SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B.* To appear.

SWENDSEN, R. H. and WANG, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

THOMPSON, E. A. and GUO, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* **8** 149–169.

TIERNEY, L. (1991). Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, Univ. Minnesota,

TÓTH, B. (1986). Persistent random walks in random environment. *Probab. Theory Related Fields* **71** 615–625.

WANG, J. S. and SWENDSEN, R. H. (1990). Cluster Monte Carlo algorithms. *Phys. A* **167** 565–579.

WEI, G. C. G. and TANNER, M. A. (1990). Calculating the content and the boundary of the highest posterior density region via data augmentation. *Biometrika* **77** 649–652.

YOUNES, L. (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré Probab. Statist.* **24** 269–294.

Comment: Monitoring Convergence of the Gibbs Sampler: Further Experience with the Gibbs Stopper

Lu Cui, Martin A. Tanner, Debajyoti Sinha and W. J. Hall

1. INTRODUCTION

Whether one follows the "multiple-run" or the "one long run" approach to implementing Markov chain methods, diagnostics for monitoring convergence will be of value. The purpose of this note is to provide further illustration of one such diagnostic, the Gibbs Stopper, originally presented in Ritter and Tanner (1992) in the multiple run context.

The basic idea behind the Gibbs Stopper is to assign the weight w to the vector $\theta = (\theta_1, \ldots, \theta_d)$, which has been drawn from the current approximation to the joint density g_i via

$$w(\theta) = \frac{q(\theta_1, \ldots, \theta_d|Y)}{g_i(\theta_1, \ldots, \theta_d)},$$

where $q(\theta_1, \ldots, \theta_d|Y)$ is proportional to the posterior density $p(\theta_1, \ldots, \theta_d|Y)$. As g_i converges toward $p(\theta_1, \ldots, \theta_d|Y)$, the distribution of the weights (associated with values of θ drawn from g_i) should converge toward a spike distribution. We have found this observation useful in assessing convergence of the Gibbs sampler, as well as in transforming a sample from g_i into a sample from the exact distribution; see Ritter and Tanner (1992). Historically, the idea of using importance weights to monitor convergence of the data augmentation algorithm was first presented in the Rejoinder of Tanner and Wong (1987) and illustrated in Wei and Tanner (1990).

To write down the functional form for g_i for the Gibbs sampler, we introduce notation following Schervish and Carlin (1990). Let $p^{(i)}(\theta) = p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, Y)$. For two vectors θ and θ', define for each $i < d$, $\theta^{(i')} = (\theta_1, \ldots, \theta_i, \theta'_{i+1}, \ldots, \theta'_d)$ and $\theta^{(d')} = \theta$. As noted in Schervish and Carlin (1990), if g_i is the joint density of the observations sampled at iteration i, then the joint density (g_{i+1}) of the observations sampled at the next iteration is given by

$$(1) \qquad \int K(\theta', \theta) g_i(\theta')\, d\lambda(\theta'), \qquad K(\theta', \theta) = \prod_{i=1}^{d} p^{(i)}(\theta^{(i')})$$

[see also Tanner and Wong (1987) and Liu, Wong and Kong (1991, 1991a)]. One may approximate the integral in (1) via the method of Monte Carlo. In particular, given the observations $\theta^1, \theta^2, \ldots, \theta^m$, use the Monte Carlo sum

$$(2) \qquad \frac{1}{m} \sum_{j=1}^{m} K(\theta^j, \theta)$$

to approximate $g_{i+1}(\theta)$. Ritter and Tanner (1992) suggest using θ values from independent chains. In this note, we use *successive* θ values from one chain to construct the Monte Carlo sum (2). Note that construction of (2) requires the normalizing constants (or good approximations to the normalizing constants) for the conditional distributions. Also note that we are examining, through $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$, the first component of each θ vector along with components

Lu Cui is a graduate student, Martin A. Tanner is Professor, Debajyoti Sinha is a graduate student and W. J. Hall is Professor, Department of Biostatistics and Department of Statistics, Box 630, University of Rochester, Rochester, New York 14642.

2 through d of the other $m - 1$ θ vectors, the first and second components of each θ vector along with components 3 through d of the other $m - 1$ θ vectors, and so on, thereby expanding the coverage of the parameter space. We feel that the effort in constructing (2) will yield useful information regarding the state of the Markov chain vis-à-vis the equilibrium distribution. An example of the potential of this approach is given in the next section.

To illustrate this convergence diagnostic, we consider a "witch's hat" distribution presented in Matthews (1991). The posterior under consideration is proportional to a mixture of a multivariate normal distribution and the uniform distribution on the open d-dimensional hypercube $(0,1)^d$ C:

$$(3) \qquad (1-\delta)\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d e^{-\sum_{i=1}^d ((y_i-\theta_i)/\sigma)^2/2} + \delta I_{(Y \in C)}.$$

We have chosen $\delta = 10^{-11}$, $\sigma = 0.03$, $d = 9$ and $Y =$ (0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9). A cursory examination of the posterior reveals a spike centered at Y, with a flat "brim" extending out to the boundary of the unit hypercube. We proceed under the assumption that the posterior at hand is analytically complicated, thereby not allowing for an easy recognition of the location of or number of spikes. Clearly, in a situation where determining the number of and location of the modes is straightforward, one would focus attention around these points, possibly along the lines suggested by Gelman and Rubin (1992).

2. RESULTS

Figure 1 presents a history of the θ_1 marginal, starting from the point (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1). The plots for the other marginals are quite similar. As can be seen from the plot, the Markov chain wanders about the hypercube until iteration 4,400 or so, at which point it locates the mode. A run shorter than 4,400 iterations would not have detected the spike.

Figure 2 presents a plot of $w(\theta^{(i)})$ versus iteration i, $i = 1,270$. The weights in this plot are based on the first 270 successive (9-dimensional) points in the chain. The 270 weights in the plot were standardized to have mean zero and unit standard deviation. As can be seen in the plot, all 270 weights are equal, with the exception of two weights.

The reason for the outlying weights is easily explained. The conditional densities $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ used in computing K in (1) are proportional to

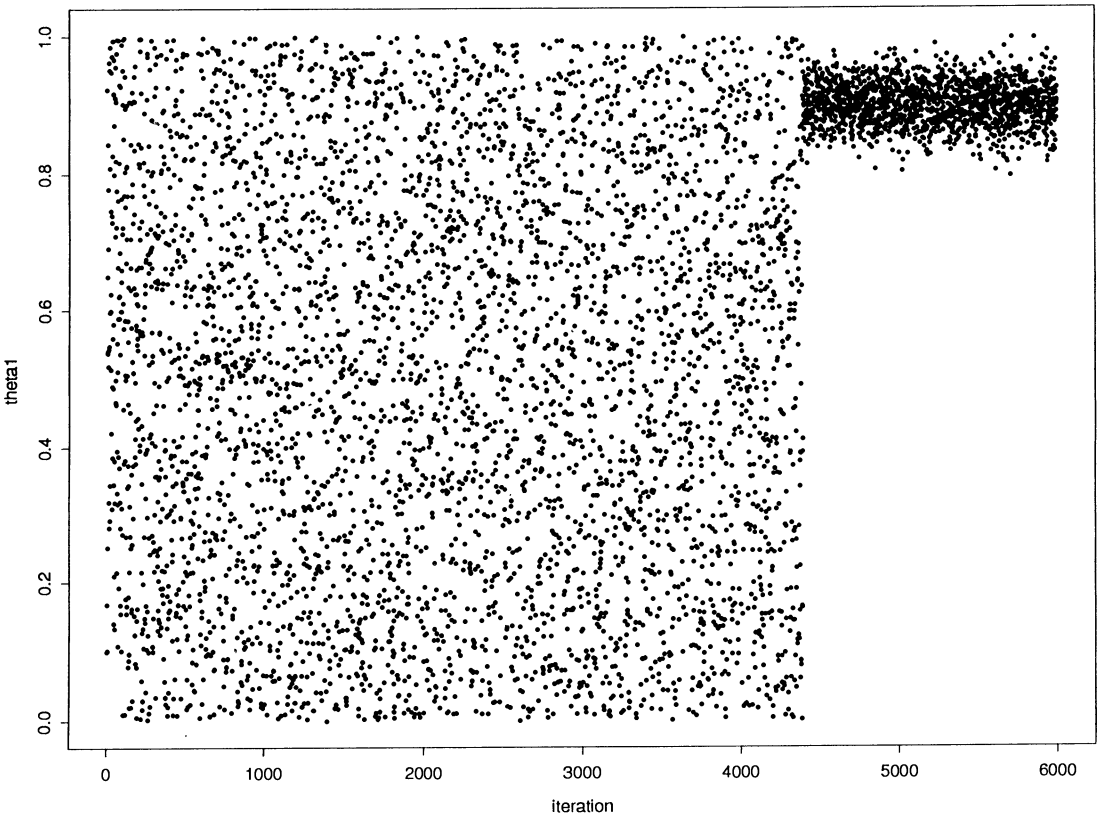

FIG. 1. θ_1 *versus iteration.*

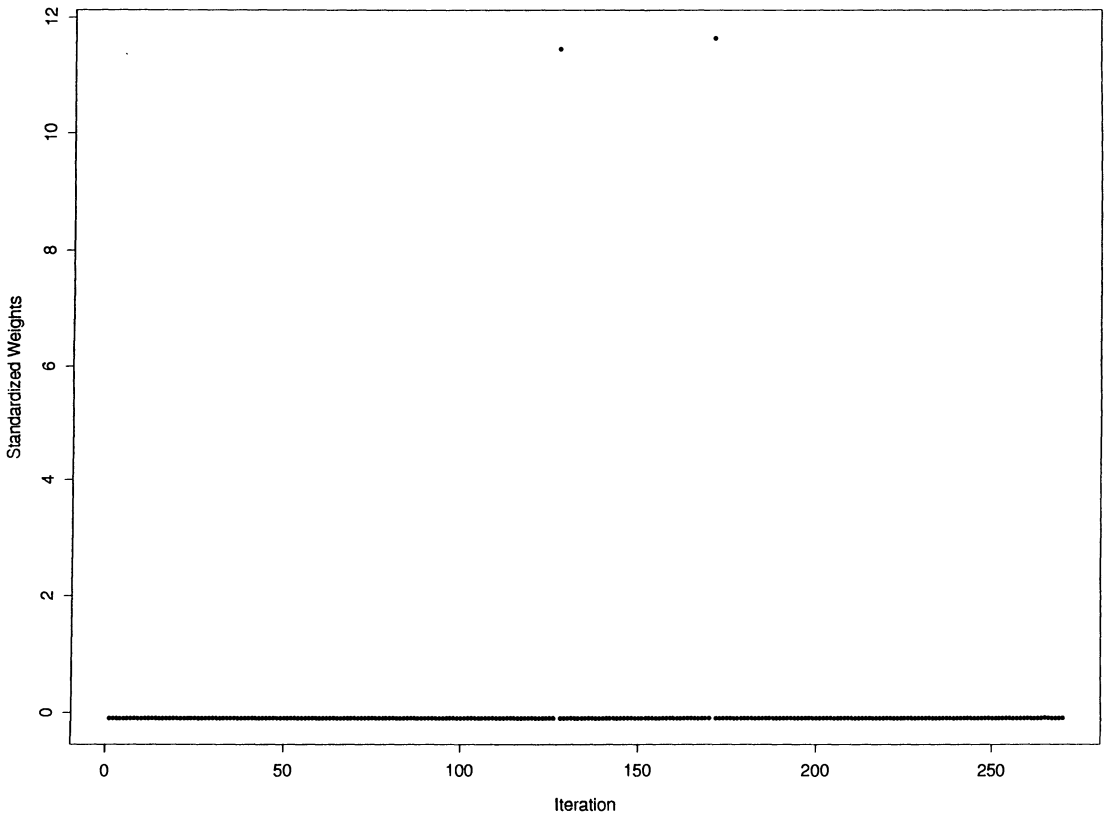

FIG. 2. *Weights versus iteration.*

(3) for specified θ_k and $0 < \theta_i < 1$. When θ_i is "far" from y_i, $i \neq k$, $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ is virtually equal to unity, independent of the value of θ_k. When θ_i is "near" y_i, $i \neq k$, $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ (as a function of θ_k) follows the normal curve. The outlying weight points noted in Figure 2 stem from the fact that in three of the 270 products in (2), which are averaged to compute g_{i+1}, one of the terms $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ has seven of the eight θ_i's within three σ's of 0.9, $i \neq k$, with θ_k "far" from 0.9. (For future reference call these components θ_i^*, $i \neq k$.) This term is nearly zero, thereby causing the entire product of terms in K to be small. Thus, rather than averaging 270 products equal to unity to compute g_{i+1}, the average was diminished by the three near-zero products, leading to the outlying weight.

As a follow-up to the investigation of the cause of the two outliers, we located the maximizer (θ_k^*) of $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$, where the θ_i's were set equal to the θ_i^* $(i \neq k)$ values identified in the previous paragraph. We then started a chain from this point. Figure 3 presents a plot of the history of the θ_1 marginal for this path; the other marginals are similar. As indicated in the figure, the chain moved immediately into the neighborhood of the spike.

In this example, we see how a careful examination of the outlying weights helps to locate the spike much earlier than the 4,400 iterations required by the original chain. Of course, this Gibbs Stopper technique is not infallible. Had we only considered the first 100 points in the Markov chain, we would have missed the outliers. Similarly, one would expect that smaller values of σ and higher values of d would require more terms in (2)—though both of these modifications would probably increase the run time of the chain as well. We feel, however, that careful use of the conditionals $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ can reveal more information about the convergence of the chain than what can be learned from the realized values of the Markov chain alone (i.e., the observed multivariate time series).

We examined additional weight plots based on nonoverlapping segments (iterations 271–810, 810–1,850, 1,850–4,010, etc.) of the original Markov chain. One does not see in this series of plots a degeneration of the distribution of the weights about a spike, as would be expected if the chain was in equilibrium—providing further indication to the data analyst of an anomaly regarding convergence of the chain. These plots highlight the slowly mixing nature of the chain.

In summary, we feel that a careful examination of

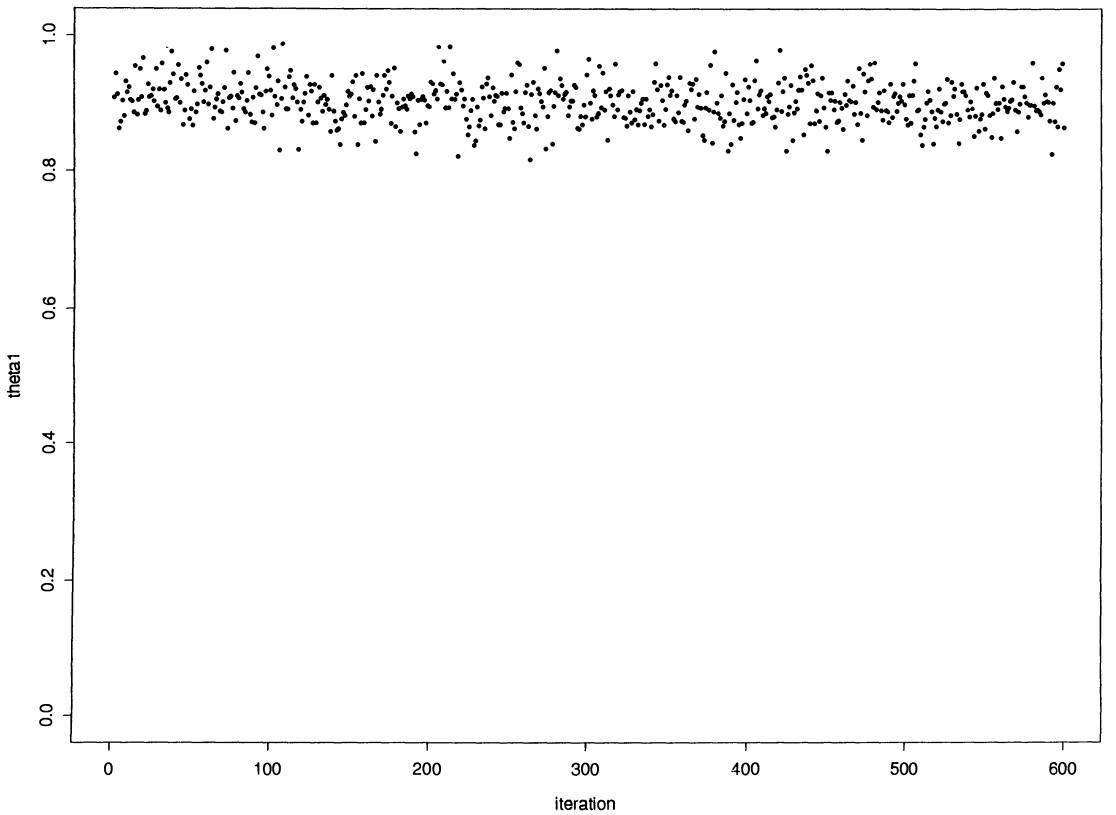

FIG. 3. θ_1 *versus iteration.*

the importance weights can yield valuable information about the convergence of the Markov chain. Further experience with this Gibbs Stopper method is warranted. Also of value would be analytical expressions that quantify the probability of outlier detection for important classes of problems.

ACKNOWLEDGMENTS

M. A. Tanner was supported by NIH Grant RO1-CA35464. D. Sinha was supported by NIH Grant RO1-CA52572.

Comment

Alan E. Gelfand

As noted by Gelman and Rubin, the problem of creating a simulation mechanism is clearly separate from the problem of using this mechanism to draw inference. Moreover, for the former problem, as observed in Green and Han (1992), the objectives of rapid convergence and good estimation performance are distinct. Translating these objectives to the latter problem, it appears that Gelman and Rubin focus on diagnosis of convergence, whereas Geyer focuses on assessing estimation performance. Again, these enterprises are not identical, accounting in part for the authors' differing views.

The two papers share a common thread in that, regardless of whether single or multiple trajectories are used, the state space of the Markov chain at each iteration is reduced to a univariate observation with trajectories thus treated as univariate time series. Though the authors' proposals can be carried out for any univariate reduction of interest, the thrust of my comments is the suggestion that, at least in certain

Alan E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, Connecticut 06269-3120.