Bayesian Analysis (2021) **16**, Number 2, pp. 667–718

Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)*†

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.

*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, the Natural Science and Engineering Research Council of Canada, Finnish Center for Artificial Intelligence, and Technology Industries of Finland Centennial Foundation 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.

†A previous version of this manuscript contained a slight omission in the paragraph under equation (3.3) and one typo in equation (4.1). More specifically, under equation (3.3) "assuming the starting distribution of the simulations is appropriately overdispersed" has been changed to "assuming the starting distributions and all intermediate distributions of the simulations are appropriately overdispersed"; in equation (4.1), the denominator was initially written as "S - 1/4" and it has now been corrected to be "S + 1/4". The article was corrected on 22 June 2021.

‡Department of Computer Science, Aalto University, Finland, Aki.Vehtari@aalto.fi
§Department of Statistics, Columbia University, New York
¶Department of Statistical Sciences, University of Toronto, Canada
‖Center for Computational Mathematics, Flatiron Institute, New York
**Department of Computer Science, Aalto University, Finland

 https://doi.org/10.1214/20-BA1221

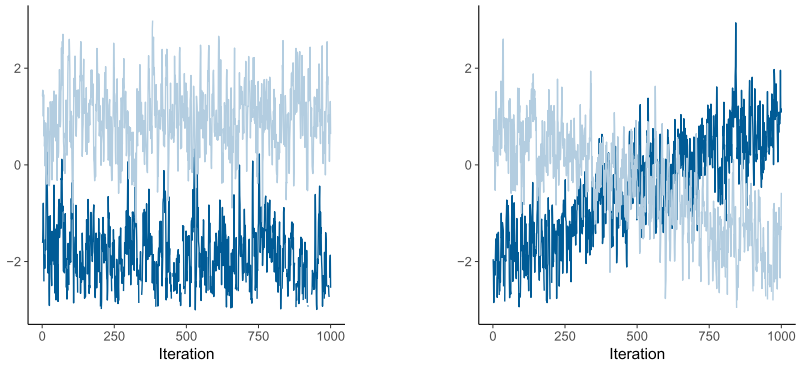

Figure 1: Examples of two challenges in assessing convergence of iterative simulations. (a) In the left plot, either sequence alone looks stable, but the juxtaposition makes it clear that they have not converged to a common distribution. (b) In the right plot, the two sequences happen to cover a common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

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; 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 thousands 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 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 Vehtari et al. (2020), 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.

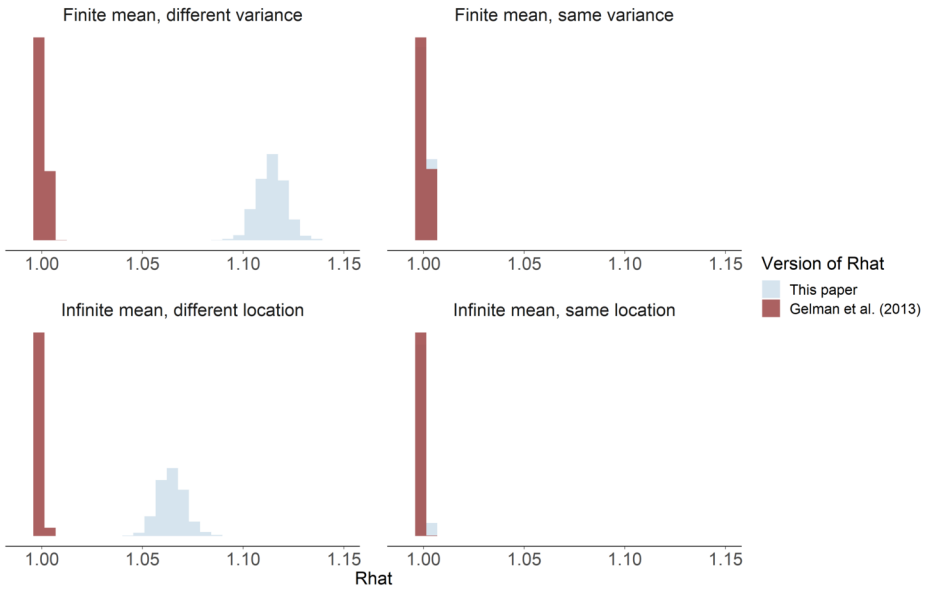

Figure 2: An example showing problems undetected by traditional $\widehat{R}$. Each plot shows histograms of $\widehat{R}$ values over 1000 replications of four chains, each with a thousand draws. In the left column, one of these four chains was incorrect. In the top left plot, we set one of the four 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.

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 used to check that Monte Carlo standard errors (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, $S = MN$ is the total number of draws from all chains, $\theta^{(nm)}$ is nth draw of mth chain, $\theta^{(.m)}$ is the average of draws from mth chain, and $\theta^{(..)}$ is average of all draws. For each scalar summary of interest θ, we compute B and W, the between- and within-chain variances:

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\overline{\theta}^{(.m)} - \overline{\theta}^{(..)})^2, \quad \text{where} \quad \overline{\theta}^{(.m)} = \frac{1}{N}\sum_{n=1}^{N}\theta^{(nm)}, \quad \overline{\theta}^{(..)} = \frac{1}{M}\sum_{m=1}^{M}\overline{\theta}^{(.m)}, \tag{3.1}$$

$$W = \frac{1}{M}\sum_{m=1}^{M}s_m^2, \quad \text{where} \quad s_m^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\theta^{(nm)} - \overline{\theta}^{(.m)})^2. \tag{3.2}$$

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,

$$\widehat{\text{var}}^{+}(\theta|y) = \frac{N-1}{N}W + \frac{1}{N}B. \tag{3.3}$$

This quantity *overestimates* the marginal posterior variance assuming the starting distributions and all intermediate distributions of the simulations are 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

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\theta|y)}{W}}, \tag{3.4}$$

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.

The effective sample size and Monte Carlo standard error

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

$$\text{Var}(\bar{\theta}) = \frac{\text{Var}(\theta|y)}{S}. \tag{3.5}$$

This generalizes to posterior expectations of functionals of parameters $\text{E}\,(g(\theta)|y)$. The square root of (3.5) 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 $\mathrm{E}\left(g(\theta)|y\right)$ or quantiles can be very different. To simplify notation, in this section we consider the effective sample size for the posterior mean $\mathrm{E}\left(\theta|y\right)$. This can be generalized in a straightforward manner to ESS estimates for $\mathrm{E}\left(g(\theta)|y\right)$. Section 4.3 deals with estimating the effective sample size of quantiles, which cannot be presented as expectations.

Estimating the effective sample size

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,

$$S_{\mathrm{eff,BDA2}} = MN\frac{\widehat{\mathrm{var}}^{+}}{B}, \tag{3.6}$$

where $\widehat{\mathrm{var}}^{+}$ 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,

$$N_{\mathrm{eff}} \;=\; \frac{N}{\sum_{t=-\infty}^{\infty}\rho_t} \;=\; \frac{N}{1+2\sum_{t=1}^{\infty}\rho_t}, \tag{3.7}$$

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

$$\rho_t = \frac{1}{\sigma^2} \int_\Theta (\theta^{(n)} - \mu)(\theta^{(n+t)} - \mu)\, p(\theta)\, d\theta. \tag{3.8}$$

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,

$$\rho_t = \frac{1}{\sigma^2} \int_\Theta \theta^{(n)}\, \theta^{(n+t)}\, p(\theta)\, d\theta. \tag{3.9}$$

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,

$$\hat{\rho}_t = 1 - \frac{W - \frac{1}{M}\sum_{m=1}^{M} s_m^2 \hat{\rho}_{t,m}}{\widehat{\text{var}}^+}. \tag{3.10}$$

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 (3.10) 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

$$S_{\text{eff}} = \frac{NM}{1 + 2\sum_{t=1}^{T} \rho_t}. \tag{3.11}$$

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

$$S_{\text{eff}} = \frac{N\,M}{\hat{\tau}}, \tag{3.12}$$

where

$$\hat{\tau} = 1 + 2\sum_{t=1}^{2k+1} \hat{\rho}_t = -1 + 2\sum_{t'=0}^{k} \hat{P}_{t'}, \tag{3.13}$$

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 (HMC) 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):

$$z^{(nm)} = \Phi^{-1}\left(\frac{r^{(nm)} - 3/8}{S + 1/4}\right). \tag{4.1}$$

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,

$$\zeta^{(mn)} = \left|\theta^{(nm)} - \text{median}(\theta)\right|. \tag{4.2}$$

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 $\Pr(\theta \leq \theta_\alpha)$ can be written as expectation which can be estimated with sample mean

$$\Pr(\theta \leq \theta_\alpha) = \mathrm{E}(\mathrm{I}(\theta \leq \theta_\alpha)) \approx \bar{\mathrm{I}}_\alpha = \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}(\theta^{(s)} \leq \theta_\alpha), \tag{4.3}$$

where $\mathrm{I}(\cdot)$ 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

$$\mathrm{Var}(\hat{\theta}_\alpha) = \mathrm{Var}(F^{-1}(\bar{\mathrm{I}}_\alpha)) = \mathrm{Var}(\bar{\mathrm{I}}_\alpha)/f(\theta_\alpha). \tag{4.4}$$

Even if we do not usually know F, this shows that the variance of θ_α is just the variance of $\bar{\mathrm{I}}_\alpha$ 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 (4.2)):

$$\Pr(\zeta \leq \zeta_{0.5}) \approx \bar{\mathrm{I}}_{\zeta,0.5} = \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}(\zeta^{(s)} \leq \zeta_{0.5}), \tag{4.5}$$

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

$$\bar{\mathrm{I}}_{\alpha,\delta} = \Pr(\hat{Q}_\alpha < \theta \leq \hat{Q}_{\alpha+\delta}), \tag{4.6}$$

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 $\mathrm{Var}(\bar{\mathrm{I}}_\alpha)$, and Liu et al. (2016) use a flat top kernel density estimate for $f(\theta_\alpha)$ and a spectral variance approach for $\mathrm{Var}(\bar{\mathrm{I}}_\alpha)$.

We propose an alternative approach which avoids the need to estimate $f(\theta_\alpha)$. Here is how we estimate, for example, a central 90% Monte Carlo error interval for $\hat{\theta}_\alpha$ (any quantiles or intervals can be computed using the same algorithm):

1. Compute the effective sample size S_{eff} for estimating the expectation $\mathrm{E}(\mathrm{I}(\theta \leq \hat{\theta}_\alpha))$.
2. Compute a and b as 5% and 95% quantiles (for other than 90% interval use corresponding quantiles) of
$$\mathrm{Beta}\left(S_{\mathrm{eff}}\alpha + 1,\ S_{\mathrm{eff}}(1-\alpha) + 1\right). \tag{4.7}$$
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 guarantees that $0 < a < 1$ and $0 < b < 1$. Asymptotically as $S_{\mathrm{eff}} \to \infty$, this beta distribution converges towards a normal distribution. Instead of drawing random sample from the beta distribution, we get sufficient accuracy for MCSE using just two deterministically chosen quantiles.
3. Propagate a and b through the nonlinear inverse transforms $A = (F^{-1}(a))$ and $B = (F^{-1}(b))$. Then A and B are corresponding quantiles in the transformed scale. As we don't know F for the quantity of interest, we use a simple numerical approximation:
$$\begin{aligned} \widehat{A} &= \theta^{(s')} \quad \text{where} \quad s' \leq Sa < s' + 1, \\ \widehat{B} &= \theta^{(s'')} \quad \text{where} \quad s'' - 1 < Sb \leq s'', \end{aligned}$$
where $\theta^{(s)}$ have been sorted in ascending order. $\widehat{A}$ and $\widehat{B}$ are then estimated 5% and 95% quantiles (or other quantiles corresponding to which quantiles a and b were chosen to be) of the Monte Carlo error interval for $\hat{\theta}_\alpha$.

The Monte Carlo standard error for $\hat{\theta}_\alpha$ can be approximated, 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. Use of deterministically chosen 16% and

84% quantiles a and b, propagating them through the nonlinear transformation and estimating the standard error from the transformed quantiles, corresponds to unscented transformation which is known to estimate the variance of the transformed quantity correct to the second order (Julier and Uhlmann, 1997).

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 (HMC) 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:

$$x_j \sim \text{Cauchy}(0, 1) \quad \text{for} \quad j = 1, \ldots, 50. \tag{5.1}$$

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 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

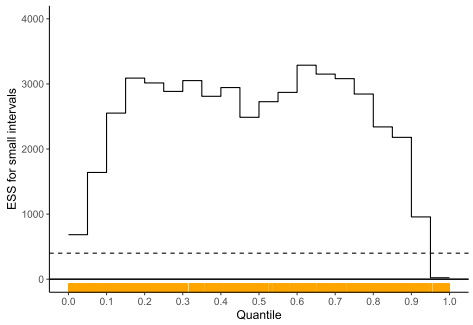

Figure 3: Local efficiency of small-interval probability estimates for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400. Orange ticks show the position of iterations that exceeded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

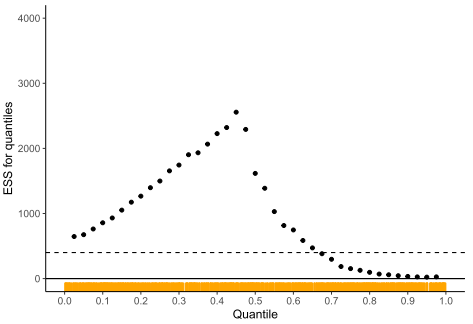

Figure 4: Efficiency of quantile estimates for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400. Orange ticks show the position of iterations that exceeded the maximum tree depth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

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:

$$a_j \sim \text{Normal}(0,1), \qquad b_j \sim \text{Gamma}(0.5, 0.5), \qquad x_j = a_j / \sqrt{b_j}. \tag{5.2}$$

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 deter-

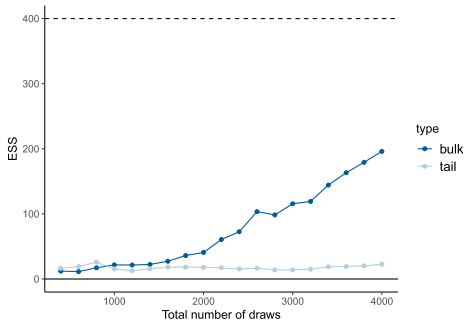

Figure 5: Estimated effective sample sizes with increasing number of iterations for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

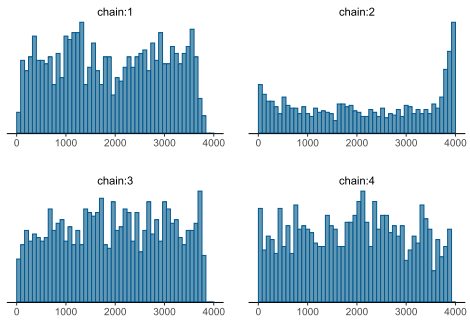

Figure 6: Rank plots of posterior draws from four chains for the Cauchy model with nominal parameterization. Results are displayed for the element of x with the smallest tail-ESS.

ministically 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.

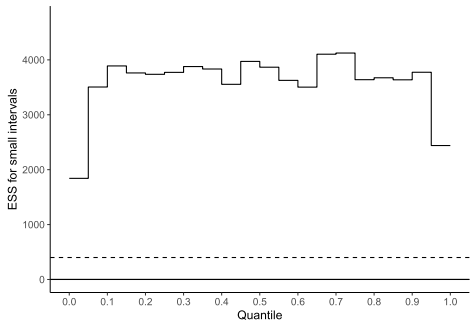

Figure 7: Local efficiency of small-interval probability estimates for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

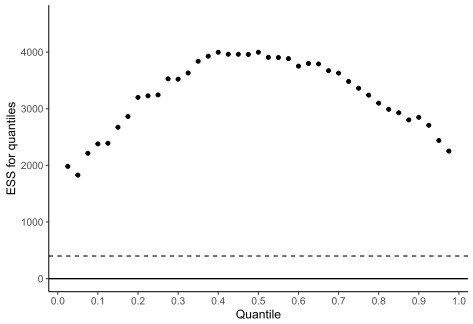

Figure 8: Efficiency of quantile estimates for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS. The dashed line shows the recommended threshold of 400.

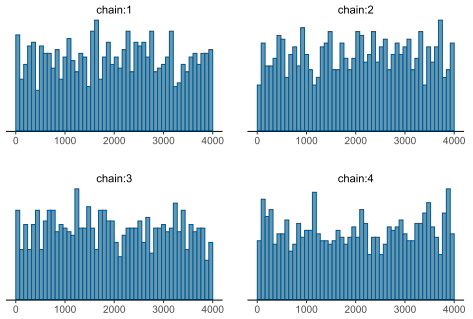

Figure 9: Rank plots of posterior draws from four chains for the Cauchy model with alternative parameterization. Results are displayed for the element of x with the smallest tail-ESS.

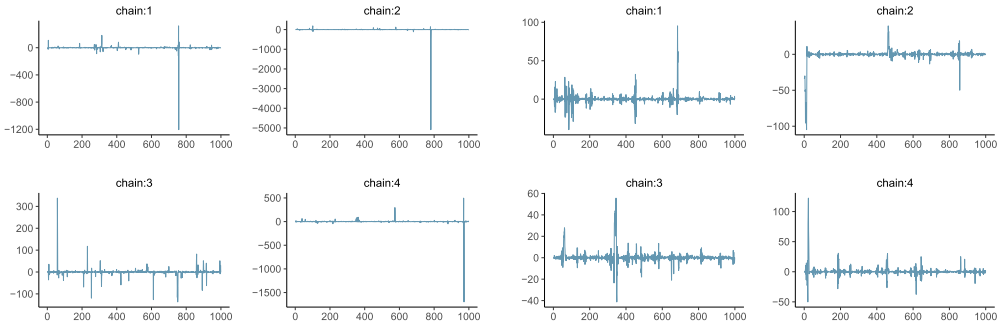

Figure 10: Trace plots of posterior draws from four chains for the Cauchy model with nominal and alternative parameterization. We do not tell which plot belongs to which model and let the reader decide themselves how easy it is to see differences in convergence from those trace plots. Results are displayed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

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.

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,

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

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.

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.

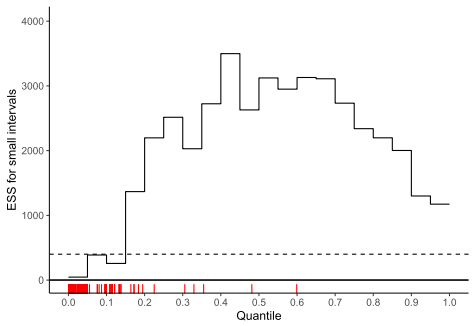

Figure 11: Local efficiency of small-interval probability estimates of τ for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400. Red ticks show the position of divergent transitions.

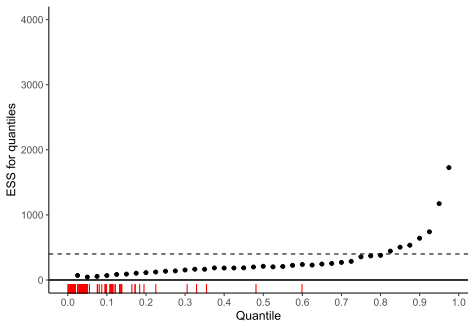

Figure 12: Efficiency of quantile estimates of τ for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400. Red ticks show the position of divergent transitions.

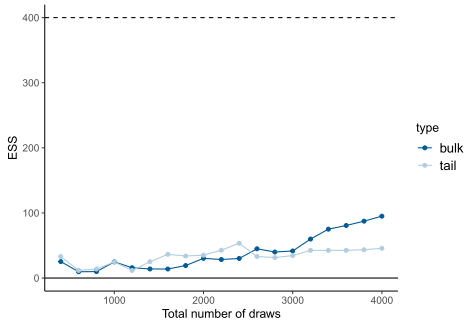

Figure 13: Estimated effective sample sizes of τ with increasing number of iterations for the eight schools model with centered parameterization. The dashed line shows the recommended threshold of 400.

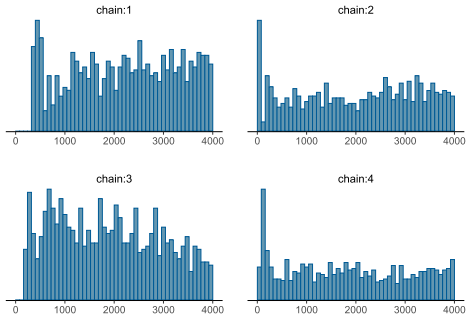

Figure 14: Rank plots of posterior draws of τ from four chains for the eight schools model with centered parameterization.

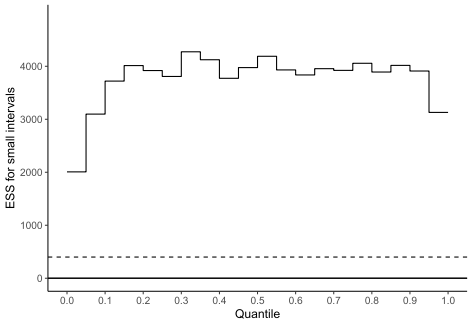

Figure 15: Local efficiency of small-interval probability estimates of τ for the eight schools model with the non-centered parameterization. The dashed line shows the recommended threshold of 400.

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.

Supplementary Material

Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC. Supplementary Material. (DOI: 10.1214/20-BA1221SUPP; .pdf).

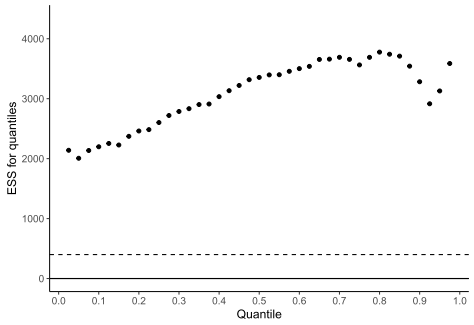

Figure 16: Efficiency of quantile estimates of τ for the eight schools model with the non-centered parameterization. The dashed line shows the recommended threshold of 400.

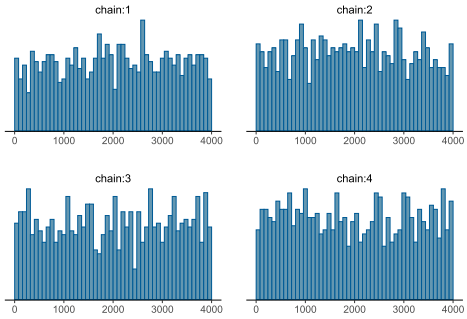

Figure 17: Rank plots of posterior draws of τ from four chains for the eight schools model with non-centered parameterization.

References

Betancourt, M. (2017). "A conceptual introduction to Hamiltonian Monte Carlo." arXiv:1701.02434. MR1699395. doi: https://doi.org/10.1017/CBO9780511470813.003. 677, 683

Betancourt, M. and Girolami, M. (2019). "Hamiltonian Monte Carlo for hierarchical models." In *Current Trends in Bayesian Methodology with Applications*, 79–101. Chapman and Hall/CRC. MR3644666. 672, 690

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Wiley; New York. MR0095553. 678

Brooks, S. P. and Gelman, A. (1998). "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics*, 7(4): 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 668, 671, 683

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M.,

Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles*, 76(1): 1–32. 668

Chernoff, H. and Savage, I. R. (1958). "Asymptotic normality and efficiency of certain nonparametric test statistics." *Annals of Mathematical Statistics*, 29(4): 972–994. MR0100322. doi: https://doi.org/10.1214/aoms/1177706436. 678

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: A comparative review." *Journal of the American Statistical Association*, 91(434): 883–904. MR1395755. doi: https://doi.org/10.2307/2291683. 668, 674, 675

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). "Programming with models: Writing statistical algorithms for general model structures with NIMBLE." *Journal of Computational and Graphical Statistics*, 26(2): 403–413. MR3640196. doi: https://doi.org/10.1080/10618600.2016.1172487. 668

Doss, C. R., Flegal, J. M., Jones, G. L., and Neath, R. C. (2014). "Markov chain Monte Carlo estimation of quantiles." *Electronic Journal of Statistics*, 8(2): 2448–2478. MR3285872. doi: https://doi.org/10.1214/14-EJS957. 676, 680, 681

Fisher, R. A. and Yates, F. (1938). *Statistical Tables for Biological, Agricultural, and Medical Research*. Oliver & Boyd; Edinburgh. MR0030288. 678

Flegal, J. M. and Jones, G. L. (2010). "Batch means and spectral variance estimators in Markov chain Monte Carlo." *Annals of Statistics*, 38(2): 1034–1070. MR2604704. doi: https://doi.org/10.1214/09-AOS735. 676

Friedman, M. (1937). "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. 678

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, third edition*. CRC Press. MR3235677. 668, 669, 670, 672, 673, 675, 676, 687

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. R. (2003). *Bayesian Data Analysis, second edition*. Chapman & Hall. MR2027492. 675

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7(4): 457–511. 667, 668, 671, 675

Geyer, C. J. (1992). "Practical Markov Chain Monte Carlo." *Statistical Science*, 7: 473–483. 672, 674, 676, 677

Geyer, C. J. (2011). "Introduction to Markov chain Monte Carlo." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X. L. (eds.), *Handbook of Markov Chain Monte Carlo*. CRC Press. MR3185067. 676, 677

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57(1): 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 672, 674, 675

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15: 1593–1623. URL http://jmlr.org/papers/v15/hoffman14a.html. MR3214779. 677

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2017). "Unbiased Markov chain Monte Carlo with couplings." *arXiv preprint* arXiv:1708.03625. MR3949304. doi: https://doi.org/10.1093/biomet/asy074. 672

Julier, S. J. and Uhlmann, J. K. (1997). "New extension of the Kalman filter to nonlinear systems." In *Proc. SPIE 3068, Signal processing, sensor fusion, and target recognition VI*, 182–193. SPIE. 682

Kong, A., Liu, J. S., and Wong, W. H. (1994). "Sequential Imputations and Bayesian Missing Data Problems." *Journal of the American Statistical Association*, 89(425): 278–288. MR3738474. 675

Laurmann, J. A. and Gates, W. L. (1977). "Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models." *Journal of the Atmospheric Sciences*, 34(8): 1187–1199. MR0496446. doi: https://doi.org/10.1175/1520-0469(1977)034⟨1187:SCITEO⟩2.0.CO;2. 675

Liu, J., Nordman, D. J., and Meeker, W. Q. (2016). "The number of MCMC draws needed to compute Bayesian credible bounds." *The American Statistician*, 70(3): 275–284. MR3535514. doi: https://doi.org/10.1080/00031305.2016.1158738. 681

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, 28(25): 3049–3067. MR2750401. doi: https://doi.org/10.1002/sim.3680. 668

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility." *Statistics and Computing*, 10(4): 325–337. 668

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). "MCMC convergence diagnostics: A review." In Bernardo, J. M., Berger, J. O., and Dawid, A. P. (eds.), *Bayesian Statistics 6*, 415–440. Oxford University Press. MR1723507. 668, 674, 675

Plummer, M. (2003). "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. 668

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, 6(1): 7–11. URL https://journal.r-project.org/archive/. 675

Raftery, A. E. and Lewis, S. M. (1992). "How many Iterations in the Gibbs Sampler?" In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 763–773. Oxford University Press. 679

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, second edition. MR2080278. doi: https://doi.org/10.1007/978-1-4757-4145-2. 668, 674, 675

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). "Probabilistic programming in Python using PyMC3." *PeerJ Computer Science*, 2: e55. 668

Sorensen, D. A., Andersen, S., Gianola, D., and Korsgaard, I. (1995). "Bayesian inference in threshold models using Gibbs sampling." *Genetics Selection Evolution*, 27(3): 229. 675

Stan Development Team (2018a). "RStanArm: Bayesian applied regression modeling via Stan. R package Version 2.17.4." URL http://mc-stan.org. 680

Stan Development Team (2018b). "Stan Modeling Language Users Guide and Reference Manual. Version 2.18.0." URL http://mc-stan.org. 673, 675, 677, 683, 688

Vats, D. and Knudson, C. (2018). "Revisiting the Gelman-Rubin Diagnostic." arXiv:1812.09384. 671, 672, 673, 674, 676

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2020). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC. Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1221SUPP. 669

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC. MR2676002. doi: https://doi.org/10.1201/EBK1439808184. 680

Invited Discussion

Dootika Vats* and Galin Jones†

We congratulate Vehtari et al. (2021) for a thought-provoking article addressing important and challenging practical problems facing users of Markov chain Monte Carlo (MCMC) algorithms. Their novel visualizations and tools will allow users to gain further insights into their simulation experiments. Vehtari et al. (2021) develop an improvement of the $\hat{R}$ statistic of Gelman and Rubin (1992). The main issues we consider are what exactly this new statistic estimates and how it might be used. We also discuss some reasons for the instability of the original $\hat{R}$ and how that might be improved.

We mostly focus on the use of a single Markov chain. However, we have no intention of rekindling the debates of one long run versus shorter parallel runs of the previous century (see, e.g., Gelman and Rubin, 1992; Geyer, 1992) except to say that any comparisons between methods should be based on the same overall computational effort. Modern computational hardware makes independent parallel implementation of MCMC nearly trivial and we see no reason not to take advantage of it. Indeed we routinely do so in our applied work. Our focus on single chain simulations below is done only to simplify the discussion.

1 Stationarity, convergence, and mixing

MCMC practitioners often speak of detecting convergence, stationarity, or mixing of the Markov chain when assessing their simulation output. It will pay dividends to define these terms and understand how they are distinct concepts.

Suppose F is a probability distribution having support X; in Bayesian statistics this is the posterior distribution. Our goal is to learn about F in order to perform statistical inference and due to the complexity of F we are forced to use Monte Carlo methods; more on this in Section 1.1. MCMC algorithms typically simulate a realization of a time-homogenous Markov chain $\{X_1, X_2, X_3, \ldots\}$ with its dynamics determined by a Markov transition kernel, $P(x, \cdot)$. Somewhat loosely, we can think of the kernel as giving the probability of moving to a set A given that the current state is x, or

$$P(x, A) = \Pr(X_{n+1} \in A \mid X_n = x) \qquad n \geq 1.$$

The distribution ν for X_1 is chosen by the practitioner. Point mass initial distributions are common, in which case the initial state is chosen deterministically. Each element of the Markov chain has a marginal distribution which we will denote $F_n(\nu, P)$ to highlight the dependence of the marginal distribution on the chosen algorithm P and the initial distribution ν. The exact form of $F_n(\nu, P)$ is typically not explicitly available.

*Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, dootika@iitk.ac.in

†School of Statistics, University of Minnesota Twin Cities, galin@umn.edu

Standard MCMC algorithms such as Metropolis-Hastings, Gibbs samplers, conditional Metropolis-Hastings, and so on, yield time-homogenous Markov chains such that the target distribution F is invariant. Thus, if $X_n \sim F$, then $X_{n+1} \sim F$ for all n. Such a Markov chain is said to be *stationary*.

If $X_1 \sim F$, then $F_n(\nu, P) = F$ for all n, but otherwise this will not be the case. While it is uncommon to be able to make an initial draw from F, it is typically the case that, for large n, the marginal distribution $F_n(\nu, P)$ will be close to F irrespective of the choice of ν. More specifically, if $\|\cdot\|$ is total variation norm, then, under standard assumptions (Douc et al., 2018; Meyn and Tweedie, 2009) the Markov chain is *ergodic*, so that for all ν,

$$\|F_n(\nu, P) - F\| \to 0 \quad \text{as} \quad n \to \infty. \tag{1}$$

This implies that the marginal distribution of X_n converges weakly to F. Thus, as the simulation experiment progresses it will produce an increasingly representative sample from F no matter the initial distribution ν.

Since the convergence in (1) is an asymptotic event, MCMC experiments often begin by trying to identify n^* so that the distribution of X_{n^*} is approximately F. Given $\varepsilon > 0$, ideally we would be able to identify some $n^* = n^*(\varepsilon)$ so that

$$\|F_{n^*}(\nu, P) - F\| \leq \varepsilon. \tag{2}$$

Because the total variation norm is non-increasing in n, if we can identify n^*, then every subsequent draw would also be approximately distributed as F. Informally, such a Markov chain is often said to have reached *stationarity* or to have *converged*, even though, in the strictest sense of these words, this it is not possible

The values before X_{n^*} are often discarded and only the remaining are used for estimation. Since n^* need not be the smallest such value, it is known as a *sufficient burn-in*. Naturally, the time to be within ϵ total variation distance of F depends on the chosen algorithm, P, and the initial distribution, ν.

Mixing may be defined in several ways. The way we will use it here, in keeping with common practice, is as an informal qualitative assessment as to whether the simulation is adequately exploring the support of F. An MCMC simulation experiment which is apparently exploring the support of F well is said to be mixing well while one that does not is said to be mixing poorly.

These concepts are illustrated in the following example.

Example 1. *Let F correspond to a two-component mixture of Gaussian distributions*

$$.7\, N(-5, 1) + .3\, N(5, .5)\,.$$

We employ a Metropolis-Hastings algorithm with proposal $N(\cdot, h^2)$ for $h \in \{1, 3, .03\}$. The results for simulations of length 10^4 are given in Figure 1. The columns in the figure correspond to the three choices of h while the rows correspond to different choices of the initial distribution. In the top row the simulation was started with a draw from F, that is, we deliberately set $\nu = F$ while in the bottom row we set ν to be the point mass distribution at -10.

Consider the top row of Figure 1. All three plots are simulations of stationary Markov chains, for which convergence trivially holds. Although the simulation in the top right panel is obviously slow mixing, it may be surprising that the simulation in the top left panel is also slowly mixing. Neither simulation is exploring the support effectively. The middle panel is the only simulation that is mixing well. It bears repeating that all three panels display simulations of stationary Markov chains.

In the bottom row of Figure 1, none of the simulations are of stationary Markov chains because we set ν to be the point mass distribution at -10. As with the top row, only the middle plot exhibits a simulation which is mixing well while the other two plots exhibit slow mixing. Moreover, identifying non-convergence looks to be challenging. For example, although both plots in the left column are nearly identical, the simulation in the bottom-left plot likely has not "converged" sufficiently close to F, even after 10^4 iterations, since it has not visited the other mode so that most of the mass of the empirical distribution would be concentrated in the bottom mode. Similar comments hold for the simulation in the bottom-right panel. However, the middle plot displays a simulation which has likely "converged." We say this because when we repeat this simulation experiment a large number of times and examine the empirical distribution of X_{500}, that is the 500th step of the simulation, it is similar to F so we suspect that convergence occurs much sooner than after 10^4 iterations.

At several points above we explicitly used our knowledge of the bimodality of the invariant distribution. This sort of knowledge is often unavailable in practically relevant applications of MCMC, which substantially complicates the assessment of the simulation output.

We see that slow mixing Markov chains can be stationary and fast mixing Markov chains can be non-stationary. Fast mixing Markov chains likely converge more quickly than those that are slow mixing. Vehtari et al. (2021, Section 3.1) use their split-$\hat{R}$ to address the question, "Did the chains mix well?" We are curious if the new split-$\hat{R}$ can be successful in diagnosing convergence? Certainly, split-$\hat{R}$ can identify situations when summary statistics of chunks of the chains disagree; indicating that the chunks may not have representatively explored the state space. On the other hand, it is not too difficult to imagine implementing several parallel simulations which are all attracted to one of the modes and produce results all look similar to the lower left panel in Figure 1. Such a simulation would seemingly mix well, but would not, in fact, be producing a representative sample from the target distribution.

1.1 Learning about F

Asymptotic arguments also justify the use of MCMC in learning about features of F. For example, if $h : \mathsf{X} \to \mathbb{R}$, we may be interested in the expectation

$$\mu_h := \mathrm{E}_F[h(X)] = \int_{\mathsf{X}} h(x)F(dx).$$

Often there are many expectations and marginal quantiles of interest, but in the interest of a concise presentation, we will focus on univariate expectations; the general case is

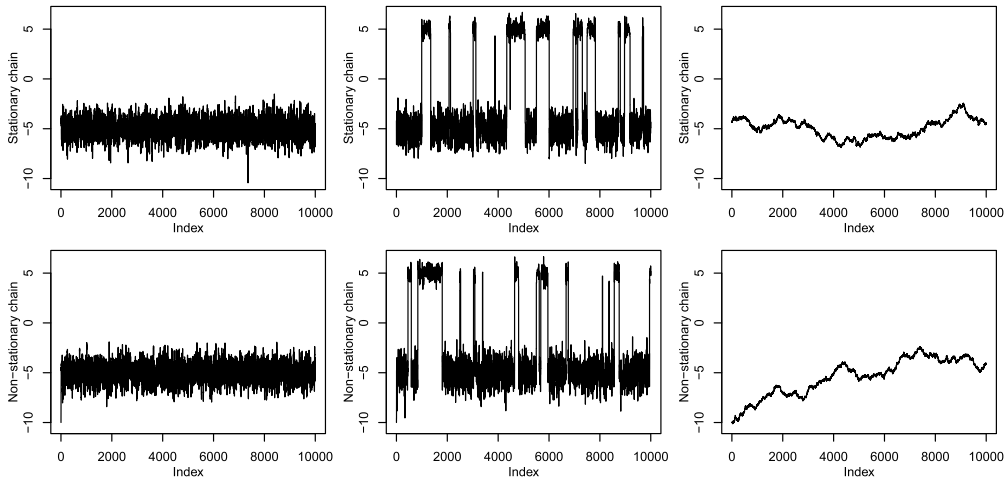

Figure 1: Trace plot of 10^4 samples of the Metropolis-Hastings algorithm for $h = 1$ (left), $h = 3$ (middle), and $h = .03$ (right). Top row are with $\nu = F$ and bottom row is ν being the point mass at -10.

more complicated but can be handled analogously (Robertson et al., 2021; Vats et al., 2019).

Suppose a sufficient burn-in, $n^* \geq 1$, is used, then we reindex the sample as

$$\{X_1, X_2, \ldots, X_n\},$$

so the total simulation effort is $n^* + n - 1$. If the expectation exists, then the sample mean converges to μ_h. That is, as $n \to \infty$, with probability 1,

$$\bar{h}_n := \frac{1}{n} \sum_{i=1}^{n} h(X_i) \to \mu_h. \tag{3}$$

In order for $\bar{h}_n$ to be reliable in the sense that another simulation of the same effort would produce a similar result, we will need to control both the bias and the variance. Choosing an initial distribution ν which is close to F, whether it is through the use of a sufficient burn-in n^* or by other means, is how we go about ensuring a representative sample and hence controlling the bias. Controlling the variance boils down to the value of n chosen, that is the size of the simulation effort after burn-in. The number of Monte Carlo samples obtained depends only on our patience, but, no matter the size of n, there will be an unknown Monte Carlo error $\hat{\theta}_n - \theta$. If a central limit theorem (CLT) holds, that is, if there exists $\sigma^2 > 0$ so that, as $n \to \infty$,

$$\sqrt{n}(\bar{h}_n - \mu_h) \xrightarrow{d} N(0, \sigma^2),$$

then the approximate sampling distribution of the Monte Carlo error is available and we can use it to assess its variability. Observe that $\sigma^2 \neq \mathrm{Var}_F h(X)$ since it must account for the serial dependence in the Markov chain.

As noted by Vehtari et al. (2021), if a CLT does not hold, the estimator $\bar{h}_n$ has unbounded variability and is unreliable. They further observe that one way to establish a Markov chain CLT is to show that the Markov chain is *geometrically ergodic*. Geometric ergodicity also has a connection to identifying a sufficient burn-in n^* which we describe in the next section.

2 Representative samples

If we are able to obtain one draw $X_1 \sim F$, then all subsequent samples are from F. Since this is mostly challenging, users are left to choose ν so that n_* is as small as possible and approximately representative samples are obtained relatively quickly. The sufficient burn-in n^* may be identified through rigorous theoretical methods or by convergence diagnostics.

2.1 Theory and its limitations

Recall that we would ideally want to find n^* satisfying inequality (2). To slightly simplify the presentation, we assume for the remainder that the initial distribution ν is a point mass distribution at x_1 and we denote the associated marginal distribution at time n as $F_n(x_1, P)$. If there exist $M : \mathsf{X} \to \mathbb{R}_+$ and $0 < \rho < 1$ satisfying

$$\|F_n(x_1, P) - F\| \leq M(x_1)\rho^n, \tag{4}$$

then the kernel P is *geometrically ergodic*. The influence of the initial value is explicit in (4). If the initial value is chosen such that $M(x_1)$ is large, then convergence could be slow until the Monte Carlo sample size is so large that ρ^n becomes small enough to offset it. Amazingly there are rigorous, constructive methods for establishing (4) and finding n^* from (2). In particular, there is a substantial literature that provide upper bounds for the right-hand side of (4) (Baxendale, 2005; Douc et al., 2004; Jerison, 2019; Roberts and Tweedie, 1999; Rosenthal, 1995). While these upper bounds could be utilized more frequently in applications to obtain values for n^*, the required analysis of the Markov chain can be challenging, and the upper bounds are sometimes so conservative as to be less than helpful (Jones and Hobert, 2004; Qin and Hobert, 2021).

2.2 Diagnostics and their limitations

Given the difficulty of finding a rigorous sufficient burn-in n^* in realistic applications, it is natural to try to assess the performance of the simulation by observing its output. This has led to an extensive literature on so-called convergence diagnostics. The interested reader may consult Roy's (2020) recent review.

As adumbrated in Section 1, using a realization of a Markov chain (or even several parallel chains) to detect convergence or non-convergence is challenging. There is

certainly a little additional solace that can be taken if there are several parallel chains started from different points and they all yield similar results. But this is far from a *guarantee* that all of the meaningful parts of the space have been explored or that a truly representative sample is being produced.

3 Interpretability and stability

Numerical diagnostics use the simulated data to construct summary statistics which estimate a feature of the process. If the summary statistic stabilizes, then the diagnostic indicates that there is no evidence of a problem (e.g. non-convergence) and victory is declared. In most cases, such summary statistics are studied in-context and have clear interpretations. So, although it is mathematically understood why a value closer to 1 is desirable, it remains unclear exactly what feature of the process Vehtari et al.'s (2021) new version of $\hat{R}$ is estimating. In other words, what is the feature R, that the new $\hat{R}$ is estimating? We would greatly appreciate it if they could help us understand this better.

An interpretation of the original $\hat{R}$ is possible via effective sample size (ESS). Recall that σ^2 is the variance from the asymptotic normal distribution of the Monte Carlo error. If $\lambda^2 = \mathrm{Var}_F h(X)$, then the ESS for estimating μ_h is $n\lambda^2/\sigma^2$. If the estimated ESS is large, then the Monte Carlo sample size is large enough that the variability due to Monte Carlo sampling is small relative to the variability of the stationary distribution. It seems clear that ESS is a measure of estimation reliability of the sample mean $\bar{h}_n$; this intuition can be made rigorous (Gong and Flegal, 2016). While the original $\hat{R}$ is commonly viewed as a convergence diagnostic, recently it has been shown to be equivalent to the ESS for estimating $\mu_x = \mathrm{E}_F X$ (Vats and Knudson, 2021). Notice that, as illustrated by Vehtari et al. (2021), this makes the original $\hat{R}$ problematic when this expectation does not exist.

Both ESS and $\hat{R}$ require the estimation of σ^2. The original $\hat{R}$ uses B/n, the sample variance of sample means from M chains to estimate σ^2, while common implementations of ESS utilize more robust alternatives. Since M is typically small, the estimator B/n is highly variable for σ^2 which often leads to premature termination of the simulation (Flegal et al., 2008). Vehtari et al. (2021, Section 3) use initial sequence estimators in estimating ESS. Given the connection between $\hat{R}$ and ESS, using this estimator in place of B/n is likely to yield reduced variability in $\hat{R}$. We also add here that a critical assumption for validity of initial sequence estimators is reversibility (Geyer, 1992). Non-reversible Markov chains are routine in applications, e.g. simple Gibbs samplers or other component-wise samplers are typically not reversible. In this case, alternatives like batch means (Glynn and Whitt, 1991) or spectral variance (Anderson, 1971) estimators will be more appropriate.

4 Final remarks

Even after decades of empirical and theoretical work on MCMC algorithms, most MCMC experiments are conducted in a black box manner because little is known about

the structure of F prior to simulation. Diagnostics based on the simulated values cannot *prove* that the simulation is providing representative samples, and the best we can hope for is that it indicates when a problem has occurred. If the diagnostics do not indicate a problem, then only a little solace should be taken by the practitioner: an absence of evidence of non-convergence is not evidence of convergence.

However, we fully understand the need for both graphical and numerical diagnostics and indeed, we use some of them in our own work. The creative methods of Vehtari et al. (2021) significantly expand the toolkit of MCMC simulation output analysis.

References

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley & Son. MR0283939. 700

Baxendale, P. H. (2005). "Renewal theory and computable convergence rates for geometrically ergodic Markov chains." *Annals of Applied Probability*, 15: 700–738. MR2114987. doi: https://doi.org/10.1214/105051604000000710. 699

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018). *Markov Chains*. Springer. MR3889011. doi: https://doi.org/10.1007/978-3-319-97704-1. 696

Douc, R., Moulines, E., and Rosenthal, J. S. (2004). "Quantitative bounds on convergence of time-inhomogeneous Markov chains." *Annals of Applied Probability*, 14: 1643–1665. MR2099647. doi: https://doi.org/10.1214/105051604000000620. 699

Flegal, J. M., Haran, M., and Jones, G. L. (2008). "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science*, 23: 250–260. MR2516823. doi: https://doi.org/10.1214/08-STS257. 700

Gelman, A. and Rubin, D. B. (1992). "Inference From Iterative Simulation Using Multiple Sequences (with discussion)." *Statistical Science*, 7: 457–472. 695

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo (with discussion)." *Statistical Science*, 7: 473–511. 695, 700

Glynn, P. W. and Whitt, W. (1991). "Estimating the asymptotic variance with batch means." *Operations Research Letters*, 10: 431–435. MR1141337. doi: https://doi.org/10.1016/0167-6377(91)90019-L. 700

Gong, L. and Flegal, J. M. (2016). "A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 25: 684–700. MR3533633. doi: https://doi.org/10.1080/10618600.2015.1044092. 700

Jerison, D. C. (2019). "Quantitative convergence rates for reversible Markov chains via strong random times." arXiv:1908.06459. MR3861371. doi: https://doi.org/10.1007/s10255-018-0779-1. 699

Jones, G. L. and Hobert, J. P. (2004). "Sufficient burn-in for Gibbs samplers for a hier-

archical random effects model." *The Annals of Statistics*, 32: 784–817. MR2060178. doi: https://doi.org/10.1214/009053604000000184. 699

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press. MR2509253. doi: https://doi.org/10.1017/CBO9780511626630. 696

Qin, Q. and Hobert, J. P. (2021). "On the limitations of single-step drift and minorization in Markov chain convergence analysis." To appear in *Annals of Applied Probability*. 699

Roberts, G. O. and Tweedie, R. L. (1999). "Bounds on regeneration times and convergence rates for Markov chains." *Stochastic Processes and their Applications*, 80: 211–229. Corrigendum (2001) **91**: 337–338. MR1807678. doi: https://doi.org/10.1016/S0304-4149(00)00074-0. 699

Robertson, N., Flegal, J. M., Vats, D., and Jones, G. L. (2021). "Assessing and visualizing simultaneous simulation error." To appear in *Journal of Computational and Graphical Statistics*. 698

Rosenthal, J. S. (1995). "Minorization conditions and convergence rates for Markov chain Monte Carlo." *Journal of the American Statistical Association*, 90: 558–566. MR1340509. 699

Roy, V. (2020). "Convergence diagnostics for Markov chain Monte Carlo." *Annual Review of Statistics and Its Application*, 7: 387–412. MR4104198. doi: https://doi.org/10.1146/annurev-statistics-031219-041300. 699

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106: 321–337. MR3949306. doi: https://doi.org/10.1093/biomet/asz002. 698

Vats, D. and Knudson, C. (2021). "Revisiting the Gelman-Rubin diagnostic." To appear in *Statistical Science*. 700

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC." To appear in *Bayesian Analysis*. 695, 697, 699, 700, 701

Invited Discussion

Christopher M. Hans*

1 Overview

Markov chain Monte Carlo (MCMC) methods continue to play an important role in statistical modeling and data analysis thirty-plus years after their introduction into the statistics research community. This is especially true for Bayesian analysis, where inference under all but the most standard of models typically requires the evaluation of integrals that do not have closed-form solutions. Sampling-based approaches to posterior inference have been popular in part due to their flexibility: rather than customizing analytical approximations to each estimand of interest, samples from a posterior distribution under one parameterization can usually be transformed easily to make inferences about functions of the parameters and to facilitate predictive inference.

The present-day landscape of MCMC methods boasts an abundance of options for posterior sampling. These options range from "bespoke" algorithms that are highly-customized for particular models to generic methods that are intended to work well for many different classes of models without requiring strong assumptions about the dimensions of the parameter spaces or the shapes of the posteriors. Examples of the latter are typically used in software that has been designed to facilitate Bayesian data analysis by providing automated methods for posterior sampling that require minimal input from the user. Stan (Carpenter et al., 2017) and NIMBLE (de Valpine et al., 2017) are two contemporary examples of such software, while WinBUGS (Lunn et al., 2000), JAGS (Plummer, 2003) and OpenBUGS (Lunn et al., 2009) have been used extensively in applied data analysis over the past two decades.

Two central computational questions face modern practitioners when employing MCMC to summarize complex Bayesian models: (i) which MCMC algorithm should one use, and (ii) for how long should one run the chain(s) before convergence is achieved and inferences are reliable? The two questions are somewhat interrelated, as a good choice of algorithm may result in fewer required samples for reliable inference. Both questions can be further refined. An algorithm in a chosen class might require tuning of some sort (e.g., selection of a step-size for a non-adaptive Metropolis algorithm); decisions about how long to run a chain may depend on the types of inferences required. Due to its practical importance, question (ii) received attention in the statistics literature as soon as MCMC methods began to be adopted for Bayesian inference. The question of whether one should run single or multiple chains—and how to monitor and diagnose convergence under both paradigms—was considered from different perspectives (see, e.g., Geyer, 1992; Gelman and Rubin, 1992a,b). While opinions—and practice—varied, it became common to see the multiple-chains approach advocated by Gelman (1996) used in applications of MCMC. Convergence diagnostics under this paradigm included use of the $\widehat{R}$ statistic (Gelman and Rubin, 1992a). Common graphical diagnostics for

*Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A., hans@stat.osu.edu

assessing convergence (or, perhaps more accurately, detecting non-convergence) included trace plots of (functions of) the parameters.

The paper "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC" (Vehtari et al., 2021) takes a modern look at aspects of question (ii). The paper introduces improvements to existing convergence diagnostics and proposes localized versions of effective sample size that are inspired by the fact that convergence of functionals of interest that are specific to different regions of the parameter space may converge at different rates. The ideas are relevant when one has a great deal of freedom over the choice of MCMC algorithm as well as when one is using automated methods for MCMC. As such, the paper's contributions can be incorporated into software such as Stan to improve MCMC practice. Moreover, the methods are conceptually straightforward and easy to implement, making them accessible to those programming MCMC algorithms from scratch.

The paper proposes several innovations aimed at improving assessment of MCMC convergence based on multiple chains. First, the authors identify deficiencies in the existing split-$\widehat{R}$ statistic (Gelman et al., 2013). They show that the existing statistic will fail to diagnose non-convergence in two specific situations: when the chains have different variances but the same mean and when the chains have different locations but infinite variances. The examples are clear and compelling and of practical importance. As noted in the paper, the former can occur when one chain is stuck in the bulk of the distribution and does not adequately explore the distribution's tails, while the latter can occur by design in complex modeling situations. The authors propose a clever solution for the heavy-tailed (latter) situation whereby convergence is monitored not based on the samples directly but based on the (normalized) *ranks* of the samples across the multiple chains. Computing split-$\widehat{R}$ after passing to the normalized ranks remedies potential problems induced by heavy tails without introducing any obvious new complications. For the former (same mean/difference variances) problematic setting, the authors propose first folding the samples about their median via the absolute value operator before rank normalizing the draws and computing the split-$\widehat{R}$ statistic. Both approaches are demonstrated to fix the deficiencies with split-$\widehat{R}$ based on the non-transformed chains, and it is recommended to use the maximum of the rank normalized split-$\widehat{R}$ and the rank normalized folded-split-$\widehat{R}$ when assessing convergence.

The second main innovation in the paper is the introduction of new localized convergence diagnostics and is also of practical importance. When analyzing data based on MCMC output, one often hears the general advice to monitor the chains for convergence based on each estimand of interest. This is wonderful advice if one knows in advance all inferences that will be made. But as data analysis and statistical modeling is an iterative endeavor, it is usually the case that one ends up making inferences about quantities that were not necessarily monitored during the initial run of the chains. Problems may arise if the initial runs were monitored based on measures of the bulk of the distribution (e.g., measures of location) but inferences about tail quantities (e.g., very small or very large quantiles) end up being desired, as longer MCMC runs may be needed to well-estimate the latter. Vehtari et al. (2021) propose monitoring effective sample size for quantiles and small probability intervals for each parameter and running the chain until convergence has been diagnosed for all.

The new methods introduced in the paper are sensible and work well in practice. I congratulate the authors on making improvements to current state-of-the-art methods for this long-standing problem. The rest of my discussion is focused on the graphical diagnostic techniques that are proposed in the paper, as they were of particular interest to me.

2 Graphical diagnostics

The paper also proposes new diagnostic plots for visualizing aspects of MCMC convergence. The use of ranks is extended to create plots to help assess whether each chain is targeting the same distribution. Plots of effective sample size for quantiles and small interval probabilities across the target distribution are used to assess whether the chains are failing to well explore specific regions of the target distribution. Plots of effective sample size per iteration for bulk and tail quantities are used to assess the relative efficiency of the sampler. The visualizations make the numerical diagnostics more user-friendly and interpretable. The "quantile and small-interval plots" are most interesting to me, as they provide a clear visualization of how well the sampler is exploring the full distribution. I expect these plots to be especially helpful when posterior samples will be used to construct density estimates. The plots may have other diagnostic uses, e.g., asymmetry in plots such as Figure 4 in Vehtari et al. (2021) might in some cases suggest specific deficiencies with the sampling algorithm.

When discussing graphical diagnostics, it is difficult not to think of the important role that graphics play in regression analysis. In the regression setting, information about residuals and measures of influence (Cook and Weisberg, 1982) can be plotted to help diagnose model misfit, though care must be taken when making such interpretations (Cook, 1994). One powerful feature of such plots is that in addition to identifying the existence of potential deficiencies with a model, the plots often suggest reasons for the deficiency as well as possible remedial measures that could be taken. The regression graphics aid the interactive model building process.

With this in mind, an interesting question is the extent to which the diagnostic visualizations proposed in Vehtari et al. (2021) might be able to serve a similar purpose: can the plots suggest specific deficiencies with a sampling algorithm in addition to indicating whether more samples should be obtained? The plots are designed to be good at the latter; any potential for achieving the former would make them even more powerful as diagnostic tools. Examples of specific deficiencies might include slow mixing due to small step sizes in a Metropolis algorithm, slow mixing in a Gibbs sampler due to strong correlation between parameters, a failure to rapidly mix across multiple modes of a multimodal distribution, etc. If the plots can suggest specific reasons for slow convergence, targeted modifications to the sampler could be made.

In thinking about this question I have focused on the rank plots that were introduced in Section 4.5 of the paper as a proposed replacement for trace plots. It might be of interest to study how specific problems with convergence manifest themselves in the rank plots, with an eye toward using the rank plots as more general diagnostic tools. If it is indeed the case that rank plots can be useful for detecting specific deficiencies,

I expect that expertise in understanding how to map features of the plots to specific deficiencies will build up over time as users begin to employ the plots across a variety of examples.

As an initial attempt at investigating the potential for such mappings, I consider below two simple examples that illustrate different ways in which a sampler may have difficulty exploring the entirety of a target distribution. The two examples were specifically chosen to be extreme so that the effect on the rank plots is clear. In both examples, a univariate target distribution $f(x)$ was explored via a random-walk Metropolis algorithm with Gaussian innovations. Four chains were run for 1000 iterations each. As intended, in both examples both the rank normalized split-$\widehat{R}$ and the rank normalized folded-split-$\widehat{R}$ statistics strongly indicate that more samples should be obtained. The MCMC samples and the rank plots were then compared to assess the extent to which features of the rank plots might be useful in indicating specific aspects of non-convergence.

2.1 Example 1

The target distribution $f(x)$ in this example is the standard normal distribution. The four chains were started at the values -10, -8, 8 and 10, which are far from the bulk of the target distribution's mass. The standard deviation for the Gaussian innovation in the proposal distribution was taken to be 0.1, which is small relative to the standard deviation of the target distribution and should discourage rapid mixing. Trace plots and rank plots for the four chains appear in Figure 1. The maximum of the two split-$\widehat{R}$ statistics is 2.35, a strong indication that more samples are needed. The rank plots confirm this, as the ranks in each chain are clearly not uniform, with strong differences observed across chains.

The ranks for chains 1 and 2 are heavily biased toward large values, while the opposite is true for chains 3 and 4. There is little overlap between the histograms for chains 1–2 and 3, while the ranks for chain 4 exhibit some overlap with the ranks for chains 1–2. While the general lack of uniformity suggests the chains need to be run longer, the patterns related to the overlap of the histograms suggest that some chains spent more time exploring regions of large values of x, while other chains spent more time exploring small values of x.

The reason for this is clear once we look at the trace plot. The starting values are well into the tails of the distribution, and with no warm-up period accounted for, the chains take a while to reach the bulk. Once there, the small standard deviation in the proposal distribution results in slow mixing of the chain and contributes to non-uniformity in the rank plots. In this example, the rank plots indicate a specific problem (the chains are only exploring specific regions of the parameter space), but there are two causes of the problem and it is not clear that the rank plots on their own can be used to identify both.

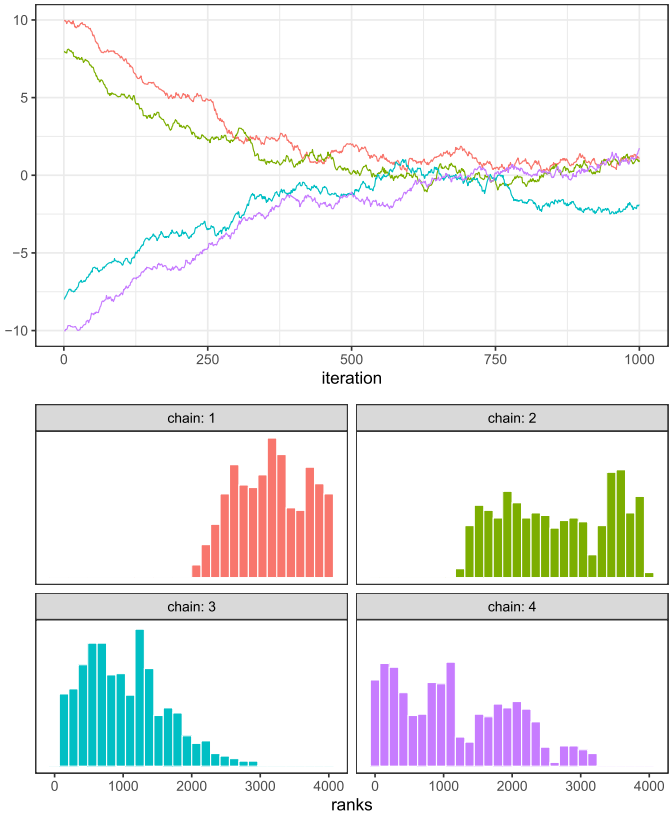

Figure 1: Trace and rank plots of MCMC output for the example in Section 2.1. The target distribution is a standard normal distribution.

2.2 Example 2

The target distribution $f(x)$ in this example is an equally-weighted mixture of a $\mathrm{N}(-4, 1)$ distribution and a $\mathrm{N}(4, 1)$ distribution. The distribution has two distinct modes that are well separated. The standard deviation for the Gaussian innovation in the proposal distribution was taken to be 1 to facilitate mixing within a mode but to discourage mixing across modes. The four chains were started at the values -5, -3, 3 and 5 so that the first two chains start in the bulk of the distribution near the lower mode while the second two chains start in the bulk of the distribution near the upper mode. Trace plots and rank plots for the four chains appear in Figure 2. The maximum of the two split-$\widehat{R}$ statistics is 1.56, a strong indication that more samples are needed. The rank plots confirm this, as the ranks in each chain are clearly not uniform, with strong differences observed across chains.

The rank plots indicate that chains 3 and 4 only explore the lower regions of the target distribution, chain 2 only explores the upper region of the distribution, and

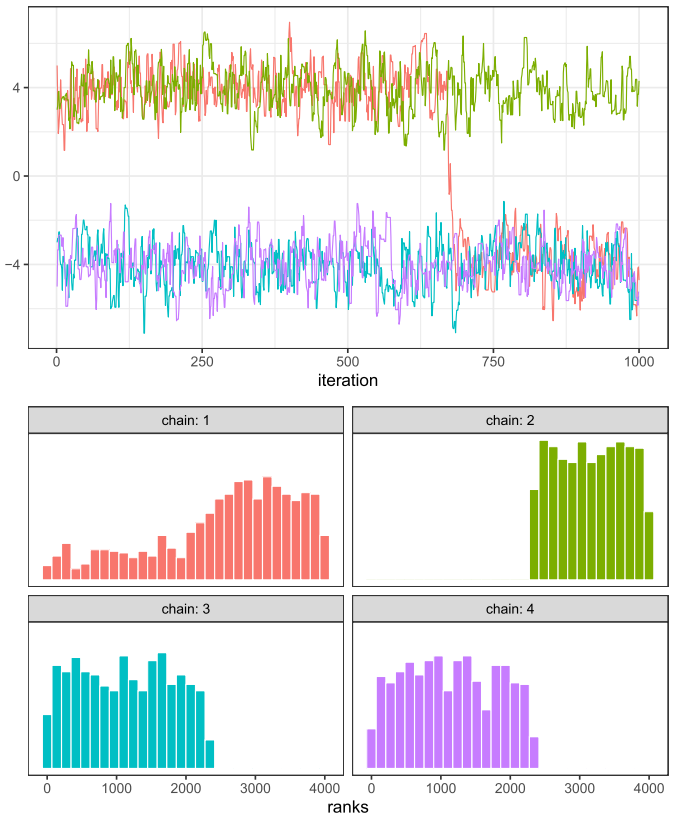

Figure 2: Plots of MCMC output for the example in Section 2.2. The target distribution is an equally-weighted mixture of a $\mathrm{N}(-4, 1)$ distribution and a $\mathrm{N}(4, 1)$ distribution.

chain 1 spends time exploring both regions, with more time spent in the upper region. The rank plots exhibit some of the same non-overlap features as did the rank plots in Example 2.1, but the reason for these features is different in this case. The chains mix well within a mode and so portions of the rank plots are fairly uniform, but the chains have difficulty jumping between modes and so the plots are not uniform across all the ranks. The cause of the difficulty is clear once one looks at the trace plot.

2.3 A note on trace plots

Vehtari et al. (2021) suggest that one might consider replacing the use of trace plots to diagnose MCMC convergence with the use of rank plots. Trace plots have been used to visualize MCMC output since the methods became popular in the early 1990s (though none appear in Gelfand and Smith, 1990). The authors provide several compelling arguments for why trace plots are not universally useful. They note that in high dimensional examples, examining large numbers of trace plots is not practical, and that trace plots

"tend to squeeze to a fuzzy mess when used with a long chain." Both are valid criticisms that illustrate problems with trace plots. The authors also provide an interesting example (summarized in Figure 10 in the paper) where MCMC is implemented based on two different parameterizations of the Cauchy distribution. The sampler converges well under one parameterization but not under the other; however, distinguishing good from bad convergence based on trace plots is not easy.

While these are excellent examples of situations where trace plots are not particularly useful, I would hesitate to completely replace my use of trace plots with rank plots. As seen in Examples 2.1 and 2.2 above, there are situations where rank plots indicate problems with convergence but trace plots provide extra information about what is causing the problems. I find it is useful to rely on a wide range of summaries when assessing convergence and look forward to using both types of plots in the future.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76(1): 1–32. 703

Cook, R. D. (1994). "On the interpretation of regression plots." *Journal of the American Statistical Association*, 89: 177–189. MR1266295. 705

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall. MR0675263. 705

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). "Programming with models: Writing statistical algorithms for general model structures with NIMBLE." *Journal of Computational and Graphical Statistics*, 26: 403–413. MR3640196. doi: https://doi.org/10.1080/10618600.2016.1172487. 703

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85: 398–409. MR1141740. 708

Gelman, A. (1996). "Inference and monitoring convergence." In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*, 131–143. Chapman & Hall/CRC. MR1397966. doi: https://doi.org/10.1007/978-1-4899-4485-6. 703

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, third edition. MR3235677. 704

Gelman, A. and Rubin, D. B. (1992a). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7: 457–511. 703

Gelman, A. and Rubin, D. B. (1992b). "A single sequence from the Gibbs sampler gives a false sense of security." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith,

A. F. M. (eds.), *Bayesian Statistics 4*, 625–631. Oxford: Oxford University Press. MR1723490. 703

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo (with discussion)." *Statistical Science*, 7: 473–483. 703

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, 28: 3049–3067. MR2750401. doi: https://doi.org/10.1002/sim.3680. 703

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility." *Statistics and Computing*, 10: 325–337. 703

Plummer, M. (2003). "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, 1–10. 703

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1221. 704, 705, 708

Contributed Discussion

Théo Moins[*], Julyan Arbel[†], Anne Dutfoy[‡], and Stéphane Girard[§]

We have highly appreciated this contribution to improve Markov chain Monte Carlo (MCMC) convergence diagnostics and would like to thank the authors for their insightful paper. In this note we discuss an adaptation of the quantile transformation introduced by the authors to the computation of some new $\hat{R}$ and analyse associated theoretical properties. More specifically, we propose to compute a *continuous version* $\hat{R}(x)$ for any level x based on indicator variables $\mathrm{I}(\theta^{(n,m)} \leq x)$, rather that on the parameter values $\theta^{(n,m)}$ themselves or their Gaussian version (4.1). This provides us with a function $\hat{R}(x)$ defining a local convergence diagnostic for any x. The rank-normalization step is circumvented since working on Bernoulli random variables $\mathrm{I}(\theta^{(n,m)} \leq x)$ ensures the existence of all moments whatever the $\theta^{(n,m)}$ distribution is. Assume that all elements $\theta^{(m,1)}, \ldots, \theta^{(m,N)}$ of a given chain m follow the same distribution F_m (without independence assumption) which may vary with m. Under this stationarity assumption, mean and variance of Bernoulli random variables $\mathrm{I}(\theta^{(n,m)} \leq x)$ can be easily written as functions of F_m, leading to an explicit expression for $R^2(x)$, the population version of $\hat{R}^2(x)$:

$$R^2(x) = 1 + \frac{1}{M} \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{M} \left(F_i(x) - F_j(x)\right)^2}{\sum_{i=1}^{M} F_i(x)(1 - F_i(x))}. \tag{1}$$

Clearly, $R^2(x) \geq 1$ for all $x \in \mathbb{R}$, with equality iff all F_m coincide. Moreover, $R^2(x) \to 1$ as $|x| \to \infty$. In order to condense the continuous index (1) into a scalar one, we may also consider its supremum over $\mathbb{R}$, denoted by R_∞, which is computed in practice at empirical quantiles. Hereafter, we illustrate how the problems of traditional $\hat{R}$ referred to as items 1. and 2. of Section 1.2 in the paper are bypassed by our proposal on two toy examples. See also Figure 1 for finite sample results.

Example 1 (Same mean and different variances)**.** Here $F_m(x) = F(x/\sigma_m)$ where σ_m is a scale parameter. As an example, we consider M chains uniformly distributed on $[-\sigma_m, \sigma_m]$ with $\sigma_m = \sigma \leq \sigma_M$ for all $m \in \{1, \ldots, M-1\}$ to model a lack of convergence. From (1), function R^2 has a maximum reached at $-\sigma/2$ and $\sigma/2$:

$$R^2_\infty := \sup_{x \in \mathbb{R}} R^2(x) = R^2(\pm\sigma/2) = 1 + \frac{M-1}{M}\left(1 - \frac{2}{1 + \frac{\sigma_M}{\sigma}}\right).$$

It appears that R^2_∞ is an increasing function of σ_M/σ with $R^2_\infty = 1$ iff $\sigma_M = \sigma$, and upper bounded by $2 - 1/M$.

Example 2 (Heavy tails and different locations)**.** Here $F_m(x) = F(x - \mu_m)$, with F a heavy-tailed distribution and μ_m a location parameter. We consider Pareto distributed

[*]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK 38000 Grenoble, France, theo.moins@inria.fr

[†]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK 38000 Grenoble, France

[‡]EDF R&D dept. Périclès 91120 Palaiseau, France

[§]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK 38000 Grenoble, France

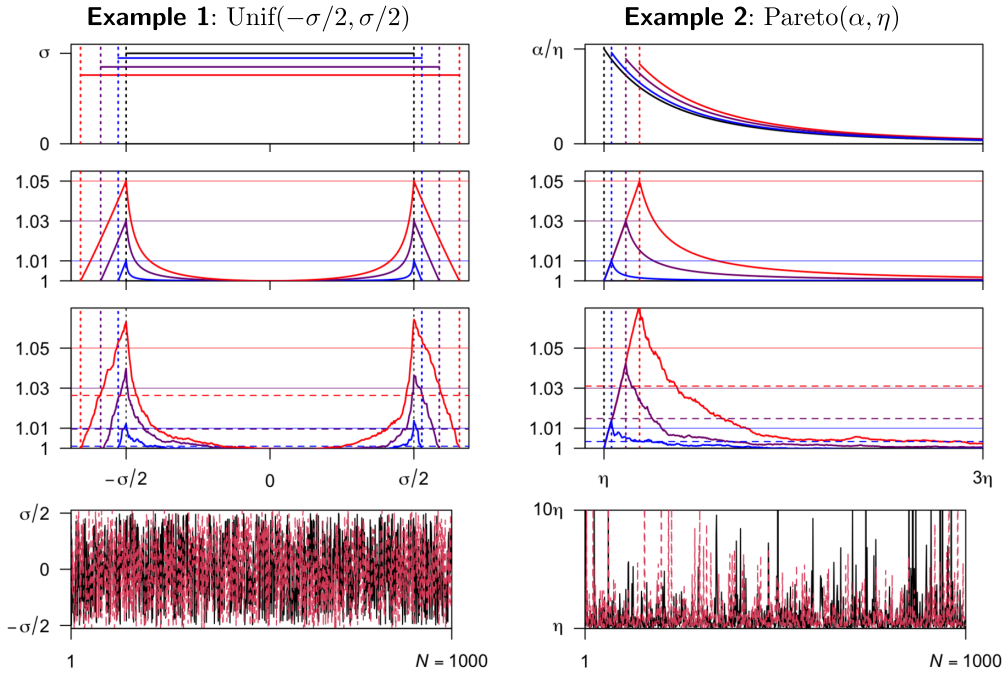

Figure 1: Illustrations with $M = 4$ chains, $N = 1000$ iterations each. Colors blue, violet and red resp. correspond to choices of 'diverging' distributions F_M such that our scalar diagnostic R_∞ matches with 1.01, 1.03 and 1.05. Top row: density of distributions F_m used, uniform (left) and Pareto (right). Colors for 'diverging' F_M and black for others. Second row: population (theoretical) version $R(x)$ of our index. Third row: empirical version $\hat{R}(x)$ of our index on simulated data; discussed rank-$\hat{R}$ shown as colored dashed lines. Bottom row: traceplots of one 'converging' chain ($m = 1$, black) and the 'diverging' chain ($m = M$, red case, $R_\infty = 1.05$).

chains with shape parameter $\alpha > 0$, and starting point $\eta > 0$ for $(M-1)$ chains and $\eta_M \geq \eta$ for the remaining one. In that case, R^2_∞ exists for any tail-index $\alpha > 0$:

$$R^2_\infty = R^2(\eta_M) = 1 + \frac{1}{M}\left(\left(\frac{\eta_M}{\eta}\right)^\alpha - 1\right).$$

Thus, R_∞ is an increasing function of η_M/η, with $R_\infty = 1$ iff $\eta_M = \eta$.

To conclude, it appears on these two examples that the proposed local version $\hat{R}(x)$ allows both to localize the convergence of the MCMC in different quantiles of the distribution, and at the same time to handle the problems not detected by classical $\hat{R}$ pointed out in the article. Compared to the rank-$\hat{R}$ proposed in the article, our scalar version $\hat{R}_\infty$ seems to be more conservative (see third row of Figure 1). Further investigation has to be done on a range of real world problems.

Rejoinder

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

We thank the discussants for their kind words and insights. In addition to these discussions, based on how our paper has been cited, it seems it would have been good to have included effective sample size (ESS) and Monte Carlo standard error (MCSE) also in the title as they are equally important as $\widehat{R}$ and the main role of $\widehat{R}$ is to handle information from multiple chains.

1 Graphical diagnostics

We agree with Hans, who wrote, "trace plots provide extra information about what is causing the problems." When proposing to replace trace plots with rank plots, we were intentionally a bit provocative. Indeed, trace plots can be useful to illustrate slow mixing or the long-lasting effect of initial values, and thus they have their role in the convergence diagnostic workflow as one potential way to find causes for problems that arise; see, e.g., Figure 32 in (Gelman et al., 2020). However, we do not think that trace plots should be the first thing inspected after running Markov chain Monte Carlo (MCMC). Rather, we recommend starting with $\widehat{R}$, ESS, and other diagnostics and then, if there seem to be problems, using graphical tools such as rank plots and trace plots to understand what went wrong. The area of graphical diagnostics for MCMC convergence is under-researched. We think it is important to value such diagnostics as investigative tools rather than treating them as ways of making yes/no recommendations.

Rank histogram plots can allow visual detection of smaller differences than can be seen in trace plots, but they also have shortcomings. The interpretation of histograms is also familiar for many users, so we intentionally focused on the visual aspect of the rank plots. After the current paper had been accepted for publication, Säilynoja et al. (2021) have developed a graphical test for testing uniformity of ranks, and they illustrate its use for comparing ranks from multiple chains. We recommend this approach as it allows us to assess whether the small differences between the chains are just due to chance or if there is a deeper problem. The approach is based on examining the empirical cumulative density function (ECDF) or its difference from the uniform baseline with a computationally efficient way to compute simultaneous confidence bands. The benefits compared to the histogram rank plots are that the ECDF doesn't require a choice of bin size and the simultaneous confidence band takes into account the dependence between empirical cumulative density values. Figure 1 shows the benefit of ECDF rank plots

*Department of Computer Science, Aalto University, Finland, Aki.Vehtari@aalto.fi
†Department of Statistics, Columbia University, New York
‡Department of Econometrics and Business Statistics, Monash University, Australia
§Center for Computational Mathematics, Flatiron Institute, New York
¶Cluster of Excellence SimTech, University of Stuttgart, Germany

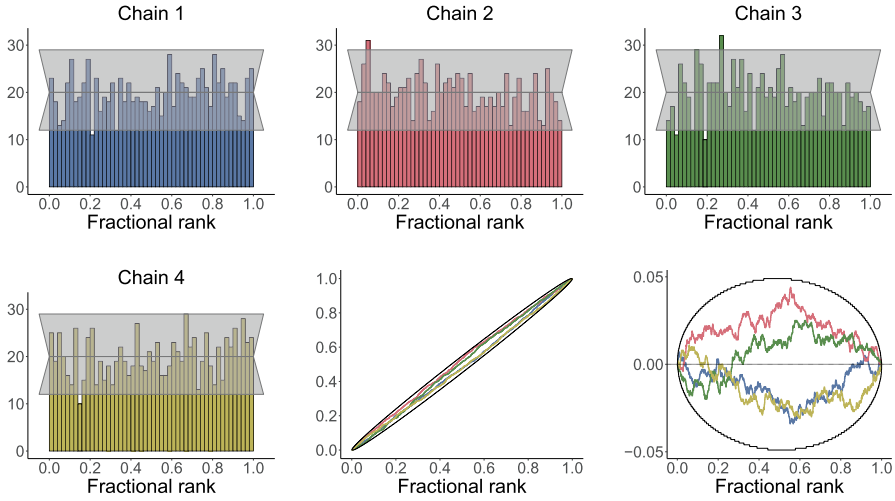

Figure 1: Histogram and ECDF rank plots. When inspecting the sampling of parameter τ, even when the 96% confidence bands of the histogram rank plots of chains 2 and 3 are exceeded by one bin each, the ECDF plot and the ECDF difference plot of the non-centered parameterization eight schools model indicate no mixing issues as the ECDF of the fractional ranks of each chain stay between the 96% simultaneous confidence bands.

over the histogram plots when analysing the convergence for the eight schools model with non-centered parameterization.

2 Empirical diagnostics do not prove convergence

We agree with Vats and Jones, who wrote, "Diagnostics based on the simulated values cannot prove that the simulation is providing representative samples, and the best we can hope for is that it indicates when a problem has occurred." The same holds also for multiple chain diagnostics. Even coupling methods that produce perfect draws can be fooled by a local mode that is well separated from others. On the other hand, we can rarely find useful qualitative guarantees for complex model classes, and using these techniques would be far beyond the ability of even advanced MCMC users. Furthermore, these qualitative bounds frequently provide insufficient information about the actual performance of the Markov chain. Hence, rather than trying to prove convergence, we use numerical and graphical diagnostics as part of our workflow to catch problems of poor mixing as quickly as possible.

3 Interpretation of $\widehat{R}$

Vats and Jones ask about the interpretation of the new version of $\widehat{R}$. Originally and inherently, the interpretation of $\widehat{R}$ is as the ratio of the scale of all the chains mixed and

the scale of each chain alone, and thus $\widehat{R}$ is inherently a multi-chain diagnostic. The original and many later versions used standard deviation as a measure of scale, whereas the rank-normalized version uses a non-parametric measure of variation, which has the benefit of working when the variance of the target is infinite while still coinciding with the variance-based $\widehat{R}$ when the target is normal.

In addition to using the rank-normalized $\widehat{R}$ as a more robust multi-chain diagnostic, when computing ESS and MCSE, we are still using the total and within variances to appropriately combine autocorrelations from several chains, which makes sense as all ESS, MCSE and autocorrelations are defined in terms of variances that thus need to be finite. We consider variance-based and rank-normalized $\widehat{R}$s as quick multi-chain diagnostics, ESS as a scale free multi-chain efficiency measure, and MCSE as interpretable with respect to domain expertise.

As we mention in the paper, we recommend MCSE and the corresponding notion of effective sample size as the diagnostics that allow interpretation in the context of the domain knowledge about the required estimation accuracy. However, we don't recommend trusting MCSE if there are other signs of bad sampling behavior, because MCSE estimates can be optimistic in these cases. If more reliability is needed, we recommend checking whether the central limit theorem is kicking in as follows: double the number of iterations and check that MCSE goes down by a factor of around $\sqrt{2}$ (correspondingly ESS goes up by a factor of 2). If that doesn't happen, there can be convergence issues or the distribution may have infinite variance which can be diagnosed, for example, with the Pareto k diagnostic (Vehtari et al., 2019). If the proposed diagnostics that work with infinite variance do not show any clear convergence issues, but the distribution for the quantity of interest has infinite variance we recommend estimating, for example, the median instead of the mean, and then using the MCSE estimate provided in the Section 4.4 of our paper.

4 Local diagnostics

The proposal by Moins, Arbel, Dutfoy, and Girard is interesting and closely related to our local efficiency plots which are also based on indicator variables. Our local ESS values use the basic $\widehat{R}$ to incorporate multichain information, and use autocorrelation information to handle within-chain correlation. They can thus can be used similarly to the estimate by Moins et al., with the additional benefit of using within-chain information more efficiently.

As the example by Moins et al. has distributions that have different supports, it seemed strange that $\widehat{R}$ was decreasing in the region of no-overlap. Investigating this helped us to recognize an unwanted behavior that we had in our diagnostic code, which we have now fixed. Following their example, we generated four "chains" of length 1000, so that three of them have independent uniform draws from $[-1, 1]$ and one of them has uniform draws from $[-1.25, -1.25]$. Figure 2 shows $\widehat{R}$, ESS, and MCSE for indicator functions $I(\theta \leq x)$, where $x \in [-1.25, 1.25]$. The $\widehat{R}$ result is very similar to the result by Moins et al. Here also we can see the easier interpretability of ESS as it is easy to

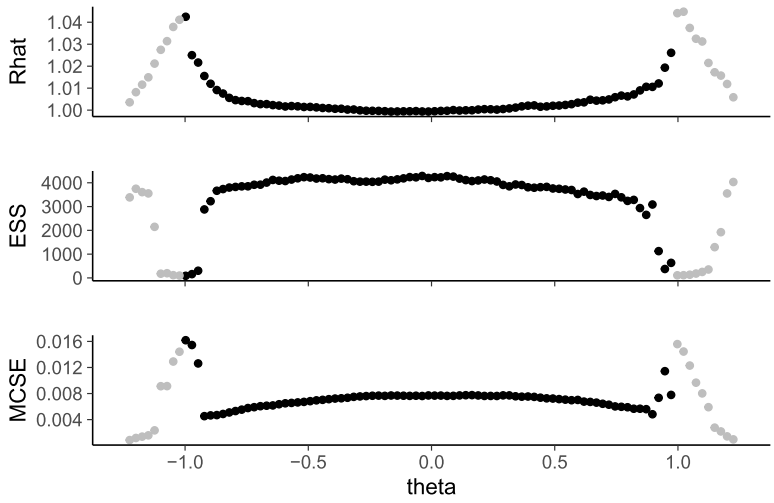

Figure 2: Illustrations with $M = 4$ chains, $N = 1000$ iterations each. Three first chains have independent uniform draws from $[-1, 1]$ and the fourth has uniform draws from $[-1.25, -1.25]$. We compute $\widehat{R}$, ESS, and MCSE for indicator functions $I(\theta \leq x)$, where $x \in [-1.25, 1.25]$. Black dots show values when the indicator function is non-constant for all chains, and grey dots show values that were computed in cases where at least one chain had constant value.

understand that an ESS very close to 0 is not good. Looking at the example, we realized that we were computing the diagnostic values also in the cases where one or more chains had constant value (shown as gray in the figure). For chains that remain constant, we can't distinguish between situations where the chain is not mixing well and where the chain is mixing appropriately, but the alternatives have small probabilities relative to the length of the chain. In such cases, we think we should err on the side of caution and report that the estimation is not reliable. This will avoid providing overly optimistic MCSE values, as shown in the bottom plot of Figure 2.

Figure 1 of Moins et al. is an excellent example of how difficult it is to use trace plots as convergence diagnostics. Rank histograms and ECDFs can much more clearly spot that the supports are not overlapping as shown in Figure 3.

Moins et al. plot their estimate with the parameter values on the x-axis. In the main paper, we plotted quantiles on x-axis. For example, with the Cauchy distribution, values of θ ranged from -5000 to 500, but most of the detail is in the range between -10 and 10, so the plot would have been hard to read. For the 8-schools example, we show local and quantile ESS with τ on x-axis in the online appendix (due to the page limit these plots were left out from the main article). For convenience, we reproduce one example in Figure 4.

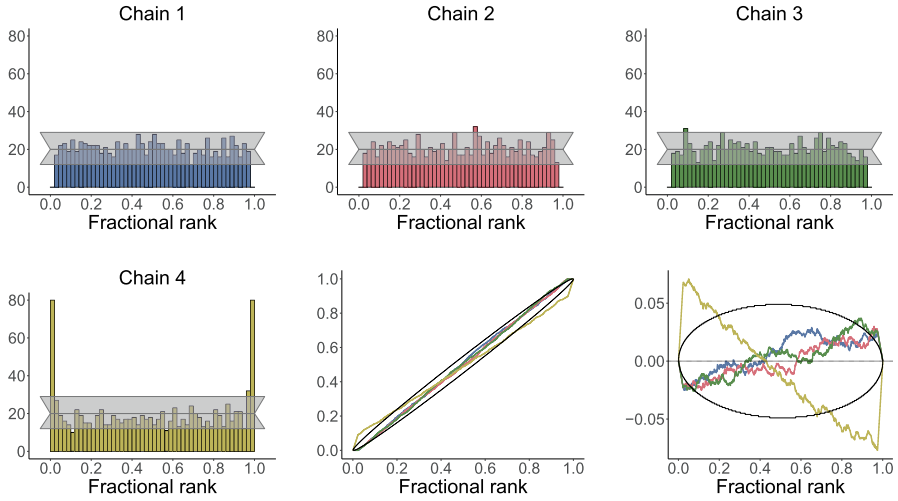

Figure 3: Histogram and ECDF rank plots for the uniform example by Moins et al. Compare these to the trace plots in Figure 1 of Moins et al.

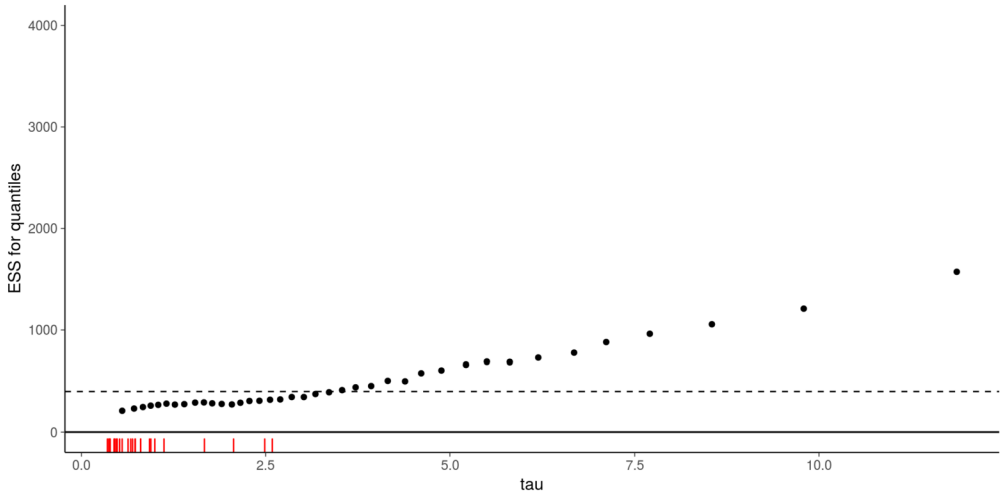

Figure 4: Sometimes using the parameter value on x-axis of the local ESS plot can provide useful additional information. Here τ is population prior scale parameter and we can see there are sampling issues when the scale is close to 0.

5 Using these tools in statistical workflow

In the paper we focused on using $\widehat{R}$, ESS, and MCSE to summarize mixing and inference for scalar quantities of interest one at a time. There is also a multivariate version of

$\widehat{R}$ (Brooks and Gelman, 1998), and more recently Lambert and Vehtari (2021) have proposed R^*, a multivariate measure of mixing that uses nonparametric classification trees. We anticipate that future workflow will involve using R^* and $\widehat{R}$ as screening tools to flag poor mixing that can then be studied more carefully with graphical diagnostics.

$\widehat{R}$ and related tools can also be used for convergence diagnostics of stochastic optimization (e.g., in variational inference) with multiple parallel optimizations or with split-$\widehat{R}$ for a single optimization (Dhaka et al., 2020). Again, this fits into workflow in two ways, first by catching many problems early and second by building some trust in results when approximate mixing has been achieved. But no diagnostic can catch everything, so we emphasize the importance of fitting multiple models, as well as simulated-data experimentation, as a way of understanding the domain of applicability of any fitting procedure. We like $\widehat{R}$, ESS, and MCSE not because they are perfect, but because they detect many problems and fit into this larger workflow.

References

Brooks, S. P. B. and Gelman, A. G. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7(4): 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 718

Dhaka, A. K., Catalina, A., Andersen, M. R., Magnusson, M., Huggins, J., and Vehtari, A. (2020). "Robust, Accurate Stochastic Optimization for Variational Inference." In *Advances in Neural Information Processing Systems*, volume 33, 10961–10973. 718

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). "Bayesian workflow." *arXiv preprint* arXiv:2011.01808. 713

Lambert, B. and Vehtari, A. (2021). "R*: A Robust MCMC Convergence Diagnostic with Uncertainty Using Decision Tree Classifiers." *Bayesian Analysis*, 1(1): 1–27. 718

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2021). "Graphical Test for Discrete Uniformity and its Applications in Goodness of Fit Evaluation and Multiple Sample Comparison." *arXiv preprint* arXiv:2103.10522. 713

Vehtari, A., Simpson, D., Gelman, A., Yuling, Y., and Gabry, J. (2019). "Pareto smoothed importance sampling." *arXiv preprint* arXiv:1507.02646. 715