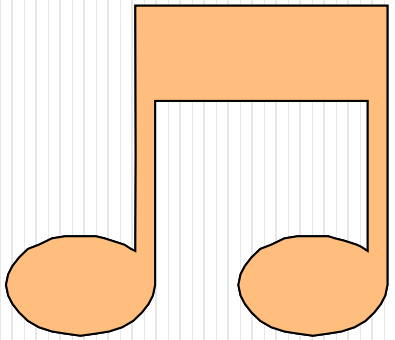
MCMC Convergence and Convergence Diagnostic

14 数理统计
张楠 2014210745
乌云
2013211749

Inference and assessing convergence

Use the collection of all the simulated draws from $p(\theta|y)$ to summarize the posterior density and to compute quantiles, moments, and other summaries of interest as needed.

Posterior predictive simulations of unobserved outcomes can be obtained by simulation conditional on the drawn values of θ.



Difficulties of inference from iterative simulation.

First, if the iterations have not proceeded long enough, the simulations may be grossly unrepresentative of the target distribution.

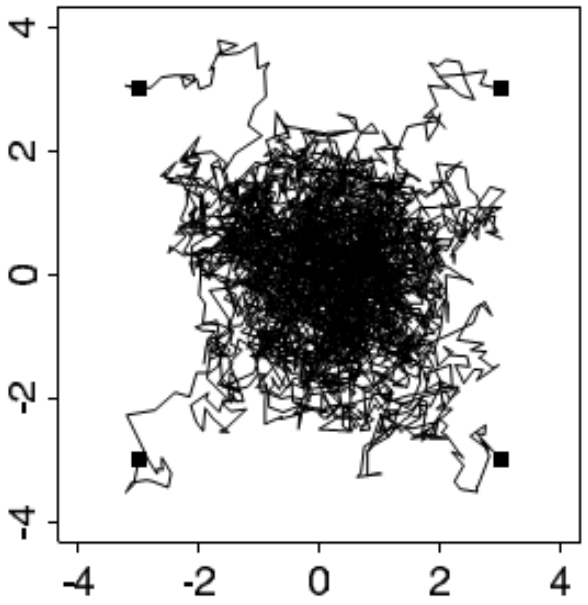


Figure 1b

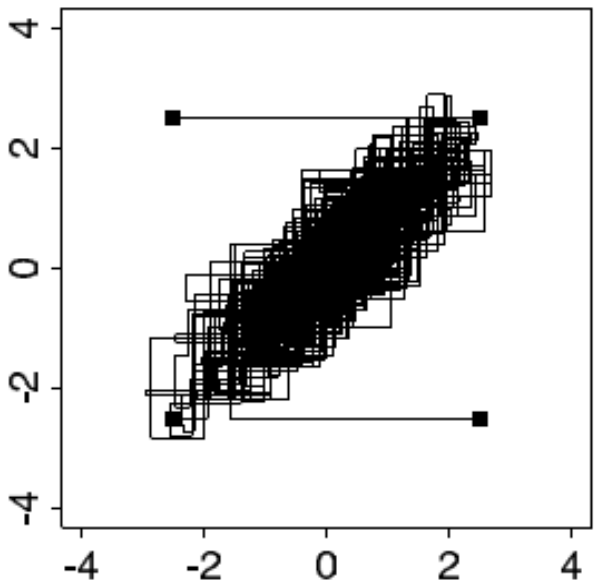
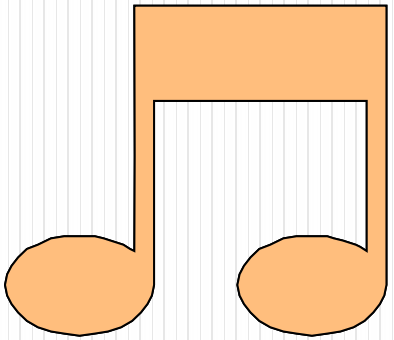


Figure 2b

Even when simulations have reached approximate convergence, early iterations still reflect the starting approximation rather than the target distribution



Difficulties of inference from iterative simulation.

The second problem with iterative simulation draws is their within–sequence correlation; aside from any convergence issues, simulation inference from correlated draws is generally less precise than from the same number of independent draws.

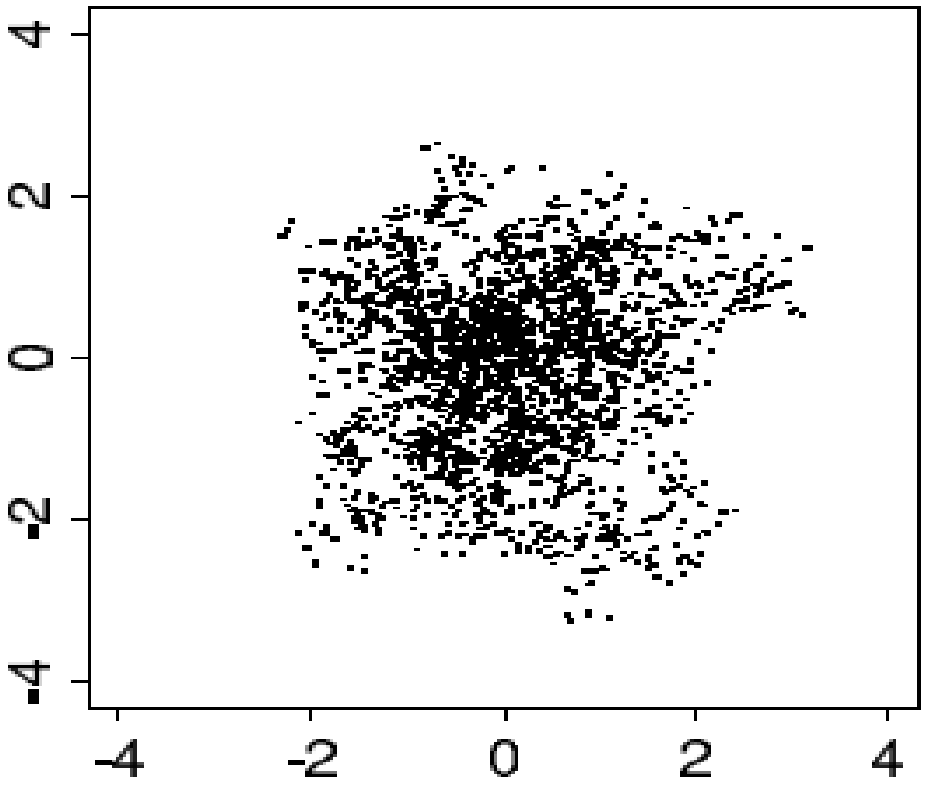
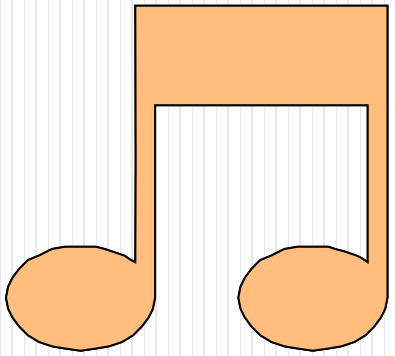


Figure 1c

Serial correlation in the simulations is not necessarily a problem because, at convergence, the draws are identically distributed as p(θ |y), and so when performing inferences, we ignore the order of the simulation draws in any case. But such correlation can cause inefficiencies in simulations.



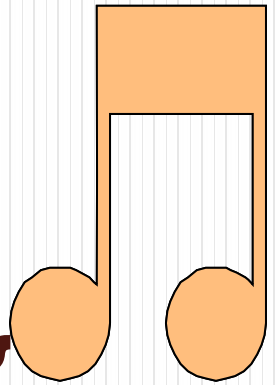
Difficulties of inference from iterative simulation

First, we attempt to design the simulation runs to allow effective monitoring of convergence, in particular by simulating multiple sequences with starting points dispersed throughout parameter space

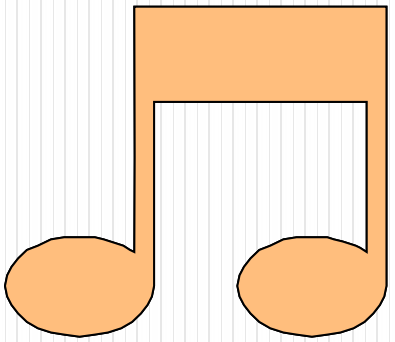
Second, we monitor the convergence of all quantities of interest by comparing variation between and within simulated sequences until 'within' variation roughly equals 'between' variation.

Third, if the simulation efficiency is unacceptably low (in the sense of requiring too much real time on the computer to obtain approximate convergence of posterior inferences for quantities of interest), the algorithm can be altered, as we discuss in Sections 12.1 and 12.2.

Discarding early iterations of the simulation runs

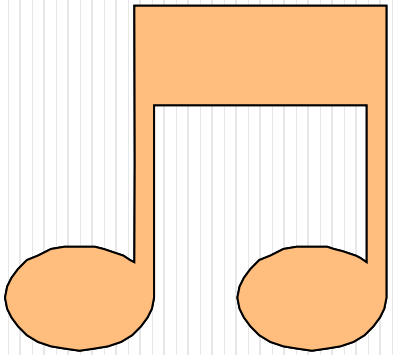


To diminish the influence of the starting values, we generally discard the first half of each sequence and focus attention on the second half. Our inferences will be based on the assumption that the distributions of the simulated values θ^t, for large enough t, are close to the target distribution, p(θ ly). We refer to the practice of discarding early iterations in Markov chain simulation as warm-up; depending on the context, different warm-up fractions can be appropriate



Discarding early iterations of the simulation runs

We adopt the general practice of discarding the first half as a conservative choice. For example, we might run 200 iterations and discard the first half. If approximate convergence has not yet been reached, we might then run another 200 iterations, now discarding all of the initial 200 iterations.

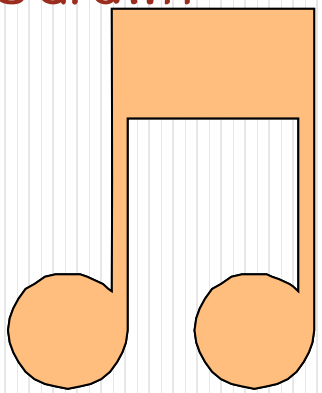


Dependence of the iterations in each sequence

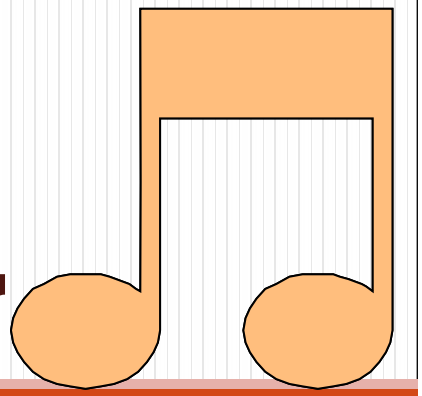
In our applications, we have found it useful to skip iterations in problems with large numbers of parameters where computer storage is a problem, perhaps setting k so that the total number of iterations saved is no more than 1000.

Multiple sequences with over-dispersed starting points

Our recommended approach to assessing convergence of iterative simulation is based on comparing different simulated sequences. To see such disparities, we clearly need more than one independent sequence. Thus our plan is to simulate independently at least two sequences, with starting points drawn from an over–dispersed distribution



Monitoring scalar estimands



We monitor each scalar estimand or other scalar quantities of interest separately. Estimands include all the parameters of interest in the model and any other quantities of interest. It is often useful also to monitor the value of the logarithm of the posterior density, which has probably already been computed if we are using a version of the Metropolis algorithm.

Thank you

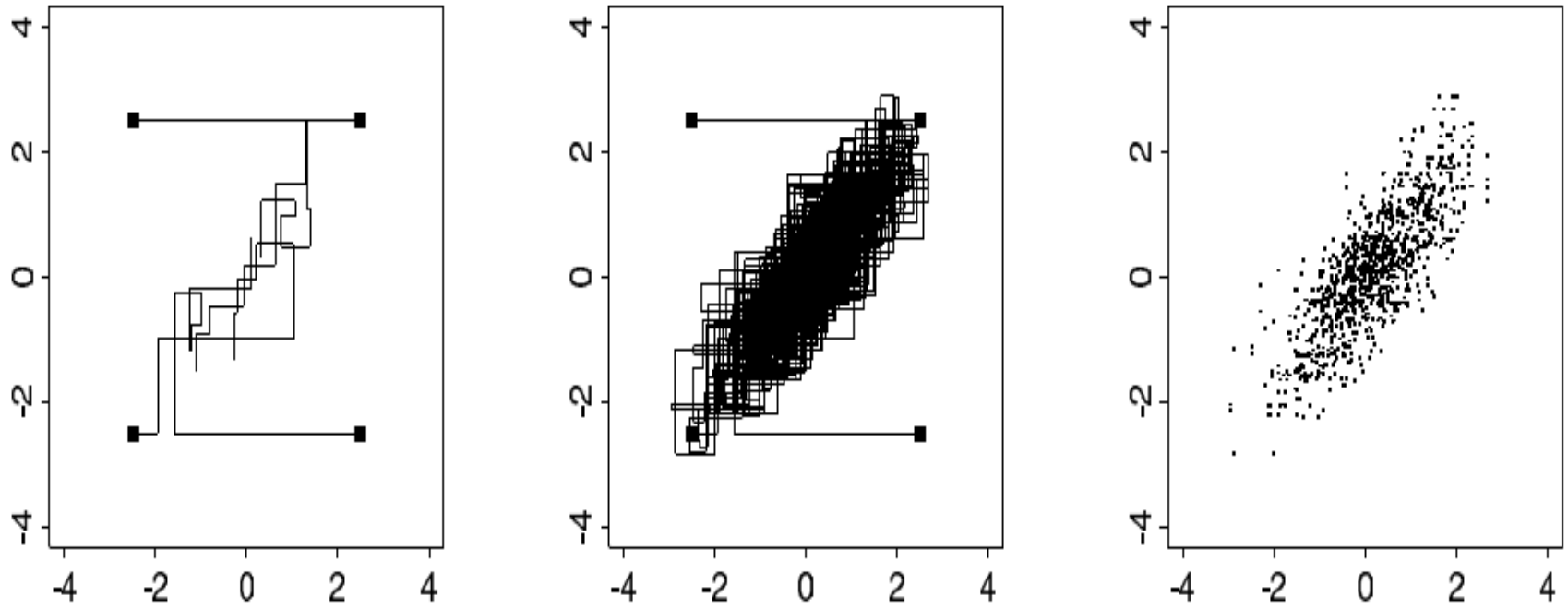
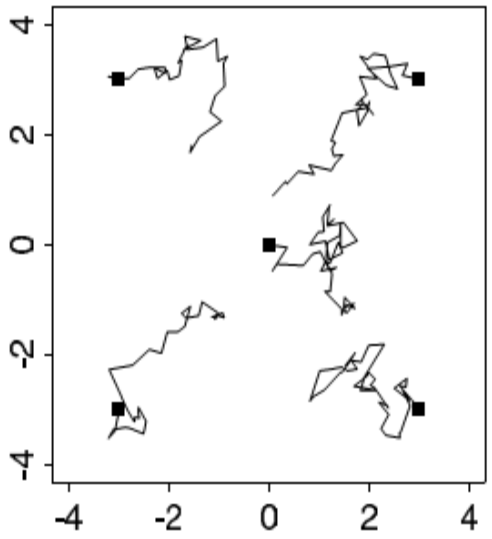
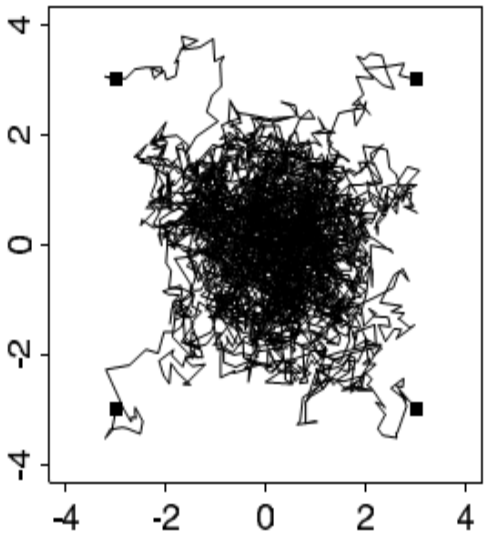


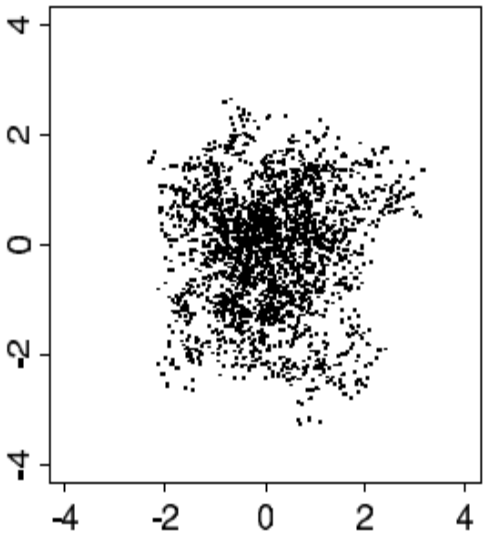
Figure 11.2



a



b



c

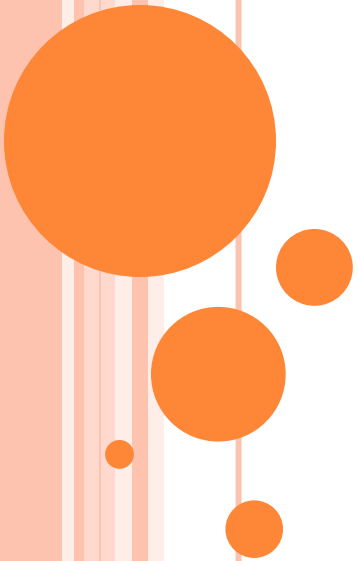
Figure 11.1

MONITORING CONVERGENCE

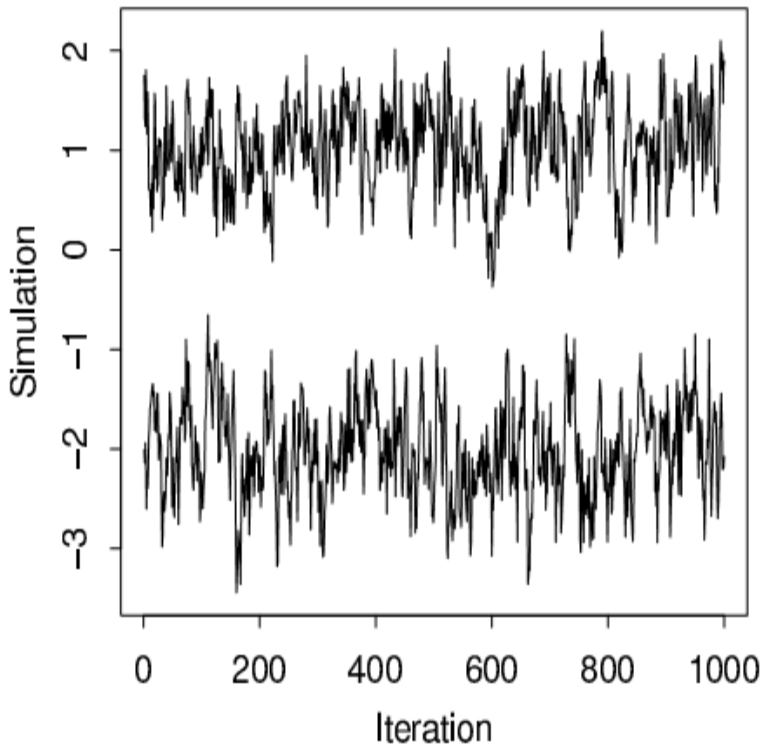
CONVERGENCE DIAGNOSTICS

ASSESSING MIXING

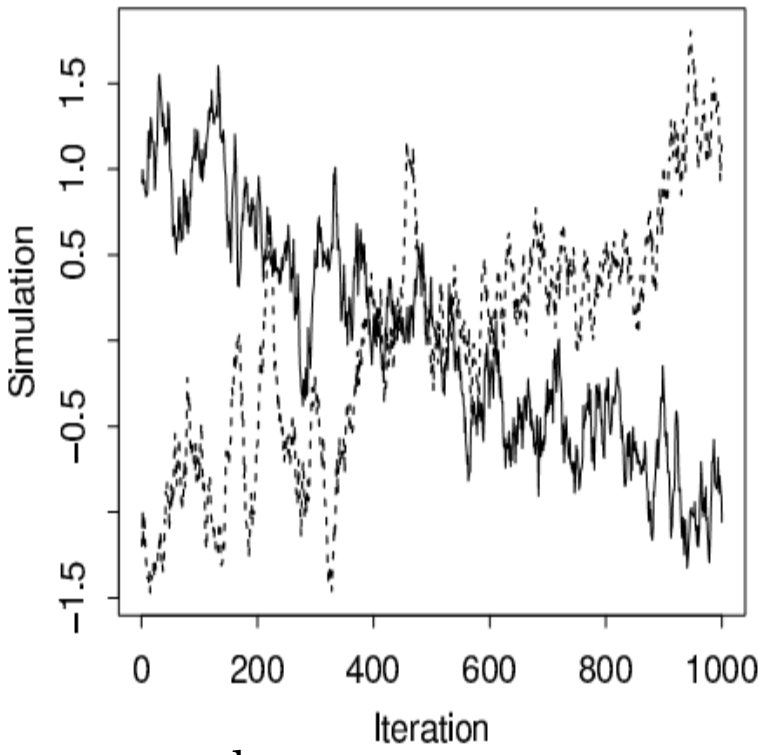
MONITORING CONVERGENCE



mixing and stationarity



a

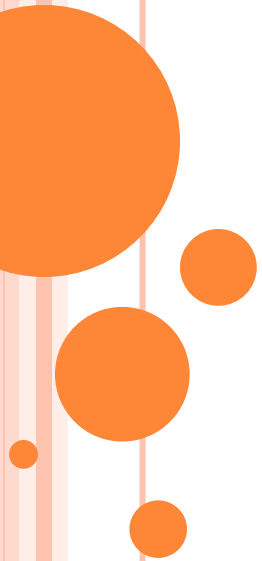


b

- The figure illustrates two of the challenges of monitoring convergence of iterative simulations. The first graph shows two sequences, each of which looks fine on its own and, indeed, when looked at separately would satisfy any reasonable convergence criterion, but when looked at together reveal a clear lack of convergence. The figure illustrates that, to achieve convergence, the sequences must together have mixed.

The second graph shows two chains that have mixed, in the sense that they have traced out a common distribution, but they do not appear to have converged. The second graph illustrates that, to achieve convergence, each individual sequence must reach stationarity.

Convergence diagnostics



Splitting each saved sequence into two parts

Splitting each saved sequence into two parts

We diagnose convergence by checking mixing and stationarity . There are various ways to do this ; we apply a fairly simple approach in which we split each chain in half and check that all the resulting half sequences have mixed. This simultaneously tests mixing and stationarity .

We start with some number of simulated sequences in which the warm-up period has already been discarded. We then take each of these chains and split into the first and second half.

Let m be the number of chains and n be the length of each chain. We always simulate at least two sequences so that we can observe mixing; thus m is always at least 4.

For example, suppose we simulate 5 chains, each of length 1000, and then discard the first half of each as warm-up. We are then left with 5 chains, each of length 500, and we split each into two parts: iterations 1–250 (originally iterations 501–750) and iterations 251–500 (originally iterations 751–1000). We now have m = 10 chains, each of length n = 250.

Assessing mixing

between- and within-sequence variances

Assessing mixing using between- and within-sequence variances

For each scalar estimand ψ, we label the simulations as ψij (i =1, . . . , n; j =1, . . . , m), and we compute B and W , the between- and within-sequence variances:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2, \text{where } \bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$$

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \; s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{ij} - \bar{\psi}_{.j}\right)^2,$$

The between-sequence variance, B, contains a factor of n because it is based on the variance of the within-sequence means, $\psi.j$, each of which is an average of n values ψij.

We can estimate var(ψ|y), the marginal posterior variance of the estimand, by a weighted average of W and B, namely

$$\widehat{var}^{+}(\psi/y)=\frac{n-1}{n}\mathrm{W}+\frac{1}{n}\mathrm{B}.$$

This quantity overestimates the marginal posterior variance assuming the starting distribution is appropriately overdispersed, but is unbiased under stationarity, or in the limit n → ∞. This is analogous to the classical variance estimate with cluster sampling.

Meanwhile, for any finite n, the ‘within’ variance W should be an underestimate of $\mathrm{var}(\psi/y)$ because the individual sequences have not had time to range over all of the target distribution and, as a result, will have less variability; in the limit as n → ∞, the expectation of W approaches $\mathbf{var}(\psi/y)$

We monitor convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for ψ might be reduced if the simulations were continued in the limit n → ∞. This potential scale reduction is estimated by

$$\hat{R}=\sqrt{\frac{\widehat{var}^{+}(\psi/y)}{W}},$$

which declines to 1 as $n \rightarrow \infty$. If the potential scale reduction is high, then we have reason to believe that proceeding with further simulations may improve our inference about the target distribution of the associated scalar estimand.

Thanks for listening!