Improving Convergence Diagnostics of Iterative Algorithms

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

Abstract

Abstract

1 Introduction

Iterative simulation, particularly Markov chain Monte Carlo (MCMC), is increasingly popular in statistics (Brooks and Gelman, 1998), especially in Bayesian applications where the goal is to represent posterior inference using a sample of posterior draws. Iterative simulation algorithms in common use typically can be proven to converge to the target distribution as the number of draws approaches infinity, but convergence is only approximate for any finite number of draws.

In practice we have two concerns:

1. The M chains may not have not mixed well, so that the simulations do not represent the target distribution because they still retain the influence of their history.
2. The effective sample size (number of effective simulation draws) is low, possibly much less than the total number of draws across chains, because of dependence (autocorrelation) within each chain.

These two issues are related. It is only possible to have a large number of effective draws if the chains have mixed well. Figure 1 illustrates two ways in which sequences of iterative simulations can fail to mix. 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. This situation may arise due to multimodal posteriors or because one chain is stuck in a region of high curvature with a step size too high to make an acceptable proposal. These two examples make it clear that any method for assessing mixing and effective sample size should use information between and within chains.

The other relevant point is that we are often fitting models with large numbers of parameters, so that it is not realistic to expect to make trace plots such as in Figure 1 for all quantities of interest. We need numerical summaries that can flag potential problems. However, as we will show in this paper, the currently existing and widely applied convergence diagnostics have serious flaws under some conditions. We will thus propose improvements to these diagnostics.

2 Convergence diagnostics for iterative algorithms

The split-$\widehat{R}$ statistic and the *effective sample size* (ESS) are routinely used to monitor the convergence of iterative simulations, which are omnipresent in Bayesian statistics in the form of Markov-Chain Monte-Carlo samples. The original $\widehat{R}$ statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) and *split-*$\widehat{R}$ (Gelman et al., 2013) are both based on the ratio of between and within-chain marginal variances of the simulations, while the latter is computed from split chains (hence the name).

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

Below, we present the computation of *split-*$\widehat{R}$ following Gelman et al. (2013), but using the notation style of Stan Development Team (2018c). These implementations represent the current de facto standard of

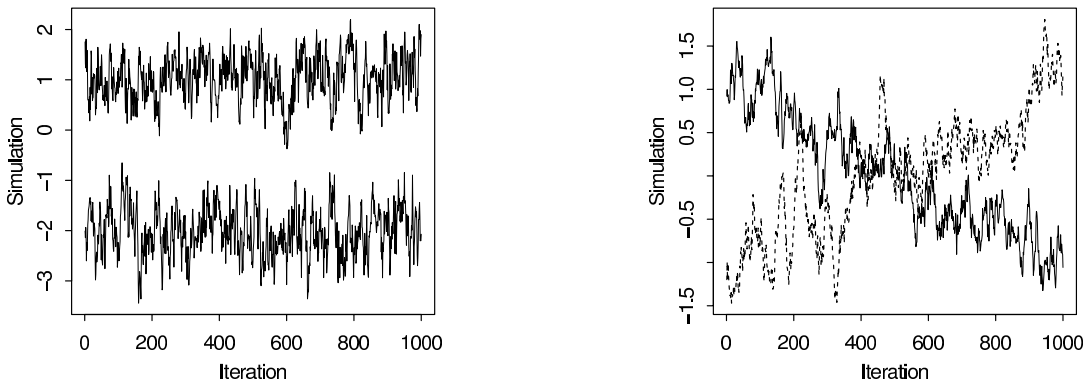

Figure 1: *Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the two sequences happen to cover a common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

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:

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

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

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

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

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* $\text{var}(\theta \mid y)$ 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 $\text{var}(\theta \mid y)$.

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 potential scale reduction is estimated as

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

which declines to 1 as $N \to \infty$. We call this *split-*$\widehat{R}$ because we are applying it to chains that have been split in half so that M is twice the number of actual chains. Without splitting, $\widehat{R}$ would get fooled by non-stationary chains (see section/appendix).

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

2.2 Effective sample size

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

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

The effective sample size of a chain is defined in terms of the autocorrelations within the chain at different lags. The autocorrelation ρ_t at lag $t \geq 0$ for a chain with joint probability function $p(\theta)$ with mean μ and variance σ^2 is defined to be

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

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

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

The effective sample size of one chain generated by a process with autocorrelations ρ_t is defined by

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

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

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

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

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

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

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

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

where

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

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

The effective sample size S_{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., the posterior is multimodal and the chains are stuck in different modes), then our S_{eff} is close to the number of chains.

2.3 Problems of current diagnostics

Split-$\widehat{R}$, and S_{eff} are well defined only if the marginal posteriors have finite mean and variance, which is not always the case. Split-$\widehat{R}$, and S_{eff} can also be unstable even if the mean and variance are finite, if the marginal distribution has thick tails. Usually split-$\widehat{R}$, and S_{eff} are computed only for the posterior mean, which can miss convergence and sampling efficiency problems in tails which affect, for example, the posterior interval estimates.

3 Improving convergence diagnostics

In this section, we discuss several measures that, together, can solve the problems of the current divergence diagnostics we identified above.

3.1 Rank normalization

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.

Rank normalized *split-*$\widehat{R}$ and S_{eff} are computed using the equations in Section 2, but replacing the original parameter values $\theta^{(nm)}$ with their corresponding rank normalized values denoted as $z^{(nm)}$. Rank normalization is done a follows: First, replace each value $\theta^{(nm)}$ by its rank $r^{(nm)}$. Average rank for ties are used to conserve

the number of unique values of discrete quantities. Ranks are computed jointly for all draws from all chains. Second, normalize ranks via the inverse normal transformation

$$z^{(nm)} = \phi^{-1}((r^{(nm)} - 1/2)/S). \tag{11}$$

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

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

3.2 Diagnostics for folded draws

Both original and rank-normalized *split-*$\widehat{R}$ can be fooled if the chains have different scales but the same location as shown in (see Appendix). To alleviate this problem, we propose to compute a rank normalized *split-*$\widehat{R}$ statistic not only for the original draws $\theta^{(nm)}$, but also for the corresponding folded draws $\zeta^{(mn)}$, that is the absolute deviations from the median

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

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

3.3 Convergence diagnostics for quantiles

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

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

For any θ_α, the ECDF gives an estimate of the cumulative probability

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

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

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

To get a better sense of the efficiency of the chains 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 section/appendix).

3.4 Efficiency estimates for the median absolute deviation

Since the marginal posterior distributions might not have finite mean and variance, by default RStan (Stan Development Team, 2018a) and RStanARM (Stan Development Team, 2018b) report median and median absolute deviation (MAD) instead of mean and standard error (SE). Median and MAD are well defined even when the marginal distribution does not have finite mean and variance. Since the median is just 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 (MAD) by computing the efficiency estimate of an indicator function based on the folded parameter values ζ (see Equation (12)):

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

where $\zeta_{0.5}$ is the median of the folded values. We see that 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 also used in the computation of the tail-ESS.

3.5 Efficiency estimates for small interval probability estimates

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

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

where $\hat{Q}_\alpha$ is an empirical α-quantile, $\delta = 1/k$ is the length of the interval with some positive integer k, and $\alpha \in (0, \delta, \ldots, 1-\delta)$ changes in steps of 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 does not depend on the shape of the distribution.

3.6 Monte Carlo error estimates for quantiles

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

1. Compute quantiles of the Beta distribution with shape parameters

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

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

2. Find indices in $s \in \{1, \ldots, S\}$ closest to the ranks of these quantiles. For example, for quantile Q, find $s = \text{round}(\text{Q} \times \text{S})$.
3. Use the corresponding $\theta^{(s)}$ from the list of sorted posterior draws as quantiles from the error distribution. These quantiles can be used to approximate the Monte Carlo standard error.

3.7 Interpreting *Split-*$\widehat{R}$, ESS, and MCSE

The ultimate focus should be in the accuracy of the estimate for the quantity of interest. This accuracy can be measure using Monte Carlo standard error (or corresponding interval) and the acceptable MCSE depends on the quantity and application. MCSE estimate is computed using the marginal posterior of the quantity and adjusted using ESS, and ESS is based on *split-*$\widehat{R}$ and autocorrelations of the chains.

In cases of easy sampling, we could compute ESS only based on the autocorrelations, but *split-*$\widehat{R}$ is very helpful in case of multimodality or other reasons causing chains to get stuck in part of the target distribution (e.g. in case of unimodal funnel and HMC, differences in step size adaptation can lead to chains to have different behavior reaching the narrow part of the funnel). In case of well separated multimodality the *split-*$\widehat{R}$ adjustment of ESS, makes the ESS estimate to be close to the number of distinct modes. If only one chain would be run, and ESS would be computed only based on autocorrelations, the ESS would be highly over-estimated. The variance of *split-*$\widehat{R}$ adjusted ESS could be improved by running more chains. We recommend running at least four chains. The *split-*$\widehat{R}$ is not alone directly determining the MCSE, as it only inflates ESS estimate which is part of MCSE computation. As a generic ad hoc rule based on the simulations we recommend to aim for *split-*$\widehat{R} < 1.01$.

The required ESS can be ultimately decided only in the context of MCSE of the quantity of interest. However, there are two reasons to look at the ESS before MCSEs. Firstly, the computation of *split-*$\widehat{R}$ and autocorrelations needed for ESS itself require estimating means and variances, and in order to get reliable convergence and ESS estimate, we recommend to aim for ESS$>$ 400. In case of running four chains, this corresponds to having at least effective sample size of 50 per each split chain to be used for estimating means, variances and autocorrelations. Often for a useful estimation accuracy ESS$>$ 100 would be sufficient, but that would require that we know that ESS$>$ 100, and when we don't know that we need higher ESS to be certain about the sampling efficiency. Secondly, effective sample sizes for different parameters are on the same scale, and thus it is easier to see which part of the model might have sampling problems.

After checking that *split-*$\widehat{R}$ and ESS fulfill the above ad hoc requirements, we recommend to take int account the application specific requirements for the accuracy of the quantity of interest and check that MCSE is low enough or continue sampling.

3.8 Diagnostic visualizations

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

Rank plots

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

Quantile and small interval plots

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

Efficiency per iteration plots

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

4 Examples

In this section, we will go through some examples to demonstrate the usefulness of our proposed methods as well as the associated workflow in determining convergence. The online appendix contains all model details, code to reproduce the results and more detailed analysis of different algorithm variants and further examples[1].

We use either dynamic Hamiltonian Monte Carlo with multinomial sampling (Betancourt, 2017) as implemented in Stan (Stan Development Team, 2018d) or Gibbs sampling as implemented in JAGS.

4.1 Cauchy: A distribution with infinite mean and variance

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

4.1.1 Nominal parameterization of Cauchy

The nominal Cauchy model with direct parameterization is

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

We set independent Cauchy distribution for 50 dimensional vector x. We use dynamic HMC and run 4 chains each with 1000 iterations of warmup and 1000 iterations stored. Dynamic HMC specific diagnostics treedepth exceedences and Bayesian fraction of missing information indicate slow mixing of the chains.

Several *split-*$\widehat{R} > 1.01$ and some ESS< 400 indicate also poor mixing. The online appendix has more results with longer chains and also with classic *split-*$\widehat{R}$. We can further analyze potential problems using local efficiency and rank plots. We specifically investigate x_{36}, which in this specific run has the smallest tail-ESS of 34. Figure 2 shows the local efficiency of small interval probability estimates (see Section Efficiency estimate for small interval probability estimates). The efficiency of sampling is worryingly low in the tails (which is caused by slow mixing in long tails of Cauchy). Figure 3 shows the efficiency of quantile estimates (see Section Efficiency for quantiles). Similar as above, the sampling efficiency is worryingly low in the tails.

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

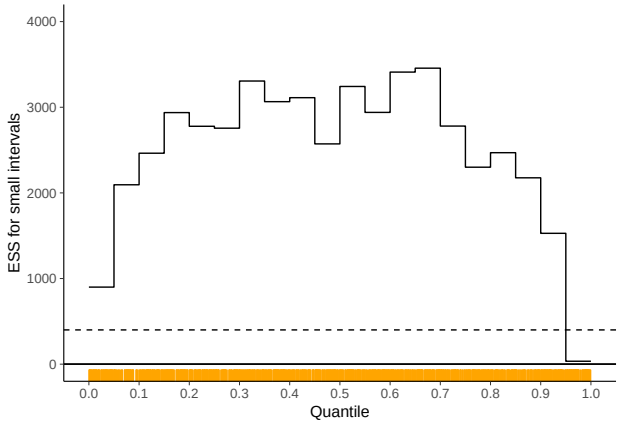

Figure 2: The local efficiency of small interval probability estimates for Cauchy model with nominal parameterization. Orange ticks show iterations that exceeded the maximum treedepth in dynamic HMC algorithm.

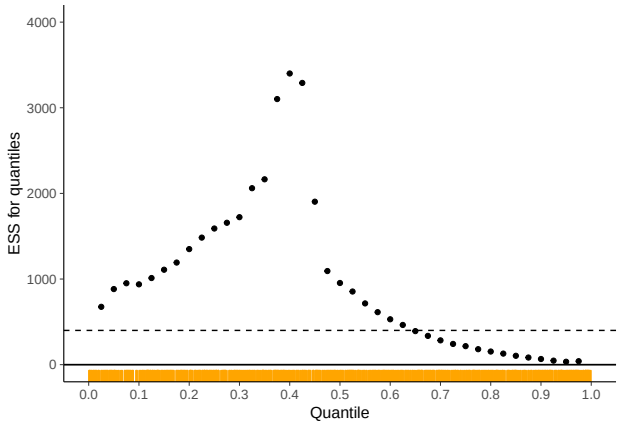

Figure 3: The efficiency of quantile estimates for Cauchy model with nominal parameterization.

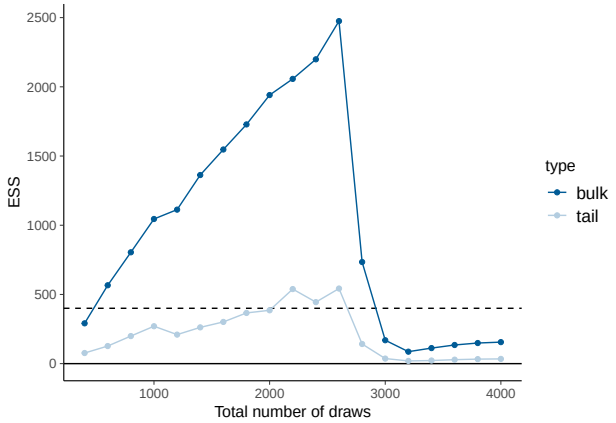

Figure 4: The estimated effective sample sizes with increasing number of iterations for Cauchy model with nominal parameterization.

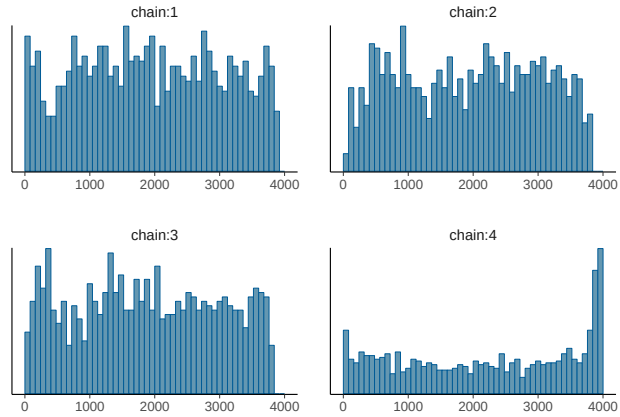

Figure 5: Rank plots of posterior draws from four chains for Cauchy model with nominal parameterization.

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

4.1.2 Alternative parameterization of Cauchy

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

$$a \sim \mathrm{N}(0,1), \qquad b \sim \mathrm{Gamma}\left(\frac{1}{2},\frac{1}{2}\right), \qquad x = \frac{a}{\sqrt{b}}. \tag{18}$$

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

We set define 50 dimensional parameter vectors a and b from which 50 dimensional quantity x is computed. We use dynamic HMC and run 4 chains each with 1000 iterations of warmup and 1000 iterations stored. There are no warnings, and the sampling is much faster.

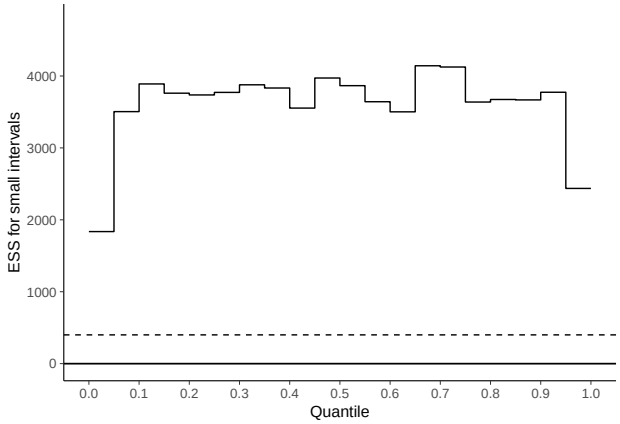

Figure 6: The local efficiency of small interval probability estimates for Cauchy model with alternative parameterization.

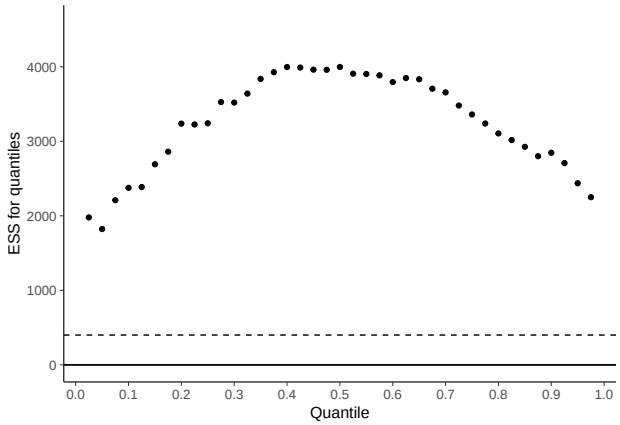

Figure 7: The efficiency of quantile estimates for Cauchy model with alternative parameterization.

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

In summary, the alternative parameterization produces results that look much better than for the nominal parameterization.

4.1.3 Half-Cauchy with nominal parameterization

Half-Cauchy priors are common and, for example, in Stan usually set using the nominal parameterization

$$x \sim \text{Cauchy}^{+}(0, 1). \tag{19}$$

However, when the constraint `<lower=0>` is used, Stan does the sampling automatically in the unconstrained $\log(x)$ space, which changes the geometry crucially.

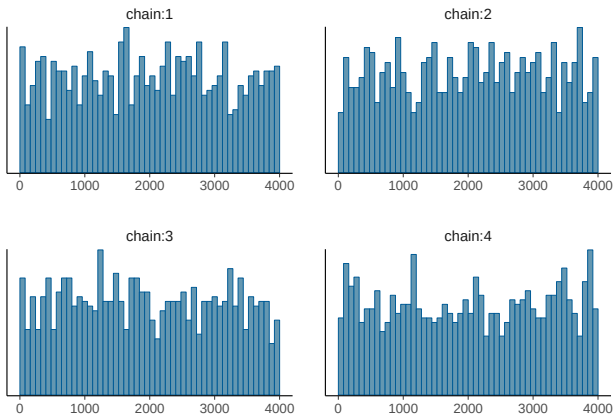

Figure 8: Rank plots of posterior draws from four chains for Cauchy model with alternative parameterization.

We set independent half-Cauchy distribution for 50 dimensional vector x with automatic transformation so that sampling is done in $\log(x)$ space. We use dynamic HMC and run 4 chains each with 1000 iterations of warmup and 1000 iterations stored. There are no warnings, and the sampling is much faster than for the Cauchy nominal model.

All *split*-$\widehat{R} < 1.01$ and ESS> 400 indicate good performance of the sampler. We see that the Stan's automatic (and implicit) transformation of constraint parameters can have a big effect on the sampling performance. More experiments with different parameterizations of the half-Cauchy distribution can be found in the online appendix.

4.2 Hierarchical model: Eight Schools

The Eight Schools data is a classic example for hierarchical models (see Section 5.5 in Gelman et al., 2013), which despite the apparent simplicity nicely illustrates the typical problems in inference for hierarchical models. The centered parameterization exhibits a funnel shape that contracts into a region of strong curvature around small values of population prior scale τ, making it difficult for most Markov chain methods to adequately explore. Online appendix contains more detailed analysis of different algorithm variants including also Gibbs sampling.

4.2.1 A Centered Eight Schools model

We use dynamic HMC and run 4 chains each with 1000 iterations of warmup and 1000 iterations stored. Instead of the default options, we run the centered parameterization model with an increased `adapt_delta` value to reduce the probability of getting divergent transitions. Despite an increased `adapt_delta`, we still observe a lot of divergent transitions, which in itself is already sufficient indicator of convergence problems. We can use *split*-$\widehat{R}$ and ESS diagnostics to recognize problematic parts of the posterior, and they can be used also in cases when other MCMC algorithms than HMC is used.

See online appendix for more details and results of longer chains.

Bulk-ESS and Tail-ESS for the between school standard deviation τ are 67 and 82 respectively. Both are less than 400, indicating we should investigate that parameter more carefully. We thus examine the local sampling efficiency in different parts of the posterior by computing the efficiency estimate for small interval estimates.

Figures 9 and 10 show the local efficiency of small interval probability estimates with respect to ranks and parameter values, respectively. The sampler has difficulties in exploring small τ values. As the sampling efficiency for estimating small τ values is practically zero, we may assume that we may miss substantial

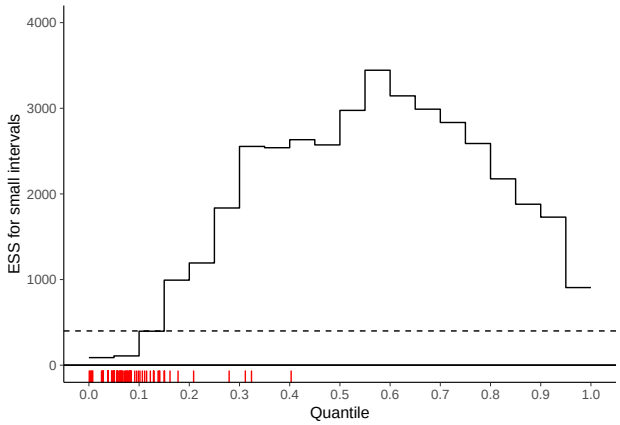

Figure 9: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization. Red ticks show divergent transitions.

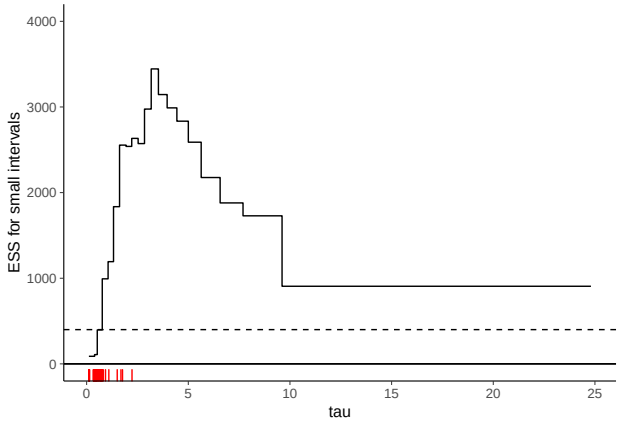

Figure 10: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization. Vertical axis is instead of ranks in scale of parameter τ. Red ticks show divergent transitions.

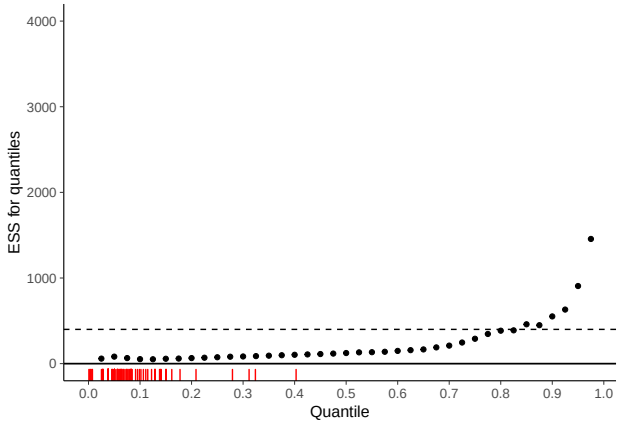

Figure 11: The efficiency of quantile estimates for 8 schools model with centered parameterization. Red ticks show divergent transitions.

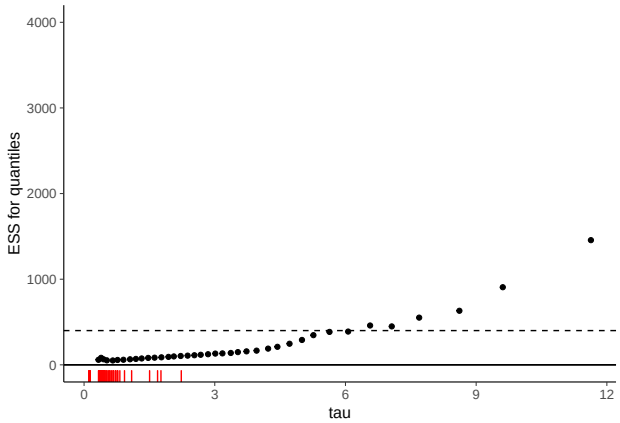

Figure 12: The efficiency of quantile estimates for 8 schools model with centered parameterization. Vertical axis is instead of ranks in scale of parameter τ. Red ticks show divergent transitions.

amount of posterior mass and get biased estimates. Red ticks, which show iterations with divergences, have concentrated to small τ values, indicate also problems exploring small values which is likely to cause bias. Figures 11 and 12 show corresponding efficiency of quantile estimates with respect to ranks and parameter values, respectively. Most of the quantile estimates have worryingly low effective sample size estimate.

Figure 13 shows how the estimated effective sample sizes change when we use more and more draws. Here we don't see sudden changes, but both bulk-ESS and tail-ESS are low. See online appendix for results of longer chains.

In line with other findings, the rank plots of τ in Figure 14 clearly show problems in the mixing of the chains.

4.2.2 Non-centered Eight Schools model

For hierarchical models, the corresponding non-centered parameterization often works better.

We use dynamic HMC and run 4 chains each with 1000 iterations of warmup and 1000 iterations stored. For reasons of comparability, we use the same `adapt_delta` as for the centered parameterization model. There are zero divergences and no other warnings which is a first good sign.

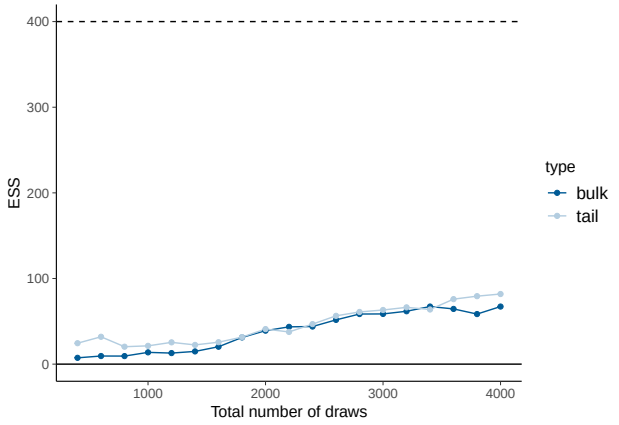

Figure 13: The estimated effective sample sizes with increasing number of iterations for 8 schools model with centered parameterization.

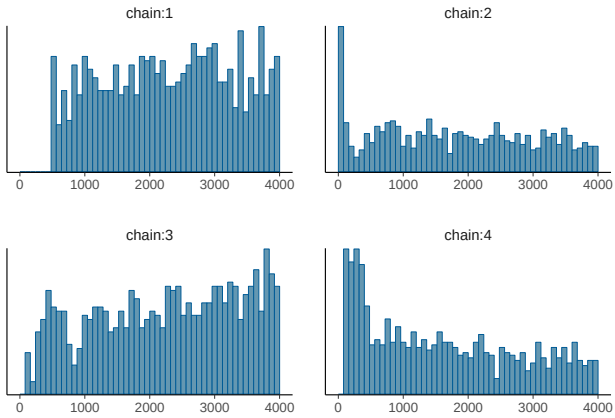

Figure 14: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization.

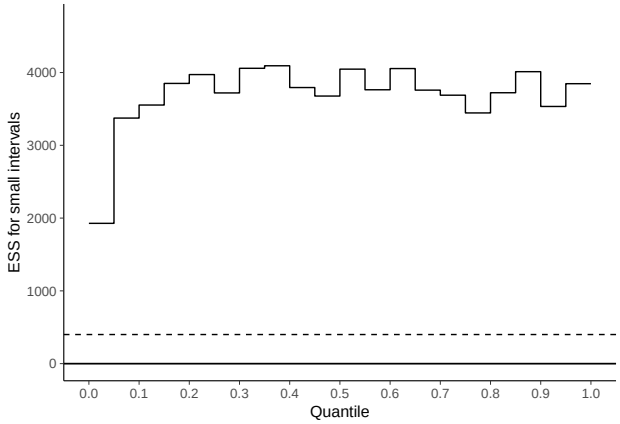

Figure 15: The local efficiency of small interval probability estimates for 8 schools model with non-centered parameterization.

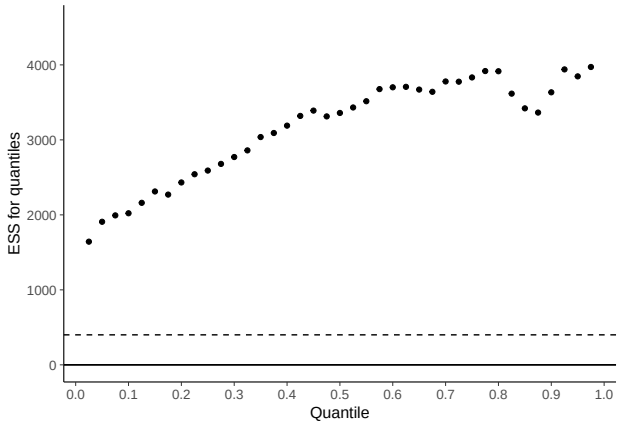

Figure 16: The efficiency of quantile estimates for 8 schools model with non-centered parameterization.

All *split*-$\widehat{R} < 1.01$ and ESS> 400 indicate a much better efficiency of the non-centered parameterization. We examine the sampling efficiency in different parts of the posterior by computing the effective sample size for small interval probability estimates for τ.

Figures 9 and 11 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.

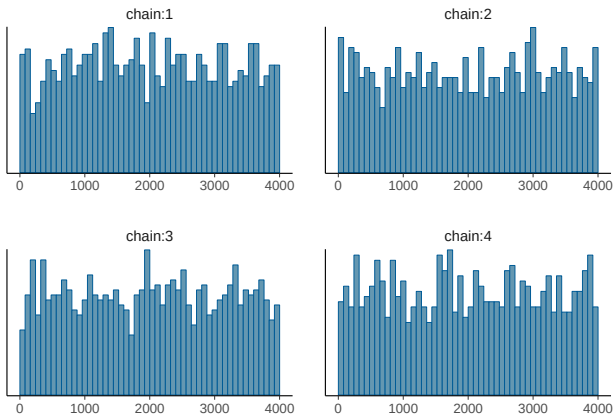

Figure 17: Rank plots of posterior draws from four chains for 8 schools model with non-centered parameterization.

Appendices

Appendix D: Normal distributions with additional trend, shift or scaling

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

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

All chains have the same trend

First all the chains are from the same Normal(0, 1) distribution plus a linear trend added to all chains. 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 $Bulk - ESSdividedbyS$ for easier comparison between different values of S. We see that split-$\widehat{R}$ is more sensitive to trends for small sample sizes, but ESS becomes more sensitive for larger samples sizes (as autocorrelations can be estimated more accurately).

Shifting one chain

Second all the chains are from the same Normal(0, 1) distribution, except one of the chains is sampled with non-zero 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 $Bulk - ESSdividedbyS$ for the same case. The effective sample size is not as sensitive as split-$\widehat{R}$,

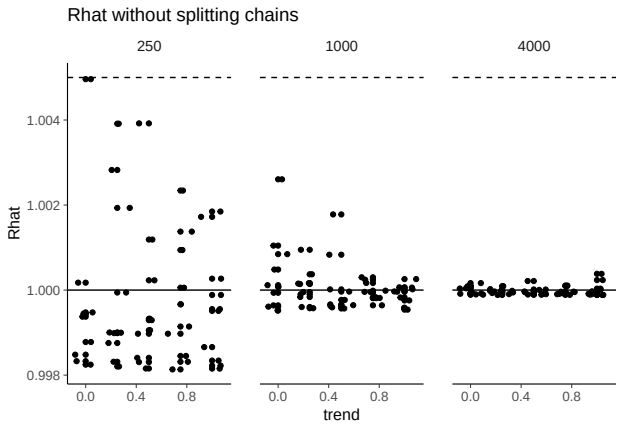

Figure 18: $\widehat{R}$ without splitting for varying chain lengths for chains which have the same trend and a similar marginal distribution.

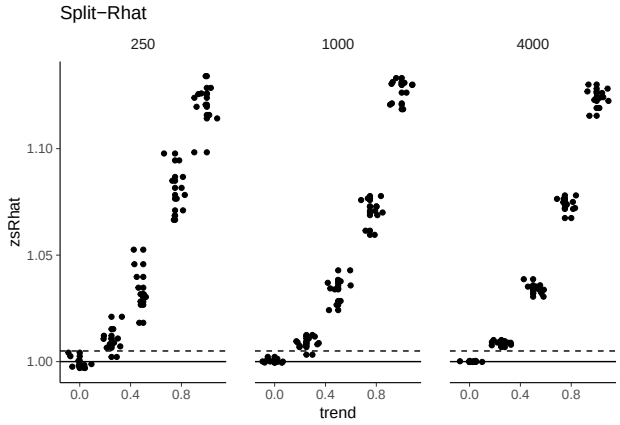

Figure 19: Split-$\widehat{R}$ for varying chain lengths for chains which have the same trend and a similar marginal distribution.

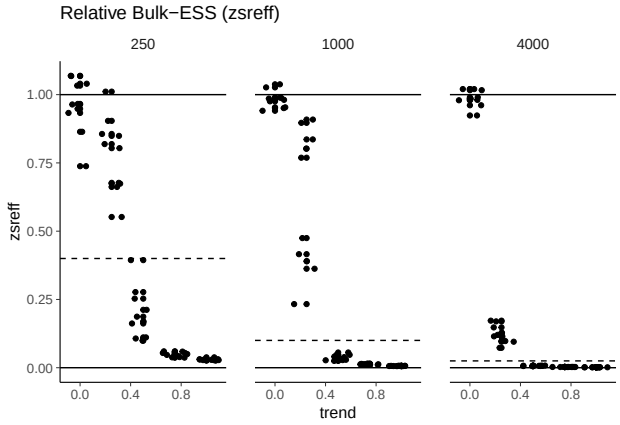

Figure 20: Relative Bulk-ESS for varying chain lengths for chains which have the same trend and a similar marginal distribution. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

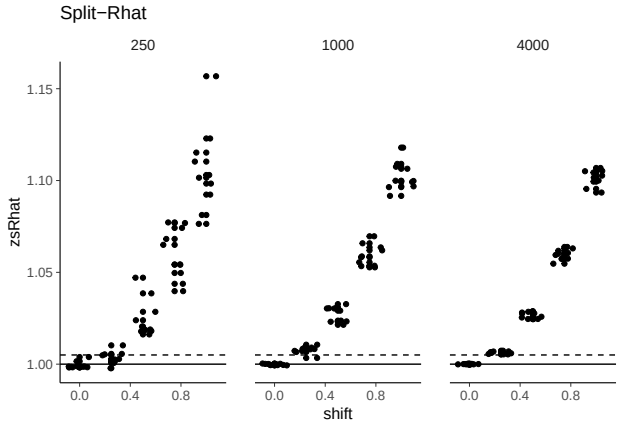

Figure 21: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different mean than the others.

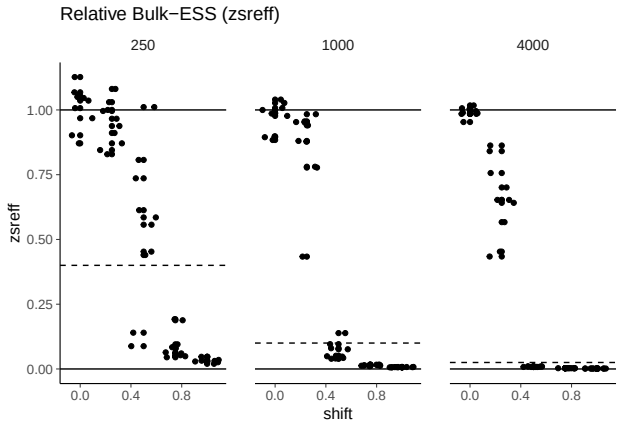

Figure 22: Relative Bulk-ESS for varying chain lengths for chains with one sampled with a different mean than the others. The dashed lines indicate the threshold $S_{\text{eff}} > 400$ at which we would consider the effective sample size to be sufficient.

but a shift with a magnitude of half the marginal standard deviation or more will lead to very low relative efficiency when the total number of draws increases.

Rank plots are practical way to visualize differences between chains. Figure 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.

Scaling one chain

Third all the chains are from the same 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 $Bulk - ESSdividedbyS$ 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.

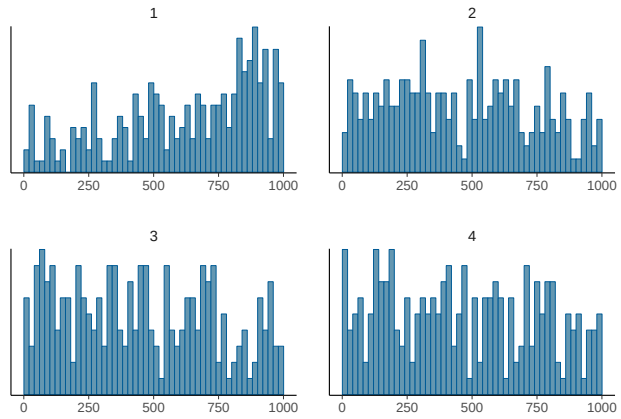

Figure 23: Rank plots of posterior draws from four chains with one sampled with a different mean than the others.

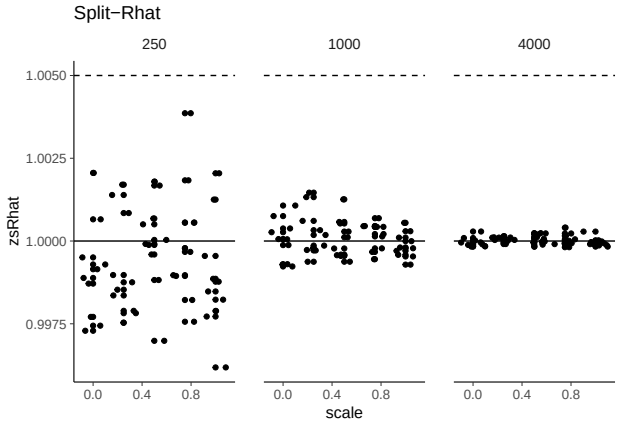

Figure 24: Split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

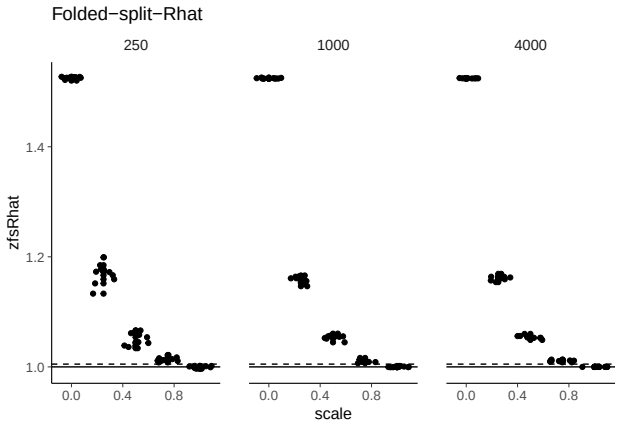

Figure 25: Folded-split-$\widehat{R}$ for varying chain lengths for chains with one sampled with a different variance than the others.

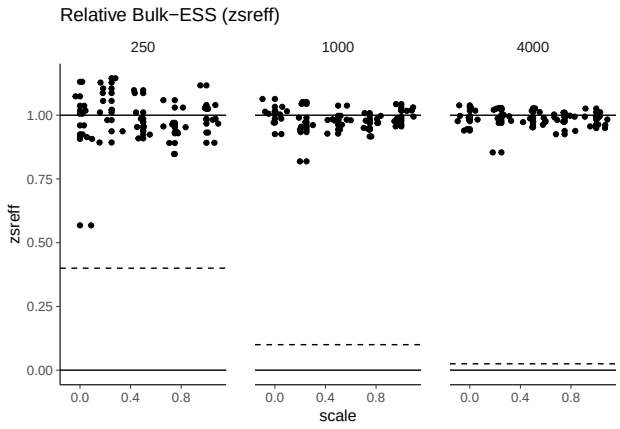

Figure 26: Relative Bulk-ESS for varying chain lengths for chains with one sampled with a different variance than the others.

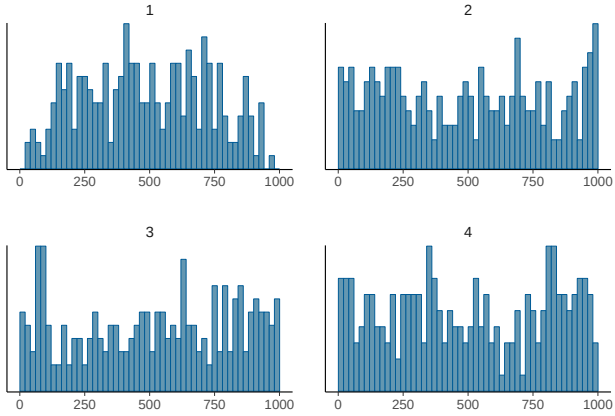

Figure 27: Rank plots of posterior draws from four chains with one sampled with a different variance than the others.

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 E: Cauchy: A distribution with infinite mean and variance

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

Figure 28 shows that trace plots for the first parameter look wild with occasional large values, and it is difficult to interpret possible convergence. Figure 29 shows classic split-$\widehat{R}$, rank normalized split-$\widehat{R}$, and rank normalized folded-split-$\widehat{R}$ for all 50 parameters. Classic split-$\widehat{R}$, which is not well-defined in this case, has much higher variability than rank normalized split-$\widehat{R}$. Rank normalized folded-split-$\widehat{R}$ has higher values than Rank normalized split-$\widehat{R}$ indicating slow mixing especially in tails.

Figure 29 shows different effective sample size estimates for all 50 parameters. Classic ESS, which is not well defined in this case, has very high variability. Bulk ESS is much more stable, and indicates that we can get reliable estimates for the location of the posterior (except for mean). Median ESS is even more stable with

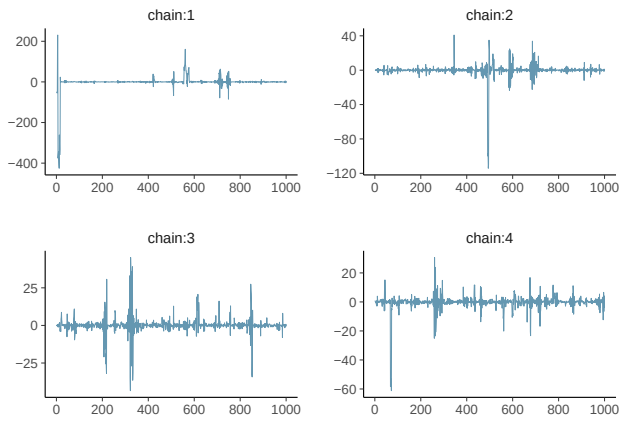

Figure 28: Trace plots of four chains for Cauchy model with nominal parameterization and `max_treedepth`=20.

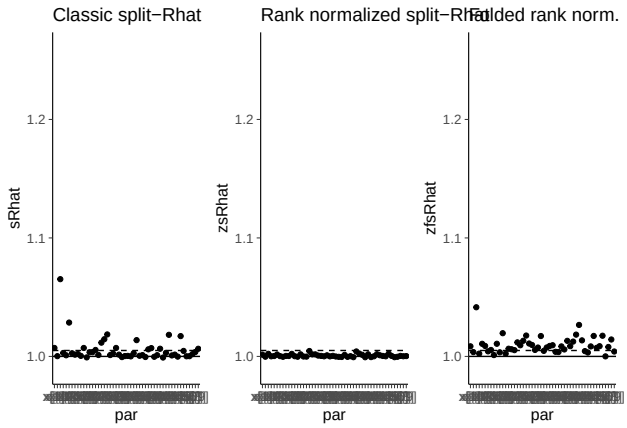

Figure 29: Classic split-$\widehat{R}$, rank normalized split-$\widehat{R}$, and rank normalized folded-split-$\widehat{R}$ for Cauchy model with nominal parameterization and `max_treedepth`=20.

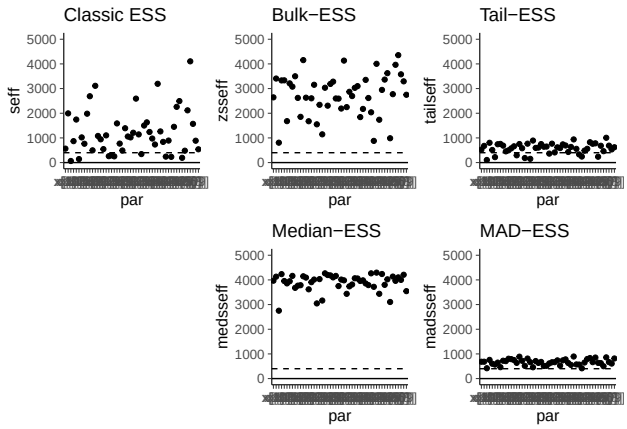

Figure 30: Classic ESS, Bulk-ESS, Tail-ESS, Median-ESS and MAD-ESS for Cauchy model with nominal parameterization `max_treedepth`=20.

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 classic ESS (which is not well defined) keep improving with more iterations. It is however recommended to use more efficient parameterization especially if the tail quantiles are quantities of interest.

A Centered Eight Schools model with very long chains and thinning

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

When autocorrelation time is high, it has been common to thin the chains by saving only a small portion of the draws. This will throw away useful information also for convergence diagnostics. We run 8 schools model with centered parameterization with 400000 iterations per chain. We thin by 200, ending up with 4000 iterations as with the default settings.

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

Figures 31, 32, and 33 show the efficiency of small probability interval estimates, efficiency of quantile estimates, and change of Bulk-SS and Tail-ESS with increasing number of iterations. Unfortunately the thinning makes split-$\widehat{R}$ and ESS estimates to miss the problems. The posterior mean is still biased, being more than 3 sigmas away from the estimate obtained using non-centered parameterization. In this case all four chains fail similarly in exploring the narrowest part of the funnel and all chains seem to "converge" to a wrong stationary distribution.

However, the rank plots shown in Figure 34 are still able to show the problem.

A centered Eight Schools with Jags

So far, we have run all models in Stan, but here we demonstrate that these diagnostics are useful also for other samplers than variants of Hamiltonian Monte-Carlo. We will fit the eight schools models also with Jags, which uses a dialect of the BUGS language to specify models. Jags uses a clever mix of Gibbs and Metropolis-Hastings sampling. This kind of sampling does not usually scale well to high dimensional

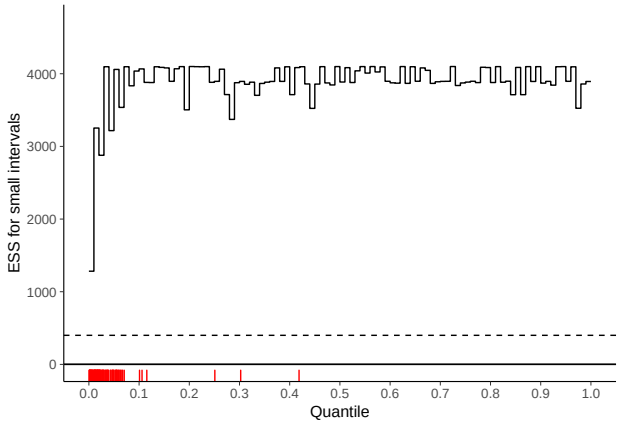

Figure 31: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization, very long chains, and thinning.

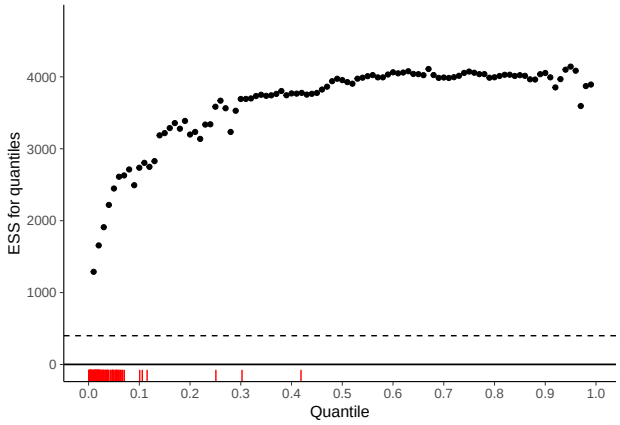

Figure 32: The efficiency of quantile estimates for 8 schools model with centered parameterization, very long chains, and thinning.

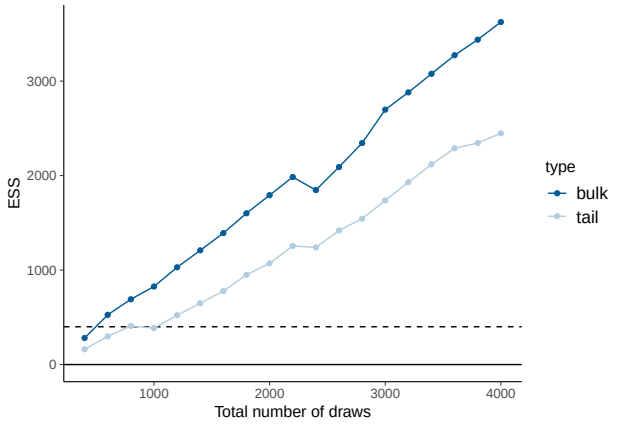

Figure 33: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization, very long chains, and thinning.

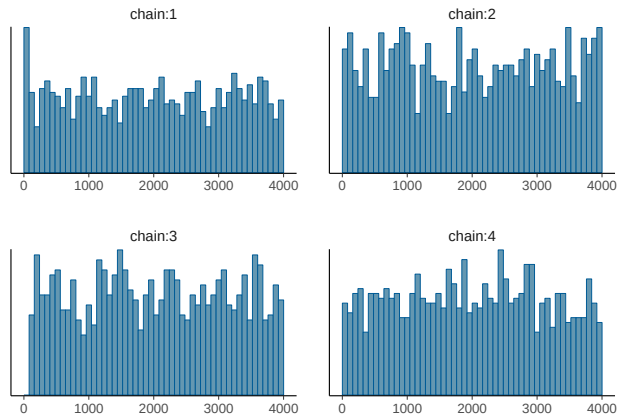

Figure 34: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization, very long chains, and thinning.

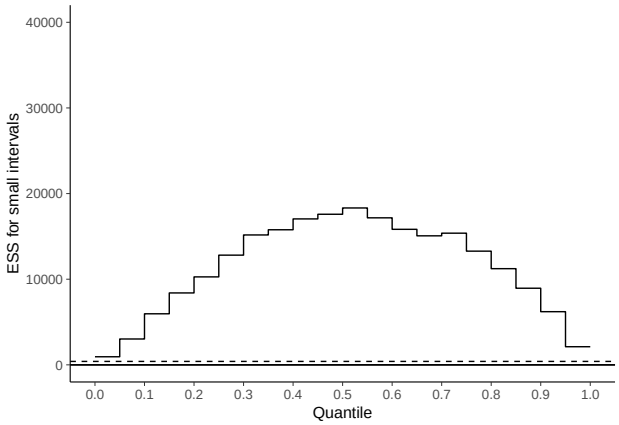

Figure 35: The local efficiency of small interval probability estimates for 8 schools model with centered parameterization and Gibbs sampling.

posteriors of strongly interdependent parameters, but for the relatively simple models discussed in this case study it works just fine.

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 next 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, however, the increase in effective sample size for τ is linear in the total number of draws. Gibbs sampler can reach the narrow part of the funnel, although the sampling efficiency is affected by the funnel. In this simple case the inefficiency of the Gibbs sampling is not dominating and good results can be achieved in reasonable time. The online appendix shows additional results for Gibbs sampling with more efficient non-centered parameterization.

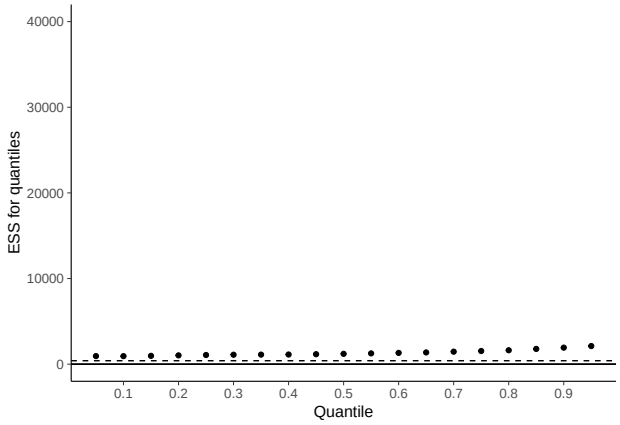

Figure 36: The efficiency of quantile estimates for 8 schools model with centered parameterization and Gibbs sampling..

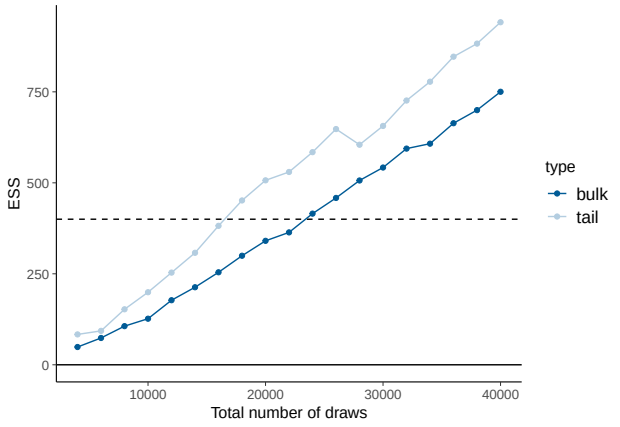

Figure 37: Rank plots of posterior draws from four chains for 8 schools model with centered parameterization and Gibbs sampling.

References

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

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

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

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

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

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

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

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

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

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

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