11: Basics of Markov chain Monte Carlo Simulation

Kai Tan

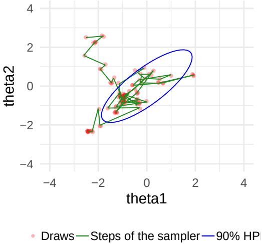
University of Kentucky

Section 1

Convergence Inference

Warm-up and convergence diagnosticse

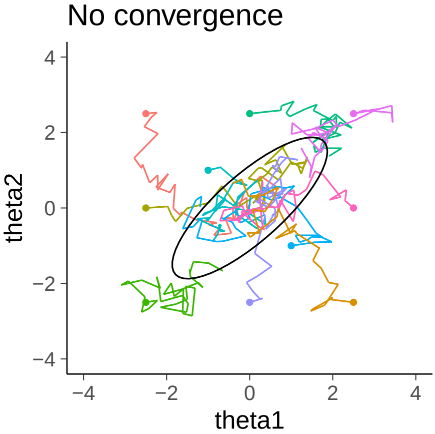
- Asymptotically chain spends the $\alpha\%$ of time where $\alpha\%$ posterior mass is
  - but in finite time the initial part of the chain may be non-representative and lower error of the estimate can be obtained by throwing it away



- **Warm-up** = remove draws from the beginning of the chain
- Convergence diagnostics
  - Do we get samples from the target distribution?

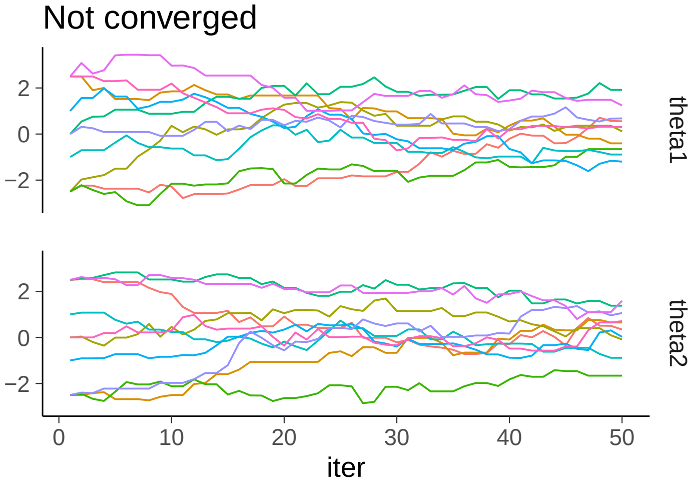
Several chains

- Use of several chains make convergence diagnostics easier
- Start chains from different starting points – preferably overdispersed

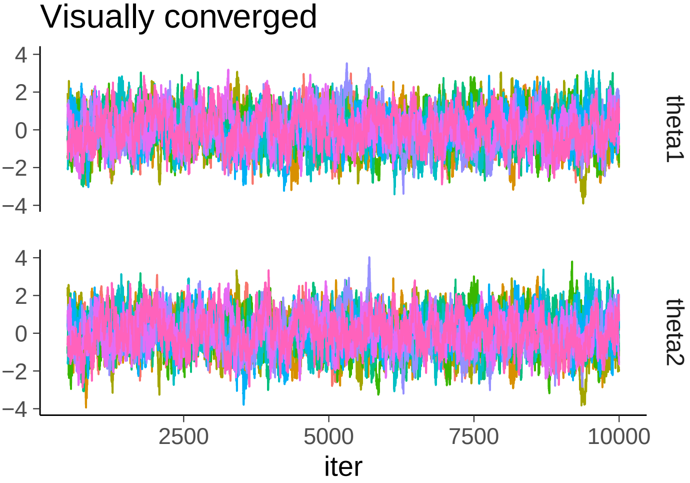


- Remove draws from the beginning of the chains and run chains long enough so that it is not possible to distinguish where each chain started and the chains are well mixed

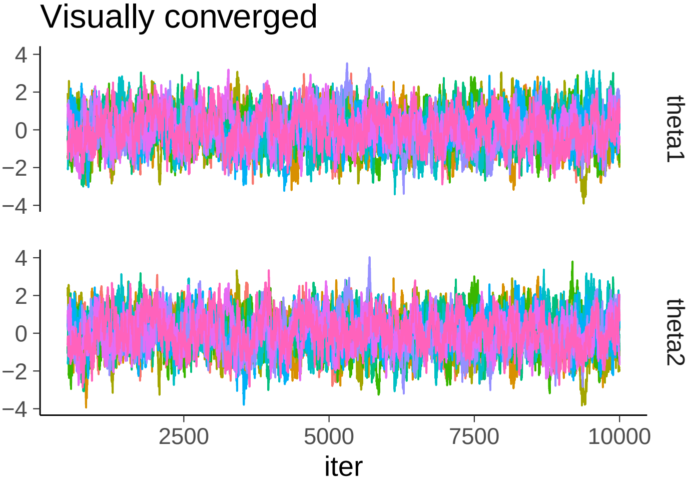
Several chains



Several chains



Several chains



Visual convergence check is not sufficient

Introduction of $\widehat{R}$ (estimated potential scale reduction)

- M chains, each having N draws.
- Within chains variance W

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2, \text{ where } s_m^2 = \frac{1}{N-1} \sum_{n=1}^{N} (\theta_{nm} - \bar{\theta}_{.m})^2$$

- Between chains variance B

$$B = \frac{N}{M-1} \sum_{m=1}^{M} (\bar{\theta}_{.m} - \bar{\theta}_{..})^2,$$

$$\text{where } \bar{\theta}_{.m} = \frac{1}{N} \sum_{n=1}^{N} \theta_{nm}, \ \bar{\theta}_{..} = \frac{1}{M} \sum_{m=1}^{M} \bar{\theta}_{.m}$$

  - B/N is variance of the means of the chains

$\widehat{R}$

- Estimate total variance $\mathrm{var}(\theta|y)$ as a weighted mean of W and B

$$\widehat{\mathrm{var}}^{+}(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B$$

  - $\widehat{\mathrm{var}}^{+}$ **overestimates** marginal posterior variance if the starting points are overdispersed

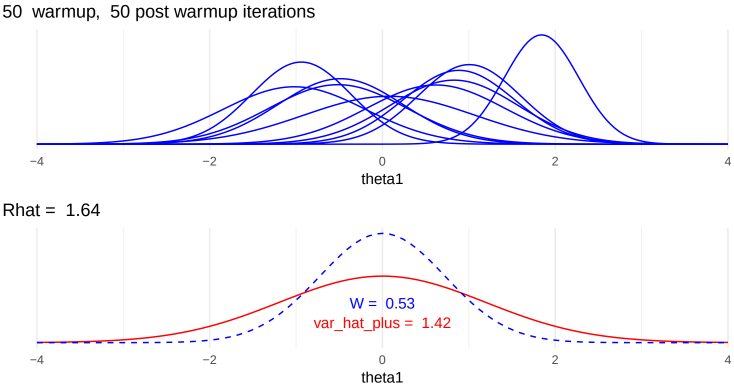
- For finite N, W **underestimates** marginal posterior variance
  - single chains have not yet visited all points in the distribution
  - when $N \to \infty, \quad \mathrm{E}(W) \to \mathrm{var}(\theta|y)$

- As $\widehat{\mathrm{var}}^{+}(\theta|y)$ overestimates and W underestimates, define

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}}$$

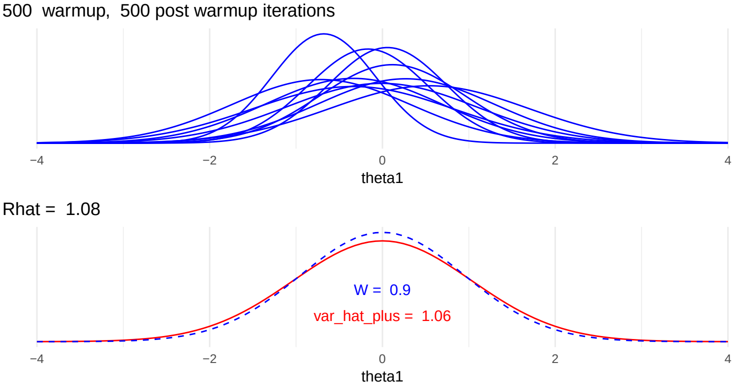
$\widehat{R}$

- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate



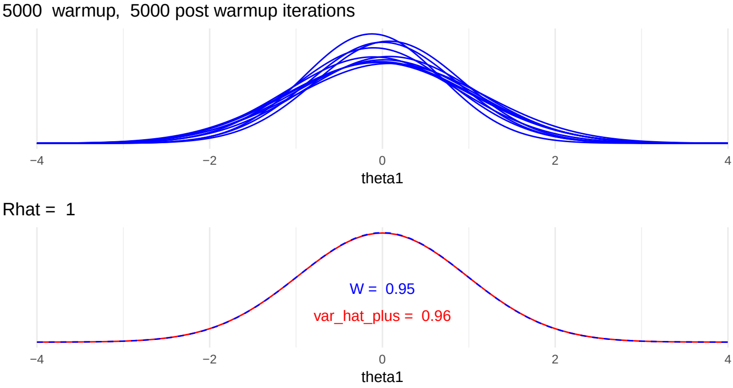
$\widehat{R}$

- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate



$\widehat{R}$

- Compare means and variances of the chains
  W = within chain variance estimate
  var_hat_plus = total variance estimate



$\widehat{R}$

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}}{W}}$$

- Estimates how much the scale of ψ could reduce if $N \to \infty$
- $\widehat{R} \to 1$, when $N \to \infty$
- if $\widehat{R}$ is big (e.g., $R > 1.1$), keep sampling
- If $\widehat{R}$ is close to 1, it is still possible that chains have not converged
  - if starting points were not overdispersed
  - distribution far from normal (especially if infinite variance)
  - just by chance when n is finite

Rank normalized $\widehat{R}$ from Vehtari, Aki, et al. (2019)

- Original $\widehat{R}$ requires that the target distribution has finite mean and variance
- Rank normalization fixes this and is also more robust given finite but high variance
- Folding improves detecting scale differences between chains
- This paper also proposes local convergence diagnostics and practical MCSE estimates for quantiles

Vehtari, Aki, et al. "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC." arXiv preprint arXiv:1903.08008 (2019).

Section 2

Effective number of simulation draws

Motivation

- If the n simulation draws within each sequence were truly independent, *mn* independent simulations from the m sequences.
- The Between-sequence variance B would be an unbiased estimate of the posterior variance $E(B) = \text{var}(\psi|y)$.
- However, the simulations of ψ within each sequence will be autocorrelated, and $E(B) > \text{var}(\psi|y)$.
- consider the statistical efficiency of the average of the simulations $\bar{\psi}_{..}$, as an estimate of the posterior mean, E$(\psi|y)$.

Asymptotic formula

$$\lim_{n\to\infty} mn\,\mathrm{var}(\bar{\psi}_{..}) = \left(1 + 2\sum_{t=1}^{\infty} \rho_t\right) \mathrm{var}(\psi|y).$$

ρ_t is the autocorrelation of the sequence ψ at lag t.

$$\rho_t = \frac{cov(\psi_i, \psi_{i-t})}{\sqrt{\mathrm{var}(\psi_i)\,\mathrm{var}(\psi_{i-t})}}$$

- If the n simulation draws from each of the m chains were independent, then $\rho_t = 0$, thus

  $$\mathrm{var}(\bar{\psi}_{..}) = \frac{1}{mn}\mathrm{var}(\psi|y).$$

  sample size will be mn.
- Consider the autocorrelation, define the **effective sample size** as

  $$n_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t}.$$

- Need to estimate the $\sum_{t=1}^{\infty} \rho_t$.
- First calculate the $\widehat{\text{var}}^+(\psi|y)$
- Then estimate the correlations $\hat{\rho}_t$ by *variogram* V_t at each lag t.

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2.$$

- Note that $\mathrm{E}(\psi_{i,j} - \psi_{i-t,j})^2 = 2(1-\rho_t)\,\text{var}(\psi|y)$, then $\mathrm{E}(V_t) = 2(1-\rho_t)\,\text{var}(\psi|y)$. We can estimate ρ_t by

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{\text{var}}^+(\psi|y)}.$$

- Infinite summation over $t = 1, 2, \ldots, \infty$?

Partial summation over $t = 1, 2, \ldots, T$

- Since the $\hat{\rho}_t$ may be too noisy for large t.
- Choose T by the following rule from Geyer (1992)

$$T = \min\{t \text{ is odd} : \hat{\rho}_{t+1} + \hat{\rho}_{t+2} < 0\}$$

- The effective sample size is thus

$$\hat{n}_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\mathbf{T}} \hat{\rho}_t}.$$

When to stop the simulation?

Rule of thumb,

- Potential scale reduction factor

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi|y)}{W}}$$

  Stop when $\widehat{R}$ is close to 1, say $\widehat{R} < 1.1$.

- Effective sample size

$$\hat{n}_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{\mathbf{T}} \hat{\rho}_t}$$

  Stop the simulation when $\hat{n}_{eff} \geq 5m$.

- Once $\widehat{R}$ is near 1 and $\hat{n}_{eff} \geq 5m$ for all scalar estimands of interest, just collect the mn simulations (deleted warm-up iterations) and treat them as a sample from the target distribution.

Section 3

Example: hierarchical normal model

Example: Hierarchical Normal Model

1. $y_{ij} \sim \text{Normal}(\theta_j, \sigma^2)$, $i = 1, 2, \ldots, n_i$, $j = 1, 2, \ldots, J$
   In total, J groups with different mean θ_j.
   - $y_{1,\mathbf{1}}, y_{2,\mathbf{1}}, \ldots, y_{n_{\mathbf{1}},\mathbf{1}} \sim N(\theta_{\mathbf{1}}, \sigma^2)$
   - $y_{1,\mathbf{2}}, y_{2,\mathbf{2}}, \ldots, y_{n_{\mathbf{2}},\mathbf{2}} \sim N(\theta_{\mathbf{2}}, \sigma^2)$
   - ...
   - $y_{1,\mathbf{J}}, y_{2,\mathbf{J}}, \ldots, y_{n_{\mathbf{J}},\mathbf{J}} \sim N(\theta_{\mathbf{J}}, \sigma^2)$
2. $\theta_j \sim \text{Normal}(\mu, \tau^2)$
3. Uniform prior: $p(\mu, \log\sigma, \log\tau) \propto \tau$.

Target distribution

$$p(\theta, \mu, \log \mu, \log \tau | y) \propto \tau \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_{ij} | \theta_j, \sigma^2).$$

- It's too complicated, and hard to make inference on it.
- Conditional posterior distribution can be calculated.
- Consider Gibbs sampling.

Gibbs sampler for this posterior distribution

Chapter 5 gives

1. Conditional posterior distribution of each θ_j.
2. Conditional posterior distribution of μ.
3. Conditional posterior distribution of σ^2.
4. Conditional posterior distribution of τ^2.

Conditional posterior distribution of θ_j

$$\theta_j|\mu, \sigma, \tau, y \sim N(\hat{\theta}_j, V_{\theta_j})$$

- The mean is a weighted mean of prior mean μ and likelihood mean $\bar{y}_{\cdot j}$.

$$\hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{\cdot j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

- The variance is

$$V_{\theta_j} = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

- Theses conditional distributions are independent.

Conditional posterior distribution of μ

$$\mu|\theta, \sigma, \tau, y \sim N(\hat{\mu}, \tau^2/J)$$

- The mean is
$$\hat{\mu} = \frac{1}{J}\sum_{j=1}^{J}\theta_j.$$

Conditional posterior distribution of σ^2

Since $y_{ij} \sim N(\theta_j, \sigma^2)$, and $\theta \sim N(\mu, \tau^2)$. It can be treated as variance of normal distribution with known mean θ_j.

$$\sigma | \mu, \theta, \tau, y \sim \text{Inv-}\chi^2(n, \hat{\sigma}^2),$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2.$$

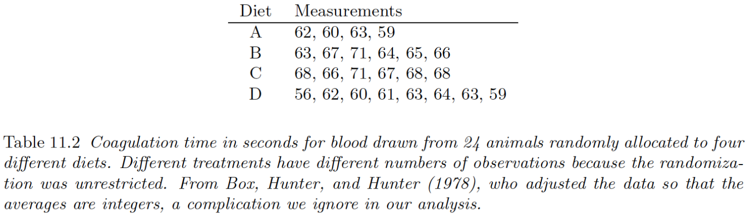
Conditional posterior distribution of τ^2

$$\tau|\theta, \mu, \sigma, y \sim \text{Inv-}\chi^2(J-1, \hat{\tau}^2),$$

where

$$\hat{\tau}^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \mu)^2.$$

Analysis of coagulation data via Gibbs sampler I



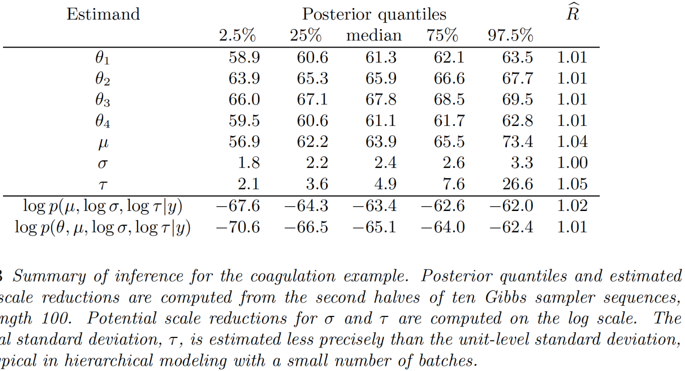
Diet	Measurements
A	62, 60, 63, 59
B	63, 67, 71, 64, 65, 66
C	68, 66, 71, 67, 68, 68
D	56, 62, 60, 61, 63, 64, 63, 59

Table 11.2 *Coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets. Different treatments have different numbers of observations because the randomization was unrestricted. From Box, Hunter, and Hunter (1978), who adjusted the data so that the averages are integers, a complication we ignore in our analysis.*

- Coagulation time (in seconds) for blood from 24 animals.
- $J = 4$ different diets, with $n_1 = 4, n_2 = 6, n_3 = 6, n_4 = 8$.
- $y_{ij} \sim N(\theta_j, \sigma^2)$, $i = 1, 2, \ldots, n_i$, $j = 1, 2, \ldots, J$

Analysis of coagulation data via Gibbs sampler II

Inference from 10 parallel Gibbs sampler sequences.
100 iterations were sufficient for approximate convergence.



Estimand	Posterior quantiles					$\widehat{R}$
	2.5%	25%	median	75%	97.5%	
θ_1	58.9	60.6	61.3	62.1	63.5	1.01
θ_2	63.9	65.3	65.9	66.6	67.7	1.01
θ_3	66.0	67.1	67.8	68.5	69.5	1.01
θ_4	59.5	60.6	61.1	61.7	62.8	1.01
μ	56.9	62.2	63.9	65.5	73.4	1.04
σ	1.8	2.2	2.4	2.6	3.3	1.00
τ	2.1	3.6	4.9	7.6	26.6	1.05
$\log p(\mu, \log\sigma, \log\tau \mid y)$	−67.6	−64.3	−63.4	−62.6	−62.0	1.02
$\log p(\theta, \mu, \log\sigma, \log\tau \mid y)$	−70.6	−66.5	−65.1	−64.0	−62.4	1.01

Table 11.3 *Summary of inference for the coagulation example. Posterior quantiles and estimated potential scale reductions are computed from the second halves of ten Gibbs sampler sequences, each of length 100. Potential scale reductions for σ and τ are computed on the log scale. The hierarchical standard deviation, τ, is estimated less precisely than the unit-level standard deviation, σ, as is typical in hierarchical modeling with a small number of batches.*