Catalano, Fasano, Fengler, Legramanti, Rebaudo

Metropolis-Hastings & Gibbs sampling for Bayesian inference

Overview

Independent MH for the mixture of two normal distributions

Gibbs sampling for a hierarchical normal model

Overview

Independent MH for the mixture of two normal distributions

Gibbs sampling for a hierarchical normal model

The model

The model is a mixture of two normals with fixed parameters

$$\mu_1 = 7, \quad \mu_2 = 10, \quad \sigma_1 = \sigma_2 = 0.5$$

and unknown (random) mixing proportion δ.

Hence the model is:

$$y_1, \ldots, y_n \mid \delta \quad \overset{i.i.d.}{\sim} \quad \delta N(\mu_1, \sigma_1^2) + (1 - \delta) N(\mu_2, \sigma_2^2)$$

We generate a sample of size $n = 100$ with $\delta = 0.7$.

We then sample from the posterior for δ with an **independent MH** using the prior as proposal distribution.

We expect the posterior to concentrate around $\delta = 0.7$.

Recall: Metropolis-Hastings (MH)

1. **Initialization:** Set $t = 0$ and sample x_0 from a starting distribution
2. **At (t+1)-th iteration:**
   - sample the candidate x^* from the proposal distribution $g(\cdot \mid x_t)$ and compute the MH ratio

$$R(x_t, x^*) = \frac{f(x^*)g(x_t \mid x^*)}{f(x_t)g(x^* \mid x_t)}$$

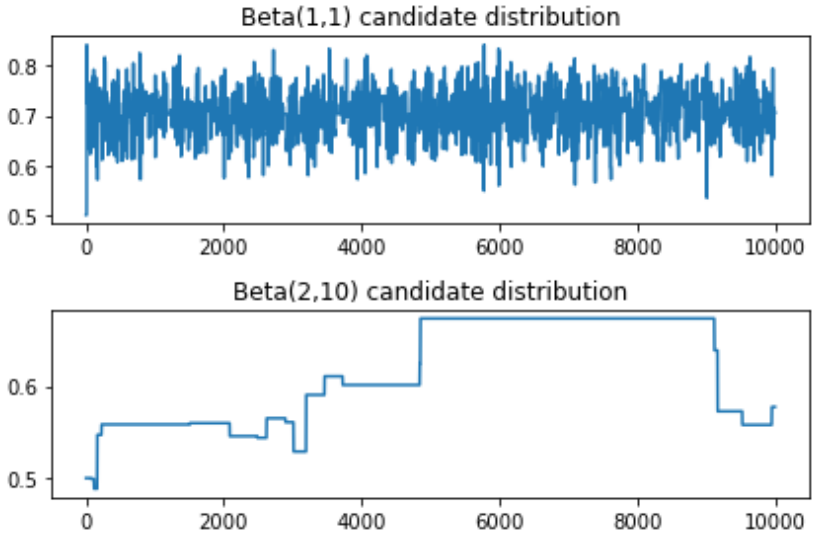
   - sample $u \sim U(0, 1)$
   - if $u < R(x_t, x^*)$ accept x^* as x_{t+1}
   - else set $x_{t+1} = x_t$

If $g(\cdot \mid x_t) = g(\cdot)$, we have *independent MH*.

Here: f = posterior, g = prior $\implies$ R = likelihood ratio.

Results: sample paths

We experimented with a $Beta(1,1) = U(0,1)$ and a more skewed $Beta(2,10)$ as priors. The start was fixed at $\delta_0 = 0.5$ for both.

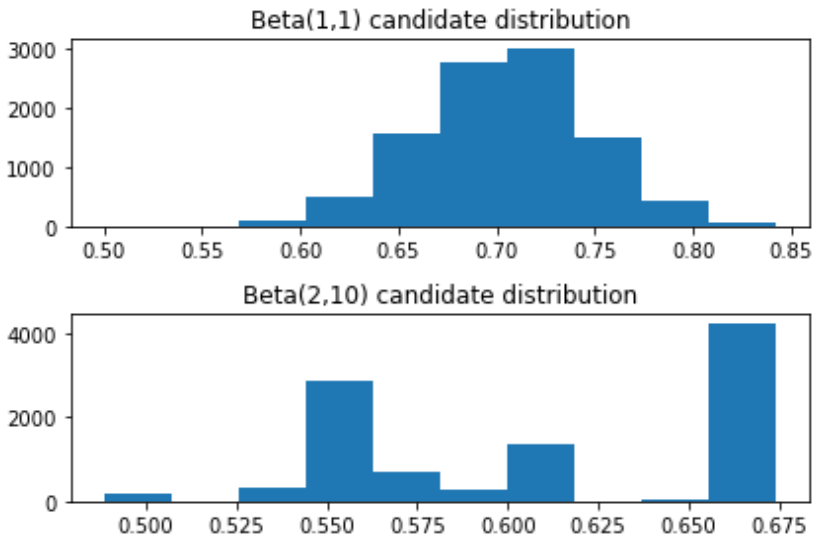


With the uniform prior, the chain moves away from δ_0 quickly and explores the posterior support well (*good mixing*).

With the $Beta(2,10)$ prior, only a few unique values are accepted and the mixing is poor.

Results: histograms

We excluded the first 5000 iterations (*burn-in*).



With the uniform prior, we get a sample with posterior mean approximately equal to 0.7, the value we used to generate the data.

With the *Beta*$(2, 10)$ prior, we have already seen that we get a lot of ties and the posterior approximation is not reliable.

Overview

Independent MH for the mixture of two normal distributions

Gibbs sampling for a hierarchical normal model

Goal

In the context of Bayesian inference, we show an example in which Gibbs sampling is the natural choice, since conditional posteriors are available (and simple).

The bayesian analysis is taken from:
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (Vol. 2)*. CRC press.

The data

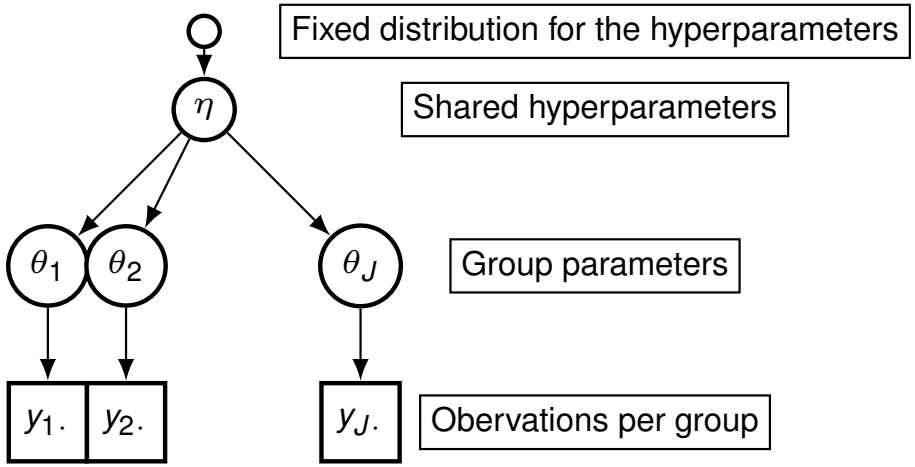
Coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets.

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

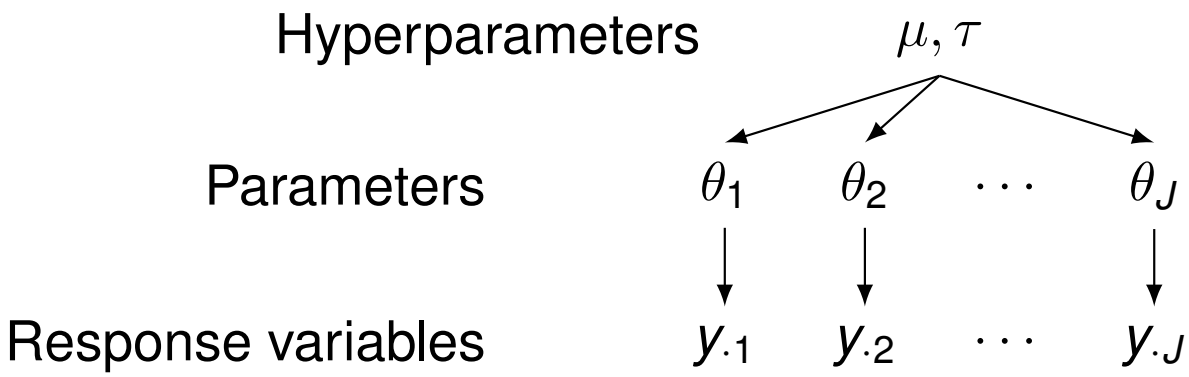
From Box, G. E., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: an introduction to design, data analysis, and model building (Vol. 1)*. New York: Wiley.

Hierarchical model

- J number of groups
- n_j number of observations for the group j
- y_{ji} i-th observation in group j
- θ_j parameters for the distribution of the samples in group j
- η hyperparameters for the distribution of the parameters, sampled from a fixed distribution



Hierarchical Normal Model and prior



$$y_{ij} \mid \theta, \sigma \sim N(\theta_j, \sigma^2) \qquad i = 1, \ldots, n_j \qquad j = 1, \ldots, J$$
$$\theta_j \mid \mu, \tau \sim N(\mu, \tau^2)$$

Prior: $\quad p(\mu, \log \sigma, \log \tau) \propto \tau$

Joint and conditional posteriors

Joint posterior:

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

Conditional posteriors:

$$\theta_j \mid \mu, \sigma, \tau, y \sim N\left(\hat{\theta}_j, V_{\theta_j}\right) \quad \text{where} \quad \hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{\cdot j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}, \quad V_{\theta_j} = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

$$\mu \mid \theta, \sigma, \tau, y \sim N\left(\hat{\mu}, \frac{\tau^2}{J}\right) \quad \text{where} \quad \hat{\mu} = \frac{1}{J}\sum_{j=1}^{J} \theta_j$$

$$\sigma^2 \mid \theta, \mu, \tau, y \sim Inv.\chi^2\left(n, \hat{\sigma}^2\right) \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2$$

$$\tau^2 \mid \theta, \mu, \sigma, y \sim Inv.\chi^2\left(J-1, \hat{\tau}^2\right) \quad \text{where} \quad \hat{\tau}^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \mu)^2$$

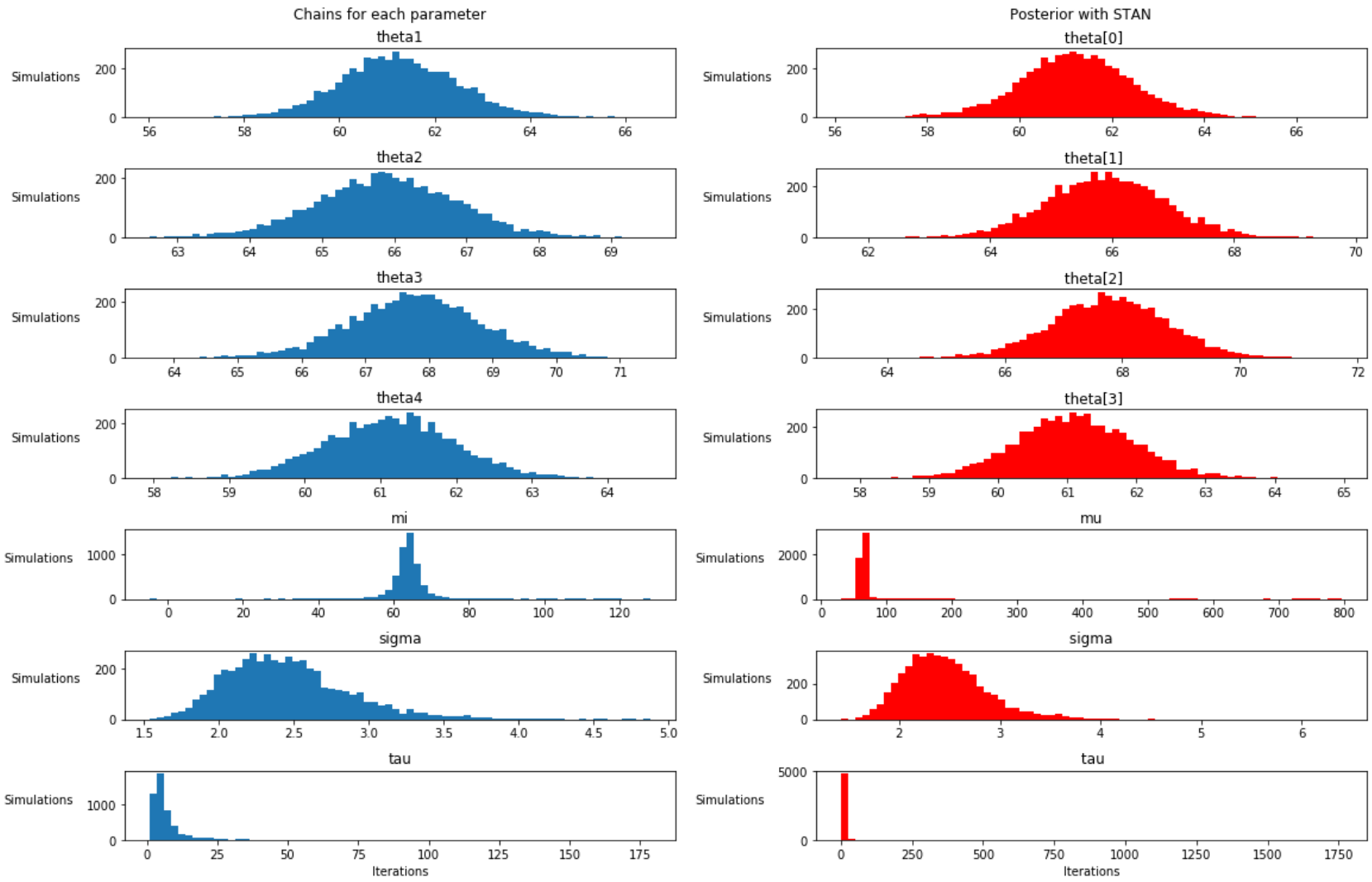
The algorithm

1. Fix initial values for $\hat{\theta}_j$ for $j \in \{1, \ldots, J\}$. In our algorithm they are sampled from the initial data.
2. Initialize $\hat{\mu}$ with the mean of $\hat{\theta}$.

$$\hat{\mu} = \frac{1}{J} \sum_{j=1}^{J} \hat{\theta}_j \tag{1}$$

3. Sample $\hat{\sigma^2}$, $\hat{\tau^2}$, $\hat{\theta}_j$, $\hat{\mu}$ from the respective conditional distributions.
4. Repeat the previous point for 1000 iterations and throw away the first half of the estimates for the parameters (warm-up).
5. Repeat from point 1 to point 4 for 10 chains.

Posterior histograms: Gibbs vs STAN



Runtime comparison: Gibbs vs STAN

		10	100	1000	10000
0	stan	0.002	0.042	0.108	0.737
1	gibbs	0.114	1.038	8.699	85.858

Diagnostic

Problems:

- early iterations reflect the starting points rather than the target distribution;
- within-sequence correlation can cause inefficient simulations.

Monitoring:

- simulating multiple sequences with starting points dispersed in the parameters space;
- comparing variation intra and inter different chains.

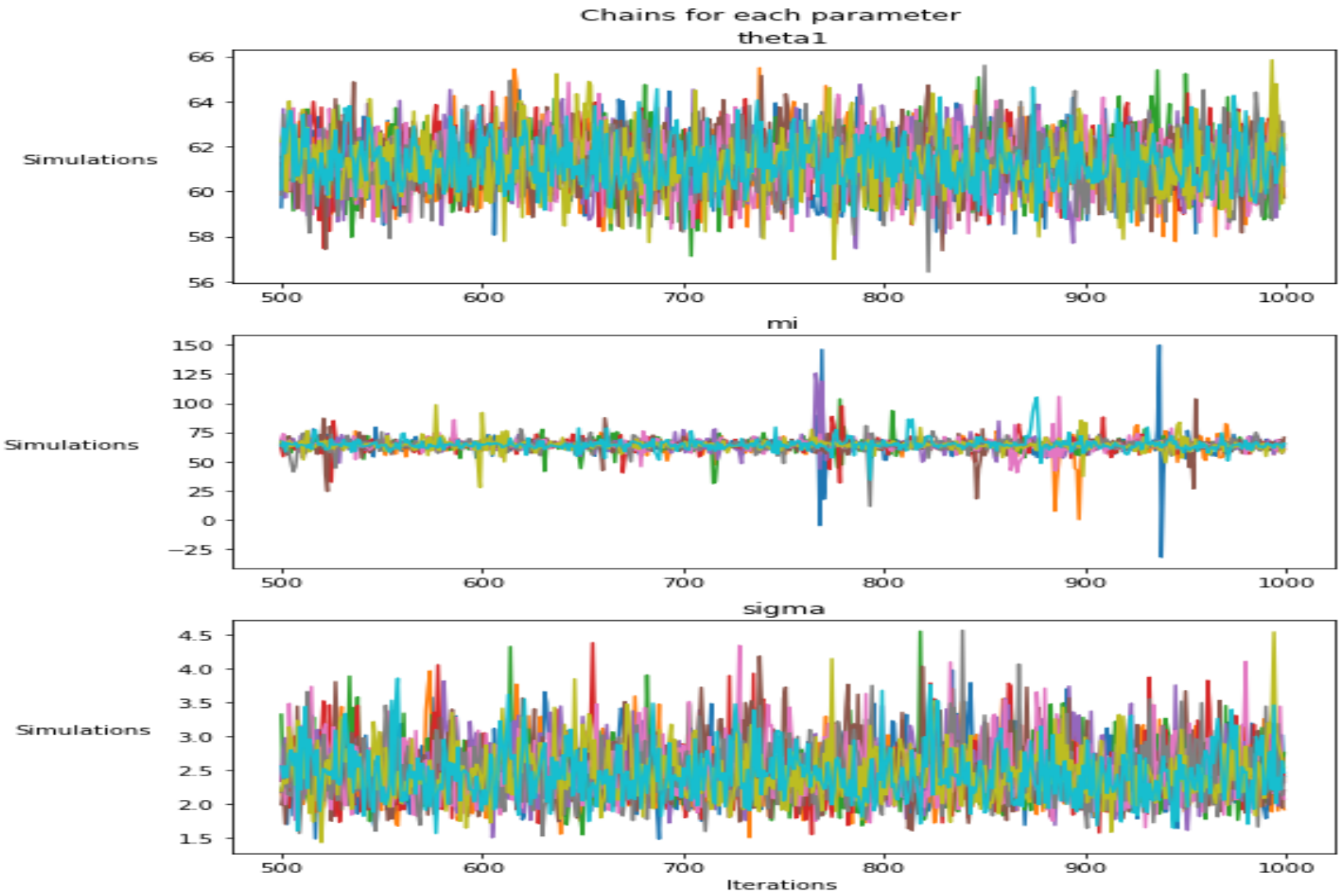
Solutions

- For initial-values dependence, **warm-up**.
  The length of the warm-up should depend on the problem. Rule of thumb: discard the first half of each chain (Gelman et al.).

- For intra-chain dependence, **thinning**.
  Keeping only every k-th draw from each sequence
  $\Rightarrow$ storage advantages.

- for both problems, **graphical comparison**.

Chains


Chains for each parameter
theta1
Simulations
66
64
62
60
58
56
500
600
700
800
900
1000
mi
Simulations
150
125
100
75
50
25
0
−25
500
600
700
800
900
1000
sigma
Simulations
4.5
4.0
3.5
3.0
2.5
2.0
1.5
500
600
700
800
900
1000
Iterations

Diagnostic indexes based on variance

We can compare the estimated variance between chains (B) and within chains (W).

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2; \;\; s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta_{ij} - \bar{\theta}_{\cdot j})^2$$

$$\hat{var}(\theta|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

$$\hat{R} = \sqrt{\hat{var}(\theta|y)/W}$$

The potential scale reduction $\hat{R}$ should be closed to 1.

Effective number of simulations

Since the approximation is not based on *i.i.d.* sample, we must consider the correlation between draws.

$$N_{eff} = \frac{m\,n}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

par	B	W	$\hat{var}$	$\hat{R}$	N_{eff}
θ_1	1.0032	1.49917	1.49719	0.999338	4254.08
θ_2	1.21471	1.00384	1.00468	1.00042	4673.61
θ_3	2.07662	1.00455	1.00884	1.00213	4276.09
θ_4	0.912064	0.742049	0.742729	1.00046	4345.45
μ	28.8446	32.1651	32.1518	0.999794	4900.98
σ	0.244715	0.164945	0.165265	1.00097	3199.1
τ	203.01	72.0278	72.5517	1.00363	1605.47

Thank you for your attention!