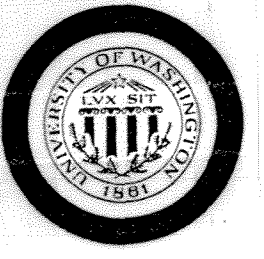
THREE SHORT PAPERS ON SAMPLING-BASED INFERENCE:
1. HOW MANY ITERATIONS IN THE GIBBS SAMPLER?
2. MODEL DETERMINATION
3. SPATIAL STATISTICS

by

Adrian E. Raftery
Steven Lewis
Jeffrey D. Banfield

TECHNICAL REPORT No. 212
May 1991

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195 USA



Three Short Papers on Sampling-based Inference: 1. How Many Iterations in the Gibbs Sampler? 2. Model Determination 3. Spatial Statistics

Adrian E. Raftery
Steven Lewis
Jeffrey D. Banfield

May, 1991

Abstract

This technical report consists of three short papers on Monte Carlo Markov chain inference. The first paper, "How many iterations in the Gibbs sampler?," proposes an easily implemented method for determining the total number of iterations required to estimate probabilities and quantiles of the posterior distribution, and also the number of initial iterations that should be discarded to allow for "burn-in".

The second paper discusses model determination via predictive distributions. The paper advocates the standard Bayesian procedure that uses Bayes factors, and points out that this can be implemented quite easily using sampling-based methods.

The third paper discusses issues in spatial statistics that use sampling-based methods. Several issues in the Bayesian image restoration approach are discussed: the modeling of spatial dependence, allowing for model uncertainty, the improper posterior distributions that arise in hierarchical Bayes modeling, and the modeling of local dependence between counts when it cannot be assumed that the observations are independent given the true rates.

How Many Iterations in the Gibbs Sampler?

Adrian E. Raftery
University of Washington *

Steven Lewis
University of Washington

April, 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.

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 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.

*This research was supported by ONR contract N-00014-88-K-0265 and by a grant from NIH. The authors are grateful to Jeremy York for providing the data for Examples 4 and 5 and for helping with the analysis. Code to implement the procedure described in this paper may be obtained from Adrian Raftery by e-mail at *raftery@stat.washington.edu*.

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

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 $P[Z_t^{(k)} = i_0 \mid Z_{t-1}^{(k)} = i_1, Z_{t-1}^{(k)}]$ and $P[Z_t^{(k)} = i_0 \mid Z_{t-1}^{(k)} = i_1]$ eventually declines exponentially 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 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).

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

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.

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 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

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	.025
4. Spatial u_1	3	1	4,052	.024
5. Spatial smoothness	40	2	24,346	-

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 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., York, J. and Mollié (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, to appear.

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.

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.*, to appear.

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

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

Figure 1 - Bimodal example

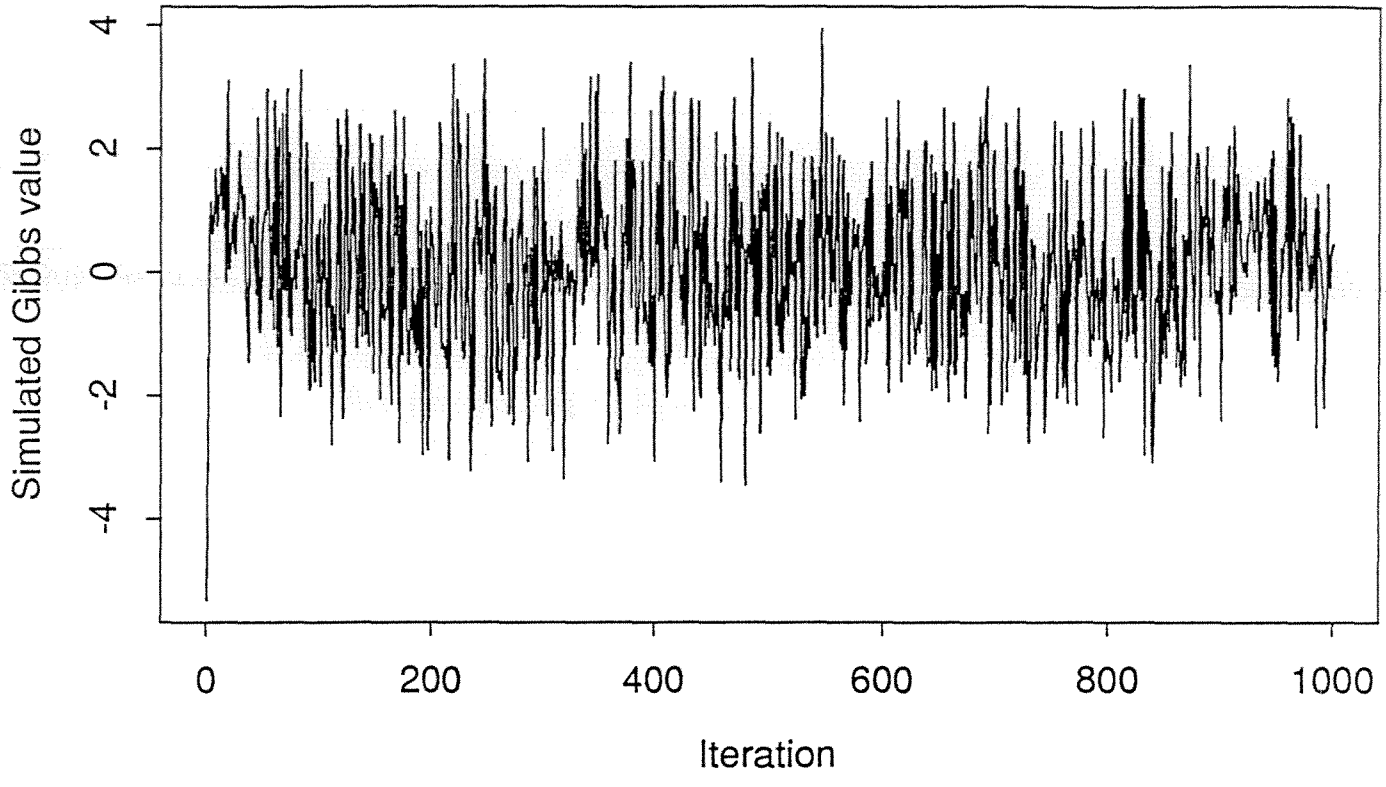


Figure 2 - Cigar in 10 dimensions example

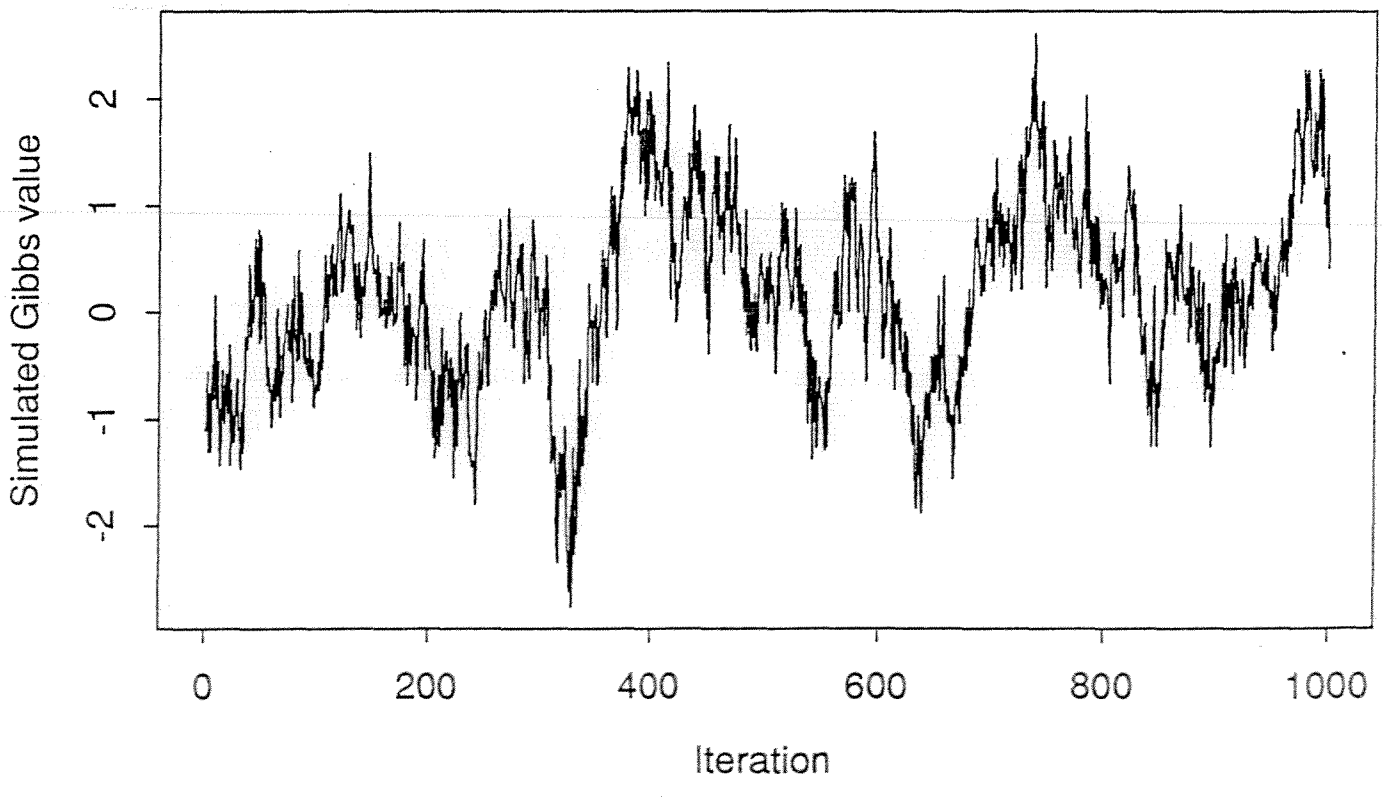


Figure 3 - Spatial example: u1

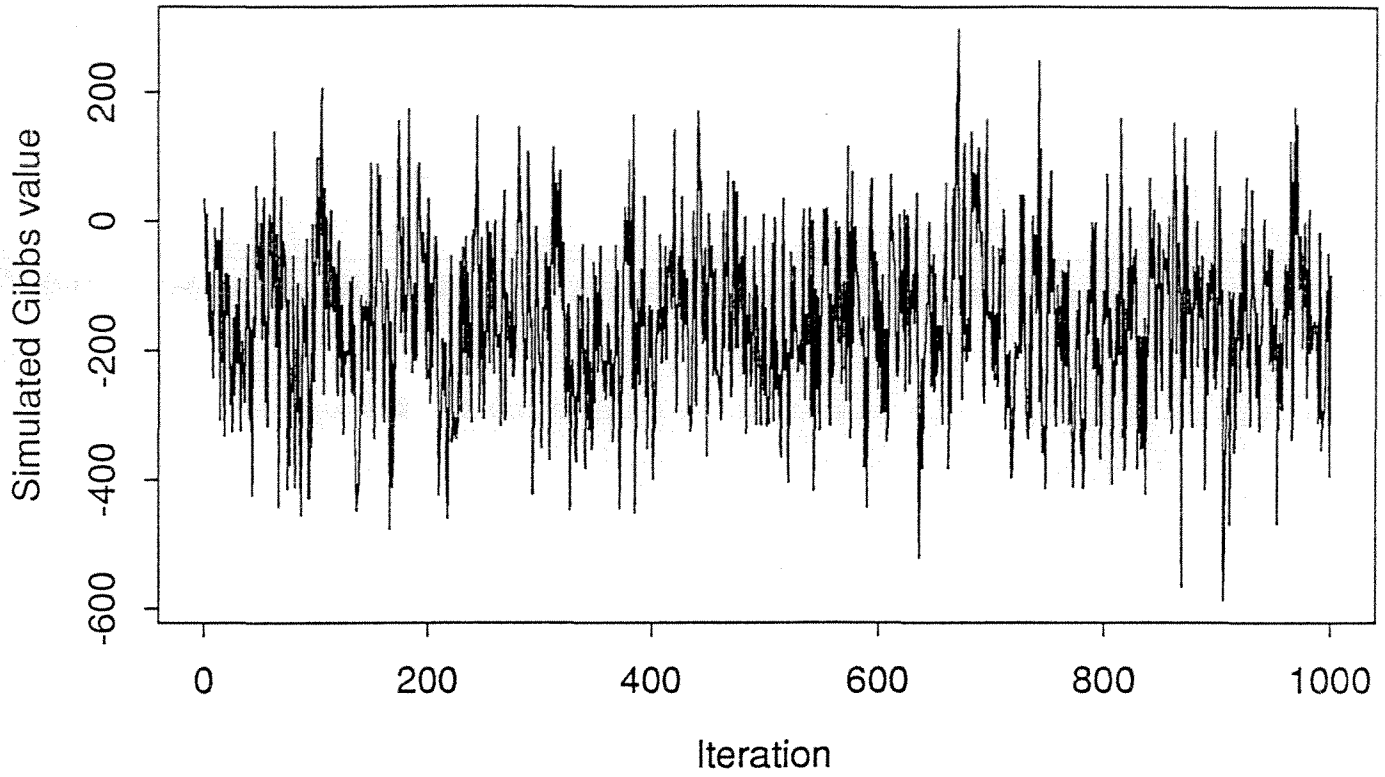
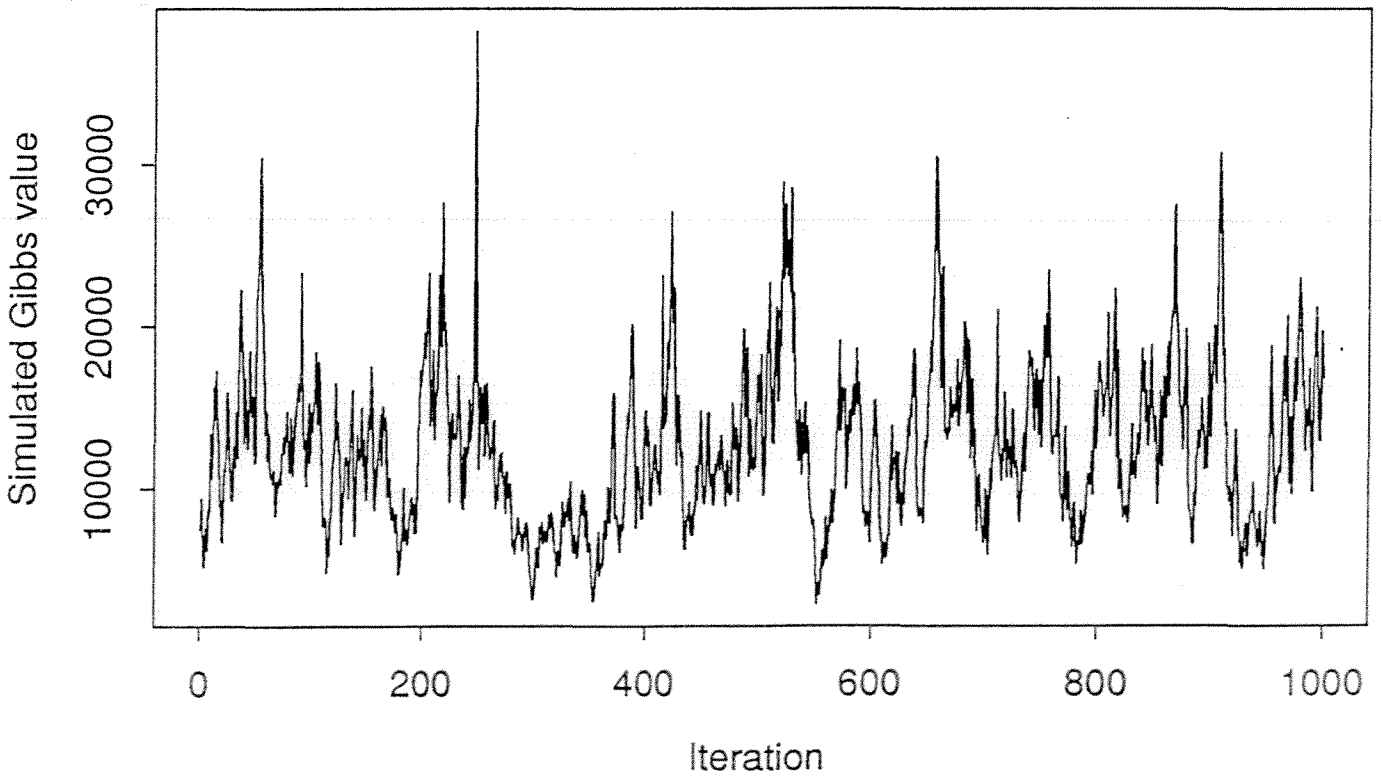


Figure 4 - Spatial example: kappa



Discussion of "Model determination using predictive distributions with implementations via sampling-based methods", by A.E.Gelfand, D.K.Dey and H.Chang

Adrian E. Raftery
University of Washington

May 28, 1991

1 Introduction and summary

It is a pleasure to congratulate the authors on an interesting and important paper that points out how sampling-based methods can make Bayesian diagnostics for model checking routinely available. Bayesian diagnostics are often similar to frequentist ones, but they have the great advantage of being systemmatically available through the predictive distribution, even for complex models. This is in contrast with frequentist diagnostics, which have to be developed from scratch for each new class of models, often requiring considerable ingenuity. The *interpretation* of Bayesian diagnostics is somewhat glossed over by the authors, however.

We part company to some extent on the issue of model choice. I am unconvinced by the arguments against the standard Bayesian procedure, namely that based on posterior model probabilities. New results indicate that posterior model probabilities *can* be readily computed using sampling-based methods. Also, the standard Bayesian procedure *is* based on predictive distributions, in a prequential rather than a cross-validation sense.

2 Bayesian diagnostics for model checking

A real achievement of this paper is to show how sampling-based methods can be used to obtain Bayesian diagnostics systemmatically and routinely for a very wide class of models. When frequentist diagnostics are available they are often similar to Bayesian diagnostics. The

great advantage of Bayesian diagnostics is that they are available quite generally from the predictive distribution, unlike their frequentist counterparts, which can require considerable ingenuity for each new class of models.

The authors have, however, rather glossed over the *interpretation* of their diagnostics. For example, in the nonlinear regression example, they conclude that points 11 and 14 are troublesome but that, all told, both models provide an adequate fit. What is the basis for this conclusion? Nothing is suggested beyond eyeballing the results, but there are certainly more precise criteria implicitly at work here, and they should be made explicit.

I would suggest that diagnostics not be used to reject the current model, but rather to guide the search for better models by indicating the direction of search, or the way in which the current model is inadequate. If this leads to the specification of an alternative model, then the current model can be compared with alternative one using the posterior odds ratio (or posterior expected utilities if these can be specified); the current model will not be rejected unless the alternative one is decisively preferred. You don't abandon a model unless you have a better one in hand.

Even viewing diagnostics this way, as an exploratory tool rather than as a basis for inference, we still need some yardstick to calibrate our inspection of the results. Here it does seem that frequentist calculations are useful, and I suspect that such calculations implicitly underly the authors' interpretation of the results in their Table 2.

3 Model comparison: In support of the standard Bayesian procedure

The standard Bayesian procedure is given by the authors' equation (3), and amounts to basing inference on the posterior model probabilities. They raise two objections to this procedure, which I will now briefly discuss.

3.1 "Bartlett's paradox"

This is the observation due to Bartlett (1957) that if under M_1 the Y_i are iid $N(0,1)$, and under M_2 they are iid $N(\theta,1)$ with $\theta \sim N(0,\tau^2)$, then $p(M_1 \mid Y) \to 1$ as $\tau^2 \to \infty$ regardless of the data; see the authors' section 2.3.1.

This has been presented by the authors and by others that they cite as a major flaw of the standard Bayesian approach, but I do not find it too disquieting. Letting $\tau^2 \to \infty$ implies that $E[|\theta|]$ also becomes arbitrarily large, so it is not too surprising that, for any

data set, $E[|\theta|]$ can be set large enough that the data prefer zero. Some prior information is almost always available that will limit the prior variance τ^2, and it is always important to investigate the sensitivity of $p(M_1 \mid Y)$ to changes in τ^2. In practice, $p(M_1 \mid Y)$ tends to be rather insensitive to changes in τ^2 over a wide range (see, e.g., Raftery, 1988). Thus, Bartlett's paradox seems to me to suggest that the use of highly diffuse priors is not a good idea for model comparison.

It may be objected that it is desirable to have a "reference" procedure for model comparison. However, in my applied experience, reasonable proper priors are often readily accepted, especially when backed up with a serious sensitivity analysis; the likelihood is often the more controversial part of the analysis.

3.2 The more serious criticism

The authors write:

> "A more serious criticism is that, in doing practical model fitting, we doubt that anyone including Bayesians would select models in this fashion [i.e. using the standard Bayesian procedure - AER.] One doesn't really believe that any of the proposed models are correct whence attaching a prior probability to an individual model's correctness seems silly. Moreover the selection process is typically evolutionary. Initially a few models (sometimes, in fact, a single one) are considered. These are examined and modified with comparisons often made in pairs until a satisfactory (in terms of both parsimony and performance) but one would certainly not say 'best' choice is made."

Attaching a prior probability to a model is not any sillier than science as traditionally practiced. Most of science is an attempt to find a model that predicts the observations to date well; it does not claim to have found the "truth" (if such a thing exists) or the "correct model". Science typically proceeds by adopting a *paradigm*, which means essentially *conditioning* on a collection of models, often with an explicit parametric form. Prior probabilities conditional on the adopted paradigm, or collection of models, do make sense.

Of course, if one does not so condition, the prior probability, and hence also the posterior probability of most models is zero. Since one does not believe the paradigm to be the "truth", this may make science as a whole seem silly, but its record of success argues in its favor. Note that the marginal likelihood, $f(Y \mid M_j)$, which is proportional to the posterior probability

of M_j, is just the (predictive) probability of the data given the model M_j, and so is precisely the right quantity for evaluating the scientific theory defined by M_j.

Consider, for example, the question of whether smoking causes lung cancer, and suppose that the currently accepted way of addressing this issue is within the framework of the logistic regression model, $\text{logit}(\Pr[\text{lung cancer}]) = \gamma 1[\text{smokes}] + \beta^T x$, where x is a vector of control variables. Conditionally on this framework (or "paradigm"), the issue becomes a comparison of the two models $M_1 : \gamma = 0$ and $M_2 : \gamma > 0$. Then a scientist's natural language statement "I am 90% sure that smoking causes lung cancer" is equivalent, given the framework, to the statement that $p(M_1) = 0.1$ and $p(M_2) = 0.9$. This does seem to make sense even if, unconditionally on the framework, $p(M_1) = p(M_2) = 0$.

Of course, the natural language statement itself can be viewed as not being about "truth", but rather about future data and trends in scientific opinion. It might mean, for example, "I am 90% sure that future data will be better predicted by M_2 than by M_1", or "I am 90% sure that within T years the belief that smoking causes lung cancer will be generally accepted"; note that the latter two statements can be given standard betting interpretations. For an example where scientists might attach substantial prior probability to the smaller ("null") model, consider cold fusion.

The authors describe the standard Bayesian procedure as a model *selection* procedure, but it is considerably richer than that. When comparing two models that genuinely represent rival scientific hypotheses, the posterior odds ratio provides a summary of the evidence for one model against the other; unless the evidence is very strong, one model will not necessarily be selected.

Often, however, model form is not the object of primary scientific interest. The authors did not say what the main scientific question was in their growth curve example, but I suspect that it was not the choice between the two models that they considered. If interest focuses instead on some other quantity, Δ, such as the next observation, Y_{16}, or the asymptote, β_0, then *model selection is a false problem*, and it is important to take account of model uncertainty. The Bayesian approach provides an immediate way of doing this using the equation

$$p(\Delta \mid Y) = \sum_{j=1}^{J} p(\Delta \mid Y, M_j) p(M_j \mid Y). \tag{1}$$

Hodges (1987) emphasized the importance of taking account of model uncertainty, pointing out that failure to do so leads to the overall uncertainty being underestimated, and hence, for example, to overly risky decisions.

If the posterior probability of one of the models is close to unity, or if the posterior distribution of Δ is almost the same for the models that account for most of the posterior probability, then $p(\Delta \mid Y)$ may be approximated by conditioning on a single model, namely by $p(\Delta \mid Y, M_i)$ for some i. This seems to be the main situation in which model selection, as such, is a valid exercise. The "evolutionary" process to which the authors refer is in reality an informal search method for finding the main models that contribute to the sum in equation (1), and in this sense may be viewed as an approximation to the full (standard) Bayesian procedure. Clearer recognition of this might lead to more satisfactory model search strategies.

4 The standard Bayesian procedure and sampling-based methods

The key quantity for the implementation of the standard Bayesian procedure is the marginal likelihood, $f(Y \mid M_j) = \int f(Y \mid \theta_j, X, M_j)\pi(\theta_j)d\theta_j$. The authors say that the Gibbs sampler does not readily produce an estimator of $f(Y \mid M_j)$. However, Newton and Raftery (1991) have recently pointed out the existence of a simple and general such estimator. They show that, given a sample from the posterior, *the marginal likelihood may be (simulation-consistently) estimated by the harmonic mean of the associated likelihood values.* This result applies no matter how the sample was obtained, whether directly using the analytic form of the posterior, by importance sampling, the Gibbs sampler, the SIR algorithm or the weighted likelihood bootstrap. There can be stability problems with this estimator, and slight modifications that avoid these are discussed in the cited reference.

The standard Bayesian procedure *is* a predictive approach since the marginal likelihood can be written

$$f(Y \mid M_j) = \prod_{r=1}^{n} f(Y_r \mid Y^{r-1}, M_j), \tag{2}$$

where $Y^{r-1} = (Y_1, \ldots, Y_{r-1})$. Note that the conditional densities on the right-hand side of equation (2) are conditional on the first $(r-1)$ observations, and *not* on all the other $(n-1)$ observations. Thus the standard Bayesian procedure is a "prequential" method in the sense of Dawid (1984), and not a cross-validation approach. Each conditional density on the right-hand side of equation (2) may be evaluated in a sampling-based way, using the same methods as the authors propose for their d_{4r}. It follows that this provides an alternative sampling-based way of calculating the marginal likelihood, and hence of implementing the standard Bayesian procedure.

Note also that equation (2) remains valid even if the observations are permuted. Thus, even if the model does not impose a natural ordering on the observations, "prequential diagnostics" may be obtained by sampling from the set of all permutations of the observations and averaging over diagnostics based on the conditional densities on the right-hand side of equation (2).

If one replaces the conditional densities on the right-hand side of equation (2) by densities conditional on all the observations except the rth one, one obtains the quantity that the authors denote by $D_4 = \prod_{r=1}^{n} d_{4r}$. This could be called a "pseudo-marginal likelihood", by analogy with the pseudo-likelihood concept introduced by Besag (1975). Using D_4 rather than $f(Y \mid M_j)$ is similar to using the pseudo-likelihood rather than the likelihood when the latter is available, which does not seem to be a very good choice. As an argument in favor of D_4, however, the authors point out that with improper priors D_4 is defined whereas $f(Y \mid M_j)$ is not. This strikes me as a disadvantage of improper priors rather than of the standard marginal likelihood.

I will attempt to summarize the various analogies and equivalences discussed in the following table.

Prequential analysis	Cross-validation
Likelihood	Pseudo-likelihood
Marginal likelihood ($f(Y \mid M_j)$)	"Pseudo-marginal likelihood" (D_4)
Posterior model probability/ Bayes factor	Fixed-level significance test
BIC (Schwarz, 1978)	AIC, C_p

Entries in the same column are regarded as being related, either by being motivated by the same approach or by being asymptotically equivalent. Entires in the same row are viewed as different approaches to the same task or concept. I prefer the entries in the left-hand column, headed "prequential analysis", while the authors seem to incline to the entries in the right-hand column. Note that the difference can be important, especially with large samples.

References

[1] Bartlett, M.S. (1957) A comment on D.V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

[2] Besag, J.E. (1975) Statistical analysis of non-lattice data. *Statistician* **24**, 179–195.

[3] Dawid, A.P. (1984) Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). *J. R. Statist. Soc.* A **147**, 178–292.

[4] Hodges, J.S. (1987) Uncertainty, policy analysis and statistics (with Discussion). *Statist. Sci.* **2**, 259–291.

[5] Newton, M.A. and Raftery, A.E. (1991) Approximate Bayesian inference by the weighted likelihood bootstrap. Technical Report no. 199, Department of Statistics, University of Washington.

[6] Raftery, A.E. (1988) Approximate Bayes factors for generalized linear models. Technical Report no. 121, Department of Statistics, University of Washington.

[7] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics

Adrian E. Raftery
University of Washington *

Jeffrey D. Banfield
Montana State University

December 5, 1990

1 Introduction

It is a pleasure to congratulate Julian Besag, Jeremy York and Annie Mollié on a superb paper that will surely take its place as yet another of Julian Besag's greatest hits, and as a first hit for the other two authors!

They argue that many spatial statistics problems can appropriately be viewed as problems in image restoration, and that image restoration problems are best solved by postulating a Markov Random Field model, and then calculating the posterior distribution of the quantities of interest using the Gibbs sampler. This is an appealing argument and the examples are encouraging. One possible difficulty arises from the fact that the models may not have the same large-scale properties as the data they are used to analyze, and this raises some questions about the status of the resulting inferences; see section 3 below.

For the practical implementation of the Bayesian image restoration approach it is important to know how many iterations of the Gibbs sampler are required, and we propose

*Adrian E. Raftery is Professor of Statistics and of Sociology, GN-22, University of Washington, Seattle, WA 98195. Jeffrey D. Banfield is Assistant Professor, Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717. This research was supported by the Office of Naval Research under Contracts N-00014-88-K-0265 and N-00014-89-J-1114. The authors are grateful to Julian Besag and Jeremy York for helpful discussions, and also to Jeremy York for computational assistance. Of course, the usual disclaimer applies, as they will be able to make clear in their rejoinder!

a method for determining this in section 2. In section 3 we consider an alternative to the Bayesian image restoration approach for the archeology example, based on mathematical morphology. In section 4 we discuss several issues in the modeling that underlies the Bayesian image restoration approach: the modeling of spatial dependence, allowing for model uncertainty, the improper posterior distributions that arise in hierarchical Bayes modeling, and the modeling of local dependence between counts when it cannot be assumed that the y_i's are independent given x.

2 How many iterations in the Gibbs sampler?

The authors point out that the Bayesian image restoration approach is not yet feasible for typical images containing 10^5 or 10^6 pixels, although it can be implemented for the problems they consider, involving 100–300 "pixels". The main reason for this is the large number of iterations required by the Gibbs sampler. For instance, in the disease risk example, the authors ran the Gibbs sampler for 11,000 iterations, discarding the first 1,000, and storing every 10th or 20th value thereafter; these numbers were fairly arbitrarily picked initially, although they appeared to give reasonable results. As a practical matter, it would seem desirable to run the Gibbs sampler for the smallest number of iterations necessary to attain a required level of accuracy, and we now outline an approximate way of determining what that is.

The validity of the Gibbs sampler stems from the fact that each cycle of the algorithm corresponds to one step of a Markov chain with stationary transition probabilities and that an ergodic theorem applies for functions of x under certain regularity conditions (Geman and Geman, 1984). This suggests that one generate a single long realization of the Markov chain and base inference on it, which is what the authors have done. By contrast, several authors who have recently applied the Gibbs sampler to more standard statistical problems (Gelfand and Smith, 1990; Gelfand, Hills, Racine-Poon and Smith, 1989) have instead adopted the

following algorithm: (i) choose a starting point; (ii) run the Gibbs sampler for T iterations and store only the last iterate; (iii) return to (i). The relationship of this latter algorithm to the underlying theory seems problemmatical, and here we consider only the case of a single long realization.

We consider the specific problem of producing results such as those in the authors' Figures 7 and 8, namely the calculation of particular quantiles of the posterior distribution of a function of x. 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 x. 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 (in their section 4 the authors use $M = 1,000$, $N = 10,000$ and $k = 10$ or 20). Our problem is to determine M, N, and k.

We first calculate U_t for each iteration t, and then form $Z_t = \delta(U_t > u)$, 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. 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 problematic in the presence of large samples.

We applied the suggested method to series of 11,000 iterations of the Gibbs sampler for u and v for each of 12 départements based on the data of the authors' Figure 4; the Gibbs sampler output was kindly supplied to us by Jeremy York. We first give illustrative results with $q = .025$, $r = .005$, $s = .95$, and $\varepsilon = .001$. For all 24 parameters considered, k was either 1 or 2, M was never more than 6, and N was always 9,034 or less. However, for the spatial smoothness parameter κ, the situation was quite different and the requirements of the Gibbs sampler were larger: $k = 5$, $M = 65$ and $N = 42,500$.

The authors' Figure 6 implicitly requires that the .1 quantile of $e^x = e^{u+v}$ be correct to one decimal place with high probability. This implies, approximately, that for each u and v we specify $q = .1$, $r = .012$ and $s = .95$, which yielded $k \leq 3$, $M \leq 12$ and $N \leq 8,300$ for all 24 parameters considered. In practice, the method would be implemented by first running, say, 1,000 iterations and then deciding on k, M and N on the basis of those. In the present case, this appeared to work quite well.

One conclusion is that the number of iterations required can vary considerably depending on what is being estimated. Here, far more iterations are required for the overall spatial smoothness parameter κ than for the relative risk at an individual node. It does not seem necessary to use only every 10th or 20th iterate, and, indeed, doing so is probably quite wasteful. Indeed, it is not clear that discarding *any* iterates is advantageous, although it does simplify the calculations here. Also, it does not seem necessary to discard the first 1,000 iterates, or anything like it; our calculations never indicated it to be necessary to discard more than the first 65.

We hope that the suggestion made here will allow the Gibbs sampler to be used more

efficiently, and hence to make Bayesian image restoration feasible for larger problems. The computer code used to carry out these calculations is available from Adrian Raftery by electronic mail at *raftery@stat.washington.edu.*

3 Using morphology to locate archeological sites: The EP algorithm

The problems of locating archeological sites in section 3 can be regarded as one of locating and finding the boundaries of objects in the image, in this case sites of previous activity. For comparative purposes, we apply a different technique based on mathematical morphology, known as the EP algorithm, that was originally developed for locating ice floes in satellite images (Banfield and Raftery, 1989).

The EP algorithm consists of two parts: erosion and propagation. The erosion part of the algorithm, which identifies the potential edge elements, is a standard application of ideas in mathematical morphology (Serra, 1982). The algorithm is iterative and operates on a binary image consisting of objects (sites of activity) on a contrasting background. At the first iteration, if a pixel is classified as "active" and a specified subset of its neighbors is inactive, the pixel is "deactivated" and becomes inactive. At the second iteration, the same operation is performed on the image resulting from the first iteration, and so on. The edge elements consist of the pixels "deactivated" at the first iteration. The propagation part of the EP algorithm keeps track of the site to which an edge pixel belongs by locally propagating the information about edge elements into the interior of the object as it is eroded.

We started the EP algorithm from the naive classification given in the authors' Figure 1(a), which is, in fact, simple thresholding. The results are shown in Figure 1. They are quite similar to those obtained from the Bayesian image restoration method, perhaps strikingly so given the noisy appearance of the naive classification in the authors' Figure 1(a). The pixels where the classifications disagree are pixels where the uncertainty is, in any event,

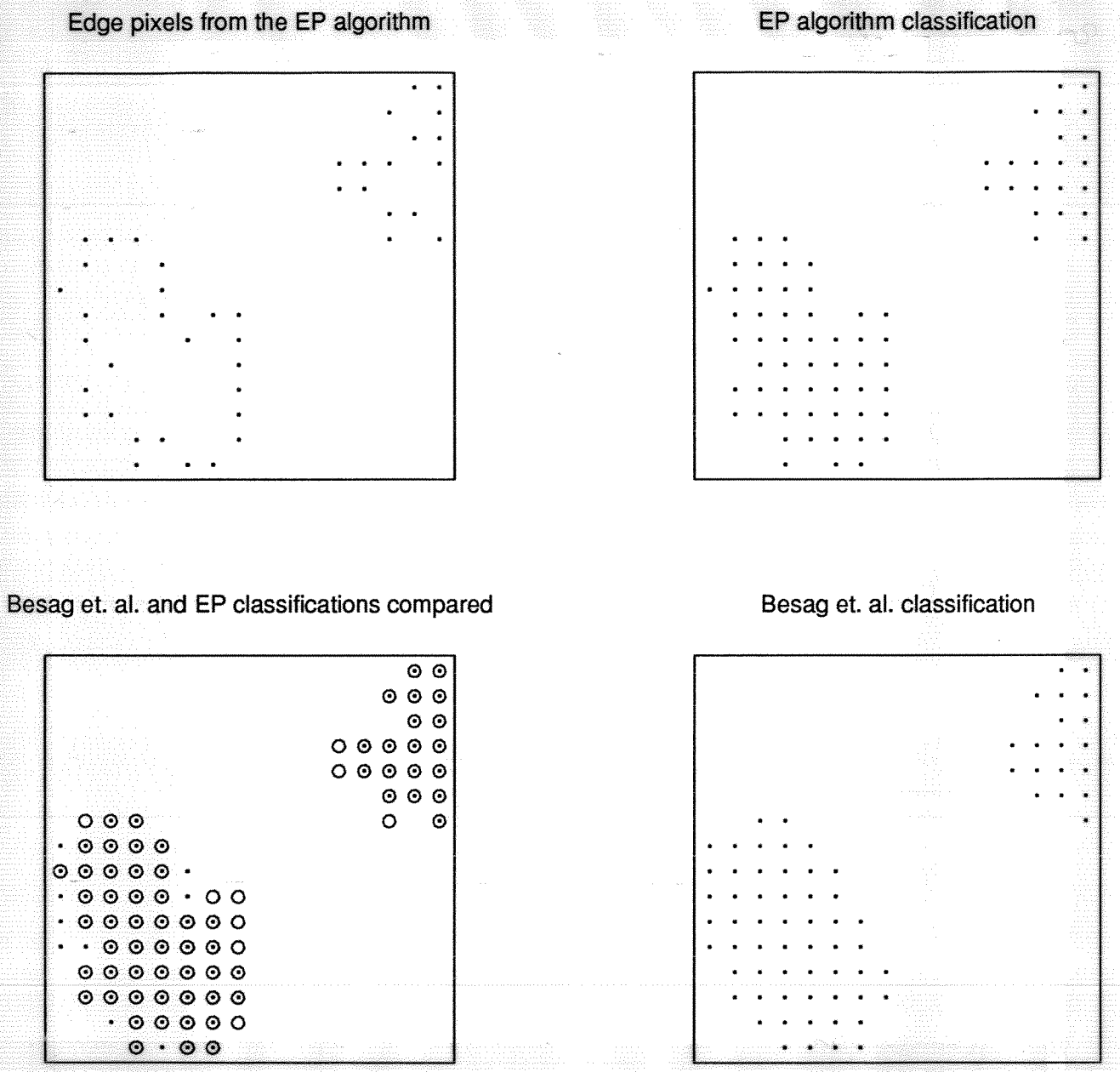


Figure 1: The EP algorithm applied to the archeology data: (a) The edge pixels identified by the EP algorithm; (b) The classification by the EP algorithm; (c) The EP and Bayesian image restoration classifications superimposed; (d) The Bayesian image restoration classification.

considerable. For almost all these pixels, the posterior probabilities in the authors' Figure 2 are well away from 0 or 1, and many of them are border pixels for which, as the authors observe, any spatial procedure is necessarily of doubtful value. Note that the EP algorithm uses only the naive classification, and does not, unlike the Bayesian image restoration method, use the full original data.

The EP algorithm has advantages and disadvantages compared to the Bayesian image restoration method: it is much faster but yields less information. The EP algorithm involves only about 10 iterations here, each of which consists only of small integer additions, while the Bayesian image restoration method uses 15,000 iterations each of which involves one exponentiation per pixel. Thus we estimate that the Gibbs iterations take at least 1,000 times, and perhaps 10,000 times as much CPU time as the EP iterations. On the other hand, the Bayesian image restoration method does have the important property of providing a statement of uncertainty in the form of posterior probabilities at each pixel.

However, we do wonder about the precise status of this statement of uncertainty. Markov random field models such as that on which the analysis is based often have a substantial probability of producing infinite one-color patches, in which case typical realizations of $\{p(x)\}$ will not resemble the true scene. This is known as the phase transition phenomenon and is discussed, for example, by Besag (1986). One consequence is that the prior may be heavily concentrated on uniform images, and one might expect this to bias the posterior towards too much uniformity. We would welcome the authors' views on these points.

4 Modeling issues

4.1 Modeling the spatial dependence

In the disease mapping example, the authors model the spatial dependence using equation (4.1). This seems sensible in the case of a spatial array that is not too dissimilar to a

rectangular array of pixels, such as the French départements. As a historical footnote, the regularity of the administrative map of France is due to Napoléon, who laid it out in the early nineteenth century in such a way that a man on horseback could reach any part of a département in a day's ride.

However, we wonder whether the specification (4.1) would be as satisfactory for much more irregularly spaced arrays. One example is the Standard Statistical Metropolitan Areas (SMSAs) of the United States, where the "neighbors" are close together in the North-East, but much further apart in the rest of the country.

An alternative but related specification has been developed in geostatistics as the basis for the so-called "kriging" method (Journel and Huijbregts, 1978). This implements the idea that dependence decreases with distance. The form of the dependence is described by the semivariogram, $\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{Var}[u(\mathbf{s}) - u(\mathbf{s}+\mathbf{h})]$, where $u(\mathbf{s})$ denotes the value of u at a location $\mathbf{s}$. If the covariance function, $C(\mathbf{h})$, exists, then $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$. If V is the resulting covariance matrix of the u_i's, and the u_i's are assumed to be jointly Gaussian, then $(u_i \mid u_{-i}) \sim N(\hat{u}_i, \sigma_i^2)$, where $\hat{u}_i = \sum_j a_{ij} u_j$ is the best linear predictor of u_i and σ_i^2 is its variance.

This may provide a more systemmatic basis for the choice of the quantities $\{a_{ij}\}$, which play a role similar to that of the $\{w_{ij}\}$ in equation (4.1). Another feature is that when, as in the disease risk example, the data correspond to areas rather than to points, the spatial dependence can take account of this explicitly. This is done by postulating a semivariogram for points, as above, and then integrating over areas to provide the corresponding values for the areas (Journel and Huijbregts, 1978). One would then proceed as before.

At first sight, it may seem that such an approach would be computationally prohibitive for even moderate data sets, since, in principle, it requires the inversion of n matrices, each of which is $(n-1) \times (n-1)$. However, if $\gamma(\mathbf{h})$ is modeled by a function with "sill", such as

the "Mathéron", or spherical, semivariogram,

$$\gamma(\mathbf{h}) = \begin{cases} \sigma^2\{\frac{3}{2}(\frac{|\mathbf{h}|}{a} - \frac{1}{2}(\frac{|\mathbf{h}|}{a})^3\} & |\mathbf{h}| \leq a \\ \sigma^2 & |\mathbf{h}| > a, \end{cases}$$

then many of the entries in V will be zero, and this can be used to reduce the computation involved in calculating the $\{a_{ij}\}$. Also, most of the $\{a_{ij}\}$ will be close to zero, and they could be set to zero without bad consequences, leading to an effective set of neighbors for each pixel, not necessarily restricted to the contiguous zones. In addition, the $\{a_{ij}\}$ have to be calculated only once for each value of (κ, λ) considered, remaining the same for each iteration of the Gibbs sampler. This suggests advantage to the strategy adopted by the authors for the archeological example, where the parameters of the prior were updated much less frequently than the values at the individual nodes.

These are tentative and untested ideas. However, the notion that the spatial modeling methods developed in geostatistics could be combined with the Bayesian image restoration methods proposed in the present paper may be a potentially fruitful one.

4.2 Model uncertainty

Several modeling choices are made in the authors' examples. These include the form of $\phi(z)$, namely whether it should be proportional to z^2 or to $|z|$, which covariates should be included in $t = A\theta$, the way the $\{w_{ij}\}$ are defined, and whether u and v should both be present. The authors, in common with most statistical modelers, have chosen a single model for each data set, and drawn conclusions conditionally on the selected model. This ignores the uncertainty associated with the model selection exercise itself. Analyses conditional on a single selected model fail to take account fully of uncertainty about structure, and so may well underestimate the uncertainty associated with their conclusions, thus, for example, biasing policy choices in favor of policies that rely on more certain information (Hodges, 1987).

Suppose that $m+1$ models $M_0, M_1, \ldots, M_m$ are being considered. In the present context, these might correspond, for example, to different choices of $\phi(\cdot)$, $\{w_{ij}\}$ and covariates. Then, if Δ is a quantity of interest in the analysis, we can take account of model uncertainty quite simply by basing inference on the unconditional posterior distribution of Δ,

$$p(\Delta \mid y) = \sum_{k=0}^{m} p(\Delta \mid y, M_k) p(M_k \mid y), \tag{1}$$

where $p(M_k \mid y)$ is the posterior probability of model M_k. This is a weighted average of the posterior densities of Δ under each of the models individually, weighted by their posterior probabilities. It will be well approximated by $p(\Delta \mid y, M_{k^\star})$, i.e. by conditioning on a single selected model $M_{k^\star}$, only if $p(M_{k^\star} \mid y) \approx 1$, or if the posterior distributions of Δ from the models with non-negligeable posterior probability are similar.

To calculate the posterior probabilities $p(M_k \mid y)$ we note that

$$p(M_k \mid y) \propto p(y \mid M_k) p(M_k). \tag{2}$$

In equation (2), $p(M_k)$ is the prior probability of M_k and

$$p(y \mid M_k) = \int p(y \mid \theta_k, M_k) p(\theta_k \mid M_k) d\theta_k, \tag{3}$$

where θ_k is the possibly vector parameter of M_k and $p(\theta_k \mid M_k)$ is its prior density. In the present context, this can be implemented by noting that x can also be included in equation (3), yielding

$$p(y \mid M_k) = \int \int p(y \mid x, \theta_k, M_k) p(\theta_k \mid M_k) dx d\theta_k. \tag{4}$$

This can be approximated by

$$p(y \mid M_k) \approx \frac{1}{T} \sum_{t=1}^{T} p(y \mid x^{(t)}, \theta_k^{(t)}, M_k), \tag{5}$$

where $\{x^{(t)}, \theta_k^{(t)}\}$ is the result of running the Gibbs sampler to obtain a sample from the *prior* distribution of (x, θ_k). A different approach to finding posterior probabilities using the

Gibbs sampler is to include a model indicator as an additional parameter (Carlin, Polson and Stoffer, 1990).

The implementation of the suggested approach to model uncertainty using equations (1), (2), (4) and (5) does not seem computationally prohibitive. At most, the computation is linear in the number of models that are fully analyzed, multiplying the required CPU time by about $2(m+1)$. However, there are several possible ways of reducing this. For example, the Gibbs sampler could be run in parallel on all the models. Also, an initial short run of equation (5) could be used to identify those models with substantial posterior probability, and a longer run restricted to those models then done to evaluate $p(\Delta \mid y)$ more precisely.

4.3 Improper posteriors in hierarchical Bayes modeling

In the authors' equation (4.5), the use of the obvious "non-informative" or scale-invariant prior for κ and λ, $p(\kappa, \lambda) \propto \kappa^{-1}\lambda^{-1}$, leads to an improper posterior distribution. As the authors point out, this is a common feature of Bayesian hierarchical models in general. It arises, for example, even in the simplest normal empirical Bayes model (Morris, 1983) where

$$(y_j \mid \theta_j, V) \sim N(\theta_j, V) \tag{6}$$

$$(\theta_j \mid \mu, A) \sim N(\mu, A) \quad (j = 1, \ldots N). \tag{7}$$

Then with the standard vague prior, $p(\mu, V, A) \propto V^{-1}A^{-1}$, the posterior $p(\theta_j \mid y)$ is improper. The authors mention the available remedy, in their case, of banning a neighborhood of $\kappa = \lambda = 0$, but instead use the improper prior (4.6), which is intended to approximate an improper uniform prior, but modified to be equal to zero at $\kappa = \lambda = 0$. The use of a uniform prior for a variance-like parameter seems somewhat unsatisfactory intuitively, as it has the disadvantages of an improper prior, without the advantages of scale invariance. Of course, it is not clear that this is really a serious problem in the present application.

Kahn (1990) analyzed this problem in the context of the normal empirical Bayes model

specified by equations (6) and (7). He reparameterized the model, setting $S = V + A$ and $T = \frac{V}{V+A}$. Then $S = \mathrm{Var}(y_j \mid \mu, S, T)$, and the prior $p(\mu, S, T) \propto S^{-1}$ leads to a proper posterior while retaining the desirable scale-invariant property of the standard prior.

By analogy, this suggests that in the present context we consider $\mathrm{Var}(y_i \mid u_{-i}, \kappa, \lambda)$, which is approximately equal to $(\frac{1}{c_i} + \frac{\kappa}{n_i} + \lambda)$ when κ and λ are small and c_i is large, as here. This suggests specifying the prior in terms of $\sigma = \frac{1}{\bar{c}} + \frac{\kappa}{\bar{n}} + \lambda$ and $\tau = \lambda/\sigma$, where an overbar denotes the average over all pixels. The natural choice is $p(\sigma, \tau) \propto \sigma^{-1}$, corresponding to $p(\kappa, \lambda) \propto \left(\frac{1}{\bar{c}} + \frac{\kappa}{\bar{n}} + \lambda\right)^{-2}$. This is an improper prior which retains, at least roughly, the desired scale-invariance properties, but does not exhibit the behavior near the origin that leads to impropriety. This prior may still lead to the Markov chain defined by the Gibbs sampler having an absorbing state, and one could multiply it by the expression in the authors' equation (4.6) to avoid this.

4.4 Local dependence between counts

The authors' model for the disease risk example assumes that, conditionally on the true relative risks x_i, the observed numbers of cases y_i are independent Poisson random variables, arguing that this is usually reasonable when the disease is non-contagious and rare. If the disease is contagious, however, it seems likely that the y_i's will be dependent, even conditionally on x. Even if the disease is non-contagious, it seems possible that the y_i's may be dependent. For example, if a disease is genetically transmitted, this could lead to spatial clustering even when the true risk is constant over space. If such dependence is present, then failing to take account of it seems likely to bias the estimated x_i's away from uniformity and hence, for example, to overstate the size and significance of the effects of covariates.

In the spirit of the authors' paper, the way to take account of such dependence is to model it explicitly. However, how to do this is not immediately obvious. The first possibility that springs to mind is the auto-Poisson model of Besag (1974). The problem with this

is that it can represent only negative dependence between neighboring pixels, producing a chessboard-like pattern, which seems unsatisfactory.

We would like to suggest another possible way of representing such spatial dependence between Poisson random variables that draws on ideas first developed in the time series context. The *mixture transition distribution* (MTD) model for a stationary time series $\{Z_t\}$ taking values in an arbitrary space $\mathbf{Z}$ is defined as follows (Raftery 1985a, 1985b; Martin and Raftery, 1987). Suppose that (V_i, W_i) $(i = 1, \ldots, p)$ is a set of bivariate random vectors taking values in $\mathbf{Z} \times \mathbf{Z}$, with conditional densities $f_i(v \mid w)$ with respect to some measure, where the marginal distribution of V_i is the same as that of W_i for each $i = 1, \ldots, p$. Then the conditional density of Z_t given $Z_{t-1}, \ldots, Z_{t-p}$ is given by

$$p(z_n \mid z_{n-1}, \ldots, z_{n-p}) = \sum_{i=1}^{p} \lambda_i f_i(z_t \mid z_{t-i}), \tag{8}$$

where $\sum \lambda_i = 1$. This can represent time series with arbitrary marginal distributions taking values in arbitrary spaces; in the discrete-valued case it fits data well, is physically motivated and is analogous in several ways to the standard autoregressive model. To specify a Poisson time series model, all that is needed is a bivariate Poisson distribution such as that of Holgate (1964) with mean μ and dependence parameter ζ, which yields

$$f_i(v \mid w) = f(v \mid w) = e^{-(\mu-\zeta)} \mu^{-w} \sum_{h=0}^{\min\{v,w\}} \frac{\binom{w}{h} \zeta^h (\mu - \zeta)^{v+w-2h}}{(v-h)}. \tag{9}$$

When the Poisson means are constant (i.e. the c_i and the x_i are constant) the obvious spatial generalization is just to replace the summation over past values in equation (8) by a summation over the neighbors of pixel n. Then the model is specified in terms of conditional distributions, and the Gibbs sampler machine can be set in motion as before. One way of generalizing this to the non-stationary situation that we have actually got, where the c_i and the x_i are not constant, is as follows. First postulate the existence of a spatial process $\{z_i^\star\}$ defined by equations (8) and (9), corresponding to constant c_i and x_i, and

let $F(\cdot)$ be the corresponding Poisson cumulative distribution function. Let $F_i(\cdot)$ be the Poisson cumulative distribution function corresponding to c_i and x_i. Then we model z_i as $z_i = F_i^{-1}(F(z_i^\star))$. If the expected counts are very small, then this will not be quite accurate due to the discreteness, and an exact solution may be obtained by allowing the dependence of z_i on $z_i^\star$ to be stochastic.

One difficulty with this suggestion is that the conditional distributions defined in this way do not define a valid joint distribution for the y_i's, by the Hammersley-Clifford theorem (Besag, 1974). However, it seems likely that any joint distribution for Poisson random variables that does satisfy the Hammersley-Clifford theorem will not allow a sufficiently broad range of positive dependence. The MTD model suggested here may well have the right *local* conditional dependence structure, while distributions that *do* satisfy the Hammersley-Clifford theorem will often have undesirable large-scale properties as well as unsatisfactory local properties.

Thus one may ask whether conditional distributions such as that specified by the MTD model that do *not* satisfy the Hammersley-Clifford theorem might not, nevertheless, provide useful operational procedures. Besag (1986) refers to this possibility, and we would appreciate the authors' current views on it.

References

Banfield, J.D. and Raftery, A.E. (1989). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Technical Report no. 172, Department of Statistics, University of Washington.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. Roy. Statist. Soc., ser. B*, **36**, 192-236.

Besag, J.E. (1986). On the statistical analysis of dirty pictures (with Discussion).

J. Roy. Statist. Soc., ser. B, **48**, 259-302.

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

Carlin, B.P., Polson, N.G. and Stoffer, D.S. (1990). A Monte Carlo approach to nonnormal and nonlinear state space modeling. Technical Report no. 486, Department of Statistics, Carnegie-Mellon University.

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

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1989). Illustration of Bayesian inference in normal data models using Gibbs sampling. Technical Report, Nottingham Statistics Group, Department of Mathematics, University of Nottingham.

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

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.

Hodges, J.S. (1987). Uncertainty, policy analysis and statistics (with Discussion). *Statist. Sci.*, **2**, 259-291.

Journel, A.G. and Huijbregts, C.J. (1978). *Mining Geostatistics.* London: Academic Press.

Kahn, M.J. (1990). Bayes empirical Bayes beta-binomial modeling with covariates, applied to health care policy. Unpublished Ph.D. dissertation, Department of Statistics, University of Washington.

Martin, R.D. and Raftery, A.E. (1987). Robustness, computation and non-Euclidean models. *J. Amer. Statist. Ass.*, **82**, 1044-1050.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications (with Discussion). *J. Amer. Statist. Ass.*, **78**, 47-65.

Raftery, A.E. (1985a). A model for high-order Markov chains. *J. Roy. Statist. Soc., ser. B*, **47**, 528-539.

Raftery, A.E. (1985b). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni*, **3-4**, 149-162.

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.

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

Serra, J.P. (1982). *Image Analysis and Mathematical Morphology.* New York: Academic Press.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 212	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Three Short Papers on Sampling-Based inference: 1. How Many Iterations in the Gibbs Sampler? 2. Model Determination 3. Spatial Statistics	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Adrian E. Raftery, Steven Lewis, Jeffrey D. Banfield	8. CONTRACT OR GRANT NUMBER(s): N00014-88-K-0265
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, University of Washington, Seattle, WA 98195	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR-661-003
11. CONTROLLING OFFICE NAME AND ADDRESS: —	12. REPORT DATE: June 1991
	13. NUMBER OF PAGES: 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): ONR Code N63374, 1107 NE 45th Street, Seattle, WA 98195	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

This technical report consists of three short papers on Monte Carlo Markov chain inference. The first paper, "How many iterations in the Gibbs sampler?," proposes an easily implemented method for determining the total number of iterations required to estimate probabilities and quantiles of the posterior distribution, and also the number of initial iterations that should be discarded to allow for "burn-in".

The second paper discusses model determination via predictive distributions. The paper advocates the standard Bayesian procedure that uses Bayes factors, and points out that this can be implemented quite easily using sampling-based methods.

The third paper discusses issues in spatial statistics that use sampling-based methods. Several issues in the Bayesian image restoration approach are discussed: the modeling of spatial dependence, allowing for model uncertainty, the improper posterior distributions that arise in hierarchical Bayes modeling, and the modeling of local dependence between counts when it cannot be assumed that the observations are independent given the true rates.