# $\widehat{R}$ does not work! Building better ways to assess the convergence of iterative algorithms

*Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, Paul Bürkner*

4 March 2019

**Abstract**

Markov chain Monte Carlo is the main computational tool for Bayesian statistics, however it is known that we must carefully monitor the convergence behaviour of the chain lest poor computation bias our estimates. In this paper we show that the most commonly used convergence diagnostic $\widehat{R}$ of Gelman and Rubin (1992) has serious flaws and we show ways around them. We also introduce a version of $\widehat{R}$ which is more sensitive to scale differences between chains, CDF and quantile based local efficiency measures, and a practical approach for computing Monte Carlo error estimates for quantiles. We suggest that common trace plots should be replaced with rank plots from multiple chains.

## 1 Introduction

Markov chain Monte Carlo (MCMC) methods form the core of many methods in computational statistics, especially in Bayesian applications where the goal is to represent posterior inference using a sample of posterior draws. While MCMC, as well as more general iterative simulation algorithms, can usually be proven to converge to the target distribution as the number of draws approaches infinity, there are rarely strong guarantees about their behaviour after a finite number of draws. In fact, decades of experience tell us that the finite sample behaviour of these algorithms can be almost arbitrarily bad.

In an attempt to assuage our fear that our MCMC algorithm may not have converged, we typically run multiple independent chains to see if the obtained distribution is similar across chains. We typically also visually inspect the sample path of the chain as well as some summary statistics such as investigating the empirical autocorrelation function.

Running multiple chains is critical to any MCMC convergence diagnostic. Figure 1 illustrates two ways in which sequences of iterative simulations can fail to converge. In the first example, two chains are in different parts of the target distribution, in the second example, the chains move but have not attained stationarity. Slow mixing can arise with multimodal target distributions or when a chain is stuck in a region of high curvature with a step size too large to make an acceptable proposal for the next step. The two examples in Figure 1 make it clear that any method for assessing mixing and effective sample size should use information between and within chains.

As we are often fitting models with large numbers of parameters, it is not realistic to expect to make and interpret trace plots such as in Figure 1 for all quantities of interest. Hence we need numerical summaries that can flag potential problems.

Probably the most widely-used attempt to construct a more formally justified convergence diagnostic is the $\widehat{R}$ statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) and split-$\widehat{R}$ (Gelman et al., 2013). These quantities monitor the ratio of the within-chain marginal variances and the between-chain marginal variances of the simulations. The idea is that if a chain has not converged to stationarity, these quantities will be quite different. The difference between the original $\widehat{R}$ and split-$\widehat{R}$ is that the latter also compares the marginal distributions of the first half of each chain with the second half of each chain, to try to ensure that each chain has itself converged. In this paper we will only consider split-$\widehat{R}$.

The split-$\widehat{R}$ is most effective if it is computed using multiple chains initialized at a diverse set of starting points. This is to reduce the dependence of diagnostic on the starting point of the chain and reduce the chance that we falsely diagnose the chain as converging when beginning at a different point would lead to a qualitatively different posterior.

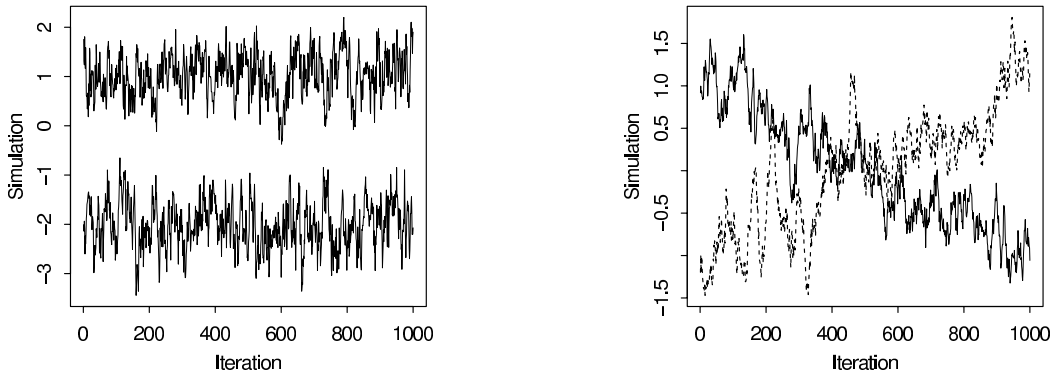


Figure 1: *Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the two sequences happen to cover a common distribution but neither sequence appears stationary. These graphs demonstrate the need to use between-sequence and also within-sequence information when assessing convergence. From Gelman et al. (2013)).*

In the context of Markov chain Monte Carlo, one can interpret split-$\widehat{R}$ with diverse seeding as an operationalization of the qualitative statement that the convergence of the Markov chain is relatively insensitive to the starting point (at least within a reasonable part of the parameter space). This is the closest we can come to verifying empirically that the Markov chain is geometrically ergodic, which is a critical property if we want a central limit theorem to hold for the approximate posterior expectations. Without this, we have no control over the large deviation behaviour of the estimates and the constructed Markov chains will be useless for practical purposes.

The problem is that $\widehat{R}$ and split-$\widehat{R}$ do not reliably work when analyzing generic iterative algorithms. This is particularly a problem when we use them within generic software packages like `Stan` or analysis tools like `R`'s `coda` library. The following example shows how the failure occurs.

**Example 1** *Figure*

We have identified two problems with split-$\widehat{R}$:

1. If the the chains don't have finite mean, $\widehat{R} \approx 1$ even if one of the chains has a different location parameter to the others;
2. If two of the chains have different variances but the same mean parameters, $\widehat{R} \approx 1$.

This is bad. It means that if we rely on the standard split-$\widehat{R}$ estimate, we incorrectly conclude that the chain is behaving well even when it is not.

In this paper, we propose improvements of split-$\widehat{R}$ that overcome the two problems described above. Furthermore, we show that the convergence of the Markov chain need not be uniform across the parameter space and propose a localized version of both split-$\widehat{R}$ and an effective sample size calculation that allow us to assess better the behaviour of localized functionals of the chain. Finally, we propose three new methods to visualize the convergence of an iterative algorithm that are more informative than standard trace plots.

The paper begins by reviewing the standard split-$\widehat{R}$ and effective samples size calculations, before proposing improvements that repair the problems noted above. We then propose some localized diagnostics and some new diagnostic plots. Finally, we show the behavior of the new diagnostics on a variety of examples.

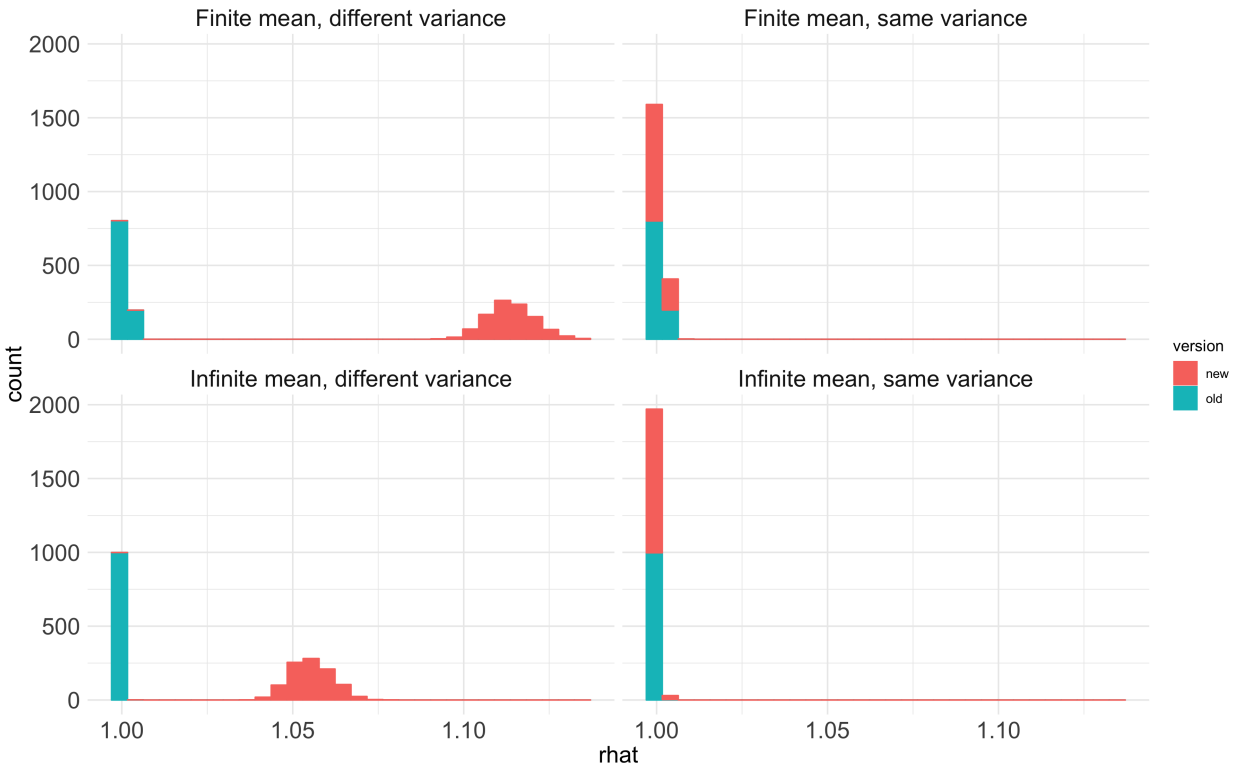


Figure 2: Caption.

# 2 Convergence diagnostics for iterative algorithms

## 2.1 Split-$\widehat{R}$

The original $\widehat{R}$ statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) and split-$\widehat{R}$ (Gelman et al., 2013) are both based on the ratio of between and within-chain marginal variances of the simulations, while the latter is computed from split chains (hence the name).

Here we present split-$\widehat{R}$, following Gelman et al. (2013), but using the notation of Stan Development Team (2018c). This implementation represents the current standard in convergence diagnostics for iterative simulations. In the equations below, $N$ is the number of draws per chain, $M$ is the number of chains, and $S = MN$ is the total number of draws from all chains. For each scalar summary of interest $\theta$, we compute $B$ and $W$, the between- and within-chain variances:

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\overline{\theta}^{(.m)} - \overline{\theta}^{(..)})^2, \quad \text{where} \quad \overline{\theta}^{(.m)} = \frac{1}{N}\sum_{n=1}^{N}\theta^{(nm)}, \quad \overline{\theta}^{(..)} = \frac{1}{M}\sum_{m=1}^{M}\overline{\theta}^{(.m)} \tag{1}$$

$$W = \frac{1}{M}\sum_{m=1}^{M}s_j^2, \quad \text{where} \quad s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta^{(nm)} - \overline{\theta}^{(.m)})^2. \tag{2}$$

The between-chain variance, $B$, also contains the factor $N$ because it is based on the variance of the within-chain means, $\overline{\theta}^{(.m)}$, each of which is an average of $N$ values $\theta^{(nm)}$. We can estimate $\text{var}(\theta|y)$, the marginal posterior variance of the estimand, by a weighted average of $W$ and $B$, namely,

$$\widehat{\text{var}}^+(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B. \tag{3}$$

This quantity *overestimates* the marginal posterior variance assuming the starting distribution of the simulations is appropriately overdispersed compared to the target distribution, but is *unbiased* under stationarity (that is, if the starting distribution equals the target distribution), or in the limit $N \to \infty$. To have an overdispersed starting distribution, independent Markov chains should be initialized with diffuse starting values for the parameters.

Meanwhile, for any finite $N$, the within-chain variance $W$ should *underestimate* $\text{var}(\theta|y)$ because the individual chains haven't had the time to explore all of the target distribution and, as a result, will have less variability. In the limit as $N \to \infty$, the expectation of $W$ also approaches $\text{var}(\theta|y)$.

We monitor convergence of the iterative simulations to the target distribution by estimating the factor by which the scale of the current distribution for $\theta$ might be reduced if the simulations were continued in the limit $N \to \infty$. This potential scale reduction is estimated as,

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\theta|y)}{W}}, \tag{4}$$

which for an ergodic process declines to 1 as $N \to \infty$. We call this split-$\widehat{R}$ because we are applying it to chains that have been split in half so that $M$ is twice the number of actual chains. Without splitting, $\widehat{R}$ would get fooled by non-stationary chains as in Figure 1b.

Split-$\widehat{R}$ is also well defined for sequences that are not Markov chains. However, for simplicity, we always refer to "chains" instead of more generally to "sequences" as the former is our primary use case for $\widehat{R}$-like measures.

## 2.2 Effective sample size

If the $N$ simulation draws within each chain were truly independent, the between-chain variance $B$ would be an unbiased estimate of the posterior variance, $\text{var}(\theta|y)$, and we would have a total of $S = MN$ independent simulations from the $M$ chains. In general, however, the simulations of $\theta$ within each chain will be autocorrelated, and thus $B$ will be larger than $\text{var}(\theta|y)$, in expectation.

One way to define effective sample size for correlated simulation draws is to consider the statistical efficiency of the average of the simulations $\bar{\theta}^{(\cdot\cdot)}$ as an estimate of the posterior mean $\text{E}(\theta|y)$. This also generalizes to posterior expectations of functionals of parameters $\text{E}(g(\theta)|y)$. We return later to how to estimate the effective sample size of quantiles which cannot be presented as expectations. For simplification, in this section we consider the effective sample size for the posterior mean.

The effective sample size of a chain is defined in terms of the autocorrelations within the chain at different lags. The autocorrelation $\rho_t$ at lag $t \geq 0$ for a chain with joint probability function $p(\theta)$ with mean $\mu$ and standard deviation $\sigma$ is defined to be

$$\rho_t = \frac{1}{\sigma^2} \int_\Theta (\theta^{(n)} - \mu)(\theta^{(n+t)} - \mu)\, p(\theta)\, d\theta. \tag{5}$$

This is just the correlation between the two chains offset by $t$ positions. Because we know $\theta^{(n)}$ and $\theta^{(n+t)}$ have the same marginal distribution in an MCMC setting, multiplying the two difference terms and reducing yields

$$\rho_t = \frac{1}{\sigma^2} \int_\Theta \theta^{(n)}\, \theta^{(n+t)}\, p(\theta)\, d\theta. \tag{6}$$

The effective sample size of one chain generated by a process with autocorrelations $\rho_t$ is defined by

$$N_{\text{eff}} = \frac{N}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}. \tag{7}$$

The effective sample size $N_{\text{eff}}$ can be larger than $N$ in case of antithetic Markov chains, which have negative autocorrelations on odd lags. The dynamic Hamiltonian Monte Carlo algorithms used in Stan (Hoffman and Gelman, 2014; Betancourt, 2017) can produce $N_{\text{eff}} > N$ for parameters with a close to Gaussian posterior (in the unconstrained space) and low dependence on the other parameters.

In practice, the probability function in question cannot be tractably integrated and thus neither autocorrelation nor the effective sample size can be directly calculated. Instead, these quantities must be estimated from the samples themselves. The rest of this section describes an autocorrelation and split-$\widehat{R}$ based effective sample size estimator, based on multiple split chains. For simplicity, each chain will be assumed to be of the same length $N$.

Computations of autocorrelations for all lags simultaneously can be done via the fast Fourier transform algorithm (FFT; see Geyer, 2011). The autocorrelation estimates $\hat{\rho}_{t,m}$ at lag $t$ from multiple chains $m \in (1, \ldots, M)$ are combined with the within-chain variance estimate $W$ and the multi-chain variance estimate $\widehat{\text{var}}^+$ introduced above to compute the combined autocorrelation at lag $t$ as,

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{m=1}^{M} \hat{\rho}_{t,j}}{\widehat{\text{var}}^+}. \tag{8}$$

If the chains have not converged, the variance estimator $\widehat{\text{var}}^+$ will overestimate the true marginal variance which leads to an overestimation of the autocorrelation and an underestimation of the effective sample size.

Because of noise in the correlation estimates $\hat{\rho}_t$ increases as $t$ increases, typically the truncated sum of $\hat{\rho}_t$ is used. Negative autocorrelations can happen only on odd lags and by summing over pairs starting from lag $t = 0$, the paired autocorrelation is guaranteed to be positive, monotone and convex modulo estimator noise (Geyer, 1992, 2011). The effective sample size of combined chains is then defined as,

$$S_{\text{eff}} = \frac{N\,M}{\hat{\tau}}, \tag{9}$$

where

$$\hat{\tau} = 1 + 2\sum_{t=1}^{2k+1} \hat{\rho}_t = -1 + 2\sum_{t'=0}^{k} \hat{P}_{t'}, \tag{10}$$

and $\hat{P}_{t'} = \hat{\rho}_{2t'} + \hat{\rho}_{2t'+1}$. The initial positive sequence estimator is obtained by choosing the largest $k$ such that $\hat{P}_{t'} > 0$ for all $t' = 1, \ldots, k$. The initial monotone sequence estimator is obtained by further reducing $\hat{P}_{t'}$ to the minimum of the preceding values so that the estimated sequence becomes monotone.

The effective sample size $S_{\text{eff}}$ described here is different from similar formulas in the literature in that we use multiple chains and between-chain variance in the computation, which typically gives us more conservative claims (lower values of $S_{\text{eff}}$) compared to single chain estimates, especially when mixing of the chains is poor. If the chains are not mixing at all (e.g., if the posterior is multimodal and the chains are stuck in different modes), then our $S_{\text{eff}}$ is close to the number of chains.

## 2.3 Problems with current diagnostics

There are two main problems with the diagnostics as described so far, which we will highlight below and illustrate in detail in Section 4 as well as in the online appendix[1]. For each problem, we propose possible

[1] https://avehtari.github.io/rhat_ess/

solutions that, together, will result in a new set of convergence diagnostics and recommendations for reasonable criteria to determine (non-)convergence.

The first problem is that split-$\widehat{R}$ and $S_{\text{eff}}$ are well defined only if the marginal distribution of the quantity of interest has finite mean and variance, which is not necessarily the case and may not even be immediately clear based on the obtained posterior draws. Split-$\widehat{R}$ and $S_{\text{eff}}$ can also be unstable when the mean and variance are finite if the marginal distribution has thick tails.

The second problem is that split-$\widehat{R}$ and $S_{\text{eff}}$ are usually computed only for the posterior mean. While this provides an estimate for the efficiency in the bulk of the distribution, it says little about the efficiencies in the tails, which is a concern for posterior interval estimates as well as for inferences about rare events.

# 3 Improving convergence diagnostics

## 3.1 Rank normalization

As split-$\widehat{R}$ and $S_{\text{eff}}$ are well defined only if the marginal posteriors have finite mean and variance, we propose to use rank normalized parameter values instead of the actual parameter values for the purpose of diagnosing convergence.

Rank normalized split-$\widehat{R}$ and $S_{\text{eff}}$ are computed using the equations in Section 2, but replacing the original parameter values $\theta^{(nm)}$ with their corresponding rank normalized values denoted as $z^{(nm)}$. Rank normalization is done a follows: First, replace each value $\theta^{(nm)}$ by its rank $r^{(nm)}$. Average rank for ties are used to conserve the number of unique values of discrete quantities. Ranks are computed jointly for all draws from all chains. Second, normalize ranks via the inverse normal transformation,

$$z^{(nm)} = \Phi^{-1}((r^{(nm)} - 0.5)/S). \tag{11}$$

For continuous variables and $S \to \infty$, the rank normalized values are normally distributed. Using normalized ranks $z^{(nm)}$ instead of ranks $r^{(nm)}$ themselves has the additional benefit that the behavior of $\widehat{R}$ and $S_{\text{eff}}$ do not change for normally distributed parameters. See online appendix for illustration of rank normalization.

We will use the term *bulk effective sample size* (bulk-ESS or bulk-$S_{\text{eff}}$) to refer to the effective sample size based on the rank normalized draws. Bulk-ESS is useful for diagnosing problems due to trends or different locations of the chains (see Appendix). Further, it is well defined even for distributions with infinite mean or variance, a case where previous ESS estimates fail. However, due to the rank normalization, Bulk-ESS is no longer directly applicable to estimate the Monte Carlo standard error of the posterior mean. We will come back to the issue of computing Monte Carlo standard errors for relevant quantities in Section 3.6.

## 3.2 Diagnostics for folded draws

Both original and rank normalized split-$\widehat{R}$ can be fooled if the chains have the same location but different scales, which can happen if one or more chains is stuck near the middle of the distribution. To alleviate this problem, we propose to compute a rank normalized split-$\widehat{R}$ statistic not only for the original draws $\theta^{(nm)}$, but also for the corresponding *folded* draws $\zeta^{(mn)}$, absolute deviations from the median,

$$\zeta^{(mn)} = \text{abs}(\theta^{(nm)} - \text{median}(\theta)). \tag{12}$$

We label as *folded-split-*$\widehat{R}$ the rank normalized split-$\widehat{R}$ measure computed on the basis of $\zeta^{(mn)}$. It measures convergence in the tails rather than in the bulk of the distribution. To obtain a single conservative $\widehat{R}$ estimate, we propose to report the maximum of rank normalized split-$\widehat{R}$ and rank normalized folded-split-$\widehat{R}$ for each parameter.

### 3.3 Convergence diagnostics for quantiles

The new $\widehat{R}$ and bulk-ESS introduced above are useful as overall efficiency measures. Next we introduce convergence diagnostics for quantiles and related quantities, which are more focused measures and help to diagnose reliability of often reported posterior intervals. Estimating the efficiency of quantile estimates has a high practical relevance in particular as we observe the efficiency for tail quantiles to often be lower than for the mean or median.

The $\alpha$-quantile is defined as the parameter value $\theta_\alpha$ for which $p(\theta \leq \theta_\alpha) = \alpha$. An estimate $\hat{\theta}_\alpha$ of $\theta_\alpha$ can thus be obtained by finding the $\alpha$-quantile of the empirical cumulative distribution function (ECDF) of the posterior draws $\theta^{(s)}$. However, quantiles cannot be written as an expectation, and thus the above equations for $\widehat{R}$ and $S_{\text{eff}}$ are not directly applicable. Thus, we first focus on the efficiency estimate for the cumulative probability $p(\theta \leq \theta_\alpha)$ for different values of $\theta_\alpha$.

For any $\theta_\alpha$, the ECDF gives an estimate of the cumulative probability,

$$p(\theta \leq \theta_\alpha) \approx \bar{I}_\alpha = \frac{1}{S} \sum_{s=1}^{S} I(\theta^{(s)} \leq \theta_\alpha), \tag{13}$$

where $I()$ is the indicator function. The indicator function transforms simulation draws to 0's and 1's, and thus the subsequent computations are bijectively invariant. Efficiency estimates of the ECDF at any $\theta_\alpha$ can now be obtained by applying rank-normalizing and subsequent computations directly on the indicator function's results.

Assuming that we know the CDF to be a certain continuous function $F$ which is smooth near an $\alpha$-quantile of interest, we could use the delta method to compute a variance estimate for $F^{-1}(\bar{I}_\alpha)$. Although we don't usually know $F$, the delta method approach reveals that the variance of $\bar{I}_\alpha$ for some $\theta_\alpha$ is scaled by the (usually unknown) density $f(\theta_\alpha)$, but the efficiency depends only on the efficiency of $\bar{I}_\alpha$. Thus, we can use the effective sample size for the ECDF (computed using the indicator function $I(\theta^{(s)} \leq \theta_\alpha)$) also for the corresponding quantile estimates. More details on the variance of the cumulative distribution function can be found in the online appendix.

To get a better sense of the sampling efficiency in the distributions' tails, we propose to compute the minimum of the effective sample sizes of the 5% and 95% quantiles, which we will call *tail effective sample size* (tail-ESS or tail-$S_{\text{eff}}$). Tail-ESS can help diagnosing problems due to different scales of the chains (see Appendix).

### 3.4 Efficiency estimates for the median absolute deviation

Since the marginal posterior distributions might not have finite mean and variance, by default rstan (Stan Development Team, 2018a) and rstanarm (Stan Development Team, 2018b) report median and median absolute deviation (MAD) instead of mean and standard error. Median and MAD are well defined even when the marginal distribution does not have finite mean and variance. Since the median is just the 50% quantile, we can get an efficiency estimate for it as for any other quantile.

Further, we can also compute an efficiency estimate for the median absolute deviation by computing the efficiency estimate of an indicator function based on the folded parameter values $\zeta$ (see Equation (12)):

$$p(\zeta \leq \zeta_{0.5}) \approx \bar{I}_{\zeta,0.5} = \frac{1}{S} \sum_{s=1}^{S} I(\zeta^{(s)} \leq \zeta_{0.5}), \tag{14}$$

where $\zeta_{0.5}$ is the median of the folded values. The efficiency estimate for the MAD is obtained by applying the same approach as for the median (and other quantiles) but with the folded parameters values.

### 3.5 Efficiency estimates for small interval probability estimates

We can get more local efficiency estimates by considering small probability intervals. We propose to compute the efficiency estimates for

$$\bar{I}_{\alpha,\delta} = p(\hat{Q}_\alpha < \theta \leq \hat{Q}_{\alpha+\delta}), \tag{15}$$

where $\hat{Q}_\alpha$ is an empirical $\alpha$-quantile, $\delta = 1/k$ is the length of the interval with some positive integer $k$, and $\alpha \in (0, \delta, \ldots, 1-\delta)$ changes in steps of $\delta$. Each interval has $S/k$ draws, and the efficiency measures the autocorrelation of an indicator function which is 1 when the values are inside the specific interval and 0 otherwise. This gives us a local efficiency measure which does not depend on the shape of the distribution.

### 3.6 Monte Carlo error estimates for quantiles

It is common practice to only report the Monte Carlo error of the mean, but not of quantiles and related quantities. As the delta method for computing the variance would require explicit knowledge of the normalized posterior density, which we don't have in most non-trivial cases, we propose the following alternative approach to compute Monte Carlo standard errors of quantiles:

1. Compute quantiles of the beta distribution with shape parameters

$$\beta_1 = S_{\text{eff}}/S \times \bar{I}_\alpha + 1 \quad \text{and} \quad \beta_2 = S_{\text{eff}}/S \times (1 - \bar{I}_\alpha) + 1. \tag{16}$$

   Including $S_{\text{eff}}/S$ takes into account the efficiency of the posterior draws.

2. Find indices in $s \in \{1, \ldots, S\}$ closest to the ranks of these quantiles. For example, for quantile $Q$, find $s = \text{round}(\text{Q} \times \text{S})$.

3. Use the corresponding $\theta^{(s)}$ from the list of sorted posterior draws as quantiles from the error distribution. These quantiles can be used to approximate the Monte Carlo standard error of quantiles.

### 3.7 Interpreting split-$\widehat{R}$

The ultimate focus should be on the accuracy of the estimate for the quantity of interest. This accuracy can be measured using the Monte Carlo standard error (MCSE) or a corresponding uncertainty interval. The MCSE estimate is obtained based on the marginal posterior of the quantity and adjusted using the effective sample size, which in turn is based on split-$\widehat{R}$ and the autocorrelation of the chains. There is no general rule that determines an acceptable value for the MCSE. Instead, this will necessarily depend on the quantity of interest and the context of application.

When there might be difficulties with mixing, it is important to use between-chain information in computing effective sample size. For instance, in the sorts of funnel-shaped distributions that arise with hierarchical models, differences in step size adaptation can lead to chains to have different behavior reaching the narrow part of the funnel. In multimodal distributions with well-separated modes, the split-$\widehat{R}$ adjustment of the ESS leads to an ESS estimate the is close to the number of distinct modes that are found. If were to run only a single chain and computed the effective sample size only based on autocorrelations, it would be highly overestimated in such cases. The robustness of a split-$\widehat{R}$ adjusted ESS can be improved by running more chains and we recommend running at least four chains by default. As a generic ad hoc rule based on the simulations we recommend to aim for split-$\widehat{R} < 1.01$ for all quantities of interest.

The required ESS can be ultimately decided based on how the fitted model will be used. However, there are two reasons to look at the ESS before looking at the MCSE. Firstly, the computation of split-$\widehat{R}$ and autocorrelations needed for ESS itself require estimating means and variances, and in order to get reliable convergence and ESS estimates, we recommend to aim for ESS $> 400$. When running four chains, this corresponds to having an effective sample size of at least 50 per split chain to be used for estimating means, variances and autocorrelations. To obtain useful estimation accuracy in practice, ESS $> 100$ will often be

sufficient. However, this would require that we know that *actually* ESS $> 100$ and we cannot be certain about that until we achieved a higher ESS estimate. Secondly, effective sample sizes for different parameters are on the same scale, and thus it is easier to see which part of the model might have sampling problems.

After checking that split-$\widehat{R}$ and ESS fulfill the above ad hoc requirements, we recommend to take into account the application-specific requirements for the accuracy of the quantity of interest and check that MCSE is low enough. Otherwise, running longer chains or reparameterizing the model may be necessary.

### 3.8 Diagnostic visualizations

In order to intuitively grasp convergence of iterative algorithms, we propose several new diagnostic visualizations in addition to the numerical convergence diagnostics discussed above. We illustrate the usage of these visualizations by means of several examples in Section 4.

**Rank plots.** Extending the idea of using ranks instead of the original parameter values, we propose to use rank plots for each chain instead of trace plots. Rank plots are nothing else than histograms of the ranked posterior samples (ranked over all chains) plotted separately for each chain. If rank plots of all chains look similar, this indicates good mixing of the chains. As compared to trace plots, rank plots don't tend to squeeze to a fuzzy mess in case of long chains.

**Quantile and small interval plots.** The efficiency of quantiles or small interval probabilities may vary drastically across different quantiles and small interval positions, respectively. We thus propose to use diagnostic plots that display efficiency of quantiles or small interval probabilities across their whole range to better diagnose areas of the distributions that the iterative algorithm fails to explore efficiently.

**Efficiency per iteration plots.** For a well explored distribution, we expect the ESS measures to grow linearly with the total number of draws $S$, or, equivalently, that the relative efficiency (ESS divided $S$) is approximately constant for different values of $S$. For small number of draws, both bulk and tail-ESS may be unreliable and cannot necessarily detect convergence problems. As a result, some convergence problems may only be detectable as $S$ increases, which then implies the ESS to grow slower then linear or even decrease with increasing $S$. Equivalently, in such a case, we would expect to see a relatively sharp drop in the relative efficiency measures. We therefore propose to plot the change of both bulk and tail ESS with increasing $S$. This can be done based on a single model without a need to refit, as we can just extract initial sequences of certain length from the original chains. However, it should be noted that some convergence problems only occur at relatively high $S$ and may thus not be detectable if the total number of draws is too small.

## 4 Examples

In this section, we will go through some examples to demonstrate the usefulness of our proposed methods as well as the associated workflow in determining convergence. The online appendix contains all model details, code to reproduce the results and more detailed analysis of different algorithm variants and further examples[2].

Unless mentioned otherwise, we use dynamic Hamiltonian Monte Carlo with multinomial sampling (Betancourt, 2017) as implemented in Stan (Stan Development Team, 2018d) and run 4 chains each with 1000 warmup iterations and 1000 post-warmup iterations used for inference.

[2]https://avehtari.github.io/rhat_ess/rhat_ess.html

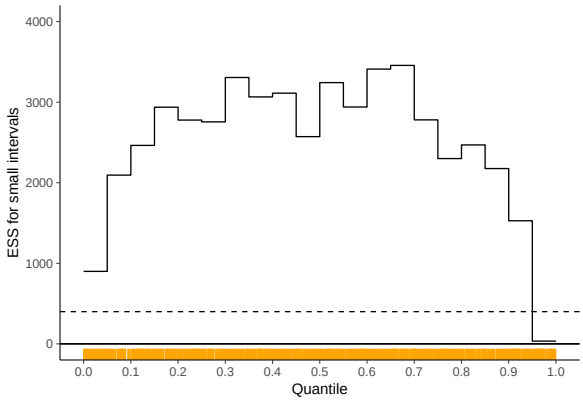


Figure 3: Local efficiency of small interval probability estimates for the Cauchy model with nominal parameterization. Orange ticks show iterations that exceeded the maximum treedepth in the dynamic HMC algorithm.

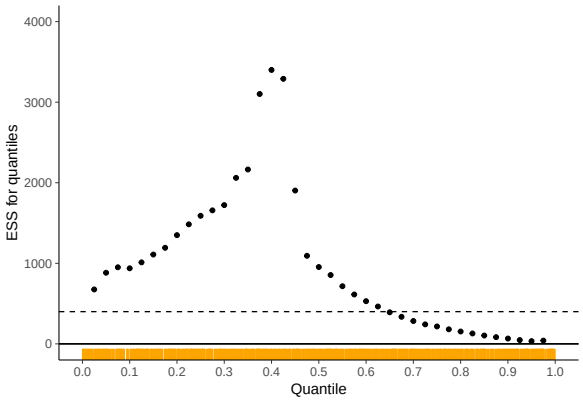


Figure 4: Efficiency of quantile estimates for the Cauchy model with nominal parameterization. Orange ticks show iterations that exceeded the maximum treedepth in the dynamic HMC algorithm.

## 4.1 Cauchy: A distribution with infinite mean and variance

The classic split-$\widehat{R}$ are based on calculating within and between chain variances. If the marginal distribution of a chain is such that the variance is not defined (i.e., infinite), the classic split-$\widehat{R}$ is not well justified. In this section, we will use the Cauchy distribution as an example of such a distribution.

### Nominal parameterization of Cauchy

The nominal Cauchy model with direct parameterization is

$$x \sim \text{Cauchy}(0, 1). \tag{17}$$

We set an independent Cauchy distribution for each element of the 50-dimensional real vector $x$. Dynamic HMC specific diagnostics, such as divergent transitions as well as iterations that exceed the maximum treedepth, indicate slow mixing of the chains.

Several split-$\widehat{R} > 1.01$ and some ESS $< 400$ also indicate convergence problems. The online appendix contains more results with longer chains and other $\widehat{R}$ diagnostics. We can further analyze potential problems using local efficiency and rank plots. We specifically investigate $x_{36}$, which, in this specific run, had the smallest tail-ESS of 34. Figure 2 shows the local efficiency of small interval probability estimates (see Section 3.5). The efficiency of sampling is very low in the tails, which is clearly caused by slow mixing in long tails of the Cauchy distribution. Figure 3 shows the efficiency of quantile estimates (see Section 3.3), which is also very low in the tails.

We may also investigate how the estimated effective sample sizes change when we use more and more draws (Brooks and Gelman (1998) proposed to use similar graph for $\widehat{R}$). If the effective sample size is highly unstable, does not increase proportionally with more draws, or even decreases, this indicates that simply running longer chains will likely not solve the convergence issues. In Figure 4, we see how unstable both bulk-ESS and tail-ESS are for this example. Rank plots in Figure 5 clearly show the mixing problem between chains. In case of good mixing all rank plots should be close to uniform.

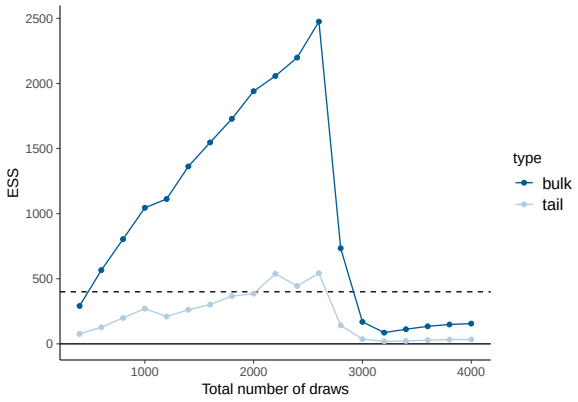


Figure 5: Estimated effective sample sizes with increasing number of iterations for the Cauchy model with nominal parameterization.

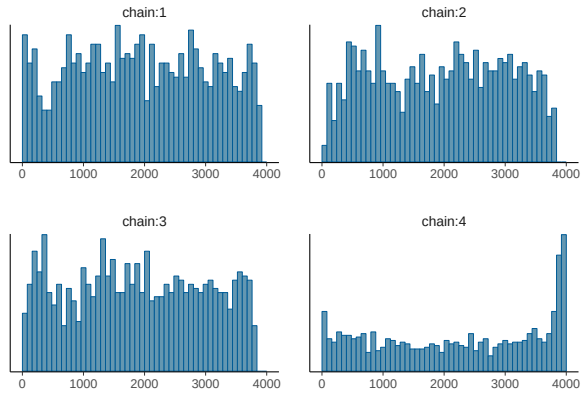


Figure 6: Rank plots of posterior draws from four chains for the Cauchy model with nominal parameterization.

**Alternative parameterization of Cauchy**

Next we examine an alternative parameterization that considers the Cauchy distribution as a scale mixture of Gaussian distributions

$$a \sim \mathrm{N}(0,1), \qquad b \sim \mathrm{Gamma}(0.5, 0.5), \qquad x = \frac{a}{\sqrt{b}}. \tag{18}$$

The model has two parameters, and the Cauchy distributed $x$ can be computed deterministically from those. In addition to improved sampling performance, the example illustrates that focusing on diagnostics matters.

We set define two 50-dimensional parameter vectors $a$ and $b$ from which the 50-dimensional quantity $x$ is computed. There are no warnings and the sampling is much faster.

All split-$\widehat{R} < 1.01$ and ESS $> 400$ indicate that sampling worked much better with the alternative parameterization. The online appendix contains more results using other parameterizations of the cauchy distribution. The vectors $a$ and $b$ used to form the Cauchy distributed $x$ have stable quantile, mean and variance values. As $x$ is Cauchy distributed it has stable quantiles, but wildly varying mean and variance estimates as the true values are not finite. We can further analyze potential problems using local efficiency estimates and rank plots. For this example. we take a detailed look at $x_{40}$, which had the smallest bulk-ESS of 2848. Figures 6 and 7 show good sampling efficiency for the small interval probability and quantile estimates. The Rank plots displayed in Figure 8 also look quite uniform across chains thus indicating convergence.

**Half-Cauchy with nominal parameterization**

Half-Cauchy priors for non-negative parameters are common and, at least in Stan, usually specified via the nominal parameterization. In this example, we set independent half-Cauchy distributions on each element of the 50 dimensional vector $x$ with positivity constraint (`<lower=0>`). Stan will then sample automatically in the unconstrained $\log(x)$ space, which changes the geometry crucially. As a result, there are no warnings and all split-$\widehat{R} < 1.01$ and ESS $> 400$ indicate good performance of the sampler despite using the nominal parameterization of the Cauchy distribution. More experiments for the half-Cauchy distribution can be found in the online appendix.

## 4.2 Hierarchical model: Eight schools

The eight schools problem is a classic example (see Section 5.5 in Gelman et al., 2013), which even in its simplicity illustrates the typical problems in inference for hierarchical models. We can parameterize this

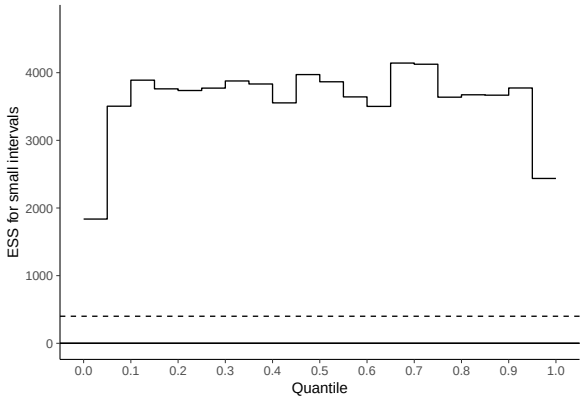


Figure 7: Local efficiency of small interval probability estimates for the Cauchy model with alternative parameterization.

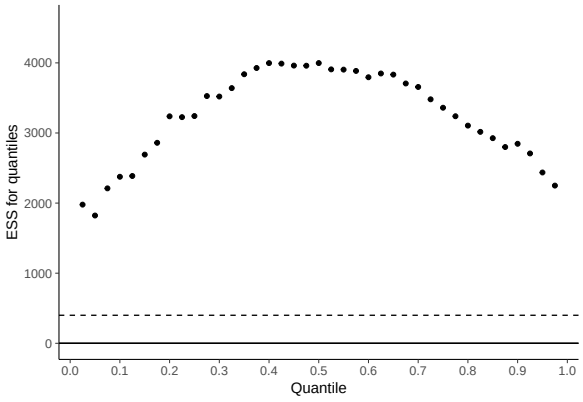


Figure 8: Efficiency of quantile estimates for the Cauchy model with alternative parameterization.

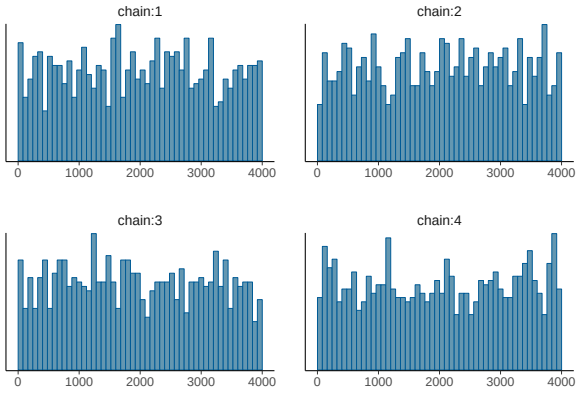


Figure 9: Rank plots of posterior draws from four chains for the Cauchy model with alternative parameterization.

simple model in at least two ways. The centered parameterization looks as follows:

$$\theta_j \sim \text{normal}(\mu, \tau)$$
$$y_j \sim \text{normal}(\theta_j, \sigma)$$

In contrast, the non-centered parameterization can be written as:

$$\tilde{\theta}_j \sim \text{normal}(0, 1)$$
$$\theta_j = \mu + \tau\tilde{\theta}_j$$
$$y_j \sim \text{normal}(\theta_j, \sigma)$$

In both parameterizations, $\theta_j$ are the school specific means, and $\mu$ as well as $\tau$ are the hierarchical mean and standard deviation of the school specific means, respectively.

Geometrically, the centered parameterization exhibits a funnel shape that contracts into a region of strong curvature around small values of the population standard deviation $\tau$, making it difficult for most Markov chain methods to adequately explore the full distribution of this paramater. The online appendix contains more detailed analysis of different algorithm variants and results of longer chains.

### A centered eight schools model

Instead of the default options, we run the centered parameterization model with more conservative settings of the HMC sample to reduce the probability of getting divergent transitions. Still, we observe a lot of divergent transitions, which in itself is already a sufficient indicator of convergence problems. We may also use the split-$\widehat{R}$ and ESS diagnostics to recognize problematic parts of the posterior. The latter two have the advantage over the divergent transitions diagnostic that they can be used with all MCMC algorithms not only with HMC.

Bulk-ESS and Tail-ESS for the between school standard deviation $\tau$ are 67 and 82 respectively. Both are much less than 400, indicating we should investigate that parameter more carefully. Figures 9 and 10 show the sampling efficiency for the small interval probability and quantile estimates. The sampler has difficulties in exploring small $\tau$ values. As the sampling efficiency for small $\tau$ values is practically zero, we may assume that we miss substantial amount of posterior mass and get biased estimates. In this case, the severe sampling problems for small $\tau$ values is reflected in the sampling efficiency for all quantiles. Red ticks, which show iterations with divergences, have concentrated to small $\tau$ values, which gives us another indication of problems in exploring small values.

Figure 11 shows how the estimated effective sample sizes change when we use more and more draws. Here we do not see sudden changes, but both bulk-ESS and tail-ESS are consistenly low. In line with the other findings, rank plots of $\tau$ displayed in Figure 12 clearly show problems in the mixing of the chains. Results for longer chains are provided in the online appendix.

### Non-centered eight schools model

For hierarchical models, the corresponding non-centered parameterization often works better. For reasons of comparability, we use the same conservative sampler setting as for the centered parameterization model. For the non-centered parameterization, we do not observe divergences or other warnings. All split-$\widehat{R} < 1.01$ and ESS $> 400$ indicate a much better efficiency of the non-centered parameterization. Figures 13 and 14 show the efficiency of small interval probability estimates and the efficiency of quantile estimates for $\tau$. Small $\tau$ values are still more difficult to explore, but the relative efficiency is good. The rank plots of $\tau$ Figure 15 show no substantial differences between chains.

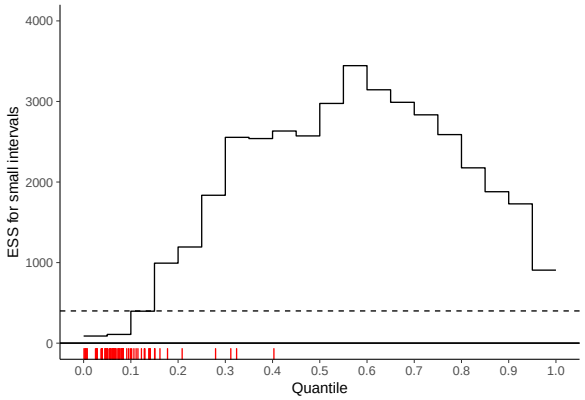


Figure 10: Local efficiency of small interval probability estimates for the eight schools model with centered parameterization. Red ticks show divergent transitions.

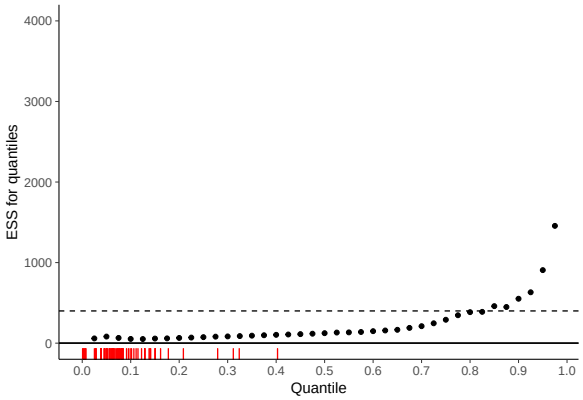


Figure 11: Efficiency of quantile estimates for the eight schools model with centered parameterization. Red ticks show divergent transitions. Red ticks show divergent transitions.

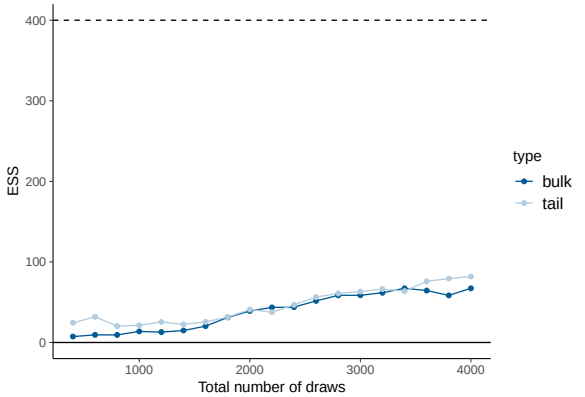


Figure 12: Estimated effective sample sizes with increasing number of iterations for the eight schools model with centered parameterization.

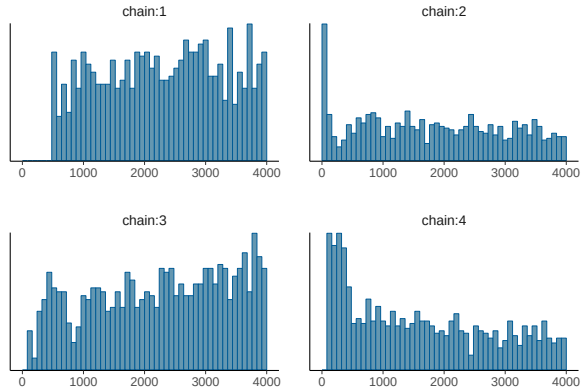


Figure 13: Rank plots of posterior draws from four chains for the eight schools model with centered parameterization.

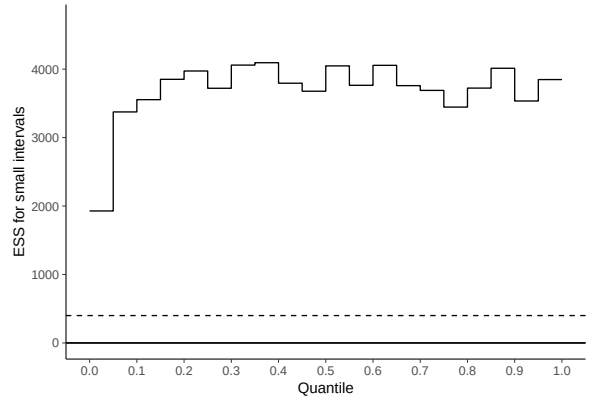


Figure 14: Local efficiency of small interval probability estimates for the eight schools model with the non-centered parameterization.

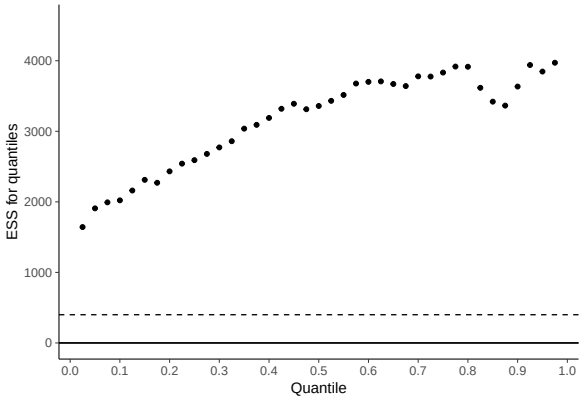


Figure 15: Efficiency of quantile estimates for the eight schools model with the non-centered parameterization.

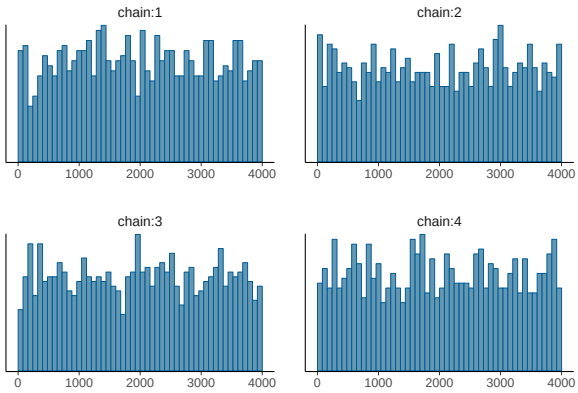


Figure 16: Rank plots of posterior draws from four chains for 8 schools model with non-centered parameterization.

# Appendices

More results and code available in online appendix[3].

## Appendix A: Normal distributions with additional trend, shift or scaling

Here we demonstrate the behavior of non-split-$\widehat{R}$, split-$\widehat{R}$, and bulk-ESS to detect various simulated cases presenting non-convergence behavior. We generate four varying length chains of iid normally distributed values, and then modify them to simulate three convergence problems:

- All chains have the same trend and a similar marginal distribution. This can happen in case of slow mixing and all chains initialized near each other far from the typical set.
- One of the chains has a different mean. This can happen in case of slow mixing, weak identifiability of one or several parameters, or multimodality.
- One of the chains having a lower marginal variance. This can happen in case of slow mixing, multimodality, or one of the chains having different mixing efficiency.

**All chains have the same trend.** First we draw all the chains are from the same $\mathrm{N}(0, 1)$ distribution plus a linear trend. Figure 16 shows that if we don't split chains, $\widehat{R}$ misses the trends if all chains still have a similar marginal distribution. Figure 17 shows that split-$\widehat{R}$ detects the trend, even if the marginals of the chains are similar. If we use a threshold of 1.01, we can detect trends which account for 2% or more of the total marginal variance. If we use a threshold of 1.1, we detect trends which account for 30% or more of the total marginal variance.

The effective sample size is based on split-$\widehat{R}$ and within-chain autocorrelation. Figure 18 shows the relative Bulk-ESS divided by $S$ for easier comparison between different values of $S$. We see that split-$\widehat{R}$ is more sensitive to trends for small sample sizes, but ESS becomes more sensitive for larger samples sizes (as autocorrelations can be estimated more accurately).

**Shifting one chain.** Second we draw all the chains are from the same $\mathrm{N}(0, 1)$ distribution, except one that is sampled with non-zero mean. Figure 19 shows that if we use a threshold of 1.01, split-$\widehat{R}$ can detect shifts with a magnitude of one third or more of the marginal standard deviation. If we use a threshold of 1.1, split-$\widehat{R}$ detects shifts with a magnitude equal to or larger than the marginal standard deviation. Figure 20

[3]https://avehtari.github.io/rhat_ess/

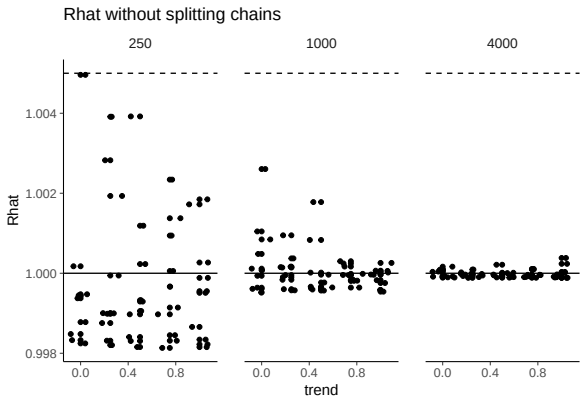


Figure 17: $\widehat{R}$ without splitting for varying chain lengths for chains which have the same trend and a similar marginal distribution.

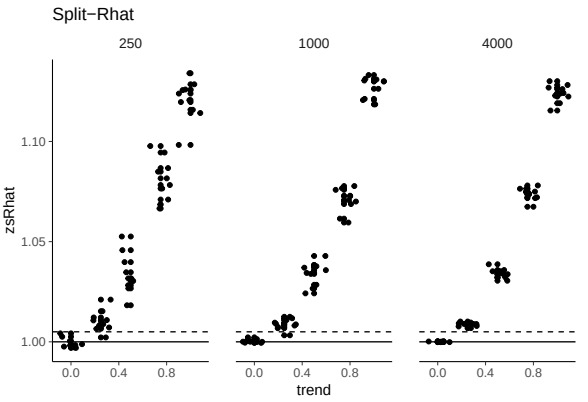


Figure 18: Split-$\widehat{R}$ for varying chain lengths for chains which have the same trend and a similar marginal distribution.

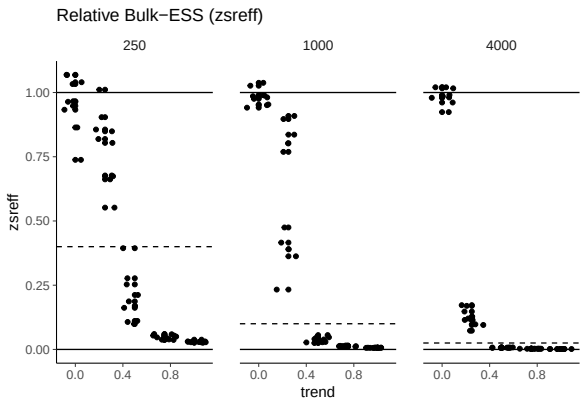


Figure 19: Relative Bulk-ESS for varying chain lengths for chains which have the same trend and a similar marginal distribution. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

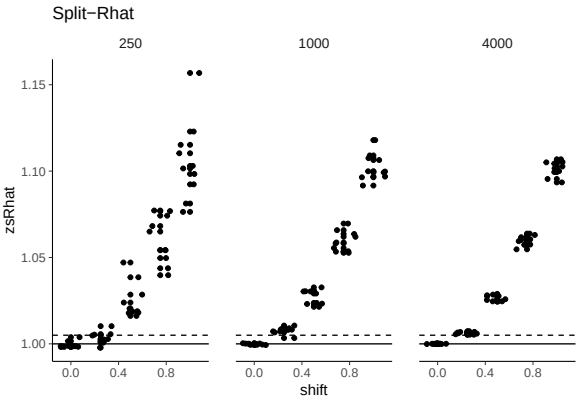


Figure 20: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different mean than the others.

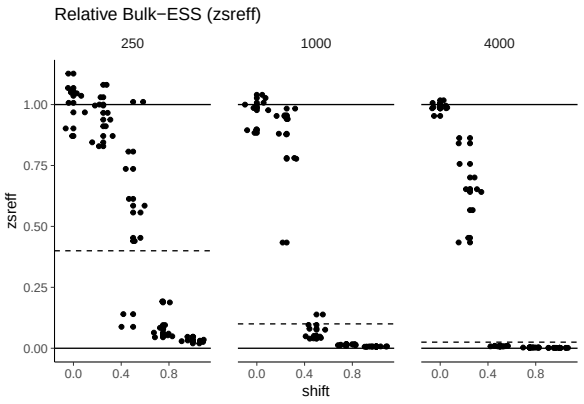


Figure 21: Relative Bulk-ESS for varying chain lengths for chains with one sampled with a different mean than the others. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

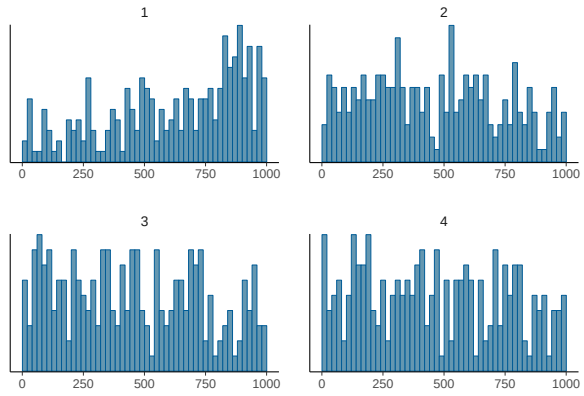


Figure 22: Rank plots of posterior draws from four chains with one sampled with a different mean than the others.

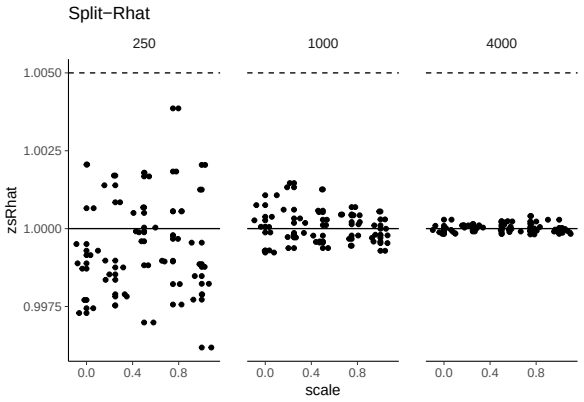


Figure 23: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

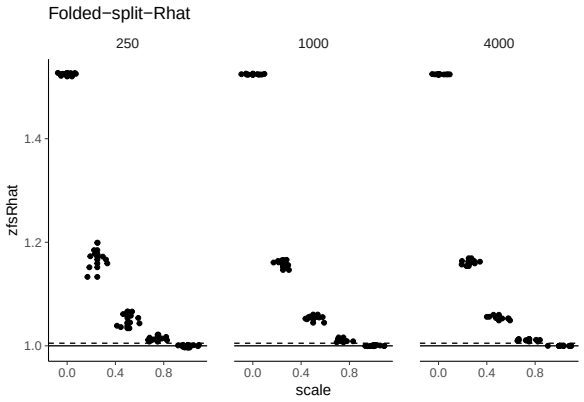


Figure 24: Folded-split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

shows the the relative Bulk-ESS for the same case. The effective sample size is not as sensitive as split-$\widehat{R}$, but a shift with a magnitude of half the marginal standard deviation or more will lead to very low relative efficiency when the total number of draws increases. Rank plots are practical way to visualize differences between chains. Figure 21 shows rank plots for the case of 4 chains, 250 draws per chain, and one chain sampled with mean 0.5 instead of 0. In this case split-$\widehat{R} = 1.05$, but the rank plots clearly show that the first chain behaves differently.

**Scaling one chain.** For our third simulation, all the chains are from the same $\text{N}(0, 1)$ distribution, except one of the chains is sampled with variance less than 1. Figure 22 shows that split-$\widehat{R}$ is not able to detect scale differences between chains. Figure 23 shows that folded-split-$\widehat{R}$ which focuses on scales detects scale differences. With a threshold of 1.01, folded-split-$\widehat{R}$ detects a chain with scale less than 3/4 of the standard deviation of the others. With a threshold of 1.1, folded-split-$\widehat{R}$ detects a chain with standard deviation less than 1/4 of the standard deviation of the others.

Figure 24 shows the the relative Bulk-ESS for the same case. The bulk effective sample size of the mean does not see a problem as it focuses on location differences between chains. Figure 25 shows rank plots for

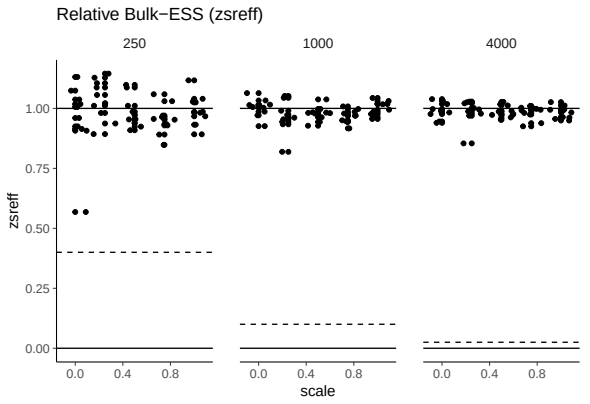


Figure 25: Relative Bulk-ESS for varying chain lengths for chains with one sampled with a different variance than the others.

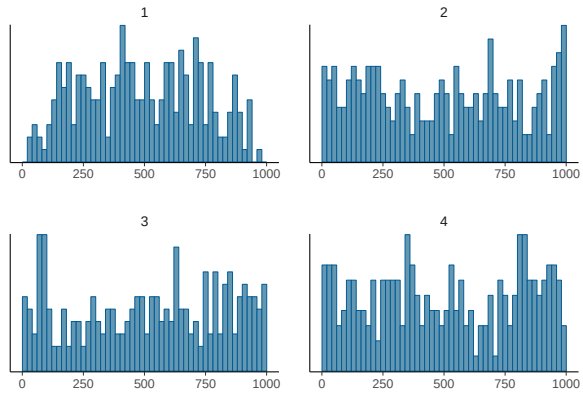


Figure 26: Rank plots of posterior draws from four chains with one sampled with a different variance than the others.

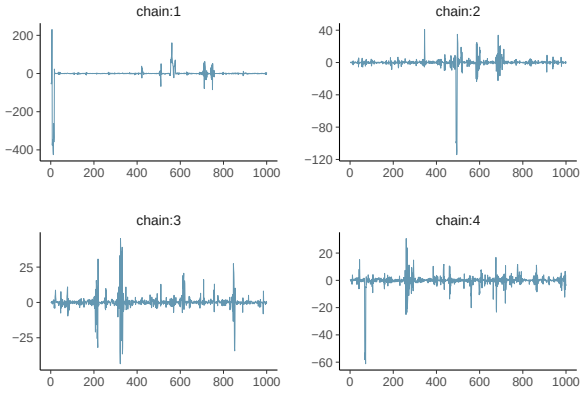


Figure 27: Trace plots of four chains for Cauchy model with nominal parameterization and `max_treedepth`=20.

the case of 4 chains, 250 draws per chain, and one chain sampled with standard deviation 0.75 instead of 1. Although folded-split-$\widehat{R} = 1.06$, the rank plots clearly show that the first chain behaves differently.

## Appendix B: Cauchy: A distribution with infinite mean and variance

Here we provide some additional results for the the nominal Cauchy model presented in the main text. Instead of the default options we increase `max_treedepth` to 20, which improves the exploration in long tails. The online appendix has additional results for the default option case and for longer chains.

Figure 26 shows that trace plots for the first parameter look wild with occasional large values, and it is difficult to interpret possible convergence. Figure 27 shows classic split-$\widehat{R}$ , rank normalized split-$\widehat{R}$, and rank normalized folded-split-$\widehat{R}$ for all 50 parameters. Classic split-$\widehat{R}$, which is not well-defined in this case, has much higher variability than rank normalized split-$\widehat{R}$. Rank normalized folded-split-$\widehat{R}$ has higher values than Rank normalized split-$\widehat{R}$ indicating slow mixing especially in tails. Figure 27 shows different effective sample size estimates for all 50 parameters. Classic ESS, which is not well defined in this case, has very high variability. Bulk ESS is much more stable, and indicates that we can get reliable estimates for the location of the posterior (except for mean). Median ESS is even more stable with relatively high values, indicating that we can estimate median of the distribution reliably. Tail-ESS has low values, indicating still too slow mixing in tails for reliable tail quantile estimates. MAD ESS values are just above our recommend threshold,

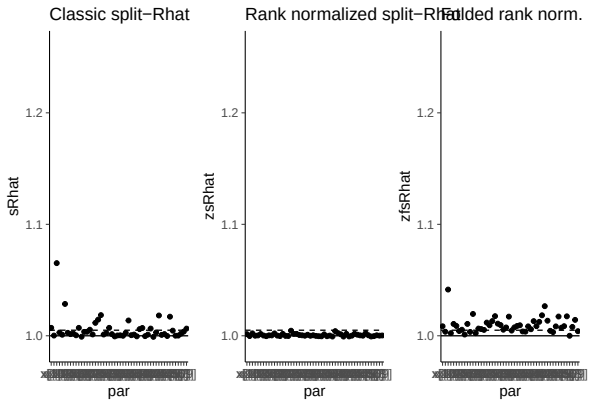


Figure 28: Classic split-$\widehat{R}$, rank normalized split-$\widehat{R}$, and rank normalized folded-split-$\widehat{R}$ for Cauchy model with nominal parameterization and `max_treedepth`=20.

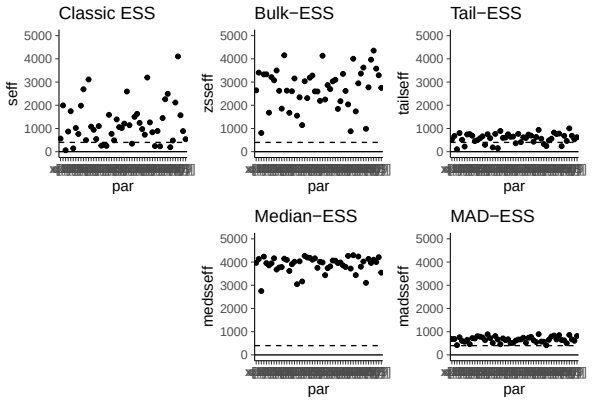


Figure 29: Classic ESS, Bulk-ESS, Tail-ESS, Median-ESS and MAD-ESS for Cauchy model with nominal parameterization `max_treedepth`=20.

indicating practically useful MAD estimates, too. The online appendix has additional results with longer chains, showing that all other ESS values except classic ESS (which is not well defined) keep improving with more iterations. It is however recommended to use more efficient parameterization especially if the tail quantiles are quantities of interest.

## Appendic C: A centered eight schools model with very long chains and thinning

Here we demonstrate a limitation of split-$\widehat{R}$ and ESS as a convergence diagnostics in case where the chains eventually converge to a common wrong stationary distribution.

When autocorrelation time is high, it has been common to thin the chains by saving only a small portion of the draws. This will throw away useful information also for convergence diagnostics. We run the eight schools model with centered parameterization with $4 \times 10^5$ iterations per chain. We remove the first half as warm-up and thin by 200, ending up with 4000 iterations as with the default settings.

We observe several divergent transitions and the estimated Bayesian fraction of missing information is also low, which still indicate convergence problems and potentially biased estimates.

Figures 29, 30, and 31 show the efficiency of small probability interval estimates, efficiency of quantile estimates, and change of Bulk-SS and Tail-ESS with increasing number of iterations. Unfortunately the thinning makes split-$\widehat{R}$ and ESS estimates to miss the problems. The posterior mean is still biased, being more than 3 sigmas away from the estimate obtained using non-centered parameterization. In this case all four chains fail similarly in exploring the narrowest part of the funnel and all chains seem to "converge" to a wrong stationary distribution. However, the rank plots shown in Figure 32 are still able to show the problem.

## Appendic D: A centered eight schools model fit using a Gibbs sampler

So far, we have run all models in Stan, but here we demonstrate that these diagnostics are useful also sampler other than Hamiltonian Monte Carlo. We will fit the eight schools models also Jags (Plummer, 2003), which uses a dialect of the BUGS language (Lunn et al., 2009) to specify models. Jags uses a clever mix of Gibbs and Metropolis-Hastings sampling. This kind of sampling does not usually scale well to high-dimensional posteriors of strongly interdependent parameters, but it can work fine for relatively simple models such as in this case study.

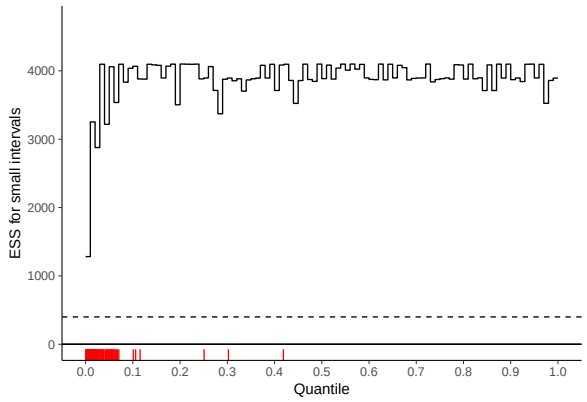


Figure 30: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization, very long chains, and thinning.

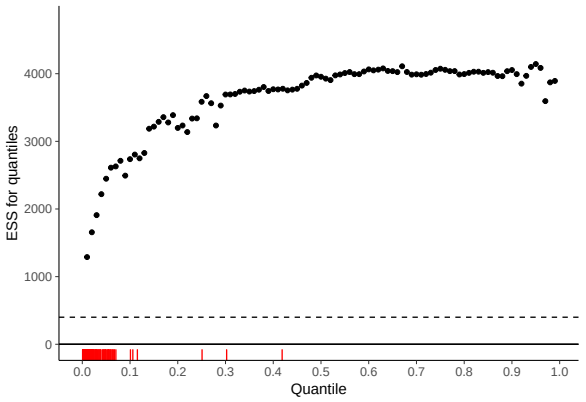


Figure 31: The efficiency of quantile estimates for 8 schools model with centered parameterization, very long chains, and thinning.

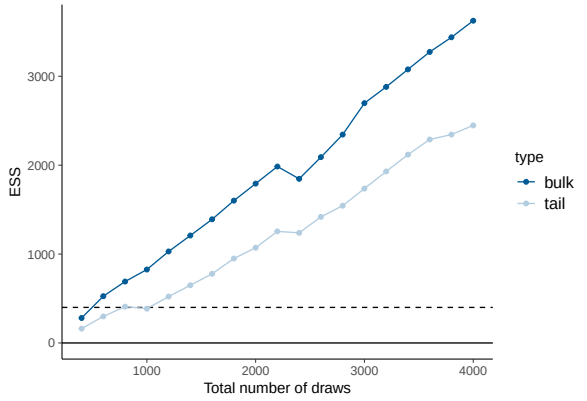


Figure 32: The estimated effective sample sizes with increasing number of iterations for 8 schools model with centered parameterization, very long chains, and thinning.

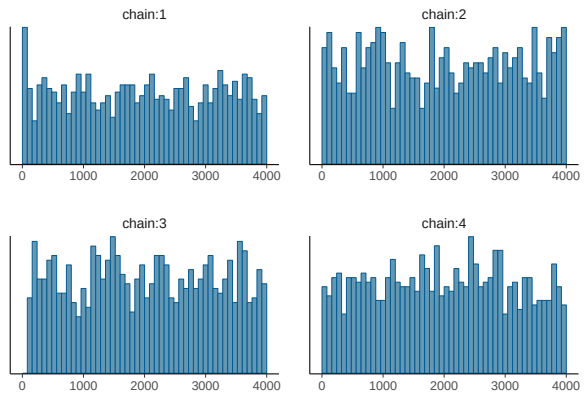


Figure 33: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization, very long chains, and thinning.

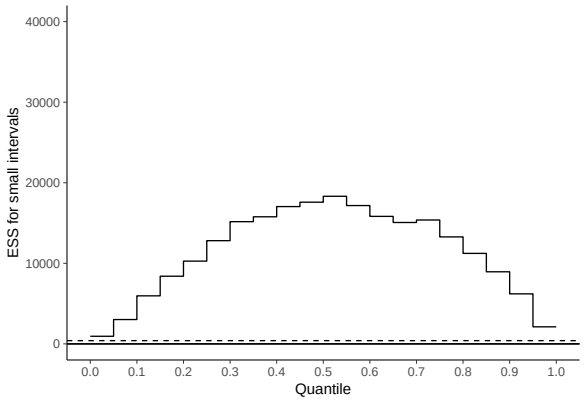


Figure 34: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization and Gibbs sampling.

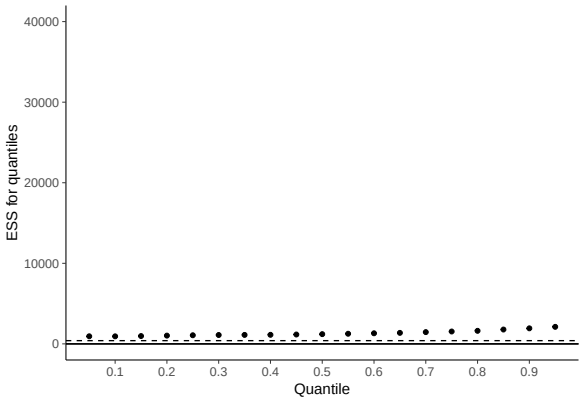


Figure 35: The efficiency of quantile estimates for 8 schools model with centered parameterization and Gibbs sampling.

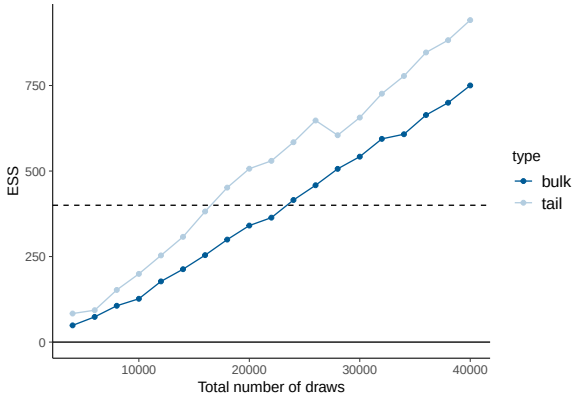


Figure 36: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization and Gibbs sampling.

First, we sample 1000 iterations for each of the 4 chains for easy comparison with the corresponding Stan results. Examining the diagnostics for $\tau$, Split-$\widehat{R} = 1.08$, Bulk-ESS= 59, and Tail-ESS= 53. 1000 iterations is clearly not enough. The online appendix shows also the usual visual diagnostics for 1000 iterations run, but here we next report the results with 10 000 iterations. Examining the diagnostics for $\tau$, now Split-$\widehat{R} = 1.01$, Bulk-ESS= 677, and Tail-ESS= 1027, which are all good.

Figures 33, 34, and 35 show the efficiency of small probability interval estimates, efficiency of quantile estimates, and change of Bulk-SS and Tail-ESS with increasing number of iterations. The relative efficiency is low, but ESS for all small probability intervals, quantiles and bulk are above the recommend threshold. Notably, however, the increase in effective sample size for $\tau$ is linear in the total number of draws. Gibbs sampler can reach the narrow part of the funnel, although the sampling efficiency is affected by the funnel. In this simple case the inefficiency of the Gibbs sampling is not dominating and good results can be achieved in reasonable time. The online appendix shows additional results for Gibbs sampling with more efficient non-centered parameterization.

# References


Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, third edition*. CRC Press, 2013.

Charlie J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–483, 1992.

Charlie J. Geyer. Introduction to Markov chain Monte Carlo. In S Brooks, A Gelman, G L Jones, and X L Meng, editors, *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, 2003.

Stan Development Team. RStan: the R interface to Stan. R package version 2.17.3, 2018a. URL http://mc-stan.org.

Stan Development Team. RStanArm: Bayesian applied regression modeling via Stan. R package version 2.17.4, 2018b. URL http://mc-stan.org.

Stan Development Team. *Bayesian Statistics Using Stan*. Stan Development Team, 2018c. URL https://github.com/stan-dev/stan-book.

Stan Development Team. Stan modeling language users guide and reference manual. version 2.18.0, 2018d. URL http://mc-stan.org.