# Monte Carlo Simulation
## The Gelman–Rubin Statistic and Recent Improvements

Zack Hassman

Department of Statistics, University of Chicago

# 1 The Gelman–Rubin Statistic

This chapter serves to introduce, derive, and motivate the Gelman–Rubin (GR) statistic, originally proposed in Ref. [3]. To streamline our presentation, we fix the test function $h(x) = x$ (and drop the $h$), but in practice any $h(x)$ can be used.

## 1.1 Statement of the Statistic

In this paper, we stick as close as possible to the notation in Chapter 4 of this course's lecture notes [15] but make some minor modifications inspired by Refs. [16, 18]. For a distribution $D$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 < \infty$, assume that we have $M$ independent chains from which we have each collected $N$ samples (after discarding the burn-in samples). Let $X_m^{(n)}$ be the $n^{th}$ draw from the $m^{th}$ chain. Define the per-chain average $\bar{X}_m$, average across all chains $\bar{X}$, and chain sample variance $s_m^2$ to be

$$\bar{X}_m := \frac{1}{N}\sum_{n=1}^{N} X_m^{(n)}, \qquad \bar{X} := \frac{1}{M}\sum_{m=1}^{M} \bar{X}_m, \qquad s_m^2 := \frac{1}{N-1}\sum_{n=1}^{N}(X_m^{(n)} - \bar{X}_m)^2. \tag{1.1}$$

Next, define the between-chain variance $B$ and within-chain variance $W$ to be

$$B := \frac{N}{M-1}\sum_{m=1}^{M}(\bar{X}_m - \bar{X})^2, \quad W := \frac{1}{M}\sum_{m=1}^{M} s_m^2 \tag{1.2}$$

An estimate for posterior variance $V$ is given by

$$V := \frac{N-1}{N}W + \frac{1}{N}B \tag{1.3}$$

Finally, the GR statistic is given by

$$\widehat{R} := \sqrt{\frac{V}{W}} = \sqrt{\frac{\frac{N-1}{N}W + \frac{1}{N}B}{W}} = \sqrt{\frac{N-1}{N} + \frac{1}{N}\frac{B}{W}} \tag{1.4}$$

## 1.2 Derivation

We first make two major assumptions. First, we assume that our MCMC draws within a single chain are independent. This is an approximation rather than an assumption since we know it is not true, but we will refer to it as an assumption. Second, we assume that asymptotically,

$$X_m^{(n)} \sim \mathcal{N}(\mu, \sigma^2) \tag{1.5}$$

and therefore,

$$\bar{X}_m \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \tag{1.6}$$

Note that we are not requiring that our target distribution is Gaussian, rather, we are assuming that after sufficient time, the samples across all of our chains will be normally distributed according to the mean and variance of the target distribution. We show that under these assumptions, the quantity $B/W$ is distributed according to

the F-distribution $F_{M-1,M(N-1)}$. The second assumption allows us to construct the following quantities:

$$Z_n = \frac{X_m^{(n)} - \mu}{\sigma} \sim \mathcal{N}(0,1), \quad \bar{Z} = \frac{1}{N}\sum_{n=1}^{N} Z_n = \frac{\bar{X}_m - \mu}{\sigma}. \tag{1.7}$$

Now consider the sum of their squared differences:

$$\sum_{n=1}^{N}(Z_n - \bar{Z})^2 = \sum_{n=1}^{N}\frac{1}{\sigma^2}(X_m^{(n)} - \bar{X}_m)^2 = \frac{(N-1)s_m^2}{\sigma^2}. \tag{1.8}$$

A standard fact is that for $Z_1, \dots Z_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, the quantity $\sum_{n=1}^{N}(Z_n - \bar{Z})^2$ has a $\chi^2$ distribution with $N-1$ degrees of freedom. In our case, the $\{Z_n\}_{n=1}^{N}$ are not truly i.i.d. (they are positively correlated since they are generated by a Markov chain), but we assume that they are for a useful simplification. Another standard fact about chi-square random variables is that the sum of independent chi-square random variables has a chi-square distribution with additive degrees of freedom. We use this to write

$$\frac{M(N-1)W}{\sigma^2} = \sum_{m=1}^{M}\frac{(N-1)s_m^2}{\sigma^2} \sim \chi^2_{M(N-1)}. \tag{1.9}$$

Again using the second assumption, we can obtain similar quantities

$$Y_m = \frac{\sqrt{N}}{\sigma}(\bar{X}_m - \mu) \sim \mathcal{N}(0,1), \quad \bar{Y} = \frac{1}{M}\sum_{m=1}^{M} Y_m = \frac{\sqrt{N}}{\sigma}(\bar{X} - \mu). \tag{1.10}$$

Constructing the sum of their squared differences,

$$\sum_{m=1}^{M}(Y_m - \bar{Y})^2 = \frac{N}{\sigma^2}\sum_{m=1}^{M}(\bar{X}_m - \bar{X})^2 = \frac{(M-1)B}{\sigma^2}. \tag{1.11}$$

The chains are independent so we have that $X_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and we may write

$$\frac{(M-1)B}{\sigma^2} \sim \chi^2_{M-1}. \tag{1.12}$$

We now take the ratio of the chi-square random variables that we have constructed

$$\frac{B}{W} = \frac{\frac{(M-1)B}{\sigma^2}\cdot\frac{1}{(M-1)}}{\frac{M(N-1)W}{\sigma^2}\cdot\frac{1}{(M(N-1))}}. \tag{1.13}$$

The quantity $B$ is only a function of $\{\bar{X}_m\}_{m=1}^{M}$ and the quantity $W$ is only a function of $\{s_m^2\}_{m=1}^{M}$. Since the chains are independent, $\bar{X}_{m_1}$ is independent of $s_{m_2}^2$ when $m_1 \neq m_2$. When $m_1 = m_2$, independence still holds due to the fact that for normally distributed random variables, the sample mean is independent of the sample variance. Therefore, the numerator and denominator are independent chi-square random variables, and we have shown that $B/W \sim F_{M-1,M(N-1)}$. $\square$

### 1.3 Motivation

A random variable $X$ follows an $F$ distribution if

$$X = \frac{Y_1/d_1}{Y_2/d_2} \tag{1.14}$$

where $Y_1, Y_2$ are independent chi-square random variables with $d_1, d_2$ degrees of freedom, respectively. $X$ has expectation and variance

$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}, \quad \text{Var}(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}. \tag{1.15}$$

As a sanity check, consider the extreme case in which $d_1, d_2 \to \infty$. By the above formulas, $\text{Var}(X) \to 0$ and $\mathbb{E}[X] \to 1$. In our context, as we increase the number of chains $M$ and the number of samples $N$ from each chain, $B/W \to 1$ and we are left with $(N - 1 + 1)/N = 1$. This is why we perform a hypothesis test for $R = 1$ to assess if our chains have reached stationarity.

### 1.4 Connection to the Effective Sample Size

In class we defined the effective sample size in the context of importance sampling to characterize weight degeneracy [15]. But effective sample size can also be used for MCMC methods to quantify the amount of information lost due to autocorrelation between samples. In Ref. [9], R.M. Neal endorses using the quantity

$$\text{ESS} := \frac{MN}{1 + 2\sum_{k=1}^{\infty} \rho_k} \tag{1.16}$$

as the effective sample size, where $\rho_k$ is the lag $k$ correlation. The appendix of the lecture notes [15] defines

$$\tau_N^2 := N \cdot \text{Var}(\bar{X}_m) = \sigma^2 \left( 1 + 2 \sum_{k=1}^{N-1} \frac{N - k}{N} \rho_k \right), \quad \tau^2 := \lim_{N \to \infty} \tau_N^2 \tag{1.17}$$

where $\tau^2$ is known as the asymptotic variance. So we can equivalently rewrite the effective sample size as

$$\text{ESS} = \frac{MN\sigma^2}{\tau^2} \tag{1.18}$$

Ref. [16] points out that

$$\widehat{R}^2 = \frac{N - 1}{N} + \frac{1}{N}\frac{B}{W} \approx 1 + \frac{1}{N}\frac{\tau^2}{\sigma^2} = 1 + \frac{M}{\text{ESS}} \tag{1.19}$$

In practice this is helpful because we may have a desired effective sample size a priori, and this can be used to specify a cutoff $\delta > 1$ for our convergence hypothesis test. Of course, one still needs to construct an estimator $\widehat{\text{ESS}}$; this is addressed in Ref. [17].

## 2 Related Variants and Improvements to Gelman-Rubin

This chapter explores some of the shortcomings of the GR statistic and modifications that seek to mitigate them. We conclude by discussing related convergence diagnostics.

## 2.1 Addressing the Independence Assumption

### 2.1.1 The Consequence of Ignoring Autocorrelation

The first of two major assumptions that we made in the derivation of GR was that the samples are independent. To see the consequences of this, we first prove the useful fact mentioned in Ref. [16] that

$$\mathbb{E}[s_m^2] = \frac{N}{N-1}\Big(\sigma^2 - \mathrm{Var}\Big(\bar{X}_m\Big)\Big) \tag{2.1}$$

Begin by writing $X_m^{(n)} - \bar{X}_m = (X_m^{(n)} - \mu) - (\bar{X}_m - \mu)$. Then

$$\sum_{n=1}^{N}(X_m^{(n)} - \bar{X}_m)^2 = \sum_{n=1}^{N}(X_m^{(n)} - \mu)^2 + N(\bar{X}_m - \mu)^2 - 2(\bar{X}_m - \mu)\sum_{n=1}^{N}(X_m^{(n)} - \mu) \tag{2.2}$$

$$= \sum_{n=1}^{N}(X_m^{(n)} - \mu)^2 + N(\bar{X}_m - \mu)^2 - 2N(\bar{X}_m - \mu)^2 \tag{2.3}$$

$$= \sum_{n=1}^{N}(X_m^{(n)} - \mu)^2 - N(\bar{X}_m - \mu)^2 \tag{2.4}$$

and

$$\mathbb{E}[s_m^2] = \frac{1}{N-1}\sum_{n=1}^{N}\mathbb{E}\Big[(X_m^{(n)} - \mu)^2\Big] - N\Big[(\bar{X}_m - \mu)^2\Big] \tag{2.5}$$

$$= \frac{N}{N-1}\Big(\sigma^2 - \mathrm{Var}(\bar{X}_m)\Big) \quad \square \tag{2.6}$$

Because our samples are generated from a Markov chain, autocorrelation causes $\mathrm{Var}(\bar{X}_m) > \frac{\sigma^2}{N}$. So the independence assumption causes $\mathrm{Var}(\bar{X}_m)$ to be underestimated which means that $\mathbb{E}[s_m^2]$ is underestimated, and the within-chain variance $W$ is underestimated. As Ref. [16] explains, the final estimator for the posterior variance $V$ corrects this by adding in the bias correction term $B/N$. Once this correction is added in, $V$ now overestimates $\sigma^2$, but since $s_m^2$ underestimates $\sigma^2$, the statistic $\widehat{R} = \sqrt{V/W}$ approaches 1 from above as $n \to \infty$.

Still, a more efficient estimator than $B$ is highly desirable, particularly one that accounts for the correlation between samples in the chain. Ref. [16] addresses this by using recent developments in MCMC variance estimation. Recall that the quantity $B$ is just the unbiased sample variance of the means of our $M$ chains. This means that $B$ is an estimator of the quantity $\tau_N^2/N = \mathrm{Var}(\bar{X}_m)$.

### 2.1.2 The Lugsail Replicated Batch Means Estimator

Vats and Knudson [16] propose replacing $B$ with a more efficient estimator known as the *lugsail replicated batch means estimator*, which we present here. The idea is that, by estimating the variance of batches means from segments of the chains, we capture

the autocorrelation within each chain better. We write $N = a \cdot b$ for batch size $b$ and number of batches $a$. The mean of batch $i \in \{1, \ldots, a\}$ is

$$\bar{X}_{mi} := \frac{1}{b} \sum_{j=(i-1)b+1}^{ib} X_m^{(j)} \tag{2.7}$$

which is used in the replicated batch means estimator of $\tau_N^2$:

$$\widehat{\tau}_b^2 := \frac{b}{aM-1} \sum_{m=1}^{M} \sum_{i=1}^{a} \left(\bar{X}_{mi} - \bar{X}\right)^2. \tag{2.8}$$

The replicated batch means estimator is strongly consistent, but one issue that it has is that it is biased from below. However, we want an estimator that is biased from above to avoid incorrectly diagnosing convergence (recall that $W$ is already being underestimated). To avoid this, the authors specifically opted for the lugsail version, given by

$$\widehat{\tau}_L^2 := 2\widehat{\tau}_b^2 - \widehat{\tau}_{\lfloor b/3 \rfloor}^2 \tag{2.9}$$

which is indeed biased from above and also strongly consistent. As desired, this estimator is more efficient than $B$. As $N \to \infty$, the variance of $B$ approaches $2\tau^4/(m-1)$ while the variance of $\widehat{\tau}_L^2$ approaches $6\tau^4/am$. So the relative efficiency of $\widehat{\tau}_L^2$ with respect to $B$ is $am/(3m-3)$, but since the number of batches $a$ grows with $N$, this ratio approaches infinity as $N \to \infty$ [16].

### 2.2 Addressing the Normality Assumption

The second major assumption we made in the derivation of the GR statistic was that the samples from our chains are asymptotically normally distributed according to the mean and variance of our target distribution. The authors of Ref. [18] use *rank-normalization* to enforce the validity of this assumption. They mention that this technique is particularly helpful when the target distribution $D$ has heavy tails. The authors first assign ranks $\{R_m^{(n)}\}$ to the samples $\{X_m^{(n)}\}$ by putting them in order. Samples that have the same value are each identically assigned the average of their ranks. Next, they apply Blom's inverse-normal transformation [1, p. 71] to the ranks to obtain normal scores $\{Z_m^{(n)}\}$:

$$Z_m^{(n)} := \Phi^{-1}\left(\frac{R_m^{(n)} - 3/8}{MN + 1/4}\right) \tag{2.10}$$

This gives us access to approximately normally distributed data, matching our second assumption in the derivation. If the samples were already normally distributed, then this transformation has no effect. Furthermore, the ranks $\{R_m^{(n)}\}$ can be visualized with histograms for each individual chain. These plots may be more interpretable than trace plots, which can become hard to read for large numbers of samples. If each chain is mixing well, the ranks of its samples should be approximately uniform with small, local fluctuations between the height of each bin. A separate and more recent work also studies quantile-based convergence diagnostics as an extension of GR [11].

## 2.3 Other Extensions of the GR Statistic

### 2.3.1 Splitting

*Split-*$\widehat{R}$, originally described in [4, p. 284], and formally presented in [18], is a very simple but powerful idea: take the $M$ chains, split each in half, and treat them as $2M$ separate chains. As explained in [4], this tests for both mixing and stationarity. If each sub-chain mixes well, then the original chain is mixing well. Similarly, if each sub-chain is exploring the same distribution, then the original chain was likely at stationarity.

### 2.3.2 Folding

The authors of Ref. [18] also consider the *folded* samples $\{\zeta_m^{(n)}\}$:

$$\zeta_m^{(n)} := \left| X_m^{(n)} - \text{median}(X) \right|. \tag{2.11}$$

This modification is designed to address the scenario in which chains are exploring the same central region of the distribution, but with large differences in variance. This may indicate that some chains are stuck and not exploring the tails. Folding allows us to measure convergence in these tails. The authors of [18] also suggest combining rank-normalization, splitting, and folding to reap the benefits of all three of these modifications. We elaborate on and experimentally validate this approach with synthetic datasets in Sec 3.2.

### 2.3.3 Nesting

Advances in computer hardware, particularly GPUs, gives researchers the ability to run large quantities of chains in parallel. This motivates the development of a GR statistic tailored to scenarios in which one has a large quantity of chains that are shorter than usual. Ref. [10] proposes *nested* $\widehat{R}$ or $\widehat{R}_\nu$. This method introduces another layer of abstraction called a *superchain*, which is a collection of chains initialized at the same point. After estimating the between-superchain variance $\widehat{B}_\nu$ and within-superchain variance $\widehat{W}_\nu$, the estimator

$$\widehat{R}_\nu := \sqrt{1 + \frac{\widehat{B}_\nu}{\widehat{W}_\nu}} \tag{2.12}$$

is used to assess convergence. The authors argue that $\widehat{R}_\nu$ is less likely to confuse short chain length for a lack of convergence.

### 2.3.4 Multivariate GR

Brooks and Gelman [2] proposed a multivariate version of the GR statistic that can be used when the target distribution has dimension $d > 1$ and one has samples of the form $\mathbf{X}_m^{(n)} = (X_{m1}^{(n)}, \ldots, X_{md}^{(n)})^T \in \mathbb{R}^d$. Accordingly, let $\bar{\mathbf{X}}_m$, $\bar{\mathbf{X}}$, and $\mathbf{S}_m$ be the multivariate generalizations of the quantities defined in Sec. 1.1. Similarly, let

$$\mathbf{B} = \frac{N}{M-1} \sum_{m=1}^{M} \left(\bar{\mathbf{X}}_m - \bar{\mathbf{X}}\right)\left(\bar{\mathbf{X}}_m - \bar{\mathbf{X}}\right)^T, \quad \mathbf{W} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{S}_m. \tag{2.13}$$

Then the $d$-dimensional GR statistic is given by

$$\widehat{R}^d = \sqrt{\frac{N-1}{N} + \frac{M+1}{MN}\lambda_{\max}(\mathbf{W}^{-1}\mathbf{B})} \tag{2.14}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Vats and Knudson [16] also update the multivariate GR statistic with a *multivariate replicated lugsail batch means estimator*, analogous to their contribution to the 1-dimensional GR statistic outlined in Sec. 2.1.2.

## 2.4 Related Works

For the purpose of breadth, we also discuss a few more notable convergence diagnostics that are not part of the GR lineage. We omit discussion of diagnostics proposed by Geweke [5] and Raftery and Lewis [14] since these are covered by the lecture notes [15].

### 2.4.1 Cusum Path Plots

Opposed to the GR statistic, which considers multiple chains, Yu and Mykland [19] propose monitoring the convergence of a single chain using any 1-dimensional summary statistic of interest. Let $X = X^{(1)}, \ldots, X^{(N)}$ denote the chain and $T(X)$ be the 1-dim summary statistic. It is important to discard some quantity $N_0$ of burn-in samples since these will have high initial bias. Compute the sample mean $\widehat{\mu}$ of $T(X)$ and then use this to calculate the centered cumulative sum (cusum) for all $t = N_0 + 1, \ldots, N$:

$$\widehat{\mu} = (N - N_0)^{-1} \sum_{n=N_0+1}^{N} T(X^{(n)}), \quad \widehat{S}_t = \sum_{n=N_0+1}^{t} \left(T(X^{(n)}) - \widehat{\mu}\right) \tag{2.15}$$

Finally, trace a curve through all $\{\widehat{S}_t\}$ to visualize the trajectory of the cusum. The mixing rate of $T(X)$ can be seen by examining the shape of this plot. Fast mixing results in a jagged path while slow mixing yields a smoother path. The cusum plot is particularly good at revealing drift and sticky regions (where the chain may get stuck). Still, the authors recommend the cusum plot as a complementary tool to trace plots that provide a second opinion on the convergence and mixing of the chain.

### 2.4.2 Regenerative Methods

Another way to study the convergence of MCMC methods is by identifying times $T_0, T_1, \ldots$ at which the chain restarts called regeneration times [12]. The segments between these times are interdependent and identically distributed and known as tours. For discrete state spaces, these times can be obtained by fixing a state and tracking whenever the chain returns to the state. Identifying these times is more challenging for continuous state spaces, and a technique called splitting from Nummelin [13] (no relation to split-$\widehat{R}$) can be used instead.

At a high level, Nummelin's splitting technique works by performing Bernoulli trials at every step of the chain. Each trial has a small probability of success, and when a success does occur, the chain is said to have entered a set of states called an *atom*, at which the chain's memory resets. This is recorded as a regeneration time. Once many

of these regeneration times have been collected, a scaled regeneration quantile (SRQ) plot can be created, in which we graph $T_i/T_R$ over $i/R$ for all $i = 1, \ldots, R$. If the chain has been run for long enough and has reached stationarity, then this plot will produce a straight line with slope 1 through the origin.

### 2.5 Heidelberger–Welch Diagnostic

The Heidelberger–Welch (HW) Diagnostic [7, 8] was originally a procedure that was designed to be performed while a MCMC simulation was running to help determine when it should stop. In practice, it is now mainly used post-hoc to determine if a chain has converged or needs to be run for longer. We present a modern interpretation of the HW diagnostic used by the Bayesian Inference package `LaplacesDemon` [6]. One unique feature of this diagnostic is that it helps the user decide the point at which to cut off the burn-in. This is different from most of the other convergence diagnostics we have discussed, in which a user provides a chain that has been truncated by hand. The HW diagnostic consists of two phases:

**(1)** Generate a chain $\{X^{(n)}\}_{n=1}^{N}$ and define five checkpoints $\{\lfloor (i-1) \cdot N/10 \rfloor\}_{i=1}^{5}$. Test the sub-chain $\{X^{(n)}\}_{n=C_i+1}^{N}$ for stationarity. If the test fails, increment $i$ and repeat until a chain is accepted. If all tests fail, keep running the chain. **(2)** If a stationary sub-chain is identified, use this to calculate the sample mean and construct a confidence interval. If the ratio of the half-length of the interval and sample mean is less than some tolerance $\epsilon$, the chain passes the accuracy test.

The details of the confidence interval and stationarity hypothesis test are given in [7] and [8], respectively. The documentation for `LaplacesDemon` raises the concern that generalizing this diagnostic to higher dimensions requires navigating the pitfalls of multiple testing. This may be one reason why this technique is not more popular. Still, the HW diagnostic may prove useful for lower-dimensional contexts in which simulation remains extremely expensive.

## 3 Assessing the Performance of Gelman-Rubin and its Extensions

In this chapter, we examine a few manufactured but realistic examples of target distributions that can fool the GR statistic $\widehat{R}$. We then apply the improvements introduced in Chapter 2 to determine which modified versions of $\widehat{R}$ can detect different types of convergence failures. We conclude with a brief discussion.

### 3.1 Benchmarking The Lugsail Replicated Batch Means Estimator

As in [16], we denote the Lugsail version of $\widehat{R}$ as $\widehat{R}_L$. For disambiguation, we refer to the original GR statistic as classical-$\widehat{R}$. To assess the performance of the Lugsail Replicated Batch Means Estimator modification to the GR statistic, we compare $\widehat{R}_L$ against classical-$\widehat{R}$ across three different scenarios using Random Walk Metropolis Hastings (RWMH), shown in Fig. 1. We use $M = 4$ chains and $N = 5,000$ samples.

1. **Low Autocorrelation** RWMH on a weakly correlated bivariate Gaussian ($\rho = .10$) at stationarity.

2. **High Autocorrelation** RWMH on a strongly correlated bivariate Gaussian ($\rho = .98$)

at stationarity.

3. **Two Modes** RWMH on a bimodal Gaussian mixture in which two chains are trapped in separate modes (non-stationary).

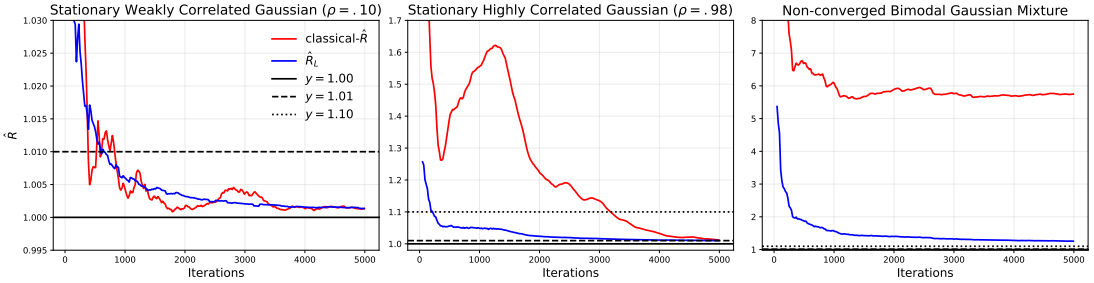


**Figure 1** A comparison of classical-$\widehat{R}$ and $\widehat{R}_L$ for RWMH on three distributions. The first two plots are for chains that have converged, and the third is for a chain that has not converged.

We make several remarks. First, in the weakly correlated setting with low autocorrelation, both statistics perform similarly after sufficiently many iterations, although $\widehat{R}_L$ has much smoother behavior. In the second plot, $\widehat{R}_L$ is substantially more robust to the autocorrelation. In the third plot, while classical-$\widehat{R}$ is much more conservative than $\widehat{R}_L$, both statistics remain safely above the generous $\delta = 1.1$ threshold that is sometimes used to assess convergence. From these experiments, $\widehat{R}_L$ is the better overall choice.

### 3.2 GR vs Split, Folded, and Rank-Normalized GR

In this section, we compare split-$\widehat{R}$, folded-$\widehat{R}$, and rank-normalized-$\widehat{R}$ with classical-$\widehat{R}$. Following Ref. [18], we combine these techniques into a signal diagnostic that we denote with $\widehat{R}_{\text{max}}$, computed with Algorithm 3.1. To compare the behavior of splitting, folding, and rank-normalization, we construct three synthetic datasets, each designed to mimic a convergence failure that a collection of MCMC chains might exhibit (shown in Fig 2). We use $M = 4$ sequences and $N = 1,000$ samples.

**Algorithm 3.1** Construction of $\widehat{R}_{\text{max}}$

1: **Input**: Chains $\{X_m^{(n)}\}$
2: $\widehat{R}_{\text{rank-normalized-split}} : \{X_m^{(n)}\} \to \text{split} \to \text{rank-normalize} \to \widehat{R}$
3: $\widehat{R}_{\text{folded-split}} : \{X_m^{(n)}\} \to \text{fold} \to \text{split} \to \text{rank-normalize} \to \widehat{R}$
4: $\widehat{R}_{\text{max}} = \max\left\{\widehat{R}_{\text{rank-normalized-split}}, \widehat{R}_{\text{folded-split}}\right\}$
5: **Output**: $\widehat{R}_{\text{max}}$

1. **Nonstationarity and Splitting** We generate $1,000$ samples from each of 4 independent Gaussians that each have a linear drift in their means. Although all sequences have the same drift behavior, splitting each sequence results in 8 subsequences

that differ substantially. This is designed to show the sensitivity of split-$\widehat{R}$ to nonstationarity.

2. **Variance Mismatch and Folding** We generate $1,000$ samples from each of 4 independent Gaussians, two of which have variance 1, and two of which have variance 2. This could resemble a situation in which two chains get stuck in the center of a distribution and do not explore the tails. Folding about the median amplifies this discrepancy.

3. **Heavy Tails and Rank-Normalization** We generate $1,000$ samples from each of 3 independent standard Cauchy distributions and $1,000$ samples from another standard independent Cauchy that has been right-shifted by 2. Since the classical GR diagnostic relies on normal asymptotics, it can struggle with heavy-tailed distributions. Rank-normalization can alleviate this and reveal that one of the sequences is behaving differently.

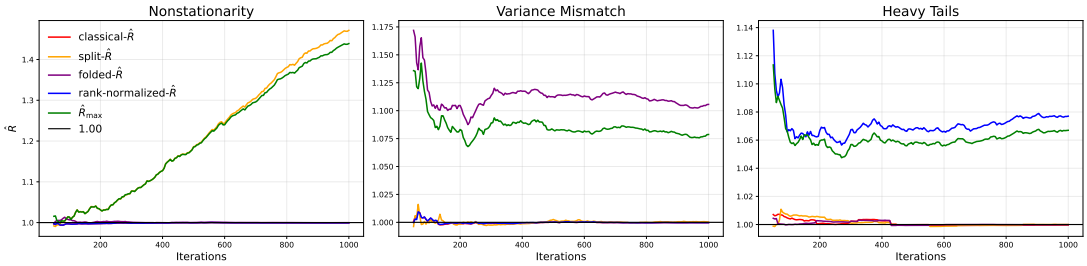


**Figure 2** A comparison of classical-$\widehat{R}$, split-$\widehat{R}$, folded-$\widehat{R}$, rank-normalized-$\widehat{R}$, and $\widehat{R}_{\max}$ on synthetic examples of non-convergence. Variants of $\widehat{R}$ that are working correctly will not be close to 1.

As we would hope, each of the individual techniques detects the convergence issue that it was designed to address. In each case, classical-$\widehat{R}$ fails to detect these issues, and remains close to 1. Although each individual technique is not effective outside of its intended settings, $\widehat{R}_{\max}$ quickly detects all three convergence issues. Crucially, $\widehat{R}_{\max}$ achieves this without being tailored to any specific failure mode.

### 3.3 Discussion and Bibliography

We have presented the GR statistic [3], some proposals to improve it [4, 16, 18, 10], as well as a few other convergence diagnostics [19, 12, 7, 8]. Furthermore, in Chapter 3, we are able to verify the merit of modifications to the GR statistic proposed by Vats [16] and Vehtari [18].

While no convergence diagnostic is free from flaws, increasingly robust tools continue to become available to researchers to characterize MCMC simulations. Motivated by the nested $\widehat{R}$ work [10], a promising future research direction could be the exploration of simulation algorithms and diagnostics designed to exploit modern parallel and distributed computer architectures.

## References


[1] G. Blom. *Statistical estimates and transformed beta-variables*. PhD thesis, Almqvist & Wiksell, 1958.

[2] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[3] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[4] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.

[5] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis, 1991.

[6] B. Hall, M. Hall, L. Statisticat, E. Brown, R. Hermanson, E. Charpentier, D. Heck, S. Laurent, Q. F. Gronau, H. Singmann, et al. Package 'laplacesdemon'. *Complete environment for Bayesian inference. Version*, 16(4), 2020.

[7] P. Heidelberger and P. D. Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24 (4):233–245, 1981.

[8] P. Heidelberger and P. D. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[9] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

[10] C. C. Margossian, M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman. Nested r $\widehat{}$: Assessing the convergence of markov chain monte carlo when running many short chains. *Bayesian Analysis*, 20(4):1587–1614, 2025.

[11] T. Moins, J. Arbel, A. Dutfoy, and S. Girard. On the use of a local r $\widehat{}$ to improve mcmc convergence diagnostic. *Bayesian analysis*, 20(1):131–156, 2025.

[12] P. Mykland, L. Tierney, and B. Yu. Regeneration in markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241, 1995.

[13] E. Nummelin. A splitting technique for harris recurrent markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318, 1978.

[14] A. E. Raftery and S. Lewis. How many iterations in the gibbs sampler? 1991.

[15] D. Sanz-Alonso and O. Al-Ghattas. A first course in monte carlo methods. *arXiv preprint arXiv:2405.16359*, 2024.

[16] D. Vats and C. Knudson. Revisiting the gelman–rubin diagnostic. *Statistical Science*, 36(4):518–529, 2021.

[17] D. Vats, J. M. Flegal, and G. L. Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337, 2019.

[18] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

[19] B. Yu and P. Mykland. Looking at markov samplers through cusum path plots: a simple diagnostic idea. *Statistics and Computing*, 8(3):275–286, 1998.