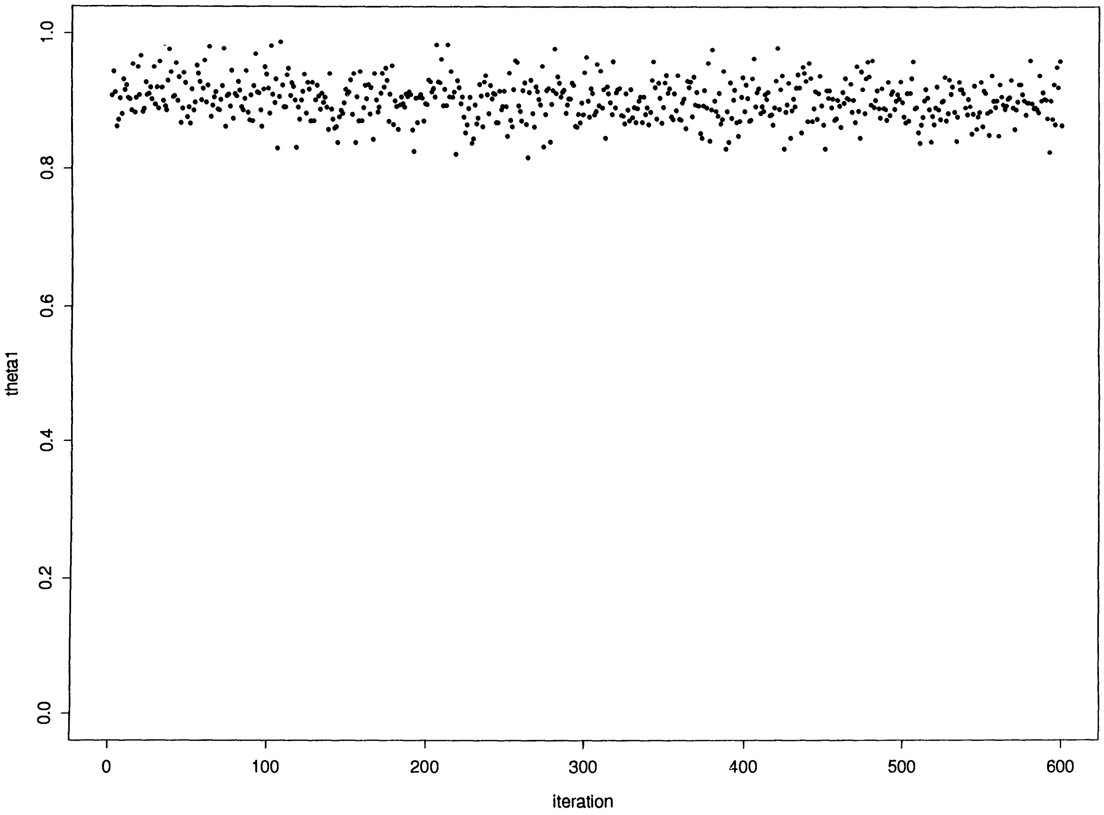

FIG. 3. θ_1 *versus iteration.*

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.

situations we can and should work with the entire state vector. The notion of investigating convergence with regard to the joint distribution of the variables is conceptually more satisfying, but, in addition, the sole analytic form we know explicitly is the nonnormalized invariant joint density that the Markov chain has been designed to have as its equilibrium distribution.

Denoting this nonnormalized density with respect to Lebesgue measure by $f(\mathbf{x})$ with $\mathbf{x}$ a $p \times 1$ vector, a convergence diagnostic built around $f(\mathbf{x})$ was proposed for the Gibbs sampler as the Gibbs stopper in Ritter and Tanner (1992). The idea is that, if at the tth iteration the marginal density of $\mathbf{x}$ is, say, $h^{(t)}(\mathbf{x})$, then under convergence, $w(\mathbf{x}) = f(\mathbf{x})/h^{(t)}(\mathbf{x})$ should be roughly constant. Of course $h^{(t)}$ is unavailable explicitly except in trivial cases. However, note that $h^{(t)}(\mathbf{x}) = \int h(\mathbf{x}|\mathbf{y}) \cdot h^{(t-1)}(\mathbf{y})\, d\mathbf{y}$, where $h(\mathbf{x}|\mathbf{y})$ is the transition kernel for the Markov chain. Hence, if $\mathbf{x}_j^{(t-1)} j = 1, 2, \ldots, m$ is a sample from $h^{(t-1)}$, a Monte Carlo approximation for $h^{(t)}$ arises as

$$\hat{h}^{(t)}(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^{m} h\left(\mathbf{x}|\mathbf{x}_j^{(t-1)}\right). \tag{1}$$

In the case of the Gibbs sampler, $h(\mathbf{x}|\mathbf{y})$ takes the form $\prod_{i=1}^{p} h(x_i|x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_p)$, where the terms in the product are the complete conditional distributions of the variables.

We can extend this idea to the case of a general Hastings-Metropolis algorithm. If f is absolutely continuous with respect to some measure μ, the transition kernel associated with such an algorithm is defined to be a mixed measure of the form

$$\begin{aligned}\Pr(\mathbf{x}^{(t)} \in A|\mathbf{x}^{(t-1)} = \mathbf{y})\\ = \int_A a(\mathbf{y}, \mathbf{x})q(\mathbf{y}, \mathbf{x})\, d\mu(\mathbf{x}) + r(\mathbf{y})\delta_{\mathbf{y}}(A),\end{aligned}$$

where $r(\mathbf{y}) = 1 - \int a(\mathbf{y}, \mathbf{x})q(\mathbf{y}, \mathbf{x})\, d\mu(\mathbf{x})$ and $\delta_{\mathbf{y}}(A) = 1$ if $\mathbf{y} \in A$, or $= 0$ if $\mathbf{y} \notin A$.

Here $q(\mathbf{y}, \mathbf{x})$ is the "proposal" transition kernel and $a(\mathbf{y}, \mathbf{x})$ is the "moving" probability [see, e.g., Tierney (1991) for further details]. Suppose $H^{(t)}$ denotes the marginal probability measure for $\mathbf{x}^{(t)}$. Then direct calculation shows that $H^{(t)}$ is absolutely continuous with respect to μ if the starting distribution is and, in fact,

$$h^{(t)} \equiv \frac{dH^{(t)}}{d\mu} = s^{(t-1)} + r \cdot \frac{dH^{(t-1)}}{d\mu}, \tag{2}$$

where $s^{(t-1)}(\mathbf{x}) = \int a(\mathbf{y}, \mathbf{x})q(\mathbf{y}, \mathbf{x})\, dH^{(t-1)}(\mathbf{y})$. Estimation of (2) is straightforward given $\mathbf{x}_j^{(t-1)}$, $j = 1, \ldots, m$, a sample from $h^{(t-1)} = dH^{(t-1)}/d\mu$. In fact, for a given $\mathbf{x}$, $\hat{s}^{(t-1)}(\mathbf{x}) = m^{-1} \sum_{j=1}^{m} a(\mathbf{x}_j^{(t-1)}, \mathbf{x})q(\mathbf{x}_j^{(t-1)}, \mathbf{x})$, $h^{(t-1)}(\mathbf{x})$ would be an appropriate kernel density estimate, and $r(\mathbf{x})$ would be computed by noniterative Monte Carlo using draws from $q(\mathbf{y}, \mathbf{x})$. These ideas can be further extended to more general hybrid Markov chain Monte Carlo algorithms such as Metropolis within Gibbs (Müller, 1992).

But then, given an estimate of $h^{(t)}$, we use $\hat{w} = f/\hat{h}^{(t)}$ in place of w. If, in fact, we have a sample $\mathbf{x}_j^{(t)}$, $j = 1, \ldots, m$ from $h^{(t)}$, we could naturally obtain the set of $\hat{w}(\mathbf{x}_j^{(t)})$ and see how "constant" they are, perhaps using a histogram or a dispersion measure.

Apart from the computational burden in computing $\hat{h}^{(t)}$, two broader problems arise in the implementation of this convergence diagnostic. First, if p is large we will require m very large in order that $\hat{h}^{(t)}$ be a good estimator of $h^{(t)}$. Second, since the normalizing constant for f is unknown, we will not know whether the $\hat{w}$'s are tending to the correct constant. It is possible that the $\hat{w}$'s are roughly constant but that some portion of the mass of f will have been missed (Roberts, 1992).

In special cases the second problem can be disposed of through a suggestion of Zellner (personal communication). Suppose we can partition $\mathbf{x}$ into $(\mathbf{x}_1, \mathbf{x}_2)$, where $h(\mathbf{x}_1|\mathbf{x}_2)$ and $h(\mathbf{x}_2|\mathbf{x}_1)$ are standard densities. This can occur when conjugacies are incorporated into the model as, for instance, in Gaussian linear models with the customary normal × inverse Wishart prior. If $\mathbf{x}_1$ is the vector of coefficients and $\mathbf{x}_2$ the variance-covariance matrix, then $h(\mathbf{x}_1|\mathbf{x}_2)$ and $h(\mathbf{x}_2|\mathbf{x}_1)$ will be updated normal and inverse Wishart densities respectively. But then, if $\mathbf{x}_j^{(t)}$, $j = 1, \ldots, m$, are a sample from $h^{(t)}$, we can obtain the marginal densities

$$\hat{h}(\mathbf{x}_1) = \frac{1}{m} \sum_{j=1}^{m} h(\mathbf{x}_1|\mathbf{x}_{2j}^{(t)}), \qquad \hat{h}(\mathbf{x}_2) = \frac{1}{m} \sum_{j=1}^{m} h(\mathbf{x}_2|\mathbf{x}_{1j}^{(t)})$$

along with $\hat{w} = h(\mathbf{x}_1|\mathbf{x}_2)\hat{h}(\mathbf{x}_2)/h\mathbf{x}_2|\mathbf{x}_1)\hat{h}(\mathbf{x}_1)$. Under convergence the function $\hat{w}$ should be approximately the constant function 1; the set $\hat{w}(\mathbf{x}_j^{(t)})$ should cluster tightly around 1.

The preceding discussion would appear to imply implementation of the Markov chain Monte Carlo algorithm through parallel strings since samples from, for example, $h^{(t)}$ are presumed. In fact, use of output from a single string can be equally well justified. Suppose we denote a post burn-in sample, possibly with spacing, from this string by $\mathbf{x}_j^*$, $j = 1, \ldots, m$. Since, under convergence, the $\mathbf{x}_j^*$'s are identically distributed, we might, for example, replace (1) by $\hat{h}(\mathbf{x}) = m^{-1} \cdot \sum_{j=1}^{m} h(\mathbf{x}|\mathbf{x}_j^*)$. Of course the (unknown) dependence among the $\mathbf{x}_j$'s muddies assessment of the precision of $\hat{h}$.

In summary, the above diagnostics will be suitable for use when p is not large, conjugacies are present or f is not pathological. Further discussion of the use of $\hat{w}$ as both a convergence diagnostic and as a convergence accelerator appears in Lee and Gelfand (1992). Lastly, considerable care is required in choosing, implementing and drawing inference from a simulation mechanism. However the reward of accommodating uncompromised modeling in many instances makes the effort worthwhile.