Comment: Extracting More Diagnostic Information from a Single Run Using Cusum Path Plot

Bin Yu

The article by Besag, Green, Higdon and Mengersen adds to a series of recent papers (Besag and Green, 1993; Geyer and Thompson, 1992; and Gelman and Rubin, 1992b) in making Markov chain Monte Carlo (MCMC) methods accessible to more statisticians, especially applied statisticians. I am glad to see that different algorithms are reviewed in a unified way and many examples are given. Although the article gives general recommendations as to which algorithms and sampling scans to choose, there is not much discussion on the empirical monitoring of convergence of the Markov chains. Since the convergence issue is very critical to the success of MCMC methods, and something close to my heart, I will make this issue my topic here. In particular, using the prostate cancer example in the article by Besag, Green, Higdon and Mengersen and the Ising model example in Gelman and Rubin (1992a), I illustrate that the cusum path plot in Yu and Mykland (1994) can effectively bring out the local mixing property of the Markov chain.

It had been believed by many MCMC researchers (including this author) that information solely from a single run of a Markov chain can be misleading since, for example, it can get trapped at a local mode of the target density. Consequently, additional information beyond that from a single run has been introduced to the convergence diagnostics. Gelman and Rubin (1992b) proposed a multiple chain approach in the MCMC context, followed by Liu, Liu and Rubin (1992) and Roberts (1992). Yu (1994) introduced additional information to a single run by taking advantage of the unnormalized target density. In the context of Gibbs samplers, Ritter and Tanner (1992) and Cui, Tanner, Sinhua and Hall (1992) suggested diagnostic statistics based on importance weights, using either multiple chains or a single chain. A priori bounds on the convergence rate can be found in Rosenthal (1993) and Mengersen and Tweedie (1993), but unfortunately these theoretical bounds are currently known only in some very special cases. For other references on existing diagnostic tools, see the recent and thorough review by Cowles (1994).

On the other hand, Yu and Mykland (1994) suggest that more information can be extracted from a single run than previously believed. The device is the cusum path plot, which brings out the local mixing behavior of the Markov chain in the direction of a chosen one-dimensional summary statistic, more effectively than the sequential plot. The cases where the cusum path plot works well are those where the mixing behavior is homogeneous across the sample space. For example, in some multimodal examples, the reason that the chain gets trapped at a local mode is because the chain moves around very slowly, even within one mode, and the cusum path plot brings out this local mixing speed even when the sampler is trapped at one mode. As shown below, the Ising model example of Gelman and Rubin (1992a) has a slow local mixing property. One situation in which the cusum path plot fails is a variant on the witch's hat (cf. Cui, Tanner, Sinhua and Hall, 1992; Yu and Mykland, 1994), where the chain has a split mixing behavior: fast in one region and slow in another.

Now we introduce the cusum path plot formally. Let $X_0, X_1, \ldots, X_n$ be a single run of a Markov chain, and let $T(X)$ the chosen one-dimensional summary statistic. Let n_0 be the "burn-in" time, and we construct our cusum statistics based on $T(X_{n_0+1}), \ldots, T(X_n)$ to avoid the initial bias of the chain. What we get out of the cusum plot is the more detailed information we cannot see in the sequential plot of $T(X)$ which MCMC users have been plotting all along.

Denote the observed cusum or partial sum as

$$\hat{S}_t := \sum_{j=n_0+1}^{t} \left[T(X_j) - \hat{\mu}\right] \quad \text{for } t = n_0 + 1, \ldots, n,$$

where

$$\hat{\mu} := \frac{1}{n - n_0} \sum_{j=n_0+1}^{n} T(X_j).$$

Bin Yu is Assistant Professor, Department of Statistics, University of California, Berkeley, California 94720-3860.

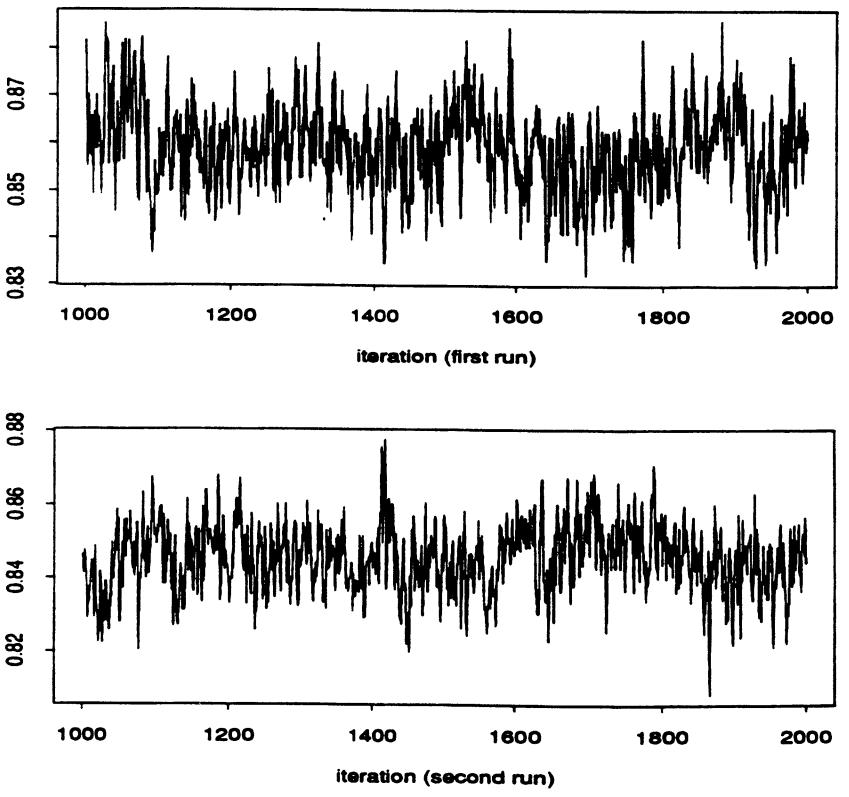

FIG. 1. *Ising model: sequential plots for two runs.*

Cusum path plot: Plot $\{\hat{S}_t\}$ against t for $t = n_0 + 1, \ldots, n$ and connect the successive points with line segments. Since $\Sigma_t \hat{S}_t = 0$, the cusum path plot ends at 0.

The mixing speed of $T(X)$ is reflected in the smoothness of the cusum plot path, that is, the more "hairy" the cusum path is, the faster the mixing speed of $T(X)$; the smoother the cusum path, the slower the mixing speed of $T(X)$. Moreover, the bigger the excursion the cusum path plot takes, the slower the mixing speed. See Yu and Mykland (1994) for the supporting arguments.

The cusum path plot should be compared to the "benchmark" cusum path plot, which is the cusum path plot of an iid sequence of normal random variables with their mean and variance matched

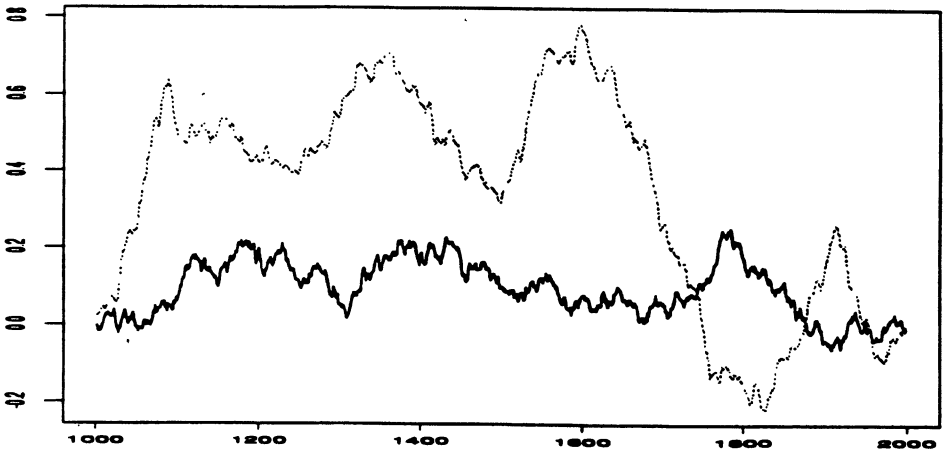

FIG. 2. *Ising model: first run (solid line, benchmark path; dotted line, Gibbs sampler path).*

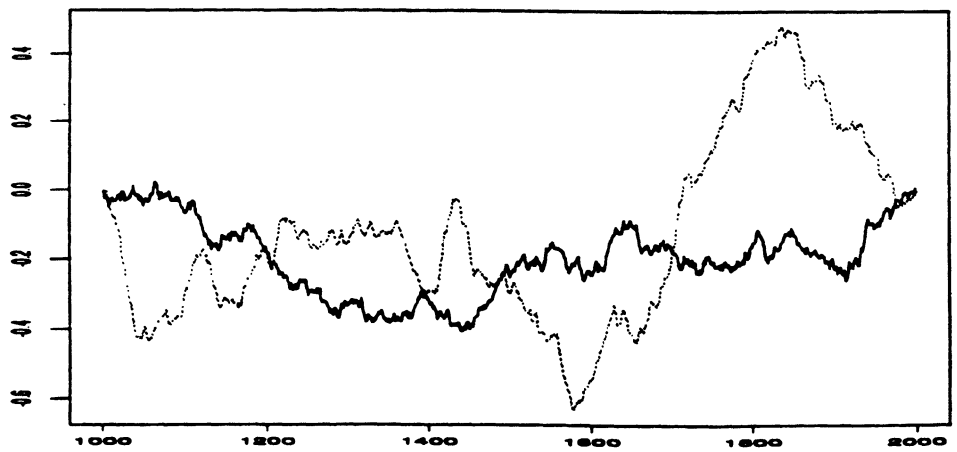

FIG. 3. *Ising model: second run (solid line, benchmark path; dotted line, Gibbs sampler path).*

with the estimated mean and variance of $\{T(X_j)$: $j = n_0 + 1, \ldots, n\}$; that is, for $t = n_0 + 1, \ldots, n$, let

$$\hat{S}_t^b := \sum_{j=n_0+1}^{t} \left[Y_j - \hat{\mu}_Y \right],$$

$$\text{where} \qquad \hat{\mu}_Y := (n - n_0)^{-1} \sum_{j=n_0+1}^{n} Y_j,$$

where $Y_{n_0+1}, \ldots, Y_n$ is an iid sequence of $N(\hat{\mu}_T, s_T^2)$ random variables with $\hat{\mu}_T$ as above and s_T^2 being the sample variance of $\{T(X_j)$: $j = n_0 + 1, \ldots, n\}$.

By the invariance principle for the partial sums of weakly dependent process (cf. Philipp and Stout, 1975), the benchmark path approximates, to the second order, the "ideal" cusum path of an iid sequence from the same target distribution. If the

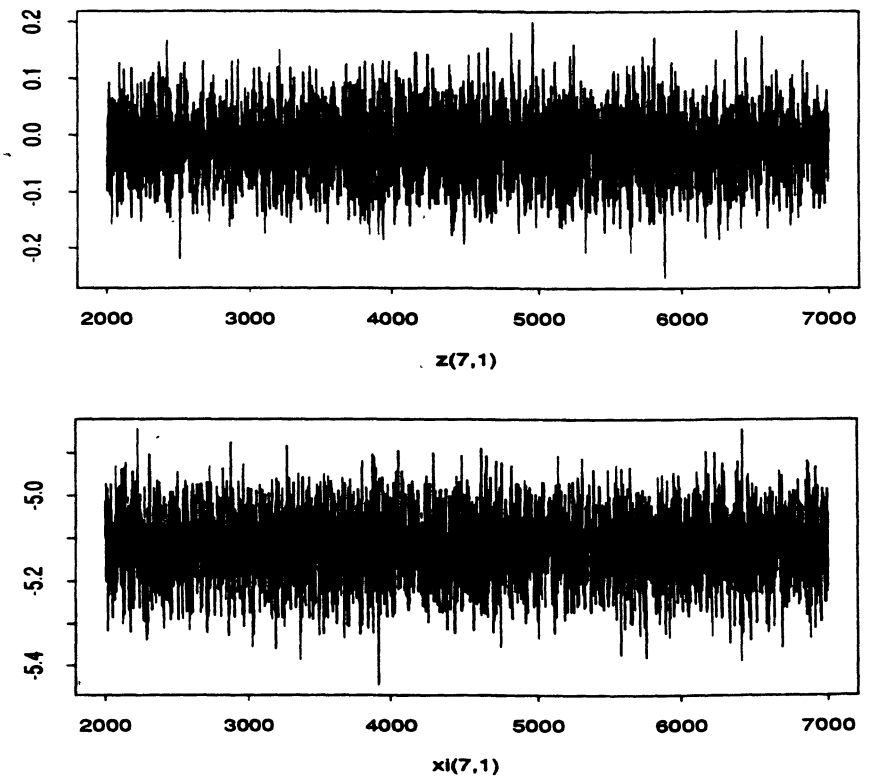

FIG. 4. *Prostate cancer example: 50-cycle gaps and block updates; sequential plots for* $z_{7,1}$ *and* $\xi_{7,1}$.

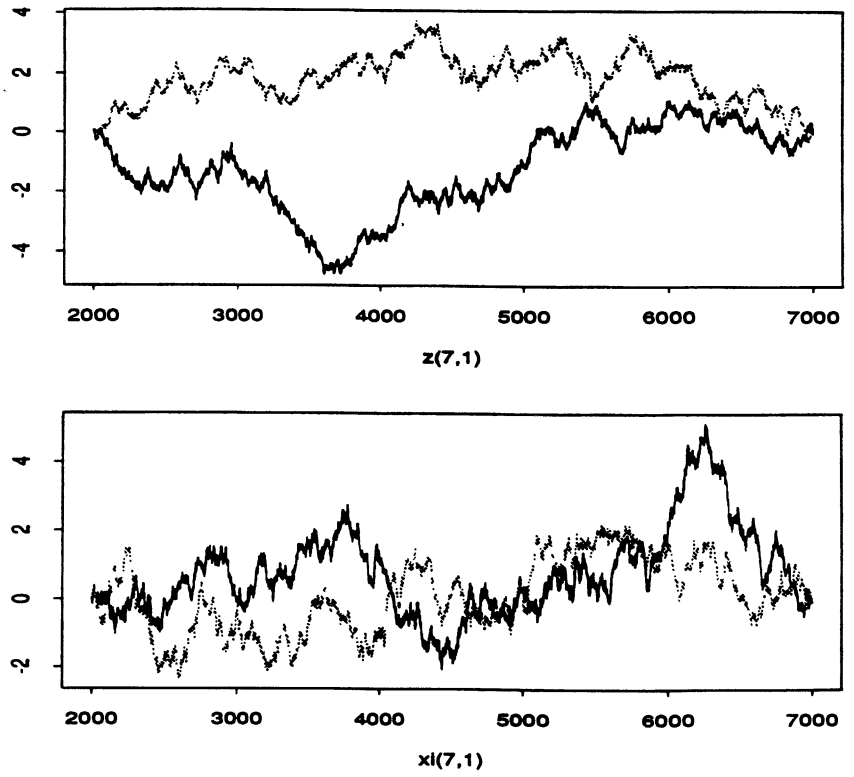

FIG. 5. *Prostate cancer example*: 50-*cycle gaps and block updates*; *cusum path plots for* $z_{7,1}$ *and* $\xi_{7,1}$ (*solid lines, benchmark paths*; *dotted lines, Gibbs sampler paths*).

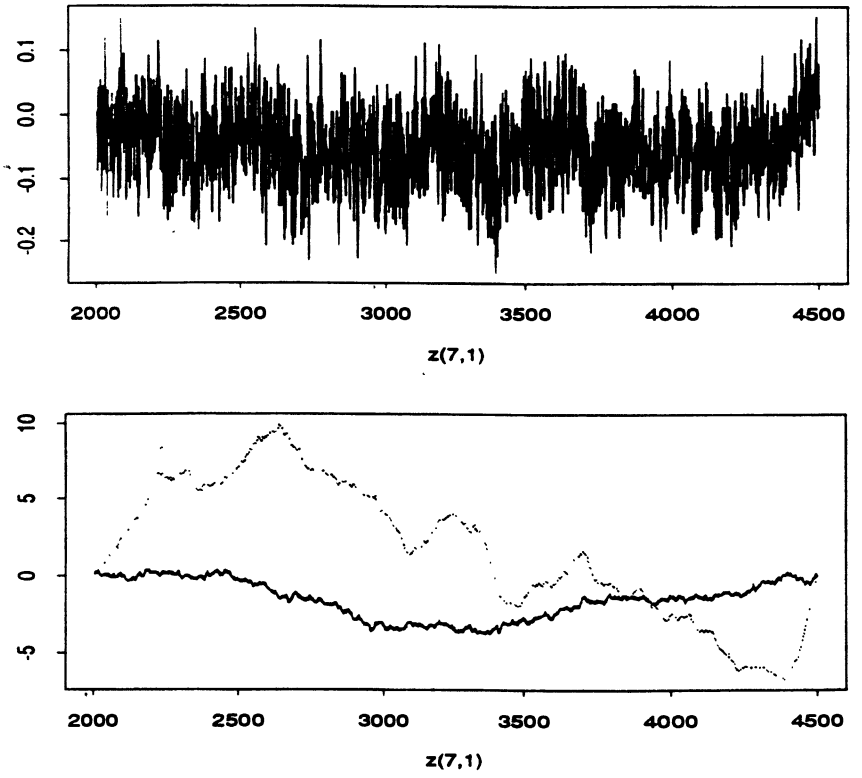

FIG. 6. *Prostate cancer example*: *equivalent model with* 10-*cycle gaps and single component updates*; *sequential and cusum path plots for* $z_{7,1}$. *In the cusum plot*: *solid line, benchmark path*; *dotted line, Gibbs sampler path.*

benchmark cusum path is comparable with the T cusum path in terms of smoothness of the path and size of the excursion, then we conclude that the sampler is mixing well [in the direction specified by $T(X)$, to be precise]. Otherwise, we conclude that the sampler is not mixing well, in the direction specified by $T(X)$. When two Markov chains are compared for the same target distribution, one may omit the "benchmark" cusum path plot.

Now we are ready to illustrate the use of the cusum path plot in the Ising model example in Gelman and Rubin (1992a) and in the prostate cancer example from the article by Besag, Green, Higdon and Mengersen. Note that we know that the mixing speed is slow in the Ising example, and Besag, Green, Higdon and Mengersen have concluded that there seems no significant multimodality problem in the prostate cancer example.

For the Ising model, professor Andrew Gelman kindly provided the two runs which appeared in Gelman and Rubin (1992a). For $n_0 = 1{,}000$ and $n = 2{,}000$, the sequential and cusum path plots are in Figures 1–3. Each of the cusum plots shows clearly that the mixing is slow, while each of the sequential plots suggests that things have stabilized.

For the prostate cancer example, the authors kindly offered the simulation data presented in their paper. For $n_0 = 2{,}000$ and $n = 7{,}000$, we monitored the 49 log-odds ratios ξ_{ij} and the corresponding reconstructed z_{ij}. The cusum path plots for all 98 parameters compare well with the benchmark plots, indicating good mixing behaviors, consistent with the claims of Besag, Green, Higdon and Mengersen. In this note, I include only the sequential and cumsum plots for two of them: $\xi_{7,1}$ and $z_{7,1}$ (Figures 4 and 5). The cusum plots display comparable paths of the data and the benchmark paths, in terms of smoothness and exclusion size. As the authors note in Section 4.2, fast mixing arises because of the block updates and a large sampling interval or gap. Note that, since the θ's, ϕ's and ψ's are themselves unidentifiable, it would be necessary to monitor them via appropriate contrasts. It is interesting to point out the effect on the cusum plots when single component updates are used and in addition the sampling interval is reduced from 50 to 10. Figure 6 shows the results for a burn-in of 20,000 cycles and data collection over a further 25,000 cycles. It is clear that the cusum plots bring out the mixing properties more explicitly than the sequential plots, and in order to obtain valid inference based on MCMC methods, extreme care is needed with convergence diagnostics.

In conclusion, MCMC users have to explore sufficiently the convergence issue before trusting the estimates that the Markov chain gives. Among other diagnostic tools such as sequential plot and autocorrelation plot, the cusum path plot is a simple and an effective device to monitor the local mixing speed of a Markov chain.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant DMS-93-22817 and Grant DAAH04-94-G-0232 from the Army Research Office.

Rejoinder

Julian Besag, Peter Green, David Higdon and Kerrie Mengersen

We thank the discussants for their contributions and insights, and for raising numerous interesting points. We shall respond to these as best we can, although obviously there are many questions for which, as yet, only partial solutions exist. We shall also try to rectify some misunderstandings that have arisen as a result of possible ambiguities in the paper. Our response is organized primarily by topic, rather than by discussant.

"ON BEING BAYESIAN"

Separation of Concerns

We have pondered Geyer's call for a separation of concerns, particularly between philosophy and computational technology, and we agree that the aim is an attractive one, but have come to a different conclusion, because in this case there are interactions that are too strong to be discounted. For example, the agricultural experiment in Section 5 of the paper is concerned with ranking and selection in comparing 75 varieties of spring barley. We contend that here it is a point of philosophy that the Bayesian paradigm provides an approach that is more useful than (indeed, we would say vastly superior to) any non-Bayesian approach. However, even in quite straightforward formulations, it is exceedingly difficult to implement a fully Bayesian analysis without MCMC. The simultaneous credible regions in the paper provide another example,