

# Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC\*

*Aki Vehtari<sup>†</sup>, Andrew Gelman<sup>‡</sup>, Daniel Simpson<sup>§</sup>, Bob Carpenter<sup>¶</sup> and Paul-Christian Bürkner<sup>||</sup>*

19 March 2019

## Abstract

Markov chain Monte Carlo is a key computational tool in Bayesian statistics, but it can be challenging to monitor the convergence of an iterative stochastic algorithm. In this paper we show that the convergence diagnostic  $\widehat{R}$  of Gelman and Rubin (1992) has serious flaws and we propose an alternative that fixes them. We also introduce a collection of quantile-based local efficiency measures, along with 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. Finally, we give concrete recommendations for how these methods should be used in practice.

## 1 Introduction

Markov chain Monte Carlo (MCMC) methods are important 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 behavior after a finite number of draws. Indeed, decades of experience tell us that the finite sample behavior of these algorithms can be almost arbitrarily bad.

### 1.1 Monitoring convergence using multiple chains

In an attempt to assuage concerns of poor convergence, we typically run multiple independent chains to see if the obtained distribution is similar across chains. We typically also visually inspect the sample paths of the chains as well as study summary statistics such as 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.

---

\*We thank Academy of Finland, the U.S. Office of Naval Research, and the Natural Science and Engineering Research Council of Canada for partial support of this research. All computer code and an even larger variety of numerical experiments are available in the online appendix at [https://avehtari.github.io/rhat\\_ess/rhat\\_ess.html](https://avehtari.github.io/rhat_ess/rhat_ess.html).

<sup>†</sup>Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland.

<sup>‡</sup>Department of Statistics, Columbia University, New York.

<sup>§</sup>Department of Statistical Sciences, University of Toronto, Canada.

<sup>¶</sup>Applied Statistics Center, Columbia University, New York.

<sup>||</sup>Department of Psychology, University of Münster, Germany

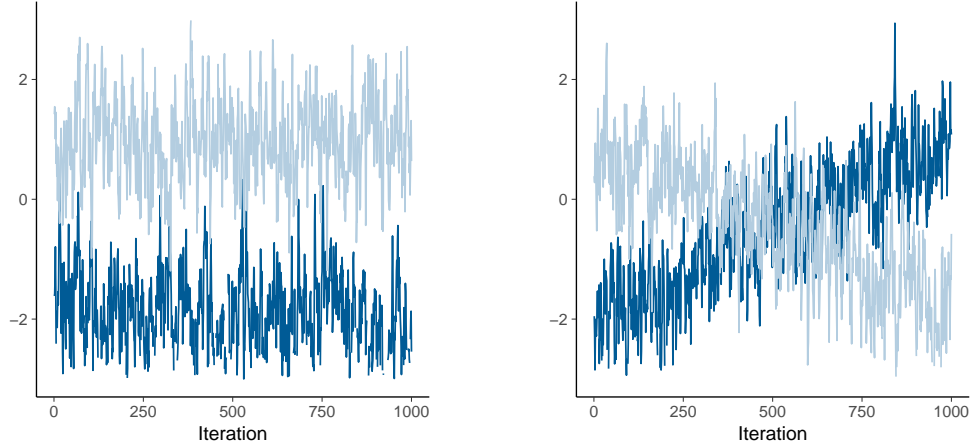


Figure 1: Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the two sequences happen to cover a common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

Probably the most widely-used convergence diagnostic is the potential scale reduction factor  $\hat{R}$  (Gelman and Rubin, 1992; Brooks and Gelman, 1998), which monitors the ratio of between- to within-chain variance for model parameters and other univariate quantities of interest. The idea is that if a set of simulations have not mixed well, the variance of all the chains mixed together will be higher than the variance of individual chains. More recently, Gelman et al. (2013) introduced split- $\hat{R}$  which also compares the first half of each chain to the second half, to try to detect lack of convergence within each chain. In this paper when we refer to  $\hat{R}$  we are always speaking of split- $\hat{R}$ .

Convergence diagnostics are most effective when computed using multiple chains initialized at a diverse set of starting points, to reduce the chance that we falsely diagnose mixing when beginning at a different point would lead to a qualitatively different posterior.

In the context of Markov chain Monte Carlo, one can interpret  $\hat{R}$  with diverse seeding as an operationalization of the qualitative statement that convergence of the Markov chain should be 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 approximate posterior expectations. Without this, we have no control over the large deviation behavior of the estimates and the constructed Markov chains may be useless for practical purposes.

Unfortunately,  $\hat{R}$  can fail to diagnose poor mixing, which can be a problem when we use them within generic software packages like `Stan` (Carpenter et al., 2017) or analysis tools like the `coda` package (Plummer et al., 2006). The following example shows failure can occur.

## 1.2 Example where traditional $\hat{R}$ fails

Figure 2 shows the distribution of  $\hat{R}$  (that is, split- $\hat{R}$  from Gelman et al. (2013)) in four different scenarios. In two of the scenarios the chains have not mixed and we would like  $\hat{R}$  to diagnose this problem, and in the other two scenarios the chains have mixed and so  $\hat{R}$  should be near 1. In this example, traditional  $\hat{R}$  fails to detect the two problematic scenarios, but a new version of  $\hat{R}$  introduced in this paper succeeds in flagging the problems. We do not intend with this example to claim that our new  $\hat{R}$  is perfect—of course, it can be defeated too. Rather, we use these simple scenarios to develop intuition about problems with the traditional

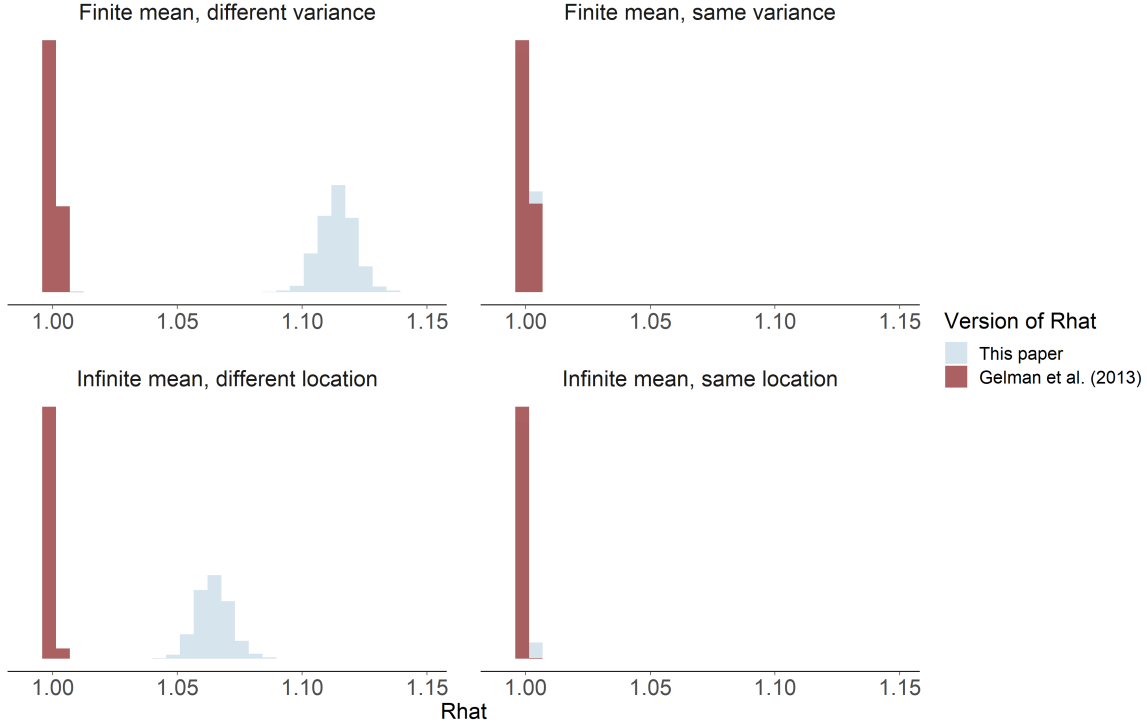


Figure 2: A synthetic problem demonstrating two modes of failure for the traditional  $\hat{R}$ . Each plot shows histograms of  $\hat{R}$  values over 1000 replications of four chains each with a thousand draws. In the left column, one of these four chains was incorrect. In the top left plot, we set one of the four chains to have a variance lower than the others. that was too low. In the bottom left plot, we took one of the four chains and shifted it. In both cases, the traditional  $\hat{R}$  estimate does not detect the poor behavior, while the new value does. In the right column, all the chains are simulated with the same distribution. The chains used for the top row plots target a normal distribution, while the chains used for the bottom row plots target a Cauchy distribution.

split- $\hat{R}$  and possible directions for improvement.

In each of the four scenarios in Figure 2, we run four chains for 1000 iterations each and then replicate the entire simulation 1000 times. The top row of the figure shows results for independent AR(1) processes with autoregressive parameter  $\rho = 0.3$ . The top left graph shows the distribution of  $\hat{R}$  when the variance of one of the four chains is set to  $1/3$  of the variance of the others. This corresponds to a scenario where one chain fails to correctly explore the parameter space. The split- $\hat{R}$  statistic defined in Gelman et al. (2013) does not detect the poor mixing, while the new variant of split- $\hat{R}$  defined later in this paper does. The top-right figure shows the same scenario but with all the chains having the same variance, and now both  $\hat{R}$  values correctly identify that mixing occurs.

The second row of Figure 2 shows the behavior of  $\hat{R}$  when the target distribution has **infinite variance**. In this case the chains were constructed as a ratio of stationary AR(1) processes with  $\rho = 0.3$ , and the distribution of the ratio is Cauchy. All of the simulated chains have unit scale, but in the lower-left figure, **we have shifted one of the four chains two units to the right**. The Gelman et al. (2013) version of  $\hat{R}$  would catch this behavior **if the chain had finite variance**, but in this case the infinite variance destroys its effectiveness—traditional  $\hat{R}$  and split- $\hat{R}$  are defined based on second-moment statistics—and it inappropriately returns a value very close to 1. On the other hand, **our new  $\hat{R}$  statistic correctly catches the error**. The bottom right figure shows that both version of  $\hat{R}$  work in this scenario when the four chains have the same distribution.

**This example identified two problems with traditional  $\hat{R}$ :**

1. If the chains have different variances but the same mean parameters,  $\hat{R} \approx 1$ .

2. If the chains have infinite variance,  $\hat{R} \approx 1$  even if one of the chains has a different location parameter to the others. This can also lead to numerical instability for thick-tailed distributions even when the variance is technically finite. It's typically very hard to assess empirically if a chain has large but finite variance or infinite variance.

Another problem is that  $\hat{R}$  is typically computed only for the posterior mean. While this provides an estimate for the convergence in the bulk of the distribution, it says little about the convergence in the tails, which is a concern for posterior interval estimates as well as for inferences about rare events.

All of these observations lead to the conclusion that the  $\hat{R}$  statistic as it is currently defined is not a good measure of convergence. However, it is easy to compute and sensitive to certain types of non-convergence. So, in this paper, we propose improvements that overcome the problems described above.

In addition, as the convergence of the Markov chain needs not be uniform across the parameter space, we propose a localized version of the effective sample size that allow us to assess better the behavior of localized functionals and quantiles 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.

## 2 Recommendations for practice

In this section we lay out practical recommendations for using the tools developed in this paper. In the interest of specificity, we have provided numerical targets for both  $\hat{R}$  and effective sample size (ESS). However, these values should be adapted as necessary for the given application.

In Section 4, we propose modifications to the  $\hat{R}$  based on rank-normalizing and folding the posterior draws. We recommend running **at least four chains** by default and only using the sample if  $\hat{R} < 1.01$ . This is a much tighter threshold than the one recommended by Gelman and Rubin (1992), reflecting lessons learnt over more than 25 years of use.

Roughly speaking, the ESS of a quantity of interest captures how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm. Clearly, the higher the ESS the better. **When there might be difficulties with mixing, it is important to use between-chain information in computing the ESS.** 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 in the neighborhood of the narrow part of the funnel. For multimodal distributions with well-separated modes, the split- $\hat{R}$  adjustment leads to an ESS estimate that is close to the number of distinct modes that are found. In this situation, ESS can be drastically overestimated if computed from a single chain.

As Vats and Knudson (2018) note, a small value of  $\hat{R}$  is not enough to ensure that an MCMC procedure is useful in practice. It also needs to have a sufficiently large effective sample size. As with  $\hat{R}$ , we recommend computing the ESS on the rank-normalized sample. This does not directly compute the ESS relevant for computing the mean of the parameter, but instead computes a quantity that is well defined even if the chains do not have finite mean or variance. Specifically, it computes the ESS of a sample from a *normalized* version of the quantity of interest, using the rank transformation followed by the normal inverse-cdf. This is still indicative of the effective sample size for computing an average, and if it is low the computed expectations are unlikely to be good approximations to the actual target expectations. We recommend requiring that the rank-normalized ESS is greater than 400. When running four chains, this corresponds to having a rank-normalized effective sample size of at least 50 per split chain.

Only when the rank-normalized and folded  $\hat{R}$  values are less than the prescribed threshold and the rank-normalized ESS is greater than 400 do we recommend computing the actual (not rank-normalized) effective sample size for the quantity of interest. This can then be used to assess the Monte Carlo standard error (MCSE) for the quantity of interest.

Finally, if you plan to report quantile estimates or posterior intervals, we strongly suggest assessing the convergence of the chains for these quantiles. In Section 4.3 we show that convergence of Markov chains is not uniform across the parameter space and propose diagnostics and effective sample sizes specifically for extreme quantiles. This is *different* from the standard ESS estimate (which we refer to as the “bulk-ESS”), which mainly assesses how well the centre of the distribution is resolved. Instead, these “tail-ESS” measures allow the user to estimate the MCSE for interval estimates.

### 3 $\hat{R}$ and the effective sample size

When coupled with an ESS estimate,  $\hat{R}$  is the most common way to assess the convergence of a set of simulated chains. There is a link between these two measures for a single chain (Vats and Knudson, 2018), but we prefer to treat these as two separate questions: “Did the chains mix well?” (split- $\hat{R}$ ) and “Is the ESS large enough?” In this section we define the  $\hat{R}$  statistic that we propose to modify. The definition of the ESS as used in the present paper is given in Appendix A. We will refer to this particular ESS measure as  $S_{\text{eff}}$ . Here we present split- $\hat{R}$ , following Gelman et al. (2013), but using the notation of Stan Development Team (2018b). This formulation 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:

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

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

The between-chain variance,  $B$ , also contains the factor  $N$  because it is based on the variance of the within-chain means,  $\bar{\theta}^{(\cdot m)}$ , 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,

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

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  $N \rightarrow \infty$ . 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  $N \rightarrow \infty$ , the expectation of  $W$  also approaches  $\text{var}(\theta|y)$ . Vats and Knudson (2018) propose a different variance estimator that is more efficient for single chains, however we are willing to trade off a slightly higher variance for increased diagnostic sensitivity (as described in the introduction) that running multiple chains brings.

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  $N \rightarrow \infty$ . This potential scale reduction is estimated as,

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

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

The effective sample size is defined as

$$S_{\text{eff}} = \frac{NM}{\hat{\tau}}, \quad (5)$$

where  $\hat{\tau}$  combines the potential scale reduction and the autocorrelations within the chains at different lags as described in detail in Appendix A.

The value of  $\hat{R}$  and ESS require reliable estimates of variances and autocorrelations, which can only occur if the chains have enough independent replicates. In particular, we only recommend relying on the  $\hat{R}$  estimate to make decisions about the quality of the chain if each of the split chains has an average ESS estimate of at least 50. In the our minimum recommended setup of four parallel chains, the total ESS should be at least 400 before we expect  $\hat{R}$  to be useful.

## 4 Improving convergence diagnostics

### 4.1 Rank normalization helps $\hat{R}$ when there are heavy tails

As split- $\hat{R}$  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- $\hat{R}$  and  $S_{\text{eff}}$  are computed using the equations in Section 3 and Appendix A, but replacing the original parameter values  $\theta^{(nm)}$  with their corresponding rank normalized values denoted as  $z^{(nm)}$ . Rank normalization proceeds as follows. First, replace each value  $\theta^{(nm)}$  by its rank  $r^{(nm)}$  within the pooled draws from all chains. Average rank for ties are used to conserve the number of unique values of discrete quantities. Second, normalize ranks via the inverse normal transformation,

$$z^{(nm)} = \Phi^{-1}((r^{(nm)} - 0.5)/S). \quad (6)$$

For continuous variables and  $S \rightarrow \infty$ , 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  $\hat{R}$  and  $S_{\text{eff}}$  do not change for normally distributed parameters. Furthermore, rank-normalized  $\hat{R}$  and  $S_{\text{eff}}$  are invariant to univariate transformations. Hence rank-normalized  $\hat{R}$  and  $S_{\text{eff}}$  are approximations to the ordinary  $\hat{R}$  and  $S_{\text{eff}}$  for a nice transformation of the parameter of interest. The effects of rank normalization are further explored in the online appendix.

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 B). 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 4.4.

### 4.2 Folding detects errors in variance and trouble exploring the tails

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

$$\zeta^{(mn)} = \left| \theta^{(nm)} - \text{median}(\theta) \right|. \quad (7)$$

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

### 4.3 Localizing convergence diagnostics: assessing the quality of quantiles, the median absolute deviation, and small-interval probabilities

The new  $\hat{R}$  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 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. This especially has implications if people are making decisions based on whether or not a specific quantile is below or above a fixed value (for example, if a posterior interval contains zero).

The  $\alpha$ -quantile is defined as the parameter value  $\theta_\alpha$  for which  $\Pr(\theta \leq \theta_\alpha) = \alpha$ . 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  $\hat{R}$  and  $S_{\text{eff}}$  are not directly applicable. Thus, we first focus on the efficiency estimate for the cumulative probability  $\Pr(\theta \leq \theta_\alpha)$  for different values of  $\theta_\alpha$ .

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

$$\Pr(\theta \leq \theta_\alpha) \approx \bar{I}_\alpha = \frac{1}{S} \sum_{s=1}^S I(\theta^{(s)} \leq \theta_\alpha), \quad (8)$$

where  $I(\cdot)$  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 ratio of variances. This means that the efficiency of  $F^{-1}(\bar{I}_\alpha)$  is well approximated asymptotically by 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 B).

Since the marginal posterior distributions might not have finite mean and variance, by default **rstanarm** (Stan Development Team, 2018a) 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 (7)):

$$\Pr(\zeta \leq \zeta_{0.5}) \approx \bar{I}_{\zeta, 0.5} = \frac{1}{S} \sum_{s=1}^S I(\zeta^{(s)} \leq \zeta_{0.5}), \quad (9)$$

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.

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

$$\bar{I}_{\alpha, \delta} = \Pr(\hat{Q}_\alpha < \theta \leq \hat{Q}_{\alpha+\delta}), \quad (10)$$

where  $\hat{Q}_\alpha$  is an empirical  $\alpha$ -quantile,  $\delta = 1/k$  is the length of the interval for some positive integer  $k$ , and  $\alpha \in (0, \delta, \dots, 1 - \delta)$  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.

#### 4.4 Monte Carlo error estimates for quantiles

It is common practice to 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 except 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

$$\beta_1 = S_{\text{eff}}/S \times \bar{I}_\alpha + 1 \quad \text{and} \quad \beta_2 = S_{\text{eff}}/S \times (1 - \bar{I}_\alpha) + 1. \quad (11)$$

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

2. Find indices in  $s \in \{1, \dots, S\}$  closest to the ranks of these quantiles. For example, for quantile  $Q$ , find  $s = \text{round}(Q \times S)$ .
3. Use the corresponding  $\theta^{(s)}$  from the list of sorted posterior draws as quantiles from the error distribution. This approximation works well except for extreme tail quantiles.

#### 4.5 Diagnostic visualizations

In order to develop intuitions around the convergence of iterative algorithms, we propose several new diagnostic visualizations in addition to the numerical convergence diagnostics discussed above. We illustrate with several examples in Section 5.

**Rank plots.** Extending the idea of using ranks instead of the original parameter values, we propose using rank plots for each chain instead of trace plots. Rank plots are histograms of the ranked posterior draws (ranked over all chains) plotted separately for each chain. If all of the chains are targeting the same posterior, we expect the ranks in each chain to be uniform, whereas if one chain has a different location or scale parameter, this will be reflected in the deviation from uniformity. 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 than 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.



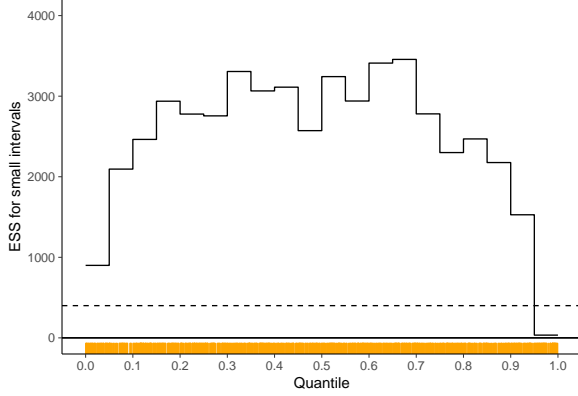


Figure 3: Local efficiency of small interval probability estimates for the Cauchy model with nominal parameterization. Orange ticks show iterations that exceeded the maximum tree depth in the dynamic HMC algorithm.

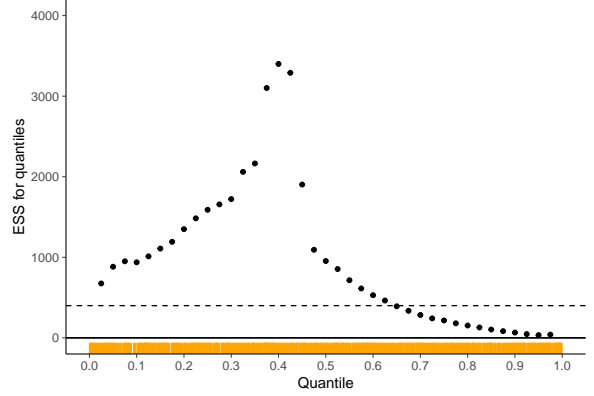


Figure 4: Efficiency of quantile estimates for the Cauchy model with nominal parameterization. Orange ticks show iterations that exceeded the maximum tree depth in the dynamic HMC algorithm.

## 5 Examples

We now demonstrate our approach and recommended workflow on several small examples. Unless mentioned otherwise, we use dynamic Hamiltonian Monte Carlo with multinomial sampling (Betancourt, 2017) as implemented in Stan (Stan Development Team, 2018b). We run 4 chains, each with 1000 warmup iterations, which do not form a Markov chain and are discarded, and 1000 post-warmup iterations, which are saved and used for inference.

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

Traditional  $\hat{R}$  is based on calculating within and between chain variances. If the marginal distribution of a chain is such that the variance is infinite, this approach is not well justified, as we demonstrate here with a Cauchy-distributed example.

#### Nominal parameterization of the Cauchy distribution

We start by simulating from independent Cauchy distributions for each element of a 50-dimensional vector  $x$ . Dynamic HMC-specific diagnostic, iterations that exceed the maximum tree depth, indicate slow mixing of the chains.

Several values of  $\hat{R}$  greater than 1.01 and some effective sample sizes less than 400 also indicate convergence problems. The online appendix contains more results with longer chains and other  $\hat{R}$  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 3 shows the local efficiency of small interval probability estimates (see Section 4.3). The efficiency of sampling is low in the tails, which is clearly caused by slow mixing in long tails of the Cauchy distribution. Figure 4 shows the efficiency of quantile estimates (see Section 4.3), which also is 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  $\hat{R}$ . 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 5, we see how unstable both bulk-ESS and tail-ESS are for this example. Rank plots in Figure 6 clearly show the mixing problem between chains. In

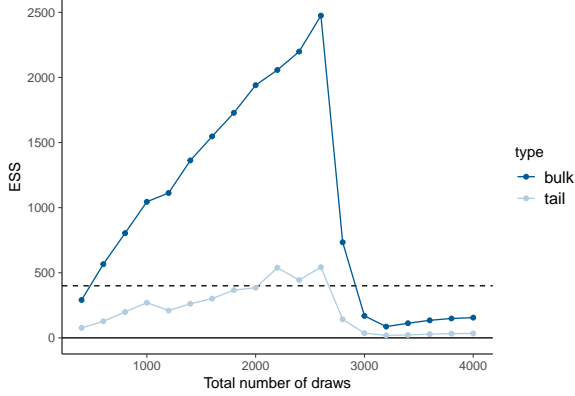


Figure 5: Estimated effective sample sizes with increasing number of iterations for the Cauchy model with nominal parameterization.

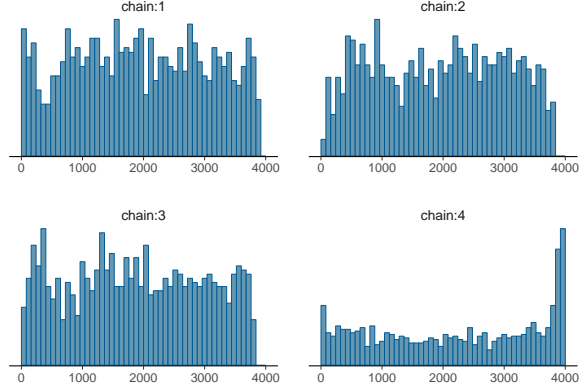


Figure 6: Rank plots of posterior draws from four chains for the Cauchy model with nominal parameterization.

case of good mixing all rank plots should be close to uniform. More experiments can be found in Appendix C and in the online appendix.

### Alternative parameterization of the Cauchy distribution

Next, we examine an alternative parameterization of the Cauchy as a scale mixture of Gaussians:

$$a_j \sim N(0, 1), \quad b_j \sim \text{Gamma}(0.5, 0.5), \quad x_j = a_j / \sqrt{b_j}. \quad (12)$$

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 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.

For all parameters,  $\hat{R}$  is less than 1.01 and ESS exceeds 400, indicating that sampling worked much better with this 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. The quantiles of each  $x_j$  are stable too, but the mean and variance estimates are widely varying. 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 7 and 8 show good sampling efficiency for the small interval probability and quantile estimates. The rank plots in Figure 9 also look close to uniform across chains, which is consistent with good mixing.

### Half-Cauchy distribution 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  $x$  constrained to be positive (in Stan, `<lower=0>`). Stan then automatically switches to the unconstrained  $\log(x)$  space, which changes the geometry crucially. With this transformations, all values of  $\hat{R}$  are less than 1.01 and ESS exceeds 400 for all parameters, indicating 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.

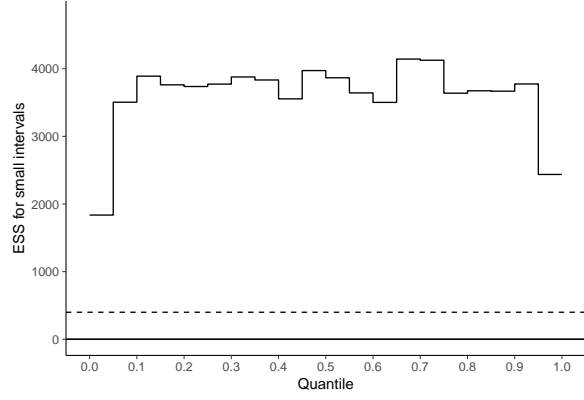


Figure 7: Local efficiency of small interval probability estimates for the Cauchy model with alternative parameterization.

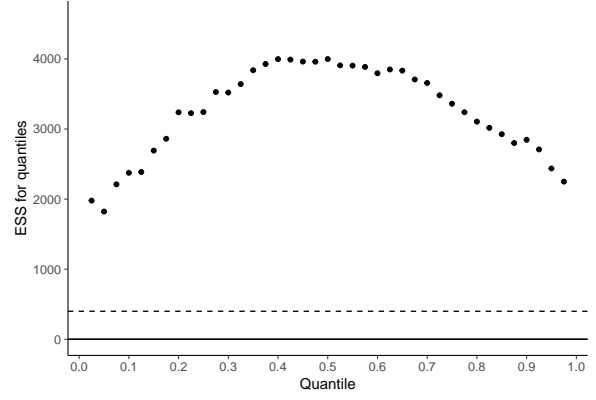


Figure 8: Efficiency of quantile estimates for the Cauchy model with alternative parameterization.

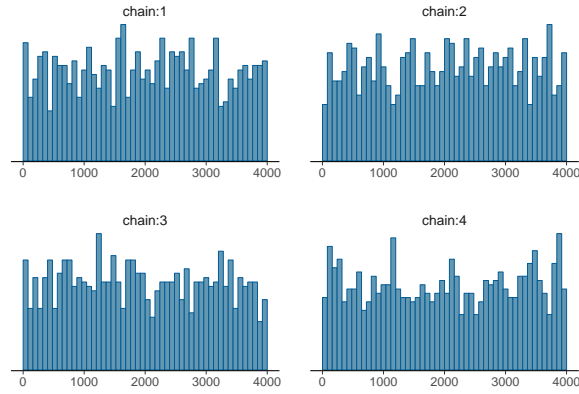


Figure 9: Rank plots of posterior draws from four chains for the Cauchy model with alternative parameterization.

## 5.2 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 simple model in at least two ways. The centered parameterization  $(\theta, \mu, \tau, \sigma)$  looks as follows:

$$\begin{aligned}\theta_j &\sim \text{normal}(\mu, \tau) \\ y_j &\sim \text{normal}(\theta_j, \sigma_j).\end{aligned}$$

In contrast, the non-centered parameterization  $(\tilde{\theta}, \mu, \tau, \sigma)$  can be written as:

$$\begin{aligned}\tilde{\theta}_j &\sim \text{normal}(0, 1) \\ \theta_j &= \mu + \tau \tilde{\theta}_j \\ y_j &\sim \text{normal}(\theta_j, \sigma_j).\end{aligned}$$

In both cases,  $\theta_j$  are the treatment effects in the eight schools, and  $\mu, \tau$  represent the population mean and standard deviation of the distribution of these effects. In the non-centered parameterization, the  $\theta$  are parameters, whereas in the centered parameterization, the  $\tilde{\theta}$  are parameters and  $\theta$  is a derived quantity.

Geometrically, the centered parameterization exhibits a funnel shape that contracts into a region of strong curvature around the population mean when faced with small values of the population standard deviation  $\tau$ , making it difficult for many simple Markov chain methods to adequately explore the full distribution of this parameter. 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 can also use  $\hat{R}$  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 10 and 11 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 12 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 consistently low. In line with the other findings, rank plots of  $\tau$  displayed in Figure 13 clearly show problems in the mixing of the chains. In particular, the rank plot for the first chain indicates that it was unable to explore the lower-end of the posterior range, while the spike in the rank plot for chain 2 indicates that it spent too much time stuck in these values. More experiments can be found in Appendix D and E as well as 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 settings as for the centered parameterization model. For the non-centered parameterization, we do not observe divergences or other warnings. All values of  $\hat{R}$  are less

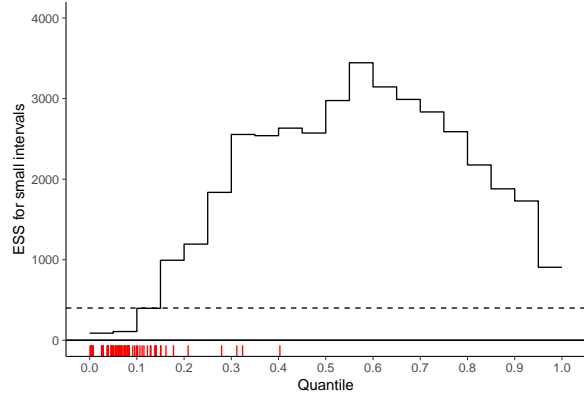


Figure 10: Local efficiency of small interval probability estimates for the eight schools model with centered parameterization. Red ticks show divergent transitions.

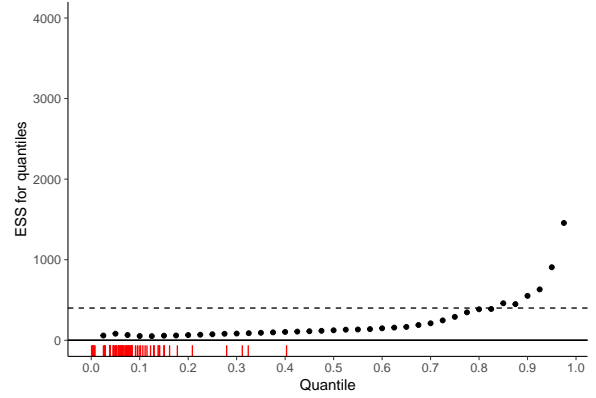


Figure 11: Efficiency of quantile estimates for the eight schools model with centered parameterization. Red ticks show divergent transitions.

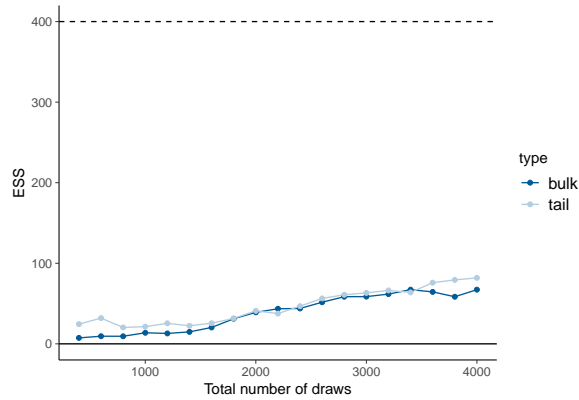


Figure 12: Estimated effective sample sizes with increasing number of iterations for the eight schools model with centered parameterization.

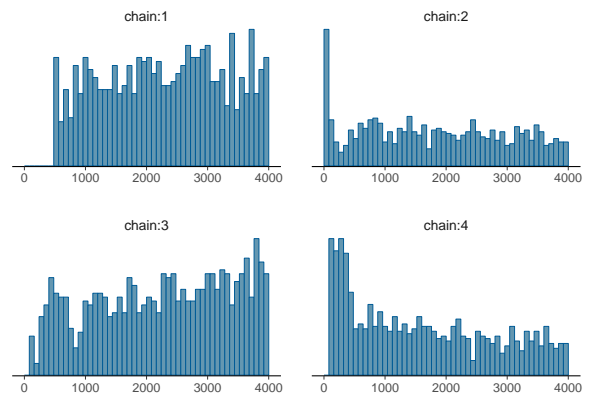


Figure 13: Rank plots of posterior draws from four chains for the eight schools model with centered parameterization.

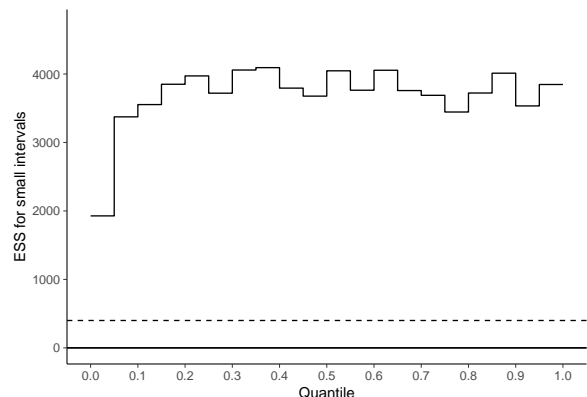


Figure 14: Local efficiency of small interval probability estimates for the eight schools model with the non-centered parameterization.

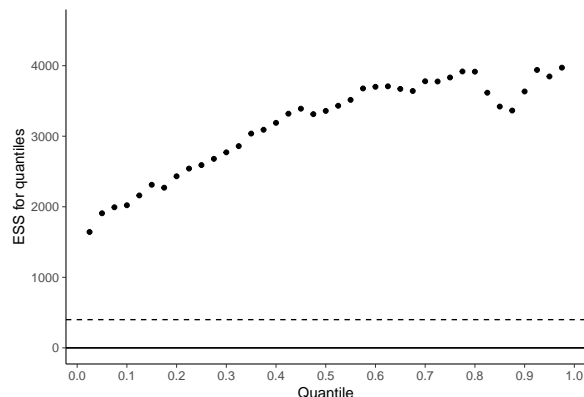


Figure 15: Efficiency of quantile estimates for the eight schools model with the non-centered parameterization.

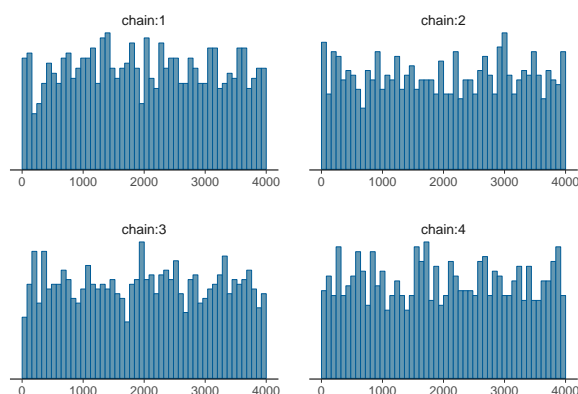


Figure 16: Rank plots of posterior draws from four chains for 8 schools model with non-centered parameterization.

than 1.01 and ESS exceeds 400, indicating a much better efficiency of the non-centered parameterization. Figures 14 and 15 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 16 show no substantial differences between chains.

## References

- Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *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.
- Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. doi: 10.18637/jss.v076.i01. URL <https://www.jstatsoft.org/v076/i01>.
- Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7(4):457–511, 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.
- Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–483, 1992.
- Charles 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 Lunnon, 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.
- Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006. URL <https://journal.r-project.org/archive/>.
- Stan Development Team. RStanArm: Bayesian applied regression modeling via Stan. R package version 2.17.4, 2018a. URL <http://mc-stan.org>.
- Stan Development Team. Stan modeling language users guide and reference manual. version 2.18.0, 2018b. URL <http://mc-stan.org>.
- Dootika Vats and Christina Knudson. Revisiting the Gelman-Rubin diagnostic. *arXiv:1812.09384*, 2018.

## Appendix A: Computing the 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  $\bar{\theta}^{(\cdot)}$  as an estimate of the posterior mean  $E(\theta|y)$ . This also generalizes to posterior expectations of functionals of parameters  $E(g(\theta)|y)$ . Section 4.3 deals with estimating 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

$$\rho_t = \frac{1}{\sigma^2} \int_{\Theta} (\theta^{(n)} - \mu)(\theta^{(n+t)} - \mu) p(\theta) d\theta. \quad (13)$$

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

$$\rho_t = \frac{1}{\sigma^2} \int_{\Theta} \theta^{(n)} \theta^{(n+t)} p(\theta) d\theta. \quad (14)$$

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

$$N_{\text{eff}} = \frac{N}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{N}{1 + 2 \sum_{t=1}^{\infty} \rho_t}. \quad (15)$$

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 sample itself. The rest of this section describes an autocorrelation and split- $\hat{R}$  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  $m \in (1, \dots, M)$  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,

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M} \sum_{m=1}^M \hat{\rho}_{t,j}}{\widehat{\text{var}}^+}. \quad (16)$$

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,

$$S_{\text{eff}} = \frac{NM}{\hat{\tau}}, \quad (17)$$

where

$$\hat{\tau} = 1 + 2 \sum_{t=1}^{2k+1} \hat{\rho}_t = -1 + 2 \sum_{t'=0}^k \hat{P}_{t'}, \quad (18)$$

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  $t' = 1, \dots, k$ . 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 distinct modes that are found.

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

Here we demonstrate the behavior of non-split- $\hat{R}$ , split- $\hat{R}$ , 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  $N(0, 1)$  distribution plus a linear trend. Figure 17 shows that if we don't split chains,  $\hat{R}$  misses the trends if all chains still have a similar marginal distribution. Figure 18 shows that split- $\hat{R}$  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- $\hat{R}$  and within-chain autocorrelation. Figure 19 shows the relative bulk-ESS divided by  $S$  for easier comparison between different values of  $S$ . We see that split- $\hat{R}$  is more sensitive to trends for small sample sizes, but ESS becomes more sensitive for larger sample sizes (as autocorrelations can be estimated more accurately).

**Shifting one chain.** Second we draw all the chains are from the same  $N(0, 1)$  distribution, except one that is sampled with nonzero mean. Figure 20 shows that if we use a threshold of 1.01, split- $\hat{R}$  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- $\hat{R}$  detects shifts with a magnitude equal to or larger than the marginal standard deviation. Figure 21 shows the the relative bulk-ESS for the same case. The effective sample size is not as sensitive as split- $\hat{R}$ , but a shift with a magnitude of half the marginal standard deviation or more will lead to low relative efficiency when the total number of draws increases. Rank plots are practical way to visualize differences between chains. Figure 22 shows rank plots for the case of 4 chains, 250 draws per chain, and one chain sampled with

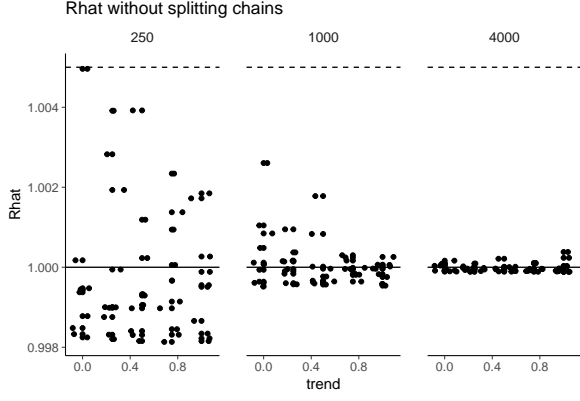


Figure 17:  $\hat{R}$  without splitting for varying chain lengths for chains which have the same trend and a similar marginal distribution.

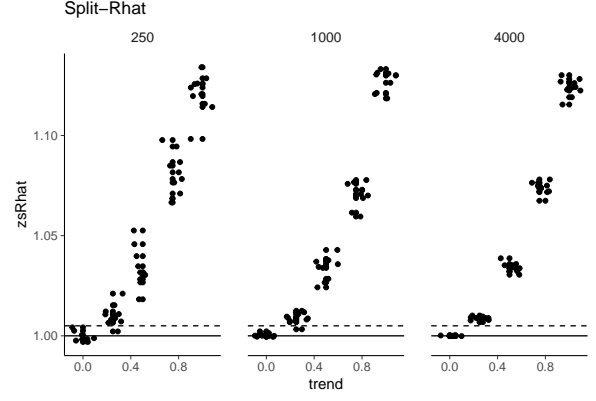


Figure 18: Split- $\hat{R}$  for varying chain lengths for chains which have the same trend and a similar marginal distribution.

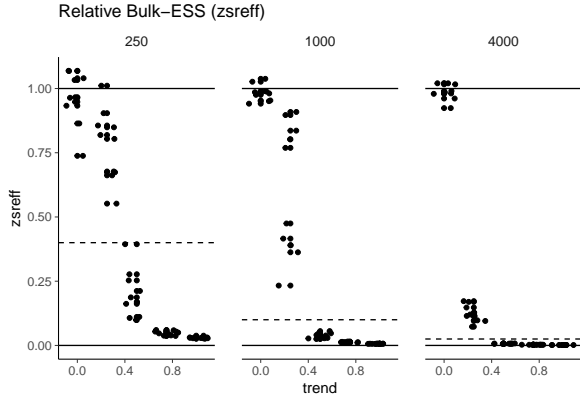


Figure 19: Relative bulk-ESS for varying chain lengths for chains which have the same trend and a similar marginal distribution. The dashed lines indicate the threshold  $S_{\text{eff}} > 400$  at which we would consider the effective sample size to be sufficient.

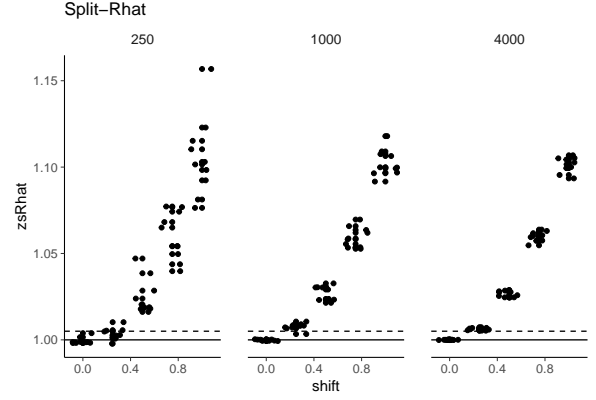


Figure 20: Split- $\hat{R}$  for varying chain lengths for chains with one sampled with a different mean than the others.

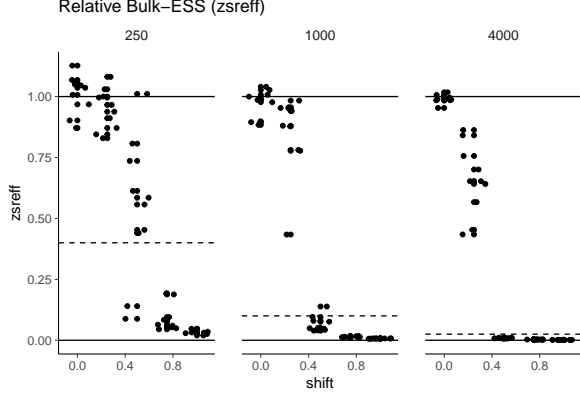


Figure 21: Relative bulk-ESS for varying chain lengths for chains with one sampled with a different mean than the others. The dashed lines indicate the threshold  $S_{\text{eff}} > 400$  at which we would consider the effective sample size to be sufficient.

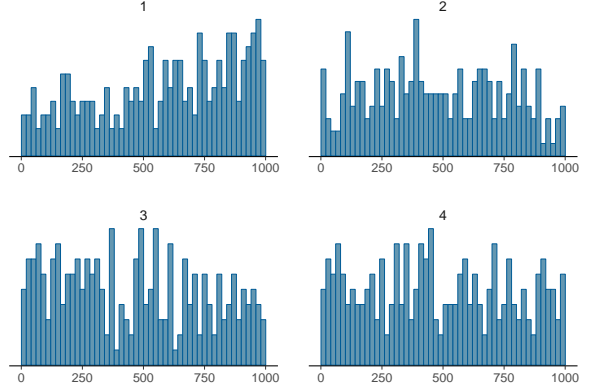


Figure 22: Rank plots of posterior draws from four chains with one sampled with a different mean than the others.

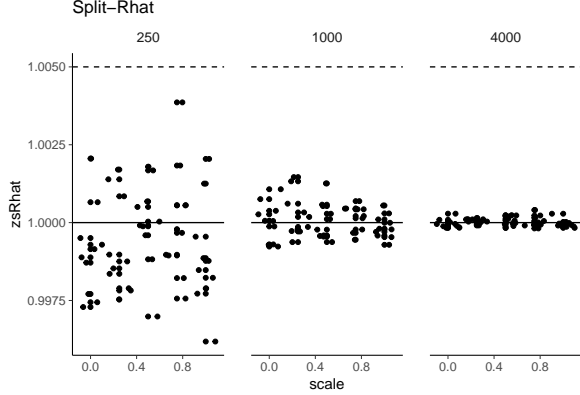


Figure 23: Split- $\hat{R}$  for varying chain lengths for chains with one sampled with a different variance than the others.

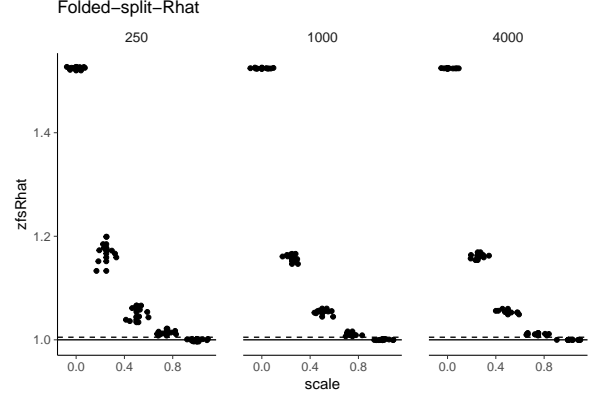


Figure 24: Folded-split- $\hat{R}$  for varying chain lengths for chains with one sampled with a different variance than the others.

mean 0.5 instead of 0. In this case split- $\hat{R} = 1.05$ , 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  $N(0, 1)$  distribution, except one of the chains is sampled with variance less than 1. Figure 23 shows that split- $\hat{R}$  is not able to detect scale differences between chains. Figure 24 shows that folded-split- $\hat{R}$  which focuses on scales detects scale differences. With a threshold of 1.01, folded-split- $\hat{R}$  detects a chain with scale less than  $3/4$  of the standard deviation of the others. With a threshold of 1.1, folded-split- $\hat{R}$  detects a chain with standard deviation less than  $1/4$  of the standard deviation of the others.

Figure 25 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 26 shows rank plots for the case of 4 chains, 250 draws per chain, and one chain sampled with standard deviation 0.75 instead of 1. Although folded-split- $\hat{R} = 1.06$ , the rank plots clearly show that the first chain behaves differently.

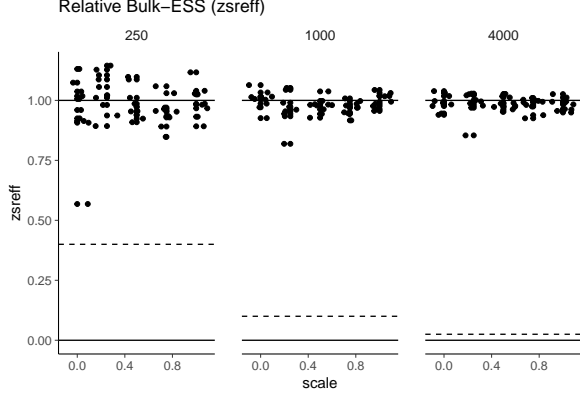


Figure 25: Relative bulk-ESS for varying chain lengths for chains with one sampled with a different variance than the others.

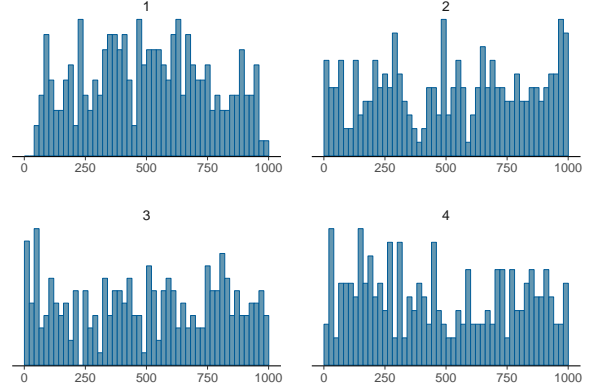


Figure 26: Rank plots of posterior draws from four chains with one sampled with a different variance than the others.

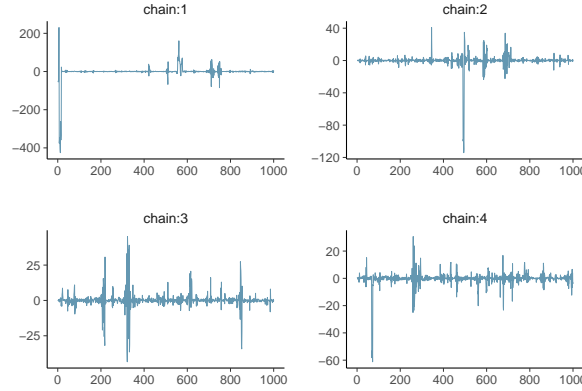


Figure 27: Trace plots of four chains for Cauchy model with nominal parameterization and `max_treedepth=20`.

## Appendix C: More experiments with the Cauchy distribution

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 27 shows that trace plots for the first parameter look wild with occasional large values, and it is difficult to interpret possible convergence. Figure 28 shows traditional  $\widehat{R}$ , rank normalized  $\widehat{R}$ , and rank normalized folded-split- $\widehat{R}$  for all 50 parameters. Traditional 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 28 shows different effective sample size estimates for all 50 parameters. Traditional ESS, which is not well defined in this case, has 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, indicating practically useful MAD estimates, too. The online appendix has additional results with longer chains, showing that all other ESS values except traditional ESS (which is not well defined) keep improving with more iterations. It is however recommended to use a more efficient parameterization especially if the tail quantiles are of interest.

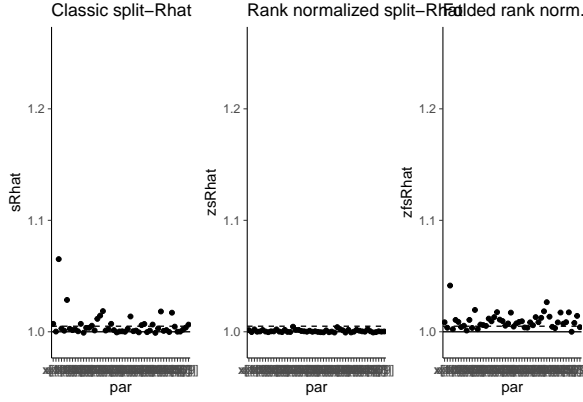


Figure 28: Traditional split- $\hat{R}$ , rank normalized split- $\hat{R}$ , and rank normalized folded-split- $\hat{R}$  for Cauchy model with nominal parameterization and  $\text{max\_treedepth}=20$ .

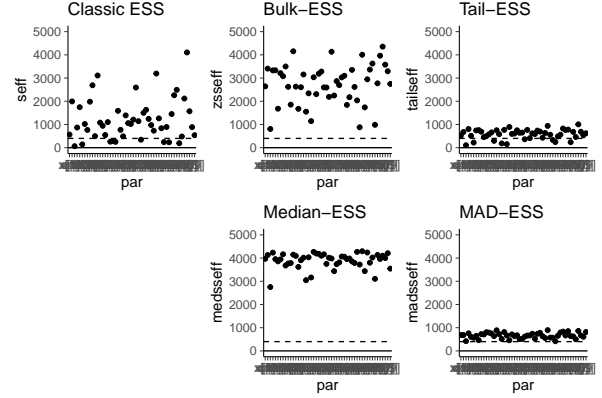


Figure 29: Traditional ESS, bulk-ESS, tail-ESS, median-ESS and MAD-ESS for Cauchy model with nominal parameterization  $\text{max\_treedepth}=20$ .

## Appendix D: A centered eight schools model with very long chains and thinning

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

When autocorrelation time is high, it is common practice to thin the chains by saving only a small portion of the draws. This can throw away useful information also for convergence diagnostics, as we demonstrate with the eight schools model with centered parameterization, run with  $4 \times 10^5$  iterations per chain, first half removed as warm-up, and thinned by taking every 200th iteration.

We observe several divergent transitions and the estimated Bayesian fraction of missing information (Betancourt, 2017) is also low, which can indicate convergence problems.

Figures 30, 31, and 32 show the efficiency of small probability interval estimates, efficiency of quantile estimates, and change of bulk-ESS and tail-ESS with increasing number of iterations. Unfortunately, after thinning, split- $\hat{R}$  and ESS miss the problems. The posterior mean is still off, being more than 3 standard deviations 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 33 are still able to show the problem.

## Appendix E: 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 also useful for samplers other than Hamiltonian Monte Carlo. We fit the eight schools models also with 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 with strongly dependent parameters (see, e.g. Hoffman and Gelman, 2014), but it can work fine for relatively simple models such as in this case study.

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- $\hat{R} = 1.08$ , 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 report the results with 10 000 iterations. Examining the diagnostics for  $\tau$ , now split- $\hat{R} = 1.01$ , bulk-ESS= 677, and tail-ESS= 1027, which are all good.

Figures 34, 35, and 36 show the efficiency of small probability interval estimates, efficiency of quantile

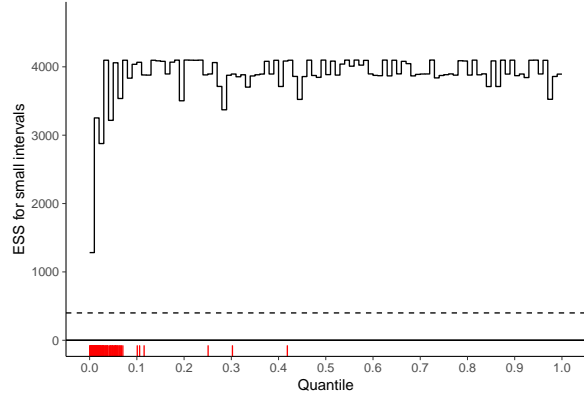


Figure 30: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization, very long chains, and thinning.

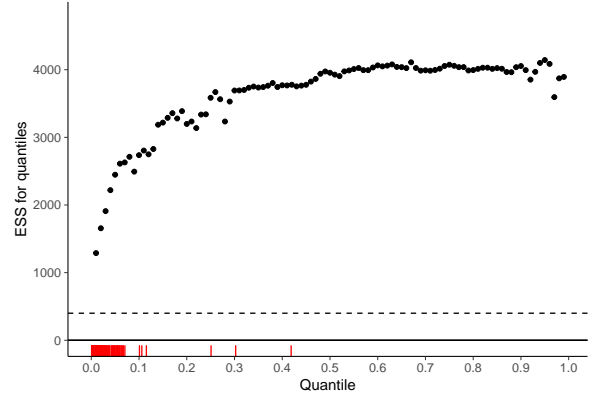


Figure 31: The efficiency of quantile estimates for 8 schools model with centered parameterization, very long chains, and thinning.

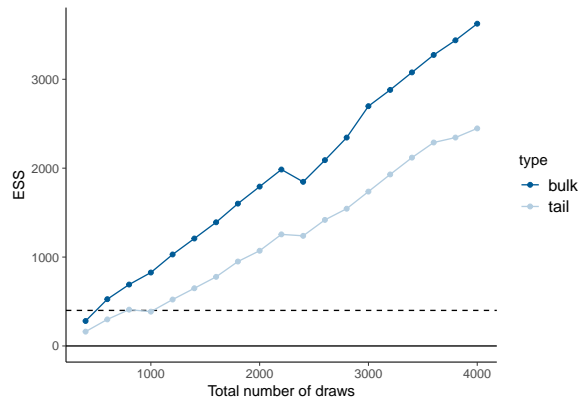


Figure 32: The estimated effective sample sizes with increasing number of iterations for 8 schools model with centered parameterization, very long chains, and thinning.

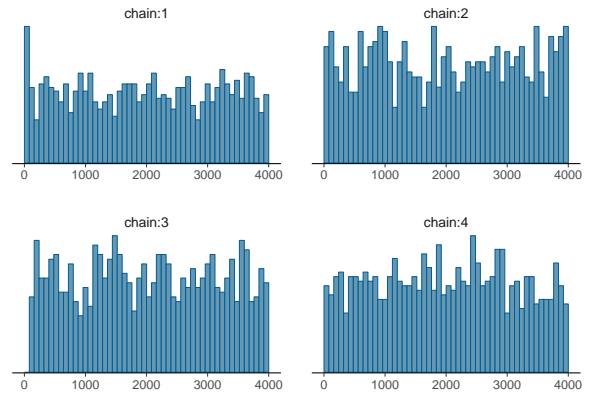


Figure 33: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization, very long chains, and thinning.

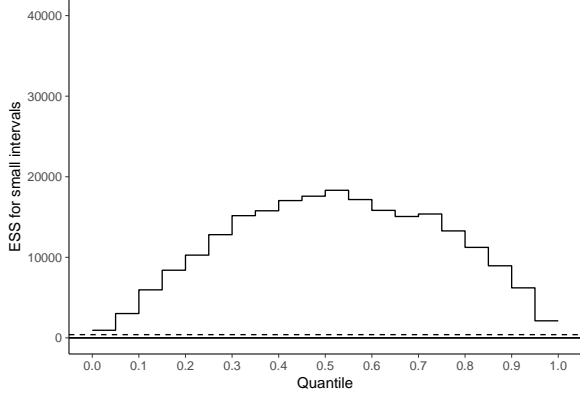


Figure 34: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization and Gibbs sampling.

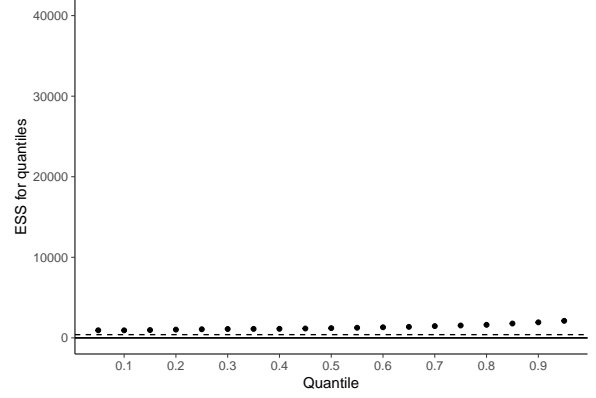


Figure 35: The efficiency of quantile estimates for 8 schools model with centered parameterization and Gibbs sampling.

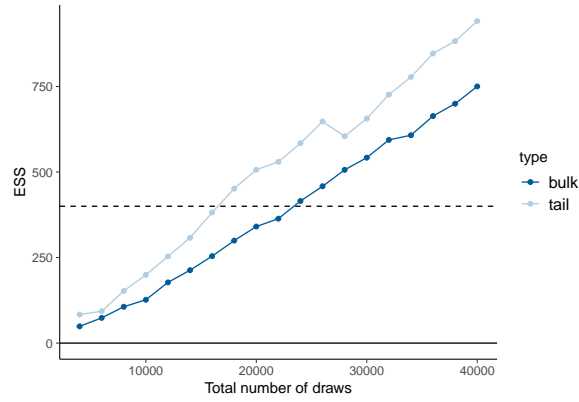


Figure 36: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization and Gibbs sampling.

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, the increase in effective sample size for  $\tau$  is linear in the total number of draws. A 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 a more efficient non-centered parameterization.