ADVANCED BAYESIAN MODELING

USING MCMC IN PRACTICE:
ASSESSING CONVERGENCE

Recall: A Markov chain must be run until convergence to avoid initial transient effects that do not properly represent the posterior (the stationary distribution).

Diagnostics for assessing convergence:

- Trace plots of sampled nodes (parameters and predictive quantities)
- Convergence statistics, such as Gelman-Rubin

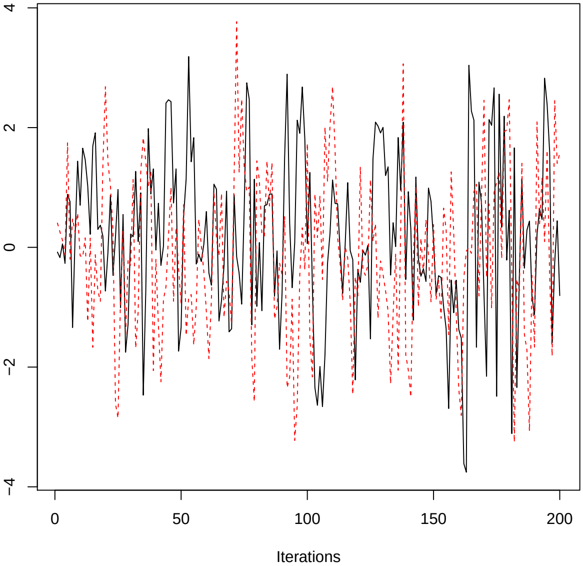
Trace Plots

A **trace plot** of a scalar sampled node (e.g., an element of θ) is a line plot of its values versus the iteration number.

When there are several chains, all chains are plotted on the same figure, each with a separate line, for comparison.

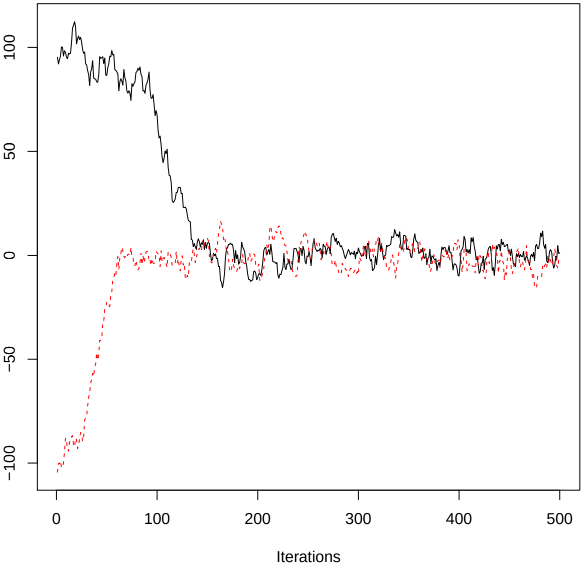
When there are several nodes to monitor, each node has its own figure.

A good-looking trace plot has all chains sampling from the same region, and fairly evenly over all iterations ...

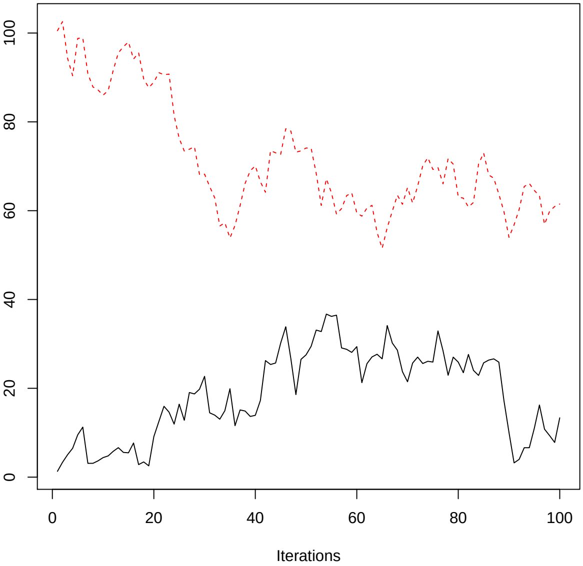

4
2
0
−2
−4
0
50
100
150
200
Iterations

A trace plot may show initial transient effects prior to convergence.

This happens especially if the initial values are overdispersed ...


100
50
0
−50
−100
0
100
200
300
400
500
Iterations

Prior to convergence, traces of different chains can fail to overlap ...



100
80
60
40
20
0
0
20
40
60
80
100
Iterations

When traces of different chains fail to (eventually) overlap, it could be a sign of nonstationarity (lack of convergence) or of very slow *mixing* (later).

Gelman-Rubin

Consider monitoring convergence using scalar ψ, with $m > 1$ chains:

$$\psi_{ij} \;=\; \text{value of } \psi \text{ from iteration } i \text{ of chain } j$$

For $n > 1$ iterations per chain, let

$$\overline{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij} \qquad \overline{\psi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\overline{\psi}_{\cdot j} \qquad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{ij} - \overline{\psi}_{\cdot j}\right)^2$$

Within- and between-chain variance estimates are

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \qquad B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{\psi}_{\cdot j} - \overline{\psi}_{\cdot\cdot}\right)^2$$

Idea: If chains are sampling different regions for ψ, between-chain variance B should be large relative to within-chain variance W.

A weighted average estimate:

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

As an estimate of $\mathrm{var}(\psi \mid y)$, this

- Tends to overestimate, when starting points are overdispersed
- Is unbiased, if all chains are stationary

When approaching stationarity, W, B, and $\widehat{\mathrm{var}}^{+}(\psi \mid y)$ should tend to $\mathrm{var}(\psi \mid y)$.

Let

$$\widehat{R} \;=\; \sqrt{\frac{\widehat{\mathrm{var}}^{+}(\psi \mid y)}{W}}$$

This is a version of the **Gelman-Rubin statistic** for ψ.

It will tend to be larger than 1, and especially large when the chains are dominated by transient effects.

It should approach 1 in the long run ($n \to \infty$).

BDA3 suggests $\widehat{R} < 1.1$ to declare convergence.

Note:

- This always requires at least two chains.
- Chains should have separate, overdispersed starting values.
- Each ψ of interest will have its own $\widehat{R}$, and all should be close to 1 to declare convergence.
- This will not work if ψ does not have a posterior variance.

Important:

- If there is an adaptation phase, convergence diagnostics (trace plots, statistics) are applied only to iterations *after* that phase.

  During the adaptation phase, there may not be a stationary distribution.

- All scalar quantities of interest should be monitored for convergence.

  (Generally, should include at least all top-level stochastic nodes.)

- Passing convergence tests does not guarantee convergence!