monitoring techniques, which, as pointed out by Dr. Geyer, have to be theoretically based. However, I remain quite worried after reading the two papers. There is no guarantee of the properties of the various estimates of the Monte Carlo variance. They just appear to work most the time. The apparent convergence of multiseries also offers no guarantee for convergence. The difficulties one faces in finding initial values remain quite open. Methods and guidelines for reparameterization to improve the mixing of the chain are still lacking. It looks like it would take some time and effort before one can automate sampling methods for use by other scientists.

Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo

Adrian E. Raftery and Steven M. Lewis

1. SUMMARY

We congratulate Andrew Gelman, Don Rubin and Charlie Geyer on a pair of articles that together summarize many of the important issues in the implementation of Markov chain Monte Carlo (MCMC) algorithms. They both make important and valid points. We do not agree fully with the recommendations in either article, however. We recommend that inference ultimately be based on a single long run, but that this be monitored using carefully chosen diagnostics, and that starting values and the exact form of the algorithm be chosen on the basis of experimentation. More complex and expensive methods such as those of Gelman and Rubin seem rarely to be necessary in standard statistical models.

Theory suggests that Markov chain Monte Carlo (MCMC) inference be based on a single long run. Gelman and Rubin, by contrast, argue that the uncertainty associated with the choice of starting value should be taken into account by using several runs with different starting values. However, this uncertainty seems to be small in most statistical problems, given a realistically large number of MCMC iterations.

Nevertheless, a bad starting value *can* lead to slow convergence. This can be diagnosed from one run and rectified by changing the starting value. Diagnostics should monitor *all* the key features of the model, such as hyperparameters in hierarchical models, as well as a selection of less essential features such as random effects. If only the quantities of interest are monitored, lack of convergence can be missed.

By the same token, Geyer's time-series variance estimation methods can give misleading results in the absence of diagnostics. There seems to be no reason to abandon standard spectral analysis methods in favor of Geyer's initial sequence estimators. Many Bayesian statistical problems boil down to the calculation of quantiles of marginal posterior distributions of quantities of interest, and then there are simpler methods that do not have the problem of sensitivity to a spectral window width. Methods based on quantiles also yield simple and effective diagnostics.

2. MULTISTART OR ONE LONG RUN?

Gelman and Rubin advocate multistart and describe a way of choosing the starting values that uses some combination of numerical optimization, EM, iterative ECM, numerical second derivatives, importance resampling and simulation from a mixture of multivariate t-distributions – all before even starting the MCMC algorithm proper. Is this feasible? And is it really necessary? The main argument for multistart is that $BV/(BV + WV)$ can be large, where BV is the between-run component of the variance of the estimate of a functional of the posterior distribution and WV is the within-run component. In our observation, this is rarely the case for standard statistical models with a realistically large number of MCMC iterations, and we would like to see at least one convincing example. As Geyer shows, a single long run works well in Gelman and Rubin's own example. (We use the term "standard statistical model" loosely but broadly; it includes, at

Adrian E. Raftery is Professor of Statistics and Sociology and Steven M. Lewis is a Ph.D. candidate. Both at the Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.

least, heirarchical models where, at each level of the hierarchy, parameters or observations are (conditionally) independently distributed with standard distributions such as normal, t, gamma, Poisson or binomial. It certainly includes the Gelman and Rubin example. It probably excludes some models with strong dependence that arise in spatial statistics, expert systems and genetic pedigree analysis.)

One long run is *not* always good enough, however. A poorly chosen starting value may well lead to slow convergence, and diagnostics based on a single run usually show this clearly (see Section 4). Changing the starting value can solve this; in our experience, simple trial and error has worked fine. One example is given in Section 4. Another example is the Ising model, analyzed by Gelman and Rubin (1992). There a poor starting value leads to slow convergence, but diagnostics for a single run based on Raftery and Lewis (1992) show this clearly, and the problem is easily solved by taking another starting value. Thus *restart* seems to work as well as *multistart* and is much easier to implement.

Many complex statistical models can be analyzed using MCMC if they are recast as hierarchical models, and there the starting value problem can be acute. If the starting value for the random effects variance or equivalent parameter (σ_a^2 in the Gelman and Rubin example) is close to zero, then componentwise MCMC (such as the Gibbs sampler) can get stuck for a long time close to the starting value. Single-run diagnostics for the random effects variance will reveal this immediately, but the series for other quantities such as the random effects themselves (a_i in the Gelman and Rubin example) can be almost uncorrelated and give no hint of trouble. Thus, diagnostics that look only at the quantities of primary interest may miss lack of convergence.

Since the starting value can have an important effect on the performance of a MCMC algorithm, there is certainly advantage to a systematic search for good starting values, and the Gelman and Rubin method achieves this for many models. Our point is simply that the Gelman and Rubin method is demanding and may not even be feasible and that much simpler *ad hoc* methods based on single-run diagnostics seem in practice to work quite well. An insistence that the Gelman and Rubin methods be used would impose a big extra burden on users and discourage the use of MCMC; the evidence so far does not seem to justify such an insistence.

Some of the hardest situations are when a discrete distribution is being simulated using a nearly reducible Markov chain in areas such as genetic pedigree analysis (Sheehan and Thomas, 1992) and expert systems (Spiegelhalter, 1988). However, the Gelman and Rubin approach is designed for continuous distributions, and we find it hard to see how it would be applied, for example, to the genetics problem of Sheehan and Thomas (1992).

3. OUTPUT ANALYSIS AND DIAGNOSTICS

Geyer suggests time-series variance estimation methods. In the absence of diagnostics, his methods can dramatically *underestimate* the variance, in spite of the asymptotic result in his Theorem 3.2; see Section 4 for an example. The variance is equal to the spectrum at zero, and Geyer uses the truncated periodogram spectral estimator. This is generally agreed to be a poor spectral estimator (Priestley, 1981, Section 7.5), and it can lead the variance estimate to be quite sensitive to the choice of window "width"; see, for example, Section 4.

One good spectral estimator is based on the Tukey-Hanning lag window, which leads to $w_n(t) = (1/2) \cdot \{(1 + \cos(\pi t/K)\}$ for $|t| \leq K$ and 0 otherwise in Geyer's Equation (3.2). One may choose K so that $\hat{\gamma}_{n,t} \sim 0$ for $|t| > K$ (Priestley, 1981, p. 539). While this sounds a little vague, the estimated variance seems to be fairly insensitive to the precise value used.

It can be argued that many Bayesian estimation problems reduce to finding posterior quantiles of quantiles of interest. In that case, the output analysis problem is simpler because we have to deal only with binary sequences. Raftery and Lewis (1992) proposed a way of finding the number of iterations needed to achieve a given precision in this case, based on an initial run. (A Fortran program to implement this method can be obtained by sending an e-mail message to *statlib@stat.cmu.edu* consisting of the single line "send gibbsit from general.") This avoids the window width selection problem. The method finds the number of iterations needed to estimate $P[U \leq u \mid \text{data}]$ to within $\pm r$ with probability s when the correct answer is q, where U is a quantity of interest. It returns the number, M, of initial iterations to be discarded, the number, N, of additional iterations required, and k, where every kth iterate is used.

This method also yields diagnostics. One can determine in advance the minimum number of iterations needed, $N_{\min}$, and so $I = (M + N)/N_{\min}$ measures the increase in the number of iterations due to dependence in the sequence. Values of I much greater than 1 indicate a high level of dependence. Our limited experience to date suggests that values of I greater than about 5 often indicate problems that can be alleviated by changing the implementation. Such dependence can be due to a bad starting value (in which case other starting values should be tried), to high posterior correlations (which can be remedied by crude correlation-removing transformations) or to "stickiness" in the Markov chain (sometimes removable by changing the MCMC algo-

rithm). It may seem surprising that a bad starting value can lead to high values of N as well as M. This happens because progress away from a bad starting value tends to be slow and gradual, leading to a highly autocorrelated sequence and high values of N.

The method should be applied to extreme (e.g., .025 and .975 or .05 and .95) posterior quantiles of quantities of interest and also to other key parameters such as hyperparameters in hierarchical models. The overall maximum value of N should be used if the starting value or the MCMC algrithm is not changed.

4. EXAMPLE

We illustrate these ideas with an example from the analysis of longitudinal World Fertility Survey data (Raftery, Lewis and Aghajanian, 1992; Lewis, 1992). The data are complete birth histories for about 5,000 Iranian women, and here we focus on the estimation of unobserved heterogeneity. Let π_{it} be the probability that woman i had a child in calendar year t. Then a simplified version of the model used is

$$\log(\pi_{it}/(1-\pi_{it})) = \beta + a_i, \qquad a_i \sim N(0, \sigma^2),\ a_i \text{ are iid.} \tag{1}$$

The prior on the hyperparameters is $p(\beta, \sigma^2) \propto \sigma^{-2}$. The a_i's are random effects representing unobserved sources of heterogeneity in fertility such as fecundability and coital frequency. There are also measured covariates in the model, but these are omitted here for ease of exposition.

Figure 1 shows a run of a MCMC algorithm starting with a value of σ^2 close to zero, namely $\sigma^2 = 10^{-5}$, and with values of the a_i's randomly generated from $N(0, 10^{-5})$. (In Figures 1 and 2, the starting value has been omitted for reasons of scaling, and β has been kept constant at a value estimated from the model without random effects.) The σ^2 series seems highly autocorrelated and the Raftery and Lewis (1992) method confirms this. With $q = 0.025$, $r = 0.0125$ and $s = 0.95$, we obtain $N = 4{,}178$, $k = 2$ and $M = 34$. Here $N_{\min} = 600$, so that $I = 7.0$. The high value of I and the trendlike appearance of Figure 1a suggest that there is a starting value problem. By contrast, the values of a_{3911} *in the same run* are much less correlated (Figure 1b) with $I = 2.4$, so that diagnostics based on that series alone would mislead.

Figure 2 shows three other series of σ^2 from different starting values, illustrating a simple trial-and-error approach to the choice of an adequate starting value.

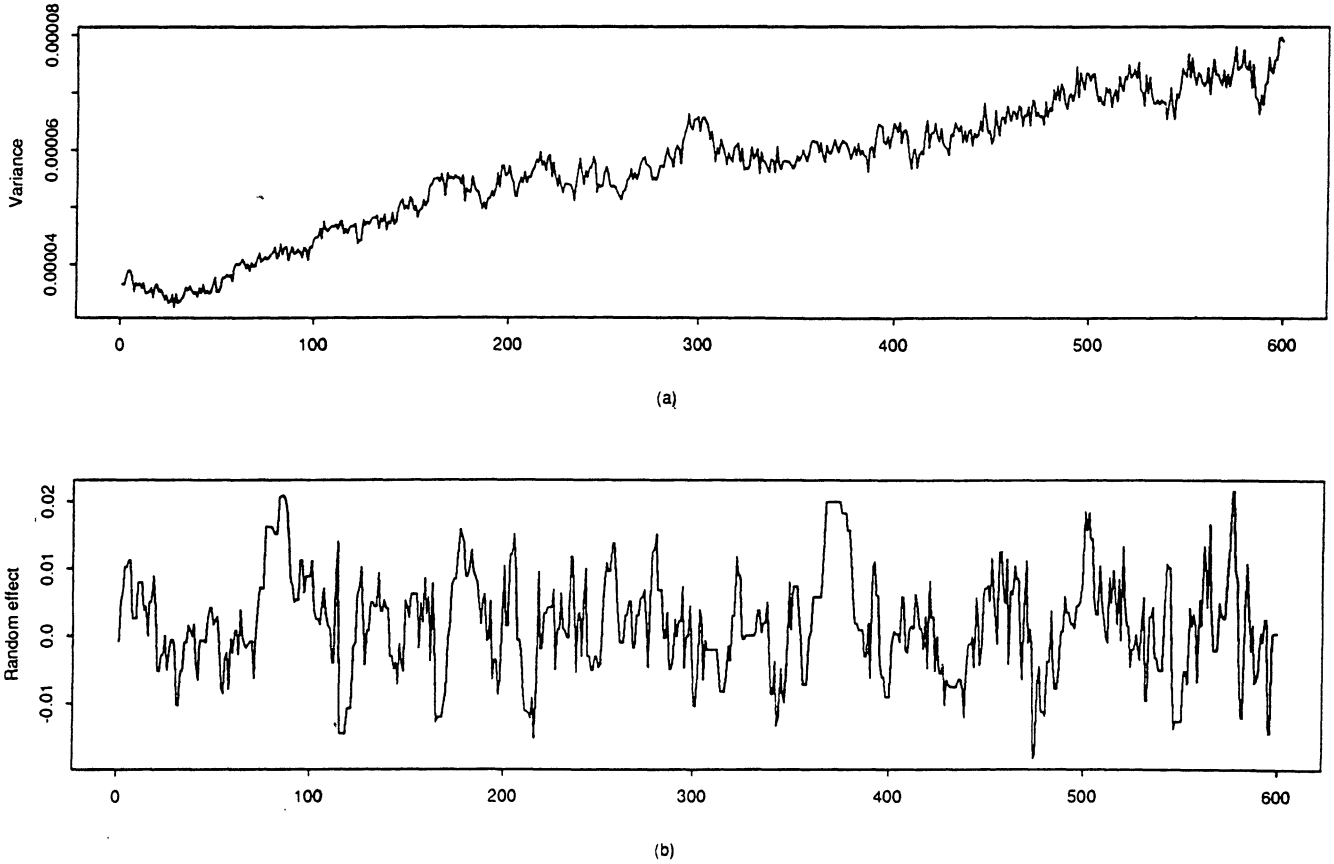

FIG. 1. *MCMC output for the model in* (1) *for the Iranian World Fertility Survey data starting with* $\sigma^2 = 10^{-5}$: (*a*) *series of* σ^2 *values;* (*b*) *series of values of* a_{3911}, *the random effect for woman* 3911 *in the survey.*

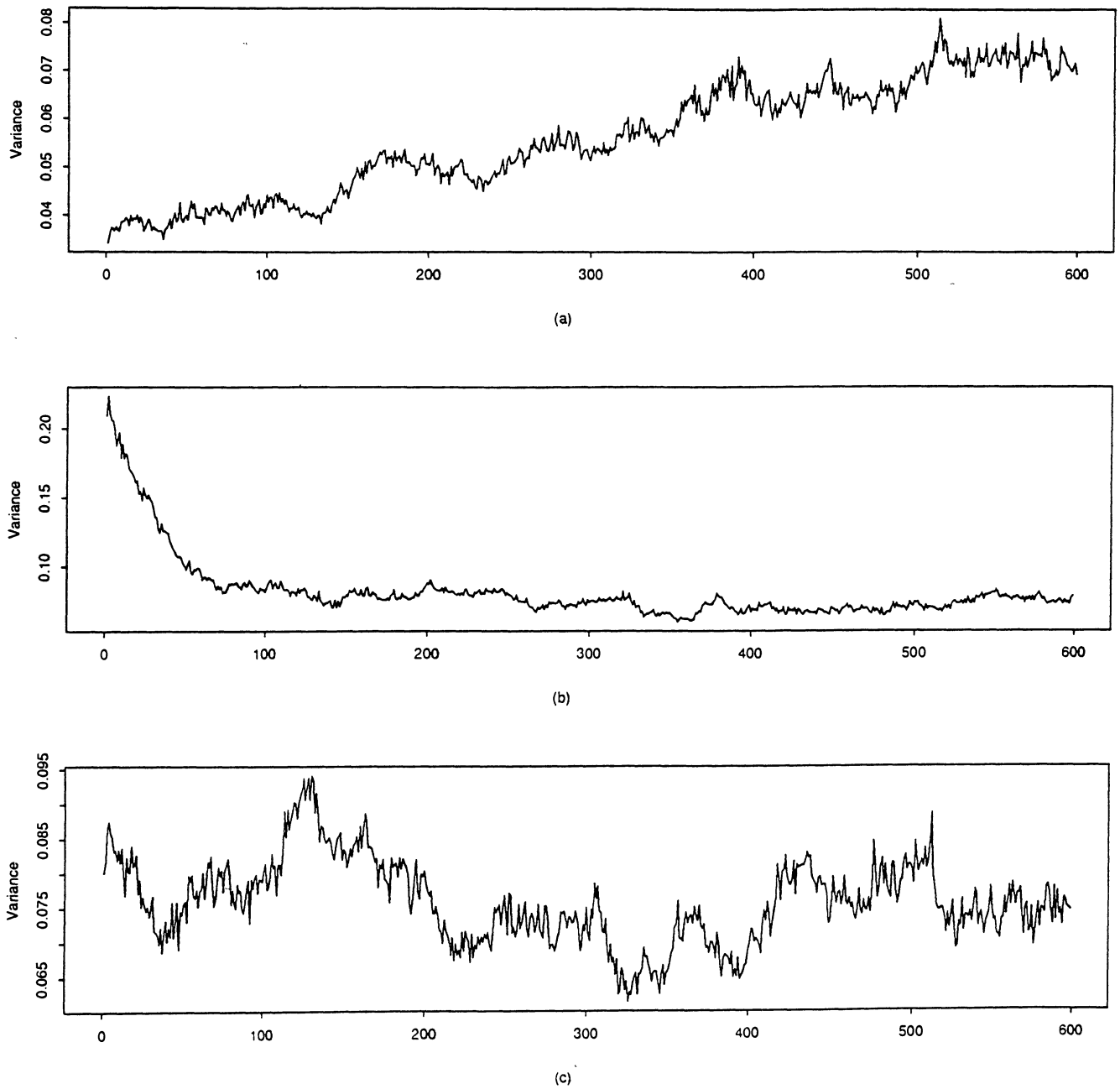

FIG. 2. *Values of σ^2 for three runs of the same MCMC algorithm as in Figure 1, with different starting values: (a) $\sigma^2 = 0.1$; (b) $\sigma^2 = 1$; (c) $\sigma^2 = 0.25$.*

Figure 2a starts with $\sigma^2 = 0.1$, and the method of Raftery and Lewis (1992) yields $I = 5.5$, which is still unsatisfactory. The plot suggests that the starting value is too low. Figure 2b starts with $\sigma^2 = 1.0$ and has $I = 5.6$; clearly the starting value is now too high. Figure 2c starts with $\sigma^2 = 0.25$ and has $I = 2.1$; the results of this trajectory all seem quite satisfactory.

This example also sheds some light on time-series variance estimation. Figure 3 shows the autocorrelation functions of the σ^2 series in Figure 1a (the bad starting value) and Figure 2c (the "good" starting value). For Figure 1a, all the initial sequence variance estimates are about 2.2×10^{-8}, while the Tukey-Hanning estimate with $K = 43$ is 0.5×10^{-8}. One would expect the series in Figure 1a eventually to converge to something like Figure 2c. For the latter, the initial positive sequence estimate is 5.9×10^{-3}, the initial monotone sequence estimate is 3.8×10^{-3} and the Tukey-Hanning estimate with $K = 27$ is 1.9×10^{-3}.

Thus, variance estimation based on Figure 1a would underestimate the variance by a factor of over 10,000! This shows the dangers of relying on stationary time-series theory without diagnostics. Of course, diagnos-

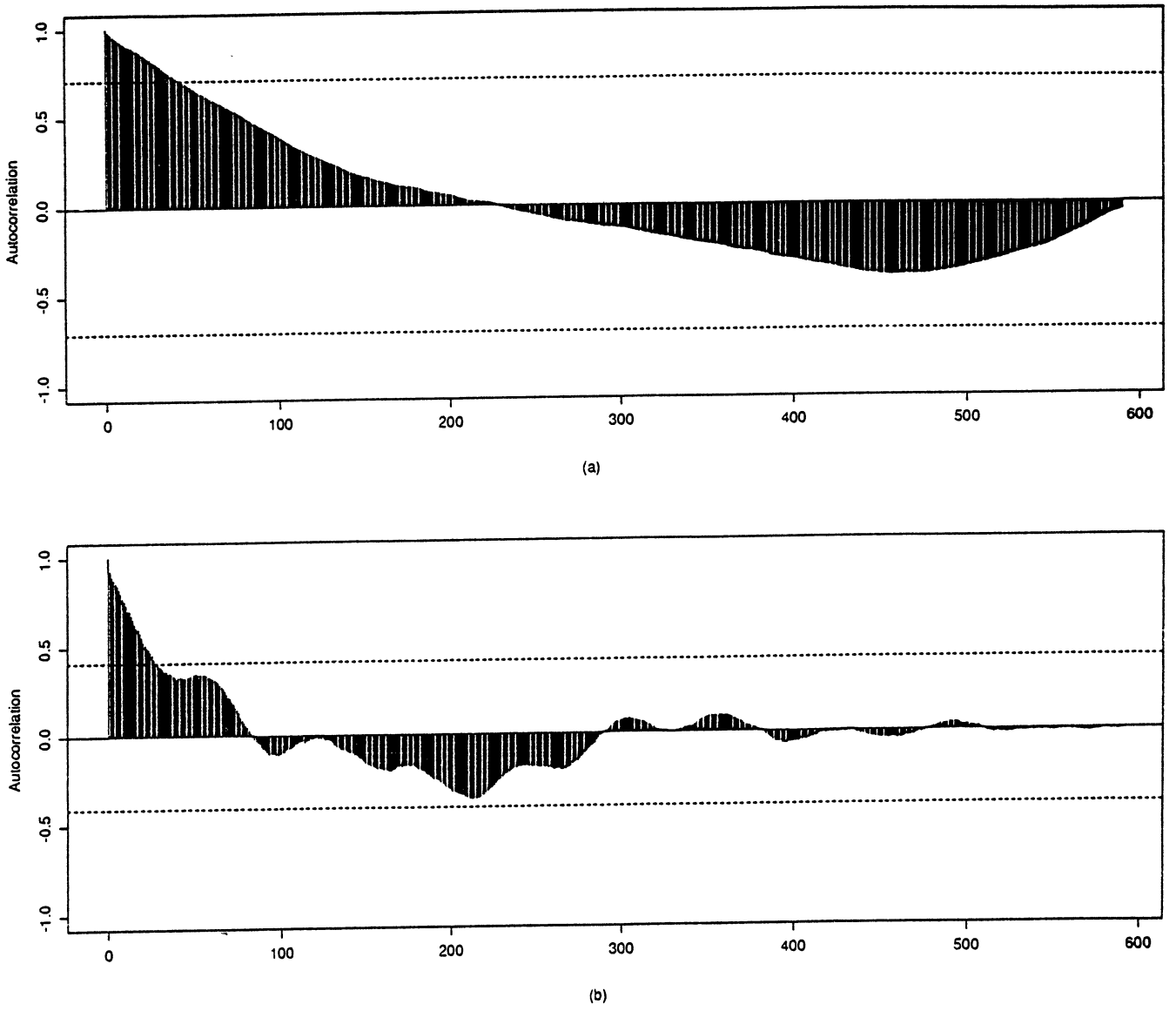

FIG. 3. *Autocorrelation functions for the σ^2 series starting at (a) 10^{-5} (Figure 1a); (b) 0.25 (Figure 2c).*

tics would immediately show that a bad starting value had been chosen. Also, the initial sequence estimators can be quite different from one another; this is because the truncated periodogram variance estimator is quite sensitive to the choice of window width. There seems to be no need to abandon the standard methods of spectral analysis, which would indicate using the Tukey-Hanning window or something similar. In this example, the initial positive sequence estimator was bigger than the Tukey-Hanning estimator by a factor of at least three.

This example bears out our main points. It is important to monitor the MCMC run for all the key parameters and to start again with different starting values when the diagnostics suggest doing this. We have not yet come across examples that convince us that complicated ways of finding starting values are necessary; simple trial and error has worked fine in almost all the cases that we are aware of, and in the cases where it has not (e.g., Sheehan and Thomas, 1992), the Gelman and Rubin method seems unlikely to help either, and special methods must be devised. For most statistical problems there seems to be little need to take account formally of uncertainty about the starting value. Output analysis methods based on quantiles yield diagnostics that seem to work well. In the absence of diagnostics, time-series variance estimation methods can be quite misleading.

ACKNOWLEDGMENTS

This research was supported by NIH Grant 5R01HD26330-02 and by ONR Contract N00014-91-J-1074.