Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review

Mary Kathryn COWLES and Bradley P. CARLIN

A critical issue for users of Markov chain Monte Carlo (MCMC) methods in applications is how to determine when it is safe to stop sampling and use the samples to estimate characteristics of the distribution of interest. Research into methods of computing theoretical convergence bounds holds promise for the future but to date has yielded relatively little of practical use in applied work. Consequently, most MCMC users address the convergence problem by applying diagnostic tools to the output produced by running their samplers. After giving a brief overview of the area, we provide an expository review of 13 convergence diagnostics, describing the theoretical basis and practical implementation of each. We then compare their performance in two simple models and conclude that all of the methods can fail to detect the sorts of convergence failure that they were designed to identify. We thus recommend a combination of strategies aimed at evaluating and accelerating MCMC sampler convergence, including applying diagnostic procedures to a small number of parallel chains, monitoring autocorrelations and cross-correlations, and modifying parameterizations or sampling algorithms appropriately. We emphasize, however, that it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution.

KEY WORDS: Autocorrelation; Gibbs sampler; Metropolis-Hastings algorithm.

1. INTRODUCTION

In a surprisingly short period, Markov chain Monte Carlo (MCMC) integration methods, especially the Metropolis–Hastings algorithm (Hastings 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) and the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990) have emerged as extremely popular tools for the analysis of complex statistical models. This is especially true in the field of Bayesian analysis, which requires evaluation of complex and often high-dimensional integrals to obtain posterior distributions for the unobserved quantities of interest in the model (i.e., unknown parameters, missing data, and data that are yet to be observed). In many such settings, alternative methodologies (such as asymptotic approximation, traditional numerical quadrature, and noniterative Monte Carlo methods) either are infeasible or fail to provide sufficiently accurate results. Properly defined and implemented, MCMC methods enable the user to successively sample values from a convergent Markov chain, the limiting distribution of which is the true joint posterior of the model unobservables. Important features of MCMC methods that enhance their applicability include their ability to reduce complex multidimensional problems to a sequence of much lower-dimensional ones and their relative indifference to the presence or absence of conjugate structure between the likelihood and the prior distribution.

Although MCMC methods have been most widely used in Bayesian analysis, they have also been used by frequentists in missing- and dependent-data settings where the likelihood itself involves complicated high-dimensional integrals (see, for example, Gelfand and Carlin 1993, and Geyer and Thompson 1992). Excellent tutorials on the methodology have recently been provided by Albert (1993) and Casella and George (1992); a more complete and advanced summary was given by Tierney (1995). The statistical applications of MCMC in just the last 5 years are far too numerous to list, covering such disparate areas as the modeling of human immunodeficiency virus (HIV) progression (Lange, Carlin, and Gelfand 1992), archaeological shape estimation (Buck, Litton, and Stephens 1993), determination of fuel economy potential in automobiles (Andrews, Berger, and Smith 1993), and the analysis of home run hitters in major league baseball (Albert 1992).

Although MCMC algorithms allow an enormous expansion of the class of candidate models for a given dataset, they also suffer from a well-known and potentially serious drawback: It is often difficult to decide when it is safe to terminate them and conclude their "convergence." That is, at what point is it reasonable to believe that the samples are truly representative of the underlying stationary distribution of the Markov chain? It is immediately clear that this is a more general notion of convergence than is usual for iterative procedures, because what is produced by the algorithm at convergence is not a single number or even a distribution, but rather a *sample* from a distribution. Worse yet, the Markov nature of the algorithm means that members of this sample will generally be *correlated* with each other, slowing the algorithm in its attempt to sample from the entire stationary distribution and muddying the determination of appropriate Monte Carlo variances for estimates of model characteristics based on the output. Much of the aforementioned applied work has shown that such high correlations, both within the output for a single model parameter (*autocorrelations*) and across parameters (*cross-correlations*) are

Mary Kathryn Cowles is Assistant Professor of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Bradley P. Carlin is Associate Professor, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455. Much of the work was done while the first author was a graduate student in the Division of Biostatistics at the University of Minnesota and then Assistant Professor, Biostatistics Section, Department of Preventive and Societal Medicine, University of Nebraska Medical Center, Omaha, NE 68198. The work of both authors was supported in part by National Institute of Allergy and Infectious Diseases FIRST Award 1-R29-AI33466. The authors thank the developers of the diagnostics studied here for sharing their insights, experiences, and software, and Thomas Louis and Luke Tierney for helpful discussions and suggestions that greatly improved the manuscript.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Review Paper

not uncommon, caused, for example, by a poor choice of parameterization or perhaps overparameterization. The latter situation can of course lead to "ridges" in the posterior or likelihood surface, long the bane of familiar statistical optimization algorithms.

Efforts at a solution to the problem of determining MCMC algorithm convergence have been concentrated in two areas. The first is theoretical, wherein the Markov transition kernel of the chain is analyzed in an attempt to predetermine a number of iterations that will ensure convergence in total variation distance to within a specified tolerance of the true stationary distribution. (Notice that this goes beyond merely proving that a certain algorithm will converge for a given problem, or even providing a rate for this convergence.) For example, Polson (1994) developed polynomial time convergence bounds for a discrete-jump Metropolis algorithm operating on a log-concave target distribution in a discretized state space. Rosenthal (1993, 1995a, 1995b) instead used Markov minorization conditions, providing bounds in continuous settings involving finite-sample spaces and certain hierarchical models. Although approaches like these hold promise, they typically involve sophisticated mathematics, as well as laborious calculations that must be repeated for every model under consideration. Moreover, in most examples analyzed thus far using these tools, the bounds obtained are quite loose, suggesting numbers of iterations that are several orders of magnitude beyond what would be considered reasonable or feasible in practice—though Rosenthal (1996) obtained tight bounds in a hierarchical normal means model related to James–Stein estimation.

As a result, almost all of the applied work involving MCMC methods has relied on the second approach to the convergence problem: applying diagnostic tools to output produced by the algorithm. Early attempts by statisticians in this regard involved comparing the empirical distributions of output produced at consecutive (or nearly consecutive) iterations and concluding convergence when the difference between the two was negligible in some sense. This led to samplers using a large number of parallel, independent chains to obtain simple moment, quantile, and density estimates. Indeed, not long ago a widely used diagnostic was the so-called "thick felt-tip pen test" of Gelfand and Smith (1990), where convergence was concluded if density estimates spaced far enough apart to be considered independent (say, five iterations) differed graphically by less than the width of a thick felt-tip pen. Besides the inherent waste of preconvergence samples in this massively parallel approach, the diagnostic often suggested convergence prematurely for slowly mixing samplers, because it measured the distance separating the sampled distribution at two different iterations rather than the distance separating either distribution from the true stationary distribution.

Of course, because the stationary distribution will always be unknown to us in practice, this same basic difficulty will plague any convergence diagnostic. Indeed, this is what leads many theoreticians to conclude that all such diagnostics are fundamentally unsound. Many researchers in other areas where MCMC methods have been used for many years (e.g., physics and operations research) have also reached this conclusion. Still, many statisticians rely heavily on such diagnostics, if for no other reason than "a weak diagnostic is better than no diagnostic at all."

In Section 2 we introduce the MCMC convergence diagnostics in our study. For each, we briefly review their theoretical bases and discuss their practicality of implementation. We also classify the methods according to whether they measure the convergence of univariate quantities or of the full joint distribution and whether their results are quantitative or qualitative (i.e., graphical) in nature. Finally, we assess the extent to which each addresses the competing issues of bias and variance in the resulting estimated features of the stationary distribution. Virtually all convergence diagnostics seek to uncover bias arising from a sample that is not representative of the underlying distribution; a few also attempt to instruct the user as to how many (autocorrelated) samples should be drawn to produce estimates with variance small enough to inspire confidence in their accuracy. For those diagnostics not addressing this latter issue, an alternative is to use batching (see, e.g., Ripley 1987, sec. 6.2), or perhaps more sophisticated time series methods (as in Geyer 1992).

In Sections 3 and 4 we apply our collection of diagnostics to some relatively simple statistical models. In so doing, we investigate whether use is indeed appropriate within the realm of common statistical practice. We find that many of the diagnostics produce results that are difficult to interpret and potentially misleading even in these idealized settings. Finally, in Section 5 we discuss our findings and offer recommendations on how to proceed in this thorny area.

2. MCMC CONVERGENCE DIAGNOSTICS

The convergence diagnostics of Gelman and Rubin (1992) and of Raftery and Lewis (1992) currently are the most popular in the statistical community, at least in part because computer programs for their implementation are available from their creators. In addition to these two, we discuss the methods of Garren and Smith (1993), Geweke (1992), Johnson (1994), Liu, Liu, and Rubin (1992), Mykland, Tiermey, and Yu (1995), Ritter and Tanner (1992), Roberts (1992, 1994), Yu (1994), Yu and Mykland (1994), and Zellner and Min (1995). Furthermore, we mention some related ideas from the operations research literature, focusing on the technique of Heidelberger and Welch (1983).

2.1 Gelman and Rubin

Based on normal theory approximations to exact Bayesian posterior inference, Gelman and Rubin's (1992) method involves two steps. Step 1, to be carried out before sampling begins, is to obtain an overdispersed estimate of the target distribution and to generate from it the starting points for the desired number of independent chains (say 10 if only one major mode was found, and more in the case of multiple modes). Step 2 is to be carried out for each scalar quantity of interest (after appropriate transformation to approximate normality, if needed) after running the Gibbs sampler

chains for the desired number of iterations, say $2n$. It involves using the last n iterations to reestimate the target distribution of the scalar quantity as a conservative Student t distribution, the scale parameter of which involves both the between-chain variance and the within-chain variance. Convergence is monitored by estimating the factor by which the scale parameter might shrink if sampling were continued indefinitely, namely

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

where B is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of the approximating t density. Slowly mixing samplers will initially have B much larger than W, because the chain starting points are overdispersed relative to the target density. Gelman and Rubin recommended an iterative process of running additional iterations of the parallel chains and redoing step 2 until the "shrink factors" for all quantities of interest are near 1; at that point, assuming that each chain has been run for a grand total of $2n$ iterations, inference may be carried out using the combined values from iterations $n+1$ to $2n$ from all chains. S-language code is available to perform step 2; it reports both the point estimate and the .975 quantile of the shrink factors, as well as empirical quantiles of sampled quantities computed from the pooled samples and estimated quantiles based on the Student t distribution.

Though created for the Gibbs sampler, Gelman and Rubin's method may be applied to the output of any MCMC algorithm. Their approach emphasizes reducing bias in estimation. They interpreted the fact that the "shrink factor" approaches 1 when the pooled within-chain variance dominates the between-chain variance to mean that at that point, all chains have escaped the influence of their starting points and have traversed all of the target distribution. They posited that there is no way to determine that this has occurred in a single chain without having additional independent chains, started from dispersed initial values, for comparison.

A number of criticisms of Gelman and Rubin's method have been made. It relies heavily on the user's ability to find a starting distribution that is indeed overdispersed with respect to the target distribution, a condition that requires knowledge of the latter to verify. Second, because the Gibbs sampler is most needed when the normal approximation to the posterior distribution is inadequate for purposes of estimation and inference, reliance on normal approximation for diagnosing convergence to the true posterior may be questionable. Also, the approach essentially is univariate. But Gelman and Rubin suggested applying their procedure to -2 times the log of the posterior density as a way of summarizing the convergence of the joint density. Advocates of running a single long chain consider it very inefficient to run multiple chains and discard a substantial number of early iterations from each. Furthermore, they point out that if one compares, for example, a single chain run for 10,000 iterations with 10 independent chains each run for 1,000 iterations, then the last 9,000 iterations from the single long chain are all drawn from distributions that are likely to be closer to the true target distribution than those reached by any of the shorter chains.

2.2 Raftery and Lewis

Raftery and Lewis's (1992) method is intended both to detect convergence to the stationary distribution and to provide a way of bounding the variance of estimates of quantiles of functions of parameters. The user must first run a single-chain Gibbs sampler for $N_{\min}$, the minimum number of iterations that would be needed to obtain the desired precision of estimation if the samples were independent. Then either Raftery and Lewis's Fortran program (available from Statlib) or the corresponding S-Plus function in the CODA package (see Sec. 5) may be run for each quantity of interest in turn, using the Gibbs chains for that quantity as input. Each program prompts the user to specify the quantile q to be estimated (e.g., .025), the desired accuracy r (e.g., $\pm$ .005), the required probability s of attaining the specified accuracy, and a convergence tolerance (explained later) δ, usually given as .001. The program then reports "nprec" (the total number of iterations that should be run), "nburn" (how many of the beginning iterations should be discarded), and "k" (where only every kth one of the remaining iterates should be used in inference). Iterations corresponding to the largest value of "nprec" obtained for any quantity tested may then be run and, if desired, the diagnostic process may be repeated to verify that they are sufficient.

The approach is based on two-state Markov chain theory, as well as standard sample size formulas involving binomial variance. A binary sequence $\{Z\}$ is formed with a 0/1 indicator for each iteration of the original Gibbs chain as to whether the value of the quantity of interest is less than a particular cutoff. "K" is the smallest skip-interval for which the behavior of the new binary sequence $\{Z^{(k)}\}$ formed by extracting every kth iterate approximates that of a first-order Markov chain. "Nburn" is the number of iterations that it takes for $\{Z^{(k)}\}$ to approach within δ (specified by the user as mentioned earlier) of its estimated stationary distribution. A large value of "nburn" suggests slow convergence to the stationary distribution, whereas a value of "nprec" much larger than $N_{\min}$ and/or "k" greater than 1 suggests strong autocorrelations within the chain. The fact that the formula for $N_{\min}$ is based on binomial variance leads to the counterintuitive result that more iterations are required for estimating quantiles near the median than extreme quantiles to obtain the same degree of accuracy. This approach may be applied to the output of any MCMC algorithm.

Raftery and Lewis emphasized that being able to pin down the accuracy of the estimation of quantiles is very useful, because they are at the heart of density estimation and also provide robust estimates of center and spread of a distribution. Critics point out that the method can produce variable estimates of the required number of iterations needed given different initial chains for the same problem and that it is univariate rather than giving information about

the full joint posterior distribution. It is also somewhat impractical in that convergence must be "rediagnosed" for every quantile of interest. Finally, recent work by MacEachern and Berliner (1994) showed that estimation quality is always degraded by discarding samples, so that the whole enterprise of estimating a skip-interval k may be inappropriate.

2.3 Geweke

Geweke (1992) recommended using methods from spectral analysis to assess convergence of the Gibbs sampler when the intent of the analysis is to estimate the mean of some function g of the parameters θ being simulated. If values of $g(\theta^{(j)})$ are computed after each iteration of the Gibbs sampler, then the resulting sequence may be regarded as a time series. Geweke's method rests on the assumption that the nature of the MCMC process and of the function g imply the existence of a spectral density $S_g(\omega)$ for this time series that has no discontinuities at frequency zero. If this assumption is met, then for the estimator of $E[g(\theta)]$ based on n iterations of the Gibbs sampler,

$$\bar{g}_n = \frac{\sum_{i=1}^{n} g(\theta^{(i)})}{n},$$

the asymptotic variance is $S_g(0)/n$. The square root of this asymptotic variance may be used to estimate the standard error of the mean. Geweke referred to this estimate as the "numeric standard error" (NSE).

Geweke's convergence diagnostic after n iterations of the Gibbs sampler is calculated by taking the difference between the means $\bar{g}(\theta)_n^A$, based on the first n_A iterations, and $\bar{g}(\theta)_n^B$, based on the last n_B iterations, and dividing by the asymptotic standard error of the difference, computed as earlier from spectral density estimates for the two pieces of the sequence. If the ratios n_A/n and n_B/n are held fixed and $n_A + n_B < n$, then by the central limit theorem, the distribution of this diagnostic approaches a standard normal as n tends to infinity. Geweke suggested using $n_A = .1n$ and $n_B = .5n$. He implied that this diagnostic may be used to determine how many initial iterations to discard. Then a sufficient number of subsequent iterations must be run to obtain the desired precision, as given by the NSE.

Geweke's method attempts to address the issues of both bias and variance. It is available in the CODA package (Best, Cowles, and Vines 1995). Like Gelman and Rubin's convergence diagnostic, Geweke's is essentially univariate, but if $g(\theta)$ were taken to be -2 times the log of the posterior density, then it also might be used to investigate convergence of the joint posterior. It requires only a single sampler chain and may be applied with any MCMC method.

Disadvantages of Geweke's method include its sensitivity to the specification of the spectral window. In addition, although his diagnostic is quantitative, Geweke does not specify a procedure for applying it but instead leaves that to the experience and subjective choice of the statistician.

2.4 Roberts

Roberts (1992) presented a one-dimensional diagnostic intended to assess convergence of the entire joint distribution. His method is applicable when the distributions of the iterates of the Gibbs sampler have continuous densities.

Roberts's method requires a symmetrized Gibbs sampler algorithm in which each iteration consists of a pass through all the full conditionals in a predetermined order, followed by a pass back through all of them in the reverse order. Roberts defined a function space and an inner product under which the transition operator induced by the kernel of such a Gibbs sampler chain is self-adjoint. He then proved that under certain regularity conditions,

$$\|f^{(n)} - f\| \stackrel{n\to\infty}{\longrightarrow} 0,$$

where $\|\cdot\|$ is the norm associated with the specified inner product, $f^{(n)}$ is the density of the values generated at the nth iteration of the Gibbs sampler, and f is the true target joint density.

Roberts's convergence diagnostic is an unbiased estimator of $\|f^{(n)} - f\| + 1$. It requires running m parallel reversible Gibbs sampler chains, all starting at the same initial values $\theta^{(0)}$, and is computed as

$$J_n = \frac{1}{m(m-1)} \sum_{l \neq p} \frac{k(\theta_l^{(1/2)}, \theta_p^{(2n-1)})}{f(\theta_p^{(2n-1)})},$$

where l and p are replications of the sampler, $\theta_l^{(1/2)}$ is the value obtained after the "forward" half of the first iteration of the lth chain, $\theta_p^{(n)}$ is the value obtained after the nth complete iteration of the pth chain, and k is the kernel of the "backward" half of the reversible sampler. Normalizing constants for all full conditionals, or good approximations to them, are needed in computing the numerator of the foregoing sum. Roberts suggested graphically evaluating the monotonic convergence of values of this diagnostic. In the usual case in which the Gibbs sampler is used, because the normalizing constant of the density f is unknown, the constant toward which the values of the diagnostic are converging is also unknown; hence the practitioner can look only for stabilization of the values, and the diagnostic does not estimate $\|f^{(n)} - f\| + 1$.

In a subsequent paper, Roberts (1994) modified his diagnostic somewhat, improving its practicality. Writing

$$\chi_n^{lp} = \frac{k(\theta_l^{(0)}, \theta_p^{(n)})}{f(\theta_p^{(n)})} \quad \text{and} \quad \chi_n^{l} = \chi_n^{ll},$$

he defined the within-chain *dependence* term $D_n = 1/m \sum_{l=1}^{m} \chi_n^l$ and the between-chain *interaction* term $I_n = 1/m(m-1)\sum_{l \neq p} \chi_n^{lp}$. Roberts showed that $E(I_n) \leq E(D_n)$, $\text{var}(I_n) \leq \text{var}(D_n)$, and most importantly, $E(I_n) = E(D_n)$ at convergence. He thus recommended choosing $m = 10$ to 20 *different* starting points (e.g., sampled from a distribution overdispersed relative to the target density) and monitoring I_n and D_n until both series are stationary and have similar locations. A reversible sampler must still be used, though Roberts noted that a random visitation scheme or storing dual iterations will also result in the required self-adjoint Gibbs transition kernel.

Roberts's approach attempts to address bias in estimation rather than variance and is evaluated in a graphical way. Although the 1992 version depends on the form of the full conditionals, the 1994 paper generalizes the technique to other MCMC algorithms. Advantages of Roberts's method of diagnosing convergence are its rigorous mathematical foundation and the fact that it assesses convergence of the entire joint distribution rather than of univariate quantities. But there are also several pragmatic disadvantages. The requirement for a reversible sampler makes for more complicated coding than in the standard Gibbs sampler algorithm for the same problem, and the special coding is problem-specific rather than generic. The variance of the statistic is large and may obscure the monotonic convergence. Two methods of stabilizing the variance—log transformation and increasing the number of replicate chains—have their own disadvantages. Log-transformed values become very volatile when the untransformed statistic is near zero, and the larger the number of replicate chains, the slower the entire computational process becomes.

2.5 Ritter and Tanner

Like Roberts's method, Ritter and Tanner's "Gibbs Stopper" (Ritter and Tanner 1992) is an effort to assess distributional convergence. The method may be applied either with multiple parallel chains or by dividing the output of a single long chain into batches. An importance weight is assigned to each vector drawn at each Gibbs sampler iteration. Histograms are drawn of the importance weights obtained either at each iteration across multiple chains or within each batch of a single chain. As with Roberts's method, convergence is assessed primarily in a graphical way by observing when the histogrammed values become tightly clustered, although Wei and Tanner (1990) noted that the standard deviations of the weights should also be monitored quantitatively.

The importance weight w assigned to the vector $(X_1^{(i)}, X_2^{(i)}, \ldots, X_d^{(i)})$ drawn at iteration i of the Gibbs sampler is calculated as

$$w = \frac{q(X_1^{(i)}, X_2^{(i)}, \ldots, X_d^{(i)})}{g_i(X_1^{(i)}, X_2^{(i)}, \ldots, X_d^{(i)})},$$

where q is a function proportional to the joint posterior density and g_i is the current Gibbs sampler approximation to the joint posterior. Whereas q is always available in Bayesian problems because the joint posterior density is always known up to a normalizing constant, g_i must be approximated by Monte Carlo integration as follows. Let $K(\theta', \theta)$ denote the probability of moving from $X^{(i)} = \theta'$ to $X^{(i+1)} = \theta$ in one iteration of the Gibbs sampler, which is the product of the full conditionals. Then if g_{i-1} is the joint density of the sample obtained at iteration $i-1$, the joint density of the sample obtained at iteration i is given by

$$g_i = \int K(\boldsymbol{\theta}', \boldsymbol{\theta}) g_{i-1}(\boldsymbol{\theta}')\, d\lambda(\boldsymbol{\theta}').$$

This integral may be approximated by the Monte Carlo sum

$$g_i(\boldsymbol{\theta}) \approx \frac{1}{m} \sum_{j=1}^{m} K(\boldsymbol{\theta}^j, \boldsymbol{\theta}),$$

where the observations $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^m$ may be obtained either from m parallel chains at iteration $i-1$ or from a batch of size m of consecutive iterates in a single chain. As in computing the Roberts diagnostic, exact or approximate normalizing constants for all full conditionals are needed in the formation of this sum.

Concerned with distributional convergence, the Gibbs Stopper purports to deal with reducing bias, rather than variance, in estimating the desired quantities. It is applicable only with the Gibbs sampler. The required coding is problem-specific, and computation of weights can be time-intensive, particularly when full conditionals are not standard distributions, so that the normalizing constants must be estimated. If the Gibbs sampler has reached equilibrium, then the value around which the Gibbs Stopper weights stabilize provides an estimate of the normalizing constant of the joint target distribution.

2.6 Zellner and Min

With the aim of determining not only whether the Gibbs sampler converges in distribution but also whether it converges to the *correct* distribution, Zellner and Min (1995) proposed two "Gibbs Sampler Convergence Criteria" (GSC^2) based on conditional probability. Both are applicable when the model parameters may be divided into two (vector or scalar) parts α and β, in terms of which the joint posterior may be written analytically and for each of which explicit and easily sampled posterior conditionals may be derived. Then, using marginals $\hat{p}(\alpha)$ and $\hat{p}(\beta)$ estimated by "Rao–Blackwellization" of the Gibbs sampler output (see Gelfand and Smith 1990) for a particular value of the parameters α_1 and β_1, the "Gibbs sampler difference convergence criterion" may be calculated as

$$\hat{p}(\alpha_1)p(\beta_1|\alpha_1) - \hat{p}(\beta_1)p(\alpha_1|\beta_1) = \hat{\eta}.$$

The components needed for the "Gibbs sampler ratio convergence criterion" may be computed using two values of the parameters (α_1, β_1) and (α_2, β_2):

$$\hat{\theta}_A = \frac{\hat{p}(\alpha_1)p(\beta_1|\alpha_1)}{\hat{p}(\alpha_2)p(\beta_2|\alpha_2)},$$

$$\hat{\theta}_B = \frac{\hat{p}(\beta_1)p(\alpha_1|\beta_1)}{\hat{p}(\beta_2)p(\alpha_2|\beta_2)},$$

and

$$\theta = \frac{\pi(\alpha_1, \beta_1)f(\alpha_1, \beta_1|\mathbf{y})}{\pi(\alpha_2, \beta_2)f(\alpha_2, \beta_2|\mathbf{y})},$$

where π is the prior and f is the likelihood. Then if the Gibbs sampler has converged, $\hat{\eta} \approx 0$ and $\hat{\theta}_A \approx \hat{\theta}_B \approx \theta$. Both Bayesian and large-sample sampling theory procedures are used to test for equality in these expressions.

These convergence diagnostics are intended to address bias rather than variance in estimation. They are quantitative and require only a single sampler chain. Coding is problem-specific, and analytical work is needed when the factorization into two sets of parameters is not obvious. Despite the name, the diagnostics may be used with MCMC samplers other than the Gibbs sampler; however, the class of problems to which these diagnostics may be applied is limited to those in which the joint posterior may be factored as indicated.

2.7 Liu, Liu, and Rubin

Like Roberts (1992, 1994), Liu et al. (1992) proposed the use of a single statistic, or "global control variable," to assess the convergence of the full joint posterior distribution. The method requires running m parallel chains started from dispersed initial values. For each pair of distinct chains i and j, the following statistic is calculated at each iteration t:

$$U^{(i,j,t)} = \frac{\pi(X^{(j,t)})}{\pi(X^{(i,t)})} \frac{K(X^{(j,t-1)}, X^{(i,t)})}{K(X^{(j,t-1)}, X^{(j,t)})},$$

where $X^{(j,t)}$ represents the vector of parameters generated at iteration t of chain j and $K(\mathbf{X}, \mathbf{Y})$ is defined as in Section 2.5. Liu et al. proved that

$$E_0(U^{(i,j,t)}) = \text{var}_\pi\left(\frac{p_t(X)}{\pi(X)}\right) + 1,$$

where the expectation on the left side is with respect to the distribution of the initial values and the variance on the right is with respect to the target distribution.

Liu et al. suggested two ways of using this "global control variable." One is to divide the parallel chains into $m/2$ independent pairs (i.e., pairs that share no common chain), and to apply the method of Gelman and Rubin described in Section 2.1 to the sequences of U values calculated at each iteration for each pair. The other is to plot the sample cumulative distribution function of the U values aggregated across parallel chains at the same iteration.

An important problem with the method is that the variance of the "control variable" may be unacceptably large, even infinite. One solution is to log transform the "global control variable" and apply the method of Gelman and Rubin to that; however, the expected value of the log transform is less interpretable than that of the original "global control variable." Another remedy is to use a very large number of parallel chains, but this of course may be infeasible due to computational time in complex models.

Liu et al.'s method is similar to the methods of Ritter and Tanner and of Roberts in that it requires problem-specific coding, is aimed at reducing bias rather than variance in estimation, and is specific to the Gibbs sampler.

2.8 Garren and Smith

Like Roberts (1992), Garren and Smith (1993) attacked the convergence diagnosis problem from a rigorous mathematical perspective. Because the convergence rate of an MCMC algorithm is governed by how close the second-largest eigenvalue of its transition matrix (or kernel density) is to 1, these authors attempted to estimate this eigenvalue directly from the sample output. Reminiscent of Raftery and Lewis (1992), they defined $Z^{(i)} = I(X^{(i)} \in E)$ for some specified subset E of the state space, so that $\rho^{(i)} \equiv E(Z^{(i)})$ is the posterior probability of E at iteration i. Assuming that the transition operator is self-adjoint and Hilbert–Schmidt (Schervish and Carlin 1992), we may write

$$\rho^{(i)} = \rho + a_2\lambda_2^i + O(|\lambda_3|^i),$$

where $\rho = \lim_{i\to\infty} \rho^{(i)}$, a_2 is some real number, and $|\lambda_3| < |\lambda_2| < \lambda_1 = 1$ are the three largest eigenvalues of the kernel density. Note that this strict inequality among the eigenvalues (a "spectral gap") need not occur for samples on uncountable state spaces; in fact, it will typically not occur for Metropolis–Hastings samplers, where candidates are not accepted with probability 1 (Chan and Geyer 1994).

Now suppose that we have m parallel sampling chains (all started from the same point in the state space), burn-in period K, and total run length N, where $1 \leq K < N$. For $i = (K+1), \ldots, N$, estimate $\rho^{(i)}$ as $\bar{Z}^{(i)}$, the sample proportion of times that $X^{(i)} \in E$ over the m replications. Then take the value $\hat{\theta} = (\hat{\rho}, \hat{a}_2, \hat{\lambda}_2)$ that minimizes

$$S(\rho, a_2, \lambda_2) = \sum_{i=K+1}^{N} (\bar{Z}^{(i)} - \rho - a_2\lambda_2^i)^2$$

as a nonlinear least squares estimate of θ. Garren and Smith showed that $\hat{\theta}$ is asymptotically normal and obtained explicit expressions for its asymptotic mean and variance. Their diagnostic procedure then involves plotting the sample values of the three components of $\hat{\theta}$ and their corresponding approximate 95% confidence bands for $K = 1, 2, \ldots$, and looking for the point at which the systematic bias in the estimates disappears. Because this may be difficult to identify, Garren and Smith also suggested looking for the point where the estimates become unstable (and the confidence bands widen dramatically). This value of K is then chosen as the proper amount of sampler burn-in.

Garren and Smith's method addresses only burn-in (bias) and not the issue of variance in estimation. It is multivariate in a limited sense, through the set of interest E. Its dependence on the existence of a spectral gap limits its applicability to the Gibbs sampler. It involves a very large number of parallel chains (the authors used $m = 250$ and 5,000 in their examples), but because they must all share the same starting point, this extra computational effort does not serve to explore different regions of the state space (though the authors stated that a multistart version of their algorithm should be possible). Finally, the authors' own empirical results are disappointing, involving substantial methodological and computational labor to provide plots that are very difficult to interpret (the point at which the estimates become "unstable" is open to question).

2.9 Johnson

Johnson (1994) used the notion of convergence as the mixing of chains initialized from an overdispersed starting

distribution, but from a nonstochastic point of view. Suppose that a Gibbs sampling chain can be created from a stream of uniform random deviates $\{u_i\}$ (e.g., using the inversion method for each full conditional distribution). For countable state spaces, Johnson showed that parallel sampling chains started from different locations but based on the same stream of uniform deviates must all eventually converge to a *single* sample path. He thus reasoned that when several chains started from an overdispersed starting distribution all agree to some tolerance $\varepsilon > 0$, the sampler has indeed "forgotten" where it started and thus must have converged. More conservatively, the process may be repeated with several different random number streams and/or sets of starting points, and convergence time defined as the median or even maximum time to overlap.

Johnson's method is quantitative, and multivariate in the sense that chain overlap can be precisely verified in arbitrarily high dimensions, but in practice convergence of univariate quantities is often monitored graphically. The method does not directly address the bias issue, though Johnson did assert that convergence time can also be thought of as an upper bound on the time required to obtain approximately independent samples. Because Johnson's method is essentially a deterministic version of Gelman and Rubin's popular approach, it also relies on an adequately overdispersed starting distribution; indeed, Johnson recommended the preliminary mode-finding approach of Gelman and Rubin for initializing the parallel chains. For slowly mixing or highly multimodal samplers, the addition of one more chain can result in a substantial increase in time to overlap. Finally, in fully conjugate sampling settings, a common stream of uniform random numbers will typically arise from the use of the same seed for the random number generator in each chain, so that the user need not perform the inversion explicitly. For nonconjugate full conditionals, however, implementation may be complicated. For Metropolis–Hastings–type algorithms, the approach is typically inapplicable, though it could be used with an independence chain Hastings algorithm by using the $\{u_i\}$ sequence in the rejection steps and a second, independent sequence $\{v_i\}$ to determine the candidates.

2.10 Heidelberger and Welch and Schruben, Singh, and Tierney

Combining the method of Schruben (1982) and Schruben, Singh, and Tierney (1983) for detecting nonstationarity in simulation output with a spectral analysis approach to estimating the variance of the sample mean, Heidelberger and Welch (1983) devised a comprehensive procedure for generating a confidence interval of prespecified width for the mean when there is an "initial transient"—that is, when the simulation does not start off in its stationary distribution. Their procedure is to be applied to a single chain. Although their approach antedates the Gibbs sampler and was designed for use in discrete-event simulation work in the operations research field, it is applicable to the output of the Gibbs sampler and other MCMC algorithms.

The approach of Schruben (1982) and Schruben et al. (1983) to diagnosing convergence is a hypothesis test based on Brownian bridge theory. The null hypothesis is that the sequence of iterates is from a stationary process that is ϕ mixing. In Markov chains, this is equivalent to uniform ergodicity, a condition that essentially applies only to compact state spaces. But it appears that the approach would still be valid for chains that are merely geometrically ergodic (Meyn and Tweedie 1993, sec. 17.4.2), a condition satisfied by many convergent Gibbs and Metropolis–Hastings algorithms on general state spaces. If $Y^{(j)}$ is the jth iterate in the output sequence, $S(0)$ is the spectral density of the sequence evaluated at zero, $[\cdot]$ is the greatest integer less than or equal to $\cdot$, n is the total number of iterations, and

$$T_0 = 0,$$

$$T_k = \sum_{j=1}^{k} Y^{(j)}, \qquad k \geq 1,$$

$$\bar{Y} = \frac{\sum_{j=1}^{n} Y^{(j)}}{n},$$

and

$$B_n(t) = \frac{T_{[nt]} - [nt]\bar{Y}}{\sqrt{nS(0)}}, \qquad 0 \leq t \leq 1,$$

then under the null hypothesis, for large n, $B_n = \{B_n(t), 0 \leq t \leq 1\}$ is distributed approximately as a Brownian bridge. Thus the Cramer–von Mises statistic,

$$\int_0^1 B_n(t)^2 \, dt,$$

may be used to test the hypothesis. Because $S(0)$ is unknown, it must be estimated from the data and the estimate used in computing $B_n(t)$.

Heidelberger and Welch incorporated this test for stationarity into the following process for detecting and eliminating an initial transient, generating confidence intervals, and controlling run length. The user of the method must specify two parameters: $j_{\max}$, the maximum number of iterations that can be run, and ε, the desired relative half-width for confidence intervals. An initial number of iterations, $j_1 = .1j_{\max}$, is run. Because a spectral density estimate of $S(0)$ based on a sequence that contained an initial transient would tend to be too large, thus reducing the size of the test statistic and consequently decreasing the power of the test, an estimate of $S(0)$ based on the second half of this run is used to perform Schruben's stationarity test on the entire run. If the null hypothesis is rejected, then the first 10% of the iterations are discarded and the stationarity test is repeated. If the null hypothesis is rejected again, then the test is repeated after an additional 10% of the iterations are discarded from the beginning of the run. The process continues until either a portion of the output of length greater than or equal to $.5j_1$ is found for which the stationarity test is passed or 50% of the iterations have been discarded and the test still rejects. In the former case, the spectral density

Table 1. Summary of Convergence Diagnostics

Method	Quantitative graphical	Single or multiple chains	Theoretical basis	Univariate/ full joint distribution	Bias/ variance	Applicability	Ease of use
Gelman and Rubin (1992)	Quantitative	Multiple	Large-sample normal theory	Univariate	Bias	Any MCMC	a
Raftery and Lewis (1992)	Quantitative	Single	2-state Markov chain theory	Univariate	Both	Any MCMC	a
Geweke (1992)	Quantitative	Single	Spectral analysis	Univariate	Both	Any MCMC	a
Roberts (1992, 1994)	Graphical	Multiple	Probability theory	Joint	Bias	Some	c
Ritter and Tanner (1992)	Graphical	Either	Importance weighting	Joint	Bias	Gibbs only	c
Zellner and Min (1995)	Quantitative	Single	Conditional probability	Joint	Bias	Some	d
Liu, Liu, and Rubin (1992)	Both	Multiple	Probability theory	Joint	Bias	Gibbs only	c
Garren and Smith (1993)	Qualitative	Multiple	Eigenvalue analysis	Univariate	Bias	Gibbs only	b
Johnson (1994)	Quantitative	Multiple	Fixed-point theorems	Joint	Bias	Some	a or c
Heidelberger and Welch (1983)	Quantitative	Single	Brownian bridge, spectral analysis	Univariate	Both	Any MCMC	a
Mykland, Tierney, and Yu (1995)	Graphical	Single	Markov chain regeneration	Joint	Bias	Some	d
Yu (1994)	Graphical	Single	L_1 distance, kernel density estimation	Joint	Bias	Some	d
Yu and Mykland (1994)	Graphical	Single	Cusum path plots	Univariate	Both	Any MCMC	b

NOTE: When all full conditionals are conjugate, Johnson's method requires no additional coding beyond that required for implementing the regular Gibbs sampler. Otherwise, Johnson's method may require special, problem-specific coding to use the inversion algorithm to generate from the nonconjugate full conditionals.

$S(0)$ is reestimated from the entire portion of the output for which the stationarity test was passed, and the standard error of the mean is estimated as $\sqrt{\hat{S}(0)/n_p}$, where n_p is the length of the retained output. If the half-width of the confidence interval generated accordingly is less than ε times the sample mean of the retained iterates, then the process stops and sample mean and confidence interval are reported.

If either the stationarity test was failed or the confidence interval was too wide, then the iterates that were removed from the beginning of the sequence during the stationarity tests are restored and more iterations are run to obtain a total run length of $j_2 = 1.5j_1$. The stationarity test/confidence interval generation procedure is then applied to the new longer sequence in the same manner as earlier, without regard for the results of the stationarity test on the initial sequence. If a stationary portion still is not found, or if a sufficiently narrow confidence interval still is not obtained, then the process may be repeated with longer and longer run lengths j_k, where each $j_{k+1} = \min(1.5j_k, j_{\max})$, until either an acceptable confidence interval is generated or the run length reaches $j_{\max}$. If the run length reaches $j_{\max}$, then the stationarity test is performed. If it is failed, then no confidence interval can be formed for the mean. If the test is passed, then a confidence interval is generated, which may or may not meet the accuracy criterion ε.

Heidelberger and Welch tested their procedure on a variety of discrete-event simulations with different strengths and lengths of initial transients and concluded that "the procedure with transient removal produced point estimates with less bias and narrower confidence intervals with more proper coverage from shorter simulations than the procedure without transient removal" (1983, p. 1143). But they found that the stationarity test had little power to detect an initial transient when the run length was shorter than the extent of the initial transient.

2.11 Mykland, Tierney, and Yu

The mathematically rigorous method of Mykland et al. (1995) is based on regenerative simulation. A stochastic process, such as a Markov chain sampler, is called "regenerative" if there exist times $T_0 \leq T_1 \leq \ldots$ such that at each T_i, the future of the process is independent of the past and identically distributed. "Tours"—sets of iterations between regeneration times—are independent and identically distributed. Assuming that a single chain has been run for a fixed number of tours n, Mykland et al. proposed plotting T_i/T_n on the y axis versus i/n on the x axis. If the chain has reached equilibrium, then this plot, which they called a "scaled regeneration quantile" (SRQ) plot, should approximate a straight line through the origin with slope equal to 1. If there are large deviations from a straight line (i.e., if some tours are much longer than others), then either the sampler simply must be run longer or the examination of the states traversed during long tours might suggest ways to improve its performance.

To implement this diagnostic, analytical work and problem-specific coding are required either to generate a Markov chain with recognizable regeneration times or, more likely, to identify regeneration times in the output of a sampler that already has been run. Let $\{X_n: n = 0, 1, \ldots\}$ be the output of an irreducible, Harris-recurrent Markov chain on a state space E with transition kernel P and invariant distribution p. Then a function $s(x)$ and a probability

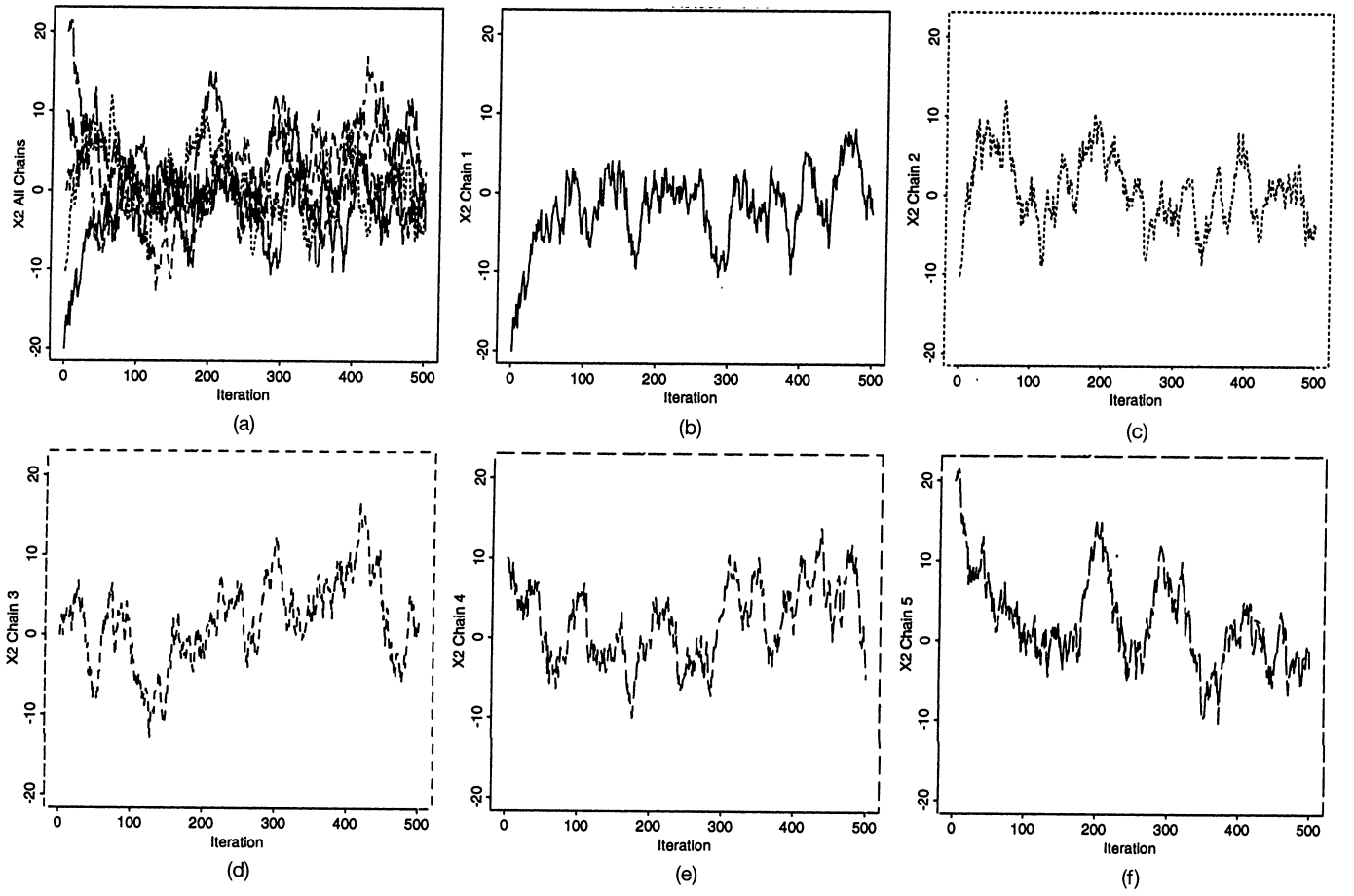


Figure 1. Plots for Trivariate Normals, High Correlations. (a) G & R = 1.23, 1.53; (b) G = −5.17, H&W = ---, lag 1 autocorrelation .95; (c) G = −4, H&W = 300, lag 1 autocorrelation .92; (d) G = −1.5, H&W = ---, lag 1 autocorrelation .96; (e) G = −11.67, H&W = ---, lag 1 autocorrelation .95; (f) G = 3.73, H&W = ---, lag 1 autocorrelation .95.

measure $\nu(dy)$ must be found satisfying

$$\pi(s) = \int s(x)\pi\,(dx) > 0 \quad \text{and} \quad P(x, A) \geq s(x)\nu(A)$$

for all points x in E and all sets A in E. Together, $s(x)$ and ν constitute an "atom" of the kernel P. Then a corresponding sequence of Bernoulli variables $\{S_n\}$ must be generated from the conditional distribution

$$P(S_n = 1 | X_n = x, X_{n+1} = Y) = \frac{s(x)\nu(dy)}{P(X, dy)}\,.$$

The iterations k such that $S_k = 1$ are regeneration times for the chain. Although this method is theoretically applicable to any MCMC algorithm, it will not always be possible to find an atom or to perform the needed variate generations. Mykland et al. provided examples of how to find an atom for certain special cases of Metropolis–Hastings and Gibbs chains. Their method addresses bias rather than variance and assesses convergence of the joint target distribution.

2.12 Yu

Yu's (1994) diagnostic approach seeks to monitor the convergence of the full joint distribution by constructing two plots based on the output of a single chain from any MCMC algorithm. She stated that although in general diagnostics based on a single chain may be misleading, running multiple parallel chains is not the only solution. She proposed as an alternative using the information contained in the unnormalized target density in combination with the output of a single chain; indeed, in many problems for which MCMC methods are used, particularly for computing Bayesian posterior densities, the target density is known up to the normalizing constant. Yu's method assumes geometric ergodicity of the Markov chain.

If $\{X_n, n = 0, 1, 2, \ldots\}$ is a Markov chain sampler with d-dimensional target density $\pi(x) = \theta g(x)$, with g known and θ the inverse of the unknown normalization constant, then Yu proposed the following procedure to construct two plots for monitoring convergence:

Step 1. Choose a one-dimensional bounded symmetric kernel $K(\cdot)$ such that $\int_{R^d} K(|x|)\,dx = 1$, and define $h(\cdot)$ to be the d-dimensional kernel with bandwidth $\sigma > 0$ such that

$$h_\sigma(x) = \frac{1}{\sigma^d} K\left(\frac{|x|}{\sigma}\right),$$

where $|\cdot|$ is the Euclidean norm in R^d. Then the kernel estimator of $\pi(\cdot)$ with bandwidth b_n is

$$\hat{\pi}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} h_{b_n}(\mathbf{x} - X_i),$$

Table 2. Raftery and Lewis's Method Applied to Trivariate Normal With High Correlations

	$q = .025$			$q = .25$			$q = .50$		
Chain	k	nburn	nprec	k	nburn	nprec	k	nburn	nprec
1	1	26	1,983	2	34	29,442	2	36	44,252
2	1	19	1,662	2	36	32,758	2	36	44,252
3	1	41	4,346	2	56	48,128	1	35	41,142
4	1	19	2,173	3	39	35,811	3	72	88,671
5	2	34	3,726	1	26	22,674	2	38	48,096

and based on this kernel estimator with fixed bandwidth σ, an efficient estimator of the inverse of the normalization constant is

$$\hat{\theta}_\sigma = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{h_\sigma(X_i - X_j)}{g(X_j)}.$$

Yu proved that $\hat{\theta}g(x)$ converges more quickly to $\pi(x)$ than does the kernel estimator $\hat{\pi}_n(x)$. With the Gibbs sampler specifically, an alternative estimator of $\pi(x)$ that converges more quickly than the kernel estimator is the product of a mixture estimator for a marginal density (Gelfand and Smith 1990) times the known form of a conditional density.

Step 2. Generate X_0 from a well-dispersed starting distribution, and run a single chain X_n for $n = 1, 2, \ldots$

Step 3. Choose a compact subset A of the support of the target distribution, which contains those points x where discrepancies between $\hat{\pi}_n(x)$ and either $\hat{\theta}g(x)$ or the Gibbs alternative are most likely. Yu suggested that Gelman and Rubin's method for choosing an overdispersed starting distribution might be helpful in choosing A.

Step 4. Based on computing resources, choose an increment $n$step defining the intervals at which convergence will be monitored.

Step 5. At intervals $n = n\text{step}, 2n\text{step}, 3n\text{step}, \ldots$, select the optimal bandwidth b_n for the kernel density estimator using a data-driven method proposed by Yu, compute $\hat{\theta}$, and use numerical integration to evaluate the following estimator of the L^1 distance over A between the kernel density estimator and π:

$$\hat{I}_n(A) = \int_A |\hat{\pi}_n(x) - \hat{\theta}g(x)|\, dx.$$

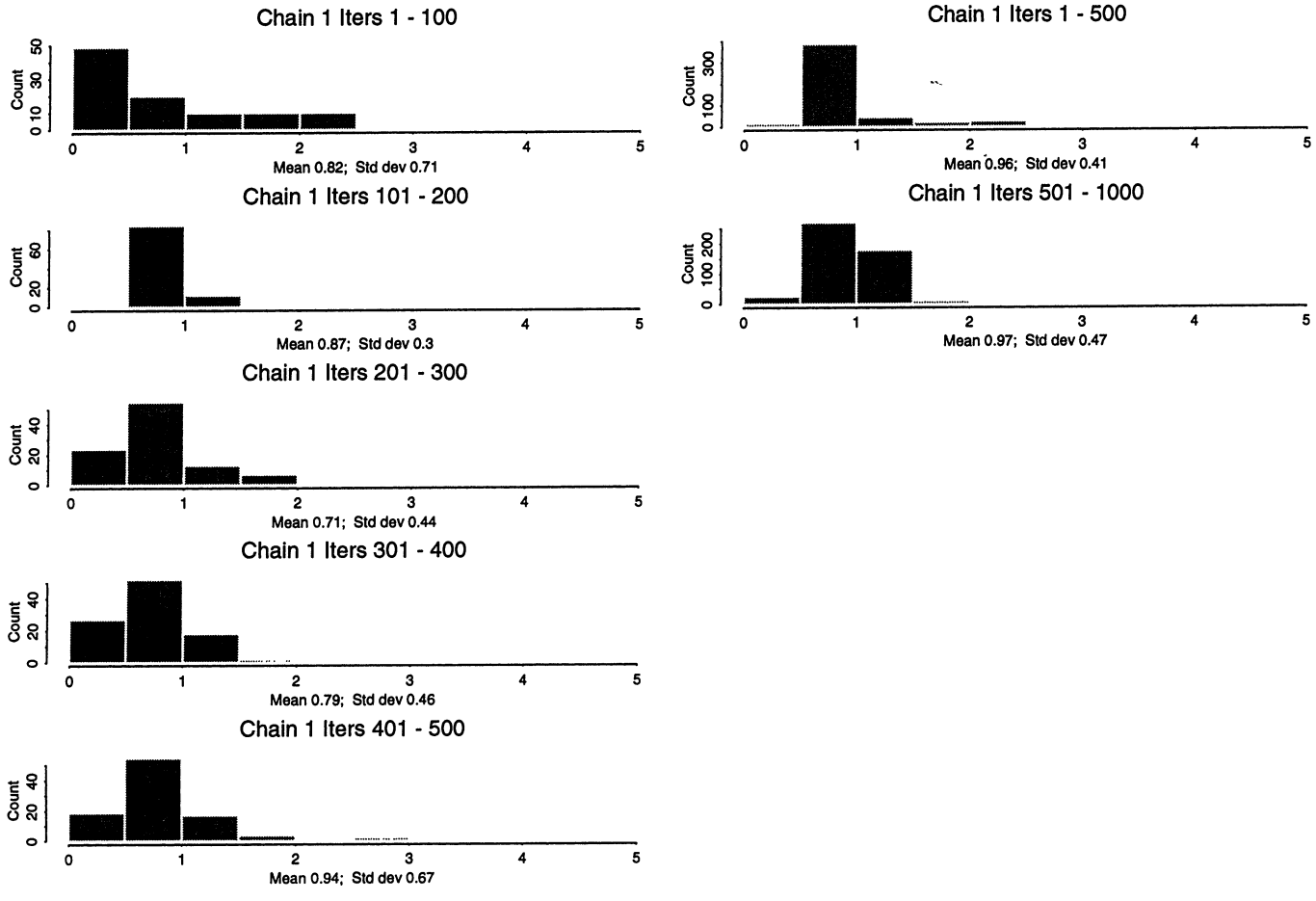
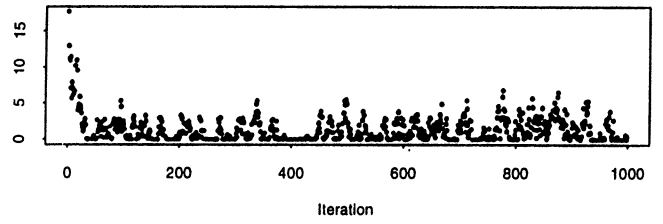
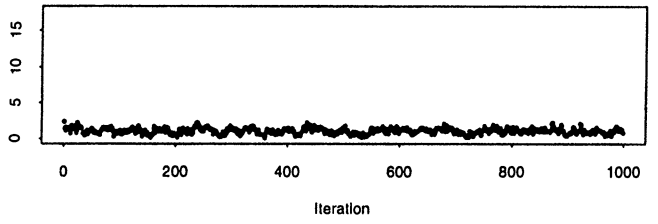


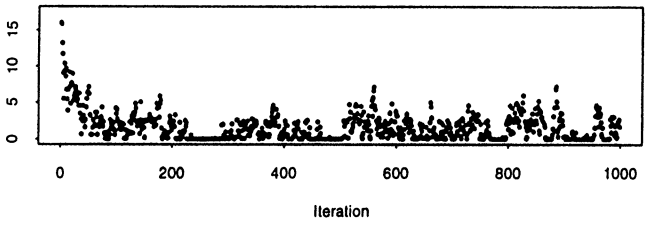
Figure 2. Histograms of Gibbs Stopper Statistics, Multivariate Normal, High Correlations.



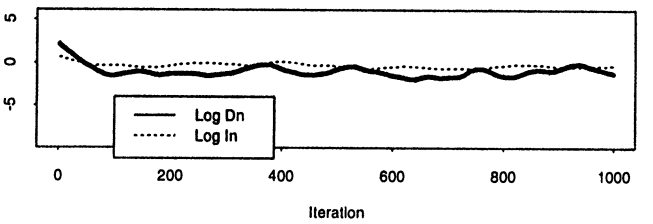
(a)



(b)



(c)



(d)

Figure 3. Roberts (1992, 1994) Convergence Diagnostics for Trivariate Normal, High Correlations. (a) Roberts (1992) convergence diagnostic, chains started 2 standard deviations below mean; (b) Roberts (1992) convergence diagnostic, chains started at mean; (c) Roberts (1992) convergence diagnostic, chains started 2 standard deviations above mean; (d) Smoothed log Roberts (1994) convergence diagnostic: solid line, log Dn; dotted line, log In.

In addition, compute $\hat{e}_\nu(A) = 2\sqrt{2/\pi}[\int_{R^d} K^2(|t|)\,dt]^{1/2} \int_A (\hat{\theta}g(x))^{1/2}\,dx$. Then

$$\widehat{eff}_n(A) = \frac{\hat{I}_n(A)}{2\hat{e}_\nu(A)}$$

is an estimator of the ratio of the expected L^1 error of the kernel density estimator based on the Markov chain output to that of the same kernel estimator based on an iid sample. With the Gibbs sampler, the appropriate product of a marginal and a conditional density would be used instead of $\hat{\theta}_g(x)$ in the foregoing expressions.

Step 6. At $n = n$step, $2n$step, $3n$step ..., construct two convergence monitoring plots: the L^1 error plot of $\hat{I}_n(A)$ versus n, and the efficiency plot of $\widehat{eff}_n(A)$ versus n. Values of $\hat{I}_n(A)$ greater than .3 show that the chain has not produced a satisfactory sample from $\pi(x)$. Failure of the efficiency plot to stabilize around a value less than 2 suggests slow mixing.

Yu warned that both plots can falsely indicate convergence if both the sample path and the chosen set A omit

Table 3. Gibbs Sampler Difference Convergence Criterion Applied to Trivariate Normal with High Correlations. All Table Entries × 10^4

Parameter values / Chain	(0.0, 0.0, 0.0)	(−1.0, −1.0, 1.0)	(0.0, −5.0, 10, 0)	(−1.0, 0.0, 10.0)
1	−6.67 (±8.46)*	.36 (±.37)*	1.21 (±5.05)*	−.060 (±.013)
2	2.83 (±8.88)*	−.25 (±.39)*	5.33 (±5.91)*	−.23 (±.13)*
3	10.04 (±7.19)*	−.75 (±.36)	.60 (±5.29)*	.002 (±.011)*
4	−6.33 (±8.77)*	−.32 (±.42)*	−4.28 (±6.24)*	−.044 (±.011)
5	−2.66 (±6.57)*	−2.00 (±.04)	14.70 (±9.29)*	.059 (±.009)

* The 95% credible set includes zero.
NOTE: Standard errors in parentheses.

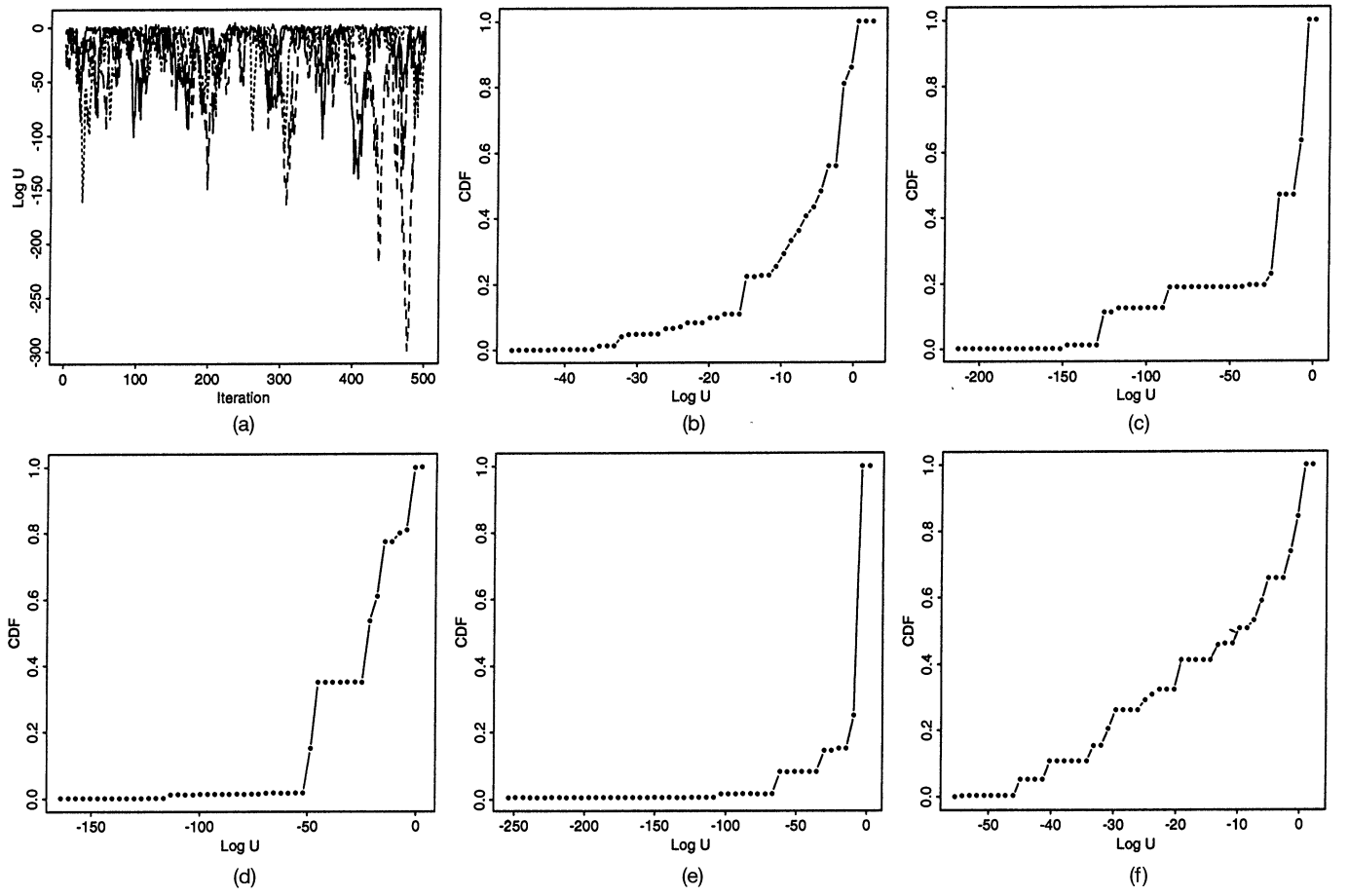


Figure 4. Liu, Liu, and Rubin Plots for Trivariate Normal, High Correlations, Log-Transformed U Statistics. (a) G&R = 1.33, 1.89; (b) iteration 100; (c) iteration 200; (d) iteration 300; (e) iteration 400; (f) iteration 500.

the same important mode of $\pi(x)$. If the dimension d is large, then computing resources may preclude integrating over a compact set A, and it may be necessary instead to choose as many points x_j as feasible at which to evaluate $(\hat{\pi}_n(x_j) - \hat{\theta}_g(x_j))$. To date, this method has not been illustrated for dimensions larger than two.

2.13 Yu and Mykland

Yu and Mykland (1994) proposed a graphical procedure based on cusum path plots applied to a univariate summary statistic (such as a single parameter) from a single chain from any MCMC sampler. Another method, such as a sequential plot of the values of the summary statistic from iteration to iteration, first must be used to determine the number of burn-in iterations n_0 to discard. Cusum path plots are then constructed as follows for iteration $n_0 + 1$ to iteration n, the last iteration generated.

If the chosen summary statistic is designated $T(X)$, then the estimate of its mean based on the retained iterates is

$$\hat{\mu} = \frac{1}{n - n_0} \sum_{j=n_0+1}^{n} T(X^{(j)}),$$

and the observed cusum or partial sum is

$$\hat{S}_t = \sum_{j=n_0+1}^{t} [T(X^{(j)}) - \hat{\mu}], \qquad t = n_0 + 1, \ldots, n.$$

The cusum path plot is obtained by plotting $\{\hat{S}_t\}$ against $t, t = n_0 + 1, \ldots n$, and connecting the successive points. Such a plot will always begin and end at zero.

Yu and Mykland showed that the slower-mixing ("stickier") the MCMC process, the smoother the cusum plot will be and the farther it will wander from zero; conversely, a "hairy" cusum path plot indicates a fast-mixing chain. They suggested comparing the cusum plot from an MCMC sampler to a "benchmark" cusum path plot obtained from iid variates generated from a normal distribution with its mean

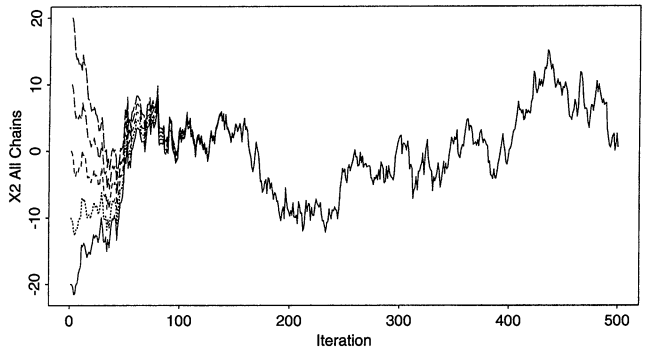


Figure 5. Johnson Plots for Trivariate Normals, High Correlations.

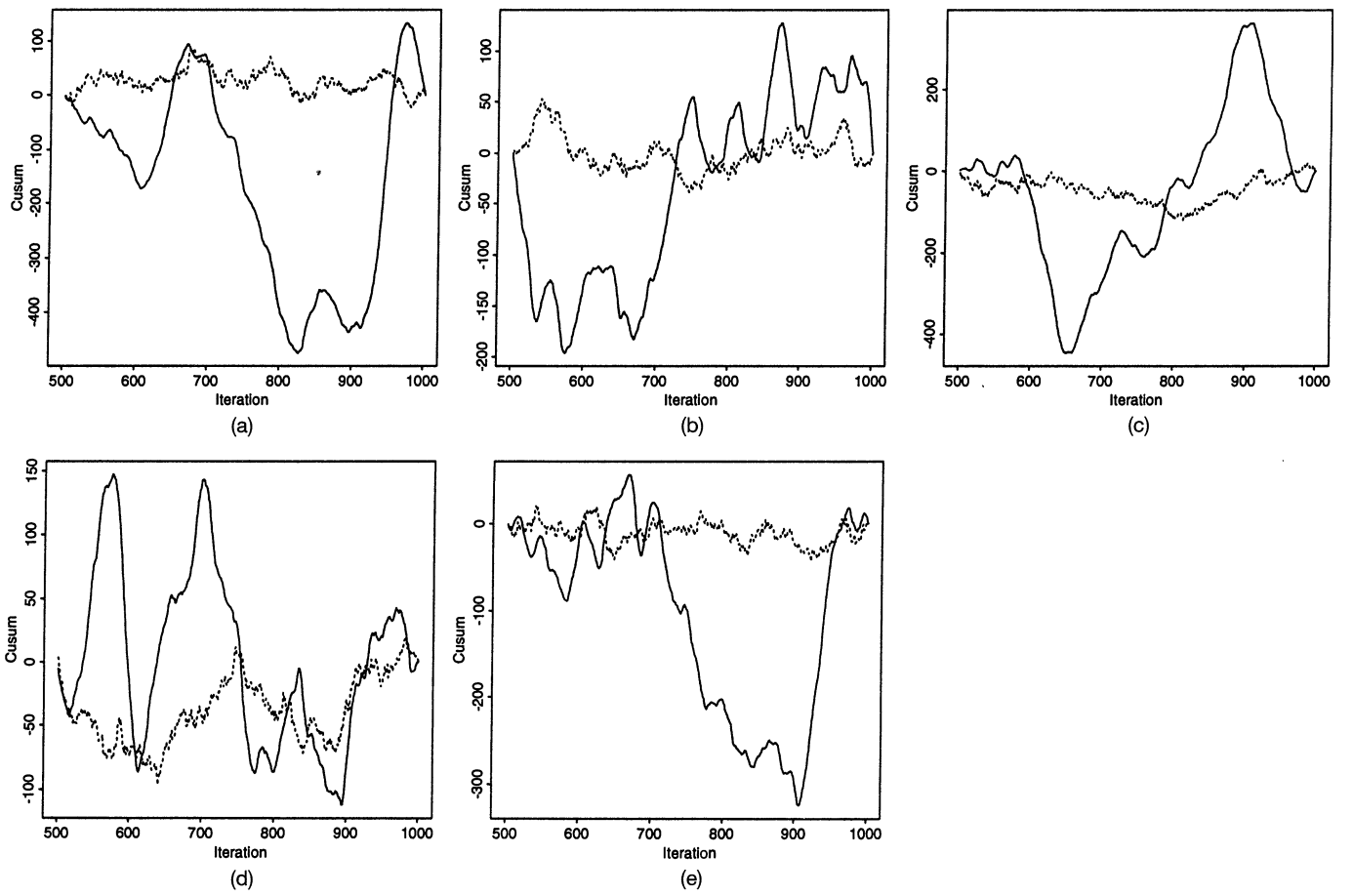


Figure 6. Cusum Path Plots for Trivariate Normals, High Correlations. (a) X_2, chain 1; (b) X_2, chain 2; (c) X_2, chain 3; (d) X_2, chain 4; (e) X_2, chain 5.

and variance matched to the sample mean and variance of the MCMC iterates.

Yu and Mykland posited that the cusum plot may obviate the need in convergence diagnosis for additional information beyond that contained in the output of a single chain. But they stated that, like other convergence diagnostics, their method may fail when some regions of the sample space are much slower-mixing than others.

Clearly, Yu and Mykland's approach is not a stand-alone diagnostic, because another method is required for determining burn-in. It may be useful in identifying samplers that are so slow-mixing that an alternative algorithm or parameterization should be sought if the entire parameter space is to be traversed in a reasonable number of iterations. Because it assess dependence between iterations, it indirectly addresses variance as well as bias in estimation.

Table 4. Means and Standard Errors Estimated from Gibbs Samples: Trivariate Normals, High Correlations, Iterations 501–1,000

Pooled sample of five chains, n = 2,500

		Batch means method			
		25 batches, size 100		10 batches, size 250	
Mean	Naive standard error	Standard error	Lag 1 autocorrelation	Standard error	Lag 1 autocorrelation
−.925	.092	.546	.493	.533	.100

Individual chains, n = 500

			25 batches, size 20		10 batches, size 50		
Chain	Mean	Naive standard error	Standard error	Lag 1 autocorrelation	Standard error	Lag 1 autocorrelation	Time series NSE
1	1.744	.238	.968	.511	1.113	.159	.579
2	−.891	.156	.507	.045	.491	−.007	.339
3	−.818	.231	.927	.550	1.281	−.028	.574
4	−1.603	.163	.547	.221	.557	−.229	.389
5	−3.056	.166	.539	.301	.677	.227	.402

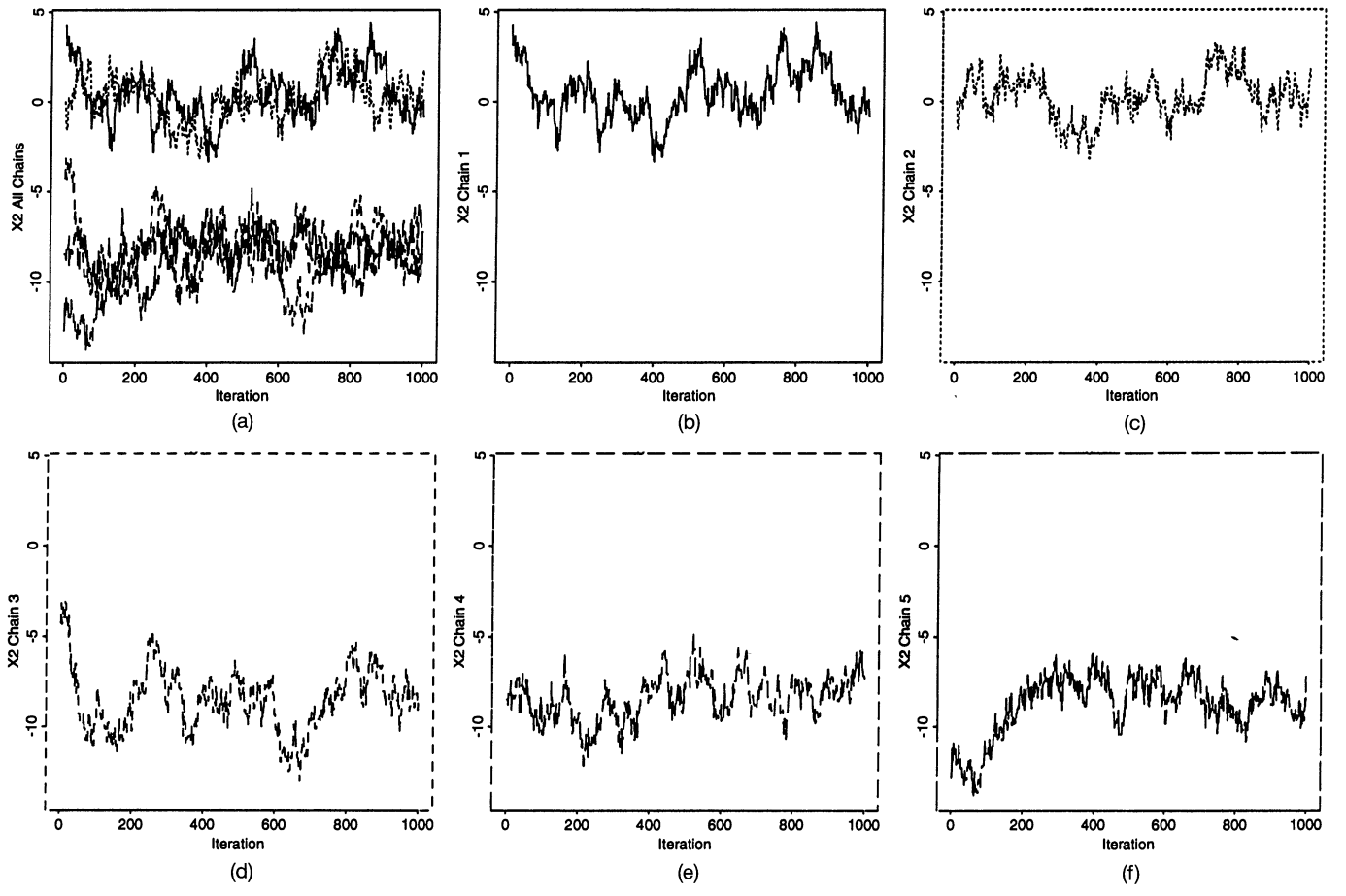


Figure 7. Plots for Bimodal Mixture of Trivariate Normals. (a) G&R = 6.09, 10.47; (b) G = −.24, H&W = ---, lag 1 autocorrelation .94; (c) G = −2.92, H&W = ---, lag 1 autocorrelation .93; (d) G = −3.9, H&W = ---, lag 1 correlation .97; (e) G = 2, H&W = 600, lag 1 autocorrelation .92; (f) G = 9.21, H&W = ---, lag 1 autocorrelation .9.

The straightforward computer code required to produce cusum path plots could be written once and applied to any problem.

Table 1 summarizes the following features of the convergence diagnostics:

- Quantitative/Graphical: Is the measure of convergence quantitative or graphical?
- Single or Multiple Chains: Does the method require a single MCMC chain or multiple parallel chains?
- Theoretical Basis: What is the theoretical basis for the method?
- Univariate/Full Joint Distribution: Does the method apply to univariate quantities or to the full joint posterior distribution?
- Bias/Variance: Is the method intended to address *bias* (i.e., the distance of the estimates of quantities of interest obtained at a particular iteration from the true values under the target distribution) or *variance* (i.e., the quality of those estimates)? We caution that a method's intent and its actual result may differ; see Sections 3.2 and 4.2 for illustrations.
- Applicability: Can the method be applied to the output of any MCMC algorithm, or is it applicable only to the Gibbs sampler? An entry of "some" in this column indicates that although the method is applicable to at least some MCMC algorithms in addition to the Gibbs sampler, there are restrictions as to type of either target distribution or generating algorithm with which it may be used.
- Ease of Use: How easy to use the method is, on the following scale:

a. Generic computer code is available to implement it.
b. Generic code may be written once and applied to the MCMC output for any problem.
c. Problem-specific code must be written.
d. Analytical work, as well as problem-specific code, is needed.

3. NUMERICAL ILLUSTRATION: TRIVARIATE NORMAL WITH HIGH CORRELATIONS

3.1 Simulation Details

We first tested 10 of the 13 convergence diagnostics on a trivariate normal with high correlations of .90, .90, and .98; specifically,

$$N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1.0 & 4.5 & 9.0\\4.5 & 25.0 & 49.0\\9.0 & 49.0 & 100.0\end{bmatrix}\right).$$

To test whether the various methods could detect convergence failure or ambiguity, we ran the samplers for rela-

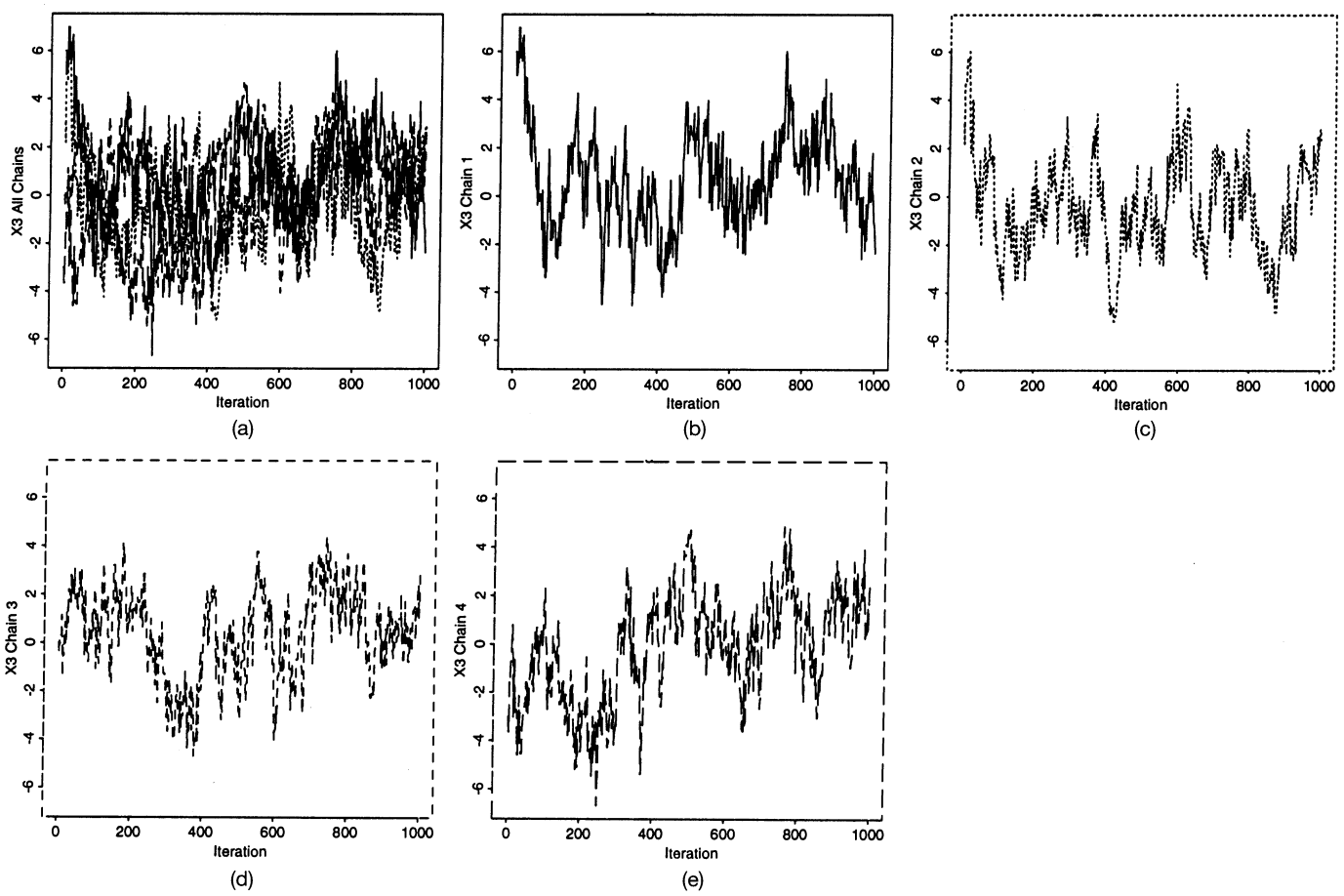


Figure 8. Plots for Bimodal Mixture of Trivariate Normals, Alternate Starting Values. (a) G&R = 1.08, 1.23; (b) G = −.42, H&W = ---, lag 1 autocorrelation .87; (c) G = 1.37, H&W = 1,000; lag 1 autocorrelation .92; (d) G = −.77, H&W = ---, lag 1 autocorrelation .91; (e) G = −.19, H&W = 600, lag 1 autocorrelation .9.

tively few iterations. Because their implementation is too complex for general use in applied work, the methods of Garren and Smith (1993), Mykland et al. (1995), and Yu (1994) were not tested.

For the parameter X_2, Figure 1 shows the traces of five parallel chains run for 500 iterations and the associated Gelman and Rubin shrink factors, Geweke convergence diagnostics, and results of Heidelberger and Welch's method. Results for the other two parameters were similar. The high correlations among the parameters cause the traces of different parameters in the same chain to be virtually identical in shape, though different in scale. Gelman and Rubin's shrink factor suggests that the chains have not completely mixed and that running additional iterations would appreciably improve the sharpness of estimation. Despite the fact that the visual impression given by the trace is that stabilization has not occurred, Geweke's diagnostic for the third chain suggests satisfactory convergence; however, for the other chains, the values of the Geweke diagnostic for all three parameters are well outside the range of .95 probability for standard normal variates. As a check, we extended the same 5 chains to a total of 1,000 iterations. Here Gelman and Rubin's shrink factors suggest improved convergence. Geweke's diagnostic applied to iterations 501–1,000 suggests satisfactory convergence of all parameters in all chains except chain 1.

We performed Heidelberger and Welch's process on each individual chain, using the 500 iterates already run as both the initial stopping point j_1 and the maximum run length $j_{\max}$. We used the time series functions in S-Plus to compute $S(0)$ and numerical integration to compute the value of the Cramer–von Mises statistic; we considered the stationarity test to have been passed if the result was less than .46, the .95 quantile for the Cramer–von Mises statistic.

Table 5. Raftery and Lewis's Method Applied to Bimodal Mixture

	$q = .025$			$q = .25$			$q = .50$		
Chain	k	nburn	nprec	k	nburn	nprec	k	nburn	nprec
1	1	15	1,788	3	45	42,111	1	37	42,762
2	1	30	3,218	1	35	28,833	1	34	39,744
3	1	18	2,012	1	59	47,828	2	54	63,652
4	1	21	2,299	4	64	58,896	3	57	71,833
5	1	30	3,218	1	47	37,657	4	64	80,180

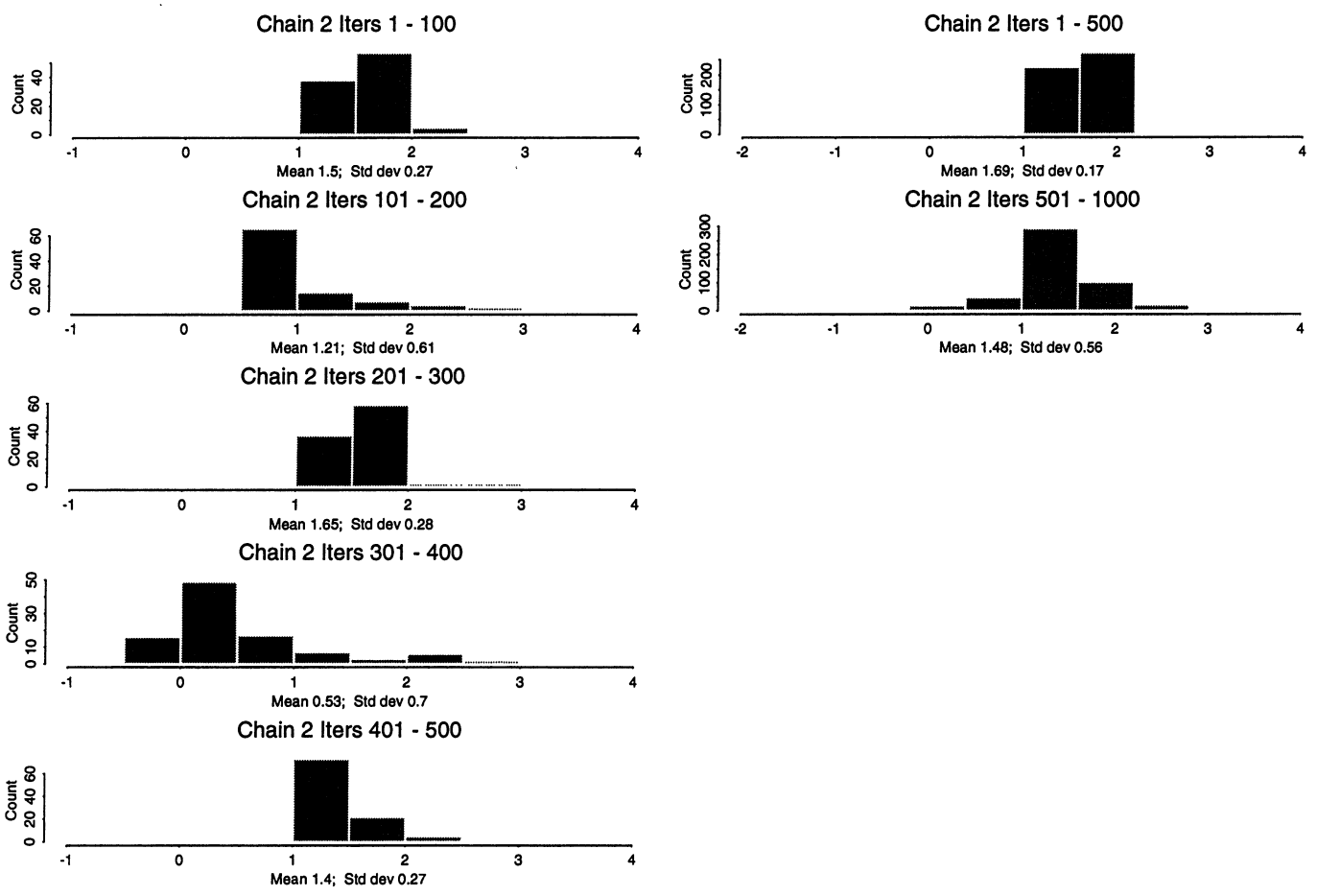


Figure 9. Histograms of Log Gibbs Stopper Statistics, Bimodal Mixture.

The number labeled "H&W=" associated with graph for each individual chain in Figure 1 is the number of iterates that remained when the stationarity test was passed. "H&W = —" indicates that a stationary portion of length at least half the total run length was not found. For each parameter in chain 2, Heidelberger and Welch's procedure determined that after discarding the initial 200 iterations, the remaining 300 iterates formed a stationary sequence. The graphs of the individual chains in Figure 1 are also annotated with the lag 1 autocorrelations within the respective chains, all of which are quite large.

Table 2 presents the results of running an S-Plus implementation of Raftery and Lewis's method on these five chains for the parameter X_2 (Best et al. 1995). The three sets of columns correspond to values of q equal to .025, .25, and .50. In all cases, $r = .01, s = .90$, and $\delta = .001$. The fact that k is usually larger than 1 and that "$n$prec" is much larger than the minimum sample size that would be required if the observations were independent, suggests high autocorrelations within the chains. Indeed, when there are high correlations among parameters, the fact that the Gibbs sampler algorithm is based on full conditionals results in high autocorrelations within chains, as had been noted in Figure 1. There is no apparent relationship between the values of Geweke's convergence diagnostic, which indicates good convergence of chain 3, and Raftery and Lewis's results.

In calculating Gibbs Stopper weights, Cui et al. (1992) advocated using small batch sizes (perhaps 50–100) for early iterations in a chain and moving to progressively larger batch sizes for later iterations. The first column of Figure 2 shows histograms as well as means and standard deviations of the Gibbs Stopper statistic for batches of size 100 for the first 500 iterations of the first chain. Although the weights are clustered around 1.0 in all histograms, the dispersion is much greater than that observed for uncorrelated samples, and the trend toward tighter clustering in later batches is not monotonic. The second column of the figure shows similar plots for batches of size 500 covering 1,000 iterations of the same chain. The means of these larger batches are much closer to the true normalization constant, 1.0. Clearly, the choice of batch size affects the interpretation of diagnostic.

Roberts's (1992) method was applied to the trivariate normal problem with high correlations by running 1,000 iterations of the reversible sampler for 10 replications of each of 3 starting points. The values of the resulting statistics are shown in Figure 3, a–c. For the chains started two standard deviations away from the mean, the visual impression is of rapid convergence in roughly 50 iterations, but from there on there are frequent excursions away from the value of 1.0. For the chains started at the mean, the statistic starts much nearer to 1.0 than in the other chains (2.2 versus 15–20), but the values do not move closer to 1.0 with successive iterations. Figure 3d shows Roberts's new (1994) diagnostic cal-

culated from 11 chains started from dispersed initial values arranged symmetrically around the mean. The plotted lines are "lowess" smooths (Cleveland 1979) of log-transformed I_n and D_n values. The fact that both sequences appear to stabilize and become equal after about 50 iterations suggests very rapid convergence.

Table 3 shows Zellner and Min's (1995) Gibbs sampler difference convergence criterion calculated for each chain at four sets of parameter values. All table entries have been multiplied by 10^4. The values in the last column are very small because the last combination of parameter values has very low probability given the mean structure and high correlations in this example; thus $p(\alpha|\beta)$ and $p(\beta|\alpha)$ are almost zero. Thus judicious choices for the values of α and β at which to evaluate this convergence diagnostic, as well as the use of credible sets rather than some predetermined criterion for "small," clearly are essential. It is not clear what conclusions should be drawn regarding convergence of chains 1 and 3–5, for each of which some of the parameter values produced 95% credible sets containing zero and others did not. Perhaps a larger posterior sample would clarify matters, because this would increase the criterion's power to detect differences from zero, but the proper size for such a sample is not clear.

Results of applying Liu et al.'s (1992) method to the trivariate normal with high correlations are presented in Figure 4. The traces of the values from four independent chains do not indicate "mixing" or "settling down," and Gelman and Rubin's diagnostic applied to them suggests a need for additional iterations. The plots of the empirical cdf, here taken every 100th iteration, likewise do not indicate convergence toward a consistent pattern.

Figure 5 shows traces of five chains run from dispersed starting values but using the same random number seed for every chain as required for Johnson's (1994) method. The five chains converge to a single sample path within 150 iterations.

Yu and Mykland's (1994) diagnostic for X_2 in all five chains is shown in Figure 6. The solid lines are the cusum path plots for the Gibbs iterates, and the dotted lines are the benchmark plots. The smoothness and large excursions away from zero in the Gibbs sampler plots are indicative of slow mixing.

Table 4 shows the means and standard errors estimated from Gibbs sampler output for X_2 from iterations 501–1,000. The distances of the estimates of means for the three parameters from their true values of zero make clear that these iterates have not converged to a sample from the true target distribution and that there is substantial bias in estimation based on them. Due to the autocorrelations within chains, the standard errors produced by the three methods are quite different. The naive standard errors approximate the target standard deviations over the square root of n, and are too small. We were unable to find a batch size that consistently kept autocorrelations within batches lower than .05, although using 10 batches came close. But the standard error estimates based on only 10 batches are likely to be inaccurate. The time series numeric standard errors may provide a good compromise, but more likely are too small. Many of the estimated means are not within two standard errors of the truth, even using the most conservative estimates of standard error.

3.2 Comparative Remarks

Raftery and Lewis's (1992) method indicates that 1,000–4,000 iterations are needed for the various chains and parameters to estimate the .025 quantile to the specified accuracy. Estimates of this quantile based on iterations 501–1,000 fell into the correct interval 9 out of the possible 15 times. In this example, Geweke's diagnostic appears to be premature in diagnosing convergence in four of the five chains (not shown for 1,000 iterations). That Gelman and Rubin's (1992) shrink factors are as near 1 as they are may be consistent with the fact that, for the pooled sample, all means are within two batch means method standard errors of the truth; however, if the starting points had not been chosen with prior knowledge of the true joint posterior, then the pooled sample might well have done no better than the individual chains. The results of some of the other diagnostics are difficult to interpret. Values of Roberts's (1992) diagnostic and the Gibbs Stopper are extremely variable and do not show a monotonic trend, whereas Roberts's (1994) diagnostic indicates very rapid convergence. Zellner and Min's (1995) diagnostic gives different results depending on the point at which it is evaluated. Liu et al.'s (1992) diagnostic, after log transformation, does appear to give clear evidence of convergence failure, whereas Johnson's (1994) method provides equally clear evidence of convergence at 150 iterations. Yu and Mykland's (1994) cusum path plots reveal "stickiness" of the chains.

The disagreements among the various methods may be explained in part by different connotations of the word "convergence." Once a single draw from the target distribution has been obtained, the sampler has "converged" in the sense that all subsequent iterates are also drawn from the target distribution. In the trivariate normal example, this probably occurs after fairly few iterations, as indicated by the methods of Roberts and Johnson. On the other hand, particularly in the presence of high correlations among the parameters as in this example, many more iterations may be required to obtain "convergence" (or "mixing") in the sense that the Markov chain has traversed the entire distribution, so that the resulting samples yield good estimates of the quantities of interest.

We compared run times of the various diagnostics. To compare the ordinary Gibbs sampler with the reversible sampler needed for Roberts' method, we timed the running of the nine parallel chains of the ordinary algorithm that were used for Liu et al.'s method and the running of three replicates from each of three starting points for the reversible sampler. All simulations were coded in the C language and run on a Sun SPARC station 10. The run times were .68 second for 1,000 iterations of the ordinary Gibbs sampler and 5.509 seconds for 2,000 iterations of the reversible sampler (needed to assess convergence at 1,000

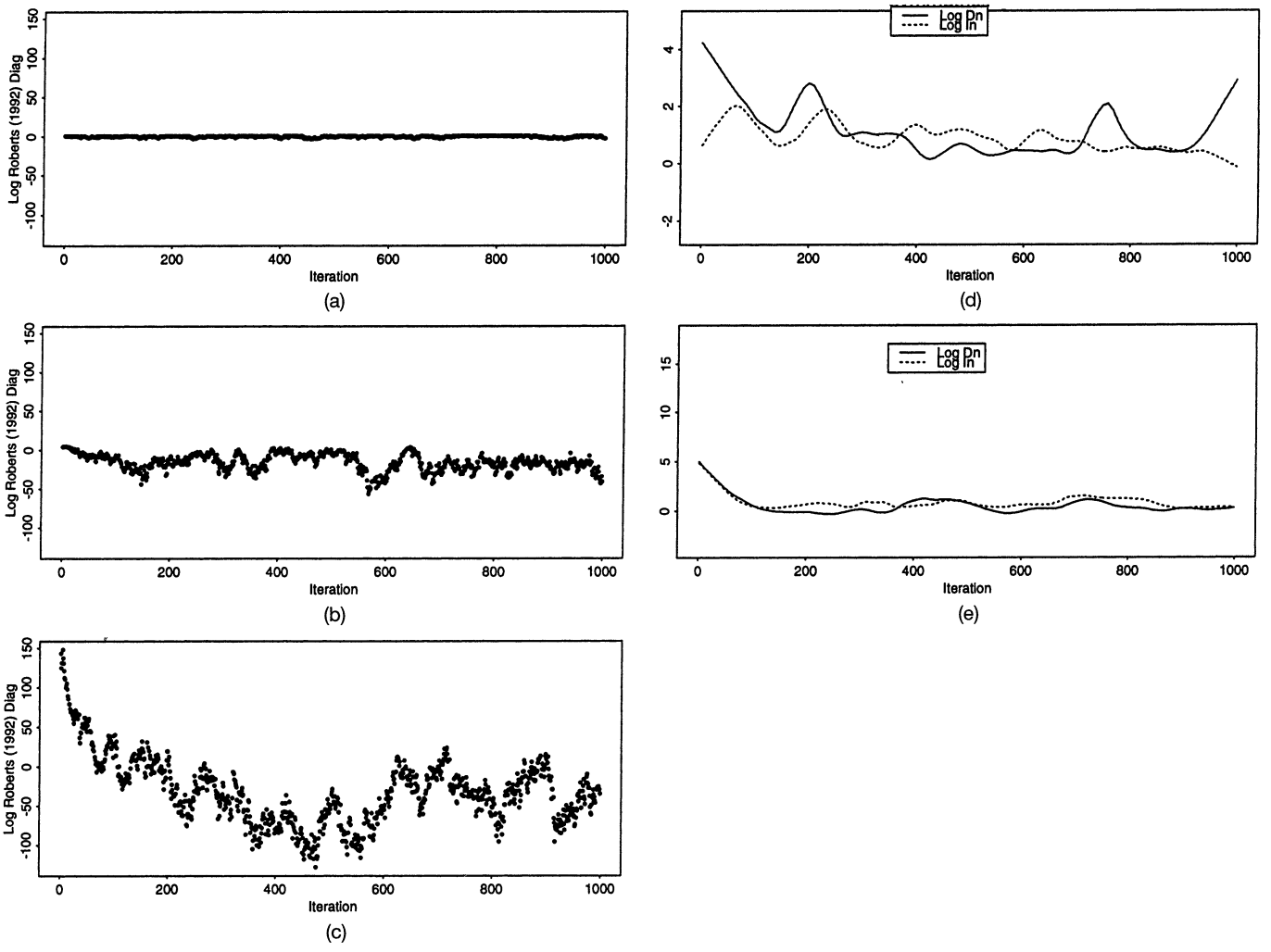


Figure 10. Roberts (1992, 1994) Convergence Diagnostics for Bimodal Mixture. (a) Log Roberts (1992) convergence diagnostic, all chains started at upper mode; (b) log Roberts (1992) convergence diagnostic, all chains started halfway between modes; (c) log Roberts (1992) convergence diagnostic, all chains started 6 standard deviations below lower mode; (d) smoothed Roberts (1994) convergence diagnostic, chains started symmetrically around both nodes: solid line, log Dn; dotted line, log In; (e) smoothed Roberts (1994) convergence diagnostic, chains started symmetrically around single mode: solid line, log Dn; dotted line, log In.

iterations), which included computation of the Roberts diagnostic for each iteration. A C program to read in the Gibbs samples from files and to compute the Gibbs Stopper weights based on all 9000 ordinary Gibbs sampler iterates took 64.55 seconds. Similar programs to apply Zellner and Min's and Liu et al.'s methods took 4 seconds and 12 seconds. S-Plus programs took 30 seconds to read-in the files of Gibbs iterates, compute Gelman and Rubin's and Geweke's diagnostics, apply Heidelberger and Welch's method, and display the graphs and diagnostics. Raftery and Lewis's program took less than 1 second to perform its computations on each given chain; however, a more convenient user interface would make it more efficient to use for multiple parameters in the same problem. The S-Plus program for Raftery and Lewis's diagnostic took 1 minute for calculations for 3 quantiles of each of 3 parameters based on 500 iterations and 5 chains. Finally, S-Plus code produced 5 cusum path plots based on 1,000 iterations in 23 seconds.

4. NUMERICAL ILLUSTRATION: BIMODAL MIXTURE OF TRIVARIATE NORMALS

4.1 Simulation Details

We next tested nine convergence diagnostics on a bimodal target density consisting of a mixture of two trivariate normals with equal probability. They shared a common covariance matrix producing high correlations,

$$\begin{bmatrix} 1.0 & 1.3 & 1.5 \\ 1.3 & 2.0 & 2.0 \\ 1.5 & 2.0 & 4.0 \end{bmatrix},$$

and their mean vectors—$(0.0, 0.0, 0.0)$ and $(-6.0, -8.49, -12.0)$—were sufficiently separated so that the marginal densities, as well as the joint, were bimodal but not so far separated that the state space would be effectively disconnected. We used this example to illustrate nonconjugate full conditionals as well as bimodality. A random-walk Metropolis algorithm (Tierney 1994) was used to generate from each unnormalized full conditional.

Table 6. Gibbs Sampler Difference Convergence Criterion Applied to Bimodal Mixture of Trivariate Normals

Parameter values chain	(0.0, 0.0, 0.0)	(−6.0, −8.5, −12.0)	(−3.0, −4.2, −6.0)	(−1.0, −1.4, −2.0)
1	12.20 (±3.03)	0 (±0)**	0 (±0)	−4.32 (±1.74)
2	5.49 (±3.11)*	0 (±0)**	0 (±0)*	−2.85 (±1.54)*
3	0 (±0)**	355.5 (±16.6)	.12 (±.16)*	0 (±0)*
4	0 (±0)**	328.5 (±14.2)	1.11 (±.88)*	0 (±0)*
5	0 (±0)**	415.9 (±15.9)	.003 (±.0009)	0 (±0)**

NOTE: All table entries have been multiplied by 10^3. An asterisk indicates that the 95% credible set includes zero; a double asterisk indicates that the estimated variance of the test statistic was zero to eight significant digits, so a credible set could not be calculated.

Nine parallel chains were run with starting values chosen at equal intervals from above the upper mode to below the lower mode. Plots for every other one of those chains for X_2 are shown in Figure 7. Obviously none of the chains traverses the entire state space; each remains in the vicinity of one of the modes. Gelman and Rubin's diagnostic clearly indicates convergence failure, but Geweke's and Heidelberger and Welch's diagnostics suggest satisfactory convergence in one case each. Figure 8 shows similar plots for a different subset of the nine chains, this time the ones started with the four highest initial values. All four of these chains "got stuck" at the same mode. Gelman and Rubin's method fails to detect the problem, Heidelberger and Welch's diagnostics imply good convergence in half the cases, and Geweke's diagnostics imply good convergence in all cases.

Raftery and Lewis's method, for which the results are given in Table 5, does not appear to find convergence in this bimodal example appreciably worse than that in the highly correlated trivariate normal problem reported in Table 2. Values for k and "nburn" are very similar in the two tables, although values for "nprec" are generally somewhat larger in the bimodal example.

The first column of Figure 9 shows histograms and means and standard deviations of log-transformed Gibbs Stopper weights for the first 500 iterations of the second chain from Figure 7. Even with the variance-reducing log transformation (recommended by Ritter and Tanner when the untrans-

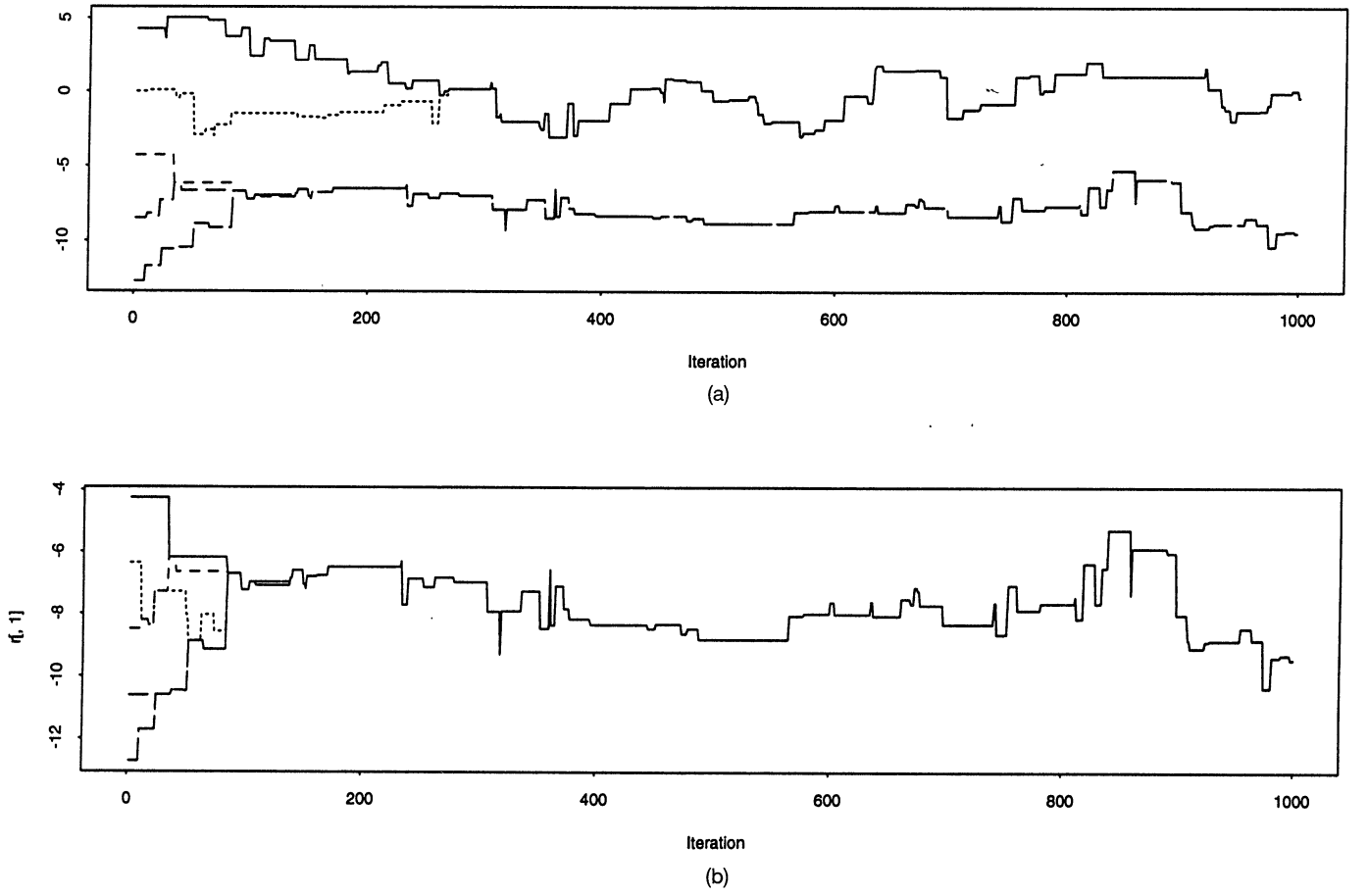


Figure 11. Johnson Plots for Bimodal Mixture. (a) X_2, starting values symmetric around both modes; (b) X_2, starting values symmetric around lower mode.

formed weights are extremely variable, as in this case), there is no apparent closer clustering around any particular value in successive batches, which might be interpreted as correctly diagnosing convergence failure. But, as shown in the second column of this figure, the mean and standard deviation are very similar in two batches of size 500 from the same chain, which might appear to imply convergence. Gibbs Stopper results are difficult to interpret, particularly because they depend on choices of batch sizes and whether or not to log-transform the weights.

Log-transformed Roberts (1992) convergence diagnostics for three sets of three reversible chains are shown in Figure 10, a–c. Because this diagnostic is intended to assess convergence of the full joint distribution, it is disappointing that it too can fail when the chains are started at one of the modes. In Figure 10d, lowess smooths of Roberts's (1994) log-transformed diagnostic are shown as calculated from 11 chains with initial values chosen symmetrically around both modes. Here convergence failure is indicated by the fact that the values of $\log D_n$ fail to stabilize and remain generally larger than those of $\log I_n$. In contrast, Figure 10e shows lowess smooths of Roberts' (1994) diagnostic with log transformation as calculated from 11 chains with initial values chosen around the upper mode, so that all chains remain in the vicinity of the same mode. When the initial values are insufficiently dispersed in this manner, even this method fails to detect the failure to sample from the full target distribution.

Table 6 shows Zellner and Min's Gibbs sampler difference convergence criterion calculated for iterations 501–1,000 of each chain at four sets of parameter values. When evaluated at points at or near a mode, the diagnostic clearly identifies convergence failure in a chain that has gotten stuck at that mode. But marginal probabilities estimated from a chain stuck at the other mode are so small that all estimates of the convergence diagnostic are identically zero. Results are less predictable at points that are of low probability under either distribution in the bimodal mixture.

Cusum path plots for the bimodal example were very similar in appearance to those shown in Figure 6 for the unimodal example, again illustrating slow mixing of the sampler.

Because Johnson's method cannot be applied when random-walk Metropolis steps are used within a Gibbs sampler, we ran a new set of chains using independence-chain Metropolis steps to generate from each nonconjugate full conditional; we used the same sequences of candidate values and uniforms used in the acceptance/rejectance step for all chains. Figure 11a shows the results for five chains with initial values chosen symmetrically around both modes. As in the Gelman and Rubin plots, three of the chains coalesce around one mode and the other two around the other mode. Convergence failure is obvious. Figure 11b is a similar plot of five chains with initial values chosen around the lower mode. From approximately iteration 100 on, all chains are traversing the same sample path. Thus with inadequate dispersion of initial values, Johnson's method also can incorrectly indicate convergence.

4.2 Comparative Remarks

This example also demonstrates the effect of nonconjugate full conditionals on computer run times of the various methods. Run time for the ordinary Gibbs sampler using the Metropolis algorithm (with either random walk or independence kernel) with one transition per iterate was 11 seconds for nine chains of 1,000 iterations each. Run times for Gelman and Rubin's, Geweke's, Raftery and Lewis's, and Yu and Mykland's methods were unchanged from those reported for the trivariate normal. But the requirement of the Gibbs Stopper and Roberts' method to estimate the normalizing constants for the full conditionals both complicated coding and radically increased execution time, which was 1 hour and 40 minutes for the Gibbs Stopper and 5 minutes for Roberts's method for 500 and 2,000 iterations. The same would have been true for Zellner and Min's method had we not in fact used the known normalizing constants for the required conditional distributions.

5. SUMMARY, DISCUSSION, AND RECOMMENDATIONS

Our summary in Section 2 shows that many of the MCMC diagnostics proposed in the statistical literature to date are fairly difficult to use, requiring problem-specific coding and perhaps analytical work. In addition, our empirical results in Sections 3 and 4 indicate that although many of the diagnostics often succeed at detecting the sort of convergence failure they were designed to identify, they can also fail in this role—even in low-dimensional idealized problems far simpler than those typically encountered in statistical practice. Taken together, our results call for caution when using these diagnostics and for continued research into both the theoretical and applied aspects of MCMC algorithms.

Concerning generic use of MCMC methods, we advocate a variety of diagnostic tools rather than any single plot or statistic. In our own work, we generally run a few (three–five) parallel chains, with starting points drawn (perhaps systematically, rather than at random) from what we believe is a distribution overdispersed with respect to the stationary distribution. We visually inspect these chains by overlaying their sampled values on a common graph for each parameter or, for very high-dimensional models, a representative subset of the parameters. We annotate each graph with the Gelman and Rubin statistic and lag 1 autocorrelations, because they are easily calculated and the latter helps to interpret the former. (Large Gelman and Rubin statistics may arise from either slow mixing or multimodality.) We also investigate crosscorrelations among parameters suspected of being nearly confounded, as high crosscorrelations may indicate a need for reparameterization.

A clever alternative to running parallel chains is to intersperse Metropolis–Hastings steps into a single Gibbs sampler chain at intervals, using a proposal density that generates candidate values from an overdispersed distribution independently of the current state of the Gibbs sampler chain (see Mykland et al. 1994). When such candidates are accepted and produce regenerations in the chain, diagnostics

requiring multiple independent chains may be applied to the tours. At the same time, such a hybrid sampler forms a single long sample path that gets closer to the stationary distribution than would many independent shorter chains.

We recommend learning as much as possible about the target density before applying an MCMC algorithm, perhaps by using mode-finding techniques or noniterative approximation methods, as well as considering multiple models for a given data set. By steadily increasing the complexity of the model under consideration, we will be more likely to detect convergence failures while at the same time gaining a deeper understanding of the data. Multiple algorithms may also be helpful, because each will have its own convergence properties and may reveal different features of the likelihood or posterior surface.

Clearly, our recommendations imply that automated convergence monitoring (as by a machine) is unsafe and should be avoided. This is something of a blow to many applied Bayesians (ourselves included) who at one time looked forward to a fully automated Bayesian data analysis package, similar to the large currently available commercial packages for likelihood analysis. Instead, we now find ourselves recommending a two-stage process, wherein model specification and associated sampling are separated from convergence diagnosis and subsequent output analysis. In fact, one of us has co-developed a collection of S-Plus routines called CODA (Best et al. 1995) for the second stage of this process, with the first stage being accomplished by BUGS, the recently-developed software package for Bayesian analysis using Gibbs sampling (Spiegelhalter, Thomas, and Best 1994, 1995; Thomas, Spiegelhalter, and Best 1992). Both of these programs and their manuals are freely available from the MRC Biostatistics Unit at the University of Cambridge (e-mail address: bugs@mrc-bsu.cam.ac.uk).

Another emerging approach to MCMC analysis is to concentrate on convergence *acceleration,* rather than on the less soluble problem of convergence diagnosis. Clever reparameterization can often substantially improve correlation structure within a model, and hence speed convergence (see Hills and Smith 1992 for a general discussion and Gelfand et al. 1995a,b for treatments specific to hierarchical random effects models). More sophisticated MCMC algorithms can also offer impressive reductions in time to convergence. Promising ideas in this regard include the use of auxiliary variables (Besag and Green 1993; Swendsen and Wang 1987), resampling and adaptive switching of the transition kernel (Gelfand and Sahu 1994), and multichain annealing or "tempering" (Geyer and Thompson 1995).

In summary, a consensus appears to be emerging that the proper approach to MCMC monitoring lies somewhere between the two extremes recently advocated by Geyer (one long chain, perhaps plotted versus iteration) and Gelman and Rubin (a single multichain diagnostic). Although it is never possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution, convergence diagnostics (along with sample correlations and plots of the samples themselves) may offer a worthwhile check on the algorithm's progress. Combined with the emerging work in determining theoretical convergence bounds and a more robust approach to algorithm and model selection, MCMC algorithms will no doubt enjoy continued popularity as computational tools for a wide array of statistical problems.

[Received September 1994. Revised September 1995.]

REFERENCES

Albert, J. H. (1992), "A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters," *The American Statistician*, 46, 246–253.

——— (1993), "Teaching Bayesian Statistics Using Sampling Methods and MINITAB," *The American Statistician*, 47, 182–191.

Andrews, R. W., Berger, J. O., and Smith, M. H. (1993), "Bayesian Estimation of Fuel Economy Potential due to Technology Improvements" (with discussion), in *Case Studies in Bayesian Statistics*, Lecture Notes in Statistics, Vol. 83, eds. C. Gatsonis, J. S. Hodges, R. E. Kass, and N. D. Singpurwalla, New York: Springer-Verlag, pp. 1–77.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 1–52.

Best, N. G., Cowles, M. K., and Vines, K. (1995), "CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30," technical report, University of Cambridge, MRC Biostatistics Unit.

Buck, C. E., Litton, C. D., and Stephens, D. A. (1993), "Detecting a Change in the Shape of a Prehistoric Corbelled Tomb," *The Statistician*, 42, 483–490.

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

Chan, K. S., and Geyer, C. J. (1995), Discussion of "Markov Chains for Exploring Posterior Distributions," by L. Tierney, *The Annals of Statistics*, 22, 1747–1758.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cui, L., Tanner, M. A., Sinha, D., and Hall, W. J. (1992), "Comment: Monitoring Convergence of the Gibbs Sampler: Further Experience With the Gibbs Stopper," *Statistical Science*, 7, 483–486.

Garren, S. T., and Smith, R. L. (1993), "Convergence Diagnostics for Markov Chain Samplers," technical report, University of North Carolina, Dept. of Statistics.

Gelfand, A. E., and Carlin, B. P. (1993), "Maximum Likelihood Estimation for Constrained or Missing Data Models," *Canadian Journal of Statistics*, 21, 303–312.

Gelfand, A. E., and Sahu, S. K. (1994), "On Markov Chain Monte Carlo Acceleration," *Journal of Computational and Graphical Statistics*, 3, 261–276.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995a), "Efficient Parameterizations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

——— (1996), "Efficient Parameterizations for Generalized Linear Mixed Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, to appear.

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.

Gelman, A., and Rubin, D. B. (1992a), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511.

——— (1992b), "Rejoinder: Replication Without Contrition," *Statistical Science*, 7, 503–507.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 721–741.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 169–193.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–483.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

——— (1995), "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference," *Journal of the American Statistical Association*, 90, 909–920.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994), "A Language and Program for Complex Bayesian Modelling," *The Statistician*, 43, 169–178.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Heidelberger, P., and Welch, P. D. (1983), "Simulation Run Length Control in the Presence of an Initial Transient," *Operations Research*, 31, 1109–1144.

Hills, S. E., and Smith, A. F. M. (1992), "Parameterization Issues in Bayesian Inference" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 227–246.

Johnson, V. E. (1994), "Studying Convergence of Markov Chain Monte Carlo Algorithms Using Coupled Sample Paths," *Journal of the American Statistical Association*, 91, 154–166.

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992), "Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4+ Counts" (with discussion), *Journal of the American Statistical Association*, 87, 615–625.

Liu, C., Liu, J., and Rubin, D. B. (1992), "A Variational Control Variable for Assessing the Convergence of the Gibbs Sampler," in *Proceedings of the American Statistical Association, Statistical Computing Section*, pp. 74–78.

Liu, J., Wong, W. H., and Kong, A. (1995), "Correlation Structure and Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society*, Ser. B, 57, 157–169.

MacEachern, S. N., and Berliner, L. M. (1994), "Subsampling the Gibbs Sampler," *The American Statistician*, 48, 188–190.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Meyn, S. P., and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*, London: Springer-Verlag.

Mykland, P., Tierney, L., and Yu, B. (1995), "Regeneration in Markov Chain Samplers," *Journal of the American Statistical Association*, 90, 233–241.

Polson, N. G. (1996), "Convergence of Markov Chain Monte Carlo Algorithms," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, to appear.

Raftery, A. E., and Lewis, S. (1992), "How Many Iterations in the Gibbs Sampler?," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 763–773.

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

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

Roberts, G. O. (1992), "Convergence Diagnostics of the Gibbs Sampler," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 775–782.

——— (1996), "Methods for Estimating L^2 Convergence of Markov Chain Monte Carlo," in *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, I. Chaloner, and J. Geweke, Amsterdam: North-Holland, pp. 373–384.

Roberts, G. O., and Hills, S. (1991), "Assessing Distributional Convergence of the Gibbs Sampler," technical report, University of Cambridge, Dept. of Mathematics.

Rosenthal, J. S. (1993), "Rates of Convergence for Data Augmentation on Finite Sample Spaces," *Annals of Applied Probability*, 3, 819–839.

——— (1995a), "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

——— (1995b), "Rates of Convergence for Gibbs Sampling for Variance Component Models," *Annals of Statistics*, 23, 740–761.

——— (1996), "Analysis of the Gibbs Sampler for a Model Related to James–Stein Estimators," *Statistics and Computing*, to appear.

Schervish, M. J., and Carlin, B. P. (1992), "On the Convergence of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics*, 1, 111–127.

Schruben, L. W., (1982) "Detecting Initialization Bias in Simulation Output," *Operations Research*, 30, 569–590.

Schruben, L., Singh, H., and Tierney, L. (1983), "Optimal Tests for Initialization Bias in Simulation Output," *Operations Research*, 31, 1167–1178.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1996), "Computation on Bayesian Graphical Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, to appear.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1994), "BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.30," technical report, University of Cambridge, MRC Biostatistics Unit.

Swendsen, R. YH., and Wang, J.-S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physics Review Letters*, 58, 86–88.

Thomas, A., Spiegelhalter, D. J., and Gilks, W. R. (1992), "BUGS: A Program to Perform Bayesian Inference Using Gibbs Sampling." in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 837–842.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

Wei, G. C. G., and Tanner, M. A. (1990), "Posterior Computations for Censored Regression Data," *Journal of the American Statistical Association*, 85, 829–839.

Yu, B. (1994), "Monitoring the Convergence of Markov Samplers Based on Estimated L^1 Error," Technical Report 409, University of California at Berkeley, Dept. of Statistics.

Yu, B., and Mykland, P. (1994), "Looking at Markov Samplers Through Cusum Path Plots: A Simple Diagnostic Idea," Technical Report 413, University of California at Berkeley, Dept. of Statistics.

Zellner, A., and Min, C.-K. (1995), "Gibbs Sampler Convergence Criteria," *Journal of the American Statistical Association*, 90, 921–927.