 2 shows the results of our multicomponent analysis of the ten sequences of length $2n = 200$, with each row of the table presenting a summary inference for a different univariate estimand. The first three columns summarize the posterior distribution by the 95% central interval based on the t distribution introduced in Section 2.2 and derived in Section 3.4. (The interval for τ corresponds to the rightmost vertical bar displayed in Figure 3.) The next two columns present the estimated potential scale reduction, $\sqrt{\hat{R}}$, along with an upper limit derived from its approximate sampling distribution. The last five columns of Table 2 will be discussed in Section 4.6.

Because the estimated potential scale reductions of Table 2 are close to 1, they suggest that further simulation will not markedly improve our estimates of the scalar estimands shown. More precise estimation of the means and variances of the target distributions, as would be achieved by further simulation, would not narrow the estimated posterior intervals much—nearly all the width of the intervals is due to the posterior variances themselves, not uncertainty due to simulation variability.

For comparison, Table 3 shows the corresponding results after only $2n = 20$ iterations of the series. The posterior means of many of the estimands (e.g., λ) are not estimated accurately, and for several of them, eventual scales could easily shrink by a factor of two or more. (The estimated scale reduction for τ and its upper limit are the same as the values in Table 1 based on $2n = 20$.) Because the simulations at $2n = 20$ are so far from convergence, we do not bother to present the simulated quantiles in Table 3. From the potential scale reductions of Table 3, we expect that further simulation beyond $n = 10$ would sharpen the posterior intervals for two

Table 3: Inference for scalar estimands based on 10 sequences, iterations 11–20

	Normal-theory posterior interval			Potential scale reduction		Coverage prob. of 95% intervals, relative to dists. from Table 2	
	2.5%	$\hat{\mu}$	97.5%	est.	97.5%	normal-theory	simulated quantiles
α_1	5.66	5.73	5.80	1.02	1.09	0.96	0.97
α_2	5.83	5.88	5.94	1.03	1.11	0.92	0.92
α_3	5.65	5.71	5.78	1.06	1.17	0.95	0.96
α_4	5.65	5.70	5.76	1.02	1.09	0.91	0.92
α_5	5.52	5.59	5.65	1.00	1.05	0.95	0.95
α_6	5.73	5.79	5.86	1.01	1.07	0.95	0.94
α_7	5.79	5.86	5.92	0.98	1.01	0.93	0.94
α_8	5.53	5.59	5.66	1.10	1.26	0.94	0.94
α_9	5.48	5.55	5.62	1.02	1.09	0.96	0.96
α_{10}	5.71	5.77	5.84	1.04	1.14	0.92	0.92
α_{11}	5.66	5.72	5.78	1.13	1.30	0.93	0.93
α_{12}	5.61	5.73	5.84	1.22	1.54	1.00	1.00
α_{13}	5.90	6.01	6.12	1.21	1.53	1.00	1.00
α_{14}	5.94	6.02	6.11	1.09	1.22	0.96	0.95
α_{15}	5.98	6.14	6.30	1.50	2.01	0.99	0.99
α_{16}	6.04	6.16	6.29	1.39	1.76	0.99	0.99
α_{17}	5.94	6.05	6.16	1.45	1.91	0.99	0.99
σ_α	0.09	0.14	0.22	1.04	1.13	0.98	0.98
β	0.13	0.30	0.48	1.18	1.44	0.98	0.98
λ	0.05	0.15	0.36	1.88	2.73	1.00	1.00
τ	0.50	0.78	1.06	1.67	2.40	1.00	1.00
σ_{obs}	0.18	0.19	0.21	1.12	1.28	0.98	0.98
$\sigma_\alpha/\sigma_{obs}$	0.45	0.73	1.16	1.04	1.13	0.98	0.98
$-2\log$(density)	642.45	775.71	908.97	1.79	3.45	1.00	1.00

reasons—more degrees of freedom for estimation and a lower estimated posterior variance. Comparison to Table 2 shows that the posterior intervals using 200 iterations do indeed narrow. But, *without ever having seen Table 2* or any simulations beyond $2n = 20$, one can tell from the high estimated potential scale reductions of Table 3 that the first 20 iterations of the simulated sequences do not summarize the target distribution as accurately as will continued simulation. In addition, it is reassuring to see that most of the posterior intervals in the first three columns of Table 3 wholly contain the more accurate intervals in Table 2. The last two columns of Table 3 give the probability coverage of the nominal 95% intervals based on $2n = 20$, using the distributions summarized in Table 2 as references.

4.6 Estimating the target distribution by the set of simulations

Now that the estimated eventual reductions in interval width are low for all 24 scalar estimands of interest, we conclude that, for each estimand, the ten sequences of length 100 (iterations 101–200) may be considered to consist of draws from the target distribution, at least under our normal-based analysis. Consequently, we might drop the normality assumption and consider the $10 \times 100 = 1000$ values for each estimand as draws from its target distribution.

The rightmost five columns of Table 2 show the simulated quantiles of the 24 estimands of interest, based on the 1000 simulated values. The 2.5%, 50%, and 97.5% quantiles agree quite well with the t-intervals presented in the left columns of the table; the discrepancies between the normal-theory and empirical 2.5% and 97.5% points for λ suggest that its marginal posterior distribution is not normal, even after the logit transformation.

At this point we also might be willing to tentatively summarize the joint distribution of *all* 22 simulated parameters by the 10×100 array of

values from the last half of the simulated sequences. Of course it is possible that some function of the parameters has not yet adequately converged, but then our normal-theory methods should detect this when applied to that estimand, and still provide a conservative distributional estimate. For example, this process was followed for $\sigma_\alpha/\sigma_{obs}$, as seen in Tables 2 and 3.

4.7 Inference about functionals of the target distribution

Once the convergence is judged to be adequate, the m independently simulated sequences can also be used to summarize any functional ψ of the multivariate target distribution, using the following method:

1. Assess the convergence of the scalar functions of the multivariate random variable that are needed to calculate ψ.

2. For each simulated sequence i, calculate the sample value ψ_i based on the empirical distribution of the n simulated iterates.

3. Create an interval for ψ based on the independent estimates $\psi_i, i = 1, \ldots, m$.

For example, let ψ be the posterior correlation between the parameters τ and λ. For ψ to be well-estimated from the simulations, the distributions of τ, λ, and $\tau\lambda$ should have approximately converged to the target distribution; τ and λ have already been judged to have adequately converged, as evidenced by their estimated potential scale reductions in Table 3. Applying our procedure to $\tau\lambda$ yields a potential scale reduction factor estimated at 1.01 with a 97.5% upper bound of 1.02, and so we are satisfied that the sequences have effectively converged for the purpose of estimating ψ. For each $i = 1, \ldots, 10$, we calculate the sample correlation of the 100 iterates $((\tau, \lambda)_{it}, t = 101, \ldots, 200)$; the results are $\{-0.282, -0.275, -0.284, -0.231, -0.256, -0.280, -0.374, -0.397, -0.413, -0.071\}$. To obtain a

simple normal-theory posterior interval for ψ, we transform the ten sample correlations to the Fisher z-scale, in which their mean is $-.297$ with standard error 0.034 on 9 degrees of freedom. Transforming back to the original scale yields an estimate of $-.288$ for ψ with a 95% posterior interval of $[-.357, -.218]$.

For another example, let ψ be the 75% quantile of the posterior distribution of τ. The distribution of τ appears, from the estimated potential scale reductions of Table 3, to have effectively converged, and so we summarize our knowledge about ψ by the values ψ_i based on iterations 101–200 of the ten simulated sequences, {0.906, 0.876, 0.881, 0.878, 0.876, 0.873, 0.872, 0.885, 0.884, 0.912}, which have a mean of 0.884 with standard error 0.004 on 9 degrees of freedom.

4.8 An example of slow convergence

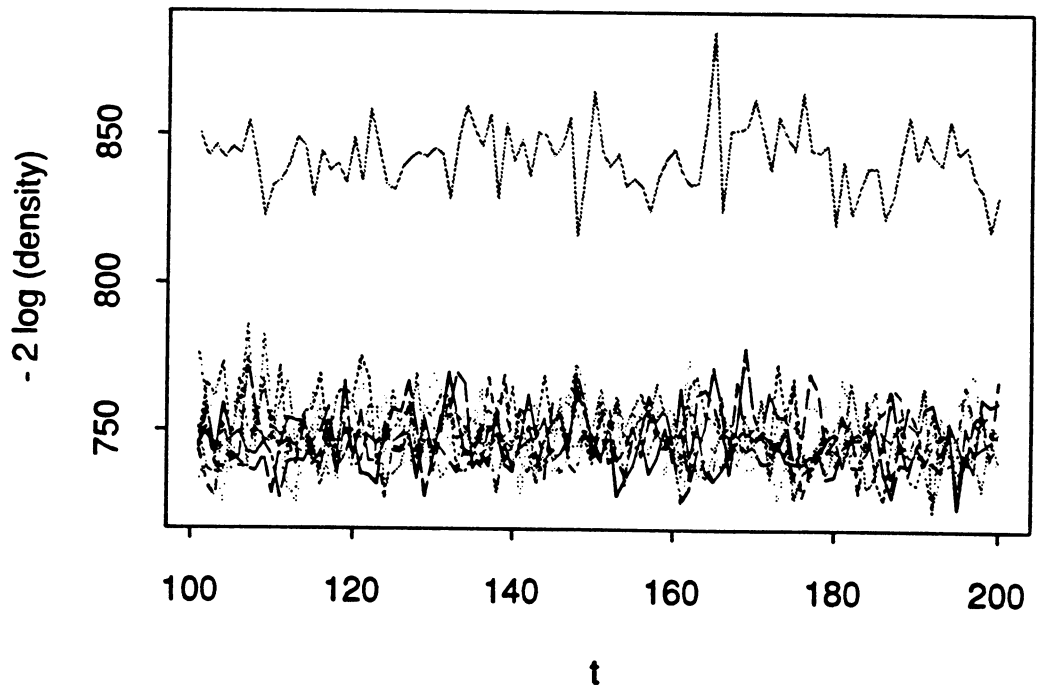
Even with the data and model of Section 4, it is possible for the Gibbs sampler to exhibit slow convergence. To illustrate this point and how our method of analysis in Section 2.2 handles slow convergence, we sample ten new sequences for 200 steps. This time, however, we draw the ten starting points directly from the initial approximate distribution described in the first paragraph of Section 4.2, without searching for modes or using importance-weighted resampling to eliminate relatively unlikely points. Table 4 presents the results; for brevity, the inferences for the components of α are omitted.

The high potential scale reductions clearly show that the simulations are far from convergence. To understand better what is happening, we plot the m sequences of log posterior densities. Figure 4 shows the last halves of the ten time series, superimposed. The single sequence that stands alone started and remains in the neighborhood of one of the minor modes found earlier by maximization. Since the minor mode is of no scientific interest and has negligible support in the data (note its relative density), we simply discard

Table 4: Inference for some scalar estimands based on a new set of 10 sequences, iterations 101–200

	Normal-theory posterior interval			Potential scale reduction	
	2.5%	$\hat{\mu}$	97.5%	est.	97.5%
σ_α	0.09	0.15	0.25	1.31	1.58
β	−0.17	0.27	0.70	2.43	3.63
λ	0.01	0.16	0.74	5.15	7.42
τ	0.70	0.84	0.98	1.19	1.35
σ_{obs}	0.18	0.19	0.20	1.10	1.21
$\sigma_\alpha/\sigma_{obs}$	0.47	0.78	1.28	1.25	1.48
−2 log(density)	681.63	757.18	832.74	3.53	5.08

Figure 4: Log posterior densities for the new set of ten simulated sequences



this sequence, as almost certainly would have occurred with importance resampling. Inference from the remaining nine series yields essentially the same results as presented earlier in Table 2. This example thus illustrates the relevance of both parts of our procedure: (1) the use of an overdispersed starting distribution, which if well-chosen can lead to conservative yet relatively efficient inferences, and (2) the analysis of multiple simulated sequences for inference and monitoring convergence.

References

Belin, T. R., and Rubin, D. B. (1990). Analysis of a finite mixture model with variance components. *Proceedings of the American Statistical Association, Social Statistics Section*, 211–215.

Belin, T. R., and Rubin, D. B. (1992). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. Technical report.

Besag, J., and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society B* **55**, to appear.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Wideman, L. (1991). Simple Bayesian methods for the analysis of logistic regression models. *Journal of the American Statistical Association* **86**, 68–78.

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

Fisher, R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

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. *Journal of the American Statistical Association* **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. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. (1992). Iterative and non-iterative simulation algorithms. In *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface.*

Gelman, A., and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4*, ed. J. Bernardo, Oxford University Press.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, ed. J. Bernardo, Oxford University Press.

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. *Journal of the Royal Statistical Society B* **55**, to appear.

Green, P. J., and Han, X. (1991). Metropolis methods, Gaussian proposals, and antithetic variables. Technical report, University of Bristol, U.K.

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

Hills, S. E., and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference. In *Bayesian Statistics 4*, ed. J. Bernardo, Oxford University Press, 627–633.

Liu, C. (1991). Qualifying paper, Department of Statistics, Harvard University.

McCulloch, R., and Rossi, P. E. (1992). An exact likelihood analysis of the multinomial probit model. Technical report.

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. (1992). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, to appear.

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

Metropolis, N., and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association* **44**, 335–341.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian Statistics 3*, 327–344, ed. J. Bernardo, Oxford University Press.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society* **B 27**, 169–203.

Raftery, A. E., and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, ed. J. Bernardo, Oxford University Press, 765–776.

Ripley, B. D. (1987). *Stochastic Simulation*, chapter 6. New York: Wiley.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. (1987b). 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. Comment on "The calculation of posterior distributions by data augmentation," by Tanner, M. A., and Wong, W. H. *Journal of the American Statistical Association* **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.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110–114.

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

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Tierney, L. (1991). Exploring posterior distributions using Markov chains. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.*

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

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.