Improving Convergence Diagnostics for MCMC Sampling Algorithms

Paul Bürkner (joint work with Aki Vehtari, Andrew Gelman, Daniel Simpson, and Bob Carpenter)

Bayesian Statistics

"If you quantify uncertainty with probability you are a Bayesian."

Micheal Betancourt

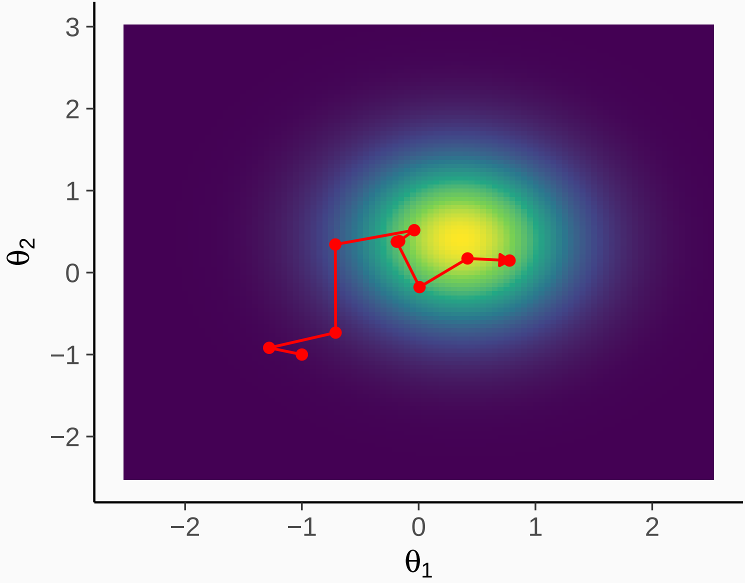
Bayes Theorem:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

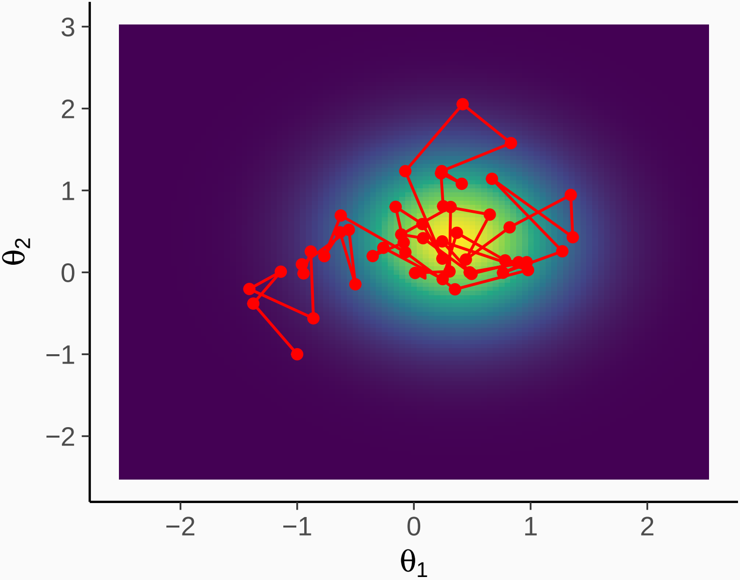
Challenge: Obtain a representation of the posterior distribution

General purpose solution: MCMC Sampling

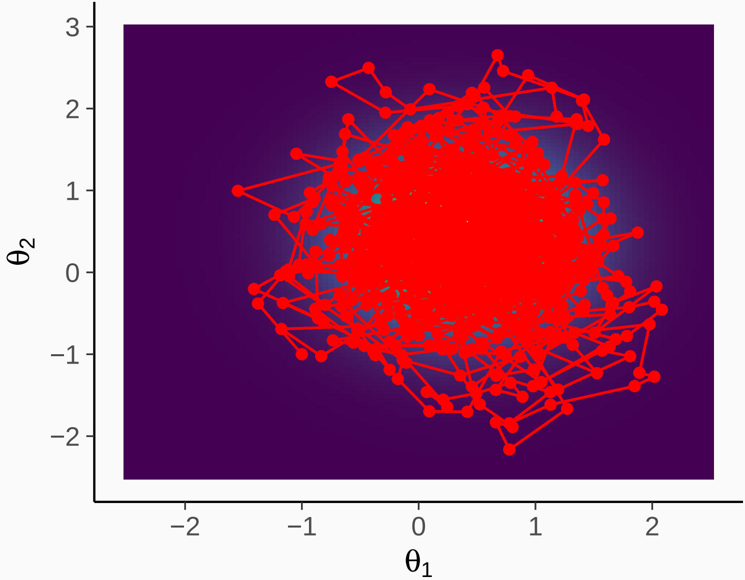
MCMC Sampling: A Single Chain (10 Iterations)



MCMC Sampling: A Single Chain (50 Iterations)



MCMC Sampling: A Single Chain (1000 Iterations)



All we care about are expectations

Expectation of some function f over the distribution $p(\theta \mid y)$:

$$\mathbb{E}_p(f) = \int f(\theta)\, p(\theta \mid y)\, \mathrm{d}\theta$$

Monto-Carlo Estimator

Having obtained exact random draws $\{\theta_s\}$ from $p(\theta \mid y)$:

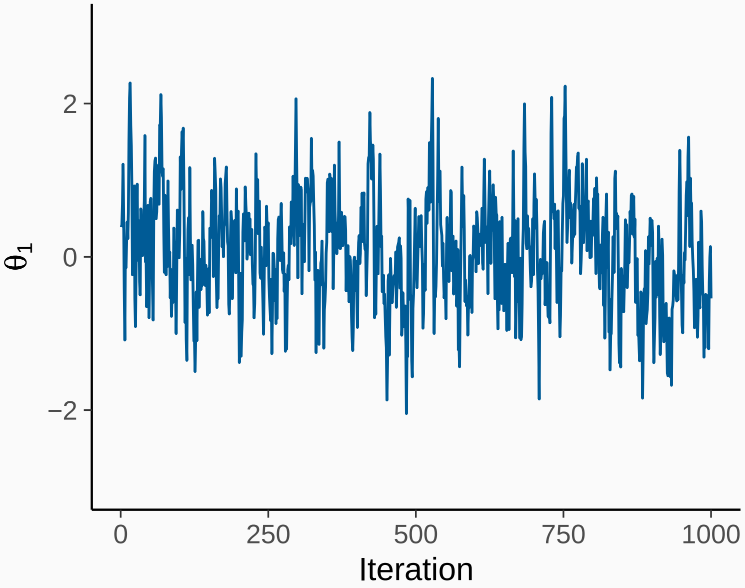
$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s) \sim \text{Normal}\left(\mathbb{E}_p(f), \sqrt{\frac{\text{Var}_p(f)}{\text{S}}}\right)$$

Markov-Chain Monto-Carlo Estimator

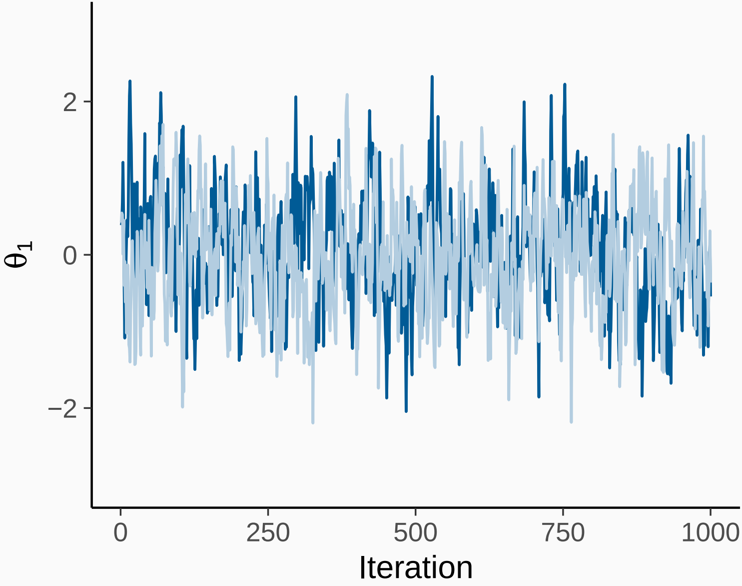
Assuming *geometric ergodicity* of a Markov Chain $\{\theta_s\}$:

$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s) \sim \mathrm{Normal}\left(\mathbb{E}_p(f), \sqrt{\frac{\mathrm{Var}_p(f)}{\mathrm{ESS}}}\right)$$

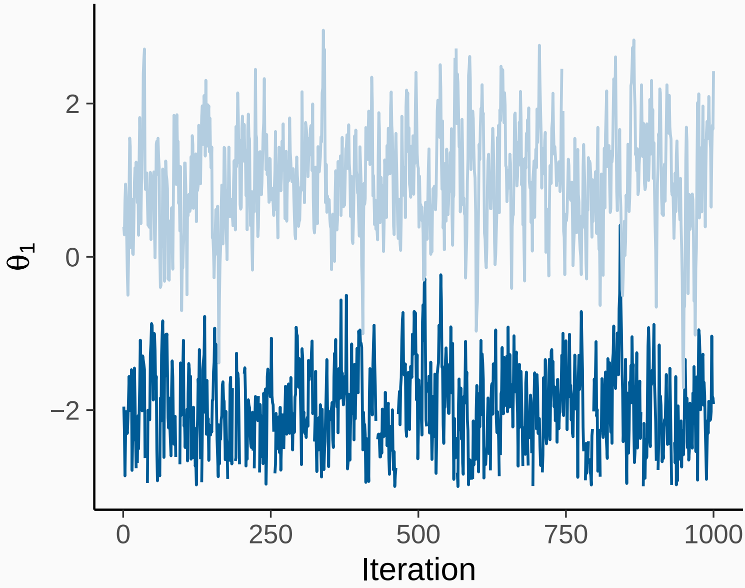
Trace Plots: Visualizing a Single Chain



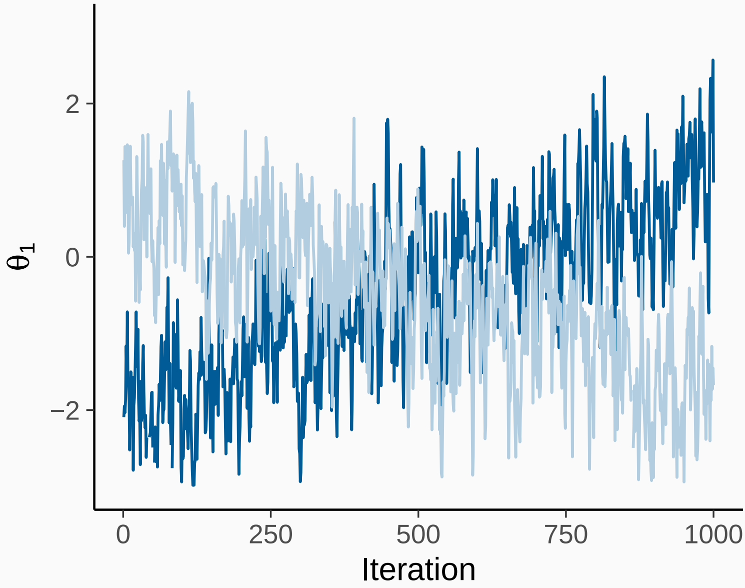
Trace Plots: Visualizing Multiple Chains



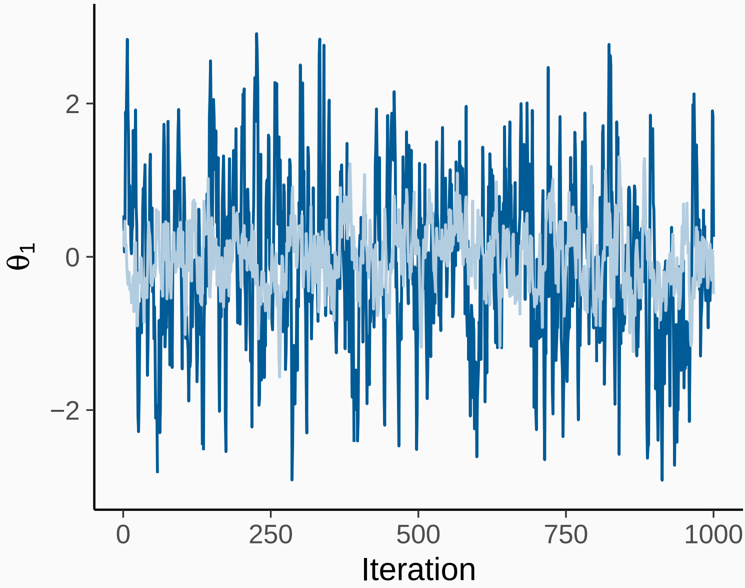
Chains with Different Locations



Non-Stationary Chains



Chains with Different Variances



Traditional MCMC Diagnostics

Between Chain Variance:

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

Within Chain Variance:

$$W = \frac{1}{M(N-1)} \sum_{m=1}^{M} \sum_{n=1}^{N} (\theta^{(nm)} - \overline{\theta}^{(.m)})^2$$

Potential Scale Reduction Factor:

$$\widehat{R} = \sqrt{\frac{\frac{N-1}{N} W + \frac{1}{N} B}{W}}$$

Effective Sample Size:

$$\text{ESS} = \frac{N\,M}{\hat{\tau}}$$

Problems with the Traditional MCMC Diagnostics

(1) We do not detect differences of chains with infinite means

(2) We do not detect non-convergence in the tails of the distribution

(3) We cannot properly localize convergence problems

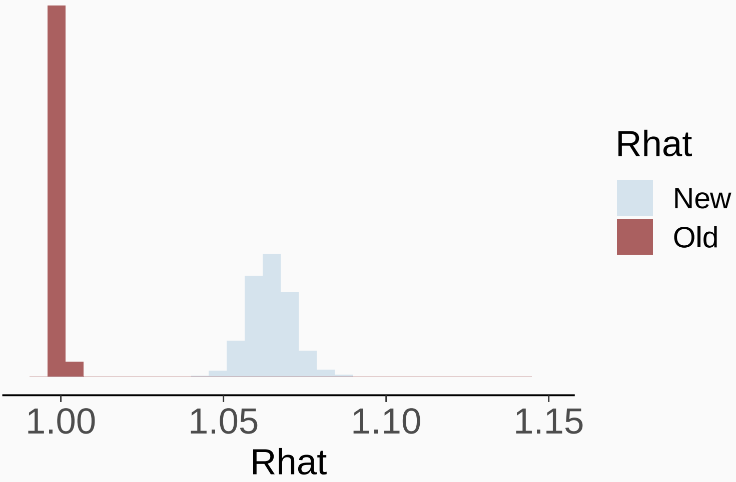
Solution to (1): Rank Normalization of Draws

- Replace the original posterior draws $\theta^{(nm)}$ with their ranks $r^{(nm)}$ computed across all chains
- Normalize the ranks via

$$z^{(nm)} = \Phi^{-1}((r^{(nm)} - 0.5)/S)$$

- Compute $\widehat{R}$ and ESS based on $z^{(nm)}$
- We call these measures bulk-$\widehat{R}$ and bulk-ESS

Chains with Infinite Mean and Different Locations



Solution to (2): Folding of Draws

- Fold the original draws $\theta^{(nm)}$ around their median

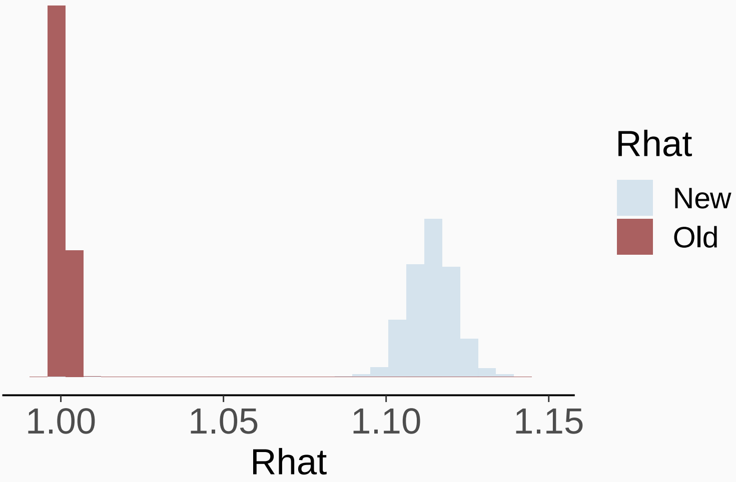
$$\zeta^{(nm)} = |\theta^{(nm)} - \text{median}(\theta^{(nm)})|$$

- Rank normalize $\zeta^{(nm)}$ in the same way as done for $\theta^{(nm)}$
- Compute $\widehat{R}$ based on rank normalized $\zeta^{(nm)}$
- We call thise measure folded-$\widehat{R}$

Proposed new version of $\widehat{R}$:

$$\widehat{R} = \text{Max}(\text{bulk-}\widehat{R}, \text{folded-}\widehat{R})$$

Chains with Finite Mean and Different Variances



Solution to (3): Efficiency of Quantiles

The empirical distribution function (ECDF) can be estimated as:

$$\Pr(\theta \leq \theta^\star) \approx \bar{I}^\star = \frac{1}{S} \sum_{s=1}^{S} I(\theta^{(s)} \leq \theta^\star), \tag{1}$$

Efficiency of the α-Quantile Q_α:

- Efficiency of the indicator $I(\theta^{(s)} \leq Q_\alpha)$

Efficiency of small intervals between Q_α and $Q_{\alpha+\delta}$:

- Efficiency of the indicator $I(\hat{Q}_\alpha < \theta^{(s)} \leq \hat{Q}_{\alpha+\delta})$

Tail-ESS: Minimum ESS of the 5% and 95% quantiles

Case Study: Eight Schools Meta-Analysis

Data:

- y_i: Mean effect of the treatment on SAT scores in school i
- σ_i: Standard deviation of the mean effect in school i

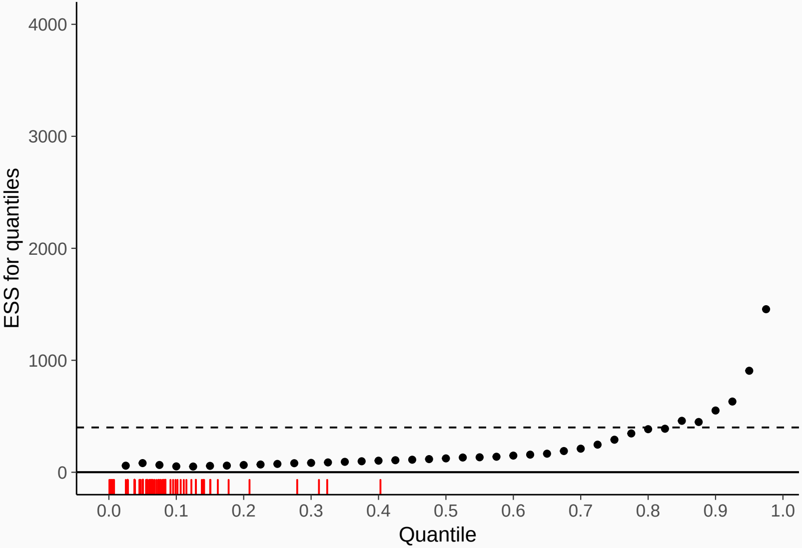
Random effects meta-analytic model:

- $y_i \sim \text{Normal}(\theta_i, \sigma_i)$
- $\theta_i \sim \text{Normal}(\mu, \tau)$

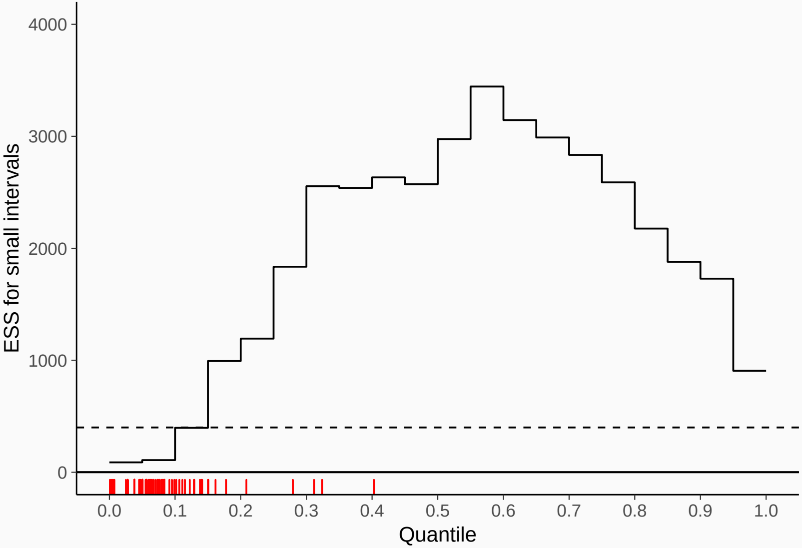
Convergence diagnostics for τ based on 4000 samples:

- $\widehat{R} = 1.02$
- bulk-ESS $= 95$
- tail-ESS $= 46$

Quantile Efficiency of τ



Small Interval Efficiency of τ



Summary

(1) MCMC Sampling is a powerful tool to estimate highly complex Bayesian models

(2) The current convergence diagnostics for MCMC algorithms have serious flaws and limitations

(3) We recommend a set of changes to alleviate these problems

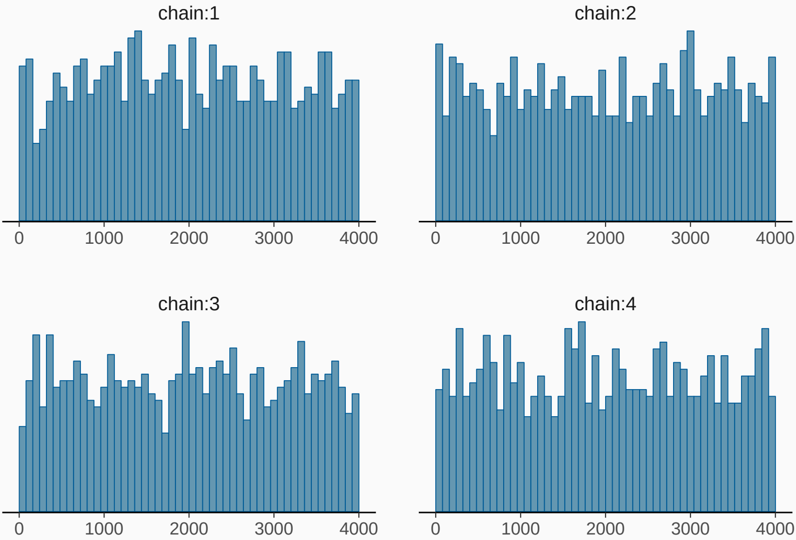
(4) We propose new visualizations for MCMC diagnostics

Thank you!

Reference: Vehtari A., Gelman A., Simpson D., Carpenter B., & Bürkner P. C. (in review). Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC. ArXiv preprint.

Appendix

Rank Plots: Good Mixing of Chains



Rank Plots: Bad Mixing of Chains