Journal of Statistical
Theory & Methods
1996, Vol. 7, pp 1~12

Convergence Diagnostics for the Gibbs Sampler[1]

Joong Kweon Sohn, Heon Joo Kim and Sang Gil Kang[2]

Abstract The Gibbs sampler is a substantially powerful tool in Bayesian analysis. However, it is necessary to choose the number of iterations and the size of random samples. This problem has been studied by many researchers. The proposed procedures by them are generally difficult to apply to a practical problem. The attraction of the sampling based approaches is their conceptual simplicity and ease of implementation for users with available computing resources but without numerical analytic efforts. In this paper we consider the problem of determining the number of iterations t, which is simple to application.

Keywords : Gibbs Sampler, Convergence Diagnostics, Stopping Rule

1. Some Stopping Rules

An important and critical problem in implementing the Gibbs sampler is how to determine when it stops sampling and uses the samples to estimate a quantity of interest. This problem has been studied by many researchers. Cowles and Carlin(1994) provided an expository review of known ten popular convergence diagnostics. They compared their performances in two simple models and concluded that all ten methods can fail to detect some sorts of convergence failure.

Among the ten convergence diagnostics, those proposed by Gelman and Rubin(1992) and Raftery and Lewis(1992) are most popular. Gelman and Rubin(1992) are strong advocates of the need for multiple parallel chains started at different initial values, m. The diagnostics proposed by Gelman and Rubin(1992) is based on normal approximations to the exact Bayesian posterior distributions and is constitutes two steps. At first an overdispersed estimate of the target distribution is obtained before sampling begins. Here 'overdispersed' is used in a sense of being more variable than the target distribution. The next is to carry

[1] The Present Studies were Supported by the Basic Science Research Institute Program, Ministry of Education, 1995, Project No. BSRI-95-1403.

[2] Department of Statistics, Kyungpook National University, Daegu 702-701, Korea

out the Gibbs sampler for each quantity for the desired number of iterations, $2n$. It involves using the last n iterations to re-estimate the target distribution. They evaluated the between-chain variance and the within-chain variance of the samples with the last n iterations. Let B be the variance between the means from the m parallel chains, W be the average of the m within-chain variances and df be the degrees of freedom of the approximating t density. Then the shrink factor,

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

converges to 1 as n goes to infinity. So by estimating the factor, convergence of the iterative simulation is monitored. This method can be used not only on the Gibbs sampler but also on other iterative methods. But since the Gibbs sampler is most needed when the normal approximation to the posterior distribution is inadequate, reliance on normal approximation for diagnosing convergence to the true posterior is questionable.

On the other hand, Raftery and Lewis(1992) claimed a single-chain Gibbs sampler. The proposed approach is based on two-state Markov chain theory and is used only for the Gibbs iteration. Also they concluded that the required number of iterations can be dramatically different for different problems and even for different quantities of interest within the same problem, so it would seem to be important to use some methods to determine the number of iterations.

Geweke(1992) also claimed a single-chain Gibbs sampler and used the methods from spectral analysis to assess convergence of the Gibbs sampler. It is assumed that the nature of the Gibbs sampling process and of the function g, which is a function of the parameters Θ, implies the existence of a spectral density, $S_g(\omega)$, for this time series. If this assumption is held then the estimator of $E(g(\Theta))$ based on n iterations is given by

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

and the asymptotic variance is $S_g(0)/n$. Also let $\bar{g}(\Theta)_n^A$ be an estimator based on the first n_A iterations and $\bar{g}(\Theta)_n^B$ based on the last n_B. The difference between $\bar{g}_n^A$ and $\bar{g}_n^B$ is calculated to diagnose convergence. But this method is sensitive to the specification of the spectral window and moreover it is quantitative rather than qualitative.

Roberts(1992) proved that under certain regularity conditions

$$\left\| f^{(t)} - f \right\| \rightarrow 0, \quad \text{as} \quad t \rightarrow \infty,$$

where $\|\cdot\|$ is the norm associated with the specified inner product, $f^{(t)}$ is the density of the values generated at the t-th iteration of the Gibbs sampler and f is the true target joint density. An unbiased estimator of $\|f^{(t)} - f\| + 1$ is used to diagnose convergence. But in a subsequent paper, Roberts(1994), instead of evaluating above quantity, defined the within-chain dependence, D_n, and the between-chain interaction, I_n, and showed that $E(D_n) = E(I_n)$ at convergence. Advantages of this method are its rigorous mathematical foundation and the fact that it assesses convergence of the entire joint distribution rather than of univariate quantities. But it requires more complicated coding than the standard Gibbs sampler algorithm and the special coding is problem-specific rather than generic.

Also Ritter and Tanner(1992) proposed the stopping rule assigning the weight ω to the vector $(U_1, U_2, \cdots, U_k)$ that has been drawn from the current approximation to the joint distribution g_i via

$$\omega = \frac{q(U_1, U_2, \cdots, U_k)}{g_i(U_1, U_2, \cdots, U_k)} \quad ,$$

where $q(U_1, U_2, \cdots, U_k)$ is proportional to true joint distribution $p(U_1, U_2, \cdots, U_k)$. By carrying along these weights, one realizes a sample from $p(U_1, U_2, \cdots, U_k)$ rather than from the approximation. Moreover, as g_i converges toward the true joint density $p(U_1, U_2, \cdots, U_k)$ the distribution of the weights $p(\omega)$ converges toward a spike distribution. That is, the distribution is degenerated about a constant. Ritter and Tanner(1992) used the feature to assess convergence of the Gibbs sampler.

This procedure is also subjective because of monitoring the distribution of the weights $p(\omega)$, which must also be plotted each time to decide whether or not to continue the procedure. Also this method is problem-specific and the computation of weights may be time-intensive, particularly when full conditional distributions are not standard distributions and thus the normalizing constants must be estimated.

Others, Zellner and Min(1994), Liu, Liu and Rubin(1992), Garren and Smith(1993) and Johnson(1994), among others, proposed Markov chain Monte Carlo(MCMC) convergence diagnostics. Although the approach proposed by Heidelberger and Welch(1983) antedates the Gibbs sampler and is 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 algorithm.

2. The Proposed Stopping Rules

Gelfand *et al.*(1990) mentioned that appropriate values for t and m depend upon a particular application and cannot be specified in advance. All examples discussed in their paper were handled with $t \leq 50$ and $m \leq 100$. Gelfand *et al.*(1990) stopped the procedure when one feels that the marginal posterior for any parameter is converged enough. For a fixed m, they increased the values of t and plotted the two marginal posterior densities obtained from samples generated from the t-th and the $t+1$-th iteration. They suggested that the procedure be stopped if two densities are visually indistinguishable. In this situation they also increased the values of m to guarantee stability of the density estimate. This procedure is called *felt-tip pen test* by Casella and George(1992). Thus the stopping procedure proposed by Gelfand *et al.*(1990) is subjective. Also the marginal posterior density obtained from the sample must be plotted to decide continuation of the algorithm. That is, it takes time to check the convergence because one must plot the marginal posterior density at each iteration.

In this section we propose a convergence criterion as follows. For a fixed m, let g_t and g_{t+1} denote the estimates of the marginal distributions based on random samples generated from the t-th iteration and the $t+1$-th iteration, respectively. Then the following fact says that g_{t+1} is almost the same as g_t as $t \to \infty$.

Let g_t denote the estimate of the marginal distribution based on random samples generated from the t-th iteration. Then since g_t and g_{t+1} are the estimated distributions with the Gibbs sequences after the t-th and the $t+1$-th iteration, respectively, g_t and g_{t+1} converge to the true distribution g as $t \to \infty$. That is, let S denote the support of g_t's, then for any ε, $0 < \varepsilon < 1$,

$$\begin{aligned} \|g_t - g\| &= \sum_{x \in S} |g_t - g| \\ &< \frac{\varepsilon}{2}, \quad \text{as } t \to \infty, \end{aligned}$$

where $\|f\| = \sum_{x \in S} |f(x)|$ on a nonempty finite set S, by Geman and Geman(1984). Thus

$$\begin{aligned} \|g_t - g_{t+1}\| &= \sum_{x \in S} |g_t - g_{t+1}| \\ &= \sum_{x \in S} |g_t - g + g - g_{t+1}| \\ &\leq \sum_{x \in S} |g_t - g| + \sum_{x \in S} |g_{t+1} - g| \end{aligned}$$

$$< \varepsilon , \quad \text{a. s.}$$

Thus $\|g_t - g_{t+1}\|$ converges to 0 as $t \to \infty$.

Stopping Procedure I

So we will iterate the procedure until there is no difference between g_t and g_{t+1}. The difference of g_{t+1} from g_t can be formulated as follows:

$$D = \int_S (g_{t+1} - g_t)dx, \tag{2.1}$$

If the value of D is very small, then one can say that g_{t+1} is almost the same as g_t. So first we will compute D in Equation (2.1) using a numerical method. To obtain more satisfactory convergence, *i.e.*, to do enough iterations, D will be overestimated as in the following figure.

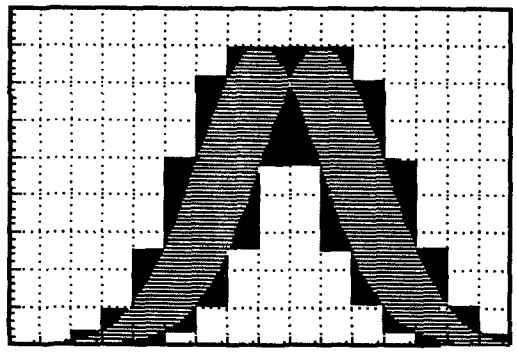

Figure 2.1 The Difference between g_{t+1} and g_t

Here the support S is divided into n subintervals and the area of the difference between two graphs is evaluated by using the rectangular rule on each pair of consecutive subintervals. But we use the vertical length which overestimates the area of the difference. In the figure the true value of the difference between g_{t+1} and g_t is the lined area and the dark area is an overestimated amount. The sum of the lined and dark areas is evaluated as follows:

$$D_x = \sum_{i=1}^{n-1} \frac{1}{n} (MX(i) - MN(i))$$

where

$$MX(i) = \max(g_t(x_i), g_{t+1}(x_i), g_t(x_{i+1}), g_{t+1}(x_{i+1}))$$

and

$$MN(i) = \min(g_t(x_i), g_{t+1}(x_i), g_t(x_{i+1}), g_{t+1}(x_{i+1})).$$

The value of D_x is less than 2 because g_{t+1} and g_t are probability density functions. Also if there is no difference between two graphs then the value of D_x would be very small. So the Gibbs sampler is stopped whenever $D_x < \varepsilon$ with a pre-specified ε, $0 < \varepsilon < 1$.

By this stopping procedure one needs not to plot the estimate of a distribution at each iteration. Also the Gibbs sampler can be automatically stopped in a computer program.

To assess the stability of the density estimate we increase the number of random samples m at the time of stopping. However, the cost of generating random variates from the full conditional distributions is not a serious problem due to high-speed computers. Thus the problem of selecting samples is less serious.

Stopping Procedure II

Now we propose another stopping rule. The procedure of Gibbs sampling is as follows. Given different initial values $U_{1j}^{(0)}, U_{2j}^{(0)}, \cdots, U_{kj}^{(0)}$, $j = 1,2,\cdots,m$, one draws

$$\begin{aligned} U_{1j}^{(1)} &\text{ from } p(U_1 | U_{2j}^{(0)}, U_{3j}^{(0)}, \cdots, U_{kj}^{(0)}) \\ U_{2j}^{(1)} &\text{ from } p(U_2 | U_{1j}^{(1)}, U_{3j}^{(0)}, \cdots, U_{kj}^{(0)}) \\ &\vdots \\ U_{kj}^{(1)} &\text{ from } p(U_k | U_{1j}^{(1)}, U_{2j}^{(1)}, \cdots, U_{k-1\,j}^{(1)}), \end{aligned}$$

for $j = 1,2,\cdots,m$, to compute one iteration of the scheme. Then after such $t+1$ iterations, the Gibbs sequences are obtained as

$$\begin{aligned} &(U_{1j}^{(0)}, U_{2j}^{(0)}, \cdots, U_{kj}^{(0)}) \\ &(U_{1j}^{(1)}, U_{2j}^{(1)}, \cdots, U_{kj}^{(1)}) \\ &\qquad\vdots \\ &(U_{1j}^{(t)}, U_{2j}^{(t)}, \cdots, U_{kj}^{(t)}) \\ &(U_{1j}^{(t+1)}, U_{2j}^{(t+1)}, \cdots, U_{kj}^{(t+1)}) \end{aligned}$$

For any s, $s = 1,2,\cdots,k$, $U_{sj}^{(t)}$, $j = 1,2,\cdots,m$, are regarded as simulated samples from the true distribution, for a sufficiently large t.

Now it is assumed that the random variables $U_{sj}^{(t)}$, $j = 1,2,\cdots,m$, are generated from the distribution $G_t(U_s)$ and $U_{sj}^{(t+1)}$, $j = 1,2,\cdots,m$, are generated from $G_{t+1}(U_s)$. Note that G_t and G_{t+1} converge to the true distribution G as $t \to \infty$. If

convergence of the Gibbs sampler is achieved then the distribution G_{t+1} may be almost equal to G_t. Thus the Gibbs iteration is stopped where G_{t+1} can be regarded as the same as G_t. So one wants to test the null hypothesis $H_0 : G_t = G_{t+1}$ *v.s.* $H_1 : G_t \neq G_{t+1}$. That is, one wants to test the hypothesis that a Gibbs sequences $U_{sj}^{(t)}$, $j = 1,2,\cdots,m$, have the distribution function G_{t+1}.

Now consider any interval $E_1, E_2, \cdots, E_M$ which are disjoint and $\bigcup_{i=1}^{M} E_i$ is the same a support of a quantity U_s. Let

$$p_l = P\{ U_{sj}^{t+1} \text{ falls in } E_l \}$$
$$= \{ \text{ number of } U_{sj}^{(t+1)} \text{ falling in } E_l \}/\ m, \quad l = 1,2,\cdots,M.$$

and N_l be the number of $U_{sj}^{(t)}$ falling in E_l, $l = 1,2,\cdots,M$. When the null hypothesis is true, N_l follows a binomial distribution with parameters m and success probability p_l because the number of $U_{sj}^{(t+1)}$ falling in E_l is approximately the same as that of $U_{sj}^{(t)}$ if $G_t = G_{t+1}$. Hence the difference $N_l - m \cdot p_l$ between the observed and the expected cell frequencies expresses the lack of fit of the data to G_{t+1}. So we adopted the function of these differences as a measure of fit. In addition the quantities $N_l - m \cdot p_l$, $l = 1,2,\cdots,M$, have approximately a normal distribution in large samples and therefore the following statistic

$$\chi^2 = \sum_{l=1}^{M} \frac{(N_l - m \cdot p_l)^2}{m \cdot p_l}$$

has approximately the χ^2_{M-1} distribution under the hypothesis in large samples. Note that in this procedure sufficiently large numbers of random samples $U_{sj}^{(t)}$ can be generated. Thus the above statistic can be used for testing $G_t = G_{t+1}$.

Note that g_{t+1} or G_{t+1} in the proposed stopping rules can be replaced by g_{t+l} or G_{t+l} for any integer l, $l \geq 2$. The comparison of g_t (or G_t) with g_{t+l} (or G_{t+l}) can prevent awkward stop when g_t converges to the true distribution very slow. Also the problem of checking the convergence is more efficient.

3. Simulation Study

Now we want to compare the performances of the proposed stopping procedures. So we apply the Gibbs sampler in the context of hierarchical model with proposed two stopping procedures.

Let X follow a normal distribution with unknown mean θ and unknown σ^2. For the prior distribution of θ and σ^2, the conjugate prior is considered, which is

For the prior distribution of θ and σ^2, the conjugate prior is considered, which is given by

$$\pi(\theta,\sigma^2) = N(\mu,\tau\sigma^2)IG(\alpha,\beta),$$

where $N(\cdot,\cdot)$ and $IG(\cdot,\cdot)$ represent a normal and an inverse gamma distributions, respectively. Here the hyperparameters τ, α and β are assumed to be known and in simulation study we will take $\tau = 1$, $\alpha = 0$ and $\beta = 1$.

To obtain normal random numbers we take $\theta = 0$, $\sigma^2 = 1$ and $\mu = 0$ and then generate 100 random variates by using subroutines in the IMSL(International Mathematical and Statistical Libraries). In this situation the marginal posterior distribution of θ, $\pi(\theta|\vec{x})$ can be evaluated directly as follows:

$$p(\theta|\vec{x}) = T\left(2\alpha+n,\ \mu(\vec{x}),\ \frac{\beta'^{-1}}{(\frac{1}{n}+n)(\alpha+\frac{n}{2})}\right),$$

where

$$\mu(\vec{x}) = \frac{\mu+n\tau\bar{x}}{n\tau+1}$$

and

$$\beta'^{-1} = \left[\frac{1}{\beta}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{n(\bar{x}-\mu)^2}{2(1+n\tau)}\right]^{-1}$$

and $T(a,b,c)$ represents a T distribution with a degrees of freedom, b location parameter and c scale parameter.

To implement the Gibbs sampler the full conditional distributions are derived as follows:

$$p(\theta|\vec{x},\sigma^2,\mu) = N\left(\frac{n\tau\bar{x}+\mu}{n\tau+1},\ \frac{\tau\sigma^2}{n\tau+1}\right)$$

$$p(\theta|\vec{x},\sigma^2,\mu) = N\left(\theta,\ \tau\sigma^2\right)$$

$$p(\sigma^2|\vec{x},\theta,\mu) = IG\left(\frac{n+2\alpha+1}{2},\ \left[\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2}+\frac{(\theta-\mu)^2}{2\tau}+\frac{1}{\beta}\right]^{-1}\right)$$

From the above full conditional distributions, one can obtain the random variates for any initial values, $\sigma_j^{2(0)}$, $\mu_j^{(0)}$ and $\theta_j^{(0)}$, $j = 1,2,\cdots m$. Here we choose 1 for $\sigma^{2(0)}$ and random variates generated from a standard normal distribution for

$\mu^{(0)}$ and $\theta^{(0)}$. Also for the number of random variates we take $m = 500$.

From the Gibbs sequences obtained after t iterations, one can estimate the marginal posterior distribution as follows:

$$\hat{p}(\theta \mid \vec{x}) = \frac{1}{m}\sum_{j=1}^{m} p(\theta \mid \vec{x}, \sigma_j^{2(t)}, \mu_j^{(t)}).$$

Figure 2.2 presents the true marginal posterior distribution of θ, $p(\theta \mid \vec{x})$ and the estimated posterior distribution with Gibbs sampler which is stopped by the stopping rule I. In Figure 2.2, one can see that the estimated marginal posterior distributions obtained after 4 iterations converge to the true distribution very well and also the proposed stopping criterion works well. Figure 2.3 presents the estimated marginal posterior distributions stopped by the stopping rule II. In Figure 2.3, one can see that the estimated distributions obtained after 3 iterations are very close to the true distribution and the proposed stopping rule also works well. In Figure 2.4, one can see that the two proposed stopping procedures are stopped at similar points.

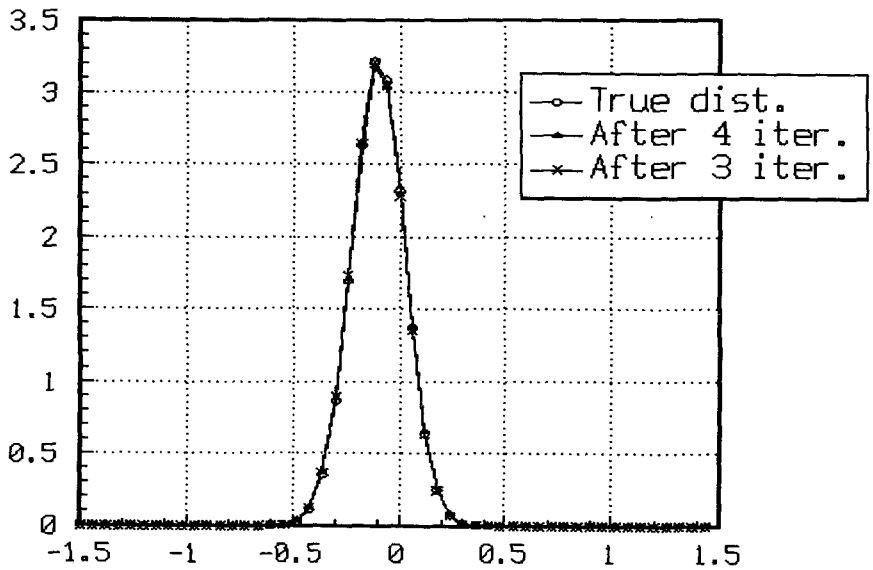

Figure 2.2 The Marginal Posterior Distribution of θ Using the Stopping Rule I

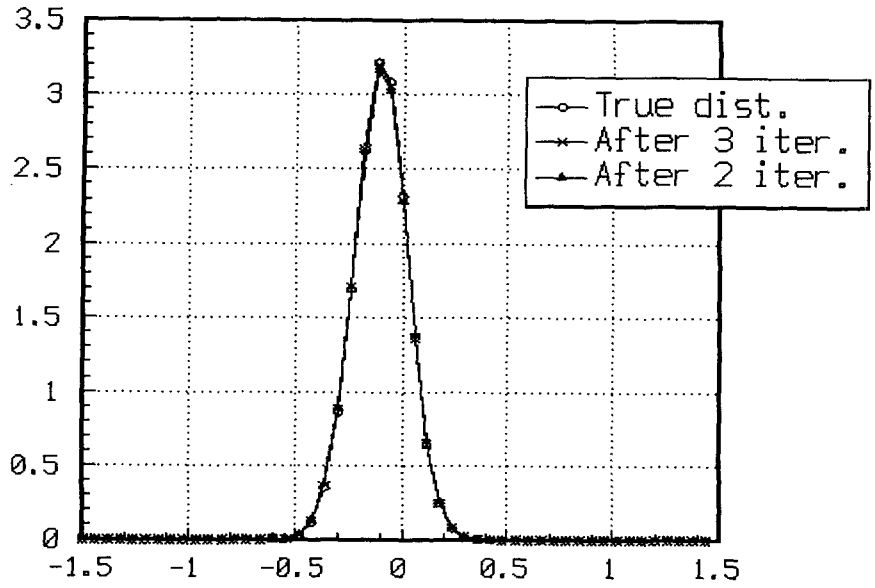

Figure 2.3 The Marginal Posterior Distribution of θ Using the Stopping Rule II

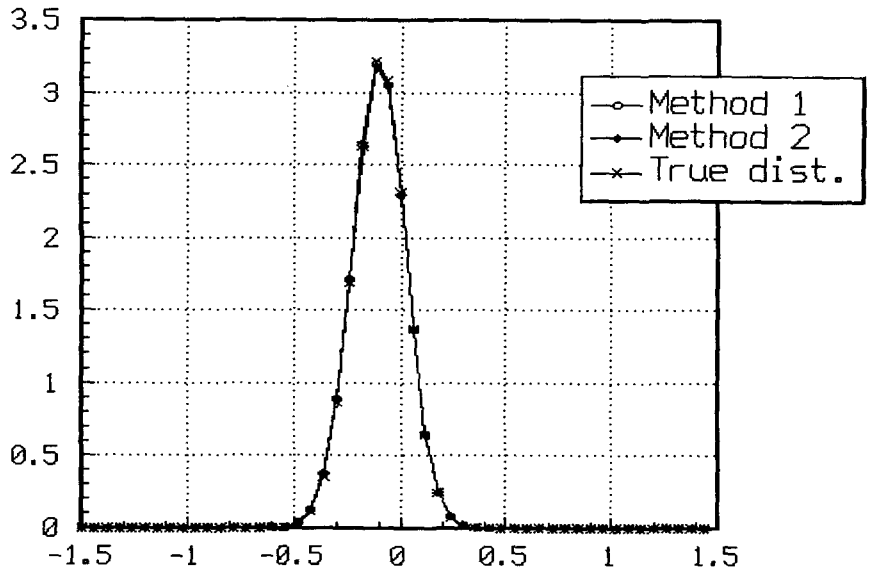

Figure 2.4 The Marginal Posterior Distribution of θ
Using the Stopping Rule I and II

References

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

Cowles, M. K. and Carlin, B. P. (1994), *Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review*, Technical Reports, Division of Biostatistics, University of Minnesota.

Garren, S. T. and Smith, R. L. (1993), *Convergence Diagnostics for Markov Chain Samplers*, Technical Report, Department of Statistics, University of North Carolina.

Gelfand, A. E. and Smith, A. F. M. (1990), *Sampling-Based Approaches to Calculating Marginal Densities*, Journal of the American Statistical Association, 85, 398-409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990), *Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling*, Journal of the American Statistical Association, 85, 972-985.

Gelman, A. and Rubin, D. B. (1992), *Inference from Iterative Simulation Using Multiple Sequences*, Statistical Science, 7, 457-511.

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

Geweke, J. (1992), *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*, Bayesian Statistics 4, Oxford: Oxford University Press, 169-193.

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

Johnson, V. E. (1994), *Testing for Convergence of Markov Chain Monte Carlo Algorithms Using Parallel Sampling Paths*, Technical Report, Institute for Statistics and Decision Sciences, Duke University.

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

Raftery, A. E. and Lewis, S. (1992), *How Many Iterations in the Gibbs Sampler?*, Bayesian Statistics 4, Oxford: Oxford University Press, 763-773.

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*, Bayesian Statistics 4, Oxford: Oxford University Press, 775-782.

Zellner, A. and Min, C-k. (1994), *Gibbs Sampler Convergence Criteria*, Submitted to Journal of the American Statistical Association.