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

Aki Vehtari[†], *Andrew Gelman*[‡], *Daniel Simpson*[§], *Bob Carpenter*[¶], *and Paul-Christian Bürkner*[†]

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. Traditional $\widehat{R}$ will fail to correctly diagnose convergence failures when the chain has a heavy tail or when the variance varies across the chains. In this paper we propose an alternative rank-based diagnostic that fixes these problems. 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 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 finite time. 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 can also visually inspect the sample paths of the chains via trace plots 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.

Of the various convergence diagnostics (see reviews by Cowles and Carlin, 1996; Mengersen et al., 1999; Robert and Casella, 2004), probably the most widely used is the potential scale reduction factor $\widehat{R}$ (Gelman and Rubin, 1992;

*We thank Ben Bales, Ian Langmore, the editor, and anonymous reviewers for useful comments. We also thank Academy of Finland, the U.S. Office of Naval Research, National Science Foundation, Institute for Education Sciences, 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.

[†]Department of Computer Science, Aalto University, Finland.
[‡]Department of Statistics, Columbia University, New York.
[§]Department of Statistical Sciences, University of Toronto, Canada.
[¶]Applied Statistics Center, Columbia University, New York.

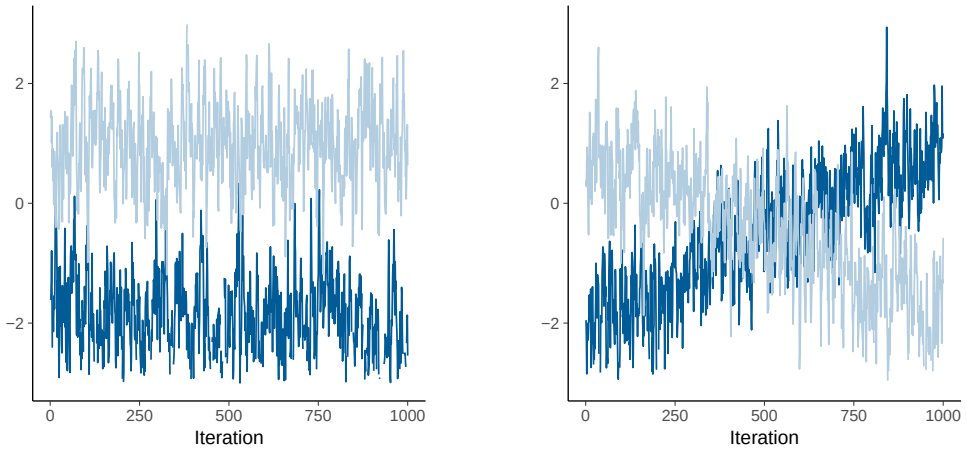

Figure 1: Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the two sequences happen to cover a common 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).

Brooks and Gelman, 1998). It is recommended as the primary convergence diagnostic in widely applied software packages for MCMC sampling such as Stan (Carpenter et al., 2017), JAGS (Plummer, 2003), WinBUGS (Lunn et al., 2000), OpenBUGS (Lunn et al., 2009), PyMC3 (Salvatier et al., 2016), and NIMBLE (de Valpine et al., 2017), which together are estimated to have hundreds of thousand of users. $\widehat{R}$ is computed for each scalar quantity of interest, as the standard deviation of that quantity from all the chains included together, divided by the root mean square of the separate within-chain standard deviations. The idea is that if a set of simulations have not mixed well, the variance of all the chains mixed together should be higher than the variance of individual chains. More recently, Gelman et al. (2013) introduced split-$\widehat{R}$ 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 $\widehat{R}$ we are always speaking of the split-$\widehat{R}$ variant.

Convergence diagnostics are most effective when computed using multiple chains initialized at a diverse set of starting points. This reduces 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 $\widehat{R}$ with diverse seeding as an operationalization of the qualitative statement that, after warmup, 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.

1.2 Example where traditional $\widehat{R}$ fails

Unfortunately, $\widehat{R}$ can fail to diagnose poor mixing, which can be a problem when it is used as a default rule. The following example shows how failure can occur.

The red histograms in Figure 2 show the distribution of $\widehat{R}$ (that is, split-$\widehat{R}$ from Gelman et al. (2013)) in four different scenarios. (Ignore the light blue histograms for now; they show the results using an improved diagnostic that we shall discuss later in this paper.) In all four scenarios, traditional $\widehat{R}$ is well under 1.1 under all simulations, thus not detecting any convergence problems—but in fact the two scenarios on the left have been constructed so that they are far from mixed. These are problems that are not detected by traditional $\widehat{R}$.

In each of the four scenarios in Figure 2, we run four chains for 1000 iterations each and then replicate the entire

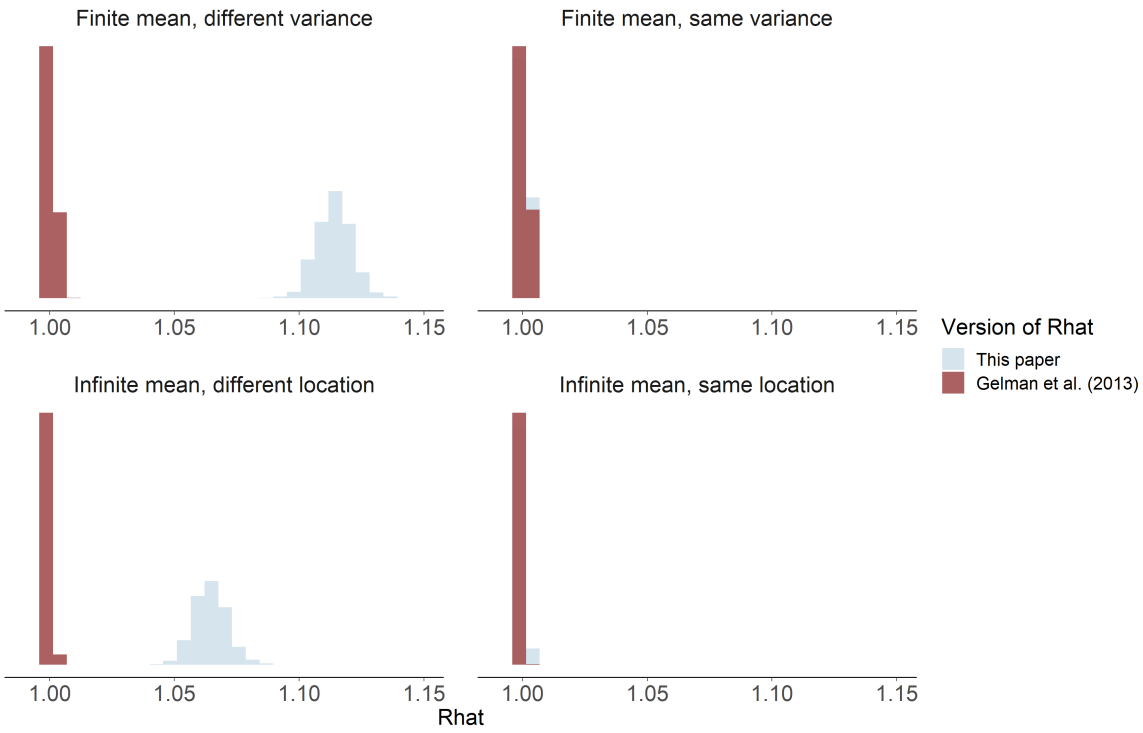

Figure 2: An example showing problems undetected by traditional $\widehat{R}$. Each plot shows histograms of $\widehat{R}$ values over 1000 replications of four chains, each with a thousand draws. In the left column, one of these four chains was incorrect. In the top left plot, we set one of the four chains to have a variance lower than the others. In the bottom left plot, we took one of the four chains and shifted it. In both cases, the traditional $\widehat{R}$ 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.

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 $\widehat{R}$ when one of the four chains is manually transformed to only have $1/3$ of the variance compared to the other three chains (see Appendix A for more details). This corresponds to a scenario where one chain fails to correctly explore the tails of the target distribution and one would hope could be identified as non-convergent. The split-$\widehat{R}$ statistic defined in Gelman et al. (2013) does not detect the poor mixing, while the new variant of split-$\widehat{R}$ 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 $\widehat{R}$ values correctly identify that mixing occurs.

The second row of Figure 2 shows the behavior of $\widehat{R}$ 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 manually shifted one of the four chains two units to the right. This corresponds to a scenario where one chain provides a biased estimate of the target distribution. The Gelman et al. (2013) version of $\widehat{R}$ would catch this behavior if the chain had finite variance, but in this case the infinite variance destroys its effectiveness—traditional $\widehat{R}$ and split-$\widehat{R}$ are defined based on second-moment statistics—and it inappropriately returns a value very close to 1.

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

1. If the chains have different variances but the same mean parameters, traditional $\widehat{R} \approx 1$.
2. If the chains have infinite variance, traditional $\widehat{R} \approx 1$ 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 hard to assess empirically if a chain has large but finite variance or infinite variance.

A related problem is that $\widehat{R}$ 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.

2 Recommendations for practice

The traditional $\widehat{R}$ statistic is general, easy to compute, and can catch many problems of poor convergence, but the discussion above reveals some scenarios where it fails. The present paper proposes improvements that overcome these problems. In addition, as the convergence of the Markov chain needs not be uniform across the parameter space, we propose a localized version of effective sample size that allows 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.

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 $\widehat{R}$ and effective sample size (ESS), which are useful as first level checks when analyzing reliability of inference for many quantities. However, these values should be adapted as necessary for the given application, and ultimately domain expertise should be use to check that Monte Carlo standard error (MCSE) for all quantities of interest are small enough.

In Section 4, we propose modifications to $\widehat{R}$ based on rank-normalizing and folding the posterior draws, only using the sample if $\widehat{R} < 1.01$. This threshold is much tighter than the one recommended by Gelman and Rubin (1992), reflecting lessons learnt over more than 25 years of use, as well as the simulation results in Appendix A. Gelman and Rubin (1992) derived $\widehat{R}$ under the assumption that, as simulations went forward, the within-chain variance would gradually increase while the between-chain variance decreased, stabilizing when their ratio was 1. The potential scale reduction factor represented the factor by which the between-chain variation might decline under future simulations, and a potential scale reduction factor of 1.1 implied that there was little to be gained in inferential precision by running the chains longer. However, as discussed by Brooks and Gelman (1998), the dynamics of MCMC are such that the between-chain variance can decrease before it increases, if the initial part of the simulation pulls all the chains to the center of the distribution, only for them to be redispersed with further simulation. As a result, $\widehat{R}$ cannot in general be interpreted as a potential scale reduction factor, and in practice and in simulations we have found that $\widehat{R}$ can dip below 1.1 well before convergence in some examples (a point also raised by Vats and Knudson (2018)), and we have found this to be much more rare when using the 1.01 threshold.

In addition, we recommend running at least four chains by default. Multiple chains are more likely to reveal multimodality and poor adaptation or mixing: we see examples for complex, misspecified or non-identifiable models in the Stan discussion forum all the time. Furthermore, most computers are able to run chains in parallel, giving multiple chains with no increase in computation time. Here we do not consider massive parallelization such as running 1000 chains or more; further research is needed in considering how to use such simulations most efficiently in such computational environments (see, for instance, the method discussed in Jacob et al. (2017)).

Roughly speaking, the effective sample size 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. The higher the ESS the better. When there might be difficulties with mixing, it is important to use between-chain as well as within-chain information in computing the ESS. A common example arises in hierarchical models with funnel-shaped posteriors, where MCMC algorithms can struggle to simultaneously adapt to a "narrow" region of high density and low volume, and a "wide" region of low density and high volume. In such a case, differences in step-size adaptation can lead to chains that have different behavior in the neighborhood of the narrow part of the funnel (Betancourt and Girolami, 2019). For multimodal distributions with well-separated modes, the split-$\widehat{R}$ 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.

A small value of $\widehat{R}$ is not enough to ensure that an MCMC sample is useful in practice (Vats and Knudson, 2018). The effective sample size must also be large enough to get stable inferences for quantities of interest. Gelman et al. (2013) proposed an ESS estimate which combines autocovariance-based single-chain variance estimates (Hastings, 1970; Geyer, 1992) from multiple chains using between- and within-chain information as in $\widehat{R}$. In Section 3.2 we

propose an improved algorithm, and as with $\widehat{R}$, 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 *rank-normalized* version of the quantity of interest, using the rank transformation followed by the inverse normal transformation. 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.

To ensure reliable estimates of variances and autocorrelations needed for $\widehat{R}$ and ESS, we recommend requiring that the rank-normalized ESS is greater than 400, a number we chose based on practical experience and simulations (see Appendix A) as typically sufficient to get a stable estimate of the Monte Carlo standard error.

Finally, when reporting 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, that is, convergence might be different in the bulk of the distribution (e.g., for the mean or median) than in the tails (e.g., for extreme quantiles). We propose diagnostics and effective sample sizes specifically for extreme quantiles. This is different from the standard ESS estimate (which we refer to as 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 $\widehat{R}$ and the effective sample size

When coupled with an ESS estimate, $\widehat{R}$ 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 (see, e.g. Vats and Knudson, 2018), but we prefer to treat these as two separate questions: "Did the chains mix well?" (split-$\widehat{R}$) and "Is the effective sample size large enough to get a stable estimate of uncertainty?" In this section we define the $\widehat{R}$ and ESS statistics that we propose to modify.

3.1 Split-$\widehat{R}$

Here we present split-$\widehat{R}$, 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 θ, we compute B and W, the between- and within-chain variances:

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

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

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 θ might be reduced if the simulations were continued in the limit $N \to \infty$. This leads to the estimator

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

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 simulated chains. Without splitting, $\widehat{R}$ would get fooled by non-stationary chains as in Figure 1b.

In cases, where we can be absolutely certain that a single chain is sufficient, $\widehat{R}$ could be computed using only single chain marginal variance and autocorrelations (see, e.g. Vats and Knudson, 2018). 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.

3.2 The effective sample size

We estimate effective sample size by combining information from $\widehat{R}$ and the autocorrelation estimates within the chains.

Given S independent simulation draws, the accuracy of average of the simulations $\bar{\theta}$ as an estimate of the posterior mean $\text{E}(\theta|y)$ can be estimated as

$$\text{Var}(\bar{\theta}) = \frac{\text{Var}(\theta|y)}{S}. \tag{5}$$

This generalizes to posterior expectations of functionals of parameters $\text{E}\left(g(\theta)|y\right)$. The square root of (5) is called the Monte Carlo standard error (MCSE).

In general, the simulations of θ within each chain tend to be autocorrelated, and $\text{Var}(\bar{\theta})$ can be larger or smaller in expectation. In the early days of using MCMC for Bayesian inference, the focus was in estimating the single chain estimate variance directly, for example, based on autocorrelations or batch means (Hastings, 1970; Geyer, 1992). See more different variance estimation algorithms in reviews by Cowles and Carlin (1996), Mengersen et al. (1999), and Robert and Casella (2004). Interpreting whether Monte Carlo standard error for a quantity of interest is small enough requires domain expertise.

Effective sample size (ESS) can be computed by dividing any variance estimate for an MCMC estimate by the variance estimate assuming independent draws. As convergence diagnostics in general started to be more popular (Gelman and Rubin, 1992; Cowles and Carlin, 1996; Mengersen et al., 1999; Robert and Casella, 2004), eventually ESS also became popular as description of the efficiency of the simulation (an early example of reporting ESS for Gibbs sampler is Sorensen et al., 1995). The term effective sample size had already been used before, for example, to describe amount of information in climatological time series (Laurmann and Gates, 1977) and the efficiency of importance sampling in Bayesian inference (Kong et al., 1994).

Although ESS is not a replacement for MCSE, it can provide a scale-free measure of information, which can be especially useful when diagnosing the sampling efficiency for a large number of variables. The downside of the term effective sample size is that it may give a false impression that the dependent simulation sample would be equivalent to an independent simulation sample with size ESS, while the equivalence is only for the estimation efficiency of the posterior mean, and the efficiency of the same dependent simulation sample for estimating another posterior functional $\text{E}\left(g(\theta)|y\right)$ or quantiles can be very different. To simplify notation, in this section we consider the effective sample size for the posterior mean $\text{E}\left(\theta|y\right)$. This can be generalized in a straightforward manner to ESS estimates for $\text{E}\left(g(\theta)|y\right)$. Section 4.3 deals with estimating the effective sample size of quantiles, which cannot be presented as expectations.

The first proposals of ESS estimates used information only from a single chain (see, e.g. Sorensen et al., 1995). The convergence diagnostic package `coda` (Plummer et al., 2006) combines (since version 0.5.7 in 2001) single chain spectral variance based ESS estimates simply by summing them, but this approach gives over-optimistic estimates if spectral variances in different chains are not equal (e.g. when different step size is used in different chains) or if chains

are not mixing well. Gelman et al. (2003) proposed an ESS estimate,

$$S_{\text{eff,BDA2}} = MN\frac{\widehat{\text{var}}^{+}}{B}, \tag{6}$$

where $\widehat{\text{var}}^{+}$ is a marginal posterior variance estimate and B is between-chain variance estimate as given in Section 3.1. This corresponds to a batch means approach with each chain being one batch. As there are usually only a small number of batches (chains), and information from autocorrelations is not used, this ESS estimate has high variance. Gelman et al. (2013) proposed an ESS estimate which appropriately combines autocorrelation information from multiple chains. Stan Development Team (2018b) made some computational improvements, and the present article provides a further improved version.

For a single chain of length N, the effective sample size of a chain can defined in terms of the autocorrelations within the chain at different lags,

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

where ρ_t is autocorrelation at lag $t \geq 0$. An equivalent approach was used by Hastings (1970) for estimating the variance of the mean estimate from a single chain. For a chain with joint probability function $p(\theta)$ with mean μ and standard deviation σ, ρ_t is defined to be

$$\rho_t = \frac{1}{\sigma^2}\int_{\Theta}(\theta^{(n)}-\mu)(\theta^{(n+t)}-\mu)\,p(\theta)\,d\theta. \tag{8}$$

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 at convergence, multiplying the two difference terms and reducing yields,

$$\rho_t = \frac{1}{\sigma^2}\int_{\Theta}\theta^{(n)}\,\theta^{(n+t)}\,p(\theta)\,d\theta. \tag{9}$$

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. Computations of autocorrelations for all lags simultaneously can be done efficiently via the fast Fourier transform algorithm (FFT; see Geyer, 2011). In our experiments, FFT-based autocorrelation estimates have also been computationally more accurate than naive autocovariance computation. As recommended by Geyer (1992) we use the biased estimate with divisor N, instead of unbiased estimate with divisor $N-t$. Also in our experiments, the biased estimate provided smaller variance in the final ESS estimate.

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 $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$ and the multi-chain variance estimate $\widehat{\text{var}}^{+} = W(N-1)/N + B/N$ to compute the combined autocorrelation at lag t as,

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{m=1}^{M} s_m^2\hat{\rho}_{t,m}}{\widehat{\text{var}}^{+}}. \tag{10}$$

If $\hat{\rho}_{t,m} = 0$ for all m, $\hat{\rho}_t = 1 - \widehat{R}^{-2}$. If in addition chains are mixing well so that $\widehat{R} \approx 1$, then $\hat{\rho}_t \approx 0$. If $\hat{\rho}_{t,m} \neq 0$ and $\widehat{R} \approx 1$, then $\hat{\rho}_t \approx \frac{1}{M}\sum_{m=1}^{M}\hat{\rho}_{t,m}$. If $\widehat{R} \gg 1$, then $\hat{\rho}_t \approx 1 - \widehat{R}^{-2}$. If chains are mixing well, this expression is equivalent to averaging autocorrelations, and if chains are not mixing well, simulations in each chain are implicitly assumed to be more correlated with each other. In our experiments, multi-chain ρ_t given by (10) and FFT-based $\hat{\rho}_{t,m}$ had smaller variance than the related multi-chain ρ_t proposed by Gelman et al. (2013).

As noise in the correlation estimates $\hat{\rho}_t$ increases as t increases, the large-lag terms need to be down weighted (lag window approach, see, e.g. Geyer, 1992; Flegal and Jones, 2010) or the sum of $\hat{\rho}_t$ can be truncated with some truncation lag T to get

$$S_{\text{eff}} \;=\; \frac{NM}{1+2\sum_{t=1}^{T}\rho_t}. \tag{11}$$

We use a truncation rule proposed by Geyer (1992), which takes into account certain properties of the autocorrelations for Markov chains. Even when the simulations are constructed using an MCMC algorithm, the time series of simulations for a scalar parameter or summary will not in general have the Markov property; nonetheless we have found these Markov-derived heuristics to work well in practice. In our experiments, Geyer's truncation had superior stability compared to flat-top (e.g. Doss et al., 2014) and slug-sail (Vats and Knudson, 2018) lag window approaches.

For Markov chains typically used in MCMC, 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

$$S_{\text{eff}} = \frac{N\,M}{\hat{\tau}}, \tag{12}$$

where

$$\hat{\tau} = 1 + 2\sum_{t=1}^{2k+1}\hat{\rho}_t = -1 + 2\sum_{t'=0}^{k}\hat{P}_{t'}, \tag{13}$$

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.

In case of antithetic Markov chains, which have negative autocorrelations on odd lags, the effective sample size S_{eff} can also be larger than S . For example, the dynamic Hamiltonian Monte Carlo algorithms used in Stan (Hoffman and Gelman, 2014; Betancourt, 2017; Stan Development Team, 2018b) is likely to produce $S_{\text{eff}} > S$ for parameters with a close to Gaussian posterior (in the unconstrained space) and low dependence on the other parameters. The benefit of this kind of super-efficiency is often limited as it is unlikely to simultaneously have super-efficiency for mean and variance (or tail quantiles) as demonstrated in our experiments.

In extreme antithetic cases, magnitude of single lag autocorrelations can stay large for a large lag t, even if the paired autocorrelations are close to zero. To improve the stability and reduce the variance of the ESS estimate, we determine the truncation lag as usual, but compute the average of truncated sum ending to usual odd lag and truncated sum ending to the next even lag. Sometimes these estimates are used for very short antithetic chains, and just by chance there can be strange estimates, and as highly antithetic chains are unlikely, in our software implementation we have restricted the ESS estimate to an upper bound of $S\log_{10}(S)$.

The effective sample size S_{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_{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_{eff} is close to the number of distinct modes that are found. Thus, our ESS estimate can be also to diagnose multimodality.

The values of $\widehat{R}$ and ESS require reliable estimates of variances and autocorrelations (in addition to the existence of these quantities; see our Cauchy examples in Section 5.1), which can only occur if the chains have enough independent replicates. In particular, we only recommend relying on the $\widehat{R}$ 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 our minimum recommended setup of four parallel chains, the total ESS should be at least 400 before we expect $\widehat{R}$ to be useful.

4 Improving convergence diagnostics

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

As split-$\widehat{R}$ and S_{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.

The use of ranks to avoid the assumption of normality goes back to Friedman (1937). Chernoff and Savage (1958) show rank based approaches have good asymptotic efficiency. Instead of using rank values directly and modifying tests for

them, Fisher and Yates (1938) propose to use expected normal scores (ordered statistics) and use the normal models. Blom (1958) shows that accurate approximation of the expected normal scores can be computed efficiently from ranks using an inverse normal transformation.

Rank normalized split-$\widehat{R}$ and S_{eff} are computed using the equations in Section 3.1 and 3.2, but replacing the original parameter values $\theta^{(nm)}$ with their corresponding rank normalized values (normal scores) 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, transform ranks to normal scores using the inverse normal transformation and a fractional offset (Blom, 1958):

$$z^{(nm)} = \Phi^{-1}\left(\frac{r^{(nm)} - 3/8}{S - 1/4}\right). \tag{14}$$

Using normalized ranks (normal scores) $z^{(nm)}$ instead of ranks $r^{(nm)}$ themselves has the benefits that (1) for continuous variables the normality assumptions in computation of $\widehat{R}$ and S_{eff} are fulfilled (via the transformation), (2) the values of $\widehat{R}$ and S_{eff} are practically the same as before for nearly normally distributed variables (the interpretation doesn't change for the cases where the original $\widehat{R}$ worked well), and (3) rank-normalized $\widehat{R}$ and S_{eff} are invariant to monotone transformations (e.g. we get the same diagnostic values when examining a variable or logarithm of a variable). 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_{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 A). 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 reveals problems with variance and tail exploration

Both original and rank normalized split-$\widehat{R}$ 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-$\widehat{R}$ statistic not only for the original draws $\theta^{(nm)}$, but also for the corresponding *folded* draws $\zeta^{(mn)}$, absolute deviations from the median,

$$\zeta^{(mn)} = \left|\theta^{(nm)} - \text{median}(\theta)\right|. \tag{15}$$

We call the rank normalized split-$\widehat{R}$ measure computed on the $\zeta^{(mn)}$ values *folded-split-*$\widehat{R}$. This 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.

Figure 1 demonstrates how our new version of $\widehat{R}$ catches some examples of lack of convergence that were not detected by earlier versions of the potential scale reduction factor. We do not intend with this example to claim that our new $\widehat{R}$ is perfect—of course, it can be defeated too. Rather, we use these simple scenarios to develop intuition about problems with traditional split-$\widehat{R}$ and possible directions for improvement.

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

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 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 α-quantile is defined as the parameter value θ_α for which $\Pr(\theta \leq \theta_\alpha) = \alpha$. An estimate $\hat{\theta}_\alpha$ of θ_α can be obtained by finding the α-quantile of the empirical cumulative distribution function (ECDF) of the posterior draws $\theta^{(s)}$.

The cumulative probabilities $\Pr(\theta \leq \theta_\alpha)$ (ECDF) can be written as expectation

$$\Pr(\theta \leq \theta_\alpha) \approx \bar{I}_\alpha = \frac{1}{S} \sum_{s=1}^{S} I(\theta^{(s)} \leq \theta_\alpha), \tag{16}$$

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 θ_α can now be obtained by applying rank-normalizing and subsequent computations directly on the indicator function's results. More details on the variance of the cumulative distribution function can be found in the online appendix. Raftery and Lewis (1992) proposed to focus on accuracy of cumulative or interval probabilities and also proposed a specific effective sample size estimate for these probability estimates.

Although the quantiles cannot be written directly as an expectation, the quantile estimate is strongly consistent and Doss et al. (2014) provide conditions for a quantile central limit theorem. Assuming that the CDF is a continuous function F which is smooth near an α-quantile of interest, we could compute

$$\mathrm{Var}(\hat{\theta}_\alpha) = \mathrm{Var}(F^{-1}(\bar{I}_\alpha)) = \mathrm{Var}(\bar{I}_\alpha)/f(\theta_\alpha). \tag{17}$$

Even if we do not usually know F, this shows that the variance of θ_α is just the variance of $\bar{I}_\alpha$ scaled by the unknown density $f(\theta_\alpha)$, and thus the effective sample size for the quantile estimate $\hat{\theta}_\alpha$ is the same as for the corresponding cumulative probability.

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_{eff}). Tail-ESS can help diagnosing problems due to different scales of the chains (see Appendix A).

Since the marginal posterior distributions might not have finite mean and variance, for example, the popular `rstanarm` package (Stan Development Team, 2018a) reports 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 same as 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 ζ (see (15)):

$$\Pr(\zeta \leq \zeta_{0.5}) \approx \bar{I}_{\zeta,0.5} = \frac{1}{S} \sum_{s=1}^{S} I(\zeta^{(s)} \leq \zeta_{0.5}), \tag{18}$$

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

$$\bar{I}_{\alpha,\delta} = \Pr(\hat{Q}_\alpha < \theta \leq \hat{Q}_{\alpha+\delta}), \tag{19}$$

where $\hat{Q}_\alpha$ is an empirical α-quantile, $\delta = 1/k$ is the length of the interval for some positive integer k, and $\alpha \in (0, \delta, \ldots, 1-\delta)$ changes in steps of δ. 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 is more localized than efficiency measure for quantiles and can be used to build intuition about what types of posterior functionals can be computed as illustrated in the examples. While the expectation of a function that only depends on intermediate values can be usually estimated with relative ease, expectations of tail probabilities or other posterior functionals that depend critically on the tail of the distribution will be usually more difficult to estimate. In addition, small probability intervals can be used in practical equivalence testing (see, e.g., Wellek, 2010).

A natural multivariate extension of small intervals would be to consider small probability volumes using a box or sphere with dimensions determined, for example, by marginal quantiles. The visualization of the multivariate results would

be easiest in 2 or 3 dimensions. In higher dimensions, for example, k-means clustering could be used to determine hyper-spheres. Even if it gets more difficult to visualize where the problematic region in the high dimensional space is, the diagnosing that sampling efficiency is low in some parts of the posterior can be useful.

4.4 Monte Carlo error estimates for quantiles

To obtain the MCSE for $\hat{\theta}_\alpha$, Doss et al. (2014) use a Gaussian kernel density estimate of $f(\theta_\alpha)$ and batch means and subsampling bootstrap method for estimating $\mathrm{Var}(\bar{I}_\alpha)$, and Liu et al. (2016) use a flat top kernel density estimate for $f(\theta_\alpha)$ and a spectral variance approach for $\mathrm{Var}(\bar{I}_\alpha)$.

We propose an alternative approach which avoids the need to estimate $f(\theta_\alpha)$. For example, here is how we estimate a central 90% Monte Carlo error interval for $\hat{\theta}_\alpha$:

1. Compute the effective sample size S_{eff} for $\mathrm{Var}(\bar{I}_\alpha)$.
2. Compute sigma points (Wan and Van Der Merwe, 2000) a and b as 5% and 95% quantiles of $\mathrm{Beta}(\beta_1, \beta_2)$ with shape parameters
$$\beta_1 = S_{\text{eff}}\bar{I}_\alpha + 1 \quad \text{and} \quad \beta_2 = S_{\text{eff}}(1 - \bar{I}_\alpha) + 1. \tag{20}$$
Using S_{eff} here takes into account the efficiency of the posterior draws. The variance of this beta distribution matches the variance of normal approximation, but using quantiles as sigma points instead of variance-based sigma points guarantees that $0 < a < 1$ and $0 < b < 1$. Asymptotically as $S_{\text{eff}} \to \infty$, this beta distribution converges towards a normal distribution.
3. Propagate the sigma points through the nonlinear inverse transforms $A = (F^{-1}(a))$ and $B = (F^{-1}(b))$. As we don't know F for the quantity of interest, we use a simple numerical approximation:
$$\widehat{A} = \theta^{(s')} \text{ where } s' \leq Sa < s' + 1$$
$$\widehat{B} = \theta^{(s'')} \text{ where } s'' - 1 < Sb \leq s''.$$
$\widehat{A}$ and $\widehat{B}$ are then estimated 5% and 95% quantiles of the Monte Carlo error interval for $\hat{\theta}_\alpha$.

The Monte Carlo standard error for $\hat{\theta}_\alpha$ can be approximated using the sigma point method (Wan and Van Der Merwe, 2000), for example, by computing $(\widehat{B} - \widehat{A})/2$, where $\widehat{A}$ and $\widehat{B}$ are estimated 16% and 84% Monte Carlo error quantiles computed with the above algorithm.

The above algorithm is useful as a default, as it is more robust than density estimation based approaches for non-smooth densities, which is common case, for example, when variables are constrained in a (semi-open) range. $\widehat{A}$ and $\widehat{B}$ are likely to have high variance in case of extreme tail quantiles and thick-tailed distributions, as there are not many $\theta^{(s)}$ in extreme tails. The approaches using a density estimate for $f(\theta_\alpha)$ can provide better accuracy when the assumptions of the density estimate are fulfilled, but they can have a high bias if the density is not smooth or the shape of the kernel doesn't match well the tail properties of the distribution. To improve accuracy of extreme tail quantile estimates, common extreme value models could be used to model the tail of the distribution.

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, such as Figure 6, 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 when used with 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 reveal convergence problems. As a result, some issues may only be detectable as S increases, if ESS grows sublinearly or even decreases 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, some convergence problems only occur at relatively high S and may thus not be detectable if the total number of draws is too small.

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 $\widehat{R}$ is based on calculating within and between chain variances. If the marginal distribution of a quantity of interest 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 standard Cauchy distributions for each element of a 50-dimensional vector x:

$$x_j \sim \text{Cauchy}(0, 1) \quad \text{for} \quad j = 1, \ldots, 50. \tag{21}$$

We monitor the convergence for each of the x_j separately. As the distribution of x has thick tails, we may expect any generic MCMC algorithm to have mixing problems. Several values of $\widehat{R}$ greater than 1.01 and some effective sample sizes less than 400 also indicate convergence problems (in addition a HMC-specific diagnostic, "iterations exceed maximum tree depth" (Stan Development Team, 2018b) also indicated slow mixing of the chains). 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 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 $\widehat{R}$. If the effective sample size is highly unstable, does not increase

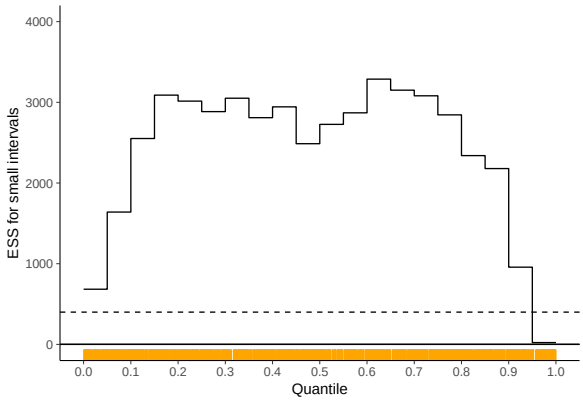

Figure 3: Local efficiency of small-interval probability estimates for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400. Orange ticks show the position of iterations that exceeded the maximum tree depth in the dynamic HMC algorithm.

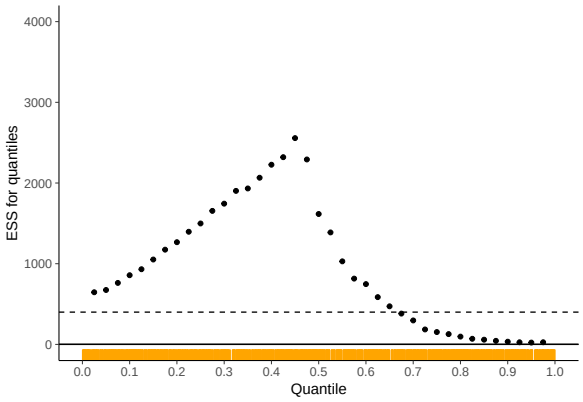

Figure 4: Efficiency of quantile estimates for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400. Orange ticks show the position of iterations that exceeded the maximum tree depth in the dynamic HMC algorithm.

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 case of good mixing all rank plots should be close to uniform. More experiments can be found in Appendix B 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:

$$a_j \sim \text{Normal}(0,1), \qquad b_j \sim \text{Gamma}(0.5,0.5), \qquad x_j = a_j/\sqrt{b_j}. \tag{22}$$

The model has two parameters which have thin-tailed distributions so that we may assume good mixing of Markov chains. Cauchy-distributed x can be computed deterministically from a and b. 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.

For all parameters, $\widehat{R}$ 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. The appearances of the plots in Figures 7, 8, and 9 are what we would expect for well mixing chains in general.

In contrast, trace plots may be much less clear in certain situations. To illustrate this point, we show trace plots of the Cauchy model in the nominal and alternative parameterizations side by side in Figure 10. Recall that the computation converged well in the alternative parameterization but not in the nominal parameterization.

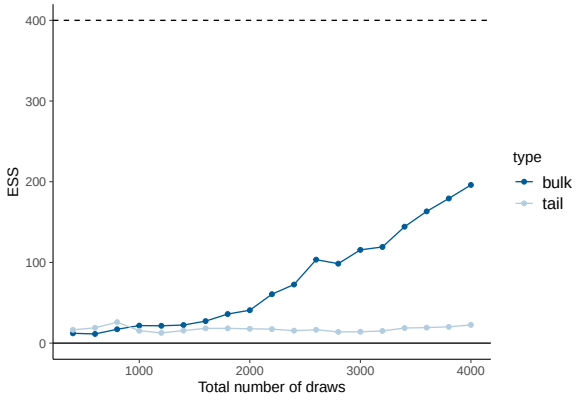

Figure 5: Estimated effective sample sizes with increasing number of iterations for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

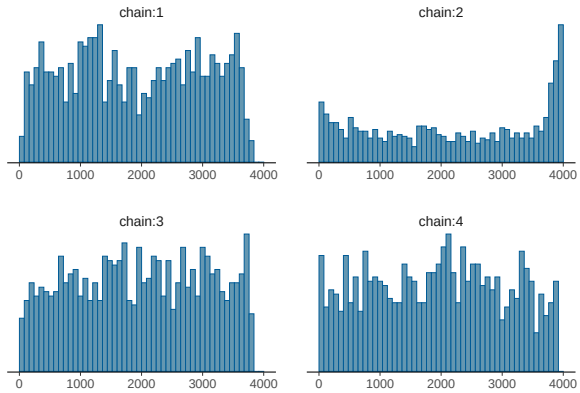

Figure 6: Rank plots of posterior draws from four chains for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS.

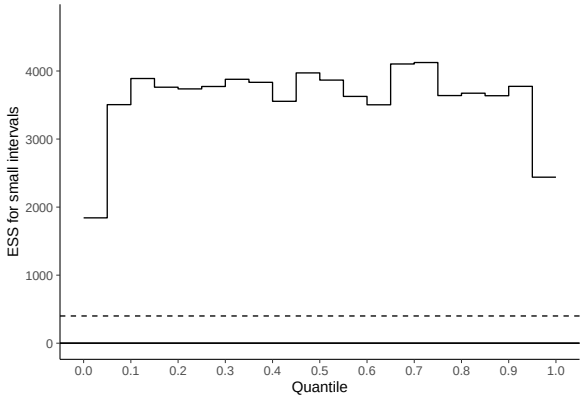

Figure 7: Local efficiency of small-interval probability estimates for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

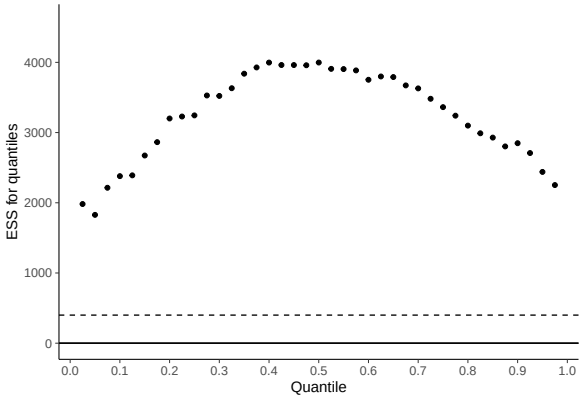

Figure 8: Efficiency of quantile estimates for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

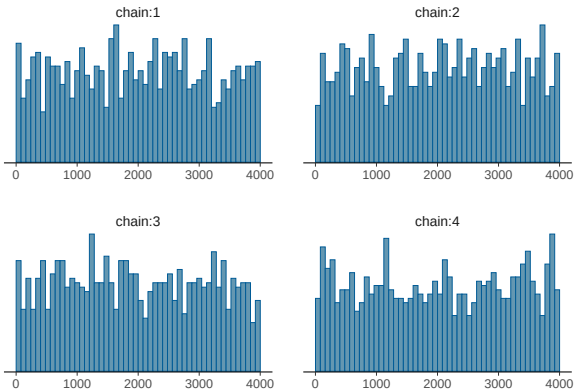

Figure 9: Rank plots of posterior draws from four chains for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS.

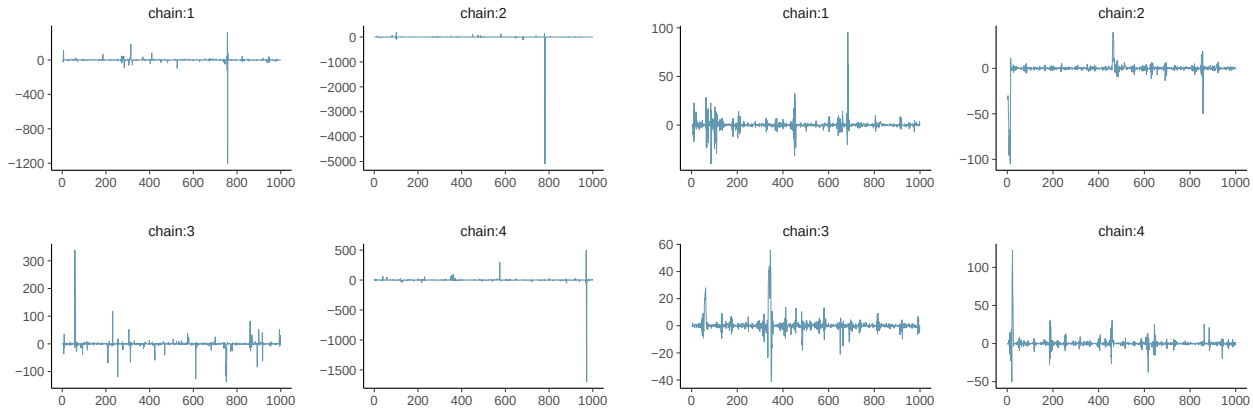

Figure 10: Trace plots of posterior draws from four chains for the Cauchy model with nominal and alternative parameterization. We do not tell which plot belongs to which model and let the reader decide themselves how easy it is to see differences in convergence from those trace plots. Results are displayed for the element of x with the smallest tail-ESS in the respective model.

Half-Cauchy distribution with nominal parameterization

Half-Cauchy priors for non-negative parameters are common and often 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. Probabilistic programming frameworks usually implement positivity constraint by sampling in the unconstrained $\log(x)$ space, which changes the geometry crucially. With this transformation, all values of $\widehat{R}$ 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.

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 typical problems in inference for hierarchical models. We can parameterize this simple model in at least two ways. The centered parameterization $(\theta, \mu, \tau, \sigma)$ is,

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

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, θ_j are the treatment effects in the eight schools, and μ, τ represent the population mean and standard deviation of the distribution of these effects. In the centered parameterization, the θ are parameters, whereas in the non-centered parameterization, the $\tilde{\theta}$ are parameters and θ 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 τ, making it difficult for many simple Markov chain methods to adequately explore the full distribution of this parameter. In the following, we will focus on analyzing convergence of τ. The online appendix contains more detailed analysis of different algorithm variants and results of longer chains.

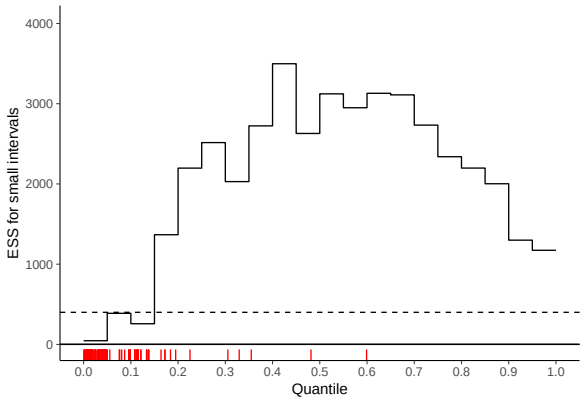

Figure 11: Local efficiency of small-interval probability estimates of τ for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400. Red ticks show the position of divergent transitions.

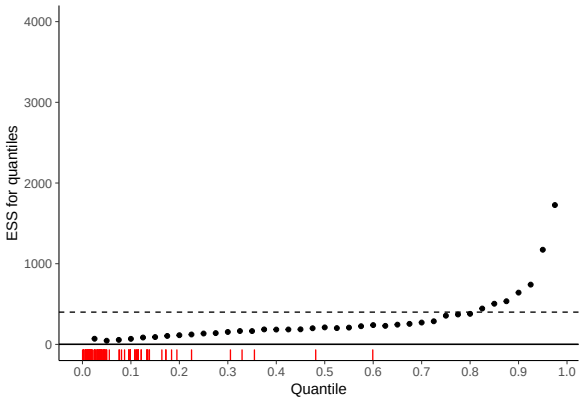

Figure 12: Efficiency of quantile estimates of τ for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400. Red ticks show the position of divergent transitions.

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, which bias the obtained estimates if they occur; for details see Stan Development Team (2018b). Still, we observe a lot of divergent transitions, which in itself is already a sufficient indicator of convergence problems. We can also use $\widehat{R}$ 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 τ are 67 and 82, respectively. Both are much less than 400, indicating we should investigate that parameter more carefully. Figures 11 and 12 show the sampling efficiency for the small-interval probability and quantile estimates. The sampler has difficulties in exploring small τ values. As the sampling efficiency for small τ 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 τ values is reflected in the sampling efficiency for all quantiles. Red ticks, which show the position of iterations with divergences, have concentrated to small τ values, which gives us another indication of problems in exploring small values.

Figure 13 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 τ displayed in Figure 14 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 Appendices C and D as well as in the online appendix.

Non-centered eight schools model

For hierarchical models, the corresponding non-centered parameterization often works better (Betancourt and Girolami, 2019). 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 $\widehat{R}$ are less than 1.01 and ESS exceeds 400, indicating a much better efficiency of the non-centered parameterization. Figures 15 and 16 show the efficiency of small-interval probability estimates and the efficiency of quantile estimates for τ. Small τ values are still more difficult to explore, but the relative efficiency is good. The rank plots of τ Figure 17 show no substantial differences between chains.

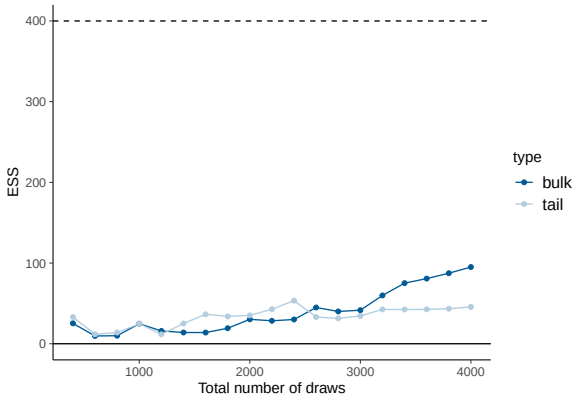

Figure 13: Estimated effective sample sizes of τ with increasing number of iterations for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400.

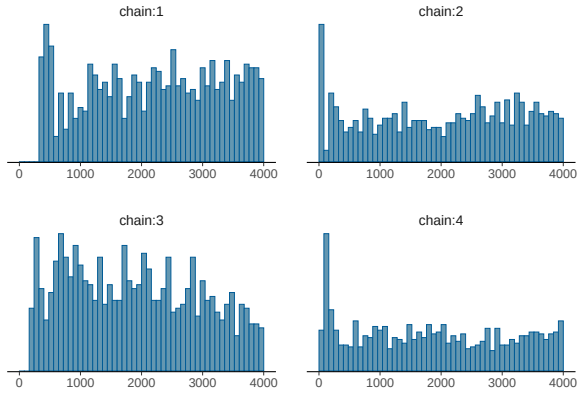

Figure 14: Rank plots of posterior draws of τ from four chains for the eight schools model with centered parameterization.

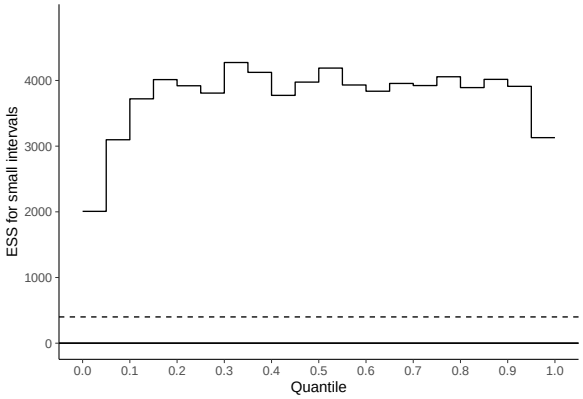

Figure 15: Local efficiency of small-interval probability estimates of τ for the eight schools model with the non-centered parameterization. The dashed line shows the recommended threshold of 400.

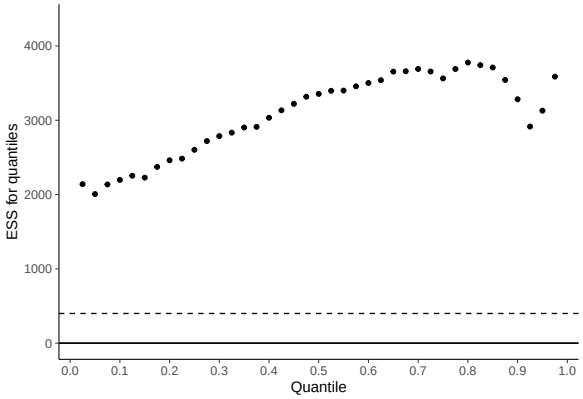

Figure 16: Efficiency of quantile estimates of τ for the eight schools model with the non-centered parameterization. The dashed line shows the recommended threshold of 400.

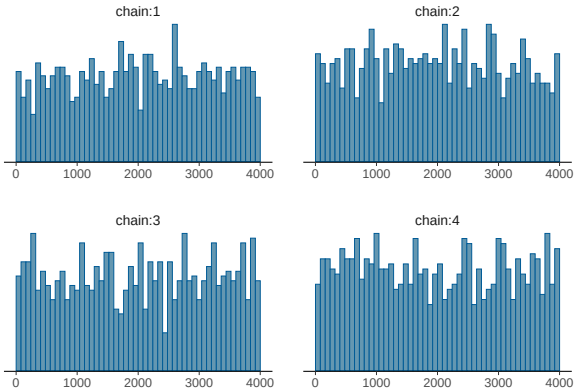

Figure 17: Rank plots of posterior draws of τ from four chains for the eight schools model with non-centered parameterization.

References

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

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, pages 79–101. Chapman and Hall/CRC, 2019.

Gunnar Blom. *Statistical Estimates and Transformed Beta-Variables*. Wiley; New York, 1958.

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.

Herman Chernoff and I. Richard Savage. Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics*, 29(4):972–994, 1958.

Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

Perry de Valpine, Daniel Turek, Christopher J. Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

Charles R. Doss, James M. Flegal, Galin L. Jones, and Ronald C. Neath. Markov chain Monte Carlo estimation of quantiles. *Electronic Journal of Statistics*, 8(2):2448–2478, 2014.

Ronald A. Fisher and Frank Yates. *Statistical Tables for Biological, Agricultural, and Medical Research*. Oliver & Boyd; Edinburgh, 1938.

James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Annals of Statistics*, 38(2):1034–1070, 2010.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

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, and Donald R. Rubin. *Bayesian Data Analysis, second edition*. Chapman & Hall, 2003.

Andrew Gelman, Zaiying Huang, David van Dyk, and W. John Boscardin. Using redundant parameters to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17:95–122, 2008.

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.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

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.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased Markov chain Monte Carlo with couplings. *arXiv preprint arXiv:1708.03625*, 2017.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

John A. Laurmann and W. Lawrence Gates. Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *Journal of the Atmospheric Sciences*, 34(8):1187–1199, 1977.

Jia Liu, Daniel J. Nordman, and William Q. Meeker. The number of MCMC draws needed to compute Bayesian credible bounds. *The American Statistician*, 70(3):275–284, 2016.

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.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

Kerrie L. Mengersen, Christian P. Robert, and Chantal Guihenneuc-Jouyaux. MCMC convergence diagnostics: A review. In Jose M. Bernardo, James O. Berger, and A. P. Dawid, editors, *Bayesian Statistics 6*, pages 415–440. Oxford University Press, 1999.

Radford M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

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/.

Adrian E. Raftery and Steven M. Lewis. How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

D. A. Sorensen, S. Andersen, D. Gianola, and I. Korsgaard. Bayesian inference in threshold models using Gibbs sampling. *Genetics Selection Evolution*, 27(3):229, 1995.

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 preprint arXiv:1812.09384*, 2018.

Eric A Wan and Rudolph Van Der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158. IEEE, 2000.

Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2010.

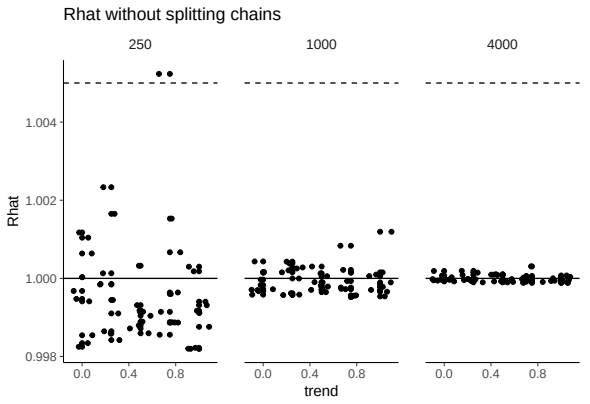

Figure 18: $\widehat{R}$ without splitting for varying chain lengths for chains which have the same linear trend and a similar marginal distribution.

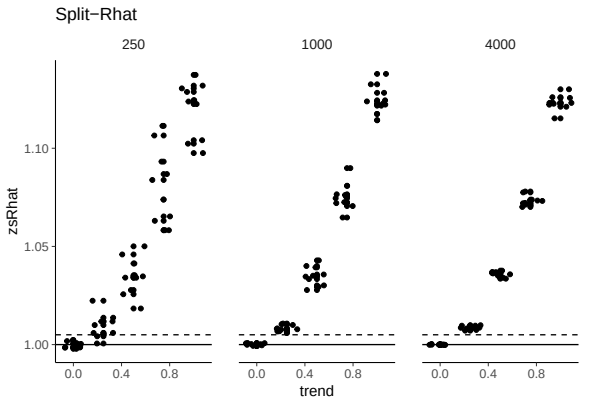

Figure 19: Split-$\widehat{R}$ for varying chain lengths for chains which have the same linear trend and a similar marginal distribution.

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.

The code for these simulations can be found in the online appendix.

All chains have the same trend. First, we draw all the chains from the same $\mathrm{Normal}(0, 1)$ distribution plus a linear trend (i.e., $\theta^{(s)} = e^{(s)} + cs$ where $e^{(s)}$ is $\mathrm{Normal}(0, 1)$ distributed, s is the iteration indicator, and c is the strength of the linear trend). Figure 18 shows that if we don't split chains, $\widehat{R}$ misses the trends if all chains still have a similar marginal distribution. Figure 19 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 20 shows the relative bulk-ESS divided by S for easier comparison between different values of S. Split-$\widehat{R}$ 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 from the same $\mathrm{Normal}(0, 1)$ distribution, except one that is sampled with nonzero mean. Figure 21 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 22 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 low relative efficiency when the total number of draws increases. Rank plots are practical way to visualize differences between chains. Figure 23 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.

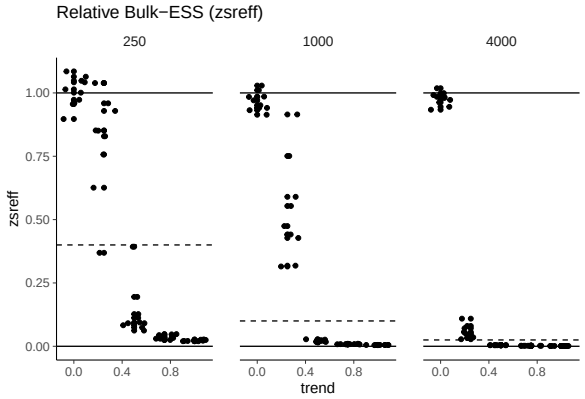

Figure 20: Relative bulk-ESS for varying chain lengths for chains which have the same trend and a similar marginal distribution. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

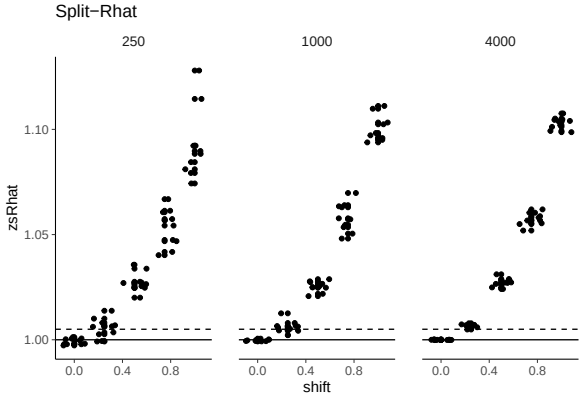

Figure 21: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different mean than the others.

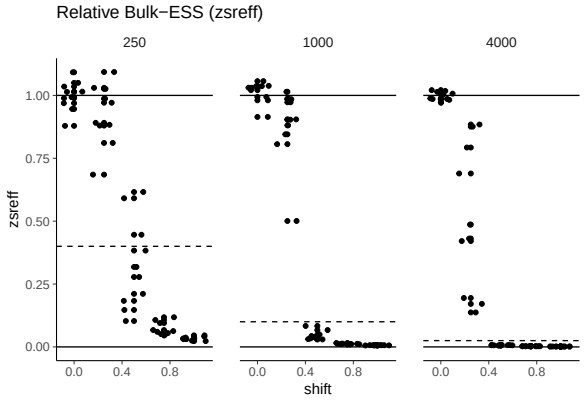

Figure 22: Relative bulk-ESS for varying chain lengths for chains with one sampled with a different mean than the others. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

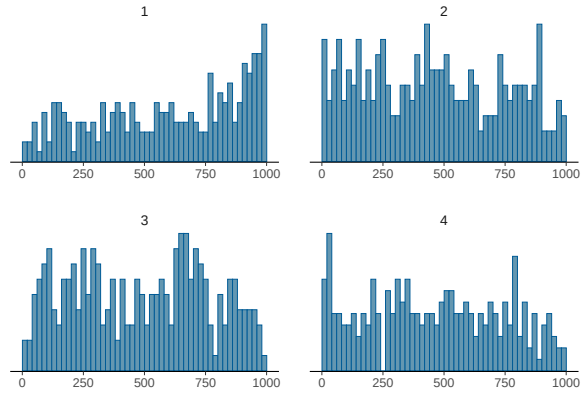

Figure 23: Rank plots of posterior draws from four chains with one sampled with a different mean than the others.

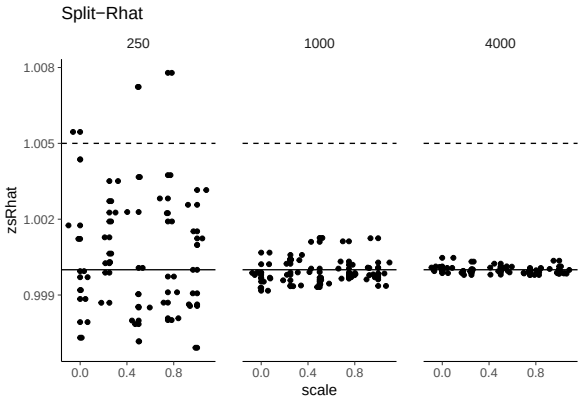

Figure 24: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

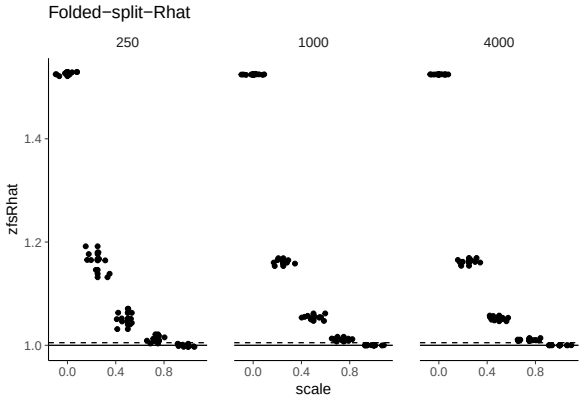

Figure 25: Folded-split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

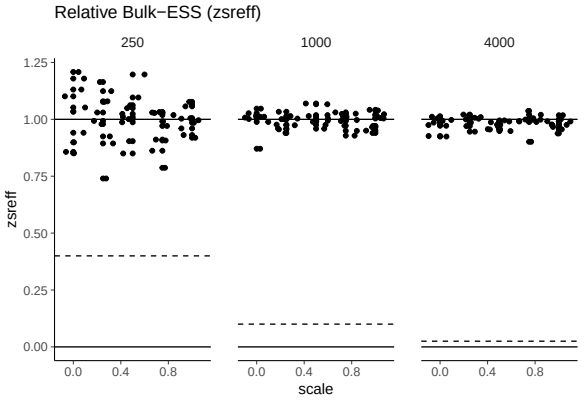

Figure 26: Relative bulk-ESS for varying chain lengths for chains with one sampled with a different variance than the others.

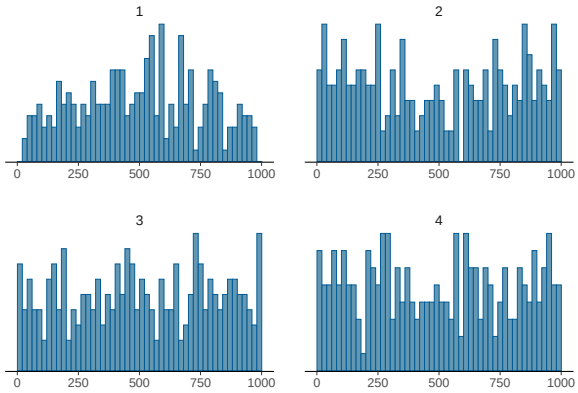

Figure 27: Rank plots of posterior draws from four chains with one sampled with a different variance than the others.

Scaling one chain. For our third simulation, all the chains are from the same $\mathrm{Normal}(0, 1)$ distribution, except one of the chains is sampled with variance less than 1. Figure 24 shows that split-$\widehat{R}$ is not able to detect scale differences between chains. Figure 25 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 26 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 27 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-$\widehat{R} = 1.06$, the rank plots clearly show that the first chain behaves differently.

Appendix B: 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 28 shows that trace plots for the first parameter look wild with occasional large values, and it is difficult to

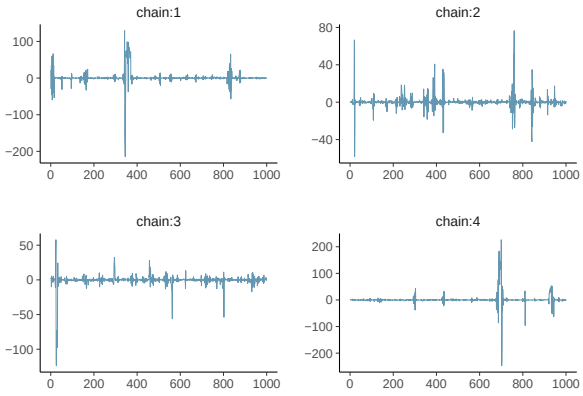

Figure 28: Trace plots of four chains for Cauchy model with nominal parameterization and `max_treedepth`=20.

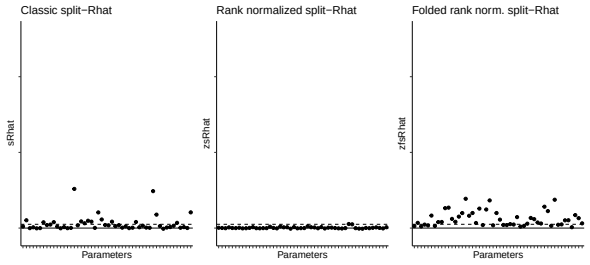

Figure 29: Traditional split-$\widehat{R}$, rank normalized split-$\widehat{R}$, and rank normalized folded-split-$\widehat{R}$ for Cauchy model with nominal parameterization and `max_treedepth`=20.

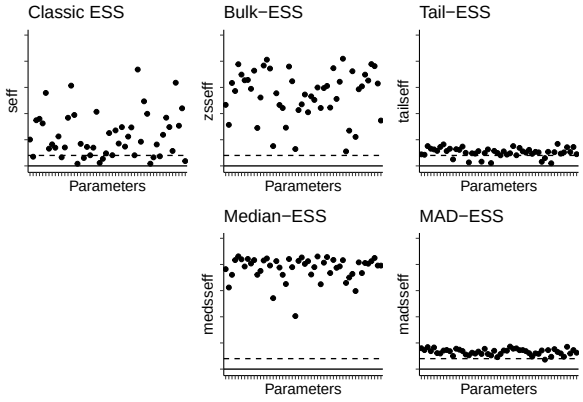

Figure 30: Traditional ESS, bulk-ESS, tail-ESS, median-ESS and MAD-ESS for Cauchy model with nominal parameterization `max_treedepth`=20.

interpret possible convergence. Figure 29 shows traditional split-$\widehat{R}$, rank normalized split-$\widehat{R}$, 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 29 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.

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

Here we demonstrate a limitation of split-$\widehat{R}$ and ESS as convergence diagnostics in a case where the chains seem to converge to a common stationary distribution, but other diagnostics can detect a likely bias.

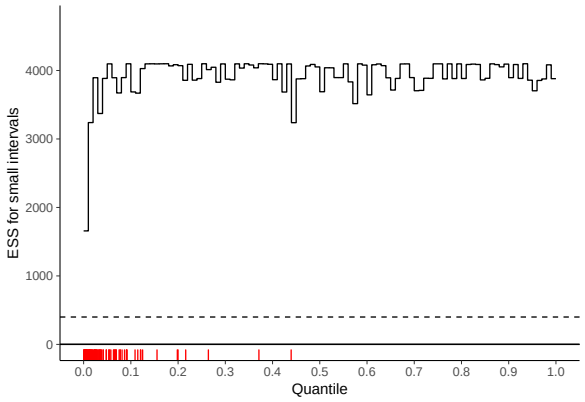

Figure 31: Local efficiency of small-interval probability estimates for eight schools model with centered parameterization, very long chains, and thinning. The dashed line shows the recommended threshold of 400.

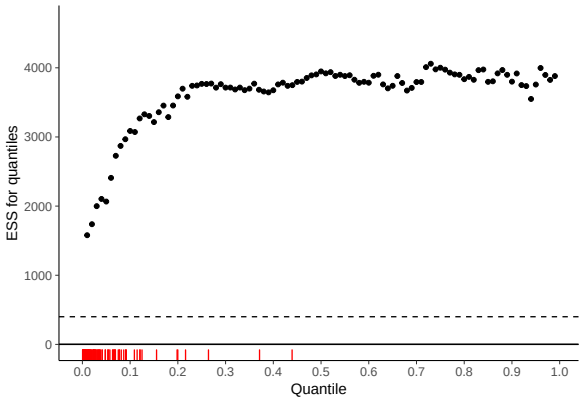

Figure 32: Efficiency of quantile estimates for eight schools model with centered parameterization, very long chains, and thinning. The dashed line shows the recommended threshold of 400.

When autocorrelation time is high, sometimes the chains are "thinned" by saving only a small portion of the draws. In general we don't recommend this approach, as it throws away useful information leading to less efficient estimates and the dependent simulation draws are not a problem for estimating the Monte Carlo error. However, we also sometimes use thinning when autocorrelation time is so high that our usual computers have memory challenges in handling unthinned chains. This example serves as warning that thinning can also throw away information useful for convergence diagnostics. As in Section 5.2 we run HMC for the eight schools model with centered parameterization, but now with 4×10^5 iterations per chain, first half removed as warm-up, and the second half thinned by keeping only every 200th iteration.

We observe several divergent transitions and the estimated Bayesian fraction of missing information (Betancourt, 2017) is also low, which indicate convergence problems. In Section 5.2 we demonstrated that the diagnostics discussed in this paper are also able to detect convergence problems.

Figures 31, 32, and 33 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-$\widehat{R}$ 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 34 are still able to show the problem.

An explanation for the changed behavior after thinning is that we are throwing away information which would make it easier to see "sticking" behavior in autocorrelations. When MCMC struggles to reach some part of the parameter space that has substantial posterior mass, it is not unusual for Markov chains to stick for several iterations (see, e.g. Neal, 2003; Betancourt and Girolami, 2019). When case sticking occurs, we usually can observe high variation in means of chains leading to high $\widehat{R}$ values. In infinite time, all chains would sample from the target distribution. With long but finite chains we can observe a situation where chains start to resemble each other, but all are still producing biased estimates. This is clearly a failure mode for $\widehat{R}$ and the failure seems to be more likely when thinning is discarding useful information about autocorrelations of the original chains. Fortunately, we have diagnostics for HMC that are specifically sensitive to cases where sticking tends to occur.

Appendix 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 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 mix of Gibbs and Metropolis-Hastings

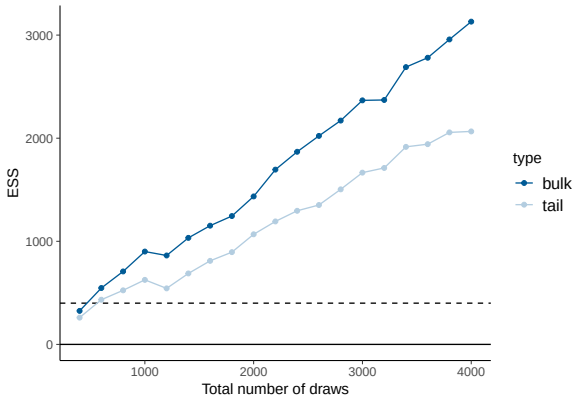

Figure 33: Estimated effective sample sizes with increasing number of iterations for eight schools model with centered parameterization, very long chains, and thinning. The dashed line shows the recommended threshold of 400.

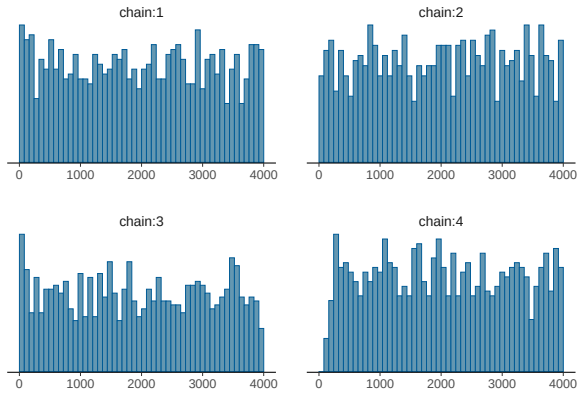

Figure 34: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization, very long chains, and thinning.

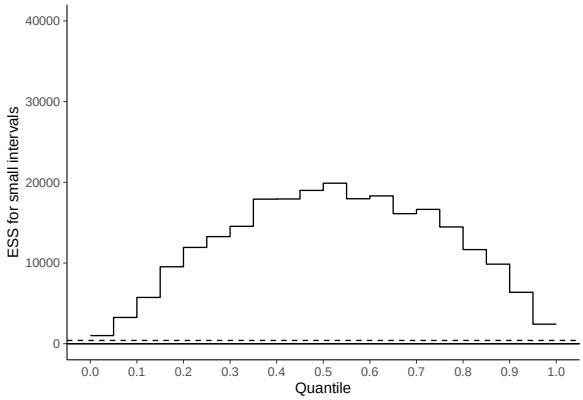

Figure 35: Local efficiency of small-interval probability estimates for the eight schools model with centered parameterization and Gibbs sampling. The dashed line shows the recommended threshold of 400.

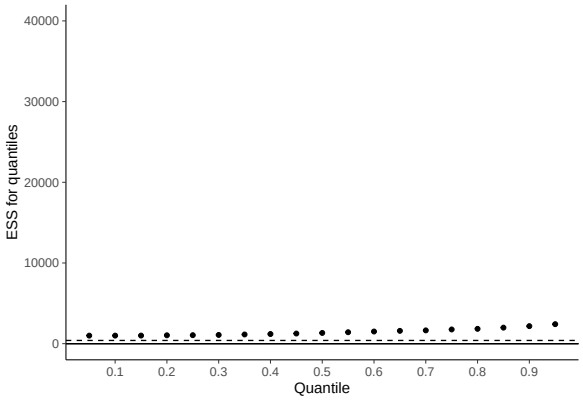

Figure 36: The efficiency of quantile estimates for the eight schools model with centered parameterization and Gibbs sampling. The dashed line shows the recommended threshold of 400.

sampling which often does not scale well to high-dimensional posteriors (see, e.g. Hoffman and Gelman, 2014) but 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 τ, 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 report the results with 10 000 iterations. Examining the diagnostics for τ, now split-$\widehat{R} = 1.01$, bulk-ESS $= 677$, and tail-ESS $= 1027$, which are all good.

Figures 35, 36, and 37 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, the increase in effective sample size for τ 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 (Gelman et al., 2008). 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.

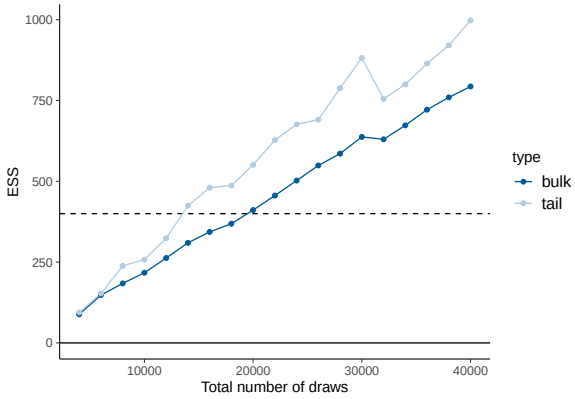

Figure 37: Change in bulk-ESS and tail-ESS with increasing number of iterations for the eight schools model with centered parameterization and Gibbs sampling. The dashed line shows the recommended threshold of 400.