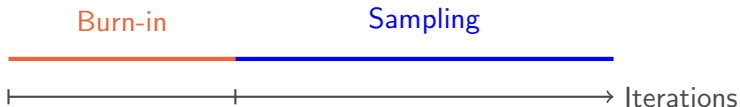


The Convergence Challenge in MCMC

- ▶ **Ideal goal:** Assess whether MCMC chains have converged
- ▶ **Fundamental problem:**
 - ▶ In general, impossible to know for sure that there is no problem
 - ▶ But we can sometimes know for sure that there *is* a problem
- ▶ **Two phases of MCMC:**
 - ▶ Transient phase (burn-in): mixing time
 - ▶ Stationary phase: Monte Carlo estimation



Why Convergence Matters

Non-converged chains:

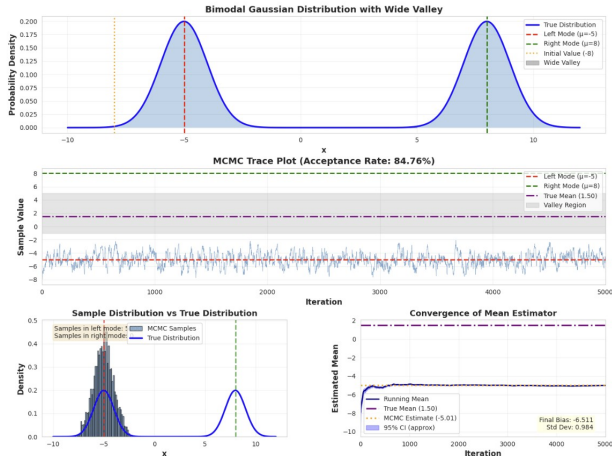
- ▶ Biased estimates
- ▶ Incorrect uncertainty quantification
- ▶ Missing important modes
- ▶ Unreliable inference

Key Question

How can we diagnose whether our MCMC chains have converged to the target distribution?

Motivating example

MCMC Failure in Bimodal Distribution: Trapped in Local Mode



The Intuition Behind Gelman-Rubin

Core Idea

If MCMC chains have converged to the target distribution, then:

- ▶ Multiple chains started from different points should look similar
- ▶ Between-chain variance \approx Within-chain variance

Compare two sources of variance:

1. **Within-chain variance (W)**
How much each chain varies
2. **Between-chain variance (B)**
How different chains are from each other

Mathematical Foundation

Consider M chains, each of length T :

Variance Decomposition

Total sum of squares = Inter-group + Intra-group

$$\sum_{m=1}^M \sum_{t=1}^T (X_{m,t} - \bar{X}_{..})^2 = \sum_{m=1}^M \sum_{t=1}^T (\bar{X}_m - \bar{X}_{..})^2 + \sum_{m=1}^M \sum_{t=1}^T (X_{m,t} - \bar{X}_m)^2$$

- ▶ **Intra-group** = Within-chain variance (W)
- ▶ **Inter-group** = Between-chain variance (B)

Key insight: After convergence, both estimate the same target variance!

The Gelman-Rubin Statistic

Definition

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}$$

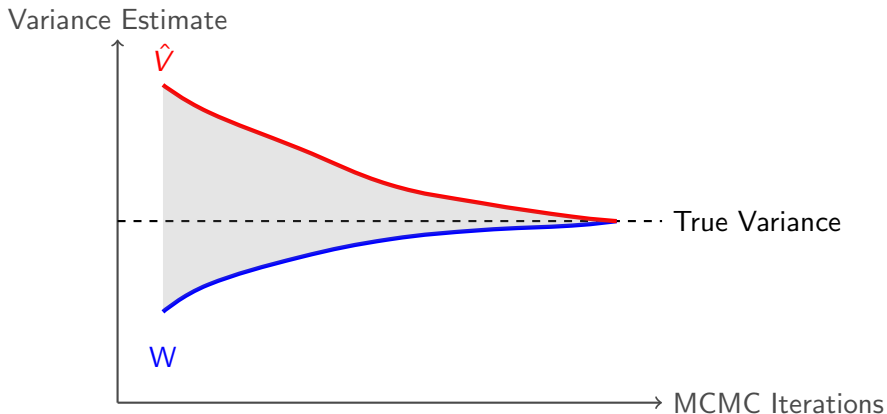
Where:

$$W = \frac{1}{M} \sum_{m=1}^M s_m^2 \quad (\text{average within-chain variance})$$

$$B = \frac{T}{M-1} \sum_{m=1}^M (\bar{X}_m - \bar{X}_{..})^2 \quad (\text{between-chain variance})$$

$$\hat{V} = \frac{T-1}{T} W + \frac{1}{T} B \quad (\text{pooled variance estimate})$$

Why It Works: The Variance Sandwich



- ▶ Initially: $W < \text{True Variance} < \hat{V}$
- ▶ As chains converge: Both W and $\hat{V} \rightarrow \text{True Variance}$
- ▶ Therefore: $\hat{P} = \sqrt{\hat{V} / W} \rightarrow 1$

Evolution of Convergence Thresholds

Historical Development

- ▶ **1992:** Gelman & Rubin propose the diagnostic
- ▶ **2004:** Gelman recommends $\hat{R} < 1.1$
- ▶ **2021:** Vehtari et al. recommend $\hat{R} < 1.01$

Why the stricter threshold?

- ▶ More computing power available
- ▶ Better understanding of convergence
- ▶ Need for more reliable inference
- ▶ Connection to effective sample size

\hat{R} threshold	ESS per chain
1.1	≈ 5
1.05	≈ 20
1.01	≈ 50

Connection to Effective Sample Size

Key Approximation (Vats & Knudson, 2021)

$$\hat{R} \approx \sqrt{1 + \frac{M}{\text{ESS}}}$$

Where:

- ▶ M = number of chains
- ▶ ESS = effective sample size (accounting for autocorrelation)

Implications:

- ▶ $\hat{R} = 1.1 \Rightarrow \text{ESS} \approx 5M$ (5 independent samples per chain)
- ▶ $\hat{R} = 1.01 \Rightarrow \text{ESS} \approx 50M$ (50 independent samples per chain)

5 effective samples per chain is too small for reliable inference!

Major Weaknesses of Gelman-Rubin

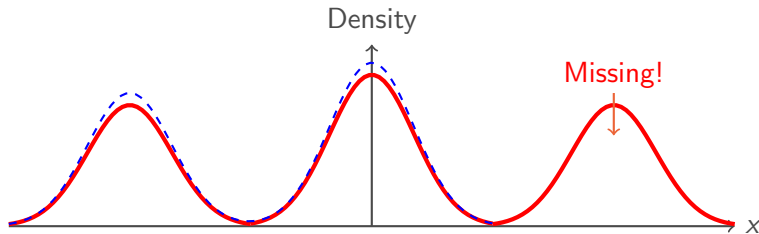
1. **Cannot detect if all modes are found**
 - ▶ Only checks if chains agree with each other
 - ▶ All chains might miss the same modes
2. **Sensitive to initialization**
 - ▶ Chains starting in the same wrong place
3. **Struggles with metastable states**
 - ▶ Chains get stuck but occasionally jump
 - ▶ Similar statistics but poor mixing
4. **Poor for heavy-tailed distributions**
 - ▶ Variance might not exist or be unstable

Remember

$\hat{R} < 1.01$ is necessary but not sufficient for convergence!

Example: Missing Modes

True distribution: Mixture of 3 Gaussians



Chains sample only 2 modes

Result: $\hat{R} < 1.01$ but completely wrong posterior!
All chains agree because they all miss the same mode.

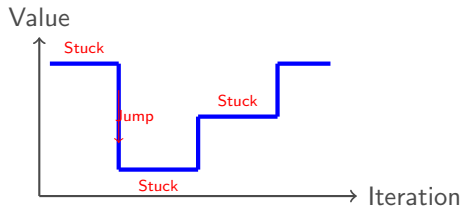
Example: Metastable States

Pathological behavior:

- ▶ Chains get “stuck” for long periods
- ▶ Occasionally jump to other regions
- ▶ All chains show same behavior
- ▶ $\hat{R} \approx 1$ despite poor mixing!

Detection:

- ▶ Very high autocorrelation
- ▶ Low effective sample size
- ▶ Visual inspection of trace plots



Despite poor mixing:

- ▶ Similar means across chains
- ▶ Similar variances
- ▶ $\hat{R} \approx 1$

Comprehensive Convergence Assessment

Use Multiple Diagnostics

1. **Gelman-Rubin statistic:** $\hat{R} < 1.01$
2. **Effective Sample Size:** $ESS > 400$ (minimum)
3. **Trace plots:** Visual inspection
4. **Autocorrelation:** Check mixing quality
5. **Geweke test:** Compare chain beginning and end

Best Practices:

- ▶ Use at least 4 chains (preferably more)
- ▶ Initialize chains from overdispersed starting points
- ▶ Run chains longer than you think necessary
- ▶ Use rank-normalized \hat{R} (more robust)
- ▶ Check both bulk and tail \hat{R}

Modern Extensions

Rank-Normalized \hat{R} (Vehtari et al., 2021)

- ▶ Transform samples to ranks (more robust to outliers)
- ▶ Split chains in half (detect within-chain problems)
- ▶ Separate bulk and tail diagnostics

Bulk- \hat{R} :

- ▶ Convergence of center
- ▶ Mean, median
- ▶ Usually converges faster

Tail- \hat{R} :

- ▶ Convergence of extremes
- ▶ 5%, 95% quantiles
- ▶ Needs more samples

Modern tools (Stan, ArviZ) implement these improvements

Summary Checklist

MCMC Convergence Checklist

1. Run at least 4 chains with dispersed starts
2. Check $\hat{R} < 1.01$ for all parameters
3. Verify ESS > 400 (bulk and tail)
4. Examine trace plots visually
5. Check autocorrelation is low
6. Run sensitivity analysis with different seeds
7. Compare results from different samplers if possible

Remember:

No single diagnostic is perfect

Key Takeaways

1. **Gelman-Rubin compares within vs between chain variance**
 - ▶ Elegant idea: converged chains should agree
2. **Modern threshold is $\hat{R} < 1.01$**
 - ▶ Old threshold (1.1) gives only 5 effective samples
 - ▶ New threshold ensures 50 effective samples
3. **\hat{R} has important limitations**
 - ▶ Can miss modes
 - ▶ Fooled by metastable states
 - ▶ Necessary but not sufficient
4. **Always use multiple diagnostics**
 - ▶ ESS, trace plots, autocorrelation
 - ▶ Visual inspection remains crucial

Good MCMC diagnostics = Reliable scientific inference

References

- ▶ Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.
- ▶ Gelman, A., et al. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman & Hall/CRC.
- ▶ Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.C. (2021). Rank-normalization, folding, and localization: An improved \hat{R} for assessing convergence of MCMC. *Bayesian Analysis*, 16(2), 667-718.
- ▶ Vats, D. and Knudson, C. (2021). Revisiting the Gelman-Rubin diagnostic. *Statistical Science*, 36(4), 518-529.
- ▶ Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434-455.