How Many Iterations in the Gibbs Sampler?

Adrian E. Raftery
University of Washington *

Steven Lewis
University of Washington

April 10, 1991; revised September 13, 1991

Abstract

When the Gibbs sampler is used to estimate posterior distributions (Gelfand and Smith, 1990), the question of how many iterations are required is central to its implementation. When interest focuses on quantiles of functionals of the posterior distribution, we describe an easily-implemented method for determining the total number of iterations required, and also the number of initial iterations that should be discarded to allow for "burn-in". The method uses only the Gibbs iterates themselves, and does not, for example, require external specification of characteristics of the posterior density.

Here the method is described for the situation where one long run is generated, but it can also be easily applied if there are several runs from different starting points. It also applies more generally to Markov chain Monte Carlo schemes other than the Gibbs sampler. It can also be used when several quantiles are to be estimated, when the quantities of interest are probabilities rather than full posterior distributions, and when the draws from the posterior distribution are required to be approximately independent.

The method is applied to several different posterior distributions. These include a multivariate normal posterior distribution with independent parameters, a bimodal distribution, a "cigar-shaped" multivariate normal distribution in ten dimensions, and a highly complex 190-dimensional posterior distribution arising in spatial statistics. In each case the method appears to give satisfactory results.

The results suggest that reasonable accuracy may often be achieved with 5,000 iterations or less; this can frequently be reduced to less than 1,000 if the posterior tails are known to be light. However, there are frequent "exceptions" when the required number of iterations is much higher. One important such exception is when there are high posterior correlations between the parameters; even crude correlation-removing

*This research was supported by ONR contract N-00014-88-K-0265 and by NIH grant no. 5R01HD26330-02. The authors are grateful to Jeremy York for providing the data for Examples 4 and 5, for helping with the analysis and for useful discussions and suggestions, and to Julian Besag and an anonymous referee for helpful comments. A Fortran program called "Gibbsit" that implements the methods described here may be obtained from StatLib by sending an e-mail message to *statlib@temper.stat.cmu.edu* containing the single line "send gibbsit from general". While the program is not maintained, questions about it may be addressed by e-mail to Adrian Raftery at *raftery@stat.washington.edu*.

reparameterizations can greatly increase efficiency in such cases. Another important exception arises in hierarchical models when the Gibbs sampler tends to get "stuck"; there it seems that the use of different Markov chain Monte Carlo schemes may improve matters. The method proposed here seems to diagnose such "exceptions" quite effectively.

1 Introduction

The Gibbs sampler was introduced by Geman and Geman (1984) as a way of simulating from high-dimensional complex distributions arising in image restoration. The method consists of iteratively simulating from the conditional distribution of one component of the random vector to be simulated given the current values of the other components. Each complete cycle through the components of the vector constitutes one step in a Markov chain whose stationary distribution is, under suitable conditions, the distribution to be simulated. Gelfand and Smith (1990) pointed out that the algorithm may also be used to simulate from posterior distributions, and hence may be used to solve standard statistical problems.

The Gibbs sampler can be extremely computationally demanding, even for relatively small-scale statistical problems, and hence it is important to know how many iterations are required to achieve the desired level of accuracy. Here we describe and investigate a simple method for doing this, first briefly mentioned in Raftery and Banfield (1991).

We focus on the situation where there is a single long run of the Gibbs sampler, as practiced by Geman and Geman (1984) and Besag, York and Mollié (1991), for example. Gelfand and Smith (1990) have instead adopted the following alogithm: (i) choose a starting point; (ii) run the Gibbs sampler for T iterations and store only the last iterate; (iii) return to (i). The choice between the two ways of implementing the algorithm has not been settled, and was the subject of considerable debate and controversy at the recent Workshop on Bayesian Computation via Stochastic Simulation in Columbus, Ohio in February, 1991.

Intuitive considerations suggest that one long run may well be more *efficient*. A heuristic argument for this might run as follows. Consider the following two ways of obtaining S values simulated from the posterior distribution. The first way consists of picking off every Tth value in a single long run of length $N = ST$. The second way is that of Gelfand and Smith (1990). In the first way, the starting point for every subsequence of length T is closer to a draw from the stationary distribution than the corresponding starting point in the second way, which is chosen by the user. Thus, the first way gives a result which is, at least, no worse than the second way. Sometimes, although not always, this may be exploited in the

first way by reducing the value of T, to obtain the same result with less total iterations. A more formal argument along similar lines was presented by R.L. Smith in the concluding discussion at the Workshop on Bayesian Computation via Stochastic Simulation.

Gelman and Rubin (1991), on the other hand, have argued that, even if the one long run approach may be more efficient, it is still important to use several different starting points. The essence of their argument is that we cannot know, in the case of any individual problem, whether a single run has converged, and that combining the results of runs from several starting points gives an honest, if conservative, assessment of the underlying uncertainty. They illustrate their argument by showing that in the Ising model convergence can be quite slow. This example refers to the 10,000-dimensional binary state-space $\{-1, 1\}^{10,000}$, and is thus untypical of the parameter spaces that arise in typical statistical problems, but it should nevertheless be taken seriously. Here we suggest that combining internal information from a partial run with properties of Markov chains may provide an alternative way of solving the problem, without sacrificing the appealing simplicity of using a single long run. In particular, Markov chain theory provides results not just about ergodicity, but also about the (geometric) rate of convergence to the stationary distribution, and the distribution of sample means. However, the method can easily be used when there are several runs from different starting points.

2 The Method

2.1 How Many Iterations to Estimate a Posterior Quantile?

We consider the specific problem of calculating particular quantiles of the posterior distribution of a function U of the parameter θ. We formulate the problem as follows. Suppose that we want to estimate $P[U \leq u \mid y]$ to within $\pm r$ with probability s, where U is a function of θ. We will find the approximate number of iterations required to do this when the correct answer is q. For example, if $q = .025$, $r = .005$ and $s = .95$, this corresponds to requiring that the cumulative distribution function of the .025 quantile be estimated to within $\pm .005$ with probability .95. This might be a reasonable requirement if, roughly speaking, we wanted reported 95% intervals to have actual posterior probability between .94 and .96. We run the Gibbs sampler for an initial M iterations that we discard, and then for a further N iterations of which we store every kth. Typical choices in the literature are $M = 1,000$, $N = 10,000$ and $k = 10$ or 20 (Besag, York and Mollié 1991). Our problem is to determine M, N, and k. Note that when $k > 1$, we may still store and use all the N iterates, and the solution given

here is then conservative.

We first calculate U_t for each iteration t, and then form $Z_t = \delta(U_t \leq u)$, where $\delta(\cdot)$ is the indicator function. $\{Z_t\}$ is a binary 0-1 process that is derived from a Markov chain by marginalization and truncation, but it is not itself a Markov chain. Nevertheless, it seems reasonable to suppose that the dependence in $\{Z_t\}$ falls off fairly rapidly with lag, and hence that if we form the new process $\{Z_t^{(k)}\}$, where $Z_t^{(k)} = Z_{1+(t-1)k}$, then $\{Z_t^{(k)}\}$ will be approximately a Markov chain for k sufficiently large.

No formal proof of this is presented here, but it does seem intuitively plausible. Here a data-based method, described below, is used to assess whether the assumption provides a reasonable approximation for the case at hand. A proof might go something as follows. The process $\{Z_t\}$ is ergodic and, if the underlying Markov chain is ϕ-mixing in the sense of Billingsley (1968), which will often be a direct consequence of the construction, then $\{Z_t\}$ is also ϕ-mixing with the same rate. Thus the maximum difference between $P[Z_t^{(k)} = i_0 \mid Z_{t-1}^{(k)} = i_1, Z_{t-2}^{(k)}]$ and $P[Z_t^{(k)} = i_0 \mid Z_{t-1}^{(k)} = i_1]$ eventually declines geometrically as a function of k, and so $\{Z_t^{(k)}\}$ is arbitrarily close to being a first-order Markov chain in that sense, for k sufficiently large.

In what follows, we draw on standard results for two-state Markov chains; see, for example, Cox and Miller (1965). Assuming that $\{Z_t^{(k)}\}$ is indeed a Markov chain, we now determine $M = mk$, the number of "burn-in" iterations, to be discarded. Let

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

be the transition matrix for $\{Z_t^{(k)}\}$. The equilibrium distribution is then $\pi = (\pi_0, \pi_1) = (\alpha+\beta)^{-1}(\beta, \alpha)$, and the ℓ-step transition matrix is

$$P^\ell = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{\lambda^\ell}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix},$$

where $\lambda = (1-\alpha-\beta)$. Suppose that we require that $P[Z_m^{(k)} = i \mid Z_0^{(k)} = j]$ be within ε of π_i for $i, j = 0, 1$. If $e_0 = (1, 0)$ and $e_1 = (0, 1)$, then $P[Z_m^{(k)} = i \mid Z_0^{(k)} = j] = e_i P^m$, and so the requirement becomes

$$\lambda^m \leq \frac{\varepsilon(\alpha+\beta)}{\max(\alpha, \beta)},$$

which holds when

$$m = m^\star = \frac{\log\left(\frac{\varepsilon(\alpha+\beta)}{\max(\alpha,\beta)}\right)}{\log \lambda}.$$

Thus $M = m^\star k$.

To determine N, we note that the estimate of $P[U \leq u \mid D]$ is $\overline{Z}_n^{(k)} = \frac{1}{n}\sum_{t=1}^{n} Z_t^{(k)}$. For n large, $\overline{Z}_n^{(k)}$ is approximately normally distributed with mean q and variance $\frac{1}{n}\frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}$. Thus the requirement that $P[q - r \leq \overline{Z}_n^{(k)} \leq q + r] = s$ will be satisfied if

$$n = n^\star = \frac{\frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}}{\left\{\frac{r}{\Phi(\frac{1}{2}(1+s))}\right\}^2},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus we have $N = kn^\star$.

To determine k, we form the series $\{Z_t^{(k)}\}$ for $k = 1, 2, \ldots$. For each k, we compare the first-order Markov chain model with the second-order Markov chain model, and choose the smallest value of k for which the first-order model is preferred. We compare the models by first recasting them as (closed-form) log-linear models for a 2^3 table (Bishop, Fienberg and Holland, 1975), and then using the BIC criterion, $G^2 - 2\log n$, where G^2 is the likelihood ratio test statistic. This was introduced by Schwarz (1978) in another context and generalized to log-linear models by Raftery (1986); it provides an approximation to twice the logarithm of the Bayes factor for the second-order model. One could also use a non-Bayesian test, but the choice of significance level is problemmatic in the presence of large samples of the size that arise routinely with the Gibbs sampler.

To implement the method, we run the sampler for an initial number of iterations, $N_{\min}$, and use this run to determine the number of additional runs required, as above. The procedure can be iterated, in that once the indicated number of iterations has been run, we may apply the method again to the entire run, reestimating α and β to determine more precisely if the number of iterations produced was in fact adequate. To determine $N_{\min}$, we note that the required N will be minimized if successive values of $\{Z_t\}$ are independent, in which case $M = 0$, $k = 1$ and

$$N = N_{\min} = \Phi^{-1}(\frac{1}{2}(1+s))^2 q(1-q)/r^2.$$

For example, when $q = .025$, $r = .005$ and $s = .95$, we have $N_{\min} = 3,748$.

We also note that the user is not *required* to use only every kth iterate; if all the iterates are used the method proposed here will be conservative in the sense of possibly overestimating the number of iterations required. On the other hand, in the majority of cases that we have examined, the preferred value of k was, in fact, 1. Also, storage considerations often point to the desirability of storing only a portion of the iterates if this is reasonable.

The user needs to give only the required precision, as specified by the four quantities q, r, s and ε. Of these, the result is by far the most sensitive to r, since $N \propto r^{-2}$. It may

Table 1: Maximum percent error in the estimated .025 quantile

r	$N_{\min}$	Percent error		
	(s=.95)	N(0,1)	t_4	Cauchy
.0025	14982	2	4	11
.005	3748	5	8	25
.0075	1665	8	13	43
.01	936	11	19	67
.0125	600	14	26	101
.015	416	19	37	150
.02	234	31	65	402

often be more natural to specify the required precision in terms of the error in the estimate of a quantile rather than the error in the cumulative distribution function at the quantile, which is what r refers to. In order to see how r relates to accuracy on the former scale, we have shown in Table 1 the approximate maximum percentage error in the estimated quantile corresponding to a range of values of r, for $q = .025$. This is defined as $100 \max\{\frac{F^{-1}(q \pm r)}{F^{-1}(q)} - 1\}$, and is shown for three distributions: normal (light-tailed), t_4 (moderate tails), and Cauchy (heavy-tailed).

Suppose we regard a 14% error as acceptable, corresponding to an estimated .975 quantile of up to 2.24 in the normal distribution, compared with the true value of 1.96. Then, if we knew $p(U \mid y)$ to have light, normal-like, tails, Table 1 suggests that $r = .0125$ would be sufficiently small. However, with the heavier-tailed t_4 distribution, $r = .0075$ is required to achieve the same accuracy, while for the very heavy-tailed Cauchy, $r = .003$ is required, corresponding to $N_{\min} \approx 10,000$.

This suggests that if we are not sure in advance how heavy the posterior tail is, $r = .005$ is a reasonably safe choice (even for the Cauchy it is not catastrophic). It also suggests that the present method could be refined by using the initial set of Gibbs iterates to estimate the asymptotic rate of decay of the posterior tail nonparametrically with methods such as those of Hall (1982), and then choosing r in light of the estimate, perhaps by referring to a t-distribution with the appropriate degrees of freedom. At first sight it might appear that a component-wise reparametrization to lighten the tails would be a good remedy. However, we suspect that this would not be a real solution, and that the problem would reappear when the results were transformed back to the scale on which the quantity of actual interest is measured.

2.2 Extensions

Several quantiles: If there are Q quantiles to be estimated ($Q > 1$), it is possible to run the Gibbs sampler for $N_{\min}$ iterations, apply the method in Section 2.1 to each quantile individually and then use the maximum values of M, k and N. This will guarantee that for each quantile marginally, at least the specified accuracy will be achieved, or, equivalently, that the expected number of estimated quantiles that lie within r of the true value is at least sQ. A different and more demanding accuracy requirement is that with probability s, all Q quantiles lie within r of the true value. A conservative solution to this uses the Bonferroni bound idea and consists of replacing s by s/Q and proceeding as before.

Estimating probabilities: The Gibbs sampler is often used to estimate probabilities rather than distributions. Examples include the probability that the ground truth at a pixel is of a particular "color" in image reconstruction (Besag, 1986), the probability that an individual is of a given genotype in genetic pedigree analysis (Sheehan, 1990), and the probability that an individual has a specific disease based on an expert system diagnosis (Lauritzen and Spiegelhalter, 1988). The present method is directly applicable to this case, and indeed takes a slightly simpler form. The $\{Z_t\}$ process is given and does not not have to be formed by truncation. One does not have to specify q, which is just the required probability. One simply specifies r and s and proceeds as before.

Independent iterates: When it is much more expensive to analyze a Gibbs iterate than to simulate it, as it may be in complex applications such as genetic pedigree analysis, it is desirable that the Gibbs iterates used be approximately independent. This can be achieved by making k big enough. To determine k, we can compare the independence model with the first-order Markov chain model for $\{Z_t^{(k)}\}$ and choose the smallest value of k for which the independence model is preferred. The models can be compared by recasting them as the independence and saturated models for a 2×2 table and using the BIC criterion. M is then calculated exactly as before, and we have $N = kN_{\min}$ by the approximate independence of the iterates.

3 Examples

We now apply the method to several examples, both simulated and real. In each case, we give the results only for $q = .025$, $r = .005$, $s = .95$ and $\varepsilon = .001$. The results are shown

Table 2: Results for the five examples

Example	M	k	N	$\hat{F}(F^{-1}(.025))$
1. Indep. normal pars.	3	1	3,914	.023
2. Bimodal	4	1	4,256	.028
3. Cigar	36	3	26,916	.024
4. Spatial u_1	3	1	4,052	.024
5. Spatial smoothness	40	2	24,346	-

in Table 2 for all the examples. The value in the column headed $\hat{F}(F^{-1}(.025))$ should be between .02 and .03 for this specification. Results for other quantiles and other accuracy requirements, not shown here, were qualitatively similar.

Example 1: Multivariate normal distribution with independent parameters

In this simulated example the method gave $k = 1$, a very small number of burn-in iterations ($M = 3$), and a value of N which is only slightly larger than the theoretical minimum (3,914 as against 3,748). Also, the result is within the specified bounds. While this is very much as one would expect, it is also a reassuring check on the performance of the method.

Example 2: A bimodal posterior distribution

Here we simulated, using the Gibbs sampler, from a mixture of two bivariate normal distributions, namely

$$\frac{1}{2}BVN(\mu_1, \Sigma) + \frac{1}{2}BVN(\mu_2, \Sigma),$$

where $\mu_1 = (-1, 1)^T$, $\mu_2 = (1, 0)^T$ and

$$\Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}.$$

The joint distribution is quite strongly bimodal, although the marginal distributions of the two components are not. The first 1,000 simulated values of the second component are shown in Figure 1. The result is surprisingly similar to that in Example 1. Again, $k = 1$, the amount of burn-in is negligible ($M = 4$), and $N = 4,256$ is not much larger than the theoretical minimum. The Gibbs iterates are slightly more highly correlated than in Example 1, and the value of N can be regarded as an index of this. Once again, the result is within the specified bounds.

Example 3: A cigar in ten dimensions

In order to investigate the effect of high posterior correlations between parameters, we used the Gibbs sampler to simulate from a 10-dimensional multivariate normal posterior distribution where each component had zero mean and unit variance, and all the pairwise correlations were equal to .9. This is a highly correlated distribution, where the first principal component (proportional to the mean of the parameters) accounts for 91% of the variance; the posterior distribution is concentrated about a thin "cigar" in 10-space. Note that this is a very poor parameterization for the Gibbs sampler.

The first 1,000 simulated values of the first parameter are shown in Figure 2. The results of applying the method are strikingly different from what we saw before. The amount of burn-in is no longer negligible, although it is not huge ($M = 36$). The dependency structure of the binary sequence is more complicated than before, leading to $k = 3$, and the level of dependency is high, so that the required N is very large, at 26,916. After that number of iterations, the result was accurate. This phenomenon seems to be due to the high level of dependency in the sequence, and not primarily to the sampler being slow to converge to the desired distribution.

It is of interest to consider the situation after 6,700 iterations; this is a large number, but substantially less than the prescribed 27,000. By that point, diagnostics based on changes in cumulative estimates suggest the Gibbs sampler to have converged. However, after 6,700 iterations, $1 - \hat{F}(F^{-1}(.975)) = .045$, compared to the true value of .025, which is well outside the prescribed tolerance, and the empirical .975 quantile was 2.22 instead of 1.96. However, the present method indicated clearly that the number of iterations was insufficient to achieve the desired accuracy.

This example also illustrates the importance of parameterization for the Gibbs sampler (see also Wakefield, 1991). A parameterization that leads to a highly correlated posterior distribution like the one considered in this example is a very poor one for the Gibbs sampler, and leads to considerable inefficiency. It seems likely that even a very simple linear reparameterization would lead to at least a five-fold reduction in the required number of iterations.

Example 4: An 190-dimensional posterior distribution from spatial statistics

Besag, York and Mollié (1991) considered the problem of mapping the risk from a disease given incidence data. Let x_i denote the unknown log relative risk in zone i and y_i the

corresponding observed number of cases. They assumed y_i to have a Poisson distribution with mean $c_i e^{x_i}$, where c_i is the expected number assuming constant risk. They let $x_i = u_i + v_i$ where the u_i have substantial spatial structure represented by the joint density

$$p(u \mid \kappa) \propto \frac{1}{\kappa^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(u_i - u_j)^2\right\},$$

where $i \sim j$ denotes the fact that zones i and j are contiguous and κ is a spatial smoothness parameter. The v_i are assumed to be generated by Gaussian white noise with parameter λ. The prior distribution was $p(\kappa, \lambda) \propto \exp\{-.005(\kappa^{-1} + \lambda^{-1})\}$. Although this is improper, the resulting posterior distribution is proper. The main aim is to find the posterior distribution of x_i, but other features of the underlying mechanism may also be of interest.

Here we show only the result for u_1 for thyroid cancer deaths in 94 departements of France; the results for the other u_i and for the v_i are similar. The Gibbs sampler here involves 190 parameters: the 94 u_i's, the 94 v_i's, κ and λ. The first 1,000 iterations are shown in Figure 3. The result is very similar to that for Examples 1 and 2. The number in the last column was obtained by running the Gibbs sampler for a total of 11,000 iterations, and treating the value obtained from this complete run as the "true" value.

Example 5: The spatial smoothness parameter

We now consider separately the spatial smoothness parameter κ from Example 4. The first 1,000 Gibbs iterations are shown in Figure 4. The results are quite different from those for u_1, and are somewhat similar to those for Example 3. The dependency structure in the induced binary sequence is complex, leading to $k = 2$, and the dependency is high, leading to $N = 24,346$. The amount of burn-in, however, while not negligible, is fairly small ($M = 40$). It was not feasible to determine the correct answer in this case.

While the difficulty with Example 3 could probably be resolved by appropriate reparameterization, the problem here seems more fundamental. Here the problem is due to the fact that κ sometimes gets "stuck" close to zero for several hundred iterations at a time. This is because having the u_i close together (i.e. high spatial smoothness) makes a small value of κ likely, while a small value of κ forces the u_i to be close together. Thus the Gibbs sampler gets caught periodically in a "vicious circle"; to escape it requires a rare event. The solution here may be the use of a different variation on Metropolis dynamics than the Gibbs sampler, perhaps involving simultaneous updating of some kind. This kind of problem seems likely to arise often in hierarchical models more generally. Note that the present method for determining the number of iterations would carry over to other forms of Metropolis dynamics.

4 Discussion

We have proposed a method for determining how many iterations are necessary in the Gibbs sampler. This is easy to implement and does not require anything beyond an initial run from the sampler itself. It appears to give encouraging results in several examples. However, much more thorough investigation is required for various kinds of difficult posterior distributions.

For "nice" posterior distributions, the examples suggest that accuracy at the level specified for illustration in this paper can be achieved by running the sampler for 5,000 iterations and using all the iterates. However, when the posterior is not "nice", the required number can be very much greater. Example 3 suggests that poor parameterization can be one reason for massive inefficiency of the Gibbs sampler, and that even simple-minded reparameterization may have the potential to lead to substantial savings. Problems may also arise in hierarchical models where the Gibbs sampler sometimes has a tendency to get "stuck"; this is illustrated in Example 5.

Our experience suggests that the present method diagnoses such problems fairly well. When the prescribed number of iterations is much larger than $N_{\min}$, there seem to be two ways to proceed. One is simply to run the sampler for the specified number of iterations; this seems the best course when iterates are computationally inexpensive. Otherwise it may well be worthwhile to reparameterize or to use a different Markov chain Monte Carlo scheme.

It has been common practice when running the Gibbs sampler to throw away a substantial number of initial iterations, often on the order of 1,000. Our results here suggest that this may not usually be necessary, and indeed, will often be quite wasteful. This is not too surprising given the geometric rate of convergence of Markov chains to the stationary distribution. When large numbers of iterations were required, this was due to the high level of dependence between successive iterates rather than to the failure of the Gibbs sampler to converge initially.

Thus, we suspect that, for typical statistical problems, the uncertainty due to the initial starting point that Gelman and Rubin (1991) capture with their methods will be a relatively small part of the overall uncertainty if the number of Gibbs iterations is realistically large. Of course, we are far from having established that conclusively here, and diagnostic checks such as those proposed by Gelman and Rubin (1991) remain important. Indeed, our method and theirs may be regarded as complementary in that our method can be viewed as determining the total number of iterations required, which will typically be little changed whether there is one long run or a small number of shorter runs from different starting points. More

specifically, if there are to be R different runs from different starting values, then each run should have $NR^{-1} + M$ iterations, of which the first M are discarded. Thus the two methods could be synthesized by using our approach to determine the total number of required iterations, and using the method of Gelman and Rubin (1991), both as a further check for convergence, and also to incorporate uncertainty about the starting point.

It has also been common practice to use only every 10th or 20th iterate and to discard the rest. The results here also suggest that in many cases this is rather prolifigate. Indeed, in the "nice" cases, the dependency between successive iterates is weak and it makes sense to use them all, even when storage is an issue.

An alternative approach to determining the number of iterations starts by viewing the sequence of Gibbs iterates as a standard time series (e.g. Geyer, 1991; Geweke, 1991; Hills and Smith, 1991). If the quantity of interest is the mean of a function of the series, then the variance of such a mean is equal to the spectrum of the corresponding series at zero, which can be estimated using standard spectral methods. This requires the user to specify both a spectral window and a window width, and the estimate of the spectrum at zero can be quite sensitive to these choices.

Obtaining posterior quantiles defining Bayesian confidence intervals is often a key goal of an analysis. When this is the case, the present method exploits the natural simplification that arises from the implied dichotomization. Thus it avoids the need to specify quantities other than the required precision (such as spectral window widths), it yields a simple estimate of the number of "burn-in" estimations, and it provides a practical lower bound, $N_{\min}$, on the number of iterations that is known before the Gibbs sampler starts running.

It may be argued that often all that is required is a posterior mean and standard deviation, and that these are not quantiles. If this is indeed the case, and there is really no interest in the shape of the posterior distribution, then there may well be little point in running the Gibbs sampler at all, as cheaper methods are frequently available for posterior means and standard deviations. However, the posterior mean and standard deviation are often used to provide a *summary* of the posterior distribution. In that case, a robust measure location, such as the median, may be preferable to the posterior mean as a descriptive measure, and the median *is* a quantile. Also, the posterior standard deviation is often used as a way of obtaining an approximate confidence interval, say by taking the posterior mean plus or minus two posterior standard deviations. However, if a sample from the posterior is available, it seems worth calculating the required interval directly—again this will be defined by quantiles. Even if a single measure of posterior dispersion is required, it may well be better to use a

more robust measure than the posterior standard deviation, such as a scaled version of the inter-quartile range; again this is defined by quantiles. Thus, appropriate summaries of the posterior distribution are often defined in terms of quantiles, even when at first sight it seems that a mean-like quantity is required.

One important message is that the required number of iterations can be dramatically different for different problems, and even for different quantities of interest within the same problem. Thus, it seems unwise to rely on a single "rule of thumb", and it would seem to be important to use some method, such as the one proposed here, to determine the number of iterations that are needed for the problem at hand.

References

Besag, J.E. (1986) On the statistical analysis of dirty pictures (with Discussion). *J. R. Statist. Soc., Ser. B,* **48**, 259–302.

Besag, J.E., York, J. and Mollié (1991) Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Ann. Inst. Statist. Math.*, **43**, 1–59.

Billingsley, P. (1968) *Convergence of Probability Measures.* New York: Wiley.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis.* Cambridge, Mass.: MIT Press.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes.* London: Chapman and Hall.

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

Gelman, A. and Rubin, D.B. (1991) An overview and approach to inference from iterative simulation. Paper presented at the Workshop on Bayesian Computation via Stochastic Simulation, Columbus, Ohio, February, 1991.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **6**, 721-741.

Geweke, J. (1991) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Paper presented to the Fourth Valencia

International Meeting on Bayesian Statistics, Valencia, Spain, April 1991.

Geyer, C. (1991) Monte Carlo maximum likelihood in exponential families. Paper presented at the Workshop on Bayesian Computation via Stochastic Simulation, Columbus, Ohio, February, 1991.

Hall, P. (1982) On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc., ser. B*, **44**, 37-42.

Hills, S.E. and Smith, A.F.M. (1991) Parametrization issues in Bayesian inference. Paper presented to the Fourth Valencia International Meeting on Bayesian Statistics, Valencia, Spain, April 1991.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with Discussion). *J. R. Statist. Soc., Ser. B.*, **50**, 157–224.

Raftery, A.E. (1986). A note on Bayes factors for log-linear contingency tables with vague prior information. *J. Roy. Statist. Soc., ser. B*, **48**, 249-250.

Raftery, A.E. and Banfield, J.D. (1991) Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 32–43.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Sheehan, N. (1990) Image processing procedures applied to the estimation of genotypes on pedigrees. Technical Report no. 176, Department of Statistics, University of Washington.

Wakefield, J. (1991) Parameterization issues in Gibbs sampling. Paper presented at the Workshop on Bayesian Computation via Stochastic Simulation, Columbus, Ohio, February, 1991.